NOUMENA RESEARCH

Same Confidence, Opposite Truth: An Atlas of Hidden Answer Geometry

Noumena

Label-blind agreement can remain nearly stationary while truth-aligned agreement inverts.

Abstract

Post-training can leave the estimated low-order shape of answer concentration nearly stationary while reversing what the model repeatedly answers. On a paired GSM8K slice, accuracy falls from 0.560 to 0.040 under DeepSeek-R1-Zero ($p = 0.000977$). Changes in total two-, three-, and four-way agreement are unresolved and small at their point estimates relative to the opposing components beneath them: 89.5–95.0% of labelled component motion cancels in the totals, gold-aligned agreement collapses, and wrong-answer agreement rises at every order, with all six component intervals excluding zero; empirical modal sets are disjoint on $23/25$ prompts. We call this estimated low-order pattern *near-isospectral epistemic inversion*. To explain it, we define the protocol-conditioned distribution over semantic answers as a hidden answer state on an answer-state manifold. Under known finite support, integer-order self-consistency measures a power-sum spectrum that is complete for distributional shape up to label permutation and gives local coordinates on regular strata; a label-aware occupancy chart restores semantic identity. Controlled activation records expose the missing coordinates directly. In one GSM8K wound, the correct greedy answer and every recorded one-pass statistic are bit-identical while four samples move from $\{260, 260, 260, \perp\}$ to $\{1.2, 260, 3, 35\}$. In a GPQA wound, even the complete empirical agreement spectrum is fixed while one sampled identity moves from a wrong answer to gold. We prove that answer geometry is recoverable from a confidence observer exactly when the answer map factors through that observer, derive a hidden-rank law for observer-blind directions, and show how repeated sampling separates states that one-pass observation cannot. The same artifacts show beneficial consolidation and a no-answer attractor. Confidence is a projection; the Atlas measures the answer geometry it can erase.

Correspondence: research@noumena.com

1 Introduction

Same confidence. Opposite truth. Epistemic inversion occurs when estimated label-blind concentration barely moves while alignment with truth reverses.

The existing frontier-scale MoE artifacts contain this transition. On the paired 25-prompt GSM8K slice, greedy accuracy falls from 0.560 to 0.040 between DeepSeek-V3-Base and DeepSeek-R1-Zero.

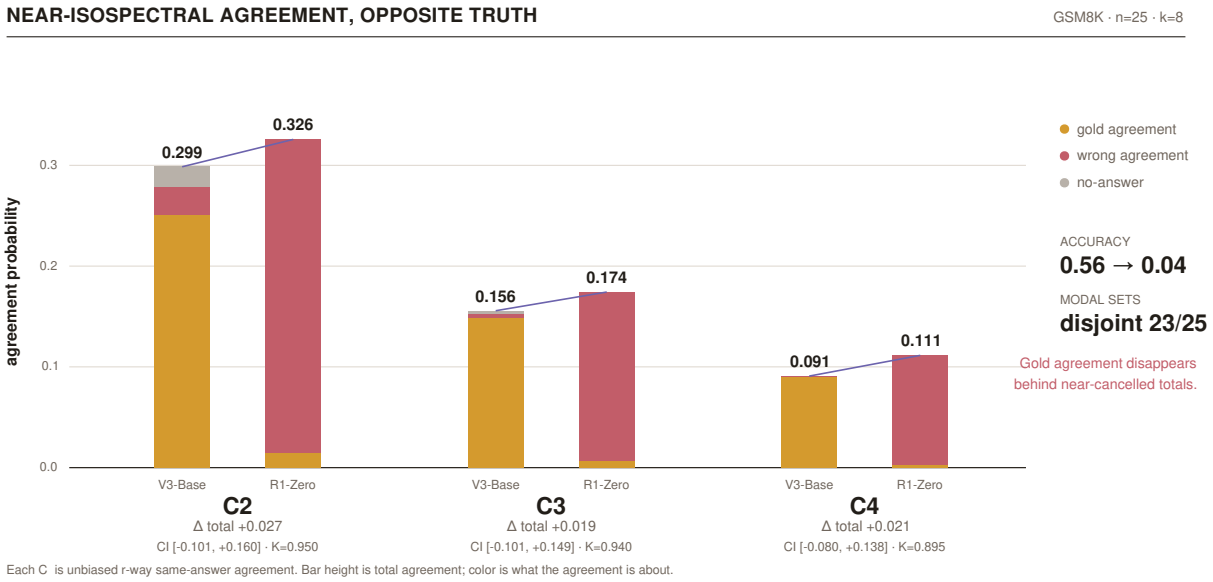

Figure 1 Near-isospectral epistemic inversion. DeepSeek-V3-Base → R1-Zero on the paired GSM8K slice. Estimated total two-, three-, and four-way agreement changes are small relative to the opposing components and remain unresolved, while gold-aligned agreement collapses and wrong-answer agreement rises at every order. Paired-prompt 95% bootstrap intervals are printed beneath each order. Label-blind totals barely move; truth alignment inverts.

Exactly 14 prompts regress and 1 improves, giving a McNemar p-value of 0.000977. Yet the point changes in total two-, three-, and four-way agreement are only 0.0271, 0.0186, and 0.0206, and every paired 95% interval crosses zero. Labelled component motion cancels by 0.950, 0.940, and 0.895 in those totals.

The total spectrum is almost still. Its semantic composition inverts. Gold-aligned agreement falls at every order by -0.2371, -0.1414, and -0.0874, while wrong-answer agreement rises by 0.2843, 0.1629, and 0.1080. All six paired intervals exclude zero. The empirical modal sets before and after post-training are disjoint on $23/25$ prompts, with mean empirical total variation 0.835.

We call this *near-isospectral epistemic inversion*: the estimated low-order total spectrum shows near-cancellation while its labelled components move from correct toward incorrect answer basins. The term describes the point-estimate pattern; it is not a population equivalence claim. Figure 1 shows the inversion. A scalar confidence score misses it. Even a multi-order, permutation-invariant description of concentration misses its semantic direction.

A greedy answer or scalar confidence does not characterize a stochastic language model. Fix a model θ, prompt x, decoding policy π, and semantic answer extractor g. For

$$Y \sim P_{\theta,\pi}(\cdot \mid x), \tag{1}$$

the induced distribution

$$p_{\theta,\pi,g,x}(a) = \Pr[g(Y) = a] \tag{2}$$

is the model's *hidden answer state*. It lives on an answer-state simplex whose interior is a statistical manifold. Its mass can consolidate around a correct answer, fragment across incompatible answers, cross into a wrong modal basin, or collapse into no answer at all.

Repeated sampling gives this state an exact measurement geometry. For integer $r \geq 2$, the probability that r independent draws agree is

$$C_r(p) = \Pr(A_1 = \cdots = A_r) = \sum_a p(a)^r. \tag{3}$$

The sequence $(C_2, C_3, \ldots)$ is the power-sum spectrum of the hidden answer state. On finite known support, enough orders recover the multiset of answer probabilities exactly. On regular strata they form a local atlas of answer-state *shape* up to label permutation. A second, label-aware occupancy chart tells us where that shape sits relative to gold, wrong, and no-answer basins. The R1-Zero artifacts measure a nearly stationary low-order prefix of the former and catastrophic movement in the latter.

Controlled activation interventions expose the same separation inside a fixed checkpoint.

Same observer. Same greedy answer. Different answer geometry. In the first wound in Figure 2, DeepSeek-V3-Base answers GSM8K prompt 6 correctly with 260. A recorded final-layer radial intervention leaves its greedy answer and every recorded prompt-step statistic (Shannon entropy, Rényi-2 entropy, both varentropies, maximum probability, and margin) bit-identical. The declared confidence surface records no change, yet four sampled answers move from

$$\{260, 260, 260, \perp\} \quad \longrightarrow \quad \{1.2, 260, 3, 35\}. \tag{4}$$

Empirical answer-space H_2 rises from 0.470 to 1.386 and total variation is 0.75. A second recorded radial run again preserves the full observer and greedy answer while producing $\{1.2, 260, 260, 35\}$.

The second wound lies in an even narrower fiber. On GPQA prompt 21, the observer and wrong greedy answer D remain bit-identical. So does the entire frequency multiset, hence every empirical agreement order and the complete quartet (H_1, H_2, V_1, V_2). Yet the sampled cloud moves from

$$\{B, D, D, \perp\} \quad \longrightarrow \quad \{C, D, D, \perp\}, \tag{5}$$

where C is gold. Both recorded radial runs contain the same identity swap. Distributional shape is unchanged; semantic identity is not.

A one-pass confidence statistic is not this state. Let $\mathcal{M}_{\theta,x}$ denote a local activation manifold. The answer map

$$P_x : \mathcal{M}_{\theta,x} \to \Delta(\overline{\mathcal{A}}_x) \tag{6}$$

and a confidence observer

$$S_x : \mathcal{M}_{\theta,x} \to \mathcal{S} \tag{7}$$

are two different images of the same internal neighborhood. The confidence surface can fold together activation states whose answer distributions differ. Infinitesimally, a direction v satisfying

$$DS_x(v) = 0, \qquad DP_x(v) \neq 0 \tag{8}$$

is invisible to the observer but active in answer space: a *hidden answer direction*.

These finite empirical clouds motivate the population-level hierarchy characterized below:

$$S(h) = S(h') \quad \text{while} \quad G(P(h)) \neq G(P(h')), \tag{9}$$

$$S(h) = S(h'),\ G(P(h)) = G(P(h')) \quad \text{while} \quad P(h) \neq P(h'). \tag{10}$$

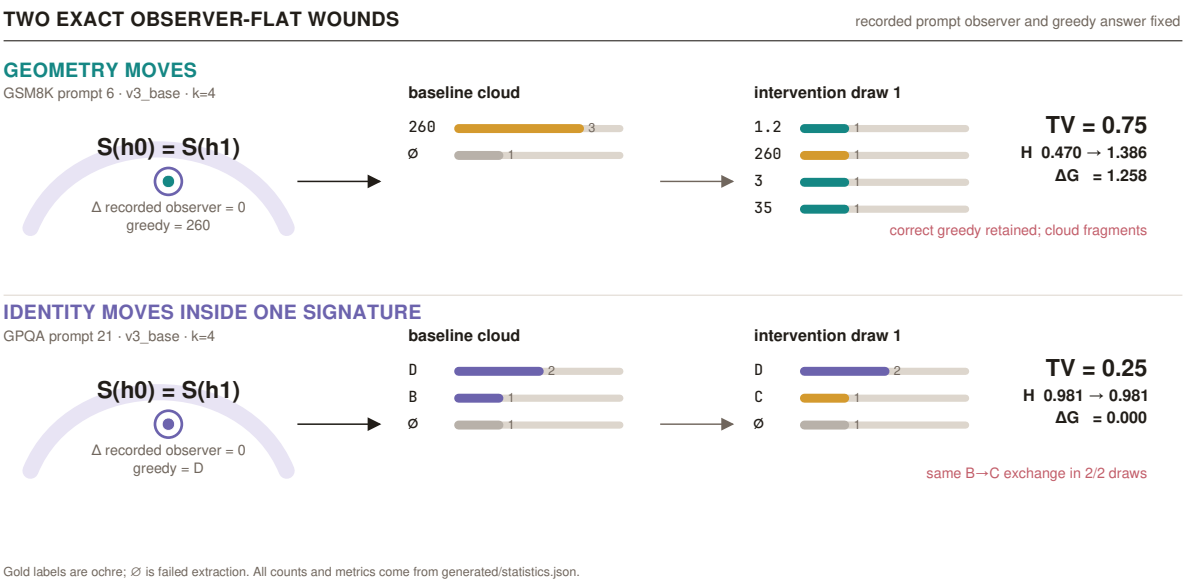

Figure 2 Two nested observer-blind wounds. Top: the declared one-pass observer and correct greedy answer are fixed while the empirical answer cloud fragments. Bottom: the observer, greedy answer, quartet, and every empirical power sum are fixed while one sampled answer moves from wrong B to gold C. Each row plots baseline and the first recorded final-radial run at $k = 4$; the bottom annotation reports that both recorded runs contain the same B→C exchange.

They are not powered certificates that the underlying population states satisfy these inequalities. The first theoretical case is *observer blindness*: the confidence surface erases answer-state motion. The second is *coordinate blindness*: every permutation-invariant shape coordinate loses which semantic answers occupy that shape. The full labelled answer state is the object; agreement geometry and semantic occupancy are complementary charts in its Atlas.

At order two, the measurement reduces to the familiar collision identity

$$\Pr(A = A') = \sum_a p(a)^2 = e^{-H_2(p)}. \tag{11}$$

The higher-order result is stronger: a single sample cloud gives unbiased U-statistics for the agreement spectrum. Self-consistency measures the hidden answer manifold through the same samples used for voting.

The remaining checkpoint branches show why the label-aware Atlas matters. DeepSeek-V3 on GSM8K raises accuracy from 0.560 to 0.680 while total collision rises by 0.359 (95% CI $[0.207, 0.504]$), almost entirely through gold-aligned agreement: beneficial consolidation. DeepSeek-R1 on GPQA has accuracy fall from 0.260 to 0.000 while collision rises by 0.522, because sampled no-answer mass grows from 0.345 to 0.810: a no-answer attractor. A single confidence axis would collapse beneficial concentration, wrong-basin inversion, and extraction failure. The Atlas separates them.

This paper makes five contributions:

1. We identify near-isospectral epistemic inversion: a post-training transition whose measured total agreement spectrum remains unresolved while truth-aligned and wrong-answer components reverse at every measured order.

2. We define the protocol-conditioned hidden answer state and an Atlas separating agreement geometry from semantic basin occupancy.
3. We prove that agreement U-statistics measure the power-sum spectrum, which is complete for finite-support shape, and derive a wrong-basin stabilization law showing that more self-consistency can lock onto a wrong mode.
4. We prove that answer geometry is recoverable from a one-pass observer exactly when the answer map factors through it, and derive a hidden-rank law characterizing unavoidable observer-blind directions.
5. We exhibit two controlled observer-flat wounds and a strict perturbation audit: $7/18$ reconstructable conditions preserve every recorded observer coordinate exactly; 5 still change their empirical answer histograms.

Our result is conditional and exact for the declared observer: confidence is a projection, answer geometry is protocol-defined, and whenever the projection collapses an answer-active direction, calibration cannot reconstruct what was erased. The artifacts suggest, but do not establish, causal predictive utility for post-training quality and test-time-compute risk. Matched branches across larger model families test prediction; matched Atlas interventions test causality.

2 Answer States, Manifolds, and Observer Surfaces

2.1 A protocol-defined answer state

Let $\kappa = (\pi, g)$ denote a complete measurement protocol: a stochastic decoding rule π and a deterministic answer extractor g. We include an explicit no-answer symbol $\perp$ and write

$$\overline{\mathcal{A}}_x := \mathcal{A}_x \cup \{\perp\}, \qquad p_{\theta,\kappa,x,\delta}(a) := \Pr_{Y \sim P_{\theta,\pi}(\cdot|x,\delta)}[g(Y) = a]. \tag{12}$$

The perturbation index δ is omitted when absent. The dependence on θ, π, and g matters: changing temperature, prompt template, or semantic normalization changes the measured state.

The answer state lives in the closed simplex

$$\mathcal{P}_x = \Delta(\overline{\mathcal{A}}_x). \tag{13}$$

Its interior is a statistical manifold; its closure is a manifold with corners. We use *hidden answer geometry* for distances, curves, basins, and charts on this simplex. The term "hidden" is observer-relative: p is the unobserved population object and the empirical histogram $\widehat{p}_k$ is its sample estimate.

The simplex has an exact answer-basin decomposition. For each answer a, define the open modal basin

$$\mathcal{C}_a := \{p \in \mathcal{P}_x : p(a) > p(b) \text{ for every } b \neq a\}. \tag{14}$$

Tie surfaces form the basin boundaries. A state can sharpen or diffuse inside one basin, cross into a different dominant answer, or enter the no-answer basin $\mathcal{C}_\perp$. These transitions are motions on the measured answer simplex.

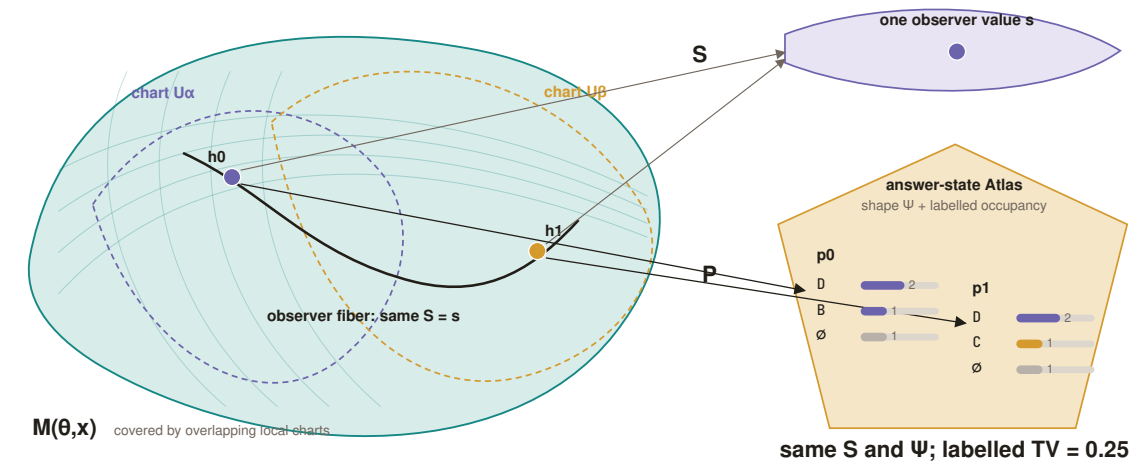

Figure 3 Activation geometry and the answer-state Atlas. A local activation manifold maps to a one-pass confidence surface through S_x and to the labelled answer-state Atlas through P_x. The latter combines permutation-invariant shape coordinates Ψ with semantic occupancy. Distinct activation points on one observer fiber can map to empirically different labelled clouds. The manifold and fiber are schematic; the endpoint histograms are receipt-backed.

2.2 Activation geometry and the answer map

Let $\mathcal{M}_{\theta,x}$ be a local reachable-state manifold in activation space. A fixed protocol induces two maps,

$$P_x : \mathcal{M}_{\theta,x} \to \mathcal{P}_x, \qquad h \mapsto p_h, \tag{15}$$

$$S_x : \mathcal{M}_{\theta,x} \to \mathcal{S}, \qquad h \mapsto \Phi(Z_h), \tag{16}$$

where P_x maps a reachable activation state to its answer state and S_x maps it to the chosen one-pass observer. The answer-state manifold and confidence surface are different images of the same local activation neighborhood.

The corresponding activation-space answer basins are the pullbacks

$$\mathcal{B}_a := P_x^{-1}(\mathcal{C}_a) \subseteq \mathcal{M}_{\theta,x}. \tag{17}$$

They need not be connected: one semantic answer may occupy several local components of the activation manifold. Controlled interventions probe these pullbacks and observe their image in answer space. The present receipts do not reconstruct every component globally. The basin object itself remains precise.

2.3 An atlas for answer-state geometry

For p on finite support, define the log power-sum curve

$$F_p(\alpha) := \log \sum_{a \in \mathrm{supp}(p)} p(a)^\alpha. \tag{18}$$

Its values and derivatives give

$$H_1(p) = -F_p'(1), \qquad H_2(p) = -F_p(2), \tag{19}$$

$$V_1(p) = F_p''(1), \qquad V_2(p) = F_p''(2). \tag{20}$$

Here V_1 is surprisal variance under p, while V_2 is surprisal variance under the order-2 escort distribution. We call

$$G(p) = (H_1, H_2, V_1, V_2) \tag{21}$$

the *quartet signature*. It summarizes spread, collision, and local curvature of F_p at orders one and two. It is permutation-invariant by construction: states with the same shape share quartet coordinates even when different semantic answers occupy that shape. The complete power-sum coordinates developed in Section 3 give a genuine local shape chart on regular finite-support strata; a label-aware occupancy chart records position among named answer basins. Their combination forms the answer-state Atlas. The quartet remains a low-cost four-functional projection of that Atlas and is not a globally valid chart.

This produces two nested fibers on the activation manifold:

$$\mathcal{F}_S(s) := S_x^{-1}(s), \tag{22}$$

$$\mathcal{F}_{S,G}(s, g) := \{h : S_x(h) = s,\ G(P_x(h)) = g\}. \tag{23}$$

Observer blindness is variation of $G \circ P_x$ inside $\mathcal{F}_S$. Coordinate blindness is variation of the full labelled state P_x inside $\mathcal{F}_{S,G}$. The GSM8K and GPQA opening wounds are empirical representatives of these two cases.

2.4 No-answer versus extracted-answer agreement

Let $r = p(\perp)$. If $r < 1$, write $\bar{p}(a) = p(a)/(1 - r)$ for the distribution conditional on extraction. Collision decomposes exactly as

$$c(p) = p(\perp)^2 + \sum_{a \neq \perp} p(a)^2 \tag{24}$$

$$= r^2 + (1 - r)^2 c(\bar{p}). \tag{25}$$

The first term is repeated no-answer agreement; the second is same-answer agreement among all draws. This decomposition prevents a format or extraction collapse from being mislabeled as semantic confidence. Both terms have unbiased pair-count estimators.

2.5 Observer-relative legibility

A confidence surface is always a specified object $S_x = \Phi(Z_x)$: for example, prompt-step logit H_2, a vector of logit statistics, a decode-path trace, or a learned probe. For each of four selected prompts, the strict perturbation audit jointly checks prompt-step H_1, H_2, V_1, V_2, maximum probability, margin, and greedy answer. The opening single-prompt wounds use the same recorded coordinates. This observer contains more information than one entropy scalar and remains a strict subset of conceivable one-pass observers.

3 The Geometry of What Confidence Erases

We establish three results: repeated sampling measures an entire power-sum spectrum of the hidden answer state; a one-pass observer recovers a target exactly if and only if the target map factors through the observer surface; and differential rank determines when invisible answer directions are unavoidable.

3.1 Self-consistency measures an agreement spectrum

Fix a finite answer alphabet with known cardinality bound m, padding to m with absent answers of zero mass. For $p \in \Delta_{m-1}$ and $\alpha > 0$, define

$$C_\alpha(p) := \sum_{a=1}^{m} p_a^\alpha, \qquad F_p(\alpha) := \log C_\alpha(p). \tag{26}$$

For integer $r \geq 2$, C_r is an r-way agreement probability.

Theorem 1 (Agreement-spectrum measurement)**.** *Let $A_1, \ldots, A_k \overset{\text{iid}}{\sim} p$ and $2 \leq r \leq k$. Define*

$$U_{r,k} := \binom{k}{r}^{-1} \sum_{\substack{I \subseteq [k] \\ |I|=r}} \mathbf{1}\{A_i = A_j \text{ for all } i, j \in I\}. \tag{27}$$

Then

$$\mathbb{E}[U_{r,k}] = C_r(p) = \Pr(A_1 = \cdots = A_r) = \exp\big((1-r)H_r(p)\big), \tag{28}$$

and, for every $t > 0$,

$$\Pr\big(|U_{r,k} - C_r(p)| \geq t\big) \leq 2\exp\left(-\frac{2kt^2}{r^2}\right). \tag{29}$$

For $r = 2$, $U_{2,k}$ is the all-pairs self-consistency estimator and $C_2 = e^{-H_2}$. Writing N_a for the sample count of answer a,

$$U_{r,k} = \frac{\sum_a \binom{N_a}{r}}{\binom{k}{r}}. \tag{30}$$

One answer cloud both produces a vote and estimates an agreement spectrum.

For a prompt with gold answer $a^\star$ and explicit no-answer symbol $\perp$, every order has the exact label-aware decomposition

$$C_r(p) = \underbrace{p(a^\star)^r}_{C_r^{\text{gold}}} + \underbrace{\sum_{a \notin \{a^\star, \perp\}} p(a)^r}_{C_r^{\text{wrong}}} + \underbrace{p(\perp)^r}_{C_r^{\text{none}}}. \tag{31}$$

The total spectrum is permutation-invariant; its three labelled components are not. Near-isospectral epistemic inversion is motion that leaves the former nearly fixed while transferring power between the latter.

Corollary 1 (Dimension-free collision measurement)**.** *For $c = C_2(p)$ and $t_3 = C_3(p)$,*

$$\mathrm{Var}(U_{2,k}) = \frac{2(c - c^2) + 4(k-2)(t_3 - c^2)}{k(k-1)} \leq \frac{1}{k}. \tag{32}$$

Consequently $\mathbb{E}|U_{2,k} - c| \leq k^{-1/2}$, independently of answer-alphabet size.

3.2 The power-sum atlas

At finite support, the agreement spectrum determines the shape of the answer state.

Theorem 2 (Power-sum atlas)**.** *For $p \in \Delta_{m-1}$, the power sums $C_1(p), \ldots, C_m(p)$ determine the multiset $\{p_1, \ldots, p_m\}$. Since $C_1 = 1$, the coordinates $C_2, \ldots, C_m$ determine answer-state shape up to permutation of answer labels.*

On the regular quotient stratum where all $p_i > 0$ and the masses are pairwise distinct, the map

$$\Psi([p]) = (C_2(p), \ldots, C_m(p)) \tag{33}$$

is a smooth local coordinate system on the simplex modulo label permutation. Equal-mass states form symmetry strata where Ψ remains a complete invariant but ceases to be a smooth chart.

The recovery is constructive. Newton's identities

$$j e_j = \sum_{r=1}^{j} (-1)^{r-1} e_{j-r} C_r, \qquad e_0 = 1, \tag{34}$$

recover the elementary symmetric polynomials. The probabilities are the roots of

$$z^m - e_1 z^{m-1} + e_2 z^{m-2} - \cdots + (-1)^m e_m = \prod_{i=1}^{m} (z - p_i). \tag{35}$$

The augmented power-sum Jacobian is a scaled Vandermonde matrix:

$$\det \left[\frac{\partial C_r}{\partial p_j} \right]_{r,j=1}^{m} = m! \prod_{i<j} (p_j - p_i), \tag{36}$$

which is nonzero exactly off equal-mass strata. A finite prefix is complete only when a finite support bound m is fixed; otherwise the full moment sequence or an explicit support assumption is required.

The quartet is a four-functional projection of the power-sum atlas,

$$G(p) = \big(-F_p'(1), -F_p(2), F_p''(1), F_p''(2)\big) \tag{37}$$

$$= (H_1, H_2, V_1, V_2), \tag{38}$$

where

$$F_p''(\alpha) = \mathrm{Var}_{q_\alpha}[\log p(A)], \qquad q_\alpha(a) \propto p(a)^\alpha. \tag{39}$$

The quartet is a four-functional jet of the complete agreement geometry and is not globally injective.

The opening wounds instantiate two nested impossibilities. The GSM8K opening wound changes $G(\widehat{p})$ while preserving the observer. The GPQA opening wound preserves the entire frequency multiset and therefore every $C_r(\widehat{p})$, not just the quartet, while exchanging semantic labels B and C. No permutation-invariant shape statistic can distinguish those two empirical states. Shape and semantic identity require different charts in the answer-state atlas.

3.3 Legibility is factorization through the observer

Let $U \subseteq \mathcal{M}_x$ be an admissible activation neighborhood, let $P : U \to \mathcal{P}_x$ be the answer map, and let $S : U \to \mathcal{S}$ be the chosen observer. For a target $T : \mathcal{P}_x \to (\mathcal{T}, d)$, write $f = T \circ P$ and define the fiber diameter

$$\Gamma_S^f(U) := \sup_{s \in S(U)} \operatorname{diam}_d f\big(U \cap S^{-1}(s)\big). \tag{40}$$

Let $\mathfrak{E}_S$ be the class of deterministic or randomized estimators whose conditional output law depends on h only through $S(h)$.

Theorem 3 (Factorization or irreducible ambiguity)**.** *The following are equivalent:*

1. $\Gamma_S^f(U) = 0$*;*
2. f *is constant on every* S*-fiber;*
3. *there exists a unique set map* $R : S(U) \to \mathcal{T}$ *satisfying* $f = R \circ S$*.*

Moreover, every deterministic or randomized estimator whose conditional output law depends only on S *obeys*

$$\inf_{\widehat{f} \in \mathfrak{E}_S} \sup_{h \in U} \mathbb{E}\, d\big(\widehat{f}(S(h)), f(h)\big) \geq \frac{1}{2}\Gamma_S^f(U). \tag{41}$$

Thus recoverability is exactly a factorization question: a confidence observer can recover answer geometry only when the answer target factors through the confidence surface. Calibration cannot reconstruct variation that the observer quotient has erased.

The factorization above is set-theoretic. If measurable recovery is required, assume f is $\sigma(S)$-measurable and the target is standard Borel; the Doob–Dynkin lemma supplies a measurable R. In the constant-rank smooth setting below, local factorization is smooth.

Corollary 2 (Quantized and near-flat observers)**.** *For a declared logging quantizer* Q_τ*, Theorem 3 applies exactly to* $\widetilde{S} = Q_\tau \circ S$*. For two activation states* h_0, h_1*, write* $S_j := S(h_j)$ *and* $f_j := f(h_j)$*. For an* L*-Lipschitz deterministic decoder* ψ*,*

$$\max_{j \in \{0,1\}} d\big(\psi(S_j), f_j\big) \geq \frac{1}{2}\big[d(f_0, f_1) - L\|S_0 - S_1\|\big]_+. \tag{42}$$

Without quantization or a regularity bound on the decoder, numerical nearness alone implies no universal lower bound.

3.4 The hidden-rank law

Dimension and rank can force fiber ambiguity.

Theorem 4 (Hidden-rank law)**.** *Assume additionally that* $\mathcal{S}$ *and* $\mathcal{T}$ *are finite-dimensional smooth manifolds and that* S *and* $f = T \circ P$ *are smooth near* h*. Set* $A = DS_h$*,* $B = Df_h$*, and define*

$$\ell_f(h) := \operatorname{rank} D(S, f)_h - \operatorname{rank} DS_h. \tag{43}$$

Then

$$\ell_f(h) = \operatorname{rank}\big(B|_{\ker A}\big) \geq \operatorname{rank} B - \operatorname{rank} A. \tag{44}$$

Hence $\ell_f(h) > 0$ if and only if there exists a tangent direction v with

$$DS_h[v] = 0, \qquad Df_h[v] \neq 0. \tag{45}$$

If rank $Df_h >$ rank DS_h*, observer-blind answer directions are unavoidable; this holds in particular when target rank exceeds observer dimension.*

Corollary 3 (Local factorization dichotomy)**.** *If S has constant rank on a neighborhood W of h, then $W \cap S^{-1}(S(h))$ is a submanifold tangent at h to* $\ker DS_h$*. When $\ell_f(h) > 0$, a curve inside this local exact observer fiber moves the target, so every sufficiently small neighborhood of h has positive local fiber diameter. Conversely, if S has constant rank on a product neighborhood, its local fibers are connected, and $\ell_f = 0$ throughout that neighborhood, then $f = R \circ S$ locally for a smooth map R.*

The hidden-rank law identifies when confidence blindness is mathematically forced: whenever the answer map exposes more independent local directions than the observer can carry, some answer geometry must disappear under projection.

3.5 Sampling penetrates an observer fiber

Corollary 4 (Finite-compute separation)**.** *Suppose h_0, h_1 share the same observer value but have collision gap*

$$\gamma = |C_2(P(h_0)) - C_2(P(h_1))| > 0. \tag{46}$$

Every S-only collision predictor has maximum expected absolute error at least $\gamma/2$. Meanwhile

$$\mathbb{E}|U_{2,k} - C_2| \leq \frac{1}{\sqrt{k}}, \tag{47}$$

so $k > 4/\gamma^2$ makes repeated sampling strictly better on the pair. Thresholding $U_{2,k}$ at the midpoint between the two collision values yields testing error under either state at most

$$2\exp\left(-\frac{k\gamma^2}{8}\right), \tag{48}$$

whereas every S-only test has equal-prior error $1/2$.

Self-consistency penetrates an observer fiber by measuring a functional that varies inside it.

Corollary 5 (Wrong-basin stabilization)**.** *Let $\widehat{a}_k$ be the plurality winner among k independent answers, with arbitrary tie-breaking. Let $a_{\max}$ be a modal answer and suppose $c = C_2(p) > 1/2$. Then $a_{\max}$ is unique,*

$$p(a_{\max}) \geq c > \frac{1}{2}, \tag{49}$$

and plurality self-consistency returns $a_{\max}$ with probability at least

$$1 - \exp\left[-2k\left(c - \frac{1}{2}\right)^2\right]. \tag{50}$$

If $a_{\max}$ is wrong, additional test-time samples stabilize the wrong basin exponentially fast.

This separates measurement from decision. The same samples that reveal a wrong-answer attractor can, under unqualified plurality voting, amplify it. The R1-Zero transition supplies the empirical warning sign: wrong-aligned agreement grows at every measured order even though the estimated total point spectrum barely moves.

3.6 RMS-normalized operational geometry

Let

$$N(h) = \frac{h}{s(h)}, \qquad s(h) = \sqrt{\|h\|_2^2/d + \epsilon}. \tag{51}$$

At $\epsilon = 0$, nonzero positive rays map to the sphere of radius $\sqrt{d}$. At $\epsilon > 0$, the image is its open ball, with Jacobian

$$DN_h[v] = \frac{v}{s(h)} - \frac{h\, h^\top v}{d\, s(h)^3}. \tag{52}$$

Tangential directions have eigenvalue $1/s(h)$; the radial direction has eigenvalue $\epsilon/s(h)^3$. Their ratio is

$$\frac{\epsilon}{s(h)^2} = \frac{\epsilon}{\|h\|_2^2/d + \epsilon}. \tag{53}$$

These sphere, ball, and eigenvalue statements describe the ungained RMS core N. With learned diagonal gain W, the normalized output is $z = WN(h)$ and $Dz_h = WDN_h$, so the image is gain-warped. The recorded `final_radial` intervention is applied *after* `model.norm`, directly in post-gain z-space; its radial direction is therefore not a raw direction subsequently suppressed by DN_h. This distinction fixes the interpretation of the interventions. The empirical question is how the implemented normalized-state directions push forward through P_x and S_x, the geometry probed by the two opening wounds.

4 Experiments and Results

4.1 Experimental protocol

Models and tasks. We compare the shared DeepSeek-V3-Base checkpoint with DeepSeek-V3, DeepSeek-R1-Zero, and DeepSeek-R1 [1, 2]. The fixed evaluation slices are the first 25 GSM8K test examples [3] and the first 50 GPQA test examples [4]. These are diagnostic slices, not full benchmark evaluations.

Answer clouds. For each prompt and checkpoint, the stage-1 receipts contain one greedy completion and $k = 8$ sampled completions. The generation harness uses temperature 0.6, a maximum of 512 generated tokens, and a deterministic per-prompt/per-sample seed schedule. The surviving local bundle contains the prompt-level JSON-derived occupancy receipts and the generator implementation. The original launch command is absent, so reproducibility is receipt-level rather than a promise of a bit-identical rerun from this directory alone.

Extraction. GPQA answers are labels A–D; failure to extract a label is $\perp$. GSM8K answers are numeric strings. Before analysis, finite Decimal-equivalent strings are merged, so `26`, `26.0`, and `26.00` denote one answer. Greedy accuracy uses the same canonicalization. This correction changes several historical receipt values and removes a spurious perturbation effect caused entirely by string formatting.

Statistics. Agreement is measured by the unbiased U-statistic in Equation (27). All checkpoint comparisons are paired by prompt. We report 95% percentile intervals from 50,000 paired-prompt bootstrap resamples with deterministic hashed seeds. Greedy-accuracy transitions are accompanied by an exact two-sided McNemar test in the text when inferentially relevant. Intervals describe uncertainty across the fixed prompt slice and generation realizations; they do not make the slice a random sample from all possible benchmark prompts.

Mechanical provenance. Every number in this section is regenerated by `build_statistics.py` from the JSON files under `occupancy_receipts/`. The build validates posterior inversion, semantic normalization, paired indices, and sample counts before emitting the TeX rows and figure data.

4.2 Result 1: near-isospectral epistemic inversion

DeepSeek-V3-Base $\rightarrow$ R1-Zero on GSM8K isolates the target phenomenon: the estimated *total* low-order agreement spectrum shows near-cancellation while its semantic composition reverses. For each prompt we decompose the unbiased r-way agreement estimator as

$$U_{r,8} = U_{r,8}^{\text{gold}} + U_{r,8}^{\text{wrong}} + U_{r,8}^{\perp}, \tag{54}$$

where the wrong component counts agreement on the same extracted non-gold answer. Table 1 averages these quantities over the 25 paired prompts.

We quantify cancellation at each order by

$$K_r := 1 - \frac{|\Delta C_r^{\text{total}}|}{|\Delta C_r^{\text{gold}}| + |\Delta C_r^{\text{wrong}}| + |\Delta C_r^{\perp}|}. \tag{55}$$

$K_r = 0$ means no cancellation of labelled component motion in the total; $K_r = 1$ means exact cancellation. The observed indices are 0.950, 0.940, and 0.895 for orders two through four. K_r is defined as zero by convention when the denominator vanishes.

The total deltas are only $+0.0271$, $+0.0186$, and $+0.0206$ for orders two through four, and every paired-prompt bootstrap interval includes zero. Yet every gold-component interval and every wrong-component interval excludes zero in the opposite direction. At order two, gold agreement changes by -0.2371 while wrong agreement rises by 0.2843; at order four, the corresponding changes are -0.0874 and $+0.1080$. Greedy accuracy falls from 0.560 to 0.040 (exact McNemar $p = 0.000977$), the empirical modal-answer sets are disjoint on 23 of 25 prompts, and mean empirical answer-state TV is 0.835.

This is an *epistemic inversion*: the estimated low-order point spectrum appears almost unchanged, while the labelled atlas shows agreement moving from the correct basin into wrong basins. The intervals do not establish population equivalence. No scalar concentration summary can see the exchange.

4.3 Complete paired checkpoint matrix

Table 2 reports every stage-1 transition. Δc is tuned minus base unbiased collision. The next two columns decompose that change into pairs that both yield $\perp$ and pairs that yield the same extracted answer. Their sum is Δc exactly.

Order	total agreement; Δ [95% CI]	Δ gold [95% CI]	Δ wrong [95% CI]	cancellation K_r
2	0.2986 $\rightarrow$ 0.3257; $\Delta = 0.0271$ [-0.1014, 0.1600]	-0.2371 [-0.3400, -0.1386]	0.2843 [0.1971, 0.3829]	0.950
3	0.1557 $\rightarrow$ 0.1743; $\Delta = 0.0186$ [-0.1014, 0.1486]	-0.1414 [-0.2250, -0.0664]	0.1629 [0.0814, 0.2657]	0.940
4	0.0909 $\rightarrow$ 0.1114; $\Delta = 0.0206$ [-0.0800, 0.1383]	-0.0874 [-0.1497, -0.0343]	0.1080 [0.0331, 0.2080]	0.895

Table 1 Near-isospectral epistemic inversion under R1-Zero. Mean unbiased r-way agreement on the paired GSM8K slice, V3-Base $\rightarrow$ R1-Zero. Labelled component motion cancels by 89.5–95.0% in the total point estimates while gold agreement is replaced by wrong agreement.

Task	Sibling	n	greedy accuracy	Δc [95% CI]	$\Delta c_{\perp}$	Δc_{ans}	greedy $\perp$
GSM8K	V3	25	0.560 $\rightarrow$ 0.680	0.359 [0.207, 0.504]	-0.020	0.379	5/25 $\rightarrow$ 0/25
GSM8K	R1-Zero	25	0.560 $\rightarrow$ 0.040	0.027 [-0.101, 0.160]	-0.020	0.047	5/25 $\rightarrow$ 0/25
GSM8K	R1	25	0.560 $\rightarrow$ 0.200	-0.081 [-0.181, 0.013]	-0.020	-0.061	5/25 $\rightarrow$ 0/25
GPQA	V3	50	0.260 $\rightarrow$ 0.140	0.415 [0.331, 0.498]	0.331	0.084	15/50 $\rightarrow$ 26/50
GPQA	R1-Zero	50	0.260 $\rightarrow$ 0.140	0.421 [0.342, 0.498]	0.284	0.136	15/50 $\rightarrow$ 25/50
GPQA	R1	50	0.260 $\rightarrow$ 0.000	0.522 [0.454, 0.587]	0.593	-0.071	15/50 $\rightarrow$ 41/50

Table 2 Paired post-training receipt ($k = 8$ per prompt). Collision uses the unbiased pair-count estimator. $\Delta c_{\perp}$ is repeated no-answer agreement and Δc_{ans} is same-extracted-answer agreement; they sum to total collision change. GSM8K numerals are semantically canonicalized.

4.4 Result 2: beneficial consolidation under V3

The V3 transition on GSM8K moves through the same atlas in the useful direction. Mean collision more than doubles from 0.299 to 0.657 ($\Delta = 0.359$, 95% CI $[0.207, 0.504]$), with increases on 20 of 25 paired prompts. The decomposition identifies the destination: no-answer agreement falls by 0.020, while agreement on extracted numeric answers rises by 0.379. Greedy accuracy moves from 0.560 to 0.680.

V3 and R1-Zero occupy different Atlas regimes. V3 concentrates mass into semantic answer basins and improves observed accuracy; R1-Zero changes basin identity while leaving the low-order total spectrum nearly fixed. The accuracy increase for V3 is not itself resolved by this small slice (exact McNemar $p = 0.453$), but the answer-state consolidation is: its collision interval excludes zero and its gain cannot be attributed to formatting or failed extraction.

4.5 Result 3: a no-answer attractor under R1

GPQA exposes a third deformation. Under R1, total collision rises from 0.259 to 0.781 ($\Delta = 0.522$, 95% CI $[0.454, 0.587]$). It rises on 49 of 50 paired prompts and never falls. At the same time, greedy accuracy goes from 0.260 to 0.000: all 13 prompts answered correctly by base are lost and no new prompt is gained (exact McNemar $p = 0.000244$).

The labelled decomposition reveals what the scalar increase hides. Repeated no-answer agreement rises by 0.593, more than the entire net collision increase, while same-extracted-answer agreement *falls* by -0.071. Sampled $\perp$ mass rises from 0.345 to 0.810, and the greedy extractor returns $\perp$ on $41/50$ prompts. Under this fixed protocol, R1 has entered a no-answer attractor. Calling it "more confident" gets the direction of the epistemic change wrong.

The three checkpoint results occupy distinct regimes: V3 produces beneficial semantic consolidation;

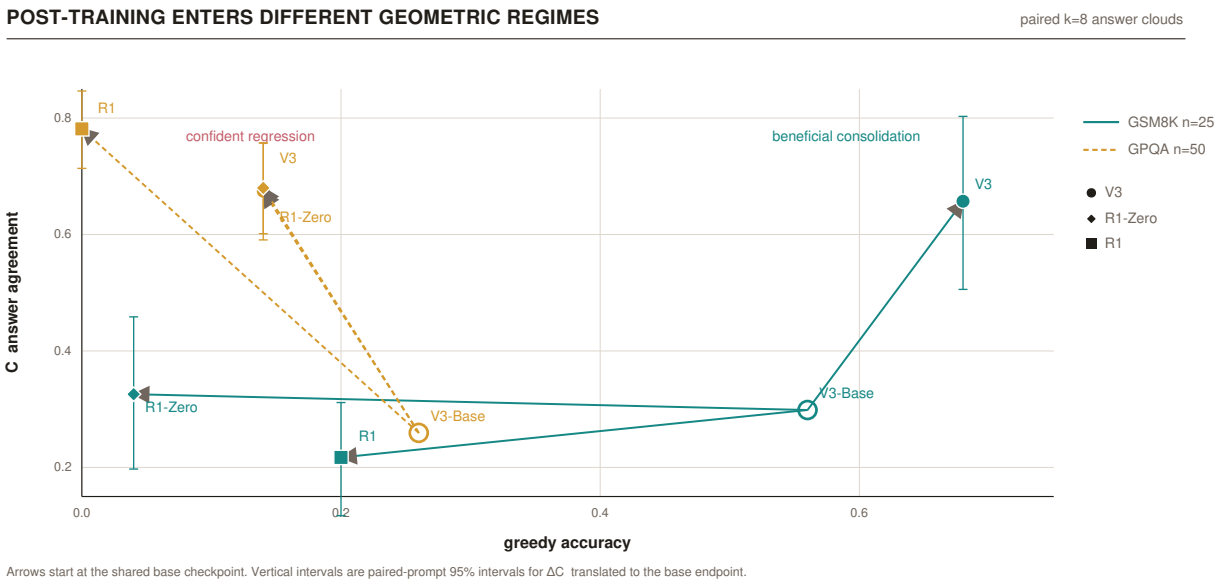

Figure 4 Three post-training deformations in one Atlas. Paired sibling checkpoints move through accuracy–collision space in operationally different directions: beneficial consolidation, near-isospectral wrong-basin inversion, and no-answer collapse. Lines connect each branch to its shared base on the declared task slice.

R1-Zero produces near-isospectral migration into wrong basins; R1 on GPQA produces destructive consolidation into $\perp$. Greater concentration is compatible with all three. The labelled answer-state atlas distinguishes them.

4.6 Result 4: dual observer-flat wounds

The checkpoint transitions require the full labelled state for this analysis; the declared one-pass observer cannot substitute for it.

Geometry changes while the observer is fixed. On GSM8K prompt 6 under V3-Base, the baseline empirical cloud is

$$\{260, 260, 260, \perp\}. \tag{56}$$

The two recorded final-radial clouds are

$$\{1.2, 260, 3, 35\}, \qquad \{1.2, 260, 260, 35\}. \tag{57}$$

In both records the correct greedy answer 260 is bit-identical to baseline. So are all recorded prompt-step coordinates: H_1, H_2, V_1, V_2, maximum probability, and margin. Nevertheless, empirical H_2 moves from 0.470 to 1.386 in the first run and 0.981 in the second; the corresponding empirical TVs are 0.750 and 0.500.

Semantic identity changes while observer and shape are fixed. On GPQA prompt 21 under V3-Base, the baseline cloud is $\{B, D, D, \perp\}$. Both recorded final-radial runs produce $\{C, D, D, \perp\}$, where C is the gold answer. The wrong greedy answer D and the full prompt-step observer remain bit-identical. The frequency multiset $(2, 1, 1)$ is also unchanged, so the quartet, $H_2 = 0.981$, and every empirical

power sum C_r are unchanged. Yet labelled empirical TV is 0.250. A complete shape spectrum cannot report which semantic answer owns each mass.

These are empirical-cloud wounds at $k = 4$. They establish the recorded observer/answer separation and the nested shape/identity distinction; population answer-state separation requires a powered replication.

4.7 Result 5: the strict observer audit

The perturbation study is separate from the checkpoint comparison. It selected two four-prompt buckets: a GSM8K bucket where V3 improved over base and a GPQA bucket where V3 and base were both wrong. For V3-Base and V3 it ran embedding Gaussian, final-layer radial, and final-layer tangent perturbations at scale 0.10, with $k = 4$ samples and two draws for radial/tangent families. Selection and small k make these design cells discriminators, not prevalence units.

The historical audit compared differences of *condition means*, allowing prompt-level changes to cancel. We instead define the recorded observer for prompt i as

$$S_i = (H_{1,i}, H_{2,i}, V_{1,i}, V_{2,i}, p_{\max,i}, m_i, a_i^{\text{greedy}}). \tag{58}$$

Exact equality requires every coordinate of $(S_1, \ldots, S_4)$ to match. For visualization, numeric discrepancy is the maximum absolute change across all six real-valued coordinates and four prompts; greedy equality is checked separately. Two of 20 nonbaseline conditions cannot be fully reconstructed locally: one lacks perturbation-side prompt metrics because of a legacy receipt orientation, and one lacks prompt rows. Among the 18 strictly auditable conditions, 7 preserve the entire four-prompt recorded observer exactly. Of those exact observer collisions, 5 still produce different semantically normalized answer histograms. This count uses promptwise coordinate equality; it does not compare condition means, so opposite prompt changes cannot cancel into a false zero.

The later layer-banded run found answer motion in $0/8$ radial/tangent conditions. That V3-only run constrains the reproducibility claim for those perturbation cells, but it does not rerun either V3-Base opening wound. The local bundle supports two strict conclusions. Exact prompt-observer collisions with different empirical clouds exist in the selected audit; a stable population perturbation law has not yet been established.

5 Implications, Limits, and the Next Test

Catastrophic forgetting has a geometry. The R1-Zero transition makes post-training forgetting directly measurable. On the fixed GSM8K slice, accuracy falls from 0.560 to 0.040 as gold-aligned two-, three-, and four-way agreement collapses and wrong-answer agreement replaces it order by order. The total agreement spectrum barely registers the exchange. Under the declared protocol, this is a geometric signature of catastrophic forgetting: probability mass that repeatedly occupied the correct basin is transported into repeatable wrong basins. Accuracy records the loss; the labelled spectrum shows what the model became.

A candidate pretraining-to-RL handoff signal. The same Atlas suggests a practical criterion for when a foundation checkpoint is becoming a good substrate for reinforcement learning. Loss and greedy benchmark accuracy are insufficient to define an RL-ready pretrain. It should preserve useful answer diversity while increasing gold-aligned agreement, keeping wrong-basin and no-answer power

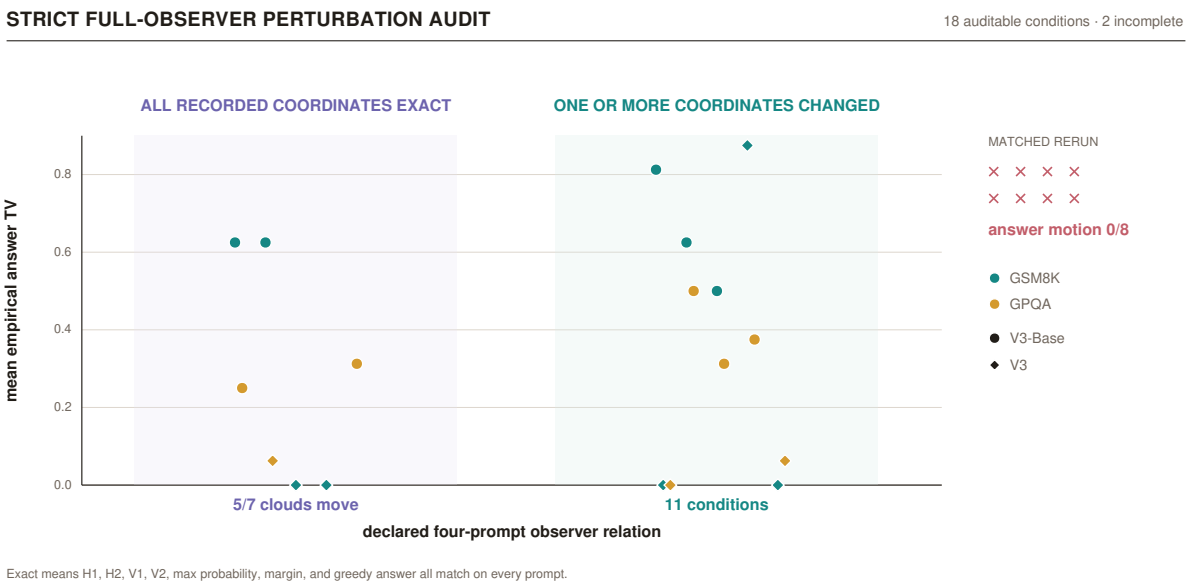

Figure 5 Strict observer audit. Each point is one auditable four-prompt condition, grouped by exact equality versus any change in the declared observer; vertical position is mean empirical total variation of the normalized $k = 4$ answer histograms. Exact means every recorded numeric coordinate and greedy answer matches promptwise. The later V3-only targeted rerun is shown separately.

visible, and avoiding observer fibers in which large answer motion is hidden from cheap diagnostics. Along a pretraining trajectory, stabilization of the labelled agreement spectrum, including its semantic components, could provide an early signal that the representation is approaching a productive handoff point. The current sibling-checkpoint artifacts suggest this causal predictive utility; they do not yet establish it. The required longitudinal test measures the Atlas across pretraining checkpoints, launches matched RL branches, and asks whether the pre-RL geometry predicts which branches consolidate competence and which invert it. That branch comparison tests prediction; a matched intervention on Atlas state is required to test causality.

This answer-state signal is complementary to the companion MoE Atlas hypothesis [5]. The companion analyzes internal routed-state geometry and its predeclared lifecycle boundary; this paper analyzes the protocol-conditioned answer distribution, agreement spectrum, and epistemic inversion. A joint longitudinal test should measure both before the same matched RL branches. No current result establishes that either Atlas determines the other.

Test-time compute is directional. Self-consistency is often treated as a monotone improvement operator. Corollary 5 shows why that intuition fails. Once collision exceeds one half, plurality voting converges exponentially to the modal answer, regardless of truth. Sampling is therefore diagnostic before it is corrective. In a gold-aware offline audit, or with an external verifier, the labelled Atlas can identify a wrong mode; it can expose no-answer dominance without either. An unqualified vote can then amplify exactly that failure. The R1-Zero spectrum is the empirical warning: high-order wrong agreement can grow behind almost unchanged totals.

The manifold and Atlas specify the mechanism. There are three different geometric objects. The local activation manifold $\mathcal{M}_{\theta,x}$ describes reachable internal states. The answer map P_x sends those states

to the labelled statistical manifold $\Delta(\overline{\mathcal{A}}_x)$. The observer map S_x sends them to a confidence surface. The power-sum atlas supplies complete local coordinates for finite-support answer-state shape on regular quotient strata; the occupancy chart restores semantic location. The hidden-rank law then states when a confidence surface must collapse answer-active directions. This vocabulary is doing exact work: it distinguishes motion in activation space, motion in distributional shape, and motion between semantic answer basins.

The receipts do not reconstruct a global latent energy landscape. Modal basins in the answer simplex are defined exactly. Activation-space basins are their pullbacks under P_x. The global topology of those pullbacks remains an empirical question, while the local factorization and rank statements already hold on the declared operational manifold.

Interventions expose the quotient. The two V3-Base wounds are fixed-checkpoint interventions, not correlations across independently trained models. In the GSM8K wound, the complete recorded observer and greedy answer are bit-identical while the empirical answer cloud fragments. In the GPQA wound, even every permutation-invariant empirical shape statistic is fixed while one answer identity moves from wrong to gold. Across the strict reconstruction, 7 of 18 conditions preserve every recorded observer coordinate exactly, and 5 still move their normalized empirical histograms. These finite-cloud observations are consistent with the quotient structure and motivate a powered population-level fiber test; they do not alone establish separated population states.

Boundary of the current evidence. The checkpoint results cover 25 GSM8K prompts and 50 GPQA prompts at $k = 8$; their paired intervals describe these fixed slices, not all prompts or model families. The intervention wounds use $k = 4$ empirical clouds, so they are exact statements about recorded samples rather than powered certificates that two population answer states differ. The perturbation buckets were selected from prior outcomes, share prompts and baselines, and contain two locally incomplete conditions. A later V3-only targeted rerun found answer motion in $0/8$ conditions; it constrains a stable V3 perturbation mechanism but does not rerun either V3-Base wound. The original launch command is absent from the local stage-1 bundle, so provenance is receipt-level rather than bit-identical-launch complete.

The extractor is part of the chart. An answer state is defined jointly by decoding and semantic extraction. Decimal normalization removes a real GSM8K false split; retaining $\perp$ reveals that GPQA R1 concentration is chiefly a no-answer attractor. A richer extractor could separate refusal, malformed formatting, ambiguity, and partial correctness. Those are not nuisance variants of one hidden state; they are different label-aware charts over the same generated text distribution. Any deployment of the Atlas should publish the ontology and report both $p(\perp)$ and conditional extracted-answer geometry.

The next experiment. The next study should pre-register full task slices, multiple base-model families, a longitudinal sequence of pretraining checkpoints, and matched post-training branches. It should increase k enough for prompt-level state separation, compare the declared logit observer with trained semantic-entropy probes, and test two prospective predictions: whether pre-RL labelled geometry forecasts post-training gains or inversions, and whether Atlas-aware test-time-compute allocation outperforms unconditional self-consistency. Matched branches would test predictive utility for model selection; a matched Atlas intervention would test causal control.

6 Related Work

Semantic uncertainty. Semantic uncertainty replaces token-string diversity with uncertainty over meaning-equivalent answer classes [6]. Kernel language entropy generalizes this idea using graded semantic similarity [7]. Our protocol-defined answer state is in this lineage; we do not claim to have invented semantic answer distributions. The contribution here is to treat the full labelled pushforward distribution as the state, separate it from its quartet signature, decompose extraction failure exactly, and study how sibling post-training checkpoints and specified observers reshape or fail to identify it.

One-pass uncertainty probes. Semantic entropy probes learn to predict sample-based semantic uncertainty from hidden states of a single generation and show that useful one-pass recovery is often possible [8]. Semantic Self-Distillation instead distills sampled semantic distributions into a lightweight student that predicts a prompt-conditioned semantic distribution before answer generation [9]. These learned observers are important counterweights to any universal insufficiency claim: in our language, they enlarge the observer class and may reduce or close a legibility gap. Our theorem is observer-relative, while the empirical audit here concerns declared logged summaries. We do not claim that learned semantic observers fail, and the current experiments compare against neither SEP nor SSD.

Self-consistency and test-time compute. Self-consistency samples diverse reasoning paths and selects the modal answer, often improving reasoning accuracy [10]. Work on test-time scaling studies when extra inference compute is worthwhile [11]. Equation (11) isolates a measurement primitive inside that procedure: independent answer agreement is an unbiased probe of order-2 answer geometry. Voting and measurement are related but distinct uses of the same samples.

Calibration, self-knowledge, and post-training. Calibration asks whether confidence predicts correctness [12]; model self-knowledge studies whether language models can report what they know [13]. Confidence elicitation after RLHF can behave differently from conditional token probability, and recent work reports post-training-related overconfidence and ownership effects [14–16]. Our object is not a calibration curve. Two states may have equal correctness and different collision, or equal collision and different modal answers. The paired DeepSeek receipts add an answer-distribution view of post-training, while the no-answer decomposition prevents confidence language from hiding extraction collapse.

Latent knowledge and bounded observers. Latent-knowledge work asks whether internal representations encode facts not expressed in outputs [17]. Bounded-observer information formalizes the dependence of usable information on a computational class [18]. Our observer fiber is a concrete version of this principle for answer-state targets. The target is behavioral geometry under a declared protocol, not an assertion that one unique latent belief exists inside the network.

Local model geometry. Weight-space neighborhood methods such as neural thickets use perturbations to expose nearby specialists [19]. Our perturbations act at fixed weights in activation space. RMS normalization supplies a precise local radial/tangential chart, but the present receipts do not recover a basin landscape. Both approaches probe neighborhoods. Their geometric objects and evidence remain different.

Companion routed-state geometry. The companion MoE Atlas manuscript formalizes the internal routed-state object of an RMS-normalized sparse MoE and a candidate lifecycle boundary for pretraining readiness and geometric forgetting [5]. The present paper studies a different object: the protocol-conditioned distribution over semantic answers, its agreement spectrum, and post-training epistemic inversion. The manuscripts share RMS-normalized local geometry and rank/fiber ingredients; here those ingredients are specialized to answer-map legibility. No theorem or artifact currently shows that the routed-state Atlas determines the answer-state Atlas, or conversely.

7 Conclusion

A language model occupies a protocol-defined answer state: a point in a labelled simplex, with modal basins, tie boundaries, shape coordinates, and semantic location. The answer-state Atlas combines complementary coordinates: under a known finite support bound, the agreement spectrum $C_2, \ldots, C_m$ coordinatizes the regular unlabeled shape manifold, while the occupancy chart restores the answer identities that every permutation-invariant statistic necessarily erases. Self-consistency uses extra samples for Monte Carlo measurement on this atlas and for voting.

The empirical centerpiece is a near-isospectral epistemic inversion. On the paired GSM8K slice, R1-Zero produces small estimated point changes in total two-, three-, and four-way agreement relative to the opposing labelled components; every total-delta interval includes zero, while agreement on the gold answer collapses and agreement on wrong answers rises at every order. Greedy accuracy falls from 0.560 to 0.040, and the empirical modal-answer sets become disjoint on 23 of 25 prompts. A scalar concentration story calls this transition almost stationary. The labelled answer atlas reveals a migration from the correct basin into wrong basins.

The sibling checkpoints expose the other directions of the same geometry. From the same base model, V3 consolidates agreement around the gold GSM8K answer, while R1-Zero consolidates around wrong answers. On GPQA, R1 enters a no-answer attractor: collision rises because mass collapses into $\perp$, not because the model has resolved the task. Post-training transports probability mass between answer basins, and transformations with similar scalar signatures can have opposite epistemic meaning.

The activation manifold explains how this information can disappear from one-pass readouts. The answer map P_x and confidence observer S_x are distinct projections of the same operational neighborhood. Recoverability is equivalent to factorization through S_x; variation inside an observer fiber imposes irreducible error. The hidden-rank law makes the obstruction inevitable whenever the answer map carries more independent local directions than the observer. The two controlled wounds provide finite-cloud representatives of both levels of loss: a fixed recorded confidence surface accompanies changing empirical shape, and even a fixed complete empirical shape spectrum accompanies changing semantic identity. They motivate, but do not by themselves certify, separated population states on one fiber.

Repeated sampling penetrates such fibers by measuring answer agreement without guaranteeing correction. Once collision exceeds one half, plurality voting stabilizes exponentially toward the modal basin whether that basin is correct or catastrophically wrong. Confidence, correctness, and legibility must therefore remain separate axes of the atlas.

Hidden answer geometry names the missing object. The confidence surface is not the answer state: it is one quotient of the model's operational manifold, while the answer atlas is another. A post-training branch can leave estimated label-blind concentration nearly stationary while inverting

the Atlas's alignment with truth.

References

[1] DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[2] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 645:633–638, 2025.

[3] 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.

[4] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

[5] Noumena. One boundary, two crossings: The MoE atlas of pretraining readiness and geometric forgetting. Unpublished companion manuscript, 2026.

[6] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

[7] Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *arXiv preprint arXiv:2405.20003*, 2024.

[8] 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.

[9] Edward Phillips, Sean Wu, Fredrik K. Gustafsson, Boyan Gao, and David A. Clifton. Semantic self-distillation for language model uncertainty. *arXiv preprint arXiv:2602.04577*, 2026.

[10] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[11] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

[13] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[14] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

[15] Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in llms: Reward calibration in rlhf. *arXiv preprint arXiv:2410.09724*, 2024.

[16] Mario Sanz-Guerrero, Manuel Mager, and Katharina von der Wense. Large language models are overconfident in their own responses. *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

[17] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

[18] Marc Finzi, Shikai Qiu, Yiding Jiang, Pavel Izmailov, J. Zico Kolter, and Andrew Gordon Wilson. From entropy to epiplexity: Rethinking information for computationally bounded intelligence. *arXiv preprint arXiv:2601.03220*, 2026.

[19] Yulu Gan and Phillip Isola. Neural thickets: Diverse task experts are dense around pretrained weights. *arXiv preprint arXiv:2603.12228*, 2026.

A Appendix

A.1 Proofs

Proof of Theorem 1. For a fixed r-subset $I \subseteq [k]$, independence gives

$$\mathbb{E}[\mathbf{1}\{A_i = A_j \text{ for all } i, j \in I\}] = \sum_a p(a)^r = C_r(p). \tag{59}$$

Averaging over the $\binom{k}{r}$ subsets proves unbiasedness. The same calculation with r fresh draws proves

$$C_r(p) = \Pr(A_1 = \cdots = A_r) = \exp\big((1-r)H_r(p)\big). \tag{60}$$

Changing one sample A_i can alter only the $\binom{k-1}{r-1}$ kernels whose index set contains i. Hence the bounded-difference constant is

$$\frac{\binom{k-1}{r-1}}{\binom{k}{r}} = \frac{r}{k}. \tag{61}$$

McDiarmid's inequality now yields

$$\Pr\big(|U_{r,k} - \mathbb{E}U_{r,k}| \geq t\big) \leq 2\exp\left(-\frac{2t^2}{k(r/k)^2}\right) \tag{62}$$

$$= 2\exp\left(-\frac{2kt^2}{r^2}\right). \tag{63}$$

Finally, grouping equal kernels by their common answer label gives the count form

$$U_{r,k} = \frac{\sum_a \binom{N_a}{r}}{\binom{k}{r}}. \tag{64}$$

□

Truth-aligned agreement decomposition. Let the answer alphabet be partitioned into the gold answer, extracted wrong answers, and $\perp$. For every integer $r \geq 2$,

$$\begin{aligned} C_r(p) &= C_r^{\text{gold}}(p) + C_r^{\text{wrong}}(p) + C_r^{\perp}(p), \\ C_r^{\text{gold}}(p) &= p(a_{\text{gold}})^r, \\ C_r^{\text{wrong}}(p) &= \sum_{a \notin \{a_{\text{gold}}, \perp\}} p(a)^r, \\ C_r^{\perp}(p) &= p(\perp)^r. \end{aligned} \tag{65}$$

The sample statistic decomposes identically by restricting the count sum $\sum_a \binom{N_a}{r}$ to the same three sets. Each component is unbiased by the kernel argument above. Table 1 uses this exact decomposition: an unchanged total can conceal equal-and-opposite transport between correct and wrong agreement.

Proof of Corollary 1. Write $X_{ij} = \mathbf{1}\{A_i = A_j\}$, $M = \binom{k}{2}$, $c = C_2(p)$, and $t_3 = C_3(p)$. Disjoint pairs of indices give independent indicators. For identical and one-index-overlapping pairs,

$$\mathrm{Var}(X_{12}) = c - c^2, \tag{66}$$

$$\mathrm{Cov}(X_{12}, X_{13}) = t_3 - c^2. \tag{67}$$

There are M variance terms and $6\binom{k}{3}$ ordered covariance terms with one shared index. Therefore

$$\mathrm{Var}(U_{2,k}) = \frac{M(c - c^2) + 6\binom{k}{3}(t_3 - c^2)}{M^2} \tag{68}$$

$$= \frac{2(c - c^2) + 4(k-2)(t_3 - c^2)}{k(k-1)}. \tag{69}$$

Moreover, $t_3 - c^2 = \mathrm{Var}_{A\sim p}[p(A)] \geq 0$, while $t_3 \leq c$ and $c - c^2 \leq 1/4$. Thus

$$\mathrm{Var}(U_{2,k}) \leq \frac{2/4 + 4(k-2)/4}{k(k-1)} = \frac{k - 3/2}{k(k-1)} \leq \frac{1}{k}. \tag{70}$$

Jensen's inequality gives

$$\mathbb{E}|U_{2,k} - c| \leq \sqrt{\mathrm{Var}(U_{2,k})} \leq k^{-1/2}. \tag{71}$$

□

Proof of Theorem 2. Let e_j denote the jth elementary symmetric polynomial in $p_1, \ldots, p_m$, with $e_0 = 1$. Newton's identities give recursively

$$j e_j = \sum_{r=1}^{j} (-1)^{r-1} e_{j-r} C_r, \qquad 1 \leq j \leq m. \tag{72}$$

Thus $C_1, \ldots, C_m$ determine $e_1, \ldots, e_m$, which determine the monic polynomial

$$z^m - e_1 z^{m-1} + e_2 z^{m-2} - \cdots + (-1)^m e_m = \prod_{i=1}^{m} (z - p_i). \tag{73}$$

Its multiset of roots is exactly $\{p_1, \ldots, p_m\}$. Since $C_1 = \sum_i p_i = 1$ on the simplex, $C_2, \ldots, C_m$ suffice. This reconstruction is permutation-invariant by construction and includes zero-mass answers when a support of size less than m is padded with zeros. Without a known finite support bound, no finite prefix is asserted to be complete.

For the local statement, consider the augmented map $(C_1, \ldots, C_m)$ on $\mathbb{R}^m$. Its Jacobian has entries

$$\frac{\partial C_r}{\partial p_j} = r p_j^{r-1}, \tag{74}$$

so factoring r from row r leaves a Vandermonde matrix and gives

$$\det\left[\frac{\partial C_r}{\partial p_j}\right]_{r,j=1}^{m} = m! \prod_{i<j} (p_j - p_i). \tag{75}$$

This determinant is nonzero when the masses are pairwise distinct. Restricting to the simplex fixes $C_1 = 1$ and removes its normal direction, so $(C_2, \dots, C_m)$ has nonsingular differential on every strictly ordered chamber. The inverse-function theorem supplies a smooth local chart there. Permuting labels moves between chambers representing the same quotient point. At equal masses the Vandermonde vanishes exactly where the permutation action has nontrivial stabilizer; these are the quotient's symmetry strata, on which the global invariant remains valid. □

Proof of Theorem 3. Because d is a metric, $\Gamma_S^f(U) = 0$ if and only if f is constant on each nonempty fiber $U \cap S^{-1}(s)$. In that case define

$$R(s) := f(h) \qquad \text{for any } h \in U \text{ with } S(h) = s. \tag{76}$$

Fiber constancy makes R well-defined, and $f = R \circ S$. Conversely, any such factorization makes f constant on every fiber. The map is unique on $S(U)$ because every $s \in S(U)$ has a preimage.

Now choose h_0, h_1 in one observer fiber and write $f_j = f(h_j)$. A possibly randomized observer-only estimator has the same conditional output law Z at both points. Coupling the two outputs by this common law and applying the triangle inequality gives

$$d(f_0, f_1) \le d(f_0, Z) + d(Z, f_1). \tag{77}$$

After expectation, at least one of the two risks is at least $d(f_0, f_1)/2$. Taking pairs whose separation approaches the fiber-diameter supremum, followed by the infimum over estimators, proves Equation (41).

The equivalence just proved is set-theoretic. Fiber constancy alone need not make R measurable for arbitrary measurable spaces. When f is $\sigma(S)$-measurable and $\mathcal{T}$ is standard Borel, the Doob–Dynkin lemma supplies a measurable version of R. The smooth local factorization used below needs neither a global quotient nor this additional measurable-space assumption. □

Proof of Corollary 2. Replacing S by $\widetilde{S} = Q_\tau \circ S$ makes points in the same recorded bucket members of an exact observer fiber, so Theorem 3 applies directly.

For the near-flat statement, let $e_j = d(\psi(S_j), f_j)$. The triangle inequality and L-Lipschitzness give

$$d(f_0, f_1) \le e_0 + d\big(\psi(S_0), \psi(S_1)\big) + e_1 \tag{78}$$

$$\le e_0 + L\|S_0 - S_1\| + e_1. \tag{79}$$

Hence $\max(e_0, e_1)$ is at least half the positive part of $d(f_0, f_1) - L\|S_0 - S_1\|$. Without the quantizer or the Lipschitz restriction, distinct observer values can be separated by a discontinuous decoder, so numerical nearness alone proves no estimator-independent bound. □

Proof of Theorem 4. Let $A = DS_h$, $B = Df_h$, and $L = (A, B)$. Projection onto the first coordinate defines a surjective linear map

$$\pi : \operatorname{im} L \longrightarrow \operatorname{im} A, \qquad \pi(Av, Bv) = Av. \tag{80}$$

Its kernel is

$$\ker \pi = \{(0, Bv) : v \in \ker A\}, \tag{81}$$

whose dimension is $\operatorname{rank}(B|_{\ker A})$. Rank–nullity therefore gives

$$\operatorname{rank} D(S,f)_h - \operatorname{rank} DS_h = \operatorname{rank}(B|_{\ker A}). \tag{82}$$

If W is a complement of $\ker A$, then $\dim W = \operatorname{rank} A$ and

$$\operatorname{im} B = B(\ker A) + B(W). \tag{83}$$

Consequently

$$\operatorname{rank} B \le \operatorname{rank}(B|_{\ker A}) + \operatorname{rank} A, \tag{84}$$

which proves the inequality. The restriction has positive rank exactly when some $v \in \ker DS_h$ satisfies $Df_h[v] \neq 0$. □

Proof of Corollary 3. If S has constant rank near h, the constant-rank theorem makes the local level set $S^{-1}(S(h))$ an embedded submanifold with tangent space $\ker DS_h$. Every v in that tangent space is the velocity of a curve γ lying entirely in the level set. When $Df_h[v] \neq 0$,

$$S(\gamma(t)) = S(h), \tag{85}$$
$$f(\gamma(t)) = f(h) + tDf_h[v] + o(t) \tag{86}$$

in a target chart. Thus the target moves inside an exact observer fiber, giving positive fiber diameter in every sufficiently small neighborhood.

Conversely, constant-rank coordinates put S locally in the form $S(u,v) = (u,0)$. The condition $\ell_f = 0$ says that Df annihilates every v-direction. Hence $f(u,v)$ is constant on each connected local fiber. Defining $R(u) = f(u,v)$ for any v on that fiber gives a well-defined smooth map with $f = R \circ S$ locally. □

Proof of Corollary 4. Apply Theorem 3 to the scalar target $f = C_2 \circ P$. The two states lie in one observer fiber and have target separation γ, so every observer-only predictor incurs maximum expected absolute error at least $\gamma/2$. Corollary 1 gives

$$\max_{j\in\{0,1\}} \mathbb{E}_j |U_{2,k} - C_2(P(h_j))| \le k^{-1/2}. \tag{87}$$

Therefore $k > 4/\gamma^2$ makes the sampling risk strictly smaller than the observer-only lower bound.

Assume without loss of generality that $c_0 < c_1$ and threshold $U_{2,k}$ at $(c_0 + c_1)/2$. An error under either state implies

$$|U_{2,k} - c_j| \ge \gamma/2. \tag{88}$$

Equation (29) with $r = 2$ bounds this probability by $2\exp(-k\gamma^2/8)$. In contrast, identical deterministic observer values induce identical observer laws under the two states; under equal priors every observer-only test therefore has error $1/2$. □

Proof of Corollary 5. Let $p_{\max} = p(a_{\max})$. Since $p(a)^2 \le p_{\max} p(a)$ for every answer,

$$c = \sum_a p(a)^2 \le p_{\max} \sum_a p(a) = p_{\max}. \tag{89}$$

Thus $c > 1/2$ implies $p_{\max} > 1/2$, so the mode is unique. If $N_{\max} \sim \text{Binomial}(k, p_{\max})$ is the number of modal samples, a plurality loss requires $N_{\max} \leq k/2$. Hoeffding's inequality gives

$$\Pr(\widehat{a}_k \neq a_{\max}) \leq \Pr(N_{\max} \leq k/2) \tag{90}$$

$$\leq \exp\left[-2k\left(p_{\max} - \frac{1}{2}\right)^2\right] \tag{91}$$

$$\leq \exp\left[-2k\left(c - \frac{1}{2}\right)^2\right]. \tag{92}$$

The argument establishes modal stability; correctness never enters it. □

A.2 RMS normalization derivation

Let $s(h) = (\|h\|_2^2/d + \epsilon)^{1/2}$ and $N(h) = h/s(h)$. Differentiation gives

$$Ds_h[v] = \frac{h^\top v}{d\, s(h)}, \qquad DN_h[v] = \frac{v}{s(h)} - \frac{h\, h^\top v}{d\, s(h)^3}. \tag{93}$$

If $h^\top v = 0$, then $DN_h[v] = v/s(h)$. For $v = h$,

$$DN_h[h] = \left(\frac{1}{s(h)} - \frac{\|h\|_2^2}{d\, s(h)^3}\right) h = \frac{\epsilon}{s(h)^3} h. \tag{94}$$

Thus the tangential and radial eigenvalues are $1/s(h)$ and $\epsilon/s(h)^3$, respectively. For $\epsilon = 0$ and $h \neq 0$, $\|N(h)\|_2 = \sqrt{d}$. For $\epsilon > 0$,

$$\|N(h)\|_2^2 = \frac{\|h\|_2^2}{\|h\|_2^2/d + \epsilon} < d, \tag{95}$$

and every norm below $\sqrt{d}$ is attained, so the image is the open ball $B_{\sqrt{d}}(0)$.

For a learned diagonal RMSNorm gain W, the post-gain output is $z = WN(h)$ and its derivative is $Dz_h = WDN_h$. The spherical or open-ball core is therefore warped by W. The recorded final-layer interventions are defined directly in this post-gain z-space.

A.3 Quartet identities

For

$$F_p(\alpha) = \log \sum_a p(a)^\alpha, \qquad q_\alpha(a) = \frac{p(a)^\alpha}{\sum_b p(b)^\alpha}, \tag{96}$$

differentiation over the positive support of p yields

$$F_p'(\alpha) = \mathbb{E}_{q_\alpha}[\log p(A)], \tag{97}$$

$$F_p''(\alpha) = \text{Var}_{q_\alpha}[\log p(A)]. \tag{98}$$

At $\alpha = 1$, $q_1 = p$, giving $H_1 = -F_p'(1)$ and $V_1 = F_p''(1)$. At $\alpha = 2$, $F_p(2) = \log \sum_a p(a)^2 = -H_2$ and q_2 is the order-2 escort, giving $V_2 = F_p''(2)$. We use the conventions $0 \log 0 = 0$ and take surprisals only over $\text{supp}(p)$.

The signature is not injective. If $p = (0.9, 0.1)$ and $p' = (0.1, 0.9)$ on two labelled answers, then $G(p) = G(p')$ but $\text{TV}(p, p') = 0.8$ and the modal answers differ. This noninjectivity allows the prompt-level GPQA label-switch witness to have nonzero state transport and zero quartet movement.

A.4 Complete normalized stage-1 summaries

Table 3 gives the plug-in $\widehat{H}_2$ summaries after GSM8K Decimal normalization. These descriptive values are kept separate from the unbiased collision analysis in the main text.

Task	Checkpoint	accuracy	wrong n	mean $\widehat{H}_2$	wrong-only $\widehat{H}_2$
GSM8K	V3-Base	0.560	11	1.055	1.246
GSM8K	V3	0.680	8	0.551	1.293
GSM8K	R1-Zero	0.040	24	1.006	1.014
GSM8K	R1	0.200	20	1.195	1.230
GPQA	V3-Base	0.260	37	1.074	1.063
GPQA	V3	0.140	43	0.399	0.354
GPQA	R1-Zero	0.140	43	0.396	0.381
GPQA	R1	0.000	50	0.256	0.256

Table 3 Semantically normalized empirical answer-cloud entropies. Values are means of per-prompt plug-in estimates at $k = 8$ and are not substituted for population H_2.

A.5 Complete perturbation audit

Table 4 enumerates all 20 nonbaseline design conditions. The observer column is the maximum absolute change across prompt-step H_1, H_2, V_1, V_2, maximum probability, and margin over the four prompts; exact-flat classification additionally requires identical greedy answers. TV is the mean prompt-level empirical total variation between the two $k = 4$ histograms after numeric normalization. A dash marks a quantity that is not recoverable from the local receipt bundle.

Task	checkpoint	perturbation	draw	max raw $\lvert\Delta S_j\rvert$	mean empirical TV	$\lVert\Delta\overline{G}\rVert_2$
gsm8k	v3_base	embed_gaussian	1	3.12	0.8125	0.1768
gsm8k	v3_base	final_radial	1	0	0.6250	0.0000
gsm8k	v3_base	final_radial	2	0	0.6250	0.4530
gsm8k	v3_base	final_tangent	1	0.00379	0.6250	0.1393
gsm8k	v3_base	final_tangent	2	0.0022	0.5000	0.1768
gsm8k	v3	embed_gaussian	1	7.68	0.8750	1.3947
gsm8k	v3	final_radial	1	0	0.0000	0.0000
gsm8k	v3	final_radial	2	0	0.0000	0.0000
gsm8k	v3	final_tangent	1	0.00115	0.0000	0.0000
gsm8k	v3	final_tangent	2	0.001	0.0000	0.0000
gpqa	v3_base	embed_gaussian	1	4.47	0.5000	0.1393
gpqa	v3_base	final_radial	1	0	0.3125	0.0000
gpqa	v3_base	final_radial	2	0	0.2500	0.1768
gpqa	v3_base	final_tangent	1	0.00272	0.3125	0.2391
gpqa	v3_base	final_tangent	2	0.00364	0.3750	0.4323
gpqa	v3	embed_gaussian	1	–	0.7500	0.6237
gpqa	v3	final_radial	1	0	0.0625	0.1768
gpqa	v3	final_radial	2	–	–	–
gpqa	v3	final_tangent	1	0.00278	0.0625	0.0000
gpqa	v3	final_tangent	2	0.00274	0.0000	0.0000

Table 4 Exhaustive selected perturbation design-cell audit. Rows share prompts and baselines and are not independent prevalence samples. The observer column is descriptive because its coordinates have different scales; only exact zero is used inferentially. $\Delta\overline{G}$ can cancel prompt-level changes, so TV is the primary empirical-cloud column.

At exact full-observer equality, $7/18$ auditable conditions are flat and 5 have nonzero sampled-

state TV. No tolerance-based count is promoted across the heterogeneous raw coordinates. Two additional conditions are observer-incomplete locally. The targeted rerun has $0/8$ conditions with answer motion.

A.6 Corrections relative to the historical summaries

The mechanical rebuild intentionally supersedes the sibling markdown summaries. The audit found:

1. GSM8K treated Decimal-equivalent strings such as `26` and `26.00` as different answers. Canonicalization removes four previously reported V3 radial/tangent movements and changes the GSM8K population table.
2. The legacy perturbation join reverses several occupancy-delta signs because most canonical files store perturbation as primary and baseline as comparison.
3. One GPQA V3 embedding file uses the opposite orientation. Its baseline prompt metrics were incorrectly relabelled as perturbation metrics, so its reported zero observer change is not locally supported.
4. Differences of condition means were called observer equality even when prompt-level changes canceled. The rebuilt audit checks every recorded coordinate of the full four-prompt observer.
5. Several hand-written witness rows mixed identifiers and values from different conditions. No such table is retained; the appendix is generated directly from receipts.

These corrections are substantive. They reduce the apparent perturbation effect and change its evidentiary status from denominator-bearing prevalence to a selected witness audit.

A.7 Receipt provenance

The population analysis reads the six files

`occupancy_receipts/{gsm,gpqa}_stage1_{v3,r1_zero,r1}.json`.

Their source paths name the canonical run roots:

- GSM8K: `/data/entropix_receipts/deepseek_stage1/20260308T222339Z_gsm8k_n25_k8_retry1`;
- GPQA: `/data/entropix_receipts/deepseek_stage1/20260309T182221Z_gpqa_n50_k8_parallel1`.

The local JSONs retain prompt indices, extracted counts, greedy predictions, gold answers, and paired posterior rows. The original JSONL files and launch scripts are not copied into this paper directory. The perturbation analysis reads the corresponding prompt-level JSONs under `occupancy_receipts/`, with the manifest in `../perturbation_bucketed_manifest_2026-03-11.md`. The generated artifact `generated/statistics.json` records the normalized analysis in a single machine-readable file.