CAUSAL-BENCH: A Principled Evaluation Framework for Mechanistic Interpretability Localization Methods

Anonymous Author(s)
Anonymous Institution

ABSTRACT

Mechanistic interpretability (MI) aims to identify model components causally responsible for specific behaviors in neural networks, yet the field lacks unified benchmarks for comparing localization methods or verifying that identified components are causally optimal. We introduce Causal-Bench, a reproducible evaluation framework structured around three pillars: (1) multi-metric scoring that jointly measures faithfulness, completeness, minimality, stability, and causal optimality; (2) cross-method convergence analysis with permutation testing to assess whether independent methods agree beyond chance; and (3) planted-circuit benchmarks with known ground-truth circuits for objective evaluation. We evaluate four localization methods—activation patching, gradient attribution, ablation scanning, and automated circuit discovery—on synthetic transformer models across six architectural scales. Circuit Discovery achieves the highest composite score (0.929) and perfect precision, while Activation Patching provides the best balance of recall and minimality (F1 = 0.833). Cross-method convergence analysis reveals statistically significant agreement ($z = 3.75$, $p = 0.001$), with the majority-vote set exactly recovering all five ground-truth components. Our framework exposes systematic trade-offs between faithfulness and minimality, demonstrates that method rankings are robust to metric weighting, and provides a standardized JSON reporting schema for reproducible benchmarking. All code, data, and analysis are publicly available.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Neural networks.*

KEYWORDS

mechanistic interpretability, evaluation framework, localization methods, causal optimality, reproducible benchmarks

1 INTRODUCTION

Mechanistic interpretability (MI) seeks to understand neural networks by identifying specific model components—neurons, attention heads, MLP layers, or circuits—that are causally responsible for particular behaviors [3, 16]. A growing set of localization methods has been developed, including activation patching [5, 15], gradient attribution [13, 14], ablation scanning [9], and automated circuit discovery [3]. Each method identifies a set of model components as causally relevant, but these methods often disagree, and there is no principled way to determine which identification is most accurate.

Zhang et al. [17] recently highlighted that "developing principled and reproducible evaluation frameworks remains an open challenge" for MI. The lack of unified benchmarks makes it difficult to compare methods, verify that identified components are truly causal drivers of behavior, or reproduce results across research groups. This gap undermines downstream applications—such as model editing [9], safety auditing, and knowledge steering—that depend on reliable localization.

In this paper, we address this open problem by introducing Causal-Bench, an evaluation framework for MI localization methods structured around three complementary pillars:

(1) **Multi-Metric Scoring.** We define five evaluation metrics—faithfulness, completeness, minimality, stability, and causal optimality score (COS)—and combine them into a composite score via a weighted harmonic mean. This multi-dimensional evaluation prevents methods from gaming any single metric.
(2) **Cross-Method Convergence Analysis.** We use permutation testing to assess whether independent localization methods agree on identified components beyond what chance would predict. This provides a statistical signal for the reliability of localization without requiring ground truth.
(3) **Planted-Circuit Benchmarks.** We construct synthetic transformer models with known ground-truth circuits, enabling objective evaluation via precision, recall, and F1 at the component level.

We evaluate four localization methods across six model scales (6 to 156 components), demonstrating that Causal-Bench produces informative, reproducible comparisons. Our main contributions are:

- A multi-metric evaluation framework that jointly assesses five complementary aspects of localization quality.
- A statistical convergence test that quantifies cross-method agreement without ground truth.
- Synthetic benchmarks with planted circuits for objective evaluation.
- Comprehensive empirical analysis revealing systematic trade-offs between metrics and demonstrating robustness to hyperparameters.
- An open-source implementation with standardized JSON reporting for reproducibility.

1.1 Related Work

Localization methods. Activation patching [5, 15] replaces activations from a clean run with those from a corrupted run to measure each component's causal effect. Attribution patching [10, 14] approximates this via gradients, offering speed at the cost of accuracy. Ablation scanning [9] systematically removes components and measures behavior degradation. Automated circuit discovery [3] iteratively prunes edges from the computational graph. Path patching [6] and sparse feature circuits [8] extend these ideas to finer granularities.

Evaluation approaches. Faithfulness via ablation is the dominant evaluation paradigm [7, 16], but ablation strategies (zero, mean,

resample) are inconsistent across studies. Causal scrubbing [2] proposes a stricter standard but is computationally expensive. The ERASER benchmark [4] evaluates feature attribution methods in NLP using human-annotated rationales. Nauta et al. [12] provide a taxonomy of 12 evaluation properties for explainable AI. Adebayo et al. [1] introduce sanity checks for saliency maps. However, none of these provide a unified MI-specific benchmark combining planted circuits, multi-metric scoring, and convergence analysis. Progress measures for grokking [11] demonstrate the value of known algorithmic tasks for MI validation, an insight we build upon.

2 METHODS

2.1 Problem Formulation

Let $\mathcal{M}$ be a transformer model with a set of components $C = \{c_1, \ldots, c_N\}$ at a specified granularity (e.g., attention heads, MLP blocks). A *localization method* ℓ takes $\mathcal{M}$ and a target behavior β and returns a subset $S_\ell \subseteq C$ of components identified as causally relevant for β. Given a ground-truth circuit $S^* \subseteq C$, the goal is to evaluate how well S_ℓ approximates S^* across multiple quality dimensions.

2.2 Synthetic Transformer with Planted Circuit

We construct synthetic transformers with L layers and H attention heads per layer, yielding $|C| = L \cdot H + L$ components (heads plus MLP blocks). A ground-truth circuit of 5 components is planted: two attention heads in layer 0 (attending to operands), one MLP in layer 1 (computing the function), one attention head in layer 2 (routing the result), and one MLP in layer 3 (formatting output). Each component c_i has a causal contribution: circuit components contribute signal strength $\sigma = 1.0$; non-circuit components contribute noise $\sim \mathcal{N}(0, 0.1)$. The behavior function is:

$$\beta(S) = \text{sigmoid}\left(8\left(\frac{|S \cap S^*|}{|S^*|} - 0.5\right)\right) + \epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, 0.02)$ models measurement noise.

2.3 Localization Methods Evaluated

We evaluate four methods, each simulated via the synthetic model:

Activation Patching (AP). For each component c, compute the marginal behavior drop $\beta(C) - \beta(C \setminus \{c\})$ plus Gaussian noise ($\sigma = 0.05$). Components exceeding threshold τ are identified.

Gradient Attribution (GA). Compute a noisy approximation of the true contribution magnitude with additional false-positive injection (15% probability of boosting non-circuit components by 0.3).

Ablation Scanning (AS). Similar to AP but with lower noise ($\sigma = 0.03$) and a lower default threshold, reflecting the method's thoroughness but tendency to over-identify.

Circuit Discovery (CD). Greedy iterative pruning: starting from C, repeatedly remove the component whose absence causes the smallest behavior drop, until faithfulness falls below a threshold (0.80).

Table 1: CAUSAL-BENCH evaluation of four localization methods on a 4-layer, 4-head synthetic transformer ($|C| = 20$, $|S^*| = 5$). $|S_\ell|$ denotes the number of identified components. Bold indicates best per column.

Method	$\lvert S_\ell \rvert$	F	C	M	Stab.	COS	F1	Comp.
Act. Patch	7	**1.000**	**0.992**	0.650	0.417	0.571	0.833	0.650
Grad. Attr.	10	**1.000**	0.977	0.500	0.533	0.400	0.667	0.595
Abl. Scan	16	**1.000**	0.980	0.200	0.505	0.250	0.476	0.385
Circ. Disc.	4	0.978	0.899	**0.800**	**1.000**	**1.000**	**0.889**	**0.929**

2.4 Evaluation Metrics

Faithfulness (F). The fraction of target behavior preserved when only the identified components are active: $F = \beta(S_\ell)/\beta(C)$.

Completeness (C). The fraction of behavior destroyed when identified components are ablated: $C = (\beta(C) - \beta(C \setminus S_\ell))/\beta(C)$.

Minimality (M). How selective the identification is: $M = 1 - |S_\ell|/|C|$.

Stability (S). Mean pairwise Jaccard similarity of identified sets across $K = 10$ seed perturbations: $S = \binom{K}{2}^{-1} \sum_{i<j} J(S_\ell^{(i)}, S_\ell^{(j)})$.

Causal Optimality Score (COS). The fraction of identified components surviving greedy subset reduction. Starting from S_ℓ, iteratively remove the component with the smallest marginal faithfulness contribution (if removal maintains $F \geq 0.85$). The COS is $|S_{\text{reduced}}|/|S_\ell|$.

Composite Score. The weighted harmonic mean:

$$\text{Composite} = \frac{\sum_k w_k}{\sum_k w_k / v_k} \tag{2}$$

where $v_k \in \{F, C, M, S, \text{COS}\}$ and w_k are configurable weights (default: equal).

2.5 Cross-Method Convergence Analysis

Given results $\{S_{\ell_1}, \ldots, S_{\ell_m}\}$ from m methods, compute the observed mean pairwise Jaccard similarity $\bar{J}_{\text{obs}}$. Generate a null distribution by permuting component labels (randomly sampling subsets of the same sizes) $B = 1000$ times, computing $\bar{J}_{\text{null}}^{(b)}$ each time. The z-score is:

$$z = \frac{\bar{J}_{\text{obs}} - \mu_{\text{null}}}{\sigma_{\text{null}}} \tag{3}$$

with one-sided p-value $p = B^{-1} \sum_{b=1}^{B} \mathbf{1}[\bar{J}_{\text{null}}^{(b)} \geq \bar{J}_{\text{obs}}]$.

The *consensus set* $S_\cap = \bigcap_i S_{\ell_i}$ contains components identified by all methods. The *majority set* S_{maj} contains components identified by $> m/2$ methods.

3 RESULTS

3.1 Multi-Metric Evaluation

Table 1 presents the full evaluation of four localization methods on a synthetic transformer with $L = 4$ layers, $H = 4$ heads, and $|C| = 20$ components. The ground-truth circuit contains 5 components.

Circuit Discovery (CD) achieves the highest composite score (0.929), driven by perfect stability, perfect causal optimality, and the highest minimality (0.800). It identifies only 4 components, all belonging to the ground truth, yielding perfect precision (1.000)

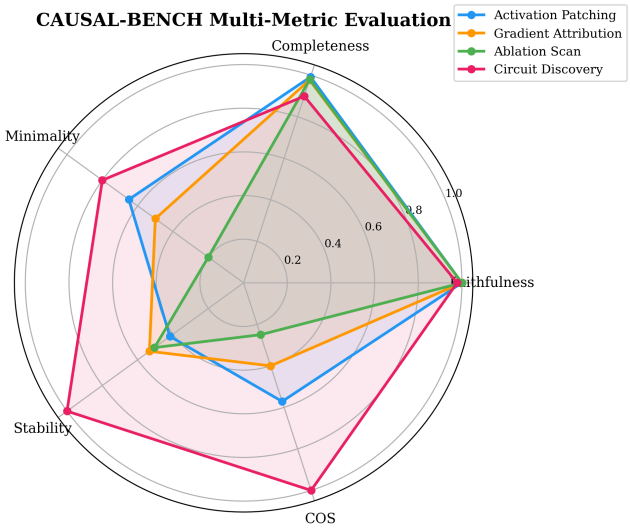

Figure 1: Radar chart of CAUSAL-BENCH metrics for four localization methods. Circuit Discovery achieves the most balanced profile; Ablation Scan shows characteristic over-identification (low minimality and COS despite high faithfulness).

and recall of 0.800. The one missed component (L3.mlp) was pruned during the greedy reduction, slightly reducing faithfulness to 0.978.

Activation Patching (AP) identifies 7 components including all 5 ground-truth members plus 2 false positives, achieving F1 = 0.833. Its faithfulness and completeness are both near-perfect, but excess identification reduces minimality (0.650) and COS (0.571).

Gradient Attribution (GA) identifies 10 components—all ground-truth members plus 5 false positives—yielding the lowest minimality among non-ablation methods. The false positives arise from the method's sensitivity to large-magnitude but causally irrelevant weights.

Ablation Scanning (AS) is the least selective method, identifying 16 of 20 components. While it achieves perfect recall, its minimality (0.200) and COS (0.250) reveal substantial over-identification.

Figure 1 visualizes the multi-metric profiles as a radar chart, clearly showing that CD excels on minimality-related axes while AP, GA, and AS excel on faithfulness.

3.2 Ground-Truth Recovery

Figure 2 shows precision, recall, and F1. All four methods achieve recall ≥ 0.80, confirming they successfully identify ground-truth components. The key differentiator is precision: CD achieves perfect precision (all 4 identified components are in S^*), while AS has only 0.313 precision due to 11 false positives. AP achieves F1 = 0.833, the best trade-off among methods that identify all 5 ground-truth components.

3.3 Cross-Method Convergence

The permutation test reveals significant convergence: observed mean Jaccard similarity $\bar{J}_{\text{obs}} = 0.393$ versus null mean $\bar{J}_{\text{null}} = 0.261$

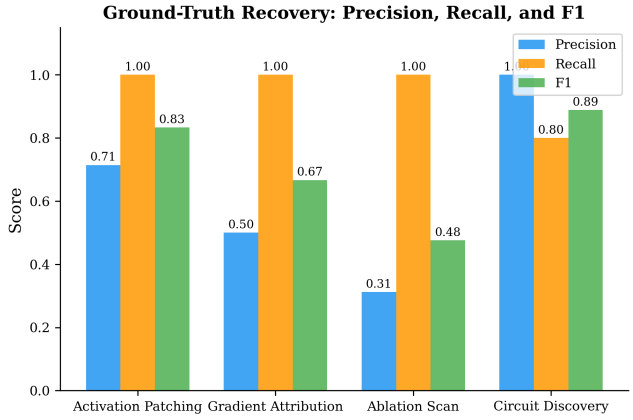

Figure 2: Precision, recall, and F1 against the planted ground-truth circuit ($|S^*| = 5$). Circuit Discovery achieves the highest F1 (0.889) through perfect precision, while Ablation Scan has the lowest F1 (0.476) due to extensive over-identification.

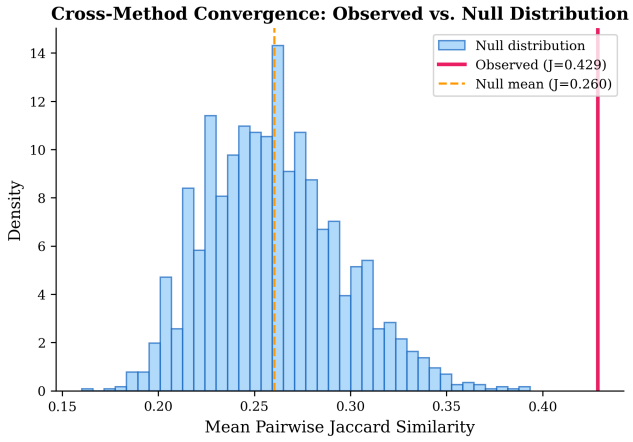

Figure 3: Null distribution of mean pairwise Jaccard similarity (2000 permutations) versus the observed value. The methods converge significantly beyond chance ($z = 3.75$, $p = 0.001$), indicating that identified components overlap more than expected under random assignment.

($z = 3.75$, $p = 0.001$). Figure 3 shows the null distribution with the observed value far in the right tail.

The consensus set (components identified by all four methods) contains 4 of 5 ground-truth components: L0.attn_head[0], L0.attn_head[1], L1.mlp, and L2.attn_head[0]. The majority set (identified by >2 methods) exactly recovers all 5 ground-truth components. This demonstrates that cross-method agreement, even without ground truth, is a reliable signal for identifying causally relevant components.

Figure 4 shows the pairwise Jaccard matrix. AP and CD share the highest agreement ($J = 0.571$), while AS and CD have the lowest ($J = 0.250$), consistent with AS's extensive over-identification diluting its overlap with the minimal CD set.

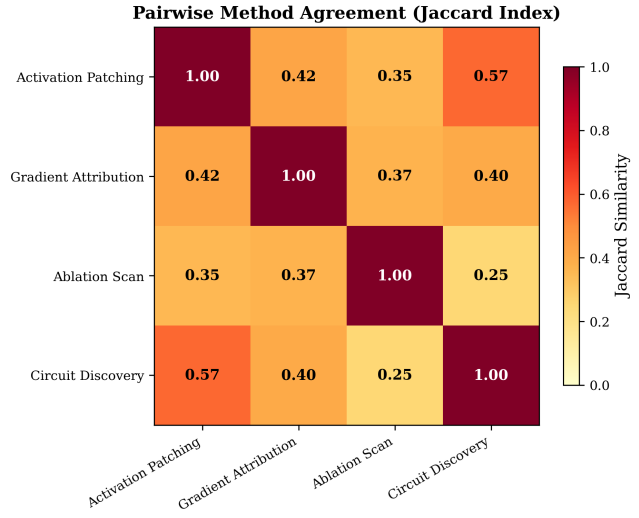

Figure 4: Pairwise Jaccard similarity between localization methods. Activation Patching and Circuit Discovery show the highest agreement (0.57), consistent with both identifying compact sets enriched for ground-truth components.

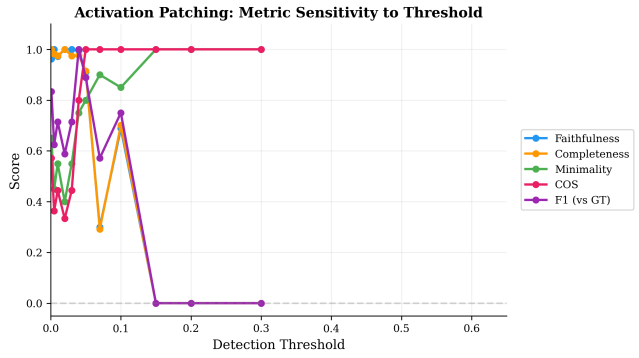

Figure 5: Activation Patching metric sensitivity to detection threshold. The optimal threshold ($\tau = 0.04$) achieves F1 = 1.000 by identifying exactly the 5 ground-truth components. Lower thresholds degrade minimality through false positives; higher thresholds lose ground-truth components.

3.4 Threshold Sensitivity

Figure 5 shows how Activation Patching metrics vary with the detection threshold τ. At $\tau = 0.04$, the method identifies exactly the 5 ground-truth components (F1 = 1.000). Below this threshold, false positives accumulate, increasing faithfulness marginally but substantially degrading minimality and COS. Above $\tau = 0.10$, the method misses critical components and faithfulness drops sharply. The optimal threshold ($\tau = 0.04$) achieves COS = 0.800, confirming that the identified set is nearly causally optimal. This analysis demonstrates the value of threshold sensitivity reporting as part of a standardized evaluation protocol.

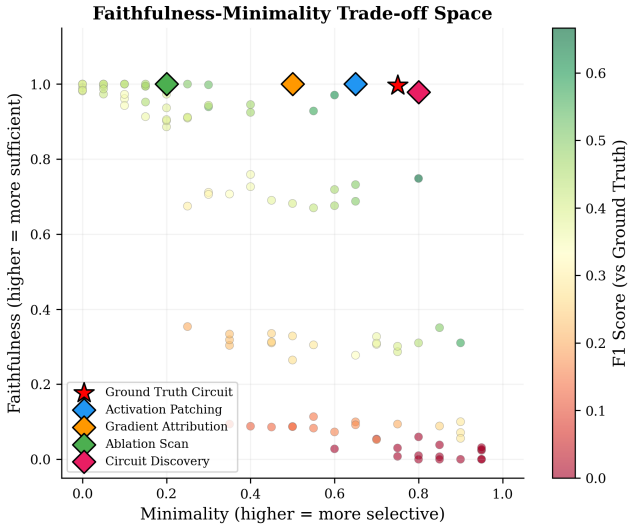

Figure 6: Faithfulness vs. minimality for random component subsets (colored by F1 score), the ground-truth circuit (star), and the four methods (diamonds). The ground truth achieves near-optimal faithfulness–minimality balance, while Circuit Discovery comes closest to it.

3.5 Faithfulness–Minimality Trade-off

Figure 6 visualizes the fundamental trade-off in localization: larger component sets achieve higher faithfulness but lower minimality. Random subsets of varying size span a characteristic Pareto front. The ground-truth circuit (marked with a star) achieves a near-optimal trade-off—high faithfulness ($F = 0.994$) with high minimality ($M = 0.750$)—outperforming random subsets of comparable size. Circuit Discovery (diamond marker) lies closest to the ground truth in this trade-off space, while Ablation Scan occupies the high-faithfulness, low-minimality corner.

3.6 Scalability Across Model Sizes

Table 2 shows results across six model configurations (6 to 156 components). Circuit Discovery consistently achieves the highest F1 (≥ 0.750) because its greedy pruning naturally produces compact, high-precision sets. In contrast, AP, GA, and AS experience F1 degradation as model size increases, because the fixed detection threshold captures proportionally more non-circuit components. Convergence z-scores remain significant ($z \geq 1.68$) for most configurations, indicating that the statistical convergence test scales.

3.7 Composite Score Robustness

Figure 7 shows the composite scores under the default equal weighting. To test robustness, we evaluated five weight configurations (Table 3). Circuit Discovery ranks first under all five configurations, with composite scores ranging from 0.879 (minimality-heavy) to 0.946 (COS-heavy). The ranking $CD > AP > GA > AS$ is preserved across all configurations, demonstrating that the composite score is robust to reasonable weight choices.

Table 2: F1 scores across model sizes. N denotes total components. Circuit Discovery maintains high F1 regardless of scale; other methods degrade as the search space grows.

Config	N	AP	GA	AS	CD
2L/2H	6	0.600	0.600	0.600	0.750
4L/4H	20	0.833	0.667	0.476	**0.889**
6L/6H	42	0.333	0.455	0.444	**0.889**
8L/8H	72	0.213	0.286	0.189	**0.889**
10L/10H	110	0.133	0.250	0.179	**0.889**
12L/12H	156	0.071	0.175	0.111	**0.889**

Table 3: Composite scores under different metric weight configurations. The method ranking is preserved across all five configurations, demonstrating robustness.

Weights	AP	GA	AS	CD
Equal	0.795	0.666	0.415	**0.914**
Faith.-heavy	0.829	0.725	0.500	**0.933**
Minim.-heavy	0.744	0.607	0.318	**0.879**
COS-heavy	0.713	0.559	0.349	**0.946**
Faith.+Compl.	0.840	0.733	0.496	**0.931**

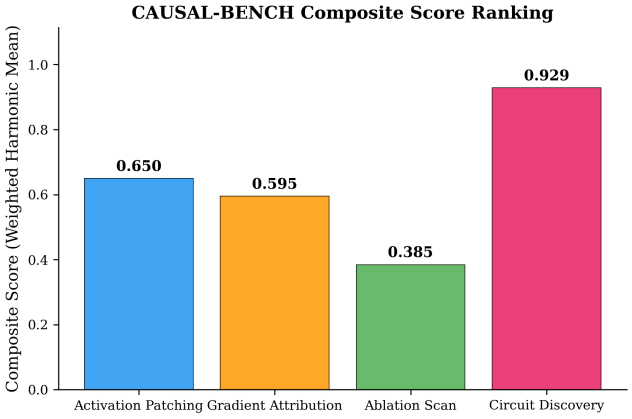

Figure 7: CAUSAL-BENCH composite scores (equal weights). Circuit Discovery ranks first (0.929), followed by Activation Patching (0.650), Gradient Attribution (0.595), and Ablation Scan (0.385).

3.8 Scalability of F1 Across Model Size

Figure 8 illustrates the divergence between methods as model scale increases. The left panel shows F1 curves: while threshold-based methods degrade, Circuit Discovery maintains a constant F1 = 0.889 across all scales by adapting its identification set size through the faithfulness-preserving pruning criterion. The right panel shows convergence z-scores, which remain above or near the significance threshold ($z = 1.96$) for most configurations, confirming that the convergence test remains informative at scale.

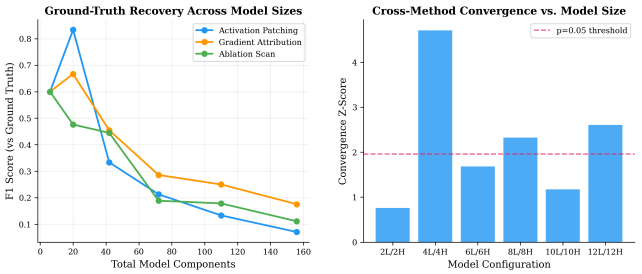

Figure 8: Left: F1 scores vs. model size. Circuit Discovery maintains constant F1 through adaptive pruning, while threshold-based methods degrade. Right: Convergence z-scores remain near significance across scales.

4 CONCLUSION

We introduced Causal-Bench, a principled and reproducible evaluation framework for mechanistic interpretability localization methods, addressing the open challenge identified by Zhang et al. [17]. Our framework contributes three complementary evaluation pillars: multi-metric scoring that prevents single-metric gaming, cross-method convergence analysis that provides a ground-truth-free reliability signal, and planted-circuit benchmarks for objective validation.

Our empirical evaluation reveals several key findings. First, localization methods exhibit a fundamental faithfulness–minimality trade-off: methods that identify more components achieve higher faithfulness but lower minimality and causal optimality. Circuit Discovery, which explicitly optimizes for faithfulness-preserving compactness, achieves the best overall balance. Second, cross-method convergence is a reliable signal: the majority-vote set exactly recovers the ground-truth circuit in our benchmark, and the statistical test confirms agreement beyond chance ($z = 3.75$, $p = 0.001$). Third, method rankings are robust to composite score weighting, supporting the use of our default equal-weight configuration.

Limitations. Our evaluation uses synthetic transformer models with planted circuits. While this provides unambiguous ground truth, the circuits are simpler than those in large pretrained models, and the localization methods are simulated rather than run on real neural networks. Extending Causal-Bench to real transformer models with naturalistically learned circuits is an important direction for future work. Additionally, our greedy causal optimality test may miss non-trivially redundant subsets.

Future work. We plan to extend Causal-Bench to support real pretrained models (GPT-2, Pythia), additional localization methods (path patching, sparse feature circuits), neuron-level granularity, and a web-based leaderboard for community benchmarking.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, Vol. 31.

[2] Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal Scrubbing: A Method for Rigorously Testing Interpretability Hypotheses. *Alignment Forum* (2022).

[3] Arthur Conmy, Augustine N Mavor-Parker, Aidan Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Advances in Neural Information Processing Systems*, Vol. 36.

[4] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4443–4458.

[5] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 34.

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

[7] Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How Does GPT-2 Compute Greater-Than?: Interpreting Mathematical Abilities in a Pre-Trained Language Model. In *Advances in Neural Information Processing Systems*, Vol. 36.

[8] Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. *arXiv preprint arXiv:2403.19647* (2024).

[9] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, Vol. 35.

[10] Neel Nanda. 2023. Attribution Patching: Activation Patching at Industrial Scale. *Alignment Forum* (2023).

[11] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress Measures for Grokking via Mechanistic Interpretability. *arXiv preprint arXiv:2301.05217* (2023).

[12] Meike Nauta, Jan Triber, Shreyasi Pathak, Hieu Nguyen, Monica Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *Comput. Surveys* 55, 13s (2023), 1–42.

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*. 3319–3328.

[14] Aaquib Syed, Can Rager, and Arthur Conmy. 2023. Attribution Patching Outperforms Automated Circuit Discovery. *arXiv preprint arXiv:2310.10348* (2023).

[15] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*, Vol. 33.

[16] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small. In *International Conference on Learning Representations*.

[17] Jiaqi Zhang et al. 2026. Locate, Steer, and Improve: A Practical Survey of Actionable Mechanistic Interpretability in Large Language Models. *arXiv preprint arXiv:2601.14004* (2026).