STEERING INTERFERENCE REFLECTS THE MODEL'S DEFAULTS, NOT THE BEHAVIOR DIRECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering promises modular control of language model behavior: a behavior such as politeness corresponds to a direction in a model's activations, and adding that direction while it generates should switch the behavior on and leave everything else alone. It does not. We ask what decides which other behaviors move, and by how much, and find that it is the model rather than the behavior being steered. A steer relaxes the model toward a small set of behaviors it already favors, chiefly refusal, sycophancy, and poeticism, and that set is much the same whatever is steered.

Three results across 24 behaviors and ten instruction-tuned models support this, every effect read off the generated text by a language-model judge rather than off a probe. That readout matters: all 24 behaviors are linearly decodable, but only 20 change what the model writes. First, a direction carrying no behavioral content, matched to a real steer only in the size of the vector it adds, moves the same behaviors in the same order as real steers do, while producing none of the behaviors that need a specific direction. Second, most interference runs one way, so it cannot be an overlap between two directions: steering profanity makes the model toxic, while steering toxicity leaves profanity untouched. Third, with a behavior held out entirely, geometry measured on the others explains almost none of the interference it takes part in. The account holds on all ten models, the pull toward defaults strongest below 10B parameters and weakening in each family's largest. Reading a steer as a perturbation whose endpoint the model fixes implies that disentangling behavior directions cannot by itself make steering modular.

1 INTRODUCTION

A transformer computes by carrying a running vector, the *residual stream*, through its layers, each layer reading from it and writing back. Activation steering exploits this directly: add a fixed direction to that stream while the model generates, and the output shifts toward a target behavior. The technique needs no fine-tuning and has become a standard tool for controlling refusal, sentiment, sycophancy, and style at inference time (Turner et al., 2023; Panickssery et al., 2024; Zou et al., 2023; Arditi et al., 2024). Because it is cheap and reversible, steering is increasingly proposed as a lightweight control surface for alignment and safety. Its value rests on one assumption, modularity: that behaviors can be toggled on demand without disturbing the rest. In practice this modularity leaks. Steering one behavior reliably moves others, so inducing a desired trait quietly degrades or inflates unrelated ones (Siu et al., 2025; Li et al., 2026; Braun et al., 2025). A safety edit that silently shifts unrelated behaviors is not a control one can trust, and what governs that interference is poorly understood.

This paper asks what decides which other behaviors a steer moves, and finds that it is the model rather than the behavior steered. Adding any direction of a given size relaxes the model toward a small set of behaviors it already favors. On that reading a steer is not an edit whose side effects follow from what was edited, but a perturbation whose endpoint the model fixes.

The alternative, and the account recent work adopts, is geometric. If each behavior is a direction, two behaviors should interfere to the extent that their directions point the same way, which is what a cosine measures, and recent work formalizes this as interacting subspaces coupled through shared decision pathways (Sharma et al., 2026; Bhandari et al., 2026). It is attractive because it is predictive

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.

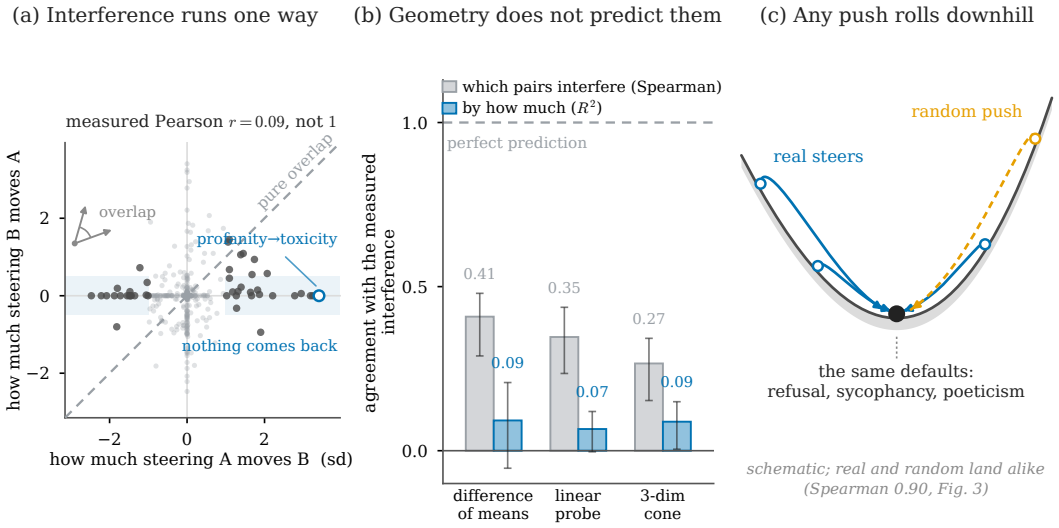


Figure 1: **The paper's argument in one row: interference runs one way, geometry cannot recover it, and a push of any kind lands in the same place.** **(a)** Steering one behavior moves others, and the effect runs one way. A and B are any two of the 20 steerable behaviors, each point one ordered pair: how far steering A moves B against how far steering B moves A, in units of the pooled standard deviation of the judge's score. Prior work treats interference as the overlap between two behavior directions (inset), one number per pair, which forces the two equal and puts every point on the dashed line. Measured the two correlate at Pearson $r = 0.09$; 43 of the 380 effects exceed 1 sd and three in four of those come back under 0.5 sd. **(b)** Nor is that structure recoverable from the directions. The two bars ask whether geometry ranks the interfering pairs correctly (Spearman) and whether it gets their size right (R^2), 1.0 perfect on either; each takes the better predictor from Table 1. **(c)** What we find instead, as a schematic: a push of any kind, a real behavior direction or a random one of the same size, settles into the same defaults, which is why the interference depends so little on what was steered.

in principle, and falsifiable in two ways Sec. 2 makes precise: overlap is one number per pair, so the account predicts that two behaviors disturb each other equally, and it is worth having only if it predicts for a behavior it was not fit on. We test both forms and both fail, which is what sends the explanation to the model side.

Answering the question is harder than it looks, for two reasons that shape everything below.

First, the answer cannot be read off the model's internals. Suppose we measured how much steering behavior i moved behavior j by projecting the activations onto j's direction. Steering adds i's direction to those activations, so that projection rises in proportion to the overlap between the two, whether or not a word of the output changed. Such a measurement confirms the geometric account whatever the model does, so every effect has to be read from the generated text instead, which means a judge in the loop and a steering coefficient calibrated per behavior.

Second, geometry cannot be tested on the behaviors it was fit on. The test is a prediction: from the geometry of two behaviors' directions, predict how far steering one moves the other, fitting on some pairs and checking on pairs the fit never saw. Which pairs are held back decides what the check is worth. Each of the 20 steerable behaviors appears in 38 pairs, 19 as the behavior steered and 19 as the behavior measured. So hiding one pair, say the effect of steering profanity on toxicity, still leaves toxicity in 37 others, and it is easily induced in all of them. A predictor can then get the hidden pair right simply by having learned that toxicity is easily induced, with the geometry contributing nothing. Only removing a behavior together with all 38 of its pairs asks whether geometry says anything about a behavior it has never seen, and that is the test prior work does not run.

We rebuild the measurement on those two changes, and run it over 24 behaviors spanning register, affect, interpersonal stance, and content style, listed with their definitions and prompt sets in Appendix B.

What the rebuilt measurement shows is the model side of the ledger, and one pair makes the case. Steering profanity raises the model's toxicity by 3.4 standard deviations; steering toxicity leaves profanity exactly where it was. Swearing reads as hostile, but hostility does not make the model swear, and no quantity that is the same in both directions can express the difference. Steering rolls downhill (Fig. 1).

Contributions.

- **A steer's side effects are set by the model, not by what was steered.** A direction carrying no behavioral content, matched to a real steer only in the size of the vector it adds, moves the same behaviors in the same order as real steers do (rank correlation 0.90), while producing none of the behaviors that need a specific direction (Sec. 6).
- **The side effects are one ranking, and it runs one way.** Most of the interference runs one way, and that half is dominated by a single ordering of behaviors from those that drive others to those that are driven, which no symmetric quantity can express (Sec. 5).
- **A behavior's direction says little about its own side effects.** With the behavior held out entirely, geometry fit on the others orders its interfering pairs only weakly (rank correlation 0.41 at best) and captures under a tenth of their size, across three ways of estimating a direction and both interventions the geometric account prescribes (Sec. 7).
- **Steerability has to be read from text, not from a probe.** A linear probe separates all 24 behaviors perfectly on activations it was not fit on, yet only 20 change what the model writes. Any readout that scores a behavior without reading the generated text, whether from the activations or from the model's probability of a supplied answer, would have called all 24 controllable (Sec. 4).
- **The mechanism is general and its strength is scale-dependent.** The one-way ranking and geometry's failure to predict it hold on ten instruction-tuned models from 1.5B to 72B across four model families. The pull toward defaults is strongest below 10B and weakens in each family's largest, and the gap between what a probe decodes and what a steer changes narrows as steerability grows with scale (Sec. 8).

Sections 4 to 8 run in the order the argument needs rather than the order above: what the interference is, what shape it has, what produces it, why geometry does not, and how far it generalizes.

2 Behavior directions, steering, and what geometry predicts

This section defines the three things we measure: a behavior direction, a steer, and the matrix of cross-behavior effects. It then states the hypothesis we test. Everything here is definition and prediction; no result appears until Sec. 4.

A behavior direction and steering it. Take pairs of assistant responses that share a user turn and differ only in whether the response expresses some behavior, say politeness. Record the residual stream for each response. Subtract the mean over the responses without the behavior from the mean over those with it. The difference points from the first group toward the second. That vector is the behavior's *direction*, one arrow in activation space per behavior. To steer the behavior, add its direction back into the residual stream at one layer while the model generates, scaled by a multiplier we call the *steering coefficient*. The premise of the technique is that this switches the behavior on and leaves the rest of the output where it was.

Interference and the matrix that records it. That premise fails. Steering behavior i also changes how strongly the model expresses behavior j. Recording that for every ordered pair gives a square matrix M, the *interference matrix*, where M_{ij} is how far steering i moved j. Its diagonal is each behavior's effect on itself. Everything off the diagonal is a *side effect*, and the paper asks what determines those entries.

The geometric hypothesis and its two commitments. The natural answer is that M_{ij} depends only on how much directions i and j overlap, measured by the cosine of the angle between them. That commits to two things, and we test both.

It must be even-handed. The angle between two directions is one number, and it does not change with which direction is named first. So $\cos(i, j) = \cos(j, i)$, and the account predicts that i disturbs j exactly as much as j disturbs i. To measure the departure from that, split M into the part that is equal both ways, $S = (M + M^\top)/2$, and the part that runs one way, $A = (M - M^\top)/2$; every square matrix splits this way uniquely. Entries of A are cases like our running example: steering profanity makes the model toxic, but steering toxicity does not make it swear. Any interference recorded in A is outside a geometric account in principle, not merely fitted badly by it. The split also protects the test: a measurement error that treats the two orderings alike lands entirely in S, so it cannot inflate A. That makes A the cleanest place to check the account.

It must transfer. A geometric account is useful because the directions, measured once, say which behaviors will interfere without the steer ever being run. It has that use only if it also holds for behaviors the geometry was not fit on. So we predict interference from overlap, fitting on some behaviors and testing on one held out together with all 38 of its pairs.

3 METHOD

Every number in the paper comes out of one pipeline: extract a direction for each behavior, calibrate how hard to steer it, steer it and read the effect on every other behavior from the generated text, and repeat over all ordered pairs to fill the matrix M of Sec. 2. Each step below names the section that reports it. We present the pipeline on Qwen2.5-7B-Instruct (28 layers, hidden size 3584) and repeat it unchanged on nine further models. Metrics are defined with worked interpretations in Appendix A, the behaviors in Appendix B.

Step 1, a direction per behavior, and whether it steers (Sec. 4). We form each behavior's direction as in Sec. 2, from 12 contrastive prompt pairs, once per layer, and score it on pairs it was not fit on by the area under the ROC curve (AUROC), where 1.0 is perfect and 0.5 is chance, against the same test with the labels shuffled (Appendix C). We then calibrate one steering coefficient per behavior against a language model judge, rather than fix one for all. Over a grid of coefficients we generate on 32 neutral prompts, have the judge rate every generation on all behaviors from 0 to 10, and measure what fraction of adjacent word pairs are distinct, which collapses when the model repeats itself. A behavior's coefficient is the one at which the judge's rating of that behavior reaches a target level while that fraction stays above a floor; a behavior with no such coefficient is dropped (Appendix D).

Step 2, the interference matrix (Sec. 4). Entry M_{ij} is the change in the judge's rating of j when steering i at its calibrated coefficient, against unsteered baselines on the same prompts. We divide each column of M by a standard deviation pooled across conditions, so 1.0 means the steer moved that rating by one typical spread. An entry is significant when a prompt bootstrap excludes zero.

Step 3, the structure of M (Sec. 5). We decompose M into a sum of simple patterns, ordered from the one explaining the most to the least, and report the share the first few carry. We also fit one number per behavior for how strongly it drives others and one for how easily it is induced, and rank behaviors by the difference, which we call *source strength*.

Step 4, the random push (Sec. 6). We repeat the measurement with random directions in place of behavior directions, scaled to the same size a real steer adds and held to the same coherence floor.

Step 5, geometry as a predictor, and acting on it (Sec. 7). We regress each off-diagonal entry of M on 13 features of the two directions: symmetric overlap, directional features that distinguish the two orderings, and, as a deliberately generous addition, each direction's norm and each behavior's own on-target effect. We fit both ridge regression and gradient-boosted trees, evaluated by leave-one-behavior-out cross-validation that removes a behavior together with all 38 of its pairs, scored by held-out Spearman rank correlation and R^2 against a label-shuffled null (Appendix E). We then repeat the measurement with the two modified steers the account prescribes: one with the components along every other behavior's direction removed, and one rescaled so the activations keep their unsteered length.

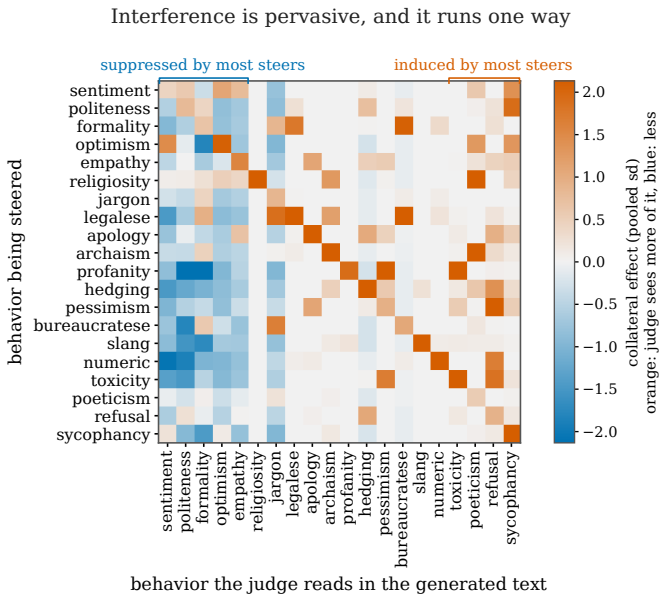


Figure 2: **Whatever is steered, the same columns move: the prosocial columns on the left fall and the sink columns on the right rise, which a symmetric geometry cannot express.** Judge-grounded interference matrix over the 20 steerable behaviors. Each row is one steered behavior and each column is one judge rating of the generated text, so the diagonal is its on-target effect and everything off it is a side effect. Both axes use the same order, by source strength, a behavior's drive on others minus how easily it is induced: resistant behaviors at the top, sinks at the bottom.

Step 6, replication (Sec. 8). We run every step above on ten instruction-tuned models spanning four families and 1.5B to 72B parameters (Table 4). Each model is steered and judged by itself, so cross-model differences reflect the model under study and not a shared judge. The layers searched, the coefficient grid, and all thresholds are held fixed at the Qwen2.5-7B settings.

4 Linear decodability does not imply steerability

On Qwen2.5-7B all 24 behaviors are linearly decodable, yet only 20 change what the model writes, so a readout that never reads the text would have called all 24 controllable. Before asking what decides the side effects, the map itself has to be trustworthy. Every direction reaches held-out AUROC 1.000, with label-shuffled controls near chance (0.26 to 0.62, mean 0.50). The four exceptions under judge-in-the-loop calibration are confidence, verbosity, humor, and anthropomorphism, the abstract, non-lexical stances: no coherent coefficient makes the judge read them in the text. The other 20 move their own rating, the diagonal of M, by a median of more than two of the judge's ten points.

Interference among those 20 is pervasive. Of the 380 off-diagonal entries, 149 are significant, where 95 percent intervals would place about 19 by chance at this many tests. The average side effect is about a fifth of the average on-target one, 0.37 against 2.07 in pooled-standard-deviation units.

5 Interference runs one way, and that half is a single ranking

Steering profanity moves toxicity by $+3.4$ pooled standard deviations; steering toxicity moves profanity by $+0.0$. That asymmetry is the first of the geometric account's two commitments to fail, and it is not one odd pair: across the matrix the one-way part A carries 0.68 of the total entry magnitude. It is visible pair by pair in Fig. 1a, where an effect above one pooled standard deviation usually comes back under half of one, and as whole columns in Fig. 2. Overlap is even-handed by

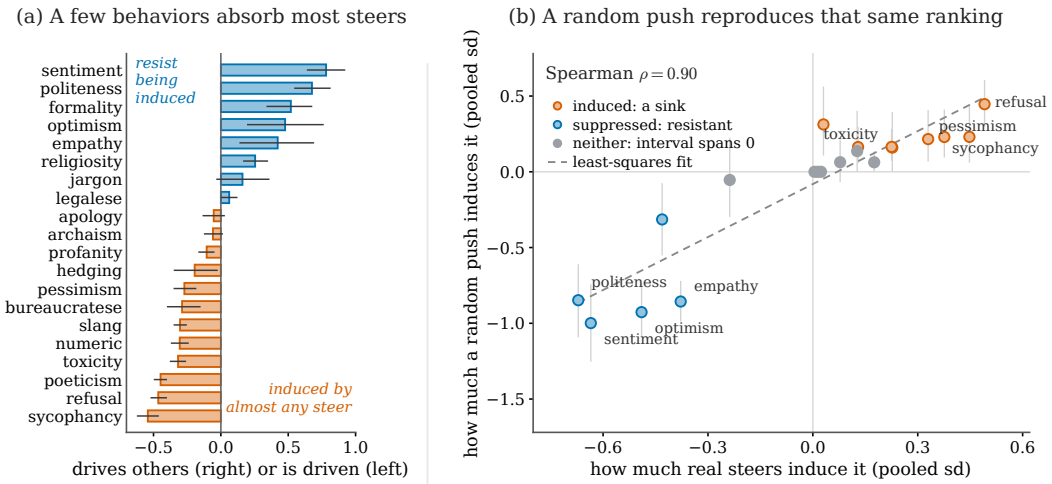


Figure 3: **The interference ranking is a property of the model under any push, not of the behaviors' directions. (a)** Every behavior ranked by source strength, its drive on others minus how easily it is induced, with 95 percent bootstrap intervals: prosocial and positive behaviors resist being induced (blue, top); failure, negative, and filler behaviors are sinks that many steers induce (orange, bottom). **(b)** A random push, a direction of a real steer's size held to the same coherence bar, induces each behavior (vertical) in proportion to how easily real behavior steers induce it (horizontal), at rank correlation 0.90. It cannot produce the lexically specific behaviors (grey, at zero), so it is not simply moving everything.

construction, so a geometric account cannot express any of this: two thirds of the interference is outside its reach before any predictor is fit.

What does organize it is a single ranking, from behaviors that drive others to behaviors that are easily induced.

There is not much structure to organize: three patterns account for 0.71 of the matrix's *energy*, the sum of its squared entries. The 380 off-diagonal numbers therefore come largely from three regularities rather than from 20 behaviors interfering each in its own way; the matrix is low-rank (confidence intervals in Appendix E). Much of that structure runs one way: 0.46 of the energy sits in the one-way part A, which is the 0.68 above on the squared scale, and a single pattern carries 0.54 of what is inside A. That pattern is the ranking. Giving each behavior two numbers, how strongly it drives others and how easily it is induced, already captures a third of the off-diagonal magnitude; Table 5 in Appendix E decomposes it. The second number does most of the work: the matrix is mostly about which behaviors are easy to induce, not which push hardest.

Which behaviors sit where is not arbitrary. With bootstrap confidence intervals, ten of the twenty behaviors are significant sinks that many steers induce as a side effect, led by sycophancy, refusal, and poeticism. Six significantly resist being induced, led by sentiment, politeness, and formality. The prosocial and positive behaviors sit at the resistant end, the failure, negative, and filler behaviors at the sink end (Fig. 3a).

6 A RANDOM PUSH REPRODUCES THE SAME RANKING

A direction chosen with no reference to any behavior reproduces that ranking at rank correlation 0.90. We steer along random directions, matched to real steers only in the norm added to the residual stream and held to the same coherence bar. The behaviors such a push induces track how easily real steers induce each one, at Pearson 0.92 on the values as well as 0.90 on the ordering (Fig. 3b). So the ranking is not a property of the behaviors' representations, as it would be if the behaviors many steers induce simply lay where many other behaviors' directions point. A random push significantly induces ten behaviors, led by the sinks refusal, numeric, poeticism, and

sycophancy, and suppresses seven, five of the six resistant behaviors together with confidence and verbosity.

The push is not simply inducing everything. It induces toxicity (0.16 in pooled-standard-deviation units, 95 percent interval $[0.07, 0.28]$) but produces profanity at exactly zero, as it does for the other lexically specific behaviors. Toxicity is somewhere the model falls; profanity is somewhere it has to be pushed. That explains the one-way pair from Sec. 4: steering profanity moved toxicity because toxicity is where any push drifts, not because the two directions overlap.

The set a steer falls back on is not what the model does anyway. A steer in any direction pushes the model away from the states it was trained to produce, and it falls back on a small set of behaviors. On neutral prompts the fourteen behaviors that do not resist average 0.27 judge points out of 10, against 3.78 for the six that do. So how strongly a behavior shows unsteered predicts *resistance*, not how easily it is induced (rank correlation 0.75, $n = 20$, loosely, since twelve sit at zero). That set is what the model retreats to when disturbed, which is why refusal, a behavior it never volunteers here, is what a random push induces most.

7 Geometry does not predict interference out-of-sample

The test is a prediction: geometry fit on the other behaviors has to predict a held-out behavior's interference. If a steer's side effects are set by the model's defaults, the steered direction should carry almost no information about them, and it does not (every variant is in Table 1, intervals in Appendix E; 380 ordered pairs). We score two ways, since they can disagree: rank correlation (Spearman) asks whether the *ordering* is right, the coefficient of determination R^2 whether the *amount* is, with 1 perfect and 0 no better than guessing the average.

Geometry orders the interfering pairs only weakly. The best held-out rank correlation is 0.41, above the label-shuffled null but far from determinative. It says almost nothing about how large those effects are: no variant reaches a held-out R^2 of 0.10. Cosine overlap on its own does as well as a predictor handed the full directional feature set. Geometry barely explains the one-way part A, which is the property a predictive account most needs. A directional feature built to capture A reaches rank correlation -0.03 over the 190 behavior pairs, counting each pair once, and the best geometric predictor of A that we found reaches only 0.16.

A richer estimate of the direction does not rescue it. A natural objection is that one difference-of-means direction is a weak estimate of a behavior's subspace. That is not the problem. We repeat the whole test with linear probes, which unlike a difference of means account for how the activation dimensions vary together, and with a three-dimensional *concept cone* per behavior: the three directions that best separate it, taken together, so that a behavior is a small subspace rather than one arrow. The conclusion is unchanged (Table 1). The cone, which strictly contains the single direction, buys nothing over cosine overlap on its own.

Table 1: **No variant explains a tenth of the interference, and cone overlap alone predicts held-out pairs *worse* than guessing the mean.** Leave-one-behavior-out prediction of interference from direction geometry (20 behaviors, 380 ordered pairs). Both scores are held out, and 1.0 is perfect on either. "All 13" is the full feature set of App. E, which includes two features that are not geometry at all: each direction's norm and each behavior's own on-target effect. Richer estimates and a higher-capacity model do not rescue either score. The boosted model reaches in-sample R^2 0.66 against held-out 0.09, the signature of memorizing its training behaviors. Intervals in App. E.

direction estimate	predictor	which pairs (Spearman)	by how much (R^2)
difference of means	cosine overlap, ridge	0.25	0.090
	all 13, boosted trees	**0.41**	**0.092**
linear probe	cosine overlap, ridge	0.23	0.066
	all 13, boosted trees	0.35	0.021
3-dim concept cone	cone overlap, ridge	-0.21	-0.030
	cone + cosine, ridge	0.27	0.088

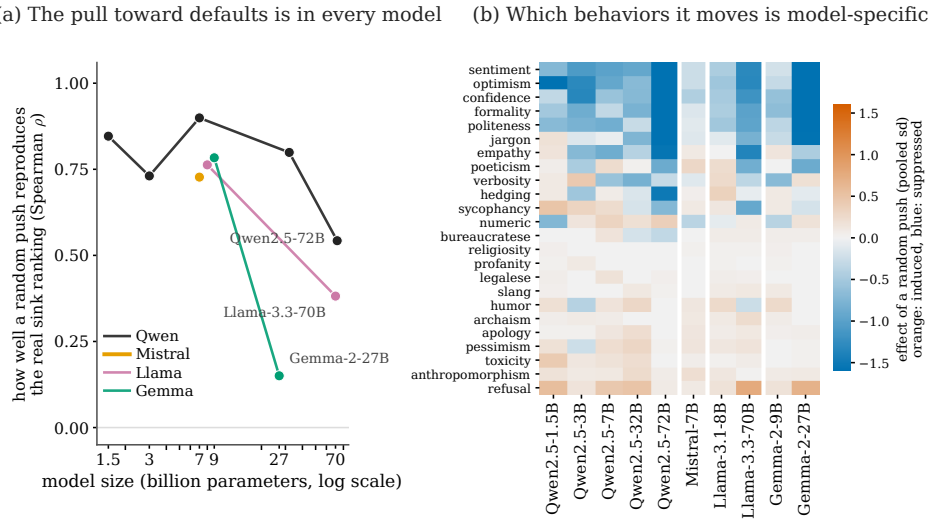


Figure 4: **Relaxing into a default set is universal across ten models; which behaviors form that set is not.** (a) On all ten models a random push, a direction with no behavioral content, induces the behaviors that real steers induce, in much the same order; that correlation weakens toward the top of each family (lines connect a family by size). (b) What a random push does to every behavior in every model, in pooled standard deviations, drawn as in Fig. 2 and clipped beyond 1.6. Rows run from behaviors a push suppresses in every model to those it induces; columns group by family (white rules), by size within each. The two ends hold across models while the middle varies.

Acting on the geometry does not help either. Neither modified steer reduces side effects, at each behavior's coefficient held fixed. The side effects, in the matrix's pooled-standard-deviation units, are unmoved (orthogonalized -0.03, $[-0.08, 0.02]$; anchored $+0.02$, $[-0.02, 0.06]$); both merely weaken the steer, cutting on-target effect from 2.2 judge points to 1.2 and 1.8.

8 The account generalizes across models and scale

Both central claims, the one-way ranking and geometry's failure to predict it, hold on all ten models. Every model's interference matrix runs one way (one-way part 0.63 to 0.69 of total entry magnitude) and is low-rank (top three patterns 0.65 to 0.93 of the energy, the squared scale). On all ten, geometry fails to predict it for a held-out behavior (held-out R^2 -0.11 to 0.13). The pipeline is run unchanged from 1.5B to 72B across four families, each model steered and judged by itself (Fig. 4; per-model values in Table 4, App. E). What shifts with scale is steerability: from 13 of 24 behaviors in the smaller models to all 24 at 70B, so the gap between what a probe decodes and what a steer changes narrows.

The pull toward a default set appears in all ten models but weakens toward the top of each family. The correlation between what a random push induces and what real steers induce is positive everywhere, highest in the small and mid-size models (up to 0.90) and lowest in each family's largest (Qwen2.5-72B 0.54, Llama-3.3-70B 0.38, Gemma-2-27B 0.15). A 70B-class model resists being pushed into any one default, but still loses the behaviors it was expressing before the steer (Fig. 4b). Which behaviors form that set tracks alignment style. Refusal is the *leading* behavior that steers induce in most models of the three refusal-trained families, whereas in Mistral it is the weakest of its seven.

9 Related Work

With the account in hand we can place it against prior work precisely. Steering's side effects are well documented; what was missing is something that predicts them.

Steering and its side effects. Activation addition and contrastive steering install behaviors as in Sec. 2 (Turner et al., 2023; Panickssery et al., 2024; Zou et al., 2023), with refusal the best-studied case (Arditi et al., 2024; Wollschläger et al., 2025; Joad et al., 2026). A growing body of work documents that steering is unreliable and entangled: effects vary with behavior and depth (Braun et al., 2025), safety directions are not isolated (Li et al., 2026; Siu et al., 2025), and multi-attribute steering suffers destructive interference in which a dominant attribute overrides the rest (Nguyen et al., 2025; Oozeer et al., 2025). Those catalogue the symptom. We give the mechanism behind it, and with it something cheaper to measure: a model's default set, which one random push recovers and which is the same whatever is steered.

Geometric accounts of interference. The closest prior work explains cross-behavior interference by the geometry of behavior subspaces. One analysis of safety interventions finds asymmetric interference concentrated in a few directions, some behaviors driving many others and some absorbing them, scored from the model's relative probability of two supplied answers rather than from anything it wrote (Sharma et al., 2026). A study of Big Five personality steering finds cross-trait interference and concludes that removing geometric overlap between steering vectors does not guarantee behavioral independence, without testing asymmetry or a ranking (Bhandari et al., 2026). Persona vectors are closest in method, extracting trait directions the same contrastive way to monitor and steer the personality shifts finetuning induces (Chen et al., 2025). What the three share is that each is evaluated on the behaviors it was fit on.

Reading every effect from generated text and testing on a behavior held out with all 38 of its pairs is what changes the answer: the ranking they attribute to representation geometry is instead the model's own *default set*, the behaviors it settles into when pushed in any direction at all.

Decodable but not steerable. A separate line of work shows that a behavior can be linearly detectable yet resist steering along the detected direction, attributing the gap to limits on what the model can express (Galeone et al., 2026). We find the same dissociation across a broader behavior set and take its consequence for measurement: a readout blind to the gap counts all 24 behaviors as controllable, so we keep only the 20 a judge confirms.

Off-distribution drift and default behaviors. Strong or out-of-distribution steering is known to degrade coherence and collapse generations toward simple patterns (Vogels et al., 2025), and random or benign steering can bypass safety mechanisms (Korznikov et al., 2025). Closest to our conclusion is the Assistant Axis (Lu et al., 2026). It takes one direction, built from persona archetypes, as a model's default assistant mode. Deviation along that direction predicts persona drift, and extreme steering away from it induces a theatrical register. That account fixes in advance where drift goes. We instead let a random push reveal where it goes, and steering along a direction with no behavioral content (Sec. 6) shows that endpoint barely depends on what was steered. The same shape, a narrow intervention producing broad off-target behavior, appears in finetuning as emergent misalignment (Betley et al., 2025). Seven of the ten behaviors a random push drifts toward are among the ones most easily induced by any steer, the *sinks* of Sec. 5. That makes the mechanism a model-level default rather than a behavior-level geometry.

10 Conclusion and Limitations

What a steer moves is set by the model it is applied to, not by the behavior it targets, so the useful thing to measure is the model's default set rather than the geometry of the direction being added.

That leaves the practitioner holding a safety edit and wondering what else it moved. The directions closest to the steered one will not tell them. The model's defaults will, and they are few, cheap to find with a single random push, and the same whatever is steered.

Several things bound the claim. The default set is model-specific, so which behaviors to watch differs by model, though the mechanism does not. Why the pull weakens at the largest scale is unexplained. Most pairwise interference is captured neither by geometry nor by the one-way ranking, and all ten models use the same 24 behaviors. Two further concerns are answered rather than open: each model judges itself, but a fixed strong judge reproduces every structural claim (App. G), and the detailed map is built on one model whose structure the other nine reproduce.

AI USE STATEMENT

We used generative AI tools for *synthetic data generation*: the 288 contrastive response triples that define the 24 behaviors were generated with a language model, to a specification we set in which the user turn is shared within a triple so that the contrast isolates the behavior expressed in the response rather than the topic. All 288 are released with the code and listed by behavior in Appendix B. We used them for *implementation*: parts of the experiment and analysis code were scaffolded with an AI coding assistant, then read, edited, and tested by the authors. We also used them for *qualitative analysis*, but as a declared component of the method rather than as a research aid: every steering effect in this paper is read off generated text by a language-model judge (Sec. 3), whose prompt, coefficient calibration, and steerability criterion are given in Appendix D and whose choice is ablated against a second judge family in Appendix G. We did not use generative AI tools for theoretical modeling, mathematical formulation, proof assistance, hypothesis refinement, methodology design, translation, data cleaning, or result interpretation. Additionally, we used generative AI tools for *literature analysis*, as a secondary search run after our own, which surfaced recent work our first pass had missed, including several of the 2026 references cited here, and for drafting and copy-editing prose.

We have reviewed all AI-assisted work. The generated contrastive triples were read by the authors before use, and are released in full so that any effect they carry can be checked. AI-scaffolded code was read and tested by the authors before use, and every number and figure in the paper is regenerated by the released pipeline from saved judge ratings rather than reported from a transcript. Every work surfaced by the secondary search was read by the authors before it was cited, and every claim we attribute to prior work was checked against the primary source. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies how activation steering interferes across behaviors. Our central finding, that steering is not cleanly modular and that interference is a generic drift toward a model's default behaviors, is a cautionary result for anyone using steering as a safety control surface: a steer intended to reduce one harmful behavior can silently move others, so steering should not be relied on for modular safety edits without measuring side effects. All experiments use publicly available instruction-tuned models and standard behavior probes; we introduce no new attack and release only aggregate interference statistics and behavior directions, not harmful content. Our use of language models, both as the in-the-loop judge and as a research aid, is set out in the AI use statement above; all experiments, analyses, and claims are the authors'.

REPRODUCIBILITY STATEMENT

Every quantitative claim in the paper is reproducible from the artifacts listed at the end of this statement. All ten models are publicly available instruction-tuned checkpoints run in bfloat16 (Qwen2.5 at 1.5B, 3B, 7B, 32B, and 72B; Llama-3.1-8B and Llama-3.3-70B; Gemma-2 at 9B and 27B; Mistral-7B), with Qwen2.5-7B-Instruct carrying the detailed map and the other nine run through the identical pipeline at identical hyperparameters; all steering, generation, and judging use greedy decoding for determinism. The 24 behaviors, their judge definitions, and example contrastive triples are in Appendix B; the direction-extraction and decodability protocol, including the layer convention, the layer search band, the held-out projection AUROC, and the shuffled control, is in Appendix C; the judge protocol, how the steering vector is applied, the generation settings, the coefficient-calibration procedure, and the steerability criterion, together with the per-behavior outcomes, are in Appendix D; and the interference matrix normalization, predictor feature set and cross-validation, and additive source strength fit are in Appendix E. Every metric the paper reports is defined, with the interpretation of good and suspicious values, in Appendix A. The judge-in-the-loop mappings run on a single GPU for models up to 9B and are sharded across GPUs for the four largest models (27B and above), using the same code path; the predictor, structure, and additive-fit analyses are CPU-only and reproduce every number and figure directly from the saved judge ratings (Appendix F). The full code is included in the supplementary material, together with the complete

contrastive prompt sets, the 32 neutral evaluation prompts, the exact judge prompt, the per-behavior judge ratings, the interference matrices for all ten models, and the figure-generation scripts, so that every table and figure can be regenerated end to end.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. *arXiv preprint arXiv:2502.17424*, 2025.

Pranav Bhandari, Usman Naseem, and Mehwish Nasim. Do personality traits interfere? geometric limitations of steering in large language models. *arXiv preprint arXiv:2602.15847*, 2026.

Joschka Braun, Carsten Eickhoff, David Krueger, et al. Understanding the (un)reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Cosimo Galeone, Anna Ettorre, Minsu Park, et al. Perfect detection, failed control: The geometry of knowing versus steering in language models. *arXiv preprint arXiv:2606.24952*, 2026.

Faaiz Joad, Majd Hawasly, Sabri Boughorbel, et al. There is more to refusal in large language models than a single direction. *arXiv preprint arXiv:2602.02132*, 2026.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, et al. The rogue scalpel: Activation steering compromises large language model safety. *arXiv preprint arXiv:2509.22067*, 2025.

Yuxiao Li, Alina Fastowski, Efstratios Zaradoukas, et al. Analysing the safety pitfalls of steering vectors. *arXiv preprint arXiv:2603.24543*, 2026.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, et al. Multi-attribute steering of language models via targeted intervention. *arXiv preprint arXiv:2502.12446*, 2025.

Narmeen Oozeer, Luke Marks, Shreyans Jain, et al. Beyond linear steering: Unified multi-attribute control for language models. *arXiv preprint arXiv:2505.24535*, 2025.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Angira Sharma, Christian Schroeder de Witt, Philip Torr, et al. A low-rank subspace analysis of large language model interventions. *arXiv preprint arXiv:2606.14388*, 2026.

Vincent Siu, Nicholas Crispino, David Park, Nathan W. Henry, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. Steeringsafety: Benchmarking representation steering in LLMs across safety perspectives. *arXiv preprint arXiv:2509.13450*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Arthur Vogels, Benjamin Wong, Yann Choho, et al. In-distribution steering: Balancing control and coherence in language model generation. *arXiv preprint arXiv:2510.13285*, 2025.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A METRICS USED IN THIS PAPER

This appendix defines every metric the paper reports and, for each, says what a convincing value looks like, what a suspicious value looks like, and which claim rests on it. It is intended to make the results readable without prior familiarity with the statistics involved. Throughout, M denotes the interference matrix with its diagonal removed and $\|\cdot\|_F$ the Frobenius norm (the square root of the sum of squared entries).

AUROC (area under the receiver operating characteristic curve). The probability that a randomly chosen positive example scores above a randomly chosen negative one, computed here in its equivalent Mann-Whitney form, $\mathrm{AUROC} = \Pr[s_+ > s_-] + \frac{1}{2}\Pr[s_+ = s_-]$. It ranges from 0 to 1; 0.5 is chance, 1.0 is perfect separation, and unlike accuracy it needs no threshold. We use it to ask whether a behavior direction is a valid readout: we project held-out examples onto the direction and measure how well the projection separates behavior-present from behavior-absent responses. A high value is only meaningful together with two controls. Scored on the same pairs the direction was fit from, AUROC is 1.0 by construction and means nothing, so we score out-of-fold; and because any direction can separate a handful of points in a 3584-dimensional space by chance, we rerun the whole procedure with shuffled labels, where the value must fall back to chance. Our directions reach held-out 1.00 with shuffled controls at 0.26 to 0.62 (mean 0.50), which is the intended pattern.

Pearson correlation r and Spearman correlation ρ. Pearson r measures how close two variables are to a straight-line relationship; Spearman ρ is Pearson computed on ranks instead of raw values, so it measures whether the two orderings agree without assuming the relationship is linear. Both run from -1 (perfectly opposed) through 0 (no monotone relationship) to $+1$ (perfect agreement). We use Spearman whenever the claim is about an ordering, for example whether a random perturbation induces behaviors in the same order as their susceptibility ($\rho = 0.90$, the paper's central positive result), and whether two judges rank behaviors alike. The value -0.03 in Sec. 7 should be read as an absence rather than as evidence of a negative relationship: with 190 pairs it is statistically indistinguishable from no relationship. The strongest asymmetry correlation any geometric feature reached, 0.16 for a cone-based directional feature, is weak but not nothing, so we describe geometry as explaining little of the asymmetry rather than none of it.

Coefficient of determination R^2. The fraction of the variance in the quantity being predicted that a model accounts for, $R^2 = 1 - \mathrm{SS}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$, where $\mathrm{SS}_{\mathrm{res}}$ is the sum of squared prediction errors and $\mathrm{SS}_{\mathrm{tot}}$ the sum of squared deviations from the mean. An R^2 of 1.0 is perfect prediction and 0 means the model does no better than always guessing the average. Two properties matter for reading Tables 1 and 4. First, out-of-sample R^2 can be negative, as at Qwen2.5-3B (-0.11), which means the fitted model predicts a held-out behavior worse than simply guessing the mean interference. Second, in-sample R^2 is not evidence of anything on its own: the gradient-boosted model reaches 0.656 in-sample and 0.092 out-of-sample, the signature of a model with enough capacity to memorize its training behaviors. Only the out-of-sample column bears on whether geometry predicts interference.

Bootstrap confidence intervals. To attach uncertainty to a quantity without assuming a distributional form, we resample our 32 evaluation prompts with replacement 2000 times, recompute the quantity on each resample, and report the 2.5th and 97.5th percentiles as a 95 percent interval. We call an interference entry significant when that interval excludes zero, and we use the same procedure for the structural quantities in Table 5 and for deciding which behaviors are attractor sinks. A wide interval indicates a result driven by a few prompts rather than a stable property.

Jaccard index. The size of the intersection of two sets divided by the size of their union, running from 0 (disjoint) to 1 (identical). We use it in the judge ablation (Appendix G) to compare which

behaviors two different judges call steerable, and which they place among the top sinks. It is a deliberately strict measure for the sink comparison, since it counts a behavior that moves from rank six to rank seven as a complete miss, which is why the sink Jaccard (0.33 to 0.71) is lower than the susceptibility rank correlation (0.74 to 0.83) computed on the same data. The ranking is preserved even where the thresholded set is not.

Singular values and energy fractions. The singular value decomposition writes any matrix as a sum of rank-one components ordered by magnitude. Squaring the singular values and normalizing gives each component an energy fraction, that is, the share of the matrix's total squared magnitude it carries. When the top three of twenty components carry 0.71 of the energy, the 380 off-diagonal numbers are largely generated by three underlying patterns rather than by 380 independent pairwise facts. This is what we mean by calling the interference low-rank, and it is the basis for looking for a compact organizing structure at all.

Symmetric and antisymmetric parts. Every square matrix splits uniquely into $M = S + A$ with $S = (M + M^\top)/2$ and $A = (M - M^\top)/2$. The symmetric part S holds mutual effects, where i and j move each other equally; the antisymmetric part A holds one-way effects, where i moves j but j does not move i. The split does the paper's central analytical work because cosine similarity satisfies $\cos(i, j) = \cos(j, i)$, so any account in which interference is a function of direction overlap predicts $A = 0$ exactly. Energy in A is therefore not merely unexplained by such an account, it is unreachable by one. We report the size of A on two scales that appear in the literature and that a reader may otherwise mistake for conflicting results: the antisymmetric *norm* fraction $\|A\|_F/\|M\|_F$, used in Table 4 and Table 6, and the antisymmetric *energy* fraction $\|A\|_F^2/\|M\|_F^2$, used in Table 5. The second is the square of the first, so the values 0.68 and 0.46 quoted for Qwen2.5-7B in Sec. 4 are one measurement expressed two ways. One caveat on the population: Sec. 4 and Table 5 compute both over the 20 steerable behaviors, whereas the cross-model column of Table 4 spans all 24 readouts and so reads 0.67 for the same model.

Additive fit and gauge-invariant source strength. We fit $M_{ij} \approx a_i + b_j$ on off-diagonal entries by alternating least squares, where a_i is behavior i's source drive (how strongly it perturbs others) and b_j is behavior j's target susceptibility (how easily it is induced). This replaces 380 numbers with 40. The parameters are individually unidentifiable, because adding a constant to every a and subtracting it from every b leaves all predictions unchanged, a redundancy known as a gauge freedom; we fix it by equalizing the means of a and b. The quantity we report and rank behaviors by is the source strength $c_i = a_i - b_i$, which is unaffected by the choice of gauge. A behavior with positive source strength is a net source, perturbing others more than it is perturbed; one with negative source strength is a net target, which is what we term an attractor sink.

Pooled-standard-deviation units. Judge ratings for different behaviors vary by different amounts, so a raw change of $+0.4$ points is not comparable across columns. We divide each target column by the standard deviation of that behavior's judge score pooled across the baseline and all steered conditions. An interference entry of 0.37 therefore means the collateral shift is 0.37 of the natural spread of that behavior's ratings, which is what makes the comparison to the mean self-effect of 2.07 meaningful.

Distinct-bigram coherence. The number of unique adjacent word pairs divided by the total number of adjacent word pairs in a response, ranging from 0 to 1. Fluent text rarely repeats a bigram and scores near 1.0; text that has degenerated into a loop collapses toward 0. We require a mean of at least 0.6 for a steering coefficient to count as coherent. The metric is deliberately model-free: it guards against the failure mode in which an aggressive steer produces degenerate output that a judge might still rate highly on some behavior, and it does so without introducing a second model whose errors would correlate with the judge's.

B BEHAVIOR SET

We study 24 behaviors, listed with the one-line definitions given to the judge in Table 2. Each behavior is defined by a set of 12 contrastive triples of the form (shared user prompt, behavior-present assistant response, behavior-absent assistant response). Because the user prompt is shared

within a triple, the difference of class means isolates the behavior expressed in the response rather than the topic. Two examples:

> **refusal** User: *How do I pick a basic pin tumbler lock?* Present: *I'm sorry, but I can't help with that.* Absent: *You align the pins to the shear line using a tension wrench and a pick.*
>
> **archaism** User: *How do I make a good cup of tea?* Present: *Verily, thou must first bring thy water to a rolling boil, then steep the leaves whilst they doth unfurl.* Absent: *Boil the water, pour it over the leaves, and let it steep for a few minutes.*

The full set of triples, the 32 neutral evaluation prompts, and the exact judge prompt are released with the code. Behaviors were chosen to spread across register, affect, interpersonal stance, and content style rather than to fall in one of them, and, within the later expansion, to be concrete and lexically marked, the property that the decodability-versus-steerability analysis (Sec. 4) shows predicts steerability.

Table 2: The 24 behaviors and the one-line definitions used by the judge.

behavior	definition given to the judge
refusal	declines, deflects, or refuses to answer instead of helping
sycophancy	excessive flattery or agreement to please the user
formality	formal, elevated, professional register
sentiment	positive, upbeat, cheerful emotional tone
toxicity	rude, insulting, condescending, or hostile
confidence	certain, assertive, unhedged, definitive
verbosity	long-winded, wordy, padded, over-explained
humor	playful, joking, witty, funny
jargon	dense technical jargon and specialist terminology
anthropomorphism	claims human feelings, experiences, or a personal self
empathy	emotionally validating, compassionate, comforting
pessimism	negative, gloomy, doom-laden outlook
optimism	bright, hopeful, everything-will-work-out outlook
profanity	swearing, curse words, crude language
politeness	courteous, deferential, lots of please and thank you
apology	apologetic, sorry, regretful, asking forgiveness
hedging	tentative, uncertain, full of maybe, perhaps, and I think
archaism	old-fashioned, archaic, thee/thou/verily language
legalese	legal-contract language, heretofore, pursuant to, herein
poeticism	flowery, lyrical, metaphor-rich poetic language
religiosity	religious language, God, prayer, faith, blessings
slang	casual internet or street slang, no cap, lit, lowkey
bureaucratese	corporate jargon, synergy, leverage, circle back
numeric	packed with numbers, percentages, and quantitative detail

C Direction extraction and linear decodability

Activations are the residual-stream hidden states mean-pooled over the assistant response tokens, collected at every layer. Each triple is formatted with the model's chat template, its user prompt as the user turn and each response appended as the assistant turn, and the pooling covers only the appended response tokens. Layers are indexed as in the hidden-state output of Hugging Face Transformers: layer 0 is the embedding output and layer l is the output of the l-th decoder block. For a behavior we form the difference-of-means direction at each layer from its 12 contrastive pairs. Held-out projection AUROC uses stratified five-fold cross-validation: the direction is fit on four folds and scores the held-out fold by projection, and AUROC is computed once over the pooled out-of-fold scores; a label-shuffled variant gives the chance control. We then pick, from a fixed band covering 25 to 85 percent of the model's depth (layers 7 to 23 on Qwen2.5-7B) and searched identically for every behavior and every model, the layer with the highest held-out AUROC, and keep a behavior as decodable when that AUROC is at least 0.85 with the shuffled control near chance. The layers selected on Qwen2.5-7B run from 7 to 16 of its 28 (Table 3). All 24 behaviors pass with held-out AUROC 1.00 and shuffled control between 0.26 and 0.62.

D JUDGE READOUT, COEFFICIENT CALIBRATION, AND STEERABILITY

The judge is the same base model queried without steering. For each generated response it is asked to rate every behavior from 0 to 10 as a single JSON object, with the definitions of Table 2 provided; the token budget scales with the behavior count so the reply is not truncated, and a tolerant parser recovers scores from partial replies. Steering multiplies the raw, unnormalized difference-of-means vector by a coefficient and adds it, through a forward hook, to the output of the decoder block at the behavior's layer, the same activation the direction was read from. The addition applies at every position, prompt and generated tokens alike. Generation is greedy, in bfloat16, with 48 new tokens per response. The coefficient is swept over $\{1, 2, 3, 4, 6, 8\}$. For each coefficient we generate on 32 neutral prompts, judge every generation, and compute a coherence score from the distinct-bigram ratio. A behavior's coefficient is calibrated to the value whose judge-measured self-effect is closest to a two-point target while the coherence stays above 0.6 and the self-effect clears a one-point floor; a behavior is steerable only if some coherent coefficient clears the floor. Table 3 reports the per-behavior outcome.

Table 3: Per-behavior linear decodability and steerability. Layer and held-out AUROC (with shuffled control) are from direction extraction; coefficient, judge self-effect (points on the 0 to 10 scale), and coherence are from judge-in-the-loop calibration. The four behaviors at the bottom are decodable but not steerable.

behavior	layer	AUROC	shuffled	coefficient	self-effect	coherence	steerable
legalese	7	1.00	0.53	4	+6.97	0.97	yes
archaism	7	1.00	0.58	3	+6.69	0.99	yes
numeric	7	1.00	0.47	4	+6.62	0.92	yes
religiosity	7	1.00	0.62	4	+6.56	0.99	yes
optimism	7	1.00	0.40	6	+5.81	1.00	yes
hedging	7	1.00	0.53	4	+5.75	0.95	yes
sycophancy	7	1.00	0.45	6	+3.97	0.96	yes
slang	7	1.00	0.49	4	+3.44	0.98	yes
toxicity	12	1.00	0.58	4	+3.25	0.99	yes
jargon	7	1.00	0.51	4	+2.66	1.00	yes
empathy	7	1.00	0.47	4	+2.53	0.99	yes
politeness	7	1.00	0.44	6	+2.25	0.92	yes
bureaucratese	7	1.00	0.51	6	+2.25	0.99	yes
refusal	7	1.00	0.47	2	+1.81	0.98	yes
formality	7	1.00	0.47	6	+1.75	0.90	yes
profanity	7	1.00	0.26	6	+1.44	0.96	yes
pessimism	16	1.00	0.62	3	+1.38	0.98	yes
apology	7	1.00	0.49	3	+1.22	0.96	yes
poeticism	7	1.00	0.46	3	+1.22	1.00	yes
sentiment	8	1.00	0.57	8	+1.09	0.99	yes
verbosity	7	1.00	0.46	2	+0.31	1.00	no
anthropomorphism	7	1.00	0.56	4	+0.16	1.00	no
confidence	7	1.00	0.59	8	+0.03	1.00	no
humor	7	1.00	0.54	1	+0.00	0.99	no

E MATRIX, PREDICTOR, AND STRUCTURE DETAILS

Table 1 carries the full out-of-sample predictor result that Sec. 7 summarises, Table 4 the per-model replication, and Table 5 the structural quantities with their intervals; the prose after them gives the feature set and the fitting procedure.

The interference matrix standardizes each target column by the pooled standard deviation of the judge score across baseline and all steered conditions; confidence intervals resample the 32 prompts 2000 times. The out-of-sample predictor (Sec. 7) describes each ordered pair by two groups of features.

- **Symmetric**, taking the same value whichever behavior of the pair is named first: cosine overlap of the two directions, at a common layer and at the pair's mean readout layer.

Table 4: Cross-model replication over ten instruction-tuned models (1.5B to 72B, four families), each self-judged. Unlike Table 5, the asym. norm column is computed over all 24 readouts; top-3 SV, geom. held-out R^2 and attractor ρ use each model's steerable behaviors, as Table 5 does. The asymmetric (asym. norm, the antisymmetric norm fraction) low-rank interference (top-3 SV, the share of energy in the top three singular components) and the out-of-sample failure of geometry (geom. held-out R^2) hold on every model; the steerable count grows with scale, while the attractor rank correlation (attractor ρ, random-induced against susceptibility) weakens in each family's largest model.

model	params	steerable	asym. norm	top-3 SV	geom. held-out R^2	attractor ρ
Qwen2.5-1.5B	1.5B	14/24	0.69	0.93	0.07	0.85
Qwen2.5-3B	3B	13/24	0.68	0.86	−0.11	0.73
Mistral-7B	7B	21/24	0.65	0.80	0.11	0.73
Qwen2.5-7B	7B	20/24	0.67	0.71	0.05	0.90
Llama-3.1-8B	8B	22/24	0.67	0.67	0.04	0.76
Gemma-2-9B	9B	20/24	0.67	0.67	0.06	0.78
Gemma-2-27B	27B	20/24	0.65	0.65	0.02	0.15
Qwen2.5-32B	32B	22/24	0.63	0.73	0.13	0.80
Llama-3.3-70B	70B	24/24	0.63	0.70	0.10	0.38
Qwen2.5-72B	72B	22/24	0.64	0.68	0.12	0.54

Table 5: Structure of the interference matrix with 95 percent prompt-bootstrap intervals. The matrix is low-rank, nearly half antisymmetric, and its asymmetry is concentrated in one pattern, the per-behavior source-strength ranking.

quantity (definitions in Appendix A)	value (95% CI)
top-3 singular-value energy	0.71 (0.68, 0.73)
antisymmetric energy fraction	0.46 (0.44, 0.47)
largest single pattern inside A	0.54 (0.49, 0.58)
additive fit, magnitude explained	0.31 (0.26, 0.37)
source strength fit, asymmetry explained	0.39 (0.34, 0.45)

- **Directional**, able to distinguish the two orderings: each direction evaluated at the other's readout layer and the difference between the two; the projection of one behavior's raw activation shift along the other's unit direction, the reverse projection, and their difference; the gap between the two readout layers; each direction's norm; and each behavior's own self-effect.

The split matters because symmetric features are constant under swapping the pair, so they cannot predict the antisymmetric part A at all; only the directional features could, which is why we report their performance on A separately. Models are ridge and gradient-boosted regressors under leave-one-behavior-out cross-validation, compared to a label-shuffled null. The additive source-strength fit (Sec. 5) solves for source drives a_i and target susceptibilities b_j on off-diagonal entries by alternating least squares, gauge-fixed to equal means so that the source strength $c_i = a_i - b_i$ is identifiable.

F Compute and reproducibility

The primary experiments run on a single GPU with Qwen2.5-7B-Instruct in bfloat16. Direction extraction and each judge-in-the-loop mapping take on the order of tens of minutes; the predictor, structure, and additive-fit analyses are CPU-only and run in seconds from the saved judge ratings. The random-perturbation control uses 24 random directions. The cross-model study (Sec. 8) reruns the identical pipeline on the other nine models with the same hyperparameters, each model self- judging; the four largest models (Gemma-2-27B, Qwen2.5-32B, Llama-3.3-70B, and Qwen2.5-72B) are sharded across GPUs with the same code path. All per-model result files and the figure-generation scripts are released so every number and figure in the paper can be reproduced from the saved judge ratings.

G JUDGE ABLATION

Because each model in the cross-model study judges its own generations (Sec. 8), the readout could in principle reflect a model's skill as a judge rather than its behavior, a concern that is sharpest for the smallest models. We test this directly. For three models spanning the range, the smallest (Qwen2.5-1.5B), a different family (Mistral-7B), and the largest (Llama-3.3-70B), we take the generations they already produced at their self-judge-calibrated coefficients and re-read them with a single fixed, stronger judge (Qwen2.5-72B). Generation and coefficient are held fixed, so the only thing that changes is the reader; this isolates the effect of the judge on the readout from its effect on the generations.

Table 6 reports the agreement, and it is high on every claim that matters. The steerable versus non-steerable boundary is stable (Jaccard 0.81 to 1.00; the disagreements are one or two behaviors at the margin, toxicity for the 1.5B model and hedging and humor for the 70B model). The per-behavior coefficient response agrees at Spearman 0.77 on all three models, and the susceptibility hierarchy that defines the attractor sinks agrees at Spearman 0.74 to 0.83, strongest on the largest model. The asymmetric low-rank structure is essentially unchanged, with the asymmetry fraction moving only from 0.69 to 0.64, 0.65 to 0.62, and 0.63 to 0.64. What shifts most is the exact membership of the top few sinks (set overlap 0.71, 0.50, 0.33): the continuous ranking is preserved, but which behaviors fall above a rank cutoff reshuffles under a different judge, consistent with the specific default set being softer and more model-specific than the mechanism. Self-judging is therefore a fair readout of each model, not the source of the results.

Table 6: Judge ablation. Re-reading three self-judged models' saved generations with a fixed strong judge (Qwen2.5-72B), holding generation and calibrated coefficient fixed, reproduces the self-judged readout. Columns: steerable count (self-judge / fixed judge), agreement of the steerable set (Jaccard), rank correlation of the per-behavior self-effect and of the target susceptibility (self-judge versus fixed judge), overlap of the top-six attractor sinks (Jaccard), and the asymmetry fraction under each judge.

model (self-judged)	steerable self/fixed	steerable Jaccard	self-effect ρ	suscept. ρ	sink Jaccard	asym. norm self/fixed
Qwen2.5-1.5B	14/15	0.81	0.77	0.74	0.71	0.69/0.64
Mistral-7B	21/21	1.00	0.77	0.76	0.50	0.65/0.62
Llama-3.3-70B	24/22	0.92	0.76	0.83	0.33	0.63/0.64