HALF-DEBIASED: TWO TEMPLATE AXES IN STEERING-VECTOR GEOMETRY, AND A METHOD THEY LICENSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Steering vectors push a large fraction of inputs in the unintended direction. A growing literature explains this using the geometry of per-sample activation differences, employing clustering, mixture-fitting, and routing techniques. Recent works identified answer-order and affirmative/negative biases in this benchmark as *behavioral* confounds. By measuring them instead as *directions* on 151 datasets across ten models, we establish three structural consequences that purely behavioral metrics cannot capture. Both axes are recoverable from individual differences at AUC 1.000 on all $2{,}114$ measurements against a controlled null. Since they are near-orthogonal, the rank-one correction applied by the field removes one axis but leaves the other intact; removing both leads to an additional 0.037 of anti-steerability on Llama-2-7B. The extent to which these axes contaminate the *deployed* vector is governed by representational resolution, falling from a cosine of 0.887 at 0.5B to 0.118 at 14B. Fed into the cluster-then-route primitive these methods share, this geometry results in letter-pure prototypes on all $48{,}320$ clusterings. While geometry predicts behavioral reliability across models, identical optimizers transfer geometric gains to behavior at different rates (17% versus 58%). Selective steering accounts for these facts: abstaining where the geometry predicts a misaligned push reduces anti-steerability by 0.071, supported by a distribution-free risk certificate and a probe requiring only the base prompt.

1 INTRODUCTION

Steering vectors (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023) modify language-model behavior by adding a fixed direction to intermediate activations. The dominant method, Contrastive Activation Addition (CAA (Rimsky et al., 2024)), averages per-sample activation differences across contrastive prompt pairs. Tan et al. (2024) showed that this control is unreliable: per-sample steerability is frequently bimodal, with up to half of inputs pushed opposite to the intended direction. Subsequent research pursued geometric explanations. Braun et al. (2025) showed that the mean cosine between training differences and the steering vector predicts which datasets steer well. A growing body of methods took the per-sample structure of $\{d_i\}$ as its object of study by clustering it, fitting mixtures to it, or routing on it (Wang et al., 2025a; Kayan & Zhang, 2025; Sharma & Raman, 2025). Two premises are implicit in this direction: that this structure reflects the concept being steered, and that improving a method's standing on it improves behavior. We evaluate both premises. We began by developing a mixture-model theory of steering reliability; however, the controls designed to validate it instead invalidated it, revealing issues that apply to this line of research as a whole. This paper reports these findings and examines what remains usable.

Zheng et al. (2024) establish that language models carry a token bias over option identifiers that dominates multiple-choice selection. Similarly, Pezeshkpour & Hruschka (2024) document order sensitivity, and Sclar et al. (2024) show that surface formatting shifts behavior by margins comparable to capability differences. Addressing this, Zhou et al. (2024) mitigate the bias through internal attention and FFN manipulation, representing the closest prior work to treating the issue representationally. Within the steering literature, Tan et al. (2024) (working on this exact corpus) documented anti-steerability and bimodality, randomized the letter assignment, and reported that both the A/B and the Yes/No splits remain predictive of per-example steerability afterwards. Subsequently, Braun et al. (2025) established directional agreement as a dataset-level predictor. All of these studies measure the effects *behaviorally*, evaluating them solely as properties of model outputs.

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.

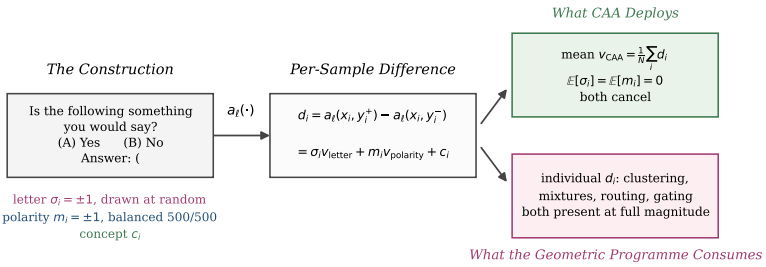


Figure 1: Eq. 2 formalizes how two balanced, zero-mean surface variables enter the per-sample difference. They cancel from the aggregate CAA vector but remain at full magnitude in the differences consumed by clustering, mixture-fitting, and gating methods. Randomization protects the aggregate mean but offers no per-sample guarantees. Above 4B parameters, the axes are near-orthogonal ($|\cos(v_{\text{letter}}, v_{\text{polarity}})| \approx 0.07$), so a rank-one projection removes an axis and leaves the other intact.

Our contribution is to measure these confounds *geometrically* in the per-sample difference space, which is the representation consumed by the methods above. From this perspective, we establish three findings that do not follow from behavioral observations: 1) both axes are recoverable from individual d_i at ceiling with a controlled null, 2) they are near-orthogonal; thus, the rank-one correction applied in the literature removes one axis while leaving the other intact, and 3) the extent to which they contaminate the *deployed* vector is a strong function of the model's representational state. While the randomization control relied on by the field holds in expectation over a dataset, the methods that consume these differences operate sample by sample (Fig. 1).

Three prior methods share the cluster-then-route primitive. ACT (Wang et al., 2025a) partitions per-head activation differences and selects a prototype per input, PDS (Kayan & Zhang, 2025) applies this to chain-of-thought contrasts, and G-ACT (Sharma & Raman, 2025) adds per-layer centroid probes. By contrast, Curveball (Raval et al., 2026) falls outside this family, replacing the linear intervention with a kernel feature space. Because these methods evaluate on structurally different data (free-text on TruthfulQA for ACT, chain-of-thought on GSM8K and AQuA-RAT for PDS, and code prompts for G-ACT), v_{letter} and v_{polarity} do not exist in their corpora. §3.3 details the implications for our scope. A second family gates the intervention rather than re-aiming it. Like our proposed probe, CAST (Lee et al., 2025), in-distribution steering (Vogels et al., 2026), DSAS (Ferrando et al., 2025), SADI (Wang et al., 2025b), and Linear-AcT (Rodríguez et al., 2025) all condition on prompt-computable quantities. Conceptually closest to our approach is MERA (Hedström et al., 2025), which fits a linear error probe and entirely declines to steer a task when no threshold confidently helps (mirroring the refusals in §5). Our method differs from MERA in two key respects. First, MERA's gate predicts task error, whereas ours relies on the specific geometric quantity validated by our audit (underscoring the necessity of §4.2). Second, MERA calibrates a threshold to improve expected performance, whereas we report a risk-coverage curve with a distribution-free bound on selective risk. While conformal risk control is standard, our contribution is not the statistical machinery itself. Rather, it is the identification of the geometric score, the provision of a distribution-free certificate, and the empirical demonstration that this exact signal fails as a reliable optimization target. The main contributions of this work are:

1. **Rank-Two Artifact and Rank-One Remedy** (§3.1, §3.2). We identify two balanced, near-orthogonal surface axes, both decodable at ceiling across all $2{,}114$ measurements against permutation controls. Consequently, the standard projection applied by the field removes one axis while leaving the other intact. Removing both reduces anti-steerability by a further 0.037 on Llama-2-7B (90 of 135 datasets, $p = 2.5 \times 10^{-7}$) and clears the letter from the cluster prototypes that the rank-one leaves behind.
2. **The Vulnerability of Mean Controls** (§3.1). The standard control corrects the mean vector deployed by CAA, but remains silent on the individual differences consumed by every clustering, mixture-fitting, and routing method.
3. **Monotonic Scaling of Surface Dimensions** (§3.2). The trend is strictly monotonic within within a model family, decreasing from 0.887 to 0.118. At 0.5B parameters, the representation is rank-one against a float32 control. We observe this exact signature in a 7B model at one-quarter of its depth; capacity and depth are not disentangled here, so we report both.

4. **Template Contamination in Clustering** (§3.3). Using the ACT estimator, the cluster-then-route primitive returns letter-pure prototypes on all $48{,}320$ clusterings across ten models.
5. **The Transfer Gap and Selective Steering** (§4.1-§5). While geometry reliably predicts behavior across models, identical optimizers transfer geometric gains at different rates (17% versus 58%). Selective steering logically follows, supported by a distribution-free certificate, a deployable probe, and explicit control that distinguishes this probe from a template reader.

2 Background and Protocol

This section establishes the method being audited, the two measured quantities, and the evaluation corpus and splits used for every result. Appendix A details the models, datasets, and hyperparameters.

2.1 Contrastive activation addition and two measures of anti-steerability

Given contrastive items $\{(x_i, y_i^+, y_i^-)\}$, CAA computes $d_i = a_\ell(x_i, y_i^+) - a_\ell(x_i, y_i^-)$ at layer ℓ and $v_{\mathrm{CAA}} = \frac{1}{N}\sum_i d_i$. At inference, the intermediate activations are updated as $h \leftarrow h + \lambda v$. Each item is a two-option multiple-choice prompt ending in `"Answer: ("`, where the behavior-matching option is assigned to letter A or B at random; we record this assignment as b_i.

The Scoring Token Must Be Resolved in Context. On SentencePiece tokenizers, standalone `"A"` encodes with a dummy prefix (id 319 for Llama-2), whereas the contextual `"Answer: ("` yields the bare letter (id 29909). Scoring the standalone string thus evaluates a token the model would never actually emit. Because the multiple-choice slopes still respond to the intervention, this discrepancy easily goes unnoticed. We therefore resolve the target token in context, verifying it as a clean single-token suffix. This tokenization split affects three of the ten models (both Llama-2 variants and Mistral, Appendix B.1; Table 6) and survives major tokenizer-library updates, confirming it as a durable property rather than a temporary artifact (Appendix A.1).

Definition 1 (Behavioral anti-steerable fraction)**.** *Let s_i be the ordinary least squares (OLS) slope of the target logit difference on the multiplier $\lambda \in \Lambda$ (Tan et al., 2024). Sample i is anti-steered if $s_i < 0$.*

Definition 2 (Directional anti-agreement fraction)**.** *Geometric proxy, defined as $|\{i : d_i^\top v < 0\}|/N$.*

Computing Definition 1 requires $|\Lambda|$ forward passes per sample per policy, whereas Definition 2 is computationally free once the differences $\{d_i\}$ are cached. Both metrics are computed on the same held-out split, ensuring that no train/test asymmetry affects their ratio. Because these are fundamentally different quantities, we label them separately throughout this work; conflating them would inaccurately present a geometric outcome as a behavioral one.

2.2 Models, corpus, and splits

Protocol. We evaluate ten instruction-tuned models (0.5B–14B parameters) across four architectural families (Appendix A.1). The corpus comprises 151 datasets from (Perez et al., 2023) (135 `persona` and 16 `advanced-ai-risk` human-generated sets), totalling 141,320 items. We partition the data into 40/10/50 train/validation/test splits. To ensure independent replicates, we derive per-dataset random seeds via CRC32 from the dataset name, preventing the identical permutations of a global seed. We evaluate steering multipliers $\Lambda = \{-1.5, \ldots, 1.5\}$ in steps of 0.5, rescaling every single-vector policy to $\|v_{\mathrm{CAA}}\|$. Extraction layers follow published indices where available (layer 13 for Llama-2-7B, layer 21 for Qwen-1.5-14B) or default to 0.45 network depth. Estimation relies strictly on training data; the validation split serves the empirical constraint and probe thresholding, while all reported metrics reflect the test split.

Corpus Composition and Preprocessing. A small subset of 643 items across two datasets contains three to eight options. We reduce these to two options by retaining only the behavior-matching choice and the labeled non-matching choice, ensuring no items are dropped. All items are utilized regardless of their `label_confidence` rating. Because the 151 datasets do not represent 151 strictly independent concepts (twelve near-duplicate name families cover 120 of the datasets), we report every statistical interval under two distinct clustering units: the individual dataset and the broader name stem (Appendix A.2).

Table 1: Surface axes across ten models (151 datasets each). Letter decodability is 1.000 and the permutation mean is 0.522 uniformly; both columns are thus omitted. "null max" denotes the highest per-dataset permutation AUC per model. Polarity applies only to the 135 `persona` sets. Near-zero $|\cos(v_{\text{letter}}, v_{\text{polarity}})|$ ensures a rank-one projection along v_{letter} leaves v_{polarity} intact.

Condition	null max	POL AUC	POL at 1.000	$\lvert\cos(v_{\text{letter}}, v_{\text{polarity}})\rvert$
Qwen-1.5-0.5B	0.631	0.789	19/135	0.360
Qwen-1.5-1.8B	0.611	0.998	132/135	0.114
Gemma-2-2B	0.616	1.000	135/135	0.074
Qwen-1.5-4B	0.626	1.000	135/135	0.075
Mistral-7B	0.627	1.000	135/135	0.078
Llama-2-7B	0.612	1.000	135/135	0.074
Qwen-1.5-7B	0.624	1.000	135/135	0.074
Gemma-2-9B	0.634	1.000	135/135	0.071
Llama-2-13B	0.624	1.000	135/135	0.073
Qwen-1.5-14B	0.626	1.000	135/135	0.072

3 The Audit: What the Per-Sample Geometry Contains

The evaluation construction used in the steering reliability literature imposes inherent structure on every contrastive pair before any conceptual information is even introduced. This section establishes the precise nature of that structure, the extent to which it contaminates the vector a policy deploys, and what a standard method returns when consuming it.

3.1 Two near-orthogonal template axes

The behavior-matching answer is randomly assigned to a letter. Per-sample difference takes the form

$$d_i = \sigma_i v_{\text{letter}} + c_i, \qquad \sigma_i = +1 \text{ if } b_i = A,\ -1 \text{ otherwise}, \tag{1}$$

where v_{letter} is item-independent and c_i is concept-dependent. If $\|v_{\text{letter}}\| \gg \|c_i\|$, the normalized population is bimodal and antipodal for reasons unrelated to the concept, while the mean remains unaffected since $\mathbb{E}[\sigma_i] = 0$.

A Second Template Axis. The answer letter is not the only balanced surface variable imposed by the benchmark construction. Across the 135 `persona` datasets, the behavior-matching answer is balanced between "Yes" and "No" by design, achieving a proportion of exactly 0.500 in every dataset. Each item therefore carries a second ± 1 label, $m_i = \mathbb{1}$[the matching answer is the affirmative], which is independent of σ_i (mutual information 0.0008 bits). Eq. 1 is therefore incomplete:

$$d_i = \sigma_i v_{\text{letter}} + m_i v_{\text{polarity}} + c_i, \tag{2}$$

where v_{polarity} represents a response-polarity direction, not the underlying concept (Fig. 1). Like the letter direction, this polarity axis has zero mean and cancels from v_{CAA} in expectation. However, it remains present at full magnitude in every individual d_i, which is precisely the object used by the geometric analysis. The remaining 16 datasets pair free-text options, meaning m_i is undefined.

Audit Design. For each surface axis, we evaluate the normalized directions $\hat{d}_i$ as follows: **(A1)** Splitting the data 50/50, we estimate v_{axis} via the difference of group means on one half and report the AUC of $\hat{d}_i^\top \hat{v}_{\text{axis}}$ for label separation on the held-out half. Cross-fitting is essential: at $d = 4096$ dimensions with $N \leq 1000$, in-sample separation is uninformative .**(A2)** We repeat this across 20 shuffled-label replicates using the identical split. A perfect AUC is meaningful only if this null estimator remains at chance. **(A3)** We compute the participation ratio $\text{PR} = (\sum_j \lambda_j)^2 / \sum_j \lambda_j^2$ of the uncentered second-moment spectrum of $\{\hat{d}_i\}$. This is not a PCA spectrum: centering would remove the CAA mean vector and modify the sign semantics governing downstream quantities.

Table 1 reports the complete audit. The answer letter achieves a cross-fit AUC of exactly 1.000 across all 1,510 measurements. The permutation control (the identical estimator on shuffled labels) yields per-dataset means of 0.512–0.546 (global mean 0.522) and maxima of 0.529–0.634, ensuring signal and null distributions never overlap. Polarity is similarly recovered at 1.000 AUC on nine

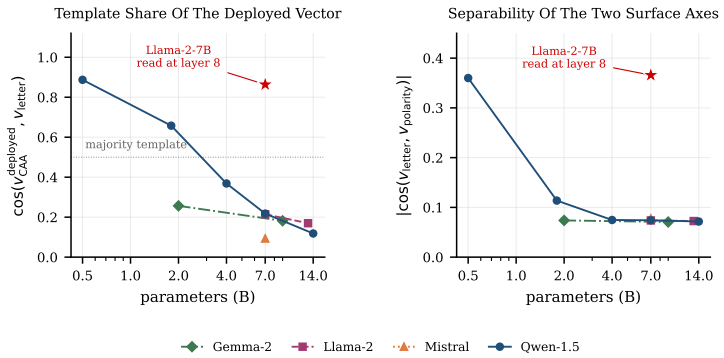


Figure 2: Two scaling laws ($n = 151$ datasets per point). **Left**: Cosine between the *deployed* (train-split) CAA vector and pure answer-letter direction. **Right**: Alignment of the two surface axes, which resolve from a single rank-one direction into two, mirroring the representation. Within the Qwen-1.5 ladder (fixed tokenizer and training pipeline), both decrease monotonically ($\rho = -1.000$). The starred point (Llama-2-7B, layer 8 of 32) matches the 0.5B model across both metrics.

Table 2: The scaling law. cos uses the all-items mean; $\cos_{\text{dep}}$ uses the train-split mean to reflect the exact deployed vector. Column "> 0.5" denotes the dataset fraction ($n = 151$) where this vector forms the majority template. For anti-agreement, *dir-anti* measures the geometric fraction on raw differences, while *anti-steer* measures the behavioral fraction under debiased CAA

Model	params	PR	cos	$\cos_{\text{dep}}$	> 0.5	dir-anti	anti-steer
Qwen-1.5-0.5B	0.5B	1.02	0.854	0.887	92.7%	48.4%	0.476
Qwen-1.5-1.8B	1.8B	1.27	0.596	0.658	68.9%	44.6%	0.406
Gemma-2-2B	2B	2.37	0.193	0.257	14.6%	27.7%	0.270
Qwen-1.5-4B	4B	1.94	0.280	0.368	29.8%	40.1%	0.415
Mistral-7B	7B	3.63	0.064	0.093	0.7%	8.3%	0.227
Llama-2-7B	7B	2.30	0.144	0.215	8.6%	24.6%	0.260
Qwen-1.5-7B	7B	2.42	0.182	0.218	9.9%	31.1%	0.405
Gemma-2-9B	9B	2.54	0.130	0.182	6.0%	16.0%	0.267
Llama-2-13B	13B	2.77	0.134	0.170	4.6%	13.5%	0.309
Qwen-1.5-14B	14B	3.43	0.100	0.118	0.0%	11.8%	0.200

of ten models. Because axis alignment drops from 0.360 at 0.5B to 0.071–0.078 at 4B and above, a rank-one projection along v_{letter} leaves v_{polarity} intact. Finally, the permutation maximum is an extreme-value statistic over folded null draws, not a per-dataset significance level; the per-dataset mean (0.522) remains the correct comparative baseline.

What the Randomization Control Protects. Tan et al. (2024) randomize letter assignment to prevent the extraction of letter direction. Eq. 1 shows that this control is mathematically correct and sufficient for its intended purpose: randomization cancels v_{letter} from the mean, which is precisely the object that CAA deploys. Also, they observe that these biases remain predictive of per-example steerability even after randomization. Our claim, however, concerns the underlying geometry, not behavior. Randomization does not affect any individual d_i, whereas geometric methods (clustering, mixture fitting, routing, and per-sample gating) operate directly on these individual differences. Consequently, a control that holds only in expectation over a dataset cannot support per-sample geometric claims. Finally, §3.2 shows that at low model capacities, this control fails to protect the mean.

3.2 Contamination of the Deployed Vector

At 0.5B, the participation ratio is 1.02 out of 1024 hidden dimensions. To numerical precision, the per-sample differences are rank-one, and the PR is below 1.20 on all 151 datasets. The deployed vector exceeds $\cos = 0.5$ on 92.7% of these datasets. At this scale, CAA does not extract a contaminated concept direction; instead, it extracts the answer-letter direction. At 14B, this same quantity drops to

0.0%. Along the Qwen-1.5 ladder (Fig. 2), where the tokenizer and training pipeline are held fixed, the deployed cosine falls monotonically from 0.887 to 0.118 across five models spanning a $28\times$ range (Spearman $\rho = -1.000$; on five points, the smallest attainable two-sided exact p is 0.0167). The separability of the two axes follows the same ordering: they are 36% collinear at 0.5B and $\approx 7\%$ from 4B upward (Fig. 2). Complete per-model values, alongside the participation ratio and both anti-steerability measures, are provided in Table 2 and Appendix B.2.

The Rank-One Geometry Is Not a Numerical Artifact. The 0.5B result relies on evaluating small residual structure at roughly three significant digits in bfloat16 precision. To confirm this is not a numerical artifact, we re-extracted the same model across all 151 datasets using float32. The PR is 1.024 in float32 versus 1.025 in bfloat16, and the deployed cosine is 0.889 versus 0.887. Every conclusion remains unchanged.

Capacity or Depth? While we can establish what the operative variable is *not*, its exact identity remains elusive. Evaluated at layer 8 of 32 (one-quarter depth), Llama-2-7B almost exactly reproduces the 0.5B empirical signature (Fig. 2, right): PR 1.041 versus 1.025, deployed cosine 0.864 versus 0.887, axis alignment 0.366 versus 0.360, polarity AUC 0.789 versus 0.789, and directional anti-agreement 0.483 versus 0.484. The within-model trend is a step rather than a gradient: PR shifts abruptly from $1.04 \to 2.30$ between layers 8 and 13, plateauing at 2.3–2.4 through layer 24. Since all ten models are tested at roughly 0.45 relative depth, representational capacity and processing depth remain confounded in the scaling ladder. Although scaling is monotonic at matched relative depths, isolating which variable drives this behavior needs a depth sweep on a second architectural family.

The Remedy Is Rank-One; the Construction Is Rank-Two. Projecting out the train-split estimate of v_{letter} reduces mean letter decodability to near chance across all models. At 0.5B, this decreases the deployed cosine from 0.887 to 0.265 and increases PR from 1.02 to 2.70, demonstrating that a rank-one correction recovers rank-2.7 structure. Nevertheless, the second template axis survives this projection at an AUC of 1.000 on all models. Measured behaviorally on the reference models, removing both axes reduces anti-steerability beyond the rank-one baseline. Across the 135 datasets with defined polarity, the anti-steerable fraction drops from 0.271 to 0.233 on Llama-2-7B (improving 90 datasets, $p = 2.5\times10^{-7}$) and from 0.214 to 0.203 on Qwen-1.5-14B (85 datasets, $p = 1.1\times10^{-6}$). This improvement is three times larger on the smaller-capacity side of the scaling ladder, where the axes are least separable (§3.1). Since the projection degrades to rank-one on the 16 label-free datasets (leading to a rank of 1.79), we exclude them from this comparison to preserve analytical accuracy.

How Much Residual is Real? A finite-data projection removes the axis only up to an error ε of order $1/\sqrt{n_{\text{fit}}}$. The surviving residual, $d_i - (d_i^\top \hat{v})\hat{v} = c_i - \|v_{\text{letter}}\|\sigma_i\varepsilon + \ldots$, remains proportional to σ_i but aligns with the estimation error. Consequently, a decoder extracts a signal even when no true structure remains. The permutation control fails to detect this artifact because shuffling labels destroys the correlation between ε and σ, leaving the true-label statistic artificially inflated.

We separate the two by refitting the projection across eight nested training fractions, scoring fixed held-out items. A residual decaying as $1/\sqrt{n_{\text{fit}}}$ reflects pure estimation error, whereas a flat residual indicates true structure. On Llama-2-7B, the median log-log slope is -0.564, closely matching the -0.500 sampling-error prediction. Dataset-level classification requires confidence intervals rather than point estimates, resulting in a three-way split: 65 intervals exclude -0.30 (Appendix B.4; Fig. 4), 26 exclude -0.50, and 60 span both, remaining unclassified at this resolution (Appendix B.4). We detail this split because a naive comparison against the dataset's permutation maximum conflates these cases and significantly overstates the first.

3.3 A Standard Primitive Recovers the Template

As a direct consequence of §3.1, if the answer letter dominates per-sample geometry, partitioning should recover it. We verify this to ensure these audited artifacts actually affect standard methods, feeding the geometry directly into the literature's current algorithmic primitive. Adopting the released ACT hyperparameters (`KMeans(K, n_init=10, random_state=42)`, $K = 4$), we cluster 32 contiguous residual-stream blocks at the extraction layer. Across all ten models, this leads to $48{,}320$ clusterings, every one of which is letter-pure (Appendix B.3; Table 7). Purity is exactly 1.000 uniformly, and 100% of blocks produce prototypes $\geq 90\%$ aligned with a single letter (Appendix B.3). Rather than uncovering latent conceptual structure, the primitive merely recovers the random letter assignment. Applying a rank-one projection reduces the fraction of letter-pure blocks to 0.0–16.9%.

Table 3: Predicting behavioral anti-steerability across ten models. The geometric predictors are computationally free: they require only cached activations, bypassing forward passes, multiplier sweeps, and behavioral evaluation.

Predictor	r	p	partial (controlling for size)
participation ratio	-0.851	0.0018	-0.691 ($p = 0.027$)
$\cos_{\text{dep}}(v_{\text{CAA}}, v_{\text{letter}})$	$+0.812$	0.0043	$+0.615$ ($p = 0.058$)
$\cos(v_{\text{CAA}}, v_{\text{letter}})$	$+0.797$	0.0057	$+0.564$ ($p = 0.089$)
$\log_{10}$(params)	-0.686	0.0285	$+0.020$ ($p = 0.956$, given PR)

However, across the four evaluated models up to 7B, between 3 and 27 datasets retain at least one such block. Projecting out the polarity axis eliminates this remainder. On the 135 datasets where polarity is defined, adjusted mutual information with the letter drops from 0.058–0.117 under rank-one to 0.002–0.009 under rank-two. Consequently, no dataset across any of the five models retains a letter-pure block (Wilcoxon $p < 10^{-13}$ on the four smaller models; Qwen-1.5-14B reaches absolute zero under rank-one alone, Appendix B.3; Table 8). The cluster statistic thus provides the rank argument of §3.2 with a third independent line of evidence.

4 What the Audited Geometry Is Good For

The audited geometry forecasts behavioral reliability well; however, it is a poor optimization target. Further geometry results are provided in Appendix B.

4.1 Geometry, not size, predicts behavior

Table 3 reports all four predictors. The participation ratio correlates with behavioral anti-steerability at $r = -0.851$ ($p = 0.0018$), outperforming $\log_{10}$ parameters (-0.686, $p = 0.0285$). Also, it proves more robust: its partial correlation survives controls for parameter count (-0.691), hidden width (-0.662), architectural family (-0.665), and read-out depth (-0.675), all at $p < 0.04$, whereas the cosine predictor fails a family control ($+0.345$, $p = 0.33$).

Three Cautions. With only ten models, partial correlation is statistically fragile: dropping any single model shifts the PR partial into $[-0.779, -0.613]$ ($p \in [0.013, 0.079]$). We thus report it to corroborate the within-family scaling ladder rather than as independent evidence. Crucially, this predictive power operates across models, not within them: intra-model PR correlates with behavioral anti-steerability at just -0.06 to -0.47.

4.2 The proxy classifies but does not optimize

Assumption 1 (Sign consistency)**.** $\text{sign}(s_i) = \text{sign}(d_i^\top v)$ *for all but a small fraction of samples.*

On held-out data, mean agreement is 0.811 (Llama-2-7B) and 0.866 (Qwen-1.5-14B). Braun et al. (2025) establish directional agreement as a dataset-level predictor (a mean cosine ranked across datasets); thus, our per-sample measurement extends their claim rather than merely replicating it. The stronger premise, implicit whenever a method is evaluated on geometry, is that improvements in the proxy imply behavioral improvements. We test this with a reliability-optimal (RO) direction: an analytical instrument yielding directions with controlled geometric improvements. It minimizes the expected anti-agreement of a von Mises-Fisher (Scott et al., 2021) mixture fitted to the debiased differences, subject to a performance floor requiring it to retain at least 90% of CAA's steering efficacy. This floor renders the problem non-degenerate; without it, the optimizer simply returns a direction with a small positive projection onto every component, attaining near-zero anti-agreement but near-zero steering power (Appendix C).

Transfer Differs Substantially Between Models. On Llama-2-7B, a 0.159 reduction in the geometric proxy yields a 0.027 reduction in the behavioral anti-steerable fraction, a transfer rate of 17%. On Qwen-1.5-14B, a 0.059 proxy reduction yields 0.034, a rate of 58%. This disparity occurs despite

Table 4: Predictive utility of sign beyond magnitude. Negative differences favor the directional rule. To control for dataset overlap (12 families span 120 of 151 datasets), confidence intervals are clustered by both dataset and name stem. All intervals strictly exclude zero.

Subset	Metric	dir. − mag.	95% CI (dataset)	95% CI (stem)
all ($n = 1510$)	risk@80%	-0.0295	$[-0.0329, -0.0261]$	$[-0.0330, -0.0228]$
all ($n = 1510$)	AURC	-0.0339	$[-0.0397, -0.0285]$	$[-0.0390, -0.0235]$
informative ($n = 1296$)	risk@80%	-0.0343	$[-0.0379, -0.0308]$	$[-0.0373, -0.0288]$
informative ($n = 1296$)	AURC	-0.0395	$[-0.0457, -0.0336]$	$[-0.0439, -0.0299]$
gain vs no abstention	risk@80%	$+0.0708$	$[+0.0657, +0.0756]$	$[+0.0621, +0.0756]$

identical optimizers, identical datasets, comparable angular displacement ($28.7°$ versus $26.1°$), and no fallbacks to CAA. Having tested and falsified four candidate mechanisms (Appendix C.4), we report this threefold difference as an empirical fact with a stark practical consequence: a transfer rate measured on one model does not generalize to another.

5 Selective Steering

The audit results in two contrasting conclusions: $d_i^\top v$ predicts which samples steer incorrectly, while optimizing the vector to improve this quantity does not improve behavior. This section develops the method motivated by these constraints and evaluates it against three abstention baselines.

5.1 Method

A method conforming to both constraints does not re-aim the steering vector; instead, it abstains. We apply the intervention where $d_i^\top v \geq \tau$ and withhold it elsewhere. For a given coverage c,

$$\text{selective risk} = P(s_i < 0 \mid \text{retained}), \qquad \text{efficacy} = c \cdot \frac{\mathbb{E}[s_i \mid \text{retained}]}{\mathbb{E}[s_i]}. \tag{3}$$

This trade-off is explicit: abstaining reduces harm but also decreases the total steering delivered. We report the risk-coverage curve, summarized by its area (AURC) so that decision rules are compared across the entire operating range rather than at a post hoc threshold.

Baselines. Since any form of abstention inherently reduces harm, a proposed rule must outperform alternatives at matched coverage. *Random* abstention serves as the performance floor. *Magnitude* abstains based on $|d_i^\top v|$, retaining confidence but discarding the sign. This acts as the primary baseline: if it matches the performance of the signed rule, then sign consistency is not the operative mechanism. *Norm* abstains based on $\|d_i\|$, with no reference to the steering direction. On datasets where almost no projections are negative, these rules coincide and yield no discriminative information. Consequently, we evaluate the method both across all datasets and strictly on this informative subset.

5.2 Results

Pooled across $1{,}510$ pairs at 80% coverage, the directional rule yields a selective risk of 0.2526, outperforming magnitude (0.2821), norm (0.3037), random (0.3230), and no abstention (0.3234). The corresponding AURC values are 0.1845, 0.2185, 0.2754, and 0.3226 (Appendix D.2; Table 9). Anti-steerability thus falls from 0.323 to 0.253 without compromising efficacy. Indeed, efficacy marginally improves (median 1.040 across $1{,}278$ defined pairs) because abstention successfully discards negative slopes. As shown in Table 4, the primary test excludes zero on all four pooled measures under both clustering units, and individually across all ten models (per-model results in Appendix D.2; full procedures in Appendix D).

The Advantage Holds at Every Tested Coverage, and Grows. The directional rule outperforms the magnitude rule at all seven tested coverages, with the margin widening monotonically from -0.010 at 95% coverage to -0.052 at 50% (Fig. 3).

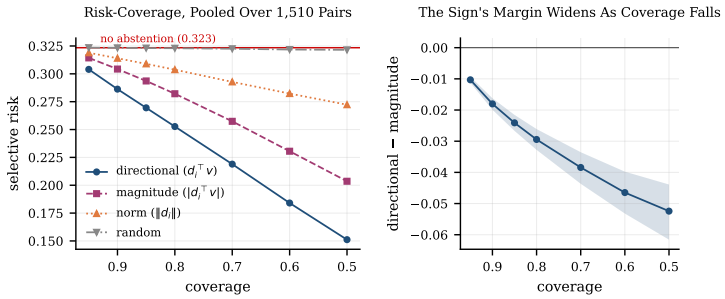


Figure 3: **Left**: Risk-coverage curves for four abstention rules across 1,510 (dataset, model) pairs. The signed rule dominates the magnitude rule (which discards sign), which itself outperforms direction-agnostic baselines. **Right**: Performance margin of the signed over the magnitude rule (95% dataset-clustered confidence band). Isolating sign consistency as the operative mechanism, this margin widens monotonically under abstention, so the score yields true discriminative information.

The Advantage Is Not Residual Template Alignment. Because d_i and the letter-logit slope s_i both derive from the answer letters, they are functions of σ_i. The performance gain could thus theoretically stem from residual template structure rather than concept-level sign prediction. The data refutes this: pooled across all pairs, the directional-over-magnitude advantage correlates *negatively* with residual letter decodability after the rank-one projection ($r = -0.117$), and never correlates positively within any model. Instead, it tracks the fraction of negative projections ($r = +0.358$) and retained-set overlap ($r = -0.392$), confirming the operative mechanism.

A Certified Operating Point. Post hoc τ selection is invalid. We instead use fixed-sequence testing across a quantile grid from conservative to permissive. Stopping at the first Clopper-Pearson upper-bound failure avoids multiplicity correction (Appendix D.1). Using disjoint test-split halves for calibration and evaluation, the procedure certified 955 of 1,510 pairs. Realized risk met the 0.10 target on 843 of these (88.3% vs a nominal 90%, well within binomial noise) at 0.591 mean coverage; otherwise, it safely declined certification. Tellingly, it refuses 143 of 151 datasets for the 0.5B model. This is the correct behaviour under high base risk, offering a stark demonstration of the scaling law.

Deployability and What the Probe Is Reading. Gating on d_i needs both candidate answers. Since these are absent in open-ended deployment, we fit a linear probe $h_i \mapsto \text{sign}(d_i^\top v)$ to the base-prompt activation, bypassing the need for answer tokens. Across all ten models, this probe achieves an AUC of 0.863–0.953, with its margin over a random baseline strictly excluding zero (Table 5).

The prompt contains "(A)…(B)…"; thus, h_i encodes the target letter, raising the concern that a sign probe merely routes its predictions through this template. We test this using an identical probe trained to predict the letter itself. Although the letter is highly decodable (AUC 0.769–0.958), the pooled per-dataset correlation between letter and sign recovery is just $r = +0.094$ ($p = 3.6 \times 10^{-4}$, $n = 1439$). This 1% shared variance contradicts the strong positive correlation required by template routing. Stratifying the datasets confirms this: in the bottom decile (letter AUC 0.682), the sign probe achieves an AUC of 0.867, barely dropping from the 0.888 achieved in the top decile (letter AUC 0.998; Fig. 5). Thus, a 0.32 collapse in letter readability yields a mere 0.021 drop in sign AUC. Sign recovery remains robust (0.766) even when letter decodability falls to random chance (0.528). Such divergence precludes the possibility that the letter mediates the sign prediction (Appendix D.3).

6 Conclusion

The control the steering literature uses against answer-order bias protects the mean but gives no per-sample guarantee. Measured as directions rather than as behavioral variance, the standard construction imposes two near-orthogonal surface axes; therefore, the field's rank-one correction removes only half of this structure: the other half remains decodable at ceiling on every model and costs 0.037 of anti-steerability on Llama-2-7B, while without any correction a standard cluster-then-route primitive recovers the letter on all 48,320 clusterings. How far these axes reach the deployed vector is governed by representational resolution. The resulting geometry is a valuable signal but not a reliable objective, and selective steering is the method that distinction licenses.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing and to retrieve and discover related work. We used generative AI tools to support research execution: auditing the code and analysis pipeline, writing the verification, and cross-checking the reported numbers against the experiment output files. We have not used generative AI tools for research ideation or for proving mathematical claims, and generating synthetic datasets does not apply to this work. We have reviewed all AI-assisted work. The research idea, the original manuscript, and the experimental design are the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The models, corpus handling, splits, and hyperparameters are specified in §2 and Appendix A, including the checkpoints and extraction layers (Appendix A.1). All data are public: the corpus is released by Perez et al. (2023), and every checkpoint is available from its public model page. Splits use per-dataset seeds, so every split is reproducible from the dataset name. The implementation code is provided as supplementary material, with a single driver, the full test suite described in Appendix E, and a provenance record of the environment used (Python 3.12.14, PyTorch 2.11.0, Transformers 5.17.0, CUDA 12.8, two NVIDIA L40 GPUs). Pooled statistics reproduce to within 2×10^{-4} across independent runs.

REFERENCES

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

Alex Ferrando, Xavier Suau, Jordi Gonzàlez, and Pau Rodríguez. Dynamically scaled activation steering. *CoRR*, abs/2512.03661, 2025.

Anna Hedström, Salim I. Amoukou, Tom Bewley, Saumitra Mishra, and Manuela Veloso. To steer or not to steer? mechanistic error reduction with abstention for language models. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Ceyhun Efe Kayan and Li Zhang. Prototype-based dynamic steering for large language models. *CoRR*, abs/2510.05498, 2025.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre L. Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *ICLR*. OpenReview.net, 2025.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger B. Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In *ACL (Findings)*, volume ACL 2023 of *Findings of ACL*, pp. 13387–13434. Association for Computational Linguistics, 2023.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *NAACL-HLT (Findings)*, volume NAACL 2024 of *Findings of ACL*, pp. 2006–2017. Association for Computational Linguistics, 2024.

Shivam Raval, Hae Jin Song, Linlin Wu, Abir Harrasse, Jeff M. Phillips, Fazl Barez, and Amirali Abdullah. Curveball steering: The right direction to steer isn't always linear. *CoRR*, abs/2603.09313, 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In *ACL (1)*, pp. 15504–15522. Association for Computational Linguistics, 2024.

Pau Rodríguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *ICLR*. OpenReview.net, 2025.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *ICLR*. OpenReview.net, 2024.

Tyler R Scott, Andrew C Gallagher, and Michael C Mozer. Von mises–fisher loss: An exploration of embedding geometries for supervised learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10592–10602. IEEE, 2021.

Vansh Sharma and Venkat Raman. Steering conceptual bias via transformer latent-subspace activation. *CoRR*, abs/2506.18887, 2025.

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 *NeurIPS*, 2024.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *CoRR*, abs/2308.10248, 2023.

Arthur Vogels, Benjamin Wong, Yann Choho, Annabelle Blangero, and Milan Bhan. In-distribution steering: Balancing control and coherence in language model generation. In *LREC*, pp. 2076–2089. ELRA Language Resource Association / ACL, 2026.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. Adaptive activation steering: A tuning-free LLM truthfulness improvement method for diverse hallucinations categories. In *WWW*, pp. 2562–2578. ACM, 2025a.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for llms via dynamic steering vectors. In *ICLR*. OpenReview.net, 2025b.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *ICLR*. OpenReview.net, 2024.

Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. Unibias: Unveiling and mitigating LLM bias through internal attention and FFN manipulation. In *NeurIPS*, 2024.

Andy Zou, Long Phan, Sarah Li Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405, 2023.