

Global Information Thresholding for Sufficient and Necessary Circuits

Jegyeong Cho*

Visual Intelligence Lab., KAIST

j2k0618@kaist.ac.kr

Abstract

Mechanistic interpretability seeks circuits that are both sufficient on their own and necessary to a model’s behavior under intervention, yet many circuit-discovery pipelines still rely on manually fixed circuit budgets. We propose a procedure that scores edges with sign-preserving integrated gradients and then chooses a single global cut-off by a performance-retention criterion for circuit selection. This makes circuit size a consequence of retained behavior rather than a manually selected budget. On the Mechanistic Interpretability Benchmark, this retention-calibrated thresholding yields circuits that are competitive on CPR/CMD across multiple tasks and models. On our GPT-2 IOI probability-space diagnostic, it improves sufficiency relative to the EAP-IG baseline while also improving the reported necessity diagnostics. Threshold ablations and stability checks further show that the selected operating point yields the best overall balance between self-sufficiency and necessity-oriented diagnostics.

1. Introduction

Understanding how deep neural networks perform specific tasks is a central goal of mechanistic interpretability [2, 15, 25, 27]. A common approach is to identify sub-networks of neurons [4] and weights that implement well-defined functions or behaviors—so-called “circuits” [4, 7, 14]. Extracting such circuits lets us describe the internal algorithm the model uses as an information flow [7] over constituent components. However, reliably isolating these circuits remains difficult [13, 18, 20, 29]: many methods either include unnecessary pieces or miss indispensable ones, weakening explanatory power and reproducibility. Small input perturbations led to substantially different sets of selected connections, and the resulting circuits did not always capture all and only the important edges. This instability motivates a more principled approach to circuit extraction.

An interpretable circuit should not merely look plausible. It should satisfy two criteria simultaneously. First, it should be *sufficient* on its own: running only the circuit should preserve almost all of the original model’s behavior. Second, it should be *necessary*: removing the circuit should cause the model’s performance to drop substantially [5]. A circuit that is self-sufficient but dispensable is not a compelling explanation, and conversely a circuit whose removal is devastating but which cannot reproduce the behavior by itself is also incomplete. Existing automatic circuit extraction methods have found it difficult to satisfy these two conditions at once [4, 7, 14]. Some methods preserve behavior by including many connections, while aggressive pruning to maximize necessity often causes the circuit alone to fail to reproduce the model’s behavior.

In practice, many circuit-discovery pipelines follow the same two-stage recipe: they assign an importance score to each edge and then impose a fixed selection budget. This design leaves the final circuit highly sensitive to the chosen budget, which is often set manually and does not transfer cleanly across tasks or models. It also induces the central trade-off of this paper: circuits that are too small cannot capture necessary and sufficient meaningful edges, whereas circuits that are too large preserve more behavior but become less selective. Recent work has shown that this sensitivity is not merely conceptual but operational, with circuit structure changing under prompt perturbations, input sampling changes, or hyperparameter choices [13, 18, 20].

This distinction between fixed top- k selection and retention-calibrated thresholding is subtle but important. A fixed top- k rule chooses circuit size *ex ante* and then asks what behavior survives under that budget, whereas our procedure reverses the order: it first specifies the retained-behavior target and then lets the observed score distribution determine the effective operating point. In settings where score mass is highly concentrated in a relatively small subset of edges, these two views can lead to markedly different circuit sizes across tasks and models. The practical question is therefore not only how to score edges, but also how to map those scores into a circuit in a way that preserves behavior without sacrificing selectivity.

*This work was carried out under the supervision of Prof. Kuk-Jin Yoon, the author’s academic advisor, at the Visual Intelligence Laboratory, KAIST (kjyoon@kaist.ac.kr).

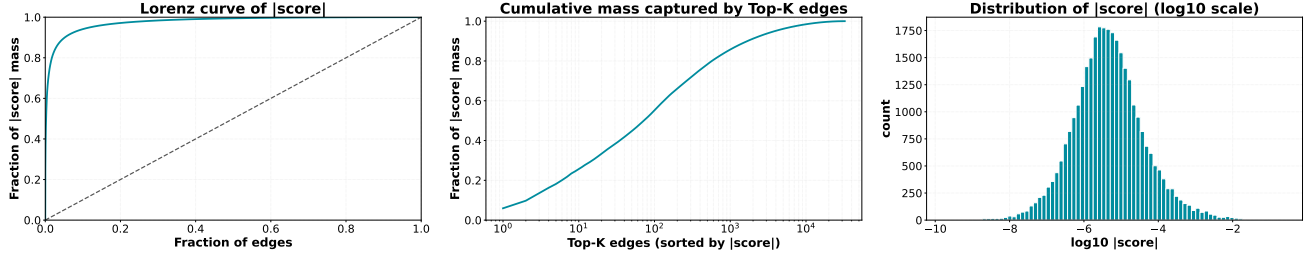


Figure 1. **Edge importance landscape produced by EAP-IG.** (Left) The Lorenz curve of $|\text{score}|$ (dashed line: uniform line) shows that the mass is extremely concentrated on a small fraction of top edges. (Middle) The cumulative mass of the top- K edges (horizontal axis in log scale) reveals a clear head-long-tail structure. (Right) The distribution of $\log_{10} |\text{score}|$ indicates a heavy-tailed (quasi log-normal) shape. This empirical distribution motivates circuit selection based on a single global threshold calibrated by retained behavior rather than by a manually fixed top- k budget: overly aggressive shrinkage degrades sufficiency, whereas overly lax thresholds reduce selectivity.

We propose an alternative. Rather than prescribing a fixed top- k circuit budget, we select the smallest thresholded circuit that satisfies a target performance-retention criterion. Concretely, we compute a signed edge score, apply a single global cutoff chosen from the observed score distribution, reinforce near-threshold edges with a narrow buffer band, and repair broken input-to-logits connectivity when necessary. The resulting pipeline is deliberately intuitive: circuit size is determined by retained behavior rather than by a manually chosen budget. Put differently, each threshold induces an effective circuit size, but unlike fixed top- k extraction the size is chosen by a retention criterion rather than imposed in advance. This distinction matters because score thresholding and fixed top- k truncation can coincide only after a budget has been chosen. Our procedure instead lets the retained-behavior target determine which operating point along the score ranking is selected for a given task and model. On the Mechanistic Interpretability Benchmark (MIB) [19], this retention-calibrated thresholding yields circuits that are sufficient, necessary, and competitive on CPR/CMD across multiple tasks and models.

Our empirical picture is correspondingly simple. First, the main benchmark results show that our approach is best or second-best on CPR and CMD across most settings. Second, probability-space diagnostics on IOI show that the method improves sufficiency relative to the strongest EAP-IG baseline while maintaining strong necessity. Third, threshold ablations reveal a non-monotonic trade-off: overly small circuits lose both sufficiency and necessity, whereas overly large circuits preserve more behavior but become less sufficient. Additional diagnostics show that this intermediate operating regime is not fragile: performance varies only mildly across retention targets, remains stable under boundary reinforcement and direct sample perturbations, and exhibits a more spatially concentrated pattern of method-specific unique edges than the strongest EAP-IG baseline.

We make three contributions:

- **A circuit-extraction perspective centered on sufficiency and necessity:** We argue that explanatory circuits should be evaluated by whether they are jointly self-sufficient and indispensable under intervention, and we operationalize this with negative log-likelihood-based sufficiency and necessity metrics.
- **A retention-calibrated alternative to fixed-budget circuit selection:** Instead of prescribing circuit size in advance, we choose, within the induced score ranking, the smallest thresholded circuit that satisfies a target performance-retention criterion, with light boundary reinforcement and connectivity repair.
- **Strong empirical performance on MIB with additional stability diagnostics:** Ours achieves comparable CPR/CMD and improves sufficiency relative to EAP-IG baselines while maintaining necessity, additional diagnostics show mild retention sensitivity, stable behavior under boundary perturbations, and a more spatially concentrated pattern of method-specific unique edges.

2. Related Work

2.1. Circuits and Causal Interventions

Mechanistic interpretability seeks to explain model behavior by reverse-engineering the internal computations implemented by a network’s parameters and activations rather than by treating the model as an opaque input-output mapping [2, 15, 25, 27]. Within this literature, a *circuit* is typically understood as a sparse subnetwork of components—neurons, channels, attention heads, or edges between them—that together implement a recognizable function. This perspective was popularized in the circuits program and related feature-visualization work, which argued that meaningful algorithms can often be localized to structured interactions among internal units rather than isolated neurons alone [21–23].

A central methodological thread in this area is *causal intervention*. Rather than only correlating internal activations

with outputs, intervention-based methods actively modify hidden states or information flow and measure the resulting behavioral change. Activation patching and path patching are representative examples: they replace or route intermediate computations so as to test whether specific components or pathways are causally important for a downstream behavior [12, 14]. More formal treatments of this causal viewpoint come from the causal-abstraction literature, which frames interpretability as aligning a high-level causal model with lower-level neural mechanisms and validating the alignment through interventions [10, 11].

These tools enabled a number of influential manual reverse-engineering case studies. Wang et al. [30] identify a circuit in GPT-2 small for indirect object identification, organizing the discovered heads into functionally distinct classes and evaluating the resulting explanation with quantitative notions of faithfulness and completeness. More recently, McDougall et al. [17] reverse-engineer the “copy suppression” motif in attention heads, showing how a recurring computation can be analyzed mechanistically rather than only behaviorally. Together, such studies demonstrate that nontrivial language-model behaviors can be decomposed into structured internal mechanisms, but they also highlight the practical cost of manual circuit construction and the difficulty of scaling such analyses across tasks, models, and prompts.

2.2. Automated Circuit Discovery, Score-to-Circuit

To reduce the manual burden of intervention studies, recent work has developed automated circuit-discovery pipelines. A prominent family is attribution-based discovery, which assigns an importance score to candidate components or edges and then extracts a sparse subgraph from those scores. Conmy et al. [4] propose an influential framework for automated circuit discovery, while Syed et al. [29] show that attribution patching can outperform several automated alternatives in recovering task-relevant structure. Hanna et al. [13] further sharpen the discussion by emphasizing that high overlap with a reference circuit does not by itself guarantee mechanistic faithfulness.

Within this automation literature, several choices recur. First, methods differ in *what* they score: nodes, edges, paths, or latent variables. Second, they differ in *how* they induce a circuit from those scores. Some approaches rely on explicit routing or graph construction, such as Information Flow Routes (IFR), which uses attribution to extract directed information-flow structure at scale [7]. Others use optimization over masks or sparse subnetworks, learning binary or soft masks that preserve task performance while minimizing the retained structure [8, 9, 26, 32]. Yet others shift attention away from edges entirely and toward features or latent variables, including distributed alignment search and sparse-autoencoder-based decompositions [11, 16, 24].

Despite their differences, many of these pipelines share a common architecture: they first compute a ranking or score over candidate structures and then apply an external sparsification rule to obtain the final circuit. In practice, the quality of the resulting explanation often depends not only on the score itself but also on how aggressively the circuit is pruned, which components are grouped together, and which budget or threshold is chosen. This makes it difficult to separate improvements in attribution quality from improvements in the score-to-circuit mapping rule.

2.3. Faithfulness, Sufficiency, Necessity, Stability

A growing body of work argues that interpretability methods should be judged by *faithfulness* rather than by plausibility or visual appeal alone. In the attribution literature, sanity checks and related evaluation criteria have shown that seemingly compelling explanations can be insensitive to model parameters or can fail to track the mechanisms that genuinely drive predictions [1, 31]. In mechanistic interpretability specifically, Hanna et al. [13] argue that circuit overlap with a hypothesized reference circuit is at best an incomplete proxy: a discovered circuit may omit genuinely necessary components or include components that are irrelevant to the true mechanism.

This concern connects naturally to the distinction between *sufficiency* and *necessity*. In explanation research more broadly, these notions are often treated separately: one asks whether an explanation alone preserves the model’s behavior, and the other asks whether removing the explanation substantially degrades that behavior [3, 5]. Mechanistic circuit discovery inherits the same tension. Highly permissive circuits can be sufficient in the sense that they preserve output behavior, yet they may fail to be necessary because they include many superfluous parts. Conversely, overly sparse circuits can look clean and necessary under ablation while failing to reproduce the model’s behavior on their own. Our paper builds directly on this distinction and treats joint sufficiency and necessity as the central target of circuit extraction rather than as secondary diagnostics.

A second major concern is *stability*. Méloux et al. [18] show that current mechanistic-interpretability pipelines can produce substantially different circuits under small perturbations, raising the question of whether the extracted mechanism is identifiable at all. Nainani et al. [20] further emphasize that circuits can be context-sensitive and can adapt across distributional conditions, complicating any notion of a single immutable mechanism. These findings make two points especially relevant for our setting: first, a high-scoring circuit is not necessarily a stable one. Second, the final selection rule matters, because small changes near the sparsification boundary can induce qualitatively different circuits even when the underlying score distribution changes only slightly.

3. Method

We propose a global-threshold-based circuit extraction method to improve mechanistic interpretability [2, 15, 25, 27]. The central goal is not simply to rank important edges, but to extract circuits that are jointly sufficient and necessary. To this end, we first define sufficiency and necessity as the criteria for a circuit explanation, then describe a signed edge-importance estimate, and finally present a retention-calibrated thresholding rule that builds a circuit.

3.1. Sufficiency and Necessity of Circuit

An interpretable circuit is a set of internal connections that accurately and succinctly explains the model’s mechanism of operation, and to this end it must satisfy two criteria: sufficiency and necessity. Sufficiency indicates whether the extracted circuit alone preserves almost all of the original model’s performance, and necessity indicates whether removing that circuit causes the model’s performance to drop substantially [5].

More formally, let the full neural network be f , and define the restricted model that uses a circuit $C \subseteq E$ as f_C (that is, we zero out the outputs of all connections not in C). Conversely, denote by f_{-C} the model obtained by removing only the connections in circuit C . Let $L(f)$ denote the average negative log-likelihood (NLL) over the task data.

Sufficiency. The circuit C alone should be able to deliver performance nearly identical to that of the original model. Expressed as the sufficiency loss,

$$\Delta L_{\text{suff}}(C) = L(f_C) - L(f) \approx 0. \quad (1)$$

The closer $\Delta L_{\text{suff}}(C)$ is to zero, the more the circuit alone reproduces the performance of the original model.

Necessity. Removing the circuit C should prevent the model from solving the task properly. This means that the task loss should increase significantly when the circuit is ablated. Expressed as the necessity loss,

$$\Delta L_{\text{nec}}(C) = L(f_{-C}) - L(f) \gg 0, \quad (2)$$

so larger values indicate stronger necessity.

An ideal explanatory circuit C is therefore one that keeps $\Delta L_{\text{suff}}(C)$ close to zero while making $\Delta L_{\text{nec}}(C)$ large. In other words, the circuit alone achieves performance comparable to the original model (high sufficiency), and the circuit is indispensable in that performance collapses without it (high necessity). However, existing automatic circuit extraction methods have found it difficult to satisfy these two conditions simultaneously [5, 18, 20].

3.2. Loss-Based Edge Importance Estimation

To identify the key connections that constitute a circuit, we first compute for each network connection (edge) a quantitative importance score for its contribution to the model’s predictions. Specifically, letting f_{-e} denote the model f with a particular connection $e \in E$ removed, we define the importance I_e of connection e as

$$I_e = L(f_{-e}) - L(f). \quad (3)$$

That is, I_e measures how much the task loss changes when edge e is removed.

This importance is estimated using integrated gradients over a continuous edge gate [6, 13, 28]. We retain the sign for analysis, but the circuit extraction rule itself operates on the score magnitude, so larger $|I_e|$ indicates a stronger task-level effect.

3.3. Circuit Extraction via a Global Threshold

After obtaining the importance I_e for all connections, we decide which connections to include in the circuit by applying a single global threshold τ . We define

$$C(\tau) = \{e \in E \mid |I_e| \geq \tau\}. \quad (4)$$

The choice of τ is the key factor that controls circuit size and the degree of performance retention.

Setting τ high selects only highly influential connections, making the circuit compact and selective by avoiding meaningless and confounding edges. If too strict, however, important connections may be omitted and both necessity and sufficiency may drop. Conversely, lowering τ includes many connections even with small effects, thereby increasing necessity but risking over-inclusion. Therefore, τ should be chosen to keep the circuit concise while preserving the model’s behavior as much as possible.

To this end, we use a target performance-retention criterion. For example, if we desire a circuit that retains at least 95% of the original model’s behavior, or whose sufficiency loss $\Delta L_{\text{suff}}(C)$ remains below a specified tolerance, we adjust τ so that the circuit $C(\tau)$ satisfies the performance condition. In practice, τ is tuned by automated binary search over the sorted score magnitudes on a calibration split. This replaces manually fixed circuit budgets-including fixed top- k variants and other budget-based sparsification rules-with a retention-calibrated global cutoff.

We also take care around the threshold boundary. If the importance of a connection lies just below τ , excluding it due to tiny changes from data sampling or numerical noise can sharply degrade sufficiency [1, 5, 13, 18, 20]. To prevent this, we introduce a buffer ratio β : near-boundary candidates satisfy

$$|I_e| \in [\tau(1 - \beta), \tau), \quad (5)$$

and a fixed fraction of these candidates ranked by $|I_e|$ is conservatively included in the circuit. This mitigates the brittleness of a hard cutoff and stabilizes the selected circuit around the threshold.

Finally, after extracting $C(\tau)$, we run a graph-theoretic reachability check on the directed circuit graph to ensure that at least one path from input to logits remains. If thresholding breaks all such paths, we repair connectivity using candidate edges until reachability is restored. The final circuit is therefore the smallest thresholded circuit within this score-induced ranking family that satisfies the retention criterion after boundary stabilization and connectivity repair.

4. Experiments

We evaluate our methods across a diverse set of models and tasks [2, 25, 27]. All experiments follow the Mechanistic Interpretability Benchmark (MIB) [19] circuit track with four task families: Indirect Object Identification (IOI), Arithmetic (addition/subtraction), Multi-Choice QA (MCQA), and ARC-Easy/Challenge (ARC), spanning syntactic binding, symbolic manipulation, and world-knowledge reasoning [19]. We consider four language models (GPT-2, Qwen-2.5, Gemma-2, Llama-3.1) to test that our single-threshold rule is not tied to a specific architecture or scale. For every method, a subgraph (“circuit”) is first extracted and then evaluated under a unified counterfactual patching harness [12, 14]: sufficiency is measured by running only the subgraph (ablating all other edges), and necessity by ablating the subgraph from the full model. This protocol is identical for all methods.

4.1. Baselines and Evaluation Metrics

Baselines. We compare *Ours* against representative circuit-extraction baselines, summarized in Table 1 and Table 2:

- **EActP (CF)**: Effective Activation Patching tuned by counterfactual held-out performance [14, 25].
- **EAP (mean/CF/OA)**: Edge Attribution Patching with different score-to-circuit mappings and ablation settings [29].
- **EAP-IG-inputs/activations (CF)**: EAP combined with Integrated Gradients on inputs or hidden activations to capture long-range edge importance [13].
- **NAP and NAP-IG (CF)**: Node-level analogues of EAP (with/without IG), operating at a coarser granularity.
- **IFR**: Information Flow Routes, which traces high-magnitude information paths without repeated counterfactual testing [7].
- **UGS**: Uniform Gradient Sampling, a sparsification baseline reported in MIB [19].

Many of these methods first produce edge scores and then convert them into circuits using manually fixed budgets or other budget-based sparsification rules. Our comparison

therefore asks whether replacing a manually fixed circuit budget with a retention-calibrated global cutoff improves circuit faithfulness under the same benchmark.

Primary metrics. Table 1 reports the integrated circuit performance ratio (CPR) and Table 2 reports the integrated circuit-model distance (CMD) and InterpBench AU-ROC, which measure how well the circuits match the full model’s behavior and faithfulness under MIB [19]. CPR measures the area under the faithfulness-versus-circuit-size curve, rewarding methods that consistently recover components with positive task-relevant effects across circuit sizes. CMD measures the area between the faithfulness curve and the ideal line, capturing how closely a discovered circuit matches the full model’s task behavior across circuit sizes, with lower being better.

Secondary metrics. We define sufficiency and necessity formally using negative log-likelihood: $\Delta\text{NLL}_{\text{suff}}$ and $\Delta\text{NLL}_{\text{nec}}$. For sufficiency, values closer to zero indicate that the circuit alone better preserves the behavior of the full model. For necessity, larger values indicate that removing the circuit causes a greater degradation relative to the full model, implying that the removed circuit was more necessary for the original behavior. We additionally report ΔBrier score and ΔECE as auxiliary calibration diagnostics under circuit-only and circuit-ablation interventions. For sufficiency, values closer to zero are preferable because they indicate that the circuit-only model stays closer to the full model’s behavior. On the necessity side, ΔBrier score and ΔECE values closer to zero are interpreted as less unfavorable auxiliary calibration shifts, not as stronger formal necessity, since circuit ablation can flatten the output distribution toward uniform and thereby trivially reduce both metrics regardless of correctness.

4.2. Main Results Across Benchmarks

We now consider the full benchmark suite in Table 1 and Table 2. Overall, *Ours* is either the best or second-best method on CPR and CMD across most task-model pairs. This headline benchmark picture is not confined to MIB alone: on InterpBench, *Ours* attains AUROC 0.79, second only to EAP-IG-activations at 0.81 and ahead of most remaining baselines. More importantly, the MIB results are consistent with the sufficiency-necessity picture developed above. Methods that chase CPR alone often pay with worse CMD or with circuits that become less selective, whereas methods that maximize necessity via aggressive pruning typically destroy sufficiency [5, 13, 18].

On IOI, *Ours* is at or near the top across all four models. For GPT-2, EActP (CF) attains the highest CPR (2.30), but *Ours* is a close second (2.01) with competitive CMD and extends to larger models where EActP is undefined. On

Table 1. **CPR** scores on the public test set across circuit localization methods and ablation types . All evaluations are assumed under counterfactual ablations. **Higher is better**. Arithmetic is averaged over addition/subtraction. We bold and underline the best and second-best per column, respectively.

Method	IOI				Arithmetic	MCQA			ARC (E)		ARC (C)
	GPT-2	Qwen-2.5	Gemma-2	Llama-3.1	Llama-3.1	Qwen-2.5	Gemma-2	Llama-3.1	Gemma-2	Llama-3.1	Llama-3.1
Random	0.25	0.28	0.30	0.25	0.25	0.27	0.32	0.26	0.32	0.26	0.25
EActP (CF)	2.30	1.21	-	-	-	0.85	-	-	-	-	-
EAP (mean)	0.29	0.71	0.68	0.98	0.35	0.29	0.33	0.13	0.26	0.34	0.80
EAP (CF)	1.20	0.26	1.29	0.85	0.55	0.85	1.49	1.00	1.08	0.80	0.82
EAP (OA)	0.95	0.70	-	-	-	0.29	-	-	-	-	-
EAP-IG-inputs (CF)	1.85	<u>1.63</u>	<u>3.20</u>	2.08	<u>0.99</u>	<u>1.16</u>	1.64	1.05	1.53	<u>1.04</u>	0.98
EAP-IG-activations (CF)	1.82	<u>1.63</u>	2.07	1.60	0.98	0.77	1.57	0.79	<u>1.70</u>	0.71	0.63
NAP (CF)	0.28	0.30	0.30	0.26	0.27	0.38	1.47	<u>1.69</u>	1.01	0.26	0.26
NAP-IG (CF)	0.76	0.29	1.52	0.42	0.39	0.77	1.71	1.87	1.53	0.26	0.26
IFR	0.58	0.31	0.25	0.09	0.89	0.40	0.38	0.52	0.34	0.36	0.24
UGS	0.97	0.98	-	-	-	1.17	-	-	-	-	-
Ours	<u>2.01</u>	1.92	3.72	<u>1.96</u>	1.01	0.73	<u>1.65</u>	1.02	1.75	1.25	1.40

Table 2. **CMD** scores on the public test set across circuit localization methods and ablation types (**lower is better**), and **AUROC** for InterpBench (**higher is better**). All evaluations are assumed under counterfactual ablations. Arithmetic scores are averaged across addition and subtraction. We bold and underline the best and second-best per column, respectively.

Method	InterpBench (↑)	IOI				Arithmetic	MCQA			ARC (E)		ARC (C)
	AUROC	GPT-2	Qwen-2.5	Gemma-2	Llama-3.1	Llama-3.1	Qwen-2.5	Gemma-2	Llama-3.1	Gemma-2	Llama-3.1	Llama-3.1
Random	0.44	0.75	0.72	0.69	0.74	0.75	0.73	0.68	0.74	0.68	0.74	0.74
EActP (CF)	0.28	0.02	0.49	-	-	-	0.36	-	-	-	-	-
EAP (mean)	0.78	0.29	0.18	0.25	0.04	0.07	0.21	0.20	0.16	<u>0.22</u>	0.28	0.20
EAP (CF)	0.73	<u>0.03</u>	0.15	0.06	0.01	<u>0.01</u>	<u>0.07</u>	0.08	0.09	0.04	0.11	<u>0.18</u>
EAP (OA)	0.77	0.30	0.16	-	-	-	0.11	-	-	-	-	-
EAP-IG-inp. (CF)	0.71	<u>0.03</u>	<u>0.02</u>	<u>0.04</u>	0.01	0.00	0.08	<u>0.06</u>	0.14	0.04	0.11	0.22
EAP-IG-act. (CF)	0.81	<u>0.03</u>	0.01	0.03	0.01	0.00	0.05	0.07	<u>0.13</u>	0.04	0.30	0.37
NAP (CF)	0.30	0.38	0.33	0.37	0.29	0.28	0.30	0.35	0.32	0.33	0.69	0.69
NAP-IG (CF)	0.62	0.27	0.20	0.26	0.19	0.18	0.18	0.29	0.33	0.28	0.67	0.67
IFR	0.71	0.42	0.69	0.75	0.83	0.22	0.60	0.62	0.48	0.66	0.64	0.76
UGS	0.74	<u>0.03</u>	0.03	-	-	-	0.20	-	-	-	-	-
Ours	<u>0.79</u>	<u>0.03</u>	0.03	<u>0.04</u>	0.01	0.00	0.14	0.05	<u>0.13</u>	0.04	<u>0.15</u>	0.13

Qwen-2.5 and Gemma-2, Ours achieves the best CPR (1.92 and 3.72) and among the lowest CMDs, and on Llama-3.1, CPR (1.96) is essentially tied with the best EAP-IG baseline (2.08) while keeping CMD at 0.01. Thus a single global threshold, without manually choosing a model-specific fixed top- k budget, is enough to remain competitive.

On Arithmetic (Llama-3.1 addition and subtraction), Ours is particularly strong, with CPR 1.01 and CMD 0.00, effectively retaining distributional behavior. Several baselines also achieve near-ideal Arithmetic behavior, but Ours remains among the strongest settings with CPR 1.01 and CMD 0.00, consistent with the retention-calibrated rule isolating a compact, low-distortion circuit in this setting [4].

On ARC-Easy and ARC-Challenge, Ours again performs strongly across the reported columns. it attains CPR 1.75, surpassing EAP-IG-activations (next best 1.70), and for Llama-3.1 it reaches CPR 1.25 on ARC-Easy and 1.40 on ARC-Challenge, while baselines remain below 1.1. CMD is likewise competitive: Ours attains joint-lowest CMD 0.04 on Gemma-2 ARC-Easy and among the lowest CMDs elsewhere. These reasoning-heavy tasks are there-

fore also handled by the same simple threshold rule [19].

The MCQA results are more mixed and show where we trade off against more permissive fixed top- k baselines [13, 29]. On Qwen-2.5 MCQA, CPR and CMD (0.73 and 0.14) lag behind the NAP-IG/EAP-IG variants. On Gemma-2 MCQA, Ours nearly matches the best CPR (1.65 vs. 1.71) while achieving the lowest CMD (0.05), and on Llama-3.1 MCQA our CPR is moderate (1.02) but CMD remains low (0.13), whereas methods reaching CPR up to 1.87 suffer significantly larger CMD (≥ 0.33). In other words, Ours trades a small amount of MCQA accuracy for circuits.

Overall, Ours is the best or second-best method on CPR and CMD across most task-model pairs, and it does so without task-specific objectives, fixed top- k budgets, or architecture tuning. The method is therefore not optimized for a single benchmark column in isolation. It is optimized to recover circuits that are jointly sufficient and necessary while remaining faithful in probability space. Read together with the retention and perturbation diagnostics above, these aggregate results suggest that the observed trade-off is a stable property of the selection rule rather than an artifact of one

Table 3. **Sufficiency and necessity metrics on GPT-2 IOI.** Comparison between the EAP-IG-inputs (CF) baseline and our circuit in terms of ΔNLL and changes in Brier score and ECE. Lower $\Delta\text{NLL}_{\text{suff}}$, $\Delta\text{Brier}_{\text{suff}}$, and $\Delta\text{ECE}_{\text{suff}}$ indicate a more self-sufficient and better calibrated circuit. For necessity, we report $\Delta\text{NLL}_{\text{nec}}$ for necessity and $\Delta\text{Brier}_{\text{nec}}$, and $\Delta\text{ECE}_{\text{nec}}$ for auxiliary metrics.

Method	$\Delta\text{NLL}_{\text{suff}} \downarrow$	$\Delta\text{NLL}_{\text{nec}} \uparrow$	$\Delta\text{Brier}_{\text{suff}} \downarrow$	$\Delta\text{Brier}_{\text{nec}} \uparrow$	$\Delta\text{ECE}_{\text{suff}} \downarrow$	$\Delta\text{ECE}_{\text{nec}} \uparrow$
EAP-IG-inputs (CF)	1.009	8.328	0.1550	-0.1977	0.2010	-0.2570
Ours	0.7579	8.431	0.1005	-0.1077	0.1415	-0.1203

Table 4. **Ablation on global threshold τ for Ours circuits.** Varying τ controls circuit size (75, 750, 1500 edges) and induces a non-monotonic trade-off between sufficiency and the reported necessity diagnostics. A very small circuit (75 edges) is highly selective but performs poorly on both sufficiency and the reported necessity metrics, while a larger circuit (1500 edges) preserves more behavior but exhibits larger calibration shifts. The intermediate setting (750 edges, $\sim 95\%$ performance retained) yields the best overall balance, simultaneously achieving low $\Delta\text{NLL}_{\text{suff}}$, the strongest NLL-based necessity score, and the least unfavorable Brier/ECE.

Edges	$\Delta\text{NLL}_{\text{suff}} \downarrow$	$\Delta\text{NLL}_{\text{nec}} \uparrow$	$\Delta\text{Brier}_{\text{suff}} \downarrow$	$\Delta\text{Brier}_{\text{nec}} \uparrow$	$\Delta\text{ECE}_{\text{suff}} \downarrow$	$\Delta\text{ECE}_{\text{nec}} \uparrow$
75/32491	2.718	5.836	0.2013	-0.1856	0.3024	-0.2434
750/32491	0.7579	8.431	0.1005	-0.1077	0.1415	-0.1203
1500/32491	1.138	8.065	0.1742	-0.1799	0.2390	-0.2395

benchmark-specific hyperparameter choice.

4.3. Circuit Sufficiency and Necessity on IOI

We first examine GPT-2 on IOI [4, 19]. Table 3 compares our circuit with the EAP-IG-inputs (CF) baseline. Ours substantially improves sufficiency: $\Delta\text{NLL}_{\text{suff}}$ decreases from 1.009 to 0.7579. Calibration shows the same pattern: the Brier and ECE are 0.1005 and 0.1415 for Ours versus 0.1550 and 0.2010 for the baseline, meaning our circuit preserves both the output distribution and calibration more faithfully.

On necessity, the NLL-based score also improves: ablating the our circuit yields $\Delta\text{NLL}_{\text{nec}} = 8.431$, compared with 8.328 for the EAP-IG baseline. Thus, under the original necessity definition, Ours is numerically higher in this comparison. For the Brier and ECE necessity gaps, Ours is still more favorable than the baseline because its reported gaps are closer to zero (from -0.1977 to -0.1077 in Brier and from -0.2570 to -0.1203 in ECE). We interpret this as more favorable auxiliary calibration behavior under ablation rather than as a standalone necessity metric. We therefore claim improved sufficiency and a slight improvement in NLL-based necessity, with auxiliary calibration diagnostics moving in the same direction.

4.4. Ablation on the Global Threshold

We study how the global threshold τ affects circuit quality by varying τ on GPT-2 IOI to produce circuits with approximately 75, 750, and 1,500 edges from 32,491 total edges. As shown in Table 4, this ablation directly reveals the central trade-off of the paper.

With a very high threshold (75 edges), the circuit is extremely small and selective. However, this regime performs poorly on both sufficiency and the reported necessity di-

agnostics: $\Delta\text{NLL}_{\text{nec}} = 5.836$ is the weakest of the three settings, and the Brier/ECE necessity gaps are the most unfavorable. At the same time, sufficiency collapses to $\Delta\text{NLL}_{\text{suff}} = 2.718$, and calibration degrades sharply. Thus, the smallest circuit is neither self-sufficient nor favorable on the reported necessity metrics since it disables too many meaningful edges to operate the target task.

At the medium setting (750 edges), our default, $\Delta\text{NLL}_{\text{suff}}$ improves sharply to 0.7579. At the same time, the NLL-based necessity score is strongest at 8.431, and the Brier/ECE gaps are the least unfavorable (-0.1077 and -0.1203). Calibration shifts are also smallest. This is therefore the best overall balance: the circuit can solve the task much more effectively on its own while also achieving the strongest reported necessity diagnostics. Tables 5 and 6 support the same conclusion from a different angle: the intermediate regime is not a fragile consequence of one particular calibrated threshold, but remains stable across retention targets, near-threshold boundary choices, and broader sample perturbations.

At the lowest threshold (1,500 edges), the circuit becomes larger. The NLL-based necessity score remains high (8.065), but sufficiency worsens relative to the 750-edge setting and calibration shifts increase again. It succeeds to contain the meaningful edges for the task but fails to reject unnecessary edges that can degrade the performance. In this regime, the circuit recovers some behavior relative to the 75-edge setting, but it is not the best operating point.

This ablation confirms that the 750-edge operating point is not chosen because it maximizes only one column in isolation, but because it provides the most favorable joint trade-off between sufficiency, calibration, and the reported necessity diagnostics. The retention sweep, the perturbation diagnostics, and the qualitative layer-pair comparison

Table 5. **Effect of retention target on IOI/GPT-2.** Retention target changes yield mild sensitivity: CPR changes by only 0.012 across retention targets 0.90/0.95/0.97/0.99.

Retention	0.90	0.95	0.97	0.99
CPR	2.0041	2.0156	2.0135	2.0119

Table 6. **Boundary-reinforcement stability on IOI/GPT-2.** CPR under two independent sample perturbations for three values of the buffer ratio β . The two perturbations align almost perfectly once even a small buffer is introduced.

	$\beta = 0.0$	$\beta = 0.2$	$\beta = 0.4$
Sample 1	2.0112	2.0117	2.0117
Sample 2	2.0205	2.0115	2.0115

all point to the same interpretation: the selected intermediate circuit is not only strong on scalar metrics, but also stable under small perturbations and structurally more concentrated than the EAP-IG baseline. Taken together, these observations strengthen the claim that our operating point is not a lucky scalar optimum but a robust compromise between self-sufficiency and indispensability.

4.5. Threshold, Stability, and Qualitative Structure.

For IOI/GPT-2, finding τ to meet a 0.95 retention target required 15 evaluations of $f_{C(\tau)}$ over sorted score magnitudes. The sensitivity to the chosen retention target is mild: as shown in Table 5, sweeping the target from 0.90 to 0.99 changes CPR by only 0.012 in total, a very small shift relative to the absolute CPR scale of roughly 2.0. This indicates that our operating regime is not critically dependent on one hand-picked value. This matters because the central methodological claim is not merely that one particular threshold works, but that retained behavior can be used to calibrate circuit size in a stable way across a nontrivial range of retention targets.

Boundary reinforcement is likewise stable. Table 6 reports CPR under two independent sample perturbations for three values of the buffer ratio β . The point of this table is not that one perturbation setting is intrinsically meaningful, but that the selected circuit is not brittle around the threshold boundary: the CPR values remain nearly unchanged across $\beta \in \{0.0, 0.2, 0.4\}$ and across both perturbations. In fact, the only visible deviation occurs for one perturbation at $\beta = 0.0$. Once even a small buffer is introduced, the two perturbations become almost identical. This is exactly the intended role of boundary reinforcement, namely to suppress unstable threshold decisions without materially changing overall circuit quality.

Direct perturbation stability tells the same story. Across six independent sample perturbations, the resulting IOI/GPT-2 circuits achieve $\text{CPR } 2.0135 \pm 0.0022$, indicat-

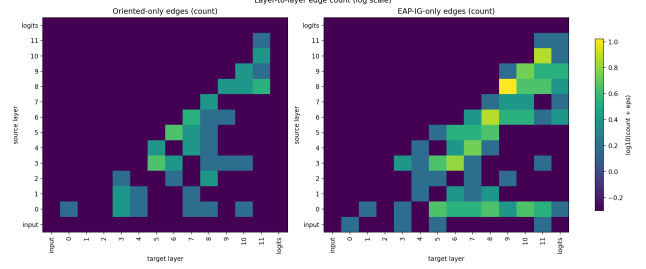


Figure 2. **Unique-edge heatmaps on GPT-2 IOI.** We compare only the edges unique to Ours and the EAP-IG baseline to visualize where the two score-to-circuit rules diverge. Our unique edges are concentrated in fewer source-target layer pairs, whereas the EAP-IG spreads its unique edges across broader layer interactions.

ing that the selected operating regime is stable not only to the exact retention target and threshold buffer, but also to broader changes in the calibration inputs. This matters because instability under small perturbations has emerged as a central concern in automated circuit discovery.

Finally, Figure 2 compares the *method-specific unique edges* discovered by Ours and the EAP-IG baseline on GPT-2 IOI. We do not compare the full circuits directly, since most salient structure is shared between the methods. Instead, we isolate the edges unique to each method to reveal where the selection rules diverge. Our unique edges are concentrated in fewer source-target layer pairs, whereas the EAP-IG baseline distributes its unique edges across a broader set of layer interactions. This suggests that Ours suppresses tail clutter while avoiding widespread expansion of the circuit. It is consistent with the quantitative threshold-ablation story: the retention-calibrated operating point appears structurally more concentrated and less diffuse than the fixed top- k baseline, while still preserving performance.

5. Conclusion

This paper argues for a design rule in automated circuit discovery: evaluate circuits by sufficiency and necessity together, and choose circuit size by retained behavior rather than by a manually fixed budget. In our setting, retention-calibrated global thresholding yields competitive CPR/CMD on Mechanistic Interpretability Benchmark and improves probability-space sufficiency and reported necessity diagnostics in GPT-2 IOI relative to EAP-IG baselines. Threshold ablations indicate that the best operating point is an intermediate regime rather than the smallest or largest circuit. The additional retention and perturbation diagnostics suggest that this regime is reasonably stable rather than a fragile consequence of a single threshold choice.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25473963). The author gratefully acknowledges Prof. Kuk-Jin Yoon (Visual Intelligence Laboratory, KAIST) for his guidance and supervision

References

- [1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 3, 4
- [2] Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. Survey Certification, Expert Certification. 1, 2, 4, 5
- [3] Beepul Bharti, Paul Yi, and Jeremias Sulam. Sufficient and necessary explanations (and what lies in between). In *The Second Conference on Parsimony and Learning (Proceedings Track)*, 2025. 3
- [4] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, pages 16318–16352. Curran Associates, Inc., 2023. 1, 3, 6, 7
- [5] 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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, 2020. Association for Computational Linguistics. 1, 3, 4, 5
- [6] Gabriel G. Erion, Joseph D. Janizek, Pascal Sturmfels, Scott M. Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3:620 – 631, 2021. 4
- [7] Javier Ferrando and Elena Voita. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA, 2024. Association for Computational Linguistics. 1, 3, 5
- [8] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2950–2958, 2019. 3
- [9] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, 2017. 3
- [10] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, pages 9574–9586. Curran Associates, Inc., 2021. 3
- [11] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, pages 160–187. PMLR, 2024. 3
- [12] Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Jun Koba Sato, and Aryaman Arora. Localizing model behavior with path patching. *ArXiv*, abs/2304.05969, 2023. 3, 5
- [13] Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. Workshop Spotlight. 1, 3, 4, 5, 6
- [14] Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *ArXiv*, abs/2404.15255, 2024. 1, 3, 5
- [15] Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, et al. A survey on mechanistic interpretability for multi-modal foundation models. *arXiv preprint arXiv:2502.17516*, 2025. 1, 2, 4
- [16] Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [17] 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: Analyzing and Interpreting Neural Networks for NLP*, 2024. 3
- [18] Maxime Méroux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 3, 4, 5
- [19] Aaron Mueller, Atticus Geiger, Sarah Wiegrefe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Forty-second International Conference on Machine Learning*, 2025. 2, 5, 6, 7
- [20] Jatin Nainani, Sankaran Vaidyanathan, AJ Yeung, Kartik Gupta, and David Jensen. Adaptive circuit behavior and generalization in mechanistic interpretability, 2024. 1, 3, 4
- [21] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 2
- [22] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

- [23] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020. [2](#)
- [24] Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision–language models. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025. [3](#)
- [25] Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models, 2024. [1](#), [2](#), [4](#), [5](#)
- [26] Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. Interpreting graph neural networks for {nlp} with differentiable edge masking. In *International Conference on Learning Representations*, 2021. [3](#)
- [27] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025. [1](#), [2](#), [4](#), [5](#)
- [28] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328, 2017. [4](#)
- [29] Aaqib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US, 2024. Association for Computational Linguistics. [1](#), [3](#), [5](#), [6](#)
- [30] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. [3](#)
- [31] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, 2019. [3](#)
- [32] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems*, 2019. [3](#)