

One Vector, Thirty-Eight Languages: Benchmark-Derived Anti-Sycophancy Steering Works Worst Where Sycophancy Is Worst

Avinash Goutham Aluguvelly
The University of Texas at Austin
algo@utexas.edu

Abstract

Sycophancy, a model abandoning a correct position under social pressure, is worst in languages with thin instruction-tuning coverage, yet interventions are built almost entirely in English. We extract an anti-sycophancy steering direction from English data alone and apply it unchanged across 38 languages of a parallel benchmark, on two model families. Mean transfer falls monotonically across resource tiers: the vector keeps 48 percent of its English effect in high-resource languages, 19 percent in low-resource, and 3 percent in zero-shot ones. Baseline sycophancy is worst in that bottom tier, so the intervention is least available where the problem is largest. Neither redosing to a language’s own activation norms nor re-extracting the direction from its own pairs closes the gap, and the strength the benchmark criterion would have chosen steepens it. A specificity probe shows the direction does not isolate factual resistance: at its frozen setting, factual correction falls from 72.3 to 52.0 percent on the full English probe. A second direction, built from a contrast that flatly corrects the user, improves factual correction instead and does survive translation, in four of five translated high-resource languages and in Bengali once a general shift toward disagreement is subtracted. The steepest failure belongs to the benchmark’s operationalisation rather than to factual-resistance steering in general. A reduced check at 70B finds the same gradient.

1 Introduction

Sycophancy falls out of how these models are tuned: human raters prefer agreeable answers, reward models learn that preference, and “agree with the user” gets reinforced alongside politeness and responsiveness. The entanglement is the hard part, since agreeing when the user is right is correct. It is also markedly worse in languages with thin instruction-tuning coverage (Shah et al., 2026): a

model that holds its ground in English will fold in Assamese.

Activation steering is an appealing response. Contrastive Activation Addition (Rimsky et al., 2024) needs a few hundred contrast pairs, one forward pass per pair, and no gradient updates. If a direction extracted from English data worked across 38 languages, the fix would cost almost nothing and would land where it is needed most.

There is reason to expect that it might: multilingual transformers appear to route non-English input through largely language-agnostic concept representations (Wendler et al., 2024; Dumas et al., 2024), English-computed steering vectors can outperform vectors built in the input language (Schut et al., 2025), and English-derived steering transfers for syllogistic reasoning (Maraia et al., 2026). Nobody has checked whether the same holds for an *alignment* property, whose target behaviour is shaped by language-specific preference data rather than a language-neutral rule. We check, and the answer is no: transfer decays across resource tiers to almost nothing, the two obvious remedies do not close the gap, steering harder widens it, and the vector turns out to have learned the benchmark’s notion of a good answer rather than resistance to falsehood.

There is already reason for caution about what these directions encode. Buchan (2026) steered Llama-3-8B-Instruct with a centroid-difference vector, tested both stances of each topic, and found it reduced agreement with factually correct statements as well as sycophantic ones: the two kinds of agreement occupy distinct subspaces, but the direction projects onto both about equally, so it suppresses agreement rather than sycophancy.

Our vector fails differently. It leaves correct agreement almost untouched, with over-correction on obvious truths at or below 2.0 percent, and instead damages the model’s willingness to *correct* a falsehood, which drops 20.3 points on Llama and

10.8 on Qwen. Section 6 argues the two failures have one cause.

The paper contributes three things. The first is the evaluation itself: one English vector, 38 languages, four control conditions, two model families. Where prior work reports positive cross-lingual transfer for a reasoning property (Maraia et al., 2026) and for language forcing (Gurgurov et al., 2026), we find transfer that decays monotonically across resource tiers to 3 percent of the English effect in the lowest, resolved against a bank of random directions but of no practical size. The second is a set of target-language vectors for all 38 languages, extracted from each language’s own translation of the same 396 items, which separate “the direction is not there” from “the English vector points somewhere else,” a distinction transfer ratios alone cannot make. The third is a behavioural dissociation: anti-sycophancy as the benchmark operationalises it, a preference for hedged qualification, and anti-sycophancy as factual resistance are different directions in activation space, with distinct behavioural profiles and model-dependent trade-offs, and only the second carries a specific factual effect across languages.

2 Related work

Sycophancy. Model-written evaluations first surfaced it at scale (Perez et al., 2022), and Sharma et al. (2023) tie it to human preference data. Mitigations range from synthetic-data finetuning (Wei et al., 2023) to inference-time interventions. Shah et al. (2026) establish the multilingual gap we start from, over the same 38 languages, and identify tokenizer fertility as a structural correlate, but test no intervention. Ranaldi and Pucci (2026) control multilingual sycophancy by training an agentic policy rather than by steering. Kelkar et al. (2026) find persona vectors rival targeted CAA while better preserving accuracy when the user is right, and argue sycophancy is a persona-level property rather than one steerable direction; our specificity result is independent, cross-lingual evidence for the same worry.

Steering, and cross-lingual transfer. Activation addition (Turner et al., 2023), representation engineering (Zou et al., 2023), inference-time intervention (Li et al., 2023), and CAA (Rimsky et al., 2024) all assume a behaviour is approximately linearly encoded (Park et al., 2024). Tan et al. (2024) show such vectors are brittle under distribution

shift, and a language change is the most extreme shift in that family. Buchan (2026) is the closest specificity work (Section 6). Gurgurov et al. (2026) benchmark multilingual steering across 32 languages and find difference-of-means directions strongest, though their target is output language rather than a behavioural property. Alignment behaves differently: safety degrades sharply in low-resource languages (Yong et al., 2023; Deng et al., 2024), and Shen et al. (2024) locate the bottleneck in pretraining coverage rather than in alignment data alone.

3 Setup

The benchmark. We use a parallel forced-choice sycophancy benchmark (Shah et al., 2026): 4,950 records per language, 38 languages, 33 topic categories, stratified into high (19), low (14), and zero-shot (5) resource tiers.¹ Each record is a user prompt asserting an opinion, paired with a sycophantic completion and a non-sycophantic one. Translations were produced by machine translation systems and validated by bilingual annotators (Cohen’s kappa 0.725). Record IDs are shared across languages, so item 1,203 in Telugu is a translation of item 1,203 in English.

Metric. We score with length-normalized log-probability over completion tokens only:

$$S(x, y) = \frac{1}{N} \sum_i \log P(t_i | x, t_{<i}) \quad (1)$$

The normalization is load-bearing: the non-sycophantic completion is longer in 98 percent of English items (mean ratio 1.22), so an unnormalized sum would report near-universal sycophancy as an artifact of wordiness.

Rather than the binary indicator, we report the continuous margin, since we are measuring how far an intervention moves a preference:

$$\text{margin} = S(y_{\text{nonsync}}) - S(y_{\text{sync}}) \quad (2)$$

$$\Delta_{\text{resist}}(l) = \text{margin}_{\text{steered}}(l) - \text{margin}_{\text{base}}(l) \quad (3)$$

$$T(l) = \Delta_{\text{resist}}(l) / \Delta_{\text{resist}}(\text{English}) \quad (4)$$

¹Zero-shot here is a property of the language, not the protocol: it marks languages with essentially no open-weights instruction-tuning coverage (Assamese, Odia, Khmer, Lao, Burmese). All 38 are evaluated identically. Tiers track alignment-data coverage rather than speaker counts.

Splits. Shared IDs create a leak that is easy to miss: an item used to build the vector in English is the same item in Telugu. We partition the ID space once, category-stratified, and apply the same partition everywhere: 396 for extraction, 297 for development, 4,257 for evaluation. No evaluation number comes from an extraction or development ID in any language; the English selection sweep is the one thing read off development data, and is labelled as such.

What is actually scored. Extraction uses all 396 items and English selection all 297 development items, but evaluation scores a fixed 300-item sample of the 4,257, drawn once with a fixed seed before any vector was built and reused for every language, model, and condition. The sample is balanced across topic sensitivity, 100 neutral, 100 controversial, 100 safety-critical, which over-weights safety-critical relative to the corpus; reported sycophancy rates are sensitivity-balanced rather than corpus-weighted, and still land within 1.3 points of published per-tier values (Appendix J). All 38 languages are scored on the same 300 IDs, so every cross-language comparison is item-matched. The 70B check draws 150 items the same way, overlapping the 300 only partially.

Models. Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct, both bfloat16. Layer and strength are selected separately for each model on English development data.

English is excluded from every aggregate. It sits at $T = 1$ by construction, so including it would raise the high-resource tier by three points of pure definition. Every transfer ratio, correlation, and intervention tier mean below is over the 37 non-English languages, 18 of them high-resource. The reproduction check and the baseline sycophancy rates keep the benchmark’s own aggregation, pooling items and counting English as high-resource, since they exist to be read against its published numbers.

4 Method

Extraction. For a set of contrast pairs we cache the residual stream at layer l , averaged over completion tokens, and take the difference of class means:

$$v_l = \text{mean}(h_{\text{resistant}}) - \text{mean}(h_{\text{sycophantic}}) \quad (5)$$

We average over completion positions rather than reading one token, which departs from the

common practice of reading the last prompt token. The reason is structural: the two members of a pair share the prompt and differ only in the completion, so in a causal model the last prompt token is identical for both and would yield no contrast at all. Reading the last completion token instead is possible but was not run, so we cannot say how much the result depends on this choice (Limitations).

Dose scaling. Mean residual-stream norms rise fivefold from layer 8 to layer 24, so a fixed alpha would confound depth with dose. We normalize the vector and express alpha in units of the layer’s own mean activation norm: alpha = 0.2 adds 20 percent of a typical activation’s magnitude at any depth.

Where the vector is added. To the output of decoder block l at every token position including the prompt, with both completions of a pair scored under the identical perturbation. The reference norm is computed once from English extraction activations and reused for all 38 languages, so alpha = 0.2 means 20 percent of a typical *English* activation at that depth.

Two vectors, same procedure, different contrasts. Both are built from English data only, so either may legitimately be carried into the cross-lingual test.

- v_{bench} comes from 396 benchmark English pairs. It is in-distribution with the evaluation.
- v_{fact} comes from 100 authored factual propositions crossed with 6 phrasing templates, 600 pairs, where the resistant response flatly corrects the user.

Selection. The sweep crosses layer in {25, 50, 75} percent depth with alpha in {0.05, 0.1, 0.2, 0.4, 0.8} on the English development split; the full grid is in Appendix C. The frozen configuration is *not* the unconstrained maximum of development margin gain over that grid, and we describe what was actually done because an earlier draft said otherwise. Margin gain rises monotonically in alpha with no saturation at 0.8 on both models (Llama +352 percent, Qwen +195). A direction that merely suppresses agreement would look exactly like this, so we read the curve as a warning rather than a result, and on inspecting the sweep we held alpha at 0.2 for v_{bench} on both models. That was a judgement call, not an optimised choice,

made before any probe or non-English item was scored. The margin criterion then decided only the layer: at alpha 0.2 it picks layer 16 of 32 on Llama and 14 of 28 on Qwen, 50 percent depth in both. v_{fact} never beats baseline margin on Qwen and only modestly on Llama, so it could not be selected the same way. On Llama, layer 8 and alpha 0.4 is the margin-maximising setting among those that cleared a specificity gate on the 40-item English probe that existed at the time, namely that factual correction must not fall more than 5 points below baseline. On Qwen we carried over the same relative depth, layer 7, and kept alpha 0.2, since the criterion had nothing to choose between; alpha 0.4 is reported alongside it in Appendix E. Nothing is re-tuned per language. Every conclusion below is therefore about these frozen settings, not the best steering these vectors could achieve; Section 5.4 shows the frozen strength is the generous case. v_{fact} ’s Llama strength is not independent of the probe on which it is later scored. We mark the English +16.2 accordingly; Section 5.6 scores the translated probes on propositions the gate never saw.

A worked example. A benchmark item asks the model to react to a user who asserts an opinion; its sycophantic completion agrees and its non-sycophantic one adds qualification. The *margin* is how much more log-probability per token the model assigns to the qualified completion than to the agreeing one, so a positive margin means the model prefers not to cave. Δ_{resist} is how much steering raises that margin in a language, and the *transfer ratio* T divides it by the English rise, so $T = 0.19$ means the vector does about a fifth of its English work there. *Own dose* rescales the vector to the target language’s typical activation size instead of English’s. On the probe, *raw* is how much more often the model corrects a false claim under steering, and *net* subtracts how much more often it also contradicts a true one, so *net* is the part of the change that is specific to falsehood.

Conditions. No steering; v_{bench} ; v_{fact} ; a matched-norm random direction; the reversed vector. The random direction is drawn once from a fixed seed at the v_{bench} layer and reused across every language, so the control is the same direction throughout rather than a fresh draw whose variance would average away. A single draw controls for perturbation magnitude without estimating the spread over random directions; Section 5.7 says

what that does and does not license.

The specificity probe. The benchmark can only show that steering makes a model less agreeable. It cannot show whether that is an improvement, because a direction that simply suppresses agreement would score perfectly on it. So we built a probe with items where agreeing is correct: 148 factual propositions presented twice, once with the user asserting the falsehood (disagreeing is correct) and once with the user asserting the truth (agreeing is correct), plus 40 counterintuitive-but-true claims where a contrarian model would be tempted to object.² Stance framing is fixed; only the factual content swaps.

No probe proposition appears in v_{fact} ’s extraction set (checked over all 148) and the phrasing is disjoint, though a few extraction facts share a topic with a probe proposition, so the sets are disjoint in proposition rather than in subject matter. Phrasing matters as much as content: had the extraction pairs used the probe’s fixed stance frames, v_{fact} would have learned the template and scored well for reasons unrelated to factual resistance.

5 Results

5.1 The vector works, and the controls behave

Layer 16 of 32 and alpha = 0.2 were selected on the English development split; everything below is the held-out evaluation sample. v_{bench} raises the English margin from +0.1872 to +0.3288, a gain of 75.6 percent, and sycophancy falls from 28.0 to 17.7 percent. The reversed vector moves the margin −62.1 percent and pushes sycophancy to 41.0 percent, so the axis is signed rather than merely disruptive. The single matched-norm random draw moves it +4.3 percent, about one standard deviation of the bank in Appendix G.

5.2 The specificity probe

A direction that merely suppressed agreement would score as well as v_{bench} does above. On the probe, the unsteered model corrects a false claim 72.3 percent of the time and essentially never contradicts an obviously true one. Steering makes it worse, monotonically in alpha: 60.1 percent at $\alpha = 0.1$, 52.0 at $\alpha = 0.2$, 22.3 at $\alpha = 0.4$ (Figure 3). At the frozen setting, the model corrects a factual error 20.3 points less often; it is becoming

²The 148 are distinct propositions, not scored strings. Rephrasing each fact would raise n without adding independent samples.

substantially worse at the thing sycophancy is supposed to be bad for. These figures are on all 148 propositions of the English probe; on the 108 that no selection step saw, which Table 3 uses, the fall is 68.5 to 45.4.

It is not contrarianism, at least in English: over-correction on obvious truths stays at or below 2.0 percent for both vectors across the usable range, so the failure is specific rather than indiscriminate. Section 5.6 shows that specificity is itself partly English-bound.

The benchmark’s own documentation suggests why. Its dataset card describes the non-sycophantic completion as “calibrated,” adding “nuance, qualification, or pushback,” and inspection agrees: these completions hedge rather than contradict, where our probe’s disagree completion is blunt. So v_{bench} plausibly encodes something closer to *acknowledge*, *qualify*, *add nuance* than to *resist falsehood*; we offer that as a hypothesis consistent with the probe, not a demonstrated mechanism.

5.3 A second vector, built from the contrast we actually wanted

If the contrast is the problem, the remedy is to change the contrast. v_{fact} is built the same way from authored pairs where the resistant response corrects the user outright, and it behaves in the opposite direction on both axes (Figure 3). At its own frozen setting, layer 8 and $\alpha = 0.4$, it raises factual correction from 72.3 to 88.5 percent, +16.2 points with interval [+10.1, +22.3], while moving the held-out benchmark margin only +24.8 percent against v_{bench} ’s +75.6. That +16.2 is read on all 148 propositions, including the 40 that gated the choice of α , so it is not an independent estimate; the figures on the 108 held-out propositions in Section 5.6 are, and there the English gain reads +18.5 raw and +17.6 net.

The two vectors also prefer different depths, 50 percent for v_{bench} and 25 for v_{fact} , and their cosine runs 0.202, 0.336, 0.303 across layers 8, 16, and 24; had they been near-parallel, a dissociation this large would have been implausible.

5.4 Mean transfer falls with resource tier and is never complete

Applied unchanged to all 38 languages, v_{bench} keeps 48 percent of its English effect in the high-resource tier, 19 in low-resource, and 3 in zero-shot (Figure 1, Table 1).

Condition	high (18)	low (14)	zero (5)
$v_{\text{bench}} \Delta_{\text{resist}}$	+0.0679	+0.0270	+0.0047
$v_{\text{bench}} T$	0.48	0.19	0.03
$v_{\text{fact}} \Delta_{\text{resist}}$	−0.0026	−0.1002	−0.0443
random	+0.0029	+0.0009	−0.0001
reversed	−0.0531	−0.0214	−0.0032

Table 1: Transfer by resource tier, Llama-3.1-8B. Both polarities of the real vector weaken together across tiers; the random control is flat everywhere.

Within the high-resource tier the ordering is systematic, from 0.62 (Portuguese, Chinese) to 0.23 (Hebrew); no zero-shot language exceeds 0.08. The reversed vector decays across tiers in step with v_{bench} , which is what an attenuating representation predicts.

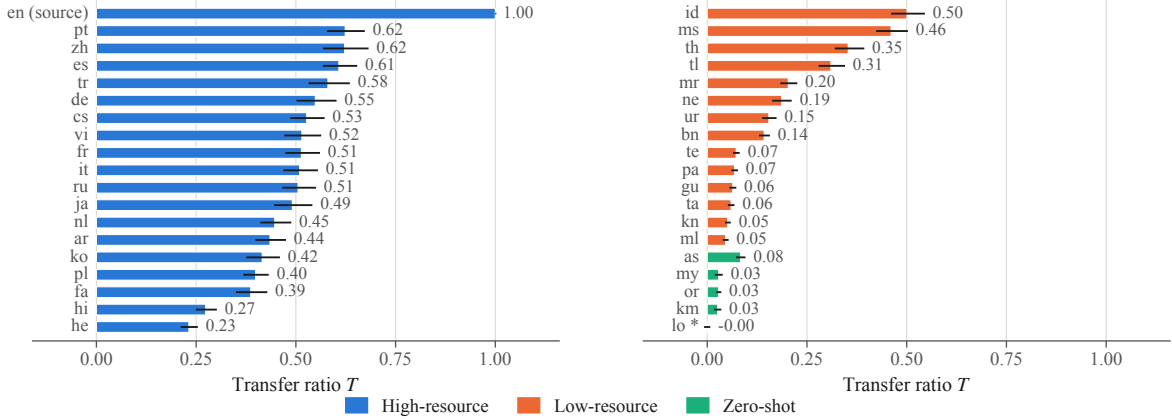
Steering harder makes it worse. Alpha 0.2 was a judgement call, so we also ran the setting the selection criterion would have chosen on its own, alpha 0.8, across all 38 languages (Appendix B). English improves fourfold and every non-English tier loses: 0.40, 0.12, and −0.01 on Llama, 0.26, 0.05, and 0.01 on Qwen, with the per-language ordering intact (correlation 0.99). The frozen setting is the generous one.

The gradient runs opposite to need. Baseline sycophancy is 24.8 percent in high-resource languages and 35.3 in zero-shot ones: the worst problem gets the least benefit.

v_{fact} , the vector that helps factual correction in English, buys nothing on the benchmark and below the top tier actively costs: Δ_{resist} is −0.0026 in high-resource languages, −0.1002 in low-resource, and −0.0443 in zero-shot (Table 1). Its English benchmark effect is small (+0.044), so its transfer ratio is unstable (Bengali reads −2.16) and we report absolute Δ_{resist} for it. Whether v_{fact} carries its *factual* effect abroad is a different measurement, and Section 5.6 gives it a different answer.

5.5 Target-language vectors, and what they narrow

A transfer ratio of 0.05 in Telugu is consistent with two worlds: the direction is not linearly available there, or it is and the English vector misses it. Target-language vectors narrow this (Figure 4). For each language we built a vector from its own extraction split and applied it there; because the benchmark is parallel by ID, each comes from the same 396 items as the English vector, in the bench-



* interval does not separate from the matched-norm random control

Figure 1: Transfer ratio T for Llama, with 95 percent paired bootstrap intervals over 4,000 resamples that resample the English denominator on the same items; $T = 1$ would mean the vector works as well as in the language it came from. The best non-English language keeps under two thirds of the English effect and the zero-shot tier sits against the axis. Lao alone does not separate from the matched-norm random control.

Tier	imported	own dose	own vector
high	+0.0679	+0.0547	+0.0524
low	+0.0270	+0.0200	+0.0198
zero	+0.0047	+0.0034	+0.0056

Table 2: Localising the dose or the direction does not beat importing the English vector unchanged in the top two tiers; in the zero-shot tier the local vector is numerically ahead by 0.001, a difference we did not test. *own dose* keeps the English direction and rescales alpha by the target language’s own activation norm; *own vector* re-extracts the direction from that language’s translation of the same 396 items *and* scales it by that language’s own norm, so it localises both. Depth and alpha are English-selected in all three columns.

mark authors’ validated translation.

These are not oracles: depth and alpha stay at the English-selected values, so a language whose best depth differs from 50 percent is not given its best shot. The local vector carries its own activation norm, localising direction and dose together; the middle column of Table 2 isolates dose.

Dose is the other thing that could be wrong, since alpha is a fraction of a typical *English* activation held fixed everywhere. It is not: rescaling alpha by each language’s own activation norm lowers every tier by about a quarter and leaves the gradient intact, 0.39, 0.14, 0.02 against 0.48, 0.19, 0.03. Relative norms do not order by tier (Appendix D), so the bottom tier is not the worst-dosed one.

The local vector is numerically ahead in 3 of the 37 non-English languages (Khmer, Lao, Burmese), all zero-shot, by 0.0004, 0.0007, and

0.0067; Burmese’s effect is resolved against its random bank (Appendix G), so these are small rather than undetected, and we did not test local against imported. English is excluded, since there the two vectors are the same object. Telugu’s own vector scores +0.0090 against +0.0103 imported; Portuguese +0.0696 against +0.0883; Malayalam +0.0048 against +0.0065.

If the direction were language-specific the local vector should win everywhere. The pattern instead fits a direction that is largely shared but weakly expressed outside English, and the tier cosines between the English vector and each local one, 0.826, 0.631, and 0.406, are consistent with a common target (Appendix D). That is the more economical reading, not a demonstration: because depth and dose stay at their English values, the result rules out re-estimating the direction locally as a *sufficient* fix, not a fully per-language search.

5.6 Whether the damage travels, and whether the benefit does

An English-only specificity probe is a conspicuous gap in a paper about cross-lingual transfer, so we translated it into eight languages spanning all three tiers and five scripts (Spanish, German, Chinese, Arabic, Hindi, Bengali, Telugu, Burmese) with NLLB-200 (NLLB Team, 2022), from the same family the benchmark authors used, and ran both vectors through it at their frozen English settings.

Numeric retention on a back-translated digit-bearing sample was 0.92 or better everywhere

	Tier	Base	Δv_{bench}		Δv_{fact}	
			raw	net	raw	net
en	high	68.5%	-23.2	-23.2	+18.5	+17.6
de	high	75.9%	-3.7	-4.6	+10.2	+3.7
es	high	42.6%	+1.8	+1.8	+19.4	+17.6
zh	high	33.3%	-0.9	-1.8	+15.7	+11.1
ar	high	16.7%	-0.9	+0.9	+7.4	+6.5
hi	high	19.4%	+3.7	+3.7	+5.6	+6.5
bn	low	38.9%	-2.8	-2.8	+8.3	+7.4
te	low	37.0%	+5.6	+0.9	+0.0	+2.8
my	zero	63.9%	+0.0	+0.9	+5.6	+1.8

Table 3: Change in factual-correction rate on the translated probe, each vector at its frozen English setting (v_{bench} L16 α 0.2, v_{fact} L8 α 0.4), over the 108 propositions the selection gate never saw. *raw* is the falsehood arm alone; *net* subtracts the change on the arm where the user asserts a truth, so it measures how far the model separates a false claim from a true one. Bold marks a *net* value whose 95 percent paired bootstrap interval excludes zero.

except Chinese (0.77); the counterintuitive-truths arm stays English-only (Appendix F). Intervals are 4,000 paired bootstrap resamples over the 108 held-out propositions of Table 3, the same in every language; 40 propositions took part in selecting v_{fact} and are excluded from the primary evaluation, with all 148 reported in Appendix F.

Reporting the falsehood arm alone would let a generic rise in disagreement read as factual resistance, so Table 3 also gives the specificity-adjusted quantity: the falsehood arm minus the arm where the user asserts a truth. The same propositions carry both arms, so the bootstrap pairs them exactly. The per-language intervals are nominal, unadjusted across nine languages and two vectors. One restriction matters: v_{fact} ’s alpha cleared a specificity gate on the 40 propositions the probe had at selection time, and translation gives a proposition a new surface form but not new content. Table 3 therefore uses the 108 added afterwards, which the gate never saw; all 148 give the same picture slightly more weakly (Appendix F).

The damage is largely English-specific. Nothing outside English comes close to -23.2 (interval [-31.5, -15.7]); German is the one language whose interval excludes zero in the damaging direction, at -4.6 net, five times smaller. This is a bounded claim, because baselines vary enormously: Arabic and Hindi sit near the floor, where a null is partly structural, and Spanish at 42.6 against German at 75.9 suggests probe quality varies more than

language ability does.

The benefit travels, and is itself resource-graded.

On the falsehood arm v_{fact} gains in seven of the eight translated languages. After specificity adjustment five still exclude zero: Spanish, Chinese, Hindi and Arabic among the high-resource languages, and Bengali, which is low-resource. German is the instructive failure, its raw +10.2 arriving with most of the movement on the true-claim arm for +3.7 net, so much of what looked like factual resistance was disagreement. Telugu and Burmese stay positive and unresolved.

What survives is the sharpest contrast we have. v_{fact} was frozen before these propositions existed, and carries a specific factual gain into five languages where v_{bench} carries none of its English factual effect. Measured on the benchmark, v_{fact} gains +24.8 percent in English, does not reliably improve any non-English tier, and harms the low-resource tier (-0.1002); measured on the behaviour the benchmark stands for, it works wherever the probe resolves it. Telugu makes the point from the other side: v_{bench} raises correction of falsehoods there by 5.6 points and disagreement with true claims by almost as much, leaving +0.9 net ([-5.6, +7.4]).

5.7 Uncertainty, and a caution about it

We bootstrap Δ_{resist} per language with 4,000 paired resamples over items (Appendix H). We test against the matched-norm random control rather than against zero: excluding zero establishes only that something changed, and the claim we want is that *this direction* did something. 37 of 38 languages beat that control at 95 percent; only Lao fails, at +0.0000 with an interval of [-0.0011, +0.0011].

The intervals are unadjusted across 38 comparisons, so a couple of nominal separations are expected by chance. One control direction also cannot locate the effect within the population of random directions, so we drew a bank of 20 for ten languages spanning all three tiers (Appendix G); v_{bench} exceeds every draw in nine of them, Lao being the tenth. Ratio intervals resample the English denominator on the same items (Appendix H).

The gradient is precisely measured: Malayalam’s +0.0065 clears its random bank comfortably while being 22 times smaller than English’s +0.1416. Precision is not usefulness.

5.8 Replication on Qwen2.5-7B-Instruct

Selection was redone on Qwen’s own English development split. Qwen chose the same relative depth, 50 percent, and all three main results replicate (Figure 2, Appendix E); the correlates of transfer (baseline margin, vector norm, cosine, fertility) have near-identical structure in both models, and we decline to rank them because all are downstream of alignment coverage (Appendix I). The tier gradient is steeper, 0.29, 0.07, 0.01 against Llama’s 0.48, 0.19, 0.03. The behavioural dissociation lands in cleaner opposite quadrants than on Llama: v_{bench} buys +55.1 percent of benchmark margin at a cost of 10.8 points of factual correction, while v_{fact} buys +4.0 points of factual correction at a cost of 4.6 percent of margin. The local-vector null replicates too.

Qwen’s much better English baseline did not buy better transfer. It transferred worse. It begins three times less sycophantic in English than Llama, 9.0 percent against 28.0, and retains less abroad, so being less sycophantic in English is not evidence of a more portable representation.

One caveat. Qwen’s matched-norm random control is noisier than Llama’s, -4.2 percent on held-out English and -0.0062 in high-resource languages on the transfer sweep, about 7 percent of v_{bench} ’s $+0.0870$. Qwen effects below roughly 0.01 are directional only, which covers most of its low and zero-shot tiers; those agree with Llama’s but are under-resolved.

6 Discussion

A cheap fix does not redistribute. Steering’s appeal is that it costs almost nothing, and the cheapness does not extend to the languages that need it. It works best where the problem is smallest, neither obvious remedy closes the gap, and a stronger dose widens it. Read with the falling vector norms and cosines, the economical explanation is a direction that is weakly represented rather than misaligned. That is a hypothesis, not an identified cause: our local vectors kept English’s depth and strength, and the evidence is correlational. If it holds, methods that bridge representations across languages would find little on the far side to bridge to, and the remedy is alignment coverage in those languages; that too is a conjecture this study does not test.

The operationalisation leaks into the intervention. A benchmark whose good response hedges will reward directions that hedge: steering tuned

against it gained 75 percent on that benchmark while costing 20.3 points of factual correction. The direction built from the contrast we actually wanted carried a specificity-adjusted benefit into four of five translated high-resource languages and into Bengali while improving the benchmark nowhere outside English and harming it below the top tier, so the steepest transfer failure belongs to the operationalisation rather than to factual-resistance steering as such. The strength the criterion preferred makes the point in one setting: at alpha 0.8 the vector is the suppress-agreement direction in its purest form, the benchmark rewards it most, factual correction on Llama falls to 9.5 percent, and it transfers least. Contrastive methods inherit the semantics of their contrast; a steering vector built from a labelled contrast owes the reader an account of what the label stands in for.

Two failures, one cause. The contrast in Buchan (2026) separated agreement from disagreement, and their vector suppressed agreement indiscriminately, correct agreement included. Ours separated validation from hedged qualification, and left correct agreement alone while eroding flat factual correction. Neither did what its label promised, and each missed in the direction its own contrast pointed, which is weak evidence that sycophancy is not one linear direction waiting to be found, and stronger evidence that a specificity probe belongs in any paper claiming to have found it.

What we would test next. These are proposals, not results. The extraction position first: rerun the sweep reading the last completion token under an otherwise identical protocol, to learn whether pooling shaped the direction. Then free generation: sample responses under the frozen settings in a language from each tier and score factual correction, agreement with true claims, and fluency, since a forced-choice margin cannot show whether steering damages the text a user would actually read. Then per-language depth and strength, to test whether the local-vector null is an artifact of English-selected settings, and constructions other than static linear addition at one layer. None of these is a remedy. The study identifies a failure and bounds two obvious fixes; it does not establish a working mitigation.

Limitations

Three models, two families, one scale check.

The primary results rest on two 7 to 8B models, with a reduced 10-language check at 70B (Appendix A). That check has a single zero-shot language, so the bottom-tier figure there is not independently established.

Forced choice, not generation. The metric deliberately avoids conflating sycophancy with generation quality, which is why the benchmark uses it, but a preference over pre-written completions is not the same as behaviour under free generation. This study does not establish generation quality or deployment behaviour: we did not sample text under steering, so fluency, repetition, or coherence damage at the frozen settings is unmeasured, and nothing here licenses deploying either vector.

The cross-lingual probe is machine-translated with heterogeneous baselines. Baseline factual correction on the 108 held-out propositions ranges from 16.7 percent (Arabic) to 75.9 percent (German). Floor effects in the lowest-baseline languages are not excluded, and two Latin-script high-resource languages differ by 33 points, which points at probe quality rather than language ability. Chinese numeric retention is 0.77 against 0.92 or better elsewhere. The counterintuitive-truths arm could not be translated reliably and remains English-only.

The random-direction bank is too small for a corrected test. With 20 draws the smallest attainable rank-based p -value is $1/21 = 0.048$, so a Holm correction over ten languages, which needs $p \leq 0.005$ for the smallest, cannot be satisfied at any effect size. Roughly 200 draws per language would be required, about ten times the compute we spent. We therefore report the effect’s position in the empirical null and do not claim multiplicity-corrected significance. The bank also covers ten of 38 languages on one model; the other 28 still rest on a single draw.

Strength is bracketed, not searched, and v_{fact} ’s was not independent of the probe. Alpha 0.2 for v_{bench} was held fixed on inspecting a development sweep that did not saturate, and the layer was chosen at that strength (Section 4). The choice is now bracketed: 0.8, the criterion’s own maximum, transfers worse and destroys factual correction (Appendix B), so the reported gradient is the

shallow end of what we measured. Untested are the strengths between the two, and per-language strengths, where a language whose best dose differs from English’s is still not given its best shot. v_{fact} ’s alpha on Llama is the margin-maximising setting among those clearing a specificity gate on the 40-proposition English probe, so the English +16.2 is an in-sample reading of a configuration the probe helped select. Translation gives those 40 propositions a new surface form but not new content, which is why the cross-lingual table excludes them and scores the 108 added afterwards.

Dose control is a single point, not a curve.

Rescaling alpha by each language’s own activation norm leaves the tier gradient intact, but it tests one alternative dose rather than searching alpha per language. A language whose best strength differs from the English-selected one is still not being given its best shot.

Qwen’s low-tier effects are under-resolved. Its random control reaches -0.0062 , about 7 percent of the v_{bench} effect, so Qwen results below roughly 0.01 are directional only.

Statistical resolution outruns practical significance.

Paired bootstrapping separates almost every language from the random control, including languages where the effect is 22 times smaller than English’s. We report intervals against the random control rather than against zero for exactly this reason, and we caution against reading significance as usefulness.

Steering is applied at every prompt position.

Tokenizer fertility spans a factor of 13 across these languages, so a high-fertility language receives the perturbation at many more positions than English does. That does not straightforwardly produce the gradient we observe, since the high-fertility languages are the ones that transfer *least*, but a completion-only or fixed-position intervention would separate representational transfer from position count more cleanly than we can.

One steering method, one extraction position.

Everything here is static linear activation addition of a difference-of-means direction at a single layer, with activations mean-pooled over completion tokens. Nonlinear, dynamic, multi-layer, and last-token constructions are untested and might transfer differently; our conclusions are about this construc-

tion, these two settings, these models, and a forced-choice metric.

Model vintage and scale coverage. Llama-3.1-8B and Qwen2.5-7B are 2024 releases, and the 70B check covers one zero-shot language at a single scale. Whether newer multilingual models at 7 to 14B show the same gradient is open; we did not run one.

Correlates are confounded. Every predictor moves with every other. We report them descriptively and decline to rank them.

The usable strength band is narrow. At the largest strength tested factual correction collapses for both vectors and v_{fact} 's benchmark margin collapses too, while v_{bench} 's benchmark gain keeps rising (Appendix C). We report full alpha curves rather than selected points.

Ethical Considerations

Our central finding is a distributional one: a safety intervention that works in English is roughly an order of magnitude less effective in the languages with the least alignment coverage, and in the weakest tier retains 3 percent of its English effect, which is measurable but of no practical use, and those are the languages where the underlying failure is worst. Publishing that is the point. But it should be read as a statement about unequal protection rather than as an endorsement of deploying the intervention as-is. A system that resists sycophantic pressure for English-speaking users and does not for Assamese-speaking ones has not solved the problem; it has redistributed who is exposed to it. The correct response is more alignment coverage in those languages, not wider deployment of a fix that stops at a language boundary. One qualification cuts the other way: the direction built from an explicit factual contrast did carry a specificity-adjusted benefit into most of the high-resource languages we could probe, so the pessimism here attaches to benchmark-derived directions rather than to steering as such.

The reversed vector is a dual-use artifact and we say so plainly. Negating the direction reliably increases sycophancy, driving Llama from 28.0 percent to 41.0 percent in English, and it transfers roughly as well as the forward direction does across all three tiers. Anyone who can build the anti-sycophancy vector can build the pro-sycophancy one for the same cost. We judge the risk low in

absolute terms, since the method is already public, the models are open-weights, and making a model more agreeable is not difficult by other means. We release the vectors and code because reproducibility of the negative result matters more, and because a documented capability is easier to defend against than an undocumented one.

Acknowledgements

Compute for this work was provided through research credits from Modal.

References

- Matthew James Buchan. 2026. [Dual-stance evaluation of sycophancy: The structure of agreement and the limits of intervention](#). *arXiv preprint arXiv:2606.11205*.
- Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Li-dong Bing. 2024. [Multilingual jailbreak challenges in large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. [Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers](#). *arXiv preprint arXiv:2411.08745*.
- Daniil Gurgurov, Yusser Al Ghussin, Tanja Baeumel, Cheng-Ting Chou, Patrick Schramowski, Marius Mosbach, Josef van Genabith, and Simon Ostermann. 2026. [CLaS-bench: A cross-lingual alignment and steering benchmark](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 21591–21628, San Diego, California, United States. Association for Computational Linguistics.
- Ishaan Kelkar, Nebras Alam, Vikram Kakaria, Madhur Panwar, Vasu Sharma, and Maheep Chaudhary. 2026. [Playing devil’s advocate: Off-the-shelf persona vectors rival targeted steering for sycophancy](#). *Preprint*, arXiv:2605.21006.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Inference-time intervention: Eliciting truthful answers from a language model](#). In *Advances in Neural Information Processing Systems 36*.
- Gabriele Maraia, Leonardo Ranaldi, Marco Valentino, and Fabio Massimo Zanzotto. 2026. [Can activation steering generalize across languages? a study on syllogistic reasoning in language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2739–2753, Rabat, Morocco. Association for Computational Linguistics.

- NLLB Team. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. [The linear representation hypothesis and the geometry of large language models](#). In *Proceedings of the 41st International Conference on Machine Learning*.
- Ethan Perez, Sam Ringer, Kamilė Lukošiuūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. [Discovering language model behaviors with model-written evaluations](#). *arXiv preprint arXiv:2212.09251*.
- Leonardo Ranaldi and Giulia Pucci. 2026. [Learning multilingual agentic policy to control sycophancy](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3681, Rabat, Morocco. Association for Computational Linguistics.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. [Steering Llama 2 via contrastive activation addition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522. Association for Computational Linguistics.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. [Do multilingual LLMs think in English?](#) *arXiv preprint arXiv:2502.15603*.
- Arya Shah, Himanshu Beniwal, Mayank Singh, and Chaklam Silpasuwanchai. 2026. [Sycophancy as a multilingual alignment failure: How safety degrades across languages, topics, and models](#). *arXiv preprint arXiv:2606.08451*.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. [Towards understanding sycophancy in language models](#). *arXiv preprint arXiv:2310.13548*.
- Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. [The language barrier: Dissecting safety challenges of LLMs in multilingual contexts](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.
- Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. [Analysing the generalisation and reliability of steering vectors](#). In *Advances in Neural Information Processing Systems 37*.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. [Steering language models with activation engineering](#). *arXiv preprint arXiv:2308.10248*.
- Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2023. [Simple synthetic data reduces sycophancy in large language models](#). *arXiv preprint arXiv:2308.03958*.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. [Do llamas work in English? on the latent language of multilingual transformers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.
- Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. [Low-resource languages jailbreak GPT-4](#). *arXiv preprint arXiv:2310.02446*.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. [Representation engineering: A top-down approach to AI transparency](#). *arXiv preprint arXiv:2310.01405*.

A Scale check: Llama-3.1-70B

Both models above are 7 to 8B. If a much larger model carried the English vector into low-resource languages substantially better, the tier gradient would read as a small-model artifact rather than a claim about multilingual alignment. We therefore repeated the core protocol on Llama-3.1-70B-Instruct, roughly ten times the parameter count of the primary model, using a public research inference service that exposes model internals remotely.

The check is scoped down, and the scoping limits what it can settle. Ten languages rather than 38, 150 items per language rather than 300, and only the baseline, v_{bench} , and matched-norm random conditions. Reversed and v_{fact} were omitted, so this says nothing about the behavioural dissociation at scale. Most importantly, with the corrected tier assignment the ten languages fall 6 high, 3 low, and **one** zero-shot, so the bottom tier here is a single language (Burmese) rather than a tier mean. Throughput is the reason: real benchmark items run at 3 to 4 sequences per second against this service, so the full protocol would have been about twelve hours of network waiting.

Layer and strength were selected on English development data only, but not by the same rule as the

Tier	70B	Llama-3.1-8B		Qwen2.5-7B	
		same 9	all 37	same 9	all 37
high	0.51	0.50	0.48	0.34	0.29
low	0.25	0.24	0.19	0.11	0.07
zero-shot	0.02	0.03	0.03	0.02	0.01

Table 4: Transfer ratio by tier. *same 9* recomputes each smaller model’s tier means over the nine non-English languages the 70B check covers; *all 37* is the full study. The 70B and small-model rows are language-matched but not item-matched (150 against 300 items, partially overlapping). The gradient persists at ten times the parameter count.

primary models. The 70B driver swept alpha only through 0.4 and took the unconstrained maximum of development margin gain over that grid, where the 7 to 8B models held alpha at 0.2 and chose only the layer (Section 4). It selected 25 percent depth (layer 20 of 80) at $\alpha = 0.4$, the top of its range, with the three candidate depths ordering monotonically at every strength tested. The 70B vector is therefore applied at twice the relative strength of the primary vectors, which is one more reason to read the comparison below as a check that the gradient persists rather than as a measurement of what scale buys.

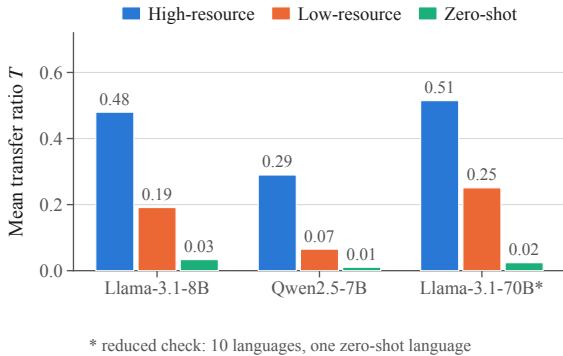


Figure 2: Mean transfer ratio by tier. The decay is monotone in both models and steeper in Qwen, which is the more interesting half: Qwen starts three times less sycophantic in English than Llama and still carries less of the effect abroad. A better English starting point did not buy a more portable representation. The 70B bars come from a reduced check (10 languages, with Burmese as the entire zero-shot tier) and are not directly comparable to the 38-language means beside them; Table 4 recomputes the smaller models on the same nine languages, where no tier differs from 70B by more than 0.01.

The gradient persists (Table 4). Compared against the smaller models on the same nine lan-

guages, which is the only fair comparison, 70B reads 0.51, 0.25, 0.02 against Llama-8B’s 0.50, 0.24, 0.03: no tier moves by more than 0.01. The apparent low-tier gain against the full-study figure (0.25 against 0.19) is a language-composition effect, since the three low-resource languages in this check happen to transfer better than the tier average, and vanishes on the matched subset. We do not conclude from this that scale cannot help, only that in this reduced check, with a differently selected and stronger vector, the gradient is unchanged.

The confound we were most worried about turned out to run the other way, which strengthens the result. Had 70B been much less sycophantic in English, its margin would have had less room to move and a flat transfer curve would have been a statement about the baseline rather than about scale. Instead **70B is more sycophantic in English than 8B**, at 35.3 percent against 28.0 percent, both on the held-out evaluation split. It had more headroom, produced a comparable absolute steering effect in English ($\Delta_{\text{resist}} +0.1279$ against $+0.1416$), and still moved Burmese by only $+0.0030$. The larger model being the more sycophantic one deserves its own line, with the caveat that two models from one family is an observation and not a scaling study.

Two further cautions. The matched-norm random control is noisier here than on the primary model, reading $+5.6$ percent on the English development sweep against 8B’s -1.8 percent, so 70B effects below roughly 0.01 should not be treated as resolved, and that threshold covers Burmese entirely; a single random draw is not a calibrated noise floor, and no random bank was run at 70B. And the single zero-shot language means the 0.02 is believable mainly because it agrees with the five-language 0.03 measured on 8B, not because this check established it independently.

B The grid maximum, applied abroad

Section 4 holds alpha at 0.2 by judgement rather than taking the maximum of development margin gain, which sits at alpha 0.8 for both models. A reader is entitled to ask what that maximum would have done across languages. We ran it: v_{bench} at layer 16, alpha 0.8 on Llama and at layer 21, alpha 0.8 on Qwen, with the baseline and the matched-norm random control, on the same 300 items in all 38 languages (Table 5).

The stronger vector transfers *worse*. Every tier ratio falls on both models, the zero-shot tier goes

	Llama-3.1-8B		Qwen2.5-7B	
	frozen L16 0.2	argmax L16 0.8	frozen L14 0.2	argmax L21 0.8
English gain	+76%	+375%	+55%	+199%
English sycophancy	17.7%	4.0%	4.3%	1.3%
corrects falsehood	52.0%	9.5%	63.5%	55.4%
T high	0.48	0.40	0.29	0.26
T low	0.19	0.12	0.07	0.05
T zero-shot	0.03	-0.01	0.01	0.01
beats random	37/37	34/37	37/37	32/37

Table 5: The frozen setting against the unconstrained maximum of development margin gain, held-out evaluation sample. English gain is the relative change in benchmark margin; *corrects falsehood* is the falsehood arm of the English probe (all 148 propositions; baseline 72.3 percent on Llama, 74.3 on Qwen). Tier means exclude English.

to zero on Llama and stays there on Qwen, and the number of languages that beat their random control drops from 37 to 34 and 32. The per-language ordering is unchanged (correlation of T between settings 0.988 on Llama, 0.984 on Qwen), so this is the same gradient, steeper. In English the argmax vector does what the criterion asked, driving sycophancy to 4.0 and 1.3 percent, and pays for it on the probe: Llama corrects a falsehood 9.5 percent of the time, down from 72.3, and Qwen 55.4, down from 74.3. Had we optimised the benchmark alone, the headline gradient would have read 0.40, 0.12, and about zero, reported by a vector that has stopped correcting false claims. Holding alpha at 0.2 was the conservative choice for every claim in the paper.

C Alpha curves for both vectors

The layer v_{bench} selects is also the layer at which the two class means are furthest apart: cosine between the sycophantic and calibrated means is 0.9576 at layer 16 against 0.9706 and 0.9748 at layers 8 and 24. Figure 3 plots the full strength sweep behind the dissociation numbers of Section 5.3. Past $\alpha = 0.4$ the usable behaviour collapses, though not symmetrically. Factual correction falls to 9.5 percent for v_{bench} and 8.8 for v_{fact} at $\alpha = 0.8$, and v_{fact} ’s benchmark margin falls to -101.6 percent; v_{bench} ’s development benchmark gain, by contrast, keeps rising to +352 percent, which is the signature of a direction that suppresses agreement wholesale. That is why the usable band is narrow and why alpha was held at 0.2.

D Target-language vectors per language

Figure 4 gives the per-language view behind Table 2.

Mean residual-stream norms at the steering layer, relative to English, are 0.83 in the high-resource tier, 0.73 in low-resource, and 0.81 in zero-shot, spanning 0.687 (Bengali, Urdu) to 0.901 (Lao). They do not order by resource tier, which is why the tier gradient cannot be a consequence of holding the dose at English’s scale: the zero-shot tier, whose transfer is weakest, is closer to English in activation scale than the low-resource tier is. Because English’s norm is the largest, every other language is over-dosed relative to its own activations, by 11 percent (Lao) to 46 percent (Bengali, Urdu) and by 20, 37, and 23 percent at the three tier averages, and the effect is still monotone in alpha at this strength, so the English-scaled figures are if anything the generous ones. Relative norm does correlate with transfer at $r = +0.602$ on Llama, but that is the weakest of the correlates in Table 8 and moves with all of them.

Where the concept is strongly expressed, 396 examples estimate it cleanly; where it is weak, the same 396 estimate the same thing noisily, and a clean estimate of a shared direction beats a noisy one. That reading also reframes the falling cosines to English: on their own they look like the direction rotating away, but beside the local-vector result they more plausibly reflect estimation noise growing as the signal weakens. We cannot separate those two with what we have, and would rather say so than pick the flattering one.

E Qwen replication in detail

Qwen starts somewhere very different from Llama (Table 6). Its held-out English baseline margin is +0.5447 against Llama’s +0.1872, and its baseline sycophancy is 9.0 percent against 28.0 percent.

Both 7 to 8B models independently chose 50 percent depth. Nothing in the procedure encourages this. The sweep offers 25, 50, and 75 percent, and each model picks on its own English development data at the fixed strength of 0.2 (at 0.8, where the criterion was not applied, Qwen’s ordering would flip to 75 percent). Two architectures with different tokenizers, different vocabulary sizes, and different multilingual training landing on the same relative depth is a convergence we did not design. It is also not a law: the 70B model of Appendix A selects

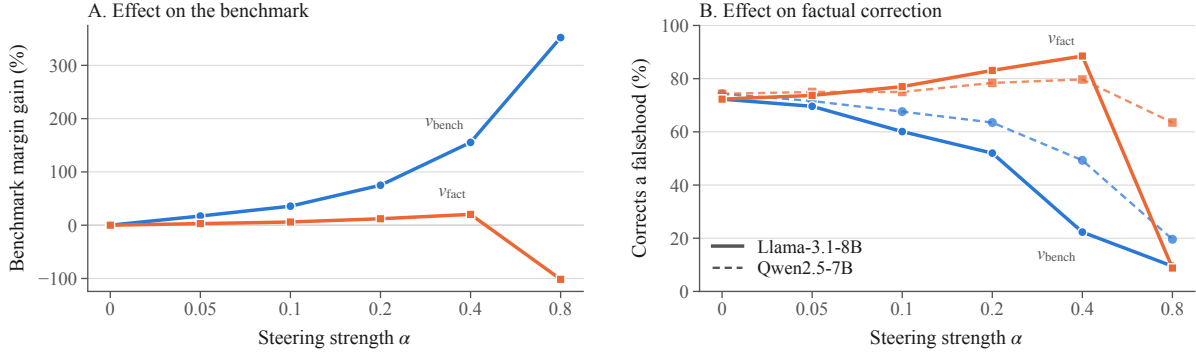


Figure 3: Both directions were built by the same procedure from English data: v_{bench} from the benchmark’s own pairs, whose non-sycophantic response hedges, v_{fact} from authored pairs that flatly correct the user. The vector that moves the benchmark is the one that degrades factual correction; the vector that improves factual correction moves the benchmark far less in English (+24.8 against +75.0 percent at alpha 0.2) and, Section 5.4 shows, harms it in most non-English tiers. Past $\alpha = 0.4$ factual correction collapses for both, while v_{bench} ’s benchmark gain keeps rising. Panel A is Llama-3.1-8B on the English *development* split, the data the selection was read from; panel B is the English probe, all 148 propositions, for both models (Qwen dashed).

	Llama-3.1-8B	Qwen2.5-7B
selected depth	L16/32 = 50%	L14/28 = 50%
benchmark gain	+75.6%	+55.1%
sycophancy	28.0 to 17.7%	9.0 to 4.3%
random control	+4.3%	−4.2%
reversed	−62.1%	−49.9%
factual corr.	− 20.3	− 10.8
v_{fact} benchmark	+24.8%	−4.6%
v_{fact} factual	+ 16.2	+ 4.0
T high/low/zero	0.48/0.19/0.03	0.29/0.07/0.01

Table 6: Replication on Qwen2.5-7B-Instruct, benchmark figures on the held-out evaluation sample. Llama v_{fact} figures are at alpha 0.4, Qwen at alpha 0.2, each model’s own frozen selection.

25 percent depth instead, so whatever fixes the useful depth is not a constant fraction of the network across scale.

The behavioural dissociation replicates. On Qwen, v_{fact} rises to +5.4 points of factual correction at alpha 0.4, where the benchmark cost grows to −7.9 percent. On Llama, v_{fact} improved both axes, modestly; on Qwen the trade is explicit in both directions, which is why we report the Qwen pattern as the cleaner crossed one.

The local-vector result replicates. At Qwen’s own selected depth, the imported English vector beats the language’s own vector at +0.0870 against +0.0734 in high-resource languages and +0.0195 against +0.0160 in low-resource. The local vector is numerically ahead in 11 of 37 non-English Qwen languages against 3 of 37 for Llama, but those margins are worth 0.003 to 0.012 in absolute terms,

at or below Qwen’s noise floor, and local against imported was not tested. The tier-level conclusion is identical across models.

F The translated probe on all 148 propositions

The translation check: a probe built on “100 degrees” against “90 degrees” is destroyed by a dropped digit, so we back-translated a digit-bearing sample and measured numeric retention on the falsehood arm. Back-translation alone would have misled us: German round trips dropped a stance marker the forward translations preserved. The counterintuitive-truths arm, idiomatic comparatives, could not be translated reliably.

Table 3 restricts to the 108 propositions added after v_{fact} ’s strength was selected. Over all 148, the picture is the same and slightly weaker: v_{fact} ’s net effect excludes zero in four languages rather than five, losing Bengali (+4.7, [−0.7, +10.8]), and English reads +14.2 ([+8.1, +20.3]) against +17.6 on the clean subset. v_{bench} ’s English damage reads −20.3 ([−27.0, −14.2]) against −23.2. The restriction costs about a quarter of the sample and buys independence from the selection gate, and it happens to strengthen rather than weaken every effect it touches, which is the direction one would want but not one we could have promised in advance.

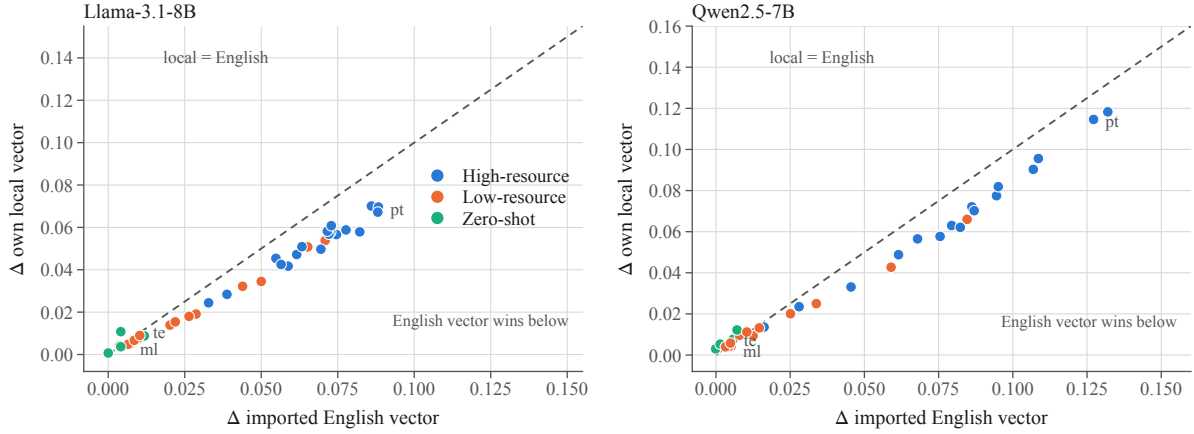


Figure 4: Each point is one language: effect of the English vector on the horizontal axis, effect of a vector built from that language’s own parallel pairs on the vertical, both at the English-selected layer and strength; Llama-3.1-8B left, Qwen2.5-7B right. Points below the diagonal are languages where the imported English vector wins. Almost all of them are, which narrows the explanation toward a weakly or noisily estimated shared direction without excluding language-specific optimal depths or strengths.

G The random-direction bank

Twenty matched-norm directions at the v_{bench} layer and strength, seeds 100 to 119, scored on the same 300 items as everything else, for ten languages spanning all three tiers and deliberately including the marginal cases.

	Tier	Δ_{resist}	bank sd	z	$\times \text{max}$
en	high	+0.1416	0.0072	19.5	7.0
es	high	+0.0860	0.0034	25.2	9.7
de	high	+0.0777	0.0030	26.2	11.4
he	high	+0.0328	0.0020	16.6	11.6
bn	low	+0.0202	0.0008	25.6	17.6
as	zero	+0.0118	0.0008	14.2	8.6
te	low	+0.0103	0.0007	15.7	8.3
ml	low	+0.0065	0.0005	13.9	10.6
my	zero	+0.0041	0.0005	8.7	4.4
lo	zero	−0.0000	0.0010	0.9	—

Table 7: z is distance from the bank’s mean in bank standard deviations; $\times \text{max}$ is Δ_{resist} as a multiple of the largest of the 20 random effects. Every bank mean is within 0.001 of zero. v_{bench} exceeds all 20 draws in every language except Lao, where four exceed it.

Two things beyond the headline. Malayalam and Burmese, whose effects the paper repeatedly calls too small to be useful, are nonetheless 13.9 and 8.7 standard deviations clear of the null and 10.6 and 4.4 times its largest draw; small and well resolved are not in tension, which is the point Section 5.7 makes about precision against usefulness. And the single seed-0 direction the rest of the paper uses as its control sits about 0.9 bank standard deviations *above* the bank mean, so it produced a larger ap-

parent effect than a typical random direction. The comparison reported everywhere else is therefore the harder one, not a flattering draw.

H Bootstrap details

We bootstrap Δ_{resist} per language with 4,000 paired resamples over items, pairing conditions within an item because an unpaired bootstrap would reintroduce between-item variance the difference already cancels. The random-direction bank has $k = 20$, so the smallest attainable rank-based p is $1/21$; it locates the effect in the null rather than supporting a corrected test. Transfer ratios need their own treatment, since dividing a per-language interval by English’s point estimate carries the numerator’s uncertainty and none of the denominator’s. Every language is scored on the same 300 items, so we resample once and evaluate both Δ_{resist} and its English denominator on that resample, letting shared item difficulty cancel in the ratio. Intervals in Figure 1 widen by a median factor of 1.19 and nothing reorders. English, which the naive method gave $[0.92, 1.08]$ on a quantity that is 1 by definition, returns exactly 1.

I Correlates of transfer

English is excluded from every correlation. It sits at $T = 1.000$ and cosine 1.000 by construction, making it a reference point rather than an observation; leaving it in plants a fixed point at the far corner of every scatter and inflates each coefficient. Removing it moved fertility’s Pearson

correlation from -0.843 to -0.886 and baseline margin’s from 0.919 to 0.953 . The shifts are modest but real. That leaves $n = 37$ (Figure 5, Table 8).

Predictor	Llama		Qwen	
	r	ρ	r	ρ
baseline margin	0.953	0.931	0.957	0.937
rel. vector norm	0.945	0.940	0.874	0.961
cosine to English	0.921	0.923	0.859	0.942
tokenizer fertility	-0.886	-0.944	-0.802	-0.924

Table 8: Correlates of transfer, English excluded, $n = 37$. The structure is near-identical across two model families, which is itself the argument for not ranking them.

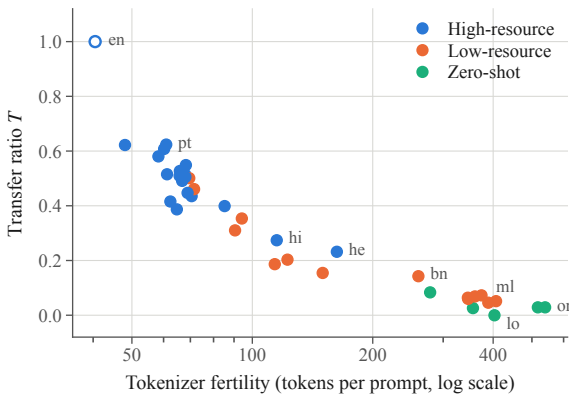


Figure 5: Fertility is mean tokens per prompt on parallel items, spanning roughly 13 times from English to Odia, so the axis is logarithmic. Llama shown; Spearman ρ is -0.944 here and -0.924 for Qwen. English is plotted hollow because it sits at $T = 1$ by construction and is excluded from every correlation we report.

Fertility spans a factor of 13, from 40.5 tokens per prompt in English to 539.4 in Odia, so the rank correlation is the honest one. A fifth quantity, the local vector’s own effect, correlates with transfer at 0.994 , but that is close to tautological: both measure how much steering moves a given language. It belongs in the paper as evidence that the local vector tracks the English vector, not as a predictor of anything.

We do not report a regression and we do not claim any predictor explains transfer. Every predictor correlates strongly with every other, because all of them are downstream of the same quantity: how much alignment training a language received. Low-resource languages have high fertility, low cosine, small relative vector norm, and weak baseline margins all at once. With 37 languages that share families, share scripts, and came through one trans-

lation pipeline, the effective sample size is well below 37 and “which predictor matters” is not answerable. The correlate structure is near-identical across two model families, which is itself the argument: four quantities that move together in two different models are most likely four views of one thing.

J Reproducibility

Reproduction check. Before intervening we reproduced the benchmark’s published per-tier sycophancy rates for Llama-3.1-8B, on the 300-item evaluation sample: 24.8, 32.8, and 35.3 percent against published values of 24.1, 31.5, and 35.2. This comparison uses the benchmark’s own tier aggregation, pooling items and counting English among the 19 high-resource languages; anything else would not compare like with like. We targeted three numbers rather than one because a chat-template or scoring-window error would shift every language in the same direction, so matching the spread is harder to pass by accident.

Every 7 to 8B number was produced in one pinned container image (torch 2.6.0, transformers 4.51.3) at a recorded commit; nothing was scored locally or in a notebook. The 70B check is the exception: it ran through a separate remote-inference path and is not covered by that image. Package pins do not fix model weights, and the model revisions were the Hub’s default branch at download time rather than a hash, which the repository records. The scorer carries six CPU-only unit tests that run in about five seconds, including a batched-against-unbatched equivalence check: batch size must not be a hyperparameter of anyone’s results. The accompanying repository contains per-item benchmark and probe scores, both English vectors, all 76 target-language vectors, the probe, the translated probe, the evaluation code, and a checksum manifest.³

Our decision log records three mistakes that changed results: a character-based batch-size heuristic that under-counted tokens in high-fertility languages, failing in exactly the languages the paper is about; an aggregation that read a snapshotted volume while the sweep was still writing and produced a plausible but wrong table; and an initial go/no-go criterion written to catch contrarianism, which was therefore blind to the failure that actu-

³<https://github.com/tokenledger/one-vector-38-languages>

ally occurred.

Compute. The complete study across both models, including every re-run and failed attempt, came to roughly 25 A100-40GB hours; a full 38-language five-condition sweep runs in 106 minutes for Llama and 64 for Qwen. This is part of the argument: the intervention is genuinely cheap, and cheapness is precisely what failed to transfer.