Transformer Circuit Faithfulness Metrics are not Robust

Joseph Miller **Bilal Chughtai**
josephmiller101@gmail.com

William Saunders

Abstract

Mechanistic Interpretability work attempts to reverse engineer the learned algorithms present inside neural networks. One focus of this work has been to discover 'circuits' – subgraphs of the full model that explain behaviour on some specific task. But how do we measure the performance of such circuits? Prior work has attempted to measure circuit 'faithfulness' – the degree to which the circuit captures the performance of the full model. In this work, we survey many considerations for designing experiments that measure circuit faithfulness by ablating portions of the model's computation. Concerningly, we find existing circuit faithfulness metrics are highly sensitive to reasonable variations in the ablation methodology. We conclude that existing faithfulness scores are somewhat arbitrary and do not provide conclusive evidence as to the validity of a given circuit. Our work emphasizes the need for more rigorous evaluations of circuits and more clarity in the precise claims being made about them. The ultimate goal of mechanistic interpretability work is to understand neural networks, so it is vital that we are not misled by our metrics. We open source a library for testing a wide range or ablation methodologies at this anonymous URL.

1 Introduction

A key metric used by Mechanistic Interpretability (MI) researchers to quantify the quality of a 'circuit' for some task is it's *faithfulness* – that is, the degree to which the circuit captures the performance of the entire model. In this work, we study various small and reasonable seeming variations on methodologies for measuring circuit faithfulness and find that such variations often lead to *significantly different* faithfulness scores.

Mechanistic Interpretability (MI) is a form of post-hoc interpretability that attempts to reverse engineer neural networks to provide faithful low-level explanations of model behaviour. One focus of interpretability work on transformer language models is identifying 'circuits' – subgraphs of the entire model's computational graph that implement an algorithm to solve some particular task; specified by a narrow input distribution. But how do we tell when we have found a 'good' circuit? Faithfulness is typically measured by performing a targeted, circuit-dependent *ablation* to the model, and measuring the effect of this on some metric of the model's output. In the context of MI, an ablation refers to a type of intervention made on the activations of a model during its forward pass with the intended purpose of 'deleting' some causal pathway(s), thereby isolating the causal effect of the circuit.

In this work, we seek to answer the questions: What do circuit faithfulness metrics actually show? Are the circuits MI researchers find robust to the methodology used to measure their quality?

We begin by reviewing the degrees of freedom in which MI researchers may vary their ablation methodology (Section 3), providing a detailed review of methods for ablating transformer circuits. Next, we test these variations on existing circuits discovered by MI researchers (Section 4). We provide detailed case studies of the 'indirect object identification' circuit by Wang et al. (2023), the 'docstring' circuit by Heimersheim & Janiak (2023) and the 'sports players' circuit by Nanda et al. (2023b). We then go on to study 'optimal circuits'

(Section 5) in the context of automated circuit discovery (Conmy et al., 2023) – an emerging paradigm of that aims to discover algorithmically without human input.

We conclude with recommendations for MI researchers (Section 6). We additionally release a library containing efficient implementations of the circuit-discovery and circuit-evaluation techniques used in this paper, that is significantly faster than all prior implementations we tested (such as Conmy et al. (2023)).

2 Related Work

Circuit Analysis. Circuit analysis a form of post-hoc interpretability focused on understanding the full end-to-end learned algorithm responsible for some specified narrow behaviour. A circuit is some subgraph of the full computational graph of the model that (is alleged to) implement some precise behavior. Circuits have historically been studied in vision models (Cammarata et al., 2021; Olah et al., 2020) and in toy transformer models (Nanda et al., 2023a; Chughtai et al., 2023). This paradigm has achieved reasonable success in transformer language models too, with a number of early papers discovering circuits implementing human understandable algorithms (Wang et al., 2023; Heimersheim & Janiak, 2023; Hanna et al., 2023) through considerable effort and manual inspection. To accelerate such studies, recent work has attempted to automate the process of discovering circuits (Conmy et al., 2023; Syed & Rager, 2023; Kramar et al., 2024), particularly in large language models, as circuits have historically required a large amount of researcher-effort to discover. Ideal circuits exist on the Pareto frontier of faithfulness, completeness and [as the whole network is already trivially optimal for the first two] simplicity (Wang et al., 2023).

Mechanistic Interpretablity attempts to reverse engineer trained machine learning models to produce faithful human understandable explanations of model predictions by analysing the low level features and algorithms implemented by the network. Circuit analysis is just one important direction in this theme of work. Besides circuit analysis, MI work more broadly seeks to understand the correct frame to interpret neural network computation (Elhage et al., 2021; Bricken et al., 2023; Cunningham et al., 2023) and to understand the learned features of models (Li et al., 2023; Tigges et al., 2023; Gurnee & Tegmark, 2024; Bills et al., 2023). MI has also inspired work in steering model outputs (Li et al., 2024; Turner et al., 2023; Rimsky et al., 2024).

3 Measuring Faithfulness

There are many choices MI researchers may choose between when designing experiments to measure circuit faithfulness. In this section, we review many important degrees of freedom.

3.1 Ablation Methodology

In the context of MI, an ablation refers to a type of intervention made on the activations of a model during its forward pass with the intended purpose of 'deleting' precise causal pathway(s). Intuitively, deleting important subcomponents for some task should damage task performance, and conversely deleting unimportant sub-components should preserve task performance. As such, ablations have arisen as a commonly used tool for *localizing* model behaviour to specific internal model components. Ablations may be used both to *find* and *evaluate* mechanistic explanations of model behavior.

The concept of ablation overlaps with a related technique, *activation patching*, in which activations are modified during a model's forward pass to some cached values from a different input. 'Corrupted' inputs are inputs which are similar to the 'clean' distribution being studied, but which have crucial differences that drastically change the output. For example, a typical 'corrupt' prompt could retain the structure of a 'clean' prompt, while switching a proper noun, such that the correct next token prediction is changed. In this work we consider activation patching to be a specific type of ablation, and use the term Resample Ablation interchangeably. But we note that in general, 'patching' means editing activations to some other value, instead of 'deleting' them, as ablation typically connotes.

Granularity	Component	Value	Token positions	Direction	Set
Heads, MLPs	Node	Resample/Patch	All tokens	Ablate Clean	Circuit
Q,K,V, MLPs	Edge	Zero	Specific tokens	Resample Clean	Complement
Neurons	Branch	Mean			
...		Noise			

Table 1: The six-tuple that defines *ablation methodology* for transformer circuits.

In the remainder of this section, we review the range of ablation techniques that exist in the literature, specifically as they relate to evaluating circuits. There exist several important degrees of freedom when evaluating transformer circuits via ablations. These are (1) the **granularity** of the computational graph used to represent the model, (2) what type of **component** in the graph ablated, (3) what type of **activation value** is used to ablate the component, (4) which **token positions** are ablated, (5) the ablation **direction** (whether the ablation destroys or restores the signal) and (6) the **set** of components (the circuit or the complement of the circuit). A circuit-based ablation methodology can therefore be specified as a six-tuple, and prior work has used many combinations. In this paper we argue that existing evaluations of circuits are sensitive to each of these variables.

3.1.1 Circuit Granularity

In this work we study circuits specified at the level of attention heads and MLPs. We also separate the input of each attention head into the Q, K and V inputs, but we omit this from our diagrams for visual simplicity. This is the most common granularity for mechanistic circuit analysis, but previous works have also studied circuits specified at the level of layers (Meng et al., 2022) and subspaces (Geiger et al., 2023).

3.1.2 Ablation Component Type (and Associated Model Views)

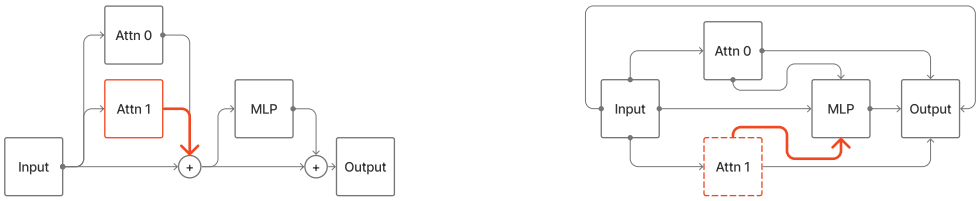
Transformers can be described as computational graphs in several different, equivalent ways. We can choose to write the graph as a residual network (Figure 6a) or a 'factorized' network in which all nodes are connected via an edge to *all* prior nodes (Figure 6b) (Elhage et al., 2021). Or we can write down a 'treeified' network that separates all paths from input to output (Figure 7a). All formulations are equivalent but the 'factorized' view allows us to isolate interactions between individual components and the 'treeified' view allows us to isolate chains of interactions from input to output.

The component type defines the type of intervention made: we detail three possibilities, with increasing granularity. The more granular approaches are generally more difficult to implement and more computationally expensive.

(1) Nodes. We may intervene on a node (in the standard, residual view) during the forward pass, replacing its activation with some other value (Figure 1a). This is the least specific form of ablation. Since all downstream nodes 'see' the change there are a large number of causal pathways affected by the ablation, which may result in unintended side-effects. This type of ablation is also known as (vanilla) activation patching (Vig et al., 2020) when we ablate with a cached activation from another input.

(2) Edges. Using the factorized view of a transformer, we may intervene on an edge between two components (Figure 1b). This is more specific than ablating nodes, as only the *specified* destination node receives the ablated activation of the source node, so a smaller number of causal pathways are affected.

(3) Branches. The previous two ablations can be applied to individual nodes or edges, or to a collection of nodes and edges. Branch ablations on the other hand can only be applied to *pathways* from input to output (Figure 7b). The causal effect of individual paths through the model is isolated by 'treeifying' the factorized model. This approach was introduced by Chan et al. (2022) and is a key component of a rigorous circuit evaluation approach known as Causal Scrubbing. However, because the number of paths in the treeified model



(a) Node Patching (often called Activation Patching) replaces the output of some component to the residual stream.

(b) Edge Patching replaces the activations of a single edge in the factorized view of a transformer.

Figure 1: The factorized formulation of transformers suggest a more specific ablation than ablating whole nodes.

is exponential in the number of layers of the model this approach to circuit evaluation is often intractable in practice. We omit treeified experiments in this work.

3.1.3 Ablation Value

When performing a causal intervention on some activation, we may choose what value we patch in. The simplest choice is to **Zero Ablate**, by replacing the activation with a vector of zeros. Prior work has noted however that the zero point is arbitrary (Wang et al., 2023). The next simplest is to apply **Gaussian Noise** (GN) to the token embeddings of the clean input to obtain corrupted activations. Prior work (Zhang & Nanda, 2024) has shown that both of these approaches can take the model significantly out of distribution, producing noisy outputs (Wang et al., 2023).

Two more principled approaches are to **Resample Ablate** (take an activation from some other corrupted input), or to **Mean Ablate** (replace with the mean activation of a node from some *distribution*). Importantly, these two ablation types do not delete all information present in a component. Instead, they delete information that *varies* across the distribution, while preserving information that is *constant*, allowing us to isolate precise language tasks, while ignoring, say, generic grammar processing. When Mean Ablating, we have an additional degree of freedom in choosing how large the mean ablation dataset is. We focus on Mean and Resample Ablations in this work.

3.1.4 Token Positions

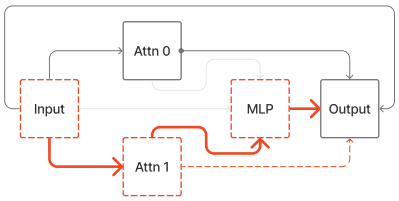
Circuits in autoregressive transformers on a narrow distribution are sometimes defined in terms of edges which each act at a given token position. When these token positions are specified, we can choose to either ablate all token positions, or only the specified set of token positions.

3.1.5 Ablation Direction and Testing Circuits

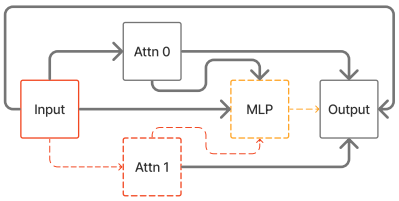
Ablation typically refers to instances where we run the model on a clean input and change activations to destroy the input signal. However, we can also run the model on a corrupt input and Resample Ablate (or Patch) in activations from the clean input. Separately, when evaluating circuits, we can choose to either ablate all the components of the circuit or we can ablate all the components *not* in the circuit (the complement).

The combination of these choices determines the target of our faithfulness metric:

Figure 2 compares the second and third rows of the table, which both measure faithfulness as the similarity of the ablation to the full model. We note that Resample Ablating clean activations for the circuit components while passing a corrupt input **allows the signal from the clean input to flow through edges not included in the circuit**. Whereas ablating with



(Row 2) Edge Patching all the edges in a circuit with clean activations allows information from the clean input to flow along paths not included in the circuit.



(Row 3) Edge Patching all the edges *not* in a circuit with corrupt activations ensures that information from the clean input only flows through edges included in the circuit.

Figure 2: Two approaches to testing a circuit that both measure faithfulness as the similarity of the output to the full model.

Model Input	Direction	Circuit Test	Faithfulness Target
Clean	Ablate Clean	Circuit	Destroy Performance
Corrupt	Resample Clean	Circuit	Restore Performance
Clean	Ablate Clean	Complement	Maintain Performance
Corrupt	Resample Clean	Complement	Maintain Inefficacy

corrupt activations on the complement of the circuit with a clean input ensures that **the signal from the input only flows through the circuit**.

3.2 Metric

One further consideration in addition to the ablation methodology is the **metric** used to evaluate the effect of the ablation. We also argue that the choice of metric is important. There are many choices used in the literature, including KL Divergence, top-k accuracy and task-specific benchmarks. In this work we will focus on the metrics used by previous authors to discover and evaluate the specific circuits that we study.

4 Faithfulness Metrics are Sensitive to Ablation Methodology

In this section, we empirically demonstrate that evaluations of a given circuit's faithfulness are highly sensitive to the experimental choices outlined in Section 3 made at evaluation time. We further argue that this sensitivity is important, and may result in practitioners finding fundamentally different algorithms.

We provide a case study here on the Indirect Object Identification (IOI) circuit identified by Wang et al. (2023), as this is the most studied language model circuit in the literature (Conmy et al., 2023; Makelov et al., 2023; Zhang & Nanda, 2024), but find similar results for other known language model circuits in Appendix C. The IOI circuit is specified as an edge-level circuit, but Wang et al. (2023) evaluate its faithfulness via a node-wise ablation methodology. We begin by testing the circuit using edge-level ablation.

The IOI circuit. The IOI circuit is a manually-identified subgraph of GPT-2 that is intended to implement the IOI algorithm. The IOI clean distribution consists of 15 sentence templates which involve two people interacting, structured such that the next word to be predicted is the indirect object `A`. Each template can be filled with names in the order `ABBA` or `BABA`, where the final `A` is the predicted token. For example: `When John and Mary went to the store, John bought flowers for Mary`. The corrupt distribution fills the same templates with names in the order `ABC` where `A`, `B` and `C` are three different names sampled independently of the corresponding clean prompt.

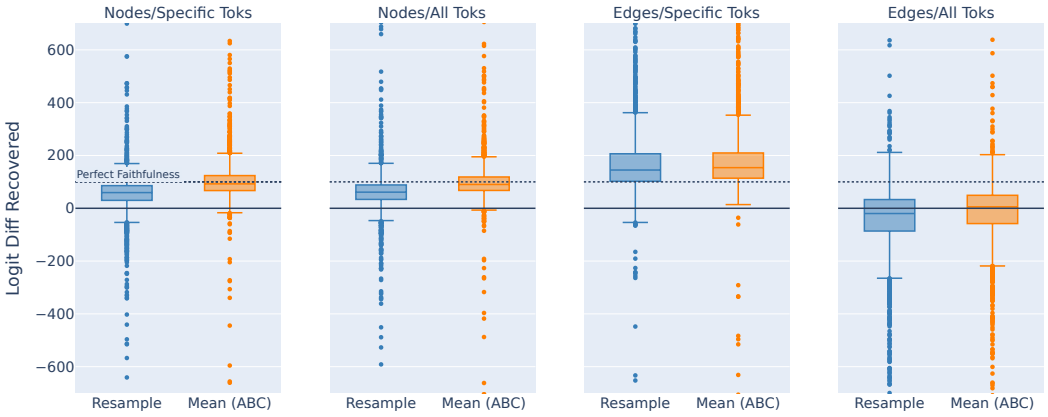


Figure 3: The IOI faithfulness metric is sensitive to (1) ablating edges/nodes, (2) the type of ablation used – we test Resample Ablations and Mean Ablations (over a dataset of 100 ABC prompts, which differs from Wang et al. (2023)) and (3) whether we differentiate between token positions in the circuit. The original IOI work evaluated with specific tokens with Mean Node Ablations and obtained a logit difference recovery of 87%. Other methodologies giving faithfulness scores above 100% or below 0% would have given the authors significantly less confidence about the IOI circuit, and may have led them to choose different edges.

Measuring IOI Circuit Faithfulness. Wang et al. (2023) define the *metric* of circuit faithfulness to be **logit difference recovered**. The logit difference is computed between the correct answer `A` and incorrect answer `B` both when the full model is run as normal and when the specified nodes are ablated. Then, the percentage of the full model's logit difference which is recovered by the ablated model is calculated. We adopt this definition of faithfulness for the remainder of this section. Note that this percentage is not constrained to the range 0 to 100 as the ablation can have a greater logit difference than the full model or the ablation can have a negative logit difference when the full model has a positive logit difference. Wang et al. (2023) test the faithfulness of their circuit by passing in a clean input and Node Ablating the complement of the circuit. They differentiate by token position – that is, they ablate nodes in the circuit at all token positions except those specified by the circuit. They use a Mean Ablation, where the mean value is computed for each token position over the `ABC` distribution, using around seven examples per template.

4.1 Variance Between Ablation Methodologies

We now show circuit faithfulness is sensitive to these choices. First we compare the faithfulness metric when we change the ablation component from nodes to edges: we ablate the complement of the set of edges specified by the circuit instead of the complement of the set of nodes in the circuit, in both cases distinguishing between different specified token positions. As shown in Figure 3, ablating at the edge level returns substantially higher percentages.

Figure 3 also evaluates the effect of ablation value. We rerun the above experiment using Resample Ablations from the ABC distribution, and find that this results in a systematically lower faithfulness as compared with mean ablations. Finally, we study the effect of ablating at every token position, instead of only those specified at the circuit. This consistently results in lower faithfulness scores. It is concerning that the edge-level circuit with specific token positions has a median score well over 100%, as this best represents the hypothesis of Wang et al. (2023).

Next, we discuss sensitivity of the faithfulness metric to both the clean distribution and intricacies of the metric calculation. For these experiments, we perform node-level Mean

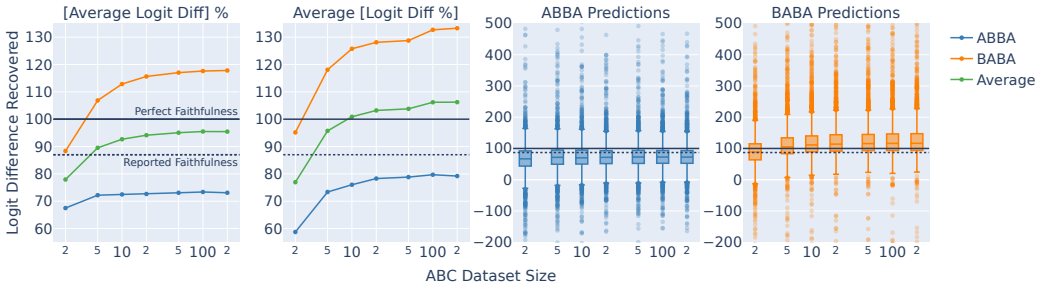


Figure 4: (Left) The IOI circuit is sensitive to the size of ABC dataset used for mean ablation. The logit difference recovered is consistently higher for prompts of the BABA format. (Middle Left) The order of computing the average and percentage affects the faithfulness metric. Wang et al. (2023) use [Average Logit Diff] %, giving lower scores than Average [Logit Diff %]. (Middle Right and Right) There is a large range of logit difference recovered, the boxplots show the interquartile range. Accordng to this faithfulness measurement methodology, The IOI circuit implements the IOI task faithfully on average, but not for many single data points.

Ablations on the complement of the circuit, split by token position, similarly to Wang et al. (2023). As shown in the left two charts of Figure 4, faithfulness is systematically greater for the prompts of form BABA than prompts of form ABBA. We also find that faithfulness monotonically increases with the size of the ABC dataset (used for computing the Mean Ablation). Finally we note that Wang et al. (2023) compute the logit difference recovered by first finding the mean logit difference for the full model and the ablated model over all prompts, and then computing the percentage (far left). If instead we compute the percent difference for each prompt and then take the mean, we return substantially higher percentages (middle left).

These are significant and important changes in evaluation. If the researchers had used a different methodology to they may have discovered a different circuit. This is important since it suggests that we cannot localize a core algorithm intrinsic to the model's completion of the task. We expand on this point in Section 5.

4.2 Variance Between Individual Datapoints

Even for a fixed ablation methodology and metric, there is significant variation in the measured faithfulness between individual prompts in the distribution.

We show this for the IOI circuit in the figures above, with results for other circuits in Appendix C. The graphs on the right of Figure 4 show a large range of faithfulness scores attained when we ablate the complement of the nodes in the IOI circuit. Note that the graphs do not show the full range of datapoints and there are several extreme outliers with a logit difference recovered in the tens of thousands of percent. The inter-quartile range (IQR) is also large, stretching up to 50% across the dataset. This is concerning: while the circuit matches the behavior on average, it does not match it for many examples. Another property of ideal circuits describing behaviour on some task is that their faithfulness *variance* should be low over the task distribution. Otherwise, the circuit is at least partially optimizied to balance out extremely high (significantly >100%) and extremely low faithfulness scores (<0%). This variance consideration is importantly missing from the mechanistic explanations of how GPT-2 implements the IOI task provided by Wang et al. (2023). We encourage MI researchers to evaluate task performance in both the average case and worst case.

5 Optimal Circuits are defined by tasks and ablation methodologies

We showed in the previous section that measurement details can greatly change the faithfulness score. However, one might ask if this difference matters. In this section we discuss

the consequences of such sensitivity for circuit discovery. We illustrate how methodologies can greatly affect the discovered circuit. We conclude that the optimal circuit for some task cannot be defined unless we also specify the ablation methodology and metric that we are using to measure it.

Tracr models (Lindner et al., 2023) are tiny transformers that are compiled instead of trained. Since the ground truth algorithm is both simple and known, they provide an excellent setup for testing circuit discovery algorithms. RASP programs (Rush & Weiss, 2023) are compiled into the weights of a transformer that implements the program exactly. Following Conmy et al. (2023), we study two Tracr models, `Reverse` and `X-Proportion`. We focus in particular on the X-Proportion model since as it is the smallest, and provides an simple example of optimal circuits being defined by ablation methodology.

The `X-Proportion` model performs the task of outputting at each token position the proportion of previous characters that are 'x's. The model has two layers, with one head in each attention layer. The first attention layer and the second MLP are not used, so we need only consider the edges between the `Input`, `MLP 0`, `Attn 1.0` and `Output`.

Conmy et al. consider the edge from `Input` to `Attn 1.0` to be part of the ground truth circuit. Inspecting the RASP program, we see that the only information in this edge's activation that is used by the model is the positional encoding of the tokens. However, this does not vary between different inputs, so if our ablation methodology uses Resample Ablations then this edge need not be included in the circuit, as ablating it will not change this positional information. However, if our ablattion methodology uses Zero Ablations, then this information will be destroyed, so the edge must be included in the circuit.

Conmy et al. test three automatic circuit discovery algorithms on this task. All three algorithms use (or approximate) Resample Ablations to discover circuits. The first method, ACDC, traverses the model in reverse topological order, ablating each edge in turn. Subnetwork Probing (SP) learns a mask parameter for each node, via gradient descent, attempting to maximize the number of nodes ablated, while minimizing the divergence from the original model. Lastly, Head Importance Scoring (HISP), uses a first order, gradient-based approximation of Node Ablation to assign attribution scores to each node. We test each circuit discovery method by sweeping over a range of importance thresholds to obtain an ordering of circuits of increasing size. We then plot pessimistic receiver operating characteristic (ROC) curves (Figure 5) and compare the area under curves.

SP and HISP, use (or approximate) Node Ablations, while ACDC uses Edge Ablations. To convert the predictions of SP and HISP to edge-based circuits, we include all edges which connect two nodes of sufficient importance. With this implementation it may be impossible for SP and HISP to correctly order edges. For example, there can be two nodes which are both individually important, but where the edge connecting them is unimportant. In this section we adjust the implementation of both SP and HISP to use (or approximate) Edge Ablations; SP learns mask parameters that ablate each edge and HISP assigns attribution scores for each edge by approximating Edge Patching. We provide a comparison between Edge and Node-based circuit discovery methods in Appendix D.

Conmy et al. considered the edges that would be required with Zero Ablations to be the correct circuits. Therefore, the algorithms fail to fully recover the "ground truth". When we instead consider the edges that are required with Resample Ablations to be the correct circuit, all three algorithms perfectly recover the "ground truth" (Figure 5).

This case study illustrates that the optimal circuit with respect to *only* a task is undefined. We must also specify the ablation methodology. That said, the overall point of circuit discovery is to localize model behaviour to model subcomponents.Ideally, such circuit discovery should not be sensitive to details of how one chooses to measure faithfulness (provided the results are directly comparable, e.g. both edge-based). This suggests localization may not in itself be a useful goal or that the granularity of the circuit is wrong. For example, features may be represented in superposition across multiple attention heads (Jermyn et al., 2023), in which case trying to localize a circuit to individual attention heads may be impossible.

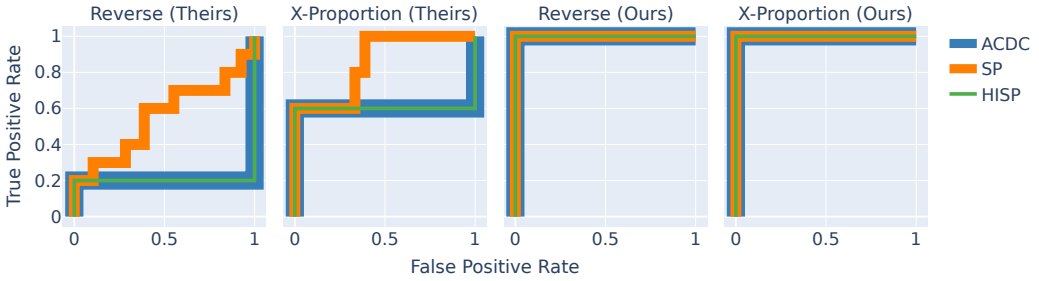


Figure 5: ROC Curves measuring the overlap between automatically discovered circuits and the two different "ground truth" circuits, for two Tracr tasks. When we match the ablation methodology of the ground truth with the ablation methodology of the circuit discovery algorithms, we can achieve perfect circuit recovery with all three methods.

6 Conclusion

In this work, we show existing transformer circuit evaluations are highly sensitive to small changes in the ablation methodology and metrics used to quantify faithfulness. We further show that the optimality of a circuit cannot be defined with respect to a task without a precise evaluation methodology.

Our work has significant consequences for circuit discovery work, particularly automated circuit discovery algorithms that aim to optimize these faithfulness scores. It also suggests that assessing the quality of automated methods by measuring the circuit overlap with some 'ground truth' can be misleading, if the ground truth was discovered using a different ablation methodology. This casts doubt on our ability to localize and understand identify consistent algorithms that implement particular capabilities in language models via optimizing faithfulness metrics.

We further highlight a need for better evaluations of mechanistic explanations. We only ultimately care about circuits in so far as they improve our ability to understand the algorithm that a model implements. This is hard to empirically measure, so researchers attempting to automatically discover circuits resort to measuring proxies of faithfulness, completeness and simplicity, and later worry about post-hoc explanation of the circuit. We are concerned this importantly misses the point of interpretability work, and risks 'Goodharting' (overfitting to a misspecified objective).

Our work suggests the need for better metrics of explanation success. One promising area of future work may consider alternative metrics of faithfulness that are more robust to methodological details.

References

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan

Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Nick Cammarata, Gabriel Goh, Shan Carter, Chelsea Voss, Ludwig Schubert, and Chris Olah. Curve circuits. *Distill*, 2021. doi: 10.23915/distill.00024.006. https://distill.pub/2020/circuits/curve-circuits.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: a method for rigorously testing interpretability hypotheses [redwood research], 2022. URL https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations, 2023.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas F. Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. *ArXiv*, abs/2303.02536, 2023. URL https://api.semanticscholar.org/CorpusID:257365438.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching, 2023.

Wes Gurnee and Max Tegmark. Language models represent space and time, 2024.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model, 2023.

Stefan Heimersheim and Jett Janiak. A circuit for Python docstrings in a 4-layer attention-only transformer, 2023. URL https://www.alignmentforum.org/posts/u6KXXmKFbXfWzoAXn/a-circuit-for-python-docstrings-in-a-4-layer-attention-only.

Adam Jermyn, Chris Olah, and Tom Henighan. Attention head superposition. https://transformer-circuits.pub/2023/may-update/index.html#attention-superposition, May 2023. Accessed: 30 March 2024.

Janos Kramar, Tom Lieberum, Rohin Shah, and Neel Nanda. Atp*: An efficient and scalable method for localizing llm behaviour to components, 2024.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task, 2023.

Maximilian Li, Xander Davies, and Max Nadeau. Circuit breaking: Removing model behaviors with targeted ablation, 2024.

David Lindner, János Kramar, Matthew Rahtz, Thomas McGrath, and Vladimir Mikulik. Tracr: Compiled transformers as a laboratory for interpretability. *arXiv preprint arXiv:2301.05062*, 2023.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023a.

Neel Nanda, Senthooran Rajamanoharan, Janos Kramar, and Rohin Shah. Fact finding: Attempting to reverse-engineer factual recall on the neuron level, Dec 2023b. URL https://www.alignmentforum.org/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024.

Alexander Rush and Gail Weiss. Thinking like transformers. In *ICLR Blogposts 2023*, 2023. URL https://iclr-blogposts.github.io/2023/blog/2023/raspy/. https://iclr-blogposts.github.io/2023/blog/2023/raspy/.

Aaquib Syed and Can Rager. Attribution patching outperforms automated circuit discovery, 9 2023.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models, 2023.

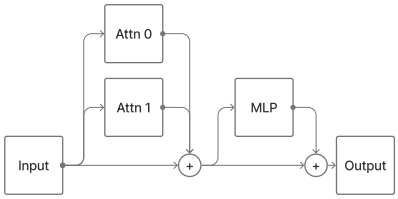
Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in neural information processing systems*, volume 33, pp. 12388–12401, 2020.

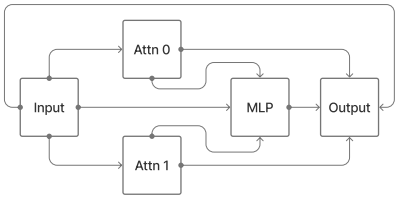
Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods, 2024.

A Further Details on Ablation Methodology

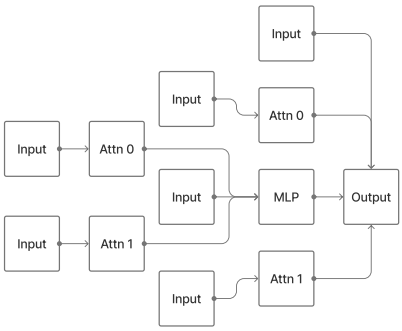


(a) The canonical formulation of a transformer. Every component reads input from the residual stream backbone.

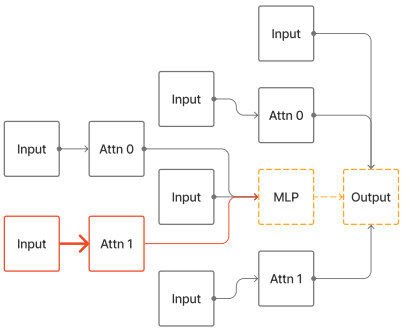


(b) The "factorized" formulation of a transformer views every component as taking input from every previous component.

Figure 6: Two equivalent formulations of the transformer architecture. We illustrate only one layer, but this extends trivially to many layers.



(a) The "treeified" formulation of a transformer separates every pathway from input to output.



(b) Branch Patching replaces the input of a single branch in the treeified view of a transformer.

Figure 7: The treeified formulation of transformers suggest a more specific approach to ablating end-to-end circuits.

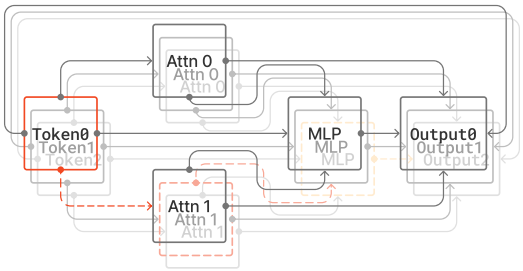


Figure 8: We can consider each token position to have a separate set of edges.

B Summary of Tasks

Name	Model	Example Clean Prompt	Example Corrupt Prompt	Correct Answer	Incorrect Answer	Faithfulness Metric
Tracr X-Proportion	Tracr X-Proportion	**y,x,z,x,w**	**z,w,w,y,x**	**0,0.5,0.333, 0.5,0.4**	**0,0,0,0,0.2**	Mean squared error
Tracr Reverse	Tracr Reverse	**1,0,2,2,2**	**1,0,0,1,2**	**2,2,2,0,1**	**2,1,0,0,1**	KL Divergence
Docstring	4 Layer Attention Only	**def error(self, create, option, file, run, client, project): """land employment camp :param file: protein author :param run: forest degree :param**	**def error(self, create, option, output, host, label, project): """land employment camp :param first: protein author :param text: forest degree :param**	**" client"**	**" size", " output", " host"," label", " first", " text", " request", " user", " file"," run", " create", " option", " project"**	Correct Prediction Proportion
Indirect Object Identification	GPT-2	Then, Scott and Jeremy went to the hospital. Jeremy gave a snack to	Then, Michael and Anderson went to the hospital. Rachel gave a snack to	**" Scott"**	**" Jeremy"**	Logit Difference Recovered
Sports Players	Pythia 2.8B	Fact: Tiger Woods plays the sport of golf\nFact: Phil Simms plays the sport of	Fact: Tiger Woods plays the sport of golf\nFact: Babe Ruth plays the sport of	**" football"**	**" basketball", " baseball"**	Top Sport Logit

Table 2: The tasks we study, which previous works have found circuits for, and the metrics used by previous works to measure their faithfulness.

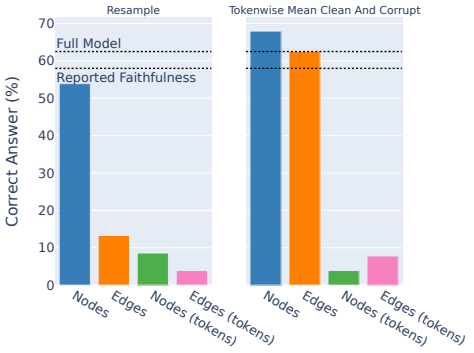
C Further Study of Faithfulness Metrics

In this section, we provide further analysis demonstrating faithfulness metrics are brittle, on two other circuits from the existing literature.

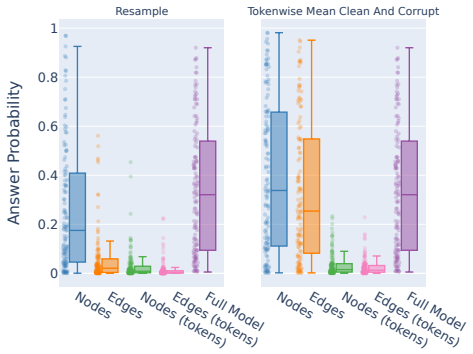
C.1 Docstring

The Docstring Task. The Docstring task (Heimersheim & Janiak, 2023) is a simple task that tests a 4 layer, attention-only, model's ability to complete a specific part of a standard Python docstring. See Table 2 for an example. All prompts follow a very similar format, with the only difference being the names of the variables in the function. The corrupt distribution follows the exact same format, using a disjoint set of variable names.

Measuring Docstring Circuit Faithfulness. Heimersheim & Janiak (2023) test their circuit using a similar methodology to the one which Wang et al. (2023) used to test the IOI circuit. They ablate all nodes in the complement of their circuit. However, unlike Wang et al. (2023) they use a Resample Ablation (also known in this context as Activation Patching), and they do not differentiate different token positions. The metric that they use for faithfulness is the percent of highest logit outputs that are the correct answer over some set of prompts.



(a) The faithfulness of the Docstring circuit according to the correct answer percent metric used by Heimersheim et al. Heimersheim & Janiak (2023) is sensitive to the type of ablation used to measure the circuit.



(b) The faithfulness of the Docstring circuit as measured by the probability of the correct answer is highly variable between individual prompts.

Figure 9: Faithfulness metrics for the Docstring circuit when ablating every node or edge *not* in the circuit, at all token positions and at token positions specified by Heimersheim et al. Heimersheim & Janiak (2023).

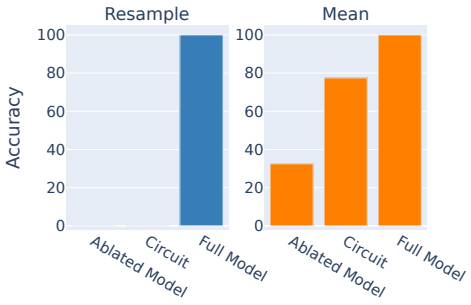
In Figure 9, we test the faithfulness of the docstring circuit with various ablation methodologies. We compare: (1) differentiating between different token positions (Heimersheim & Janiak (2023) specify their circuit with token positions, even though they do not use this information in their faithfulness evaluations), (2) ablating at the edge-level and node-level (they also specify edges, even though they evaluate only with nodes), (3) ablating with resample and mean ablations and (4) two different faithfulness metrics: correct answer percentage and answer probability.

We measure various significant changes in faithfulness in response to these adjustments. Most importantly, Edge Ablations perform significantly better using a Mean Ablation instead of a Resample Ablation. Had Heimersheim & Janiak (2023) performed edge-level Resample Ablations instead of node-wise Resample Ablations, they may have trusted their circuit significantly less (and if they had used edge-level Mean Ablations, they may have trusted it more).

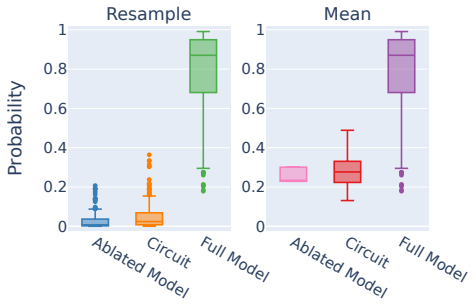
Differentiating by token position also had a large effect on faithfulness scores for both node-wise and edge-wise ablations. These low scores suggest the circuit is in fact performing significant computation on token positions outside of the circuit specified by Heimersheim & Janiak (2023).

When we measure the probability of the correct answer we find that, similar to IOI, the variance between individual prompts is high. This is important for reasons outlined in Section 4.

C.2 Sports Players



(a) The percentage of prompts for which the correct sport has the highest output logit with Mean and Resample Ablations.



(b) The output probability of the correct sport with Mean and Resample Ablations.

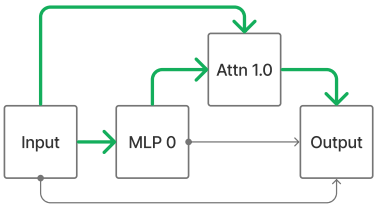
Figure 10: The faithfulness of the Sports Players circuit is reduced when using Resample Ablations.

The Sports Players Task. The Sports Players task (Nanda et al., 2023b) is a simple task that tests the Pythia-2.8b model's (Biderman et al., 2023) ability to recall the sports of famous football, baseball and basketball players. See Table 2 for an example. All prompts follow a very similar format, with the only difference being the name of the sports player in question. The corrupt distribution follows the exact same format, with each clean/corrupt pair having two players of different sports.

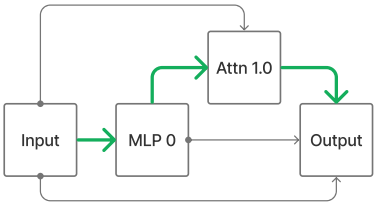
Measuring Sports Players Circuit Faithfulness. In Figure 10, we test the faithfulness of the edge-level sports players circuit, distinguishing token positions while (1) ablating the complement with both Resample and Mean Ablations and (2) calculating two different faithfulness metrics: correct answer percentage (considering only the three possible sports, following Nanda et al. (2023b)) and answer probability.

We find a dramatic difference in correct answer percentage between Resample and Mean Ablation. Note that random guessing would achieve 33% accuracy as there are 3 possible sports, and this is roughly what we see when Mean Ablating the whole model. But Resample Ablating adds signal from the corrupt prompt, which is always a different sport, explaining the 0% accuracy score for the Ablated Model and the Circuit.

C.3 Further Detail on the X-Proportion Tracr Ground Truth Circuits



(theirs) The "ground truth" circuit for the Tracr X-Proportion task using Zero Ablations.



(ours) The "ground truth" circuit for the Tracr X-Proportion task using Resample Ablations.

Figure 11: For the Tracr X-Proportion circuit, the edge from `Input` to `Attn 1.0` is only used to transfer the positional encoding, so it is not required when using Resample Ablations, since these preserve information that is constant between the clean and corrupt distribution. This illustrates the principle that optimal circuits cannot be defined without an ablation methodology. (Nodes `Attn 0.0` and `MLP 1` are not shown as they are not used in this model.)

D Edge-Based vs. Node-Based Circuit Discovery Methods

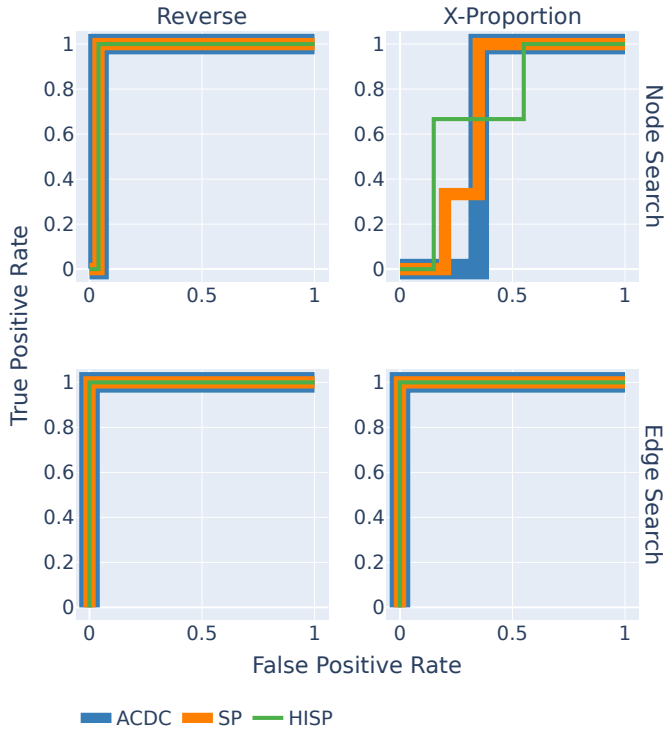


Figure 12: ROC Curves for Edge-Based and Node-Based circuit discovery methods, using the Resample Ablation edges as the ground truth (ours).

In Section 5, we adapted the Subnetwork Probing (SP) and Head Importance Scoring (HISP) circuit discovery methods to use (or approximate) Edge Ablation. ACDC (Conmy et al., 2023) already uses Edge Ablations, but we can similarly adapt ACDC to use Node Ablations. We compare the performance of the Node Patching versions of ACDC, SP and HISP to the Edge Patching versions, for the Resample Ablation based "ground truth" circuit introduced in Section 5 (Figure D).

E Clarifying Nomenclature

Some authors have used different terms for some of the concepts introduced in 3. Activation patching has previously also been called Causal Tracing or Interchange Intervention. In this section, we summarise how our nomenclature relates to the terminology used by Redwood Research in their series of early mechanistic interpretability transformer-circuits papers. Chronologically, these are Wang et al. (2023); Chan et al. (2022); Goldowsky-Dill et al. (2023).

We first discuss the final, most comprehensive work (Chan et al., 2022), which we refer to as Causal Scrubbing. Causal Scrubbing is a very general approach for evaluating circuits together with explanations of the role of nodes within the circuit. It generically comprises performing specific branch-based Resample Ablations on the treeified model on both the circuit *and* its complement.

Causal Scrubbing randomly replaces activations with those that your hypothesis predicts will not change the model output. For instance, if we claim that a given node detects whether the input is even, causal scrubbing will patch in an activation from a different even input. In general, Causal Scrubbing permits an arbitrary number of possible counterfactual inputs.

Goldowsky-Dill et al. (2023) simplify this setup, dropping the strict requirement of requiring an explanation for each node. This reduces the hypothesis class to the classical circuit discovery problem; does some path matter for task performance or not?

Finally Wang et al. (2023) perform a further simplified version of path patching to discover their circuit. This is equivalent to Edge Resample Ablation in our terminology but which they call Path Patching. They patch paths one at a time, to establish which edges are important for task performance. Importantly, Wang et al. (2023) reason that the IOI task should be an attention-only task, as it only comprises moving information between tokens. As such, they take nodes to only be attention heads, with MLPs considered to be part of the direct path between nodes. This approach of one-hop path patching is extended and automated by Conmy et al. (2023).