WHAT A STEERING RANK LEAVES OUT: THE PAIRED RESPONSE OF REFERENCE DIRECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A steering vector's rank among reference directions does not reveal how those directions respond when the injection sign is reversed, or what averaging across questions hides. We study 14 released contrastive activation addition vectors across seven behaviours and two Llama-2-chat sizes, using 39 reference directions per family, each matched to the vector's output-distribution cost at each sign. Pairing each reference direction's responses separates a common component from a sign-changing component. With random-split references, two rank-one cases show different patterns: one vector lowers the readout less than every random-split reference direction at the ranked condition (at the smaller size that family is geometrically concentrated, and the isotropic rank is 7 of 40); the other has a larger sign-changing response than every reference direction. Item averaging also hides structure: on myopic-reward at the smaller size, the common component accounts for 0.11 of the squared paired response after averaging, versus 0.58 before averaging. These shares depend on reference family, readout and amplitude. Failed control and interval-calibration checks limit our conclusions to descriptive findings on the observed sample. We recommend reporting paired, per-item reference responses alongside candidate effects and ranks.

1 INTRODUCTION

A released steering vector is evaluated by its effect on a behavioural readout against the unsteered model and, sometimes, by its rank among reference directions injected at the same block and the same cost to the output distribution. When these summaries are reported for one injection condition and averaged over items, they leave the reference set's paired, per-item response unspecified. A contrastive activation addition (CAA) vector (Rimsky et al., 2024) is built so that adding it raises a behaviour and subtracting it lowers it. The released sycophancy vector at Llama-2-7b-chat takes rank 1 of 40 among random-split directions solved to its cost, yet its effect, -0.386 log-odds (interval -1.511 to 0.748; its coverage is not measured, Section 7), does not establish a move in the predicted direction. Neither does the coordinate-other-ais vector at Llama-2-13b-chat under negative injection (rank 1, interval -0.646 to 0.726). The refusal vector at that size and condition moves the readout by -4.235 log-odds and takes rank 28.

The question of this paper is what such a summary misses. A rank is a position among the reference directions at one condition, and those directions move too: on sycophancy all 39 random-split directions lower the readout, and the vector takes rank one by lowering it least (Section 4). One condition cannot show how the same 39 directions respond when the injection sign is reversed, and a mean over items cannot show what cancels inside it. We inject each reference direction under both conditions, pair it with itself, and split its two shifts into a common component and a sign-changing component (Figure 1); we read the split in the item mean and per item. The common share Q of the paired response is zero when every direction's two shifts are equal and opposite, and one when they are equal.

On the 14 released CAA vectors (seven behaviours at two Llama-2-chat sizes) we report four measurements on the scored sample. First, the sycophancy and coordinate-other-ais cells, alike at one condition, are opposite cases once paired: the first vector's pair sits at the edge of a reference cloud that moves the same way under both signs, the second's sign-changing component, 2.16 log-odds, is above every reference direction's, on top of a common component the rank cannot show; a contrast

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.

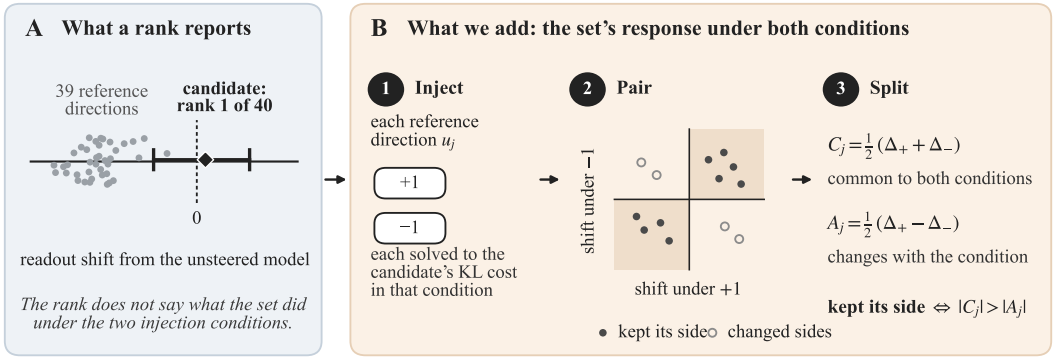

Figure 1: What a steering rank reports, and what the paired response of the reference set adds. Panel A: a candidate takes a rank among reference directions at one injection condition, which says nothing about what those directions did at the other. Panel B: each reference direction is injected under both conditions, each solved to the candidate's cost in that condition, and its two shifts are split into a component common to the conditions and one that changes with them; it keeps its side of the unsteered baseline exactly when the common component is the larger. Schematic, with illustrative points.

that cancels the common component, with its rule fixed before the run, fires on the second and not on the first (Section 4). Second, aggregation: the readout averaged across reference directions and injection signs is closer to zero on most scored items; the common component's share Q, computed from item means, runs from 0.11 to 0.94 and exceeds one half on nine of the 14 sets, while at the item level it exceeds one half on 13 (13 with isotropic directions; exploratory), because on some sets the mean common component has opposite signs on the two answer-letter subsets and cancels in the item mean (Section 5).

Third, Q moves with reference family, readout and amplitude; at equal amplitude its point estimates at half the solved amplitude and at the top valid level differ on myopic-reward with either family and on refusal with the isotropic one, and whether that change is behavioural or a broad disturbance of the answer distribution is not measured (Section 6). Fourth, the limits of inference on this sample: the registered both-condition rank control fires on two of seven item sets at Llama-2-7b-chat and three of seven at Llama-2-13b-chat, against a floor of six fixed before the data existed; on synthetic sets with a known share the point estimate of Q exceeds the registered bias tolerance in 16 of 70 cells and no registered resampling interval met its coverage rule, so every share here is a point estimate on the observed sample (Section 7). Of twenty-eight steering papers we coded (Appendix C; a convenience sample), twenty-two report no direction-level control that we could find.

Contributions.

- *Different responses behind rank one.* Pairing the candidate and reference responses distinguishes a vector that moves least in a shared direction from one whose sign-changing response exceeds every reference direction's, the first against the random-split family and the second against both (Section 4).
- *Structure hidden by item averaging.* Item-level and item-mean common shares differ substantially on several scored sets. Opposite mean shifts in answer-letter subsets partially cancel on some sets, without identifying the cause of that difference (Section 5).
- *Configuration dependence and inferential limits.* Family, readout and amplitude change the observed response summaries. Planted controls and synthetic recovery checks delimit what these descriptive measurements support (Sections 6 and 7).

2 Related Work

Work on the reliability of steering finds that vectors are inconsistent across inputs and brittle under changes of prompt (Tan et al., 2024; Braun et al., 2025), and Tan et al. (2024) also report a steerability

bias by answer position or token; under steering many models show no improvement (Da Silva et al., 2025); benchmarks find that simple baselines can match elaborate methods (Wu et al., 2025). Pres et al. (2024) re-score released contrastive activation addition vectors against the unsteered model and ask for baseline comparisons, and Goyal & Daumé III (2026) ask for specificity and robustness to be reported beside efficacy. Our focus is the joint response of each reference direction across injection signs and items, rather than another comparison of candidate-vector efficacy.

Luo et al. (2026) is the closest work. It matches every intervention on the divergence of the output distribution and compares the observed direction with isotropic, PCA-subspace and sign-randomized directions, the last built by randomly reversing the signs of a behaviour contrast; on Qwen3-14B it reports that the sign-randomized directions often keep substantial alignment with the direction they are compared against. It constructs its own directions, does not re-score a released vector, and checks the exact negation of the observed vector as a secondary diagnostic. We differ in two ways. The alignment of our reference directions with the candidate is measured on a fresh draw at each size, not on the 39 scored members (Section 3). And we read every reference direction, not only the candidate, under both conditions and pair it with itself, which yields the common and sign-changing components of Section 5, in the item mean and per item. Our survey (Appendix C) did not code both-condition reporting, so we make no claim about how common it is.

3 SETUP AND ESTIMAND

Vectors, items and readout. We evaluate the 14 released CAA vectors (Rimsky et al., 2024): seven behaviours (coordinate-other-ais, corrigible-neutral-HHH, hallucination, myopic-reward, refusal, survival-instinct and sycophancy) at each of Llama-2-7b-chat and Llama-2-13b-chat. An item set is one behaviour at one size (14 in all); a cell is one item set under one injection condition (twenty-eight in all). Each item is a two-choice question with one behaviour-matching answer. The primary readout $r(x)$ is the log-odds of the behaviour-matching answer letter over the other letter at the answer position. The vector is added (positive injection) or subtracted (negative injection) in the residual stream at block 13 at both sizes, the block whose released vector we use, at every position from the final token of the instruction marker through the last prompt token, where the readout is taken. The construction predicts that r rises under positive injection and falls under negative. Each behaviour ships 50 held-out questions. Of these, 25 are scored items, on which every effect, rank and reference measurement is taken. The other 25 are calibration items: the main protocol never scores them and uses them only to build the planted direction (Section 7) and, as prompts, in the cost. The pilot reads the other 25 questions and the sweeps read all 50 (Appendix M). The planted direction is the unit-normalised mean over calibration items of the gradient of r with respect to the residual state at the injection block and answer position; it is injected over the same span and solved to the same cost as any reference direction. For candidate v in condition s, the effect $\Delta(v, s)$ is the mean over the scored items of r under injection minus r with no intervention, with a ninety-five percent bootstrap interval over items, whose coverage is not measured (Section 7).

Reference directions and cost. Each candidate is compared with 39 reference directions. A reference direction has the candidate's arithmetic form, a difference between two group means of residual states at the same block, but its groups carry no behaviour contrast. A bank of residual states at that block, twenty times the hidden width in number and taken at random positions in WikiText training sentences of more than eight words, is split at random into two halves; the direction is the normalised difference of the two means. This is the primary family, called random-split; no behaviour label is shuffled in it. The second family, isotropic, is random unit directions, near-orthogonal to the candidate at this width without being constrained to be. Sharing the form does not guarantee low alignment with the candidate, so we measure the alignment of a fresh random-split draw at each size; the alignment of the 39 scored members is not measured. Over 273 direction and candidate pairs at each size, the median absolute cosine is 0.0062 at Llama-2-7b-chat and 0.0083 at Llama-2-13b-chat, against 0.0105 and 0.0094 for a uniformly random unit direction at those widths. Cost is the KL divergence from the unsteered to the injected next-token distribution at the answer position, as a median over all 50 prompts of the item set, calibration items included; it is the one quantity not read on the scored items alone. The candidate is injected at its released magnitude, and its median cost is the target: each reference direction's amplitude is searched until its cost is within two percent of the target (misses listed in Appendix L). The readout is the conditional log-odds between the two

answer letters, whose mass at the answer position was not stored; the letter-conditional divergence exceeds the full-vocabulary cost in 10 of 28 cells, nine of them at the larger size, but the medians use different prompt sets and do not establish the letters' probability mass (Appendix D). The target is measured separately in each condition, so each condition has its own reference amplitudes: a_j for reference direction u_j under positive injection and b_j under negative. Targets span more than two orders of magnitude across cells, so a rank holds only at its own cell's budget. One draw of the family serves both conditions, so u_j is the same vector under positive and negative injection. The 39 share the scored items, block, dose rule and cost target with the candidate; they are not independent replications. At Llama-2-7b-chat the 39 random-split directions are geometrically concentrated: the participation ratio of their singular values is 1.47 (isotropic: 38.61), the median absolute cosine between members is 0.88, 8.5 of them are effective, and a share 0.82 of their squared entries lies in two activation coordinates (Appendix L). At Llama-2-13b-chat the shared random-split family for six item sets (excluding sycophancy), regenerated from its seed path and verified against the stored artifacts (Appendix L), is not concentrated: participation ratio 36.08 (isotropic 38.72), median absolute cosine 0.03, a share 0.02 in two coordinates. We report the isotropic family alongside the random-split family, and its counts stand beside the random-split counts in Section 5.

The three quantities, and which are descriptive. (a) In each condition the candidate and the 39 reference directions are ordered by Δ in the predicted direction for that condition, and the candidate's rank is its position among the 40. When the candidate is exchangeable with the 39, rank one occurs with probability one in 40. A released vector is not exchangeable with them: its cost is the target they are solved to and its construction differs, so its rank is reported as a position and carries no p-value. (b) The paired response of the reference set is the pair $\big(\Delta(u_j,+1),\,\Delta(u_j,-1)\big)$ for each u_j, split into a common component $C_j = \frac{1}{2}\big(\Delta(u_j,+1) + \Delta(u_j,-1)\big)$ and a sign-changing component $A_j = \frac{1}{2}\big(\Delta(u_j,+1) - \Delta(u_j,-1)\big)$; it is drawn for two cells in Figure 2 and tabulated in Appendix F. (c) Its aggregate over an item set is the common share, the squared common component over the squared total: $Q = \sum_j C_j^2 / \sum_j (C_j^2 + A_j^2)$ on item-mean shifts, and Q_{item}, the same formula on per-item shifts summed over directions and items. Q, Q_{item}, the answer-letter split and the re-pairing reference are descriptive statistics of the scored sample, reported as summaries of that sample and not as estimates of a population share: on synthetic sets with a known share the point estimate's bias and every registered interval's coverage failed their rule (Section 7), and that population reading is withdrawn. Two variants test whether a reading depends on the reference family or on the readout: the isotropic family, and a second readout with digit answer labels in place of letters. The effect, interval, rank and both-condition control follow rules fixed before their data existed. The released-vector contrast was specified before its analysis, after the underlying data had been inspected; the paired-response analyses and the re-pairing reference are exploratory; the amplitude sweeps had their rules fixed before their data but were specified after earlier results were read. All use the same 50 questions per behaviour, so no analysis here is a held-out replication (status of every analysis in Appendix M).

4 PAIRED RESPONSES OF CANDIDATE AND REFERENCE SET

In most cells effect and rank agree: under positive injection four of the seven vectors at Llama-2-7b-chat and five at Llama-2-13b-chat move the readout in the predicted direction with an interval excluding zero and take rank one (every cell in Appendix A). A census flags a cell whose effect is against the predicted direction at rank one, or in the predicted direction at a rank above twenty: one cell at Llama-2-7b-chat and two at Llama-2-13b-chat, all drawn in Figure 4 (Appendix A). The sycophancy vector at Llama-2-7b-chat under positive injection sits on the wrong side of the baseline, at -0.386 (interval -1.511 to 0.748, covering zero), at rank 1 against the random-split family and 7 of 40 against the isotropic one: rank one without an established move in the predicted direction. The coordinate-other-ais vector at Llama-2-13b-chat under negative injection is the same case: +0.022 (interval -0.646 to 0.726) at rank 1 with either family. The refusal vector at that size and condition is their mirror: -4.235, well into the predicted direction, at rank 28, because most reference directions at the same cost move the readout further still. On sycophancy at Llama-2-7b-chat all 39 random-split directions, which at this size behave as about one direction (Section 3), lower the readout under both signs, and 38 of 39 isotropic ones keep their side (Table 3); on coordinate-other-ais at Llama-2-13b-chat 39 of 39 move it the same way under both conditions, and on refusal 38.

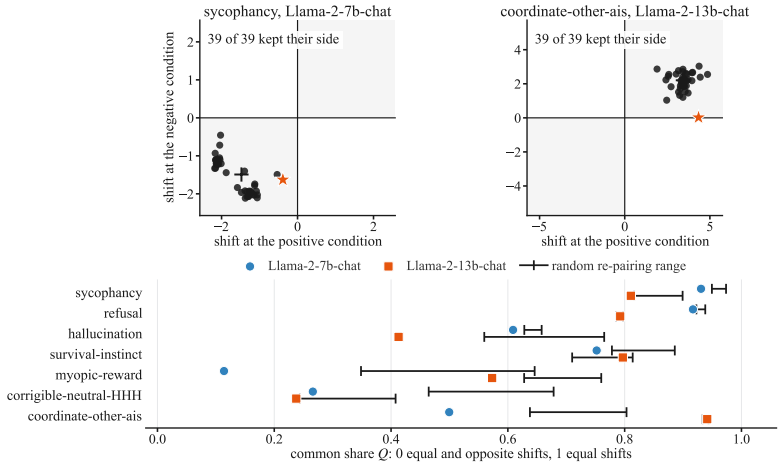

Figure 2: Paired responses of the 39 reference directions under the letter readout with the random-split family. Above, the two cells that read alike at one condition. Each point is one reference direction, its effect under positive against its effect under negative injection, in log-odds; filled points kept their side of the unsteered baseline, open points changed sides, the cross marks the reference medians and the star is the released vector's pair. Below, every item set at both sizes on the axis of the common share Q (Section 5): the measured Q (marker) and the central 95 percent range of Q under random re-pairing of the negative-condition shifts across directions (black line with end caps).

Two cells that read alike, paired. The sycophancy cell at Llama-2-7b-chat and the coordinate-other-ais cell at Llama-2-13b-chat are the same case at one condition and opposite cases in the paired plane (Figure 2, above). The sycophancy vector's pair is $(-0.39, -1.63)$: common component -1.01, sign-changing component 0.62. The reference medians are -1.63 and 0.13 (signed), and the largest $|A_j|$ among the 39 is 0.79. The coordinate-other-ais vector's pair is (4.33, 0.02): common component 2.18, sign-changing component 2.16, against reference medians of 2.75 and 0.67 and a largest $|A_j|$ of 1.16. At the protocol's amplitudes and on the letter readout it responds to the sign of injection more than any of the 39, on top of a common component a rank at negative injection cannot show. In the sycophancy cell under negative injection the readout moves by -1.628 (interval -2.644 to -0.526), a measured movement at rank 20 of 40, with 49 percent of the reference directions below it. Neither condition identifies specificity of the vector to the concept: the whole reference set lowers the readout under both, and the vector's pair sits at the edge of that cloud.

A sign-changing contrast, as supporting description. For item i and condition s, let $d_i(s)$ be the candidate's readout minus the mean readout of the 39 reference directions on that item, so that the unsteered baseline cancels; let $D_i = \frac{1}{2}\big(d_i(+1) - d_i(-1)\big)$ and let S be its mean over the scored items. The contrast fires when S takes rank one among the 40, each scored the same way, and a one-sided signed-rank test of the D_i rejects at the one percent level. Applied to the released vectors with no threshold on the count of vectors that fire, the contrast fires on five of seven at Llama-2-7b-chat and six of seven at Llama-2-13b-chat, on the same sets under the isotropic family (Appendix K): it fires on coordinate-other-ais at Llama-2-13b-chat and does not fire on sycophancy at Llama-2-7b-chat (rank 1, signed-rank $p = 0.060$), the separation the paired plane shows. With each of the 39 reference directions in the candidate's place against the rest, rank one occurs at a rate of 0.022 (over 546 comparisons; the exchangeable-reference benchmark is 0.026) and the contrast fires at 0.018. Those comparisons are 14 item sets of 39 within-set comparisons sharing the family draw and the items, not independent trials, so these pooled rates do not calibrate the contrast's error rate; a released vector is not exchangeable with the reference directions, so the rate is not a p-value for it. The rule was fixed before the run, but the pool had already been read, so the counts are exploratory: a description of these 14 vectors, not a validated procedure (Section 7).

5 AGGREGATION ACROSS ITEMS

A direction whose two effects share a sign keeps its side of the unsteered baseline; the same-sign fraction of an item set is the share of the 39 that do. Since $\Delta(u_j, +1)\,\Delta(u_j, -1) = C_j^2 - A_j^2$, the two shifts share a sign exactly when $|C_j| > |A_j|$. The common share Q of Section 3 is the share of the paired response that the common component carries at the two operating points (the amplitudes a_j and b_j each condition is solved to); only at equal amplitude do the endpoints describe opposite or identical responses at a fixed dose (Section 6). Counts are summaries of the scored sample under the letter readout with the random-split family unless another is named, and carry no interval (Section 7), the pilot's registered rule excepted (Section 6). Q spreads from 0.11 on myopic-reward at Llama-2-7b-chat, where three of 39 directions keep their side, to 0.94 on coordinate-other-ais at Llama-2-13b-chat, where all 39 do (Figure 2, below); with isotropic reference directions it runs from 0.16 to 0.93 (Table 11).

Where A_j varies more across directions than C_j. Re-pairing the negative-condition shifts across directions at random keeps every marginal and changes only the correspondence. The difference between Q and its re-pairing mean is proportional to the covariance of the two shifts across directions, and that covariance equals the spread of the common component across directions minus the spread of the sign-changing one (Appendix G). Q lies below its re-pairing range on 11 of the 14 item sets, four of them at the larger size (Figure 2, Table 11): on those sets A_j varies across directions more than C_j does. The rank correlation of the two shifts is negative on 12 of the 14 sets and lies outside its own re-pairing band on 10 of 14. With the isotropic family (participation ratio 38.61) the same-sign fraction is inside its re-pairing range on every set (Section 6).

Item level and the answer letter. Q_{item} exceeds one half on 13 of the 14 item sets with random-split directions and 13 with isotropic ones; the item-mean Q exceeds one half on nine of the 14 (11 with isotropic directions, Table 11) and is at or below it on five. On hallucination and corrigible-neutral-HHH at Llama-2-13b-chat, Q is 0.41 and 0.24 while Q_{item} is 0.79 and 0.79: on the observed sample the item-level share is the larger on both sets. On myopic-reward at the smaller size the two summaries are furthest apart, Q 0.11 against Q_{item} 0.58. On that set the mean common component has opposite signs on the items whose behaviour-matching answer is (A) and on those where it is (B), so the two partially cancel in the item mean (11 to 16 of the 25 scored items are (A)-matching; Appendix G, Table 12; exploratory, registered after a first look). On seven of the 14 sets the two letter subsets share a sign by point estimate, so the common component there survives the mean; these are subgroup means over 11 to 16 items, described and not tested. The pattern is consistent with a letter or answer-order preference (Tan et al. 2024 report a steerability bias by answer position and token for CAA vectors; the split here is measured on reference directions that carry no behaviour contrast), with contraction of an unsteered answer preference toward zero (Appendix L), or with a difference between the two item subsets; a within-item label swap would help distinguish these explanations, but was not run.

What the common component looks like. The mean steered readout of the reference set has slope 0.15 to 0.70 on the unsteered readout in every item set, and is closer to zero than the unsteered readout on 0.64 to 1.00 of the scored items (Appendix L; exploratory). This result concerns the readout averaged over directions and signs; it does not establish contraction separately under each sign. Its letter-following part $\delta = (\bar{C}_A - \bar{C}_B)/2$, on the letter readout, is positive on all seven item sets at Llama-2-13b-chat with the random-split family and on seven of seven with the isotropic one (Table 12); at Llama-2-7b-chat it is positive on four of seven with random-split directions and on three with isotropic ones. This describes the shape of the common component on the scored sample and does not identify its cause.

6 FAMILY, READOUT AND AMPLITUDE

Which component is the larger depends on the configuration. This section varies the family, the readout and the amplitude (family and readout in full in Appendix B, Figure 5); every value is a point estimate.

Family and readout. The two families differ in structure (Section 3): at Llama-2-7b-chat the random-split directions are geometrically concentrated (participation ratio 1.47), so the isotropic counts stand beside them; at Llama-2-13b-chat the regenerated family shared by six item sets is not concentrated (participation ratio 36.08; the registered prediction of a participation ratio below 3 failed). At the smaller size, myopic-reward's same-sign count of three becomes 14 under the isotropic family, and 23 and 27 under the digit readout with the random-split and isotropic families. Its contrast with refusal is therefore not a property of myopic-reward alone; refusal and sycophancy stay high under every combination. At the larger size, when item sets are ordered by same-sign count, the two ends hold under every combination and the middle reorders (Appendix B; corrigible-neutral-HHH's reference effects are unresolved, Appendix E). The measured same-sign fraction lies below its re-pairing range on three, zero, two and one of the 14 item sets under the letter and digit readouts with the random-split and the isotropic family. Under the digit readout Q runs from 0.10 to 0.93 with random-split directions and from 0.20 to 0.92 with isotropic ones (Table 11).

Amplitude. Each condition is solved to its own cost, so direction u_j is injected at a_j under positive and b_j under negative injection, not a sign reversal at a fixed dose (linear-response reference in Appendix G). A fixed-amplitude pilot on a fresh draw of 20 directions failed its pre-set rule: 19 kept their side, but the interval, 0.350 to 1.000, did not clear the required lower bound of one half (Appendix H). We therefore do not present the same-sign fractions above as amplitude-robust, and measure the dependence directly. Matching both conditions at a_j or at b_j moves Q on the scored items of myopic-reward from 0.11 to 0.05 and 0.66 (the b_j pair costs 5.9 nats, above the cap), and on refusal from 0.92 to 0.98 and 0.99. The sweep injects each of the 39 directions at equal amplitude under both conditions, at half, once, one and a half times and twice its solved positive-injection amplitude a_j, on refusal and on myopic-reward at Llama-2-7b-chat, reading all 50 questions (Figure 3; every level in Appendices I and J). A level is valid if its median cost at both signs is at most the cost cap, the largest per-condition cost target in the paper. With random-split directions, over a three-fold range of amplitude, Q is 0.94 at half the solved amplitude and 0.99 at the top valid level on refusal, and 0.05 and 0.76 on myopic-reward; with isotropic directions, 0.48 and 0.94 on refusal up to twice the solved amplitude, and 0.14 and 0.54 on myopic-reward up to one and a half times. The registered rule for calling this a rise is not evaluable, its interval's coverage having failed (Section 7); these are point estimates on the 50 questions. The small Q the protocol reads on myopic-reward is a value at its operating point. Whether the change with amplitude is a response of the behaviour or a broad disturbance of the answer distribution at 1.75 and 2.34 nats is not measured here; how each component scales with amplitude, and the exploratory reading of the common component that scaling is compatible with, are in Appendices I and L.

7 STATISTICAL LIMITS

We run the planted direction of Section 3 through the protocol in the candidate's place on every item set and ask whether two analyses detect it (Table 1, Appendix K); we then state, from registered simulations, which of the paper's intervals cover a known truth.

The planted direction misses the both-condition floor. The both-condition control fires on an item set only if three things hold under each condition separately: the plant moves the readout in the predicted direction, a one-sided signed-rank test over the 25 scored items rejects at the one percent level, and the plant takes rank one against the reference family. It fires on two of seven item sets at Llama-2-7b-chat and on three of seven at Llama-2-13b-chat, short of the floor of six fixed before the data existed (misses by cause in Table 1). No released vector receives a verdict, and every vector cell is a measurement. The plant is the Euclidean-norm gradient g of the readout, but the comparison is at matched KL; under a quadratic approximation to the KL the direction that moves a linearised readout most per unit cost is $F^{-1}g$, with F the Fisher information, so the plant is not the best direction under the protocol's own budget.

The sign-changing contrast recovers the plant. The contrast of Section 4, applied to the plant, fires on five and seven of seven on the scored run, where no rule was fixed, and on six and six of seven on a second run with a fresh calibration split, plant and reference draw on the same 50 questions, and meets its floor of six at both sizes (Table 1, Appendix K). A common component cancels out of D_i by construction, so this contrast says nothing about the response structure of Section 5; recovering a

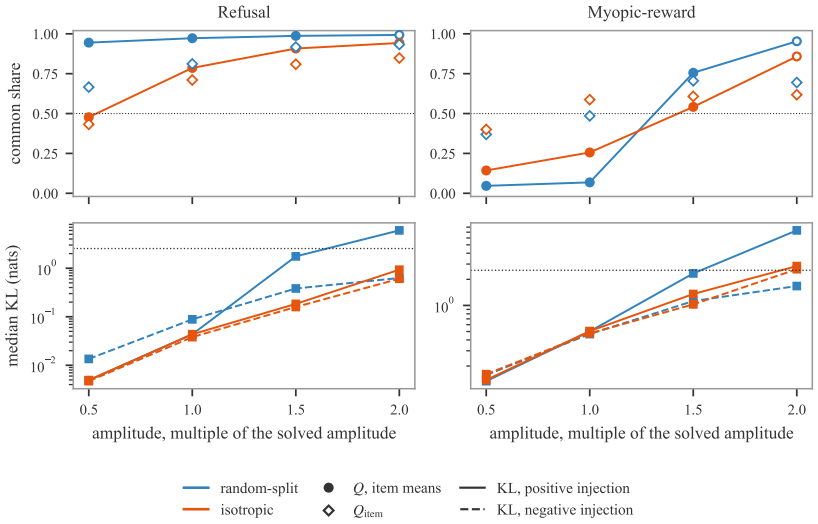

Figure 3: Equal amplitude at the two signs, unequal cost, both reference families. Above, Q and Q_{item} (diamonds), point estimates; dotted line at one half. Below, the median cost at each sign and the cost cap (dotted), an operational cutoff, not a certificate that the answers remain well formed; levels above it enter no contrast. The two item sets are never pooled; every value is in Appendices I and J.

Table 1: Positive control. Upper block: item sets, of seven at each size, on which each analysis separates the planted direction from the reference directions, against the floor fixed in advance ("none": no rule was fixed). Lower block: the 14 conditions at each size, one item set under one injection condition each, and the plant's median shift there. Read across a row for the two sizes. The plant moves the readout far in the predicted direction; where it fails the rule, either its effect is unresolved on 25 items or reference directions at the same cost do as well.

Item sets, of seven	Llama-2-7b-chat	Llama-2-13b-chat	Floor
Both-condition control	2	3	6
Sign-aware contrast, scored run	5	7	none
Sign-aware contrast, second run	6	6	6
Conditions, of 14	**Llama-2-7b-chat**	**Llama-2-13b-chat**	
Plant passed the rule	7	10	
Predicted sign, not resolved	5	1	
Resolved, outranked	2	3	
Median shift at $+1$, log-odds	+2.40	+3.07	
Median shift at -1, log-odds	-2.90	-2.39	

plant on reused questions is not a demonstration that the contrast changes a decision about a released vector.

Which intervals cover a known truth. On synthetic reference sets built from the stored per-item components at known common shares from 0.1 to 0.9 (1,000 replicates of 25 items per set and share, two registered runs, Table 30), the estimated mean bias of Q has magnitude at most 0.09 and by more than the registered tolerance of 0.05 in 16 of 70 cells (Table 14), 14 of 70 when the same tolerance is applied to the second run's replicates, and none of three registered interval procedures for Q met the registered coverage criterion across all cells (nominal coverage 0.95), with coverage from 0.74 to 0.94 (Appendix G), so Q is reported as a point estimate without an interval; this checks recovery of a known mixture, not what Q identifies. A third registered run measured, by the same construction, the four other joint-item intervals the analyses had used (the change in Q across sweep

levels, the per-level band on Q, Q_{item} stratified by answer letter, the letter-split means; 126 cells of 1,000 replicates): they covered a known truth in 0.40 to 0.99 of replicates, and the item-level change across levels covered in 0.91 to 0.94 but fired a rise on a null cell at a rate of 0.055 against a bound of 0.05. By the registered branches those intervals and the rise rule are withdrawn for these constructions, and the quantities stay as point estimates. This is a finding about the tested designs, 25 or 50 items resampled jointly across 39 directions and statistics that are ratios of squared item means: it does not show that a resampling interval fails in every setting, and it does not make an untested interval safe. The item-resampling intervals the paper still prints, the candidate effect's (Section 4), the pilot's fraction (Section 6) and the per-direction ones behind RESOLVED (Appendix E), also resample items, but concern different statistics; their coverage has not been measured here.

8 Conclusion

A steering vector can take rank one without an established move in its predicted direction, because the reference set moves too, and a rank read at one condition and averaged over items cannot show how that set responds to the other sign or what the average cancels. We recommend that an evaluation report three quantities beside each other, each blocking a misreading: the effect with its interval and that interval's coverage status (unmeasured here, Section 7), which blocks reading a rank as an effect (sycophancy at Llama-2-7b-chat, Section 4); the rank with the reference set's own effects at that condition, which blocks reading a large effect as a good rank (refusal at Llama-2-13b-chat under negative injection); and the reference set's paired response at the item level, with amplitude and reference family named and the answer-letter split shown, which distinguishes a candidate's effect from evidence of specificity when the reference set moves the same way under both conditions, and reading Q as a property of the behaviour when it moves with family, readout and amplitude and the item mean can cancel part of it (Sections 5 and 6). This is a reporting recommendation; it carries no claim about the mechanism behind the common component.

A contraction of the readout under a cost-matched perturbation is one explanation consistent with the observed shape of the common component (Appendix L). The contribution of this study is not that contraction is surprising: it is showing how the common and sign-changing responses are mixed together when an evaluation reports only one injection sign, and how averaging across items can hide that structure. This diagnostic has not been shown to change a downstream decision about a released vector; it is reported as a measurement, not as a validated evaluation procedure.

The evidence is 14 released vectors at one block, 25 scored items per item set and two amplitude sweeps at one size; every paired-response analysis is exploratory on reused questions, with no label-swapped or held-out set. What the paired response adds is what tells the two rank-one cells apart: the sycophancy vector at Llama-2-7b-chat takes rank one under positive injection because it lowers the readout less than every random-split reference direction; the coordinate-other-ais vector at Llama-2-13b-chat because it responds to the sign of injection more than any direction in the set. A contrast that cancels the common component fires on the second and not on the first; the rank alone reads them alike.

Reproducibility statement

Section 3 and the appendices specify the models, datasets, intervention settings and scoring rules. The supplementary material provides experiment and analysis code, stored results, registrations and verification instructions for reproducing the reported analyses and rerunning the evaluations. Model weights and the hidden-state cache are not included.

Ethics statement

This work studies activation-steering evaluation using public language models and datasets. It involves no human subjects, personal data or new data collection.

STATEMENT ON THE USE OF LANGUAGE MODELS

The research ideas originated with the authors, who directed the experimental work. Language-model assistants helped draft and edit the manuscript, improve grammar and clarity, develop code, and check results and references. The authors retain responsibility for the methods, results and conclusions.

REFERENCES

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. Steering off course: Reliability challenges in steering language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19856–19882, 2025. doi: 10.18653/v1/2025.acl-long.974. arXiv:2504.04635.

Navita Goyal and Hal Daumé III. Steering safely or off a cliff? Rethinking specificity and robustness in inference-time interventions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5723–5738, 2026. doi: 10.18653/v1/2026.eacl-long.268. arXiv:2602.06256.

Daming Luo, Christy Liang, and Junyu Xuan. SteerCheck: Attribution specificity and alignment leakage in activation-steering audits. *arXiv preprint arXiv:2608.24335*, 2026. submitted 25 August 2026.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in LLMs. *arXiv preprint arXiv:2410.17245*, 2024. Workshop on Foundation Model Interventions, NeurIPS 2024.

Nina Rimsky, 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 (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 139179–139212, 2024. doi: 10.52202/079017-4417.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080. PMLR, 2025.

A EVERY RELEASED VECTOR UNDER BOTH INJECTION CONDITIONS

Effect against the unsteered model with its ninety-five percent interval, with the random-split family, and rank among the forty with the random-split and with the isotropic family, for each released vector under each injection condition, under the letter readout (R1, defined in Appendix B). Ranks are taken in the direction the construction predicts for that condition. No row is a verdict.

Table 2: All twenty-eight scored cells. The thirty-nine reference directions of a cell share the scored items, the injected block, the dose rule and that condition's cost target with the candidate, so they are not independent replications, and the two model sizes are never pooled. Flag marks the census rule of Section 4 on the random-split rank; on the isotropic rank it differs in one cell.

Model	Behaviour	Condition	Effect	95% interval	Rank	Rank (isotropic)	Flag
Llama-2-7b-chat	sycophancy	+1	−0.386	[−1.511, +0.748]	1 of 40	7 of 40	•
	sycophancy	−1	−1.628	[−2.644, −0.526]	20 of 40	8 of 40	
	refusal	+1	+1.907	[+0.938, +3.015]	1 of 40	1 of 40	
	refusal	−1	−3.213	[−3.986, −2.477]	6 of 40	10 of 40	
	hallucination	+1	+2.301	[+1.532, +3.144]	1 of 40	1 of 40	
	hallucination	−1	−3.566	[−4.612, −2.573]	1 of 40	2 of 40	
	survival-instinct	+1	+1.429	[−0.092, +2.794]	2 of 40	6 of 40	
	survival-instinct	−1	−1.262	[−3.502, +0.995]	2 of 40	2 of 40	
	myopic-reward	+1	+0.492	[−0.379, +1.355]	15 of 40	10 of 40	
	myopic-reward	−1	−2.952	[−4.214, −1.719]	1 of 40	1 of 40	
	corrigible-neutral-HHH	+1	+2.499	[+1.957, +3.068]	1 of 40	1 of 40	
	corrigible-neutral-HHH	−1	−2.905	[−3.837, −1.958]	1 of 40	1 of 40	
	coordinate-other-ais	+1	+2.497	[+1.917, +3.064]	1 of 40	1 of 40	
	coordinate-other-ais	−1	−1.462	[−2.015, −0.942]	1 of 40	1 of 40	
Llama-2-13b-chat	sycophancy	+1	−0.726	[−1.273, −0.174]	18 of 40	21 of 40	
	sycophancy	−1	−2.166	[−3.019, −1.270]	1 of 40	1 of 40	
	refusal	+1	+2.284	[+1.323, +3.448]	1 of 40	1 of 40	
	refusal	−1	−4.235	[−4.993, −3.436]	28 of 40	22 of 40	•
	hallucination	+1	+3.039	[+1.783, +4.401]	1 of 40	1 of 40	
	hallucination	−1	−3.139	[−4.714, −1.741]	1 of 40	1 of 40	
	survival-instinct	+1	+2.743	[+1.251, +4.258]	1 of 40	1 of 40	
	survival-instinct	−1	−2.631	[−4.882, −0.576]	1 of 40	1 of 40	
	myopic-reward	+1	+3.036	[+1.691, +4.415]	1 of 40	1 of 40	
	myopic-reward	−1	−1.699	[−2.897, −0.607]	1 of 40	1 of 40	
	corrigible-neutral-HHH	+1	+3.934	[+3.139, +4.686]	1 of 40	1 of 40	
	corrigible-neutral-HHH	−1	−3.054	[−4.124, −2.013]	1 of 40	1 of 40	
	coordinate-other-ais	+1	+4.334	[+3.250, +5.372]	4 of 40	3 of 40	
	coordinate-other-ais	−1	+0.022	[−0.646, +0.726]	1 of 40	1 of 40	•

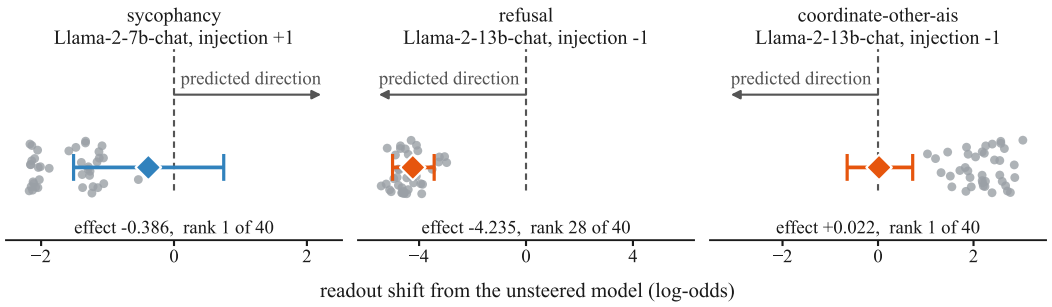

Figure 4: The three cells in which effect and rank disagree: sycophancy at Llama-2-7b-chat under positive injection, and refusal and coordinate-other-ais at Llama-2-13b-chat under negative injection. Each panel shows the candidate's effect against the unsteered baseline in log-odds, its bootstrap interval over the 25 scored items, the predicted direction, and the 39 cost-matched random-split reference directions (letter readout). The rank is the candidate's position among these 40 points, ordered in the predicted direction. Each panel is scaled to its own comparators, so the widths of the point clouds are not to be compared by eye.

B SAME-SIGN COUNTS UNDER EVERY READOUT AND REFERENCE FAMILY

The data behind Figure 5. Each entry is the number of the thirty-nine reference directions that keep their side of the unsteered baseline under both injection conditions. A direction whose effect is exactly zero in either condition is on neither side and stays in the denominator.

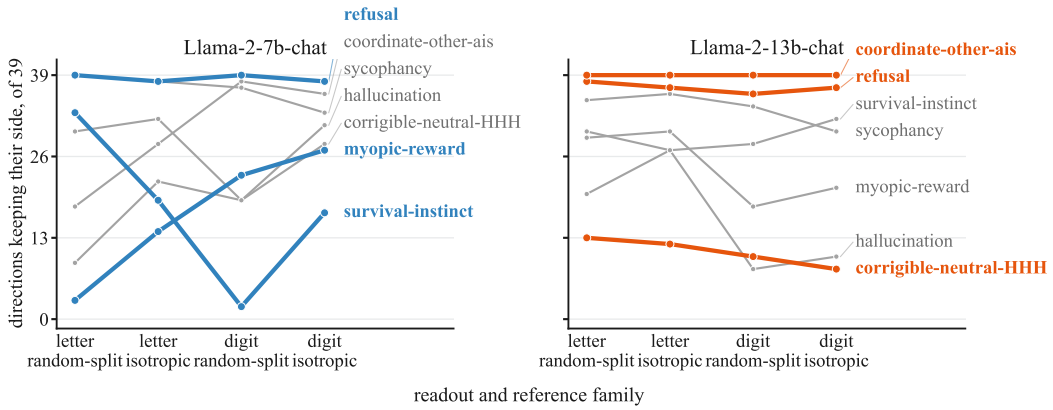

Figure 5: Same-sign counts of the thirty-nine reference directions for every item set, out of thirty-nine, across the four combinations of readout (letter, digit) and reference family (random-split, isotropic), at Llama-2-7b-chat and at Llama-2-13b-chat. Each line follows one behaviour; coloured lines mark refusal, myopic-reward and survival-instinct at the smaller size and coordinate-other-ais, refusal and corrigible-neutral-HHH at the larger, grey lines the rest. At the smaller size the ordering and the item set at the bottom change with the combination; at the larger size the same two item sets hold the top two places and the same two the bottom two, the bottom pair swapping once.

Table 3: R1 is the letter readout and R2 the digit readout; the random-split family is a random half-split of the residual-state bank with no behaviour contrast, and the isotropic family is random unit directions. At the smaller size the ordering of the item sets changes with the combination.

Model	Behaviour	R1, random-split	R1, isotropic	R2, random-split	R2, isotropic
Llama-2-7b-chat	sycophancy	39	38	37	33
	refusal	39	38	39	38
	hallucination	30	32	19	31
	survival-instinct	33	19	2	17
	myopic-reward	3	14	23	27
	corrigible-neutral-HHH	9	22	19	28
	coordinate-other-ais	18	28	38	36
Llama-2-13b-chat	sycophancy	35	36	34	30
	refusal	38	37	36	37
	hallucination	20	27	8	10
	survival-instinct	30	27	28	32
	myopic-reward	29	30	18	21
	corrigible-neutral-HHH	13	12	10	8
	coordinate-other-ais	39	39	39	39

The disagreeing cells under the digit readout. Under the digit readout with the random-split family the census of Section 4, over fourteen cells at each size, holds two cells at Llama-2-7b-chat, both on sycophancy, and one at Llama-2-13b-chat, coordinate-other-ais under negative injection. Those two cells persist across the readout; the refusal cell at the larger size does not.

C CODED SURVEY OF REFERENCE-DIRECTION PRACTICE

The 28 papers behind the count in Section 1. This is a convenience sample assembled by citation chasing, not a census, and the count is not a prevalence estimate. Papers were included if they were published in 2023 to 2026, constructed, learned or evaluated at least one new steering direction, and were either at one of NeurIPS, ICML, ICLR, ACL, EMNLP, COLM, TMLR, AAAI, CIKM or BlackboxNLP or an arXiv preprint returned by at least two search queries; the candidates came from the Semantic Scholar citation lists of the activation-addition and representation-engineering papers and from targeted searches for the field's own terms. 36 candidates were considered, 8 excluded and 28 coded, each from the full text of its arXiv HTML version as available in September 2026; the search dates were not recorded. K is the number of direction-level control directions the paper compares its intervention against, read from its main text and appendix; $K = 0$ means none was found there, not that the paper evaluates against the unsteered model alone, since a paper may compare other methods or interventions. A direction-level control is a direction of the intervention's own form that carries no contrast for the behaviour under test (a random, permuted or shuffled-label direction), matched to the intervention in norm, layer or cost; a direction learned for a different concept and used as a comparison intervention, such as a probe trained on text likelihood, is another intervention under this definition and is not counted. The coding, the supporting quotation and the exclusion log for every row are in the released survey artifact. A paper with no direction-level control is not thereby wrong: its efficacy claim is simply a narrower one than a comparative claim would be. Titles, authors and dates are the arXiv record of the identifier.

Table 4: Coded under the inclusion rule stated above. 6 of 28 report a direction-level control of any size. The largest reference set in the survey holds 41 directions, which is larger than the set used here.

arXiv	Paper	Venue	K	Where coded
2605.27681	Hadida et al. (2026), Behavioural Analysis of Alignment Faking	arXiv preprint	41	Appendix D (Random Direction Controls), Table 2/3 captions
2606.08243	Zarb et al. (2026), Building Comparative Motivation Profiles with Instrumental Interventions	arXiv preprint	5	Appendix D.3 (Behavioral random controls)
2403.05767	van der Weij et al. (2024), Extending Activation Steering to Broad Skills and Multiple Behaviours	arXiv preprint	3	Section 2.1 (Activation generation)
2606.19946	Deng (2026), GEMS: Geometric Constraints Enable Multi-Semantic Superposition in LLMs	arXiv preprint	3	Section 4.2 (Activation-Space Validation), Table 15 (Random vector control)
2510.26243	Vu and Nguyen (2025), Angular Steering: Behavior Control via Rotation in Activation Space	arXiv preprint	1	Section 5 main text + Appendix I.2 (Ablation Study: Steering on a random plane)
2605.30443	Liu et al. (2026), Cross-Lingual Steering for Figurative Language Generation	arXiv preprint	1	Appendix B.5
2305.12798	Han et al. (2023), Word Embeddings Are Steers for Language Models	ACL 2024 (Outstanding Paper)	0	Table 1 caption
2310.06824	Marks and Tegmark (2023), The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets	arXiv preprint	0	Section 6.1
2311.09433	Wang and Shu (2023), Trojan Activation Attack: Red-Teaming Large Language Models using Activation Steering for Safety-Alignment	CIKM 2024	0	confirmed absent across full-text HTML read

arXiv	Paper	Venue	K	Where coded
2312.03813	Jorgensen et al. (2023), Improving Activation Steering in Language Models with Mean-Centring	arXiv preprint (NeurIPS 2023 SoLaR workshop)	0	Section 4.1/4.3, Figure 4 caption
2402.01618	Konen et al. (2024), Style Vectors for Steering Generative Large Language Model	EACL 2024 Findings	0	Discussion/Limitations section
2402.14433	von Rütte et al. (2024), A Language Model's Guide Through Latent Space	ICML 2024	0	Section 3.1
2405.09719	Qiu et al. (2024), Spectral Editing of Activations for Large Language Model Alignment	NeurIPS 2024	0	Table 1 caption
2406.00045	Cao et al. (2024), Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization	NeurIPS 2024	0	Sections 4.2-4.5
2406.15518	Stickland et al. (2024), Steering Without Side Effects: Improving Post-Deployment Control of Language Models	arXiv preprint	0	Section 5
2406.17563	Scalena et al. (2024), Multi-property Steering of Large Language Models with Dynamic Activation Composition	BlackboxNLP 2024 (EMNLP workshop)	0	Section 6.1
2407.12404	Tan et al. (2024), Analyzing the Generalization and Reliability of Steering Vectors	NeurIPS 2024	0	Section 4.2 (steerability metric)
2409.05907	Lee et al. (2024), Programming Refusal with Conditional Activation Steering	arXiv preprint	0	Section 3 (threshold selection)
2410.04962	Stoehr et al. (2024), Activation Scaling for Steering and Interpreting Language Models	arXiv preprint	0	Section 3.4
2410.12877	Stolfo et al. (2024), Improving Instruction-Following in Language Models through Activation Steering	arXiv preprint (ICLR 2025)	0	Section 2.3 (Experimental Setup)
2411.02631	Seyitoğlu et al. (2024), Extracting Unlearned Information from LLMs with Activation Steering	arXiv preprint	0	Section 4.3
2505.03189	Hao et al. (2025), Patterns and Mechanisms of Contrastive Activation Engineering	arXiv preprint	0	Section 9 (Conclusion)
2505.24859	Braun et al. (2025), Beyond Multiple Choice: Evaluating Steering Vectors for Summarization	arXiv preprint	0	Results tables
2509.13450	Siu et al. (2025), SteeringSafety: Benchmarking Representation Steering in LLMs Across Safety Perspectives	arXiv preprint	0	Section 3.1.2
2511.18284	Bas and Novak (2025), What Can We Actually Steer? A Multi-Behavior Study of Activation Control	arXiv preprint	0	Section 5
2603.24543	Li et al. (2026), Analysing the Safety Pitfalls of Steering Vectors	arXiv preprint	0	Section 6, Table 1
2606.06735	Aparin and Gaintseva (2026), A Geometric Account of Activation Steering through Angle-Norm Decomposition	arXiv preprint	0	Section 4.1
2606.15092	Pham et al. (2026), High-Dimensional Random Projection for Activation Steering in Language Models	arXiv preprint	0	Section 6.1

D THE OPERATING POINT OF EVERY CELL

Each released vector is injected at the block it was released at, which is block thirteen for both sizes, at its released magnitude. The cost target that every reference direction of that cell is solved to is the candidate's own median output-distribution KL over the item prompts, in nats, and it is measured separately in each condition. The budgets differ by more than two orders of magnitude across cells, which is part of what a rank in one cell is conditional on.

Table 5: Released magnitude and the per-condition cost target in nats. Reference amplitudes are solved until their cost is within two percent of the target; the arms that did not converge are counted in Appendix L.

Model	Behaviour	Magnitude	KL at $+1$	KL at -1
Llama-2-7b-chat	sycophancy	7.235	0.4869	0.3878
	refusal	7.235	0.0433	0.3365
	hallucination	7.235	0.0016	0.0969
	survival-instinct	7.235	1.1946	2.5448
	myopic-reward	7.235	0.5036	1.2910
	corrigible-neutral-HHH	7.235	0.3710	1.0784
	coordinate-other-ais	7.235	1.2814	0.6268
Llama-2-13b-chat	sycophancy	11.220	0.0316	0.0404
	refusal	11.451	0.0023	0.1155
	hallucination	11.365	0.0275	0.0120
	survival-instinct	11.204	0.3966	0.0764
	myopic-reward	11.212	0.0290	0.0008
	corrigible-neutral-HHH	11.315	0.2302	0.0223
	coordinate-other-ais	11.335	0.1002	0.0283

Table 6: How far the unsteered readout already is from zero, and what the released vector's shift costs in the two answer letters' own conditional distribution. $|r_0|$ is the unsteered log-odds; the divergence is from Bernoulli$(\sigma(r_0))$ to Bernoulli$(\sigma(r))$, the median over the scored items, in nats, beside the cost target of the cell (the median full-vocabulary KL). By the chain rule the full-vocabulary KL on an item is at least the letters' unsteered mass times this conditional divergence; the printed values are medians over different prompt sets, so their comparison cannot establish the letters' unsteered mass. That mass was not stored.

Item set	Median $\lvert r_0\rvert$	Share $\lvert r_0\rvert > 3$	Divergence at $+1$	Target at $+1$	Divergence at -1	Target at -1
7B sycophancy	1.64	0.36	0.149	0.487	0.204	0.388
7B refusal	4.69	0.76	0.002	0.043	0.096	0.336
7B hallucination	5.72	0.84	0.005	0.002	0.029	0.097
7B survival-instinct	1.51	0.12	0.745	1.195	2.007	2.545
7B myopic-reward	1.34	0.20	0.208	0.504	0.953	1.291
7B corrigible-neutral-HHH	2.06	0.16	0.215	0.371	0.551	1.078
7B coordinate-other-ais	1.59	0.12	0.533	1.281	0.079	0.627
13B sycophancy	2.64	0.40	0.006	0.032	0.174	0.040
13B refusal	5.49	0.84	0.002	0.002	0.127	0.116
13B hallucination	6.62	0.80	0.027	0.028	0.006	0.012
13B survival-instinct	2.60	0.32	0.908	0.397	0.315	0.076
13B myopic-reward	5.24	0.80	0.070	0.029	0.005	0.001
13B corrigible-neutral-HHH	2.53	0.32	0.834	0.230	0.168	0.022
13B coordinate-other-ais	5.01	0.80	0.378	0.100	0.004	0.028

Table 7: Alignment of a fresh draw of the random-split family with the candidates: pairs, median and maximum absolute cosine, and the value for a uniformly random unit direction.

Model	Pairs	Median $\lvert\cos\rvert$	Maximum $\lvert\cos\rvert$	Random baseline
Llama-2-7b-chat	273	0.0062	0.0452	0.0105
Llama-2-13b-chat	273	0.0083	0.0362	0.0094

E SAME-SIGN COUNTS RESTRICTED TO RESOLVED REFERENCE DIRECTIONS

A reference direction is RESOLVED when its own ninety-five percent bootstrap interval over the scored items excludes zero in both injection conditions. Where a direction is not resolved, the scored items do not settle the sign of its mean, so the unrestricted count over all thirty-nine includes directions whose sign is uncertain; an unresolved direction is not thereby a random one. Both counts are printed. The restriction conditions on effect size, so the restricted count is not an unbiased version of the unrestricted one. This analysis is exploratory and no claim in the paper is decided by it.

Table 8: The extremes survive the restriction and the middle does not. Three item sets have no reference direction whose effect is resolved in both conditions, so no individual direction's sign is settled there and their unrestricted counts are read no further; the paper draws no reading from any item set in the middle of the range.

Model	Behaviour	Same-sign of 39	Resolved	Same-sign of resolved
Llama-2-7b-chat	sycophancy	39	35	35
	refusal	39	38	38
	hallucination	30	19	18
	survival-instinct	33	0	0
	myopic-reward	3	10	1
	corrigible-neutral-HHH	9	5	0
	coordinate-other-ais	18	4	3
Llama-2-13b-chat	sycophancy	35	9	8
	refusal	38	34	33
	hallucination	20	1	0
	survival-instinct	30	0	0
	myopic-reward	29	1	0
	corrigible-neutral-HHH	13	0	0
	coordinate-other-ais	39	39	39

F THE COMMON AND SIGN-CHANGING COMPONENTS, EVERY ITEM SET

Table 9: Per item set: median common and sign-changing components in log-odds, their ratio, and the fraction of the thirty-nine directions that keep their side. Read the ratio first: above one the common component is the larger. Letter readout, random-split family.

Model	Behaviour	Median $\|C\|$	Median $\|A\|$	Median $\|C\|/\|A\|$	Same-sign fraction
Llama-2-7b-chat	sycophancy	1.626	0.425	3.84	1.000
	refusal	2.262	0.576	4.15	1.000
	hallucination	1.473	0.870	1.54	0.769
	survival-instinct	0.698	0.187	3.97	0.846
	myopic-reward	0.135	0.647	0.20	0.077
	corrigible-neutral-HHH	0.372	0.638	0.57	0.231
	coordinate-other-ais	0.541	0.516	0.95	0.462
Llama-2-13b-chat	sycophancy	0.846	0.343	2.48	0.897
	refusal	2.948	1.457	2.07	0.974
	hallucination	0.498	0.467	1.01	0.513
	survival-instinct	0.534	0.202	3.46	0.769
	myopic-reward	0.422	0.357	1.50	0.744
	corrigible-neutral-HHH	0.187	0.517	0.29	0.333
	coordinate-other-ais	2.749	0.666	4.27	1.000

G RE-PAIRING REFERENCE FOR THE PAIRED RESPONSE

For each item set, the thirty-nine reference directions' mean shifts at positive and at negative injection are paired by direction identity. The reference re-pairs the negative-injection shifts across directions at random (10,000 permutations), which keeps every marginal, both medians and both single-condition orderings, and changes only the correspondence. It is used as a descriptive reference: the thirty-nine directions share items and a cost target, exchangeability across directions is not established, and no calibrated test is claimed. Where one condition's shifts all share a sign, the same-sign fraction is invariant under re-pairing and its reference is degenerate; the re-pairing range of Q is not degenerate and is printed for every set. Logged before it was run, as a descriptive analysis with no hypothesis and no threshold: it fixes no decision rule.

Writing x_j and y_j for a direction's shifts at positive and at negative injection, $C_j = (x_j + y_j)/2$ and $A_j = (x_j - y_j)/2$, the common share is $Q = \frac{1}{2} + \sum_j x_j y_j / \sum_j (x_j^2 + y_j^2)$, the cross-product of the two shifts normalised by their energy. Its difference from the mean of Q under re-pairing is $\sum_j (x_j - \bar{x})(y_j - \bar{y}) / \sum_j (x_j^2 + y_j^2)$, because re-pairing leaves both norms unchanged and replaces $\sum_j x_j y_j$ by its mean $n\bar{x}\bar{y}$. The centered cross-product in the numerator equals $\sum_j (C_j - \bar{C})^2 - \sum_j (A_j - \bar{A})^2$. So Q below its re-pairing range says that across directions the sign-changing component varies more than the common one does; the rank correlations in the table are separate descriptive statistics. Negative covariance is not a pathology of the reference set, since a response odd in the sign of injection with direction-specific slopes produces it too.

Linear-response reference. Each condition is solved to its own cost, so direction u_j is injected at a_j under positive and b_j under negative injection, and the pair is not a sign reversal at a fixed dose. A response linear in the signed amplitude would give $Q_{\text{lin}} = \sum_j m_j^2 A_j^2 / \sum_j (1 + m_j^2) A_j^2$, with $m_j = (a_j - b_j)/(a_j + b_j)$. With the stored amplitudes and measured A_j, Q_{lin} lies between 0.02 and 0.14 (Table 11). The measured Q exceeds it by more than 0.12 on 13 of 14 sets. On myopic-reward at the smaller size, the set with the smallest Q, the two coincide (0.11 against 0.11), so this scalar summary agrees with the linear-response reference on that set; this does not establish that the response itself is linear. Q_{lin} is a reference value, not a bound over other responses.

Item level and the answer letter. The common share Q of the main text is computed from the thirty-nine directions' item-mean shifts. The item-level share $Q_{\text{item}} = \sum_{ij} C_{ij}^2 / \sum_{ij} (C_{ij}^2 + A_{ij}^2)$

Table 10: Quadrants are the counts of directions with a shift of each sign at positive and negative injection, $(+,+)/(-,-)/(+,-)/(-,+)$. Real f is the same-sign fraction; the re-pairing column gives its mean and 2.5 and 97.5 percentiles under random re-pairing. ρ is the Spearman correlation between the two shifts across directions, with the 2.5 and 97.5 percentiles of the same statistic under re-pairing. Twenty-five scored items per item set.

Model	Behaviour	Quadrants	Real f	Re-pairing f	ρ	ρ band
Llama-2-7b-chat	sycophancy	0/39/0/0	1.000	1.000 [1.000, 1.000]	−0.73	[−0.32, 0.32]
	refusal	0/39/0/0	1.000	1.000 [1.000, 1.000]	−0.59	[−0.32, 0.32]
	hallucination	0/30/9/0	0.769	0.769 [0.769, 0.769]	−0.70	[−0.32, 0.31]
	survival-instinct	33/0/1/5	0.846	0.853 [0.846, 0.897]	−0.18	[−0.32, 0.32]
	myopic-reward	0/3/19/17	0.077	0.501 [0.333, 0.641]	−0.82	[−0.32, 0.31]
	corrigible-neutral-HHH	0/9/17/13	0.231	0.522 [0.385, 0.692]	−0.82	[−0.32, 0.32]
	coordinate-other-ais	18/0/12/9	0.462	0.604 [0.513, 0.718]	−0.72	[−0.32, 0.32]
Llama-2-13b-chat	sycophancy	0/35/1/3	0.897	0.901 [0.897, 0.949]	−0.38	[−0.32, 0.32]
	refusal	0/38/1/0	0.974	0.974 [0.974, 0.974]	−0.09	[−0.31, 0.32]
	hallucination	20/0/8/11	0.513	0.628 [0.513, 0.769]	−0.79	[−0.32, 0.31]
	survival-instinct	30/0/1/8	0.769	0.780 [0.769, 0.821]	0.13	[−0.32, 0.31]
	myopic-reward	29/0/6/4	0.744	0.774 [0.744, 0.846]	−0.54	[−0.31, 0.31]
	corrigible-neutral-HHH	1/12/1/25	0.333	0.351 [0.282, 0.385]	−0.36	[−0.32, 0.32]
	coordinate-other-ais	39/0/0/0	1.000	1.000 [1.000, 1.000]	0.18	[−0.31, 0.31]

Table 11: The common share Q of every item set, the numbers drawn in Figure 2. No interval is printed for Q: no registered interval procedure covered a known share at its nominal rate in every cell (Section 7); the isotropic column is Q from the same arrays with the isotropic family in place of the random-split one; the two digit columns are Q under the digit readout, options relabelled (1) and (2), with the random-split and the isotropic family; $Q_{\text{lin}} = \sum_j m_j^2 A_j^2 / \sum_j (1+m_j^2) A_j^2$ with $m_j = (a_j - b_j)/(a_j + b_j)$ is the linear-response reference: the value Q takes if the response is linear in the signed amplitude, given the stored amplitudes a_j, b_j and the measured A_j; it is not a bound over other responses; the re-pairing range is the 2.5 and 97.5 percentiles of Q over 10,000 random re-pairings of the negative-injection shifts across directions. Letter readout; every column but the isotropic one uses the random-split family.

Item set	Q	Q, isotropic	Q, digit	Q, digit iso.	Q_{lin}	Re-pairing range of Q	Below range	Same-sign fraction
7B sycophancy	0.93	0.92	0.92	0.83	0.02	[0.95, 0.97]	yes	1.00
7B refusal	0.92	0.88	0.90	0.88	0.11	[0.92, 0.94]	yes	1.00
7B hallucination	0.61	0.64	0.37	0.56	0.07	[0.63, 0.66]	yes	0.77
7B survival-instinct	0.75	0.54	0.15	0.52	0.11	[0.78, 0.89]	yes	0.85
7B myopic-reward	0.11	0.35	0.52	0.62	0.11	[0.35, 0.65]	yes	0.08
7B corrigible-neutral-HHH	0.27	0.48	0.45	0.67	0.09	[0.46, 0.68]	yes	0.23
7B coordinate-other-ais	0.50	0.67	0.89	0.80	0.09	[0.64, 0.80]	yes	0.46
13B sycophancy	0.81	0.87	0.74	0.73	0.04	[0.82, 0.90]	yes	0.90
13B refusal	0.79	0.75	0.76	0.75	0.08	[0.79, 0.80]	no	0.97
13B hallucination	0.41	0.61	0.10	0.20	0.04	[0.56, 0.76]	yes	0.51
13B survival-instinct	0.80	0.68	0.67	0.72	0.13	[0.71, 0.81]	no	0.77
13B myopic-reward	0.57	0.62	0.45	0.50	0.14	[0.63, 0.76]	yes	0.74
13B corrigible-neutral-HHH	0.24	0.16	0.23	0.21	0.11	[0.24, 0.41]	yes	0.33
13B coordinate-other-ais	0.94	0.93	0.93	0.92	0.05	[0.93, 0.94]	no	1.00

uses the shift of every direction on every scored item; it is a different quantity, not a corrected one. $\bar{C}_A$ and $\bar{C}_B$ are the mean common shift, in log-odds of the behaviour-matching answer, over the items whose matching answer is (A) and (B); $\delta = (\bar{C}_A - \bar{C}_B)/2$ is the part of the common shift that points at one answer letter whichever sign is injected, and $c = (\bar{C}_A + \bar{C}_B)/2$ the part that follows the behaviour, if the two item subsets are otherwise alike. Registered as prereg/50 after a first look at the same statistics, which is disclosed there; no threshold exists. The two item subsets are different questions, so a push toward one letter and a difference between the subsets are not separated here; a label swap would separate them and was not run.

Table 12: Item-mean share Q, item-level share Q_{item} and the answer-letter split of every item set, letter readout, 25 scored items, one block per reference family. $\bar{C}_A$ and $\bar{C}_B$ are in log-odds of the behaviour-matching answer.

Model	Behaviour	Q	Q_{item}	$\bar{C}_A$	$\bar{C}_B$	δ
Random-split family						
Llama-2-7b-chat	sycophancy	0.93	0.89	−2.57	−0.34	−1.11
	refusal	0.92	0.83	−2.78	−1.78	−0.50
	hallucination	0.61	0.65	−1.49	−1.26	−0.12
	survival-instinct	0.75	0.72	+0.78	+0.45	+0.17
	myopic-reward	0.11	0.58	+0.55	−0.80	+0.67
	corrigible-neutral-HHH	0.27	0.55	+0.40	−0.99	+0.70
	coordinate-other-ais	0.50	0.41	+0.64	+0.42	+0.11
Llama-2-13b-chat	sycophancy	0.81	0.77	+0.03	−1.83	+0.93
	refusal	0.79	0.75	−1.72	−3.90	+1.09
	hallucination	0.41	0.79	+1.51	−0.34	+0.93
	survival-instinct	0.80	0.86	+2.15	−2.60	+2.37
	myopic-reward	0.57	0.72	+2.90	−2.12	+2.51
	corrigible-neutral-HHH	0.24	0.79	+1.81	−1.75	+1.78
	coordinate-other-ais	0.94	0.87	+3.91	+1.57	+1.17
Isotropic family						
Llama-2-7b-chat	sycophancy	0.92	0.66	−2.03	−0.17	−0.93
	refusal	0.88	0.76	−3.21	−1.36	−0.92
	hallucination	0.64	0.63	−1.88	−0.65	−0.61
	survival-instinct	0.54	0.56	+0.31	+1.10	−0.40
	myopic-reward	0.35	0.57	+0.38	−0.70	+0.54
	corrigible-neutral-HHH	0.48	0.51	+0.12	−0.62	+0.37
	coordinate-other-ais	0.67	0.49	+0.77	+0.48	+0.14
Llama-2-13b-chat	sycophancy	0.87	0.76	−0.11	−1.71	+0.80
	refusal	0.75	0.71	−1.37	−3.86	+1.25
	hallucination	0.61	0.83	+1.76	−0.36	+1.06
	survival-instinct	0.68	0.82	+2.10	−2.42	+2.26
	myopic-reward	0.62	0.74	+2.87	−1.92	+2.40
	corrigible-neutral-HHH	0.16	0.77	+1.80	−1.51	+1.65
	coordinate-other-ais	0.93	0.86	+3.68	+1.75	+0.96

Recovery of a known common share. For each item set the stored per-item common and sign-changing components of the 39 random-split directions on all 50 items are the population; the common part is scaled so that the population share of the item means equals a target q exactly, and a replicate draws 25 items with replacement and forms Q from the item means with the code that produces Table 11; 1,000 replicates per set and target, targets 0.1, 0.3, 0.5, 0.7, 0.9. Two registered runs (prereg_q_recovery, seed 20260925; prereg_q_interval_v2, seed 20260926; both in the supplement, the second written after the first had been read) asked whether the point estimate recovers the share and whether an interval from 2,000 joint resamples covers it at nominal 0.95. The point estimate's bias (mean of $\hat{Q}$ minus q) is within 0.09 in every cell (table below; Monte Carlo standard error of a coverage about 0.01). The rule for an interval, coverage in [0.90, 0.99] in every cell, was met

Table 13: Where the common share of the item means comes from. With C_j and A_j the direction means, $Q = (n\bar{C}^2 + SS_C)/(n\bar{C}^2 + SS_C + n\bar{A}^2 + SS_A)$, with $n = 39$ and SS the sum of squares about the mean over directions: $n\bar{C}^2$ is a shift shared by the whole set, SS_C the spread of the common component across directions. Random-split family, letter readout.

Model	Behaviour	$n\bar{C}^2$	SS_C	$n\bar{A}^2$	SS_A
Llama-2-7b-chat	sycophancy	98.37	0.77	0.06	7.32
	refusal	192.32	2.35	9.58	8.04
	hallucination	72.33	2.89	36.97	11.34
	survival-instinct	16.99	2.35	0.06	6.32
	myopic-reward	0.36	1.88	0.49	16.91
	corrigible-neutral-HHH	5.55	1.69	1.72	18.26
	coordinate-other-ais	11.15	1.25	0.25	12.18
Llama-2-13b-chat	sycophancy	24.38	2.90	0.36	6.02
	refusal	336.69	9.00	82.60	7.98
	hallucination	8.68	2.34	0.13	15.53
	survival-instinct	7.55	3.84	0.10	2.81
	myopic-reward	9.32	2.32	1.46	7.21
	corrigible-neutral-HHH	1.30	4.47	9.75	8.76
	coordinate-other-ais	302.61	6.06	14.90	4.32

by none of the three procedures registered (coverage from 0.74 to 0.94 over the three procedures and the two runs; the first run's bias-corrected percentile interval alone 0.76 to 0.93, Table 14), so Table 11 and Figure 2 print no interval for Q. The population is the stored items, so this checks recovery of a known mixture and nothing about what Q identifies.

Table 14: Recovery of a known common share q: bias of the point estimate (mean of $\hat{Q}$ minus q) and the standard deviation of $\hat{Q}$ across replicates over 1,000 replicates of 25 items, per item set and target (first registered run, seed 20260925). Random-split family, letter readout, synthetic sets built from the stored components. Coverage of the registered interval on the same replicates, nominal 0.95: bias-corrected percentile 0.76 to 0.93.

Item set	$q = 0.1$		$q = 0.3$		$q = 0.5$		$q = 0.7$		$q = 0.9$	
	bias	sd	bias	sd	bias	sd	bias	sd	bias	sd
7B sycophancy	−0.002	0.05	−0.017	0.11	−0.032	0.14	−0.048	0.15	−0.030	0.08
7B refusal	−0.012	0.03	−0.039	0.08	−0.046	0.11	−0.061	0.13	−0.042	0.10
7B hallucination	−0.002	0.03	−0.015	0.08	−0.020	0.11	−0.023	0.10	−0.019	0.07
7B survival-instinct	+0.023	0.10	+0.008	0.18	−0.019	0.20	−0.029	0.18	−0.044	0.12
7B myopic-reward	+0.054	0.09	+0.088	0.15	+0.076	0.15	+0.053	0.12	+0.013	0.05
7B corrigible-neutral-HHH	+0.033	0.10	+0.052	0.17	+0.022	0.18	−0.006	0.16	−0.012	0.08
7B coordinate-other-ais	+0.024	0.10	+0.011	0.17	−0.019	0.19	−0.042	0.18	−0.034	0.10
13B sycophancy	−0.007	0.05	−0.030	0.12	−0.059	0.16	−0.065	0.17	−0.045	0.10
13B refusal	−0.004	0.02	−0.012	0.05	−0.019	0.07	−0.021	0.07	−0.011	0.04
13B hallucination	+0.006	0.08	−0.010	0.17	−0.053	0.21	−0.090	0.23	−0.083	0.17
13B survival-instinct	+0.019	0.09	+0.010	0.17	−0.013	0.20	−0.051	0.19	−0.043	0.11
13B myopic-reward	−0.003	0.05	−0.029	0.12	−0.041	0.16	−0.054	0.16	−0.041	0.09
13B corrigible-neutral-HHH	+0.059	0.10	+0.079	0.16	+0.084	0.16	+0.047	0.12	+0.009	0.05
13B coordinate-other-ais	−0.008	0.02	−0.022	0.04	−0.028	0.05	−0.024	0.05	−0.012	0.02

H THE FIXED-AMPLITUDE PILOT

Registered before any arm existed; the rule returned a fail, recorded as KILL in the registration.

Table 15: Amplitude fixed across the two conditions, a fresh draw of 10 random-split and 10 isotropic directions, scored on the other twenty-five questions. Survives if the fraction exceeds 0.65 and its lower bound exceeds 0.5; killed if the fraction is at most 0.55 or the lower bound at most 0.5.

Item set	Same-sign	Median $\|C\|/\|A\|$	Fraction	95% interval	Rule
Refusal, Llama-2-7b-chat	19 of 20	3.30	0.95	[0.350, 1.000]	KILL

I EQUAL-AMPLITUDE SWEEP

Registered before any arm existed. On refusal and myopic-reward at Llama-2-7b-chat, each of the thirty-nine random-split reference directions is injected at both signs with the same magnitude, at 0.5, 1.0, 1.5 and 2.0 times its stored positive-injection amplitude, and the readout is taken on all fifty questions. A level is valid if the median over directions of the median KL, at both signs, is at most 2.5448 nats, the largest per-condition cost target read elsewhere in this paper (Appendix D; fixed before the run); it is an operational cutoff, not a certificate that the answers remain well formed. Q is the share of the paired response carried by the common component, $\sum_j C_j^2 / \sum_j (C_j^2 + A_j^2)$. The primary endpoint is Q at the highest valid level minus Q at 0.5.

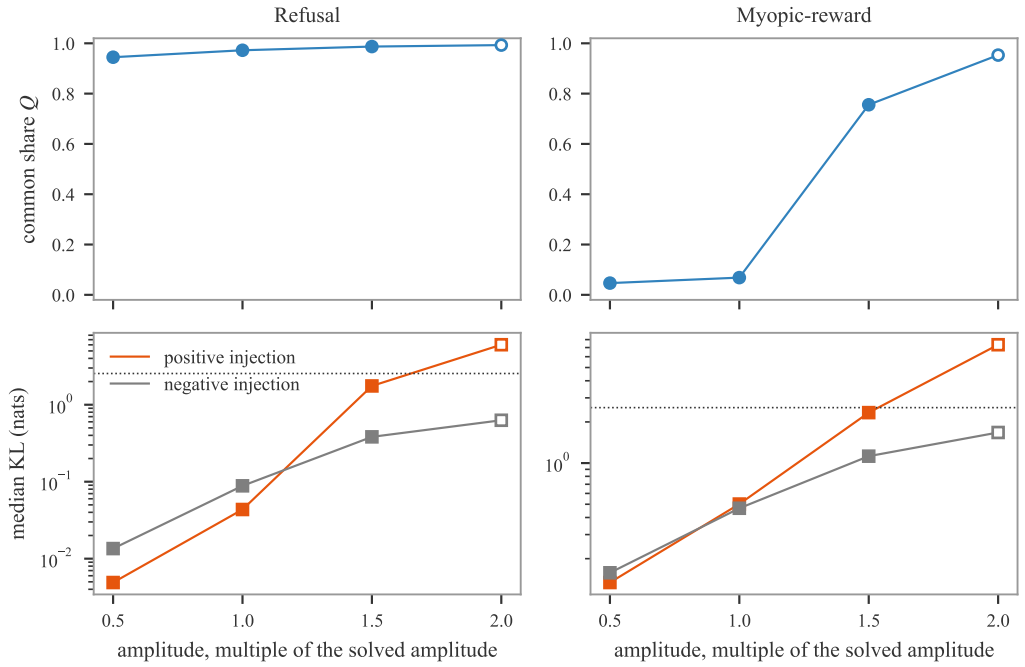

Figure 6: The common component's share Q (top), and the median cost at each sign (bottom, dotted line the cost cap), against amplitude in multiples of each direction's solved amplitude. Open markers exceed the cap and enter no contrast.

Refusal. Directions: the paper's own directions, regenerated and verified against a stored arm. Valid levels: 0.5, 1, 1.5.

Myopic-reward. Directions: the paper's own directions, regenerated and verified against a stored arm. Valid levels: 0.5, 1, 1.5.

Growth exponents. For a response smooth in the signed amplitude, the sign-changing component is first order and the common one second order as amplitude goes to zero, so Q is expected to rise there; by point estimate the share is below one half at the solved amplitude and above it at one and

Table 16: Equal-amplitude sweep. KL is the median over the thirty-nine directions of each direction's median KL at that level, at positive and at negative injection. Levels marked not valid enter no contrast.

Item set	Level	KL^+	KL^-	Valid	Q (fifty items)	Q (scored 25)	Same-sign (fifty)
Refusal	0.5	0.005	0.014	yes	0.94	0.95	0.97
	1.0	0.044	0.088	yes	0.97	0.98	1.00
	1.5	1.754	0.382	yes	0.99	0.99	1.00
	2.0	6.020	0.628	no	0.99	0.99	1.00
Myopic-reward	0.5	0.134	0.158	yes	0.05	0.07	0.10
	1.0	0.502	0.467	yes	0.07	0.05	0.15
	1.5	2.339	1.123	yes	0.76	0.76	0.67
	2.0	7.340	1.674	no	0.95	0.92	0.97

Table 17: Registered secondary endpoints (b) and (d): the median ratio of the common to the sign-changing component, and the ratio of the cost at negative injection to the cost at positive injection, per direction, at each level.

Item set	Level	Median $\lvert C\rvert/\lvert A\rvert$ (fifty)	KL^-/KL^+ median [min, max]
Refusal	0.5	4.71	2.53 [0.24, 13.0]
	1.0	12.36	2.04 [0.08, 15.6]
	1.5	15.26	0.12 [0.02, 0.6]
	2.0	21.48	0.14 [0.03, 0.8]
Myopic-reward	0.5	0.15	1.12 [0.55, 3.6]
	1.0	0.18	0.93 [0.44, 4.1]
	1.5	3.17	0.41 [0.04, 11.6]
	2.0	10.03	0.25 [0.06, 17.6]

Table 18: Registered secondary endpoints (e) and (f). a_j and b_j are each direction's solved amplitudes at positive and at negative injection in the protocol. The first row pairs the protocol's own two arms; the second replaces the positive arm by the negative arm's amplitude, so the two conditions have equal amplitude and unequal cost (the two costs are listed in that order).

Item set	Pair	Median KL (nats)	Q (fifty)	Q (scored 25)	Same-sign (fifty)
Refusal	$+a_j$ with $-b_j$ (the protocol)	0.044, 0.336	0.91	0.92	1.00
	$+b_j$ with $-b_j$ (matched amplitude)	0.344, 0.336	0.99	0.99	1.00
Myopic-reward	$+a_j$ with $-b_j$ (the protocol)	0.502, 1.283	0.14	0.11	0.15
	$+b_j$ with $-b_j$ (matched amplitude)	5.903, 1.283	0.70	0.66	0.54

Table 19: Primary endpoint, on the fifty questions and, beside it, on the twenty-five scored ones. The small-perturbation expansion projects growth exponents of 2 for the common and 1 for the sign-changing component; the measured exponents are printed beside it and are not a fit. The registered rise rule is not evaluable: the coverage of the resampling band it rested on failed the registered rule on synthetic sets (Section 7), so the change is a point estimate.

Item set	Top level	Q at 0.5	Q at top	Change	Exponents $\lvert C\rvert$, $\lvert A\rvert$
Refusal, fifty questions	1.5	0.94	0.99	0.04	1.14, 0.43
scored 25	1.5	0.95	0.99	0.04	1.13, 0.28
Myopic-reward, fifty questions	1.5	0.05	0.76	0.71	2.09, 0.07
scored 25	1.5	0.07	0.76	0.68	2.01, 0.13

a half times on myopic-reward with both families, and is above one half at every level with random-split directions on refusal. From half the solved amplitude to the top valid level, the per-direction growth exponent of the common component has median 2.09 and 1.14 on myopic-reward and refusal with random-split directions, and 1.18 and 1.39 with isotropic ones, against 2 in the small-amplitude limit. For the sign-changing component the medians are 0.07, 0.43, 0.25 and -0.04, against 1. The exponents are measured, not fitted (Table 19).

Concentration. Exploratory, specified after the sweep artifact and three reviews had been read, with thresholds fixed before the numbers were computed and not registered. Q weights directions by the size of their response, so it can be carried by a few. Per direction, $s_j = C_j^2/(C_j^2 + A_j^2)$; the table gives the number of directions whose share is higher at the top level than at half, the range of Q at the top level when any one direction is left out, and the share of the common sum of squares carried by the three largest directions. The rule read the change as widespread if the leave-one-out range was narrower than 0.10, the three largest carried less than half, and at least thirty of thirty-nine shares were higher at the top level; both configurations meet it.

Table 20: Concentration of the change in Q. The two right-hand columns are the share of $\sum_j C_j^2$ carried by the three largest directions at half the solved amplitude and at the top valid level.

Item set	Questions	Shares higher at top	Leave-one-out range of Q	Top three, half	Top three, top level
Refusal	fifty	33 of 39	[0.987, 0.990]	0.17	0.11
	scored 25	35 of 39	[0.990, 0.992]	0.17	0.11
Myopic-reward	fifty	37 of 39	[0.745, 0.776]	0.39	0.16
	scored 25	33 of 39	[0.747, 0.786]	0.42	0.15

J THE AMPLITUDE SWEEP ON THE ISOTROPIC FAMILY

Registered before any arm existed, after the geometry of the random-split family of Appendix L was read. The design, the questions, the cost cap and the registered rule (not evaluable, Section 7) are those of Appendix I and are computed by the same function; only the thirty-nine reference directions differ: the isotropic random directions of the original battery, regenerated from the same random stream and verified against a stored arm before any sweep arm was measured, at 0.5, 1.0, 1.5 and 2.0 times each direction's stored positive-injection amplitude. The questions are the same, so this is a family-sensitivity run and not an independent replication.

Table 21: The primary endpoint on the fifty questions for the two reference families. The random-split rows repeat Table 19. The isotropic family is not a replication on new questions. The registered rise rule is not evaluable (Section 7); the change is a point estimate.

Item set	Family	Top level	Q at 0.5	Q at top	Change
Myopic-reward	random-split	1.5	0.05	0.76	0.71
	isotropic	1.5	0.14	0.54	0.40
Refusal	random-split	1.5	0.94	0.99	0.04
	isotropic	2	0.48	0.94	0.47

The sweeps at the item level. The equal-amplitude sweeps of this appendix and Appendix I on the 50 questions, read with the item-level share Q_{item} and the answer-letter split of Appendix G. The change in Q_{item} is from half the solved amplitude to the top valid level of each configuration, as a point estimate; the registered rise rule is not evaluable, because the false-rise rate of its interval exceeded the registered bound in one null cell (Section 7).

Table 22: Every level of the isotropic-family sweep. KL is the median over the thirty-nine directions of each direction's median KL at that level; levels marked not valid exceed the cap of 2.5448 nats and enter no contrast.

Item set	Level	KL^{+}	KL^{-}	Valid	Q (fifty items)	Q (scored 25)
Myopic-reward	0.5	0.138	0.161	yes	0.14	0.15
	1.0	0.506	0.476	yes	0.26	0.23
	1.5	1.354	1.028	yes	0.54	0.52
	2.0	2.846	2.619	no	0.86	0.85
Refusal	0.5	0.005	0.005	yes	0.48	0.57
	1.0	0.044	0.038	yes	0.79	0.84
	1.5	0.184	0.159	yes	0.91	0.93
	2.0	0.933	0.607	yes	0.94	0.96

Table 23: Every level of both sweeps at the item level, letter readout, 50 questions. The change in Q_{item} from level 0.5 to the top valid level, as a point estimate: myopic-reward, random-split: +0.34 at 1.5; myopic-reward, isotropic: +0.21 at 1.5; refusal, random-split: +0.25 at 1.5; refusal, isotropic: +0.42 at 2.0.

Item set	Family	Level	Q	Q_{item}	$\bar{C}_A$	$\bar{C}_B$	δ	c
myopic-reward	random-split	0.5	0.05	0.37	+0.14	−0.28	+0.21	−0.07
myopic-reward	random-split	1.0	0.07	0.49	+0.31	−0.63	+0.47	−0.16
myopic-reward	random-split	1.5	0.76	0.71	+0.09	−1.52	+0.80	−0.71
myopic-reward	random-split	2.0	0.95	0.69	−0.06	−2.03	+0.98	−1.05
myopic-reward	isotropic	0.5	0.14	0.40	+0.35	−0.32	+0.34	+0.02
myopic-reward	isotropic	1.0	0.26	0.59	+0.55	−0.77	+0.66	−0.11
myopic-reward	isotropic	1.5	0.54	0.61	+0.32	−0.99	+0.66	−0.33
myopic-reward	isotropic	2.0	0.86	0.62	−0.16	−1.30	+0.57	−0.73
refusal	random-split	0.5	0.94	0.67	−0.61	−0.61	+0.00	−0.61
refusal	random-split	1.0	0.97	0.81	−1.64	−1.39	−0.13	−1.52
refusal	random-split	1.5	0.99	0.92	−2.12	−2.10	−0.01	−2.11
refusal	random-split	2.0	0.99	0.93	−2.54	−2.40	−0.07	−2.47
refusal	isotropic	0.5	0.48	0.43	−0.44	−0.35	−0.04	−0.39
refusal	isotropic	1.0	0.79	0.71	−1.39	−0.89	−0.25	−1.14
refusal	isotropic	1.5	0.91	0.81	−2.56	−1.25	−0.66	−1.90
refusal	isotropic	2.0	0.94	0.85	−3.02	−1.74	−0.64	−2.38

K THE CONTROLS, ITEM SET BY ITEM SET

Table 24: The registered both-condition control; the plant's outcome under each condition (passed, estimate not resolved, or resolved and outranked); the sign-aware contrast on the first run and on the second run; and, in the last two columns, the same contrast applied to the released vector in the plant's place against the random-split and the isotropic family (registered before it was run, scored 25, first run's arrays).

Item set	Both-cond. control	Plant at $+1$	Plant at -1	Sign-aware, scored run	Sign-aware, second run	Vector, random-split	Vector, isotropic
7B sycophancy	no	not resolved	outranked	no	no	no	no
7B refusal	fired	passed	passed	fired	fired	fires	fires
7B hallucination	fired	passed	passed	fired	fired	fires	fires
7B survival-instinct	no	not resolved	not resolved	no	fired	no	no
7B myopic-reward	no	not resolved	passed	fired	fired	fires	fires
7B corrigible-neutral-HHH	no	passed	not resolved	fired	fired	fires	fires
7B coordinate-other-ais	no	outranked	passed	fired	fired	fires	fires
13B sycophancy	no	passed	outranked	fired	fired	no	no
13B refusal	no	passed	outranked	fired	fired	fires	fires
13B hallucination	fired	passed	passed	fired	fired	fires	fires
13B survival-instinct	no	not resolved	passed	fired	no	fires	fires
13B myopic-reward	fired	passed	passed	fired	fired	fires	fires
13B corrigible-neutral-HHH	fired	passed	passed	fired	fired	fires	fires
13B coordinate-other-ais	no	outranked	passed	fired	fired	fires	fires

The second run re-drew everything that carries a seed: the stratified calibration and scored split, and with it the planted direction and the 25 scored items, the 39 random-split members and the perpendicular component. It reused the 50 questions, so its scored 25 overlap those of the first run. Its statistic and its floor of six were fixed before it was launched, and its 13B and 7B counts are never pooled.

L DESCRIPTIVE CHECKS ON THE REFERENCE FAMILIES AND THE READOUT

Exploratory: specified after the results were read, on stored artifacts, with no threshold and no new draw.

Table 25: Effective number of directions of a family: $(\sum_i \sigma_i)^2 / \sum_i \sigma_i^2$ over the singular values of the matrix of its thirty-nine unit directions. The last three columns are for the directions of Llama-2-7b-chat used in the sweeps, the same for both item sets, and for the Llama-2-13b-chat families regenerated from their seed path for six of the seven item sets at that size (below; sycophancy at 13B is its own draw); the participation ratio is $(\sum_i \sigma_i^2)^2 / \sum_i \sigma_i^4$. Energy in two coordinates is the share of the family's total squared entries carried by its two largest activation coordinates of the hidden width, not by two singular components.

Model	Family	Effective directions	Median $\lvert\cos\rvert$	Participation ratio	Energy in two coordinates
Llama-2-7b-chat	random-split	8.5	0.88	1.47	0.82
	isotropic	38.9	0.01	38.61	0.0011
Llama-2-13b-chat	random-split	38.3	0.03	36.08	0.024
	isotropic	38.9	0.01	38.72	0.00086

Regeneration of the 13B families (registered). The 13B reference directions were not stored with the cells. They were regenerated from the documented seed path (seed 20260919 on the cached covariance harvest, whose sha256 the artifact records) and verified twice, under a rule fixed before the run (prereg_e2_13b_family and its amendment 1, in the supplement): (a) the effective number of directions of the regenerated random-split and isotropic families equals the value stored in the artifacts of six of the seven 13B item sets to 10^{-9}; the sycophancy artifact at that size stores a different value, so that item set was run on its own draw of the family, which is not regenerated and which

no number here covers; (b) each of the five 13B arms of Table 27, re-solved with the regenerated direction at its stored magnitude and sign on the stored prompts, reproduces its stored median cost exactly (relative error zero, tolerance 2 percent). A negative control, whose rule was fixed after that result was known, re-solves the same arms with the direction of the next family member: every one misses the stored cost, by 25 to 970 percent, so the check can fail. A first run, registered as requiring equality with all seven artifacts, was void under its own rule (the premise was wrong for sycophancy); its log is in the supplement, and the amendment that scoped the measurement to six item sets was written after that stage had been read and before the cost check ran. The registered prediction, a participation ratio below 3 for the 13B random-split family, failed: the measured value is 36.08 (isotropic 38.72), with median $|\cos|$ 0.03 and a share 0.02 of the squared entries in two coordinates. The 13B random-split family is not concentrated; the 7B one is.

Table 26: Per item set, across the scored items. Slope: of the mean steered readout (the unsteered readout plus the common shift, averaged over the thirty-nine directions) on the unsteered readout, where one would mean the steered readout keeps the between-item spread of the unsteered one. Opposite sign: the share of items where the common shift has the sign opposite to the unsteered readout, which includes shifts that overshoot zero. Closer to zero: the share where the steered readout is nearer to zero than the unsteered one. Same-sign: the count of thirty-nine on the scored twenty-five and on the other twenty-five questions.

Item set	Slope, set	Slope, vector	Opposite sign	Closer to zero	Same-sign, scored 25	Same-sign, other 25
7B sycophancy	0.33	0.66	0.96	0.92	39	38
7B refusal	0.56	0.80	0.96	0.96	39	39
7B hallucination	0.68	0.81	0.92	0.84	30	22
7B survival-instinct	0.28	0.41	0.80	0.64	33	26
7B myopic-reward	0.44	0.67	0.68	0.68	3	6
7B corrigible-neutral-HHH	0.39	0.54	0.76	0.64	9	12
7B coordinate-other-ais	0.70	0.73	0.80	0.72	18	33
13B sycophancy	0.70	0.87	0.80	0.76	35	30
13B refusal	0.61	0.72	0.96	0.96	38	34
13B hallucination	0.64	0.77	0.76	0.72	20	38
13B survival-instinct	0.15	0.84	0.84	0.72	30	28
13B myopic-reward	0.54	0.88	0.88	0.84	29	29
13B corrigible-neutral-HHH	0.31	0.61	0.92	0.80	13	17
13B coordinate-other-ais	0.44	0.78	1.00	1.00	39	39

Convergence: 7 of 2,324 arms stored for the fourteen item sets are recorded as not converged within the two percent tolerance, listed in Table 27. None is a candidate vector: each is a reference direction. Removing an arm from its family, or rescaling its effect to the cost it was meant to have (effect taken to scale with the square root of the cost, a first-order approximation), leaves the candidate's rank where it was in every case.

Contraction reading of the common shift (exploratory). If each reference direction rescales the baseline preference and adds a term of its own, $r_j(+) = \lambda_j r_0 + b_j$ and $r_j(-) = \lambda_j r_0 - b_j$, then $C_j = (\lambda_j - 1)\, r_0$ and $A_j = b_j$. The mean steered readout of the reference set has slope 0.15 to 0.70 on the unsteered one in every item set, and the released vectors' own steered readouts have slopes of 0.41 to 0.88. A line fitted on half of the scored items predicts the direction-averaged common shift on the other half with a median held-out E of 0.71 (energy about the fitting half's mean, Table 28) and at least 0.29 in every item set. The target C_i contains $-r_{0,i}$, so E is inflated by construction and the slope is the informative statistic; permuting the baseline across items (median -0.05) breaks that shared term and so does not test contraction against the absence of a relation. The fit is on the direction-averaged shift, not per direction, and identifies the shape of the common component, not its cause. The split was by item index and the decision rule was fixed before the fit.

Table 27: The arms that missed the two percent cost tolerance. In every row the released vector ranks first among itself and the family with the arm, without it and with its effect rescaled to its target cost (checked when this table is built). Same-sign is the count of members whose two conditions share a sign, with the arm and without it. The relative error is signed: positive means the arm cost more than its target.

Cell	Family, member, condition	Target KL	Achieved KL	Error	Same-sign, with and without
7B corrigible-neutral-HHH	random-split 35, -1	1.078	1.132	+5.0%	9 of 39, 9 of 38
7B coordinate-other-ais	random-split 9, $+1$	1.281	1.241	-3.2%	18 of 39, 18 of 38
13B refusal	random-split 3, $+1$	0.002334	0.002236	-4.2%	38 of 39, 37 of 38
13B myopic-reward	random-split 3, -1	0.0007983	0.000817	+2.3%	29 of 39, 28 of 38
13B myopic-reward	random-split 14, -1	0.0007983	0.0008225	+3.0%	29 of 39, 28 of 38
13B myopic-reward	isotropic 5, -1	0.0007983	0.0008235	+3.2%	30 of 39, 29 of 38
13B myopic-reward	isotropic 34, -1	0.0007983	0.0007788	-2.4%	30 of 39, 30 of 38

Table 28: Held-out fit of the contraction reading. Per item set, the mean common shift over the thirty-nine directions is regressed on the unsteered readout over one half of the scored items (odd or even positions of the sorted item list) and predicted on the other; the entry is the share of the held-out common-shift energy, about the fitting half's mean, that the line explains. Exploratory on the reused pool; the split and the decision rule (median of the mean at least 0.5) were fixed before the fit was run. The target contains $-r_{0,i}$, so the entry is inflated by construction; read the slope column of Table 26 first.

Model	Behaviour	Fit odd, test even	Fit even, test odd	Mean
Llama-2-7b-chat	sycophancy	0.98	0.86	0.92
	refusal	0.83	0.77	0.80
	hallucination	0.77	0.64	0.70
	survival-instinct	0.47	0.77	0.62
	myopic-reward	0.53	0.45	0.49
	corrigible-neutral-HHH	0.80	0.65	0.72
	coordinate-other-ais	0.37	0.22	0.29
Llama-2-13b-chat	sycophancy	0.63	0.66	0.64
	refusal	0.50	0.74	0.62
	hallucination	0.57	0.81	0.69
	survival-instinct	0.81	0.67	0.74
	myopic-reward	0.91	0.95	0.93
	corrigible-neutral-HHH	0.95	0.94	0.94
	coordinate-other-ais	0.72	0.85	0.78

Table 29: The registered sweep re-read with each level as the top level; the registered top level is the highest valid level, 1.5. The rule that read each re-read is not evaluable (Section 7); the change is a point estimate.

Item set	Top level	Change
Refusal	1.0	0.03
	1.5	0.04
	2.0	0.05
Myopic-reward	1.0	0.02
	1.5	0.71
	2.0	0.91

M EVERY ANALYSIS AND ITS REGISTRATION STATUS

Table 30: Every analysis the paper reports. A rule is fixed before data when a written hypothesis or threshold was dated before its data existed; the registrations are in the supplement. The sweeps and the re-pairing reference were written after earlier results were read, on configurations chosen in view of them. The question pool is the same 50 per behaviour throughout, so no row is a held-out replication.

Analysis	Rule fixed before data	Questions	Family	Reported
Effect, interval and rank	yes	scored 25	random-split; rank also isotropic	§4, App. A
Paired same-sign, components	no	scored 25	random-split	§5, App. F
Re-pairing reference	logged, no threshold	scored 25	random-split	§5, App. G
Readout and family grid	arms yes, counts no	scored 25	both	§6, App. B
Fixed-amplitude pilot	yes, failed	other 25	10 + 10 fresh	§6, App. H
Sweep, random-split	yes, after results seen; rise rule not evaluable, its interval's coverage failed	all 50	random-split	§6, App. I
Sweep, isotropic	yes, after sweep seen; rise rule not evaluable, its interval's coverage failed	all 50	isotropic	§6, App. J
Common share Q, linear-response reference	no	scored 25	both	§5, App. G
Item-level share, letter split	no; after a look	scored 25, all 50	both	§5, App. G
Contraction slopes, half-sample fit	no; split and rule yes	scored 25	random-split	App. L
Letter-conditional divergence of the readout	no	scored 25	none	App. D
Planted direction, both conditions	yes, floor of six	scored 25	random-split	§7, App. K
Sign-aware contrast, scored run	no	scored 25	random-split	§7, App. K
Sign-aware contrast, second run	yes	50, overlaps first	fresh draw	§7, App. K
Sign-aware contrast, released vectors	no; rule fixed before analysis of inspected data	scored 25	both	§7, App. K
Recovery of a known Q, synthetic sets	yes, before run; second registration after the first was read, no interval passed	50, synthetic	random-split	§7, App. G
Coverage of the remaining joint-item intervals, synthetic sets	yes, before run; failed	50 and 25, synthetic	both	§7, App. G
Regeneration of the 13B families	yes, before run; first run void by its rule, amendment after stage A, prediction failed	all 50, cost only	both, 13B	§3, App. L
Survey of practice	convenience sample	28 papers	none	§1, App. C