COORDINATES WITHOUT A COMPASS: GAUGE SYMMETRY AND TRUST ANCHORS IN LLM VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Relational evidence can recover every pairwise truth relation while leaving absolute truth unresolved. We study verification systems whose judgments are correct only up to a shared binary orientation. Flipping both latent truth and this orientation preserves every observation, so no aggregation of such evidence can certify an individual claim. We formalize this ambiguity as a gauge symmetry and show that certification is possible exactly when truth is constant across observationally indistinguishable states. A trusted anchor resolves the ambiguity only when it fixes the target mechanism's orientation. Consequently, in noiseless relational systems, trust cost is determined by the number of independently orientable components, not by the amount of evidence. Experiments support these predictions. On MMLU-Pro, DeepSeek recovers the truth partition with 98.5% accuracy, yet unanchored absolute accuracy remains 50%; one held-out anchor raises it to 97.0%. HumanEvalPlus audits show that anchors help aligned mechanisms but can hurt oppositely oriented ones. A label-free router topology predicts the same within-component advantage in two frozen cohorts. Relational coordinates become a compass only when a mechanism-compatible anchor supplies their orientation.

1 INTRODUCTION

A verifier may determine every relation among a collection of claims and still fail to certify a single one. Across claims $i = 1, \ldots, n$, let

$$G_i = \tau_i \oplus \theta, \qquad \tau_i, \theta \in \{0, 1\}, \tag{1}$$

where τ_i records whether answer i is correct and θ orients the shared verification mechanism. If θ were known, the reports would determine every τ_i. Simultaneously flipping every τ_i and θ preserves the joint law of all G_i while reversing truth. The reports recover every pairwise truth relation, yet the two global orientations remain observationally identical. Agreement fixes the coordinate system, not the direction called true.

Behavioral verification methods infer correctness from repeated samples, semantic agreement, elicited confidence, self-checks, or cross-model judgments (Wang et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Lyu et al., 2025; Manakul et al., 2023; Xiong et al., 2024; Liu, 2026). These signals can improve decisions, but self-consistent and correlated errors show that agreement may reveal relative structure without determining absolute correctness (Tan et al., 2025; Kim et al., 2025; Denisov-Blanch et al., 2026; Brilliant, 2026). Models of latent confounding make the remaining problem explicit: without an external assumption or label, aggregation can recover structure while leaving its sign unresolved (Zhao et al., 2026).

Statistical identification provides a complementary view: which decisions remain determined when the same observations are compatible with different worlds? Blackwell comparison formalizes the decisions supported by an experiment (Blackwell, 1953; Le Cam, 1964; Torgersen, 1991); selective classification studies the coverage–risk tradeoff when a predictor may abstain (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2019). Related work studies label symmetries and the anchors that resolve them (Stephens, 2000; Wei, 2026), while statistical verification can lose its guarantee when proxy data or learned evaluators violate its assumptions (Metevier et al., 2026). Our question lies at the intersection: when does recovered relational structure determine absolute truth?

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.

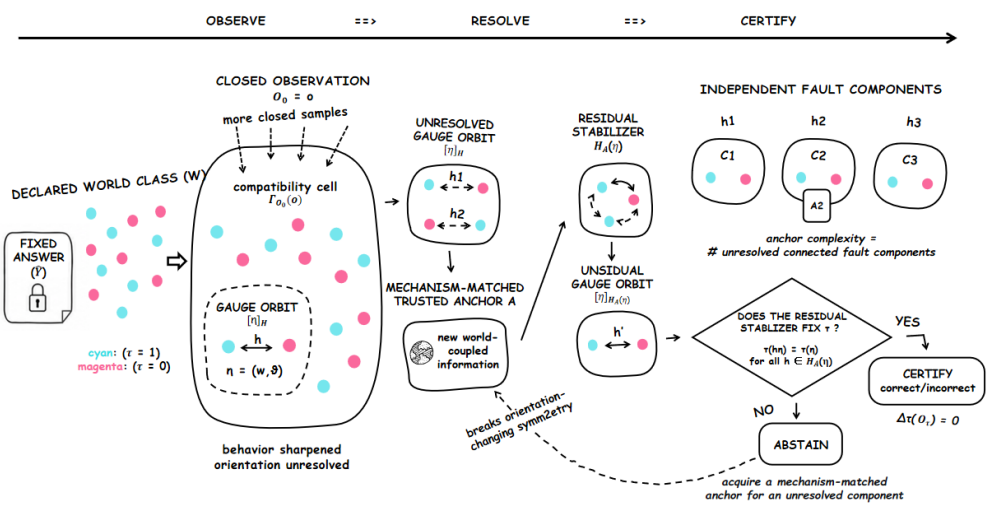

Figure 1: **Latent recovery is not yet certification.** Relative evidence can determine a complete coordinate system while leaving opposite truth orientations observationally equivalent. A mechanism-matched anchor fixes the semantic orientation; evidence from an unrelated fault domain need not.

We answer by separating relational recovery from truth certification. We model an observation-preserving flip as a gauge symmetry. If such a flip also reverses the target truth, no rule using those observations alone can certify the target; certification becomes possible once all observationally compatible states agree. A trusted anchor removes the ambiguity only for mechanisms whose orientation is linked to it. In noiseless relational systems, the number of independently orientable components, rather than the amount of evidence, determines how many anchors are required.

Experiments test these consequences. On MMLU-Pro (Wang et al., 2024), DeepSeek recovers the truth partition with 98.5% accuracy, yet unanchored absolute accuracy remains 50%; one held-out anchor raises it to 97.0%. HumanEvalPlus (Liu et al., 2023) audits show that additional correct labels help aligned mechanisms but can hurt oppositely oriented ones, extending execution-based code verification (Ni et al., 2023). Finally, a label-free router topology predicts the same within-component advantage in two frozen cohorts.

2 Why Perfect Evidence Can Fail

Fix a deployed answer $\widehat{Y}_f$ and a declared world class $\mathcal{W}$. The binary target is

$$\tau(w) = \mathbf{1}\{\widehat{Y}_f = Y(w)\}. \tag{2}$$

An observation o from channel O induces the compatibility cell

$$\Gamma_O(o) = \{w \in \mathcal{W} : K_O(o \mid w) > 0\}. \tag{3}$$

A cell is *truth-homogeneous* when $|\{\tau(w) : w \in \Gamma_O(o)\}| = 1$; write $\mathrm{Cert}(O; \mathcal{W})$ for the set of such observations. An exact certificate may output 0, 1, or abstain and must be correct in every compatible world.

Exact certification has a sharp feasibility condition. A certificate may decide at o precisely when $\Gamma_O(o)$ is truth-homogeneous. Universal decisiveness therefore requires truth homogeneity in every reachable compatibility cell.

2.1 From Compatibility Cells to Gauge Orbits

Equation 3 defines compatible worlds but rarely permits their enumeration. Observation-preserving symmetries expose a certification obstruction without constructing the full cell. Let the expanded

state be $\eta = (w, \vartheta) \in \mathcal{W} \times \Theta$, where ϑ indexes the evidence mechanism and the target remains $\tau(\eta) = \tau(w)$. A group $\mathcal{H}$ acts on expanded states.

An element $h \in \mathcal{H}$ is an *observational gauge symmetry* of experiment O when

$$P^O_{h\eta} = P^O_\eta \qquad \text{for every expanded state } \eta. \tag{4}$$

The orbit $[\eta]_\mathcal{H} = \{h\eta : h \in \mathcal{H}\}$ is its gauge class. Because exact certification is support-sensitive, we call the experiment *orbit-complete* when

$$\operatorname{supp}(P^O_\eta) \cap \operatorname{supp}(P^O_{\eta'}) \neq \varnothing \quad \Longleftrightarrow \quad \eta' \in [\eta]_\mathcal{H}. \tag{5}$$

Thus distinct orbits have disjoint observational supports.

The minimax decision error measures the robust gap to a truth oracle:

$$\Delta_\tau(O) = \inf_Q \sup_\eta \Pr_\eta\{Q(O) \neq \tau(\eta)\}, \tag{6}$$

where Q ranges over randomized binary decoders. Thus $\Delta_\tau(O) = 0$ means robust certification is possible, whereas $1/2$ is the maximal binary gap.

Theorem 1 (Gauge-orbit obstruction)**.** *Every compatibility cell reachable from η contains the reachable part of $[\eta]_\mathcal{H}$. Hence exact certification is impossible if $\tau(h\eta) \neq \tau(\eta)$ for some observational symmetry h. For a binary target, such a truth-reversing symmetry implies the maximal gap*

$$\Delta_\tau(O) = \frac{1}{2}. \tag{7}$$

If O is orbit-complete, every reachable compatibility cell is exactly one gauge orbit, so truth constancy on the orbit is necessary and sufficient for exact certification without explicit world enumeration.

The elementary channel in Equation 1 has $\mathcal{H} = \mathbb{Z}_2$: its nonidentity element maps (τ, θ) to $(1-\tau, 1-\theta)$. Each fixed mechanism is a perfect verifier, but the robust experiment is maximally deficient because its one bit has no grounded orientation. Conditional on θ, the channel carries one full bit, yet its robust certification gap remains $1/2$.

Orbit recovery preserves every relation invariant to the unknown mechanism but leaves target-reversing transformations unresolved. Certification requires an observation that excludes those transformations, not a reconstruction of the entire world.

3 Trust as Symmetry Breaking

As illustrated in Figure 1, an added channel has certification value only insofar as it excludes transformations that still change the target. Certification value therefore measures target-changing symmetry broken, not evidence accumulated. One world-coupled bit may certify what unlimited ungrounded evidence cannot.

3.1 Anchors and stabilizers

Generalized anchor–stabilizer identifiability becomes target-specific: admissible transformations preserve target truth (Wei, 2026). Let an equivariant, world-coupled value $A(\eta)$, revealed independently of the untrusted experiment, serve as a trusted anchor. After its true value is observed, the compatible gauge transformations form the stabilizer

$$\mathcal{H}_A(\eta) = \{h \in \mathcal{H} : h \cdot A(\eta) = A(\eta)\} = \{h \in \mathcal{H} : A(h\eta) = A(\eta)\}. \tag{8}$$

Theorem 2 (Anchor–stabilizer criterion)**.** *Conditioned on a trusted anchor, the residual gauge class is $[\eta]_{\mathcal{H}_A(\eta)}$. The anchor removes the gauge obstruction for target τ_j if and only if*

$$\tau_j(h\eta) = \tau_j(\eta) \quad \textit{for every } h \in \mathcal{H}_A(\eta). \tag{9}$$

When the unanchored experiment is orbit-complete, this condition is also sufficient for exact identification of τ_j.

Anchor value is therefore mechanism-relative. An anchor calibrated under one channel restricts the stabilizer of another channel only if their orientations are linked by a declared invariance assumption. Source identity, similar format, or agreement does not establish that link.

3.2 FAULT TOPOLOGY AND ANCHOR COMPLEXITY

For relational evidence, let $\tau = (\tau_1, \ldots, \tau_n)$ and suppose a graph E releases only

$$R_{ij} = \tau_i \oplus \tau_j, \qquad (i,j) \in E. \tag{10}$$

These relations can reveal the entire truth partition without labeling either side correct.

Corollary 3 (Anchor complexity)**.** *If the relation graph has C connected components, its gauge group is $\mathbb{Z}_2^C$, acting by an independent global flip in each component. Complete edge relations are orbit-complete. Certifying all n targets requires and is achieved by one trusted label in every component. Thus the noiseless anchor complexity is C, not the number of relation judgments or verifiers.*

Proposition 4 (Unbounded anchor value)**.** *For n claims in one connected component, complete noiseless relations have minimax per-claim Hamming error $1/2$ without an anchor, regardless of repeated observation. One trusted label recovers all n truths exactly. Consequently the number of expected errors removed by one anchor is $n/2$ and is unbounded in the number of claims sharing a stable orientation. If every claim is its own component, an anchor removes no error on any remaining unlabeled claim.*

Under noise, relations estimate a partition only up to componentwise flips, while held-out anchors select its orientation.

3.3 NOISY ANCHORS AND THE ORIENTATION-TRANSFER LAW

Let source mechanism s agree with truth on C_s of N trusted items. For an odd number k sampled without replacement, the probability that their majority retains the source's nominal orientation is

$$\pi_s(k) = \sum_{j=(k+1)/2}^{k} \frac{\binom{C_s}{j}\binom{N-C_s}{k-j}}{\binom{N}{k}}. \tag{11}$$

This quantity measures confidence in a *sign*, not confidence in any individual target judgment.

Proposition 5 (Orientation-transfer law)**.** *Let a disjoint target mechanism t have nominal accuracy a_t. Orient all target judgments by the majority of k trusted source anchors. Its exact expected accuracy is*

$$\begin{aligned} A_{s\to t}(k) &= \pi_s(k)a_t + [1-\pi_s(k)](1-a_t), \\ A_{s\to t}(k) - \tfrac{1}{2} &= 2\big[\pi_s(k) - \tfrac{1}{2}\big]\big[a_t - \tfrac{1}{2}\big]. \end{aligned} \tag{12}$$

Consequently transfer is above chance exactly when the anchor majority and target mechanism have the same orientation. Changes in k affect target accuracy only through $\pi_s(k)$; whenever $\pi_s(k)$ rises, accuracy moves toward a_t, improving a target with $a_t > 1/2$ and degrading one with $a_t < 1/2$.

Corollary 6 (Trust amplification is relational)**.** *Suppose $\pi_s(k) > 1/2$ increases with the odd anchor count k. Then $A_{s\to t}(k)$ increases when $a_t > 1/2$, decreases when $a_t < 1/2$, and equals $1/2$ when $a_t = 1/2$. If $\pi_s(k) \to 1$, transfer converges to a_t: the target's better orbit representative when source and target orientations agree, and its worse representative when they disagree. Consequently no scalar property of s alone can guarantee beneficial transfer to every target mechanism.*

No source–target coefficient is fitted. A stable source drives an aligned target toward its better orbit representative and an opposite target toward its worse one; a weak source sign leaves either near chance. If a target-only flip preserves the complete source experiment, the relative sign is unidentified: the same rule helps one compatible world and harms its pair, so its uniform minimax gain is zero. A missing fault-graph edge therefore means no guarantee, not negative transfer. Held-out edge signs turn a transfer matrix into a signed graph: connected sign-consistent mechanisms share one orientation, while unstable regions require separate anchors.

Corollary 7 (Noisy anchor propagation)**.** *For component c of size n_c, let a frozen estimate have orbit error $e_c \le 1/2$. Let each of an odd number k_c of independent or without-replacement anchors vote for the better orientation with probability at least $1/2 + \gamma_c$, where $\gamma_c > 0$. Then the expected Hamming risk of the oriented estimate satisfies*

$$\mathbb{E}[R] \le \sum_{c=1}^{C} \frac{n_c}{n}\left[e_c + (1-2e_c)e^{-2k_c\gamma_c^2}\right], \qquad n = \sum_{c=1}^{C} n_c. \tag{13}$$

Moreover, if

$$k_c \geq \frac{\log(C/\delta)}{2\gamma_c^2} \quad \textit{for every } c, \tag{14}$$

all component orientations are correct with probability at least $1-\delta$*, and the realized global error is the size-weighted orbit error* $\sum_c (n_c/n)e_c$*.*

Noise changes replication within a component, not the number of independent signs: simultaneous orientation costs $O(\sum_c \gamma_c^{-2} \log(C/\delta))$. Stable world-coupled mechanisms can transport orientation to disjoint claims; unlinked mechanisms admit no uniform transfer sign. Approximate symmetry replaces this binary obstruction with a continuous Le Cam decision gap.

4 Experiments

Freezing each judgment isolates the evidential intervention: the claim, model output, and correctness label remain fixed while the certificate receives relations, anchors, or an unreduced world-coupled channel. Correlated model views then reveal whether nominal multiplicity corresponds to distinct fault domains.

Every answer and evidence outcome was frozen before analysis. MMLU-Pro truth comes from the answer key; HumanEvalPlus correctness comes from the unexposed remainder of the EvalPlus tests. Trust labels are used only to orient a component and are removed from its evaluation set. All anchor expectations are exact finite-population hypergeometric averages over odd-sized anchor sets, not fitted estimates.

The MMLU relational audit uses 70 frozen answers; the code audits use 158 eligible HumanEvalPlus programs for each of Qwen2.5-3B (Yang et al., 2024), Phi-3.5-mini (Abdin et al., 2024), and DeepSeek-V4-Flash (DeepSeek-AI, 2026). Evidence mechanisms are fixed before evaluation: real executable probes, probes with corrupted expected outputs, and compile/load checks. Whenever one split identifies direction and another evaluates it, task identifiers are deterministically partitioned into 79 discovery and 79 evaluation programs. Anchor counts are $k \in \{1, 3, 5, 9\}$, all odd to avoid a fitted tie rule. Reported quantities include raw accuracy, orbit accuracy $\max(a, 1-a)$, and expected oriented accuracy in Equation 12; no benchmark item used as an anchor contributes to its own evaluation.

4.1 Near-perfect partitions without an absolute certificate

A controlled *release intervention* isolates orientation in two existing, frozen evidence channels. For MMLU-Pro, J_i is DeepSeek's binary certificate at the fixed 0.9 agreement threshold under either an item-matched or equal-format mismatched reference. For HumanEvalPlus, J_i is whether a frozen program passes three real or three corrupted executable probes. The released experiment contains every equality relation

$$\widetilde{R}_{ij} = \mathbf{1}\{J_i = J_j\} \tag{15}$$

but withholds the individual bits J_i. No gold label enters this release. Because the complete relation graph determines J only up to a global flip, the analysis separates (i) relation accuracy against $\mathbf{1}\{\tau_i = \tau_j\}$, (ii) orbit accuracy $\max\{\Pr(J=\tau), \Pr(1-J=\tau)\}$, and (iii) absolute accuracy after orienting with trusted labels that are excluded from evaluation.

The separation in Figure 2 is sharpest for matched DeepSeek evidence: it recovers the MMLU truth partition with 98.5% orbit accuracy, and 97.0% of released pairwise relations are correct. Yet the unanchored experiment has two observationally identical global orientations, so its absolute minimax accuracy is exactly 50%. One uniformly selected trusted label raises leave-anchor-out accuracy to 97.0%; three labels reach 98.5%. Across the four executable channels, orbit accuracy ranges from 73.4% to 93.0%, while one-anchor accuracy ranges from 60.7% to 87.0%.

4.2 Trusted anchors are mechanism-relative

The same relational evidence can support opposite decisions under different anchor mechanisms. We test all 27 source–target pairs among real probes, corrupted probes, and compile/load checks for three

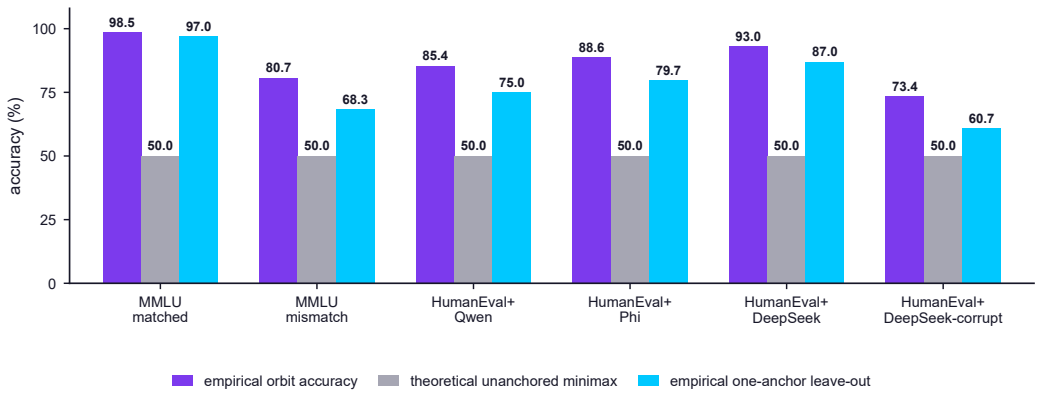

Figure 2: **Truth structure without an absolute certificate.** Purple and cyan bars are empirical accuracies; the gray bar is the theorem-implied minimax accuracy of the unanchored relational release. Across MMLU-Pro mechanisms and three HumanEvalPlus model families, strong recovery up to a global flip coexists with 50% absolute certification, while one held-out trusted label recovers much of the orientation.

code models on 158 frozen HumanEvalPlus programs per model. Without a fitted pairwise parameter, the orientation law predicts above-chance transfer exactly when source and target orientations agree. A separate natural audit uses trusted labels from 79 programs to orient unmodified channels on 79 disjoint programs.

In the natural audit, matched nine-anchor accuracy reaches 93.7% for DeepSeek, 92.1% for Phi, and 85.8% for Qwen. Discovery orientations predict all $12/12$ held-out transfer signs; exact finite-population averaging over the disjoint split separates transfer from calibration on the anchors (Appendix Table 7; $p = 1/10001$).

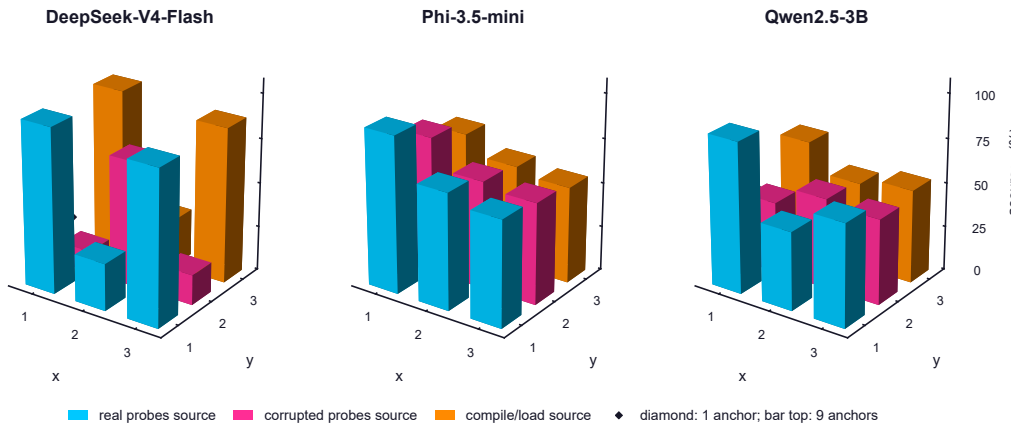

Figure 3: **Orientation transfer across 27 mechanism pairs.** The x and y axes index target and anchor-source mechanisms: 1 = real probes, 2 = corrupted probes, and 3 = compile/load. Bars report exact nine-anchor leave-out accuracy; black diamonds mark one anchor. Correct labels amplify either the aligned orientation or its complement.

Figure 3 confirms the predicted sign structure: after nine anchors, aligned pairs average 70.1% accuracy and opposite pairs 32.9%. Eight curves worsen as correct labels are added, with the largest drop from 29.9% to 11.9%. All $9/9$ orientations learned on one half persist on the other and predict $23/27$ held-out signs; the four misses are near-random Qwen cells within 0.71 points of chance. Thus aligned signs amplify trust, opposite signs amplify error, and unidentified signs provide no guarantee (Corollary 6).

4.3 SHARED-COORDINATE FAILURE IN MODEL VIEWS

Eight Qwen views share one persistent mechanism across 100 BBH (Suzgun et al., 2023) items. Their mean agreement is 90.1%, and their consensus follows the persistent injected anchor on 96% of items, although that anchor is correct only 21% of the time. Figure 4 exposes the consequence: among the 79 items with at least $7/8$ agreement, 62 consensuses are wrong; unanimity still leaves 50 errors among 66 items. Agreement has precisely stabilized a shared coordinate system without grounding it. The injected anchor makes this a failure witness, not a prevalence estimate. Both curves apply the same plurality threshold, controlling coverage; only whether the views share a persistent coordinate changes. The answer scorer and acceptance thresholds remain fixed, so neither coverage nor scoring explains the separation. Physical multiplicity therefore does not imply distinct fault domains; dependence structure, rather than view count, is the natural unit of trust. The eight views therefore constitute one fault component, not eight independent votes.

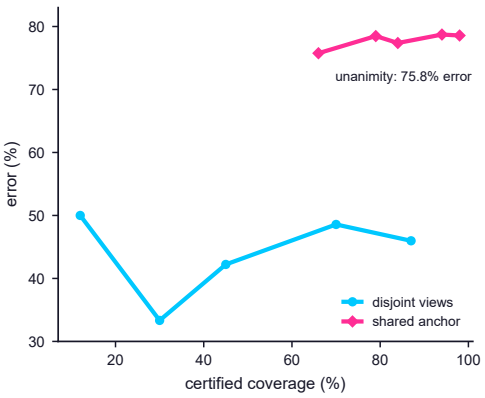

Figure 4: **Shared-coordinate failure.** Agreement lowers error for disjoint views but leaves eight views sharing one persistent anchor at 75.8% error under unanimity.

4.4 TRUST COST FOLLOWS FAULT TOPOLOGY

Holding the matched MMLU judgments fixed, we alter only the connectivity of their relation graph. As one connected component, zero versus one anchor gives 50.0% versus 97.0% leave-out accuracy. Splitting the same 67 accepted judgments into 14 subject components gives 50.0%, 52.7%, 71.0%, and 97.5% after orienting 0, 1, 7, and 14 components, respectively. Under taskwise drift, the graph has 70 singleton components. Anchoring 1, 3, 5, or 10 of them leaves every other task at exactly 50%. The costs in Table 1 therefore follow the number of unresolved components, as Corollary 3 predicts, not the amount of agreement inside them.

Table 1: **Anchor cost follows fault topology, not verifier count.** Expected leave-anchor-out accuracy (%) after changing only the connectivity of the same frozen MMLU judgments. The final column is the number of trusted labels required to orient every independent gauge component.

Fault topology	Components	0 anchors	1 anchor	7 anchors	Required
Globally connected	1	50.0	97.0	98.5	1
Subject domains	14	50.0	52.7	71.0	14
Taskwise drift	70	50.0	50.0	50.0	70

The intervention establishes the scaling law exactly but fixes its topology by construction. A prospective natural-domain audit instead inferred topology and anchor roles from 400 historical items before Qwen and Phi processed 280 previously unused MMLU-Pro questions. The primary mechanism used 30 deployment labels and reached 76.2%, 7.1 points above independent domainwise calibration with 70 labels. A single global orientation also reached 76.2%. The inferred domains therefore save labels relative to separate calibration, while the fresh data do not distinguish their signs from a single global orientation.

4.5 NATIVE INTERNAL TOPOLOGY AND ANCHOR PORTABILITY

Granite-3.0-3B-A800M-Instruct (Granite Team, IBM, 2024) provides a native test of whether label-free topology predicts anchor portability. We extract the pre-top-k router logits from all 40 experts in each of its 32 layers. Each claim pairs the original correct MMLU-Pro option with one preselected public distractor; candidate order is balanced, while labels and model outputs remain unchanged. For each layer, the first principal axis of the 40-dimensional left–right router difference is fitted

on 560 topology claims. Absolute correlation of these axes defines a label-free layer graph. Its maximal persistent sets are layers 9–19 and 21–29 under all three thresholds $\{.55, .575, .60\}$ and both predetermined halves; all 59 strong edge signs agree across the halves. This graph, its within-component signs, and its decoder are frozen before anchor or evaluation truth is opened.

Table 2: **Natural router topology predicts anchor portability.** Held-out percentages from two disjoint cohorts under a topology and decoder inferred without labels. Matched and cross use 17 and 51 anchors in cohorts I and II; intervals are paired-bootstrap 95% CIs. Control reports the orbit accuracy loss under category-matched wrong-question inputs.

Cohort	Claims	Target orbit	Source orbit	Matched	Cross	Δ [95% CI]	Control
I	980	65.6	52.2	59.7	51.0	$+8.7\ [2.4, 15.0]$	+8.1
II	420	65.2	53.1	64.3	55.4	$+8.9\ [2.6, 17.1]$	+9.0

Two subsequent cohorts contain no claims used to discover the graph and no claims shared with each other. The later component retains target orbit accuracy 65.6% and 65.2%; the earlier component remains statistically indistinguishable from chance. Anchoring the target component with its own labels exceeds transfer from the earlier component by 8.7 points at $k = 17$ and 8.9 points at $k = 51$ (Table 2). Paired bootstrap intervals, $[2.4, 15.0]$ and $[2.6, 17.1]$, exclude zero, and all 14 subject categories have a positive gap in both cohorts. Replacing each question with a category-matched wrong question reduces target orbit accuracy by 8.1 and 9.0 points, respectively.

The same frozen topology predicts differential anchor portability twice on disjoint claims: within-component anchoring outperforms cross-component transfer in both cohorts.

A model-held-out audit tests whether this transport is specific to sparse routing. With extraction and graph rules frozen before Qwen2.5-1.5B inference, hidden-state coordinates at layers 14, 21, and 28 form a component on TruthfulQA. On disjoint claims, its held-out orbit accuracy is 70.4%, and expected oriented accuracy rises from 58.1% with one anchor to 66.6% with nine. A deranged-question control retains a weaker coordinate. Native computation can therefore transport orientation across unseen claims without fixing what that orientation means (Appendix D.1.4).

4.6 World coupling supplies the compass

Table 3: **World coupling, not evidence volume.** Certificate error (%; lower is better) before relational release. Matched evidence is coupled to the target instance; mismatch preserves format and volume.

Setting	Closed / compile	Mismatch / corrupt	Mechanism-matched
MMLU-Pro ($\gamma = .9$)	20.0	19.3	**1.5**
HumanEvalPlus Qwen	41.8	55.7	**14.6**
HumanEvalPlus Phi	39.9	34.2	**11.4**
HumanEvalPlus DeepSeek	12.7	73.4	**7.0**

Unlike the relational release, these channels expose their original decisions. Matched and mismatched conditions preserve format, domain, and evidence volume; they differ in whether the added evidence remains coupled to the target instance. The comparison uses original judgments before any pairwise release. MMLU holds model, format, domain, and sampling budget fixed while replacing an item-matched rationale with a cyclically mismatched one. HumanEvalPlus holds each frozen program and probe count fixed while replacing executable expected outputs with corrupted ones.

Across the unreduced channels in Table 3, the matched MMLU rationale collapses error from 20.0% to 1.5%; equal-format same-domain evidence leaves it at 19.3%. Real executable probes likewise reduce error for all three model families, while corrupting the expected outputs attenuates or reverses the benefit. Equal format, domain, and evidence volume do not substitute for item-level coupling. The corrupted channels nevertheless preserve the relational structure quantified by the transfer matrix.

5 DISCUSSION

For orientation transfer, the natural unit of trust is a fault component, not a verifier. Additional judgments within an oriented component can improve precision; they cannot resolve a sign that remains free. Under stable noiseless relations, one trusted bit can orient an arbitrarily large connected component. Its value is nonlocal: the same label can remove one error or thousands, depending on the topology through which its orientation propagates.

Anchor portability is therefore a relation between source and target mechanisms, not an intrinsic property of the source. Correct labels identify the source orientation; they do not identify its relation to every target. An aligned relation converts increasing source certainty into increasing target accuracy. An opposite relation converts the same correct labels into increasing target error. Without a stable relation, both directions remain compatible with the observations and neither has a uniform guarantee. This trichotomy separates epistemic grounding from reputation, independence from mere plurality, and world coupling from physical source separation.

The controlled relational releases establish the identification separation, while the native-internal audits ask whether topology also predicts transfer in unreduced models. The Granite topology, frozen without labels, predicts higher within-component than cross-component anchor portability in two disjoint cohorts, by 8.7 and 8.9 points, with both intervals excluding zero, although the stronger prespecified decision rule was not met. Together with the model-held-out hidden-state replication, these results show that native internal coordinates can transport orientation across claims without grounding it in the world. They establish a mechanism rather than its deployment prevalence, and their anchor counts remain conditional on the declared symmetry class. Binary correctness makes the residual group especially transparent; multiclass and continuous decisions preserve the stabilizer criterion but induce different anchor complexities.

6 CONCLUSION

Gauge orbits expose a gap that evidence volume cannot close: every pairwise truth relation may be known while no individual claim is certifiable. We show that certification is possible exactly when every observationally compatible transformation preserves the target truth. A trusted anchor can enforce this condition, while fault topology determines how far its orientation travels: aligned relations preserve the direction, opposite relations reverse it, and unidentified relations provide no guarantee. Experiments on MMLU-Pro, HumanEvalPlus, and frozen internal representations exhibit this separation between relational recovery and absolute certification. Beyond binary correctness, the same criterion applies with richer transformation groups. Coordinates become a compass only when a world-coupled anchor fixes the target's orientation.

AI USE STATEMENT

Generative AI tools were used to assist with writing and polishing the manuscript to improve its readability, as well as with code review and correction. The authors reviewed all AI-assisted work and take responsibility for the final content of the manuscript.

REPRODUCIBILITY STATEMENT

The formal assumptions and complete proofs are provided in the appendix. The appendix also documents dataset construction, model and decoding settings, splits, evaluation rules, statistical procedures, and supporting empirical results needed to reproduce the reported analyses.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. doi: 10.48550/arXiv.2404.14219.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953.

Andrew Michael Brilliant. Limits of self-correction in LLMs: An information-theoretic analysis of correlated errors. *SSRN Electronic Journal*, 2026. doi: 10.2139/ssrn.6044795. URL https://ssrn.com/abstract=6044795. Preprint.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

DeepSeek-AI. DeepSeek-V4 preview: Entering the era of affordable million-token context, 2026. URL https://deepseek.com/en/news/v4-preview/.

Yegor Denisov-Blanch, Joshua Kazdan, Jessica Chudnovsky, Rylan Schaeffer, Sheng Guan, Soji Adeshina, and Sanmi Koyejo. Consensus is not verification: Why crowd wisdom strategies fail for LLM truthfulness. *arXiv preprint arXiv:2603.06612*, 2026. doi: 10.48550/arXiv.2603.06612. URL https://arxiv.org/abs/2603.06612.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2151–2159, 2019.

Granite Team, IBM. Granite 3.0 language models, October 2024. URL https://github.com/ibm-granite/granite-3.0-language-models/.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30038–30066. PMLR, 2025. URL https://proceedings.mlr.press/v267/kim25e.html.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35 (4):1419–1455, 1964. doi: 10.1214/aoms/1177700372.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? Rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ning Liu. LLMs as a jury: Cross-model consensus can outperform process reward models for LLM reasoning. *arXiv preprint arXiv:2607.10139*, 2026. doi: 10.48550/arXiv.2607.10139. URL https://arxiv.org/abs/2607.10139.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18): 19260–19268, 2025. doi: 10.1609/aaai.v39i18.34120.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Blossom Metevier, Max Springer, Bohdan Turbal, and Aleksandra Korolova. Towards statistical verification for trustworthy AI. In *ICLR 2026 Workshop on Principled Design for Trustworthy AI*, 2026. URL https://openreview.net/forum?id=kzFpgeTfJs.

Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-Tau Yih, Sida Wang, and Xi Victoria Lin. LEVER: Learning to verify language-to-code generation with execution. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26106–26128, 2023.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000. doi: 10.1111/1467-9868.00265.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023. doi: 10.18653/v1/2023.findings-acl.824.

Hexiang Tan, Fei Sun, Sha Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. Too consistent to detect: A study of self-consistent errors in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4755–4765. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.238. URL https://aclanthology.org/2025.emnlp-main.238/.

Erik Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991. doi: 10.1017/CBO9780511666353.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or Fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020.

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 *International Conference on Learning Representations*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3018.

Yuan-Hao Wei. Anchor–stabilizer theory for identifiable representation learning: Generalized anchors as symmetry breakers. *Preprints.org*, 2026. doi: 10.20944/preprints202606.0188.v1. URL https://www.preprints.org/manuscript/202606.0188. Preprint.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115.

Jitian Zhao, Changho Shin, Tzu-Heng Huang, Satya Sai Srinath Namburi GNVV, and Frederic Sala. CARE: Confounder-aware aggregation for reliable LLM evaluation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3WPDFjZ1UT.

A Decision-Theoretic Extensions

A.1 Operational frontier

For coverage target γ, define

$$\mathcal{F}_{O,\Pi}(\gamma) = \inf_{c:\Pr_\Pi(c(O)\neq\perp)\geq\gamma} \Pr_\Pi(c(O) \neq \tau \mid c(O) \neq \perp). \tag{16}$$

For an implementable rule family $\mathcal{C}$, let $\mathcal{F}^{\mathcal{C}}_{O,\Pi}$ restrict the infimum to $c \in \mathcal{C}$. The restricted frontier upper-bounds $\mathcal{F}_{O,\Pi}$ and admits direct estimation from a finite benchmark sweep.

Theorem 8 (Operational frontier)**.** *Let $e_\Pi(o) = 1 - \max_t \Pr_\Pi(\tau = t \mid O = o)$. The optimal certificate labels by posterior majority and accepts cells in increasing order of $e_\Pi(o)$, randomizing only at a boundary tie. Moreover,*

$$\sup\{\gamma : \mathcal{F}_{O,\Pi}(\gamma) = 0\} = \Pr_\Pi(O \in \mathrm{Cert}(O; \mathcal{W})).$$

If O_1 is a world-independent garbling of O_2, then $\mathcal{F}_{O_2,\Pi}(\gamma) \leq \mathcal{F}_{O_1,\Pi}(\gamma)$; redundant augmentation leaves the frontier unchanged.

Proposition 9 (Overlap–error identity)**.** *Let $\mu_t(o) = \Pr_\Pi(O = o, \tau = t)$. For any acceptance set A, the minimum conditional error among certificates deciding exactly on A is*

$$\frac{\sum_{o\in A} \min\{\mu_0(o), \mu_1(o)\}}{\sum_{o\in A}[\mu_0(o) + \mu_1(o)]}.$$

It is zero exactly when no accepted observation has positive mass under both truth values. Under a full-support prior, this is equivalent to $A \subseteq \mathrm{Cert}(O; \mathcal{W})$.

A.2 Deficiency to the truth oracle

The compatibility cell gives an exact set characterization but need not be enumerated. Instead view O as the statistical experiment $\mathcal{E}_O = (\mathcal{O}, \{P^O_w : w \in \mathcal{W}\})$ and truth itself as the oracle experiment $\mathcal{E}_\tau = (\{0,1\}, \{\delta_{\tau(w)} : w \in \mathcal{W}\})$. Exact certification is the Blackwell comparison $\mathcal{E}_\tau \preceq_{\mathrm{B}} \mathcal{E}_O$. Its quantitative relaxation is the one-sided Le Cam deficiency (Le Cam, 1964; Torgersen, 1991)

$$\Delta_\tau(O) = \inf_Q \sup_{w\in\mathcal{W}} \mathrm{TV}\left(QP^O_w, \delta_{\tau(w)}\right), \tag{17}$$

where Q ranges over Markov kernels from observations to binary decisions. Because the target law is a point mass,

$$\Delta_\tau(O) = \inf_c \sup_{w\in\mathcal{W}} \Pr_w\{c(O) \neq \tau(w)\}. \tag{18}$$

Thus deficiency is the residual worst-case decision error after the best randomized use of the entire experiment, not the performance of a chosen confidence score.

Theorem 10 (Truth-oracle deficiency)**.** *For finite $\mathcal{W}$ and $\mathcal{O}$: (i) $\Delta_\tau(O) = 0$ exactly when every reachable compatibility cell is truth-homogeneous; (ii) $O_1 \preceq O_2$ implies $\Delta_\tau(O_2) \leq \Delta_\tau(O_1)$; and (iii) $\Delta_\tau(O)$ is the value of the linear program*

$$\begin{aligned} \min_{q,\epsilon} \quad & \epsilon \\ \text{s.t.} \quad & \sum_o P^O_w(o) q_o \leq \epsilon && (\tau(w) = 0), \\ & \sum_o P^O_w(o)(1 - q_o) \leq \epsilon && (\tau(w) = 1), \qquad 0 \leq q_o \leq 1. \end{aligned} \tag{19}$$

Here q_o is the probability of deciding one after observing o. The LP dual selects least-favorable mixtures of opposite-truth worlds whose induced observation laws are hardest to distinguish.

Equation 19 is the computational replacement for explicit cell construction. It also exposes the limit of empirical computation: if a finite trace contains an item identifier and the same labeled items are used to optimize Q, the LP can memorize the benchmark and spuriously return zero. A trace coarsening or restricted decoder family yields a computable *upper* bound on unrestricted deficiency; a positive lower bound requires declared counterfactual worlds or a robust ambiguity set.

A.3 CLOSED-CHANNEL IMPOSSIBILITY

Let O_0 be the complete closed observation. Any augmentation satisfying

$$W \longrightarrow O_0 \longrightarrow U \tag{20}$$

is a genuine garbling of O_0; deterministic post-processing provides the canonical case. Fresh closed samples need not satisfy this Markov chain and may estimate behavior more precisely. Let $B(w) = \mathcal{L}_w(O_0^\infty)$ be the full behavioral law revealed in the unlimited-sampling limit.

Theorem 11 (Closed-channel impossibility)**.** *A universally correct function of $B(w)$ exists if and only if*

$$B(w) = B(w') \Longrightarrow \tau(w) = \tau(w'). \tag{21}$$

Hence opposite-truth worlds with the same complete closed law cannot be certified by any amount of closed observation.

Closed sampling may exhaust uncertainty about behavior while leaving truth ambiguity intact.

A.4 TRANSFORMER COMPUTATION AS A CLOSED CHANNEL

Let θ be a frozen autoregressive Transformer, x its complete input, and ξ decoding randomness independent of the world conditional on x. Define the *computation trace*

$$Z_\theta(x, \xi) = \left\{h_{\ell j}, a_{\ell j}, m_{\ell j}, z_j, \widehat{y}_j\right\}_{\ell, j}, \tag{22}$$

containing every residual state h, attention output a, MLP output m, pre-softmax logit vector z, and generated token. This deliberately grants a white-box certificate more access than ordinary deployment.

Corollary 12 (Mechanistic closure)**.** *Suppose opposite-truth worlds w, w' expose the same x to a frozen model and do not otherwise affect its computation. Then*

$$\mathcal{L}_w(Z_\theta \mid x) = \mathcal{L}_{w'}(Z_\theta \mid x).$$

Consequently no statistic of activations, attention, logits, entropy, sampled reasoning, or any randomized probe of this trace exactly certifies τ on a world class containing w, w'.

The proof is operational: conditioned on x, both worlds execute the same transition equations with the same law for ξ. Hence they occupy the same behavioral fiber in Theorem 11; every internal score is a garbling of that fiber. This avoids enumerating $\mathcal{W}$ or deciding whether a tiny probability is strictly positive. To refute closed certification it is enough to exhibit an opposite-truth pair whose difference is absent from the model input. Pretraining may make the trace correlate strongly with truth under a benchmark distribution, but correlation is not a channel through which the current world can update the frozen computation.

A.5 ONE-SIDED INCONSISTENCY WITNESS

If T preserves truth and $\widehat{Y}(Tx) \neq \widehat{Y}(x)$, both predictions cannot be correct. Inconsistency can therefore falsify joint correctness; consistent error shows why the converse fails.

A.6 CONDITIONAL SUPPORT SEPARATION

For $O_0 = o$, define the evidence supports under the two residual truth values:

$$S_t(o) = \bigcup_{\substack{w \in \Gamma_{O_0}(o) \\ \tau(w) = t}} \{g : K_G(g \mid o, w) > 0\}, \qquad t \in \{0, 1\}. \tag{23}$$

Theorem 13 (Support separation)**.** *After observing (o, g), exact certification is possible exactly when $g \in S_0(o) \triangle S_1(o)$. The added channel is decisive throughout $\Gamma_{O_0}(o)$ exactly when $S_0(o) \cap S_1(o) = \varnothing$.*

A.7 Externality under mechanism uncertainty

An external source is not automatically an external statistical experiment. Let its response law be $K_{G,\vartheta}(\cdot \mid w)$, where $\vartheta \in \Theta$ indexes source corruption, shared training data, tool failure, stale state, or dependence between nominally distinct verifiers. If ϑ is not observed by the certificate, the relevant experiment is indexed by (w, ϑ) while its target remains $\tau(w)$. Define

$$\Delta_\tau^\Theta(G) = \inf_Q \sup_{w \in \mathcal{W}, \vartheta \in \Theta} \mathrm{TV}\big(QK_{G,\vartheta}(\cdot \mid w), \delta_{\tau(w)}\big)\,. \tag{24}$$

Theorem 14 (Robust externality)**.** *For finite spaces, $\Delta_\tau^\Theta(G) = 0$ if and only if every robust cell*

$$\Gamma_G^\Theta(g) = \{w : \exists \vartheta \in \Theta, K_{G,\vartheta}(g \mid w) > 0\}$$

is truth-homogeneous. Its value is computed by Equation 19 with one constraint for each (w, ϑ). Hence provenance does not determine epistemic externality. A physically separate source may retain positive robust deficiency under mechanism uncertainty or a shared fault, while a tool-mediated channel within one system may introduce world-coupled information.

This criterion separates *provenance* from *epistemic externality*. Ten verifiers in one fault domain may add context and agreement while inducing nearly the same experiment. A source is useful only insofar as its entire mechanism class moves the experiment toward the truth oracle. Theorem 1 gives a structural certificate of robust failure: finding one truth-reversing symmetry of (w, ϑ) proves $\Delta_\tau^\Theta(G) = 1/2$ without enumerating the robust cell or solving the LP. The LP remains useful when symmetry is only approximate.

A.8 Minimal truth resolution

Let $\mathcal{P}_\tau(o)$ partition $\Gamma_{O_0}(o)$ by the value of τ.

Theorem 15 (Minimal evidence principle)**.** *$\mathcal{P}_\tau(o)$ is the coarsest partition supporting a universally decisive exact certificate. Every decisive augmentation refines it, and a redundant channel cannot resolve an ambiguous truth partition.*

Evidence need not recover the world; it need only resolve the truth property being certified.

A.9 Approximate and adaptive consequences

An exact symmetry is the endpoint of a continuous decision boundary. For any opposite-truth pair with laws P_0, P_1, full-coverage error is at least $(1 - \mathrm{TV}(P_0, P_1))/2$; approximate symmetry therefore leaves an error floor. In sequential acquisition, the residual stabilizer compresses the state: a useful action removes a generator that still changes the target, whereas another observation inside an already oriented component may add precision but no certification value. Appendix B.2 gives the TV boundary, an adaptivity separation, and simultaneous finite-sample stopping guarantees used by the secondary acquisition audits.

B Proofs and Auxiliary Results

B.1 Gauge orbits, stabilizers, and anchor complexity

Proof of Theorem 1. If $P_{h\eta}^O = P_\eta^O$, every observation with positive probability under η has the same positive probability under $h\eta$. Hence both expanded states occur in the same compatibility cells. If their binary targets are opposite, let $q(o)$ be any randomized probability of deciding one. Because the observation laws coincide, the two statewise error probabilities sum to one. Their maximum is therefore at least $1/2$. A world-independent fair coin attains error $1/2$ in every state, proving $\Delta_\tau(O) = 1/2$. Under orbit completeness, any observation in the support of P_η^O is impossible under states outside $[\eta]_\mathcal{H}$ and possible under every state inside it. Each reachable compatibility cell is therefore exactly one orbit, so truth is exactly certifiable precisely when it is constant on every orbit. □

Proof of Theorem 2. Equivariance gives $A(h\eta) = h \cdot A(\eta)$, so a transformed state $h\eta$ remains compatible with the revealed anchor exactly when $h \in \mathcal{H}_A(\eta)$. This set is the group stabilizer of $A(\eta)$, and the residual gauge class is its orbit. A target varying on that orbit is not identifiable by Theorem 1; a target fixed by the stabilizer is well defined on it. If the unanchored experiment is orbit-complete, conditioning on the anchor refines each full orbit into precisely these stabilizer orbits, establishing sufficiency. □

Proof of Corollary 3. Flipping every τ_i in one connected component preserves every edge XOR, so C components give an action of $\mathbb{Z}_2^C$. Conversely, if two label vectors have the same edge XORs, their pointwise XOR is constant across each edge and therefore across each connected component. They differ by exactly one of these componentwise flips, making the relation experiment orbit-complete. Without an anchor, each component has a truth-reversing invisible flip. One trusted label in that component fixes its flip and propagates along all paths. Thus one anchor per component is necessary and sufficient. □

Proof of Proposition 4. On a connected relation graph, τ and $1-\tau$ induce identical complete relations. For any estimated label vector, its Hamming losses against these two states sum to n, so worst-case loss is at least $n/2$; independent random labels attain per-coordinate risk $1/2$. Given one trusted τ_j, every other label equals $\tau_j \oplus R_{ij}$ along any path, yielding zero error. Repeated observation of an already known relation experiment does not remove the global flip. With singleton components no path connects an anchored claim to any unlabeled claim, so the latter retain risk $1/2$. □

Proof of Proposition 5. The source majority retains the target's nominal coordinate with probability $\pi_s(k)$. Target accuracy is then a_t; under the complementary coordinate it is $1 - a_t$. Total probability gives the first equality in Equation 12; subtracting $1/2$ and factoring gives the second. □

Proof of Corollary 6. Treating $A_{s\to t}$ as a function of π_s yields

$$\frac{\partial A_{s\to t}}{\partial \pi_s} = 2a_t - 1.$$

Its sign gives the three monotonicity regimes, and substitution gives the limit as $\pi_s \to 1$. For any reliable source, two admissible targets with nominal accuracies a and $1 - a$, $a > 1/2$, respond in opposite directions to the same increase in source certainty. A source-only trust scalar therefore cannot guarantee improvement on both without an assumption linking source and target orientations. □

Proof of Corollary 7. Let q_c denote the probability that the anchor majority selects the wrong orientation in component c. The component error is e_c under the better orientation and $1 - e_c$ under its complement, hence its expected error equals

$$(1 - q_c)e_c + q_c(1 - e_c) = e_c + q_c(1 - 2e_c).$$

Hoeffding's inequality for independent votes, and its without-replacement counterpart for a finite anchor pool, gives $q_c \le e^{-2k_c\gamma_c^2}$. Weighting by n_c/n proves Equation 13. The budget in Equation 14 makes $q_c \le \delta/C$ for every component. A union bound then orients all C components correctly with probability at least $1 - \delta$; on that event only the within-orbit errors remain. □

B.2 Auxiliary decision and acquisition results

Proposition 16 (Two-world full-coverage error boundary)**.** *For finite-observation laws P_0, P_1 from opposite-truth worlds, every non-abstaining binary certificate satisfies*

$$\inf_c \max_i \Pr_{P_i}(c(O) \neq \tau(w_i)) \ge \frac{1 - \mathrm{TV}(P_0, P_1)}{2}.$$

The right side is the equal-prior Bayes error and TV cannot increase under garbling.

Proposition 17 (Shared-bias failure)**.** *If $G_i = s_\tau + B + \epsilon_i$, where $s_1 = -s_0 \neq 0$, the noise law is shared across truth states, and B is unrestricted, no rule based on $(G_1, \ldots, G_k)$ uniformly identifies τ.*

Proposition 18 (Unbounded adaptivity gap)**.** *For every m, some finite resolution problem has optimal static exact cost $m + 1$ and optimal adaptive exact cost 2.*

Theorem 19 (Safe adaptive selection)**.** *Let $\mathcal{A}$ be M policies fixed before an i.i.d. calibration sample. For each π, form a one-sided exact error upper bound U_π and coverage lower bound L_π, each at tail $\delta/(2M)$. Then, with probability at least $1 - \delta$, both bounds hold simultaneously for every policy. Consequently the least-cost policy satisfying $U_\pi \le \alpha$ and $L_\pi \ge \gamma$ meets both population targets with probability at least $1 - \delta$.*

B.3 Proofs for the core characterization

Proof of Theorem 8. Represent a randomized selective rule by its acceptance probability $a(o) \in [0, 1]$ and, when accepted, a truth label. For fixed a, the posterior-majority label minimizes the error contribution at each o, so the error numerator is $\mathbb{E}_\Pi[a(O)e_\Pi(O)]$ and coverage is $\mathbb{E}_\Pi a(O)$. A rule with coverage above γ can be thinned independently to coverage γ without changing conditional error. The resulting linear program is minimized by assigning acceptance mass to the smallest values of $e_\Pi(o)$ first; an exchange of acceptance mass between any inverted pair strictly improves the objective. Boundary randomization handles the final tie.

If O_1 is a garbling of O_2, a rule observing O_2 can simulate the garbling and then apply any rule for O_1, preserving both coverage and error. Taking infima proves Blackwell monotonicity. When $W \to O \to U$, (O, U) trivially reveals O, while it can also be generated from O using the world-independent kernel for U. The two observations are therefore Blackwell-equivalent, proving equality of their full frontiers.

The lower partial mean of the nonnegative variable $e_\Pi(O)$ is zero exactly up to the probability mass on which $e_\Pi(O) = 0$. Under full support, posterior error is zero exactly when the reachable compatibility cell contains one truth value, which establishes the stated zero-error endpoint. □

Proof of Proposition 9. At an accepted observation o, predicting t contributes $\mu_{1-t}(o)$ error mass, so the smaller contribution is $\min\{\mu_0(o), \mu_1(o)\}$. Summing these pointwise optima over A and dividing by its probability gives the identity. The numerator vanishes exactly when every accepted o has mass under at most one truth value. With finite or countable $\mathcal{W}$ and a full-support prior, positive channel probability in a compatible world contributes positive joint mass; hence this condition is equivalent to truth-homogeneity of every accepted cell. □

Proof of Theorem 10. A kernel from o to a binary decision is specified by $q_o = Q(1 \mid o)$. Total variation from QP_w^O to the point mass at $\tau(w)$ equals the probability that this randomized decision is wrong. Writing that error separately for $\tau(w) = 0$ and $\tau(w) = 1$ gives Equation 19.

If two opposite-truth worlds assign positive mass to the same o, zero error would require both $q_o = 0$ and $q_o = 1$, a contradiction. Conversely, when every support cell is homogeneous, set q_o to its common truth value. This proves the zero-deficiency equivalence. If O_1 is obtained from O_2 by a Markov kernel, composing that kernel with any decision rule for O_1 produces the same risks using O_2; taking infima proves monotonicity. Finite LP duality gives an adversarial distribution over the world-indexed risk constraints, equivalently least-favorable mixtures of the two composite truth hypotheses. □

Proof of Theorem 14. Treat (w, ϑ) as the parameter of one enlarged statistical experiment and assign it target value $\tau(w)$. Theorem 10 applies verbatim. An observation g belongs to its support cell precisely when some ϑ makes $K_{G,\vartheta}(g \mid w)$ positive, which yields $\Gamma_G^\Theta(g)$ and the stated LP indexing. □

Proof of Theorem 19. Let N_π and K_π denote accepted and erroneous accepted calibration counts. Fix π. Because its entire history-to-action map is independent of the calibration set, its item-level acceptance and error indicators are i.i.d. Conditional on N_π, K_π is binomial with parameter $\mathrm{Err}(\pi)$, while N_π is binomial with parameter $\mathrm{Cov}(\pi)$. One-sided exact-binomial coverage therefore fails with probability at most $\delta/(2M)$ for each of the two bounds. A union bound over both bounds and all M policies gives the simultaneous event. On that event every empirically feasible policy is population

feasible, so arbitrary data-dependent selection among them, including by observed cost, preserves both inequalities. □

B.4 Auxiliary mutual-information resolution

For a non-degenerate full-support prior Π, define

$$\mathcal{R}_\Pi(O) = \frac{I_\Pi(\tau; O)}{H_\Pi(\tau)}. \tag{25}$$

Normalized mutual information measures average dependence under Π rather than robust certifiability across compatible worlds.

Theorem 20 (Resolution hierarchy)**.** $0 \le \mathcal{R}_\Pi(O) \le 1$*; Blackwell garbling cannot increase it; redundant augmentation leaves it unchanged; and it equals one exactly when every reachable cell is truth-homogeneous.*

Proof of Theorem 20. The first claim follows from $0 \le I_\Pi(\tau; O) \le H_\Pi(\tau)$. If $W \to O_2 \to O_1$, then also $\tau \to O_2 \to O_1$ under Π, so data processing gives $I_\Pi(\tau; O_1) \le I_\Pi(\tau; O_2)$. For a redundant U, $I_\Pi(\tau; U \mid O) = 0$, and the chain rule gives $I_\Pi(\tau; O, U) = I_\Pi(\tau; O)$. Moreover, $\mathcal{R}_\Pi(O) = 1$ exactly when $H_\Pi(\tau \mid O) = 0$. Full support ensures that a reachable cell containing both truth values gives both values positive posterior mass and hence positive conditional entropy; the converse is immediate. □

Proof of Theorem 11. If a correct function of B exists, equal behavioral laws receive the same output and therefore must share a truth value. If truth is constant on every fiber of B, labeling each fiber by that value defines the required function. □

Proof of Theorem 13. The refined cell at (o, g) contains a world of truth value t exactly when $g \in S_t(o)$. It is nonempty and truth-homogeneous precisely when g lies in one support but not the other. Every reachable g is decisive precisely when no g belongs to both supports. □

Proof of Theorem 15. A decisive certificate is constant and correct on each observation cell, so no cell can contain elements of both truth blocks. Thus its partition must refine $\mathcal{P}_\tau(o)$. Conversely, labeling the two truth blocks by 0 and 1 defines an exact decisive certificate. The redundancy claim follows from the world-independent kernel factorization. If two decisive representations are both minimal, neither can split a truth block, so both induce the same partition up to output labels. □

Proof of Proposition 17. Choose biases B_0 and B_1 so that $s_0 + B_0 = s_1 + B_1$. The complete joint observation law is then identical under the two truth values, including in the limit of infinitely many views. □

Proof of Proposition 18. Let a world contain an index $I \in [m]$ and bits $B_1, \dots, B_m$, with target $\tau = B_I$. A unit-cost index channel reveals I, and unit-cost channel j reveals B_j. A universally exact static bundle must acquire the index and every bit channel, hence costs $m + 1$. An adaptive policy first observes I and then acquires only B_I, hence costs two. Both lower bounds are attained by these policies. □

B.5 Proof of the two-world full-coverage boundary

Proof. The certificate induces a test between P_0 and P_1. The standard two-point testing bound lower-bounds its maximal error by half of one minus the total variation distance. Under equal priors, the likelihood-ratio test attains average error $(1 - \mathrm{TV}(P_0, P_1))/2$. Total variation cannot increase under a Markov kernel, which proves the refinement inequality. □

When $P_0 = P_1$, the lower bound is $1/2$. Repeated sampling can increase the total variation between distinct product laws, but it cannot separate equal laws. The distinction isolates sampling error from non-identifiability.

B.6 Informational redundancy cannot refine cells

Lemma 21. *Suppose U is generated from O_0 through a world-independent kernel. For any realized (o, u) with positive conditional probability,*

$$\Gamma_{(O_0,U)}(o, u) = \Gamma_{O_0}(o).$$

Proof. The Markov property gives $K_{O_0,U}(o, u \mid w) = K_{O_0}(o \mid w)K_U(u \mid o)$. For fixed (o, u) with $K_U(u \mid o) > 0$, this joint probability is positive exactly for worlds in which $K_{O_0}(o \mid w) > 0$. □

A redundant augmentation can therefore reformat or resample evidence without increasing certifiability.

C Supporting Empirical Results

C.1 Closed repetition estimates behavior, not truth

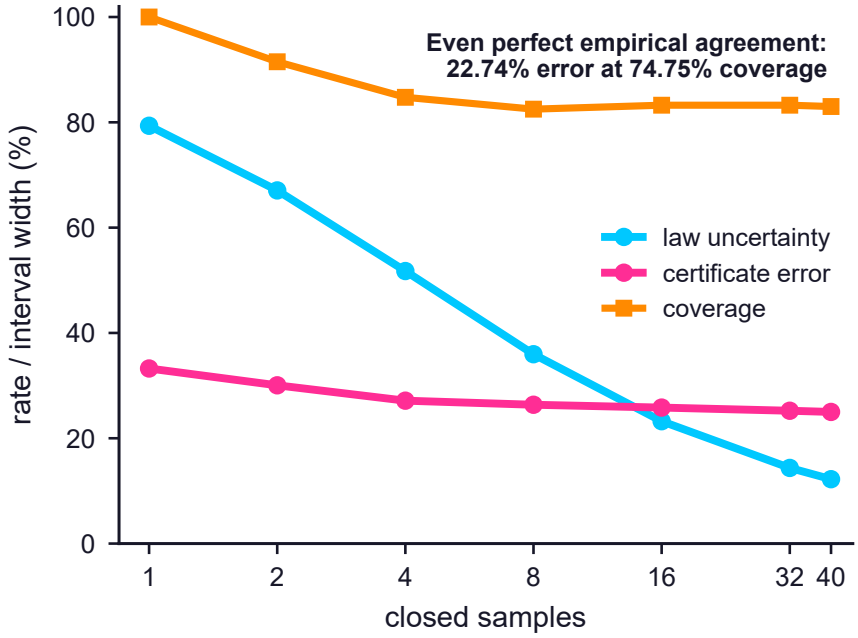

Figure 5: **More closed samples identify the response law, not its truth orientation.** Uncertainty contracts with additional samples, yet unanimity retains 68 errors among 299 accepted answers.

The sample prefixes in Figure 5 separate uncertainty about the response law from uncertainty about truth. For every MMLU-Pro answer, the analysis takes the first $n \in \{1, 2, 4, 8, 16, 32, 40\}$ DeepSeek samples under the original prompt. Modal-frequency uncertainty uses a two-sided 95% Wilson interval, and acceptance uses the fixed 0.9 agreement threshold at every prefix. Increasing n contracts the mean response-law interval from 79.3 to 12.2 percentage points, but the high-agreement error remains 25.0%; even unanimity leaves 68 wrong certificates among 299 accepted answers. Repetition therefore improves estimation within a behavioral fiber without orienting that fiber to truth.

C.2 Internal readouts and repetition predict risk but do not certify

The closed-evidence audit fixes 400 category-balanced MMLU-Pro questions. Qwen2.5-3B-Instruct, Phi-3.5-mini-instruct, and DeepSeek V4-Flash answer the same items. Qwen and Phi consistency is stability under a truth-preserving distractor swap; DeepSeek consistency is at least 0.9 agreement

across 40 samples. The criteria are not numerically equivalent; each exposes a closed high-coherence subset.

A decoder-output audit sums the API-exposed top-logprob mass of every token variant of the frozen modal answer. The score is a conservative lower bound because omitted vocabulary items contribute zero. It predicts risk, but it does not orient truth: among 133 wrong modal answers, 64 have mean answer-token mass at least .99, and 33 have lower-bound log odds at least 10. Threshold .99 still accepts 72.5% of items at 22.1% error. Thus the wrong alternative can already be a high-margin readout of the residual computation rather than an accident of temperature or finite resampling (protocol in Appendix D.2.1).

Closed diagnostics also fail to identify deployment composition. On held-out local-model items, a sharp reweighting LP preserves 25% coverage and the means and empirical deciles of confidence, agreement, entropy, and first–second distance. Under the modest constraint $1/4 \leq nq_i \leq 4$, the accepted-risk identified sets remain $[14.7, 89.7]\%$ for Qwen and $[22.3, 90.5]\%$ for Phi. Correctness labels enter only the retrospective extrema, never the acceptance rule.

Frozen DeepSeek prefixes separate estimation from resolution at 1, 2, 4, 8, 16, 32, and 40 samples. From one to 40 samples, the mean interval width for modal response probability contracts from 79.3 to 12.2 percentage points, yet error among ≥ 0.9-agreement answers remains 25.0%. Even perfect empirical agreement retains 22.7% error. The model's behavioral law sharpens without acquiring the missing world-sensitive distinction.

Across the matched 400-item audit, high-consistency errors number 150 for Qwen, 166 for Phi, and 83 for DeepSeek; these subsets retain more than half of each model's total errors. Even perfect empirical DeepSeek agreement accepts 299 answers with 68 errors (22.7% risk). These observations are consistent with Corollary 12: internal precision ranks risk but does not break an absent world symmetry.

C.3 Truth-sensitive grounding moves the frontier

The grounding audit fixes the majority answer from 40 closed DeepSeek samples on all 70 official MMLU-Pro validation items. A second 40-sample channel receives either the item's reference rationale or an equal-format rationale cyclically shifted from another item in the same domain. Reference text is fixed in advance and stripped of answer-letter statements. Format, model, and sampling budget are identical; only truth-matched content changes.

Stricter closed agreement trades coverage for lower error but does not approach a low-error certificate. At $\gamma = 0.9$, the reference-conditioned channel improves both coverage (95.7% versus 92.9%) and certificate error (1.5% versus 20.0%; Appendix Table 9).

The mismatch returns the frontier close to closed repetition. On the 55 items accepted by both added channels, matched evidence wins nine discordant certificates to zero (two-sided exact $p = 0.0039$). The matched minus mismatched risk difference is -17.8 percentage points under a 14-domain-stratified bootstrap (95% CI $[-27.8, -8.6]$). Equal context length and format leave item correspondence as the source of resolution.

C.4 Resolution-aware acquisition of executable evidence

One greedy code answer is frozen for each model on 158 format-compatible HumanEvalPlus tasks (Chen et al., 2021; Liu et al., 2023). Ten tests, ranked by a task-ID-only deterministic order, form the candidate pool; all remaining tests define held-out correctness. Each candidate is executed separately, so policies are replayed on identical code and outcomes. They observe only prior probe outcomes and test structure, never the held-out suite.

Executable probes couple a frozen program to its realized behavior. Three matched probes reduce binary certificate error to 14.6%, 11.4%, and 7.0% for Qwen, Phi, and DeepSeek, whereas cyclically corrupted expected outputs raise it to 55.7%, 34.2%, and 73.4%. A frozen resolution-aware order reaches matched errors with 44.4%, 32.8%, and 65.3% fewer executions than a fixed order.

The same acquisition principle transfers to open-corpus SciFact discovery (Wadden et al., 2020): on 174 claims whose complete rationale is retrievable, resolution-aware ranking requires 10.15 sentences on average, versus 13.78 for cited-source relevance and 19.03 for a sentence-level generic reranker.

C.5 Approximate decision boundary

The separately frozen coherent-error cohort makes the same failure instancewise. One hundred fresh closed samples falsely support all 13 fixed answers, whereas reference-conditioned evidence correctly contradicts 12 ($p = 4.88 \times 10^{-4}$; Appendix D.4.1). Repetition sharpens the same behavioral fiber; grounding cuts across it.

Exact support separation is the endpoint, not the only informative regime. Across 13 noise levels in the controlled five-view experiment, empirical likelihood-test error remains within 0.15 percentage points of the $(1 - \mathrm{TV})/2$ boundary. It rises from 0.01% at $\mathrm{TV} = 0.9998$ to 46.21% at $\mathrm{TV} = 0.0749$ (Appendix Figure 7). Evidence therefore composes to the extent that its joint law separates opposite-truth worlds, not to the extent that its components agree with one another.

D Supplementary Experimental Details

Frozen splits, model outputs, analysis artifacts, and extended result tables are provided in the supplementary material. The account below states the estimands, controls, and statistical procedures needed to interpret them.

D.1 Relational evidence and fault topology

D.1.1 Relational release and leave-anchor-out evaluation

The MMLU release uses binary predictions frozen by the original grounding audit. Acceptance is fixed at agreement .9 separately within each channel. The HumanEvalPlus release records whether each frozen program passes three real or three corrupted executable probes on the 158 tasks eligible for every model. The release operation compares prediction bits without reading benchmark truth.

For n released judgments, let C agree with truth. The complete equality graph has relation accuracy

$$\frac{\binom{C}{2} + \binom{n-C}{2}}{\binom{n}{2}} = 1 - \frac{C(n-C)}{\binom{n}{2}}, \tag{26}$$

and orbit accuracy $\max(C, n - C)/n$. For an odd set of k anchors sampled uniformly without replacement, let a be the number on which the released orientation agrees with truth. The orientation maximizing anchor agreement is retained if $a > k/2$ and flipped otherwise. Crucially, anchors are removed from the target set. The exact leave-anchor-out accuracy is

$$\sum_{a=0}^{k} \frac{\binom{C}{a}\binom{n-C}{k-a}}{\binom{n}{k}} \frac{\mathbf{1}\{a > k/2\}(C-a) + \mathbf{1}\{a < k/2\}(n-C-k+a)}{n-k}. \tag{27}$$

For cross-mechanism transfer we use the analogous exact four-cell hypergeometric sum over whether each candidate anchor is correct under the source and target mechanisms. No asymptotic approximation or fitted decoder enters these numbers.

For the 14-component intervention, accepted MMLU items are connected only within their frozen subject domain. We average exactly over every subset of k oriented domains and over a uniformly selected anchor inside each selected domain, again excluding anchors from evaluation. The singleton intervention places each task in a separate component; consequently every unlabeled target retains accuracy $1/2$ for every $k < n$.

The natural held-out positive control in Table 7 reuses the frozen 79/79 task split but restricts both source and target to unmodified real probes and compile/load. For each odd anchor count, it computes the exact hypergeometric probability of the discovery mechanism's orientation and applies that sign to the disjoint evaluation mechanism. The one-sided permutation test shuffles evaluation truths within model, preserving predictions and class counts. The complete curves and item-level test results are included in the supplementary material.

Table 4: **Empirical tests of relational trust.** Every held-out effect uses labels disjoint from the claims used to fix topology or orientation. "Cross" denotes transfer from a different inferred component, not an incorrect label.

Audit	Frozen intervention	Held-out result	Identified claim
HumanEvalPlus, unmodified channels	79 programs fix direction; 79 new programs evaluate it	$12/12$ transfer signs; real→real rises from 80.6% to 90.6% between one and nine anchors	Orientation transports within a stable executable mechanism.
HumanEvalPlus, 27 mechanism pairs	Real, corrupted, and compile/load channels; no pairwise fit	$27/27$ full-matrix and $23/27$ split-held-out signs; aligned cells average 70.1%, opposite cells 32.9% at nine anchors	Relative sign, not source correctness alone, determines transfer.
Granite native router, cohort I	Two label-free persistent depth components; all claims fresh after topology freeze	Target-matched 59.7% vs cross-component 51.0%; gap $+8.7$ points, 95% CI $[2.4, 15.0]$	Natural component membership predicts anchor portability.
Granite native router, cohort II	Same frozen topology; zero claim overlap; $k = 51$ fixed in advance	Target-matched 64.3% vs cross-component 55.4%; gap $+8.9$ points, 95% CI $[2.6, 17.1]$	An independent cohort reproduces the topology–portability ordering.
TruthfulQA (Lin et al., 2022), model-held-out layers	The graph rule is frozen before Qwen2.5-1.5B inference	Layers 14, 21, and 28 form a component; both primary edges retain their signs, and nine-anchor accuracy reaches 66.6%	Native layer coordinates carry orientation across unseen claims.
MMLU-Pro prospective domains	Six components inferred from 400 historical items; 210 new evaluations	Topology and global orientation both 76.2%; topology exceeds domainwise calibration by 7.1 points using 30 rather than 70 labels	Label efficiency improves, but independent natural flips are not identified.
MMLU-Pro cross-domain stress test	Five apparently opposite domain pairs selected on discovery halves only	Target raw accuracy shifts from 27.3% on discovery to 51.5% held out; nine-anchor cross transfer is 51.5%	A discovery sign without a stable edge does not license destructive-transfer claims.

Table 5: Complete relational-release results. All entries are accuracy (%). "None" is the theorem-implied absolute minimax accuracy; anchor columns exclude the trusted tasks from evaluation.

Benchmark	Evidence mechanism	Relation	Orbit	None	1 anchor	5 anchors
MMLU	matched reference	97.0	98.5	50.0	97.0	98.5
MMLU	mismatched reference	68.3	80.7	50.0	68.3	77.6
HumanEval+	Qwen real probes	75.0	85.4	50.0	75.0	83.8
HumanEval+	Phi real probes	79.7	88.6	50.0	79.7	87.7
HumanEval+	DeepSeek real probes	87.0	93.0	50.0	87.0	92.8
HumanEval+	DeepSeek corrupted probes	60.7	73.4	50.0	60.7	67.6

D.1.2 NATURAL-DOMAIN TOPOLOGY: DEVELOPMENT AUDIT AND PROSPECTIVE EVALUATION

The development audit uses the same 400 frozen MMLU-Pro items but does not reuse the controlled 1/14/70 connectivity intervention. For each of eight binary mechanisms, every subject domain is split into historical discovery, deployment-anchor, and evaluation sets in proportions $1/2$, $1/4$, and $1/4$. On the historical split, a domain orientation is identified only when its Wilson interval excludes $1/2$. Identified domains form one signed component with their empirically estimated relative orientations; unresolved domains remain singletons. The absolute historical signs are then discarded.

For every inferred component, the historically strongest domain is the root. An odd set of deployment labels from that root chooses the component's global orientation, and all root labels are excluded from

Table 6: Mechanism-relative anchor transfer on HumanEvalPlus DeepSeek tasks. Entries are exact leave-anchor-out accuracy (%). Clean and corrupted mechanisms have opposite signs, so clean anchors increasingly select the complement of the corrupted coordinate.

Anchor source	Relation target	1 anchor	3 anchors	5 anchors	9 anchors
Real probes	Real probes	87.0	91.9	92.8	93.0
Corrupted probes	Corrupted probes	60.7	65.0	67.6	70.5
Real probes	Corrupted probes	29.9	27.2	26.7	26.6

Table 7: **Natural held-out anchor transfer.** Expected accuracy (%) on 79 unseen HumanEvalPlus programs after orienting a frozen, unmodified channel with trusted labels from 79 disjoint programs. "Gain" is the change from one to nine anchors.

Anchor source → target	1 anchor	9 anchors	Gain
Real probes → real probes	80.6	**90.6**	+10.0
Compile/load → real probes	67.2	79.5	+12.4
Compile/load → compile/load	60.3	64.8	+4.5

evaluation. The per-domain baseline instead purchases the same number of labels independently in all 14 subjects. Both methods are evaluated on the identical untouched third split. Results average 1,000 stratified repetitions; the reported intervals describe the central 95% range over split assignments rather than population confidence intervals.

Table 8: Threshold sensitivity of the natural-domain audit with five anchors per inferred component. Label counts are per mechanism; accuracy is averaged over all eight mechanisms.

Identification rule	Inferred C	Labels	Saving	Topology acc.	Per-domain acc.
$z = 1.64$	11.21	56.1	19.9%	59.0	59.1
$z = 1.96$	11.22	56.1	19.8%	59.0	59.2
$z = 2.58$	12.85	64.3	8.2%	59.2	59.1

The primary analysis uses $z = 1.96$ and includes all eight mechanisms.

A second prospective audit evaluated 280 questions excluded from all development runs, with five anchor and 15 evaluation items in each of 14 domains. The Qwen answer on the unperturbed prompt is the fixed claim, and Phi agreement is the primary verifier. Historical Wilson intervals implied six signed components; their absolute orientations were discarded before fresh inference.

With five labels at each component root, topology calibration used 30 labels and achieved 76.2% on the 210 untouched evaluation items. Five labels in every domain used 70 labels and achieved 69.0%; however, the frozen global and raw orientations also achieved 76.2%. The primary criterion required topology to beat this global baseline; the observed tie provides no evidence that six independent signs were needed. The 40-label saving relative to domainwise calibration remains, but on this cohort the inferred components share the same absolute orientation. Counting the 400 historical labels, the deployment saving breaks even only after ten repeated batches (700 labels under either policy). Four secondary mechanisms inferred 11–13 components and yielded only 7.1–21.4% nominal deployment savings. The comparison separates label efficiency from evidence that the inferred domains carry independently variable signs.

An exploratory cross-domain stress test uses the earlier 400 frozen natural MMLU-Pro judgments without corruption. Across 2,000 stratified splits, each discovery half selects the five source–target domain pairs with the strongest apparent opposite signs; the other halves remain untouched. Mean source and target discovery accuracies are 82.4% and 27.3%. On held-out target items, however, raw accuracy returns to 51.5%, and expected cross-domain accuracy remains 51.5% after nine source anchors (central split interval $[42.3, 62.3]\%$). The selected opposition therefore does not persist on held-out claims, and the inferred source orientation does not transfer. The complete split distribution is provided in the supplementary material.

D.1.3 NATIVE-ROUTER TOPOLOGY AND HELD-OUT TRANSFER

The Granite audit observes the model's own sparse router rather than a post-hoc probe. For each of 32 layers and 40 experts, the observation is the left-candidate minus right-candidate pre-top-k router logit at the final prompt token. A deterministic rule fixed before labels are accessed partitions the 1,400-item discovery bank into 560 topology, 280 anchor, and 560 evaluation claims. The design fixes ordinary MMLU-Pro labels, one public distractor per question, balanced candidate order, and a category-matched wrong-question control. Topology is frozen before any anchor or evaluation label is accessed.

PCA compresses each layer's 40 expert differences to one topology axis. Maximum-spanning-tree signs align axes within every persistent absolute- correlation component. Components 0 and 1 contain layers 9–19 and 21–29; they recur under thresholds $.55$, $.575$, and $.60$ in both topology halves. Of 59 edges above the weakest full-sample threshold, all 59 retain their sign in each half. Fresh claims reuse the PCA axes, topology-set means and standard deviations, tree signs, components, and decoder without refitting.

The first prospective cohort excludes 2,800 previously touched MMLU identifiers and contains 420 anchor and 980 evaluation claims. At $k = 17$, within-target orientation yields 59.73% expected accuracy and cross-component orientation 51.01%, a paired-bootstrap difference of 8.71 points (95% CI $[2.39, 15.03]$). A second powered cohort was registered after this pilot. Its initial list contained 204 pilot identifiers and was replaced, before inference, by a disjoint cohort of 560 anchor and 420 evaluation claims. At the fixed $k = 51$, the two accuracies are 64.27% and 55.37%, a difference of 8.90 points (95% CI $[2.63, 17.12]$).

The prespecified composite criteria also required a 10-point gap and specified lower confidence bounds. The observed gaps of 8.71 and 8.90 points fall below those magnitude thresholds, and the powered cohort exceeds the proposed cross-component ceiling by 0.37 points. In both cohorts, the positive matched-minus-cross ordering recurs under a topology frozen without labels.

D.1.4 MODEL-HELD-OUT LAYER MECHANISMS

A separate replication freezes the extraction and graph rule before loading Qwen2.5-1.5B. Each of 790 TruthfulQA claims pairs the best truthful answer with the best false answer in balanced order. At layers 7, 14, 21, and 28, the observation is the last-token hidden-state difference between the two candidates. Leading axes and the absolute-correlation graph use 215 topology claims without truth labels; 196 disjoint claims supply anchors, and 379 claims remain for evaluation. The primary graph threshold is 0.4.

The graph separates layer 7 from the component containing layers 14, 21, and 28. Both primary component edges retain their signs on the evaluation claims; the third edge, included at the weaker 0.3 threshold, does so as well. The component reaches 70.4% held-out orbit accuracy (95% CI $[65.7, 74.8]\%$). Exact finite-population orientation increases its expected accuracy from 58.1% with one anchor to 66.6% with nine and 69.0% with seventeen. The anchor-pool sign and evaluation sign agree.

Pairing the same candidates with deterministically deranged questions produces a different component, containing layers 7, 14, and 21, with 62.0% orbit accuracy. Stable internal orientation therefore survives model-held-out evaluation, but does not by itself establish that the coordinate is grounded in the question. The result complements the router audit: native computation can transport an orientation before any observation fixes what that orientation means.

D.2 CLOSED CHANNELS AND GROUNDED EVIDENCE

D.2.1 MATCHED ITEMS AND MODEL OBSERVATIONS

Matched item construction. The original candidate bank contains 1,400 MMLU-Pro questions. Excluding 460 items previously opened during permutation development, confirmatory smokes, and a fresh 40-item study leaves 940 items. A deterministic alphabetical-category round-robin rule selects 400. The selected identifiers, questions, transformations, and evaluation code are frozen before the matched target-model analyses.

For the distractor transformation, we exchange the positions of two incorrect choices and update their letters in the prompt. The correct choice and its displayed position remain fixed. The operation preserves question semantics and the target answer letter while perturbing irrelevant alternatives.

Local inference. Qwen2.5-3B-Instruct and Phi-3.5-mini-instruct use their native chat templates. The next-token logits for answer letters A–J are renormalized within this set, and the maximum determines the answer. No chain of thought, prompt tuning, training, or post-hoc calibration enters the main 400-item audit. Ten thousand item-bootstrap resamples estimate confidence intervals for risk differences.

API sampling and output scores. DeepSeek V4-Flash runs in non-thinking mode at temperature 0.4. Each of the 400 original prompts and 400 positive twins receives 40 independent responses. The run records the returned model identifier and system fingerprint. Thirteen of 32,000 responses are unparsable; item majorities remain well-defined. The main agreement score is the empirical frequency of the modal answer on the original prompt.

For the mechanistic output-head audit, each response also records the top 20 first-token log probabilities. We strip surrounding whitespace from exposed answer-letter token variants, exponentiate their log probabilities, and sum the mass associated with the frozen modal letter. Any unexposed token is assigned zero; the score is consequently a conservative lower bound on the modal answer class's vocabulary mass. Scores are averaged over the same 40 original-prompt samples. Thresholds 0.9, 0.99, 0.999, and 0.9999 are fixed independently of correctness.

D.2.2 Closed-sample scaling

Figure 5 uses the first $n \in \{1, 2, 4, 8, 16, 32, 40\}$ responses according to the frozen sample index for every original prompt. At each prefix, the modal-frequency interval is the two-sided 95% Wilson interval and certificate acceptance requires modal agreement of at least 0.9. No prefix or threshold is chosen from the resulting correctness labels.

D.2.3 Grounded-channel threshold frontier

Table 9: Certificate frontier on 70 fixed MMLU-Pro answers. Error is conditional on non-abstention. "Mismatch" supplies another same-domain item's reference under the identical 40-sample protocol.

	Closed repetition		Mismatch control		Item-matched reference	
γ	Cov.	Error	Cov.	Error	Cov.	Error
0.80	95.7%	22.4%	85.7%	20.0%	98.6%	2.9%
0.90	92.9%	20.0%	81.4%	19.3%	95.7%	1.5%
0.95	88.6%	17.7%	80.0%	19.6%	95.7%	1.5%
1.00	81.4%	12.3%	74.3%	13.5%	94.3%	1.5%

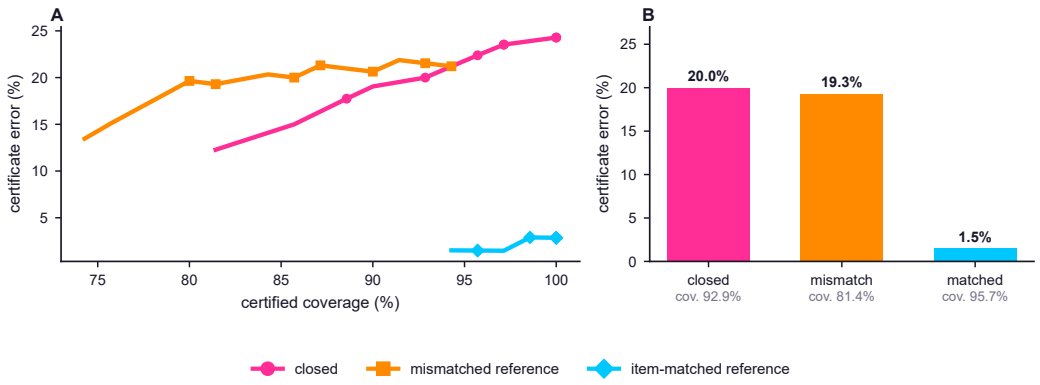

Figure 6: An equal-format mismatched reference fails to improve the closed frontier; only the item-matched reference sharply reduces certificate error.

Table 10: Full matched-400 counts.

Model	High correct	High wrong	Lower correct	Lower wrong
Qwen2.5-3B	109	150	15	126
Phi-3.5-mini	126	166	18	90
DeepSeek V4-Flash	249	83	18	50

D.2.4 Controlled decision boundaries

The resolution hierarchy uses a balanced Bernoulli truth variable and 120,000 draws. The four observations are nested tuples: a truth-independent anchor; the anchor plus a binary symmetric semantic view with accuracy 0.65; those variables plus three conditionally independent grounded views with accuracy 0.80; and the preceding tuple plus truth itself. Empirical cell frequencies estimate Equation 25 and determine the optimal cellwise majority error.

The full-coverage-boundary experiment uses five conditionally independent binary views and 160,000 draws at each of 13 equally spaced noise levels from 0.02 to 0.48. The exact observation laws are $\mathrm{Binom}(5, \rho)$ and $\mathrm{Binom}(5, 1-\rho)$. Their total variation is computed by summing the six probability-mass differences.

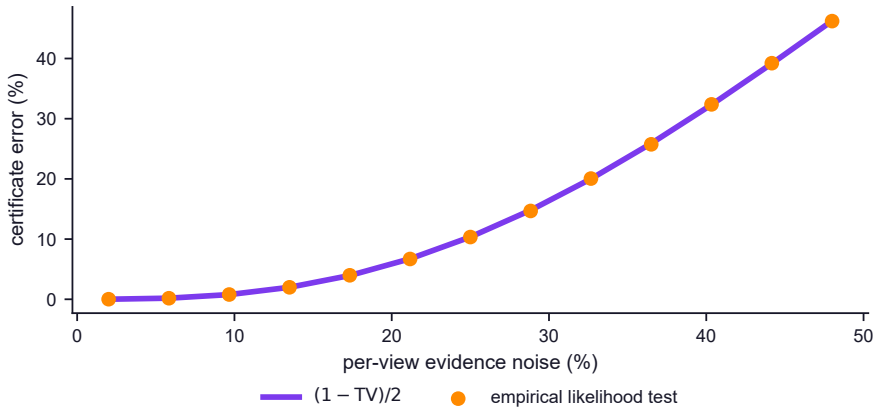

Figure 7: Controlled verification of Proposition 16. Empirical likelihood-test error follows the exact equal-prior $(1 - \mathrm{TV})/2$ boundary across 13 evidence-noise levels.

D.3 Evidence acquisition and verification

D.3.1 Adaptive executable evidence

Table 11: HumanEvalPlus binary certificate error (%). Δ_3 compares three probes with compile-only evidence; intervals are paired 95% bootstrap intervals.

Model	Accuracy	Compile	1 probe	2 probes	3 probes	Δ_3
Qwen	54.4	41.8	19.0	15.2	14.6	-27.2 $[-34.2, -20.3]$
Phi	32.9	39.9	13.9	12.0	11.4	-28.5 $[-35.4, -21.5]$
DeepSeek	80.4	12.7	7.6	7.0	7.0	-5.7 $[-9.5, -2.5]$

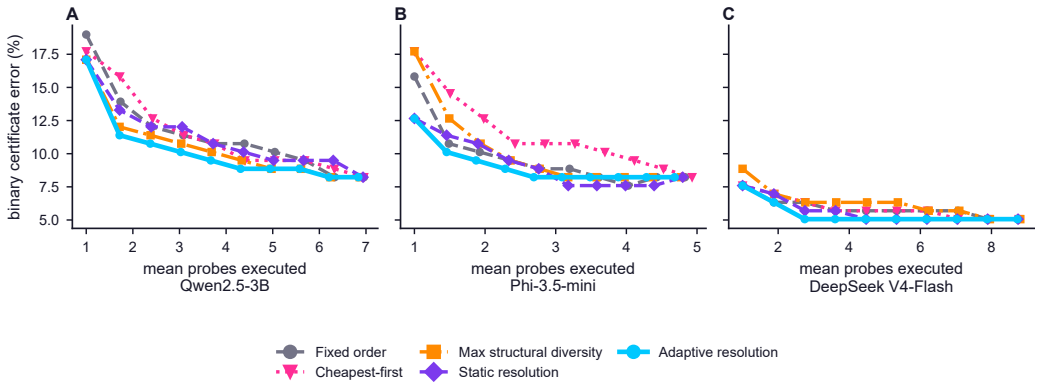

Figure 8: Complete HumanEvalPlus acquisition frontiers for Qwen, Phi, and DeepSeek. Curves compare the frozen non-oracle acquisition policies at identical candidate pools; lower certificate error and fewer mean executions are preferred. Matched-error cost is the primary comparison because the fixed-budget discordant counts are small.

D.3.2 Fourteen-domain observation-channel audit

The breadth suite consists of the complete 70-item MMLU-Pro validation split, with five questions from each of 14 domains. The protocol freezes items, prompts, temperature, sample counts, parse rules, the error-cohort criterion, and the paired test before target inference.

Reference rationales come from the pre-existing benchmark material. A deterministic sanitizer removes explicit answer-letter declarations while preserving task-relevant content. Stage 1 uses 40 responses per item in both closed and grounded conditions. The original closed majority is then frozen as the answer being certified. For each threshold, we compute the two ternary rules in Section C.3 directly from the modal labels and empirical agreements. Once the 13-item coherent-error cohort is fixed, Stage 2 obtains 100 new closed responses per item.

For the negative control, each domain's five items are sorted by frozen ID and their sanitized rationales are cyclically shifted by one position. The resulting derangement assigns no item its own rationale, while domain, prompt format, instruction, model, temperature, and 40-response budget remain matched. The derangement is fixed before evaluation. The primary comparison uses $\gamma = 0.9$; the plotted curve reports the complete threshold sweep from 0.5 to 1. A paired exact test uses items accepted by both reference channels, and 10,000 bootstrap replicates sample five items with replacement inside each of the 14 domains.

Statistical test. For paired majority outcomes, the exact two-sided sign test equals the exact McNemar test when ties are discarded. Stage 1 contains 14 grounding-only gains and no closed-only gain, so $p = 2/2^{14} = 0.00012207$. Stage 2 contains 12 grounding-only corrections and no repetition-only correction, so $p = 2/2^{12} = 0.00048828$. At $\gamma = 0.9$, the matched-versus-mismatched certificate comparison on 55 jointly accepted items contains nine matched-only correct decisions and no mismatched-only correct decision, giving $p = 2/2^9 = 0.00390625$.

Table 12: Domain coverage of the frozen coherent-error cohort.

Domain	Errors	Domain	Errors
Business	1	Chemistry	1
Computer science	1	Engineering	1
Health	1	History	1
Law	1	Mathematics	2
Philosophy	2	Physics	2

D.3.3 EXECUTABLE-VERIFIER AUDIT

HumanEvalPlus augments HumanEval with large automatically generated test suites (Chen et al., 2021; Liu et al., 2023). Each task in the official 164-task release receives one greedy continuation from each of two local models and DeepSeek V4-Flash. DeepSeek uses the official API in non-thinking mode with temperature zero. The shared sanitizer removes Markdown fences, removes any common presentation margin, and adds exactly one four-space function-body level while preserving relative indentation. It never uses tests or correctness. Generated code is parsed before execution; filesystem, process, network, dynamic-execution, and non-whitelisted imports are rejected. Each candidate runs in a fresh isolated Python process with a 12-second timeout.

For each task, a deterministic ordering fixed before evaluation selects the first three test indices as probes; the complement defines truth. The one- and two-probe bundles are prefixes of this ordering, so costs are nested and no output selects tests. Six tasks cannot be sliced by the literal-I/O parser and are excluded for all three models by this format-only rule. Corrupted probes retain inputs but rotate the three expected outputs. Ten thousand paired item-bootstrap replicates form risk-difference intervals; exact sign tests use discordant certificate errors.

For adaptive acquisition, an independent deterministic ordering selects ten candidate probes; its complement defines truth. Eighteen literal-AST features describe size, depth, numeric magnitude, strings, containers, and special values for inputs and outputs. A depth-three histogram gradient booster estimates probe-failure probability. Qwen scores are five-fold task-cross-fitted. Target scores average this feature prediction with the frozen Qwen outcome for the same task and probe. The adaptive diversity coefficient is selected from $\{0, .05, .10, .25, .50\}$ on Qwen only; 0.10 is then frozen. Random-order curves average 100 deterministic permutations. An independent split assigns 40 design and 118 calibration tasks. It selects coefficient 0.50 on design data, uses no task-specific source outcome, and applies Clopper–Pearson tails $\delta/(2 \times 4)$ to the declared budgets $\{1, 2, 4, 8\}$.

D.3.4 SCIFACT OPEN-CORPUS ACQUISITION

The official SciFact release retains its original train/development split: 809 training claims, 300 development claims, and 5,183 evidence abstracts. Word unigram/bigram TF–IDF retrieves a common top-20 abstract set from the full corpus. Claim-disjoint training fits ordinary cited-source relevance and expert-rationale resolution scores from lexical overlap, title similarity, number/negation match, position, and length features. All policies then rank the same retrieved abstract sentences. The sentence-level generic reranker has the same logistic model class and fit claims, but its positive target is the maximum-TF–IDF sentence within each train cited source; neither rationale membership nor SUPPORT/CONTRADICT labels enter that target. The held-out evaluator computes the first acquisition step at which every sentence in any one annotated rationale set has been collected; partial sets count as incomplete. The offline completion metric remains unavailable to the acquisition policy as a stopping signal. The dev split contains 188 evidence-bearing claims, of which 174 have a complete rationale in the common retrieved set. No dev rationale label is used to train, calibrate, or order a policy.

Table 13: SciFact discovery on 174 retrievable claims. Cost is mean sentences to complete a rationale; @k is completion within budget k (%).

Acquisition rule	Cost ↓	@3 ↑	@10 ↑
Lexical sentence	20.49	37.8	54.8
Cited-source relevance	13.78	33.5	66.5
Generic sentence relevance	19.03	39.4	57.4
Random acquisition	67.07	3.2	11.2
Resolution-aware	**10.15**	**56.4**	**74.5**

Table 14: Certificate decisions for the 13 preregistered coherent errors. The target is the truth of the original fixed answer.

Evidence channel	Correct "incorrect"	False "correct"
100 fresh closed samples	0/13	13/13
40 reference-conditioned samples	12/13	1/13

D.4 Failure witnesses and robustness

D.4.1 Frozen coherent-error certificate witness

Every target answer in this cohort is wrong by construction. Fresh closed samples preserve and therefore support the same wrong majority on all 13; reference-conditioned evidence contradicts it on 12. The one false certificate is important: source access alone does not satisfy Theorem 13.

D.4.2 Persistent shared-anchor audit

The transferred shared-bias audit uses 100 BBH multiple-choice items, stratified into 50 correct and 50 incorrect direct predictions and sampled round-robin across task subsets. Qwen2.5-3B-Instruct produces eight stochastic rationales per item at temperature 0.7 and top-$p = 0.9$. An item-specific cyclic successor of the direct argmax is inserted as a persistent anchor during both rationale generation and final option scoring. The output layer is evaluated in FP32.

The manipulation intentionally creates a common cause across views. Consequently, its 62/79 high-agreement error count is evidence about a failure mode of channel composition, not an estimate of the frequency of shared bias in ordinary prompting.

D.4.3 Robust partial-identification analysis

A finite-sample linear program isolates calibration from certification without claiming to estimate the prevalence of natural shifts. For each model, isotonic calibration is fitted on 100 development items and a top-25% acceptance rule, including uniform randomization at the boundary tie, is frozen before the 100 test labels enter the audit.

The ambiguity set contains test weights q_i satisfying $\sum_i q_i = 1$, $1/\rho \le n q_i \le \rho$, and $\sum_i q_i a_i = .25$. For each of raw confidence, agreement, normalized entropy, and first–second distance, it also preserves the source first moment and the mass of every empirical decile; tied values are never split into distinguishable groups. The source uniform weights are feasible for every $\rho \ge 1$. Because coverage is fixed, minimizing or maximizing the accepted error numerator is a linear program. Its sharp identified set is

$$\mathcal{I}_\rho(a) = \left[\min_{q \in \mathcal{A}_\rho} \frac{\sum_i q_i a_i e_i}{.25},\ \max_{q \in \mathcal{A}_\rho} \frac{\sum_i q_i a_i e_i}{.25} \right]. \tag{28}$$

The intervals are identified sets under explicitly bounded reweighting, not confidence intervals and not estimates of natural shift frequency. Labels are used adversarially to attain the endpoints, so the result establishes non-identification under the declared observables rather than a deployable shift detector.

Table 15: Complete density-ratio sensitivity of the sharp accepted-risk interval (%). All diagnostic constraints have numerical residual below 2×10^{-15}.

Model	$\rho = 1$	$\rho = 1.5$	$\rho = 2$	$\rho = 4$	$\rho = 8$
Qwen2.5-3B	58.7	[42.4,72.4]	[32.7,79.3]	[14.7,89.7]	[7.3,94.8]
Phi-3.5-mini	62.1	[48.2,73.5]	[39.4,80.3]	[22.3,90.5]	[9.1,95.3]

D.4.4 A POSITIVE CONTROL FOR CONSISTENCY

A small free-text GSM8K control provides a regime in which consistency substantially improves probability quality. Qwen2.5-3B-Instruct receives the standard eight-shot chain-of-thought prompt and produces 20 full generations for each of ten items at temperature 0.4. The official extractor parses the answers.

Table 16: Free-text GSM8K consistency control. Lower Brier is better.

Quantity	Value
Accuracy	70%
Parse rate	100%
Raw reciprocal-perplexity Brier	0.2504
Agreement Brier	0.0865
Entropy Brier	0.0693
First–second distance Brier	0.0433

The control is too small for population inference, but it shows that consistency scores can be informative in this regime.

E SCOPE OF THE CLAIMS

The exact results quantify over a declared per-instance world class $\mathcal{W}_x$. Truth-homogeneous compatibility cells characterize certification within that class, conditional support separation marks its zero-error endpoint, and an equal-law fiber containing both truth values makes the endpoint unattainable through closed evidence.

For multiple choice, $\mathcal{W}_x^{\mathrm{MC}} = \{w_y : y \in \mathcal{A}_x\}$ with $Y(w_y) = y$. For code, $\mathcal{W}_x^{\mathrm{code}}$ contains total input–output specifications consistent with the task and execution semantics. The released HumanEvalPlus held-out suite samples this class but does not exhaust it; finite probes can leave specifications that agree on observed inputs and diverge elsewhere. MMLU and HumanEvalPlus therefore measure partial resolution under finite evaluators. Simultaneous confidence bounds protect calibration-distribution selection, not arbitrary transfer distributions.

SciFact isolates acquisition from a frozen corpus. Lower complete-rationale cost demonstrates more efficient discovery of benchmark-annotated evidence, while gold completion remains an offline evaluator rather than a deployable certificate, stopping rule, or source-trust guarantee. Ranking, calibration, selective risk, and certification remain distinct objectives. Only the last requires information that separates the relevant truth alternatives.