
How to use and interpret activation patching

Stefan Heimersheim

stefan.heimersheim@gmail.com

Neel Nanda

Abstract

Activation patching is a popular mechanistic interpretability technique, but has many subtleties regarding how it is applied and how one may interpret the results. We provide a summary of advice and best practices, based on our experience using this technique in practice. We include an overview of the different ways to apply activation patching and a discussion on how to interpret the results. We focus on what evidence patching experiments provide about circuits, and on the choice of metric and associated pitfalls.

1 Introduction

1.1 What is activation patching?

Activation patching (also referred to as Interchange Intervention, Causal Tracing, Resample Ablation, or Causal Mediation Analysis) is the technique of replacing internal activations of a neural net. It is also known as Causal Tracing, Resample Ablation, Interchange Intervention or more generally as Causal Mediation Analysis. Variants of this technique have been widely used in the literature (Vig et al., 2020; Geiger et al., 2021a; Geiger, Richardson, and Potts, 2020; Soulos et al., 2019; Finlayson et al., 2021; Geiger et al., 2021b; Meng et al., 2022; Wang et al., 2022; Chan et al., 2022; Hase et al., 2023; Hanna, Liu, and Variengien, 2023; Conmy et al., 2023; Todd et al., 2023; Hendel, Geva, and Globerson, 2023; Feng and Steinhardt, 2023; Lieberum et al., 2023; Cunningham et al., 2023; Stolfo, Belinkov, and Sachan, 2023; Goldowsky-Dill et al., 2023; McDougall et al., 2023; Geva et al., 2023; Huang et al., 2023; Merullo, Eickhoff, and Pavlick, 2023; Tigges et al., 2023). Here we focus on the technique where we overwrite some activations during a model run with cached activations from a previous run (on a different input), and observe how this affects the model’s output.

1.2 How is this related to ablation?

Ablation is the common technique of zeroing out activations. Activation patching is more targeted and controlled: We replace activations with other activations rather than zeroing them out. This allows us to make targeted manipulations to locate specific model behaviours and circuits.

1.3 An example

For example, let’s say we want to know which model internals are responsible for factual recall in ROME (Meng et al., 2022). How does the model complete the prompt “The Colosseum is in” with the answer “Rome”? To answer this question we want to manipulate the model’s activations. But the model activations contain many bits of information: This is an English sentence; The landmark in question is the Colosseum; This is a factual statement about a location.

Ablating some activations will affect the model if these activations are relevant for *any* of these bits. But activation patching allows us to *choose* which bit to change and *control* for the others. Patching with activations from “Il Colosseo è dentro” locates where the model stores the language of the prompt but may use the same factual recall machinery. Patching with activations from “The Louvre is in” locates which part of the model deals with the landmark and information recall. Patching between

“The Colosseum is in the city of” and “The Colosseum is in the country of” locates the part of the model that determines *which* attributes of an entity are recalled.

A simple activation patching procedure typically looks like this:

1. Choose two similar prompts that differ in some key fact or otherwise elicit different model behaviour:

E.g. “The Colosseum is in” and “The Louvre is in” to vary the landmark but control for everything else.

2. Choose which model activations to patch

E.g. MLP outputs

3. Run the model with the first prompt—the source prompt—and save its internal activations

E.g. “The Louvre is in” (source)

4. Run the model with the second prompt—the destination prompt—but overwrite the selected internal activations with the previously saved ones (patching)

E.g. “The Colosseum is in” (destination)

5. See how the model output has changed. The outputs of this patched run are typically somewhere between what the model would output for the un-patched first prompt or second prompt

E.g. observe change in the output logits for “Paris” and “Rome”

6. Repeat for all activations of interest

E.g. sweep to test all MLP layers

1.4 What is this document about

We want to communicate useful practical advice for activation patching, and warn of common pitfalls to avoid. We focus on three areas in particular:

1. What kind of patching experiments provide which evidence? (Section 2)
2. How should you interpret activation patching results? (Section 3)
3. What metrics you can use, what are common pitfalls? (Section 4)

For a general introduction to mechanistic interpretability in general, and activation patching in particular we refer to ARENA¹ chapter 1 (in particular activation patching in chapter 1.3) as well as the corresponding glossary entries on Neel Nanda’s website².

2 What kind of patching experiments should you run?

2.1 Exploratory and confirmatory experiments

In practice we tend to find ourselves in one of two different modes of operation: In exploratory mode we run experiments to find circuits and generate hypotheses. In confirmatory mode we want to verify the circuit we found and check if our hypothesis about its function is correct.

In *exploratory* patching we typically patch components one at a time, often in a sweep over the model (layers, positions, model components). We do this to get an idea of which parts of a model are involved in the task in question, and may be part of the corresponding circuit.

In *confirmatory* patching we want to confirm a hypothesised circuit by verifying that it actually covers all model components needed to perform the task in question. We typically do this by patching many model components at once and checking whether the task performance behaves as expected. A well-known example of patching for circuit verification is Causal Scrubbing (Chan et al., 2022).

¹ARENA: <https://arena3-chapter1-transformer-interp.streamlit.app/>

²<https://neelnanda.io/glossary>

2.2 Which components should you patch

Patching can be done on different levels of granularity determining the components to patch. For example, we may patch the residual stream at a certain *layer* and *position*, or the output of a certain MLP [*layer, position*] or Attention Head [*layer, head, position*]. At even higher granularity we could patch individual neurons or sparse autoencoder features.

An even more specific type of patching is path patching. Usually, patching any component will affect *all* model components in later layers (“downstream”). In path patching instead we let each patch affect only a single target component. We call this patching the “path” between two components. For details on patch patching we refer to ARENA section 1.3.4.

Path Patching can be used to understand whether circuit components affect each other directly, or via mediation by another component. For example if we want to distinguish between *mediation* (component A affects output C via component B), and *amplification/calibration* (component A affects output C directly, but component B reads from A and also affects output C by boosting or cancelling the signal to amplify or calibrate component A). These two structures look identical in default component patching, but different in path patching: a direct connection (composition) between A and C exists only in the second case.

As a rule of thumb, you want to start with low-granularity patching (e.g. residual stream patching), then increase granularity, and finally use path patching to test which components interact with each other. Fast approximations to activation patching, such as attribution patching (see Nanda, 2023, and also AtP*, **atpstar**) can help speed up this process in large models.

2.3 Noising and Denoising

There are multiple ways to do activation patching. The techniques differ in what the source (source of activations / model run from which the activations are copied) and destination prompt (destination that is overwritten / model run in which the activations will be inserted, this is called base in Interchange Interventions language, Geiger et al., 2021b) are. *The use of words “source” and “destination” is unrelated to their meaning in Transformer attention.*

The two main methods are Denoising and Noising (see the next section for other methods).

1. **Denoising:** We can patch activations from a clean first prompt into a corrupted second prompt “clean → corrupt”. That is running the model on the clean prompt while saving its activations, then running the model on the corrupted prompt while overwriting some of its activations with previously saved clean-prompt activations. We observe which patch restores the clean-prompt behaviour, i.e. patching which activations were **sufficient to restore** the behaviour.

2. **Noising:** Or you can patch activations from a corrupted first prompt into a clean second prompt “corrupt → clean”. That is running the model on the corrupted prompt while saving its activations, then running the model on the clean prompt while overwriting some of its activations with previously saved corrupt-prompt activations. We observe which patch breaks the clean-prompt behaviour, i.e. patching which activations were **necessary to maintain** for the behaviour.

An important and underrated point is that these two directions can be very different, and are not just symmetric mirrors of each other. In some situations denoising is the right tool, and in others it’s noising, and understanding the differences is a crucial step in using patching correctly.

Technique	Source (saved)	Source run input	Destination / Base (overwritten)	Destination / Base run input	Observation
Clean → corrupted (Denoising, Causal Tracing ²)	First run activations (clean)	Clean tokens	Second run activations (corrupted)	Corrupt tokens	What restores behaviour
Corrupted → clean (Noising, Resample Ablation)	First run activations (corrupted)	Corrupt tokens	Second run activations (clean)	Clean tokens	What breaks behaviour

For now we round patching effects to “if I patch these activations the model performance is / isn’t affected”. We discuss metrics and measuring patching effects in the last section.

2.4 Example: AND gate vs OR gate

Consider a hypothetical circuit of three components A, B, and C that are connected with an AND or an OR gate. They are embedded in a much larger network, and of the three just C is connected to the output. We run an experiment where we patch all components using the denoising or noising technique.

AND circuit: $C = A \text{ AND } B$

- Denoising (clean \rightarrow corrupt patching): Denoising either A or B has no effect on the output, only denoising C restores the output. This is because denoising A still leaves B at the corrupted (incorrect) baseline, and vice versa. *Denoising found only one of the circuit components.*
- Noising (corrupt \rightarrow clean patching): Noising either A or B has an effect, as well as noising C.

Noising works better in this case, as it finds all circuit components in the first pass.

OR circuit: $C = A \text{ OR } B$

- Denoising (clean \rightarrow corrupt patching): Denoising either A or B has an effect, as well as denoising C.
- Noising (corrupt \rightarrow clean patching): Noising either A or B has no effect on the output, only denoising C restores the output. This is because noising A still leaves B at the clean (correct) baseline, and vice versa. *Denoising found only one of the circuit components.*

Denoising works better in this case, as it finds all circuit components in the first pass. These AND and OR structures can appear in real-world transformers as serial-dependent components (e.g. a later attention head depending on an earlier one) or parallel components (such as redundant backup attention heads).

2.5 Comparison to ablations and other patching techniques

There are activation patching techniques based on a single prompt. The original Causal Tracing (ROME, Meng et al., 2022) falls into this category, and also zero- and mean-ablation can be seen as patching techniques.

1. **Zero ablation:** Overwrite (“ablate”) the targeted activations with zeros and observe ablating which component breaks the model behaviour.
2. **Mean ablation:** Same as above but overwrite targeted activations with their dataset mean value rather than zero. This is slightly more principled than zero ablating since there is no special meaning to activations being zero.
3. **Gaussian noise patching** (also called Causal Tracing*): This is a clean \rightarrow corrupt patching variant that uses as its corrupt run input the embeddings of the clean prompt with added Gaussian noise. The idea is to thereby automatically generate the corrupted “prompt”. It was originally used in ROME (called Causal Tracing there) but has not been used much recently, especially because the corruption can sometimes be ineffective.³

³The success of Gaussian noise corruption is highly sensitive to the noise level. Zhang and Nanda (2023) that if the noise level is just slightly lower than used in ROME, the model can recover the correct completion despite the corruption.

Technique	Source (saved)	Source run input	Destination / Base (over-written)	Destination / Base run input	Observation
Clean → corrupted (Denoising, Causal Tracing ²)	First run activations (clean)	Clean tokens	Second run activations (corrupted)	Corrupt tokens	What restores behaviour
Corrupted → clean (Noising, Resample Ablation)	First run activations (corrupted)	Corrupt tokens	Second run activations (clean)	Clean tokens	What breaks behaviour
Zero ablation	Zero activations	N/A	Clean run activations	Clean tokens	What breaks behaviour
Mean ablation	Dataset mean activations	N/A	Clean run activations	Clean tokens	What breaks behaviour
Gaussian Noise patching (Causal Tracing ²)	First run activations (clean)	Clean tokens	Second run activations (corrupted from modified clean input)	Clean token embedding + Gaussian noise	What restores behaviour

Generally we recommend corrupted-prompt-based techniques, noising and denoising. Their advantage is that one can run very precise experiments, editing some features while controlling for others. They allow us to trace *the difference between clean and corrupted prompt*. To illustrate this consider the prompts “Angela Merkel is the leader of” → “Germany” vs “Joe Biden is the leader of” → “America”. Patching will find components that deal with Angela Merkel vs Joe Biden, but not components that would be indifferent to this change, such as the “answer is a country circuit” or the “political leader circuit”. A secondary advantage of noising and denoising is that they tend to bring the model less out-of-distribution than ablation techniques (as pointed out in Chan et al. (2022), as well as in e.g. Hase, Xie, and Bansal, 2021)

2.6 Choosing corrupted prompts

Having a corrupted prompt is great because it can tell us what model components care about, but also a possible pitfall if we don’t notice what our prompts trace and don’t trace. We give some examples for the Indirect Object Identification (IOI, Wang et al., 2022) demo sentence “John and Mary went to the store. John gave a bottle of milk to”. Different corruptions which highlight different properties the model might care about include:

Corruption	Example	Property traced in model
None (Clean)	John and Mary ... John ...	
Replace the value of one or multiple names, without changing the grammatical structure	John and Alice ... John ... Alice and Mary ... Alice ... Alice and Bob ... Alice ...	Where the model represents the name values
Change which name is direct and indirect object without changing the names or positions	John and Mary ... Mary ...	The value and position of the indirect object
Change the position of the names without changing which one is subject and indirect object	Mary and John ... Mary ...	The value, but not the position of the indirect object (position is fixed)
Change a name to break the behaviour	John and Mary ... Alice ... Alice and Mary ... John ...	Specifics about IOI setting (e.g. that a name is duplicated at all)
Change all the names	Alice and Bob ... Charlie ...	Finding and confirming all relevant components

What kind of prompt should you choose? No matter which you choose, keep in mind what properties your prompt does and does not change, and take this into account when interpreting patching results. As a rule of thumb you want to choose small (narrow) variations for exploratory patching, this will help you narrow down what each component is tracking. Choosing a narrow prompt distribution also helps increase the (typically low) sensitivity of denoising, and decrease the (typically high) sensitivity of noising. For confirmatory patching you need to choose a wide distribution of prompts that varies all variables of the hypothesised circuit. Then you can noise (corrupt → clean patch) all non-circuit components, and check that the model still shows the behaviour in question.

²Causal Tracing has been used to describe ROME-style Gaussian noise patching in particular, but also to describe clean → corrupted patching in general. We recommend avoiding the name to avoid confusion.

3 How do you interpret patching evidence?

In the previous section we said that denoising (clean → corrupt patching) tests whether the patched activations are **sufficient to restore** model behaviour. And noising (corrupt → clean patching) tests whether the patched activations are **necessary to maintain** model behaviour. These two are usually not complements of each other, nor does one imply the other. In this section we will walk through a made-up example experiment.

3.1 Walkthrough a stylized example

Consider the hypothetical “Nobel Peace Prize” circuit. The model correctly completes “Nobel Peace” with “Prize”, using the following circuit:

- Attention head L0H0 is a “Previous Token Head” and copies the embedding of “Nobel” to the position of “Peace”
- Neuron L1N42 maps the mix of Nobel and Peace embeddings to the Prize logit
- Everything else doesn’t matter (of course a real circuit is typically much messier)

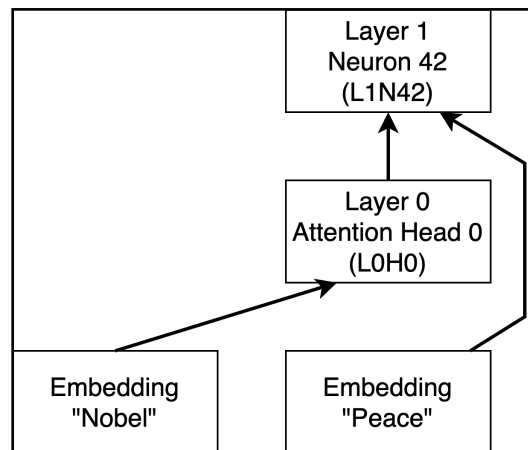


Figure 1: Toy “Nobel Peace Prize” circuit

Now let us run the standard patching examples, take a distribution of random English words for the corrupted prompt. We would find

- Noising (corrupt → clean patching) suggests that the outputs of head L0H0, the output of neuron L1N42, and the embeddings (Nobel & Peace) are all necessary components.
- Denoising (clean → corrupt patching) suggests that the output of neuron L1N42 is sufficient to restore the circuit.

What happened here? Denoising finds only the neuron output L1N42, because the other two components individually are not sufficient to restore the circuit behaviour! We’re dealing with an AND circuit between the attention head output and the “Peace” embedding. Noising finds all three components here.

Nonetheless denoising L1N42 alone restored the model behaviour. This is a crucial intuition to keep in mind about denoising: If you patch component A in layer N, it has seen clean versions of every component in layers 0 to N-1. If there’s an important component B in layer N-1 that is mediated by component A, the model can be restored without denoising B.

Patching experiments are sensitive to what precisely are the changes between the corrupt and clean prompt. If we created two additional corrupt distributions where we replace only either “Nobel” or “Peace” with a random word (i.e. distributions “X Peace” and “Nobel Y”) we could narrow down which component depends on which input.

Alternatively we could use path patching to confirm the precise interactions. Say we want to test whether the Peace embedding is necessary as an input to L0H0, as an input to L1N42, or both. For this we could patch only the corresponding paths, and find that denoising (1) “Nobel → L0H0” and (2) “Peace → L1N42” paths is sufficient. Alternatively we might find that noising every path except for (1) “Nobel → L0H0”, (2) “L0 → L1N42”, and (3) “Peace → L1N42” does not break performance. Note again that denoising only required restoring two paths (restoring a cross-section of the circuit) while noising required leaving 3 paths clean (the full circuit).⁴

3.2 Concepts & gotchas

The walkthrough above presents a typical circuit discovery workflow. We want to highlight a couple of additional concepts and common issues.

Sensitivity & prompt choice: A positive patching result implies you have found activations dealing with the *difference between the clean and corrupt prompt*. Make sure to consider all degrees of freedom in a task, and consider multiple sets of corrupted prompts if necessary.

Scope of activation patching: More generally, activation patching is always based on prompt distributions, and does not make statements for model behaviour outside these specific distributions. For more discussion on the limitations of patching, and the specificity of prompt-based interpretability in general, see Neel Nanda’s writing on [What Can\(’t\) Activation Patching Teach Us](#).

No minimality: Here, and in many parts of the literature, a circuit is treated as a collection of model components that are responsible for a particular model behaviour. We typically make no claims that we have found the *smallest* such collection of components, we only test that this collection is sufficient.

Backup behaviour & OR-gates: In some cases researchers have discovered “Backup heads”, components that are not normally doing the task but jump into action if other components are disrupted (Hydra effect, McGrath et al., 2023). For example, in IOI when one ablates a name mover head (a key component of the circuit) a backup name mover head will activate and then do the task instead (Wang et al., 2022).

It can be helpful to think of these as OR-gates where either component is sufficient for the model to work. This does not fit well into our attempts of defining a circuit, nor plays well with the circuit finding methods above. Despite the name mover heads being important, if we ablate them then, due to backup heads compensating, the name movers look less important. Fortunately, backup behaviour seems to be lossy, i.e. if the original component boosted the logits by $+X$, the backup compensates for this by boosting less than X (the Hydra effect paper found $0.7 \cdot X$). Thus these backup component weaken the visibility of the original component, but it is usually still visible since even $0.3 \cdot X$ is a relatively large effect.

Negative components: Some work in this area (e.g. Wang et al., 2022; Heimersheim and Janiak, 2023) noticed attention heads that consistently negatively affected performance, and noising them would increase performance. This is problematic, because it makes it hard to judge the quality of a circuit analysis: it may look like we’ve fully recovered (or more than fully recovered!) performance, by finding half the positive components but excluding all negative ones. This is an unsolved problem. Conmy et al. (2023) propose using Kullback Leibler (KL) divergence as a metric to address this, which penalises any deviation (positive or negative), at the cost of also tracking lots of variation we may not care about.

⁴This method doesn’t yet confirm which information is carried in the different paths. We can go a step further and noise (corrupt → clean patch) even some of the important circuit connections, namely “Nobel → L0H0 → L1N42” path from the “Nobel Y” distribution, and the “Peace → L1N42” path from the “X Peace” distribution. Doing that is essentially Causal Scrubbing (Chan et al., 2022).

4 Metrics and common pitfalls

So far we talked about “preserving” and “restoring” performance, but in practice, model performance is not binary but a scale. Typically we find some components matter a lot, while others provide a small increase in performance. For the best interpretability we might look for a circuit restoring e.g. 90% of the model’s performance, rather than reaching exactly 100% (for examples see Chan et al., 2022). A useful framing is the “pareto frontier” of circuit size vs. performance recovered - recovering 80% of performance with 1% of the components is more impressive than 90% of the performance with 10% of the components, but there will always be a minimum circuit size to recover a given level of performance.

It’s easy to treat metrics as an after-thought, but we believe that the right or wrong choice of a metric can significantly change the interpretation of patching results. Especially for exploratory patching, the wrong metric can be misleading. The choice of metric matters less for confirmatory patching, where you expect a binary-ish answer (“have I found the circuit or not”) and all metrics should agree. We’ll go through a couple of metric choices in this section:

Based on	Example
Logit difference (= Logprob difference)	Logit(Mary) - Logit(John)
Logarithmic probability (logsoftmax)	Logprob(Mary)
Probability (softmax)	Prob(Mary)
Accuracy / Rank of correct answer	Rank(Mary)==0

An honourable mention goes to the KL divergence. Unlike the previous metrics, this metric aims to compare the full model output, rather than focusing on a specific task. KL divergence is a good metric in such cases.

In addition to these output based metrics, in some cases it makes sense to consider some model internals as metrics themselves. For example, one might use the attention paid by the name mover head to the indirect object as a metric to identify the subcircuit controlling this head, or the activation of a key neuron or SAE feature, or the projection onto a probe (Nanda et al., 2023).

In our experience, it’s worth implementing many metrics and briefly analysing all of them. Computing a metric is cheap (compared to the cost of the forward pass), and they all have different strengths and weaknesses, and can illuminate different parts of the big picture. And if they all agree that’s stronger evidence than any metric on its own. Where they disagree, we personally trust **logit difference** (or equivalently logprob difference) the most.

4.1 The logit difference

Logit difference measures to what extent the model knows the correct answer, and it allows us to be specific: We can control for things we don’t want to measure (e.g. components that boost both Mary and John, in the IOI example) by choosing the right logits to compare (e.g. Mary vs John, or multiple-choice answers). The metric also is a mostly linear function of the residual stream (unlike probability-based metrics) which makes it easy to directly attribute logit difference to individual components (“direct logit attribution”, “logit lens”). It’s also a “softer” metric, allowing us to see partial effects on the model even if they don’t change the rank of the output tokens (unlike e.g. accuracy), which is crucial for exploratory patching. We discuss problems with this and other metrics in the next section.

Intuition for why logits and logit differences (LDs) are a natural unit for transformers: The residual stream and output of a transformer is a sum of components. Every component added to the residual stream corresponds to an addition to the LD (as the LD corresponds to a residual stream direction, up to layer norm). A model component can easily change the LD by some absolute amount (e.g. +1 LD). It cannot easily change the LD by a relative amount ($LD \times = 1.5$), or change the probabilities by a specific amount ($prob \neq 0.20$). For example consider a model component that always outputs -1 logit to duplicated names (assume “John and Mary ... John ...”). This component then always writes +1 LD in favour of Mary, and gets a score of 1 in terms of LD. Other metrics (such as probability) judge this component differently, depending on what the baseline was (e.g. due to other patches). We would argue that logits and logit differences are closer to the mechanistic process happening in the

transformer, and thus feel like a more natural unit. This is of course not a requirement, and also does not hold in all places (e.g. if a component's output depends on the input LD), but it seems to work well in practice.

4.2 Flaws & advantages of different metrics

It is essential to be aware of what a metric measures and is sensitive to. A key thing to track is whether the metric is discrete vs continuous, and whether it's exponential vs linear (in the logits) - continuous. Linear metrics are usually more accurate, which is crucial when doing exploratory patching and assigning "partial credit" to model components. Here we list common pitfalls of popular metrics.

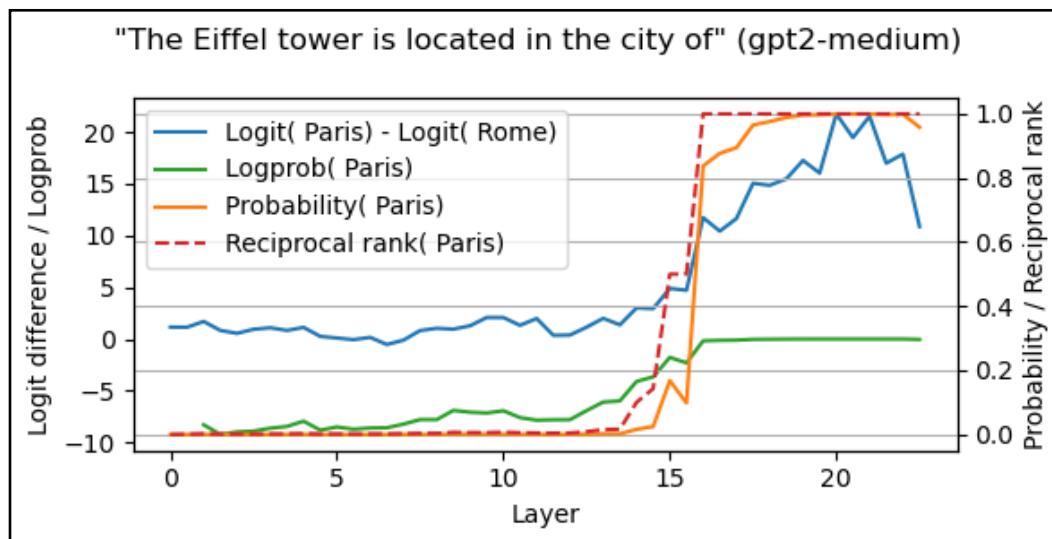


Figure 2: Illustration of different metrics for an example patching experiment with GPT-2 medium.

- **Logit difference / logprob difference:** The difference between the logit of the correct answer and the incorrect answer(s). This metric specifically measures the difference between the selected logits, and is not sensitive to components which affect all of them. For example, in IOI it measures the model's confidence in Mary vs John which encapsulates the IOI-circuit well without being sensitive to the "is the next token a name?"-circuit.
- **Potential false-positive:** Because the metric is a *difference* it may be driven by either getting better at the correct answer or worse at the incorrect answer. Thus it is worth checking the logits or logprobs of individual answers to confirm. This is particularly concerning because the corrupted model likely puts a high probability on the incorrect answer. This means that any patch that indiscriminately damages the model and gets it closer to uniform will damage the incorrect answer logprob and so boost the logit diff.
- **Logprobs:** This metric measures the logprob of the correct answer. It is sensitive to absolute change in logarithmic probabilities (i.e. relative change in probabilities) and captures our intuition for what good model performance means. We broadly think it is a good metric. Its main flaws are
 - **Saturation:** Once the correct answer becomes the model's top guess, the logprob stops increasing meaningfully, even though the confidence can increase much more. We can see this in Figure 2, where the green line saturates after layer 17.
 - **Unspecificity:** We lose the ability to control for other properties, e.g. in IOI we cannot distinguish between components that increase both $P(\text{John})$ and $P(\text{Mary})$ from components that only increase $P(\text{Mary})$. This can be intended, or unintended, it's just important to keep in mind.
 - **Inhibition:** To increase the logprob on John, the model can either increase the John logit, or decrease other top logits, and it is hard to distinguish which is happening. This

may be desirable or not because the two operations likely have different mechanisms and may be better tracked separately.

- Probabilities: This metric measures the probability of the right answer, or the difference in probabilities of different answers. The main issue with such metrics is

- Probabilities are non-linear, in the sense that they track the logits exponentially. For example, a model component adding +2 to a given logit can create a 1 or 40 percentage point probability increase, just depending on what the baseline was.

As an example of the non-linearity consider the orange line in the figure above: A modest increase in logit difference around layer 17 converts to a jump in probability.

- Probabilities also inherit the problems of the logprob metric, namely saturation and unspecificity.

The figure shows the saturation effect for the orange line at layer >18.

- Binary and discrete metrics (Accuracy / top-k performance / rank / etc): These metrics round off each input to a discrete metric (and then tend to average over a bunch of inputs).

- The problem with these is that generally many components contribute to a model's performance, with no single decisive contributor. Discrete metrics may suggest that some significant contributors are unimportant, because they aren't enough to cross a threshold. Alternatively, these metrics may suggest that one contributor among many is *all* that matters because it happens to be the one that pushes the model over the threshold. We generally recommend using continuous metrics instead.

As an example consider the Figure above: The rank-based metric (red line) jumps around layer 15 when the corresponding logit passes the rank 1 and 0 thresholds, while it is not sensitive to any of the other changes.

- Discrete metrics can be a good fit for confirmatory patching rather than exploratory patching, as in some sense accuracy *is* the metric we care about - can the model get the question right or not?

- Logits: We could just take the answer *logit* as a metric. This is somewhat unprincipled because logits have an arbitrary baseline (adding +1 to all logits would not affect the output) but tend to work in practice. Logit(John) often matches Logprob(John) without being affected by the downsides of the logprob metric.

This metric can incorrectly pick up on components that just contribute to many logits. Ensuring that the residual stream and logits have mean zero (default in TransformerLens) can help address this.

Desired property:	Continuous (not discrete)	Track model confidence linearly	Measure John vs Mary independent of P(name)	False-positive when "breaking the model"
Logit Difference:	Yes	Yes	Yes	Yes
Logit:	Yes	Yes (typically)	Yes	Maybe
Logprob:	Yes	Not close to $p=1$	No	No
Probability:	Yes	No	No	No
Binary metrics:	No	No	No	No

5 Summary

In most situations, use activation patching instead of ablations. Different corrupted prompts give you different information, be careful about what you choose and try to test a range of prompts.

There are two different directions you can patch in: denoising and noising. These are *not* symmetric. Be aware of what a patching result implies!

- Denoising (a clean \rightarrow corrupt patch) shows whether the patched activations were **sufficient to restore** the model behaviour. This implies the components make up a cross-section of the circuit.
- Noising (a corrupt \rightarrow clean patch) shows whether the patched activations were **necessary to maintain** the model behaviour. This implies the components are part of the circuit.

Be careful when using metrics that are (i) discrete, (ii) overly sharp, or (iii) sensitive to unintended information. Ideally use a range of metrics, and try to have at least one metric that is continuous and roughly linear in logits such as logit difference or logprob. We recommend representing patching results in a big dataframe with a column per metric and row per patching experiment, and making a bunch of plots from this.

- Model top-k accuracy is discrete and can overrepresent changes at thresholds and shows no change for large effects that don't cross thresholds.
- Most effects from patching are linear and additive in logit space. Probability is exponential in logit space, so it overemphasises effects near a threshold and suppresses effects elsewhere, creating overly sharp patching plots
- Logprob can saturate, and cannot control for a patch that boosts both the correct and incorrect answer(s)

Acknowledgements

Thanks to Arthur Conmy, Chris Mathwin, James Lucassen, and Fred Zhang for comments on a draft of this manuscript.

References

- Chan, Lawrence, Adria Garriga-Alonso, Nix 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. URL: <https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing>.
- Conmy, Arthur, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso (Apr. 2023). “Towards Automated Circuit Discovery for Mechanistic Interpretability”. In: *arXiv e-prints*, arXiv:2304.14997, arXiv:2304.14997. DOI: 10.48550/arXiv.2304.14997. arXiv: 2304.14997 [cs.LG].
- Cunningham, Hoagy, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey (Sept. 2023). “Sparse Autoencoders Find Highly Interpretable Features in Language Models”. In: *arXiv e-prints*, arXiv:2309.08600, arXiv:2309.08600. DOI: 10.48550/arXiv.2309.08600. arXiv: 2309.08600 [cs.LG].
- Feng, Jiahai and Jacob Steinhardt (Oct. 2023). “How do Language Models Bind Entities in Context?”. In: *arXiv e-prints*, arXiv:2310.17191, arXiv:2310.17191. DOI: 10.48550/arXiv.2310.17191. arXiv: 2310.17191 [cs.LG].
- Finlayson, Matthew, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov (June 2021). “Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models”. In: *arXiv e-prints*, arXiv:2106.06087, arXiv:2106.06087. DOI: 10.48550/arXiv.2106.06087. arXiv: 2106.06087 [cs.CL].
- Geiger, Atticus, Hanson Lu, Thomas Icard, and Christopher Potts (June 2021a). “Causal Abstractions of Neural Networks”. In: *arXiv e-prints*, arXiv:2106.02997, arXiv:2106.02997. DOI: 10.48550/arXiv.2106.02997. arXiv: 2106.02997 [cs.AI].

Geiger, Atticus, Kyle Richardson, and Christopher Potts (Apr. 2020). “Neural Natural Language Inference Models Partially Embed Theories of Lexical Entailment and Negation”. In: *arXiv e-prints*, arXiv:2004.14623, arXiv:2004.14623. DOI: 10.48550/arXiv.2004.14623. arXiv: 2004.14623 [cs.CL].

Geiger, Atticus, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts (Dec. 2021b). “Inducing Causal Structure for Interpretable Neural Networks”. In: *arXiv e-prints*, arXiv:2112.00826, arXiv:2112.00826. DOI: 10.48550/arXiv.2112.00826. arXiv: 2112.00826 [cs.LG].

Geva, Mor, Jasmijn Bastings, Katja Filippova, and Amir Globerson (Apr. 2023). “Dissecting Recall of Factual Associations in Auto-Regressive Language Models”. In: *arXiv e-prints*, arXiv:2304.14767, arXiv:2304.14767. DOI: 10.48550/arXiv.2304.14767. arXiv: 2304.14767 [cs.CL].

Goldowsky-Dill, Nicholas, Chris MacLeod, Lucas Sato, and Aryaman Arora (Apr. 2023). “Localizing Model Behavior with Path Patching”. In: *arXiv e-prints*, arXiv:2304.05969, arXiv:2304.05969. DOI: 10.48550/arXiv.2304.05969. arXiv: 2304.05969 [cs.LG].

Hanna, Michael, Ollie Liu, and Alexandre Variengien (Apr. 2023). “How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model”. In: *arXiv e-prints*, arXiv:2305.00586, arXiv:2305.00586. DOI: 10.48550/arXiv.2305.00586. arXiv: 2305.00586 [cs.CL].

Hase, Peter, Mohit Bansal, Been Kim, and Asma Ghandeharioun (Jan. 2023). “Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models”. In: *arXiv e-prints*, arXiv:2301.04213, arXiv:2301.04213. DOI: 10.48550/arXiv.2301.04213. arXiv: 2301.04213 [cs.LG].

Hase, Peter, Harry Xie, and Mohit Bansal (June 2021). “The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations”. In: *arXiv e-prints*, arXiv:2106.00786, arXiv:2106.00786. DOI: 10.48550/arXiv.2106.00786. arXiv: 2106.00786 [cs.LG].

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

Hendel, Roei, Mor Geva, and Amir Globerson (Oct. 2023). “In-Context Learning Creates Task Vectors”. In: *arXiv e-prints*, arXiv:2310.15916, arXiv:2310.15916. DOI: 10.48550/arXiv.2310.15916. arXiv: 2310.15916 [cs.CL].

Huang, Jing, Atticus Geiger, Karel D’Oosterlinck, Zhengxuan Wu, and Christopher Potts (Sept. 2023). “Rigorously Assessing Natural Language Explanations of Neurons”. In: *arXiv e-prints*, arXiv:2309.10312, arXiv:2309.10312. DOI: 10.48550/arXiv.2309.10312. arXiv: 2309.10312 [cs.CL].

Lieberum, Tom, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik (July 2023). “Does Circuit Analysis Interpretability Scale? Evidence from Multiple Choice Capabilities in Chinchilla”. In: *arXiv e-prints*, arXiv:2307.09458, arXiv:2307.09458. DOI: 10.48550/arXiv.2307.09458. arXiv: 2307.09458 [cs.LG].

McDougall, Callum, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda (Oct. 2023). “Copy Suppression: Comprehensively Understanding an Attention Head”. In: *arXiv e-prints*, arXiv:2310.04625, arXiv:2310.04625. DOI: 10.48550/arXiv.2310.04625. arXiv: 2310.04625 [cs.LG].

McGrath, Thomas, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg (July 2023). “The Hydra Effect: Emergent Self-repair in Language Model Computations”. In: *arXiv e-prints*, arXiv:2307.15771, arXiv:2307.15771. DOI: 10.48550/arXiv.2307.15771. arXiv: 2307.15771 [cs.LG].

Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov (Feb. 2022). “Locating and Editing Factual Associations in GPT”. In: *arXiv e-prints*, arXiv:2202.05262, arXiv:2202.05262. DOI: 10.48550/arXiv.2202.05262. arXiv: 2202.05262 [cs.CL].

Merullo, Jack, Carsten Eickhoff, and Ellie Pavlick (Oct. 2023). “Circuit Component Reuse Across Tasks in Transformer Language Models”. In: *arXiv e-prints*, arXiv:2310.08744, arXiv:2310.08744. DOI: 10.48550/arXiv.2310.08744. arXiv: 2310.08744 [cs.CL].

Nanda, Neel (2023). *Attribution Patching: Activation Patching At Industrial Scale*. Blogpost. Section “How to Think About Activation Patching”. URL: <https://www.neelnanda.io/mechanistic-interpretability/attribution-patching>.

Nanda, Neel, SenR, János Kramár, and Rohin Shah (2023). *Fact Finding: Attempting to Reverse-Engineer Factual Recall on the Neuron Level (Post 1)*. Alignment Forum. URL: <https://www.alignmentforum.org/posts/u6KXXmKFbXfWzoAXn/a-circuit-for-python-docstrings-in-a-4-layer-attention-only>.

alignmentforum.org/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall]

Soulos, Paul, Tom McCoy, Tal Linzen, and Paul Smolensky (Oct. 2019). “Discovering the Compositional Structure of Vector Representations with Role Learning Networks”. In: *arXiv e-prints*, arXiv:1910.09113, arXiv:1910.09113. DOI: 10.48550/arXiv.1910.09113. arXiv: 1910.09113 [cs.LG].

Stolfo, Alessandro, Yonatan Belinkov, and Mrinmaya Sachan (May 2023). “A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis”. In: *arXiv e-prints*, arXiv:2305.15054, arXiv:2305.15054. DOI: 10.48550/arXiv.2305.15054. arXiv: 2305.15054 [cs.CL].

Tigges, Curt, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda (Oct. 2023). “Linear Representations of Sentiment in Large Language Models”. In: *arXiv e-prints*, arXiv:2310.15154, arXiv:2310.15154. DOI: 10.48550/arXiv.2310.15154. arXiv: 2310.15154 [cs.LG].

Todd, Eric, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau (Oct. 2023). “Function Vectors in Large Language Models”. In: *arXiv e-prints*, arXiv:2310.15213, arXiv:2310.15213. DOI: 10.48550/arXiv.2310.15213. arXiv: 2310.15213 [cs.CL].

Vig, Jesse, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber (Apr. 2020). “Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias”. In: *arXiv e-prints*, arXiv:2004.12265, arXiv:2004.12265. DOI: 10.48550/arXiv.2004.12265. arXiv: 2004.12265 [cs.CL].

Wang, Kevin, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt (Nov. 2022). “Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small”. In: *arXiv e-prints*, arXiv:2211.00593, arXiv:2211.00593. DOI: 10.48550/arXiv.2211.00593. arXiv: 2211.00593 [cs.LG].

Zhang, Fred and Neel Nanda (Sept. 2023). “Towards Best Practices of Activation Patching in Language Models: Metrics and Methods”. In: *arXiv e-prints*, arXiv:2309.16042, arXiv:2309.16042. DOI: 10.48550/arXiv.2309.16042. arXiv: 2309.16042 [cs.LG].