# NORM SHARE PREDICTS WHETHER A SUBSPACE ABLATION REMOVES THE FEATURE OR THE MODEL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Feature ablation is used to isolate what an internal representation contributes to a language model. We show that two operations commonly called ablation produce widely different effects on the same selected features, ranging from language-specific control to broad model failure. For sparse-autoencoder features, latent zeroing subtracts the selected latents' contributions and preserves reconstruction error while directional projection removes every residual-stream component along their decoder directions at every token. Because a language-specific feature's decoder span can carry shared computation, projection can cause broad model failure. We predict this failure using norm share, $\rho$, the average fraction of a token's squared residual-stream norm in the ablated subspace. This requires one forward pass without labels, generation or intervention. All $52$ measured subspaces with $\rho$ above one half raise the other languages' cross-entropy by at least $6$ nats. In language-identity tests across eleven languages of parallel text, Gemma 3's mid-network subspaces carry most of a typical token's norm. At layer 17, projection raises target cross-entropy by 7–31 nats (up to several times baseline), with similarly large increases on the other ten languages. Zeroing the same latents concentrates changes on the target and switches output language, with no switches under matched random-feature controls. The same subspace construction at the Llama 3.1 site has low $\rho$ and selective projection, consistent with the decline in projection selectivity as $\rho$ rises across measured subspaces in both families. Manipulating the dominant shared directions changes this outcome: removing them from Gemma's basis makes projection selective, and adding them to Llama's basis lowers selectivity. Adding a random direction of the same rank leaves projection selectivity unchanged. The diagnostic also separates destructive from harmless refusal interventions. We then use selective language removal to test whether language-specific representation interferes with multilingual reasoning. Across three model families and four removal constructions, no setting significantly raises pooled MGSM accuracy. The gains reported by Zhao et al. (2025) are per-language best points of a grid searched on the evaluated problems. Applying this rule to our reproduction, where no setting raises pooled accuracy, gives $+2.1$ points in-sample and $0.0$ when selection and scoring use different halves of the problems. Subspace interventions should therefore report $\rho$ and collateral effects before attributing behavioural changes to removal of the targeted property.

## 1 INTRODUCTION

Removing a representation is a common way to test its contribution to a language model's behaviour. Interpretation depends on collateral effects: a language-selective ablation of Chinese features should affect Chinese more than unrelated languages. Disruption across languages prevents attributing the outcome to removing Chinese. A growing family of interventions projects refusal, style, concept and language subspaces out of the residual stream. For multilingual reasoning (Zhao et al., 2025), the premise is that language-specific representation competes with reasoning in the shared stream; we call this the *interference account*. A parallel literature reports clean, property-specific effects by zeroing sparse-autoencoder latents and preserving reconstruction error (Deng et al., 2025; Chou et al., 2025; Andrylie et al., 2025).

Both operations are called "ablation". However, projection removes every component along the



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

selected decoder directions at every token, while latent zeroing removes each latent's contribution only where it fires. These directions can also carry other features' contributions and a shared offset, so projection can remove far more than the selected latents contribute.

We measure this overlap with $\rho$, the fraction of a typical token's squared residual norm that projection removes on the evaluation distribution. It takes one forward pass, no labels and no generation. At large $\rho$, projection removes the model along with the feature: cross-entropy rises by tens of nats on every language, making downstream effects unattributable to the selected property. A small $\rho$ is necessary for a selective projection (§4).

Large models carry a few directions of massive activation shared by every token (Sun et al., 2024). At Gemma 3 layer 17 the dominant direction lies along the SAE decoder bias; zeroing such a direction is known to break a model. Neither literature measures how much of it an ablation subspace carries, yet the two draw opposite conclusions from "ablation". Mean-centring the projection or restoring each token's norm does not make projection selective (Appendix B.11). The unreported quantity $\rho$ predicts the outcome.

We establish this on language identity, using the other languages in a parallel corpus as controls. The two model families span Gemma 3's mid-network ($\rho = 0.72$–$0.94$) and Llama 3.1's layer 19 ($\rho = 0.05$–$0.12$). Every disagreement between the two literatures that we measure follows from this difference. Changing $\rho$ within one model, with features held fixed, changes projection selectivity. A removal is *selective* to the degree that it changes the target language more than the others; §2 defines the index.

The paper makes five contributions.

1. Across $53$ subspaces, projection selectivity falls with the one-pass diagnostic $\rho$ (§4). Changing $\rho$ within a model at fixed features changes selectivity; a random direction of the same rank has no effect.
2. $\rho$ predicts the collateral of refusal-direction removal (Arditi et al., 2024) (Spearman $+0.93$, twenty directions, two models). On Gemma 3 the deflated layer-17 direction cuts refusal from $75\%$ to $18\%$ at its own layer for $0.01$ nats or less (§4.1).
3. At Gemma 3 layer 17, identical latents give $19.5\times$ language selectivity under zeroing and $1.05\times$ under projection (§5). The gap survives ten dictionaries and six selection criteria.
4. Removing a language's latents switches output language for six of eleven languages, establishing their causal role (§6). No random-feature control switches any.
5. Across three model families, four constructions and two position policies, no removal setting significantly raises pooled MGSM accuracy (§7). On our reproduction of the grid, the per-language best-point rule behind the numbers Zhao et al. (2025) report gives $+2.1$pp in-sample and $0.0$ on held-out problems.

## 2 SETUP

**Models and dictionaries.** We use Gemma 3 4B IT (34 layers, $d_{\text{model}} = 2560$) with Gemma Scope 2 JumpReLU residual-stream SAEs (McDougall et al., 2025) at layers $\{9, 17, 22, 29\}$. Widths $\{16\text{k}, 65\text{k}, 262\text{k}\}$ crossed with the release's three sparsity targets give nine dictionaries at layer 17, ten with a seed replicate. We also use Gemma 3 12B IT (layer 24), Llama 3.1 8B Instruct with Goodfire's ReLU SAE (layer 19), and two models the multilingual-reasoning method covers: Qwen-2.5-7B-Instruct ($\rho$, §4.1) and Qwen-2.5-3B-Instruct ($\rho$ and accuracy, §7).

**Feature selection.** We select the top-$k$ language features by intersecting the monolinguality criterion $\nu$ (Deng et al., 2025) with a supervised probe, computed on last-token residuals over $250$ MGSM prompts per language. On Llama we use $\nu$ on FLORES sentences disjoint from those scored (Appendix E). Selection uses no answers, model outputs or accuracy numbers. The tests of §5–§6 score unseen text. The accuracy test of §7 scores MGSM problems whose prompts entered selection; Appendix B.8 repeats it with FLORES-selected features. We use $k = 20$ for the primary SAE experiments to remove a whole subspace at a small enough scale that single-latent instability (Appendix C) does not dominate.

**Corpora.** We score cross-entropy on FLORES-200 devtest (NLLB Team, 2022), using 200 parallel sentences per language. We score reasoning accuracy on MGSM (Shi et al., 2023), using 250 problems per language, greedy decoding and 384 new tokens. Numeric answer extraction is blind to the output language (Appendix B.9). Both corpora are parallel across the eleven MGSM languages {en, es, fr, de, ru, zh, ja, th, sw, bn, te}, so every comparison is paired at the item level.

**Selectivity index.** For a target language, $\Delta_T$ is the change in its cross-entropy under the removal and $\Delta_O$ the mean change over the other languages. The concentration index $C = (|\Delta_T| - |\Delta_O|)/(|\Delta_T| + |\Delta_O|) \in [-1, +1]$ indicates no mean collateral at $+1$, equal magnitudes of target and mean collateral effects at 0, and greater mean collateral magnitude below 0. A *cell* is one (model, site, $k$, target language) combination with its own subspace and $\rho$; $\bar{C}$ is the mean over a table's cells. A cell is *inert* if removal changes target cross-entropy by under 0.1 nats. Since $C$ is undefined for no effect, inert cells enter no mean, count or correlation. Every inert cell in this paper also leaves the other languages' mean within 0.1 nats, so the rule excludes no cell with collateral.

**Random-feature controls.** To test whether an effect depends on language-based selection, we compare each selected set with randomly sampled sets of equally many active SAE latents (latents that fire at least once on the selection inputs), matched on layer, intervention positions, dictionary and number of latents, with every selected latent excluded from the pool. Each random set is evaluated on every language, so one set is a control for each target; each experiment states its number of sets.

## 3 Two ways to remove a feature

Let $h \in \mathbb{R}^d$ be the residual stream at the site an SAE reads, and let $S$ be a set of $k$ selected latents with decoder rows $d_j$ and JumpReLU activations $a_j(h)$. Two operations are both described in the literature as "ablating" those features.

**Directional projection.** Orthonormalise $\{d_j\}_{j \in S}$ into $Q \in \mathbb{R}^{k \times d}$ with $QQ^\top = I_k$. Then $Q^\top Q$ is the rank-$k$ orthogonal projector onto their span. Subtract its output:

$$h \longleftarrow h - Q^\top Q\, h. \tag{1}$$

This deletes every component along those directions at every token, regardless of latent activity. Arditi et al. (2024) apply this map as written. Zhao et al. (2025) subtract a fraction $\lambda \leq 0.4$ of the component, $h - \lambda Q^\top Q h$, and add it back at later layers. Equation (1) sets $\lambda = 1$; §7 runs their pipeline at their strengths.

**SAE-native removal.** Subtract only what the latents themselves contribute:

$$h \longleftarrow h - \sum_{j \in S} a_j(h)\, d_j. \tag{2}$$

Since $h = \sum_j a_j(h) d_j + b_{\text{dec}} + \varepsilon(h)$, this zeroes the selected latents while preserving the decoder bias and reconstruction error, as in feature-level interpretability (Deng et al., 2025).

Equation (1) applies a fixed linear map at every token. The removal in equation (2) vanishes wherever the latents do not fire. Latents that fire only on their own language therefore leave other languages unchanged by construction. The importance of this difference at a site is empirical; §4 reduces it to one measurable quantity.

## 4 Norm share predicts projection selectivity

Projection fails to isolate language at Gemma's mid-network sites (§5) and is nearly as selective as latent zeroing on Llama (§B.3). Norm share predicts both outcomes before an intervention is run.

**Norm share.** For an ablation subspace with orthonormal basis $Q \in \mathbb{R}^{k \times d}$ and residual activations $h_t$ over the evaluation distribution, norm share is the share of a typical token's squared norm that projection would remove, estimated in one forward pass over a few thousand tokens:

$$\rho \;=\; \mathbb{E}_t \left[ \frac{\|Q h_t\|^2}{\|h_t\|^2} \right]. \tag{3}$$

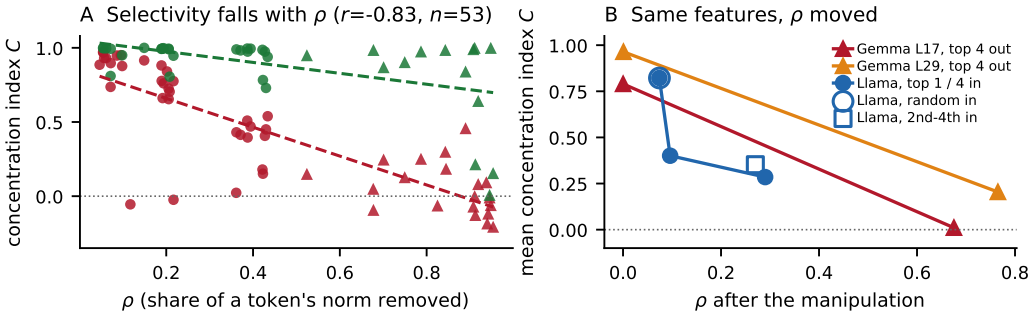



**Figure 1: $\rho$ predicts whether a projection isolates a property. (A)** Every measured cell, projection in red and SAE-native removal in green (circles Llama, triangles Gemma), with least-squares fits; SAE-native removal is shown on the 39 cells it affects, and $r$ is Pearson's for projection. **(B)** The within-model test of §4.1: the same $k$=20 features with $\rho$ moved down (Gemma) or up (Llama), means over the languages the manipulated basis affects; open markers add a random direction (circle) or the second to fourth directions alone (square).

**Evidence across subspaces.** At Gemma 3 4B layers 9–29 the selected $k$=20 subspaces have $\rho = 0.72$–$0.94$, averaged over languages at each layer (0.31–0.96 per language). Projection removes most of the representation and raises cross-entropy 7–31 nats on every language, without selectivity. At Llama 3.1 8B layer 19 the same construction has $\rho = 0.046$–$0.12$: projection removes the feature, a small part of each token. We estimate each $\rho$ on FLORES tokens; independent samples agree to about $\pm 0.03$.

Within Llama, increasing $k$ raises $\rho$ and lowers projection selectivity while SAE-native removal remains strongly selective. Model, site and selection criterion are fixed, with features selected on disjoint FLORES sentences (Appendix E; full sweep in Appendix B.13).

The 53 distinct selected subspaces in Table 11 span $\rho = 0.05$–$0.95$. Directional $C$ falls with $\rho$ at Spearman $-0.845$ ($p = 1.8 \times 10^{-15}$). The cells share geometry within five (model, layer) sites (four Gemma, one Llama) and are not independent. Resampling sites gives a 95% interval of $[-0.87, -0.28]$. Within families, Spearman is $-0.79$ over 33 Llama cells and $-0.51$ over 20 Gemma cells (Appendix B.13). Within a (site, $k$) stratum, $\rho$ varies little and the ordering is weak.

Across all 102 measured projection subspaces, the pooled evidence primarily establishes a sharp regime split: all 52 with $\rho$ above one half cost the other languages at least 6.0 nats (median 16.6) and reach at most $C = +0.46$. For $\rho < 0.15$, median $C$ is 0.8. These subspaces include the eleven-language selections, the rank-1 mean-difference directions of Table 18 and Gemma 3 12B. For the 12B model we predicted $\rho$ above one half before measuring it (0.73–0.96). Section 4.1 tests the relation by manipulating individual subspaces; Appendix B.13 gives further summary statistics.

SAE-native selectivity also falls with $\rho$, at a third of projection's slope (Appendix B.13). It subtracts only the latents' contributions and preserves the rest of their span.

**Scope of the diagnostic.** $\rho$ is a diagnostic measured on two model families and one property in 53 cells, plus refusal in twenty subspaces, where the destructive regime begins well below $\rho = 0.5$ (§4.1). Despite the sharp regime split, we claim no universal constant; the SAE-native slope shows the boundary is soft. $\rho$ predicts how damage is distributed, with its magnitude measured separately.

A low $\rho$ is required and not sufficient: on Llama, adding the top direction costs more selectivity at $\rho = 0.10$ ($C = +0.40$) than the $k$=100 subspace at $\rho = 0.20$ ($+0.70$). That direction carries 0.95 of the second moment on a few massive-activation tokens and 0.02 of a typical one. Single mean-difference directions with $\rho < 0.03$ still cost up to 22 nats on other languages (Table 18).

### 4.1 Shared directions and tests within each model

**Shared activation directions.** At Gemma layer 17, the top principal component of the token second-moment matrix carries 0.997 of the total squared norm. A typical token has $\rho = 0.967$ along this "massive activation" direction (Sun et al., 2024), which aligns with the SAE decoder bias ($|\cos| = 0.998$). Almost any 20 decoder rows overlap it. The eight random-feature controls of §5

carry $\rho = 0.57$ and fifty further random draws $0.60$, compared with $0.72$ for the eleven selected sets (Mann–Whitney $p = 0.09$). High $\rho$ across selected and random subspaces provides an early warning about the site. Each intervention still requires measuring $\rho$ on its actual subspace; the selection criterion cannot substitute for it.

For every non-English target, selected latents contribute under $0.1\%$ of the norm projection removes. The removed norm is dominated by the decoder bias and other latents' components, which overlap along the shared direction (Table 14).

**Changing the subspace.** We test the effect of changing $\rho$ with model, site and selected features held fixed. On Gemma, we orthogonalise the selected span against the top four principal directions of the second-moment matrix. This preserves rank 20 and lowers $\rho$ to $0.00$ from $0.68$ on the languages compared here. Projection becomes selective: mean $C$ over languages affected by at least $0.1$ nats rises from $+0.01$ to $+0.79$ at layer 17 and from $+0.20$ to $+0.96$ at layer 29. Each pair uses the same languages, with $C$ higher in all eight and all nine (Wilcoxon $p = 0.008$ and $0.004$).

Effect size falls as selectivity rises. At layer 17, the target effect falls from 21 nats to $0.7$ and collateral from 20 nats to zero (Table 13). This is the trade-off in §B.2: directions carrying enough of the token to move cross-entropy are shared across languages. The repaired projection still removes the property. It changes output language in five of the six languages affected by latent zeroing; no control changes it by more than $3.1$ points (Table 22).

On Llama, adding the site's top directions to the same $k$=20 features (rank 21 or 24) raises $\rho$. Mean $C$ falls from $+0.82$ to $+0.40$ with one direction and $+0.29$ with four, falling in all eleven languages (Wilcoxon $p = 0.001$). Adding a random direction of the same rank leaves $C$ at $+0.82$. Adding only the second to fourth directions also lowers selectivity, with almost no change in the fraction of the first token's squared norm removed. Across these manipulated subspaces, selectivity follows $\rho$; neither rank nor the massive-activation direction alone explains the changes (Appendix B.13).

**Constructions without an SAE.** The subspace method also permits a test without an SAE. It uses centred per-language means (rank $L-1$ by SVD; Zhao et al., 2025, Algorithm 1). Rebuilding it at each site gives $\rho = 0.95$ at Gemma layer 17, costing $9.6$ nats on every language, and $0.06$ at Llama layer 19, costing $0.10$. At the method's targeted mid-network layers, $\rho$ is $0.04$–$0.07$ on Qwen-2.5-7B and $0.04$–$0.10$ on Qwen-2.5-3B, used for accuracy in §7; Appendix B.13 gives the cross-entropy costs. Thus $\rho$ identifies where this construction destroys the model and where it mildly perturbs it.

LEACE (Belrose et al., 2023) erases a concept's linear trace with an oblique projection weighted by the activation covariance, which spares the dominant shared directions by construction. Fitted for each one-vs-rest language concept on disjoint FLORES text, it is selective in all twenty-two cells of the two sites (mean $C = +0.99$ at each, none inert). Target cross-entropy rises $0.4$–$4.6$ nats, with no other language moving by more than $0.08$. The edit removes little of a typical token's squared norm. Orthogonal projection of the same mean-difference directions costs 9–37 nats on other languages at Gemma (Table 18). Where $\rho$ is high the remedy is an operator that spares the shared directions: latent zeroing, a deflated basis, or LEACE.

**Refusal directions.** Refusal interventions test the relation on another property. Arditi et al. (2024) remove refusal with the difference of mean activations between harmful and harmless instructions, extracted at one layer and projected out at every layer. Repeating the recipe at ten extraction layers per model gives twenty single-direction subspaces spanning $\rho = 0.000$–$0.929$. $\rho$ at the extraction layer predicts the collateral cross-entropy of the published all-layer protocol at Spearman $+0.928$ ($p = 3.9 \times 10^{-9}$; Spearman $+0.709$ within Gemma alone) and that of projecting the direction out at its own layer only at $+0.835$ ($p = 4.8 \times 10^{-6}$; $+0.35$ within Gemma, $p = 0.33$).

Most of this is a regime split: every direction with $\rho \geq 0.17$ costs at least $4.3$ nats at its own layer and every direction with $\rho < 0.04$ under $0.6$ (Table 12).

Every Llama 3.1 8B direction carries $\rho \leq 0.03$, and its layer-8 direction cuts refusal from $88\%$ to $18\%$ at $0.02$ nats with harmless prompts unchanged, reproducing the published result. Directions extracted at layers 12–26 leave refusal at or above baseline ($96$–$100\%$). Thus $\rho$ diagnoses collateral; a direction's ability to remove refusal requires a separate test.

On Gemma 3 4B, directions from layers 4 to 24 carry $\rho = 0.17$–$0.93$ and cost 20–40 nats under the

| | Directional projection | | | SAE-native removal | | | |
|---|---|---|---|---|---|---|---|
| Lang | $\Delta_T$ | $\Delta_O$ | $C$ | $\Delta_T$ | $\Delta_O$ | ratio | $C$ |
| en | 7.53 | 10.14 | $-0.148$ | 0.711 | 0.515 | $1.4\times$ | $+0.159$ |
| es | 16.95 | 15.06 | $+0.059$ | $-0.011$ | 0.000 | | inert |
| fr | 25.97 | 22.65 | $+0.068$ | 3.277 | 0.001 | $\gg 100\times$ | $+1.000$ |
| de | 23.63 | 21.50 | $+0.047$ | $-0.016$ | $-0.002$ | | inert |
| ru | 27.26 | 25.98 | $+0.024$ | $-0.010$ | 0.002 | | inert |
| zh | 26.47 | 20.15 | $+0.135$ | 1.909 | 0.057 | $33.4\times$ | $+0.942$ |
| ja | 18.75 | 19.20 | $-0.012$ | 3.032 | 0.001 | $\gg 100\times$ | $+0.999$ |
| th | 23.76 | 20.84 | $+0.065$ | 1.432 | $-0.018$ | | $+0.975$ |
| sw | 10.13 | 13.14 | $-0.129$ | 1.126 | 0.017 | $66.1\times$ | $+0.970$ |
| bn | 10.97 | 14.29 | $-0.131$ | 0.560 | $-0.010$ | | $+0.966$ |
| te | 30.82 | 28.50 | $+0.039$ | 0.385 | 0.073 | $5.3\times$ | $+0.682$ |
| mean, all eleven | 20.20 | 19.22 | $+0.002$ | 1.127 | 0.058 | $\mathbf{19.5}\times$ | |
| mean, the eight affected | 19.30 | 18.62 | $\mathbf{-0.014}$ | 1.554 | 0.080 | $19.5\times$ | $\mathbf{+0.837}$ |

**Table 1:** Cross-entropy change in nats on FLORES-200, layer 17, $k$=20, all eleven MGSM languages; baseline 4.3–6.4 nats. Both halves apply *identical* feature sets at an *identical* site; inert marks $|\Delta_T| < 0.1$ nats. In the affected-language row, both $C$ means use the same eight languages. Per-language ratios are $\Delta_T/\Delta_O$, omitted where $\Delta_O$ is negative; summary ratios divide the corresponding column means (Deng et al. (2025) report 5–10×).

all-layer protocol. Refusal falls to zero because the model no longer produces text. The layer-29 direction ($\rho = 0.003$) costs 0.6 nats at its own layer and 11 when projected out at every layer.

Removing each direction's overlap with the top four second-moment directions applies the language repair and brings the cost at its own layer to at most 0.01 nats at every layer tested. Refusal at layers 4 to 14 then stays near baseline (68–78%), identifying the original effect as collateral. At layers 17 and 20 it falls to 18% and 45% against 75% in the same run, with harmless prompts unchanged. The original directions cost 19 and 11 nats at their own layers. Deflating at every layer brings the layer-29 direction's all-layer cost to 0.1 nats (Table 21). An intervention applied at many sites needs $\rho$ at each.

## 5 Selectivity under the two operations

### 5.1 Measurement

For a target language $L$ we ablate its selected latents, score cross-entropy on all eleven languages of FLORES-200 devtest with item-level pairing, and report signed $\Delta_T$, $\Delta_O$ and the concentration index $C$ (§2). Averaging signed changes could hide collateral through cancellation. We therefore recompute $C$ using the mean of $|\Delta_\ell|$ over the other ten languages in place of $|\Delta_O|$. No $C$ in Tables 1 or 9 changes by more than 0.04. SAE-native removal changes any single other language by at most 0.8 nats (English's features; 0.4 for any other target). Projection changes every language by tens of nats. Thresholds of 0.05 and 0.2 nats identify the same inert cells. We report $\Delta_T/\Delta_O$ for comparison with Deng et al. (2025); it has a pole at $\Delta_O = 0$, and we never use it for inference. Each site has eight random-feature controls (§2).

### 5.2 Result

Table 1 details the high-$\rho$ regime of §4. SAE-native removal is strongly language-selective at layer 17, where subspace interventions operate. The ratio of column means is $19.5\times$, compared with $1.05\times$ for projection of identical features. Mean $C$ over the eight affected languages is $+0.837$ for zeroing and $-0.014$ for projection on those same eight. At layer 29, $C = +0.90$ over ten affected languages, compared with $+0.21$ for projection on the same ten. SAE-native $C$ is higher in all eighteen affected cells across the two layers ($18/18$, Wilcoxon $p = 7.6 \times 10^{-6}$; Appendix B.13).

Averaging $\Delta_O$ over four other languages, as Deng et al. do, gives $6.9\times$ on the same features ($8.6\times$ with five-language selection), inside their range, and moves $C$ by at most 0.01.

Scaling the projection down to bracket SAE-native removal's effect size leaves the gap intact (§B.12).

**English features.** English is the only language whose selected features are not selective under either operation ($C = +0.159$ at layer 17, inert at layer 29, $-0.143$ on the 12B model). Its selected features carry general competence, consistent with its role as the unmarked language of an English-centric model (§7 shows this directly). They are also the selected set most aligned with the dominant shared direction (mean $\cos^2$ 0.72 at Gemma layer 17 and 0.90 on Llama, against at most 0.27 and 0.002 for the other languages).

**Robustness checks.** Appendix B tests dictionaries, feature selection, model size and intervention site. At layer 17, SAE-native removal is language-selective in every dictionary ($10/10$); projection is selective in none ($0/9$; Table 3). The dictionaries span three widths, three sparsity targets and a seed replicate. Across six selection criteria, including one optimised for projection, projection achieves at most weak selectivity at layer 17. The projection-optimised criterion is selective at layer 29 (Table 4). Gemma 3 12B at layer 24 reproduces the gap between operations (Table 5); Table 6 gives the layer-29 accuracy test.

## 6 Behavioural test: output language

If these are language-identity features, removing them should change the output language. We prompt with 128 MGSM problems per language, generate 96 tokens greedily, and label the output language by Unicode script or, for the five Latin-script languages, by `langdetect`.

**Selected features and controls.** Six of eleven selected sets move the output language, by 69.8 points on average (range 37.5–99.2). Removing Chinese latents makes the model answer Chinese problems in English 99.2% of the time. For every language that switches, outputs leaving the prompt's language are in English (Table 22). None of the three random-feature controls lowers any language's share by more than 2.3 points across thirty-three (draw $\times$ language) cells. For each of the six switching languages, the selected set differs from every control draw at $p < 10^{-16}$ (Fisher exact over 128 prompts). With language as the unit, six of eleven switch under selected removal and none under any control (exact sign test, $p = 0.031$).

The model remains coherent when it switches. Given a Chinese problem, it writes *"Let $E$ be the number of eggs that Janet's duck lays each day…"* and reasons through it correctly. English, Spanish, German, Russian and Telugu do not change output language. The $\nu$ score measures language specificity relative to the sampled languages; it does not guarantee that removing the selected features will switch output language.

**Directional projection.** Projecting the same features at the same site reduces the share of outputs in the prompt's language to 0% in every selected and control arm. Output is in German, Arabic or Tagalog regardless of the prompt's language. One random-feature control emits 128 empty strings in every language.

## 7 Reasoning under clean removal

With a removal that isolates language ($19.5\times$ at layer 17) and controls output language, we test whether removing a language's selected features improves its MGSM accuracy.

**SAE-native removal.** English, Chinese, Thai and Bengali lose accuracy relative to their controls, with intervals excluding zero (the first three also at the Bonferroni level for eleven tests; Table 2). All remaining seven intervals contain zero, so no language improves significantly at this budget. The apparent gains for Spanish, French and German shrink against their random-feature controls, which also improve. Chinese, Thai and Bengali switch language and lose accuracy; French, Japanese and Swahili switch without losing accuracy. Features selected on a disjoint corpus reproduce the verdict (pooled $-2.1$pp $[-3.6, -0.5]$, Appendix B.8).

Over the eleven languages the effect is $-7.7$pp $[-9.1, -6.4]$, and $-3.0$pp $[-4.5, -1.6]$ with English excluded. Zhao et al. (2025) report average gains of 0.9 to 3.5 points across ten models.[1] Every

[1] Computed from the per-language scores of their Table 1; some printed averages differ from their row means.

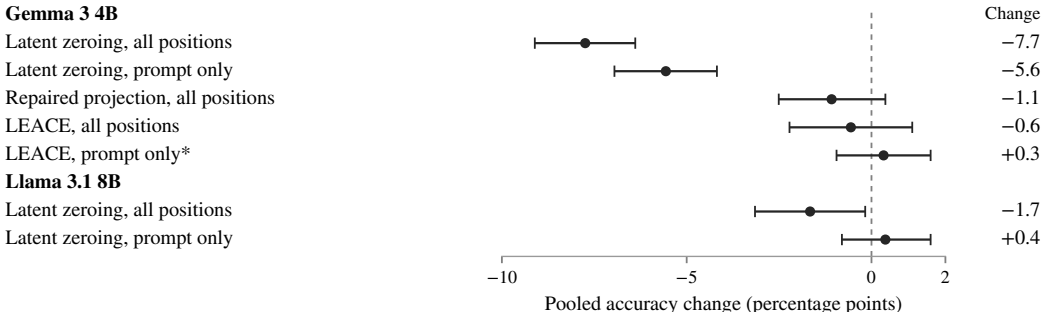



**Figure 2:** Pooled MGSM effects for the primary Gemma and Llama removals, with paired-bootstrap 95% intervals. Changes are in percentage points. SAE-native and repaired removals use matched random-feature controls. For LEACE, all-position effects are measured against other-language erasers, and prompt-only effects against the unmodified baseline (asterisk). No interval lies wholly above zero.

| Lang | base | selected | $\Delta$ | McNemar $p$ | control mean | sel. − control | 95% CI |
|---|---|---|---|---|---|---|---|
| en | 60.0 | 7.6 | $\mathbf{-52.4}$ | $\mathbf{4.4 \times 10^{-35}}$ | $+2.4$ | $\mathbf{-54.8}$ | $[-60.6, -49.1]$ |
| es | 60.4 | 64.4 | $+4.0$ | 0.087 | $+2.7$ | $+1.3$ | $[-2.4, +5.0]$ |
| fr | 50.4 | 54.8 | $+4.4$ | 0.18 | $+4.3$ | $+0.1$ | $[-5.3, +5.5]$ |
| de | 57.6 | 62.0 | $+4.4$ | 0.09 | $+2.4$ | $+2.0$ | $[-2.1, +6.1]$ |
| ru | 63.6 | 63.6 | $+0.0$ | 1.0 | $+0.7$ | $-0.7$ | $[-4.4, +3.0]$ |
| zh | 62.8 | 52.8 | $\mathbf{-10.0}$ | $\mathbf{0.0026}$ | $-1.3$ | $\mathbf{-8.7}$ | $[-14.5, -3.1]$ |
| ja | 59.6 | 56.8 | $-2.8$ | 0.41 | $-2.2$ | $-0.6$ | $[-5.3, +4.2]$ |
| th | 60.4 | 48.0 | $\mathbf{-12.4}$ | $\mathbf{4.7 \times 10^{-5}}$ | $+2.2$ | $\mathbf{-14.6}$ | $[-19.9, -9.4]$ |
| sw | 38.8 | 38.0 | $-0.8$ | 0.90 | $+0.1$ | $-0.9$ | $[-6.2, +4.4]$ |
| bn | 60.8 | 54.4 | $\mathbf{-6.4}$ | $\mathbf{0.040}$ | $+1.1$ | $\mathbf{-7.5}$ | $[-12.5, -2.6]$ |
| te | 57.2 | 56.4 | $-0.8$ | 0.89 | $+0.0$ | $-0.8$ | $[-5.9, +4.3]$ |
| pooled over eleven languages | | | | | | | $\mathbf{-7.7}$ $[-9.1, -6.4]$ |

**Table 2:** MGSM accuracy (%), layer 17, $k$=20, SAE-native removal at all positions, 250 problems per language, four random-feature controls; paired at the problem level (MGSM is parallel), bootstrap intervals over problems.

pooled estimate in this section is below their smallest reported gain.

**Alternative removals.** The repaired projection of §4.1 yields no significant improvement against controls in any language ($0/11$; Table 17, Figure 2). Chinese falls 13.2pp against its control (McNemar $p = 1.7 \times 10^{-4}$ against baseline). At Llama's low-$\rho$ site no language improves significantly (§B.3).

Confined to prompt tokens, as in Zhao et al. (2025), latent zeroing gives no significant pooled gain on either model. Telugu improves on Llama (Table 19). LEACE erasure also gives no significant pooled gain under either position policy (Table 20).

**The Qwen pipeline.** On Qwen-2.5-3B-Instruct we run their full pipeline in two forms. It builds a subspace from centred per-language means (Zhao et al., 2025, Algorithm 1), removes it from prompt tokens at layers 12–26 and re-injects it at layers 27–35. Applied to the residual stream, it lowers pooled accuracy at all twelve removal settings in the grid. Changes range from $-0.9$pp $[-2.6, +0.8]$ to $-13.1$pp, with ten significant costs (Table 15). At $(0.3, -0.2)$ the language subspace costs 6.9pp; three random rank-10 subspaces cost nothing (selected minus control $-7.5$pp $[-9.1, -5.8]$). Only Telugu improves, rising from a $1.6\%$ baseline by $+2.9$pp $[+0.1, +6.0]$.

We also apply the pipeline to each layer's MLP output, the edit site in the released code, using their prompts and greedy decoding. The example setting in their code, $(0.3, -0.1)$, changes pooled accuracy by $-0.5$pp $[-1.8, +0.8]$, compared with their $+2.8$. None of the four settings we ran there significantly changes pooled accuracy or any language's accuracy.

**Selection on evaluated problems.** Their reported numbers are per-language best points of a grid searched on the evaluated problems (their Appendix F: "we perform a grid search to find the best results"). On our residual-stream grid, where no setting raises pooled accuracy, this rule reports $+2.1$pp in-sample. Choosing on half the problems and scoring on the other half gives 0.0pp

$[-2.0, +1.7]$, and $-0.3$pp $[-1.6, +1.0]$ on the code-faithful runs (Appendix B.10).

No construction, position policy or model yields a significant pooled gain. Estimates range from $-13.1$ to $+0.4$pp, and the largest upper bound of any pooled interval is $+1.7$pp. Every pooled effect distinguishable from zero is a cost.

## 8 Related work

**Language features.** Deng et al. (2025) introduce the monolinguality criterion and obtain 5–10$\times$ cross-entropy selectivity by zeroing language-specific latents. Chou et al. (2025) and Andrylie et al. (2025) use them to steer output language, and Deng et al. (2026) to suppress code-switching. We compare both operations on identical latents at the same site, explaining our agreement with latent zeroing and disagreement with subspace interventions. Wong et al. (2026) report selective directional projection at late layers, consistent with our layer-dependent results (§B.2).

**Multilingual reasoning.** Zhao et al. (2025) attribute average MGSM gains across ten models to relieving interference by removing a language subspace at middle layers and re-injecting it at higher layers. Our four clean-removal constructions on three families give no significant pooled gain (§7). The example setting in their code, at its released edit site, does not reproduce their Qwen-2.5-3B gain (harness differences: Appendix B.10). Their construction's norm share is low at the targeted Qwen-2.5 layers and high at Gemma layer 17 (§4.1).

**Activation geometry.** Erasing a linear concept uses iterated nullspace projection (Ravfogel et al., 2020) or LEACE's covariance-weighted projection, which changes the representation least (Belrose et al., 2023). LEACE is selective at both sites, including where orthogonal projection fails (§4.1). Nguyen et al. (2026) reduce steering collateral through constrained optimisation weighted by the activation second moment. For removal, $\rho$ measures the subspace's share of this mass after normalising each token's residual. Our repair (§4.1) is a simpler use of this geometry.

**SAE reliability.** Multilingual models compute in a largely shared space (Qi et al., 2023; Wendler et al., 2024; Wu et al., 2025b; Dumas et al., 2025; Brinkmann et al., 2025) with language-specific neurons (Tang et al., 2024). Dictionaries trained on the same data learn different features (Chanin et al., 2024; Chanin & Garriga-Alonso, 2025; Leask et al., 2025; Paulo & Belrose, 2025); their value for steering is questioned (Wu et al., 2025a; Kantamneni et al., 2025; Korznikov et al., 2026; Heap et al., 2025; Jørgensen & Hansen, 2026). All ten dictionaries we test agree at a fixed operation (§B.1).

## 9 Limitations

$\rho$ is measured on language identity in Gemma 3 and Llama 3.1, on the mean-difference construction in Qwen-2.5, and on the refusal direction at ten layers of each model. The SAE-native slope shows the boundary is soft, and we claim no universal threshold. English's $-54.8$pp pulls down the pooled $-7.7$pp. Accuracy is measured on MGSM with greedy decoding only (Zhao et al. (2025) also report XWinograd and M-MMLU). The primary accuracy experiments target one layer at $k = 20$.

## 10 Conclusion

"Ablation" names two operations with opposite failure modes. One forward pass identifies destructive projection: a subspace carrying most of a token's norm removes the model. Removing shared directions from the subspace repairs the intervention for language and refusal. Language features are real, and their clean removal gives no significant gain in pooled reasoning accuracy in any setting we measure. Subspace interventions should report $\rho$ and collateral effects before attributing a change to a removed property.

### REPRODUCIBILITY

Code, the pre-registration document, per-cell result files, the script that recomputes every number in this paper from them, and the exact commands are provided as an anonymised archive in the supplementary material. Every experiment runs on one GPU with 40 to 80GB of memory, and each model's full sweep completes within a few GPU-hours. Models run in bfloat16; ablation bases are orthonormalised by QR in float32, and $\rho$ and every cross-entropy are accumulated in float32.

### AI USE STATEMENT

A large language model (Anthropic's Claude) was used under the authors' direction to help draft and polish sections of the text, to write and debug experiment and analysis code, and to critique experiments. The authors set the research question, the design and the claims, wrote and revised the paper, checked every number against the result files with the verification script in the supplementary material, and take full responsibility for the content.

## REFERENCES


W. Zhao et al., "When Less Language is More: Language-Reasoning Disentanglement Makes LLMs Better Multilingual Reasoners," *NeurIPS*, 2025. *arXiv:2505.15257*.

C. McDougall et al., "Gemma Scope 2: Sparse Autoencoders for Gemma 3," *Google DeepMind Technical Report*, 2025.

B. Deng et al., "Unveiling Language-Specific Features in Large Language Models via Sparse Autoencoders," *ACL*, 2025. *arXiv:2505.05111*.

B. Deng et al., "SASFT: Sparse Autoencoder-guided Supervised Finetuning to Mitigate Unexpected Code-Switching in LLMs," *ICLR*, 2026. *arXiv:2507.14894*.

C.-T. Chou et al., "Causal Language Control in Multilingual Transformers via Sparse Feature Steering," *ACL Student Research Workshop*, 2025. *arXiv:2507.13410*.

L. M. Andrylie et al., "Sparse Autoencoders Can Capture Language-Specific Concepts Across Diverse Languages," *arXiv:2507.11230*, 2025.

S. H. Wong, H. Sajjad, and A. B. Siddique, "LangFIR: Discovering Sparse Language-Specific Features from Monolingual Data for Language Steering," *COLM*, 2026. *arXiv:2604.03532*.

D. Chanin et al., "A is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders," *arXiv:2409.14507*, 2024.

D. Chanin and A. Garriga-Alonso, "Sparse but Wrong: Incorrect L0 Leads to Incorrect Features in Sparse Autoencoders," *arXiv:2508.16560*, 2025.

P. Leask et al., "Sparse Autoencoders Do Not Find Canonical Units of Analysis," *arXiv:2502.04878*, 2025.

G. Paulo and N. Belrose, "Sparse Autoencoders Trained on the Same Data Learn Different Features," *arXiv:2501.16615*, 2025.

Z. Wu et al., "AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders," *arXiv:2501.17148*, 2025a.

S. Kantamneni et al., "Are Sparse Autoencoders Useful? A Case Study in Sparse Probing," *arXiv:2502.16681*, 2025.

A. Korznikov et al., "Sanity Checks for Sparse Autoencoders: Do SAEs Beat Random Baselines?" *arXiv:2602.14111*, 2026.

T. Heap et al., "Sparse Autoencoders Can Interpret Randomly Initialized Transformers," *arXiv:2501.17727*, 2025.

M. G. Jørgensen and L. K. Hansen, "Steering LLMs? Actually, Sparse Autoencoders can outperform simple baselines," *arXiv:2605.31183*, 2026.

J. Qi, R. Fernández, and A. Bisazza, "Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models," *EMNLP*, 2023. *arXiv:2310.10378*.

T. Tang et al., "Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models," *ACL*, 2024.

C. Wendler et al., "Do Llamas Work in English? On the Latent Language of Multilingual Transformers," *ACL*, 2024.

Z. Wu et al., "The Semantic Hub Hypothesis: Language Models Share Semantic Representations Across Languages and Modalities," *ICLR*, 2025b.

C. Dumas et al., "Separating Tongue from Thought: Activation Patching Reveals Language-Agnostic Concept Representations," *ACL*, 2025.

J. Brinkmann et al., "Large Language Models Share Representations of Latent Grammatical Concepts Across Typologically Diverse Languages," *NAACL*, 2025.

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

F. Shi et al., "Language Models are Multilingual Chain-of-Thought Reasoners," *ICLR*, 2023.

T. Nguyen, T. A. Nguyen, S. Alemohammad, and R. G. Baraniuk, "Minimizing Collateral Damage in Activation Steering," *arXiv:2605.01167*, 2026.

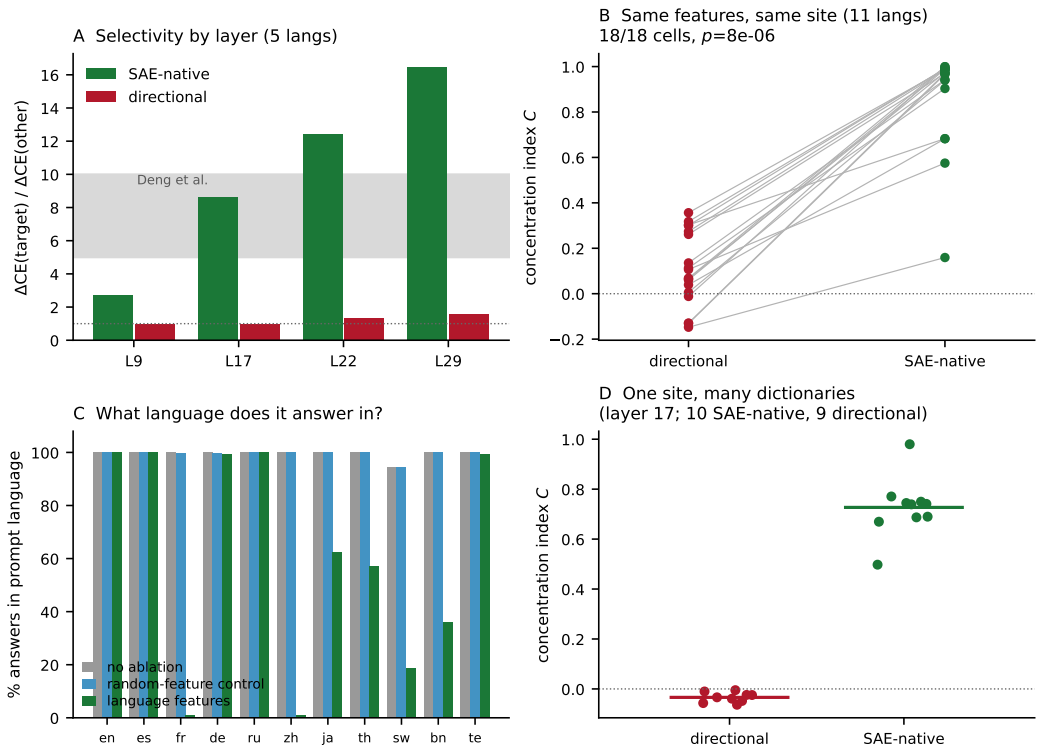



**Figure 3: The high-$\rho$ case study (Gemma 3 4B). (A)** Cross-entropy selectivity by layer; grey band is the 5–10× reported by Deng et al. (2025). **(B)** Concentration index for every (layer × language) cell, the same features under both operations. **(C)** Percentage of answers still in the prompt's language after ablation. **(D)** One activation site read by many dictionaries.

A. Arditi et al., "Refusal in Language Models Is Mediated by a Single Direction," *NeurIPS*, 2024. *arXiv:2406.11717*.

M. Sun, X. Chen, J. Z. Kolter, and Z. Liu, "Massive Activations in Large Language Models," *COLM*, 2024. *arXiv:2402.17762*.

N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman, "LEACE: Perfect Linear Concept Erasure in Closed Form," *NeurIPS*, 2023. *arXiv:2306.03819*.

S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg, "Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection," *ACL*, 2020.

# A The Gemma case study in one figure

Figure 3 collects four Gemma 3 4B results in one panel each: the layer sweep of §B.13, the per-cell comparison of §5, the output-language test of §6, and the dictionaries of §B.1.

# B Robustness in full

## B.1 Ten dictionaries reading one site

Sparse dictionaries trained on the same data learn different features (Leask et al., 2025; Paulo & Belrose, 2025). To test whether our conclusions depend on the dictionary, we hold the site at layer 17 and vary width, sparsity and training seed under both operations.

Every dictionary agrees that the features are language-selective under SAE-native removal, and every dictionary agrees they are non-selective under directional projection, with $C$ within 0.06 of zero for all nine.

The ratio ranges from $3.9\times$ to $57\times$ across dictionaries, with $C$ inside $[+0.50, +0.98]$. This instability is why we do not use the ratio for inference.

| Operation | mean $C$ | sd | range | sign agreement |
|---|---|---|---|---|
| SAE-native | $+\mathbf{0.727}$ | 0.118 | $[+0.498, +0.980]$ | $\mathbf{10/10}$ positive |
| Directional | $-0.034$ | 0.020 | $[-0.063, -0.005]$ | $0/9$ positive |

**Table 3:** Layer 17, $k$=20, widths $\{16\text{k}, 65\text{k}, 262\text{k}\}$ crossed with three sparsity levels, plus a seed replicate. Means are over the languages each removal affects. Paired over the nine dictionaries carrying both operations, Wilcoxon $p = 0.004$.

| | | Directional projection | | | SAE-native removal | | |
|---|---|---|---|---|---|---|---|
| Layer | criterion | $\Delta_T$ | $\Delta_O$ | $C$ | $\Delta_T$ | $\Delta_O$ | $C$ |
| 17 | $\nu$ (monolinguality) | 20.38 | 19.70 | $+0.004$ | 1.569 | 0.146 | $+0.771$ (4) |
| 17 | **dirspec** | 0.920 | 0.557 | $+\mathbf{0.220}$ | $-0.004$ | 0.000 | inert (0) |
| 17 | $\nu$, $\tau$=0.01 | 20.97 | 21.00 | $-0.013$ | 0.215 | 0.189 | $+0.092$ (1) |
| 17 | $\nu$, $\tau$=0.05 | 21.16 | 21.13 | $-0.012$ | 0.428 | 0.133 | $+0.673$ (3) |
| 17 | $\nu$, $\tau$=0.2 | 19.73 | 19.47 | $-0.008$ | 0.682 | 0.143 | $+0.741$ (4) |
| 17 | $\nu$, $\tau$=0.5 | 20.33 | 19.89 | $-0.004$ | 1.361 | 0.143 | $+0.679$ (4) |
| 29 | $\nu$ (monolinguality) | 18.95 | 11.95 | $+0.210$ | 1.887 | 0.114 | $+0.897$ (4) |
| 29 | **dirspec** | 3.808 | 0.365 | $+\mathbf{0.777}$ | 0.166 | 0.001 | $+0.997$ (2) |
| 29 | $\nu$, $\tau$=0.05 | 20.24 | 13.37 | $+0.200$ | 1.049 | 0.122 | $+0.699$ (4) |
| 29 | $\nu$, $\tau$=0.5 | 19.27 | 12.00 | $+0.214$ | 1.535 | 0.113 | $+0.877$ (4) |

**Table 4:** Six selection criteria, both operations, $k$=20, mean over five languages; the SAE-native $C$ is the mean over the languages the removal affects ($|\Delta_T| \geq 0.1$ nats), with that count in parentheses. *dirspec* selects features to maximise the concentration of the projection itself, the strongest case available for directional ablation. $\tau$ rows are random-token filtering at four thresholds at layer 17 and two at layer 29.

The seed replicate (same width, same sparsity target, different training seed) gives $C = 0.741$ against $0.750$ for its twin. Functional structure is stable across seeds even when the learned vectors differ.

### B.2 Selection criteria for directional projection

Wong et al. (2026) report selective directional ablation of SAE language features, directly challenging §5. We test whether differences in feature selection explain the disagreement.

Their pipeline differs from ours in two ways. They select by activation frequency and discard latents that fire on random-token sequences, treating these as responses to properties beyond language identity. Our selection uses Deng et al.'s monolinguality score $\nu$ without this filter. They also ablate the top two features; subspace interventions use larger $k$.

We test criteria most favourable to projection: $\nu$, a four-point sweep of the random-token threshold $\tau$, and a criterion that maximises projection's own objective. Since directional ablation removes $Q^\top Qh$, selectivity requires this projection to concentrate on the target language. For latent $j$ and language $L$, we measure over FLORES tokens

$$\text{dirspec}(j, L) \;=\; \frac{\mathbb{E}_{t \in L}\,(\hat{d}_j \cdot h_t)^2}{\mathbb{E}_{t \notin L}\,(\hat{d}_j \cdot h_t)^2}, \tag{4}$$

and select the top-$k$. Choosing features by the quantity projection depends on gives the strongest single-direction case with this dictionary. The criterion is greedy and claims no optimum over subspaces.

At layer 17, the best directional concentration across all six criteria is $+0.220$, obtained by *dirspec*. SAE-native removal of the monolinguality-selected features gives $+0.771$. Zeroing the projection-optimised features is inert, as expected for features chosen by direction that need not fire. Every other criterion, including all four filtering thresholds, stays within $\pm 0.05$ of zero. Filtering helps at no $\tau$. Failure persists even with selection optimised for projection, ruling out the monolinguality criterion as its cause.

At layer 29, *dirspec* reaches $C = +0.777$ with $\Delta_T = 3.8$ nats, selective by our criterion, and the only place in our results where directional projection behaves as Wong et al. describe. They also locate the strongest effects at late layers. Their Figure 4 calls English "a notable exception", consistent with English being the one non-selective language under both operations in both our models.

At layer 17, projection trades effect size for concentration. Selecting by $\nu$ gives $20.4$ nats at $C = +0.004$; *dirspec* gives $0.92$ nats at $C = +0.220$. Projection cannot combine a large effect with concentration here because directions carrying enough variance to matter are shared across languages. SAE-native removal avoids this trade-off: on $\nu$-selected features it changes cross-entropy by $1.57$ nats at $C = +0.771$. Its input-dependence selects the contribution at inference time. Our data support this trade-off at layer 17 and its absence at layer 29, leaving no basis for a claim that projection never works.

### B.3 A SECOND FAMILY: LLAMA 3.1 8B

Goodfire's ReLU SAE for Llama 3.1 8B Instruct reads layer 19 of 32, changing the architecture, tokeniser, SAE trainer and SAE type. We select features on disjoint FLORES sentences from the evaluation distribution; Appendix E explains why selection on chat-template last tokens does not transfer. Latent zeroing reaches $\bar{C} = +0.958$ over six affected languages, above Gemma's $+0.837$. Per-language firing ratios range from 2 to $172\times$, measured as selected latents' mean activation on the target divided by their mean on the other languages. The features and SAE-native operation generalise across families.

Directional projection also works here ($\bar{C} = +0.822$ over all eleven), as §4 predicts at $\rho = 0.05$–$0.12$. This supports a per-site criterion for when projection isolates a property.

Llama provides the cleanest test of the interference account: removal is selective by cross-entropy, so a reasoning benefit from deleting language features would be visible. The full accuracy protocol uses MGSM, 250 problems $\times$ eleven languages, three random-feature controls and item-level pairing. No language improves significantly ($0/11$; the largest apparent gain, Telugu $+3.6$pp, has CI $[-2.0, +9.2]$). Swahili alone degrades significantly ($-16.0$pp against its control, CI $[-22.4, -9.7]$, McNemar $p = 2.6 \times 10^{-5}$). The pooled effect is $-1.7$pp $[-3.2, -0.2]$. Swahili's features also carry the largest single-latent contribution on this model ($\Delta$CE $1.7$ nats), placing the decisive accuracy cost where the most language machinery is removed.

Confined to prompt tokens (Appendix B.9), removal gives a Swahili effect of $-4.9$pp $[-10.7, +0.8]$ and a pooled effect of $+0.4$pp $[-0.8, +1.6]$; Telugu improves. High $\rho$ confounds a projection-based test. At this low-$\rho$ site the test is clean, and the pooled gain predicted by the interference account is absent under both position policies.

### B.4 A LARGER MODEL

We repeat the cross-entropy comparison on Gemma 3 12B IT at layer 24 of 48, the same relative depth, using its own Gemma Scope 2 release and the same protocol with eight random-feature controls.

| | Directional | | SAE-native | |
|---|---|---|---|---|
| Lang | ratio | $C$ | ratio | $C$ |
| en | $1.0\times$ | $-0.024$ | $0.7\times$ | $-0.143$ |
| zh | $1.3\times$ | $+0.121$ | $\mathbf{97.2}\times$ | $+0.980$ |
| es | $1.2\times$ | $+0.075$ | inert | inert |
| bn | $1.2\times$ | $+0.082$ | $\mathbf{1251}\times$ | $+0.998$ |
| sw | $1.1\times$ | $+0.054$ | inert | inert |
| mean | | $+0.061$ | | $\mathbf{+0.612}$ |

**Table 5:** Gemma 3 12B IT, layer 24, $k$=20; inert marks $|\Delta_T| < 0.1$ nats and the SAE-native mean is over the three affected languages. The 4B model gives $-0.034$ and $+0.771$ for the same two columns on these five languages.

The pattern reproduces: mean concentration $+0.061$ under directional projection against $+0.612$ under SAE-native removal. Scoring the 4B model on the same five languages at layer 17, from the sweep of Figure 3A, gives $-0.034$ and $+0.771$. SAE-native is higher in every affected language except English. This recurring exception reflects English's role in these models.

### B.5 A SECOND INTERVENTION SITE

Repeating the accuracy experiment at layer 29 separates the patterns of failure of §7 cleanly. Baselines drift by up to 2pp between runs (bfloat16 generation depends on batching and library version); every contrast in this paper is formed inside one run, so the drift cancels.

| Lang | base | selected | $\Delta$ | McNemar $p$ | sel. − control | 95% CI |
|---|---|---|---|---|---|---|
| en | 60.0 | 64.0 | $+4.0$ | 0.17 | $+3.0$ | $[-1.7, +7.9]$ |
| zh | 62.8 | 53.2 | $\mathbf{-9.6}$ | $\mathbf{0.0071}$ | $\mathbf{-8.8}$ | $[-15.0, -2.5]$ |
| es | 60.4 | 60.0 | $-0.4$ | 1.0 | $-1.0$ | $[-5.9, +3.8]$ |
| bn | 60.8 | 60.8 | $+0.0$ | 1.0 | $-1.5$ | $[-7.0, +4.1]$ |
| sw | 38.8 | 36.8 | $-2.0$ | 0.62 | $-1.7$ | $[-7.7, +4.4]$ |
| pooled | | | | | | $-2.0$ $[-4.6, +0.5]$ |

**Table 6:** MGSM accuracy (%) at layer 29 on the five languages of the layer sweep, same protocol as Table 2.

Chinese, the language whose features are cleanest, degrades significantly at both sites ($-8.7$pp against its control at layer 17, $-8.8$pp here). English's $-54.8$pp does not replicate: at layer 29 the effect is $+3.0$pp and indistinguishable from controls. This follows the interpretation in §5: English's layer-17 features carry general competence, whose removal explains the collapse. At layer 29 its features behave like the other languages'. No language improves at either site.

### B.6 SPANISH AT LAYER 17

Spanish barely moves on any measure at layer 17: its selected latents change Spanish cross-entropy by $-0.011$ nats (Table 1), do not change the output language (Table 22), and leave accuracy inside the control band ($+1.3$pp against its control, Table 2). The selection procedure returns features that are nearly inert for Spanish at this site. We keep Spanish in every test; the limitation lies in feature selection, while the operation contrast of Table 1 is unaffected.

### B.7 SPARSITY OF THE ABLATION

At $k{=}1$ the same qualitative pattern holds but single features are unstable: the top Spanish latent is near-dead, and per-language $C$ swings between $-1$ and $+1$ on effects of order $10^{-3}$ nats. We therefore report $k{=}20$ throughout the main text and give $k{=}1$ in Appendix C.

### B.8 ACCURACY WITH FEATURES SELECTED ON A DISJOINT CORPUS

Section 7 selects features on last-token residuals of the MGSM prompts later scored for accuracy. To test selection on independent inputs, we apply the same criterion to FLORES sentences 300–549 per language, each placed in the chat template's question slot. These share no text with the scored problems or the FLORES sentences 0–199 scored in §5. The new sets share 17% of their latents with the MGSM-selected sets, ranging from 1 of 20 for Spanish to 10 for Swahili. The test therefore changes both corpus and features.

No language improves significantly against its control ($0/11$). The largest apparent gains, Swahili $+5.4$pp $[-0.2, +10.9]$ and French $+4.9$pp $[-0.7, +10.5]$, have intervals containing zero. The effects on English and Japanese are $-8.0$pp $[-13.5, -2.8]$ and $-29.7$pp $[-35.7, -23.8]$ against their controls. The pooled effect is $-2.1$pp $[-3.6, -0.5]$ (Table 16). Different features change which languages lose accuracy relative to Table 2; the negative pooled effect and absence of significant gains persist.

### B.9 REMOVAL ON THE PROMPT TOKENS ONLY, AND UNDER LEACE

Zhao et al. (2025) edit only prompt tokens. Tables 2 and 10 zero latents at every position, including generation. Table 19 repeats both with the hook active only during prefill and applies the same policy to random-feature controls. On Gemma, the pooled effect is $-5.6$pp $[-7.0, -4.2]$, or $-0.7$pp $[-2.1, +0.8]$ without English. English ($-54.5$pp), Chinese ($-8.7$pp) and Spanish ($-3.4$pp $[-6.2, -0.6]$) fall against their controls; no language improves. Prompt-only removal still changes

output language: only $4\%$ of Chinese, $16\%$ of French and $50\%$ of Bengali answers retain the prompt's language.

On Llama the pooled effect is $+0.4$pp $[-0.8, +1.6]$, and no language degrades significantly. Telugu improves by $+8.3$pp $[+3.3, +13.2]$, with 30 problems gained and 11 lost (McNemar $p = 0.0043$, $0.048$ after Bonferroni correction over eleven languages). Its three controls score $35.2$–$36.0\%$ against a $36.4\%$ baseline, and every output remains in Telugu.

Table 20 erases each language with LEACE in place of latent zeroing. The eraser is fitted on token residuals of $400$ translated MATH prompts per language (Zhao et al., 2025), which are disjoint from MGSM; scored on FLORES it is selective (mean $C = +0.96$ over ten affected languages, target changes of $0.1$–$0.8$ nats, no other-language mean above $0.03$). Applied at every position, its pooled effect against erasers of three other languages is $-0.6$pp $[-2.2, +1.1]$. No language's accuracy changes significantly. Erasure still changes output language: only $36\%$ of English, $44\%$ of Spanish and $13\%$ of Swahili answers retain the prompt's language. On prompt tokens alone, the pooled change against baseline is $+0.3$pp $[-0.9, +1.6]$. German rises ($+5.6$pp $[+1.2, +10.0]$) and Russian falls ($-3.6$pp $[-7.2, -0.4]$); neither survives correction over eleven tests.

We extract the number after an answer marker when present, otherwise the output's last number. The two rules agree on $99.7\%$ of the $16{,}500$ Gemma outputs of Table 19 and on all $13{,}750$ Llama outputs. Using the last number for every output, regardless of language, changes no pooled estimate in this section by more than $0.1$pp and no language by more than $0.8$pp.

### B.10 Reproducing the pipeline of Zhao et al.

Table 7 compares our two implementations of the pipeline of Zhao et al. (2025) on Qwen-2.5-3B-Instruct with their released code. Their hook edits each decoder layer's first output, which is the MLP output in their vLLM version. Their subspace is fitted on the same tensor. Our code-faithful run applies this edit through a hook on each layer's MLP in transformers.

The residual-stream removal grid uses $\lambda_{\text{mid}} \in \{0.1, \ldots, 0.4\}$ and $\lambda_{\text{high}} \in \{0, -0.2, -0.4\}$.

The example setting in their code, given in their README for another model, yields a pooled change of $-0.5$pp $[-1.8, +0.8]$ against baseline; their per-language scores give $+2.8$. The other settings we ran give $+0.4$ $[-0.9, +1.7]$ at $(0.2, -0.1)$, $-0.5$ $[-1.9, +0.9]$ at $(0.3, -0.2)$ and $-0.6$ $[-1.8, +0.7]$ at $(0.3, 0)$. No language changes significantly in any of them.

Their reported numbers average per-language best points from a grid searched on the evaluated problems, using accuracy and fidelity (their Appendix F and `grid_search_analyzer.py`). MGSM has no separate development split. Using accuracy alone, that rule reports $+2.1$pp on our fourteen residual-stream settings, none of which raises pooled accuracy (best $-0.5$pp). It reports $+1.0$pp on the four code-faithful settings. Choosing on a random half of the problems and scoring on the other half ($1{,}000$ splits) gives $0.0$pp $[-2.0, +1.7]$ and $-0.3$pp $[-1.6, +1.0]$.

Our baseline remains $3.0$ points below theirs, with the gap concentrated in English, Spanish and French. Answer extraction does not explain it: both extractors fall back to the last number, and only 3 of 250 English outputs contain a `\boxed` answer.

### B.11 Mean-centred projection

Defining $\rho$ as the share of centred variance in a subspace changes little on token-level activations. A few massive-activation tokens also dominate the centred spectrum: its top component carries $0.996$ of centred variance at Gemma layer 17. The $k$=20 selected sets score within $0.04$ of their norm share at all four layers. Projecting the centred activation, $h \leftarrow h - Q^\top Q(h - \mu)$ with $\mu$ the mean residual at the site, is the standard remedy for a shared offset. At Gemma layer 17 it lowers $\rho$ from $0.72$ to $0.29$ and leaves the operation non-selective: the selected sets cost $1.5$–$12.3$ nats averaged over languages at mean $C = +0.14$, and random-feature controls cost 6–9 nats. Decoder directions share components beyond the mean. Removing the top four principal directions reaches $\rho \approx 0$. At Llama layer 19 centring changes little ($\rho = 0.08$, $C$ between $+0.69$ and $+0.98$ for every language except English).

Restoring each token's norm after projection, $h' = (h - Q^\top Qh)\,\|h\|/\|h - Q^\top Qh\|$, tests whether

| | their released code | residual-stream runs | code-faithful run |
|---|---|---|---|
| edit site | MLP output | residual stream | MLP output |
| fitted on | last-token MLP output | last-token residual | last-token MLP output |
| fitting prompts | translated MATH | same, 3,000 per language | same, 3,000 per language |
| layers | 12–26 out, 27–35 in | same | same |
| prompt (chat format) | localised template | question only | localised template |
| decoding | greedy | greedy | greedy |
| new tokens | 16,384 | 384 | 1,024 |
| scoring | boxed, marker, last number | marker, last number | marker, last number |
| reported | best point per language | every grid point | every setting completed |
| baseline, mean (en) | 55.7 (86.0) | 47.6 (70.8) | 52.7 (76.4) |

**Table 7:** Protocols for the pipeline of Zhao et al. (2025) on Qwen-2.5-3B-Instruct. The localised template is their `mgsm_0shots` prompt ("Question: … Step-by-Step Answer:"); first-column baselines are the means of the per-language scores in their Table 1. Of the code-faithful runs, $(0.1, -0.1)$ and one random subspace were stopped after three languages and one.

the damage is a loss of magnitude. On identical features in the same run, plain projection gives mean $C = +0.00$ with 19 nats of mean collateral. Norm-restored projection gives $-0.01$ with 24, and random-feature controls cost 12–26 nats under both. Rescaling fails to repair the damage because projection changes the token's direction; it cannot recover the removed massive-activation component.

### B.12 Matched magnitude

A smaller perturbation could look more concentrated for that reason alone, so we scale the projection, $h \leftarrow h - \alpha Q^\top Q h$, on identical features to bracket SAE-native removal's target effect. At $\alpha = 0.1$ and $0.2$ the mean target change is $0.57$ and $2.8$ nats, bracketing SAE-native's $1.13$, and mean $C$ is $+0.16$ and $+0.15$ against $+0.84$. The gap survives at both scales.

### B.13 The paired comparison and the layer sweep

Against the eight random-feature controls, SAE-native selectivity is significant at both layers (per-language $z$-scores of the selected set's $C$ against the draws, combined by Stouffer's method: $z = 6.94$ and $3.54$), with the controls' mean $C$ at $+0.118$ and $+0.211$. Pairing the two operations across the eighteen (layer $\times$ language) cells where SAE-native removal has an effect:

$$\text{directional } \bar{C} = +0.11 \quad \text{vs} \quad \text{SAE-native } \bar{C} = +0.87, \quad \text{higher in } 18/18 \text{ cells},$$

Wilcoxon signed-rank $p = 7.6 \times 10^{-6}$, paired $t = 14.2$, $p = 7.8 \times 10^{-11}$.

A four-layer sweep $\{9, 17, 22, 29\}$ on the five-language subset {en, zh, es, bn, sw}, with twelve random-feature controls, gives the same picture with the same sign at every depth ($\bar{C} = +0.08$ vs $+0.75$, higher in $14/14$ affected cells, Wilcoxon $p = 1.2 \times 10^{-4}$; Figure 3A). We use the eleven-language, two-layer sweep as the primary because it has more languages per cell and a lower $\Delta_O$ floor.

**Norm-share summary statistics.** In the Llama sweep of §4, $k = 20$, 100 and 500 give $\rho = 0.074$, $0.197$ and $0.400$. Directional $C$ falls from $+0.822$ to $+0.697$ to $+0.361$; SAE-native $C$ stays at $+0.958$, $+0.968$ and $+0.935$. Across the pooled subspaces, Pearson $r = -0.828$. Within Llama's single site, across its three values of $k$, Spearman is $-0.79$ over 33 cells ($p = 5.5 \times 10^{-8}$); it is $-0.51$ over the 20 Gemma cells ($p = 0.02$).

Among the projection subspaces with $\rho$ above one half, median collateral is $16.6$ nats and the largest $C$ is $+0.46$. For $\rho$ below $0.15$, median $C$ is $0.8$.

SAE-native selectivity also falls with $\rho$: $r = -0.49$ and $R^2 = 0.24$ over the 39 affected cells, with mean $C$ falling to $+0.74$ in the highest-$\rho$ bin. Its slope is a third as steep ($-0.37$ versus $-0.97$ per unit $\rho$).

At layer 29, the ratios of column means in §5 are $33.0\times$ for SAE-native removal and $1.52\times$ for projection.

For the centred-mean construction on Qwen-2.5-7B and Qwen-2.5-3B, full-strength cross-entropy costs are 2.0 and 2.2 nats; the default strength costs 0.7 on both.

LEACE removes $0.2\%$ and $1.5\%$ of a typical token's squared norm at the Gemma and Llama sites, respectively (Table 18).

Adding Llama's second to fourth directions gives $\rho = 0.27$ without increasing the first, massive-activation token's contribution. It lowers $C$ to $+0.35$ in all eleven languages ($p = 0.001$). Across the 44 distinct subspaces in this manipulation, $C$ tracks $\rho$ at Spearman $-0.71$ ($p = 7.9 \times 10^{-8}$).

For the refusal repairs in §4.1, the original directions at Gemma layers 17 and 20 cost 19 and 11 nats at their own layers.

## C $k = 1$ RESULTS

Table 8 repeats the layer-17 comparison with a single latent per language. Directional projection of one decoder direction still costs 9–23 nats on every language and is non-selective wherever it has an effect; SAE-native removal of one latent moves cross-entropy by under 0.1 nats for three of five languages, so most cells are inert and the two that remain sit at the extremes. Single latents are too small a unit for this measurement, which is why $k$=20 is the main-text setting.

| | Directional projection | | | SAE-native removal | | |
|---|---|---|---|---|---|---|
| Lang | $\Delta_T$ | $\Delta_O$ | $C$ | $\Delta_T$ | $\Delta_O$ | $C$ |
| en | 11.38 | 17.56 | −0.21 | 0.882 | 0.534 | +0.25 |
| zh | 23.47 | 12.61 | +0.30 | 0.030 | −0.001 | inert |
| es | −0.16 | 0.29 | −0.30 | 0.001 | 0.000 | inert |
| bn | 19.57 | 23.32 | −0.09 | −0.201 | 0.000 | +1.00 |
| sw | 9.43 | 10.70 | −0.06 | 0.023 | 0.003 | inert |

**Table 8:** Layer 17, $k$=1, five languages, FLORES-200 cross-entropy in nats; inert marks $|\Delta_T| < 0.1$.

## D SUPPLEMENTARY TABLES

Tables 9–22 give the per-cell numbers behind §4, §5, §6 and §7.

| | Directional projection | | | SAE-native removal | | |
|---|---|---|---|---|---|---|
| Lang | $\Delta_T$ | $\Delta_O$ | $C$ | $\Delta_T$ | $\Delta_O$ | $C$ |
| en | 11.13 | 13.12 | −0.082 | −0.029 | −0.040 | inert |
| es | 23.77 | 18.94 | +0.113 | 1.155 | 0.059 | +0.904 |
| fr | 18.09 | 17.88 | +0.006 | 1.734 | −0.054 | +0.940 |
| de | 21.04 | 18.63 | +0.061 | 0.965 | 0.006 | +0.988 |
| ru | 54.81 | 26.00 | +0.356 | 2.959 | −0.020 | +0.987 |
| zh | 36.03 | 18.66 | +0.318 | 2.332 | 0.012 | +0.990 |
| ja | 18.32 | 9.86 | +0.300 | 2.730 | 0.011 | +0.992 |
| th | 27.30 | 14.61 | +0.303 | 1.204 | 0.227 | +0.683 |
| sw | 10.20 | 5.97 | +0.261 | 1.236 | 0.019 | +0.970 |
| bn | 23.55 | 13.42 | +0.274 | 2.560 | 0.018 | +0.986 |
| te | 21.95 | 17.73 | +0.106 | 1.137 | 0.307 | +0.575 |
| mean over the 10 affected | | | +0.210 | | | +0.901 |

**Table 9:** Layer 29, $k$=20, eleven languages, FLORES-200 cross-entropy in nats; the companion to Table 1. Inert marks $|\Delta_T| < 0.1$.

## E FEATURE SELECTION

**Procedure.** For each dictionary we encode the last-prompt-token residual of 250 MGSM problems per language and score every latent twice. The monolinguality score of latent $j$ for language $L$ is $\nu_{j,L} = \mu_{j,L} - \max_{L' \neq L} \mu_{j,L'}$, with $\mu_{j,L}$ its mean activation on $L$ (Deng et al., 2025). The probe is

| Lang | base | selected | $\Delta$ | McNemar $p$ | control mean | sel. $-$ control | 95% CI |
|---|---|---|---|---|---|---|---|
| en | 88.4 | 88.0 | $-0.4$ | 1 | $-0.3$ | $-0.1$ | $[-2.0, +1.6]$ |
| es | 72.8 | 71.2 | $-1.6$ | 0.6 | $-1.1$ | $-0.5$ | $[-5.1, +4.0]$ |
| fr | 68.8 | 66.0 | $-2.8$ | 0.28 | $-0.7$ | $-2.1$ | $[-6.4, +2.0]$ |
| de | 68.4 | 68.4 | $+0.0$ | 1 | $-0.4$ | $+0.4$ | $[-3.7, +4.5]$ |
| ru | 68.0 | 67.2 | $-0.8$ | 0.84 | $-0.4$ | $-0.4$ | $[-4.0, +3.2]$ |
| zh | 64.4 | 65.2 | $+0.8$ | 0.86 | $+1.5$ | $-0.7$ | $[-5.2, +3.7]$ |
| ja | 55.6 | 54.8 | $-0.8$ | 0.89 | $-1.1$ | $+0.3$ | $[-4.9, +5.5]$ |
| th | 60.4 | 62.0 | $+1.6$ | 0.64 | $+0.0$ | $+1.6$ | $[-3.3, +6.5]$ |
| sw | 56.4 | 42.0 | $-14.4$ | $2.6 \times 10^{-5}$ | $+1.6$ | $-16.0$ | $[-22.4, -9.7]$ |
| bn | 47.2 | 42.4 | $-4.8$ | 0.18 | $-0.5$ | $-4.3$ | $[-10.4, +1.9]$ |
| te | 36.4 | 39.2 | $+2.8$ | 0.4 | $-0.8$ | $+3.6$ | $[-2.0, +9.2]$ |
| pooled | | | | | | $-1.7$ | $[-3.2, -0.2]$ |

**Table 10:** MGSM accuracy (%) on Llama 3.1 8B Instruct, layer 19, $k$=20, SAE-native removal at all positions, 250 problems per language, three random-feature controls; the companion to Table 2.

| Site | lang | $\rho$ | $C_{\text{dir}}$ | $C_{\text{sae}}$ |
|---|---|---|---|---|
| Gemma L17 | zh | 0.524 | $+0.150$ | $+0.950$ |
| Gemma L17 | es | 0.677 | $+0.050$ | inert |
| Gemma L17 | bn | 0.677 | $-0.092$ | $+0.986$ |
| Gemma L17 | sw | 0.908 | $-0.070$ | $+0.991$ |
| Gemma L17 | en | 0.953 | $-0.206$ | $+0.156$ |
| Gemma L22 | zh | 0.843 | $+0.300$ | $+0.986$ |
| Gemma L22 | es | 0.845 | $+0.185$ | inert |
| Gemma L22 | sw | 0.911 | $+0.000$ | $+1.000$ |
| Gemma L22 | en | 0.912 | $-0.127$ | $+0.215$ |
| Gemma L22 | bn | 0.919 | $+0.082$ | $+0.642$ |
| Gemma L29 | bn | 0.701 | $+0.247$ | $+0.869$ |
| Gemma L29 | es | 0.750 | $+0.128$ | $+0.904$ |
| Gemma L29 | sw | 0.787 | $+0.252$ | $+0.976$ |
| Gemma L29 | en | 0.825 | $-0.063$ | inert |
| Gemma L29 | zh | 0.890 | $+0.458$ | $+0.840$ |
| Gemma L9 | sw | 0.938 | $+0.094$ | inert |
| Gemma L9 | zh | 0.940 | $-0.184$ | inert |
| Gemma L9 | es | 0.942 | $-0.118$ | inert |
| Gemma L9 | en | 0.945 | $-0.008$ | $+0.009$ |
| Gemma L9 | bn | 0.948 | $-0.060$ | $+1.000$ |
| Llama L19, $k$=100 | de | 0.149 | $+0.915$ | $+1.000$ |
| Llama L19, $k$=100 | ru | 0.189 | $+0.881$ | $+0.993$ |
| Llama L19, $k$=100 | es | 0.189 | $+0.777$ | $+0.994$ |
| Llama L19, $k$=100 | fr | 0.192 | $+0.661$ | $+0.994$ |
| Llama L19, $k$=100 | zh | 0.194 | $+0.760$ | $+0.998$ |
| Llama L19, $k$=100 | th | 0.202 | $+0.841$ | inert |
| Llama L19, $k$=100 | ja | 0.205 | $+0.724$ | $+0.994$ |
| Llama L19, $k$=100 | sw | 0.206 | $+0.654$ | $+0.991$ |
| Llama L19, $k$=100 | te | 0.208 | $+0.704$ | $+0.806$ |
| Llama L19, $k$=100 | bn | 0.217 | $+0.775$ | $+0.947$ |
| Llama L19, $k$=100 | en | 0.217 | $-0.024$ | inert |
| Llama L19, $k$=20 | de | 0.046 | $+0.886$ | inert |
| Llama L19, $k$=20 | ja | 0.054 | $+0.970$ | $+0.999$ |
| Llama L19, $k$=20 | zh | 0.055 | $+0.970$ | $+0.998$ |
| Llama L19, $k$=20 | th | 0.057 | $+0.931$ | inert |
| Llama L19, $k$=20 | sw | 0.061 | $+0.931$ | $+0.996$ |
| Llama L19, $k$=20 | te | 0.072 | $+0.737$ | $+0.810$ |
| Llama L19, $k$=20 | fr | 0.072 | $+0.897$ | $+0.994$ |
| Llama L19, $k$=20 | es | 0.086 | $+0.948$ | inert |
| Llama L19, $k$=20 | ru | 0.095 | $+0.950$ | inert |
| Llama L19, $k$=20 | bn | 0.098 | $+0.877$ | $+0.950$ |
| Llama L19, $k$=20 | en | 0.117 | $-0.055$ | inert |
| Llama L19, $k$=500 | en | 0.361 | $+0.023$ | inert |
| Llama L19, $k$=500 | de | 0.362 | $+0.431$ | $+0.989$ |
| Llama L19, $k$=500 | fr | 0.371 | $+0.414$ | $+0.982$ |
| Llama L19, $k$=500 | ru | 0.387 | $+0.509$ | $+0.995$ |
| Llama L19, $k$=500 | es | 0.388 | $+0.395$ | $+0.974$ |
| Llama L19, $k$=500 | zh | 0.394 | $+0.470$ | $+0.994$ |
| Llama L19, $k$=500 | ja | 0.422 | $+0.180$ | $+0.987$ |
| Llama L19, $k$=500 | te | 0.423 | $+0.151$ | $+0.783$ |
| Llama L19, $k$=500 | sw | 0.427 | $+0.407$ | $+0.974$ |
| Llama L19, $k$=500 | th | 0.430 | $+0.450$ | $+0.730$ |
| Llama L19, $k$=500 | bn | 0.434 | $+0.539$ | $+0.939$ |

**Table 11:** The 53 cells behind Figure 1A: one selected subspace per (model, site, $k$, language), its $\rho$ on FLORES tokens, and its concentration index under each operation. Gemma cells use the five-language selection (en, zh, es, bn, sw) at $k$=20; Llama cells use the eleven-language selection on FLORES sentences disjoint from the scored ones. A cell is inert under SAE-native removal when $|\Delta_T| < 0.1$ nats; inert cells enter no mean, count or correlation.

one multinomial logistic regression over the standardised latent activations, with the language as the label (scikit-learn defaults: $\ell_2$ penalty, $C = 1$, L-BFGS), fitted on a random half of the problems and 98.3% accurate on the other half at layer 17. A latent's probe score for $L$ is the absolute value of its coefficient in $L$'s row. This ranks latents directly without a map from residual space to latents. Each criterion shortlists the top 0.305% of the dictionary (50 latents at 16k). The selected set is the intersection of the two shortlists in order of $\nu$, truncated to $k$. If the intersection holds fewer than $k$ latents, we complete it by summed rank under both criteria so every cell has exactly $k$. On Llama the FLORES selection ranks active latents by $\nu$ alone.

**Selection site on Llama.** MGSM last-token selection uses the final formatting token of the chat template. Evaluating these features on FLORES left only 2,846 of 65,536 Goodfire latents active and made Llama's latent zeroing nearly inert ($\Delta$CE $\leq 0.04$ nats). The same procedure transferred successfully on Gemma. Selecting on FLORES tokens from the scored distribution raised the active count to 40,573 and produced the selective features in §B.3. Features should therefore be selected on the intended evaluation distribution.

| | | ΔCE (nats) | | refusal rate (%) | | |
|---|---|---|---|---|---|---|
| Layer | $\rho$ | site only | all layers | harmful, site | harmful, all | harmless, all |
| *Gemma 3 4B IT* (baseline refusal 71% harmful, 1% harmless) | | | | | | |
| 4 | 0.173 | 4.34 | 31.88 | 0 | 0 | 0 |
| 6 | 0.670 | 10.84 | 40.09 | 0 | 0 | 0 |
| 8 | 0.928 | 4.66 | 31.92 | 0 | 0 | 0 |
| 10 | 0.929 | 7.74 | 33.06 | 0 | 0 | 0 |
| 12 | 0.871 | 21.88 | 28.15 | 0 | 0 | 0 |
| 14 | 0.321 | 23.72 | 23.92 | 0 | 0 | 0 |
| 17 | 0.429 | 19.17 | 23.66 | 0 | 0 | 0 |
| 20 | 0.248 | 11.08 | 20.68 | 0 | 0 | 0 |
| 24 | 0.280 | 7.29 | 19.79 | 0 | 0 | 0 |
| 29 | 0.003 | 0.59 | 10.99 | 78 | 0 | 0 |
| *Llama 3.1 8B Instruct* (baseline refusal 88% harmful, 2% harmless) | | | | | | |
| 4 | 0.001 | 0.00 | 0.00 | 94 | 85 | 2 |
| 6 | 0.001 | 0.00 | 0.00 | 92 | 60 | 2 |
| 8 | 0.001 | 0.00 | 0.02 | 69 | 18 | 2 |
| 10 | 0.000 | 0.00 | 0.00 | 85 | 39 | 2 |
| 12 | 0.001 | 0.01 | 0.01 | 91 | 96 | 1 |
| 14 | 0.011 | 0.01 | 0.03 | 90 | 99 | 5 |
| 16 | 0.024 | 0.01 | 0.05 | 97 | 100 | 17 |
| 19 | 0.025 | 0.00 | 0.05 | 86 | 100 | 5 |
| 22 | 0.025 | 0.01 | 0.05 | 80 | 97 | 5 |
| 26 | 0.032 | 0.02 | 0.05 | 81 | 97 | 3 |

**Table 12:** The refusal direction of Arditi et al. (2024) extracted at each layer: its $\rho$ on FLORES tokens, mean cross-entropy change when projected out at that layer only and at every layer, and refusal rates under both protocols on 100 held-out harmful and 100 harmless prompts.

| Model, site | basis ($k$=20 selected features) | mean $\rho$ | mean $\Delta_T$ | mean $\Delta_O$ | mean $C$ ($n$) |
|---|---|---|---|---|---|
| Gemma 3 4B, L17 | unchanged basis | 0.675 | +21.31 | +20.03 | +0.010 (8) |
| Gemma 3 4B, L17 | top-4 directions removed | 0.000 | +0.67 | −0.03 | +0.790 (8) |
| Gemma 3 4B, L29 | unchanged basis | 0.765 | +27.20 | +17.29 | +0.204 (9) |
| Gemma 3 4B, L29 | top-4 directions removed | 0.001 | +0.75 | +0.01 | +0.964 (9) |
| Llama 3.1 8B, L19 | unchanged basis | 0.074 | +5.52 | +0.61 | +0.822 (11) |
| Llama 3.1 8B, L19 | top-1 direction added | 0.096 | +8.59 | +3.41 | +0.401 (11) |
| Llama 3.1 8B, L19 | top-4 directions added | 0.290 | +8.79 | +4.53 | +0.285 (11) |
| Llama 3.1 8B, L19 | 2nd–4th directions added | 0.270 | +8.02 | +3.45 | +0.352 (11) |
| Llama 3.1 8B, L19 | random direction added | 0.074 | +5.54 | +0.61 | +0.822 (11) |

**Table 13:** The fixed-$k$ manipulations of §4.1: the same selected features with the site's top principal directions removed from the basis (Gemma) or added to it (Llama). Means are over the languages the manipulated basis affects by at least 0.1 nats, $n$ in parentheses, and the unchanged basis is averaged over the same languages. Principal directions are eigenvectors of the uncentred token second-moment matrix at the site; at Gemma layer 17 three stand above a flat bulk (shares 0.997, $1.0 \times 10^{-3}$, $4.9 \times 10^{-4}$, then $5.0 \times 10^{-5}$), and removing four covers them. Every repaired basis has rank 20, its smallest singular value at least 0.48. Each row is measured within one run; the 2nd–4th row comes from a second run that reproduces the other Llama rows exactly.

## F PRE-REGISTRATION

A pre-registration document (included in the supplementary material; its version history will be made public with the code) was committed before any grid result was computed. It fixed the feature selection (the monolinguality criterion intersected with a supervised probe on MGSM last-token residuals, with no accuracy number entering selection), $k = 20$, the dictionary grid at layer 17, the accuracy protocol (250 problems, greedy decoding, 384 new tokens, paired bootstrap), and six predictions about MGSM accuracy under projection at a single token position, with a mixed-effects model as the headline test. Of these, P3 and P4 held (dictionaries, and seed replicates, disagree under that operation), P2 held for English and not for Bengali, and P1, P5 and P6 failed (baselines within ±4pp of a pilot run; an accuracy-free purity statistic predicting accuracy, and doing so better than simpler statistics). Everything this paper reports was outside that design: the operation contrast, SAE-native removal, $\rho$, the manipulations, the random-feature controls and the 0.1-nat inert rule are

| Lang | $\rho$ | $\eta$ | own | $b_{\text{dec}}$ | other latents | error | cross | $\eta$, top-4 removed |
|---|---|---|---|---|---|---|---|---|
| en | 0.989 | 0.935 | 76.1% | 115% | 25% | 0.1% | −41% | 0.0003 |
| es | 0.872 | 0.847 | 0.0% | 43% | 61% | 0.0% | −4% | 0.0001 |
| fr | 0.315 | 0.306 | 0.0% | 43% | 61% | 0.0% | −4% | 0.0001 |
| de | 0.427 | 0.416 | 0.0% | 43% | 61% | 0.0% | −5% | 0.0001 |
| ru | 0.688 | 0.664 | 0.0% | 42% | 61% | 0.0% | −3% | 0.0001 |
| zh | 0.545 | 0.526 | 0.0% | 42% | 62% | 0.0% | −4% | 0.0002 |
| ja | 0.851 | 0.853 | 0.0% | 44% | 61% | 0.0% | −4% | 0.0002 |
| th | 0.911 | 0.910 | 0.0% | 43% | 61% | 0.0% | −4% | 0.0002 |
| sw | 0.937 | 0.947 | 0.0% | 43% | 61% | 0.0% | −4% | 0.0002 |
| bn | 0.630 | 0.616 | 0.0% | 44% | 61% | 0.0% | −6% | 0.0001 |
| te | 0.949 | 0.950 | 0.0% | 43% | 61% | 0.0% | −4% | 0.0002 |

**Table 14:** What projection removes at Gemma 3 4B layer 17, decomposed. With $P = Q^\top Q$ and $s(h)$ the selected latents' own reconstruction, $\eta = \mathbb{E}_t \, \|P(h_t - s(h_t))\|^2 / \|h_t\|^2$ is the share of a token's squared norm that projection removes beyond what SAE-native removal would; *own* is the latents' own contribution as a share of the removed norm. Writing $h - s(h) = u + b_{\text{dec}} + \varepsilon$ (unselected latents, decoder bias, reconstruction error), the next four columns give each term's share of $\mathbb{E}\|P(h - s)\|^2$ and the cross terms; the terms overlap along the shared direction, so shares can exceed one and the cross terms are negative. The last column is $\eta$ after the basis is orthogonalised against the site's top four directions. 44,000 FLORES tokens, eleven-language selection, $k$=20.

| $\lambda_{\text{mid}}$ | $\lambda_{\text{high}}$ | en | es | fr | de | ru | zh | ja | th | sw | bn | te | pooled | 95% CI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | −0.2 | −0.4 | −0.8 | +3.6 | +1.2 | −0.8 | +0.8 | +0.4 | −4.0 | −1.2 | −5.6 | +1.6 | −0.5 | [−1.9, +0.9] |
| 0.0 | −0.4 | +0.0 | −0.8 | +2.4 | +0.0 | −2.4 | +1.2 | −8.0 | −7.2 | −3.6 | −8.4 | +3.6 | −2.1 | [−3.6, −0.5] |
| 0.1 | 0.0 | +5.2 | −8.8 | +0.4 | −2.8 | −6.0 | −1.6 | +0.4 | +2.4 | −2.4 | −12.4 | +0.8 | −2.3 | [−3.8, −0.7] |
| 0.1 | −0.2 | +2.0 | −6.8 | +3.6 | +2.0 | −6.4 | +3.2 | +0.0 | +3.2 | −2.0 | −12.4 | +2.4 | −1.0 | [−2.7, +0.7] |
| 0.1 | −0.4 | +2.8 | −4.4 | −0.4 | −4.0 | −2.8 | −2.0 | +0.0 | −0.8 | +3.2 | −3.2 | +1.6 | −0.9 | [−2.6, +0.8] |
| 0.2 | 0.0 | +2.4 | −8.8 | −0.8 | −3.2 | −8.4 | −3.2 | −0.8 | −0.4 | −3.6 | −11.2 | −0.8 | −3.5 | [−5.2, −1.8] |
| 0.2 | −0.2 | +2.8 | −7.6 | −3.2 | −3.6 | −13.6 | −6.0 | −5.2 | −1.6 | −2.8 | −11.2 | +4.4 | −4.3 | [−6.0, −2.5] |
| 0.2 | −0.4 | −4.0 | −10.0 | −2.8 | −4.8 | −9.2 | −10.4 | −8.0 | −3.6 | −1.2 | −2.4 | +2.8 | −4.9 | [−6.7, −3.1] |
| 0.3 | 0.0 | −4.0 | −10.4 | −2.0 | −9.6 | −12.0 | −6.4 | −11.6 | −8.4 | −2.4 | −17.2 | +1.2 | −7.5 | [−9.3, −5.7] |
| 0.3 | −0.2 | −1.2 | −10.0 | −2.0 | −12.0 | −10.8 | −10.4 | −8.8 | −9.2 | −0.4 | −14.8 | +3.6 | −6.9 | [−8.8, −5.1] |
| 0.3 | −0.4 | −3.2 | −14.4 | −2.0 | −13.6 | −18.0 | −14.0 | −8.8 | −9.2 | −1.2 | −11.6 | +6.0 | −8.2 | [−10.0, −6.3] |
| 0.4 | 0.0 | −10.0 | −16.0 | −2.4 | −11.6 | −18.4 | −16.0 | −13.6 | −12.4 | −1.6 | −14.4 | +2.0 | −10.4 | [−12.4, −8.5] |
| 0.4 | −0.2 | −17.2 | −14.8 | −11.6 | −16.0 | −17.6 | −18.0 | −12.0 | −17.6 | −4.0 | −16.4 | +6.4 | −12.6 | [−14.5, −10.7] |
| 0.4 | −0.4 | −5.2 | −24.4 | −12.4 | −21.2 | −19.6 | −20.0 | −14.8 | −18.4 | −2.8 | −9.2 | +4.0 | −13.1 | [−15.0, −11.2] |

**Table 15:** MGSM accuracy change (pp) against baseline under the full pipeline of Zhao et al. (2025) on Qwen-2.5-3B-Instruct: removal of the rank-10 language subspace at layers 12–26 with strength $\lambda_{\text{mid}}$ and re-injection at layers 27–35 with strength $\lambda_{\text{high}}$ (negative values inject), on prompt tokens, over their grid; 250 problems per language, paired bootstrap over problems for the pooled interval. $\lambda_{\text{mid}} = 0$ rows are re-injection alone.

exploratory.

| Lang | base | selected | $\Delta$ | McNemar $p$ | control mean | sel. $-$ control | 95% CI |
|---|---|---|---|---|---|---|---|
| en | 61.6 | 52.0 | $-9.6$ | 0.0018 | $-1.6$ | $-8.0$ | $[-13.5, -2.8]$ |
| es | 60.4 | 61.2 | $+0.8$ | 0.81 | $+0.5$ | $+0.3$ | $[-2.9, +3.4]$ |
| fr | 52.0 | 57.2 | $+5.2$ | 0.11 | $+0.3$ | $+4.9$ | $[-0.7, +10.5]$ |
| de | 58.8 | 60.8 | $+2.0$ | 0.5 | $-0.4$ | $+2.4$ | $[-1.7, +6.5]$ |
| ru | 64.8 | 67.2 | $+2.4$ | 0.46 | $-0.1$ | $+2.5$ | $[-2.4, +7.5]$ |
| zh | 63.6 | 64.8 | $+1.2$ | 0.78 | $-2.4$ | $+3.6$ | $[-1.3, +8.5]$ |
| ja | 57.6 | 28.8 | $-28.8$ | $1.3 \times 10^{-15}$ | $+0.9$ | $-29.7$ | $[-35.7, -23.8]$ |
| th | 62.0 | 62.0 | $+0.0$ | 1 | $-0.2$ | $+0.2$ | $[-5.1, +5.6]$ |
| sw | 38.4 | 43.6 | $+5.2$ | 0.12 | $-0.2$ | $+5.4$ | $[-0.2, +10.9]$ |
| bn | 61.6 | 56.4 | $-5.2$ | 0.15 | $-0.1$ | $-5.1$ | $[-11.3, +1.0]$ |
| te | 58.8 | 58.0 | $-0.8$ | 0.88 | $-1.4$ | $+0.6$ | $[-4.4, +5.4]$ |
| pooled over eleven languages | | | | | | | $-2.1$ $[-3.6, -0.5]$ |

**Table 16:** MGSM accuracy (%) at Gemma 3 4B layer 17, $k$=20, SAE-native removal at all positions, with features selected on FLORES sentences 300–549 per language (Appendix B.8) in place of the MGSM prompts; 250 problems per language, four random-feature controls, paired at the problem level, bootstrap intervals over problems.

| | Repaired projection, Gemma 3 4B L17 | | | | Zhao construction, Qwen-2.5-3B | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | $\lambda$=0.3 | | $\lambda$=0.4 | |
| Lang | base | repaired | $-$ control | 95% CI | $-$ control | 95% CI | $-$ control | 95% CI |
| en | 58.4 | 58.4 | $-3.3$ | $[-8.0, +1.5]$ | $-6.8$ | $[-12.4, -1.2]$ | $-9.6$ | $[-14.8, -4.3]$ |
| es | 60.0 | 65.6 | $+1.5$ | $[-3.5, +6.3]$ | $-5.5$ | $[-11.1, +0.4]$ | $-8.7$ | $[-14.8, -2.7]$ |
| fr | 52.4 | 56.0 | $+0.5$ | $[-4.5, +5.6]$ | $-5.3$ | $[-11.1, +0.4]$ | $-6.5$ | $[-12.7, -0.3]$ |
| de | 58.0 | 62.4 | $+3.9$ | $[-0.4, +8.1]$ | $-4.5$ | $[-10.4, +1.3]$ | $-13.3$ | $[-19.6, -7.1]$ |
| ru | 63.6 | 61.2 | $+0.3$ | $[-3.7, +4.4]$ | $-14.0$ | $[-20.3, -7.9]$ | $-15.2$ | $[-21.7, -8.7]$ |
| zh | 61.6 | 48.8 | $-13.2$ | $[-18.9, -7.6]$ | $-10.9$ | $[-16.3, -5.7]$ | $-20.9$ | $[-27.1, -14.8]$ |
| ja | 58.8 | 57.6 | $+0.9$ | $[-4.0, +6.0]$ | $-12.9$ | $[-18.5, -7.2]$ | $-16.9$ | $[-22.3, -11.3]$ |
| th | 61.2 | 57.6 | $-2.5$ | $[-7.7, +2.5]$ | $-8.4$ | $[-14.8, -2.1]$ | $-14.8$ | $[-20.8, -8.7]$ |
| sw | 38.8 | 40.0 | $+0.7$ | $[-4.5, +5.9]$ | $+1.1$ | $[-1.9, +4.0]$ | $-1.7$ | $[-4.3, +0.8]$ |
| bn | 60.4 | 57.2 | $-3.6$ | $[-8.1, +0.8]$ | $-13.7$ | $[-18.1, -9.5]$ | $-14.1$ | $[-18.7, -9.6]$ |
| te | 58.8 | 59.2 | $+3.1$ | $[-1.7, +8.0]$ | $+0.8$ | $[-1.7, +3.5]$ | $+2.8$ | $[-0.1, +6.1]$ |
| pooled | | | $-1.1$ | $[-2.5, +0.4]$ | $-7.3$ | $[-8.9, -5.7]$ | $-10.8$ | $[-12.5, -9.2]$ |

**Table 17:** MGSM accuracy (%). Left: the repaired projection of §4.1 at Gemma layer 17, against four random-feature draws. Right: the mean-difference construction of Zhao et al. (2025) projected out of the prompt tokens at the mid-network layers of Qwen-2.5-3B-Instruct, against three random rank-10 subspaces at the same layers and strength. No language improves significantly in any column.

| Site | operator | rank | edit share | mean $\Delta_T$ | mean $\Delta_O$ | mean $C$ ($n$) |
|---|---|---|---|---|---|---|
| Gemma L17 | LEACE, one language vs. rest | 1 | 0.002 [0.000, 0.007] | $+2.38$ | $+0.00$ | $+0.988$ (11) |
| Gemma L17 | orthogonal proj., same directions | 1 | 0.616 [0.003, 0.958] | $+20.19$ | $+24.30$ | $-0.152$ (11) |
| Gemma L17 | LEACE, eleven-way label | 10 | 0.017 | mean $\Delta$CE $+2.03$ over all eleven | | |
| Gemma L17 | orthogonal proj., same span | 10 | 0.963 | mean $\Delta$CE $+22.65$ over all eleven | | |
| Llama L19 | LEACE, one language vs. rest | 1 | 0.015 [0.006, 0.039] | $+1.32$ | $+0.00$ | $+0.986$ (11) |
| Llama L19 | orthogonal proj., same directions | 1 | 0.018 [0.009, 0.030] | $+6.05$ | $+3.50$ | $+0.306$ (11) |
| Llama L19 | LEACE, eleven-way label | 10 | 0.148 | mean $\Delta$CE $+1.27$ over all eleven | | |
| Llama L19 | orthogonal proj., same span | 10 | 0.116 | mean $\Delta$CE $+9.83$ over all eleven | | |

**Table 18:** Covariance-aware erasure at the two sites. LEACE (Belrose et al., 2023) is fitted on FLORES sentences 300–549 of every language, for the concept "the language is $L$" (rank 1, one eraser per language) and for the eleven-way language label (rank 10), and scored on the 200 sentences of Table 1. The orthogonal rows apply equation (1) to the same cross-covariance directions, which for a single language is the difference between its mean and the mean of the rest. Edit share is the mean fraction of a token's squared norm the edit removes, $\mathbb{E}_t[\|h_t - h'_t\|^2/\|h_t\|^2]$, with its range over languages; for an orthogonal projection it equals $\rho$. Means are over all eleven languages; no cell is inert.

| | Gemma 3 4B, layer 17 | | | | | Llama 3.1 8B, layer 19 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Lang | base | sel. | − control | 95% CI | in lang. | base | sel. | − control | 95% CI | in lang. |
| en | 61.6 | 7.6 | −54.5 | [−60.5, −48.6] | 100 | 88.8 | 88.4 | +0.1 | [−1.7, +2.1] | 100 |
| es | 60.4 | 60.0 | −3.4 | [−6.2, −0.6] | 100 | 71.6 | 69.6 | −1.3 | [−4.5, +1.9] | 100 |
| fr | 52.0 | 56.0 | +1.3 | [−4.0, +6.8] | 16 | 69.2 | 68.4 | +0.5 | [−2.7, +3.7] | 100 |
| de | 58.8 | 60.4 | +1.6 | [−1.6, +4.9] | 100 | 68.4 | 68.8 | −0.0 | [−3.3, +3.3] | 99 |
| ru | 64.8 | 63.2 | −0.6 | [−3.3, +1.9] | 100 | 67.6 | 67.2 | −1.3 | [−4.3, +1.5] | 100 |
| zh | 63.6 | 54.0 | −8.7 | [−14.5, −3.0] | 4 | 65.2 | 66.0 | +0.8 | [−2.7, +4.3] | 100 |
| ja | 57.6 | 59.2 | +1.6 | [−3.2, +6.4] | 75 | 54.4 | 57.2 | +2.5 | [−1.3, +6.5] | 100 |
| th | 62.0 | 62.8 | +0.9 | [−3.1, +4.9] | 73 | 59.6 | 61.2 | +1.6 | [−2.0, +5.3] | 100 |
| sw | 38.4 | 38.4 | +0.9 | [−3.9, +5.9] | 92 | 56.4 | 51.6 | −4.9 | [−10.7, +0.8] | 99 |
| bn | 61.6 | 59.6 | −0.7 | [−6.1, +4.4] | 50 | 46.0 | 45.2 | −2.1 | [−8.7, +4.4] | 32 |
| te | 58.8 | 58.4 | +0.4 | [−4.3, +5.3] | 100 | 36.4 | 44.0 | +8.3 | [+3.3, +13.2] | 100 |
| pooled | | −5.6 | [−7.0, −4.2] | | | | +0.4 | [−0.8, +1.6] | | |
| no English | | −0.7 | [−2.1, +0.8] | | | | +0.4 | [−0.9, +1.7] | | |

**Table 19:** MGSM accuracy (%) with SAE-native removal applied to the prompt tokens only (the prefill pass; every generated position is left alone), $k$=20, 250 problems per language, four random-feature controls on Gemma and three on Llama under the same position policy, paired bootstrap over problems. "In lang." is the percentage of the selected arm's outputs still written in the prompt's language.

| | | every position | | | | | prompt tokens only | | |
|---|---|---|---|---|---|---|---|---|---|
| Lang | base | erased | control mean | − control | 95% CI | in lang. | erased | Δ | 95% CI |
| en | 62.8 | 60.0 | −1.5 | −1.3 | [−6.8, +3.9] | 36 | 63.6 | +0.8 | [−3.2, +4.8] |
| es | 58.8 | 62.4 | +3.1 | +0.5 | [−4.1, +5.5] | 44 | 62.4 | +3.6 | [−0.4, +8.0] |
| fr | 54.0 | 54.8 | −2.0 | +2.8 | [−1.5, +7.2] | 72 | 51.6 | −2.4 | [−7.2, +2.0] |
| de | 57.6 | 61.2 | +0.3 | +3.3 | [−1.5, +8.1] | 67 | 63.2 | +5.6 | [+1.2, +10.0] |
| ru | 65.6 | 60.8 | −1.2 | −3.6 | [−8.3, +0.9] | 99 | 62.0 | −3.6 | [−7.2, −0.4] |
| zh | 62.0 | 64.4 | +0.1 | +2.3 | [−2.9, +7.6] | 94 | 66.4 | +4.4 | [−0.8, +9.6] |
| ja | 58.0 | 55.2 | +0.0 | −2.8 | [−8.5, +2.9] | 64 | 56.4 | −1.6 | [−6.0, +2.8] |
| th | 64.4 | 60.4 | −3.1 | −0.9 | [−5.7, +4.1] | 99 | 61.6 | −2.8 | [−7.6, +2.0] |
| sw | 36.0 | 33.6 | +1.2 | −3.6 | [−9.6, +2.4] | 13 | 36.8 | +0.8 | [−4.8, +6.0] |
| bn | 62.0 | 61.6 | +0.1 | −0.5 | [−5.5, +4.4] | 98 | 57.2 | −4.8 | [−10.0, +0.4] |
| te | 57.6 | 55.6 | +0.3 | −2.3 | [−7.5, +2.9] | 100 | 61.2 | +3.6 | [−1.2, +8.4] |
| pooled | | | | −0.6 | [−2.2, +1.1] | | | +0.3 | [−0.9, +1.6] |

**Table 20:** MGSM accuracy (%) at Gemma 3 4B layer 17 with each language's identity erased by LEACE, the eraser fitted on the translated MATH prompts of Zhao et al. (2025), which are disjoint from MGSM. Controls are the erasers of three other languages applied to the same problems; the prompt-only columns are against baseline. 250 problems per language, paired bootstrap over problems; "in lang." is the percentage of outputs still in the prompt's language.

| | | own layer | | | | every layer | | |
|---|---|---|---|---|---|---|---|---|
| | | ΔCE | | refusal, deflated (%) | | ΔCE | | refusal (%) |
| Layer | $\rho$ | original | deflated | harmful | harmless | original | deflated | deflated |
| 4 | 0.173 | 4.34 | 0.01 | 76 | 1 | 31.88 | 27.91 | 25 |
| 6 | 0.670 | 10.84 | 0.00 | 76 | 1 | 40.09 | 19.48 | 12 |
| 8 | 0.928 | 4.66 | 0.00 | 78 | 1 | 31.92 | 10.84 | 36 |
| 10 | 0.929 | 7.74 | 0.00 | 68 | 1 | 33.06 | 11.09 | 3 |
| 12 | 0.871 | 21.88 | 0.00 | 72 | 1 | 28.15 | 10.27 | 3 |
| 14 | 0.321 | 23.72 | 0.00 | 69 | 1 | 23.92 | 9.52 | 7 |
| 17 | 0.429 | 19.17 | 0.00 | 18 | 1 | 23.66 | 8.72 | 0 |
| 20 | 0.248 | 11.08 | 0.00 | 45 | 1 | 20.68 | 4.70 | 0 |
| 24 | 0.280 | 7.29 | 0.00 | 66 | 1 | 19.79 | 5.93 | 0 |
| 29 | 0.003 | 0.59 | 0.00 | 71 | 1 | 10.99 | 0.11 | 63 |

**Table 21:** Gemma 3 4B refusal directions with their overlap with the top four second-moment directions removed at each layer where they are projected out. Baseline refusal in this run is 75% on harmful and 1% on harmless prompts; original columns are Table 12. Cross-entropy in nats on FLORES-200.

| Lang | baseline | **selected** | $\text{ctrl}_0$ | $\text{ctrl}_1$ | $\text{ctrl}_2$ | switches? | repaired proj. |
|---|---|---|---|---|---|---|---|
| en | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | no | 100.0 |
| es | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | no | 97.7 |
| fr | 100.0 | **0.8** | 100.0 | 100.0 | 99.2 | **yes** | **21.1** |
| de | 100.0 | 99.2 | 99.2 | 100.0 | 100.0 | no | 99.2 |
| ru | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | no | 85.9 |
| zh | 100.0 | **0.8** | 100.0 | 100.0 | 100.0 | **yes** | **26.6** |
| ja | 100.0 | **62.5** | 100.0 | 100.0 | 100.0 | **yes** | **10.9** |
| th | 100.0 | **57.0** | 100.0 | 100.0 | 100.0 | **yes** | **66.4** |
| sw | 94.5 | **18.8** | 93.0 | 92.2 | 98.4 | **yes** | 82.8 |
| bn | 100.0 | **35.9** | 100.0 | 100.0 | 100.0 | **yes** | **69.5** |
| te | 100.0 | 99.2 | 100.0 | 100.0 | 100.0 | no | 96.9 |

**Table 22:** Percentage of answers still written in the prompt's language, layer 17, $k$=20, 128 problems per cell, eleven-language feature selection; $\text{ctrl}_0$–$\text{ctrl}_2$ are the three random-feature controls. A cell counts as switching if it falls more than 25 points below its own baseline. The last column is the selected set under the repaired projection of §4.1, from its own run (baseline 100.0, Swahili 95.3; no control falls more than 3.1 points below its baseline); bold marks a switch by the same rule.