PLACEMENT, NOT QUANTITY: HOW TRANSFORMERS LOCK INTO THE WRONG RULE, AND WHY THE REPAIR SIGNAL IS INVISIBLE TO THE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

A model trained under ambiguous supervision can lock into a wrong rule: a shortcut that gives the same labels as the intended rule on the training distribution, but different labels outside it. We ask what repairs such a model, and we find that the answer is placement, not quantity. A few counterexamples repair the lock-in only when they are placed to cover the region where the two rules disagree. How many counterexamples we add, how uncertain the model is about them, and how diverse they are do not matter. We show this in a controlled three-valued-logic testbed, where an intended rule and an inductively simpler shortcut agree on the training distribution by construction, so any preference of the model is inductive. Across three lock-in families, selecting counterexamples that cover the disagreement region beats selecting the same number at random or by diversity, with paired-bootstrap CIs that exclude zero. A direct-sampling intervention shows that region coverage is causally necessary at every budget we test: selection without coverage never closes the gap, even at four times the minimal pair. Eight placed counterexamples outperform an unplaced dose of 3200 on the held-out disagreement-region test, because off-target examples dilute the few that cover the region. Lock-in also follows a directional regularity: in all four families the model commits to the inductively simpler rule, whichever rule we declare correct, and re-scoring one family's weights under a flipped ground truth confirms the symmetry. Finally, in this testbed the repairing counterexamples cannot be found in advance. The signal is invisible to the model's confidence, to value ranking, and to ensemble disagreement: every locked seed commits to the same rule, so a committee is unanimous exactly where repair is needed (ranking AUC 0.500), and the model is confidently wrong where it errs. On a disagreement region of this kind, diagnosis must be interactive. A CelebA demonstration carries the placement signal to natural data: greedy failure-region selection beats budget-matched random with a paired CI excluding zero, on ten seeds pooled and on the five pre-registered extension seeds alone. There the failure region is one-sided and the locked model's own loss also finds it, so the invariant is the coverage requirement, not the selector.

1 INTRODUCTION

Neural networks, including neuro-symbolic models that pair a learned perception module with a symbolic reasoner, often reach high in-distribution accuracy by committing to an unintended rule, and then generalize poorly out of distribution. Such *reasoning shortcuts* (Marconato et al., 2023b) appear whenever the training signal is consistent with more than one rule and the model settles on one of them.

Prior work studies when these shortcuts form and prevents them with concept-level supervision. We ask what happens after a shortcut has formed: can a committed wrong rule be repaired, and by what? Our answer is that repair depends on where the counterexamples sit, not on how many there are, how uncertain the model is about them, or how diverse they are. The surprising part is uncertainty. In standard practice one ranks candidate examples by the model's uncertainty; in our setting this ranking carries no signal.

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.

Our contributions are the following.

- **Placement, not quantity (Section 3).** Across three lock-in families, selecting a few counterexamples that cover the disagreement region beats selecting the same number at random or by diversity on held-out OOD, with paired-bootstrap CIs that exclude zero. A handful of placed counterexamples beats an unplaced dose two orders of magnitude larger.
- **Region coverage is causally necessary (Section 4).** A pre-registered direct-sampling intervention imposes coverage by construction. A cross-region pair beats a within-region pair in every seed at every budget, and the advantage grows with the budget instead of closing. Selection without coverage never catches up.
- **The repair signal is invisible to the model (Section 5).** In the testbed, no confidence-based, structural, value-based, or committee selector that we pre-registered finds the repairing counterexamples. A committee of locked seeds is unanimous exactly on the disagreement region, so diagnosis must be interactive. On a one-sided natural region (CelebA) the model's loss does find the region, so the boundary is the geometry of the region, not the method.
- **Lock-in tracks inductive preference (Section 6).** Across four families the model commits to the inductively simpler rule, whichever rule we declare correct. Re-scoring one family's weights under a flipped ground truth confirms the symmetry.

The evidence comes from more than 1800 training runs on the logic testbed (four rule families, 10 seeds each, greedy, random, diversity, direct-sampling and test-blind selector arms, budgets up to k = 8 and doses up to 3200 examples) and 49 CelebA repair runs. Every confirmatory experiment was pre-registered in one of nine hash-frozen documents, all results are reported per seed, and the released code recomputes the headline numbers from the released per-seed files (Appendix D).

We use a controlled synthetic testbed, propositional logic under a three-valued semantics, because isolating inductive preference needs a training distribution that is strictly neutral between the two rules (Section 2). A CelebA demonstration (Appendix B) is a first step beyond it. Figure 1 sketches the geometry.

2 SETUP: ISOLATING INDUCTIVE PREFERENCE

Our testbed is propositional logic under strong Kleene's three-valued semantics (Kleene, 1952): a formula over an expression tree evaluates to true, false, or unknown (Figure 1). With three values we can construct an intended rule R^* (the full strong-Kleene semantics) and a shortcut rule $R^\dagger$ that are provably equivalent on the training distribution but differ on a designated disagreement region. A model trained only where R^* and $R^\dagger$ agree has no distributional reason to prefer either rule. Any preference it shows is inductive, and isolating that preference is the purpose of the testbed.

We train a compact transformer (transformer_t64_l2) under ambiguous supervision, that is, on labels that are consistent with both R^* and $R^\dagger$, and we ask which rule it adopts. A model has locked in when it commits to $R^\dagger$. We define this at two levels. Behavioral lock-in is the commitment itself: on the disagreement region and on a held-out OOD test (kept disjoint from everything used to select counterexamples), the model generalizes by the shortcut and not by the intended rule. The strict lock-in signature adds a mechanical gate: in-distribution accuracy ≥ 0.90, disagreement-region accuracy ≤ 0.80 or mean OOD accuracy ≤ 0.80, and predictions closer to $R^\dagger$ than to R^* (thresholds fixed before the runs; Appendix D). The two levels need not coincide. In the always-abstain family the model locks in behaviorally in 10/10 seeds but meets the strict signature in 8/10; this is the one family whose shortcut is only description-length-*tied* with R^*, the weakest simplicity gap we measure (Appendix C). The depth-shortcut and unknown-as-false families lock in behaviorally in 10/10 seeds. Lock-in is not specific to transformers: two GRU variants lock in 5/5 seeds in the depth-shortcut and always-abstain families, and placement replicates on the GRU (greedy$-$random $+0.105$ at k=2, CI $[+0.060, +0.157]$, 10/10 seeds; Appendix A). The full characterization below is on the transformer. Architecture, serialization, data sizes, and optimization are given in Appendix D.

The testbed is synthetic because natural data cannot guarantee this neutrality; synthesis guarantees it by construction and allows the intervention of Section 4, where we impose the suspected cause and measure its effect. The cost is external validity (Appendix B, Section 7).

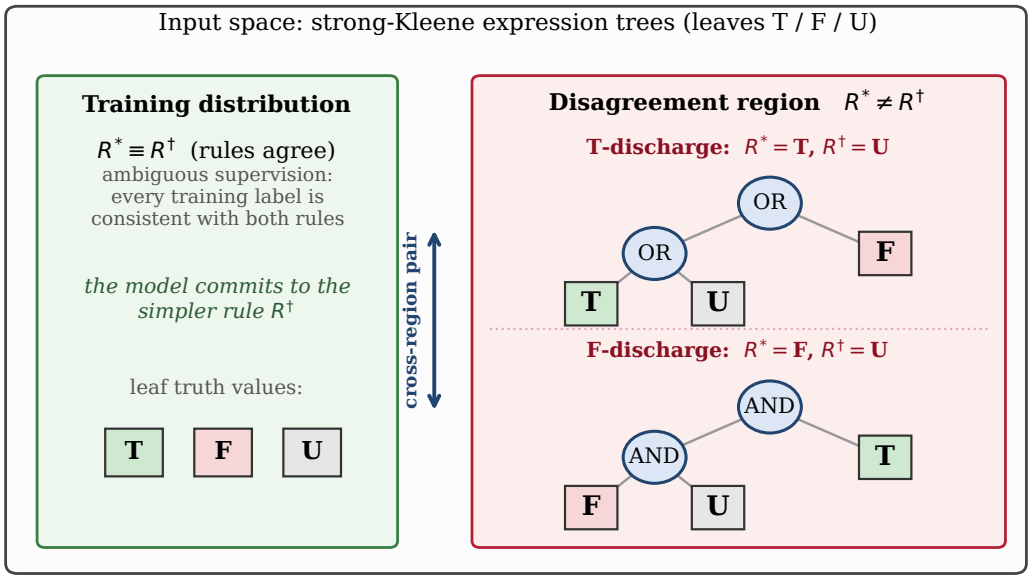

Figure 1: The testbed. The intended rule R^* (strong-Kleene) and the simpler shortcut $R^\dagger$ agree on the training distribution and differ only on the disagreement region, which splits into two discharge sub-regions; a cross-region pair places one counterexample in each.

Box 1 — A worked example (the depth-shortcut family). The intended rule R^* is strong-Kleene three-valued logic. The simpler shortcut $R^\dagger$ that the model locks into, *shallow-root dominance*, applies strong-Kleene's *discharge* rule at the root only: every child subtree that contains an unknown is collapsed to *unknown* (no discharge inside subtrees), and then the root operator is evaluated by strong Kleene on the two collapsed child values. On depth-1 expressions the two rules agree (OR(U, T) is T under both), so they disagree only on nested discharge cases, and these come in two independent directions. In the **T-discharge** sub-region a disjunction discharges an unknown to *true* through its true branch: for OR(OR(T, U), F), R^* = T, but $R^\dagger$ = U, because the left subtree contains an unknown and is collapsed to U, and OR(U, F) = U at the root. In the **F-discharge** sub-region a conjunction discharges an unknown to *false* through its false branch: for AND(AND(F, U), T), R^* = F, but $R^\dagger$ = U, because AND(U, T) = U at the root. Discharging to true and discharging to false are separate behaviors, so a **cross-region pair**, one counterexample from each sub-region, teaches both. Training on both directions moves the model from $R^\dagger$ toward R^*; training on two trees from a single sub-region teaches only one direction and leaves the model wrong on the other. This is why covering both sub-regions is necessary and not only helpful (Section 4: cross-region beats within-region in 10/10 seeds).

3 Placement, not quantity: repairing wrong-rule lock-in

Repairing a model that has locked onto the wrong rule depends on *where* a handful of counterexamples sit in input space, not on how many there are. The counterexamples that matter cover the regions where $R^\dagger$ and R^* disagree; the same number of counterexamples that do not cover those regions do not repair the model. We call this the placement thesis: repair is governed by placement, and not by quantity, by the model's uncertainty (Section 5), or by diversity. We test it on the three families whose shortcut $R^\dagger$ is inductively simpler (depth-shortcut, always-abstain, and unknown-as-false). In all three, ambiguous supervision brings the model to high in-distribution accuracy while it generalizes by $R^\dagger$ and not by R^*.

Placement beats quantity, and beats volume. Repair means retraining from the same initialization on the ambiguous training set plus the k selected counterexamples (Appendix D). Greedy is forward selection from a fixed pool of 20 structurally clustered candidates per (family, seed); at each step it adds the candidate that most improves accuracy on a held-out counterexample-fit set, and it never

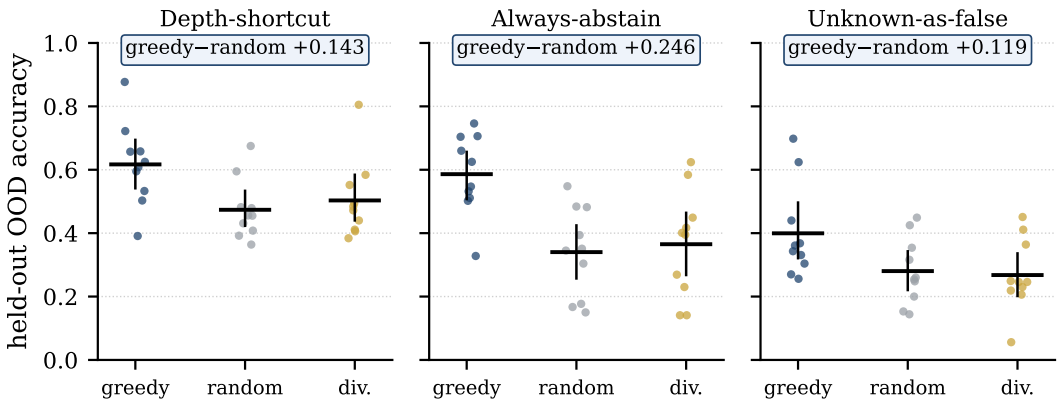

Figure 2: Placement, not quantity. Across three lock-in families, greedy region-coverage selection beats random and diversity on held-out OOD at $k = 2$ (per-seed dots, bootstrap 95% CIs; greedy−random means annotated).

sees the OOD test or the sub-region labels. Diversity picks the k candidates of the same pool that are farthest apart in a structural feature space (Appendix A). Selecting k counterexamples greedily to cover the disagreement region outperforms selecting the same k at random on held-out OOD. At k = 2 the greedy advantage is +0.143 [+0.077, +0.208] in the depth-shortcut family (9/10 seeds), +0.246 [+0.193, +0.308] in the always-abstain family (10/10 seeds), and +0.119 [+0.081, +0.164] in the unknown-as-false family (10/10 seeds). Every paired-bootstrap 95% CI excludes zero (Figure 2). The count is the same in all arms; only the placement differs. The contrast with volume is even larger. Here we score counterexample-test accuracy (CE: accuracy on the held-out disagreement-region test set of Appendix D, which is separate from the OOD firewall), so that both arms are read on one set. A few placed oracle counterexamples reach CE 0.905 at 8 examples and CE 0.995 at 25 examples. A broad, unplaced dose of 3200 examples, about two orders of magnitude more, reaches only CE 0.732. The reason is dilution, not rare coverage: the broad dose already contains ≈640 on-target items, so coverage is saturated, but its ≈2560 off-target examples pull it below what 8 placed examples achieve. More data does not replace placed data. Both arms use the same unweighted recipe; we have not tested whether reweighting the broad dose (Idrissi et al., 2022) narrows the gap.

Placement is not diversity. Greedy does more than choose varied examples: it beats a diversity baseline that maximizes example variety directly. In the two families where we pre-registered this comparison, greedy wins with paired-bootstrap 95% CIs that exclude zero: +0.221 [+0.153, +0.290] at k = 2 and +0.130 [+0.053, +0.199] at k = 3 in the always-abstain family, and +0.132 [+0.079, +0.191] (10/10 seeds) at k = 2 in the unknown-as-false family. The depth-shortcut family shows the same ordering (greedy over diversity in 8/10 seeds); its main evidence is the mechanism test in Section 4. The IID guardrail holds throughout: no selection arm reduces IID accuracy by more than 0.05 in any family.

Durable versus diffuse repair. What greedy places are cross-region pairs, counterexamples that span both sub-regions where $R^\dagger$ and R^* disagree, so placement has a geometric structure. This structure is visible in how the advantage behaves with the budget (Figure 3). Where region coverage is discrete, in the depth-shortcut family, the advantage is durable: the greedy−random gap is +0.143 at k = 2 and still +0.127 at k = 5, and random never closes it. Where coverage is diffuse, in the always-abstain family, random eventually supplies it by chance and the gap closes from +0.246 to +0.083. A region-coverage account predicts this split: discrete coverage must be placed at every budget, while diffuse coverage is eventually hit by chance. Section 4 tests the causal claim directly.

4 The mechanism: region coverage is causally necessary

Region coverage is not only the structure that greedy happens to place (Section 3). In the depth-shortcut family, whose disagreement region splits into two discrete sub-regions, it is causally necessary for repair: without it, repair plateaus. Its benefit also grows with the counterexample budget

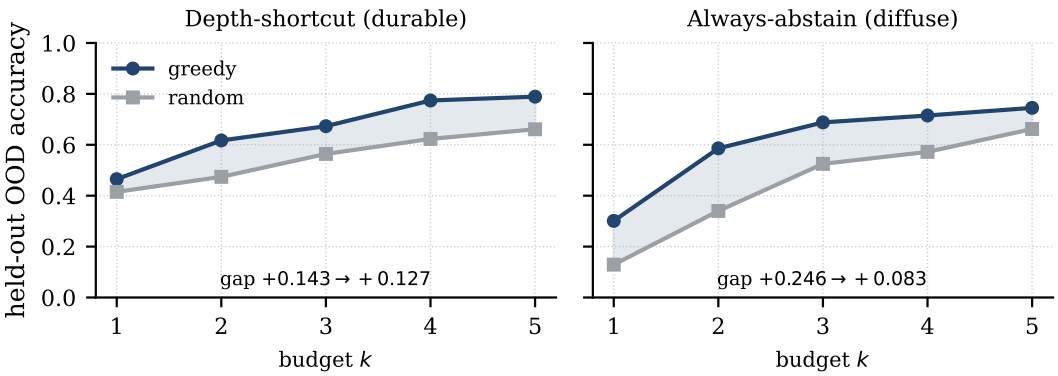

Figure 3: Durable vs. diffuse repair. Greedy−random gap over $k = 1..5$: durable in the depth-shortcut family, closing in the always-abstain family.

through k = 6, and the within-region alternative never catches up, even at k = 8. We establish both claims on the depth-shortcut family, the one family whose repair structure is geometrically characterized, with a pre-registered test (S2) that samples counterexamples directly by region instead of letting greedy choose them. Direct sampling turns a correlation into a cause: if region coverage still drives repair when we impose it by construction, the effect cannot be an artifact of how greedy ranks examples.

The test samples a cross-region set (at k = 2, one counterexample from each disagreement sub-region) and compares it with a within-region set drawn from a single sub-region. For larger budgets the cross-region set is filled round-robin over the two sub-regions in random order (an odd budget gives one sub-region one extra item), the within-region set draws all k from one randomly chosen sub-region, and both are re-drawn in each of five repeats per seed. The cross-region pair wins on held-out OOD in 10/10 seeds at every budget k = 2, 3, 4 (Figure 4). The within-region pair, with the same count but confined to one sub-region, does not. Region coverage is the only factor that varies and the only one we impose, so the advantage is due to it and not to the greedy procedure.

S2 also overturned a bet that we had pre-registered. In the first experiment we gated on G2 (N1 in Table 1), the hypothesis that the advantage concentrates at the smallest budget, with a peak at k = 2. G2 failed: the dissociation appeared in only 4 of 10 seeds, against a pre-registered bar of 7 of 10. S2 shows why: the cross-region advantage grows instead of peaking. Cross-region OOD rises from 0.531 to 0.590 to 0.646 across k = 2, 3, 4, within-region stays flat at 0.412, 0.414, 0.442, and the greedy selection rises in parallel, 0.617 to 0.673 to 0.774. An extended sweep (Appendix A) continues the rise to 0.685 and 0.729 at k = 5 and 6. The cross-region *level* saturates by k = 8 (0.727), but the within-region arm never moves (0.443–0.450), so the *advantage* widens instead of closing: +0.28 at k = 8, with every paired-bootstrap CI excluding zero, and greedy continuing to 0.840. Greedy also sits above direct cross-region sampling at every budget (0.617 vs. 0.531 at k = 2), so coarse coverage is the necessary condition, and instance selection inside the covered sub-regions adds a further lift. There is no k = 2 peak: coverage keeps paying off as the budget grows.

A second direct-sampling check, run independently in the earlier experiment, confirms the same ordering on the training-side metric: cross-region beats within-region in 10/10 seeds, again with greedy out of the loop. This clean characterization holds only for the depth-shortcut family. In the always-abstain and unknown-as-false families greedy still beats random and diversity (Section 3), but the repair structure is diffuse: single-cell partitions of the disagreement region do not localize it. Per-seed comparisons at each budget are in Appendix A, Figure 7.

Pre-registered negatives. We registered five hypotheses before running. Four were falsified at their frozen bars, and the fifth (N4) gives only a weak, non-deployable signal. Each is reported where it arises, and all are collected with their bars and sources in Table 1 (Appendix A).

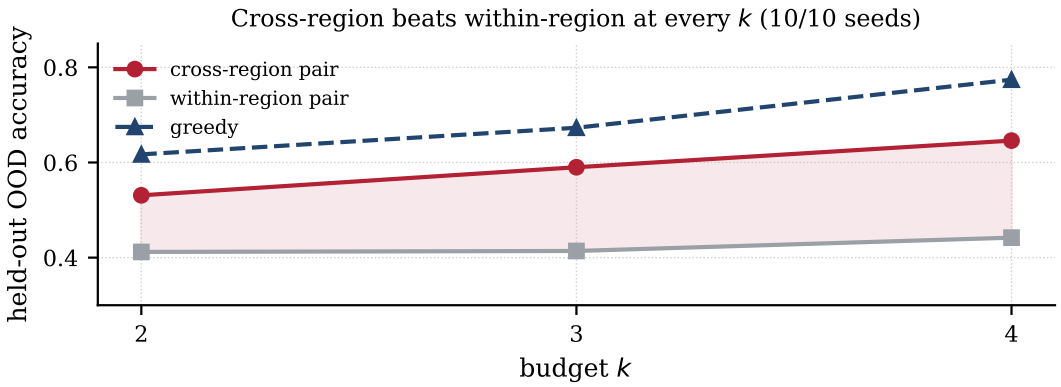

Figure 4: Region coverage is causally necessary. Directly sampled cross-region beats within-region on held-out OOD at every k (10/10 seeds); the greedy envelope tracks cross-region.

5 The repair signal is invisible to the model

Sections 3 and 4 show that a few well-placed counterexamples repair a locked-in model and that placement means region coverage. The operational question is whether those counterexamples can be found in advance: ranked by the model's own uncertainty, by a structural feature, or by any rule that does not run the held-out OOD test. Our answer is a clear negative. The value of a counterexample is **model-relative and confidence-invisible**: no static or confidence-based selector that we pre-registered recovers it. This forces the diagnosis of wrong-rule lock-in to be *interactive*, driven by behavior on held-out tests, and not by anything one can read off the model or the example beforehand. We show it along four independent axes, on both characterized lock-in families (depth-shortcut and always-abstain). The claim is scoped to regions of this kind, on which the locked model is confidently wrong on every branch; on a one-sided natural region its loss can find the region (Appendix B, Section 7).

A locked model is most confident where it is systematically wrong. Its softmax margin on the disagreement region is 0.984–0.997 across every selection, yet its prediction there never agrees with the intended rule R^*. Confidence is flat, high, and pointed in the wrong direction, so it cannot point to the model's own errors. Predictive entropy has the same blindness: it ranks disagreement-region examples at an AUC of only 0.643 ± 0.032 and 0.586 ± 0.044 on the two families, a weak signal that is different from 0.5 but far below any deployable selection threshold. The model is not uncertain where it is wrong, and margin does not separate the counterexamples greedy chooses from random ones (standardized mean difference -0.098, p = 1.000; +0.109, p = 0.697).

Specialness is relational, not a property of particular examples. The counterexamples greedy selects share no canonical example across seeds: the cross-seed Jaccard overlap of greedy's picks is exactly 0.000 (0 of 20 at k = 2, 0 of 30 at k = 3, both families). There is nothing universal to memorize. The value of a counterexample is defined relative to the specific locked model and the region coverage it still lacks, not by any intrinsic property of the example.

Every test-blind selector fails. We pre-registered three deployable, CE-blind selectors against a bar of recovering at least 60% of the greedy–random OOD gap in at least 8 of 10 seeds. None clears it. The strongest, a cross-region selector, recovers about half the gap (0.552; 55%) but in only 7 of 10 seeds, which is promising but too unreliable to deploy. A min-margin selector is actively harmful (-54%), and the one structural feature with a raw effect, num_u, recovers just 17% and does not survive multiple-comparison correction (Holm p = 0.117). A good selection *exists*: greedy finds one interactively, by forward selection on a held-out counterexample-fit test of 4000 disagreement-region items, never on the firewalled OOD evaluation (protocol: Appendix A). But no rule that avoids interactive behavioral testing recovers it. Interactive selection therefore assumes a labeled slice of the disagreement region: it optimizes a small repair budget given a behavioral test, and it does not by itself discover an uncharacterized shortcut.

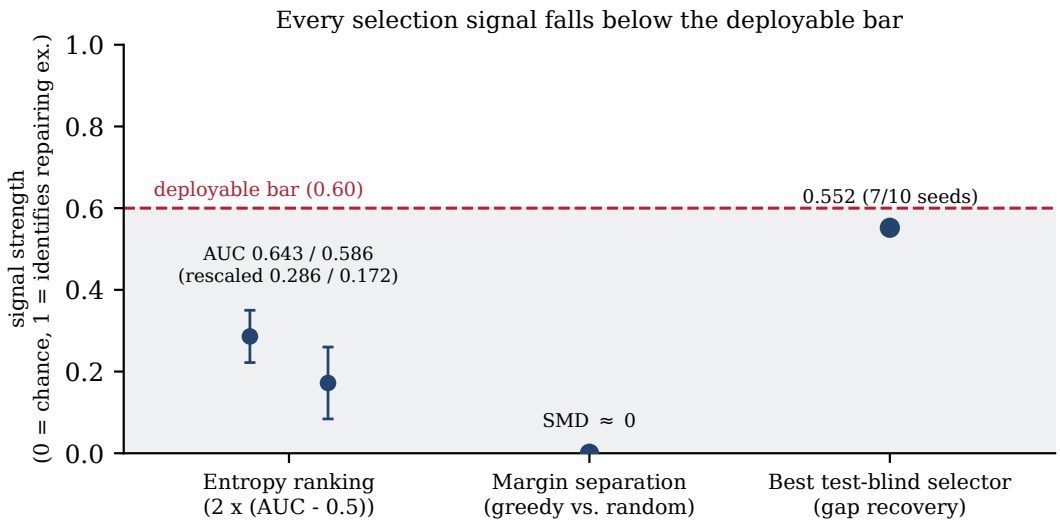

Figure 5: Three candidate selection signals on one axis: 0 is chance, 1 a signal that would perfectly identify the repairing counterexamples; the shaded band is the sub-deployable region. Entropy ranking AUCs (0.643/0.586) are plotted as $2(\mathrm{AUC} - 0.5)$, i.e. 0.286/0.172; margin separation is SMD ≈ 0; the best test-blind selector recovers 0.552 of the gap (7/10 seeds). Every mark falls inside the band. Cross-seed pick overlap (Jaccard 0.000, text) is instance interchangeability within the covered regions, not a ranking signal, and is not plotted.

The disagreement/committee family is blind too. The one acquisition family not covered by single-model confidence is disagreement-based active learning: query-by-committee and ensemble vote-entropy (Seung et al., 1992; Beluch et al., 2018), and Bayesian BALD (Houlsby et al., 2011; Gal et al., 2017). On the disagreement region this signal is structurally absent. Across all 10 seed-models and all 400 released disagreement-region candidates, every model commits to the same shortcut rule $R^{\dagger}$ (agreement-with-R^{*} rate 0/400) at a mean softmax margin of 0.986 and 0.997 on the two families; the whole committee is confidently wrong on every item. A committee built from these members is therefore unanimous, its hard vote-entropy is identically 0 (bootstrap 95% CI [0, 0]), and it cannot rank repairing counterexamples above the agreement region (AUC 0.500). Even the weak single-model entropy signal (0.643 / 0.586) collapses to exact chance. This is a member-level analysis computed from the released per-seed predictions; the full select→retrain→OOD committee loop is left as a compute-heavier confirmation. The reason is mechanical, and it is what the directional regularity (Section 6) predicts: seeds locked into the same simpler rule have no disagreement to exploit where repair is needed. We also tested heterogeneous committees (Xiong et al., 2021) (10 transformer + 10 GRU seeds; Appendix A). Heterogeneity removes the unanimity (vote entropy 0.506) but restores no signal: the disagreement is region-unspecific (AUC 0.630, about the same as single-model entropy), and every unanimous item is unanimous on $R^{\dagger}$. Cross-architecture disagreement does not land where repair is needed.

The failure is therefore not limited to confidence: the structure of an example, the diversity of a set, and ensemble disagreement fail independently, so the missing signal is a property of the problem and not of any one method (Figure 5). Wrong-rule lock-in cannot be diagnosed by inspecting the model; it must be probed.

6 Directional regularity: lock-in tracks inductive preference

A final observation reframes "wrong-rule lock-in": across our four families the model commits to whichever consistent rule is inductively simpler, and lock-in is this commitment seen from a rule that we declared correct. The evidence base is four constructed families; the findings of Section 3–Section 5 stand independently of it.

Lock-in direction tracks the inductively simpler rule (n = 4 families)

family	shortcut $R^\dagger$	simpler	lock-in /10	post-flip /10
depth-shortcut	shallow-root	$R^\dagger$	10/10	—
unknown-as-false	unknown → false	$R^\dagger$	10/10	—
always-abstain	unknown → abstain	$R^\dagger$	10/10*	—
depth-threshold	truncation	R^*	0/10	10/10 the flip

* 10/10 behavioral, 8/10 strict

Figure 6: Directional regularity over four families: the model commits to the inductively simpler rule; the depth-threshold counter-case does not lock in (0/10) and flips to 10/10 under truncation truth.

The pattern is consistent across all four families (Figure 6). In three families the simpler rule is the shortcut $R^\dagger$, and the model commits to it: the depth-shortcut and unknown-as-false families in 10/10 seeds, the always-abstain family in 10/10 behaviorally and 8/10 by the strict signature. In the fourth, the depth-threshold family, the simpler rule is the intended rule R^* (full strong-Kleene), and the model commits to that: it does not lock in (0/10 seeds). The direction follows inductive preference, not the rule we declared correct. We ground "inductively simpler" in five model-independent complexity proxies, computed in the same way for R^* and each $R^\dagger$, with every rule re-implemented bit-identically to the released label generator (Appendix C, Table 4). The depth-threshold inversion that the flip depends on holds under *every* description-length and memory proxy: there the committed rule is R^*, with 3 vs. 13 register states and 24 vs. 97 truth-table rows against truncation, whose depth counter is exactly the excess. In every family the committed rule is simpler than the alternative under at least one fixed measure and never more complex under all of them, and the axis that carries the simplicity matches the inductive bias each family is built around. The proxies do not give one monotone ordering (depth-shortcut: simpler only on discharge depth; depth-threshold: the reverse), so which axis dominates is not predicted by the proxies alone. We report a four-family empirical pattern and leave the architectural account to future work.

The flip makes the symmetry concrete. We take the depth-threshold checkpoints, weights that never lock in when scored against strong-Kleene, and re-score them against truncation as the declared truth. The training data is unchanged and stays ambiguous (0 violations in 100,000 samples), so the same weights can be scored against either truth. Against truncation-truth the same weights show a textbook lock-in signature in 10/10 seeds: in-distribution accuracy 0.977, near-zero accuracy and near-certain commitment on the disagreement region. One set of weights reads as no-lock-in from one declared truth and as lock-in from the other, because the actual commitment of the model, to the simpler rule, never moved. The model is also able to represent truncation: given disambiguating supervision it learns it to a median CE of 1.000, so the commitment is a preference, not an incapacity.

The placement thesis survives the flip as well: greedy region coverage beats random selection on the truncation-scored held-out OOD in 10/10 seeds at k = 2. We use the flip for this second-direction confirmation and for this relative placement transfer; its absolute repair is not counted as efficacy, because truncation is hard to teach few-shot, and a few placed counterexamples may not move a model toward a rule more complex than the one it committed to (Section 7).

7 Related work, limitations, and conclusion

Related work. Our phenomenon is a *reasoning shortcut*: high training accuracy reached through unintended features, as in shortcut learning (Geirhos et al., 2020), Clever-Hans predictors (Lapuschkin et al., 2019), and underspecified models where many rules fit equally well (D'Amour et al., 2022), surveyed by Steinmann et al. (2024). Marconato et al. (2023b) formalize reasoning shortcuts as

unintended optima of the learning objective; wrong-rule lock-in is the same failure, and we ask the question they leave open. They mitigate shortcuts with concept-level supervision (Marconato et al., 2023a); related repairs use counterexample-guided augmentation (Dreossi et al., 2018); and train-time robustness methods (IRM (Arjovsky et al., 2019), group DRO (Sagawa et al., 2020), JTT (Liu et al., 2021), and data-centric variants that curate the training set (Fazla & Hashemi, 2026)) prevent adoption when group structure is given. We address the post-hoc question these do not: a model *already* committed, repaired through placement alone, and the geometry of that repair. Post-hoc repair itself has precedent: last-layer re-training on a small group-balanced set (Kirichenko et al., 2023) and model patching (Goel et al., 2021) both update a committed model. Our repair retrains the whole network with the placed counterexamples added; whether the coverage requirement also holds for readout-only repair is open.

Why the model locks into one rule and not the other is simplicity bias: networks prefer the simpler of several consistent hypotheses even when a more complex one predicts equally well (Shah et al., 2020). Our directional regularity makes this prior precise for rule commitment and shows it symmetrically: the locked-in rule is the inductively simpler one across four families, and re-scoring one family's weights under a flipped ground truth shows that the direction follows simplicity, not the declared-correct rule, in both directions.

That uncertainty cannot pick out the repairing counterexamples has precedent, but the precedent had no mechanism. Standard acquisition strategies (Settles, 2009; Sener & Savarese, 2018) are unreliable in reality checks of deep active learning and often no better than random on real benchmarks (Gashi et al., 2024; Ren et al., 2023). Our result also reaches the disagreement-based branch, query-by-committee (Seung et al., 1992; Beluch et al., 2018), BALD (Houlsby et al., 2011; Gal et al., 2017), and influence-function selection (Koh & Liang, 2017), which Section 5 shows to be blind in our setting for a structural reason: an ensemble of locked seeds is unanimous exactly on the disagreement region. The theory branch of active learning is disagreement-based in a different sense: CAL-style selective sampling and its agnostic descendants (Cohn et al., 1994; Dasgupta, 2005; Balcan et al., 2006; Hanneke, 2014) query from the disagreement region of a maintained version space to bound the label complexity of learning from scratch. Our question is instead which sub-regions must be covered, at a fixed budget, to repair a model that is already committed; what is zero here is the empirical disagreement of a committee of gradient-trained networks, not the version-space disagreement itself, which stays positive because both rules fit the training data. Value-based ranking (Data Shapley (Ghorbani & Zou, 2019), TracIn (Pruthi et al., 2020)) is the adjacent family: Section 5 tests its cheapest proxies (min-margin and a structural feature) and both fail. Full influence-based selection in this setting remains open, as do hybrid acquisition in gradient space (Ash et al., 2020) and training that diversifies committee members on purpose (Lee et al., 2023); our heterogeneous-committee test (Appendix A) is the closest probe and did not recover the signal. What is new is not that uncertainty is unreliable, but that confidence, value, diversity, and ensemble disagreement fail together, for a stated reason: the locked model is most confident exactly where it is systematically wrong.

Limitations. Four boundaries limit these claims. First, our core evidence is synthetic by design (Section 2). A CelebA demonstration (Appendix B) transfers the selection-level placement signal: greedy beats budget-matched random with a paired CI excluding zero on ten seeds (pre-registered extension; the five new seeds separate alone), and diversity again does not. There, a budget-matched top-loss selector matches greedy, because the one-sided region of CelebA makes loss a region proxy; which signal finds the region depends on the setting, the coverage requirement does not. The demonstration draws from a candidate pool enriched to 50% failure-region items, and a Waterbirds attempt did not pass the pre-registered lock-in gate under three recipes (Appendix B). External validity beyond this benchmark, including natural-language shortcut benchmarks (McCoy et al., 2019) and the wider WILDS suite (Koh et al., 2021), is future work. Second, the directional regularity rests on four constructed families (Section 6). Third, the repair mechanism is characterized cleanly only for the depth-shortcut family; the structure of the unknown-rule families is diffuse and we do not localize it (Section 4). Fourth, the flip gives second-direction confirmation and relative placement transfer only; its absolute repair is left out of the efficacy claims (Section 6).

Conclusion. Repairing a model that has locked onto the wrong rule is governed by placement, not quantity: a few counterexamples repair lock-in when they cover the disagreement region, a requirement shown to be causally necessary (by direct sampling, where the region structure is discrete) at every budget we test. Two further results sharpen established priors: the locked-in rule is the

inductively simpler one, made symmetric and measurable for rule commitment; and on a symmetric disagreement region the repair signal is invisible to confidence, value ranking, and committee disagreement, so diagnosis must be interactive, while on a one-sided natural region the model's loss can stand in for the region but the coverage requirement remains. Which counterexamples to add is a question that only the behavior of the model can answer.

REPRODUCIBILITY STATEMENT

Every headline empirical claim in this paper was registered before it was tested, with one exception: the member-level committee analysis of Section 5 was computed post hoc from released per-seed predictions, and its heterogeneous-committee follow-up (Appendix A) was then pre-registered before inference. The hypotheses, gates, and decision tree of each experiment were frozen in a pre-registration file and SHA-256-hashed before any training began: the main repair experiment (hash 1EBD489A...), the mechanism strengthening round (C5835CE8...), the directional-flip campaign (8C60AFCD...), and the third-family replication (6E476111...). The hash is checked again at analysis time, and a frozen gate is never relaxed after the fact. All repair and mechanism results are reported over ten seeds with per-seed values; the released files contain one row per training run, more than 1800 runs for the logic testbed and 49 for CelebA (Appendix D). Selection optimizes a counterexample-fit objective only; generalization is measured on a held-out, exact-depth OOD set that is disjoint by canonical id from everything used to select examples. Every number in the paper is traced to a row in a number list tied to its source report. Runs use canonical-id caching of every (family, seed, arm) cell, so a re-run gives identical results and resumes after an interruption. We release the code, the configurations, the pre-registration files with their full hashes, and the per-seed outputs; the supplementary `reproduce.py` recomputes the headline numbers from the released files.

ETHICS STATEMENT

The controlled part of this work uses synthetic data (propositional-logic expression trees generated under a three-valued semantics), with no human subjects and no personal data. The natural-data demonstration (Appendix B) uses the public CelebA face-attribute dataset (Liu et al., 2015) under its non-commercial research terms. We collect no new data, we identify no individuals, and we use the standard blond-hair-versus-male split only as an established spurious-correlation benchmark (Sagawa et al., 2020), not to draw conclusions about demographic groups. We see no direct ethical risk in the study. Its aim, to understand why models commit to unintended rules and how a committed model is repaired, is meant to make learned systems easier to inspect and to correct.

AI USE STATEMENT

In this work, we used generative AI tools for two tasks: drafting and editing the manuscript, and writing code (analysis, figure-generation, and experiment-orchestration scripts). We did not use generative AI tools to formulate the research questions, the hypotheses, the experimental designs, the theoretical or conceptual framing, or any claim in this paper. The synthetic datasets are produced by the released label-generator code from the rule definitions in Section 2, not by an AI tool; the other required-disclosure categories do not apply to this work. All scientific decisions are the authors'. Every confirmatory experiment was pre-registered in a human-reviewed document (SHA-256-hashed before the run), and no criterion was relaxed afterwards. No reported number comes from the text output of an AI tool: every statistic is computed by released, human-reviewed code from the released per-seed result files, and the supplementary artifact recomputes the headline numbers from those files with one command. AI-assisted code was verified by running it against the pre-registered readouts, and the citations were checked against their published metadata. We take responsibility for the final content of this work, including text and artifacts produced with the aid of generative AI.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019. URL `https://arxiv.org/abs/1907.02893`.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1906.03671.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 65–72, 2006.

William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9368–9377, 2018. doi: 10.1109/CVPR.2018.00976.

David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://jmlr.org/papers/v23/20-1335.html.

Sanjoy Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Kurt Keutzer, Alberto Sangiovanni-Vincentelli, and Sanjit A. Seshia. Counterexample-guided data augmentation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 2071–2078, 2018. doi: 10.24963/ijcai.2018/286. URL https://www.ijcai.org/proceedings/2018/286.

Arda Fazla and Abolfazl Hashemi. Mitigating spurious correlations with memorization-guided dataset de-biasing. *arXiv preprint arXiv:2606.02830*, 2026.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1183–1192, 2017. URL https://proceedings.mlr.press/v70/gal17a.html.

Edrina Gashi, Jiankang Deng, and Ismail Elezi. Deep active learning: A reality check. arXiv preprint arXiv:2403.14800, 2024. URL https://arxiv.org/abs/2403.14800.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2008.06775.

Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2–3):131–309, 2014.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745, 2011. URL https://arxiv.org/abs/1112.5745.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Proceedings of the First Conference on Causal Learning and Reasoning (CLeaR)*, 2022. URL https://arxiv.org/abs/2110.14503.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2204.02937.

Stephen Cole Kleene. *Introduction to Metamathematics*. North-Holland, Amsterdam, 1952.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1885–1894, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. URL https://arxiv.org/abs/2012.07421.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019. doi: 10.1038/s41467-019-08987-4. URL https://www.nature.com/articles/s41467-019-08987-4.

Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Learning from underspecified data. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2202.03418.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. URL https://arxiv.org/abs/2107.09044.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

Emanuele Marconato, Stefano Teso, and Andrea Passerini. Neuro-symbolic reasoning shortcuts: Mitigation strategies and their limitations. arXiv preprint arXiv:2303.12578, 2023a. URL https://arxiv.org/abs/2303.12578.

Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023b. URL https://arxiv.org/abs/2305.19951.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. URL https://arxiv.org/abs/1902.01007.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Simiao Ren, Saad Lahrichi, Yang Deng, Willie J. Padilla, Leslie Collins, and Jordan Malof. Does deep active learning work in the wild?, 2023. URL https://arxiv.org/abs/2302.00098.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1911.08731.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009. URL https://research.cs.wisc.edu/techreports/2009/TR1648.pdf.

H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT)*, pp. 287–294, 1992. doi: 10.1145/130385.130417.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL https://arxiv.org/abs/2006.07710.

David Steinmann, Felix Divo, Maurice Kraus, Antonia Wüst, Lukas Struppek, Felix Friedrich, and Kristian Kersting. Navigating shortcuts, spurious correlations, and confounders: From origins via detection to mitigation. *arXiv preprint arXiv:2412.05152*, 2024.

Ruibin Xiong, Yimeng Chen, Liang Pang, Xueqi Cheng, Zhiming Ma, and Yanyan Lan. Uncertainty calibration for ensemble-based debiasing methods. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021. URL https://arxiv.org/abs/2111.04104.

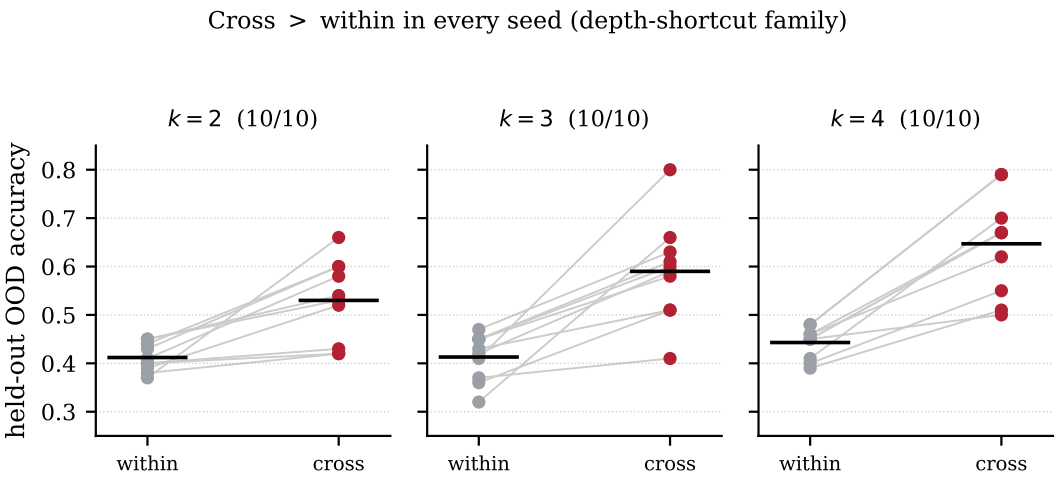

Figure 7: Region coverage is necessary in every seed (depth-shortcut family). At each k, the cross-region pair exceeds the within-region pair on held-out OOD in all 10 seeds.

A Additional per-seed results and the full negatives table

Selection protocol (pool, objective, firewall). All selection arms draw from the *same* candidate pool per (family, seed): 20 representatives clustered by *structural* features only (tree depth, U-leaf positions, operator/leaf counts, generator mode), never by the embeddings or predictions of the locked model, so the arms are matched in budget and in pool by construction. The greedy arm is forward selection that maximizes accuracy on a held-out *counterexample-fit* test (disagreement-region items disjoint from the pool). It has *no* access to the $R^*/R^\dagger$ region partition (region coverage is what the selection finds, not an input), and it never touches the exact-depth OOD sets, which are disjoint by canonical id from everything used in selection and are scored only after the selection is frozen. The partition is given only to the direct-sampling arms of Section 4, where it is the imposed treatment variable that turns the correlational comparison into a causal one. Overfitting the counterexample-fit objective cannot produce the reported effect: every headline lift is scored only on the firewalled OOD sets, so a selection that only overfits its fit set would fail to transfer and would *reduce* the measured advantage.

GRU replication of the placement effect. A pre-registered (SHA-frozen before the run) architecture-robustness check reruns the arm contrast of Section 3 on a GRU (h64) in the depth-shortcut family, with the same entry point, candidate pool, counterexample-fit greedy objective, and OOD firewall, on 10 seeds. Behavioral lock-in holds in 9/10 seeds at baseline. The committed contrast, greedy vs. budget-matched random at $k{=}2$ on held-out OOD, replicates: $+0.105$, paired-bootstrap 95% CI $[+0.060, +0.157]$, greedy above random in $10/10$ seeds (descriptively, greedy also beats diversity in $10/10$, $+0.167$ $[+0.097, +0.244]$; post-repair IID never drops below 0.989). Committed readout: **REPLICATES**. Placement is therefore a property of the lock-in phenomenon and not of the transformer.

Heterogeneous committees. A natural question is whether committees of *different architectures* recover a disagreement signal on the rule-disagreement region. We built a 20-member heterogeneous committee from existing ambiguous-only (dose-0) checkpoints of the depth-shortcut family: 10 `transformer_t64_l2` seeds (the e3 baselines of the paper) and 10 `gru_h64_legacy` seeds (the GRU-replication baselines; byte-identical training data per seed), plus a 25-member extension that adds 5 `gru_h256` Phase-2A models. All members vote (argmax) on the depth-controlled diagnostic pool of §5 (8,000 trees padded to depth 6; A_f is the per-item agreement indicator between R^* and $R^\dagger$: 5,233 disagreement-region items $A_f{=}0$, 2,767 agreement-region items $A_f{=}1$; design and thresholds pre-registered before inference). The committee does *not* collapse to unanimity: mean normalized hard-vote entropy is 0.506 on $A_f{=}0$, and only 7.2% of disagreement-region items are unanimous. But the signal is not restored either. Entropy is high everywhere off-distribution (0.359

Table 1: Pre-registered negatives (full): each hypothesis registered before running, then evaluated at its frozen bar (four falsified; N4 yields only a weak, non-deployable signal).

#	Hypothesis	Frozen bar	Result	Reported
N1	The greedy advantage concentrates at k = 2 (dissociation D > 0)	D > 0 in $\geq$ 7/10	4/10 — falsified	Section 4
N2	A test-blind selector recovers the greedy−random gap	$\geq$ 60% gap recovery and beats random in $\geq$ 8/10	cross-region 0.552 (55% recovery), 7/10; top-num_u 17%; min-margin −54% — all falsified	Section 5
N3	A single disagreement-cell localizes repair (Option B)	own-cell CE $\geq$ 0.80; cross−within cell $\geq$ +0.20	0.276 / 0.253; −0.027 / +0.067 — falsified	Section 4
N4	Predictive entropy separates repairing counterexamples	AUC clearly above the 0.5 no-information line	0.643 / 0.586 — weak signal, far below the N2 deployable bar	Section 5
N5	depth_threshold is a fourth lock-in family	the strict lock-in signature	0/10 — does not lock in (the counter-case)	Section 6

on the depth-matched agreement region vs. 0.058 on shallow IID), so vote entropy ranks $A_f{=}0$ above $A_f{=}1$ with an AUC of only 0.630 $[0.616, 0.644]$ (25-member: 0.626), which is no more useful than the single-model predictive-entropy AUC of 0.643 and far below deployability. The structure that survives is the damaging one: *every* unanimous disagreement-region item is unanimous on $R^\dagger$ (100%; likewise $274/274$ unanimous items on the counterexample split), and the architecture-level majorities co-commit item by item: the transformer and GRU majorities agree on 69.0% ($3{,}609/5{,}233$) of the $A_f{=}0$ items and, whenever they agree, they agree on $R^\dagger$ ($3{,}609/3{,}609$); all three architecture majorities equal $R^\dagger$ on 64.5%. Member-level analysis shows why: transformers commit to $R^\dagger$ almost deterministically (97.2% of votes; within-architecture pairwise agreement 0.964), while GRUs commit noisily (61.3%; within-architecture agreement only 0.540), so heterogeneous vote entropy is mostly the mutual off-distribution inconsistency of the GRUs, concentrated on U-heavy constructions from *other* families' counterexamples and not on the disagreement items of this family. A depth-confounded design would mislead here: against shallow IID items the entropy of the same committee reaches an AUC of 0.920, an artifact of depth and not of rule disagreement, which disappears (0.630) on the depth-matched control. Heterogeneity therefore refutes our conjectured unanimity mechanism as stated, but it strengthens the operative claim: even cross-architecture disagreement is test-blind and does not place counterexamples.

Extended budget sweep (k to 8). Run after the main campaign with a three-branch readout fixed in advance (keeps-rising / plateaus / declines; the slope-attenuation bar was frozen before the data were read), on the same 10 seeds, arms, and held-out OOD firewall as Section 4.

B Preliminary natural-benchmark demonstration (CelebA)

This appendix reports a pre-registered transfer demonstration of the selection-level placement thesis on a natural spurious-feature benchmark. *Provenance.* A Waterbirds attempt failed its frozen lock-in gate in three recipe probes (default; JTT-style; literature-exact batch size), which triggered the retarget branch fixed in advance. The CelebA design below was then frozen and SHA256-fixed *before* any CelebA training, with a four-branch readout (placement-transfers / partial / null / excluded), and no criterion was relaxed afterwards. A pre-registered extension (SHA256-frozen before any stage-2 training) then added seeds 5–9 under the identical recipe on rented cloud hardware (the within-seed

Table 2: Direct-sampling contrast over extended budgets (depth-shortcut family, 10 seeds). The cross-region *level* saturates by $k = 8$; the within-region arm never moves, so the advantage widens. Committed readout: PLATEAUS-AT-$k5$ (late $k4 \to k8$ slope $+0.020$ per unit k, CI $[+0.012, +0.027]$, $0.35\times$ the early rate, below the frozen keeps-rising bar of 0.5); the necessity precondition (cross $>$ within in $\geq 8/10$ seeds at every k) passes throughout.

k	cross-region OOD	within-region OOD	cross$>$within (seeds)	paired-boot 95% CI
2	0.531	0.412	10/10	$[+0.072, +0.170]$
3	0.590	0.414	10/10	$[+0.119, +0.243]$
4	0.646	0.442	10/10	$[+0.153, +0.255]$
5	0.685	0.443	9/10	$[+0.171, +0.306]$
6	0.729	0.450	10/10	$[+0.229, +0.322]$
8	0.727	0.445	10/10	$[+0.216, +0.347]$

pairing is not affected) and a test-blind top-loss arm. The pooled estimate therefore comes from two stages, and both cohorts are reported separately.

Design. CelebA blond×gender (WILDS): an ImageNet-pretrained ResNet-50 is ERM-fine-tuned and locks onto the gender shortcut in 10/10 seeds (test average 0.949–0.954; blond-male failure-region accuracy 0.272–0.600, gate ≤ 0.60). The analogue of the disagreement region is *one-sided* (blond-male), so the cross-region causal contrast of Section 4 is degenerate by structure and is *not* claimed; a region-oracle arm (a uniform draw of all B counterexamples from the failure region) serves as the unweighted within-region reference. Five budget-matched arms ($B = 8$ from a 40-candidate pool: 20 failure-region, 20 from the other groups; the pool is thus enriched to 50% failure-region items, far above the natural base rate of blond males in CelebA, which makes acquisition easier than on a raw data stream): random, diversity (the exact Section 3 selector on penultimate features), greedy (forward selection on a held-out failure-region selection split, OOD-blind), region-oracle, and (stage-2) maxloss, the top-B pool items by the cross-entropy loss of the locked model, which uses only the locked model and the pool's own labels (no region labels, no selection split, no OOD access). Repair is identical across arms (fine-tuning with a 32-image replay guardrail; no arm drops test average accuracy by more than 0.01). The readout metric is held-out *test*-split failure-region accuracy, firewalled from selection.

Table 3: CelebA repair by selection arm: held-out failure-region (blond-male) accuracy per seed. Seeds 0–4 are the frozen stage-1 record; seeds 5–9 are the pre-registered extension (identical recipe, cloud hardware); maxloss is the stage-2 test-blind top-loss arm (seeds 1–9; the seed-0 checkpoint was not retained; every maxloss contrast is computed over the nine matched seeds). Pooled greedy $-$ random separates with a CI excluding zero and the five extension seeds separate on their own; diversity again does not beat random. Committed readouts: **G1 SEPARATES / G2 HOLDS / G3 LOSS-SUFFICES**.

arm	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9	mean
random	0.350	0.600	0.522	0.511	0.389	0.494	0.389	0.600	0.378	0.439	0.467
diversity	0.311	0.633	0.494	0.528	0.300	0.406	0.406	0.511	0.406	0.500	0.449
greedy	0.311	0.706	0.606	0.633	0.394	0.494	0.500	0.622	0.572	0.539	0.538
region-oracle	0.417	0.617	0.511	0.556	0.394	0.422	0.444	0.656	0.456	0.489	0.496
maxloss	—	0.700	0.622	0.656	0.428	0.550	0.428	0.628	0.533	0.494	0.560

greedy $-$ random: pooled $+0.071$ $[+0.029, +0.112]$; extension seeds 5–9 alone $+0.086$ $[+0.027, +0.143]$; beats diversity in 9/10 seeds
region-oracle $-$ random: $+0.029$ $[+0.000, +0.053]$ · maxloss $-$ random: $+0.080$ $[+0.052, +0.110]$; $-$ greedy: -0.003 $[-0.029, +0.022]$; $-$ oracle: $+0.055$ $[+0.019, +0.090]$

Reading. At a budget of eight examples on a 25M-parameter model, interactive greedy selection separates from budget-matched random with a paired CI excluding zero, on the pooled ten seeds and on the five pre-registered extension seeds alone, and the diversity baseline again fails, so the selection-level signature of the synthetic results transfers. The unweighted within-region reference separates only weakly ($+0.029$, lower bound at zero), so instance selection inside the region carries most of the lift. The stage-2 top-loss arm matches greedy (-0.003 $[-0.029, +0.022]$) and beats random: on this one-sided region the loss of the locked model is itself a region proxy (6–8 of its 8 picks are failure-region items), so budget-matched top-loss selection gives the same coverage. This is a boundary on selector invisibility for natural data: which signal finds the region depends on the setting (the test-blind selectors of the synthetic families still fail, Section 5), the coverage requirement does not. The transfer is at the *selection level* only; it is not a naturalization of the

causal geometry, because the one-sided region does not allow the cross-region intervention. The stage-1 PARTIAL record and the two-stage extension are both released.

C RULE COMPLEXITY: OPERATIONALIZING "INDUCTIVELY SIMPLER"

We quantify the "inductively simpler" claim of Section 6 with five model-independent complexity proxies, computed in the same way for the intended rule R^* (full strong-Kleene) and for each family's committed shortcut $R^\dagger$ over three-valued expression trees: symbolic-program length (AST node count of the smallest implementation), truth-table specification size (distinct input→output rows), register-transducer state count, number of unknown-discharge special cases, and discharge-recursion depth. Every rule is re-implemented from its symbolic definition and verified to be *bit-identical* to the released label generator on $40{,}000$ random expression trees (0 mismatches); the recomputation script and its CSV output are released (`reproducibility/analysis/rule_complexity_proxy.py`).

Table 4: Static complexity of the intended rule R^* (full strong-Kleene) versus each family's committed shortcut $R^\dagger$, under five model-independent proxies. $\Delta = c(\text{committed}) - c(\text{other})$; $\Delta < 0$ (bold) means that the *committed* rule is the simpler one. In families 1–3 the committed rule is $R^\dagger$; in the depth-threshold family it is R^* (the inversion). Lower is simpler; the discharge-depth axis is unbounded for strong-Kleene, so $-\infty$ means that the committed rule has finite (or no) discharge depth against an unbounded alternative, and $+\infty$ the reverse.

Family	Proxy	$c(R^*)$	$c(R^\dagger)$	Δ	committed
depth-shortcut	program len.	198	371	$+173$	$R^\dagger$
	truth-table	24	48	$+24$	
	reg. states	3	6	$+3$	
	discharge cells	4	4	0	
	discharge depth	unbounded	root	$-\infty$	
always-abstain	program len.	198	198	0	$R^\dagger$
	truth-table	24	24	0	
	reg. states	3	3	0	
	discharge cells	4	0	$\mathbf{-4}$	
	discharge depth	unbounded	none	$-\infty$	
unknown-as-false	program len.	198	167	$\mathbf{-31}$	$R^\dagger$
	truth-table	24	13	$\mathbf{-11}$	
	reg. states	3	2	$\mathbf{-1}$	
	discharge cells	4	0	$\mathbf{-4}$	
	discharge depth	unbounded	none	$-\infty$	
depth-threshold	program len.	198	224	$\mathbf{-26}$	R^*
	truth-table	24	97	$\mathbf{-73}$	
	reg. states	3	13	$\mathbf{-10}$	
	discharge cells	4	16	$\mathbf{-12}$	
	discharge depth	unbounded	3	$+\infty$	

The depth-threshold inversion, on which the flip experiment depends, is confirmed under every description-length and memory proxy: the committed rule there is R^* itself, strictly simpler than truncation on every axis (3 vs. 13 register states, 24 vs. 97 truth-table rows, 198 vs. 224 program nodes); the excess is exactly the depth counter that truncation must maintain. The unknown-as-false shortcut is also strictly simpler than R^* under all five measures. For the remaining two families the simplicity lies on the unknown-handling axis and not in gross program size: always-abstain is description-length-identical to strong-Kleene (its sticky-unknown operators are the formal dual of the strong-Kleene ones) but removes every discharge special case (0 vs. 4), and shallow-root applies the discharge rule at one level only (discharge depth 1 vs. unbounded), which is the depth shortcut that its name denotes. That the description-length-*tied* family (always-abstain) is also the weakest-locking one (strict signature in $8/10$ seeds, vs. $10/10$ elsewhere) is consistent with this account. So in every family the committed rule is strictly simpler than the alternative under at least one fixed, model-independent measure, and in no family is it more complex under all of them: a directional regularity, not an artifact of one metric.

D TESTBED SPECIFICATION

Grammar and serialization. Expressions are trees over the binary connectives AND and OR, the unary connective NOT, and the leaves T, F, U. A tree is serialized in prefix (Polish) notation as a token sequence over the vocabulary {T, F, U, NOT, AND, OR} plus a padding token; no parentheses are needed in prefix form. Training trees have depth ≤ 4; the OOD sets are generated at exact depths 8 and 12 (2000 trees per depth); the diagnostic pool of Section 5 is generated at depth 12.

Data. Per (family, seed): 10000 ambiguous training trees (labels identical under R^* and $R^\dagger$), a validation split drawn from the same ambiguous distribution, 2000 IID test trees, and a 4000-item counterexample test set drawn from the disagreement region (this is the counterexample-fit set that greedy selection scores on; it is disjoint by canonical id from the candidate pool and from both OOD sets). All splits are deduplicated by canonical tree id.

Model. `transformer_t64_l2`: token embedding of width 64 with learned positional embeddings (maximum length 512), two post-norm Transformer encoder layers with 4 heads, feed-forward width 128, ReLU, no dropout; the encoder output is mean-pooled over the non-padding positions and mapped by a linear layer to the three classes. The GRU variants (`gru_h64`, `gru_h256`) replace the encoder by a GRU of hidden width 64 or 256 over the same token sequence.

Optimization. AdamW, learning rate 10^{-3}, weight decay 0, batch size 128, cross-entropy loss, at most 40 epochs with early stopping on validation loss (patience 10) and restoration of the best-validation checkpoint. The locked baseline is trained on the ambiguous set alone. *Repair* retrains a fresh model from the same seed on the ambiguous set plus the k selected counterexamples under the identical recipe (no reweighting, no balanced batching); the broad 3200-example dose of Section 3 is trained in the same way. The CelebA demonstration (Appendix B) instead fine-tunes the locked ResNet-50.

Inventory of runs. Each row of the released per-seed files is one training run. The logic-testbed campaigns are: the main repair experiment (540 runs: two families, 10 seeds, greedy / random / diversity / direct-sampling arms, budgets $k = 1..3$ plus the cross- and within-region pairs), the third-family replication (220 runs), the budget-curve completion (200 runs, $k = 4, 5$), the direct-sampling sweep to $k = 8$ (300 runs, five repeats per cell), the extended greedy / random / diversity sweep (270 runs), the GRU replication (90 runs), the directional-flip campaign (190 runs), and the lock-in baselines (10 runs per family). Together this is more than 1800 trained models on the testbed. The CelebA demonstration adds 49 repair fine-tunes of a ResNet-50 over five arms and ten seeds, and the heterogeneous-committee analysis scores 25 checkpoints on 8000 diagnostic trees (200000 votes). Nine pre-registration documents (SHA256-frozen before the corresponding runs), 18 analysis scripts, and `reproduce.py` are released with the data.

Lock-in gate. A seed counts as strictly locked in when its IID accuracy is ≥ 0.90, its counterexample-test accuracy is ≤ 0.80 or its mean OOD accuracy over depths 8 and 12 is ≤ 0.80, and its predictions are closer to $R^\dagger$ than to R^*; behavioral lock-in requires only the last two conditions.