WHEN IS A STEERING DIRECTION STATISTICALLY SUPPORTED?

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering uses directions constructed from examples to steer model behavior and interpret representations. Recent work shows that variability across directions from small construction samples can obscure differences between prompt formats. When does variability in a steering vector translate into variability in model responses? In the evaluated A/B constructions, we find much of the sampling variability lies along an answer-position axis. For myopic reward, proportional stratification uses this structure to improve response repeatability over ordinary sampling at fixed target and sample size. In other experiments, an answer's probability can initially increase along an estimated direction but decrease along its target direction. To assess what construction data support about the target direction, we introduce SteerCert. It reports a verdict for claims holding throughout a confidence set for the *population contrast*, the expected activation difference. Valid confidence-set coverage controls false population claims. Our theory bounds the independent data needed to estimate the target direction precisely: necessary and sufficient sample-size scales match at fixed confidence under stated assumptions.

1 INTRODUCTION

Activation steering often constructs vectors by averaging activation differences between contrasting examples (Turner et al., 2024; Rimsky et al., 2024; Li et al., 2023; Zou et al., 2023). These vectors are used to steer model behavior and interpret representations through projections and comparisons between directions (Arditi et al., 2024). Evaluating one such vector does not reveal how model responses depend on its construction examples. With small construction samples, variation across estimated directions can obscure differences between prompt formats (Braun et al., 2025). When does variability in a steering vector translate into variability in model responses?

We find structure in the sampling variability of estimated directions that can be used to improve response repeatability. In the evaluated A/B constructions, a substantial share of this variability lies along the answer-position axis. This axis contrasts examples whose behavior-matching answer is A versus B. Proportional answer-position stratification reduces angular error versus ordinary sampling while preserving the fixed dataset's mean contrast and sample size. For myopic reward, it also improves response repeatability (Secs. 5.1–5.2). An exploratory diagnostic shows that response deviations depend on the direction of estimation error at fixed angular error (Sec. 5.2).

Improved repeatability alone does not establish that properties observed at an estimated direction hold at its target. SteerCert assesses which such claims are statistically supported. Its target is the *population contrast*, the expected activation difference under a fixed construction recipe and sampling distribution. Analyses that reuse an estimated direction share uncertainty about this contrast. SteerCert (i) *constructs* a confidence set for the population contrast; (ii) *assesses* which claims hold for every mean in the set; and (iii) *reports* the corresponding verdict. Valid confidence-set coverage controls false population claims (Sec. 3).

How much independent data are needed to attain the directional precision assessed by SteerCert? Our theory shows that necessary and sufficient sample-size scales match at fixed confidence under stated covariance and local population conditions (Thm. 1). We also analyze a model in which examples generated from the same prompt template share a component of variation. Increasing examples per template reduces within-template variability but leaves between-template variability, limiting uniform directional precision at a fixed number of independent templates (Thm. 2).

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.

We examine how directional uncertainty affects claims about local responses, the derivatives of answer probabilities at zero intervention strength. For fixed datasets with known mean contrasts, local responses are sometimes positive at estimated directions but negative at target directions (Sec. 5.3). For public constructions, the target is the direction of the unknown population contrast under the data-generating distribution. We assess positivity over all *compatible directions*, the directions of the nonzero means in the confidence set. This is statistically supported in only a small fraction of evaluated cases, with bootstrap/approximate calibration status (Sec. 5.4).

Contributions.

- **Sampling design can improve response repeatability at a fixed target and sample size.** We find answer-position structure in sampling variability; proportional stratification reduces angular error versus ordinary sampling and improves response repeatability for myopic reward (Secs. 5.1–5.2).
- **SteerCert identifies statistically supported population claims.** It reports a verdict for claims valid throughout a confidence set; valid coverage controls false claims (Sec. 3).
- **Theory characterizes the independent data required for directional precision.** At fixed confidence and under stated assumptions, we match necessary and sufficient sample-size scales and show the limits of adding examples within templates (Thms. 1–2).
- **Positive local responses can reverse from estimated to target directions**, as seen in fixed datasets with known mean contrasts. In evaluated public constructions, positivity over all compatible directions is rarely statistically supported under bootstrap/approximate calibration (Secs. 5.3–5.4).

2 Setup and Motivation

We introduce the setup and use two examples to distinguish detection from directional precision.

2.1 Preliminaries and Setup

Statistical setup. We specify the population target by fixing the model, layer, token rule, contrast construction, and target sampling distribution. For steering constructions based on mean contrasts, let $W_t \in \mathbb{R}^d$ denote the activation contrast from one sampling unit. The construction and sampling scope define the unit. In the baseline setup, $W_1, \ldots, W_T$ are iid samples from the declared distribution P on $\mathbb{R}^d$. We assume the first-moment condition $\mathbb{E}_P\|W_1\|_2 < \infty$ so that the population target is defined. Define the empirical contrast, population contrast, and population magnitude by

$$\widehat{\Delta} := \frac{1}{T}\sum_{t=1}^{T} W_t, \qquad \Delta := \mathbb{E}_P W_1, \qquad a := \|\Delta\|_2. \tag{1}$$

When $a > 0$, the population direction is $u_\star := \Delta/a$. Fresh samples under the same construction and distribution estimate the same Δ; changing the source or construction can change the target. When P assigns equal weight to each row's contrast in a fixed file, Δ is the finite-file mean. Additional assumptions for confidence-set coverage and theoretical bounds accompany each result.

Sampling units and scope. For inference within a fixed prompt template, W_t may be a matched-pair difference $\phi(x_t^+) - \phi(x_t^-)$, where ϕ denotes the activation under the fixed model, layer, and token rule. For inference across templates, W_t may summarize contrasts from one independently sampled template. The summary rule and template sampling distribution are part of the target specification; the average above weights template summaries equally. In either case, T counts independent sampling units. Pairs sharing a template do not supply independent observations of variation across templates. Confidence-set coverage must match the declared sampling unit, distribution, and sampling design.

Directional targets and claims. When $a > 0$, the population line is represented by $\{u_\star, -u_\star\}$, whereas the population direction $u_\star$ retains the sign specified by the contrast construction. At $a = 0$, neither is defined. Fix an angular resolution $0 \le \varepsilon_0 < \pi/2$ and a possibly data-dependent reported center $\widehat{u} \in \mathbb{S}^{d-1}$. A *unit axis* is a unit vector representing a line; $\widehat{u}$ and $-\widehat{u}$ represent the same estimated line. A *signal claim* asserts $a > 0$; the following claims additionally require

$$\text{line claim:} \quad \arccos\left|\widehat{u}^\top u_\star\right| \le \varepsilon_0, \qquad \text{oriented-direction claim:} \quad \arccos\left(\widehat{u}^\top u_\star\right) \le \varepsilon_0.$$

Verdict	Condition	Interpretation at resolution ε_0
ABSTAIN	S^c	Insufficient evidence of a nonzero mean.
SIGNAL ONLY	$S \cap L^c$	Nonzero mean; a line claim about $\widehat{u}$ is unsupported.
UNORIENTED LINE	$S \cap L \cap O^c$	The mean is nonzero and its direction is localized up to sign; orientation about $\widehat{u}$ is unsupported.
ORIENTED DIRECTION	$S \cap L \cap O$	The mean is nonzero and its direction lies within ε_0 of the reported center $\widehat{u}$.

Table 1: Four verdicts at fixed angular resolution. The three positive verdicts express increasingly specific population claims; validity depends on calibration. Abstention asserts neither a zero mean nor an ineffective intervention. The confidence-ball rule cannot report the unoriented-line verdict.

The line claim bounds the *line error*, allowing proximity to $\pm\widehat{u}$; the oriented-direction claim requires proximity to $\widehat{u}$. Sec. 3 specifies when these claims are statistically supported.

2.2 Motivating Examples

An arrow without a population direction. Consider $W_t = \Delta + \sigma\xi_t$ for $\sigma > 0$, where $\xi_1, \ldots, \xi_T$ are iid samples from $N(0, I_d)$. At $\Delta = 0$, the sample mean is nonzero almost surely, so normalization returns $\widehat{u} := \widehat{\Delta}/\|\widehat{\Delta}\|$ even though $u_\star$ is undefined. By isotropy, the normalized direction is uniform on the sphere for every T. For two independent samples of size T with normalized means $\widehat{u}_1$ and $\widehat{u}_2$,

$$\mathbb{E}\left\|\widehat{\Delta}\right\|^2 = \frac{d\sigma^2}{T}, \qquad \|\widehat{u}\| = 1, \qquad \mathbb{E}(\widehat{u}_1^\top \widehat{u}_2)^2 = \frac{1}{d}. \tag{2}$$

Increasing T reduces the expected squared norm of the mean, but normalization removes this scale information. The expected squared cosine between independent reconstructions therefore remains $1/d$. This motivates abstention about a population direction; the perturbation's behavioral effect remains a separate question. Randomizing contrast signs likewise gives a zero target mean, but rerandomizations of one fixed pool share observations and need not satisfy the $1/d$ identity (Appendix D.1.1).

A detectable contrast with an unresolved direction. For $\Delta \neq 0$ in the same model, a population direction exists. The component of the sample-mean error perpendicular to Δ rotates the estimate, and normalization preserves this angular error. A confidence ball barely excluding zero contains means pointing in substantially different directions (Fig. 1). Excluding zero supports a nonzero contrast, but locating its direction within the requested angle requires a sufficiently narrow range of compatible directions. The angular resolution specifies the desired precision around a reported center; the continuous angular radius describes the uncertainty and allows assessment at other resolutions.

3 Statistical Support for a Steering Direction

The motivating examples leave a decision problem: which population claim do the construction data support at a chosen angular resolution? We define SteerCert's decision rule, specialize it to a confidence ball, and explain how confidence-set coverage controls false claims.

SteerCert. We propose SteerCert, a decision rule that assesses population claims in three steps:

1. **Construct.** From the construction sample D, construct a confidence set using the chosen uncertainty procedure for the declared population target, sampling unit, and design.
2. **Assess.** At the reported center $\widehat{u}$ and resolution ε_0, evaluate the signal (S), line (L), and orientation (O) conditions defined below.
3. **Report.** Report the *verdict* in Table 1 corresponding to the most specific claim holding for every mean in the set, or ABSTAIN if none holds, together with the calibration status.

Four possible verdicts. Let $\mathfrak{C}(D) \subseteq \mathbb{R}^d$ be the chosen confidence set for Δ; a population claim is *statistically supported* when it holds for every mean in this set. The compatible directions are

$$\mathcal{U}(D) := \left\{ \frac{\theta}{\|\theta\|} : \theta \in \mathfrak{C}(D),\ \theta \neq 0 \right\}.$$

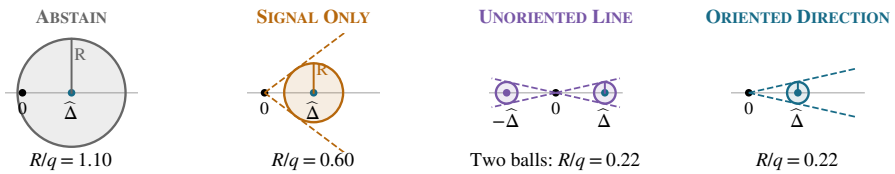

Figure 1: **Illustration of four verdicts.** Radius-R balls at $\widehat{\Delta}$, or their union at $\pm\widehat{\Delta}$ for the unoriented-line region; $q = \|\widehat{\Delta}\| = 1$ and resolution $\varepsilon_0 = 0.35$ rad. Dashed rays delimit compatible directions.

For the reported center $\widehat{u}$ and resolution $0 \le \varepsilon_0 < \pi/2$ from Sec. 2.1, define the events

$$\begin{aligned} \text{signal:} \quad & S := \{0 \notin \mathfrak{C}(D)\}, \\ \text{line:} \quad & L := \{|\widehat{u}^\top v| \ge \cos\varepsilon_0 \text{ for every } v \in \mathcal{U}(D)\}, \\ \text{orientation:} \quad & O := \{\widehat{u}^\top v \ge \cos\varepsilon_0 \text{ for every } v \in \mathcal{U}(D)\}. \end{aligned}$$

Condition S excludes zero; L and O impose the line and oriented-direction inequalities on every compatible direction, with $O \subseteq L$. The verdict conditions in Table 1 are mutually exclusive and cover every data realization, with superscript c denoting event complementation.

Decision rule for a confidence ball. For a Euclidean ball about the sample mean, the rule compares radius and center norm. For $0 < \delta < 1$, let $R \ge 0$ satisfy $\mathbb{P}\{\|\widehat{\Delta} - \Delta\| \le R\} \ge 1 - \delta$ and put $q := \|\widehat{\Delta}\|$. If $q > R$, the ball's directions form a cap with center $\widehat{u} := \widehat{\Delta}/q$ and radius $\widehat{\varepsilon} := \arcsin(R/q)$. The rule reports ORIENTED DIRECTION when $q > R$ and $R \le q\sin\varepsilon_0$, SIGNAL ONLY when $q\sin\varepsilon_0 < R < q$, and ABSTAIN when $q \le R$. For $q > R$, the cap lies in the open hemisphere facing $\widehat{u}$, so L implies O at $\varepsilon_0 < \pi/2$; the ball never reports an unoriented-line verdict (Thm. 6). Fig. 1 shows these verdicts and a schematic unoriented-line example: two separated balls.

Coverage and calibration. Suppose the chosen confidence-set procedure satisfies $\mathbb{P}\{\Delta \in \mathfrak{C}(D)\} \ge 1 - \delta$ under the declared sampling distribution, with $0 < \delta < 1$. If we report only supported claims, a false report requires the confidence set to miss Δ, giving

$$\mathbb{P}\{\text{a false claim is reported}\} \le \mathbb{P}\{\Delta \notin \mathfrak{C}(D)\} \le \delta.$$

The bound covers data-dependent selection of multiple claims supported by the same set (Neyman, 1937; Scheffé, 1953). This unconditional bound does not imply the same error bound conditional on reporting a particular positive verdict; Appendix E gives an empirical example. Asymptotic coverage yields an asymptotic false-claim bound. Nominal error settings alone do not establish coverage, so each verdict is reported with its confidence set's calibration status. Joint-bootstrap and bounded-data balls have asymptotic and finite-sample guarantees, respectively, under their own assumptions (Appendix D; see Appendix D.4 for the joint multiplier ball algorithm).

4 THEORETICAL LIMITS ON DIRECTIONAL PRECISION

Sec. 3 defines when population claims are statistically supported by a confidence set; we now ask how much independent data directional precision requires. We compare independent observations with observations that share template effects under the local population setting below.

Local population setting. We consider a population direction whose approximate location is known. Fix an integer $d \ge 2$, a contrast magnitude $a > 0$, and a base direction $u_0 \in \mathbb{S}^{d-1}$. For a resolution $0 < \varepsilon < \pi/2$ and a common universal neighborhood constant $K > 0$ used in the results below, restrict the unknown direction u to

$$\mathcal{U}_\varepsilon := \{u \in \mathbb{S}^{d-1} : \angle(u, u_0) \le K\varepsilon\}.$$

The notation $\mathcal{U}_\varepsilon$ suppresses K. For each sampling scheme, a *local population setting* specifies independent sampling units and a class of laws with finite second moments, population contrasts $\Delta = au$ for $u \in \mathcal{U}_\varepsilon$, and declared covariance restrictions. The estimation procedure is supplied with a, u_0 and the covariance bounds or matrices specifying the class. Covariance conditions determine the noise scale; independence conditions determine how averaging reduces uncertainty.

Required sample size. We use the angle in the oriented-direction claim defined in Sec. 2.1. A possibly randomized procedure using all observations succeeds if it reports a unit vector $\widehat{w}$ with $\arccos(u^\top \widehat{w}) \le \varepsilon$; abstention counts as failure. For $0 < \delta < 1$, the same procedure must succeed with probability at least $1 - \delta$ under every law in the class. The *required sample size* is the least positive integer n allowing some procedure to achieve this guarantee from n independent units.

4.1 Sample Size and Angular Precision

Resolving a direction requires controlling noise perpendicular to the population contrast. We first consider independent observations under a common covariance bound and give necessary and sufficient sample-size scales for direction estimation.

Theorem 1 (Directional precision under a covariance bound)**.** *There exist universal constants $c > 0$, $K \ge 1$, $\delta_0 \in (0, 1)$, and $0 < \varepsilon_{\max} \le \pi/(4K)$ such that the following holds whenever $0 < \delta < \delta_0$ and $0 < \varepsilon \le \varepsilon_{\max}$. In the local population setting above, fix $\sigma > 0$ and let $\mathcal{P}_\varepsilon$ consist of all distributions with $\mathbb{E}_P W = au$ for $u \in \mathcal{U}_\varepsilon$ and $\mathrm{Cov}_P(W) \preceq \sigma^2 I_d$. For iid sampling from this class, let $m_\star$ be the required sample size. Then*

$$\max\left\{1, \frac{c\sigma^2(d-1)}{a^2 \sin^2 \varepsilon}\right\} \le m_\star \le \left\lceil \frac{\sigma^2 d}{\delta a^2 \sin^2 \varepsilon} \right\rceil. \tag{3}$$

The covariance condition bounds variance in every unit direction by σ^2, while allowing covariance matrices to vary across distributions. The lower bound concerns a uniform guarantee over the stated class and neighborhood; individual distributions may permit accurate estimation with fewer observations. At fixed confidence, the necessary and sufficient sample-size scales match up to constants. They grow linearly with dimension, and halving a sufficiently small target angle approximately quadruples these scales. Detecting a nonzero population contrast can require fewer samples than achieving uniform directional accuracy. Appendix B establishes this separation under the same covariance bound and gives the supporting Gaussian lower bound for procedures that may abstain.

4.2 Variation Across Independent Templates

With T independent templates and m examples each, we ask how far increasing m alone can improve directional precision. Consider the additive random-effects model (Laird & Ware, 1982),

$$z_{it} = \Delta + \delta_t + \varepsilon_{it}, \qquad \delta_t \overset{\text{iid}}{\sim} Q_b, \qquad \varepsilon_{it} \overset{\text{iid}}{\sim} Q_w.$$

Let $\mathcal{F}(\Sigma_b, \Gamma)$ be the class of all pairs (Q_b, Q_w) of centered laws with finite second moments and respective covariances $\Sigma_b \succ 0$ and $\Gamma \succeq 0$. All template effects and within-template errors are mutually independent. A template mean has covariance $\Sigma_b + \Gamma/m$.

Theorem 2 (Directional precision across independent templates)**.** *The constants $c, K, \delta_0, \varepsilon_{\max}$ in Thm. 1 can be chosen so that the following holds for the same ranges of δ and ε. For the template sampling scheme above with an integer $m \ge 1$ fixed, use the population class $(Q_b, Q_w) \in \mathcal{F}(\Sigma_b, \Gamma)$ and $u \in \mathcal{U}_\varepsilon$ from the local population setting above. Let $T_\star$ be the required sample size in templates. Put $\Sigma_m := \Sigma_b + \Gamma/m$ and let the columns of $B_0 \in \mathbb{R}^{d \times (d-1)}$ be an orthonormal basis of $u_0^\perp$. If*

$$\frac{\|B_0^\top \Sigma_m u_0\|_2}{u_0^\top \Sigma_m u_0} \tan(K\varepsilon) \le \frac{1}{2}, \tag{4}$$

then, setting $\tau_F^{(m)}(u_0) := \mathrm{tr}[(B_0^\top \Sigma_m^{-1} B_0)^{-1}]$, we have

$$\max\left\{1, \frac{c\tau_F^{(m)}(u_0)}{a^2 \sin^2 \varepsilon}\right\} \le T_\star \le \left\lceil \frac{32\, \tau_F^{(m)}(u_0)}{\delta a^2 \sin^2 \varepsilon} \right\rceil. \tag{5}$$

At fixed confidence, necessary and sufficient template-count scales match up to constants under these conditions. Here $\tau_F^{(m)}(u_0)$ summarizes the covariance dependence: when $\Sigma_m = \sigma^2 I_d$, it equals $\sigma^2(d-1)$, matching the dimension and noise dependence in Thm. 1. The upper bound is attainable with the supplied parameters under condition (4). This sufficient condition holds for sufficiently

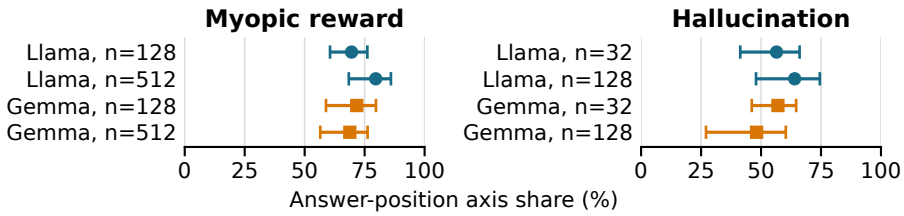

Figure 2: **Sampling variability is concentrated along the answer-position axis.** Points show variance shares from 30 normalized directions per condition under ordinary sampling; bars are pointwise 95% bootstrap intervals conditional on the fixed files (Appendix E).

small resolutions at fixed Σ_m and u_0, and whenever u_0 is an eigenvector of Σ_m. More examples per template reduce Γ/m but leave Σ_b unchanged, so

$$\tau_F^{(m)}(u_0) \downarrow \tau_F^{(b)}(u_0) := \mathrm{tr}[(B_0^\top \Sigma_b^{-1} B_0)^{-1}] > 0 \qquad (m \to \infty).$$

Thus increasing m alone cannot ensure arbitrarily fine uniform directional accuracy at fixed T; the resulting lower bound on $\sin\varepsilon$ decreases as $T^{-1/2}$ with a, Σ_b, and u_0 fixed. Appendix C gives the estimator construction, proofs, scope of both bounds, and extensions to procedures that may abstain.

Scope of directional guarantees. These results concern estimation of the population direction. Inferring an intervention effect additionally requires assumptions on the downstream response or separate behavioral measurements. Given a specified response gradient, Sec. 5.4 examines when a positive local response is supported over all compatible directions.

5 Empirical Evaluation

We study the structure of sampling variability in steering directions and its consequences for model responses. Using known finite-file means as references, we compare directional error and response repeatability across sampling designs. We assess which local-response claims hold over compatible directions for finite-file targets and unknown population contrasts in public constructions.

Data and models. We use contrast pairs from four CAA behaviors (Rimsky et al., 2024) and a fixed collection of 135 persona and 17 advanced-AI-risk (xrisk) evaluation files (Perez et al., 2023). The CAA response-repeatability experiments use Llama-2-7B-chat at block 13 and Gemma-2-2B-IT at block 12. The 152-file directional-precision comparisons use these model–layer settings and additionally include Qwen2.5-0.5B-Instruct and Qwen2.5-7B-Instruct at block 11. The CAA local-response analyses use Llama-2-7B-chat at blocks 11, 12, and 13. All block indices are zero-based.

5.1 Structure and Reduction of Sampling Variability

In the following two experiments, we ask whether sampling variability has structure that can be used to improve directional precision without changing the target or sample size.

Variance decomposition. We ask how much sampling variance lies along the answer-position axis. We repeatedly sample each fixed file and decompose the variance across normalized directions into the component along the answer-position axis and its orthogonal remainder. The *answer-position axis* contrasts the A- and B-matching stratum means in the subspace perpendicular to the full-file mean direction. For CAA myopic reward ($n = 128, 512$) and hallucination ($n = 32, 128$), we use 30 ordinary-sampling constructions per condition on Llama and Gemma (Appendix E).

Fig. 2 summarizes the results. In the evaluated CAA constructions, a substantial share of sampling variance lies along the answer-position axis. The share is 69–80% for myopic reward. For hallucination, it is 56–64% for Llama and 48–57% for Gemma. This decomposition concerns repeated estimates of each fixed target; answer position and content are not independently manipulated.

Table 2: **Angular error falls at a fixed target and sample size.** Mean angular MSE (rad^2), $n = 500$; equal file weights within families. Reductions from ordinary sampling have pointwise 95% Monte Carlo intervals conditional on files (paired resampling of 200 construction repetitions per file).

Model	Family	Ordinary MSE	Stratified MSE	MSE reduction (%) [95% MC]
Llama-2-7B	persona	0.0790	0.0289	63.3 [62.6, 64.0]
	xrisk	0.0077	0.0030	61.1 [59.5, 62.8]
Gemma-2-2B	persona	0.2259	0.0999	55.8 [55.0, 56.6]
	xrisk	0.0071	0.0032	54.8 [53.2, 56.5]
Qwen2.5-0.5B	persona	1.8407	0.4420	76.0 [75.8, 76.2]
	xrisk	0.3965	0.0404	89.8 [89.5, 90.1]
Qwen2.5-7B	persona	1.2134	0.2440	79.9 [79.6, 80.1]
	xrisk	0.1281	0.0181	85.9 [85.3, 86.4]

Table 3: **Response repeatability can improve at a fixed sample size.** Response variance ratios (answer-position stratification / ordinary sampling) at fixed intervention strength, with pointwise 95% paired bootstrap intervals (30 constructions; 40 questions in both answer orders).

Behavior	n	Llama	Gemma
Myopic reward	128	0.067 [0.038, 0.111]	0.425 [0.258, 0.679]
	512	0.052 [0.026, 0.113]	0.515 [0.341, 0.748]
Hallucination	32	0.230 [0.100, 0.503]	0.495 [0.276, 1.170]
	128	0.159 [0.078, 0.329]	0.310 [0.178, 1.067]

Sampling-design comparison. We compare ordinary sampling with proportional answer-position stratification on all 152 persona and xrisk files at the four model–layer settings above. Both designs draw $n = 500$ items without replacement in 200 paired repetitions per file. Weighting the A- and B-matching stratum sample means by their original file proportions preserves the finite-file mean target (Appendix E). Angular MSE uses signed angles to the exact file-mean direction, with equal file weights within families.

Proportional answer-position stratification reduces angular MSE by 54.8–89.8% relative to ordinary sampling in all eight model–family conditions (Table 2). For Llama persona, angular MSE falls from 0.0790 to 0.0289 rad^2. Appendix E reports resolution-specific verdict counts.

5.2 Response Repeatability and Directional Error

We examine how model responses vary across reconstructed directions. We compare response repeatability across sampling designs and relate angular error to deviations from a reference response.

Response repeatability. We ask whether controlling answer-position proportions makes responses more repeatable across reconstructions. We compare the two sampling designs at fixed sample size, inputs, and intervention strength. Myopic reward includes stratification by a fixed random partition. Response variance averages the variance across reconstructions of the behavior-matching answer letter's full-vocabulary probability over inputs. CAA comparisons (Sec. 5.1) use 30 paired constructions and 40 questions in both answer orders (Appendix F.1).

For myopic reward ($n = 128$, coefficient -1), stratified/ordinary response variance ratios are 0.067 on Llama and 0.425 on Gemma (Table 3). Mean probability reductions from no intervention differ by $+0.32$ pp $[-0.49, 1.14]$ and $+0.014$ pp $[0.005, 0.029]$, respectively (stratified minus ordinary; pointwise 95% paired bootstrap intervals). Llama's variance ratio to random stratification is $0.063\,[0.037, 0.103]$. For hallucination, response variance decreases on Llama; both Gemma intervals include one. Appendix F.2 gives mean effects, original response endpoints, and the Qwen comparison.

Angular error and response deviation. We ask whether constructions with larger angular error also produce larger deviations from the full-frame reference response. Using the same myopic-reward constructions and responses, we compute Spearman correlations separately for each model, sampling

Table 4: **Angular error correlates with response MSE under ordinary sampling.** Within-condition Spearman correlations for myopic reward at fixed intervention strength, with pointwise 95% paired bootstrap intervals (30 constructions; 40 questions in both answer orders).

Sampling design	n	Llama	Gemma
Ordinary	128	$0.868\,[0.684, 0.945]$	$0.851\,[0.574, 0.929]$
Stratified	128	$0.500\,[0.054, 0.757]$	$0.525\,[0.130, 0.741]$
Ordinary	512	$0.950\,[0.857, 0.982]$	$0.609\,[0.237, 0.825]$
Stratified	512	$0.106\,[-0.310, 0.492]$	$0.121\,[-0.268, 0.462]$

Table 5: **A positive estimated derivative need not hold at the target.** Denominator: 240,000 dependent evaluations/design (50 questions/behavior, both answer orders). Confidence sets have bootstrap/approximate calibration status; text intervals use paired bootstrap over 200 reconstructions.

Sampling design	Require positivity at / over	Correct positive (%)	Incorrect positive (%)	Withheld (%)
Ordinary	Estimated direction	75.233	2.420	22.347
	All compatible directions	0.000	0.000	100.000
Answer-position stratified	Estimated direction	77.883	1.008	21.109
	All compatible directions	3.058	0.000	96.942

design, and sample size. Response MSE averages squared differences from the full-frame mean direction's response probabilities over questions and answer orders (Appendix F.3).

For myopic reward under ordinary sampling, less accurate direction estimates tend to produce larger response deviations. At $n = 128$, the correlations are 0.868 for Llama and 0.851 for Gemma (Table 4). For stratified sampling at $n = 512$, the association is uncertain in both models, with intervals spanning zero. A separate evaluation on held-out constructions did not establish that squared angular error improves prediction of each input's signed probability deviation from the reference over an intercept-only predictor (Appendix F.3).

In an exploratory Llama myopic-reward diagnostic using the full-frame reference, answer-position-axis removal at matched angular error and intervention norm yields $0.225\,[0.156, 0.309]$ times the response MSE of random-axis removal (Appendix F.4).

5.3 Directional Uncertainty and Local Response Claims

We ask whether an initial increase in answer probability along a reconstructed direction also holds along the mean direction it estimates. We measure this initial response by the derivative of the behavior-matching answer letter's full-vocabulary probability at zero intervention strength. With the model and input fixed, the derivative along u is $g^\top u$, where g is the common response gradient. In a separate finite-file experiment, we compare derivatives along reconstructed directions and the full-frame mean direction, which is used only for evaluation. We reconstruct directions from 128 examples using ordinary sampling and proportional answer-position stratification on Llama-2-7B-chat, across four CAA behaviors and blocks 11–13. We assess local positivity over the same compatible directions, using the input's gradient; this is a response claim, not an additional SteerCert verdict. For a cap centered at $\widehat{u}$ with radius $\widehat{\varepsilon} < \pi/2$ and $c = g^\top \widehat{u} > 0$, positivity throughout is equivalent to $\widehat{\varepsilon} < \varepsilon_{\text{crit}} := \arcsin(c/\|g\|)$ (Appendix D.5).

A positive local derivative along a reconstructed direction can correspond to a negative derivative along the target direction. Under ordinary sampling, this sign mismatch occurs in 5,809 of the 186,367 evaluations with a positive estimated derivative, or $3.117\%\,[2.932, 3.308]$ (95% interval; Table 5). No positive claims were supported under ordinary sampling; answer-position stratification supported $3.058\%\,[2.965, 3.153]$ of all evaluations. For myopic reward at block 11, stratification reduces the median radius from 0.280 to 0.107 rad, yet supports no positive claims. The median across reconstructions of the median boundary over inputs with $c > 0$ is about 0.025 rad under both designs; support uses individual comparisons (Appendix G). The finite-file analysis withholds many positive local-response claims that are correct at the target, so limited support does not establish an unavoidable lack of information.

5.4 Supported Claims from Public Construction Data

We ask which positive local-response claims the twelve public CAA constructions support using all construction rows. Treating these rows as a sample from the generating distribution, we assess its population direction using confidence balls with bootstrap/approximate calibration status (Appendix H.1). The input panel and dose-zero probability derivative are those of Sec. 5.3.

Of 1,200 fixed, dependent input–layer evaluations, 956 have a positive center derivative, but only 50 retain positivity throughout the compatible directions. Support requires the cap radius to fall below the input-specific boundary $\varepsilon_{\mathrm{crit}}$ (Sec. 5.3); a positive center derivative alone is insufficient. Withholding support does not establish a negative response at the population direction.

6 Related Work

Steering directions. Activation steering modifies behavior along directions in activation space (Turner et al., 2024; Zou et al., 2023). Mean contrasts underlie CAA (Rimsky et al., 2024), refusal directions (Arditi et al., 2024), and persona vectors (Chen et al., 2025). Pahde et al. (2025) show that classifier-based CAVs can diverge from concept directions and propose pattern-based CAVs.

Construction variability and model responses. Braun et al. (2025) report variability in directions from small samples and link agreement among activation differences to steering effectiveness. Gao et al. (2026) examine answer-encoding effects on steering attribution and evaluate mapping-balanced construction. We decompose sampling variability in the evaluated A/B constructions and compare directional precision and response repeatability under proportional answer-position stratification versus ordinary sampling at fixed finite-file mean and sample size (Secs. 5.1–5.2).

Concept-direction uncertainty. CAV analyses quantify sampling variability, including for mean-difference estimators, and its consequences for sensitivity and thresholded scores (Wenkmann & Garreau, 2026; Schnoor et al., 2026a). We ask which signal, line, or orientation claims about a fixed recipe's population contrast are statistically supported at a specified resolution. SteerCert reports a verdict based on a confidence set; valid coverage controls false population claims. Under stated assumptions, we also bound the independent sample size required for directional precision. MERA (Hedström et al., 2025) calibrates intervention strength and when to abstain; SteerCert's abstention withholds a population claim. Appendix A gives detailed comparisons.

7 Conclusion

Studying construction variability can inform both sampling design and claims about the target direction. In our myopic-reward experiments, proportional answer-position stratification uses the observed structure to improve response repeatability over ordinary sampling at the same target and sample size. Improved repeatability alone does not establish that a property observed at an estimated direction holds at its target. SteerCert assesses which claims hold throughout a confidence set. In the evaluated public constructions, positive local responses are statistically supported over all compatible directions in only a small fraction of cases, with bootstrap/approximate calibration status. Supporting such claims depends on directional precision. Our theory characterizes the independent data needed to attain a given precision under stated assumptions.

Limitations, scope, and future work. Our empirical findings concern the evaluated models, inputs, and construction recipes; these comparisons do not isolate model-size or family effects. Greater repeatability of answer probabilities does not by itself establish more desirable behavior or improvements in free generation. The sampling-design results do not establish a benefit from selecting designs with SteerCert. The local-response analysis concerns derivatives at zero intervention strength and does not establish finite-dose effects. Guarantees for the declared population contrast require valid confidence-set coverage; our empirical confidence sets have bootstrap/approximate calibration status (finite-file coverage: Appendix G.4). We have not established whether the public-construction confidence sets are unnecessarily wide. Future work can test how broadly the observed structure occurs and evaluate confidence-set coverage and width across models and construction sources.

Reproducibility Statement

Appendix D specifies the confidence-set procedures and their assumptions. Appendices E–H give the settings, comparisons, and additional results for the main-text experiments; Appendix I.1 describes the compute environment. The experimental code is provided in the supplementary material. Full activation extraction also requires the cited weights and public construction data.

AI use statement

Generative AI tools assisted with literature search, research design, proof development, code development, data analysis and interpretation, and manuscript preparation. The authors take full responsibility for the final content.

References

Alaa Anani, Tobias Lorenz, Bernt Schiele, Mario Fritz, and Jonas Fischer. Certified circuits: Stability guarantees for mechanistic circuits. In *Proc. ICML*, 2026. URL https://arxiv.org/abs/2602.22968.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Proc. NeurIPS*, 2024. URL https://arxiv.org/abs/2406.11717.

R. R. Bahadur and L. J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Stat.*, 27(4):1115–1122, 1956. URL https://doi.org/10.1214/aoms/1177728077.

Thomas Ball, Shuo Chen, and Cormac Herley. Can we count on LLMs? the fixed-effect fallacy and claims of GPT-4 capabilities. arXiv preprint arXiv:2409.07638, 2024. URL https://arxiv.org/abs/2409.07638.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Proc. NeurIPS*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *Ann. Stat.*, 41(2):802–837, 2013. URL https://doi.org/10.1214/12-AOS1077.

Prajjwal Bhattarai and Tuka Alhanai. Signatures of steerability in activation space of language models. arXiv preprint arXiv:2609.14151, 2026. URL https://arxiv.org/abs/2609.14151.

Jayadev Billa. Predicting where steering vectors succeed. arXiv preprint arXiv:2604.15557, 2026. URL https://arxiv.org/abs/2604.15557.

Gilles Blanchard, Alexandra Carpentier, and Maurilio Gutzeit. Minimax Euclidean separation rates for testing convex hypotheses in $\mathbb{R}^d$. *Electron. J. Stat.*, 12(2):3713–3735, 2018. URL https://doi.org/10.1214/18-EJS1472.

Joschka Braun. Understanding unreliability of steering vectors in language models: Geometric predictors and the limits of linear approximations. Master's thesis, University of Tübingen, 2026. URL https://arxiv.org/abs/2602.17881.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models. In *Proc. ICLR FM-Wild Workshop*, 2025. URL https://arxiv.org/abs/2505.22637.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Z. Wahrscheinlichkeitstheor. Verwandte Geb.*, 47(2):119–137, 1979. URL https://doi.org/10.1007/BF00535278.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. In *Proc. NeurIPS*, volume 37, pp. 49519–49551, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/58cbe393b4254da8966780a40d023c0b-Abstract-Conference.html.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. arXiv preprint arXiv:2507.21509, 2025. URL https://arxiv.org/abs/2507.21509.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, 16(1):41–46, 1970. URL https://doi.org/10.1109/TIT.1970.1054406.

Herbert H. Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *J. Verbal Learn. Verbal Behav.*, 12(4):335–359, 1973. URL https://doi.org/10.1016/S0022-5371(73)80014-3.

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, 3 edition, 1977. ISBN 9780471162407. URL https://www.wiley-vch.de/en/areas-interest/mathematics-statistics/sampling-techniques-978-0-471-16240-7.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. Steering off course: Reliability challenges in steering language models. In *Proc. ACL*, pp. 19856–19882. Association for Computational Linguistics, 2025. URL https://doi.org/10.18653/v1/2025.acl-long.974.

Gilles R. Ducharme, Myoungshic Jhun, Joseph P. Romano, and Kinh N. Truong. Bootstrap confidence cones for directional data. *Biometrika*, 72(3):637–645, 1985. URL https://doi.org/10.1093/biomet/72.3.637.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *Ann. Stat.*, 7(1):1–26, 1979. URL https://doi.org/10.1214/aos/1176344552.

R. A. Fisher. Dispersion on a sphere. *Proc. R. Soc. Lond. A Math. Phys. Sci.*, 217(1130):295–305, 1953. URL https://doi.org/10.1098/rspa.1953.0064.

Zhiwei Gao, Shaowen Peng, Shoko Wakamiya, and Eiji Aramaki. What does activation steering control? attribution across answer encodings and output-sensitive subspaces. arXiv preprint arXiv:2608.22985, 2026. URL https://arxiv.org/abs/2608.22985.

Alexander Giessing and Jianqing Fan. Bootstrapping ℓ_p-statistics in high dimensions. arXiv preprint arXiv:2006.13099, 2020. URL https://doi.org/10.48550/arXiv.2006.13099.

Navita Goyal and Hal Daumé III. Steering safely or off a cliff? rethinking specificity and robustness in inference-time interventions. In *Proc. EACL*, pp. 5723–5738. Association for Computational Linguistics, 2026. URL https://doi.org/10.18653/v1/2026.eacl-long.268.

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 *Proc. ICML*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22924–22945. PMLR, 2025. URL https://proceedings.mlr.press/v267/hedstrom25a.html.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.*, 19 (3):293–325, 1948. URL https://doi.org/10.1214/aoms/1177730196.

Thomas Hotz and Florian Kelma. Non-asymptotic confidence sets for extrinsic means on spheres and projective spaces. arXiv preprint arXiv:1602.04117, 2016. URL https://doi.org/10.48550/arXiv.1602.04117.

Stuart H. Hurlbert. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.*, 54(2):187–211, 1984. URL https://doi.org/10.2307/1942661.

Shawn Im and Sharon Li. A unified understanding and evaluation of steering methods. arXiv preprint arXiv:2502.02716v2, 2026. URL https://arxiv.org/abs/2502.02716v2.

Yuri I. Ingster and Irina A. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169 of *Lecture Notes in Statistics*. Springer, 2003. URL https://doi.org/10.1007/978-0-387-21580-8.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. Improving activation steering in language models with mean-centring. arXiv preprint arXiv:2312.03813, 2023. URL https://arxiv.org/abs/2312.03813.

Etaash Katiyar and Qingyuan Zhao. On testing mean proportionality of multivariate normal variables. arXiv preprint arXiv:2103.05574, 2021. URL https://arxiv.org/abs/2103.05574.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proc. ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Siegfried Kropf, Kai Antweiler, and Ekkehard Glimm. Use of multivariate distance measures for high-dimensional data in tests for difference, superiority, equivalence and non-inferiority. *Biom. J.*, 64(3):577–597, 2022. URL https://doi.org/10.1002/bimj.202000367.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4): 963–974, 1982. URL https://doi.org/10.2307/2529876.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Proc. NeurIPS*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Yuxiao Li, Alina Fastowski, Efstratios Zaradoukas, Bardh Prenkaj, and Gjergji Kasneci. Analysing the safety pitfalls of steering vectors. arXiv preprint arXiv:2603.24543, 2026. URL https://arxiv.org/abs/2603.24543.

Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *Ann. Stat.*, 47(2):783–794, 2019. URL https://doi.org/10.1214/17-AOS1639.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Proc. COLM*, 2024. URL https://arxiv.org/abs/2310.06824.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989. URL https://doi.org/10.1017/CBO9781107359949.008.

Maxime Méloux, Giada Dirupo, François Portet, and Maxime Peyrard. The dead salmons of AI interpretability. arXiv preprint arXiv:2512.18792, 2025. URL https://arxiv.org/abs/2512.18792.

Maxime Méloux, François Portet, and Maxime Peyrard. Mechanistic interpretability as statistical estimation: A variance analysis. arXiv preprint arXiv:2510.00845, 2026. URL https://arxiv.org/abs/2510.00845v4.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. arXiv preprint arXiv:2411.00640, 2024. URL https://arxiv.org/abs/2411.00640.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Trans. Assoc. Comput. Linguist.*, 12:933–949, 2024. URL https://doi.org/10.1162/tacl_a_00681.

Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. Lond. A Math. Phys. Sci.*, 236(767):333–380, 1937. URL https://doi.org/10.1098/rsta.1937.0005.

Angus Nicolson, Lisa Schut, J. Alison Noble, and Yarin Gal. Explaining explainability: Recommendations for effective use of concept activation vectors. *Trans. Mach. Learn. Res.*, 2025. URL https://arxiv.org/abs/2404.03713.

Simon Ostermann, Daniil Gurgurov, Tanja Baeumel, Michael A. Hedderich, Sebastian Lapuschkin, Wojciech Samek, and Vera Schmitt. From weights to activations: Is steering the next frontier of adaptation? In *Proc. ACL*, pp. 29854–29879. Association for Computational Linguistics, 2026. URL https://doi.org/10.18653/v1/2026.acl-long.1377.

Frederik Pahde, Maximilian Dreyer, Moritz Weckbecker, Leander Weber, Christopher J. Anders, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. Navigating neural space: Revisiting concept activation vectors to overcome directional divergence. In *Proc. ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/852870d56e20a22fcd52ea14523fb820-Abstract-Conference.html.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proc. ICML*, 2024. URL https://arxiv.org/abs/2311.03658.

Andrej Pázman and Luc Pronzato. Optimum design accounting for the global nonlinear behavior of the model. *Ann. Stat.*, 42(4):1426–1451, 2014. URL https://doi.org/10.1214/14-AOS1232.

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, Noemi 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 Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In *Proc. ACL Findings*, pp. 13387–13434. Association for Computational Linguistics, 2023. URL https://doi.org/10.18653/v1/2023.findings-acl.847.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Proc. NAACL Findings*, pp. 2006–2017. Association for Computational Linguistics, 2024. URL https://doi.org/10.18653/v1/2024.findings-naacl.130.

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. URL https://arxiv.org/abs/2410.17245.

Prashant C. Raju. The geometric canary: Predicting steerability and detecting drift via representational stability. arXiv preprint arXiv:2604.17698, 2026. URL https://arxiv.org/abs/2604.17698.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proc. ACL*, pp. 15504–15522. Association for Computational Linguistics, 2024. URL https://doi.org/10.18653/v1/2024.acl-long.828.

Henry Scheffé. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1–2): 87–110, 1953. URL https://doi.org/10.1093/biomet/40.1-2.87.

Ekkehard Schnoor, Jawher Said, Malik Tiomoko, Wojciech Samek, and Alexander Jung. α-TCAV: A unified framework for testing with concept activation vectors. arXiv preprint arXiv:2605.15688, 2026a. URL https://doi.org/10.48550/arXiv.2605.15688.

Ekkehard Schnoor, Malik Tiomoko, Jawher Said, Alex Jung, and Wojciech Samek. Concept activation vectors: A unifying view and adversarial attacks. In *Proc. ICASSP*, 2026b. URL https://arxiv.org/abs/2509.22755.

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 *Proc. ICLR*, 2024. URL https://arxiv.org/abs/2310.11324.

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 *Proc. NeurIPS*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fb3ad59a84799bfb8d700e56d19c231b-Abstract-Conference.html.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. arXiv preprint arXiv:2308.10248, 2024. URL https://arxiv.org/abs/2308.10248.

Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. arXiv preprint arXiv:2602.06801, 2026. URL https://arxiv.org/abs/2602.06801v4.

Julia Wenkmann and Damien Garreau. On the variability of concept activation vectors. In *Proc. ICML*, 2026. URL https://doi.org/10.48550/arXiv.2509.24058.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *Proc. ICML*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66945–66970. PMLR, 2025. URL https://proceedings.mlr.press/v267/wollschlager25a.html.

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 *Proc. ICML*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080. PMLR, 2025. URL https://proceedings.mlr.press/v267/wu25a.html.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997. URL https://doi.org/10.1007/978-1-4612-1880-7_29.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *Proc. ICLR*, 2024. URL https://arxiv.org/abs/2309.03882.

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

A Extended related work

Estimation and uncertainty for concept directions. CAVs have already been studied as statistical estimators: Wenkmann & Garreau (2026) analyze variability from random reference examples and its effects on sensitivity and thresholded scores; Schnoor et al. (2026b) characterize CAV means, covariances, and dependence on the non-concept distribution. With both classes random, Schnoor et al. (2026a) derive the mean-contrast covariance $\Sigma_1/n_1 + \Sigma_2/n_2$ and project uncertainty onto a fixed gradient. Related work analyzes circuit-estimation variance and advocates explicit statistical models for interpretability (Méloux et al., 2026; 2025); Certified Circuits instead studies stability under bounded dataset edits (Anani et al., 2026). Our confidence-set approach asks which signal, line, and orientation claims hold across compatible means at a specified resolution. Valid coverage controls false claims, whereas the sample-size bounds also require directional accuracy, counting abstention as failure. The public CAA audit's approximate bootstrap has no established joint population coverage; data-dependent selection also requires care (Berk et al., 2013), since unconditional error control need not hold conditional on a non-abstaining verdict.

Representation geometry and construction. TCAV and pattern-based CAVs connect directions to explanations (Kim et al., 2018; Pahde et al., 2025); truth-representation studies, the linear representation hypothesis, and concept erasure connect geometry to probing and intervention (Marks & Tegmark, 2024; Park et al., 2024; Belrose et al., 2023). Layer choice, entanglement, and the negative probe set affect a CAV's meaning (Nicolson et al., 2025). Mean-centred steering subtracts a background activation mean from a target mean (Jorgensen et al., 2023), whereas BiPO optimizes a vector through preferences over generation probabilities (Cao et al., 2024). The choice of background defines the contrast. Subtracting the same centre from both class means leaves their difference unchanged. Our results concern the mean contrast defined by a fixed model, layer, token position, and sampling recipe, without requiring an entire concept to be one-dimensional.

Behavioral identifiability. Different directions can have similar behavioral effects, as studied through orthogonal perturbations, trait scores, and logits by Venkatesh & Kurapath (2026). The refusal directions and concept cones of Wollschläger et al. (2025) likewise describe behavioral structure, with orthogonality distinct from independence under intervention. Our compatible directions describe uncertainty about a fixed population contrast, so estimating it precisely need not identify a unique behavioral mechanism. The matched-angle diagnostic (Appendix F.4) compares selected perturbations at nonzero intervention strength, without assuming that all orthogonal errors affect behavior.

Steering reliability and geometric predictors. Steering studies document input and prompt sensitivity (Tan et al., 2024), model-dependent effectiveness (Da Silva et al., 2025), specificity failures under shift (Goyal & Daumé III, 2026), and changes in jailbreak susceptibility (Li et al., 2026). Predictors include agreement among activation differences (Braun et al., 2025), activation-space separation (Bhattarai & Alhanai, 2026), unembedding-based layer diagnostics (Billa, 2026), and task-aligned representational stability (Raju, 2026). Different prompt variants can yield directionally different but similarly effective vectors (Braun, 2026); in synthetic experiments, separation tracks empirical–true-feature alignment (Bhattarai & Alhanai, 2026). Our directional error concerns a recipe's population contrast, which need not be a latent feature direction. Response repeatability across reconstructions is therefore evaluated separately from mean effects, specificity, and safety.

Behavioral evaluation and abstention. Steering evaluations cover multiple-choice and open-ended tasks (Im & Li, 2026) and compare prompting and fine-tuning baselines (Wu et al., 2025); a complementary taxonomy situates steering among adaptation methods (Ostermann et al., 2026). Pres et al. (2024) emphasize task-relevant contexts, likelihoods, cross-behavior comparisons, and baselines. Our response variance and response MSE measure matching-answer probability repeatability and deviation from a reference response, rather than free-generation quality or method rankings. MERA calibrates whether and how strongly to intervene (Hedström et al., 2025), and classical rejection trades prediction error against rejection (Chow, 1970); SteerCert's ABSTAIN verdict withholds a population claim.

Answer encoding, prompts, and sampling. Option-identifier biases (Zheng et al., 2024) and answer-order sensitivity (Pezeshkpour & Hruschka, 2024) concern answer selection. Gao et al. (2026) examine steering attribution across encodings and average directions from six semantic mappings, rescaling to the canonical CAA norm. Their mapping-balanced construction generally targets a different contrast from a fixed mapping. We study how a construction sample's A/B composition changes its mean contrast, without identifying this variation with selection bias. Standard proportional stratification (Cochran, 1977), weighted by original frame proportions, preserves our finite-frame target while allowing comparison with ordinary sampling at equal sample size. This comparison does not evaluate the relative behavioral effects of different targets. Formatting sensitivity (Sclar et al., 2024) and instruction-dependent performance and rankings (Mizrahi et al., 2024) motivate accounting for template variation; they do not establish our random-effects assumptions. Miller (2024) discuss clustered and paired evaluation uncertainty. Our response bootstrap preserves question identities and paired conditions while resampling constructions; its intervals are pointwise and conditional on the selected frames and settings.

Confidence sets for directional means. Classical directional inference (Fisher, 1953) includes consistent nonparametric bootstrap confidence cones for sphere-valued observations (Ducharme et al., 1985). Hotz & Kelma (2016, Corollary 3 and Prop. 6) project Euclidean confidence balls onto spheres, obtaining an arcsine angular radius or the whole sphere when zero is interior; they also treat projective spaces. Thm. 6 uses this geometry, but estimates the direction of a Euclidean mean with variable-norm observations. Normalizing observations before averaging generally changes that target.

Independent data requirements and testing. Thm. 1 combines a sample-mean second-moment bound (Lugosi & Mendelson, 2019) with an Assouad lower bound (Yu, 1997). Under the random-effects model (Laird & Ware, 1982), within-template averaging gives covariance $\Sigma_b + \Gamma/m$. Thm. 2 combines nonlinear least-squares stability (Pázman & Pronzato, 2014) and a Gaussian lower bound to obtain matching template-count scales through $\tau_F^{(m)}(u_0)$, at fixed confidence under the stated moment and locality conditions, with magnitude, base direction, and covariances supplied. Classical minimax testing (Ingster & Suslina, 2003; Blanchard et al., 2018) provides context for Prop. 5's covariance-bounded detection–direction comparison. Gaussian mean-proportionality tests (Katiyar & Zhao, 2021) and multivariate equivalence tests (Kropf et al., 2022) address neighboring questions: failure to reject equality or proportionality does not localize all compatible directions up to sign, as an unoriented-line verdict requires.

B Independent observations: Theorem 1

This appendix gives the shared Gaussian lower-bound ingredients, the proof of Thm. 1, and its scope and detection–direction implications. Appendix C reuses the Gaussian lower bound for template sampling.

B.1 Auxiliary results

Directional notation. For unit vectors v, w, write $\angle_{\text{line}}(v, w) := \min\{\angle(v, w), \angle(v, -w)\}$. For a unit vector w, let

$$\mathcal{C}_\varepsilon(w) := \{v \in \mathbb{S}^{d-1} : \angle(v, w) \le \varepsilon\}, \qquad \mathcal{L}_\varepsilon(w) := \mathcal{C}_\varepsilon(w) \cup \mathcal{C}_\varepsilon(-w).$$

B.1.1 Local Gaussian tangent estimation

We use the classical hypercube reduction for minimax lower bounds reviewed by Yu (1997), with weights chosen for the tangent covariance.

Lemma 3 (Local Gaussian tangent estimation lower bound)**.** *Consider $X_t \overset{\text{iid}}{\sim} N(au, \Sigma)$, $t = 1, \dots, T$, with $d \ge 2$, $T \ge 1$, fixed $a > 0$, known $\Sigma \succ 0$, and $u \in \mathbb{S}^{d-1}$. Fix $u_0 \in \mathbb{S}^{d-1}$, let B have orthonormal columns spanning $u_0^\perp$, and put*

$$\tau_F(u_0) := \operatorname{tr}\left[(B^\top \Sigma^{-1} B)^{-1}\right].$$

There are universal constants $c, K_0, r_0, \delta_0 > 0$ such that, for every $0 < \varepsilon \le r_0$, there is a finite set $\mathcal{V}_\varepsilon \subseteq \{u : \angle(u, u_0) \le K_0\varepsilon\}$ for which

$$a^2 \le c\,\frac{\tau_F(u_0)}{T\varepsilon^2} \quad \Longrightarrow \quad \inf_{\widetilde{u}} \sup_{u \in \mathcal{V}_\varepsilon} \mathbb{P}_{au}\{\angle(\widetilde{u}, u) > \varepsilon\} \ge \delta_0,$$

where the infimum is over possibly randomized $\mathbb{S}^{d-1}$-valued estimators.

Proof of Lemma 3. The sample mean $\bar{X} \sim N(au, \Sigma/T)$ is sufficient in the Gaussian location model. Use the base direction u_0 and orthonormal-column matrix B from the lemma statement. Set

$$I_\perp := B^\top \Sigma^{-1} B, \qquad C_\perp := I_\perp^{-1}.$$

Let ν_j and e_j be the eigenvalues and eigenvectors of $C_\perp$, and let $\tau_F := \sum_j \nu_j$. For sign vectors $\theta \in \{\pm 1\}^{d-1}$, define

$$h_\theta := r \sum_{j=1}^{d-1} \sqrt{\frac{\nu_j}{\tau_F}}\,\theta_j e_j, \qquad u_\theta := \sqrt{1 - r^2}\,u_0 + B h_\theta.$$

Neighboring vertices that differ in coordinate j have KL divergence

$$D_{\mathrm{KL}}(P_\theta \| P_{\theta^{(j)}}) = \frac{T}{2}\left(a(u_\theta - u_{\theta^{(j)}})\right)^\top \Sigma^{-1}\left(a(u_\theta - u_{\theta^{(j)}})\right) = 2\frac{Ta^2 r^2}{\tau_F}.$$

If $a^2 \le c\tau_F/(Tr^2)$, all neighboring KL divergences are bounded by a small constant. For any direction estimator $\tilde{u}$, decode signs by

$$\widehat{\theta}_j := \mathrm{sign}(e_j^\top B^\top \tilde{u}),$$

with arbitrary tie-breaking. If $\widehat{\theta}_j \ne \theta_j$, then the jth tangent coordinate differs from the truth by at least $r\sqrt{\nu_j/\tau_F}$. Hence

$$\angle^2(\tilde{u}, u_\theta) \ge \|\tilde{u} - u_\theta\|_2^2 \ge r^2 \sum_{j=1}^{d-1} \frac{\nu_j}{\tau_F} \mathbf{1}\{\widehat{\theta}_j \ne \theta_j\},$$

Pair neighboring vertices along each coordinate and average over the hypercube. The Bretagnolle–Huber inequality (Bretagnolle & Huber, 1979) bounds the pairwise testing error below by a positive constant because the neighboring KL divergences are bounded. Summing with weights $w_j := \nu_j/\tau_F$, whose sum is one, gives the weighted Assouad bound (Yu, 1997) and hence

$$\sup_\theta \mathbb{E}_\theta\left[\angle^2(\tilde{u}, u_\theta)\right] \ge c_1 r^2.$$

The constant-probability statement follows by a bounded-loss argument. Let $\tilde{u}_{\mathrm{cl}}$ be a closest point of the parameter class $\{u_\theta\}$ to $\tilde{u}$; by the triangle inequality $\angle(\tilde{u}_{\mathrm{cl}}, u_\theta) \le 2\angle(\tilde{u}, u_\theta)$, and since every class member lies within angle $\arcsin r \le 2r$ of u_0, the loss of $\tilde{u}_{\mathrm{cl}}$ is bounded: $\angle^2(\tilde{u}_{\mathrm{cl}}, u_\theta) \le 16r^2$. The in-expectation bound applies to $\tilde{u}_{\mathrm{cl}}$ as well. For the corresponding vertex, put $Z := \angle^2(\tilde{u}_{\mathrm{cl}}, u_\theta)/r^2$ and $p := \mathbb{P}(Z \ge c_1/2)$. Since $0 \le Z \le 16$ and $\mathbb{E}Z \ge c_1$, $c_1 \le 16p + (1-p)c_1/2$, which gives

$$\sup_\theta \mathbb{P}_\theta\left(\angle(\tilde{u}_{\mathrm{cl}}, u_\theta) \ge \sqrt{c_1/2}\,r\right) \ge \delta_0 := \frac{c_1}{32 - c_1},$$

and $\angle(\tilde{u}, u_\theta) \ge \angle(\tilde{u}_{\mathrm{cl}}, u_\theta)/2$ transfers the bound to $\tilde{u}$. Take $r := C'\varepsilon$, where $C' := 4/\sqrt{c_1/2}$, and absorb C'^2 into the lower-bound constant. For sufficiently small ε, $\angle(u_\theta, u_0) = \arcsin(C'\varepsilon) \le 2C'\varepsilon$; thus $K_0 := 2C'$ and the displayed class prove the lemma. □

B.1.2 Selective Gaussian line lower bound

Thm. 1 counts abstention as failure. For procedures that may abstain, the selective lower bound instead requires both control of false claims and a sufficiently high probability of reporting a direction up to sign at the requested resolution, uniformly over the same local neighborhood. A procedure that always abstains makes no false claims but supplies no direction estimate, so false-claim control alone cannot imply the bound.

The following theorem states the selective lower bound for a general known Gaussian covariance.

Theorem 4 (Selective local Gaussian line lower bound)**.** *Consider the Gaussian location experiment*

$$X_t = \Delta + \omega_t, \qquad \omega_t \overset{\text{iid}}{\sim} N(0, \Sigma), \qquad t = 1, \dots, T,$$

with $d \geq 2$, $T \geq 1$, $\Delta = au$, fixed $a > 0$, and known $\Sigma \succ 0$. Fix $u_0 \in \mathbb{S}^{d-1}$, let $B_0 \in \mathbb{R}^{d\times(d-1)}$ have orthonormal columns spanning $u_0^{\perp}$, and define

$$\tau_F(u_0) := \operatorname{tr}[(B_0^{\top}\Sigma^{-1}B_0)^{-1}]. \tag{6}$$

There are universal constants $c, K, r_0, \delta_0 > 0$ such that the following holds for $0 < r \leq r_0$. Let $\mathcal{U}_r := \{u \in \mathbb{S}^{d-1} : \angle(u, u_0) \leq Kr\}$. A possibly randomized procedure either abstains or returns an axis $\widehat{w} \in \mathbb{S}^{d-1}$ and sine radius $\widehat{\rho} \in [0, 1]$. Suppose, for some $\alpha, \beta \in [0, 1]$ with $\alpha + \beta < \delta_0$, it satisfies both

$$\sup_{u\in\mathcal{U}_r} \mathbb{P}_{au}\{\textit{output and } u \notin \mathcal{L}_{\arcsin\widehat{\rho}}(\widehat{w})\} \leq \alpha$$

and

$$\inf_{u\in\mathcal{U}_r} \mathbb{P}_{au}\{\textit{output and } \widehat{\rho} \leq r\} \geq 1 - \beta.$$

The first condition bounds the probability of a false claim by α. The second requires a report at the requested resolution with probability at least $1 - \beta$, uniformly over the same neighborhood. Then necessarily

$$a^2 \geq c\,\frac{\tau_F(u_0)}{Tr^2}.$$

For isotropic $\Sigma = \sigma^2 I_d$, this is $a^2 \geq c\sigma^2(d-1)/(Tr^2)$. The conclusion is local, selective, and explicit about its quantifiers; it makes no claim about a procedure's radius on samples where that procedure abstains.

Proof of the radius lower-bound claim in Theorem 4. Suppose such a procedure exists with $\alpha+\beta < \delta_0$, with δ_0 from Lemma 3. Let $\varepsilon := \arcsin r$, and take the Assouad class $\mathcal{V} := \mathcal{V}_\varepsilon$ from that lemma. For a sufficiently large universal K and sufficiently small r_0, $\mathcal{V} \subseteq \mathcal{U}_r$. Define $\tilde{u} := u_0$ on abstention. On output, define $\tilde{u} := \hat{w} \cdot \operatorname{sign}(\langle \hat{w}, u_0\rangle)$, the reported axis sign-corrected toward the base point, with a zero inner product assigned sign $+1$. On the joint event that the procedure outputs, covers u, and $\widehat{\rho} \leq r$, which has probability at least $1 - \alpha - \beta$ under every $u \in \mathcal{V}$, we have $\angle_{\text{line}}(\hat{w}, u) \leq \arcsin\widehat{\rho} \leq \arcsin r = \varepsilon$. Sign correction turns the line bound into an oriented bound: either $\angle(\tilde{u}, u) \leq \varepsilon$, or $\angle(\tilde{u}, u) \geq \pi - \varepsilon$; the latter is impossible because $\angle(\tilde{u}, u_0) \leq \pi/2$ by construction while $\angle(\tilde{u}, u) \geq \pi - \varepsilon$ would force $\angle(\tilde{u}, u_0) \geq \pi - \varepsilon - \angle(u, u_0) \geq \pi - (1 + K_0)\varepsilon > \pi/2$. Hence $\sup_{u\in\mathcal{V}} \mathbb{P}_u(\angle(\tilde{u}, u) > \varepsilon) \leq \alpha + \beta < \delta_0$, contradicting Lemma 3 whenever $a^2 \leq c\,\tau_F(u_0)/(T\varepsilon^2)$. Since $r \leq \varepsilon \leq 2r$ for small r, this is the stated condition after absorbing constants. □

B.2 Proof of Theorem 1

The upper bound converts the standard second-moment bound for the sample mean (Lugosi & Mendelson, 2019, Sec. 1) to angular precision. The lower bound specializes the local Gaussian construction proved in Appendix B.1 to covariance $\sigma^2 I_d$.

Proof. For necessity, take the neighborhood constant K, radius cutoff r_0, and error-level constant $\bar{\delta}_0$ from Thm. 4, enlarging K to at least 1 and shrinking r_0 to at most $1/2$. Choose the common constants $\varepsilon_{\max} := \min\{r_0, \pi/(4K)\}$ and $\delta_0 := \bar{\delta}_0/2$ for both main theorems. Restrict the covariance-bounded class to $N(au, \sigma^2 I_d)$, for which $\tau_F(u_0) = \sigma^2(d-1)$. Regard each reported direction as an axis and attach the fixed sine radius $r := \sin\varepsilon$. Since line error is no greater than the angle to the population direction, on $\{u : \angle(u, u_0) \leq Kr\}$, which is contained in $\mathcal{U}_\varepsilon$, accuracy with probability at least $1 - \delta$, counting abstention as failure, implies both

$$\mathbb{P}_{au}\{\text{a report with line error} > \varepsilon\} \leq \delta, \qquad \mathbb{P}_{au}\{\text{a report at sine radius } r\} \geq 1 - \delta.$$

The selective lower bound applies with $\alpha = \beta = \delta$ and 2δ below its error-level constant. It gives $a^2 \geq c\sigma^2(d-1)/(m\sin^2\varepsilon)$. Taking the minimum over procedures and using $m_\star \geq 1$ proves the lower bound in (3). The alternatives share a, u_0, σ, so supplying these quantities does not remove the lower bound.

For sufficiency, take $m := \lceil \sigma^2 d/(\delta a^2 \sin^2 \varepsilon) \rceil$ and write $\bar{W} := m^{-1} \sum_{i=1}^{m} W_i$, $\Sigma_P := \mathrm{Cov}_P(W)$, and $e := \bar{W} - au$. Report $\widehat{w} := \bar{W}/\|\bar{W}\|_2$ when $\bar{W} \neq 0$, and any fixed unit vector otherwise. Let $\theta := \arccos(u^\top \widehat{w})$. Independence and the covariance bound imply

$$\mathbb{E}_P \|e\|_2^2 = \frac{\mathrm{tr}\, \Sigma_P}{m} \leq \frac{\sigma^2 d}{m}.$$

On the event $\|e\|_2 \leq a \sin \varepsilon$, we have $u^\top \bar{W} \geq a - \|e\|_2 > 0$, so $\bar{W} \neq 0$ and $\theta < \pi/2$. Orthogonal projection gives

$$a \sin \theta = \left\| (I - \widehat{w}\widehat{w}^\top) a u \right\|_2 = \left\| (I - \widehat{w}\widehat{w}^\top) e \right\|_2 \leq \|e\|_2 \leq a \sin \varepsilon.$$

Thus $\theta \leq \varepsilon$. A zero sample mean has $\|e\|_2 = a > a \sin \varepsilon$ and is included in the failure event used below. Markov's inequality yields, uniformly over $P \in \mathcal{P}_\varepsilon$,

$$\mathbb{P}_P\{\theta > \varepsilon\} \leq \mathbb{P}_P\{\|e\|_2 > a \sin \varepsilon\} \leq \frac{\sigma^2 d}{m a^2 \sin^2 \varepsilon} \leq \delta$$

by the choice of m. This proves the upper bound on $m_\star$ in (3). The upper bound holds for every direction on the sphere, and hence on the specified local neighborhood. The matching-order interpretation fixes $\delta > 0$; the upper bound does not assert optimal confidence dependence. □

B.3 Scope and implications

Scope of the paired bounds. Although the local population setting in Sec. 4 supplies a, u_0, σ, the normalized sample mean uses only the observations, so both bounds in Thm. 1 remain valid without this information.

Selective reporting with a trace bound. For m i.i.d. units with a declared $V > 0$ satisfying $\mathrm{tr}\, \mathrm{Cov}(W) \leq V$, independence and Markov's inequality give

$$R_\gamma := \sqrt{\frac{V}{m\gamma}}, \qquad \mathbb{P}\{\|\bar{W} - \Delta\|_2 \leq R_\gamma\} \geq 1 - \gamma, \quad \gamma \in (0,1).$$

The confidence-ball decision rule of Sec. 3, applied to the ball centered at $\bar{W}$ with radius R_α, controls false claims by α, including when $\Delta = 0$. For $a := \|\Delta\|_2 > 0$, $\alpha, \beta \in (0,1)$, and $0 < \varepsilon < \pi/2$, the condition

$$m \geq \frac{V}{a^2} \left(\frac{1}{\sqrt{\beta}} + \frac{1}{\sqrt{\alpha} \sin \varepsilon} \right)^2 \quad \Longleftrightarrow \quad a \geq R_\beta + \frac{R_\alpha}{\sin \varepsilon}$$

ensures $\|\bar{W}\|_2 \geq a - R_\beta \geq R_\alpha / \sin \varepsilon > R_\alpha$ on an event of probability at least $1 - \beta$. The rule therefore reports an ORIENTED DIRECTION verdict at resolution ε with at least that probability. For $\mathrm{Cov}(W) \preceq \sigma^2 I_d$, take $V = \sigma^2 d$; the sufficient sample-size order at fixed positive α, β is $\sigma^2 d/(a^2 \sin^2 \varepsilon)$.

B.3.1 Detection is not direction

Under a declared covariance bound, a test based on cross-pair inner products can detect a nonzero contrast before the sample resolves its direction up to sign at the requested angular resolution.

Proposition 5 (Detection does not establish a direction)**.** *Let $X_1, \ldots, X_T \overset{\text{iid}}{\sim} P$, where $T \geq 2$, $\Delta := \mathbb{E}_P X$, and $\mathrm{Cov}_P(X) \preceq \sigma^2 I_d$ for a declared $\sigma > 0$. Consider a sequence with $d \to \infty$, allowing $T, \sigma, a > 0$ to vary. There are tests of $H_0 : \Delta = 0$ against $\|\Delta\|_2 = a$ whose type-I and type-II error probabilities tend to zero uniformly over the respective null and alternative distributions satisfying the covariance bound when*

$$\frac{T a^2}{\sigma^2 \sqrt{d}} \longrightarrow \infty.$$

In contrast, fix $\alpha, \beta \in (0,1)$ with $\alpha + \beta < \delta_0$ for the constant in Thm. 4, and a sufficiently small angular resolution ε. Any line procedure satisfying that theorem's local false-claim and reporting

conditions with $r = \sin\varepsilon$, uniformly over all distributions with means au in the stated neighborhood and covariance at most $\sigma^2 I_d$, must have

$$a^2 \geq c\,\frac{\sigma^2(d-1)}{T\sin^2\varepsilon} \asymp \frac{\sigma^2 d}{T\varepsilon^2}.$$

A confidence ball calibrated from the covariance bound attains this directional order at fixed positive α, β (Appendix B.3). Hence a regime $\sqrt{d} \ll Ta^2/\sigma^2 \ll d/\varepsilon^2$ supports reliable signal detection but not a procedure that both controls false line claims locally and reports at that resolution with the required probability uniformly over the covariance-bounded class.

Proof. For $T \geq 2$, use the cross-pair energy statistic

$$\widehat{A} := \frac{1}{T(T-1)}\sum_{t\neq s} X_t^\top X_s.$$

Write $\xi_t := X_t - \Delta$ and $\Sigma_P := \mathrm{Cov}_P(X)$. The Hoeffding decomposition (Hoeffding, 1948) is

$$\widehat{A} - \|\Delta\|^2 = \frac{2}{T}\sum_t \Delta^\top \xi_t + \frac{1}{T(T-1)}\sum_{t\neq s}\xi_t^\top \xi_s.$$

The two terms are uncorrelated. Independence and finite second moments give $\mathbb{E}(\xi_t^\top \xi_s)^2 = \mathrm{tr}(\Sigma_P^2)$ for $t \neq s$; cross terms involving an index appearing once vanish by conditioning. Consequently,

$$\mathbb{E}_P\widehat{A} = \|\Delta\|_2^2, \qquad \mathrm{Var}_P(\widehat{A}) = \frac{4}{T}\Delta^\top\Sigma_P\Delta + \frac{2}{T(T-1)}\,\mathrm{tr}(\Sigma_P^2).$$

Since $\Sigma_P \preceq \sigma^2 I_d$, the test $\varphi := \mathbf{1}\{\widehat{A} > a^2/2\}$ satisfies

$$\sup_{P:\Delta=0}\mathbb{P}_P(\varphi=1) \leq \frac{8\sigma^4 d}{T(T-1)a^4}, \qquad \sup_{P:\|\Delta\|_2=a}\mathbb{P}_P(\varphi=0) \leq \frac{8\sigma^4 d}{T(T-1)a^4} + \frac{16\sigma^2}{Ta^2},$$

where both suprema are over laws satisfying the covariance bound. These are Chebyshev bounds, uniform over the two distribution classes; the test is specified for the separation magnitude a in the proposition. With $\lambda := Ta^2/\sigma^2$ and $T/(T-1) \leq 2$, the bounds are at most $16d/\lambda^2$ and $16d/\lambda^2 + 16/\lambda$, respectively. Both vanish when $\lambda/\sqrt{d} \to \infty$.

For directional necessity, restrict to $X_t \overset{\text{iid}}{\sim} N(\Delta, \sigma^2 I_d)$ within the covariance-bounded class and apply Thm. 4 with $r = \sin\varepsilon$. In the isotropic case $\tau_F(u_0) = \sigma^2(d-1)$, giving $a^2 \gtrsim \sigma^2 d/(T\varepsilon^2)$ for small ε. For attainment, Appendix B.3 with $m = T$ constructs a confidence ball with false-claim probability at most α and reporting probability at least $1-\beta$ at the requested resolution over the full covariance-bounded class, at the stated sample-size order. □

C Independent templates: Theorem 2

This appendix gives the auxiliary results and proof for Thm. 2, followed by the role of its assumptions. The independent-template requirement reflects the distinction between repeated observations and independent units in random-effects models and pseudoreplication (Laird & Ware, 1982; Hurlbert, 1984). The same generalization issue arises when linguistic materials are treated as fixed (Clark, 1973; Ball et al., 2024).

C.1 Auxiliary results

C.1.1 Template means and reporting requirements

Averaging the additive model gives $\bar{z}_t = \Delta + \delta_t + m^{-1}\sum_i \varepsilon_{it}$, with covariance $\Sigma_m = \Sigma_b + \Gamma/m$ by independence. With dependent within-template errors, the second term instead equals $m^{-2}\sum_{i,j}\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{jt})$; arbitrary dependence need not give $1/m$ decay.

For each pair $(Q_b, Q_w) \in \mathcal{F}(\Sigma_b, \Gamma)$ defined in Sec. 4.2, write $\mathbb{P}_{u,Q_b,Q_w}$ for the law of the full dataset with $\Delta = au$, including any independent randomization used by the procedure. The covariances,

magnitude, and design parameters are fixed across this class. Take K, r_0, δ_0 from Thm. 4, and fix $0 < r \le r_0$ and $\alpha, \beta \in [0,1]$ with $\alpha + \beta < \delta_0$. Let $\mathcal{U}_r := \{u \in \mathbb{S}^{d-1} : \angle(u, u_0) \le Kr\}$ and let A denote the event that the procedure reports a unit axis $\widehat{w} \in \mathbb{S}^{d-1}$ and a sine radius $\widehat{\rho} \in [0,1]$. The selective reporting requirements are

$$\sup_{\substack{u \in \mathcal{U}_r \\ (Q_b, Q_w) \in \mathcal{F}(\Sigma_b, \Gamma)}} \mathbb{P}_{u, Q_b, Q_w}\{A,\ \arccos|u^\top \widehat{w}| > \arcsin \widehat{\rho}\} \le \alpha, \tag{7}$$

$$\inf_{\substack{u \in \mathcal{U}_r \\ (Q_b, Q_w) \in \mathcal{F}(\Sigma_b, \Gamma)}} \mathbb{P}_{u, Q_b, Q_w}\{A,\ \widehat{\rho} \le r\} \ge 1 - \beta. \tag{8}$$

The same procedure and fixed error levels must satisfy both requirements over the entire class and neighborhood. Independence concerns the effect and error vectors; independent coordinates are not required.

C.1.2 Gaussian reduction and lower bound at finite m

Using Σ_m and $\tau_F^{(m)}$ from Thm. 2, the selective reporting requirements (7)–(8) imply

$$a^2 \ge c\,\frac{\tau_F^{(m)}(u_0)}{Tr^2} \ge c\,\frac{\tau_F^{(b)}(u_0)}{Tr^2}. \tag{9}$$

Proof. Restrict to $Q_b = N(0, \Sigma_b)$ and $Q_w = N(0, \Gamma)$ in the moment class, including singular Γ, and reveal the covariances. The T independent template means have covariance Σ_m. Within-template residuals $z_{it} - \bar{z}_t$ are jointly Gaussian and uncorrelated with these means, hence independent of them; their known law does not involve Δ. A randomized procedure using only the means can therefore simulate the residuals and reproduce any full-data procedure. This sufficiency argument is asserted only in the known-covariance Gaussian experiment: with unknown covariance, residuals can carry information about Γ. Thm. 4 at this covariance gives the first inequality. Since $\Sigma_m \succeq \Sigma_b$, inversion, compression to $u_0^\perp$, and a second inversion give

$$B_0^\top \Sigma_m^{-1} B_0 \preceq B_0^\top \Sigma_b^{-1} B_0, \qquad (B_0^\top \Sigma_m^{-1} B_0)^{-1} \succeq (B_0^\top \Sigma_b^{-1} B_0)^{-1}.$$

Taking traces proves the second inequality. Both compressed matrices are positive definite because $\Sigma_b \succ 0$. □

C.2 Proof of Theorem 2

Proof. For necessity, use the common constants chosen in Appendix B.2, and write r_0 for the cutoff from Thm. 4. Complete any abstention of a procedure meeting the stated accuracy requirement by an arbitrary unit vector. Regard each reported vector as an axis and attach the deterministic sine radius $r := \sin\varepsilon$ to every report. Since line error is no greater than the angle to the population direction, the resulting procedure has false line report probability at most δ and reporting probability one over the same class and neighborhood. Since $r \le r_0$ and $Kr \le K\varepsilon$, it satisfies (7)–(8) with $\alpha = \delta$, $\beta = 0$ on the cap of radius Kr. Equation (9), together with $T_\star \ge 1$, proves the lower bound. All lower-bound alternatives have the same magnitude and covariances, so supplying those quantities and u_0 to the procedure does not alter the lower bound.

For sufficiency, put $X_t := \bar{z}_t$, $S := \Sigma_m$, and $\bar{X} := T^{-1}\sum_t X_t$. The template means are independent, with $\mathbb{E}X_t = au$ and $\mathrm{Cov}(\bar{X}) = S/T$ throughout the moment class. In the basis $[u_0, B_0]$, write

$$[u_0, B_0]^\top S [u_0, B_0] = \begin{pmatrix} s & c^\top \\ c & A \end{pmatrix}, \qquad b = \frac{c}{s}, \qquad C = A - \frac{cc^\top}{s}. \tag{10}$$

Block inversion gives $C = (B_0^\top S^{-1} B_0)^{-1}$ and $\tau := \mathrm{tr}\, C = \tau_F^{(m)}(u_0)$. Set $R := K\varepsilon$; the theorem ensures $R \le \pi/4$ and $\|b\| \tan R \le 1/2$. Parameterize the cap by

$$H_R := \{h \in \mathbb{R}^{d-1} : \|h\| \le \sin R\}, \qquad u(h) := \sqrt{1 - \|h\|^2}\, u_0 + B_0 h.$$

Define

$$D := B_0^\top - b u_0^\top, \qquad Y := \frac{D\bar{X}}{a} + b, \qquad F(h) := h + b\big(1 - \sqrt{1 - \|h\|^2}\big).$$

Then $\mathbb{E}Y = F(h)$ and $DSD^\top = C$, so $\xi := Y - F(h)$ has mean zero and

$$\mathrm{Cov}(\xi) = \frac{C}{Ta^2}, \qquad \mathbb{E}\,\|\xi\|^2 = \frac{\tau}{Ta^2}.$$

These identities use only the stated second moments. Since $h \mapsto \sqrt{1-\|h\|^2}$ has gradient norm at most $\tan R$ on H_R, the locality condition gives

$$\|F(h) - F(k)\| \geq (1 - \|b\| \tan R)\,\|h-k\| \geq \tfrac{1}{2}\,\|h-k\|, \qquad h,k \in H_R.$$

Choose a measurable minimizer $\widehat{h} \in \arg\min_{k\in H_R}\|Y - F(k)\|$ and report $\widehat{u} := u(\widehat{h})$. Compactness and continuity ensure existence; a fixed measurable selection resolves ties. The deterministic least-squares stability argument (Pázman & Pronzato, 2014, Thm. 1 and Eq. (8)) gives

$$\left\|\widehat{h} - h\right\| \leq 2\left\|F(\widehat{h}) - F(h)\right\| \leq 2\Big(\left\|F(\widehat{h}) - Y\right\| + \|Y - F(h)\|\Big) \leq 4\,\|\xi\|.$$

Both u and $\widehat{u}$ lie in the cap of radius $R \leq \pi/4$ around u_0, so $\theta := \arccos(u^\top \widehat{u}) \leq 2R \leq \pi/2$. The Jacobian of $u(h)$ has operator norm at most $1/\cos R \leq \sqrt{2}$. Hence $\sin\theta \leq \|\widehat{u} - u\| \leq 4\sqrt{2}\|\xi\|$. Since sine is increasing on $[0, \pi/2]$, Markov's inequality yields, uniformly over the moment class and cap,

$$\mathbb{P}\{\theta > \varepsilon\} \leq \frac{32\tau_F^{(m)}(u_0)}{Ta^2 \sin^2\varepsilon}.$$

Taking $T := \lceil 32\tau_F^{(m)}(u_0)/(\delta a^2 \sin^2\varepsilon)\rceil$ proves the upper bound with a procedure that always reports a unit vector. □

C.3 Scope and role of the assumptions

C.3.1 Known parameters and attainable guarantees

The upper-bound construction requires the parameters supplied in the local population setting of Sec. 4, including the numerical value of a in $Y = D\bar{X}/a + b$, and condition (4); it gives no reporting guarantee for an unknown, varying magnitude. The lower bound holds throughout the universal resolution range without that additional condition and is proved with a, u_0, Σ_b, Γ supplied, even though this information makes estimation easier. It therefore also holds when those parameters are not supplied; the upper-bound construction uses the supplied information.

Fisher trace and sample-mean tangent noise. With the blocks in (10), $\tau_F^{(m)}(u_0) = \mathrm{tr}\,A - \|c\|^2/s \leq \mathrm{tr}\,A$, whereas the normalized sample mean has first-order tangent covariance $A/(Ta^2)$ at u_0. The Fisher-trace upper bound is attained by the regression-adjusted local estimator; substituting estimated parameters into the Fisher trace does not calibrate the normalized sample mean or the approximate public bootstrap.

Confidence ball with a trace bound. For a declared $V_m \geq \mathrm{tr}\,\Sigma_m$, apply Appendix B.3 to the T template means with $V = V_m$. This gives a calibrated confidence ball and its reporting guarantee without knowledge of a or a locality restriction. The bound uses $\mathrm{tr}\,\Sigma_m$ rather than the smaller Fisher trace and requires a valid V_m, not an unvalidated covariance estimate.

Scope of the bounds. The bounds concern the uniform accuracy requirement in Sec. 4, not every individual distribution or realized reported radius. Gaussianity is used only for the lower-bound construction; the upper bound uses the stated second moments.

D Confidence-set methods and calibration

This appendix gives the confidence-ball geometry, calibration procedures and proofs, and the local-response minimum used in Secs. 5.3–5.4.

D.1 Confidence-ball geometry

The containment rule of Sec. 3 gives the following confidence-ball implementation.

Theorem 6 (Decision rule for a confidence ball). *Suppose $R \geq 0$ satisfies $\mathbb{P}\{\|\widehat{\Delta}-\Delta\|_2 \leq R\} \geq 1-\delta$, with $\delta \in (0,1)$, and set $\mathfrak{B} := \{\theta : \|\theta - \widehat{\Delta}\|_2 \leq R\}$, $m := \|\widehat{\Delta}\|_2$. Every point in $\mathfrak{B}$ has norm at least $(m-R)_+$. When $m > R$, its directions form exactly the spherical cap centered at $\widehat{u} := \widehat{\Delta}/m$ with half-angle*

$$\varepsilon_{\text{ball}} := \arcsin(R/m).$$

The ball supports an oriented ε_0-direction exactly when $m > R$ and $R \leq m \sin \varepsilon_0$, signal only when $m > R > m \sin \varepsilon_0$, and abstention when $m \leq R$. It never reports an unoriented-line verdict.

Proof of Theorem 6. For every $\theta \in \mathfrak{B}$, the reverse triangle inequality gives $\|\theta\|_2 \geq (m-R)_+$; on the stated coverage event this applies to Δ. If $m > R$, it also excludes zero. For a unit vector v, the ray $\{rv : r > 0\}$ intersects $\mathfrak{B}$ exactly when its distance from $\widehat{\Delta} = m\widehat{u}$ is at most R. Because $R < m$, every such ray has positive projection on $\widehat{u}$; its distance is $m \sin \angle(v, \widehat{u})$. Hence intersection is equivalent to $\angle(v, \widehat{u}) \leq \arcsin(R/m)$, proving the radial-image identity. The three verdict conditions are then precisely the set containments in Sec. 3. If $m \leq R$, the ball contains zero and cannot support a signal claim. □

D.1.1 Null example: normalization without a population direction

For fixed pairs and independent Rademacher signs s_t, $\mathbb{E}_s[s_t z_t \mid z_t] = 0$, although $\widehat{\Delta}_s := T^{-1}\sum_t s_t z_t$ is generically nonzero. For two independent isotropic Gaussian pools, their normalized signed means are independent and uniform on the sphere, giving both identities in (2). The $1/d$ squared-cosine identity does not apply to two rerandomizations of one fixed pool: at $T = 1$ their squared cosine is always one. In either null construction, the target direction is undefined.

D.2 Calibration assumptions and population targets

The geometric claim map is agnostic to the calibration method, but nontrivial uniform finite-sample inference for an unrestricted population mean faces the Bahadur–Savage obstruction (Bahadur & Savage, 1956). Our finite-sample calibrations therefore use declared distributional bounds or the mean of bounded observations; merely assuming that each distribution has finite moments supplies no common bound.

Independent sampling units. For i.i.d. clusters $\mathcal{Z}_t$ and a fixed summary g, the units $W_t = g(\mathcal{Z}_t)$ are i.i.d.; independence within a cluster is unnecessary. Calibration applies to their declared mean under its own moment or support assumptions. Equal-cluster and observation-weighted means are different targets.

Clipping and the population target. For fixed $B > 0$, define $\psi_B(w) := w \min\{1, B/\|w\|\}$, with $\psi_B(0) = 0$, and let $\Delta_B := \mathbb{E}\psi_B(W)$. If $\mathbb{E}\|W\| < \infty$, the triangle inequality gives

$$\|\Delta - \Delta_B\| \leq \mathbb{E}\,\|W - \psi_B(W)\| = \mathbb{E}(\|W\| - B)_+.$$

A declared bound $\|\Delta - \Delta_B\| \leq \bar{\beta}_B$ therefore enlarges the radius in Prop. 7 to $R_B + \bar{\beta}_B$ for a raw-mean claim. Without such a bound, coverage concerns Δ_B. If $\|W\| \leq B$ almost surely, clipping preserves the original target.

D.3 Bounded-data calibration

Clipping bounds the observations and gives a computable finite-sample certificate for their mean. Appendix D.2 explains the assumptions and the distinction from the raw population mean.

Proposition 7 (Clipped certificate). *Let $W_1, \ldots, W_T$ be i.i.d. declared units. Fix $B > 0$ before seeing the certification sample and set $X_t := W_t \min\{1, B/\|W_t\|_2\}$, with $X_t := 0$ when $W_t = 0$. For $\Delta_B := \mathbb{E}X_t$, $\bar{X} := T^{-1}\sum_t X_t$, and $\delta \in (0,1)$, the ball centered at $\bar{X}$ with radius*

$$R_B(T, \delta) := \frac{B}{\sqrt{T}}\left(1 + \sqrt{2\log(1/\delta)}\right)$$

covers Δ_B with probability at least $1-\delta$. Thm. 6 then supplies finite-sample verdicts.

Proof of Proposition 7. Radial clipping gives $\|X_t\| \le B$ almost surely. For $f := \|T^{-1}\sum_t X_t - \Delta_B\|$, replacing one observation changes f by at most $2B/T$. Independence gives

$$\mathbb{E}f \le \sqrt{\mathbb{E}f^2} = \sqrt{T^{-2}\sum_t \mathbb{E}\,\|X_t - \Delta_B\|^2} \le B/\sqrt{T}.$$

McDiarmid's inequality (McDiarmid, 1989) adds a deviation of at most $B\sqrt{2\log(1/\delta)/T}$ with probability at least $1-\delta$. Their sum is $R_B(T,\delta)$; Thm. 6 then gives the verdict. □

These verdicts concern Δ_B. A raw-mean claim requires a declared bias bound $\|\Delta - \Delta_B\|_2 \le \bar{\beta}_B$, which enlarges the radius to $R_B + \bar{\beta}_B$ (Appendix D.2).

D.4 Empirical calibration: one joint multiplier ball

Bootstrap calibration estimates a sampling distribution from the observed data (Efron, 1979). For raw activation means, we use a single residual multiplier ball; high-dimensional norm-bootstrap theory requires additional approximation conditions (Giessing & Fan, 2020).

Computational procedure. The inputs are the declared independent units $W_1, \dots, W_T$, resolution $0 \le \varepsilon_0 < \pi/2$, error settings $\delta, \eta_{\mathrm{MC}} \in (0,1)$, and a draw count B_{boot} satisfying the quantile requirement in step 3.

1. Compute the sample mean $\bar{W} := T^{-1}\sum_t W_t$.
2. For each $b = 1, \dots, B_{\mathrm{boot}}$, sample independent $\xi_{bt} \sim N(0,1)$ and compute

$$G_b^* := \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \xi_{bt}(W_t - \bar{W}), \qquad Q_b^* := \|G_b^*\|_2^2.$$

3. Sort $Q_{(1)}^* \le \cdots \le Q_{(B_{\mathrm{boot}})}^*$. To account for Monte Carlo quantile error, put

$$k := 1 + F^{-1}_{\mathrm{Bin}(B_{\mathrm{boot}},1-\delta)}(1-\eta_{\mathrm{MC}}), \quad q^* := Q^*_{(k)}, \quad R^* := \sqrt{q^*/T},$$

   where the draw count must be large enough that $k \le B_{\mathrm{boot}}$.
4. Form $\mathfrak{B}^* := \{\theta : \|\theta - \bar{W}\|_2 \le R^*\}$ and set $m := \|\bar{W}\|_2$. If $m \le R^*$, report ABSTAIN. Otherwise set $\widehat{u} := \bar{W}/m$ and $\widehat{\varepsilon} := \arcsin(R^*/m)$, and report ORIENTED DIRECTION if $R^* \le m\sin\varepsilon_0$, or SIGNAL ONLY otherwise. Report the verdict with the calibration status justified by the assumptions below.

We refer to this frozen specification as `bootstrap-total-joint-ball-v1`. The verdict conditions follow from Thm. 6; coverage is asymptotic under the following proposition's assumptions.

Proposition 8 (Residual multiplier-ball coverage)**.** *Suppose d is fixed, the W_t are i.i.d. with mean Δ and finite second moment, and the cdf of $\|Z\|_2$ is continuous at its $(1-\delta)$-quantile for $Z \sim N(0, \mathrm{Cov}(W_t))$. Then, jointly over the data and the Monte Carlo draws,*

$$\liminf_{T\to\infty} \mathbb{P}\{\Delta \in \mathfrak{B}^*\} \ge 1 - \delta - \eta_{\mathrm{MC}}.$$

Conditional geometry is finite-sample exact once $\mathfrak{B}^*$ is returned; only coverage of the raw mean is asymptotic. For growing d, additional high-dimensional approximation conditions are required (Giessing & Fan, 2020) and are not assumed to hold automatically for activations. Thus deterministic propagation of this ball yields claims with asymptotic support, without a finite-sample guarantee.

Proof of Proposition 8. The multivariate central limit theorem gives $\sqrt{T}(\bar{W} - \Delta) \Rightarrow Z$. Conditional on the data, G_b^* is exactly Gaussian with covariance $T^{-1}\sum_t (W_t - \bar{W})(W_t - \bar{W})^\top$, which converges in probability to $\mathrm{Cov}(W_t)$. Conditional weak convergence, the continuous mapping theorem, and continuity at the target quantile therefore make the exact conditional multiplier quantile consistent for the sampling quantile of $\|\sqrt{T}(\bar{W} - \Delta)\|_2$.

For the Monte Carlo step, write q_{cond} for the exact conditional $(1-\delta)$-quantile of Q_b^*. Conditional on the data, the number of draws strictly below q_{cond} is binomial with success probability $p_< := \mathbb{P}\{Q_b^* < q_{\text{cond}} \mid W_1, \ldots, W_T\} \le 1-\delta$. It equals $1-\delta$ for a continuous conditional distribution; otherwise its count is stochastically dominated by $\text{Bin}(B_{\text{boot}}, 1-\delta)$. The event $Q^*_{(k)} < q_{\text{cond}}$ requires at least k such draws, so the chosen rank bounds its probability by η_{MC}. A union bound with the bootstrap approximation error gives the displayed liminf coverage. □

A nominal 95% card budgets both errors, for example $\delta = 0.0475$, $\eta_{\text{MC}} = 0.0025$, and records the information in Appendix I. Separate cards require familywise allocation: for J predeclared choices, Bonferroni uses $\sum_j(\delta_j + \eta_j) \le \alpha$. A max-bootstrap instead uses matched units, shared Gaussian multipliers, and a recorded joint maximum statistic. Selection otherwise uses a split disjoint from certification (Berk et al., 2013). Resample whole independent pairs or clusters.

D.5 Exact local minimum over a spherical cap

For a declared response and additive intervention, let g be the gradient with respect to the shared additive vector at dose zero. Fix a unit center $\widehat{u} \in \mathbb{R}^d$, $d \ge 2$, and $0 \le \varepsilon \le \pi/2$. Write

$$C_\varepsilon(\widehat{u}) = \{u : \|u\| = 1,\ \widehat{u}^\top u \ge \cos\varepsilon\}, \quad c = g^\top \widehat{u}, \quad B = \left\|(I - \widehat{u}\widehat{u}^\top)g\right\|.$$

The worst local derivative over this cap is

$$M := \min_{u \in C_\varepsilon(\widehat{u})} g^\top u = \begin{cases} 0, & g = 0, \\ -\|g\|, & g \ne 0 \text{ and } -g/\|g\| \in C_\varepsilon(\widehat{u}), \\ c\cos\varepsilon - B\sin\varepsilon, & \text{otherwise.} \end{cases} \tag{11}$$

Indeed, writing $u = \widehat{u}\cos\theta + v\sin\theta$ with $v \perp \widehat{u}$ and $\|v\| = 1$ reduces the minimum at each angle to $c\cos\theta - B\sin\theta$. The unconstrained minimizer is $-g/\|g\|$; if it lies outside the cap, the minimum occurs on its angular boundary. This proves the piecewise formula, including centers with nonpositive derivatives.

For $c > 0$, the unconstrained minimizer lies outside every such cap, and we define

$$\varepsilon_{\text{crit}} := \arcsin(c/\|g\|), \qquad M > 0 \iff \varepsilon < \varepsilon_{\text{crit}}. \tag{12}$$

At equality, the minimum is zero, so strict positivity requires a strict inequality. For $c \le 0$, the center itself has nonpositive derivative; no shrinking of a cap that retains this center can support positivity everywhere. The boundary depends on the input, response, intervention location, and center; it is not a generally recommended resolution. When a confidence set's compatible directions form this cap, $M > 0$ supports the local sign over those directions, with the confidence set's own calibration status. This geometric calculation does not calibrate an approximate cone, establish finite-dose effects, or make claims about a test-question population.

E Sampling variability and sampling design

This appendix specifies the fixed files, sampling designs, and variance decomposition used in Sec. 5.1.

Fixed files and construction recipe. The collection contains 135 persona and 17 advanced-AI-risk (xrisk) files from the model-written evaluations of Perez et al. (2023). File sizes are 518–1,000 for persona and 934–1,000 for xrisk, totaling 150,136 rows. Persona sources are pinned to `anthropics/evals` revision `84fcc677`; xrisk sources are pinned to `Anthropic/model-written-evals` revision `d533f626`. The fixed local copies match upstream row order in 151 files; `corrigible-less-HHH` retains 532 additional row copies. All original multiplicities enter every reported construction and finite-file target. For zero-based row index i, persona's matching answer is placed at B when $\lfloor i/2 \rfloor$ is even and at A otherwise; xrisk preserves the source A/B assignment. Each contrast subtracts the nonmatching-answer activation from the matching-answer activation. Llama uses the CAA chat rendering; Gemma and Qwen use their native chat templates, with the answer appended and no extra EOS or truncation. All use token index -2, verified to contain the answer letter, including Qwen's fused answer token. Llama extraction uses FP16/SDPA, Gemma BF16/eager attention, and Qwen BF16/SDPA; model, rendering, and numerical conditions vary together.

Table 6: **Sampling-design verdicts at resolution 0.35.** Ordinary sampling and answer-position stratification target the same finite-file mean at $n = 500$, using 200 paired constructions per file. Expected oriented-direction verdict counts average over repetitions. Calibration status is bootstrap/approximate.

Model	Family	ORIENTED DIRECTION ordinary → stratified
Llama-2-7B	persona	49.0 → 95.5 / 135
	xrisk	16.0 → 17.0 / 17
Gemma-2-2B	persona	12.0 → 44.3 / 135
	xrisk	16.6 → 17.0 / 17
Qwen2.5-0.5B	persona	0.0 → 2.0 / 135
	xrisk	0.0 → 14.2 / 17
Qwen2.5-7B	persona	0.0 → 0.2 / 135
	xrisk	0.0 → 16.2 / 17

Sampling design for the finite-file mean. Let stratum h contain N_h rows and have weight $w_h = N_h/N$. Proportional sampling rounds nw_h by the largest-remainder rule and samples n_h rows without replacement within each stratum; ordinary sampling is uniform without replacement over the entire file. Common random row priorities pair the two designs in each of 200 repetitions per file and sample size. The stratified estimator is $\widehat{\Delta} = \sum_h w_h \bar{z}_h$; both estimators target the original row-weighted mean before normalization. Every positive-weight stratum must have $n_h \geq 1$, and a non-census stratum requires $n_h \geq 2$ for the multiplier below; smaller quotas leave the specified procedure undefined and require stopping. For each Gaussian multiplier draw, form

$$e^* = \sum_h w_h \sqrt{\frac{1 - n_h/N_h}{n_h(n_h - 1)}} \sum_{i=1}^{n_h} \xi_{hi}(z_{hi} - \bar{z}_h), \qquad \xi_{hi} \overset{\text{iid}}{\sim} N(0,1).$$

A census stratum ($n_h = N_h$) contributes zero multiplier noise, including a singleton census, whose contribution is set directly to zero without evaluating $0/0$. The reported runs require $n_h \geq 2$ in every stratum. Ordinary sampling uses one stratum. The radius is the `higher` 0.95 quantile of 399 norms $\|e^*\|$, and both designs use the confidence-ball rule in Sec. 3. When $\|\widehat{\Delta}\| \leq R$, directional coverage is scored against the whole sphere, rather than a hemisphere. These are bootstrap/approximate balls; the finite-population correction does not establish exact coverage.

For each file, angular MSE averages squared signed angles to its exact file-mean direction over construction repetitions. A zero file mean or construction mean makes this angle undefined and stops the comparison, without excluding the file or replacing the construction. Family reductions are $1 - \sum_f \text{MSE}_{f,\text{stratified}} / \sum_f \text{MSE}_{f,\text{ordinary}}$, assigning equal weight to files. The 95% intervals use 1,999 paired resamples of repetitions within each fixed file; they quantify Monte Carlo uncertainty, not generalization across behaviors. Expected verdict counts sum the per-file reporting frequencies. For Llama, Gemma, and Qwen-0.5B, angular MSE uses separate paired draws with FP64 activation subtraction; verdicts use FP32 subtraction followed by FP64 scoring. Qwen-7B's two metrics use the latter draws. Thus the MSE estimates in Table 2 and verdict counts in Table 6 do not claim identical individual constructions across metrics.

Sampling-design verdicts at resolution 0.35. The recorded $n = 500$ comparison uses ordinary sampling and proportional answer-position stratification without replacement across all 152 files at the four model–layer settings. Persona answer positions follow a fixed row-index rule; xrisk retains the source assignments. Both designs estimate the original finite-file mean and use a common multiplier ball with finite-population correction and bootstrap/approximate calibration status.

Table 6 complements the angular MSE comparison in Table 2 with expected oriented-direction verdict counts. Stratification lowers angular MSE in all eight model–family conditions, with different effects on verdicts. For Llama persona, MSE falls by 63.3% and the expected oriented-direction count rises from 49.0 to 95.5 out of 135. For Qwen-7B persona, MSE falls by a larger 79.9%, but the count rises only from zero to 0.2 out of 135. Stratification can improve estimation of the same contrast, but even large relative gains need not reach the declared resolution.

Calibration status and error denominators. In the finite-file sampling-design comparison, Qwen-7B persona at $n = 500$ under stratification has 95.1% ball coverage and 96.5% directional coverage. SteerCert reports 47 oriented-direction verdicts among 27,000 constructions; 21 have actual angular error exceeding 0.35 rad. The observed unconditional false-report rate is $21/27000 \approx 0.078\%$, whereas the error fraction conditional on an oriented-direction report is $21/47 \approx 44.7\%$. These are descriptive frequencies from an approximate procedure, not validation of the finite-sample guarantee. The guarantee under valid confidence-set coverage controls unconditional false reports and does not give the same bound conditional on reporting an oriented-direction verdict.

CAA variance decomposition. Sec. 5.1 and Fig. 2 examine normalized directions from the fixed CAA myopic-reward and hallucination frames. The myopic-reward frame contains 908 construction examples after removing the 40 evaluation questions and two smoke questions; the hallucination frame contains 958 after the corresponding exclusions. Both source datasets had been used in earlier analyses. For each frame, Llama-2-7B-chat uses block 13 and Gemma-2-2B-IT block 12. The plotted estimates use 30 ordinary-sampling constructions without replacement per condition: $n = 128, 512$ for myopic reward and $n = 32, 128$ for hallucination. Each construction normalizes the mean activation difference between behavior-matching and nonmatching answer tokens from a sample drawn without replacement. The A- and B-matching examples define two strata. The axis is the normalized component of the difference between the full-frame A- and B-matching stratum means perpendicular to the full-frame mean direction. It is defined without fitting to response outcomes. The axis share divides the sample variance of the normalized directions' scalar projections onto that axis by their total sample variance across constructions. Variance is centered at the mean of the constructed directions; it is distinct from mean squared error relative to the full-frame direction. Orthogonal projection separates the total variance into the axis component and its orthogonal remainder. The plotted pointwise 95% intervals use the saved 1,999-replicate construction-bootstrap estimates and condition on the fixed frames. The myopic-reward decomposition is a retrospective exploratory analysis of existing constructions. Answer position and content are not independently manipulated, so the decomposition does not identify a causal answer-position effect.

F Response repeatability and directional error

This appendix gives the construction and response measurements for Sec. 5.2, additional comparisons, and the exploratory matched-angle diagnostic.

F.1 Construction and response measurement

Frames and sampling. We use the myopic-reward and hallucination frames described in Appendix E, with 908 and 958 construction examples, respectively. For each behavior, the 40 evaluation questions are balanced by the original position of the behavior-matching answer (20 A and 20 B); these questions and two separate smoke questions are excluded from the construction frame. Evaluation eligibility excludes canonical question matches to the historical response panels named in the experiment records. The source datasets had been used previously, so the exclusion does not make them previously unused datasets. Each evaluation question is rendered in both answer orders.

Ordinary sampling draws n examples without replacement. Proportional answer-position stratification assigns stratum counts by largest-remainder rounding and weights the two sample stratum means by their original frame proportions. Both designs therefore estimate the same raw finite-frame mean before normalization. Common random row priorities pair the designs and nest the sample sizes within each of 30 independent construction repetitions. The myopic-reward random-stratification control uses one fixed, count-preserving random permutation of the stratum labels, with the same allocation and weighting rule. It is distinct from the random-axis control in Appendix F.4. Myopic reward uses $n = 128, 512$; hallucination uses $n = 32, 128$. For hallucination, the Llama stratified-32 condition has 60 prespecified constructions for the correlation analysis, while sampling-design ratios use only the common 30 paired constructions.

Intervention. Llama-2-7B-chat uses block 13 and Gemma-2-2B-IT block 12, with zero-based indices. Each example contributes a post-block, final-answer-token activation difference between the behavior-matching and nonmatching answers. Native chat rendering is followed by the assistant prefix

(; literal A/B continuation token IDs are appended for the construction contrasts. For a constructed unit direction $\widehat{u}$, the injected vector is $\kappa c\widehat{u}$, with $c = 7.6808801498667405$ and coefficient κ. The intervention is applied only at the final prompt token, at the same block used for construction. The full-frame reference uses the same norm, coefficient, and injection position. A common norm across models does not imply a common intervention effect. Inference uses FP32, with SDPA for Llama and eager attention retaining attention soft-capping for Gemma.

Response variance and response MSE. Let p_{rjo} be the behavior-matching answer letter's full-vocabulary probability for construction r, question j, and answer order o. Let p_{ujo} be the corresponding probability under the unit full-frame mean direction u. With $J = 40$ questions, two orders, and R constructions, we report

$$V = \frac{1}{2J}\sum_{j,o}\frac{1}{R-1}\sum_{r=1}^{R}(p_{rjo} - \bar{p}_{jo})^2, \qquad D_r = \frac{1}{2J}\sum_{j,o}(p_{rjo} - p_{ujo})^2, \qquad \bar{p}_{jo} = \frac{1}{R}\sum_{r=1}^{R} p_{rjo}.$$

Response variance V measures variation across reconstructions; response MSE D_r measures deviation from the full-frame reference response for one construction. Comparisons of mean response MSE average D_r over constructions. The full-frame direction is an evaluation reference, not a construction method with the same sample budget. Angular error is the signed-direction angle $\theta_r = \arccos(u^\top \widehat{u}_r)$, without taking the absolute value of the inner product. The correlation analysis uses $\mathrm{Spearman}(\theta_r^2, D_r)$, which equals $\mathrm{Spearman}(\theta_r, D_r)$ because squaring preserves the ranks of nonnegative angles. Neither response variance nor response MSE measures the sign of the intervention effect or free-generation quality.

Uncertainty and analysis status. The response comparisons use 1,999 paired two-way percentile bootstrap repetitions, resampling construction IDs and question identities within their original A/B strata. Both answer orders and paired sampling designs are retained together. Intervals are pointwise and condition on the fixed source frames and selected model, layer, and coefficient settings; they are not simultaneous guarantees. The studies are exploratory, and the Gemma myopic-reward study was added after the Llama/Qwen parent study. The myopic-reward protocols fixed the model–layer settings, intervention norm, three coefficients $(-0.25, -0.5, -1)$, and sample sizes $(128, 512)$ before each study's response-inference run. Gemma inherited the parent's norm, coefficients, sample sizes, and question panel and used its established construction-audit block 12. These choices were made in an ongoing exploratory program with prior results available; fixing them for execution does not make them independent of earlier development. The main-text display of full-vocabulary probability at $\kappa = -1$ harmonizes saved conditions retrospectively. The original myopic-reward primary endpoint was conditional A/B probability, whereas the hallucination primary endpoint was full-vocabulary probability at $\kappa = -1$.

F.2 Additional response-repeatability comparisons

Random stratification and absolute scale. For Llama myopic reward at $n = 128$ and $\kappa = -1$, response variances are 2.305×10^{-3} under ordinary sampling, 1.537×10^{-4} under answer-position stratification, and 2.443×10^{-3} under random stratification. The answer-position/random-stratification variance ratio is $0.063\,[0.037, 0.103]$. For Gemma, the corresponding variances are 3.579×10^{-6}, 1.522×10^{-6}, and 3.756×10^{-6}, with ratio $0.405\,[0.241, 0.669]$. These comparisons support a benefit from the answer-position partition in this setup; they do not establish equivalence between random stratification and ordinary sampling or a causal mediation fraction.

Mean effects. We measure mean probability reduction from the unsteered baseline separately from repeatability. At myopic-reward $n = 128$ and $\kappa = -1$, the reductions under ordinary and answer-position-stratified sampling are 19.024 and 19.346 percentage points for Llama, and 1.720 and 1.734 for Gemma. The paired differences (stratified minus ordinary) are $0.322\,[-0.493, 1.139]$ percentage points for Llama and $0.0143\,[0.0046, 0.0292]$ for Gemma. The comparisons do not establish equivalence or noninferiority of mean effects.

Different sample sizes. For myopic reward at $\kappa = -1$, the response variance ratio for stratified 128 versus ordinary 512 is $0.303\,[0.150, 0.670]$ on Llama and $3.716\,[2.129, 6.195]$ on Gemma. The

Table 7: **Myopic-reward repeatability across response endpoints and coefficients.** Response variance ratios compare answer-position stratification with ordinary sampling at the same n. Intervals are pointwise 95% paired bootstrap intervals; each condition uses 30 constructions and 40 questions in both answer orders.

Coefficient	n	Llama	Gemma
Conditional A/B probability			
-0.25	128	$0.056\ [0.026, 0.099]$	$0.393\ [0.264, 0.610]$
-0.25	512	$0.040\ [0.019, 0.087]$	$0.470\ [0.334, 0.681]$
-0.5	128	$0.054\ [0.026, 0.097]$	$0.380\ [0.249, 0.602]$
-0.5	512	$0.037\ [0.017, 0.080]$	$0.454\ [0.317, 0.661]$
-1	128	$0.051\ [0.026, 0.093]$	$0.359\ [0.220, 0.592]$
-1	512	$0.034\ [0.018, 0.069]$	$0.429\ [0.279, 0.638]$
Full-vocabulary probability			
-0.25	128	$0.069\ [0.040, 0.114]$	$0.438\ [0.282, 0.678]$
-0.25	512	$0.050\ [0.025, 0.103]$	$0.526\ [0.352, 0.764]$
-0.5	128	$0.065\ [0.034, 0.110]$	$0.434\ [0.275, 0.676]$
-0.5	512	$0.046\ [0.023, 0.095]$	$0.523\ [0.349, 0.758]$
-1	128	$0.067\ [0.038, 0.111]$	$0.425\ [0.258, 0.679]$
-1	512	$0.052\ [0.026, 0.113]$	$0.515\ [0.341, 0.748]$

Table 8: **Qwen myopic-reward repeatability at both response endpoints.** Response variance ratios (answer-position stratification / ordinary sampling) at the same n, with pointwise 95% paired bootstrap intervals; 30 constructions and 40 questions in both answer orders.

Coefficient	n	Conditional A/B	Full-vocabulary
-0.25	128	$0.146\ [0.076, 0.327]$	$0.392\ [0.228, 0.780]$
	512	$0.124\ [0.067, 0.304]$	$0.228\ [0.141, 0.678]$
-0.5	128	$0.135\ [0.068, 0.316]$	$0.377\ [0.220, 0.741]$
	512	$0.111\ [0.062, 0.285]$	$0.226\ [0.139, 0.639]$
-1	128	$0.114\ [0.059, 0.292]$	$0.327\ [0.193, 0.672]$
	512	$0.080\ [0.046, 0.198]$	$0.201\ [0.126, 0.532]$

Llama angular-MSE ratio for that comparison is 1.596, so its ordering by angular MSE differs from its ordering by response variance. These results do not establish a general fourfold reduction in sample or compute requirements.

Original endpoint and other coefficients. Table 7 retains the conditional A/B endpoint and all three evaluated myopic-reward coefficients alongside the full-vocabulary endpoint. Conditional A/B probability divides the matching-answer probability by the sum of the A and B probabilities. The two endpoints use the same stored responses but answer different measurement questions.

Qwen response comparison. The parent myopic-reward study also evaluated Qwen2.5-7B-Instruct at zero-based block 13 using the same 908-example frame, 40 questions in both answer orders, 30 paired constructions, intervention norm, three coefficients, and sample sizes as Llama. It used the native chat rendering with the assistant prefix `(`, explicit literal A/B continuation IDs, FP32/SDPA, and intervention at the final prompt token. The mean unsteered A+B probability mass was 0.0272% in this setting. Table 8 reports the completed comparison at both endpoints; the small full-vocabulary scale motivates presenting it separately from the main-text Llama/Gemma comparison, not excluding observations from the experiment. At $n = 128$ and $\kappa = -1$, full-vocabulary response variances were 2.749×10^{-9} under ordinary sampling, 9.002×10^{-10} under answer-position stratification, and 2.871×10^{-9} under random stratification; the answer-position/random ratio was $0.314\ [0.195, 0.659]$. For stratified 128 versus ordinary 512, the variance ratio was $2.043\ [1.223, 4.157]$, so the lower-budget comparison favored ordinary sampling despite the same-budget reduction. The mean probability-reduction difference at $n = 128$, stratified minus ordinary, was $-0.0000883\ [-0.000425, 0.000177]$ percentage points. All intervals are pointwise 95% paired bootstrap intervals under the procedure above; these results concern this response format and do not establish free-generation quality.

Table 9: **Additional associations between angular error and response MSE.** Within-condition Spearman correlations at coefficient -1, with pointwise 95% paired bootstrap intervals. Each condition uses 30 constructions, except Llama hallucination stratified 32, which uses all 60 prespecified constructions; all use 40 questions in both answer orders.

Sampling design	n	Llama	Gemma
Myopic reward			
Random stratification	128	$0.898\,[0.724, 0.956]$	$0.817\,[0.511, 0.918]$
Random stratification	512	$0.947\,[0.857, 0.978]$	$0.619\,[0.190, 0.849]$
Hallucination			
Ordinary	32	$0.748\,[0.493, 0.893]$	$0.612\,[-0.026, 0.831]$
Stratified	32	$0.517\,[0.278, 0.678]$	$0.183\,[-0.213, 0.591]$
Ordinary	128	$0.800\,[0.537, 0.911]$	$0.615\,[-0.040, 0.825]$
Stratified	128	$0.518\,[0.199, 0.750]$	$0.257\,[-0.223, 0.580]$

F.3 Angular error and response deviation

The myopic-reward associations in Table 4 are computed within each model, sampling design, sample size, and coefficient condition, using 30 constructions. They are a retrospective analysis of the existing response panel. Their bootstrap resamples constructions within the fixed prediction folds described below and resamples questions within the original A/B strata, retaining both answer orders. We do not test differences between the correlations reported for different conditions. In particular, intervals spanning zero for stratified 512 do not establish absence of an association, and the ordinary-512 results do not support a claim that increasing sample size alone weakens the association. Table 9 gives the random-stratification and hallucination results at coefficient -1. The 60-construction Llama hallucination stratified-32 analysis is distinct from the 30-construction paired design comparison in Table 3.

Association and prediction. We separately evaluate whether squared angular error improves prediction of response deviations for constructions excluded from fitting. Within each myopic-reward condition, alternating construction IDs define two fixed folds of 15 constructions. For each question and answer order, a linear model of the probability deviation from the full-frame response is fitted on one fold and evaluated on the other, then the roles are reversed. The comparator is an intercept-only model using the same training constructions. Bootstrap resampling retains the fixed fold membership and refits the models, preventing an original construction from appearing in both training and evaluation folds. For ordinary 128 at coefficient -1, the prediction-MSE reductions relative to the intercept-only model are $-7.5\%\,[-98.2, 21.3]$ for Llama and $-57.8\%\,[-176, 24.7]$ for Gemma. Thus the observed rank associations do not establish improved prediction by this squared-angle model. This evaluation separates construction repetitions, not questions, datasets, or models. The angular error uses the known finite-frame reference and is distinct from SteerCert's estimated angular radius. For example, in Llama hallucination stratified 32, the correlation of the estimated ball-radius-to-contrast-norm ratio with response MSE is $0.045\,[-0.222, 0.284]$, despite the positive association with realized angular error.

F.4 Matched-angle diagnostic

Design. We use Llama-2-7B-chat at block 13, myopic reward, all 30 ordinary-128 constructions, and the existing 40 questions in both answer orders at coefficient -1. Let u be the unit mean direction of the 908-example frame and q its answer-position axis from Appendix E. For an original unit construction v with angle $\theta = \arccos(u^\top v)$, removing a tangent axis a and restoring the original angle gives

$$w_a = \frac{v - u(u^\top v) - a(a^\top v)}{\|v - u(u^\top v) - a(a^\top v)\|}, \qquad v_a = \cos(\theta)u + \sin(\theta)w_a.$$

We compare the original direction, removal of $a = q$, and removal of a random tangent axis. Each construction's random axis is generated by projecting an independent standard Gaussian vector onto the tangent space perpendicular to u and normalizing, with axes fixed before response evaluation. The random axes are uniform in the 4,095-dimensional tangent space. All three conditions share the

Table 10: **Response MSE differs at matched angular error and intervention norm.** Llama myopic reward, ordinary $n = 128$, coefficient -1; mean response MSE uses full-vocabulary matching-answer probabilities and the full-frame reference response. Intervals are pointwise 95% paired bootstrap intervals.

Direction	Mean response MSE $\times 1{,}000$
Original	$2.293\,[1.402, 3.362]$
Answer-position-axis removal	$0.517\,[0.306, 0.787]$
Random-axis removal	$2.296\,[1.404, 3.370]$

intervention norm, and the two operations preserve each construction's original angular error. The reference direction and axis use the complete frame, so this is an exploratory diagnostic rather than a construction method with the same sample budget.

Results and interpretation. Table 10 reports mean response MSE under the three conditions. The answer-position/random-axis MSE ratio is $0.225\,[0.156, 0.309]$, using 1,999 paired two-way bootstrap repetitions over construction/axis pairs and questions, with both answer orders retained. The intervals are pointwise, and the reused question panel is not an independent replication. This comparison shows that response deviation can differ at the same angular error and intervention norm in this setting. The operations do not match the amount of error removed: median removed tangent-error energy is 64.7% for the answer-position axis and 0.0106% for the random axis, while median rotations from the original direction are $6.27°$ and $0.0747°$. The comparison therefore does not isolate sensitivity per unit of removed error or identify a causal mediation fraction for stratification.

Separate hallucination diagnostic. For Llama hallucination with 60 already-stratified $n = 32$ constructions, the same angle-restoring answer-position-axis operation gives a mean response MSE ratio of $1.027\,[0.928, 1.133]$ relative to the original directions. This comparison does not establish an improvement or equivalence. It differs from the myopic-reward comparison in behavior, sampling design, and comparator, so their difference cannot be attributed to stratification alone.

G Directional uncertainty and local response claims

This appendix gives the finite-file construction comparison and the exploratory, post-hoc sign analysis in Sec. 5.3. The target is the mean direction of a fixed construction file, evaluated using saved activation differences and response gradients.

G.1 Construction and response measurement

We use Llama-2-7B-chat and four CAA construction files: sycophancy, hallucination, myopic reward, and coordination with other AIs. Their matching-answer A/B counts are respectively 489/511, 504/496, 474/476, and 179/181. For each file and block $\ell \in \{11, 12, 13\}$, the target is the full-frame mean direction $u_F = \Delta_F / \|\Delta_F\|$, where Δ_F averages all of that file's activation differences. The reference is used only for evaluation, not for constructing the confidence set or selecting a sampling design.

Each of 200 paired repetitions draws $n = 128$ items without replacement under ordinary sampling, proportional answer-position stratification, and stratification by a fixed random partition. The random partition has the same stratum sizes as the answer-position partition and is sampled once per file. Both stratified estimators use the original file weights $w_h = N_h/N$ and the weighted mean $\widehat{\Delta} = \sum_h w_h \overline{W}_h$. Sample counts n_h use proportional allocation with the largest-remainder rule, breaking ties in favor of the first stratum (A for answer-position stratification). Within a repetition, the designs share item-level random priorities, and the same selected items are used at all three blocks. The random seed is 20260925, with separate streams for sampling, multipliers, and interval estimation.

The response panel uses all 50 released test questions per behavior in file order, each in its released and swapped answer orders. These are previously opened, fixed evaluation panels. We reuse the

model, prompt rendering, and saved gradients specified in Appendix H: the assistant prefix is `(`, and the shared additive vector is applied at the selected block from the final token of `[/INST]` onward, inclusively. For input x and block ℓ, let $f_{x,\ell}(v)$ be the behavior-matching answer letter's full-vocabulary probability under additive vector v and let $g = \nabla_v f_{x,\ell}(0)$. The derivative at zero intervention strength along a Euclidean unit direction u is $g^\top u$. This experiment uses the original complete construction files and this shared-vector intervention; the response-repeatability experiment in Appendix F.1 uses its own construction frames and intervention specification.

G.2 Confidence sets and positive local-response claims

For each reconstruction, we compute a nominal 95% multiplier confidence ball using only the sampled activation differences. For stratum h, write W_{hi} for the sampled differences and $\overline{W}_h$ for their mean. Each of 399 standard Gaussian multiplier draws produces

$$Z^{(b)} = \sum_h w_h \sqrt{\frac{1 - n_h/N_h}{n_h(n_h-1)}} \sum_{i=1}^{n_h} \xi_{hi}^{(b)} (W_{hi} - \overline{W}_h), \qquad \xi_{hi}^{(b)} \overset{\text{iid}}{\sim} N(0,1).$$

Ordinary sampling uses the same expression with one stratum. The ball radius R is the empirical 95th percentile of $\|Z^{(b)}\|$, using the `higher` quantile convention and FP64 calculations. If $R < \|\widehat{\Delta}\|$, its compatible directions form a cap centered at $\widehat{u} = \widehat{\Delta}/\|\widehat{\Delta}\|$ with angular radius $\widehat{\varepsilon} = \arcsin(R/\|\widehat{\Delta}\|)$. Otherwise, the evaluation conservatively uses all directions and withholds a positive local-response claim. Calibration status is bootstrap/approximate.

For each fixed input, let $c = g^\top \widehat{u}$ and $t = g^\top u_F$. The first rule attributes a positive derivative to the target whenever $c > 0$. The second requires $M > 0$, where M is the exact minimum of $g^\top u$ over the compatible directions (Appendix D.5). A positive claim is correct when $t > 0$ and incorrect when $t \le 0$. Withholding a positive claim does not assert a negative derivative; these rules evaluate local-response claims and do not define additional SteerCert verdicts. The 4 behaviors, 3 blocks, 100 input renderings, and 200 repetitions yield 240,000 dependent evaluations per sampling design.

G.3 Sign discrepancies and uncertainty

Table 5 divides every count by all 240,000 evaluations within its sampling design. The 3.117% in the main text instead divides 5,809 incorrect positive claims by the 186,367 positive estimated derivatives under ordinary sampling. All 5,809 have strictly negative reference derivatives; none has $t = 0$. For answer-position stratification, the corresponding fraction is $2,419/189,339 = 1.278\%$ $[1.189, 1.368]$. The random-partition control yields 75.312% correct positive claims, 2.436% incorrect positive claims, and 22.252% withheld claims under the first rule; the second rule withholds all claims.

The intervals in Sec. 5.3 and this subsection are pointwise 95% intervals from 1,999 paired bootstrap resamples of the 200 construction repetitions, keeping methods, behaviors, blocks, and both answer orders paired. They measure Monte Carlo uncertainty conditional on the fixed files, questions, saved gradients, and random partition. Inputs and layers are not resampled as independent observations. The later sign reanalysis reproduces the earlier construction comparison's point estimates; its fixed bootstrap stream gives the interval $[2.965, 3.153]$ for the 3.058% supported fraction.

The reference derivative is positive in 956 of the 1,200 fixed input–layer rows, or 191,200 evaluations over all repetitions. The 7,339 supported claims under answer-position stratification recover 3.838% of these positive evaluations. No incorrect positive claims were observed under the compatible-direction rule, but zero observed errors do not establish zero error probability. Under ordinary sampling this rule makes no positive claims, so its error fraction conditional on a positive claim is undefined. For the ordinary design, replacing the first rule by the second changes incorrect positive claims by -2.420 percentage points $[-2.569, -2.274]$ and correct positive claims by -75.233 percentage points $[-75.497, -74.960]$.

Among the ordinary design's incorrect positive claims, the absolute reference derivative has median 0.333, 90th percentile 3.740, and maximum 19.044; its median among correct positive claims is 1.355. Units are probability percentage points per unit coefficient along a Euclidean unit direction, evaluated at zero intervention strength. These summaries include small derivatives and do not specify a practical-loss threshold or a finite-strength probability change.

Table 11: **Smaller cone radii support positivity for some inputs.** Median radii across 200 reconstructions; support percentages use all 20,000 input–reconstruction evaluations per cell. Stratification is by answer position. Ordinary sampling supports no positive claims in any cell. "All directions" retains reconstructions without a finite cap in the median calculation.

Behavior	Block	Ordinary radius (rad)	Stratified radius (rad)	Stratified support (%)
Sycophancy	11	All directions	1.168	0.000
	12	All directions	0.770	0.000
	13	All directions	0.675	0.000
Hallucination	11	0.236	0.119	2.040
	12	0.232	0.125	2.700
	13	0.241	0.128	2.840
Myopic reward	11	0.280	0.107	0.000
	12	0.241	0.099	1.405
	13	0.267	0.103	13.105
Coordination	11	0.218	0.092	4.330
	12	0.213	0.092	6.025
	13	0.231	0.094	4.250

G.4 Cone radii, input-specific boundaries, and calibration

For $c > 0$, Equation (12) gives the boundary $\varepsilon_{\text{crit}} = \arcsin(c/\|g\|)$ and the equivalence $M > 0$ if and only if $\widehat{\varepsilon} < \varepsilon_{\text{crit}}$. For $c \leq 0$, the center itself prevents positivity throughout any cap that retains it. The center, derivative, and boundary are recomputed for every reconstruction and sampling design; no boundary from an earlier center is reused. The boundary concerns the specified input and response, rather than a universal resolution.

Table 11 summarizes the radii and support rates across all twelve behavior–block conditions. For myopic reward at block 13, answer-position stratification reduces the median radius from 0.267 to 0.103 rad and increases the supported fraction from 0 to 13.105%. At block 11, the median radius decreases from 0.280 to 0.107 rad, but no positive claims are supported. In that condition the boundary summary is approximately 0.025 rad under both designs: within each reconstruction we take the median over inputs with $c > 0$, then take the median across reconstructions. These medians describe the geometry; support counts use the individual input–reconstruction comparisons, not a comparison of medians.

Directional support and confidence-set coverage are evaluated separately. Under answer-position stratification, the cell-level coverage of Δ_F by the nominal 95% balls ranges from 91.5% to 95.5%; coverage of u_F by the compatible directions ranges from 93.0% to 99.5%. For sycophancy at blocks 12 and 13, ball coverage is $183/200 = 91.5\%$ in each cell, with a Wilson 95% interval of $[86.81, 94.63]\%$. The exact cap geometry and the absence of observed incorrect supported claims do not change the bootstrap/approximate calibration status. This fixed-file experiment evaluates attribution to u_F; the public-construction analysis in Appendix H uses a different target, construction size, and confidence procedure.

H Directional precision and local response

This exploratory analysis supplements Sec. 5.4. It evaluates whether the local derivative of a specified response is positive over the joint-ball compatible directions at each fixed evaluation input.

H.1 Joint-ball assessment of public constructions

The CAA construction data use the released generate sets from `nrimsky/CAA`, revision `5dabbbd9a0bca5f25e174501e959de378806aa48`. The four behaviors have 1,000 sycophancy, 1,000 hallucination, 950 myopic-reward, and 360 coordination examples, each evaluated at Llama-2-7B-chat blocks 11–13. The recipe appends the answer in the chat template and reads

the answer-letter activation at token index -2. Each contrast subtracts the nonmatching-answer activation from the matching-answer activation.

The main-text assessment uses the joint multiplier confidence-ball rule for all twelve CAA constructions, with the original construction rows, activation differences, and reconstructed centers. For differences $z_1, \ldots, z_n$ with mean Y, the Gaussian perturbations are $n^{-1} \sum_i \xi_i (z_i - Y)$ with independent standard Gaussian weights. The radius is the 19,133rd ordered norm among 19,999 draws, using error allocations $\delta = 0.0475$ and $\eta_{\mathrm{MC}} = 0.0025$ for the nominal 95% calculation. For zero-based block index ℓ, PCG64 uses seed $\ell + 17$. Scoring uses FP64, and no finite-population correction is applied: the target is the generating-distribution contrast under the row-independence working assumption. The compatible directions form a cap with angular radius $\arcsin(R/\|Y\|)$ when $R < \|Y\|$; otherwise the ball does not support a direction at any resolution below $\pi/2$. The procedure has bootstrap/approximate calibration status; the quantile calculation does not establish coverage of the generating-distribution contrast.

H.2 Inputs, response, and gradient

The 12 public CAA constructions combine four behaviors with Llama-2-7B-chat blocks 11, 12, and 13. Each behavior uses all 50 released test questions in file order, in both the released and swapped answer orders; the same questions are evaluated at all three blocks. These previously opened test files are fixed panels, not new samples for a confirmatory evaluation. The model revision is `f5db02db724555f92da89c216ac04704f23d4590` and the CAA source revision is `5dabbbd9a0bca5f25e174501e959de378806aa48`. For input x and block ℓ, let $f_{x,\ell}(v)$ be the full-vocabulary softmax probability of the behavior-matching answer letter when the shared additive vector v is applied at that block from the final token of `[/INST]` onward, inclusively. The prompt ends with the assistant prefix `(`. The declared gradient is $g = \nabla_v f_{x,\ell}(0)$, so the dose-zero local derivative along a unit direction u is $g^\top u$. This is not the probability conditional on the A/B tokens.

Gradients are computed in FP32 with TF32 disabled, SDPA, and batch size one. Simultaneous zero perturbations at the three blocks yield three partial gradients in one backward pass; this does not describe a simultaneous three-block intervention. The 400 inputs yield 1,200 item–layer rows, not 1,200 independent questions. The reconstructed FP64 center uses the unchanged construction cache, and the cone radius comes from the joint-ball assessment in Appendix H.1. All public cones in this analysis have radii below $\pi/2$; no origin-interior case is treated as a hemisphere.

H.3 Comparisons and results

Appendix D.5 defines the center derivative c, the exact cap minimum M, and the boundary $\varepsilon_{\mathrm{crit}}$, including nonpositive centers and the zero-gradient case. We compare positivity at the reconstructed center with positivity over the joint-ball compatible directions. The released vector supplies a separate center-only comparison; its derivative signs agree with those of the reconstructed center in all evaluated rows.

Positive center derivatives rarely remain positive throughout the cone (Table 12): among 1,200 rows, the center derivative is positive in 956, and the joint-ball M in 50. The 50 rows occur in five cells. Only hallucination has questions supported in both answer orders: 1, 2, and 4 of 50 at blocks 11, 12, and 13, respectively. Although myopic reward has positive M in individual answer orders, it has no question supported in both orders at any evaluated block. With answer orders pooled, the cell-level median $\varepsilon_{\mathrm{crit}}$ ranges from 0.025 to 0.060 rad conditional on a positive center derivative (Table 13); it is not a recommended resolution for general steering use.

Uncertainty and interpretation. These are complete finite-file descriptive counts, with no sampling interval for an unposited population of test questions. Construction uncertainty has the inherited bootstrap/approximate calibration status; there is no simultaneous 95% guarantee across all cells. The analysis does not measure the response at the unknown population direction or a finite-dose effect. A nonpositive cone minimum does not establish that the true population direction has a negative response or that the released vector is ineffective. For each construction, we use the same confidence set for the population contrast across all evaluated inputs. We do not compare this analysis with confidence intervals constructed directly for $g^\top \Delta$ at a prespecified input.

Table 12: **Local-response signs for every public CAA cell.** c is the center derivative; M is the minimum local derivative over the joint-ball cap. Triples give released order / swapped order / their questionwise intersection, each out of 50. Reused questions, layers, and answer orders are not independent trials.

Behavior	Block	Radius	Center $c > 0$	Joint-ball $M > 0$
Sycophancy	11	0.971	45 / 37 / 36	0 / 0 / 0
Sycophancy	12	0.609	43 / 40 / 36	0 / 0 / 0
Sycophancy	13	0.557	44 / 36 / 32	0 / 0 / 0
Hallucination	11	0.093	38 / 38 / 27	4 / 4 / 1
Hallucination	12	0.091	41 / 46 / 38	6 / 5 / 2
Hallucination	13	0.096	42 / 46 / 38	10 / 8 / 4
Myopic reward	11	0.114	48 / 49 / 47	0 / 0 / 0
Myopic reward	12	0.097	41 / 42 / 33	0 / 1 / 0
Myopic reward	13	0.108	35 / 33 / 19	5 / 7 / 0
Coordination	11	0.165	35 / 29 / 22	0 / 0 / 0
Coordination	12	0.162	38 / 37 / 28	0 / 0 / 0
Coordination	13	0.173	40 / 33 / 30	0 / 0 / 0

Table 13: **Input-specific precision boundaries.** Each triple gives the 10th percentile / median / 90th percentile of $\varepsilon_{\text{crit}}$ in radians, conditional on $c > 0$. The count is positive-center rows out of 100; pooled orders remain paired renderings of 50 questions. These are empirical quantiles, not uncertainty intervals.

Behavior	Block	Count	Released order	Swapped order	Pooled orders
Sycophancy	11	82	0.023 / 0.053 / 0.078	0.015 / 0.050 / 0.078	0.016 / 0.051 / 0.079
Sycophancy	12	83	0.021 / 0.054 / 0.084	0.015 / 0.043 / 0.076	0.016 / 0.046 / 0.082
Sycophancy	13	80	0.009 / 0.053 / 0.083	0.011 / 0.065 / 0.107	0.008 / 0.060 / 0.101
Hallucination	11	76	0.010 / 0.045 / 0.094	0.012 / 0.052 / 0.091	0.009 / 0.049 / 0.096
Hallucination	12	87	0.010 / 0.048 / 0.096	0.007 / 0.040 / 0.092	0.007 / 0.045 / 0.095
Hallucination	13	88	0.013 / 0.053 / 0.106	0.007 / 0.050 / 0.105	0.007 / 0.052 / 0.106
Myopic reward	11	97	0.007 / 0.026 / 0.044	0.008 / 0.024 / 0.043	0.007 / 0.025 / 0.044
Myopic reward	12	83	0.011 / 0.049 / 0.086	0.009 / 0.039 / 0.082	0.009 / 0.042 / 0.085
Myopic reward	13	68	0.006 / 0.058 / 0.115	0.004 / 0.039 / 0.119	0.005 / 0.046 / 0.119
Coordination	11	64	0.005 / 0.029 / 0.076	0.003 / 0.018 / 0.082	0.003 / 0.027 / 0.083
Coordination	12	75	0.011 / 0.051 / 0.088	0.007 / 0.027 / 0.080	0.008 / 0.029 / 0.086
Coordination	13	73	0.013 / 0.039 / 0.086	0.005 / 0.031 / 0.084	0.009 / 0.035 / 0.085

Numerical and implementation checks. Source/card/cache identities, reconstructed centers, and released vectors were checked before scoring. The numerical checks compare single-block and simultaneous-block gradients, the shared gradient and the sum of per-position gradients, scalar chain rules, and centered finite differences. The initial padded-batch smoke failed its frozen numerical check; batch size one was adopted and checked before the main run without changing the scientific inputs or scoring. An independent implementation verified all saved geometry, counts, and quantiles.

I Reporting format and reproducibility

Report the construction recipe, population target, independent sampling unit, sample size, resolution, verdict, angular radius, and calibration status. Specify the calibration method and its assumptions, statistical and Monte Carlo error budgets, selection or multiplicity treatment, and any support or clipping bound. Record implementation settings and code/data provenance needed to reproduce the assessment. Nominal error settings alone do not establish coverage. For a local-response claim, additionally specify the response, intervention locations, worst local derivative M over the compatible directions, and, when defined, $\varepsilon_{\text{crit}}$ (Appendix D.5).

I.1 COMPUTE ENVIRONMENT

Activation extraction and public-vector audits ran locally on macOS using PyTorch's MPS backend; finite-pool resampling used the local CPU. Behavioral inference ran on Runpod using an NVIDIA GeForce RTX 4090, PyTorch 2.8.0+cu128, and Transformers 5.16.1.