FAITHFULNESS ABOVE ONE: HOW A CIRCUIT BENCHMARK REWARDS LEAVING OUT SUPPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

MIB, the Mechanistic Interpretability Benchmark, ranks circuit-discovery methods by two areas of a faithfulness curve. On its grid, the higher-is-better CPR is a constant plus the overshoot of the curve above the full model minus its undershoot below it, and the lower-is-better CMD is their sum, so every published pair can be checked, and CPR rewards circuits for exceeding the model they are meant to explain. MIB computes CPR on circuits of the highest-scoring edges. For additive edge effects and exact scores this rule maximizes CPR, and its most faithful circuits contain no suppressive component. Among MIB's own entries, the leading CPR in 9 of its 11 model and task cells can only come from a curve above the model, and on GPT-2 small the CPR spread of five public submissions is overshoot. Removing the logit edges of the two negative name mover heads, part of the known IOI circuit, from the MIB example's magnitude-selected circuits makes them less complete and raises their CPR from 0.97 to 1.58, above the 1.24 of the circuits whose CPR MIB reports. Rescored under the ABC counterfactual, EAP, the MIB example's method, gives value-selected circuits that exceed the model at the sizes that carry most of the weight when evaluated under the answer flip that MIB runs and fall below it under ABC. The largest overshoot, MAttr's, is not explained by the two edges and arises in patched states far outside the model's range. We recommend computing both scores on one declared circuit family, with a per-example distance from the model as the headline behavioral score.

1 INTRODUCTION

Circuit discovery asks which subgraph of a neural network's computation graph carries out a given behavior, such as indirect object identification, IOI (Wang et al., 2023; Conmy et al., 2023). Methods differ mainly in how they score edges and how they turn scores into a subgraph (Syed et al., 2024; Hanna et al., 2024; Bhaskar et al., 2024), and they are now compared on the circuit-localization track of the Mechanistic Interpretability Benchmark, MIB (Mueller et al., 2025). Its leaderboard hosted the BlackboxNLP 2025 shared task (Arad et al., 2025), and method papers report its scores as headline evidence (Edin et al., 2026; Shao et al., 2026; Arora et al., 2026) or in appendices (Zhang & Wang, 2026).

MIB scores a method through a faithfulness curve $f(k)$, the share of the model's task behavior that the method's circuit of size k recovers, and reports two areas of it, the integrated circuit performance ratio CPR, higher is better, and the integrated circuit-model distance CMD, lower is better. The benchmark introduces them for two goals it calls "valid but complementary", circuits that contribute to higher task performance and circuits that capture every component with a measurable effect. It builds a separate circuit for each goal from the same edge scores, the highest-scoring edges for CPR and the largest-magnitude edges for CMD (Mueller et al., 2025, Section 3.1 and Appendix E.1). We ask what a high CPR says about the circuit it is computed on, and what the rule that builds that circuit optimizes.

Our answer is that the CPR column rewards circuits for exceeding the model they are meant to explain, and that its selection rule earns this reward by leaving out the components that work against the task. The benchmark's first goal invites such a reward. Under the idealizations of Section 4, the model's own suppression and the counterfactual cap it, and a method comes closer to the cap the more exactly its positive scores pick out the edges that support the task. With U the undershoot, the

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.

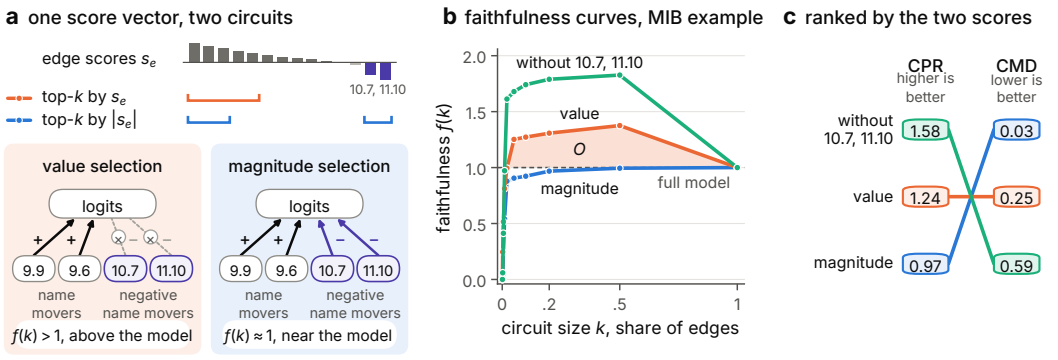


Figure 1: **The negative name mover edges under the two selection rules, GPT-2 small on IOI.** **a**, Schematic. From one score vector the benchmark builds its CPR circuit from the highest scores and its CMD circuit from the largest magnitudes, so only the second keeps the negatively scored logit edges of the negative name movers 10.7 and 11.10. **b**, The MIB example's curves under value selection, magnitude selection and magnitude selection without the two edges, with the overshoot of the value-selected curve shaded on the linear axis on which the benchmark integrates. **c**, The three families ranked by each score.

area where the circuit recovers less than the model, and O the overshoot, the area where it recovers more, $\text{CPR} = W + O - U$ and $\text{CMD} = O + U$ on the benchmark's grid of width W. Figure 1 shows the consequence for the benchmark's own example submission, which we call the MIB example. Removing the logit edges of the negative name mover heads 10.7 and 11.10, which Wang et al. (2023) count as part of the mechanism, from every circuit of its magnitude-selected family raises its CPR from 0.97 to 1.58, above the 1.24 of the value-selected family whose CPR the benchmark reports.

Parts of this picture were known. Conmy et al. (2023, Appendices E.1 and F.3) found that optimizing the logit difference did not recover the negative name movers, and preferred KL divergence because a task metric can be over-optimized until the circuit performs the task more than the model. Hanna et al. (2024) noted that selecting the highest signed scores would yield a better-performing circuit but discard components that matter through their negative effect, and selected edges by magnitude. A user of the benchmark expected the two areas of one curve to add up to at least one, and a maintainer replied that the two scores come from different circuits and need not add up to one unless all scores share a sign.[1] Nikankin et al. (2025) changed their shared-task pipeline for CPR because negatively scored edges are penalized there, and named the negative name mover heads as negatively scored components. Concurrently, Arora et al. (2026, Section 4) argue that because the logit difference is unbounded, faithfulness far above one at some sizes can average out faithfulness below one at others, and that an area taken on a linear axis in k lets large circuits dominate CPR, and they propose a compactness score based on interchange intervention accuracy. Their first point is the qualitative form of the CPR identity in Lemma 1. What was missing is an exact account and a measurement of what the known design of the two scores does to the benchmark's results: how many published leaders exceed the model, how much of a reported spread is overshoot, which part of a known mechanism a higher score leaves out, where the largest overshoot arises, and whether the answer depends on the counterfactual. Our contributions are the following.

- An exact decomposition of both scores into overshoot and undershoot in Section 3, which turns every published pair into a check and holds for any score normalized by the model's effect, and a theory of the two selection rules in Section 4, under which value selection rewards scores that separate supporting from suppressive edges and then scores the circuit without the suppressive ones.
- A model-free check of all 11 published cells in Section 5 and a re-evaluation of five public submissions on GPT-2 small in Section 6. The leading CPR can only come from a curve above the model in 9 cells among MIB's entries and in 9 of 9 among the shared task's, and on GPT-2 the reported CPR column orders submissions by overshoot, which arises from left-out suppression and, for the largest, in patched states far outside the model's range.

[1] https://github.com/hannamw/MIB-circuit-track/issues/19, July and August 2025.

- A causal test in Section 7 on the IOI circuit, with predictions written down before the runs. Removing the logit edges of the two negative name mover heads makes magnitude-selected circuits exceed the model, and removing them from the full model lifts it above one under both counterfactuals and on a second model, and the most negatively scored edges act the same way on a second task. Rescoring EAP under each counterfactual shows that the evaluation's counterfactual sets the side of one at the sizes that carry most of the weight.

2 THE CIRCUIT-LOCALIZATION TRACK

A model's computation graph has nodes for the input, attention heads, MLPs and the logits, and one edge for each path from an upstream output to a downstream input. A method assigns each edge e a real score s_e. For attribution methods such as EAP-IG (Hanna et al., 2024), a positive score estimates that the edge supports the behavior and a negative score that it works against it. The benchmark turns scores into a circuit of size k, a share of all E edges, by one of two rules,

$$\mathcal{C}_k^{\text{val}} = \text{top}_{\lfloor kE \rfloor}\{s_e\}, \qquad \mathcal{C}_k^{\text{mag}} = \text{top}_{\lfloor kE \rfloor}\{|s_e|\}, \tag{1}$$

followed by pruning of edges that no longer connect the input to the logits. We call the first value selection and the second magnitude selection. A sequence of circuits built by one rule from one score vector is a circuit family. Edges outside the circuit carry their activations on a counterfactual input, and faithfulness at size k is

$$f(k) = \frac{m(\mathcal{C}_k) - m(\emptyset)}{m(\mathcal{M}) - m(\emptyset)}, \tag{2}$$

where m is the task metric averaged over examples, here the logit difference between the correct and the incorrect answer, $\mathcal{M}$ is the full model and $\emptyset$ the empty circuit. The benchmark evaluates f on the grid $\mathcal{K} = \{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$ and integrates with the trapezoid rule. With weights $w_i > 0$ that sum to the grid width $W = 0.999$,

$$\text{CPR} = \sum_{i=1}^{10} w_i \, f(k_i), \qquad \text{CMD} = \sum_{i=1}^{10} w_i \, \big|1 - f(k_i)\big|. \tag{3}$$

At $k = 1$ the circuit is the full model and $f(1) = 1$, so the endpoint adds the constant $w_{10} = 0.25$ to every CPR and nothing to any CMD. The sizes 0.2 and 0.5 carry 80% of the remaining weight.

Stated intent and released code. MIB states that CPR "prioritizes methods that locate components that improve model performance" and CMD those "with any effect on model performance, including negative" (Mueller et al., 2025, Section 3.1), and assigns the rules accordingly (Mueller et al., 2025, Appendix E.1). Its EAP-IG library defaults to magnitude selection, and so did its evaluation script until May 2025, which has defaulted to value selection since, as Section I details. The leaderboard shows each score in its own tab, sorted by its mean over tasks and models, next to a mean taken after a sigmoid (Mueller et al., 2025, Appendix D.6) that its code centers at one, so overshoot enters the default order uncompressed. In the GitHub issue cited in the introduction, a maintainer wrote that the leaderboard would use different rules for the two scores "to give a more optimistic picture of each method's performance", and the user who opened the issue replied that for NAP the pairing lowers CPR, from 0.62 to 0.28. For IOI, the circuit track replaces the subject's second mention with the indirect object, which flips the answer (Mueller et al., 2025, Appendix D.1), as the released code does by default, while an appendix of the preprint names the ABC counterfactual (Wang et al., 2023), which replaces it with a third name (Mueller et al., 2025, Appendix C.1 of arXiv:2504.13151v2). Scores of edge attribution patching, EAP (Syed et al., 2024), that we computed with the benchmark's attribution settings agree with the MIB example's at Pearson $r = 0.998$ and Spearman $\rho = 0.78$ and share its most negatively scored edges, and the benchmark's stored EAP curve matches our evaluation of the MIB example within 0.06 at every size, so we take the MIB example to be EAP. We analyze the benchmark as run and repeat the key tests under ABC.

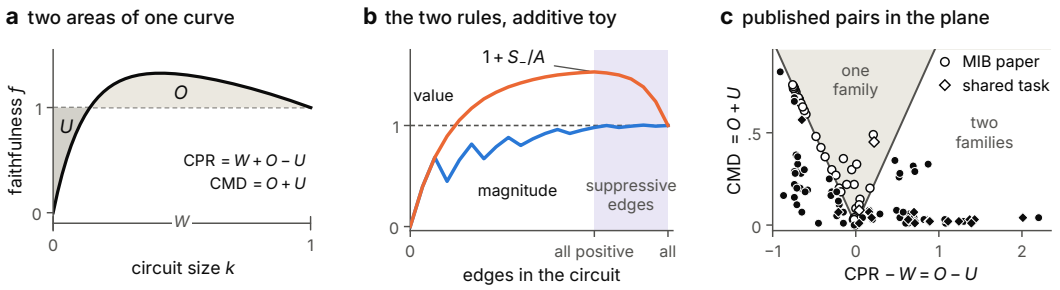


Figure 2: **Two areas behind both scores.** **a**, Schematic. A curve accrues undershoot U below the model and overshoot O above it, and, by Lemma 1, CPR adds O while CMD penalizes it. **b**, A toy example of Proposition 1 with hypothetical additive effects. The value-selected family peaks at $1+S_-/A$ once it contains every positive edge, while the magnitude-selected family rises toward one. **c**, The 97 published MIB pairs ○ and the 20 shared-task pairs ◇ in the plane of the two areas. A pair read off one family lies in the shaded cone $|\text{CPR} - W| \le \text{CMD}$, and filled markers lie outside it by more than rounding error.

3 Two Areas Behind Both Scores

Where $f < 1$ a circuit recovers less of the behavior than the model, and where $f > 1$ it produces more. We call the two weighted areas of Figure 2a the undershoot and the overshoot,

$$U = \sum_{i=1}^{10} w_i \left(1 - f(k_i)\right)_+, \qquad O = \sum_{i=1}^{10} w_i \left(f(k_i) - 1\right)_+. \tag{4}$$

Lemma 1. *For one circuit family,* $\text{CPR} = W + O - U$ *and* $\text{CMD} = O + U$. *Hence* $|\text{CPR} - W| \le \text{CMD}$, *with equality if and only if* $O = 0$ *or* $U = 0$.

Lemma 2. *Let two families have areas* (O_a, U_a) *and* (O_b, U_b), *and write* $\delta_O = O_a - O_b$ *and* $\delta_U = U_a - U_b$. *If* $|\delta_U| > |\delta_O|$, CPR *and* CMD *both prefer the family with the smaller undershoot. If* $|\delta_O| > |\delta_U|$, CPR *prefers the family with the larger overshoot and* CMD *the other one. If* $|\delta_O| = |\delta_U| > 0$, *exactly one of the two scores ties. Over* n *families without such ties, Kendall's* τ *between the two orderings is* $(N_U - N_O)/\binom{n}{2}$, *where* N_U *and* N_O *count the pairs separated more by undershoot and more by overshoot.*

Both follow from $x = x_+ - (-x)_+$ and $|x| = x_+ + (-x)_+$ summed with positive weights, and Section A gives the details. The two scores are the diagonal coordinates of one point in the plane of the two areas, as Figure 2c shows. Both penalize undershoot, and they disagree about overshoot. Three consequences hold for any reported numbers.

Corollary 1. *A reported pair with* $|\text{CPR} - W| > \text{CMD}$ *comes from two families, since one family would need a negative area. A reported* CPR *alone certifies* $O \ge \text{CPR} - W$ *when it exceeds* W *and* $U \ge W - \text{CPR}$ *when it falls short. For a reported pair taken from two families, the* CMD *family's own* CPR *lies within* $W \pm \text{CMD}$ *and the* CPR *family's own* CMD *is at least* $|\text{CPR} - W|$. *For values rounded to two decimals each bound loosens by* 0.005 *per value.*

4 Theory of the Two Selection Rules

The regimes below come from the rules, not from Lemma 1, and Section A proves the three results that follow.

Additive effects. Let every edge have a fixed effect a_e and let the metric add, $m(\mathcal{C}) = m(\emptyset) + \sum_{e \in \mathcal{C}} a_e$, with net effect $A = \sum_e a_e > 0$, so that $f(\mathcal{C}) = \sum_{e \in \mathcal{C}} a_e / A$. Write $S_- = \sum_{a_e < 0} |a_e|$ for the total effect of the edges that work against the task, and let the scores be exact up to a positive factor. Figure 2b shows a toy example.

Proposition 1. *Under additive effects and exact scores, before pruning,* ***a*** *at every size the value-selected circuit is the most faithful circuit with as many edges, so value selection maximizes* CPR *over*

all selection rules. ***b*** *Along the value-selected family f rises until every positive edge is included and falls afterwards, with maximum $1 + S_-/A$. When $S_- > 0$ it stays above one from the first size at which it exceeds one up to the last edge, so the family overshoots if and only if the model has an edge that works against the task and the grid reaches that range.* ***c*** *The magnitude-selected circuit minimizes the omitted absolute effect $\sum_{e \notin \mathcal{C}} |a_e|$, an upper bound on $A\,|1 - f(\mathcal{C})|$, but need not minimize $|1 - f|$.* ***d*** *For any circuit, and so under magnitude selection, $A\,(1 - f)$ is the net signed effect of the omitted edges, so omitted effects of opposite sign cancel in* CMD.

Monotone effects. Additivity is too strong for real circuits. A weaker assumption allows interactions: the edges split into a set $\mathcal{P}$ and a set $\mathcal{N}$ such that adding an edge of $\mathcal{P}$ to any circuit never lowers m and adding an edge of $\mathcal{N}$ never raises it, and a method scores exactly the edges of $\mathcal{P}$ positively.

Proposition 2. *Under monotone effects and sign-correct scores,* ***a*** *no circuit is more faithful than $\mathcal{P}$, and $f(\mathcal{P}) \geq 1$.* ***b*** *The value-selected family rises to $f(\mathcal{P})$ and falls afterwards, and every family's* CPR *is at most $w_{10} + (W - w_{10})\, f(\mathcal{P})$.* ***c*** *Removing negatively scored edges from a circuit never lowers its faithfulness.*

The ceiling of CPR is thus set by the model's own suppression, a property of the model, the task and the counterfactual, and the circuit that attains it is the one without suppressive components. Part c is testable, and Section 7 tests it. The last fact holds without any assumption on how edges combine.

Proposition 3. *If a method scores at least $\lfloor kE \rfloor$ edges positively, its value-selected circuit of size k contains no negatively scored edge, whatever values its negative scores take. When at least half of the edges are scored positively, the method's* CPR *under value selection is therefore unchanged by any change to its negative scores that keeps them negative.*

For four of the five submissions below, whose scores are about half positive, this holds up to $k = 0.2$, and at $k = 0.5$ negatively scored edges make up at most 2.1% of their value-selected circuits. For MAttr, whose scores are 80% negative, it holds up to $k = 0.1$. At $k = 0.5$, the size that carries the most weight, the value-selected circuit of such a method is thus close to the set of edges it scores positively, which by Proposition 2 is most faithful when it equals $\mathcal{P}$. CPR under value selection therefore rewards scores whose signs separate the edges that support the task from those that work against it, a fair reading of the benchmark's first goal, and credits the circuit without the second kind. A method whose signs err loses supporting edges and undershoots. Under additive effects, a circuit's overshoot $A(f - 1)$ is at most the suppression it leaves out.

5 Model-free checks of the published cells

Corollary 1 applies to every published number. We transcribed the CPR and CMD tables of the MIB paper for counterfactual ablations on the public test split, all models and tasks, and the pairs of the new entries in the shared-task report, which are computed on the private test split (Arad et al., 2025, Section 3.2). Table 1 summarizes each of the 11 cells, and Figure 2c shows the pairs.

Overshoot among leading entries. In 10 of the 11 cells the entry with the highest published CPR certifies overshoot, by up to $O \geq 2.19$ and in 8 cells by at least 0.1. Within one split, MIB's leading entry on the public split does so in 9 cells and the shared task's on the private split in 9 of its 9. On IOI with GPT-2 small, 10 entries report CMD ≤ 0.03, so the circuits behind their CMD track the model, yet they report CPR from 0.97 to 2.43, where one circuit per entry would confine them to 0.96 to 1.04, and 9 certify overshoot, as do 27 of 35 such entries over 7 cells.

Undershoot among weaker entries. Of the published CPR values, 59 certify undershoot, 32 of them by at least 0.5, and 10 of the 11 cells contain both kinds, so a published CPR column orders entries by overshoot minus undershoot. For 35 entries with two circuits the value-selected circuit certifiably scores a lower CPR than the magnitude-selected one, by up to 0.69, as in the case the benchmark's user reported.

Two circuits per row. Of the 97 MIB pairs in Section C, 58 lie outside the cone of Lemma 1 by more than rounding error, 24 with a CPR too high for their CMD and 34 too low, as do 18 of the 20

Table 1: **MIB's cells from published numbers alone.** The leading entry has the highest CPR, and $O \geq$ gives a lower bound on its overshoot. For the entries with CMD ≤ 0.05, the table gives their number, the range of the CPR they report, the range one circuit per entry would confine it to, and how many certify overshoot. Below counts entries with two circuits whose value-selected circuit certifiably scores a lower CPR than their magnitude-selected one. ILP is ILP+PNR+Boot, shared-task entries are scored on the private test split and MIB's on the public one, and Section C gives the sources. *On IOI-Qwen, Hybrid-Ens ties with P-Ens.

	Leading entry			Entries with CMD ≤ 0.05				
Cell	Method	CPR	$O \geq$	Number	Reported CPR	One circuit	Certified	Below
IOI-GPT2	Hybrid-Ens	2.43	1.42	10	0.97 to 2.43	0.96 to 1.04	9	2
IOI-Qwen	P-Ens*	1.88	0.87	7	0.98 to 1.88	0.96 to 1.04	6	6
IOI-Gemma	EAP-IG-inp	3.20	2.19	3	2.07 to 3.20	0.95 to 1.05	3	2
IOI-Llama	ILP	2.39	1.38	5	0.85 to 2.39	0.95 to 1.05	3	3
Arith-Llama	ILP	1.04	0.03	4	0.55 to 1.04	0.98 to 1.02	1	5
MCQA-Qwen	Hybrid-Ens	1.19	0.18	2	0.77 to 1.19	0.94 to 1.06	1	7
MCQA-Gemma	NAP-IG (CF)	1.71	0.70	0	–	–	–	1
MCQA-Llama	NAP-IG (CF)	1.87	0.86	0	–	–	–	2
ARCE-Gemma	EAP-IG-act	1.70	0.69	4	1.08 to 1.70	0.95 to 1.05	4	1
ARCE-Llama	EAP-IG-inp	1.04	0.03	0	–	–	–	4
ARCC-Llama	EAP-IG-inp	0.98	–	0	–	–	–	2

Table 2: **Both scores and both areas under both rules, GPT-2 small, IOI.** By Lemma 1, CPR $= W + O - U$ and CMD $= O + U$ within each rule. Shaded, the pair that MIB reports for a signed submission. Peak is the largest faithfulness under value selection.

	CPR ↑		CMD ↓		Overshoot O		Undershoot U		
Entry	value	magnitude	value	magnitude	value	magnitude	value	magnitude	Peak
MIB example	1.236	0.969	0.248	0.030	0.242	0.000	0.005	0.030	1.38
Warm-start EP	2.451	1.003	1.457	0.024	1.455	0.014	0.002	0.010	3.02
Submission A	2.202	0.970	1.206	0.030	1.205	0.000	0.002	0.030	2.67
Submission B	2.527	0.980	1.531	0.030	1.529	0.006	0.002	0.025	3.12
MAttr	9.468	1.027	8.469	0.128	8.469	0.078	0.000	0.050	12.58
IRP	0.265	0.265	0.734	0.734	0.000	0.000	0.734	0.734	1.00
Random	0.214	0.242	0.785	0.757	0.000	0.000	0.785	0.757	1.00

shared-task pairs, and all but one, which Section C traces to a transcription error, belong to methods with two circuits.

6 Re-evaluation on GPT-2 small

Setup. We evaluated seven entries under both rules on GPT-2 small (Radford et al., 2019) and IOI with MIB's released code. Five are public submissions with signed scores: the MIB example, MAttr (Arora et al., 2026), and three entries of one shared-task team (Mondorf et al., 2025), Warm-start EP, the team's sequential ensemble S-Ens, and two entries we call Submissions A and B. On GPT-2 the scores of Submission B are three quarters of those of Submission A plus one quarter of those of Warm-start EP, as for the team's Hybrid-Ens, an equal-weight average of S-Ens and the three EAP variants that its P-Ens averages, so we take A and B to be P-Ens and Hybrid-Ens, with repositories in Table 14. By the rank correlation of their scores they form three groups, the MIB example with Submission A at Spearman $\rho = 0.94$, Submission B with Warm-start EP, whose far larger scores dominate the average, at $\rho = 0.996$, and MAttr, whose scores are defined only up to a common additive constant, since shifting them all leaves the masks it learns unchanged (Arora et al., 2026, Appendix A.4), so that their sign is a convention. Adding a constant to them leaves its value-selected circuits unchanged and changes its magnitude-selected ones. Two are controls. IRP, the public entry for Information Relevance Propagation, has nonnegative scores, so both rules agree. Random assigns seeded Gaussian scores. All evaluations use the same 96 validation examples, and Section I repeats everything on 100 test examples from other templates.

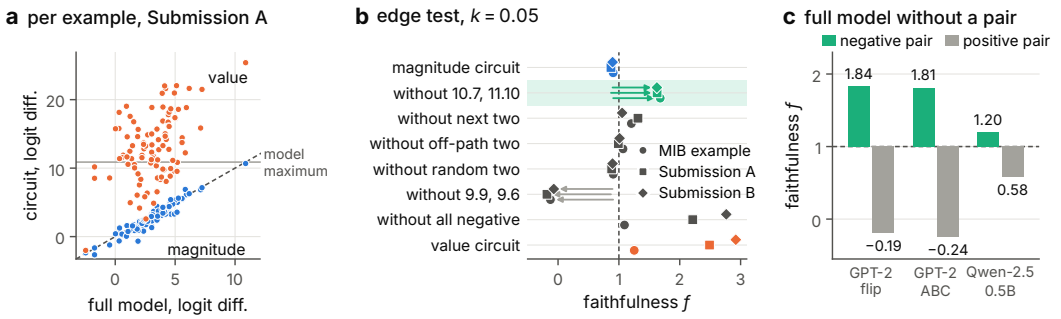


Figure 3: **Overshoot per example and the edge test on IOI. a**, Each validation example's logit difference under Submission A's circuits of size 0.2 against the full model's, GPT-2 small. **b**, Faithfulness at $k = 0.05$ after removing pairs of edges from the magnitude-selected circuit on GPT-2 small, as in Table 6 in Section E. **c**, The full model without the negative pair ■ or a positive pair ■, on GPT-2 under both counterfactuals and on Qwen-2.5 0.5B.

Overshoot in the reported CPR column. Under value selection the five submissions undershoot by at most $U = 0.006$ and overshoot by 0.24 to 8.47, as Table 2 shows. Their clipped CPR, the area under $\min(f, 1)$, which equals $W - U$, lies between 0.993 and 0.999, so the whole spread of the reported column, from 1.24 to 9.47, is overshoot. By Lemma 2, CPR and CMD then order the five in exact reverse under value selection, $\tau_b = -1.00$, an arithmetic consequence rather than a finding about methods. Under magnitude selection the overshoot stays below 0.079 and the differences between submissions are small. Reported as MIB reports them, CPR from one family and CMD from the other, the two columns rank different quantities of different circuits, and Section B gives the effect of this pairing and of the benchmark's logarithmic weighting option.

Overshoot example by example. The value-selected circuits of Warm-start EP, Submissions A and B and MAttr exceed the full model on the same example in 99% to 100% of example and size pairs with $k \geq 0.05$, and those of the MIB example in 65%. Those of Warm-start EP and Submissions A and B exceed 10.9, the largest logit difference the model produces on any validation example, in 63% to 81% of these pairs, and those of MAttr in 100%, against at most 1.1% under magnitude selection, as Figure 3a shows for Submission A.

Two sources of overshoot. We split the logit difference into a projection onto the two names and the final normalization scale, as Section G details. Removing the negative name mover pair from the full model multiplies the projection by 2.63 at a normalization factor of 1.00, with a KL divergence of 0.66 from the model's outputs. MAttr's value-selected circuit at $k = 0.2$ reaches a mean logit difference of 71.5, against 3.16, with a projection 14.2 times the model's at a normalization factor of 1.6 and a KL divergence of 23.07, against at most 3.29 for the other four submissions. The MIB example's overshoot matches left-out suppression, as Section 7 shows, while MAttr's arises in patched states far outside the model's range, and CPR rewards both. Under mean ablation neither circuit recovers the behavior, with $f = 0.02$ for MAttr's and $f = -0.06$ for the MIB example's.

7 Leaving out suppression

The best-known negatively scored components in IOI are the negative name mover heads 10.7 and 11.10, which attend to the correct name and write against it (Wang et al., 2023) and perform copy suppression (McDougall et al., 2024). In the MIB example and in Submission A, the two most negatively scored of all 32,491 edges are the logit edges of these two heads, and in Submission B the logit edge of 11.10 is the most negative. For the MIB example and Submission A, magnitude selection keeps the pair at every size, and value selection at no size below the full model.

Two edges. Figure 3b and Table 6 show edits of magnitude-selected circuits at $k = 0.05$. In the MIB example, removing only the two negative name mover edges raises faithfulness from 0.91 to 1.68, above the 1.25 of the value-selected circuit of the same size, and in Submission A from 0.87 to 1.63. The pair raises faithfulness 2.6 and 1.7 times as much as the next two most negative edges,

the comparison named in the predictions, and the two most negative edges off the pathway raise it by 0.11 to 0.16. In Submission B, whose scores place the logit edge of 10.7 at rank 145 among the negatively scored edges, removing the pair still adds 0.73. Across the three entries and the sizes 0.05, 0.2 and 0.5, removing the pair adds 0.73 to 0.84, and removing it from the full model gives $f = 1.84$. The rise exceeds the whole gap between the MIB example's magnitude- and value-selected circuits at these sizes and covers 47% to 50% of the corresponding gap for Submission A and 36% to 41% for Submission B. Removing the pair thus suffices to produce the MIB example's overshoot, produces part of that of Submissions A and B, and none of MAttr's, whose scores rank the two edges among the most positive. Of the four predictions written down before the test and listed in Section E, the first two held, the third held for Submissions A and B and is indeterminate for the MIB example, and the fourth did not hold. Section E shows where larger edits depart from the idealizations.

Two edges from every circuit. Removing the pair from every circuit of the MIB example's magnitude-selected family, as Figure 1b shows, raises its CPR from 0.97 to 1.58, above 1.24 for its value-selected family by 0.34 with 95% interval 0.24 to 0.44, and its CMD from 0.030 to 0.59. For the three entries the edited families score CPR from 1.57 to 1.59, whatever their scores, and only the MIB example's exceeds its value-selected family. CPR rewards this edit, which removes edges of two components that the IOI and copy-suppression analyses count as part of the mechanism, and CMD penalizes it, the discordant case of Lemma 2. Conversely, restoring the two heads with their inputs to the value-selected circuits of the three entries at $k = 0.05$, 0.2 and 0.5 lowers faithfulness by 0.51 to 1.03.

The ABC counterfactual. With the uploaded scores and five predictions written down beforehand, as Section F reports, the negative name mover edges keep their effect under ABC. Removing them lifts the full model to 1.81 against 1.84 under the flip, as Figure 3c shows. The MIB example's value-selected circuits, whose scores were computed under the flip, reach 1.38 at $k = 0.5$ under the flip but 0.70 under ABC. To separate the counterfactual of scoring from that of evaluation, we recomputed EAP's scores under each and evaluated both under each. Scores computed under ABC give value-selected circuits that reach 1.07 and 1.16 at $k = 0.2$ and 0.5 under the flip but 0.78 and 0.95 under ABC, as the flip scores do. At these sizes the counterfactual of the evaluation, not a mismatch with that of the scoring, sets the side of one, although under the flip the ABC scores fall short of the model at $k = 0.05$ and cut the overshoot over the three sizes from 0.20 to 0.08.

A second model. On Qwen-2.5 0.5B (Yang et al., 2024) the two most negatively scored edges of the MIB example and of Submission A are the logit edges of heads 23.12 and 22.6, and measured without the scores, head 23.12 has the most negative direct effect on the logit difference of all 336 heads and attends to the indirect object, the signature of a negative name mover. On 32 validation examples, removing the pair from the full model lifts faithfulness to 1.20 and removing the logit edges of heads 20.9 and 22.0 lowers it to 0.58, as Figure 3c shows. Submission A's value-selected circuits reach 2.10 and 2.11 at $k = 0.2$ and 0.5, and Section H gives its edge test and the outcomes of five predictions written down beforehand.

A second task. On MCQA with Qwen-2.5 0.5B no suppressive component is known in advance. In the MIB example's magnitude-selected circuit at $k = 0.05$, with $f = 0.80$, the two most negatively scored edges are the edge from MLP 23 to the logits and the edge from MLP 21 to MLP 23. Removing them raises faithfulness to 1.10, against 0.71 for the two most negative edges off their pathway and 0.80 for a random pair, and removing them from the full model gives $f = 1.59$, against 0.38 without the two most positively scored logit edges, as Table 13 in Section H shows.

8 Related work

Patching results and faithfulness depend on the counterfactual, the metric and the ablation (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024; Miller et al., 2024). Edin et al. (2026) suggest that the lead of their method GIM on CPR, which CMD does not show, could come from separating positive from negative scores, a distinction that value selection keeps and magnitude selection discards, and in their Appendices A.1.1 and A.6 deem CMD flawed because under magnitude selection correctly detected negative components enter a circuit and lower its f. By Proposition 1 this objection holds

where the circuit without those components falls short of one, but where it exceeds one, as our value-selected circuits of Section 6 do, admitting the components brings f back toward one and gives a CMD 8.4 to 66 times smaller. The CPR values above one that Edin et al. (2026) report for GIM certify by Corollary 1 that its value-selected circuits exceed one at some sizes. Circuit claims face non-identifiability, estimator variance and analytic multiplicity (Méloux et al., 2025; 2026; Wu et al., 2026; Mahale, 2026), and construct validity asks whether a score measures what it is named for (Bean et al., 2025). Section D discusses further work.

9 Recommendations and discussion

Recommendations. First, compute CPR and CMD on one declared circuit family, which by Lemma 1 reports its overshoot and undershoot and removes the cone violations of Section 5. Second, make a two-sided distance from the model, computed example by example, the headline behavioral score. Miller et al. (2024, Section 4.2) warn that a circuit can balance out examples far above and far below the model on average, and Wu & Barez (2026) score query circuits by a faithfulness score that penalizes deviations above and below the model equally for each query. On our data the per-example distance of magnitude-selected circuits is 1.4 to 2.6 times their CMD. KL divergence from the model, which Conmy et al. (2023, Appendix E.1) prefer because a task metric can be pushed past the model, is one such distance, and it puts MAttr's value-selected circuit at $k = 0.2$ at 23.07 against at most 3.29 for the other four submissions. The per-example distance penalizes the edit that CPR rewards: removing the two negative name mover edges from the MIB example's magnitude-selected circuits raises their per-example distance from 0.06 to 0.59, above the 0.41 of its value-selected family. At our sample size the per-example distance under magnitude selection does not rank the four submissions other than MAttr. Clipped CPR, $W - U$, is no substitute, since it cannot tell a circuit that matches the model from one that exceeds it by any amount. Third, name the selection rule, the counterfactual and the weighting of every column. No score computed from outputs shows that a circuit is complete, since omitted effects of opposite sign also cancel within one example, so completeness needs tests on components, such as the edge test of Section 7.

Faithfulness above one. Circuits that exceed the full model have been reported as a strength (O'Neill & Bui, 2024), exempted from the penalty of a variant of CMD (Rai et al., 2026) and reported as a concern (Miller et al., 2024; Conmy et al., 2023; Arora et al., 2026). Four observations argue against rewarding faithfulness above one. The reward depends on the counterfactual as well as on the method, and it also pays for patched states far outside the model's range. It is unbounded, an overshoot of 8.47 in a CPR of 9.47 for MAttr, and the states it rewards lie beyond the model example by example. By Proposition 3, the column cannot credit a method for ranking suppressive components correctly, only for excluding them, while the IOI circuit includes them and its completeness criterion asks for every component that matters, whatever its sign (Wang et al., 2023).

Limitations. The controlled experiments cover GPT-2 small on IOI with five submissions in three groups and 96 validation and 100 test examples, and reduced designs on Qwen-2.5 0.5B with 32 IOI and 50 MCQA examples, whereas the benchmark reports on 1,000 test examples. Only EAP was rescored under both counterfactuals. The account of MAttr's overshoot rests on measurement alone. The published numbers show how often leading entries overshoot, not why, and do not reach private scores. Only 22 of the 70 stored leaderboard curves with a published counterpart reproduce the published CPR, so we use them only for calibration. We do not claim that any method is better than another, and the pairing we study is disclosed by the benchmark.

10 Conclusion

MIB's two faithfulness scores are exact readings of overshoot and undershoot taken from two circuits, and when edge effects add and scores are exact the rule behind the CPR circuit earns its reward by leaving out suppression. Leaving out two edges of a known circuit lifts the CPR of the MIB example's magnitude-selected family from 0.97 to 1.58 while making its circuits less complete, and the counterfactual and patched states outside the model's range change the size of such rewards. Scoring one circuit per submission, reporting its two areas and naming the counterfactual make each result the property of one circuit.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing the text. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The authors have read and adhered to the ICLR Code of Ethics. This work reanalyzes published benchmark results and public leaderboard submissions. It involves no human subjects and no personal data. We name public submissions by their repositories so that the analysis can be checked. The checks rely on MIB's public tables, stored curves, released code and the maintainers' replies.

REPRODUCIBILITY STATEMENT

Section A proves every lemma, proposition and corollary. Section I describes the data, the code change and the replication on the test split. The supplementary material contains the transcribed published tables, the list of public score files with their repositories, the evaluation and edge-removal code, a checker that prints the certified areas and bounds of any reported pair, the recorded predictions with the hashes of the scripts they refer to, the per-example evaluation outputs from which every resampling result is computed, the rescored EAP scores with their curves, the normalization and MCQA runs, and the scripts that generate every measured value in the paper. These scripts also check Lemma 1 numerically on every curve of Section 6, and the transcribed tables were compared with the source papers.

REFERENCES

Alaa Anani, Tobias Lorenz, Bernt Schiele, Mario Fritz, and Jonas Fischer. Certified Circuits: Stability Guarantees for Mechanistic Circuits. In *International Conference on Machine Learning*, 2026. arXiv:2602.22968.

Dana Arad, Yonatan Belinkov, Hanjie Chen, Najoung Kim, Hosein Mohebbi, Aaron Mueller, Gabriele Sarti, and Martin Tutek. Findings of the BlackboxNLP 2025 Shared Task: Localizing Circuits and Causal Variables in Language Models. In *Proceedings of the 8th BlackboxNLP Workshop*, 2025. arXiv:2511.18409.

Aryaman Arora, Kirill Acharya, Nathan Hu, Yanzhe Zhang, Noah Goodman, Dan Jurafsky, and Christopher Potts. Matryoshka attribution: Learning to attribute language model outputs to representations and weights. *arXiv preprint arXiv:2609.25518*, 2026.

Amir Asiaee. Certified Interventional Fidelity: Anytime-Valid, Adaptive Evaluation of Causal Claims in Mechanistic Interpretability. In *Conference on Uncertainty in Artificial Intelligence*, 2026. arXiv:2607.08349.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025. arXiv:2511.04703.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding Transformer Circuits with Edge Pruning. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.16778.

Stefan Blücher, Johanna Vielhaben, and Nils Strodthoff. Decoupling Pixel Flipping and Occlusion Strategy for Consistent XAI Benchmarks. *Transactions on Machine Learning Research*, 2024. arXiv:2401.06654.

Nicolò Brunello, Andrea Cerutti, Andrea Sassella, and Mark Carman. BlackboxNLP-2025 MIB Shared Task: IPE: Isolating Path Effects for Improving Latent Circuit Identification. In *Proceedings of the 8th BlackboxNLP Workshop*, pp. 528–536, 2025. doi: 10.18653/v1/2025.blackboxnlp-1.30.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With Little Power Comes Great Responsibility. In *Conference on Empirical Methods in Natural Language Processing*, 2020. arXiv:2010.06595.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. Evaluating and Characterizing Human Rationales. In *Conference on Empirical Methods in Natural Language Processing*, 2020. arXiv:2010.04736.

Xi Chen, Mingyu Jin, Jingcheng Niu, Yutong Yin, Jinman Zhao, Bangwei Guo, Dimitris N. Metaxas, Zhaoran Wang, Yutao Yue, and Gerald Penn. All Circuits Lead to Rome: Rethinking Functional Anisotropy in Circuit and Sheaf Discovery for LLMs. In *International Conference on Machine Learning*, 2026. arXiv:2605.12671.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2304.14997.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Annual Meeting of the Association for Computational Linguistics*, 2020. arXiv:1911.03429.

Joakim Edin, Casper L. Christensen, Róbert Csordás, Tuukka Ruotsalo, Zhengxuan Wu, Maria Maistro, Jing Huang, and Lars Maaløe. Correcting Gradient-Based Circuit Localization via Interaction-Aware Backpropagation. *arXiv preprint arXiv:2505.17630v4*, 2026.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing Model Behavior with Path Patching. *arXiv preprint arXiv:2304.05969*, 2023.

Zhiren Gong, He Lu, Tiantong Wang, Yichi Zhang, Yixin Wang, Zihao Zeng, Ming Xiao, Chau Yuen, and Wei Yang Bryan Lim. Conditional Co-Ablation: Recovering Self-Repair Backups in Transformer Circuits. *arXiv preprint arXiv:2607.01940*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

Rohan Gupta, Iván Arcuschin, Thomas Kwa, and Adrià Garriga-Alonso. InterpBench: Semi-Synthetic Transformers for Evaluating Mechanistic Interpretability Techniques. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. arXiv:2407.14494.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms. In *Conference on Language Modeling*, 2024. arXiv:2403.17806.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Chia-Ying Hsieh, Hsin-Yuan Fang, and Chun-Shu Wei. AIM: Adversarial Information Masking for Faithfulness Evaluation of Saliency Maps. *arXiv preprint arXiv:2605.16905*, 2026.

M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93, 1938.

M. G. Kendall. The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3):239–251, 1945. doi: 10.1093/biomet/33.3.239.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024.

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying Variance in Evaluation Benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

Ajay Pravin Mahale. Explanation Multiplicity: Circuit-Level Interpretability Evidence Does Not Survive Defensible Analytic Variation. *arXiv preprint arXiv:2608.13754*, 2026.

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 *International Conference on Learning Representations*, 2024. arXiv:2311.17030.

Alireza Bayat Makou, Jingcheng Niu, Subhabrata Dutta, and Iryna Gurevych. Many Circuits, One Mechanism: Input Variation and Evaluation Granularity in Circuit Discovery. *Transactions on Machine Learning Research*, 2026. arXiv:2606.06267.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy Suppression: Comprehensively Understanding a Motif in Language Model Attention Heads. In *Proceedings of the 7th BlackboxNLP Workshop*, pp. 337–363, 2024.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The Hydra Effect: Emergent Self-repair in Language Model Computations. *arXiv preprint arXiv:2307.15771*, 2023.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, Everywhere, All at Once: Is Mechanistic Interpretability Identifiable? In *International Conference on Learning Representations*, 2025. arXiv:2502.20914.

Maxime Méloux, François Portet, and Maxime Peyrard. Mechanistic Interpretability as Statistical Estimation: A Variance Analysis. In *International Conference on Machine Learning*, 2026. arXiv:2510.00845.

Evan Miller. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer Circuit Evaluation Metrics Are Not Robust. In *Conference on Language Modeling*, 2024. arXiv:2407.08734.

Philipp Mondorf, Mingyang Wang, Sebastian Gerstner, Ahmad Dawar Hakimi, Yihong Liu, Leonor Veloso, Shijia Zhou, Hinrich Schütze, and Barbara Plank. BlackboxNLP-2025 MIB Shared Task: Exploring Ensemble Strategies for Circuit Localization Methods. In *Proceedings of the 8th BlackboxNLP Workshop*, 2025. arXiv:2510.06811.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden 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 *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 2025.

Yaniv Nikankin, Dana Arad, Itay Itzhak, Anja Reusch, Adi Simhi, Gal Kesten-Pomeranz, and Yonatan Belinkov. BlackboxNLP-2025 MIB Shared Task: Improving Circuit Faithfulness via Better Edge Selection. In *Proceedings of the 8th BlackboxNLP Workshop*, 2025. arXiv:2510.25786.

Charles O'Neill and Thang Bui. Sparse Autoencoders Enable Scalable and Reliable Circuit Identification in Language Models. *arXiv preprint arXiv:2405.12522*, 2024.

Wenbo Pan, Zhichao Liu, Xianlong Wang, Haining Yu, and Xiaohua Jia. Towards Long-Horizon Interpretability: Efficient and Faithful Multi-Token Attribution for Reasoning LLMs. In *International Conference on Machine Learning*, 2026. arXiv:2602.01914.

Yotam Perlitz, Ariel Gera, Ofir Arviv, Asaf Yehudai, Elron Bandel, Eyal Shnarch, Michal Shmueli-Scheuer, and Leshem Choshen. Do These LLM Benchmarks Agree? Fixing Benchmark Evaluation with BenchBench. *arXiv preprint arXiv:2407.13696*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI, 2019.

Daking Rai, Mor Geva, and Ziyu Yao. Data-driven Circuit Discovery for Interpretability of Language Models. *arXiv preprint arXiv:2605.09129*, 2026.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, et al. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*, 2024.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A Consistent and Efficient Evaluation Strategy for Attribution Methods. In *International Conference on Machine Learning*, 2022. arXiv:2202.00449.

Cody Rushing and Neel Nanda. Explorations of Self-Repair in Language Models. In *International Conference on Machine Learning*, 2024. arXiv:2402.15390.

Shun Shao, Binxu Wang, Shay B. Cohen, Anna Korhonen, and Yonatan Belinkov. Differentiable Faithfulness Alignment for Cross-Model Circuit Transfer. In *Conference on Language Modeling*, 2026. arXiv:2604.24302.

Yang Sheng and Jie Fu. Circuit Claims Depend on What Is Extracted and How It Is Compared. *arXiv preprint arXiv:2607.18921*, 2026.

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 *Advances in Neural Information Processing Systems*, 2024. arXiv:2410.13032.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis. In *Conference on Empirical Methods in Natural Language Processing*, 2023. arXiv:2305.15054.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution Patching Outperforms Automated Circuit Discovery. In *Proceedings of the 7th BlackboxNLP Workshop*, 2024. arXiv:2310.10348.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The Curse of Multiple Mediators: Hidden Interaction Effects in Activation Patching. *arXiv preprint arXiv:2606.27510*, 2026.

Han-yu Wang. Function-Vector Heads Are Two Populations: Writers and Cancellers in In-Context Learning. *arXiv preprint arXiv:2606.07560*, 2026.

Kevin 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 *International Conference on Learning Representations*, 2023. arXiv:2211.00593.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, Assemble, Ace: Understanding How LMs Answer Multiple Choice Questions. In *International Conference on Learning Representations*, 2025. arXiv:2407.15018.

Frank Zhengqing Wu, Francesco Tonin, and Volkan Cevher. Demystifying Variance in Circuit Discovery of LLMs. *arXiv preprint arXiv:2606.16920*, 2026.

Tung-Yu Wu and Fazl Barez. Query Circuits: Explaining How Language Models Answer User Prompts. In *International Conference on Machine Learning*, 2026. arXiv:2509.24808.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*, 2024. arXiv:2309.16042.

Guanhua Zhang and Moritz Hardt. Inherent Trade-Offs between Diversity and Stability in Multi-Task Benchmarks. In *International Conference on Machine Learning*, 2024. arXiv:2405.01719.

Luyang Zhang and Jialu Wang. When Attribution Patching Lies: Diagnosis and a Second-Order Correction. *arXiv preprint arXiv:2606.09899*, 2026.

A PROOFS

Proof of Lemma 1. Write $d_i = f_i - 1$ and $\Delta_i = k_{i+1} - k_i$. The trapezoid rule applied to any function g of the curve gives $\sum_i w_i\, g(d_i)$ with $w_1 = \Delta_1/2$, $w_{10} = \Delta_9/2$ and $w_i = (\Delta_{i-1} + \Delta_i)/2$ otherwise, so $\sum_i w_i = W$ and every w_i is positive. Hence $\text{CPR} - W = \sum_i w_i d_i$, $\text{CMD} = \sum_i w_i |d_i|$, $O = \sum_i w_i (d_i)_+$ and $U = \sum_i w_i (-d_i)_+$. Since $d = d_+ - (-d)_+$ and $|d| = d_+ + (-d)_+$, $\text{CPR} - W = O - U$ and $\text{CMD} = O + U$. Then $|\text{CPR} - W| = |O - U| \le O + U = \text{CMD}$, with equality exactly when $O = 0$ or $U = 0$. The areas are sums at the grid points. They equal the geometric areas above and below one of the interpolated curve except where it crosses one inside an interval. □

Proof of Lemma 2. By Lemma 1, $\text{CPR}_a - \text{CPR}_b = \delta_O - \delta_U$ and $\text{CMD}_a - \text{CMD}_b = \delta_O + \delta_U$. CPR strictly prefers a when the first is positive and CMD when the second is negative. Both strictly prefer a exactly when $\delta_U < -|\delta_O|$ and both strictly prefer b exactly when $\delta_U > |\delta_O|$. When $|\delta_O| > |\delta_U|$ the two differences share the sign of δ_O, so CPR prefers the family with the larger overshoot and CMD the other. When $\delta_O = \delta_U \neq 0$, CPR ties and CMD prefers the family with the smaller undershoot, and when $\delta_O = -\delta_U \neq 0$ the roles swap. Concordant pairs are those of the first case, discordant pairs those of the second. □

Proof of Corollary 1. A pair from one family has $O = (\text{CMD} + \text{CPR} - W)/2 \ge 0$ and $U = (\text{CMD} - \text{CPR} + W)/2 \ge 0$. Since $O, U \ge 0$, $O \ge O - U = \text{CPR} - W$ and $U \ge W - \text{CPR}$. For the CMD family, $|\text{CPR} - W| \le \text{CMD}$ by Lemma 1, and for the CPR family its own CMD is at least $|\text{CPR} - W|$ for the same reason. The inequalities are preserved under averaging over runs because the cone is convex. Each bound involves the reported values only through $\text{CPR} - W$, $|\text{CPR} - W|$ and CMD, and each of these changes by at most as much as the one reported value it contains. Rounding to two decimals moves each reported value by at most half a unit in the second decimal place, so each bound loosens by at most that amount per value. □

Proof of Proposition 1. Order the effects $a_{(1)} \ge \cdots \ge a_{(E)}$ and let $F(n) = \sum_{i \le n} a_{(i)}/A$, the faithfulness of the value-selected circuit of n edges. **a** Any circuit of n edges has $\sum_{e \in \mathcal{C}} a_e \le \sum_{i \le n} a_{(i)}$ by exchanging a smaller effect inside for a larger one outside, and the weights are positive, so CPR is maximal. **b** The increments $a_{(n)}$ are nonincreasing, so F is concave, rises while $a_{(n)} > 0$ and falls afterwards. Its maximum is $F(P) = S_+/A = 1 + S_-/A$ with P the number of positive effects and S_+ their sum, since $S_+ - S_- = A$. Because $F(E) = 1$ and $F(E-1) = 1 - a_{(E)}/A > 1$ when $S_- > 0$, concavity keeps $F > 1$ from the first n with $F(n) > 1$ up to $E - 1$. When $S_- = 0$, $F \le 1$ everywhere. **c** $A\,|1 - f(\mathcal{C})| = |\sum_{e \notin \mathcal{C}} a_e| \le \sum_{e \notin \mathcal{C}} |a_e|$, and keeping the n largest magnitudes minimizes the right side. For effects $0.9, -0.8, 0.5, 0.4$ and two edges, $A = 1$, magnitude selection keeps 0.9 and -0.8 with $|1 - f| = 0.9$, while $\{0.5, 0.4\}$ gives 0.1. **d** $A\,(1 - f(\mathcal{C})) = \sum_{e \notin \mathcal{C}} a_e$. □

Proof of Proposition 2. **a** For any circuit, removing its edges of $\mathcal{N}$ does not lower m and adding the edges of $\mathcal{P}$ it lacks does not lower it either, so $f(\mathcal{C}) \leq f(\mathcal{P})$. The full graph is $\mathcal{P}$ plus edges of $\mathcal{N}$, so $f(\mathcal{P}) \geq f(\mathcal{M}) = 1$. **b** The value-selected circuits are nested. Up to size $|\mathcal{P}|$ they grow by edges of $\mathcal{P}$ and reach $\mathcal{P}$, and afterwards they grow by edges of $\mathcal{N}$. Every $f(k_i)$ with $i < 10$ is at most $f(\mathcal{P})$ by **a** and $f(1) = 1$, which bounds CPR. **c** This follows from the definition of $\mathcal{N}$. Pruning does not change m, because an edge whose source receives no clean input carries the source's counterfactual activation whether or not it is in the circuit, and a node with no path to the logits in the circuit feeds nothing that reaches them. □

Proof of Proposition 3. The top $\lfloor kE \rfloor$ scores are all positive when at least that many are positive, and the order among positive scores does not depend on the negative ones. Pruning only removes edges. With at least half of the edges positive this holds at every grid size below one, and $f(1) = 1$. □

B Resolution and rankings

Table 3 compares the orderings that pairs of criteria induce over the five submissions.

Table 3: **Rank agreement between criteria.** Kendall τ_b (Kendall, 1945) between the orderings that two criteria induce over the five submissions, on the validation and test examples, with the 5th to 95th percentile over 2,000 resamples of the validation examples. Rows that compare CPR and CMD within one rule follow from Lemma 2 and serve as checks. Orderings use unrounded values, and three magnitude-selected CMD values in Table 2 agree to the digits shown.

Comparison	Held fixed	validation	test	resampling range
CPR vs. CMD	value selection	-1.00	-1.00	-1.00
CPR vs. CMD	magnitude selection	-0.40	-0.20	-0.40 to $+0.20$
value vs. magnitude	CPR	$+0.80$	$+0.80$	$+0.60$ to $+0.80$
value vs. magnitude	CMD	$+0.60$	0.00	0.00 to $+0.80$
CPR value vs. CMD magnitude	neither	-0.60	0.00	-0.80 to 0.00

Resampling by template. MIB's validation and test splits use disjoint sets of nine templates each. We pool both splits, 196 examples from 18 templates, and resample templates 2,000 times. The reported pair then orders the five at $\tau_b = 0.20$, with 5th and 95th percentiles of -0.20 and 0.20, and the best of the four submissions other than MAttr by CMD under magnitude selection is Submission B in 68% of resamples and Warm-start EP in 28%. With the same resampling, the entry best by the reported CPR is best by the reported CMD in none of the resamples, because MAttr leads the first and trails the second, and in 68.0% without MAttr. Of the ten pairwise orders by CMD under magnitude selection, three hold in fewer than 95% of resamples. With 18 templates as clusters, percentile intervals can cover less than their nominal level.

Effect of the pairing. Warm-start EP reports CPR $= 2.45$ where one circuit would give 1.00 at the same CMD, and MAttr 9.47 where one circuit would give 1.03, an option that submissions with nonnegative scores or uploaded circuits lack. On value-selected circuits CMD would rank Random above four of the five submissions.

Logarithmic weights. The regimes above belong to the benchmark's integration, which is linear in k. Under MIB's logarithmic option, rescaled to the same width, the MIB example's value-selected family undershoots more than it overshoots, $U = 0.21$ against $O = 0.14$, MAttr's magnitude-selected family overshoots more than it undershoots, and the best submission by CMD under magnitude selection changes from Warm-start EP to Submission B.

Rank agreement at leaderboard sizes. For n untied items the permutation null of Kendall's τ is the distribution of inversions of a random permutation (Kendall, 1938). Table 4 lists the smallest τ that reaches one-sided $p \leq \alpha$. These are standard values, and they answer whether two orderings are associated, not whether a leaderboard can separate its entries.

Table 4: **Critical values of rank correlation.** Smallest τ with one-sided $p \leq \alpha$ under the exact null, and the same for Spearman ρ where the null is enumerated.

n	τ, α=.05	τ, α=.01	ρ, α=.05
4	1.00	–	1.00
5	0.80	1.00	0.90
6	0.73	0.87	0.83
7	0.62	0.81	0.71
8	0.57	0.71	0.64
9	0.50	0.67	0.60
10	0.47	0.60	–
11	0.42	0.56	–
12	0.39	0.55	–
13	0.36	0.51	–
14	0.36	0.47	–
15	0.33	0.47	–

The benchmark's own columns. Within each of the 11 cells of MIB's tables, the CPR and CMD orderings of the eight to eleven reported methods agree at a median τ_b of 0.69, ranging from 0.07 to 0.87. Each column comes from its own circuit family, so Lemma 2 does not decompose this agreement, and in 6 cells the method with the best CPR is not among those with the best CMD. MIB also reports an AUROC against the known circuits of an InterpBench model (Gupta et al., 2024). Across the 11 methods, its ordering and the CMD ordering on IOI with GPT-2 agree at $\tau_b = 0.12$, and at this size the one-sided exact test at level 0.05, which needs $\tau \geq 0.42$, has power 0.53 when the two orderings follow a bivariate normal model with Kendall's τ of 0.4.

C Published pairs by method

MIB's cells cover GPT-2 small (Radford et al., 2019), Qwen-2.5 0.5B (Yang et al., 2024), Gemma-2 2B (Riviere et al., 2024) and Llama-3.1 8B (Grattafiori et al., 2024) on IOI (Wang et al., 2023), arithmetic, Arith (Stolfo et al., 2023), multiple-choice question answering, MCQA (Wiegreffe et al., 2025), and the easy and challenge sets of the AI2 Reasoning Challenge, ARCE and ARCC (Clark et al., 2018). The CPR and CMD numbers in MIB's tables are the same in its arXiv v2 and its proceedings version. Figure 2c shows the published pairs in the plane of the two areas, and Table 5 gives, for each method, the number of its pairs that lie outside the cone. By Corollary 1, the circuits behind 25 published CPR values, the leading ones of 7 cells among them, are farther from the model by CMD than the random baseline of their cell.

One pair cannot come from the pairing. IFR on IOI-Llama reports CPR $= 0.09$ with CMD $= 0.83$, outside the cone by at least 0.06 beyond rounding, yet IFR's scores are nonnegative, so both rules return one circuit. MIB's private-test tables report 0.11 and 0.89 for the same cell, on the edge of the cone, which points to a transcription error in the public table. The same check on MIB's private-test tables finds 62 of 97 pairs outside it, 23 with the CPR too high for its CMD and 39 with the CPR too low. In the shared-task report one entry has its CMD in the arithmetic column and its CPR in the MCQA column, so we treat the two values as unpaired. Nikankin et al. (2025) build the circuits of their entry with an integer program that keeps a minimum share of positive edges for CMD and uses bootstrapping in place of that share for CPR, so the entry's pairs also come from two circuits in the cells their paper reports, which include no Llama-3.1 cell. MIB's table of arithmetic scores by operator lists the random baseline with CPR and CMD that no curve can produce, with the two columns apparently transposed.

D Further related work

Beyond this benchmark. Lemmas 1 and 2 hold for any curve, reference level and positive weights, so any score that normalizes an explanation's effect by the model's and can exceed its reference inherits the same reward for overshoot. ERASER sufficiency rewards subsets that exceed the full input in the same way (DeYoung et al., 2020), and Carton et al. (2020) propose a clipped version.

Table 5: **Pairs that lie outside the cone of Lemma 1 by more than 0.01, by method.** Entries marked † are from the shared-task report (Arad et al., 2025), namely P-Ens, S-Ens and Hybrid-Ens (Mondorf et al., 2025), ILP+PNR+Boot (Nikankin et al., 2025) and IPE (Brunello et al., 2025), and all others are from the MIB paper.

Method	Outside the cone	Pairs
Random	0	11
EActP (CF)	1	3
EAP (mean)	9	11
EAP (CF)	9	11
EAP (OA)	2	3
EAP-IG-inp	7	11
EAP-IG-act	9	11
NAP (CF)	10	11
NAP-IG (CF)	10	11
IFR	1	11
UGS	0	3
P-Ens†	3	3
S-Ens†	3	3
Hybrid-Ens†	3	3
ILP+PNR+Boot†	7	9
IPE (CF)†	2	2

Evaluation of circuits. The shared-task organizers suggested that removing components with inconsistent effects and mixing positive and high-magnitude components may regularize a circuit toward the full model (Arad et al., 2025, Section 3.4), and by Lemma 1 moving a circuit toward the model lowers CMD from either side but lowers CPR where the circuit exceeds the model. Negative heads suppress the answer that positive heads promote (Wang et al., 2023; McDougall et al., 2024), and ablations are partly compensated by other components (McGrath et al., 2023; Rushing & Nanda, 2024; Gong et al., 2026), so edits like ours measure net effects. Shared-task entries tuned an ensemble against CMD minus CPR (Mondorf et al., 2025) or chose a different edge selection for each score (Nikankin et al., 2025). Hypothesis tests for circuits (Shi et al., 2024), interaction effects between mediators (Vaidyanathan et al., 2026) and anytime-valid fidelity tests (Asiaee, 2026) refine how one circuit is evaluated, and a second-order correction reduces the error of the attribution patching scores from which circuits are built (Zhang & Wang, 2026). Stability of circuits can be certified (Anani et al., 2026), claims change with the extracted object and the comparison level (Sheng & Fu, 2026), several faithful circuits can carry out one task (Chen et al., 2026), and evaluation granularity alone can change a verdict (Makou et al., 2026). For function-vector heads on synthetic in-context tasks, magnitude ranking recovers most sign-opposed heads on one task and few on another, and ablating these heads raises the logit difference of the correct label (Wang, 2026). For benchmarks in general, rankings trade diversity against stability (Zhang & Hardt, 2024), benchmarks that agree across many models often disagree on the top-ranked ones (Perlitz et al., 2024), benchmark scores are often compared without the variance that decides whether a difference is meaningful (Madaan et al., 2024; Miller, 2024), and small comparisons are often underpowered (Card et al., 2020). Deletion and insertion curves and signed and unsigned attributions raise questions related to ours (Rong et al., 2022; Blücher et al., 2024; Hsieh et al., 2026).

E EDGE-REMOVAL TEST

Before running the test we recorded four predictions, which the supplementary material contains with their timestamp and the hash of the script. The records are local files, not deposited with a third party. The starting circuit is the magnitude-selected circuit at $k = 0.05$ of the MIB example, Submission A and Submission B.

- Removing the negative name mover pair raises faithfulness more than removing the next two negative edges and more than removing two random negative edges, for the MIB example and Submission A. Outcome: held.
- Removing the positively scored logit edges of the name movers 9.9 and 9.6 lowers faithfulness for all three entries. Outcome: held.

- Removing every negatively scored edge raises faithfulness above one and closes most of the gap to the value-selected circuit, for all three entries. Outcome: held for Submissions A and B and is indeterminate for the MIB example, with the share of the gap closed 54% for the MIB example, 95% interval 40% to 62% and below half in 24% of resamples, against 83% and 92% for Submissions A and B.
- The pair closes only part of that gap, less than all negative edges do. Outcome: did not hold, because in the MIB example the pair alone exceeds the value-selected circuit.

Intervals and Submission B. The next two most negative edges include query inputs to the same heads, one for the MIB example and both for Submission A. With templates resampled, the 95% intervals of the difference between the pair's effect and theirs are 0.41 to 0.55 and 0.26 to 0.37. In Submission B the pair adds 0.73, against 0.11 for its off-path pair, scored down to -0.27, so the effect follows the components, not the ranking. For the MIB example the rise is 218% to 242% of the gap between its two families at the sizes 0.05, 0.2 and 0.5.

Limits of the idealizations. Removing ever more negatively scored edges, which part c of Proposition 2 says never lowers faithfulness, raises it at every step for Submission B and up to a dip of 0.01 for Submission A, but lowers it for the MIB example after 32 edges, as Table 7 shows. Part a of Proposition 1 fails too: the MIB example's magnitude-selected circuit without its 32 most negative edges scores 2.05, 2.43 and 2.53 at $k = 0.05$, 0.2 and 0.5, above its value-selected circuits.

Under additive effects the value-selected family would peak at $1 + S_-/A$, which the scores put at 5.0, 3.9 and 4.5 for the MIB example and Submissions A and B, against observed peaks of 1.38, 2.67 and 3.12. The off-path control was added after the MIB example's first results. The 95% interval of the amount by which the pair's effect exceeds it is 0.52 to 0.73 for the MIB example and 0.54 to 0.76 for Submission A. Table 7 reports every dose. For the MIB example faithfulness falls after 32 edges, to 1.44 and 1.09, so some of its negatively scored edges carry signal that positively scored paths need, and its magnitude-selected circuit without the 32 most negative edges stays above the value-selected circuits' 1.25, 1.31 and 1.38 at $k = 0.05$, 0.2 and 0.5.

Limits of patching. For our edits, edge-level patching also cuts the path through which the name movers drive the queries of the negative name movers (McDougall et al., 2024), attribution patching can miss components that feed a saturated attention softmax or whose direct and indirect effects cancel (Kramár et al., 2024), and patching can reach states the model never visits (Makelov et al., 2024; Goldowsky-Dill et al., 2023).

Table 6: **Faithfulness at $k = 0.05$ after removing edges from the magnitude-selected circuit, GPT-2 small, IOI.** NNM pair, logit edges of heads 10.7 and 11.10. Next two, the next two most negative edges, the comparison named in the predictions. Off-path two, the two most negative edges not touching the two heads, added later. Random two, mean of three random negative pairs. NM pair, logit edges of the name movers 9.9 and 9.6.

		Magnitude selection, after removing						
Entry	magnitude	NNM pair	next two	off-path two	random two	NM pair	all negative	value
MIB example	0.91	1.68	1.20	1.07	0.91	−0.13	1.09	1.25
Submission A	0.87	1.63	1.31	0.99	0.88	−0.18	2.21	2.49
Submission B	0.90	1.62	1.05	1.01	0.90	−0.07	2.76	2.92

Table 7: **Edge-removal doses.** Faithfulness at $k = 0.05$ after removing the given number of most negatively scored edges from the magnitude-selected circuit, GPT-2 small, IOI.

		Most negative edges removed					
Entry	magnitude	2	8	32	256	all	value
MIB example	0.91	1.68	2.00	2.05	1.44	1.09	1.25
Submission A	0.87	1.63	1.90	2.15	2.22	2.21	2.49
Submission B	0.90	1.16	1.30	1.39	2.24	2.76	2.92

Family edit and add-back. The family edit removes whichever of the two logit edges the magnitude-selected circuit contains at each of the nine sizes below one, taking the sizes 0.05, 0.2 and 0.5 from the test above, and keeps $f(1) = 1$. With the full model's own edit at $k = 1$, $f = 1.84$, the MIB example's edited family would score CPR $= 1.79$. The add-back test starts from the value-selected circuit and adds the two logit edges with every edge into the two heads that the magnitude-selected circuit of the same size contains, so that the heads receive clean inputs. Of its two predictions, recorded before it ran, the prediction that every restoration lowers faithfulness held, and the prediction that the drop is at least 0.3 for the MIB example and Submission A held. Both edits were added after the four predictions above.

F ABC COUNTERFACTUAL

Under the flip, a left-out edge carries the activation of a prompt whose answer is the other name, so a left-out suppressive edge suppresses the wrong name, and under ABC it carries that of a prompt with little preference between the two. The ABC runs use the same 96 validation examples, where the empty circuit has a mean logit difference of -0.73 against -2.79 under the flip. The flip multiplies the pair's effect in logit difference by 1.6 and the model's own effect over the empty circuit by 1.5, so the normalized effect barely changes. We recorded five predictions before the runs, with O' and U' summed over the sizes evaluated for each entry.

B1. Under ABC, each of the four entries of Table 8 has $O' > U'$ under value selection.

B2. Under ABC, each of the four entries has a smaller O' under magnitude selection than under value selection.

B3. Under ABC at $k = 0.05$, removing the negative name mover pair from the magnitude-selected circuit of the MIB example and of Submission A raises faithfulness, and removing their two most positively scored logit edges lowers it.

B4. Under ABC, the full model has faithfulness above one without the negative name mover pair and below one without the two most positively scored logit edges.

B5. Under ABC, removing the negative name mover pair from the full model raises faithfulness by half to twice as much as under the flip, and raises the logit difference by a quarter to three quarters as much as under the flip.

A design change, recorded before the affected runs, restricts B3 to the MIB example and evaluates Submission A in B1 and B2 at $k = 0.2$ and 0.5 only. Removing the 8 or 32 most negative edges from the MIB example's magnitude-selected circuit at $k = 0.05$, a control added afterwards, gives 1.47 and 1.69, so under ABC too the most negatively scored edges work against the task, and leaving out the top 8 already lifts the circuit above the model while value selection, which leaves out every negatively scored edge, stays below it. Table 8 and Figure 4 compare the two counterfactuals at the two sizes that carry most of the weight. Of the five predictions, B1 and B2 failed on the MIB example, and B3, B4 and B5 held. B2 failed there because neither family of the MIB example overshoots under ABC, so magnitude selection cannot lower an overshoot of zero.

Table 8: **Flip and ABC counterfactuals.** Value-selected faithfulness under the flip and ABC counterfactuals at $k = 0.2$ and 0.5, and partial areas O' and U' over the sizes evaluated under ABC, from 0.05 to 0.5 for the MIB example and 0.2 and 0.5 for the others. The last two columns give the undershoot of both families under ABC.

	$k = 0.2$		$k = 0.5$		O', value		U', ABC	
Entry	flip	ABC	flip	ABC	flip	ABC	value	magnitude
MIB example	1.31	0.67	1.38	0.70	0.242	0.000	0.228	0.051
Submission A	2.66	1.62	2.67	1.66	1.001	0.389	0.000	0.016
Submission B	3.05	1.63	3.12	1.52	1.258	0.334	0.000	0.049
MAttr	12.49	11.90	12.58	12.25	6.932	6.680	0.000	0.199

Scoring and evaluating under each counterfactual. We recomputed EAP's scores with the benchmark's attribution settings, the first 100 training examples in batches of 20, patching, and

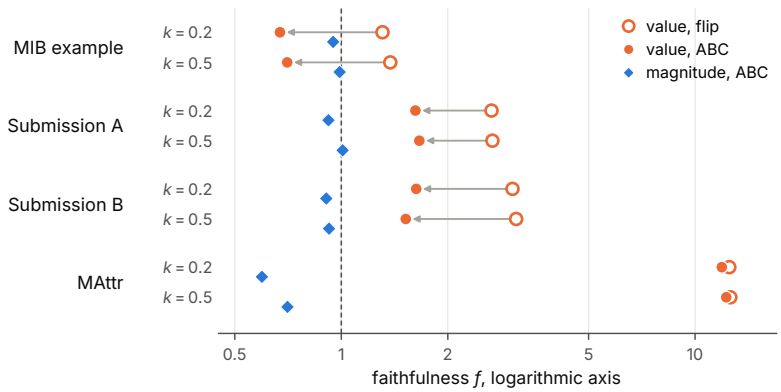


Figure 4: **Faithfulness under the answer flip that MIB runs and under the ABC counterfactual, GPT-2 small, IOI.** Arrows run from the flip to ABC for the value-selected circuits of Table 8, and diamonds give the magnitude-selected circuits under ABC.

the logit difference as the loss, once under the flip and once under ABC, and evaluated both score vectors under both counterfactuals on the same 96 validation examples. The flip scores agree with the MIB example's at Pearson $r = 0.998$ and share its four most negatively scored edges. The ABC scores place the logit edges of 10.7 and 11.10 at ranks 1 and 3 from the most negative, and the Spearman correlation between the ABC and the flip scores is 0.50. Table 9 gives faithfulness at the three sizes that carry most of the weight and the two areas. Under ABC, the value-selected circuits of the ABC scores undershoot by $U' = 0.08$ over the three sizes, more than the 0.06 of their magnitude-selected circuits, because at $k = 0.2$ the magnitude-selected circuit is more faithful, 0.92 against 0.78, contrary to part a of Proposition 1. Areas marked $*$ cover only these three sizes, because the other sizes were not evaluated in time.

Table 9: **EAP scored and evaluated under each counterfactual, GPT-2 small, IOI.** Faithfulness at three sizes and the overshoot and undershoot of each family, under value selection and magnitude selection. Areas marked $*$ are the partial areas O' and U' over the three sizes.

		value					magnitude				
Scored	Evaluated	0.05	0.2	0.5	O	U	0.05	0.2	0.5	O	U
flip	flip	1.18	1.28	1.34	0.20*	0.00*	0.91	0.96	1.00	0.00*	0.01*
flip	ABC	0.70	0.74	0.76	0.00*	0.16*	0.52	0.86	0.93	0.00*	0.08*
ABC	flip	0.74	1.07	1.16	0.08*	0.01*	0.69	0.99	1.00	0.00*	0.01*
ABC	ABC	0.57	0.78	0.95	0.00*	0.08*	0.50	0.92	0.95	0.00*	0.06*

G FINAL NORMALIZATION OF PATCHED STATES

With the final LayerNorm folded into the unembedding, the logit difference of an example is $c + p/\sigma$, where σ is the normalization scale at the answer position, p the projection of the centered residual stream onto the difference between the unembeddings of the two names, and c the difference between the folded biases. Relative to the full model, a circuit's mean logit difference net of c is thus close to the product of a projection factor, the ratio of the mean p to the model's, and a normalization factor, the ratio of the model's mean σ to the circuit's. Table 10 gives both for the value-selected circuits of the five submissions at $k = 0.2$ on the 96 validation examples, with the KL divergence of the circuit's next-token distribution from the model's at the answer position. Under mean ablation, with the full model and the mean-ablated empty circuit, whose mean logit difference is 0.22, as references, MAttr's value-selected circuit at $k = 0.2$ has $f = 0.02$ and the MIB example's has $f = -0.06$, so the overshoot does not survive mean ablation.

Table 10: **Final normalization of value-selected circuits at $k = 0.2$, GPT-2 small, IOI.** Faithfulness, mean logit difference, normalization and projection factors relative to the full model, and mean KL divergence from the full model.

Circuit	f	Logit difference	Normalization	Projection	KL
Full model	1.00	3.16	1.00	1.00	0.00
Full model without the NNM pair	1.84	8.14	1.00	2.63	0.66
MAttr	12.49	71.52	1.62	14.17	23.07
MIB example	1.31	4.99	1.04	1.49	0.54
Submission A	2.66	13.02	1.11	3.71	2.22
Submission B	3.05	15.33	1.03	4.74	3.29
Warm-start EP	2.94	14.67	1.03	4.54	3.04

H SECOND MODEL AND SECOND TASK

Qwen-2.5 0.5B runs in float32 on CPU instead of bfloat16 on GPU, with MIB's loading, patching and pruning, on the first 32 validation examples, where the full model has a mean logit difference of 4.97 and the empty circuit -4.96. The curves use the sizes 0.05, 0.2 and 0.5, which carry most of the weight, and the edge test uses one random pair. Tables 11 and 12 and Figure 5 give the results. The head measurement is one forward pass per example. A head's direct effect is its output at the final position, divided by the final normalization scale and projected onto the difference of the unembeddings of the two names. Head 23.12 has a direct effect of -0.74 and attends to the indirect object with weight 0.44 and to the subject with 0.06, head 22.6 has -0.40, 0.15 and 0.19, and the three heads with the most positive direct effect, 20.9, 22.0 and 21.0, are those whose logit edges the MIB example ranks most positive. We recorded five predictions beforehand for the MIB example and Submission A, with O' and U' summed over the three sizes.

Q1. Under value selection, each entry has $O' > U'$ and $O' > 0.1$.

Q2. Under magnitude selection, each entry has $O' < 0.1$ and a smaller O' than under value selection.

Q3. At $k = 0.05$, removing the two most negatively scored edges, the negative pair, from each entry's magnitude-selected circuit raises faithfulness more than removing the random pair or the two most negative edges off the pathway.

Q4. At $k = 0.05$, removing the two most positively scored logit edges, the positive pair, from each entry's magnitude-selected circuit lowers faithfulness.

Q5. For each entry, the full model has faithfulness above one without the negative pair and below one without the positive pair.

Of the five predictions, Q1 did not hold, failing for the MIB example and holding for Submission A, Q2 did not hold, failing for the MIB example and holding for Submission A, Q3 held for Submission A and was indeterminate for the MIB example, whose effect is below 0.01, Q4 held for Submission A and was indeterminate for the MIB example, whose effect is below 0.01, and Q5 held. An edit counts as holding only when the effect or margin a prediction names reaches 0.01, the resolution of the tables, a rule we set after the runs. The MIB example's value-selected circuits recover none of the behavior, $|f| \leq 0.02$, and at $k = 0.05$ its magnitude-selected circuit recovers none either, so its edits cannot separate the pairs. MIB reports a CPR of 0.26 on this cell for EAP scored with counterfactual ablation, the variant it marks CF (Mueller et al., 2025, Table 14). By Corollary 1 this value certifies undershoot, and the benchmark's stored curve for it lies within 0.02 of our curve for the MIB example, so Q1 and Q2 could not hold for the MIB example, as a check of the published value before recording them would have shown, and its edits at $k = 0.05$ change faithfulness by at most 0.006.

MCQA. The MCQA run uses the first 50 validation examples of the benchmark's colored-object questions with four choices and its default counterfactual, which relabels the choices with numbers. The full model has a mean logit difference of 6.90 and the empty circuit -5.85. The edge test repeats that of Table 12 with the MIB example's scores, reduced for time to the controls in Table 13, and without predictions written down beforehand. Table 13 gives the results.

Table 11: **Qwen-2.5 0.5B on IOI.** Faithfulness at three sizes under both rules and partial areas over them.

	Faithfulness f						Partial areas			
	value			magnitude			value		magnitude	
Entry	0.05	0.2	0.5	0.05	0.2	0.5	O'	U'	O'	U'
MIB example	0.02	0.01	0.02	−0.01	0.97	1.00	0.000	0.631	0.000	0.049
Submission A	2.07	2.10	2.11	0.88	0.95	0.99	0.704	0.000	0.000	0.020

Table 12: **Qwen-2.5 0.5B on IOI, edge removal.** Faithfulness at $k = 0.05$ after removing edges from the magnitude-selected circuit, the value-selected circuit, and the full model without either pair.

		after removing a pair					full model without a pair	
Entry	magnitude	negative	off path	random	positive	value	negative	positive
MIB example	−0.01	−0.01	−0.01	−0.01	0.00	0.02	1.20	0.58
Submission A	0.88	1.06	1.02	0.88	0.44	2.07	1.20	0.58

Table 13: **Qwen-2.5 0.5B on MCQA, edge removal.** Faithfulness at $k = 0.05$ after removing edges from the MIB example's magnitude-selected circuit, the value-selected circuit, and the full model without either pair.

		after removing						full model without a pair	
Entry	magnitude	negative	next two	off path	random	positive	value	negative	positive
MIB example	0.80	1.10	–	0.71	0.80	0.65	0.62	1.59	0.38

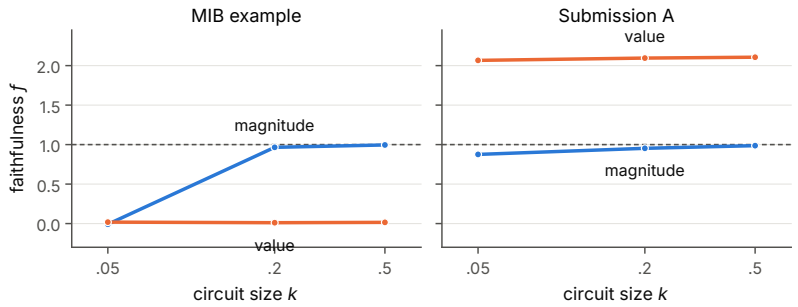


Figure 5: **Qwen-2.5 0.5B on IOI.** Faithfulness under both rules at the three evaluated sizes, the curves of Table 11.

I DATA, CODE AND REPLICATION

Score files and evaluation. All score files were downloaded from the public Hugging Face repositories of the submissions, listed in the supplementary material. Selection and pruning use `Graph.apply_topn` from the benchmark's EAP-IG library at the commit the benchmark repository pins. We used the benchmark repository at commit b759df3 and changed only the eighteen tensor allocations that hard-code a CUDA device. The evaluation set is the first 96 examples of the validation split after the benchmark's length filter, with the benchmark's default counterfactual. The random entry uses one Gaussian seed, whereas MIB averages three uniform seeds. For independent continuous scores both rules then select a uniformly random subset. Ranges give the smallest and the largest value, each rounded to the digits shown, and bounds stated with *at most*, *at least* or *up to* are rounded away from the data. A 95% interval spans the central 95% of the values a quantity takes over 2,000 resamples of the validation templates, drawn with replacement, and differences and shares are computed within each resample. The intervals in Section B use the percentiles stated there.

Public score files. Table 14 lists the submissions whose score files we used, with the share of negative scores and the mean Jaccard overlap between the two rules' circuits, and Figure 6 plots the overlap for each file. If a share q of the n largest-magnitude edges is positive and at least n edges are positive, the two circuits share qn of $2n - qn$ edges before pruning, a Jaccard overlap of $q/(2-q)$, which matches the measured overlap within 0.03 at every $k \geq 0.05$ on the GPT-2 files.

Table 14: **Public score files by submission.** Counts and ranges over all model and task cells each submission covers. The file of Submission B for MCQA with Qwen-2.5 is a copy of its GPT-2 file. The FlashTrace files score edges with the package of the attribution method FlashTrace (Pan et al., 2026).

Submission	Repository	Files	Negative, %	Jaccard
MIB example	`mib-bench/mib-circuits-example`	12	26 to 51	0.29 to 0.35
Warm-start EP	`PMon/mib-circuits-warmstart-edge-pruning-z-attribution`	3	50	0.30 to 0.36
Submission A	`mingyang26/mib-circuit-localization`	3	50 to 51	0.33 to 0.35
Submission B	`my26w/mib-circuit-localization-v2`	2	50	0.38
MAttr	`aryaman/mattr-mib-circuits`	12	79 to 91	0.47 to 0.68
IRP	`stanleyh12/mib-irp-circuits`	12	0	1.00
FlashTrace	`stanleyh12/mib-flashtrace-circuits`	11	0	1.00

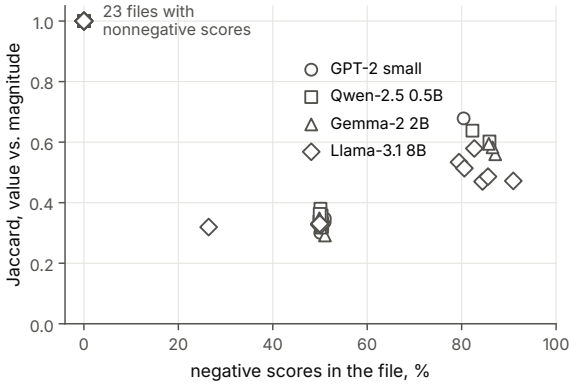


Figure 6: **Overlap of the circuits of the two rules.** Mean Jaccard overlap between the value-selected and magnitude-selected circuits over the nine sizes below one, one point per public score file, against the share of negative scores in the file. Files with nonnegative scores give identical circuits by definition.

Calibration against the benchmark's files. The leaderboard repository stores 84 baseline curves. Leaving out the 14 arithmetic curves, which MIB publishes as one column for two operations,

70 have a published CPR, and 22 of them reproduce it to two decimals. The one for EAP with counterfactual ablation on IOI with GPT-2 gives CPR $= 1.20$ with $O = 0.21$, $U = 0.006$ and peak 1.32, while our evaluation of the MIB example gives 1.24, $O = 0.24$, $U = 0.005$ and peak 1.38. The same stored curve gives CMD $= 0.21$, whereas the published CMD is 0.03, so the published CMD comes from another circuit. Of the matching curves, 18 undershoot more than they overshoot, in line with the two-sided reading of Section 5. Two changes in the benchmark's code history can separate a stored curve from a published number. Until commit a150644 of 10 May 2025 the evaluation integrated by default with logarithmic weights over the realized edge counts, and until commit 90efa38 of 17 May 2025 the evaluation script parsed its selection flag with `type=bool`, so that any value it received, `False` included, selected edges by magnitude. Since then the script has defaulted to value selection.

Circuit sizes, curves and averaging. Table 15 gives the realized edge counts at three sizes and Figure 7 all curves. The benchmark averages the metric over examples before normalizing, so deviations of opposite sign cancel. For magnitude-selected circuits, the share of example and size pairs on which the circuit exceeds the model on that example ranges from 31% to 47%, although their averaged curves stay near one.

Table 15: **Edges of the evaluated circuit under value selection and magnitude selection at three nominal sizes, GPT-2 small, IOI.** The library's pruning can leave edges whose source has lost its inputs, and `evaluate_graph` prunes again before patching, so the counts can fall below the selected $\lfloor kE \rfloor$. Pruning does not change the metric.

	$k = 0.005$		$k = 0.05$		$k = 0.5$	
Entry	value	magnitude	value	magnitude	value	magnitude
MIB example	87	138	1,260	1,580	13,548	16,245
Warm-start EP	138	143	1,413	1,596	14,407	16,245
Submission A	96	135	1,498	1,569	15,636	16,245
Submission B	138	145	1,412	1,596	14,405	16,245
MAttr	158	156	1,619	1,601	16,245	16,245
IRP	0	0	1,159	1,159	16,245	16,245
Random	0	0	121	583	16,038	15,215

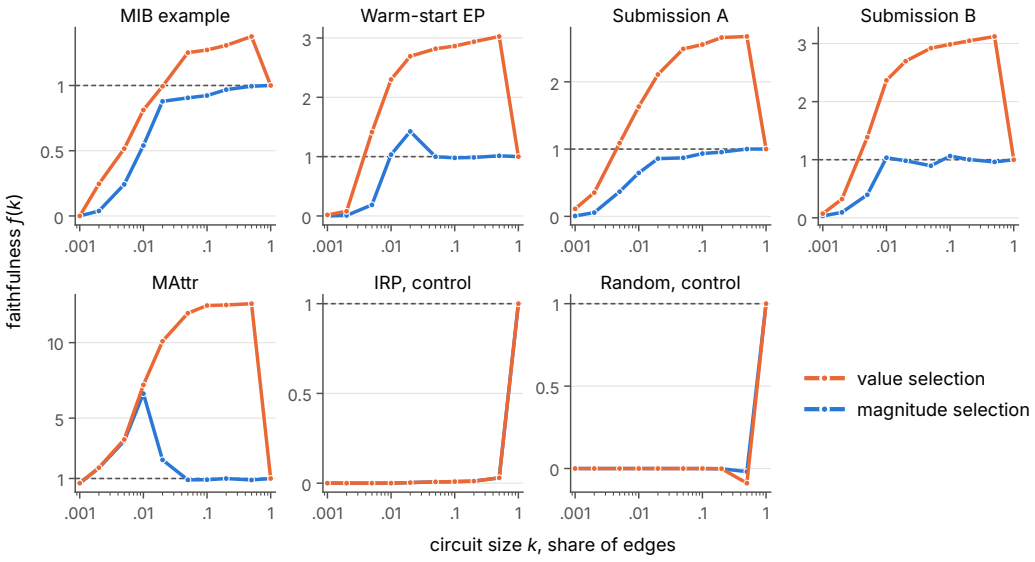


Figure 7: **Faithfulness curves of all seven entries under both rules, GPT-2 small, IOI, validation examples.** For IRP the two curves coincide.

Transfer to the test templates. The benchmark reports on its public test split, whose templates differ from the validation templates. We repeated the evaluation on the first 100 examples of that

split for all seven entries, taking $f(1) = 1$ from its definition. Table 16 reports the results. The value-selected circuits of Warm-start EP, Submissions A and B and MAttr exceed the model on the same example in 100% of example and size pairs with $k \geq 0.05$, and those of the MIB example in 68%. Those of the five signed submissions undershoot by at most $U = 0.006$ and overshoot by 0.25 to 7.11, magnitude-selected circuits overshoot by at most 0.070, and the reported pair lies outside the cone for all five submissions, as on the validation examples. Averaging over examples hides deviations of opposite sign. Measured example by example and then averaged, the distance of the magnitude-selected circuits from the model is 1.4 to 2.6 times their CMD on the validation examples.

Table 16: **Replication on the test split.** Columns as in Table 2.

	CPR ↑		CMD ↓		Overshoot O		Undershoot U		
Entry	value	magnitude	value	magnitude	value	magnitude	value	magnitude	Peak
MIB example	1.242	0.964	0.254	0.035	0.248	0.000	0.005	0.035	1.38
Warm-start EP	2.223	1.006	1.229	0.027	1.227	0.017	0.002	0.010	2.70
Submission A	2.064	0.975	1.069	0.024	1.067	0.000	0.002	0.024	2.48
Submission B	2.279	0.995	1.283	0.018	1.281	0.007	0.002	0.011	2.77
MAttr	8.111	1.024	7.112	0.115	7.112	0.070	0.000	0.045	10.72
IRP	0.253	0.253	0.746	0.746	0.000	0.000	0.746	0.746	1.00
Random	0.223	0.247	0.776	0.752	0.000	0.000	0.776	0.752	1.00

The recommended distance. Table 17 gives the per-example distance for every entry, the trapezoid area over the grid of the mean absolute difference between the circuit's and the model's logit difference on each example, normalized like f. Under magnitude selection it separates the four submissions other than MAttr by at most 0.006 on the validation examples and 0.012 on the test examples, in different orders, so at this sample size it does not rank them. For value-selected circuits that exceed the model on nearly every example it matches their CMD.

Table 17: **Per-example distance.** CMD and per-example distance under both rules, GPT-2 small, IOI, validation examples, and the per-example distance under magnitude selection on the test examples.

	value		magnitude		magnitude, test
Entry	CMD	distance	CMD	distance	distance
MIB example	0.248	0.407	0.030	0.057	0.049
Warm-start EP	1.457	1.457	0.024	0.063	0.060
Submission A	1.206	1.209	0.030	0.058	0.048
Submission B	1.531	1.531	0.030	0.063	0.058
MAttr	8.469	8.469	0.128	0.175	0.150
IRP	0.734	0.734	0.734	0.734	0.746
Random	0.785	0.785	0.757	0.757	0.752