WE ARE NOT PROBING OR STEERING CONCEPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The *Linear Representation Hypothesis* associates high-level concepts with directions in language models, but it remains unclear how these concept-related linear structures are organized within the model. We propose the *Answer-Basin Representation Hypothesis*: the probability measure induced over answers by the model's continuation distribution organizes these linear structures, with its statistics represented along linear directions shared across questions. All continuations yielding the same answer form an answer basin, whose mass is their total probability. These basin masses define the pushforward probability measure over answers. We posit that concept-related linear structure emerges from differences in the answer measure rather than being determined by changes in concept labels. Experiments across models and tasks link concept-consistent effects and their reversals in probing and steering to the alignment between concept labels and the answer measure.

1 INTRODUCTION

The Linear Representation Hypothesis associates high-level concepts with directions in language models (Park et al., 2024). This view has inspired extensive interpretability research. Researchers use linear probes to identify concepts such as correctness and safety (Marks & Tegmark, 2024; Cencerrado et al., 2025; Zou et al., 2023), and apply activation steering to influence the corresponding model behavior (Turner et al., 2023; Arditi et al., 2024). However, it remains unclear how these concept-related linear structures are organized within the model.

To understand this organization, we start from the generative distribution of a language model. An autoregressive language model predicts tokens sequentially, and the conditional probabilities at successive steps determine the probability of a complete continuation. For example, a math problem can be solved in different ways that lead to the same final answer (Wang et al., 2023). Therefore, for a given question, we define an answer basin as the set of all continuations yielding the same answer and define its mass as their total probability under the model. These basin masses define the pushforward probability measure over answers (Kallenberg, 2021). We call this the answer probability measure, or answer measure.

Our central observation is that **concept-related linear structure emerges with differences in the answer measure rather than changes in concept labels.** Across correctness, social bias, and safety, and across Qwen and Gemma models at different scales, the correct–wrong distinction remains fixed while representational separation varies with the answer-basin mass gap (Figure 1), which means that concept labels alone do not explain why the observed separation varies across these groups. We propose the *Answer-Basin Representation Hypothesis*: statistics of the answer measure organize linear structure in hidden states along directions shared across questions, and their alignment with concept labels determines when these directions exhibit concept-consistent behavior.

Probing experiments show that linear probes capture answer concentration before generation and relative basin mass after generation. Their concept-level behavior, however, depends on the association between labels and basin mass in the training data. Reversing this association reverses their concept predictions on unseen questions, even though the labeling rules and training objective remain unchanged. More interestingly, probes trained solely on basin-mass comparisons between two incorrect answers to the same question still recover candidate mass orderings on unseen questions.

Steering experiments show that reversing the correct–wrong mass relationship in the source data produces vectors with nearly opposite orientations in a shared PCA projection, even though the labels remain unchanged. When a fixed vector is applied across target questions, answers with greater

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

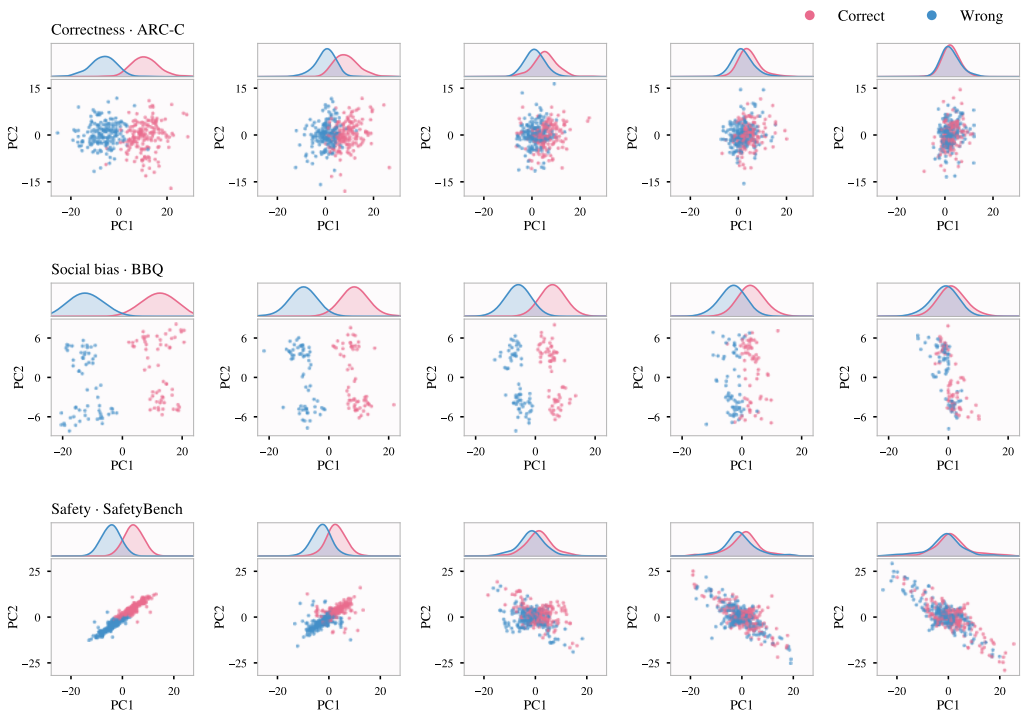


Figure 1: **Larger answer-basin mass gaps accompany clearer separation of correct and wrong answers.** Qwen2.5-7B representations on three attributes: correctness (ARC-C), social bias (BBQ), and safety (SafetyBench). Columns show correct–wrong answer pairs with decreasing mass gaps under the candidate-answer measure. Points are answer representations in a shared PCA space within each row; curves show their PC1 distributions.

initial basin mass receive larger relative gains. An intervention oriented toward correct answers can therefore produce opposite concept-level effects across questions. The answer measure thus connects vector geometry with intervention behavior, explaining when steering produces concept-consistent control and when its effects reverse.

Our main contributions are:

1. We formalize answer basins and the answer measure, and propose the *Answer-Basin Representation Hypothesis* to explain how concept-related linear structure emerges and is organized.
2. Through probing and steering experiments, we show how concept-consistent effects and their reversals depend on the relationship between labels and the answer measure.

2 THE *Answer-Basin Representation Hypothesis*

Under a fixed evaluation protocol, model outputs induce a probability measure over answer identities. The *Answer-Basin Representation Hypothesis* connects this measure's statistics to hidden representations, probing, and steering to explain the geometry in Figure 1 and Appendix D.

2.1 ANSWER MEASURES AND ANSWER BASINS

Throughout, an evaluation instance fixes a question q, a countable answer space $\mathcal{A}_q$, and an answer-measure protocol. The protocol maps model outputs to a probability measure μ_q on $\mathcal{A}_q$; we suppress the protocol in the notation and write $m_q(a) = \mu_q(\{a\})$ for answer mass. We construct this measure using either a generation protocol or a finite-choice protocol.

Under a generation protocol, let $P_q = P_\theta(\cdot \mid q)$ be the continuation measure on $\mathcal{Y}_q$, and let the measurable map $\phi_q : \mathcal{Y}_q \to \mathcal{A}_q$ assign each continuation its answer identity. The fiber $\mathcal{B}_q(a) = \phi_q^{-1}(\{a\})$, the answer basin of a, contains all continuations producing it. The pushforward answer

measure and basin masses are

$$\mu_q = (\phi_q)_{\#} P_q, \qquad m_q(a) = \mu_q(\{a\}) = P_q(\mathcal{B}_q(a)) \,. \tag{1}$$

The map ϕ_q therefore converts probability over token trajectories into probability over answer identities. Basin mass aggregates all trajectories in a fiber and defines the model's answer-level probability landscape (Wang et al., 2023; Farquhar et al., 2024).

A finite-choice protocol constructs an answer measure directly on the listed answers. The elements of $\mathcal{A}_q$ are the complete candidate answers; option letters serve as their labels under a fixed presentation. In our letter-based evaluations, the model scores candidates through their assigned letters. We vary the letter–answer assignments and map scores back to the corresponding answers. A fixed scoring-and-aggregation rule converts the model-derived scores into nonnegative masses summing to one, indexed by the underlying answers. This construction defines a candidate measure on $\mathcal{A}_q$. Section 2.3 specifies the generation and finite-choice constructions used in our experiments. The answer-level mass, concentration, and signed-contrast quantities below are defined for either fixed construction.

For $A, A' \overset{\text{iid}}{\sim} \mu_q$, the collision concentration (Simpson, 1949) summarizes how the answer measure is concentrated and equals the expected mass of an answer drawn from it:

$$C_2(\mu_q) := \Pr(A = A' \mid q) = \sum_{a \in \mathcal{A}_q} m_q(a)^2 = \mathbb{E}_{A \sim \mu_q}[m_q(A)]. \tag{2}$$

For two positive-mass answers, define the log-mass contrast

$$\Delta_q^{\log m}(a, b) := \log \frac{m_q(a)}{m_q(b)}. \tag{3}$$

2.2 THE *Answer-Basin Representation Hypothesis*

The preceding section defines the answer measure and its two constructions. We examine hidden states before answer generation at the final prompt token and after answer generation at the final answer token specified by the protocol. These positions are the question and answer boundaries, respectively. Let $h_Q(q)$ denote the question-boundary state and $H_A(q, a)$ the answer-state random variable for answer a under the fixed protocol. Its generation and finite-choice constructions are given in Appendix A.1.

Definition 2.1 (Answer-conditioned representation)**.** *For each positive-mass answer, define*

$$\bar{h}_q(a) := \mathbb{E}[H_A(q, a)] \,. \tag{4}$$

We now ask whether shared linear directions reveal answer mass in these means and concentration in question-boundary states.

Answer-Basin Representation Hypothesis. *Fix a model, a residual-stream layer, an answer-measure construction, and a state-extraction protocol. Hidden-state differences are consistently oriented with signed changes in statistics of the same answer measure at the question and answer boundaries. There exist directions $w_Q, w_A \in \mathbb{R}^d$, shared across questions, such that:*

At the question boundary. *For any questions q, q' with $C_2(\mu_q) \neq C_2(\mu_{q'})$,*

$$\big[C_2(\mu_q) - C_2(\mu_{q'})\big] w_Q^\top \big[h_Q(q) - h_Q(q')\big] > 0, \tag{5}$$

At the answer boundary. *For any question q and positive-mass answers a, b with $m_q(a) \neq m_q(b)$,*

$$\Delta_q^{\log m}(a, b) w_A^\top \big[\bar{h}_q(a) - \bar{h}_q(b)\big] > 0. \tag{6}$$

In other words, increasing concentration corresponds to positive displacement along w_Q, while increasing answer mass corresponds to positive displacement along w_A.

The *Answer-Basin Representation Hypothesis* describes an idealized linear ordering of hidden representations by answer-measure statistics. We learn shared directions from finite data to test how closely models realize this structure and how well these directions generalize across questions.

Answer-measure transport under intervention. The readout hypothesis motivates a corresponding prediction for intervention. An answer-mass-oriented direction v averages within-question differences of the answer-conditioned means, each pointing toward the higher-mass answer (Appendix A.2). At layer ℓ, we apply $h_\ell^{(\alpha)} = h_\ell + \alpha v$ and reapply the same answer-measure protocol to the intervened model outputs. Denote the resulting measure by $\mu_{q,\alpha}^{(v)}$. Under generation, this is the pushforward of the intervened continuation measure; under finite-choice scoring, it is obtained by applying the same scoring-and-aggregation rule to the intervened candidate scores. For answers with positive baseline and intervened masses, define

$$m_{q,\alpha}^{(v)}(a) := \mu_{q,\alpha}^{(v)}(\{a\}), \quad G_{q,v}^{(\alpha)}(a) := \log \frac{m_{q,\alpha}^{(v)}(a)}{m_q(a)}. \tag{7}$$

For such a direction, we posit an additional transport hypothesis: positive steering gives initially higher-mass answers greater relative amplification. For answers a, b with positive baseline and intervened masses and $m_q(a) \neq m_q(b)$,

$$\boxed{\Delta_q^{\log m}(a,b)\left[G_{q,v}^{(\alpha)}(a) - G_{q,v}^{(\alpha)}(b)\right] > 0, \qquad \alpha > 0.} \tag{8}$$

Positive steering is predicted to increase relative preference for the initially higher-mass answer.

Concept-level orientation. An external attribute assigns each answer a label $C_q : \mathcal{A}_q \to \{0, 1\}$, such as correctness or truthfulness. For a pair (a, b), labels and mass ordering align when $[C_q(a) - C_q(b)]\Delta_q^{\log m}(a, b) > 0$, with greater mass on the attribute-positive answer. A negative product indicates conflict, with greater mass on the attribute-negative answer. Answers with the same label can still have different basin masses. On binary questions, positive steering therefore favors the desired answer under alignment and its alternative under conflict.

For generation protocols, the measure-change identities are developed in Appendix A.3.

2.3 Instantiating Answer Measures

We instantiate both answer-measure constructions: repeated sampling estimates the measure induced by generation, while candidate scoring defines a measure over listed options.

Sampling-based estimation. Under a fixed generation protocol, draw $K \geq 2$ independent continuations and let $N_q(a)$ count those assigned to answer a by ϕ_q. We estimate mass and concentration by

$$\widehat{m}_q(a) := \frac{N_q(a)}{K}, \qquad \widehat{C}_{2,q} := \frac{\sum_a N_q(a)\big(N_q(a) - 1\big)}{K(K-1)}. \tag{9}$$

The second quantity is the fraction of agreeing sample pairs and is unbiased for $C_2(\mu_q)$.

Candidate-answer measures. Given a question q and a finite candidate-answer set $\mathcal{A}_q$, we use a specified choice-response protocol: the model reads the question and candidate answers and outputs an option label. For each option permutation, we map the label outputs back to the corresponding answers. The protocol's fixed rules for scoring, aggregation, and normalization then yield answer masses $\widetilde{m}_q(a)$ satisfying

$$\widetilde{m}_q(a) \geq 0, \qquad \sum_{a \in \mathcal{A}_q} \widetilde{m}_q(a) = 1.$$

These masses define the answer measure under this protocol and its concentration:

$$\widetilde{\mu}_q := \sum_{a \in \mathcal{A}_q} \widetilde{m}_q(a)\, \delta_a, \qquad \widetilde{C}_{2,q} := C_2(\widetilde{\mu}_q) = \sum_{a \in \mathcal{A}_q} \widetilde{m}_q(a)^2. \tag{10}$$

Here, δ_a is the unit point mass at answer a. Labels from different permutations map to the same candidate-answer set, so the measure is indexed by answer identity. Option presentation, score aggregation, and normalization jointly determine this answer measure; the specific protocols are given in Appendix B. Hats denote sampling estimates under a generation protocol; tildes denote quantities defined by a candidate-scoring protocol.

Table 1: **Probing answer concentration.** Linear scores before answer generation track future answer concentration on held-out questions.

Model	GSM8K		BBQ		SafetyBench	
	Order Acc. ↑	ρ ↑	Order Acc. ↑	ρ ↑	Order Acc. ↑	ρ ↑
Qwen2.5-7B base	74.84	0.676	70.52 ± 1.40	0.582	67.95 ± 1.64	0.515
Qwen2.5-14B base	77.60	0.720	73.92 ± 1.86	0.657	67.95 ± 0.86	0.521
Gemma-3-4B-PT	75.45	0.700	63.48 ± 3.92	0.384	70.42 ± 1.18	0.571
Gemma-3-12B-PT	78.20	0.745	74.48 ± 0.86	0.677	63.42 ± 0.82	0.387

3 Probing Answer-Basin Representations

The *Answer-Basin Representation Hypothesis* predicts that shared linear directions reflect concentration ordering across questions before answer generation and basin-mass ordering within a question after answer generation. We test these predictions on held-out questions using the sampling estimates and candidate measures in Section 2.3. For readouts after answer generation, we examine performance across datasets, model families, and parameter scales, its relation to training-pair mass gaps, and whether readout persists when the paired training answers share the same concept label.

For $w = w_A$, Eq. 6 gives higher mean probe scores to higher-mass answers within each question; the linearity derivation is given in Appendix A.1.

3.1 Experimental Setup and Metrics

Before answer generation, probes predict concentration from the final prompt-token state on GSM8K (Cobbe et al., 2021), BBQ (Parrish et al., 2022), and SafetyBench (Zhang et al., 2024). Each probe uses 500 training questions; targets are Eq. 9 for repeated generations or $\widetilde{C}_{2,q}$ for candidate measures derived from option scores. We report pairwise ordering accuracy (Order Acc.) and Spearman correlation (ρ); see Appendix B.1 for details.

After answer generation, probes use the hidden state at the last token of each candidate answer appended to the question and learn basin-mass ordering on ARC-Challenge (Clark et al., 2018) and SafetyBench. Permutation-aggregated option preferences define $\widetilde{\mu}_q$. Low, Medium, and High vary the training-pair mass gap, $\delta_q(a, b) := |\widetilde{m}_q(a) - \widetilde{m}_q(b)|$, to test its relation to readout generalization. Wrong-only uses pairs of answers that are both incorrect according to the dataset's gold answer but have different basin masses, using their ordering as supervision. Basin ordering averages probe–mass ranking agreement over all candidate pairs within each test question, then across questions. Top-1 correctness evaluates the highest-scoring answer against the gold label. Baseline definitions, training budgets, and evaluation splits are given in Appendix B.2.

3.2 Before Answer Generation: Answer Concentration

Eq. 2 identifies the concentration of the future answer distribution with the expected basin mass of the answer eventually realized. We therefore test whether a linear readout from the hidden state before answer generation preserves concentration ordering across questions, as posited in Eq. 5.

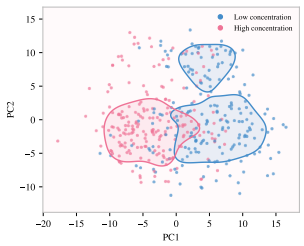


Figure 2: PCA projection of the question-final hidden states for 200 high- and 200 low-concentration GSM8K questions (Qwen2.5-7B). Contours enclose 50% density regions.

Concentration is readable before answer generation. Table 1 evaluates concentration readout on held-out questions. Order Acc. measures how often probe scores correctly order two test questions by estimated answer concentration. It exceeds chance across all evaluated models and tasks. The moderate to strong positive Spearman correlations (ρ) show that higher probe scores tend to correspond to higher estimated concentration across the test set. Both metrics support the prediction in Eq. 5 that a shared linear readout captures concentration ordering before answer generation.

Figure 2 visualizes question-final representations from Qwen2.5-7B for 200 high- and 200 low-concentration GSM8K questions, grouped by estimated answer concentration $\widehat{C}_2$. Contours enclose 50% density regions. The two groups have different distributions in the PCA projection.

3.3 After Answer Generation: Answer-Basin Mass

Alignment and conflict reverse concept predictions. We examine how the relationship between correctness labels and basin mass in the training data shapes concept predictions on held-out questions. We train correctness probes on answer pairs with the largest positive (Align) or most negative (Conflict) signed mass gaps, $\widetilde{m}_q(a_{\text{correct}}) - \widetilde{m}_q(a_{\text{wrong}})$, keeping correct answers as the positive label in both conditions. On the same held-out questions, Align exceeds the mean-logp baseline and Conflict falls below it in all twelve model–dataset combinations on ARC-C, BBQ, and SafetyBench (Figure 3). The relationship between basin mass and correctness in training determines the probe's concept orientation. Additional GSM8K results are reported in Appendix E.

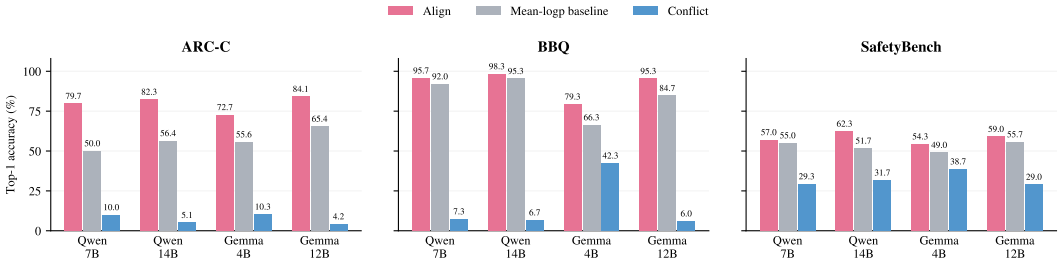


Figure 3: **Probing under label–basin alignment and conflict.** With the same correctness supervision, answer-selection accuracy is above the mean-logp baseline under aligned training and below it under conflict training.

Eq. 6 predicts that mean answer representations admit a shared linear ordering by basin mass. We next examine their geometry on GSM8K using Qwen2.5-7B, averaging the final hidden states of responses that yield the same answer. Figure 4 compares correct and wrong answer groups across decreasing estimated basin-mass gaps. Each point is an answer group's mean representation; g denotes the median estimated log mass ratio in each group of questions, and contours enclose 50% density regions. Larger-gap groups show clearer separation in the PCA projection; as the gap narrows, the distributions increasingly overlap.

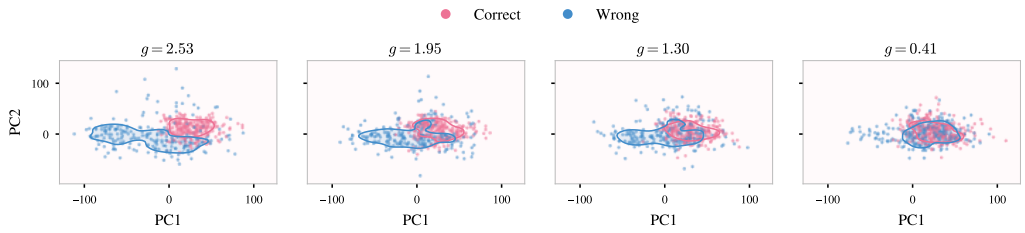


Figure 4: **Answer-group geometry across basin-mass gaps.** Qwen2.5-7B on GSM8K: each point is an answer group's mean representation; g is the median estimated log mass ratio in each group of questions. Contours enclose 50% density regions.

We then test basin-mass ordering directly by training linear probes on answer-pair mass comparisons. Table 2 evaluates their ordering and answer-selection accuracy on ARC-Challenge across model families and training-pair gaps. SafetyBench results are reported in Appendix F, Table 4.

Clearer probability contrasts support readout. Across the evaluated datasets, model families, and parameter scales, basin-ordering accuracy exceeds chance, supporting linearly readable candidate-mass information after answer generation. Larger-gap training conditions generally yield higher basin-ordering accuracy and Top-1 correctness. This fits the intuition that a more decisive preference between training answers provides a clearer probability contrast for learning a shared direction. The accompanying increase in Top-1 indicates that, on these tasks, improved basin readout also produces selections more consistent with gold answers.

Table 2: **Probing basin mass across training conditions.** Larger training-pair mass gaps generally support more accurate basin readout and answer selection. Probes trained only on incorrect answers still recover basin-mass ordering on held-out questions.

Model	Metric (%)	Mean logp	Basin-supervised linear probes			
			Low	Medium	High	Wrong-only
ARC-Challenge						
Qwen2.5-7B base	Basin ordering	—	69.85 ± 2.33	77.77 ± 0.56	80.43 ± 0.27	80.00 ± 0.31
	Top-1 correctness	50.04	55.50 ± 4.81	73.72 ± 1.47	79.36 ± 0.63	78.03 ± 0.54
Qwen2.5-14B base	Basin ordering	—	72.10 ± 2.05	79.04 ± 0.48	82.17 ± 0.31	81.62 ± 0.29
	Top-1 correctness	56.39	59.80 ± 4.10	76.94 ± 1.28	82.06 ± 0.72	80.95 ± 0.68
Gemma-3-4B-PT	Basin ordering	—	62.98 ± 3.72	80.80 ± 0.11	81.19 ± 0.12	79.84 ± 0.27
	Top-1 correctness	55.62	42.78 ± 5.71	71.88 ± 0.41	73.03 ± 0.44	69.70 ± 0.72
Gemma-3-12B-PT	Basin ordering	—	66.42 ± 3.15	83.58 ± 0.06	83.39 ± 0.10	82.60 ± 0.23
	Top-1 correctness	65.41	48.36 ± 5.12	85.70 ± 0.44	86.56 ± 0.20	83.28 ± 0.22

3.4 Probing without Concept Supervision

Wrong-only trains probes on pairs of incorrect answers with different basin masses, using their ordering as supervision. These probes still rank answers in held-out questions by $\widetilde{m}_q$ (Table 2 and Appendix F, Table 4), showing that the signal persists without a contrast in external concept labels. This transfer supports a shared scoring rule that assigns higher scores to answers with higher basin mass within each question and extends this relation to candidates in new questions.

4 Steering Answer-Basin Representations

Eq. 8 predicts that positive basin-oriented steering increases the relative preference for initially higher-mass answers. We test how this response affects answer selection when basin ordering aligns or conflicts with concept labels, then examine how source mass separation shapes the steering direction.

Experimental setup and metrics. For each dataset, we construct a label-defined steering vector by averaging correct-minus-wrong answer-conditioned representation contrasts, $v = \mathbb{E}_q[\bar{h}_q(a_T) - \bar{h}_q(a_O)]$, where $C_q(a_T) = 1$ and $C_q(a_O) = 0$. Representations are extracted at the answer boundary. We hold v fixed across alignment and conflict, defined by the positive and negative signs of $\Delta_q^{\log m}(a_T, a_O)$, respectively. We measure the behavioral response as the change in length-normalized answer log-probability. For source geometry, we relate the signed projection of $\bar{h}_q(a_T) - \bar{h}_q(a_O)$ to the log-mass contrast $\Delta_q^{\log m}(a_T, a_O)$ and compare individual and aggregate contrast norms across basin-mass gaps. Appendix B.4 gives the detailed protocols.

4.1 Why Does Steering Succeed or Fail?

We test Gemma-3-4B-PT on BBQ and SafetyBench. We add αv to the residual stream after block 19 for BBQ and block 16 for SafetyBench, at the final token of the `Answer:` prefix and every subsequent answer-prediction position.

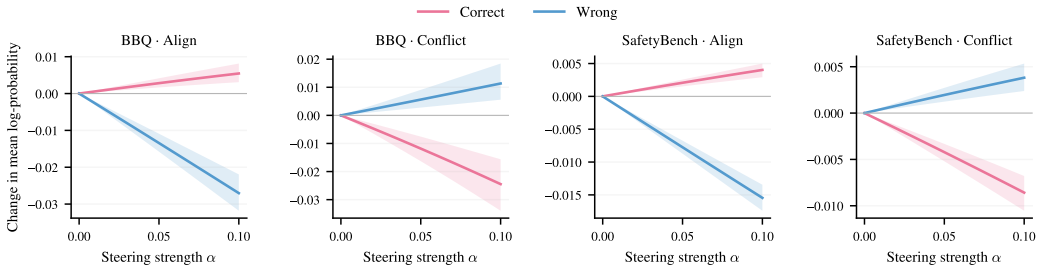


Figure 5: **A fixed label-defined steering vector produces opposite answer preferences.** Changes in length-normalized answer log-probability on Gemma-3-4B-PT under alignment and conflict. Shaded bands denote 95% bootstrap intervals.

Target alignment shapes steering. Figure 5 shows that the correct answer receives the larger behavioral response under alignment, whereas the incorrect answer receives the larger response under conflict. The relative effect reverses on both datasets despite the fixed vector and unchanged label orientation. Thus, a direction constructed to favor correct answers can favor incorrect answers when they carry greater initial answer mass. This pattern connects the behavioral effect of label-defined steering to the answer measure.

The result complements the probing dissociation in Section 3.3: source label–basin alignment shapes learned concept predictions, while target label–basin alignment shapes the relative effect of a fixed intervention. Appendix C reports the corresponding answer-letter logit responses using the same vectors and question pairs.

4.2 How Does Basin Separation Shape Steering Vectors?

The basin-dependent reversal in Section 4.1 motivates examining how source basin structure shapes the vectors themselves.

Source alignment changes projected steering directions. Using the Align and Conflict source pools from Section 3.3, we select 80 question pairs per condition and average their correct-minus-wrong answer-conditioned representation contrasts. Thus, both vectors retain the same label orientation. Figure 6 shows their joint PCA projections for Qwen2.5-7B base and Gemma-3-4B-PT on ARC-C, BBQ, and SafetyBench. Switching the sign of the source log-mass contrast produces opposing projected directions while preserving the correct–wrong labels. This connects source basin alignment to the geometry of label-defined steering vectors.

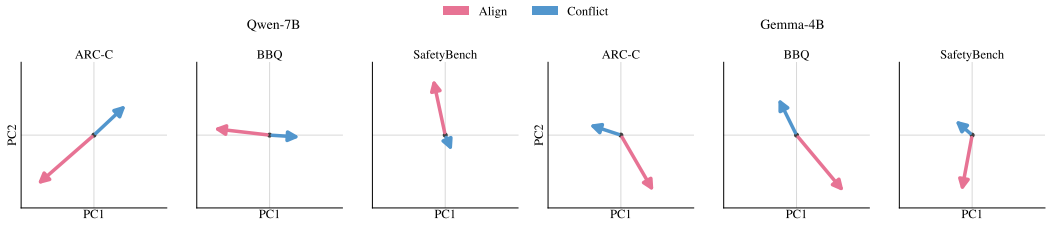


Figure 6: **Steering-vector projections under source alignment and conflict.** Each vector averages 80 correct-minus-wrong answer-conditioned representation contrasts. Within each panel, PCA is fitted jointly to the two sets of contrasts; arrows show the unnormalized vectors projected onto PC1 and PC2 from a common origin.

We next examine how gap magnitude shapes individual contrasts and their aggregate. Using Qwen2.5-7B answer-letter states on BBQ and SafetyBench, we compare signed representation contrasts with answer log-mass contrasts.

Larger gaps, more coherent contrasts. Signed projection grows with the answer log-mass contrast (Figure 7). Grouping source questions by absolute basin-mass gap reveals longer contrasts and more consistent directions at larger gaps. These contrasts accumulate into larger aggregate vectors, linking basin separation to both displacement magnitude and directional coherence. Sample-wise normalization preserves the aggregate direction, so magnitude weighting has little effect on its orientation.

Vector magnitude contributes to steering strength. Raw averaging gives longer answer-conditioned representation contrasts greater weight. Vectors constructed from larger-gap sources produce stronger mean transport effects, and matching aggregate-vector norms reduces much of this difference. Together with the directional-coherence results, these observations connect source basin separation to the geometry of the constructed intervention. Basin structure thus enters the account at both ends of steering: source separation is reflected in vector magnitude and coherence, while target alignment predicts which answer receives the greater relative gain.

5 Related Work and Limitations

Concept Representations and Information Geometry. Park et al. (2024) take concept variables and their linear representations as a starting point for connecting readout, geometry, and steering.

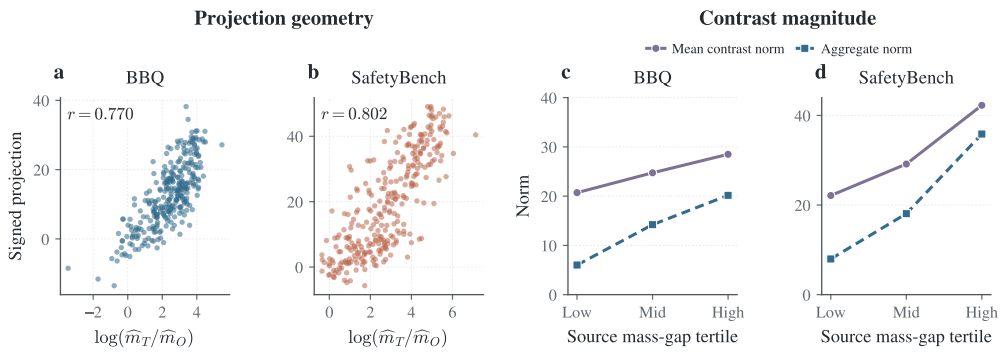


Figure 7: **Basin separation and steering-vector geometry.** Signed representation projections increase with the answer log-mass contrast. Larger mass gaps are associated with longer, more directionally coherent contrasts that form larger aggregate steering vectors.

Further developments are discussed in Appendix G. The *Answer-Basin Representation Hypothesis* takes the answer measure as the organizing object of answer-related representations and interprets concept-related linear structure through its statistics. Within this account, the alignment between external labels and the answer measure determines when statistical directions exhibit concept-consistent readout and control effects.

Answer Distributions and Semantic Uncertainty. Self-consistency aggregates reasoning paths by answer for answer selection (Wang et al., 2023); semantic entropy aggregates generations by meaning to estimate uncertainty (Kuhn et al., 2023; Farquhar et al., 2024). Semantic entropy probes establish a predictive relationship between these statistics and hidden states (Kossen et al., 2024). Grouping outputs by answer or meaning connects these methods to answer-level probability measures. The *Answer-Basin Representation Hypothesis* links concentration readout before generation and within-question basin-mass ordering after generation through a common answer measure.

Concept Probing and Activation Steering. Hidden-state probes predict truthfulness and correctness (Azaria & Mitchell, 2023; Marks & Tegmark, 2024; Cencerrado et al., 2025). Analyses of disagreement between probe predictions and model outputs examine the interpretation of such readouts (Liu et al., 2023). Activation steering uses attribute or contrastive directions to modify behavior (Turner et al., 2023; Li et al., 2023; Zou et al., 2023; Rimsky et al., 2024; Arditi et al., 2024). We examine how the relationship between concept labels and answer mass shapes the predictions learned by probes and the effects of fixed steering vectors.

Limitations. In open-ended question answering, estimating the answer measure through free-form sampling faces finite-sample error and uncertainty in grouping responses by answer. Many of our experiments therefore use candidate-answer measures defined over fixed candidate-answer spaces. Extending these results to a broader range of open-ended tasks requires more accurate methods for answer grouping and answer-measure estimation.

6 Conclusion

The *Answer-Basin Representation Hypothesis* connects the semantics of answer-level linear directions to the probability structure of the model's answer basins. It brings linear probing and activation steering into a common account: hidden states expose statistics of the answer measure, and interventions reshape that measure. The complementary probing and steering dissociations support a relational interpretation of concept semantics, in which a direction's concept-level predictions and effects depend on how external labels align with endogenous answer-basin structure. Together, these findings place the model's answer distribution at the center of understanding what linear directions represent and how they influence behavior.

AI USE STATEMENT

In this work, we used generative AI tools from OpenAI (ChatGPT/Codex) and Anthropic (Claude/Claude Code) to help implement experiment code and to automate the execution of our experiments across multiple random seeds. We also used these tools for retrieval and discovery, such as finding related work, and for polishing the English writing to improve readability. All AI-assisted work underwent author review: we tested the AI-assisted code, checked the automated runs before reporting their results, checked every retrieved work against its original source before citing it, and verified the technical accuracy of the polished text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyzes the hidden representations of publicly released open-weight language models from the Qwen2.5 and Gemma-3 families on public benchmarks: GSM8K, ARC-Challenge, BBQ, and SafetyBench. It involves no human subjects, crowdsourced annotation, or personal data, and we use every model and dataset under its license and terms of use. BBQ and SafetyBench include socially biased and unsafe answer options by design; we use them only to measure model representations and answer preferences. Our findings bear on the reliable use of interpretability tools for bias and safety. Probes trained to detect such attributes can reverse their predictions when concept labels conflict with answer mass, and a steering vector built to favor correct answers can raise the relative preference for incorrect answers on BBQ and SafetyBench. The same mechanism could shift an open-weight model toward undesirable answers. Exploiting it, however, requires white-box access to activations and uses only standard activation addition, so it adds little capability beyond existing steering methods.

REPRODUCIBILITY STATEMENT

Section 2 defines answer basins, answer measures, and the statistics that our probes and steering vectors target. Appendix A states the underlying measure-theoretic assumptions, derives the relation between probe scores and answer-conditioned representations, and proves Proposition A.1. Section 2.3 and Appendix B specify how we estimate answer measures and run each experiment, including prompts, sampling and answer-parsing rules, option permutations, extraction layers, intervention positions and strengths, regularization, and data splits. The same appendix gives the number of source-set selections and bootstrap resamples behind the reported standard deviations and confidence intervals. Appendix D describes the answer-pair selection and PCA procedures behind the geometry figures. All models and datasets we use are publicly available. The supplementary material contains anonymized implementations of the core methods used in our experiments. We will release the complete experimental code publicly upon publication.

REFERENCES

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

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Iván Vicente Moreno Cencerrado, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi. No answer needed: Predicting LLM answer accuracy from question-only linear probes. *arXiv preprint arXiv:2509.10625*, 2025. URL https://arxiv.org/abs/2509.10625.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

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

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21879–21911. PMLR, 2024. URL https://proceedings.mlr.press/v235/jiang24d.html.

Olav Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, 3 edition, 2021. doi: 10.1007/978-3-030-61871-1. URL https://link.springer.com/book/10.1007/978-3-030-61871-1.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024. URL https://arxiv.org/abs/2406.15927.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09664.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://arxiv.org/abs/2306.03341.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4797. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.291. URL https://aclanthology.org/2023.emnlp-main.291/.

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

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.01506.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.15293.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

E. H. Simpson. Measurement of diversity. *Nature*, 163:688, 1949. doi: 10.1038/163688a0. URL https://www.nature.com/articles/163688a0.

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*, 2023. URL https://arxiv.org/abs/2308.10248.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. SafetyBench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15537–15553. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.830. URL https://aclanthology.org/2024.acl-long.830/.

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 Measure-Theoretic Details

Let $(\mathcal{Y}_q, \mathcal{F}_q, P_q)$ be the continuation probability space, $(\mathcal{A}_q, \mathcal{G}_q)$ a countable answer space, and $\phi_q : (\mathcal{Y}_q, \mathcal{F}_q) \to (\mathcal{A}_q, \mathcal{G}_q)$ measurable (Kallenberg, 2021). The fibers $\mathcal{B}_q(a) = \phi_q^{-1}(\{a\})$ form a measurable partition of $\mathcal{Y}_q$, so $\sum_a m_q(a) = 1$. For $\Delta_q = \{(a, a) : a \in \mathcal{A}_q\}$, Eq. 2 follows:

$$(\mu_q \otimes \mu_q)(\Delta_q) = \sum_a m_q(a)^2 = \int_{\mathcal{A}_q} m_q(a)\, \mathrm{d}\mu_q(a). \tag{11}$$

A.1 Basin-Conditioned Hidden-State Measures

Let $\tau_q(y)$ be the final token position of the answer in continuation y, as specified by the fixed evaluation protocol. The hidden state after answer generation is defined by

$$h_A(q, y) := h_{\tau_q(y)}(q, y). \tag{12}$$

The joint hidden-state and answer measure is

$$\Lambda_q := \big(h_A(q, \cdot), \phi_q\big)_{\#} P_q. \tag{13}$$

For $m_q(a) > 0$, its answer-conditioned hidden-state measure is

$$\nu_q^a(D) := \frac{\Lambda_q(D \times \{a\})}{m_q(a)}, \qquad D \in \mathcal{B}(\mathbb{R}^d). \tag{14}$$

For generation protocols, $H_A(q, a)$ has distribution ν_q^a, so its first moment is $\bar{h}_q(a) = \int h\, \mathrm{d}\nu_q^a(h)$. For finite-choice protocols, $H_A(q, a)$ is the state extracted while scoring candidate a, with any averaging over presentation variants specified by the fixed state-extraction protocol. Both constructions yield Equation 4.

A linear probe scores individual hidden states after answer generation. To interpret these scores at the level of the answer measure, we average over the states associated with each answer under the

fixed extraction protocol. Using $H_A(q, a)$ from Section 2.2, for $f_w(h) = w^\top h + \beta$ and positive-mass answers a, b whose associated states have finite first moments, linearity gives

$$\mathbb{E}[f_w(H_A(q, a))] - \mathbb{E}[f_w(H_A(q, b))] = w^\top \big[\bar{h}_q(a) - \bar{h}_q(b)\big]. \tag{15}$$

For $w = w_A$, Eq. 6 gives higher mean scores to higher-mass answers within each question.

A.2 Basin-Oriented Direction Construction

Let $\mathcal{S}$ be a source distribution over question–answer triples (q_s, a_s, b_s) with positive, unequal basin masses and integrable representation differences. The ideal basin-oriented displacement is

$$v_{\mathcal{S}} := \mathbb{E}_{(q_s,a_s,b_s)\sim\mathcal{S}}\big[\operatorname{sgn}(\Delta_{q_s}^{\log m}(a_s, b_s))\big(\bar{h}_{q_s}(a_s) - \bar{h}_{q_s}(b_s)\big)\big]. \tag{16}$$

Eq. 6 implies $w_A^\top v_{\mathcal{S}} > 0$, so the source construction has a positive component along the shared basin-mass direction. The remaining transformer layers and autoregressive rollout map the translated state nonlinearly into a new continuation measure.

A.3 Answer-Level Change of Measure

Assume $P_{q,\alpha}^{(v)} \ll P_q$. Measurability of ϕ_q gives $\mu_{q,\alpha}^{(v)} \ll \mu_q$. Define

$$r_{q,\alpha}^{(v)} := \frac{\mathrm{d}P_{q,\alpha}^{(v)}}{\mathrm{d}P_q}, \qquad g_{q,\alpha}^{(v)} := \frac{\mathrm{d}\mu_{q,\alpha}^{(v)}}{\mathrm{d}\mu_q}. \tag{17}$$

Proposition A.1 (Answer-level change of measure)**.** *For μ_q-almost every answer a,*

$$\frac{\mathrm{d}\mu_{q,\alpha}^{(v)}}{\mathrm{d}\mu_q}(a) = \mathbb{E}_{P_q}\left[\frac{\mathrm{d}P_{q,\alpha}^{(v)}}{\mathrm{d}P_q}(Y) \;\middle|\; \phi_q(Y) = a\right]. \tag{18}$$

Proof. For every $B \in \mathcal{G}_q$,

$$\begin{aligned}
\mu_{q,\alpha}^{(v)}(B) &= P_{q,\alpha}^{(v)}\big(\phi_q^{-1}(B)\big) \\
&= \int_{\phi_q^{-1}(B)} r_{q,\alpha}^{(v)}(y)\,\mathrm{d}P_q(y) \\
&= \int_B \mathbb{E}_{P_q}\Big[r_{q,\alpha}^{(v)}(Y) \mid \phi_q(Y) = a\Big]\,\mathrm{d}\mu_q(a).
\end{aligned}$$

The conditional expectation is the Radon–Nikodym derivative of the intervened answer measure with respect to the baseline answer measure. □

For every positive-mass atom,

$$g_{q,\alpha}^{(v)}(a) = \frac{\mu_{q,\alpha}^{(v)}(\{a\})}{\mu_q(\{a\})}, \tag{19}$$

whose logarithm, when the intervened mass is also positive, is the basin amplification in Equation 7.

A.4 Answer Odds and Evaluation Attributes

By the definition of $G_{q,v}^{(\alpha)}$, steering updates pairwise answer log odds as

$$\log \frac{m_{q,\alpha}^{(v)}(a)}{m_{q,\alpha}^{(v)}(b)} = \Delta_q^{\log m}(a, b) + \Big[G_{q,v}^{(\alpha)}(a) - G_{q,v}^{(\alpha)}(b)\Big]. \tag{20}$$

Under the transport hypothesis in Eq. 8, positive steering increases the log odds of a over b when $m_q(a) > m_q(b)$.

For the evaluation attribute C_q defined in Section 2.2, its probability and intervention-induced change are obtained by integrating over the answer measure:

$$\pi_q(C = 1) = \int C_q\,\mathrm{d}\mu_q, \qquad \Delta\pi_q^{(v,\alpha)}(C) = \int C_q\,\mathrm{d}\big(\mu_{q,\alpha}^{(v)} - \mu_q\big). \tag{21}$$

B Experimental Details for Answer-Basin Analysis

B.1 Concentration Probes Before Answer Generation

For each GSM8K question, we estimate future answer concentration from $K = 64$ independent generations using Eq. 9. A linear probe maps the final prompt-token state to $\widehat{C}_{2,q}$ and is evaluated on 200 question-disjoint GSM8K questions, with the same generation template and split across models.

For BBQ and SafetyBench, we construct $\widetilde{\mu}_q$ by exhaustively permuting option-to-letter assignments, mapping the letter logits back to the original options, and applying softmax to their averaged logits. The target is $\widetilde{C}_{2,q} = \sum_a \widetilde{m}_q(a)^2$, the concentration of this candidate measure. BBQ uses disambiguated questions with the two substantive options; SafetyBench uses English questions with their original two, three, or four options. Each probe uses the state before answer generation from an option-free prompt with standardized ridge regression and validation-selected regularization. We evaluate on 5,512 held-out BBQ questions and 2,667 held-out SafetyBench questions. Table 1 reports their means over five training sets and sample standard deviations for Order Acc.

B.2 Probes After Answer Generation and Mean-Logp Baseline

For pairs with unequal masses, random ranking is correct with probability $1/2$. For a question with n_q candidates and one correct answer, random Top-1 accuracy is $1/n_q$. Mean logp separately ranks individual candidate strings by length-normalized log probability.

Matched Low, Medium, and High conditions share a source-question budget and a question-disjoint test set. On SafetyBench, each Low, Medium, and High fit for Qwen2.5-7B and Qwen2.5-14B uses 100 source questions, with one answer pair per question. The Gemma Low, Medium and High conditions use 100 source questions. Wrong-only uses 100 training questions, each contributing its highest- and lowest-mass incorrect answers under the candidate measure.

For Table 2, each model scores a candidate independently under `Question: {question}` followed by a newline and `Answer: {candidate}`. We average the conditional log probabilities of the candidate continuation tokens, including its leading space. The question prefix and special tokens are excluded from the average. Tokenization retains the checkpoint's default BOS behavior, and the full prompt is checked to preserve the prefix token sequence. Top-1 answer selection uses the highest scalar score, with no fitted parameters.

Evaluation uses all 1,165 four-choice ARC-Challenge test questions. The baseline reports one Top-1 accuracy per fixed test set. Probe accuracies in Table 2 are means and sample standard deviations over five source-set selections.

B.3 GSM8K Instruct–CoT Alignment and Conflict

We additionally test the alignment–conflict comparison under free-form chain-of-thought generation on GSM8K (Cobbe et al., 2021). We use `Qwen/Qwen2.5-3B-Instruct` with its chat template and an instruction to reason step by step before placing the final numerical answer in `\boxed{}`. For each question, $K = 128$ continuations are sampled at temperature 0.7 with a maximum of 512 generated tokens. A fixed numerical parser maps each continuation to its final boxed answer, and answer counts provide the empirical basin masses.

Both training conditions use the same correctness labels. In the aligned condition, the correct-answer basin is strictly larger than every incorrect basin; one correct trajectory is paired with one trajectory from the smallest incorrect basin. In the conflict condition, the correct-answer basin is strictly smaller than every incorrect basin; one correct trajectory is paired with one trajectory from the largest incorrect basin. The aligned pool is downsampled to match the size of the conflict pool. Each probe uses the complete layer-24 hidden state at the final numerical-answer token, followed by feature standardization and L_2-regularized logistic regression with $C = 0.001$.

Evaluation uses the same 188 question-disjoint test questions for both probes. For each test question, the probe scores all 128 sampled continuations; Top-1 is correct when the highest-scoring continuation gives the gold answer. Table 3 shows that aligned training yields higher held-out correctness than conflict training under otherwise identical concept supervision.

B.4 Fixed-Vector Steering and Basin Geometry

Label-defined option-text steering. For Figure 5, each input is `Question: {question}\nAnswer: {option}`. We extract the final option-token state and aver-

Table 3: **GSM8K Instruct–CoT alignment and conflict.** Mean basin counts describe the paired training trajectories; both probes are evaluated on the same 188 held-out questions.

Training condition	Correct basin	Wrong basin	Top-1 (%)
Aligned	108.22	1.00	78.72
Conflict	26.22	101.78	70.74

age the correct-minus-wrong contrasts without normalization, using 100 aligned BBQ source questions and 300 aligned SafetyBench source questions. Evaluation uses 100 aligned and 100 conflict BBQ questions, and 300 of each condition for SafetyBench. The candidate-answer measure determines the sign of each correct–wrong mass gap. All original options are retained, and complete-option scores average token log-probabilities over the option text without EOS. Intervention is applied after blocks 19 and 16, respectively, at the final `Answer:` token and subsequent answer-prediction positions, for $\alpha = 0, 0.01, \ldots, 0.10$. Pointwise 95% intervals use 2,000 context-group bootstrap resamples (Efron, 1979).

Answer-letter geometry. For Figure 7, we use Qwen2.5-7B base and 300 source questions per dataset. We present both options, append each answer letter, and extract block-20 states, averaging semantically matched contrasts over AB/BA orders. The candidate measure averages binary choice probabilities across AB/BA orders after normalization within each order.

Projection estimates. For Figure 7, we form correct-minus-wrong contrasts from the same block-20 answer-letter states and use the AB/BA-averaged baseline probabilities. Within each dataset, we split the 300 source questions into five folds by context group. For each held-out fold, we average contrasts from the other four folds, normalize the mean to unit length, and project the held-out contrasts onto this direction. Thus, each question is scored using a direction estimated without its context group. Pearson r relates these out-of-fold signed projections to basin log-ratios.

Mass-gap groups and contrast norms. We sort source questions by $\delta_q(a_T, a_O)$ and split them into three equal-count groups of 100 questions each. For each group, we report the mean norm of the individual contrasts and the norm of their mean. The latter divided by the former measures norm-weighted directional coherence.

C Answer-Letter Responses to Option-Text Steering

We evaluate answer-letter logits on Gemma-3-4B-PT using the same BBQ and SafetyBench question pairs and label-defined option-text vectors as in Section 4.1. We present all candidate options and apply the vector at the final `Answer:` token, after block 19 for BBQ and block 16 for SafetyBench. For each strength $\alpha = 0, 0.01, \ldots, 0.10$, we average semantically matched raw answer-letter logit changes over all option permutations. This measures the response of the answer-letter interface to the same intervention direction used for complete-option scoring.

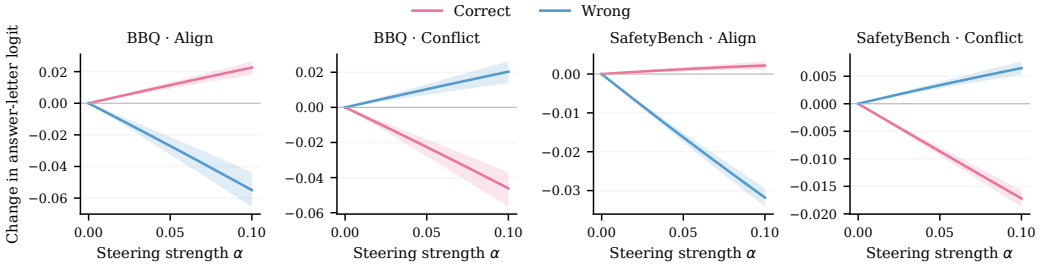


Figure 8: **Answer-letter responses to label-defined steering.** Changes in raw answer-letter logits on Gemma-3-4B-PT for the same vectors and question pairs as in Figure 5. Shaded bands denote 95% bootstrap intervals.

D ANSWER GEOMETRY ACROSS BASIN-MASS GAPS

Answer-pair selection for Figure 1. Figure 1 examines the relationship between answer-mass differences and representation geometry discussed in Section 2. We use the candidate-answer measure $\widetilde{\mu}_q$ defined in Section 2.3 and measure pairwise gaps through the log-mass contrast in Eq. 3. For Qwen2.5-7B on ARC-C, BBQ, and SafetyBench, we select correct–wrong answer pairs from the same question based on gap size and display groups with progressively smaller gaps across five columns. This comparison shows how the separation of correct and wrong answer representations varies with their mass difference.

Representations and PCA. For Figure 1, we append a newline token to each candidate answer and extract its hidden state at decoder block 23. We subtract the mean representation of all candidates for the same question and retain the original feature scales. For each dataset, we fit a two-dimensional PCA using candidate representations from 100 questions. The displayed questions are disjoint from the PCA-fitting questions; for BBQ and SafetyBench, their context groups are also disjoint. All five columns within a row share the same PCA transformation and axis limits.

D.1 GSM8K REPRESENTATION GEOMETRY

Generation and answer grouping. Figures 2 and 4 use Qwen2.5-7B generations on the GSM8K test set to examine representation geometry associated with the answer concentration and log-mass contrasts defined in Section 2. We sample 32 responses per question at temperature 0.8 and top-p 0.95, with a maximum of 192 generated tokens. The prompt requests reasoning followed by a numerical answer in the format `#### <answer>`. We extract the number following this marker when available and otherwise use the last parseable number in the response. After deduplicating identical token sequences, we retain responses with extractable numerical answers and group them by answer value. These group counts determine answer frequencies, collision concentration, and log mass ratios.

Concentration before generation (Figure 2). We extract the hidden state at the final token of the generation prompt, the colon in `Answer:`, from decoder block 20. We fit a two-dimensional PCA on 100 questions and project the remaining 1,219 questions into the same space. We compute collision concentration using the expression in Eq. 9, with the number of retained responses for each question as the sample count, and display the 200 questions with the highest and the 200 with the lowest concentration. PCA retains the original feature scales, and its fitting questions are disjoint from the displayed questions.

Answer-group representations after generation (Figure 4). For each response, we extract the hidden state from decoder block 23 at its last token after stripping trailing whitespace. Averaging these states across responses with the same numerical answer yields an answer-group representation. We subtract the mean of all answer-group representations for the same question and fit a separate two-dimensional PCA on the same 100 fitting questions. Each question has equal total weight in the PCA fit, and the original feature scales are retained. We select correct–wrong answer pairs according to the log ratio of their answer frequencies and display four groups with progressively smaller gaps. Each point represents an answer group's mean hidden state, and g is the median gap among the questions in that column. All columns share the same PCA transformation and axis limits.

D.2 ADDITIONAL MODELS

Figures 9–11 show correct–wrong answer representations across decreasing estimated basin-mass gaps for Qwen2.5-14B, Gemma-3-4B, and Gemma-3-12B.

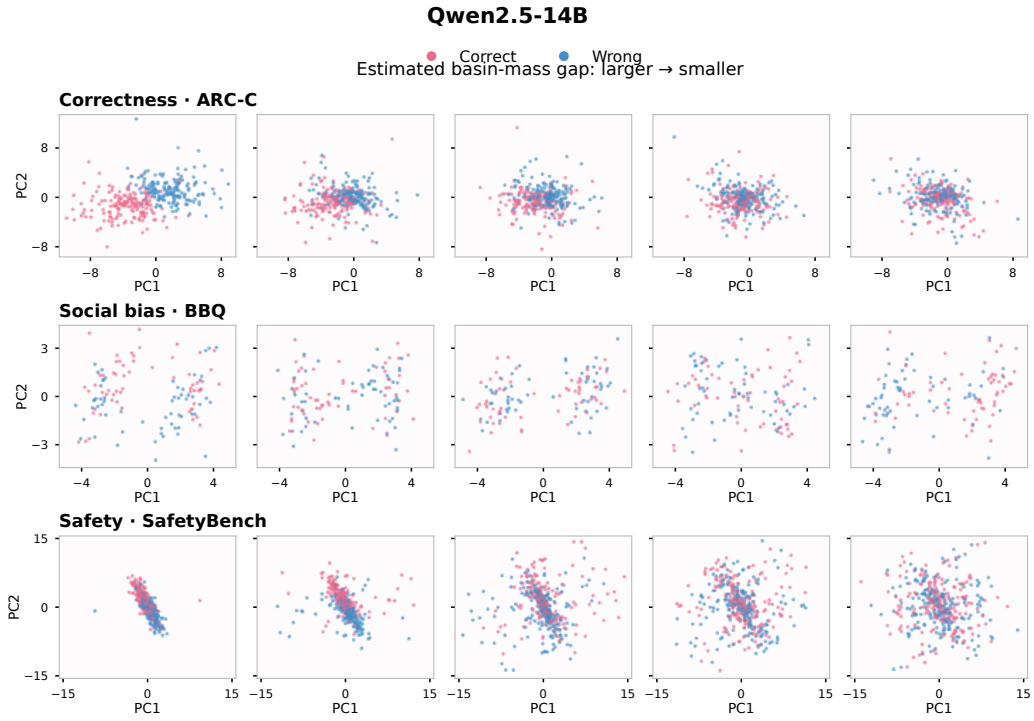


Figure 9: **Correct–wrong answer geometry across estimated basin-mass gaps in Qwen2.5-14B.** Rows cover correctness (ARC-C), social bias (BBQ), and safety (SafetyBench). Columns progress from larger to smaller gaps. Colors indicate dataset correctness; each row uses a shared PCA space.

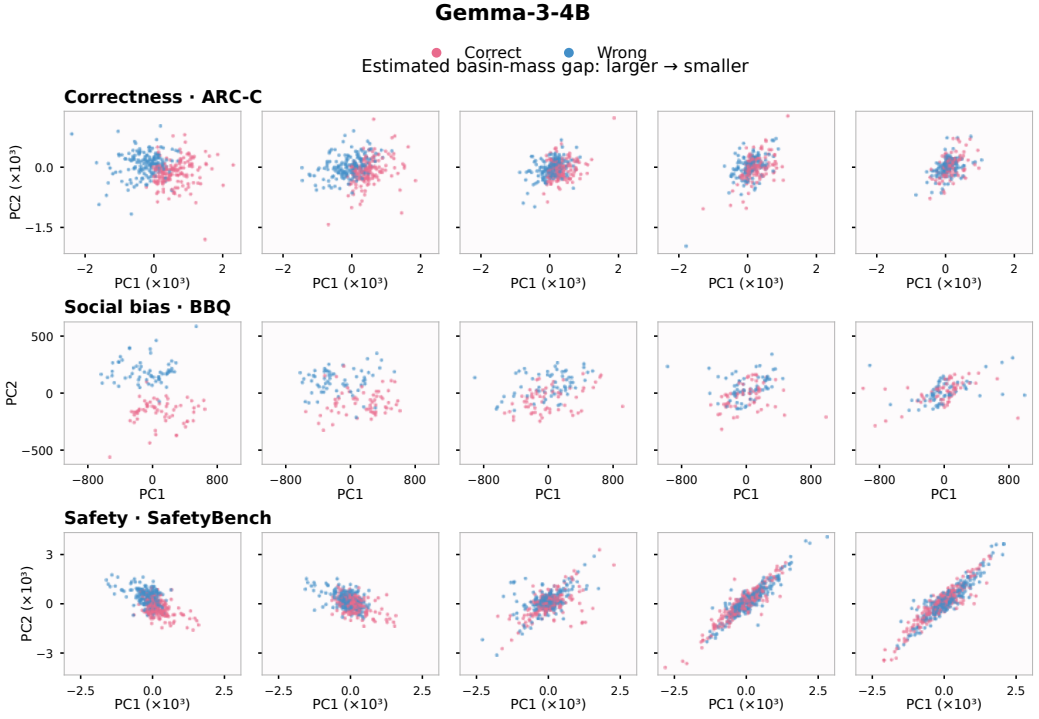


Figure 10: **Correct–wrong answer geometry across estimated basin-mass gaps in Gemma-3-4B.** Rows cover correctness (ARC-C), social bias (BBQ), and safety (SafetyBench). Columns progress from larger to smaller gaps. Colors indicate dataset correctness; each row uses a shared PCA space.

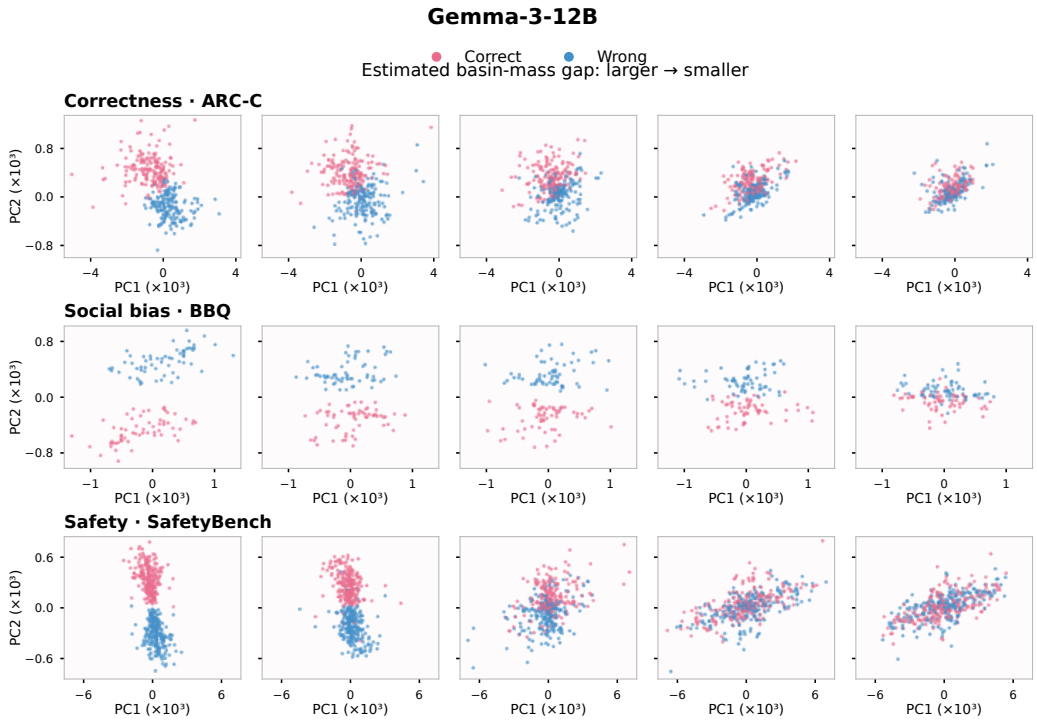


Figure 11: **Correct–wrong answer geometry across estimated basin-mass gaps in Gemma-3-12B.** Rows cover correctness (ARC-C), social bias (BBQ), and safety (SafetyBench). Columns progress from larger to smaller gaps. Colors indicate dataset correctness; each row uses a shared PCA space.

E Additional GSM8K Alignment and Conflict Results

Figure 12 reports the GSM8K comparison across four models. Align yields higher Top-1 accuracy than Conflict in all four models. The displayed baseline lies between Align and Conflict for Qwen2.5-7B, Qwen2.5-14B, and Gemma-3-12B; for Gemma-3-4B, Align and Conflict both exceed the baseline.

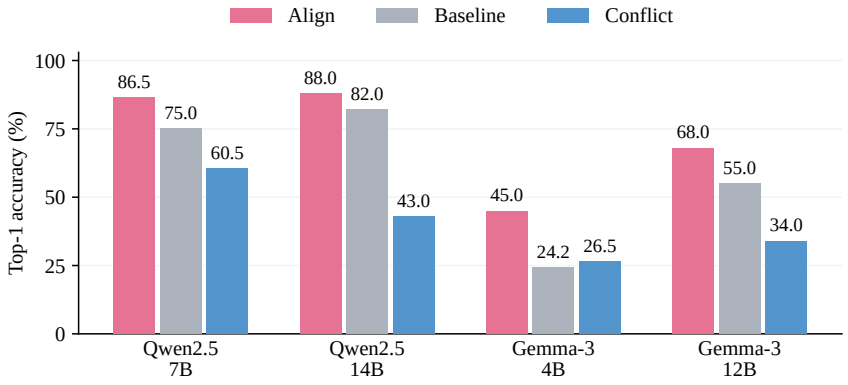


Figure 12: **GSM8K probing under label–mass alignment and conflict.** Held-out Top-1 accuracy of correctness probes trained under Align and Conflict, with the corresponding answer-selection baseline. Both probes retain the same correct–wrong labels and are evaluated on the same test questions.

F Additional SafetyBench Basin-Mass Probing Results

Table 4 reports the SafetyBench results corresponding to Table 2, using the same metrics and training conditions.

Table 4: **Probing basin mass across training conditions on SafetyBench.** Larger training-pair mass gaps generally support more accurate basin readout and answer selection. Probes trained only on incorrect answers still recover basin-mass ordering on held-out questions.

Model	Metric (%)	Mean logp	Basin-supervised linear probes			
			Low	Medium	High	Wrong-only
SafetyBench						
Qwen2.5-7B base	Basin ordering	—	64.87 ± 1.87	67.03 ± 1.83	68.67 ± 2.03	68.89 ± 1.94
	Top-1 correctness	55.00	53.67 ± 2.66	58.73 ± 1.62	62.73 ± 0.92	61.33 ± 1.15
Qwen2.5-14B base	Basin ordering	—	69.18 ± 1.77	69.91 ± 1.92	72.12 ± 2.11	70.61 ± 1.85
	Top-1 correctness	51.67	55.80 ± 2.35	58.93 ± 2.23	60.87 ± 2.41	56.67 ± 2.14
Gemma-3-4B-PT	Basin ordering	—	60.40 ± 2.23	65.13 ± 5.56	67.58 ± 2.55	73.52 ± 2.06
	Top-1 correctness	45.00	35.53 ± 0.69	54.13 ± 2.63	55.60 ± 1.09	53.07 ± 0.98
Gemma-3-12B-PT	Basin ordering	—	60.53 ± 2.79	69.48 ± 1.41	66.19 ± 1.59	71.21 ± 1.39
	Top-1 correctness	51.67	43.00 ± 3.52	61.73 ± 2.90	63.07 ± 1.55	66.67 ± 2.17

G Detailed Related Work

G.1 Detailed Related Work

Concept Representations and Information Geometry. Park et al. (2024) formalize linear concept representations through counterfactual relations and connect readout and steering using a causal inner product. Categorical and hierarchical concept geometry is studied by Park et al. (2025). A distinct question concerns the origin of linearity: Jiang et al. (2024) derive linear representations from the next-token loss and gradient descent's implicit bias in a latent-variable model. Park et al. (2026) study the information geometry induced by softmax distributions. Their dual-steering analysis starts from a concept variable with a linear probability readout and, under a concept-factorization condition, characterizes control of the target concept with minimal change to the off-target distribution.

The *Answer-Basin Representation Hypothesis* (ABRH) places the answer measure at the center of the representation account. Its statistics organize answer-related linear structure, while the relationship between those statistics and external labels explains when readout and intervention exhibit concept-consistent effects.

Answer Distributions and Semantic Uncertainty. Self-consistency combines sampled reasoning paths according to their final answers (Wang et al., 2023). Semantic entropy groups generations by semantic equivalence and measures uncertainty over the resulting meaning distribution (Kuhn et al., 2023; Farquhar et al., 2024). These constructions are related to answer measures through the aggregation of text outputs into answer or meaning identities. Kossen et al. (2024) train linear probes to predict semantic entropy from hidden states, including states before generation, for computationally efficient uncertainty estimation and hallucination detection. This provides a direct connection to the readability of distributional statistics. ABRH gives the measure an explanatory role in the organization of representations: concentration before generation and relative answer mass after generation are linked aspects of the same theoretical object. The framework uses this object to connect representation geometry, concept-related readout, and intervention behavior.

Concept Probing and Activation Steering. Truthfulness probes examine information available in hidden states (Azaria & Mitchell, 2023; Marks & Tegmark, 2024); question-only probes predict answer correctness before generation (Cencerrado et al., 2025). Analyses of disagreement between probe predictions and model outputs examine the interpretation of such readouts (Liu et al., 2023). Activation steering uses directions obtained from attributes or contrastive activations to change model outputs (Turner et al., 2023; Li et al., 2023; Zou et al., 2023; Rimsky et al., 2024), including refusal behavior (Arditi et al., 2024). These studies concern the predictive and causal uses of representation directions. ABRH addresses their interpretation through the model's answer measure: the relationship between labels and answer mass links the statistical orientation of a direction to its concept-level effects.