ARE WE RECOVERING THE MECHANISM? OBJECTIVE-LEVEL RECOVERY GAPS IN MECHANISTIC INTERPRETABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic interpretability aims to recover the internal computations responsible for model behavior. Progress in automated circuit discovery is often framed as a search problem: better attribution or optimization should identify better mechanisms. This assumes that the evaluation objective can recognize a better circuit once it is found. We show that intervention-defined faithfulness can instead prefer an equally sized circuit that reproduces the model's behavior less well, creating an *objective-level recovery gap*. Across four human-reference tasks and InterpBench, we compare validation faithfulness with behavior on held-out prompts under fixed ordinary resampling. The behavioral criterion is agreement with the intact model, including its mistakes, except on Greater-Than, where we use semantic accuracy. Controlled reference edits reveal misranking without any discovery algorithm, and outputs of EAP, EAP-IG, ACDC, and Edge-SP exhibit the same failure. Under resampling, KL misranks 9.4%–41.2% of candidate pairs across these methods on the human-reference tasks. We investigate *context distortion* as an explanation: replacing excluded signals changes the inputs on which retained components operate. Restoring selected signals from the recipient's intact-model execution repairs 96 of 100 persistent KL misrankings from the discovery pool on both validation and held-out prompts. The circuits and their original behavioral scores remain unchanged. These findings show why better discovery alone is insufficient when its objective rewards the wrong candidate.

1 INTRODUCTION

Mechanistic interpretability seeks to identify the internal computations responsible for model behavior. Circuit analysis approaches this goal by representing a computation as a sparse subgraph of interacting model components (Olah et al., 2020; Elhage et al., 2021). Detailed studies have reconstructed circuits for behaviors such as indirect object identification and numerical comparison (Wang et al., 2023; Hanna et al., 2023), while automated methods aim to find such circuits at scale (Conmy et al., 2023; Syed et al., 2023; Hanna et al., 2024). Much of this effort follows a natural view of the problem: better attribution, more data, or improved search should recover a better account of the computation used by the intact model. But successful recovery also depends on recognizing a better circuit once it has been found.

The usual test is intervention-defined *faithfulness*: how well a circuit preserves model outputs when computation outside it is ablated or replaced (Conmy et al., 2023; Hanna et al., 2024). This test is central to both discovery and evaluation. It can determine which connections are retained during search and which candidate is ultimately presented as an explanation. A high score is therefore more than a description of a circuit; it is a reason to choose that circuit over another. This raises the question we study:

Can faithfulness prefer an equally sized circuit that reproduces the model's behavior less well?

Executing a circuit requires replacing excluded signals with donor activations, means, or zeros (Miller et al., 2024; Heimersheim & Nanda, 2024). Retained components then receive inputs that differ from those in the intact model and recompute their activations from these altered inputs. We call this

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.

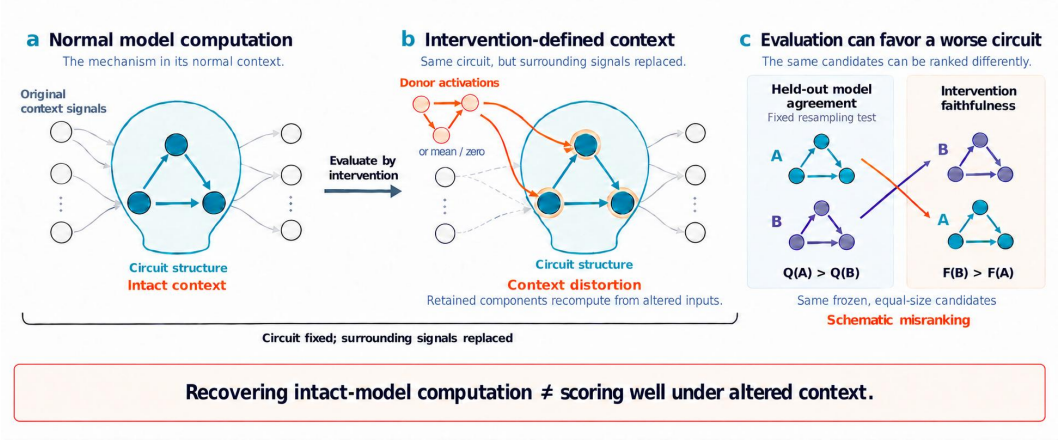


Figure 1: **The objective-level recovery gap.** Replacing excluded signals with donor activations, means, or zeros changes a circuit's computational context without changing its structure. This context distortion can make faithfulness prefer B even when an equally sized A better reproduces intact-model behavior under a fixed behavioral test.

change in computational context *context distortion*. Different circuits can respond differently to the replacement signals, allowing the score to favor a circuit that works well under the intervention over one that better reproduces the model under a fixed behavioral test, as illustrated in Figure 1.

This is a problem for the objective as well as for discovery. Search may fail to find a high-scoring candidate. But if the score itself prefers a behaviorally worse candidate, searching more effectively does not resolve that preference. We refer to this misalignment as an *objective-level recovery gap*. We test it through pairwise *misranking*: a validation score favors one circuit while independent-test behavior favors the other. Comparing fixed candidates lets us examine what the objective rewards separately from how those candidates were obtained. All primary comparisons use circuits of the same size, ruling out differences in the number of retained components as an explanation for misranking. Even under this control, faithfulness can prefer a circuit that reproduces model behavior less well. This failure persists among outputs of established discovery methods on human-studied language-model tasks, making it relevant to practical circuit selection.

Using a fixed behavioral evaluation across interventions, we compare references with candidate circuits and candidates with one another. Preferring a worse candidate over a stronger reference is a particularly stringent failure. Yet discovery usually involves choosing among imperfect candidates, so even a low reference–candidate misranking rate cannot establish that faithfulness reliably guides this choice.

We study this problem in two complementary settings: InterpBench, which provides known mechanisms in smaller, semi-synthetic models (Gupta et al., 2024), and the Human suite, which contains human-studied circuits in pretrained language models, including IOI (Wang et al., 2023) and Docstring (Heimersheim & Janiak, 2023). The latter more closely reflects common applications of circuit discovery. Its higher aggregate candidate–candidate misranking in our main discovery evaluation (Section 4) highlights the importance of assessing faithfulness beyond semi-synthetic benchmarks.

We begin with controlled candidates, then examine discovery outputs and an intervention on the evaluation context. Our main findings are:

- **Faithfulness can misrank circuits without a discovery algorithm.** Equal-size modifications of reference circuits produce misrankings under resampling, mean, and zero replacement. The failures occur both against references and between candidates, and some score preferences select circuits with substantially worse held-out behavior. Replacing KL with logit or probability differences changes the failures but does not eliminate them (Section 3).
- **The same failure occurs in actual discovery outputs.** Candidate pools from EAP, EAP-IG, ACDC, and Edge-SP all contain misranked pairs. Under resampling alone, candidate–

candidate KL misranking ranges from 9.4% to 41.2% across the four methods on the human-reference tasks. These comparisons reuse the validation and test prompts from the controlled study (Section 4).

- **Restoring part of the intact context can repair the preference.** For 100 persistent KL misrankings selected from the discovery pool, we restore selected recipient-intact signals while keeping the circuits, original behavioral scores, and donor assignments elsewhere fixed. At least one partial restoration corrects both validation and independent-test rankings in 96 cases. This supports context distortion as a contributor to the observed failures (Section 5).

2 Related work and background

2.1 Related work

Circuit discovery and recovery benchmarks. ACDC uses activation patching to prune connections (Conmy et al., 2023); EAP approximates edge effects with gradients (Syed et al., 2023), and EAP-IG improves attribution using integrated gradients (Hanna et al., 2024; Sundararajan et al., 2017). Edge-SP learns an edge mask through subnetwork optimization (Conmy et al., 2023; Sanh et al., 2020). Other approaches optimize sparse edge masks at larger scales (Bhaskar et al., 2024) or discover circuits over interpretable features (Marks et al., 2025). These methods address how to find candidates. Known-computation benchmarks address how to assess them: Tracr compiles programs into transformers (Lindner et al., 2023), InterpBench supplies semi-synthetic models with specified circuits (Gupta et al., 2024), and MIB standardizes evaluations of circuit and causal-variable localization (Mueller et al., 2025). We use reference circuits to construct controlled candidates and discovery methods to test practical relevance, while assessing selection through independently measured behavior.

Reliability of circuit evaluation. Faithfulness measurements can change substantially with ablation methodology (Miller et al., 2024), and activation-patching conclusions depend on corruption and scoring choices (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). Adversarial Circuit Evaluation finds inputs on which published circuits diverge from the full model (uit de Bos & Garriga-Alonso, 2024); statistical circuit tests examine preservation, localization, and minimality (Shi et al., 2024). Subspace-patching studies further show that an intervention can change outputs through a dormant pathway, complicating its interpretation as evidence about the original computation (Makelov et al., 2024). We study how this unreliability affects circuit selection: a score can favor an equally sized candidate with worse held-out behavior even when a better candidate is available. Our context-restoration experiment tests whether altered inputs to retained components contribute to this preference. This treats the reliability of the selection objective as a separate requirement for mechanism recovery.

2.2 Background

Circuit execution and interventions. Let M be the intact model and S an executable circuit mask. The execution $M_{S,a}$ retains signals in S and replaces excluded signals according to intervention a. Following standard ablation choices (Miller et al., 2024; Heimersheim & Nanda, 2024), resampling uses a saved donor assignment; mean replacement uses a frozen mean bank; zero replacement inserts zeros. *All primary comparisons are size-controlled:* each pair contains the same number of components in the declared circuit representation, so ranking reversals cannot be attributed simply to differences in circuit size. We remove duplicate masks before forming pairs and compare circuits only within the same task and, for discovery outputs, the same method. Circuit size is measured in the components used by that task's executor.

Validation faithfulness. For validation prompts V, our primary score is negative output-distribution KL:

$$F^{\mathrm{KL}}(S) = -\mathbb{E}_{x \in V}\left[D_{\mathrm{KL}}(p_M(\cdot \mid x) \,\|\, p_S(\cdot \mid x))\right]. \tag{1}$$

Larger scores are better. Scores are averaged over the valid output positions within each prompt and then over prompts; resampling additionally averages three shared donor draws. Logit difference (LD) and probability difference (PD) provide additional task-based margin checks. For task-correct classes

Table 1: The benchmark suites used for controlled and discovery comparisons. The first four rows are human circuits; the final row is the corrected ten-task InterpBench panel.

Task	Behavioral essence	Source
IOI	Indirect-object identification: recover the name that fills the repeated syntactic role.	(Wang et al., 2023)
Greater-Than	Numerical comparison: determine whether one two-digit year ending is strictly greater than another.	(Hanna et al., 2023)
Docstring	Documentation retrieval: predict the token sequence associated with a function's docstring behavior.	(Heimersheim & Janiak, 2023)
Acronym	Acronym completion: map a multiword description to its abbreviated form.	(García-Carrasco et al., 2024)
InterpBench suite	Ten semi-synthetic tasks with native-closure-verified references: 113, 97, 2, 82, 111, 45, 58, 93, 103, and 25.	(Gupta et al., 2024)

P and remaining valid task classes N, at each scored position,

$$F^{\mathrm{LD}} = \frac{1}{|P|}\sum_{j\in P} z_j - \frac{1}{|N|}\sum_{j\in N} z_j, \qquad F^{\mathrm{PD}} = \sum_{j\in P} p_j - \sum_{j\in N} p_j. \tag{2}$$

Task mappings define the targets; candidate predictions do not. Probabilities use the original output space, without renormalization to $P \cup N$. LD and PD test whether changing the faithfulness metric resolves misranking while leaving Q fixed; task-specific definitions are in Appendix A.

Independent behavioral criterion. We denote behavioral quality by Q. For each circuit, we measure agreement with the intact model's answers on held-out prompts T, using *ordinary resampling* and averaging over the same donor draws. For categorical sequence tasks, valid-position agreement is aggregated within each prompt before averaging prompts. Agreement includes intact-model mistakes. Q remains fixed when we change the validation intervention or score.[1] For categorical answer agreement,

$$Q(S) = \mathbb{E}_{x\in T}\left[\frac{1}{|I_x|}\sum_{i\in I_x} \mathbf{1}\{\hat{y}_S(x,i) = \hat{y}_M(x,i)\}\right]. \tag{3}$$

Misranking and the recovery objective. For a fixed pair, let B be the circuit with higher Q and W the circuit with lower Q. A *misranking* occurs when faithfulness prefers W:

$$F(W) > F(B), \qquad \Delta Q = Q(B) - Q(W) > 0, \qquad \Delta F = F(W) - F(B) > 0. \tag{4}$$

R–C compares a reference with an edited or discovered candidate; C–C compares two candidates. Either member of an R–C pair may have higher Q. We determine the better circuit from Q rather than assuming that the reference is always better. Small absolute and relative tolerances exclude numerical ties; Appendix A gives the thresholds, denominators, and paired uncertainty estimates.

The distinction between the two objectives can also be expressed over a family $\mathcal{S}_K$ of size-K circuits:

$$S^\star_{F,a} \in \arg\max_{S\in\mathcal{S}_K} F_a(S), \qquad S^\star_Q \in \arg\max_{S\in\mathcal{S}_K} Q(S). \tag{5}$$

If $Q(S^\star_{F,a}) < Q(S^\star_Q)$, optimizing F_a more effectively cannot recover the best circuit under Q. Our experiments do not compute these global optima. They test the ranking property that motivates this concern: whether the score can prefer a worse member of a fixed pair.

3 Faithfulness Can Misrank Controlled Circuit Candidates

We first remove discovery from the experiment. Starting from a reference circuit, we generate alternatives of the same size and ask how faithfulness ranks them. If a score prefers the worse member of a pair, the failure cannot be attributed to the search procedure: both candidates are already available for evaluation.

[1] Greater-Than uses semantic accuracy: the fraction of predicted two-digit year endings strictly greater than the prompt's starting ending, rather than exact-token agreement.

Table 2: KL misranking. Each cell reports the percentage of total pairs that misrank, followed in parentheses by mean ΔQ (percentage points) among those misrankings for that intervention and pair type. Q ties and score ties count as non-misranking outcomes.

Task	Resampling		Mean		Zero	
	R–C	C–C	R–C	C–C	R–C	C–C
Human suite						
IOI	6.7 (0.33)	7.3 (3.56)	33.3 (8.82)	23.4 (8.05)	3.3 (0.50)	21.1 (15.90)
Greater-Than	0.0 (–)	3.4 (1.85)	0.0 (–)	5.9 (1.28)	56.7 (28.87)	54.0 (27.20)
Docstring	8.0 (0.92)	3.3 (0.99)	14.0 (4.86)	12.9 (5.00)	32.0 (4.40)	16.0 (6.03)
Acronym	5.0 (0.28)	5.8 (3.94)	15.0 (11.50)	20.6 (14.44)	10.0 (18.36)	29.2 (20.46)
InterpBench suite						
113	1.7 (0.04)	1.0 (0.44)	15.0 (0.37)	9.0 (1.24)	15.0 (0.22)	10.1 (2.28)
97	0.0 (–)	5.2 (10.09)	0.0 (–)	14.5 (23.24)	0.0 (–)	22.0 (21.74)
2	5.0 (0.07)	1.7 (0.13)	1.7 (0.07)	1.1 (0.82)	36.7 (27.25)	23.9 (27.03)
82	0.0 (–)	1.2 (6.43)	5.0 (0.21)	5.1 (8.79)	6.7 (19.22)	27.2 (32.99)
111	3.3 (0.22)	3.4 (0.76)	18.3 (1.15)	9.5 (1.03)	80.0 (5.24)	72.5 (4.33)
45	0.0 (–)	1.1 (1.95)	1.7 (0.04)	3.2 (4.39)	10.0 (15.91)	10.7 (21.04)
58	5.0 (0.38)	3.6 (0.47)	3.3 (0.37)	4.0 (2.17)	36.7 (48.39)	40.5 (38.28)
93	0.0 (–)	0.6 (2.22)	3.3 (0.04)	2.9 (8.50)	70.0 (63.55)	68.7 (56.59)
103	3.3 (0.22)	2.0 (0.76)	5.0 (0.37)	5.0 (9.67)	18.3 (12.21)	19.7 (24.06)
25	6.0 (0.27)	17.5 (2.40)	10.0 (0.36)	25.3 (6.79)	32.0 (49.03)	47.4 (40.96)

Experimental setup. For each reference, we replace retained components with the same number of excluded components. The six edit bands are one component, 5%, 10%, 20%, 50%, and 75% of reference size, with ten seeds per attainable distinct band. Identical masks and bands resolving to the same edit count are deduplicated. We compare references with candidates (R–C) and candidates with one another (C–C) under resampling, mean, and zero replacement. Candidate masks, prompts, and the test Q ordering remain fixed across evaluations. InterpBench references are checked against native execution before use; Appendix B details conversion validation, candidate generation, and the unattainable 75% band for task 25.

Misranking depends on the task and intervention. Table 2 reports KL misranking rates and the mean behavioral deficit among failures. The pattern is task-dependent. On Greater-Than, C–C misranking rises from 3.4% under resampling to 54.0% under zero replacement, while R–C misranking rises from zero to 56.7%. IOI shows a different pattern: mean replacement misranks 33.3% of R–C pairs, compared with 3.3% under zero replacement. Because the circuits and Q ordering do not change, these differences show that the intervention can change which candidate the score rewards. Rates use all eligible pairs, including ties, as the denominator.

Larger edits do not lead to a consistent change in misranking. One possible explanation is that ranking becomes unreliable only after a circuit has been changed substantially. Figure 2 tests this by separating the six edit bands. Misranking does not increase or decrease uniformly with edit magnitude. On Greater-Than, zero replacement misranks every R–C comparison in the 50% band; on Docstring, its R–C rate falls from 100% for the smallest edits to zero in the two largest bands. A circuit's distance from the reference therefore does not by itself predict whether faithfulness will recognize the better member of a pair. Band-specific counts and merged candidate pools are reported in Appendix B.4.

The failures include substantial behavioral deficits. InterpBench exhibits both near-null and high-misranking settings despite reference Q values of 97.33%–100%. Task 97 has no R–C KL misranking under any intervention, whereas task 111 reaches 80.0% R–C and 72.5% C–C under zero replacement. For task 93, zero replacement misranks 70.0% of R–C pairs. Among these pairs, the preferred circuit loses an average of 63.55 percentage points (pp) of Q relative to the alternative. This is a substantial cost of following the score. Other failures involve near-ties: task 113's resampling

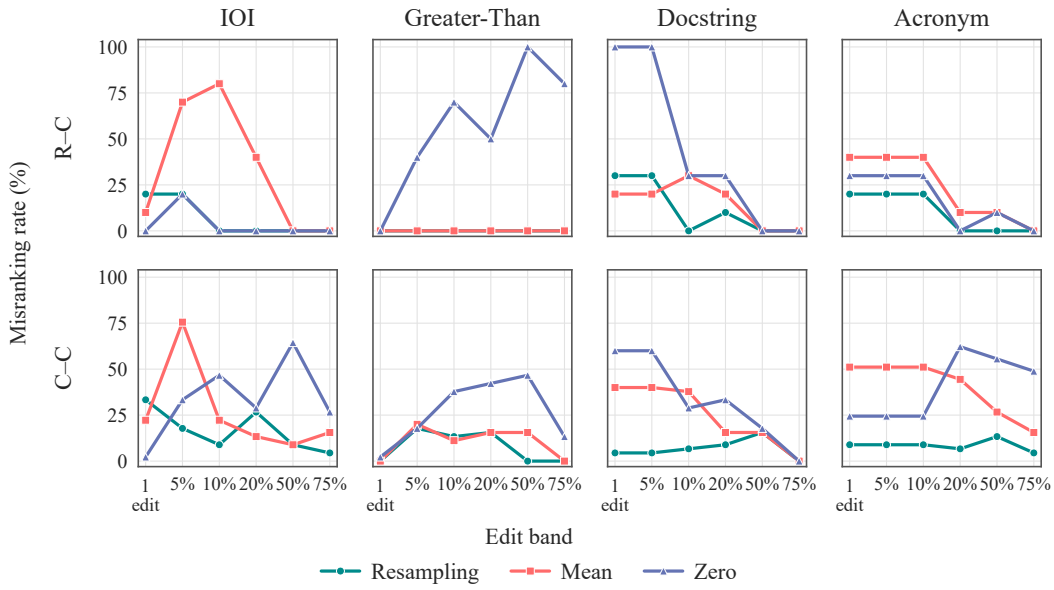


Figure 2: **Edit magnitude does not determine misranking monotonically.** Columns show human-reference tasks; rows show R–C and within-band C–C comparisons. The y-axis is the KL misranking rate over all pairs. Colors and markers denote interventions. The first band changes one component. Repeated points for merged bands reuse the same candidate set. Overall C–C results in Table 2 also include between-band pairs. Exact denominators are in Appendix B.4.

Table 3: KL/LD/PD on common pair identities across the 14 controlled tasks. Each cell reports percentage misranking followed by mean ΔQ in percentage points for that metric, intervention, and pair type; the denominator is the total common-pair pool.

Metric	Resampling		Mean		Zero	
	R–C	C–C	R–C	C–C	R–C	C–C
KL	3.0 (0.36)	3.7 (3.03)	8.8 (4.28)	9.3 (8.31)	29.5 (28.83)	33.3 (28.03)
LD	6.1 (1.07)	6.2 (2.70)	6.4 (3.22)	7.7 (5.68)	18.1 (13.05)	19.0 (16.37)
PD	6.5 (12.13)	5.0 (9.16)	6.1 (1.33)	8.8 (8.72)	25.1 (20.73)	25.1 (20.39)

R–C failures have a mean deficit of only 0.04 pp. The two cases illustrate why we report the size of the behavioral deficit alongside the frequency of misranking.

The ranking failure extends beyond a particular faithfulness metric. Faithfulness describes the goal of preserving model behavior under an intervention; KL, LD, and PD are different ways to score that execution. Our concern is whether such a score recognizes the behaviorally better circuit. Table 3 therefore repeats the same comparisons with all three metrics, keeping candidates, Q, prompts, and intervention fixed. Every metric exhibits misranking. The rates and failed pairs differ, but the objective-level problem is not specific to KL. The result supports a concern across the tested measures of faithfulness, rather than an objection to one formula. Appendix E reports which failures the metrics share.

Low reference–candidate rates can coexist with candidate-selection failures. A reference supplies an established circuit against which to test the score, but discovery also requires distinguishing between alternative candidates. InterpBench task 97 illustrates the distinction: it has no R–C KL misranking under any intervention, yet its C–C rates are 5.2%, 14.5%, and 22.0% under resampling, mean, and zero replacement. Recognizing the reference in these comparisons does not ensure reliable ordering of the candidates. We next test that ordering in the outputs of actual discovery methods.

Table 4: Human suite discovery: KL misranking among primary size-controlled pairs. Each cell reports percentage misranking followed by mean ΔQ in percentage points for that intervention and pair type; Q ties count as non-misranking outcomes.

Method	Resampling		Mean		Zero	
	R–C	C–C	R–C	C–C	R–C	C–C
EAP	13.8 (1.92)	20.0 (2.18)	48.3 (6.01)	31.0 (3.67)	20.7 (5.17)	40.0 (3.76)
EAP-IG	35.5 (1.33)	30.4 (1.36)	38.7 (1.17)	53.0 (1.49)	29.0 (0.67)	45.2 (1.25)
ACDC	57.1 (1.67)	41.2 (1.11)	50.0 (1.76)	39.2 (1.47)	28.6 (1.83)	43.1 (1.86)
Edge-SP	0.0 (–)	9.4 (1.73)	20.0 (7.92)	14.4 (1.91)	17.5 (34.14)	31.1 (8.42)

Table 5: InterpBench suite discovery: KL misranking among primary size-controlled pairs. Each cell reports percentage misranking followed by mean ΔQ in percentage points for that intervention and pair type; Q ties count as non-misranking outcomes.

Method	Resampling		Mean		Zero	
	R–C	C–C	R–C	C–C	R–C	C–C
EAP	5.0 (0.68)	8.7 (0.14)	16.0 (0.66)	13.3 (0.79)	49.0 (1.56)	25.6 (5.34)
EAP-IG	0.0 (–)	2.9 (0.07)	6.1 (1.10)	2.9 (0.10)	46.5 (1.44)	5.0 (0.06)
ACDC	10.5 (0.48)	7.9 (0.43)	18.4 (0.63)	24.1 (1.88)	30.3 (2.48)	46.0 (2.10)
Edge-SP	0.0 (–)	18.7 (0.15)	6.0 (0.51)	27.1 (0.25)	40.0 (1.36)	37.6 (0.27)

4 Faithfulness Can Misrank Discovered Circuits

Random edits need not resemble the circuits returned by a discovery method. To test practical relevance, we evaluate candidates from four established circuit-discovery methods: EAP, EAP-IG, ACDC, and Edge-SP. Each comparison stays within one task and one method. The question is whether faithfulness recognizes the better circuit among the outputs that a method actually returns.

Experimental setup. We use ten discovery seeds, with 100 discovery prompts per task and seed shared across methods and disjoint from evaluation prompts. We reuse *all validation and independent-test prompts from Section 3*, together with its mean banks, valid-position masks, and donor assignments. This keeps the evaluation fixed while changing how candidates are obtained. ACDC and Edge-SP have a 300-second discovery cap per run. Primary comparisons require equal actual executable size; other valid outputs, extraction adaptations, and runtime limits are documented in Appendix C. InterpBench reference comparisons use the validated native-reference closures.

Misranking occurs in the outputs of all four methods. The results in Tables 4 and 5 show that the failures are not limited to random edits. On the human-reference tasks, resampling KL misranks 20.0% of EAP C–C pairs, 30.4% for EAP-IG, 41.2% for ACDC, and 9.4% for Edge-SP. Misranking therefore occurs under ordinary resampling as well as mean and zero replacement. InterpBench resampling C–C rates range from 2.9% for EAP-IG to 18.7% for Edge-SP. These methods return better candidates, but faithfulness sometimes ranks them below worse alternatives.

Intervention choice affects selection within the same pool. On the human-reference tasks, EAP-IG C–C misranking rises from 30.4% under resampling to 53.0% under mean replacement. On InterpBench, ACDC C–C misranking rises from 7.9% under resampling to 46.0% under zero replacement. The scores also sometimes favor candidates with sizable Q deficits: for Human Edge-SP R–C pairs under zero replacement, the mean deficit among misrankings is 34.14 pp. By contrast, InterpBench EAP-IG resampling C–C failures average 0.07 pp. These preferences differ sharply in their consequences for behavior, even though both count as misrankings.

The rates should be read within each method's pool. Methods differ in task coverage, duplicate outputs, and the circuit sizes they reach, which prevents a direct comparison of their pooled percentages. In particular, the Human ACDC primary pool includes Docstring and Acronym; its capped IOI and

Greater-Than outputs belong to other size strata. Appendix C reports the complete inventories and denominators.

The Human suite reveals more frequent candidate-selection failures. Pooling the common discovery pairs within each suite, Human C–C KL misranking is 20.9%, 30.9%, and 38.1% under resampling, mean, and zero replacement, compared with 9.8%, 16.2%, and 26.8% on InterpBench (Appendix C.5). The Human rates are higher under each intervention. This pattern is not uniform across individual methods, but it shows why the smaller semi-synthetic benchmark is insufficient on its own to assess the reliability of discovery in pretrained language models.

InterpBench's smaller models and semi-synthetic tasks may make candidates easier to distinguish, although our comparison does not isolate these factors as causes. The Human suite provides more direct evidence about selection in pretrained models; low rates on synthetic benchmarks alone cannot establish reliability there.

The additional KL/LD/PD comparisons for discovery outputs are reported in Appendix C.5. They support the same conclusion as Table 3: changing the metric alters the failures but does not eliminate them. We now examine whether changing the intervention context can repair a fixed misranking.

5 Restoring Computational Context Can Repair Misranking

The preceding experiments show that faithfulness can prefer a worse circuit. We now investigate context distortion as an explanation. If the replacement signals contribute to the wrong preference, restoring some of the original signals should be able to correct it. We test this prediction on existing KL misrankings, keeping the two circuits and their original Q values fixed throughout.

Cases and controls. We select 100 equal-size circuit-pair/intervention cases from the Section 4 discovery pool: 50 Human and 50 InterpBench cases, spanning 50 resampling, 25 mean, and 25 zero cases. Each has a Q deficit of at least 0.5 pp and a strict KL misranking on both validation and independent-test prompts. Appendix D reports pair-type allocations and complete outcomes.

Partial context restoration. At selected excluded incoming connections, we replace intervention signals with activations cached from the recipient prompt's intact-model execution. Retained connections compute normally. Masks, weights, unrestored donor assignments, mean banks, prompts, and the original Q ordering remain fixed. Only selected replacement signals change, isolating whether context restoration can correct the preference.

Searching for a restoration that repairs the ranking. We group excluded incoming connections by receiving node and search at 80%, 40%, 20%, and 10% of the *original* eligible group count, with up to five predetermined attempts per level. The first validation-successful subset becomes the sampling parent for the next level; a failed level resets sampling to the original universe. Each qualifying subset is tested on independent prompts using their own intact signals and original donors. Test outcomes never guide the search. A repair requires at least one subset to correct both rankings beyond the predefined tolerance; later failures do not cancel earlier success. Search ends after the 10% level, within 20 attempts. Appendix D gives the full procedure.

The preference is repaired in 96 of 100 cases. Table 6 reports independently confirmed repairs, including 50/50 Human cases and 46/50 InterpBench cases. At the nominal 10% level alone, 82 cases have a confirmed repair. Success need not be monotonic in the restoration fraction because the search selects different subsets and can restart from the original universe. The any-level result counts each repaired case once.

The before (after) column reports $Q(\text{nonpreferred}) - Q(\text{preferred})$: positive when faithfulness prefers the worse circuit and negative after repair. The mean is $+10.33$ pp across all 100 baseline failures and -9.30 pp among the 96 repaired cases; those same 96 cases have a baseline mean of $+9.30$ pp. Original Q values remain unchanged. The four unconfirmed cases remain in repair-rate denominators and have no after-repair value.

Table 6: **Context restoration repairs persistent misrankings.** Entries are independent-test-confirmed repair rates in percent. Columns 80%–10% denote nominal restoration levels; successes can overlap across levels. Any level counts each case once. Mean ΔQ is shown as before (after) in pp, with signed deficit $Q(\text{nonpreferred}) - Q(\text{preferred})$. Before means include all selected cases; after means include only confirmed repairs, once per case. The underlying Q values remain fixed.

			Confirmed repair rates (%)				
Intervention	Pair	Mean ΔQ (pp) before (after)	80%	40%	20%	10%	Any level
Human suite (50 cases)							
Resampling	R–C	1.50 (−1.50)	100.0	100.0	50.0	50.0	100.0
	C–C	4.67 (−4.67)	87.0	95.7	87.0	91.3	100.0
Mean	R–C	14.21 (−14.21)	100.0	100.0	80.0	80.0	100.0
	C–C	25.96 (−25.96)	66.7	100.0	100.0	100.0	100.0
Zero	R–C	19.18 (−19.18)	90.9	90.9	100.0	81.8	100.0
	C–C	40.00 (−40.00)	100.0	100.0	100.0	100.0	100.0
InterpBench (50 cases)							
Resampling	R–C	4.19 (−4.19)	91.7	100.0	91.7	91.7	100.0
	C–C	2.42 (−2.42)	84.6	100.0	100.0	76.9	100.0
Mean	R–C	1.90 (−2.32)	66.7	50.0	50.0	33.3	66.7
	C–C	0.94 (−0.94)	83.3	83.3	100.0	100.0	100.0
Zero	R–C	27.96 (−27.96)	100.0	71.4	85.7	85.7	100.0
	C–C	26.23 (−4.82)	16.7	33.3	66.7	66.7	66.7
All 100 cases		10.33 (−9.30)	84.0	88.0	87.0	82.0	96.0

6 Discussion and conclusion

Faithfulness can misrank equally sized circuits, both in controlled comparisons and among actual discovery outputs. The failure persists under ordinary resampling, so avoiding zero ablation is insufficient. Higher pooled misranking in the Human suite highlights the importance of testing pretrained language models alongside semi-synthetic benchmarks. Restoring part of the intact computational context repairs 96 of 100 selected persistent failures while leaving circuits and their original behavioral scores fixed. These results support context distortion as a contributor to misranking and expose an objective-level recovery gap: improving circuit discovery does not ensure mechanism recovery if the evaluation objective prefers the worse candidate.

Implications for evaluation. Our results suggest reporting a fixed behavioral criterion alongside faithfulness, together with the intervention and circuit-size controls. Misranking rates should be accompanied by behavioral deficits: a reversal between nearly tied circuits and one that loses tens of percentage points have different consequences. Testing several metrics and interventions can reveal whether a preferred circuit depends on a particular evaluation choice (Miller et al., 2024; Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). Causal abstraction formalizes mechanistic explanations through correspondence under interventions (Geiger et al., 2021; 2025). Together with structural recovery and statistical circuit tests (Shi et al., 2024), it provides complementary evidence beyond output agreement.

Limitations. Our behavioral criterion Q depends on ordinary resampling and does not by itself establish mechanism identity. Tasks and search budgets are limited, and circuit pairs are dependent. Restoration results concern a selected cohort of persistent KL failures; they neither identify minimal causal sets nor establish superiority over random restoration. Restoration can reintroduce useful information, and successful repairs do not establish whether the same approach would work for other metrics or unselected failures. Pairwise reversals demonstrate local ranking failures, not differences between globally optimal circuits.

AI USE STATEMENT

The authors originated the research questions, methods, algorithms, and experimental design. Generative AI tools supported implementation and debugging, experiment execution, result summarization, and refinement of the manuscript and its presentation under the authors' supervision. The authors inspected and debugged the assisted code, checked the reported results, and made all methodological and interpretive decisions. They accept full responsibility for the final manuscript, code, and results.

ETHICS STATEMENT

This study evaluates saved model circuits and benchmark prompts. It does not introduce a human-subject study. Its principal integrity concern is overinterpreting operational circuit scores as evidence of complete mechanisms; we explicitly distinguish these claims and preserve nulls, ties, and incomplete strata.

REPRODUCIBILITY STATEMENT

Appendix A specifies shared models, hardware, prompt splits, scores, behavioral criteria, and uncertainty. Appendix B details reference conversion, controlled candidate generation, and additional Section 3 results. Appendix C provides discovery settings, executable-size checks, and task–method results. Appendix D specifies the 100-case cohort, restoration algorithm, and complete outcomes. Appendix E reports same-pair metric overlap and deficit support.

REFERENCES

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/20fdaf67581e6d7157376d1ed584040a-Abstract-Conference.html.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Jorge García-Carrasco, Alejandro Maté, and Juan C. Trujillo. How does GPT-2 predict acronyms? extracting and understanding a circuit via mechanistic interpretability. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3322–3330. PMLR, 2024. URL https://proceedings.mlr.press/v238/garcia-carrasco24a.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 9574–9586, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah D. Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *J. Mach. Learn. Res.*, 26: 83:1–83:64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

Rohan Gupta, Iván Arcuschin Moreno, Thomas Kwa, and Adrià Garriga-Alonso. Interpbench: Semi-synthetic transformers for evaluating mechanistic interpretability techniques. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/a8f7d43ae092d9a5295775eb17f3f4f7-Abstract-Datasets_and_Benchmarks_Track.html.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/efbba7719cc5172d175240f24be11280-Abstract-Conference.html.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *CoRR*, abs/2403.17806, 2024. doi: 10.48550/ARXIV.2403.17806. URL https://doi.org/10.48550/arXiv.2403.17806.

Stefan Heimersheim and Jett Janiak. A circuit for python docstrings in a 4-layer attention-only transformer. In *Alignment Forum*, 2023.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *CoRR*, abs/2404.15255, 2024. doi: 10.48550/ARXIV.2404.15255. URL https://doi.org/10.48550/arXiv.2404.15255.

David Lindner, János Kramár, Sebastian Farquhar, Matthew Rahtz, Tom McGrath, and Vladimir Mikulik. Tracr: Compiled transformers as a laboratory for interpretability. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/771155abaae744e08576f1f3b4b7ac0d-Abstract-Conference.html.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. *CoRR*, abs/2407.08734, 2024. doi: 10.48550/ARXIV.2407.08734. URL https://doi.org/10.48550/arXiv.2407.08734.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Victor Sanh, Thomas Wolf, and Alexander M. Rush. Movement pruning: Adaptive sparsity by fine-tuning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/eae15aabaa768ae4a5993a8a4f4fa6e4-Abstract.html.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M. Blei. Hypothesis testing the circuit hypothesis in llms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/abccb8a90b30d45b948360ba41f5a20f-Abstract-Conference.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings*

of Machine Learning Research, pp. 3319–3328. PMLR, 2017. URL http://proceedings.mlr.press/v70/sundararajan17a.html.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. *CoRR*, abs/2310.10348, 2023. doi: 10.48550/ARXIV.2310.10348. URL https://doi.org/10.48550/arXiv.2310.10348.

Niels uit de Bos and Adrià Garriga-Alonso. Adversarial circuit evaluation. *CoRR*, abs/2407.15166, 2024. doi: 10.48550/ARXIV.2407.15166. URL https://doi.org/10.48550/arXiv.2407.15166.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

A SHARED EXPERIMENTAL AND ANALYSIS DETAILS

A.1 MODELS, HARDWARE, AND FROZEN PROMPT FIXTURES

Table 1 summarizes the task behaviors and original circuit sources. IOI, Greater-Than, and Acronym use GPT-2 small; Docstring uses the four-layer attention-only, width-512 code model studied with its reference circuit. IOI, Greater-Than, and Docstring are represented by factorized attention/residual edges; Acronym is represented by attention heads. InterpBench supplies a separate compiled transformer and native ground-truth computation for each task. Component counts are meaningful within these representations, not comparable units across models.

Inference used Ubuntu 22.04, Python 3.12.3, PyTorch 2.3.0 with CUDA 12.1, an NVIDIA vGPU with 48 GiB VRAM, and an Intel Xeon Platinum 8481C host with 1 TiB RAM. PyTorch used four CPU threads and TF32 was disabled. Sections 3 and 4 share exactly the same frozen 100 validation prompts, 100 independent-test prompts, 100-prompt mean bank, and three donor assignments. Validation and test prompt identities are disjoint. Discovery prompts are separate from both evaluation splits. The exact reuse of both evaluation splits, including every validation and held-out prompt, controls comparisons between Sections 3 and 4. Independence refers to the disjoint validation and test prompts, not a new test sample for each section.

Human prompt selection preserves the original task strata: IOI uses the saved ABBA template, Greater-Than preserves the century strata, Docstring uses the fixed 41-token stratum, and Acronym uses four-word prompts. For InterpBench, the saved ordered prompts and their valid-position masks are retained. Three donor draws are averaged within each recipient prompt before averaging prompts; they are not three independent observations. The frozen mean bank is reused for mean intervention, and zero intervention replaces the same excluded signals by zero.

A.2 SCORES AND BEHAVIORAL TARGETS

KL compares the intact and intervened output distributions at the saved scored positions. Human tasks use the original vocabulary distribution; Acronym scores its three teacher-forced letter positions. InterpBench uses its original categorical output space and nine valid positions per prompt. Position-level contributions are averaged within a prompt before prompt averaging.

LD and PD targets are specified by the dataset, independently of candidate predictions. IOI and Docstring use the correct answer option as P and the remaining frozen options as N. Acronym uses the correct letter token and the other 25 alphabet tokens at each teacher-forced position. Greater-Than uses valid two-digit year endings strictly above the prompt's starting ending as P, with the remaining valid endings as N. InterpBench uses the dataset-correct class as P and the remaining original classes as N. Its class counts, in task order 113, 97, 2, 82, 111, 45, 58, 93, 103, and 25, are 30, 64, 26, 16, 11, 16, 11, 11, 11, and 56. PD probabilities come from the full original softmax, without renormalizing to the answer support.

Ordinary-task Q compares task-native answers with intact-model answers under ordinary resampling, using the same three donor draws and prompt aggregation. Greater-Than Q instead measures whether the predicted two-digit ending strictly exceeds the prompt's starting ending. Its reference semantic Q is 99.67%; exact-token agreement is not substituted for this criterion. The same Q values orient KL, LD, and PD pairs, regardless of the validation intervention.

A.3 TOTAL-PAIR DENOMINATORS, TIES, AND UNCERTAINTY

After deduplicating masks within a task and candidate universe, R–C contains each reference–candidate identity once and C–C contains each unordered distinct candidate pair once. Primary pairs have equal actual executable size. The misranking rate is the number of strict misrankings divided by the *total* pair pool, including Q ties and score ties. Ties are non-misranking outcomes, but are not evidence of a correct strict ordering. An empty pool is unavailable, not a zero rate. For common-metric analyses, the denominator is the intersection of exact pair identities available for all three scores.

Numerical comparisons use the following absolute and relative tolerances: absolute and relative tolerances of 10^{-6} for the Human suite and 10^{-12} and 10^{-9}, respectively, for the corrected InterpBench

Table 7: InterpBench reference-conversion checks. Projected and closure columns count executable edges, not candidate variants. Logit errors compare the native node intervention with the outgoing-edge closure on the saved eight-prompt resampling probe; they are representation checks, not held-out Q estimates.

Task	Projected edges	Closure edges	Max logit error	RMS error
113	30	587	1.42e-14	2.42e-15
97	20	127	7.11e-14	1.25e-14
2	14	177	2.13e-14	5.46e-15
82	10	176	5.68e-14	1.32e-14
111	14	112	1.14e-13	2.97e-14
45	9	110	4.26e-14	8.22e-15
58	13	111	2.49e-14	7.63e-15
93	13	111	3.91e-14	1.09e-14
103	13	111	4.26e-14	9.40e-15
25	10	72	2.56e-13	5.09e-14

controlled analysis. Rounding displayed values never determines pair eligibility. Every positive Q deficit reported with a misranking rate is averaged only over strict misrankings in that cell. No 0.5-pp deficit threshold is imposed in Sections 3 or 4; that threshold is specific to the Section 5 cohort.

Paired bootstrap calculations resample prompts while preserving paired circuit contributions and within-prompt donor averaging. The Human controlled analysis uses 1,000 bootstrap draws and the corrected InterpBench analysis uses 10,000. We report paired support only when it was computed for the Q definition and score direction used in that comparison. In particular, semantic-Q intervals for Greater-Than do not provide joint LD/PD uncertainty, and support computed for exact-token agreement cannot be used for semantic Q. Pooled pair counts and deficit summaries are descriptive because circuits recur across pairs and interventions.

B Section 3: controlled candidates, reference conversion, and additional results

B.1 Converting InterpBench references without dropping their computation

The authoritative reference is InterpBench's native node computation. A direct projection of the listed native graph edges into the factorized edge executor can omit outgoing contributions of retained nodes. Such a projection changes execution even when the nominal reference nodes are unchanged. We therefore use an outgoing-edge closure, rather than interpreting the projected edge list as the ground-truth executable mask.

The conversion maps each retained attention output from the native head-output hook to its residual-stream result, and each retained MLP postactivation to its MLP output. The input embedding is retained as the root source. For every retained native source, the closure keeps all of its outgoing connections in the factorized Q/K/V and residual executor. Complementary sources are intervened upon. This preserves the source contribution at every destination where the native node execution supplies it. The resulting edge count can greatly exceed the directly projected graph count: for task 113, 30 projected edges become a 587-edge executable closure.

We validate the adapted reference against native reference execution under resampling, mean, and zero replacement, comparing logits and Q under matching intervention semantics. Unresolved conversions are excluded rather than treated as ground truth; this excludes `ioi_next_token`, whose adapted reference closure remains unresolved. Table 7 reports eight-prompt, one-donor resampling probes for all ten tasks. The largest absolute discrepancy is 2.56×10^{-13}. These probes check equivalence between the two circuit representations on the evaluated prompts. Agreement with the intact model is a separate measurement, reported as reference Q below. Both the controlled and discovery analyses use the validated closures.

Table 8: Controlled candidate inventory. Entries give the number of swapped components followed by the number of unique candidates in parentheses. Columns with the same edit count refer to the same deduplicated masks. A dash means the band is unattainable. Reference size is measured in factorized edges, except Acronym, which uses attention heads.

Task	Reference size	1 edit	5%	10%	20%	50%	75%
IOI	963	1 (10)	48 (10)	96 (10)	193 (10)	482 (10)	722 (10)
Greater-Than	235	1 (10)	12 (10)	24 (10)	47 (10)	118 (10)	176 (10)
Docstring	24	1 (10)	1 (10)	2 (10)	5 (10)	12 (10)	18 (10)
Acronym	8	1 (10)	1 (10)	1 (10)	2 (10)	4 (10)	6 (10)
113	587	1 (10)	29 (10)	59 (10)	117 (10)	294 (10)	440 (10)
97	127	1 (10)	6 (10)	13 (10)	25 (10)	64 (10)	95 (10)
2	177	1 (10)	9 (10)	18 (10)	35 (10)	89 (10)	133 (10)
82	176	1 (10)	9 (10)	18 (10)	35 (10)	88 (10)	132 (10)
111	112	1 (10)	6 (10)	11 (10)	22 (10)	56 (10)	84 (10)
45	110	1 (10)	6 (10)	11 (10)	22 (10)	55 (10)	83 (10)
58	111	1 (10)	6 (10)	11 (10)	22 (10)	56 (10)	83 (10)
93	111	1 (10)	6 (10)	11 (10)	22 (10)	56 (10)	83 (10)
103	111	1 (10)	6 (10)	11 (10)	22 (10)	56 (10)	83 (10)
25	72	1 (10)	4 (10)	7 (10)	14 (10)	36 (10)	–

Table 9: Section 3 Human suite: reference Q and mean candidate Q (percent), with unique candidate count in parentheses per band. Merged bands share the same candidates; those are not duplicated in pooled results. Greater-Than uses semantic accuracy; other tasks use intact-model answer agreement. All Q evaluations use ordinary resampling.

Task	Reference	1 edit	5%	10%	20%	50%	75%
IOI	99.00	99.00 (10)	97.80 (10)	83.33 (10)	59.40 (10)	58.20 (10)	49.30 (10)
Greater-Than	99.67	99.33 (10)	97.67 (10)	85.03 (10)	84.47 (10)	55.00 (10)	53.70 (10)
Docstring	34.00	32.57 (10)	32.57 (10)	28.20 (10)	25.03 (10)	12.63 (10)	9.67 (10)
Acronym	87.67	80.67 (10)	80.67 (10)	80.67 (10)	70.66 (10)	43.92 (10)	23.51 (10)

B.2 Size-preserving candidate generation

The task panel is exactly 113, 97, 2, 82, 111, 45, 58, 93, 103, and 25. Given a reference with K components, the requested swap count is one component or $k = \max(1, \lfloor fK + 0.5 \rfloor)$ for $f \in \{0.05, 0.10, 0.20, 0.50, 0.75\}$. Each candidate removes k retained components and adds k components from outside the reference, preserving size K. Thus a "20%" band specifies 20% reference replacements, not a 20% reduction in size. Ten seeds, 810–819, generate candidates at each attainable distinct swap count; the InterpBench generator offsets each seed by $100000k$.

Canonical mask identities are deduplicated, and bands that round to the same k share the same masks. This occurs in Docstring and Acronym and explains repeated figure points. A swap count is attainable only if both the retained and complementary pools contain at least k components. Task 25 has only 50 candidates because its 54-swap, 75% band is unattainable. Table 8 records realized counts and unique candidate inventories for every task.

B.3 Reference and edited-candidate behavioral performance

Candidate Q means average unique circuit identities within a band, rather than pairs containing those identities. Repeated merged bands therefore show the same circuit-level mean. Reference Q is measured using the same ordinary-resampling independent-test criterion. Tables 9 and 10 separate candidate quality from ranking reliability. Larger edits generally reduce mean Q, but sensitivity varies: at 75% replacement, InterpBench task 111 retains 90.19% mean Q, whereas task 113 falls to 3.11%. The Human references also differ in baseline Q, so low absolute agreement and misranking are distinct observations.

Table 10: Section 3 InterpBench suite: reference Q and mean candidate Q (percent), with unique candidate count in parentheses per band. Merged bands share the same candidates; those are not duplicated in pooled results. Greater-Than uses semantic accuracy; other tasks use intact-model answer agreement. All Q evaluations use ordinary resampling.

Task	Reference	1 edit	5%	10%	20%	50%	75%
113	99.93	99.92 (10)	96.59 (10)	92.73 (10)	60.46 (10)	3.59 (10)	3.11 (10)
97	100.00	100.00 (10)	75.40 (10)	44.49 (10)	25.01 (10)	19.49 (10)	7.37 (10)
2	97.78	92.18 (10)	84.39 (10)	60.45 (10)	53.60 (10)	3.50 (10)	3.48 (10)
82	99.44	97.07 (10)	67.29 (10)	65.94 (10)	40.41 (10)	15.60 (10)	10.22 (10)
111	97.33	96.79 (10)	94.75 (10)	92.88 (10)	92.93 (10)	90.35 (10)	90.19 (10)
45	98.22	98.17 (10)	69.28 (10)	55.97 (10)	38.63 (10)	10.58 (10)	10.62 (10)
58	98.78	93.66 (10)	89.00 (10)	56.79 (10)	31.77 (10)	10.40 (10)	10.47 (10)
93	99.89	94.63 (10)	90.63 (10)	60.99 (10)	38.74 (10)	11.23 (10)	11.16 (10)
103	99.41	93.77 (10)	86.54 (10)	49.79 (10)	33.46 (10)	10.46 (10)	10.15 (10)
25	99.33	99.31 (10)	78.27 (10)	46.94 (10)	43.90 (10)	21.29 (10)	–

B.4 Task-level, within-band, and margin-score results

Table 11 gives exact KL misranked / total pair counts for all available R–C and C–C comparisons. Figure 2 uses R–C pairs within each band and C–C pairs whose two candidates belong to that band. Overall task-level C–C counts also include between-band pairs. Table 12 provides the counts underlying the figure, so a high rate from a small pool can be distinguished from a frequent failure in a large pool. For example, IOI mean-replacement R–C failures rise from 1/10 for one edit to 7/10 at 5% replacement, then fall to 0/10 at 50%, illustrating the nonmonotonic pattern.

Table 11: Complete Section 3 KL misranking counts. Every cell is misranked / total pairs, with Q ties included in the denominator.

Task	Resampling		Mean		Zero	
	R–C	C–C	R–C	C–C	R–C	C–C
Human suite						
IOI	4/60	130/1770	20/60	414/1770	2/60	373/1770
Greater-Than	0/60	61/1770	0/60	104/1770	34/60	956/1770
Docstring	4/50	41/1225	7/50	158/1225	16/50	196/1225
Acronym	2/40	45/780	6/40	161/780	4/40	228/780
InterpBench suite						
113	1/60	17/1770	9/60	160/1770	9/60	179/1770
97	0/60	92/1770	0/60	257/1770	0/60	389/1770
2	3/60	30/1770	1/60	20/1770	22/60	423/1770
82	0/60	22/1770	3/60	91/1770	4/60	482/1770
111	2/60	61/1770	11/60	169/1770	48/60	1284/1770
45	0/60	20/1770	1/60	57/1770	6/60	190/1770
58	3/60	63/1770	2/60	71/1770	22/60	716/1770
93	0/60	10/1770	2/60	51/1770	42/60	1216/1770
103	2/60	35/1770	3/60	88/1770	11/60	349/1770
25	3/50	214/1225	5/50	310/1225	16/50	581/1225

Table 12: Exact misranked / total pair counts underlying Figure 2. C–C comparisons in this table are within the same band; between-band pairs appear in the overall task table. Tied pairs remain in the denominator.

Task	Band	Resampling		Mean		Zero	
		R–C	C–C	R–C	C–C	R–C	C–C
IOI	1 edit	2/10	15/45	1/10	10/45	0/10	1/45
IOI	5%	2/10	8/45	7/10	34/45	2/10	15/45
IOI	10%	0/10	4/45	8/10	10/45	0/10	21/45
IOI	20%	0/10	12/45	4/10	6/45	0/10	13/45
IOI	50%	0/10	4/45	0/10	4/45	0/10	29/45
IOI	75%	0/10	2/45	0/10	7/45	0/10	12/45
Greater-Than	1 edit	0/10	0/45	0/10	0/45	0/10	1/45
Greater-Than	5%	0/10	8/45	0/10	9/45	4/10	8/45
Greater-Than	10%	0/10	6/45	0/10	5/45	7/10	17/45
Greater-Than	20%	0/10	7/45	0/10	7/45	5/10	19/45
Greater-Than	50%	0/10	0/45	0/10	7/45	10/10	21/45
Greater-Than	75%	0/10	0/45	0/10	0/45	8/10	6/45
Docstring	1 edit	3/10	2/45	2/10	18/45	10/10	27/45
Docstring	5%	3/10	2/45	2/10	18/45	10/10	27/45
Docstring	10%	0/10	3/45	3/10	17/45	3/10	13/45
Docstring	20%	1/10	4/45	2/10	7/45	3/10	15/45
Docstring	50%	0/10	7/45	0/10	7/45	0/10	8/45
Docstring	75%	0/10	0/45	0/10	0/45	0/10	0/45
Acronym	1 edit	2/10	4/45	4/10	23/45	3/10	11/45
Acronym	5%	2/10	4/45	4/10	23/45	3/10	11/45
Acronym	10%	2/10	4/45	4/10	23/45	3/10	11/45
Acronym	20%	0/10	3/45	1/10	20/45	0/10	28/45
Acronym	50%	0/10	6/45	1/10	12/45	1/10	25/45
Acronym	75%	0/10	2/45	0/10	7/45	0/10	22/45

Table 13: Section 3 task-level LD and PD misranking counts. Each entry is misranked pairs / total pairs within its own fixed task universe, including Q and score ties in the denominator. A dash means no primary pair pool. The pooled common-metric comparison uses the intersection of pair identities rather than pooling these denominators.

Task	Metric	Resampling		Mean		Zero	
		R–C	C–C	R–C	C–C	R–C	C–C
IOI	LD	1/60	49/1770	2/60	121/1770	46/60	1406/1770
IOI	PD	2/60	172/1770	2/60	105/1770	45/60	1402/1770
Greater-Than	LD	1/60	159/1770	1/60	180/1770	6/60	398/1770
Greater-Than	PD	0/60	43/1770	0/60	77/1770	12/60	479/1770
Docstring	LD	4/50	44/1225	9/50	175/1225	8/50	167/1225
Docstring	PD	38/50	598/1225	9/50	700/1225	39/50	911/1225
Acronym	LD	1/40	22/780	2/40	40/780	7/40	148/780
Acronym	PD	0/40	17/780	5/40	49/780	12/40	231/780
113	LD	4/60	87/1770	4/60	88/1770	4/60	88/1770
113	PD	3/60	44/1770	3/60	68/1770	2/60	65/1770
97	LD	0/60	92/1770	0/60	158/1770	0/60	97/1770
97	PD	0/60	46/1770	0/60	188/1770	0/60	151/1770
2	LD	2/60	30/1770	1/60	22/1770	2/60	58/1770
2	PD	3/60	31/1770	1/60	18/1770	8/60	196/1770
82	LD	1/60	25/1770	1/60	37/1770	1/60	45/1770
82	PD	0/60	2/1770	3/60	82/1770	4/60	196/1770
111	LD	14/60	444/1770	10/60	406/1770	46/60	1078/1770
111	PD	2/60	27/1770	11/60	157/1770	33/60	897/1770
45	LD	1/60	35/1770	2/60	47/1770	5/60	88/1770
45	PD	0/60	6/1770	2/60	61/1770	5/60	128/1770
58	LD	3/60	68/1770	3/60	66/1770	5/60	105/1770
58	PD	1/60	31/1770	2/60	66/1770	10/60	254/1770
93	LD	3/60	42/1770	3/60	39/1770	5/60	125/1770
93	PD	0/60	5/1770	2/60	53/1770	6/60	166/1770
103	LD	2/60	31/1770	2/60	34/1770	3/60	93/1770
103	PD	1/60	12/1770	4/60	94/1770	6/60	150/1770
25	LD	12/50	279/1225	11/50	344/1225	7/50	422/1225
25	PD	2/50	106/1225	5/50	276/1225	19/50	480/1225

Table 14: Section 3 KL deficits among misranked primary pairs. Distributions are median [first quartile, third quartile]; Q gaps are in pp and score gaps are in KL units. Counts pool dependent pair records descriptively; a circuit may occur in many pairs.

Task	Intervention	Pair	Misranked	ΔQ (pp)	ΔF	$\Delta Q \geq 5$
Human suite	Resampling	R–C	10	0.33 [0.33, 0.33]	0.034 [0.019, 0.055]	0
Human suite	Resampling	C–C	277	1.00 [0.33, 2.33]	0.079 [0.031, 0.173]	43
Human suite	Mean	R–C	33	4.33 [1.67, 9.00]	0.304 [0.082, 0.471]	16
Human suite	Mean	C–C	837	3.33 [1.00, 10.33]	0.177 [0.061, 0.353]	350
Human suite	Zero	R–C	56	6.33 [1.33, 45.67]	0.716 [0.120, 2.252]	31
Human suite	Zero	C–C	1753	17.11 [2.67, 42.33]	1.054 [0.293, 2.678]	1193
InterpBench suite	Resampling	R–C	14	0.11 [0.04, 0.39]	0.001 [0.000, 0.013]	0
InterpBench suite	Resampling	C–C	564	0.52 [0.11, 4.84]	0.524 [0.016, 4.100]	139
InterpBench suite	Mean	R–C	37	0.33 [0.07, 0.78]	0.000 [0.000, 0.001]	0
InterpBench suite	Mean	C–C	1274	1.44 [0.33, 13.66]	0.047 [0.000, 1.533]	484
InterpBench suite	Zero	R–C	180	7.15 [1.20, 78.19]	1.937 [0.371, 35.915]	114
InterpBench suite	Zero	C–C	5809	13.48 [3.30, 55.67]	3.677 [0.825, 25.036]	3981

Table 13 reports LD and PD within each task's own candidate universe, while the main common-metric table uses only pair identities shared by all three metrics. Table 14 complements mean KL deficits with their distributions. Human C–C failures under zero replacement have a median Q deficit of 17.11 pp, with 1,193 of 1,753 deficits at least 5 pp; these failures therefore include substantial behavioral losses. Small gaps remain in the analysis; none are discarded to favor a particular intervention or metric. Cross-metric overlap and existing uncertainty support are reported in Appendix E.

C Section 4: discovery setup, executable-size control, and additional results

C.1 Discovery objectives and budgets

Our discovery runs use the authors' released implementations of EAP, EAP-IG, ACDC, and Edge-SP (Syed et al., 2023; Hanna et al., 2024; Conmy et al., 2023), with the task-specific adaptations and executable-size checks described below. Discovery uses seeds 810–819 and 100 prompts per task and seed, shared across methods and disjoint from evaluation; validation and independent-test scoring reuse the frozen Section 3 fixtures, mean bank, valid positions, and three donor draws. Human EAP-IG runs use 30 integration steps. InterpBench attribution uses the categorical KL discovery objective, batches of ten prompts, and 30 EAP-IG steps. These choices are held fixed when changing the subsequent validation metric to LD or PD.

Configured ACDC and Edge-SP searches have a 300-second cap, excluding setup, extraction, and scoring. ACDC follows its recorded sequential deletion trajectory. Edge-SP freezes model weights and optimizes mask parameters with Adam at learning rate 0.01, sparsity regularization coefficient 1, and at most 300 full-100-prompt updates, using microbatches of four. A sampled mask is shared across the microbatches of a full update. Native mask optimization uses FP32 and saved InterpBench evaluation uses FP64. Acronym requires a head-level adaptation: its masks act on attention-head outputs rather than factorized edges.

C.2 Corrected extraction and actual-size eligibility

InterpBench reference sizes and the conversion checks are those in Appendix B. Attribution candidates are extracted at the corrected reference size from their saved full-universe scores, ordered by absolute attribution with canonical edge names breaking ties. EAP-IG uses its integrated scores. ACDC uses the first recorded deletion checkpoint at the required size, replaying accepted native deletions. Edge-SP uses the discovery-objective checkpoint and its top-K retention parameters. Evaluation prompts and Q do not select extraction checkpoints.

Table 15: Section 4 Human suite: reference Q and mean saved-candidate Q (percent), with candidate count in parentheses. This performance inventory includes non-primary sizes, which are excluded from the primary misranking tables. Methods are not ranked by this table. Greater-Than uses semantic accuracy; other tasks use intact-model answer agreement. All Q evaluations use ordinary resampling.

Task	Reference	EAP	EAP-IG	ACDC	Edge-SP
IOI	99.00	90.70 (10)	98.10 (10)	98.50 (10)	58.80 (10)
Greater-Than	99.67	96.38 (8)	99.48 (9)	100.00 (4)	59.87 (10)
Docstring	34.00	30.37 (9)	34.47 (10)	34.90 (10)	35.57 (10)
Acronym	87.67	84.53 (4)	83.00 (4)	85.17 (4)	23.80 (10)

Table 16: Section 4 InterpBench suite: reference Q and mean saved-candidate Q (percent), with candidate count in parentheses. This performance inventory includes non-primary sizes, which are excluded from the primary misranking tables. Methods are not ranked by this table. Greater-Than uses semantic accuracy; other tasks use intact-model answer agreement. All Q evaluations use ordinary resampling.

Task	Reference	EAP	EAP-IG	ACDC	Edge-SP
113	99.93	66.51 (10)	100.00 (10)	97.30 (10)	96.83 (10)
97	100.00	25.90 (10)	100.00 (10)	97.62 (4)	100.00 (10)
2	97.78	100.00 (10)	100.00 (10)	98.14 (6)	99.34 (10)
82	99.44	80.53 (10)	100.00 (10)	94.71 (10)	96.03 (10)
111	97.33	96.99 (10)	99.85 (10)	98.01 (10)	99.01 (10)
45	98.22	98.32 (10)	99.75 (10)	97.15 (10)	93.96 (10)
58	98.78	97.61 (10)	100.00 (9)	97.38 (8)	99.96 (10)
93	99.89	99.95 (10)	100.00 (10)	97.92 (6)	97.06 (10)
103	99.41	95.88 (10)	100.00 (10)	97.03 (8)	99.83 (10)
25	99.33	98.67 (10)	99.99 (10)	99.49 (4)	99.88 (10)

Before scoring, candidate extraction is matched to the validated native-reference closure and the corresponding executable component universe. We check executable keys, actual mask size, canonical mask identity, and finite outputs. Only equal-actual-size reference–candidate and candidate–candidate pairs enter primary misranking tables. Duplicate masks are collapsed before pairs are enumerated.

The Human suite ACDC primary stratum contains Docstring and Acronym; the capped IOI and Greater-Than outputs have other actual sizes. We retain these outputs in the performance inventory and exclude them from the equal-size comparisons. The same actual-size requirement applies to the other methods. A dash denotes an unavailable pair pool; zero denotes an available pool with no misranking.

C.3 Reference and discovered-candidate performance

Tables 15 and 16 report candidate Q, averaging unique circuit identities represented in each task–method inventory, not their repeated appearances across pairs. The inventory also includes outputs excluded from the equal-size comparisons, so its means should be interpreted with the reported circuit sizes. Greater-Than uses semantic Q throughout, including ACDC and Edge-SP LD/PD comparisons. Each margin score is matched to the semantic-Q value of the same circuit mask before determining the pair ordering. Candidate quality varies across tasks and methods. For example, EAP-IG reaches mean Q of 100% on seven InterpBench tasks, while its Human-task means range from 34.47% on Docstring to 99.48% on Greater-Than. These values describe the available candidates, not whether faithfulness orders them correctly.

C.4 Complete task–method misranking and deficit summaries

Tables 17 and 18 report KL and LD/PD counts by task, method, intervention, and pair type. On IOI EAP resampling C–C pairs, KL misranks 11/45, compared with 5/45 for LD and 13/45 for PD: changing the score changes the failures without eliminating them. Their total-pair denominators include Q and score ties. The main method summaries are conditional on the method's own candidate distribution; the common-metric summaries additionally intersect exact pair identities. Table 19 distinguishes frequency from severity. Under resampling, median C–C deficits are 1.00 pp for Human

and 0.11 pp for InterpBench; 4/93 and 0/158 failures, respectively, exceed or equal 5 pp. Thus frequent reversals need not imply uniformly large losses.

Table 17: Section 4 task-level KL misranking counts (misranked/total pairs). Ties count as non-misranking outcomes; dashes denote unavailable pair pools.

Task	Method	Resampling		Mean		Zero	
		R–C	C–C	R–C	C–C	R–C	C–C
IOI	EAP	0/10	11/45	9/10	22/45	0/10	14/45
	EAP-IG	7/10	15/45	7/10	31/45	2/10	23/45
	ACDC	—	—	—	—	—	—
	Edge-SP	0/10	8/45	1/10	8/45	0/10	16/45
Greater-Than	EAP	1/8	4/28	1/8	3/28	3/8	13/28
	EAP-IG	2/9	12/36	4/9	18/36	5/9	19/36
	ACDC	—	—	—	—	—	—
	Edge-SP	0/10	2/45	0/10	6/45	6/10	15/45
Docstring	EAP	2/7	3/21	1/7	4/21	3/7	11/21
	EAP-IG	2/8	8/28	1/8	9/28	2/8	6/28
	ACDC	4/10	18/45	3/10	17/45	4/10	20/45
	Edge-SP	0/10	5/45	7/10	9/45	1/10	13/45
Acronym	EAP	1/4	2/6	3/4	2/6	0/4	2/6
	EAP-IG	0/4	0/6	0/4	3/6	0/4	4/6
	ACDC	4/4	3/6	4/4	3/6	0/4	2/6
	Edge-SP	0/10	2/45	0/10	3/45	0/10	12/45
113	EAP	0/10	2/45	0/10	6/45	1/10	16/45
	EAP-IG	0/10	0/45	0/10	0/45	0/10	0/45
	ACDC	0/10	2/45	0/10	13/45	0/10	11/45
	Edge-SP	0/10	19/45	0/10	15/45	0/10	16/45
97	EAP	0/10	15/45	0/10	13/45	0/10	6/45
	EAP-IG	0/10	0/45	0/10	0/45	0/10	0/45
	ACDC	0/4	0/6	0/4	1/6	0/4	1/6
	Edge-SP	0/10	0/45	0/10	0/45	0/10	0/45
2	EAP	0/10	0/45	0/10	0/45	10/10	0/45
	EAP-IG	0/10	0/45	0/10	0/45	10/10	0/45
	ACDC	0/6	1/15	4/6	1/15	4/6	6/15
	Edge-SP	0/10	6/45	0/10	13/45	10/10	13/45
82	EAP	0/10	0/45	4/10	4/45	2/10	16/45
	EAP-IG	0/10	0/45	0/10	0/45	1/10	0/45
	ACDC	0/10	9/45	0/10	33/45	9/10	28/45
	Edge-SP	0/10	8/45	0/10	18/45	0/10	26/45
111	EAP	3/10	7/45	4/10	12/45	5/10	26/45
	EAP-IG	0/10	11/45	0/10	9/45	10/10	13/45
	ACDC	2/10	5/45	4/10	9/45	2/10	24/45
	Edge-SP	0/10	7/45	0/10	4/45	9/10	31/45
45	EAP	1/10	3/45	4/10	11/45	1/10	15/45
	EAP-IG	0/10	2/45	5/10	4/45	0/10	1/45
	ACDC	2/10	2/45	1/10	3/45	4/10	19/45
	Edge-SP	0/10	3/45	0/10	20/45	0/10	22/45
58	EAP	0/10	12/45	0/10	7/45	9/10	11/45
	EAP-IG	0/9	0/36	0/9	0/36	9/9	0/36
	ACDC	0/8	1/28	3/8	0/28	3/8	8/28
	Edge-SP	0/10	0/45	0/10	0/45	10/10	0/45
93	EAP	0/10	0/45	1/10	0/45	6/10	8/45
	EAP-IG	0/10	0/45	0/10	0/45	6/10	0/45
	ACDC	0/6	0/15	0/6	2/15	0/6	9/15
	Edge-SP	0/10	9/45	0/10	18/45	1/10	23/45
103	EAP	0/10	0/45	2/10	1/45	7/10	5/45
	EAP-IG	0/10	0/45	0/10	0/45	10/10	0/45
	ACDC	3/8	1/28	0/8	2/28	0/8	17/28
	Edge-SP	0/10	13/45	0/10	17/45	10/10	14/45
25	EAP	1/10	0/45	1/10	6/45	8/10	12/45
	EAP-IG	0/10	0/45	1/10	0/45	0/10	8/45
	ACDC	1/4	1/6	2/4	3/6	1/4	5/6
	Edge-SP	0/10	19/45	6/10	17/45	0/10	24/45

Table 18: Section 4 task-level LD and PD misranking counts. Each entry is misranked pairs / total pairs within its own fixed task and method universe, including Q and score ties in the denominator. A dash means no primary pair pool. The pooled common-metric comparison uses the intersection of pair identities rather than pooling these denominators.

Task	Method	Metric	Resampling		Mean		Zero	
			R–C	C–C	R–C	C–C	R–C	C–C
IOI	EAP	LD	0/10	5/45	0/10	21/45	10/10	25/45
IOI	EAP	PD	0/10	13/45	0/10	23/45	10/10	23/45
IOI	EAP-IG	LD	2/10	9/45	2/10	10/45	7/10	19/45
IOI	EAP-IG	PD	2/10	10/45	2/10	11/45	7/10	20/45
IOI	ACDC	LD	–	–	–	–	–	–
IOI	ACDC	PD	–	–	–	–	–	–
IOI	Edge-SP	LD	0/10	5/45	0/10	4/45	10/10	25/45
IOI	Edge-SP	PD	0/10	5/45	0/10	8/45	10/10	29/45
Greater-Than	EAP	LD	2/8	8/28	5/8	15/28	1/8	11/28
Greater-Than	EAP	PD	1/8	2/28	1/8	8/28	1/8	11/28
Greater-Than	EAP-IG	LD	2/9	8/36	4/9	13/36	4/9	13/36
Greater-Than	EAP-IG	PD	2/9	6/36	4/9	15/36	4/9	9/36
Greater-Than	ACDC	LD	–	–	–	–	–	–
Greater-Than	ACDC	PD	–	–	–	–	–	–
Greater-Than	Edge-SP	LD	0/10	4/45	0/10	11/45	1/10	30/45
Greater-Than	Edge-SP	PD	0/10	2/45	0/10	7/45	0/10	31/45
Docstring	EAP	LD	4/7	5/21	1/7	7/21	3/7	7/21
Docstring	EAP	PD	6/7	6/21	1/7	15/21	3/7	9/21
Docstring	EAP-IG	LD	2/8	10/28	5/8	11/28	2/8	16/28
Docstring	EAP-IG	PD	4/8	10/28	5/8	9/28	5/8	20/28
Docstring	ACDC	LD	2/10	19/45	3/10	14/45	6/10	28/45
Docstring	ACDC	PD	2/10	21/45	3/10	14/45	4/10	29/45
Docstring	Edge-SP	LD	0/10	1/45	7/10	6/45	6/10	12/45
Docstring	Edge-SP	PD	3/10	8/45	7/10	35/45	9/10	32/45
Acronym	EAP	LD	0/4	1/6	0/4	2/6	0/4	3/6
Acronym	EAP	PD	0/4	1/6	0/4	3/6	4/4	4/6
Acronym	EAP-IG	LD	0/4	1/6	0/4	3/6	0/4	3/6
Acronym	EAP-IG	PD	0/4	1/6	0/4	3/6	3/4	2/6
Acronym	ACDC	LD	0/4	2/6	0/4	3/6	0/4	0/6
Acronym	ACDC	PD	2/4	1/6	0/4	2/6	1/4	5/6
Acronym	Edge-SP	LD	0/10	2/45	0/10	5/45	0/10	12/45
Acronym	Edge-SP	PD	0/10	1/45	0/10	4/45	0/10	12/45
113	EAP	LD	0/10	1/45	0/10	19/45	0/10	22/45
113	EAP	PD	0/10	1/45	0/10	7/45	0/10	18/45
113	EAP-IG	LD	10/10	0/45	10/10	0/45	10/10	0/45
113	EAP-IG	PD	0/10	0/45	10/10	0/45	10/10	0/45
113	ACDC	LD	0/10	15/45	0/10	16/45	0/10	13/45
113	ACDC	PD	0/10	3/45	0/10	12/45	0/10	10/45
113	Edge-SP	LD	10/10	11/45	10/10	11/45	10/10	11/45
113	Edge-SP	PD	0/10	5/45	7/10	14/45	10/10	26/45
97	EAP	LD	0/10	21/45	0/10	24/45	0/10	8/45
97	EAP	PD	0/10	10/45	0/10	14/45	0/10	9/45
97	EAP-IG	LD	0/10	0/45	0/10	0/45	0/10	0/45
97	EAP-IG	PD	0/10	0/45	0/10	0/45	0/10	0/45
97	ACDC	LD	0/4	2/6	0/4	2/6	0/4	1/6
97	ACDC	PD	0/4	0/6	0/4	1/6	0/4	0/6
97	Edge-SP	LD	0/10	0/45	0/10	0/45	0/10	0/45
97	Edge-SP	PD	0/10	0/45	0/10	0/45	0/10	0/45
2	EAP	LD	0/10	0/45	0/10	0/45	10/10	0/45
2	EAP	PD	0/10	0/45	0/10	0/45	10/10	0/45
2	EAP-IG	LD	0/10	0/45	0/10	0/45	10/10	0/45
2	EAP-IG	PD	0/10	0/45	0/10	0/45	10/10	0/45
2	ACDC	LD	1/6	2/15	1/6	2/15	4/6	1/15
2	ACDC	PD	0/6	0/15	4/6	1/15	4/6	4/15
2	Edge-SP	LD	0/10	29/45	0/10	29/45	10/10	30/45
2	Edge-SP	PD	0/10	9/45	0/10	32/45	10/10	25/45

Task	Method	Metric	Resampling		Mean		Zero	
			R–C	C–C	R–C	C–C	R–C	C–C
82	EAP	LD	2/10	3/45	4/10	3/45	1/10	5/45
82	EAP	PD	0/10	1/45	4/10	4/45	2/10	16/45
82	EAP-IG	LD	0/10	0/45	0/10	0/45	0/10	0/45
82	EAP-IG	PD	0/10	0/45	0/10	0/45	0/10	0/45
82	ACDC	LD	4/10	6/45	2/10	7/45	4/10	8/45
82	ACDC	PD	0/10	3/45	0/10	33/45	8/10	26/45
82	Edge-SP	LD	10/10	18/45	10/10	24/45	0/10	23/45
82	Edge-SP	PD	0/10	5/45	0/10	18/45	0/10	29/45
111	EAP	LD	5/10	17/45	5/10	10/45	5/10	12/45
111	EAP	PD	4/10	4/45	4/10	12/45	5/10	17/45
111	EAP-IG	LD	3/10	18/45	0/10	19/45	0/10	10/45
111	EAP-IG	PD	0/10	7/45	0/10	9/45	9/10	15/45
111	ACDC	LD	2/10	15/45	1/10	17/45	2/10	15/45
111	ACDC	PD	0/10	4/45	4/10	9/45	2/10	19/45
111	Edge-SP	LD	1/10	28/45	1/10	23/45	1/10	34/45
111	Edge-SP	PD	0/10	6/45	0/10	4/45	9/10	25/45
45	EAP	LD	1/10	6/45	2/10	11/45	1/10	16/45
45	EAP	PD	1/10	3/45	4/10	14/45	1/10	15/45
45	EAP-IG	LD	0/10	29/45	0/10	32/45	0/10	22/45
45	EAP-IG	PD	0/10	2/45	6/10	4/45	0/10	1/45
45	ACDC	LD	2/10	3/45	3/10	5/45	4/10	8/45
45	ACDC	PD	0/10	1/45	2/10	4/45	7/10	20/45
45	Edge-SP	LD	10/10	19/45	10/10	19/45	10/10	18/45
45	Edge-SP	PD	0/10	3/45	0/10	13/45	0/10	24/45
58	EAP	LD	0/10	7/45	0/10	6/45	9/10	6/45
58	EAP	PD	0/10	12/45	0/10	7/45	9/10	9/45
58	EAP-IG	LD	0/9	0/36	0/9	0/36	9/9	0/36
58	EAP-IG	PD	0/9	0/36	0/9	0/36	9/9	0/36
58	ACDC	LD	3/8	0/28	3/8	0/28	3/8	14/28
58	ACDC	PD	2/8	0/28	3/8	0/28	3/8	11/28
58	Edge-SP	LD	0/10	0/45	0/10	0/45	10/10	0/45
58	Edge-SP	PD	0/10	0/45	0/10	0/45	10/10	0/45
93	EAP	LD	0/10	0/45	0/10	0/45	8/10	1/45
93	EAP	PD	0/10	0/45	1/10	0/45	8/10	5/45
93	EAP-IG	LD	0/10	0/45	0/10	0/45	10/10	0/45
93	EAP-IG	PD	0/10	0/45	0/10	0/45	10/10	0/45
93	ACDC	LD	3/6	3/15	3/6	3/15	0/6	4/15
93	ACDC	PD	0/6	0/15	0/6	2/15	0/6	9/15
93	Edge-SP	LD	10/10	26/45	10/10	24/45	0/10	21/45
93	Edge-SP	PD	0/10	13/45	10/10	16/45	0/10	20/45
103	EAP	LD	0/10	0/45	0/10	1/45	7/10	8/45
103	EAP	PD	0/10	0/45	2/10	1/45	7/10	7/45
103	EAP-IG	LD	0/10	0/45	0/10	0/45	10/10	0/45
103	EAP-IG	PD	0/10	0/45	0/10	0/45	10/10	0/45
103	ACDC	LD	2/8	3/28	0/8	3/28	0/8	2/28
103	ACDC	PD	1/8	0/28	0/8	2/28	0/8	19/28
103	Edge-SP	LD	0/10	26/45	0/10	27/45	10/10	29/45
103	Edge-SP	PD	0/10	13/45	0/10	0/45	10/10	20/45
25	EAP	LD	2/10	13/45	1/10	13/45	8/10	13/45
25	EAP	PD	1/10	6/45	1/10	6/45	8/10	12/45
25	EAP-IG	LD	10/10	9/45	9/10	9/45	0/10	6/45
25	EAP-IG	PD	0/10	0/45	1/10	0/45	0/10	8/45
25	ACDC	LD	1/4	2/6	1/4	2/6	4/4	6/6
25	ACDC	PD	0/4	2/6	2/4	3/6	1/4	5/6
25	Edge-SP	LD	7/10	6/45	5/10	3/45	1/10	19/45
25	Edge-SP	PD	1/10	20/45	5/10	14/45	0/10	21/45

Table 19: Section 4 KL deficits among misranked primary pairs. Distributions are median [first quartile, third quartile]; Q gaps are in pp and score gaps are in KL units. Counts pool dependent pair records descriptively; a circuit may occur in many pairs.

Task	Intervention	Pair	Misranked	ΔQ (pp)	ΔF	$\Delta Q \geq 5$
Human suite	Resampling	R–C	23	1.33 [0.50, 2.33]	0.084 [0.057, 0.120]	0
Human suite	Resampling	C–C	93	1.00 [0.33, 2.00]	0.034 [0.012, 0.089]	4
Human suite	Mean	R–C	41	3.00 [1.33, 7.33]	0.251 [0.110, 0.496]	14
Human suite	Mean	C–C	138	1.33 [0.67, 2.67]	0.060 [0.024, 0.135]	9
Human suite	Zero	R–C	26	2.33 [0.75, 12.33]	1.286 [0.334, 3.245]	9
Human suite	Zero	C–C	170	2.00 [1.00, 4.00]	0.685 [0.207, 2.712]	36
InterpBench suite	Resampling	R–C	13	0.52 [0.33, 0.81]	0.033 [0.011, 0.051]	0
InterpBench suite	Resampling	C–C	158	0.11 [0.04, 0.19]	0.005 [0.001, 0.014]	0
InterpBench suite	Mean	R–C	42	0.59 [0.33, 0.96]	0.000 [0.000, 0.001]	0
InterpBench suite	Mean	C–C	262	0.22 [0.07, 0.70]	0.000 [0.000, 0.004]	16
InterpBench suite	Zero	R–C	158	1.22 [0.59, 2.22]	4.997 [0.973, 12.401]	6
InterpBench suite	Zero	C–C	434	0.37 [0.11, 1.37]	0.403 [0.041, 3.255]	37

Pairs are dependent when they share candidates. Pooled counts and deficit distributions therefore summarize the observed candidate pools and should not be interpreted as independent-observation estimates. Appendix E reports which failures persist across scoring metrics on the same pairs.

C.5 Faithfulness metrics and comparison between suites

Tables 20 and 21 compare KL, LD, and PD on the intersection of available pair identities within each suite. Q, circuit masks, prompts, donor assignments, and intervention are held fixed when the score changes. Greater-Than uses semantic Q throughout. Each cell reports the fraction of all common pairs that misrank and the mean Q deficit among those failures. Ties remain in the denominator.

All three metrics exhibit misranking in discovery outputs. The Human suite has higher pooled C–C KL rates than InterpBench under resampling, mean, and zero replacement. This suite comparison is descriptive: the task distributions, models, candidate pools, and tie frequencies differ. The benchmark contrast therefore does not isolate the effect of model scale or semi-synthetic task construction. Its practical implication is that the InterpBench results alone do not characterize the failures observed in pretrained-model circuits.

Table 20: KL/LD/PD on common pair identities in the Human suite. Each cell reports percentage misranking followed by mean ΔQ in percentage points for that metric, intervention, and pair type; the denominator is the total common-pair pool.

Metric	Resampling		Mean		Zero	
	R–C	C–C	R–C	C–C	R–C	C–C
KL	20.2 (1.55)	20.9 (1.55)	36.0 (4.24)	30.9 (2.05)	22.8 (10.90)	38.1 (4.28)
LD	12.3 (1.79)	17.9 (1.43)	23.7 (3.22)	28.0 (2.69)	43.9 (11.86)	45.7 (6.40)
PD	19.3 (3.46)	19.5 (1.72)	20.2 (2.93)	35.2 (4.72)	53.5 (10.16)	52.9 (7.09)

Table 21: KL/LD/PD on common pair identities in the InterpBench suite. Each cell reports percentage misranking followed by mean ΔQ in percentage points for that metric, intervention, and pair type; the denominator is the total common-pair pool.

Metric	Resampling		Mean		Zero	
	R–C	C–C	R–C	C–C	R–C	C–C
KL	3.5 (0.56)	9.8 (0.18)	11.2 (0.69)	16.2 (0.79)	42.1 (1.61)	26.8 (2.14)
LD	26.4 (1.95)	20.9 (0.59)	24.3 (1.90)	22.5 (2.09)	45.6 (1.36)	23.8 (2.53)
PD	2.7 (0.38)	8.2 (0.17)	18.7 (1.18)	15.8 (0.80)	48.5 (1.45)	27.5 (2.22)

C.6 Connection to the context-restoration study

Section 5 selects its 100 persistent misrankings from this discovery candidate pool. Selection retains the circuits, baseline intervention, evaluation prompts, original donor assignments, and independent-test Q ordering. Appendix D reports the cohort's task allocation and complete outcomes.

D Section 5: cohort eligibility, restoration search, and complete outcomes

D.1 The frozen 100-case cohort

A case is a task, circuit pair, and baseline intervention. The 100 cases are selected from the discovery candidate pool used in Section 4. Selection requires equal executable circuit size, a Q deficit of at least 0.5 pp, a strict validation KL misranking, and the same misranking on independent-test prompts before restoration. Q and score ties are ineligible. Greater-Than uses semantic accuracy. The frozen cohort contains 96 distinct underlying circuit pairs, with some pairs evaluated under more than one intervention.

The cohort contains 50 Human cases and 50 InterpBench cases, with 50 resampling, 25 mean, and 25 zero cases overall. The Human allocation is 23 R–C and 27 C–C; the InterpBench allocation is 25 R–C and 25 C–C. Selection is conditional on persistent failures and the specified quotas, rather than a random sample for estimating population repairability. Table 22 reports the task allocation.

D.2 Eligible groups and recipient-intact restoration

Let B and W be the fixed higher-Q and lower-Q circuits. Eligible sites are connections excluded by either mask, grouped by receiving node. Let $\mathcal{G}$ be the original eligible group universe and $G = |\mathcal{G}|$. Group counts and edge counts are distinct because groups can contain different numbers of excluded incoming connections.

For each recipient prompt, we cache signals from its intact-model execution. At a selected excluded connection, we replace the original donor, mean, or zero signal with the corresponding recipient-intact signal. A selected site affects a circuit only when that circuit excludes it. Retained connections continue computing normally. Masks and weights are unchanged, and all unrestored sites preserve the original intervention and donor assignments. The same subset is tested on independent recipients

Table 22: Restoration cohort by task. Entries count task–pair–intervention cases; the cohort is selected from the Section 4 discovery candidate pool. Test repair is confirmation at any level. The Human suite contributes 50/50 repairs and the InterpBench cohort 46/50.

Task	Cases	Test repairs	R–C	C–C
IOI	17	17	8	9
Greater-Than	10	10	7	3
Docstring	11	11	6	5
Acronym	12	12	2	10
15	1	1	1	0
18	4	4	0	4
25	3	2	1	2
45	3	3	3	0
56	1	0	0	1
67	2	2	2	0
71	6	4	4	2
93	1	1	1	0
97	28	28	12	16
124	1	1	1	0

using their own intact caches and their original test donor assignments. Neither the search nor restoration changes the original independent-test Q values or their ordering.

D.3 Validation-guided search at four restoration levels

Before searching, we verify that the lower-Q circuit is preferred on both validation and independent-test prompts. A restoration must reverse the score ordering beyond the scale-aware tolerance. For scores $F(B)$ and $F(W)$ on the split being evaluated, define

$$\epsilon(F(B), F(W)) = \max\{10^{-6},\ 10^{-6}\max(1, |F(B)|, |F(W)|)\}. \tag{6}$$

A corrected ordering requires $F(B) - F(W) > \epsilon(F(B), F(W))$. Test ties or reversals below this tolerance do not count as confirmed repairs.

The search visits fractions $f \in \{0.8, 0.4, 0.2, 0.1\}$ in descending order, with at most five predetermined attempts per level:

1. Set the sampling universe to the original eligible groups $\mathcal{G}$. At fraction f, the target subset size is approximately fG, with integer rounding and at least one group. Every percentage refers to the *original* group count G, not the size of the current parent subset.
2. Sample restoration subsets from the current universe and evaluate their validation rankings. If a level finds a validation-successful subset, use its ***first*** successful subset as the parent sampling universe for the next smaller level. For example, a successful 80% subset is the universe from which to sample a subset containing approximately 40% of the original groups.
3. If all five attempts at a level fail on validation, reset the universe to $\mathcal{G}$ for the next smaller fraction. Thus a failed 40% refinement is followed by 20% attempts drawn from the original universe. If a fallback succeeds, resume refinement within its successful subset.
4. Apply each qualifying validation-successful subset unchanged on independent-test prompts. Test outcomes must not determine the parent subset, subsequent sampling, refinement, or stopping. A case has a confirmed repair if at least one searched subset corrects both validation and test rankings under the required tolerance.
5. Continue through the 10% level within the budget of at most 20 validation attempts per case. Preserve every confirmed success: a failed smaller restoration never cancels a successful larger restoration.

The first validation success determines the next sampling universe; it does not invalidate another qualifying repair already found. The case-level endpoint is the existence of at least one subset that repairs both rankings. Nominal levels describe groups restored, not circuit size or necessarily the

Table 23: Validation selection and independent-test confirmation by nominal restoration level. The first validation-successful subset determines the next sampling universe; confirmation requires a subset that corrects both validation and test rankings. Test outcomes never guide the search. Effective group and site counts are median [first quartile, third quartile] among confirmed cases. Fractions are nominal group-search levels, relative to the original eligible group count, not fractions of circuit size.

Level	Validation repair	Test confirmed	Groups restored	Sites restored
80%	86/100	84/100	10.0 [10.0, 39.0]	200.5 [119.0, 24447.2]
40%	88/100	88/100	5.0 [5.0, 20.0]	111.5 [60.8, 11786.5]
20%	89/100	87/100	2.0 [2.0, 10.0]	39.0 [24.0, 4182.5]
10%	83/100	82/100	1.0 [1.0, 5.0]	15.5 [12.0, 1795.8]

exact fraction of excluded edges. Rounding, unequal group sizes, and resets after failure explain why effective site counts and repair rates need not change monotonically.

The search seeks a sufficient partial restoration within a fixed budget. It starts at 80% without requiring full restoration first. We do not test whether the selected set is globally minimal, uniquely causal, or better than random restoration. Table 23 separates validation repairs from test confirmations. At 80%, 86 validation repairs yield 84 confirmations; at 10%, 83 yield 82. Among confirmed 10% cases, the median restoration covers one receiving-node group but 15.5 sites, illustrating why group fractions cannot be read as edge fractions.

D.4 Confirmation, signed deficits, and complete outcomes

All 100 selected cases have a persistent original test misranking, and 96 have at least one independently confirmed repair. Nominal-level confirmations are 84, 88, 87, and 82 for 80%, 40%, 20%, and 10%, respectively. These successes are non-exclusive. Human cases contribute 50/50 confirmations and InterpBench cases 46/50. All four cases without confirmation remain in the relevant denominators.

A search can fail because no validation repair is found or because a validation repair does not transfer to test, including a test tie. Table 25 records confirmation for every case at each level but does not distinguish these failure reasons. Every case in the frozen cohort has an original persistent test misranking. The reported 96/100 measures success within the search budget, not the fraction of failures that could ever be repaired.

Q remains fixed. The signed before/after deficit follows score preference: $Q(\text{nonpreferred}) - Q(\text{preferred})$. It is positive before repair and negative after a confirmed correction. Baseline means include all selected cases; after-repair means include confirmed cases once each, regardless of how many levels succeed. Consequently, the overall means, $+10.33$ pp before and -9.30 pp after, use different conditioning sets. On the same 96 repaired cases, the baseline mean is $+9.30$ pp. No after-repair value is imputed for the four failures.

Table 24 gives the exact counts behind Table 6. Table 26 shows that all 50 resampling cases are repaired, versus 23/25 under both mean and zero replacement. Repairs occur in both pair types (46/48 R–C and 50/52 C–C); all four unconfirmed cases belong to InterpBench. The any-level endpoint counts the union of independently confirmed restorations; it is not a single preselected-level endpoint.

Table 24: Independent-test-confirmed repairs as n/N, with the same rows and denominators as Table 6. Columns 80%–10% denote nominal restoration levels; successes can overlap across levels. Any level counts each case once. Mean ΔQ is shown as before (after) in pp, with signed deficit $Q(\text{nonpreferred}) - Q(\text{preferred})$. Before means include all selected cases; after means include only confirmed repairs, once per case. The underlying Q values remain fixed.

Intervention	Pair	Mean ΔQ (pp) before (after)	Confirmed repairs (n/N)				
			80%	40%	20%	10%	Any level
Human suite (50 cases)							
Resampling	R–C	1.50 (−1.50)	2/2	2/2	1/2	1/2	2/2
	C–C	4.67 (−4.67)	20/23	22/23	20/23	21/23	23/23
Mean	R–C	14.21 (−14.21)	10/10	10/10	8/10	8/10	10/10
	C–C	25.96 (−25.96)	2/3	3/3	3/3	3/3	3/3
Zero	R–C	19.18 (−19.18)	10/11	10/11	11/11	9/11	11/11
	C–C	40.00 (−40.00)	1/1	1/1	1/1	1/1	1/1
InterpBench (50 cases)							
Resampling	R–C	4.19 (−4.19)	11/12	12/12	11/12	11/12	12/12
	C–C	2.42 (−2.42)	11/13	13/13	13/13	10/13	13/13
Mean	R–C	1.90 (−2.32)	4/6	3/6	3/6	2/6	4/6
	C–C	0.94 (−0.94)	5/6	5/6	6/6	6/6	6/6
Zero	R–C	27.96 (−27.96)	7/7	5/7	6/7	6/7	7/7
	C–C	26.23 (−4.82)	1/6	2/6	4/6	4/6	4/6
All 100 cases		10.33 (−9.30)	84/100	88/100	87/100	82/100	96/100

Table 25: Complete Section 5 per-case catalog. Gaps use the same score-preference orientation as Table 6; Q values remain fixed. Y/N indicate independent-test confirmation. A dash in the after-repair column denotes no confirmed repair, rather than a negative gap for a failed case.

Task	Intervention	Pair type	Baseline ΔQ (pp)	After-repair ΔQ (pp)	80	40	20	10
Human suite								
Acronym	Mean	R–C	15.44	-15.44	Y	Y	Y	Y
Acronym	Resampling	C–C	11.33	-11.33	Y	Y	Y	Y
Acronym	Resampling	C–C	0.56	-0.56	Y	Y	Y	Y
Acronym	Resampling	C–C	12.33	-12.33	Y	Y	Y	Y
Acronym	Resampling	C–C	7.44	-7.44	Y	Y	N	Y
Acronym	Resampling	C–C	8.78	-8.78	Y	Y	Y	Y
Acronym	Resampling	C–C	1.22	-1.22	Y	Y	Y	Y
Acronym	Mean	C–C	69.56	-69.56	Y	Y	Y	Y
Acronym	Resampling	C–C	0.67	-0.67	Y	Y	Y	Y
Acronym	Resampling	C–C	0.56	-0.56	Y	Y	Y	Y
Acronym	Resampling	C–C	10.56	-10.56	Y	Y	N	N
Acronym	Resampling	R–C	2.33	-2.33	Y	Y	N	N
Docstring	Mean	R–C	7.33	-7.33	Y	Y	Y	Y
Docstring	Zero	R–C	7.33	-7.33	Y	Y	Y	Y
Docstring	Resampling	R–C	0.67	-0.67	Y	Y	Y	Y
Docstring	Mean	C–C	1.00	-1.00	N	Y	Y	Y
Docstring	Mean	R–C	6.33	-6.33	Y	Y	Y	Y
Docstring	Resampling	C–C	1.33	-1.33	Y	Y	Y	Y
Docstring	Resampling	C–C	0.67	-0.67	N	N	Y	Y
Docstring	Resampling	C–C	1.33	-1.33	Y	Y	Y	Y
Docstring	Resampling	C–C	2.00	-2.00	Y	Y	Y	Y
Docstring	Zero	R–C	6.33	-6.33	Y	Y	Y	Y
Docstring	Zero	R–C	3.00	-3.00	Y	Y	Y	Y
Greater-Than	Zero	R–C	41.00	-41.00	N	N	Y	N

Task	Intervention	Pair type	Baseline ΔQ (pp)	After-repair ΔQ (pp)	80	40	20	10
Greater-Than	Zero	R–C	9.33	-9.33	Y	Y	Y	Y
Greater-Than	Zero	R–C	41.67	-41.67	Y	Y	Y	Y
Greater-Than	Resampling	C–C	0.67	-0.67	N	Y	Y	Y
Greater-Than	Zero	R–C	42.00	-42.00	Y	Y	Y	Y
Greater-Than	Zero	C–C	40.00	-40.00	Y	Y	Y	Y
Greater-Than	Resampling	C–C	1.00	-1.00	Y	Y	Y	Y
Greater-Than	Zero	R–C	9.00	-9.00	Y	Y	Y	Y
Greater-Than	Zero	R–C	10.00	-10.00	Y	Y	Y	Y
Greater-Than	Zero	R–C	40.67	-40.67	Y	Y	Y	N
IOI	Mean	R–C	29.67	-29.67	Y	Y	Y	Y
IOI	Mean	R–C	2.00	-2.00	Y	Y	Y	Y
IOI	Resampling	C–C	0.67	-0.67	Y	Y	Y	Y
IOI	Resampling	C–C	20.67	-20.67	Y	Y	Y	Y
IOI	Mean	C–C	7.33	-7.33	Y	Y	Y	Y
IOI	Mean	R–C	40.00	-40.00	Y	Y	N	N
IOI	Zero	R–C	0.67	-0.67	Y	Y	Y	Y
IOI	Resampling	C–C	3.00	-3.00	N	Y	Y	Y
IOI	Mean	R–C	8.00	-8.00	Y	Y	Y	Y
IOI	Mean	R–C	5.00	-5.00	Y	Y	Y	Y
IOI	Mean	R–C	11.00	-11.00	Y	Y	Y	Y
IOI	Resampling	C–C	2.00	-2.00	Y	Y	N	N
IOI	Resampling	C–C	0.67	-0.67	Y	Y	Y	Y
IOI	Resampling	C–C	11.33	-11.33	Y	Y	Y	Y
IOI	Resampling	C–C	0.67	-0.67	Y	Y	Y	Y
IOI	Resampling	C–C	8.00	-8.00	Y	Y	Y	Y
IOI	Mean	R–C	17.33	-17.33	Y	Y	N	N
InterpBench suite								
124	Zero	R–C	47.00	-47.00	Y	Y	Y	Y
15	Zero	R–C	61.75	-61.75	Y	Y	Y	Y
18	Resampling	C–C	2.22	-2.22	Y	Y	Y	N
18	Resampling	C–C	5.07	-5.07	Y	Y	Y	Y
18	Resampling	C–C	5.07	-5.07	N	Y	Y	N
18	Resampling	C–C	1.41	-1.41	Y	Y	Y	Y
25	Zero	R–C	76.81	-76.81	Y	N	N	N
25	Zero	C–C	76.93	–	N	N	N	N
25	Zero	C–C	10.07	-10.07	Y	Y	Y	Y
45	Mean	R–C	0.85	-0.85	Y	Y	N	N
45	Mean	R–C	0.85	-0.85	Y	Y	Y	Y
45	Zero	R–C	0.85	-0.85	Y	Y	Y	Y
56	Zero	C–C	61.15	–	N	N	N	N
67	Mean	R–C	6.56	-6.56	Y	N	Y	N
67	Resampling	R–C	6.56	-6.56	Y	Y	Y	Y
71	Resampling	C–C	0.63	-0.63	N	Y	Y	Y
71	Mean	R–C	1.04	–	N	N	N	N
71	Mean	C–C	0.63	-0.63	Y	Y	Y	Y
71	Mean	R–C	1.04	–	N	N	N	N
71	Mean	R–C	1.04	-1.04	Y	Y	Y	Y
71	Zero	R–C	1.04	-1.04	Y	Y	Y	Y
93	Zero	R–C	4.33	-4.33	Y	N	Y	Y
97	Resampling	R–C	4.78	-4.78	Y	Y	Y	Y
97	Resampling	C–C	0.56	-0.56	Y	Y	Y	Y
97	Mean	C–C	0.56	-0.56	Y	Y	Y	Y
97	Resampling	R–C	4.07	-4.07	Y	Y	Y	Y
97	Mean	C–C	0.52	-0.52	Y	Y	Y	Y
97	Resampling	R–C	4.81	-4.81	Y	Y	Y	Y
97	Zero	R–C	3.96	-3.96	Y	Y	Y	Y
97	Resampling	C–C	3.30	-3.30	Y	Y	Y	Y
97	Resampling	R–C	3.96	-3.96	N	Y	Y	Y
97	Resampling	C–C	4.78	-4.78	Y	Y	Y	Y
97	Resampling	C–C	2.04	-2.04	Y	Y	Y	Y
97	Zero	C–C	4.81	-4.81	N	Y	Y	Y
97	Resampling	R–C	2.07	-2.07	Y	Y	Y	Y

Task	Intervention	Pair type	Baseline ΔQ (pp)	After-repair ΔQ (pp)	80	40	20	10
97	Resampling	R–C	3.30	-3.30	Y	Y	Y	N
97	Resampling	R–C	4.81	-4.81	Y	Y	N	Y
97	Resampling	R–C	3.30	-3.30	Y	Y	Y	Y
97	Resampling	C–C	0.56	-0.56	Y	Y	Y	Y
97	Resampling	C–C	2.07	-2.07	Y	Y	Y	N
97	Resampling	C–C	1.74	-1.74	Y	Y	Y	Y
97	Resampling	R–C	4.07	-4.07	Y	Y	Y	Y
97	Mean	C–C	1.33	-1.33	Y	Y	Y	Y
97	Resampling	R–C	4.48	-4.48	Y	Y	Y	Y
97	Resampling	R–C	4.07	-4.07	Y	Y	Y	Y
97	Mean	C–C	0.56	-0.56	Y	Y	Y	Y
97	Mean	C–C	2.04	-2.04	N	N	Y	Y
97	Zero	C–C	1.67	-1.67	N	N	Y	Y
97	Resampling	C–C	2.04	-2.04	Y	Y	Y	Y
97	Zero	C–C	2.74	-2.74	N	N	Y	Y

Table 26: Section 5 100-case cohort breakdown. Entries are confirmed successes / selected cases; levels are non-exclusive.

Stratum	Any level	80%	40%	20%	10%
Human suite	50/50	45/50	48/50	44/50	43/50
InterpBench suite	46/50	39/50	40/50	43/50	39/50
Resampling	50/50	44/50	49/50	45/50	43/50
Mean	23/25	21/25	21/25	20/25	19/25
Zero	23/25	19/25	18/25	22/25	20/25
R–C	46/48	44/48	42/48	40/48	37/48
C–C	50/52	40/52	46/52	47/52	45/52

E SAME-PAIR METRIC OVERLAP AND AGREEMENT DEFICITS

For each task, method, intervention, and pair type, we join KL, LD, and PD by exact circuit identities and require identical Q values. The common pool joins identical pair identities across all three metrics; ties remain in each metric's denominator and count as non-misranking. For each common-pool pair we report one of eight disjoint outcomes: all three misrank; KL+LD, KL+PD, or LD+PD only; KL, LD, or PD only; or none. "None" means no metric strictly misranks; it includes Q ties and score ties as well as correct strict rankings. Ties remain in the total denominator and are distinguished from correct rankings in the analysis. Table 27 distinguishes shared failures from metric-specific ones. In the Human discovery C–C pool, 128 records fail under all three metrics, while 156 fail only under KL. Table 28 further separates ties from correct rankings: 1,899 of the 3,267 InterpBench discovery C–C records in "None" are Q ties, so absence of misranking does not always demonstrate successful discrimination.

Table 29 shows one shared-failure example per Human task, selected by a fixed lexicographic rule rather than deficit size. All three scores favor the worse candidate, with Q deficits ranging from 2 to 8.778 pp. The overlap calculation matches records by task, method, intervention, and circuit pair before comparing the three metrics. Table 30 reports Q deficits and available joint uncertainty support. Among the 597 Human controlled C–C records misranked by all three metrics, 271 have deficits of at least 5 pp. Agreement between metrics therefore does not ensure a behaviorally sound preference; incomplete uncertainty records are distinguished from unsupported failures. A pair can recur under multiple interventions, so the pooled entries are dependent. Shared failures indicate agreement between the metrics on the same wrong preference.

Table 27: Same-pair metric overlap in the total common primary pool, including ties, pooled across interventions and methods in Section 4. Columns name exactly the metrics that misrank. None includes ties as well as correct strict rankings. Counts are dependent descriptive records.

Task / pair	Pool	All 3	KL+LD	KL+PD	LD+PD	KL	LD	PD	None
S3 Human suite R–C	630	15	10	12	58	62	5	79	389
S3 Human suite C–C	16635	597	332	286	1566	1652	414	2335	9453
S3 InterpBench suite R–C	1770	60	20	54	12	97	65	12	1450
S3 InterpBench suite C–C	51465	1784	673	1595	351	3595	1765	326	41376
S4 Human suite R–C	342	22	8	6	50	54	11	28	163
S4 Human suite C–C	1338	128	63	54	151	156	67	147	572
S4 InterpBench suite R–C	1125	143	5	50	60	15	153	9	690
S4 InterpBench suite C–C	4857	356	52	318	104	128	576	56	3267

Table 28: Decomposition of the none-misrank category in Table 27. Q ties are counted first; score ties then count pairs with non-tied Q and at least one tied score. The remaining pairs are correctly strictly ranked by all three metrics. These mutually exclusive counts sum to None.

Task / pair	None	Q ties	Score ties	All three correct
S3 Human suite R–C	389	66	0	323
S3 Human suite C–C	9453	1023	0	8430
S3 InterpBench suite R–C	1450	240	4	1206
S3 InterpBench suite C–C	41376	6069	73	35234
S4 Human suite R–C	163	15	0	148
S4 Human suite C–C	572	66	0	506
S4 InterpBench suite R–C	690	69	0	621
S4 InterpBench suite C–C	3267	1899	7	1361

Table 29: Shared-failure examples: lexicographically first all-three failure per Human suite task (no gap filter). Score gaps favor the worse-Q candidate. Q gaps are in percentage points.

Task	Intervention	Pair	KL gap	LD gap	PD gap	Q gap
IOI	mean	k193_s810 / k193_s816	0.1192	0.1585	0.0006241	2
Greater-Than	mean	k12_s811 / k12_s813	0.07244	0.03427	0.01234	2
docstring	mean	k12_s811 / k12_s814	0.06245	0.2988	0.027	2
acronym	mean	k2_s818 / k6_s812	0.02785	0.4786	0.1154	8.778

Table 30: Agreement deficits among all-three-metric failures. Joint support: supported / available saved paired-support records, requiring support for all three metrics. A dash means no complete joint-uncertainty record; it does not mean zero supported failures.

Panel / pair	Shared	Median [Q1, Q3] (pp)	$\geq$ 5 pp	Joint support
S3 Human suite R–C	15	2.33 [0.50, 4.67]	4	2/13
S3 Human suite C–C	597	3.67 [1.00, 18.67]	271	152/472
S3 InterpBench suite R–C	60	2.19 [0.16, 7.15]	25	38/60
S3 InterpBench suite C–C	1784	4.31 [0.74, 6.97]	799	1158/1784
S4 Human suite R–C	22	1.00 [0.33, 4.83]	6	5/16
S4 Human suite C–C	128	1.00 [0.67, 2.33]	12	13/106
S4 InterpBench suite R–C	143	1.00 [0.48, 1.52]	1	97/143
S4 InterpBench suite C–C	356	0.19 [0.07, 0.74]	14	78/356