

Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching

Aleksandar Makelov*
 aleksandar.makelov@gmail.com
 SERI MATS

Georg Lange*
 mail@georglange.com
 SERI MATS

Neel Nanda
 neelnanda27@gmail.com

Abstract

Mechanistic interpretability aims to understand model behaviors in terms of specific, interpretable features, often hypothesized to manifest as low-dimensional subspaces of activations. Specifically, recent studies have explored subspace interventions (such as activation patching) as a way to simultaneously manipulate model behavior and attribute the features behind it to given subspaces.

In this work, we demonstrate that these two aims diverge, potentially leading to an illusory sense of interpretability. Counterintuitively, even if a subspace intervention makes the model’s output behave *as if* the value of a feature was changed, this effect may be achieved by activating a *dormant parallel pathway* leveraging another subspace that is *causally disconnected* from model outputs. We demonstrate this phenomenon in a distilled mathematical example, in two real-world domains (the indirect object identification task and factual recall), and present evidence for its prevalence in practice. In the context of factual recall, we further show a link to rank-1 fact editing, providing a mechanistic explanation for previous work observing an inconsistency between fact editing performance and fact localization.

However, this does not imply that activation patching of subspaces is intrinsically unfit for interpretability. To contextualize our findings, we also show what a success case looks like in a task (indirect object identification) where prior manual circuit analysis informs an understanding of the location of a feature. We explore the additional evidence needed to argue that a patched subspace is faithful.

1 Introduction

Recently, large language models (LLMs) have demonstrated impressive (Vaswani et al., 2017; Devlin et al., 2019; OpenAI, 2023; Radford et al., 2019; Brown et al., 2020), and often surprising (Wei et al., 2022), capability gains. However, they are still widely considered ‘black boxes’: their successes – and failures – remain largely a mystery. It is thus an increasingly pressing scientific and practical question to understand *what* LLMs learn and *how* they make predictions.

This is the goal of machine learning interpretability, a broad field that presents us with both technical and conceptual challenges (Lipton, 2016). Within it, mechanistic interpretability (MI) is a subfield that seeks to develop a rigorous low-level understanding of the mechanisms and

*Equal Contribution.

learned algorithms behind a model’s computations. MI frames these computations as collections of narrow, task-specific algorithms – *circuits* (Olah et al., 2020; Geiger et al., 2021; Wang et al., 2023) – whose operations are grounded in concrete, atomic building blocks akin to variables in a computer program (Olah, 2022) or causal model (Vig et al., 2020; Geiger et al., 2023a). MI has found applications in several downstream tasks: removing toxic behaviors from a model while otherwise preserving performance by minimally editing model weights (Li et al., 2023b), changing factual knowledge encoded by models in specific components to e.g. enable more efficient fine-tuning in a changing world (Meng et al., 2022a), improving the truthfulness of LLMs at inference time via efficient, localized inference-time interventions in specific subspaces (Li et al., 2023a) and studying the mechanics of gender bias in language models (Vig et al., 2020).

A central question in MI is: what is the proper definition of these building blocks? Any satisfying mechanistic analysis of high-level LLM capabilities must rest on a rigorous and comprehensive answer to this question (Olah, 2022). Many initial mechanistic analyses have focused on mapping circuits to collections of *model components* (Wang et al., 2023; Heimersheim & Janiak), such as attention heads and MLP layers. A workhorse of these analyses is *activation patching*¹ (Vig et al., 2020; Geiger et al., 2020; Meng et al., 2022a; Wang et al., 2023), which intervenes on model computation on an input by replacing the activation of a given component with its value when the model is run on another input. By seeing which components lead to a significant task-relevant change in outputs compared to running the model normally, activation patching aims to pinpoint tasks to specific components.

However, localizing features to entire components is not sufficient for a detailed understanding. A plethora of empirical evidence suggests that the features LLMs represent and use are more accurately captured by *linear subspaces* of component activations (Nanda, 2023a; Li et al., 2021; Abdou et al., 2021; Grand et al., 2018). Complicating matters, phenomena like superposition and polysemyticity (Elhage et al., 2022) suggest that these subspaces are not easily enumerable, like individual neurons – so searching for them can be non-trivial. This raises the question:

Does the success of activation patching carry over from component-level analysis to finding the precise subspaces corresponding to features?

In this paper, we demonstrate that naive generalizations of subspace activation patching can lead to misleading interpretability results. Specifically, we argue empirically and theoretically that a subspace seemingly encoding some feature may be found in the MLP layers on the path between two model components in a transformer model that communicate this feature as part of some circuit.

As a concrete example of how this illusion can happen in the practice of interpretability, recent works such as Geiger et al. (2023b); Wu et al. (2023) have sought to identify interpretable subspaces using gradient descent, with training objectives that optimize for a subspace patch with a causal effect on model predictions. While this kind of end-to-end optimization has promise, we show that, instead of localizing a variable used by the model, subspace interventions such as subspace activation patching can create such a variable by *activating a dormant pathway*.

Counterintuitively, the mathematics of subspace interventions makes it possible to activate another, ‘dormant’, direction, which is ordinarily inactive, but can change model outputs when activated (see Figure 1), by exploiting the variation of model activations in a direction correlated with a feature even if this second direction does not causally affect the output. An equivalent view of this phenomenon that we explore in Appendix A.3 is that the component contains two subspaces

¹also known as ‘interchange intervention’ (Geiger et al., 2020) and sometimes referred to as ‘resample ablation’ (Chan et al., 2022) or ‘causal tracing’ (Meng et al., 2022a).

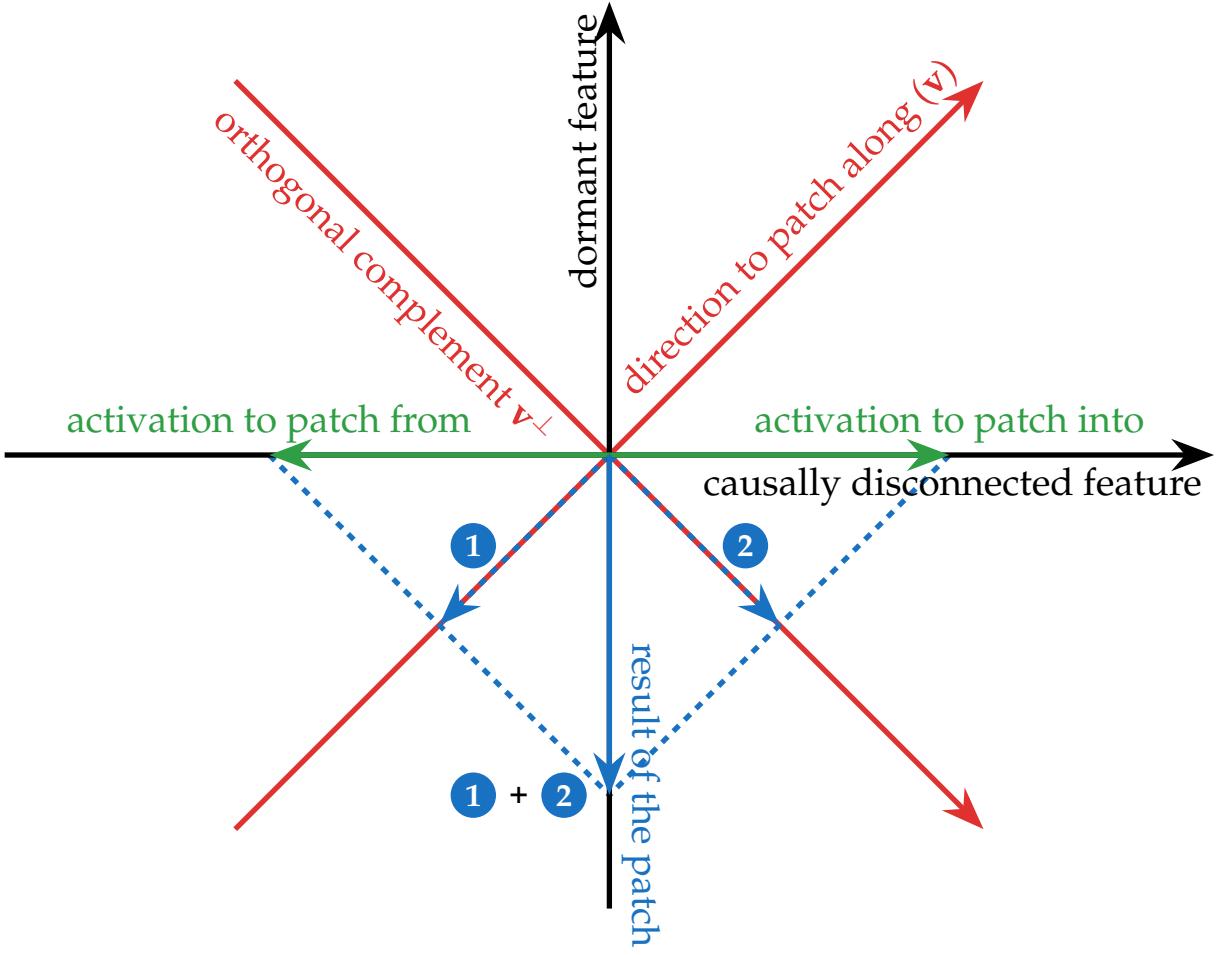


Figure 1: The key mathematical phenomenon behind the activation patching illusion illustrated for a 2-dimensional activation space. We intervene on an example's **activation** (green, right) by setting its orthogonal projection on a **1-dimensional subspace v of activation space** (red, top-right) to equal the orthogonal projection of another example's **activation** (green, left) on v . The result is a **patched activation vector** orthogonal to both activations. Specifically, to form the patched activation we take the v component (①) of the activation we are patching from, and combine it with the v^\perp component (②) of the original activation. This results in the patched activation ① + ②. This can lead to counterintuitive results when the original and new directions have fundamentally different roles in a model's computation; see Section 3 for details, and Figure 14 for a step-by-step guide through this figure.

that mediate the variable, but whose effects normally cancel each other out (hence, there’s no total effect, making the component as a whole appear ‘dormant’). The activation patching intervention decouples these two subspaces by altering an activation only along one of them. Under this perspective, our contribution is to show that model components are likely to contain such pairs of subspaces that perfectly cancel out. When this phenomenon is realized in the hidden activations of an MLP layer, it leads to causally meaningful subspaces which have a substantial and crucial component that is causally disconnected from model outputs, owing to the high-dimensional kernel of an MLP layer’s down-projection in a transformer (see Figure 3).

While it is, in principle, possible that subspaces that represent some variable but cancel each other out exist in many model components, we find this unlikely. Specifically, our results suggest that every MLP layer between two components communicating some feature through residual connections is likely to contain a subspace which appears to mediate the feature when activation patched. We find this implausible on various grounds that we revisit in Section 8. Thus, we consider at least some of these subspaces to exhibit a kind of *interpretability illusion* (Bolukbasi et al., 2021; Adebayo et al., 2018).

Our contributions can be summarized as follows:

- In Section 3, we provide the key intuition for the illusion, and construct a distilled mathematical example.
- In Section 4, we find a realization of this phenomenon ‘in the wild’, in the context of the indirect object identification task (Wang et al., 2023), where a 1-dimensional subspace of MLP activations found using DAS (Geiger et al., 2023b) can seem to encode position information about names in the sentence, despite this MLP layer having negligible contribution to the circuit as argued by Wang et al. (2023).
- To contextualize our results, in Section 5 we also show how DAS can be used to find subspaces that faithfully represent a feature in a model’s computation. Specifically, we find a 1-dimensional subspace encoding the same position information in the IOI task, and validate its role in model computations via mechanistic experiments beyond end-to-end causal effect. We argue that activation patching on subspaces of the residual stream is less prone to illusions.
- Going beyond the IOI task, in Section 6 we also exhibit this phenomenon in the setting of *fact editing* (Meng et al., 2022a). We show that 1-dimensional activation patches imply approximately equivalent rank-1 model edits (Meng et al., 2022a). In particular, this shows that rank-1 model edits can also be achieved by activating a dormant pathway in the model, without necessarily relying on the presence of a fact in the weight being edited. This suggests a mechanistic explanation for the observation of (Hase et al., 2023) that rank-1 model editing ‘works’ regardless of whether the fact is present in the weights being edited.
- In Section 7, we collect arguments and evidence for why this interpretability illusion ought to be prevalent in real-world language models.
- Finally, in Section 8, we provide conceptual discussion of these findings.

We have also released code to reproduce our findings².

²<https://github.com/amakelov/activation-patching-illusion>

2 Related Work

2.1 Discovering and Causally Intervening on Representations with Activation Patching

Researchers have been exploring increasingly fine-grained ways of reverse-engineering and steering model behavior. In this context, *activation patching* (Vig et al., 2020; Geiger et al., 2020) is a widely used causal intervention, whereby the model is run on an input A, but chosen activations are ‘patched in’ from input B. Motivated by causal mediation analysis (Pearl, 2001) and causal abstraction Geiger et al. (2023a), activation patching has been used to localize model components causally involved in various behaviors, such as gender bias (Vig et al.), factual recall (Meng et al., 2022a), multiple choice questions (Lieberum et al., 2023), arithmetic (Stolfo et al., 2023) and natural language reasoning (Geiger et al., 2021; Wang et al., 2023; Geiger et al., 2023b; Wu et al., 2023), code (Heimersheim & Janiak), and (in certain regimes) topic/sentiment/style of free-form natural language (Turner et al., 2023).

Activation patching is an area of active research, and many recent works have extended the method, with patching paths between components (Goldowsky-Dill et al., 2023), automating the finding of sparse subgraphs (Conmy et al., 2023), fast approximations (Nanda, 2023b), and automating the verification of hypotheses (Chan et al., 2022).

In particular, *full-component activation patching* – where the entire activation of a model component such as attention head or MLP layer is replaced – is not the end of the story. A wide range of interpretability work (Mikolov et al., 2013; Conneau et al., 2018; Hewitt & Manning, 2019; Tenney et al., 2019; Burns et al., 2022; Nanda et al., 2023) suggests the *linear representation hypothesis*: models encode features as linear subspaces of component activations that can be arbitrarily rotated with respect to the standard basis (due to phenomena like superposition, polysemanticity (Arora et al., 2018; Elhage et al., 2022) and lack of privileged bases (Smolensky, 1986; Elhage et al., 2021)).

Motivated by this, recent works such as Geiger et al. (2023b); Wu et al. (2023); Lieberum et al. (2023) have explored *subspace activation patching*: a generalization of activation patching that operates only on linear subspaces of features (as low as 1-dimensional) rather than patching entire components.

Our work contributes to this research direction by demonstrating both (i) a common illusion to avoid when looking for such subspaces and (ii) a detailed case study of successfully localizing a binary feature to a 1-dimensional subspace.

2.2 Interpretability Illusions

Despite the promise of interpretability, it is difficult to be rigorous and easy to mislead yourself. A common theme in the field is identifying ways that techniques may lead to misleading conclusions about model behavior (Lipton, 2016). In computer vision, Adebayo et al. (2018) show that a popular at the time class of pixel attribution methods is not sensitive to whether or not the model used to produce is has actually been trained or not. In Geirhos et al. (2023), the authors show how a circuit can be hardcoded into a learned model so that it fools interpretability methods; this bears some similarity to our illusion, especially its fact editing counterpart.

In natural language processing, Bolukbasi et al. (2021) show that interpreting single neurons with maximum activating dataset examples may lead to conflicting results across datasets due to subtle polysemanticity (Elhage et al., 2022). Recently, McGrath et al. (2023) demonstrated that full-component activation patching in large language models is vulnerable to false negatives due to (ordinarily dormant) backup behavior of downstream components that activates when a

component is ablated.

We contribute to the study of interpretability illusions by demonstrating a new kind of illusion which can arise when intervening on model activations along arbitrary subspaces, by demonstrating it in two real-world scenarios, and providing recommendations on how to avoid it.

2.3 Factual Recall

A well-studied domain for discovering and intervening on learned representations is the localization and editing of factual knowledge in language models (Wallat et al., 2020; Meng et al., 2022b; Dai et al., 2022; Geva et al., 2023; Hernandez et al., 2023). A work of particular note is Meng et al. (2022a), which localizes factual information using a variation of full-component activation patching, and then edits factual information with a rank-1 intervention on model weights. However, recent work has shown that rank-1 editing can work even on weights where the fact supposedly is not encoded (Hase et al., 2023), and that editing a single fact often fails to have its expected common-sense effect on logically related downstream facts (Cohen et al., 2023; Zhong et al., 2023).

We contribute to this line of work by showing a formal and empirical connection between activation patching along 1-dimensional subspaces and rank-1 model editing. In particular, rank-1 model edits can work by creating a dormant pathway of an MLP layer, regardless of whether the fact is stored there. This provides a mechanistic explanation for the discrepancy observed in Hase et al. (2023).

3 A Conceptual View of the Illusion

3.1 Preliminaries: (Subspace) Activation Patching

Activation patching (Vig et al., 2020; Geiger et al., 2020; Wang et al., 2023; Chan et al., 2022) is an interpretability technique that intervenes upon model components, forcing them to take on values they would have taken if a different input were provided. For instance, consider a model that has knowledge of the locations of famous landmarks, and completes e.g. the sentence $A = \text{'The Eiffel Tower is in'}$ with 'Paris' .

How can we find which components of the model are responsible for knowing that 'Paris' is the right completion? Activation patching approaches this question by

- (i) Running the model on A ;
- (ii) Storing the activation of a chosen model component \mathcal{C} , such as the output of an attention head, the hidden activations of an MLP layer, or an entire residual stream (a.k.a. bottleneck) layer;
- (iii) Running the model on e.g. $B = \text{'The Colosseum is in'}$, but with the activation of \mathcal{C} taken from A .

If we find that the model outputs 'Paris' instead of 'Rome' in step (iii), this suggests that component \mathcal{C} is important for the task of recalling the location of a landmark.

The linear representation hypothesis proposes that *linear subspaces* of vectors will be the most interpretable model components. To search for such subspaces, we can adopt a natural generalization of full component activation patching, which only replaces the values of a subspace U (while leaving the projection on its orthogonal complement U^\perp unchanged). This was proposed in Geiger et al. (2023b), and closely related variants appear in Turner et al. (2023); Nanda et al. (2023); Lieberum et al. (2023).

For the purposes of exposition, we now restrict our discussion to activation patching of a 1-dimensional subspace (i.e., a *direction*) spanned by a unit vector \mathbf{v} (i.e., $\|\mathbf{v}\|_2 = 1$). We remark that the illusion also applies to higher-dimensional subspaces (see Appendix A.1 for theoretical details; later on, in Appendix B.6, we also show this empirically for the IOI task). If $\mathbf{act}_A, \mathbf{act}_B \in \mathbb{R}^d$ are the activations of a model component \mathcal{C} on examples A, B and $p_A = \mathbf{v}^\top \mathbf{act}_A, p_B = \mathbf{v}^\top \mathbf{act}_B$ are their projections on \mathbf{v} , patching from A into B along \mathbf{v} results in the patched activation

$$\mathbf{act}_B^{\text{patched}} = \mathbf{act}_B + (p_A - p_B)\mathbf{v}. \quad (1)$$

For a concrete scenario motivating such a patch, consider a discrete binary feature used by the model to perform a task, and prompts A, B which only differ in the value of this feature. A 1-dimensional subspace can easily encode such a feature (and indeed we explore an example of this in great detail in Sections 4 and 5).

3.2 Intuition for the Illusion

What would make activation patching a good attribution method? Intuitively, an equivalence is needed: an activation patch should work *if and only if* the component/subspace being patched is indeed a *faithful to the model’s computation* representation of the concept we seek to localize. Revisiting Equation 1 with this in mind, it is quite plausible that, if \mathbf{v} indeed encodes a binary feature relevant to the task, the patch will essentially overwrite the feature with its value on A , and this would lead to the expected downstream effect on model predictions³.

Going in the other direction of the equivalence, when will the update in Equation 1 change the model’s output in the intended way? Intuitively, two properties are necessary:

- **correlation with the concept:** \mathbf{v} must be activated differently by the two prompts. Otherwise, $p_A \approx p_B$, and the patch has no effect;
- **potential for changing model outputs:** \mathbf{v} must be ‘causally connected’ to the model’s outputs; in other words, it should be the case that changing the activation along \mathbf{v} can at least in some cases lead to a change in the next-token probabilities output by the model. Otherwise, if, for instance, \mathbf{v} is in the nullspace of all downstream model components, changing the activation’s projection along \mathbf{v} alone won’t have any effect on the model’s predictions.

For example, if the component \mathcal{C} we are patching is the post-nonlinearity activation of an MLP layer, the only way this activation affects the model’s output is through matrix multiplication with a down-projection W_{out} . So, if $\mathbf{v} \in \ker W_{out}$, we will have

$$W_{out}\mathbf{act}_B^{\text{patched}} = W_{out}\mathbf{act}_B + (p_A - p_B)W_{out}\mathbf{v} = W_{out}\mathbf{act}_B.$$

In other words, the activation patch leads to the exact same output of the MLP layer as when running the model on B without an intervention. So, the patch will leave model predictions unchanged.

³It is in principle possible that, even if the value of the feature is overwritten, this has no effect on model predictions. For example, it is possible that \mathbf{v} is not the only location in the model’s computation where this feature is represented; or, it may be that there are backup components that are normally inactive on the task, but activate when the value of the subspace \mathbf{v} is changed, as explored in McGrath et al. (2023). Such scenarios are beyond the scope of this simplified discussion.

The crux of the illusion is that \mathbf{v} may obtain each of the two properties from two ‘unrelated’ directions in activation space (as shown in Figure 1) which ‘happen to be there’ as a side effect of linear algebra. Specifically, we can form

$$\mathbf{v}_{\text{illusory}} = \frac{1}{\sqrt{2}} (\mathbf{v}_{\text{disconnected}} + \mathbf{v}_{\text{dormant}}), \quad (2)$$

for orthogonal unit vectors $\mathbf{v}_{\text{disconnected}}^\top \mathbf{v}_{\text{dormant}} = 0$ such that

- $\mathbf{v}_{\text{disconnected}}$ is a **causally disconnected direction** in activation space: it distinguishes between the two prompts, but is in the nullspace of all downstream model components (e.g., a vector in $\ker W_{\text{out}}$ for an MLP layer with down-projection W_{out});
- $\mathbf{v}_{\text{dormant}}$ is a **dormant direction** in activation space: it can *in principle* steer the model’s output in the intended way, but is not activated differently by the two prompts (in other words, $\mathbf{v}_{\text{dormant}}^\top \mathbf{act}_A \approx \mathbf{v}_{\text{dormant}}^\top \mathbf{act}_B$).

To illustrate this algebraically, consider what happens when we patch along $\mathbf{v}_{\text{illusory}}$. We have

$$\begin{aligned} p_A &= \mathbf{v}_{\text{illusory}}^\top \mathbf{act}_A = \frac{1}{\sqrt{2}} (\mathbf{v}_{\text{disconnected}}^\top \mathbf{act}_A + \mathbf{v}_{\text{dormant}}^\top \mathbf{act}_A) \\ p_B &= \mathbf{v}_{\text{illusory}}^\top \mathbf{act}_B = \frac{1}{\sqrt{2}} (\mathbf{v}_{\text{disconnected}}^\top \mathbf{act}_B + \mathbf{v}_{\text{dormant}}^\top \mathbf{act}_B) \end{aligned}$$

By assumption, $\mathbf{v}_{\text{dormant}}^\top \mathbf{act}_B = \mathbf{v}_{\text{dormant}}^\top \mathbf{act}_A$. Thus,

$$p_A - p_B = \frac{1}{\sqrt{2}} (\mathbf{v}_{\text{disconnected}}^\top \mathbf{act}_A - \mathbf{v}_{\text{disconnected}}^\top \mathbf{act}_B)$$

so the patched activation is

$$\mathbf{act}_B^{\text{patched}} = \mathbf{act}_B + \frac{1}{\sqrt{2}} (\mathbf{v}_{\text{disconnected}}^\top \mathbf{act}_A - \mathbf{v}_{\text{disconnected}}^\top \mathbf{act}_B) \mathbf{v}_{\text{illusory}}.$$

If for example $\mathbf{v}_{\text{illusory}}$ is in the space of post-nonlinearity activations of an MLP layer with down-projection matrix W_{out} , and $\mathbf{v}_{\text{disconnected}} \in \ker W_{\text{out}}$, the new output of the MLP layer after the patch will be

$$\begin{aligned} W_{\text{out}} \mathbf{act}_B^{\text{patched}} &= W_{\text{out}} \mathbf{act}_B + \frac{1}{\sqrt{2}} (\mathbf{v}_{\text{disconnected}}^\top \mathbf{act}_A - \mathbf{v}_{\text{disconnected}}^\top \mathbf{act}_B) W_{\text{out}} \mathbf{v}_{\text{illusory}} \\ &= W_{\text{out}} \mathbf{act}_B + \frac{1}{2} (\mathbf{v}_{\text{disconnected}}^\top \mathbf{act}_A - \mathbf{v}_{\text{disconnected}}^\top \mathbf{act}_B) W_{\text{out}} \mathbf{v}_{\text{dormant}} \end{aligned} \quad (3)$$

where we used that $W_{\text{out}} \mathbf{v}_{\text{disconnected}} = 0$. From this equation, we see that, by patching along the sum of a disconnected and dormant direction, the variation in activation projections on the disconnected part (which we assume is significant) ‘activates’ the dormant part: we get a new contribution to the MLP’s output (along $W_{\text{out}} \mathbf{v}_{\text{dormant}}$) which can then possibly influence model outputs. This contribution would not exist if we patched only along $\mathbf{v}_{\text{disconnected}}$ (because it would be nullified by W_{out}) or $\mathbf{v}_{\text{dormant}}$ (because then we would have $p_A \approx p_B$).

We make the concepts of causally disconnected and dormant subspaces formal in Subsection 3.5. We also remark that, under the assumptions of the above discussion, the optimal illusory patch will provably combine the disconnected and dormant directions with equal weight $\frac{1}{\sqrt{2}}$ as in Equation 2; the proof is given in Appendix A.2.

3.3 The Illusion in a Toy Model

With these concepts in mind, we can construct a distilled example of the illusion in a toy (linear) neural network. Specifically, consider a network \mathcal{M} that takes in an input $x \in \mathbb{R}$, computes a three-dimensional hidden representation $\mathbf{h} = x\mathbf{w}_1$, and then a real-valued output $y = \mathbf{w}_2^T \mathbf{h}$. Define the weights to be

$$\mathbf{w}_1 = (1, 0, 1), \quad \text{and} \quad \mathbf{w}_2 = (0, 2, 1)$$

and observe that $\mathcal{M}(x) = x$, i.e. the network computes the identity function:

$$x \mapsto \mathbf{h} = (x, 0, x) \mapsto y = 0 \times x + 2 \times 0 + 1 \times x = x.$$

This network is illustrated in Figure 2. We can analyze the 1-dimensional subspaces (directions) spanned by each of the three hidden activations:

- the h_1 direction is causally disconnected: setting it to any value has no effect on the output;
- the h_2 direction is dormant: it is constant (always 0) on the data, but setting it to some other value will affect the model’s output;
- the h_3 direction mediates the signal through the network: the input x is copied to it, and is in turn copied to the output⁴.

As expected, patching along the direction h_3 overwrites the value of the x feature (which in this example is identical to the input). That is, patching along h_3 from x' into x makes the network output x' instead of x .

However, patching along the sum of the causally disconnected direction h_1 and the dormant direction h_2 represented by the unit vector $\mathbf{v}_{\text{illusory}} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0\right)$ has the same effect: using Equation 1, patching from x' into x along $\mathbf{v}_{\text{illusory}}$ results in the hidden activation

$$\mathbf{h}^{\text{patched}} = \left(\frac{x+x'}{2}, \frac{x'-x}{2}, x \right)^T$$

which when multiplied with \mathbf{w}_2 gives the final output $2 \times \frac{x'-x}{2} + 1 \times x = x'$.

3.4 Detecting the illusion in practice

How can we tell if this kind of phenomenon occurs for a given subspace activation patch? Given a subspace spanned by a unit vector \mathbf{v} , suppose that activation patching along this subspace has an effect on model outputs consistent with changing the property that varies between the examples being patched. We can attempt to decompose it orthogonally into a causally disconnected part and a dormant part, and argue that each of these parts has the properties described in the above sections.

⁴An important note on this particular example is that the distinction between causally disconnected, dormant and faithful to the computation directions is artificial, and here it is only used for exposition. In particular, we show in Appendix A.3 that re-parametrizing the hidden layer of the network via a rotation makes $\mathbf{v}_{\text{illusory}}$ take the role of the faithful direction \mathbf{e}_3 , and the two other (rotated) basis vectors become a disconnected/dormant pair. By contrast, when we exhibit the illusion in real-world scenarios, a reparametrization of this kind would need to combine activations between different model components, such as MLP layers and residual stream activations. We return to this point in Section 8.

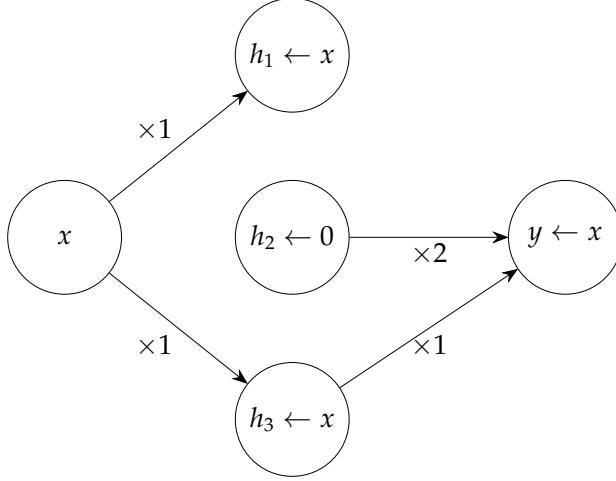


Figure 2: A network \mathcal{M} illustrating the illusion. The network computes the identity function: $\mathcal{M}(x) = x$. The activation of the input, output and each hidden neuron for a generic input x are shown in the circles, with arrows indicating the weight of the connections (no arrow means a weight of 0). The hidden unit h_3 stores the value of the input and passes this to the output, while the unit h_2 is dormant and h_1 is disconnected from the output. However, activation patching the 1-dimensional linear subspace spanned by the sum of the h_1 and h_2 basis vectors (defined by the unit vector $\mathbf{v} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$) has the same effect on model behavior as patching just the unit h_3 .

Specifically, when \mathbf{v} is in the post-GELU activations of an MLP layer in a transformer with down-projection W_{out} (almost all examples in this paper are of this form), it is clear that the orthogonal projection of \mathbf{v} on $\ker W_{out}$ is causally disconnected. This suggests writing $\mathbf{v} = \mathbf{v}_{\text{null}} + \mathbf{v}_{\text{row}}$ where $\mathbf{v}_{\text{null}} \in \ker W_{out}$ is the orthogonal projection on $\ker W_{out}$, and \mathbf{v}_{row} is the remainder, which is in $(\ker W_{out})^\perp$, the rowspace of W_{out} . Using this decomposition, we can perform several experiments:

- compare the strength of the patch to patching along the subspace spanned by \mathbf{v}_{row} alone, obtained by removing the causally disconnected part of \mathbf{v} . If \mathbf{v}_{row} is indeed dormant as we hope to show, this patch should have no effect on model outputs; in reality, \mathbf{v}_{row} may only be approximately dormant, so the patch may have a small effect. Conversely, if this patch has an effect similar to the original patch along \mathbf{v} , this is evidence against the illusion;
- check how dormant \mathbf{v}_{row} is compared to \mathbf{v}_{null} by comparing the spread of projections of the examples on both directions.

We use these experiments, as well as others, throughout the paper in order to rule out or confirm the illusion.

3.5 Formalization of Causally Disconnected and Dormant Subspaces

For completeness, in this subsection we give a (somewhat) formal treatment of the intuitive ideas introduced in the previous subsection. Readers may also want to consult Appendix A.1 for background on patching higher-dimensional subspaces, which is used to define these concepts.

Let $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{O}$ be a machine learning model that on input $x \in \mathcal{X}$ outputs a vector $y \in \mathcal{O}$ of probabilities over a set of output classes. Let \mathcal{D} be a distribution over \mathcal{X} , and \mathcal{C} be a component of

\mathcal{M} , such that for $x \sim \mathcal{D}$ the hidden activation of \mathcal{C} is a vector $c_x \in \mathbb{R}^d$. For a subspace $U_{\mathcal{C}} \subset \mathbb{R}^d$, we let u_x be the orthogonal projection of c_x onto $U_{\mathcal{C}}$. Finally, let $\mathcal{M}_{U_{\mathcal{C}} \leftarrow u_y}(x)$ be the result of running \mathcal{M} with the input x and setting the subspace $U_{\mathcal{C}}$ patched to u_y .

We say $U_{\mathcal{C}}$ is *causally disconnected* if $\mathcal{M}_{U_{\mathcal{C}} \leftarrow u'}(x) = \mathcal{M}(x)$ for all $u' \in U_{\mathcal{C}}$. In other words, intervening on the model by setting the orthogonal projection of \mathcal{C} 's activation on $U_{\mathcal{C}}$ to any other value does not change the model's outputs. For a concrete example of a causally disconnected subspace, consider an MLP layer in a transformer model with an output projection matrix W_{out} ; then, $\ker W_{out}$ is a causally disconnected subspace of the hidden (post-nonlinearity) activations of the MLP layer.

We say $U_{\mathcal{C}}$ is *dormant* if $\mathcal{M}_{U_{\mathcal{C}} \leftarrow u_y}(x) \approx \mathcal{M}(x)$ with high probability over $x, y \sim \mathcal{D}$, but there exists $u' \in \mathbb{R}^d$ such that $\mathcal{M}_{U_{\mathcal{C}} \leftarrow u'}(x)$ is substantially different from $\mathcal{M}(x)$ (e.g., significantly changes the model's confidence on the task's answer). In other words, a dormant subspace is approximately causally disconnected when we patch its value using activations realized under the distribution \mathcal{D} , but can have substantial causal effect if set to other values.

4 The Illusion in the Indirect Object Identification Task

4.1 Preliminaries

In Wang et al. (2023), the authors analyze how the decoder-only transformer language model GPT-2 Small (Radford et al., 2019) performs the *indirect object identification* task. In this task, the model is required to complete sentences of the form ‘When Mary and John went to the store, John gave a bottle of milk to’ (with the intended completion in this case being ‘Mary’). We refer to the repeated name (John) as **S** (the subject) and the non-repeated name (Mary) as **IO** (the indirect object). For each choice of the **IO** and **S** names, there are two patterns the sentence can have: one where the **IO** name comes first (we call these ‘ABB examples’), and one where it comes second (we call these ‘BAB examples’). Additional details on the data distribution, model and task performance are given in Appendix B.1.

Wang et al. (2023) suggest the model uses the algorithm ‘Find the two names in the sentence, detect the repeated name, and predict the non-repeated name’ to do this task. In particular, they find a set of four heads in layers 7 and 8 – the **S-Inhibition heads** – that output the signal responsible for *not* predicting the repeated name. The dominant part of this signal is of the form ‘Don’t attend to the name in first/second position in the first sentence’ depending on where the **S** name appears (see Appendix A in Wang et al. (2023) for details). In other words, this signal detects whether the example is an ABB or BAB example. This signal is added to the residual stream⁵ at the last token position, and is then picked up by another class of heads in layers 9, 10 and 11 – the **Name Mover heads** – which incorporate it in their queries to shift attention to the **IO** name and copy it to the last token position, so that it can be predicted (Figure 3).

4.2 Finding Subspaces Mediating Name Position Information

How, precisely, is the positional signal communicated? In particular, ‘don’t attend to the first/second name’ is plausibly a binary feature represented by a 1-dimensional subspace. In this subsection, we present methods to look for such a subspace.

⁵We follow the conventions of Elhage et al. (2021) when describing internals of transformer models. The residual stream at layer k is the sum of the output of all layers up to $k - 1$, and is the input into layer k .

Gradient of name mover attention scores.

As shown in Wang et al. (2023), the three name mover heads identified therein will attend to one of the names, and the model will predict whichever name is attended to. The position feature matters mechanistically by determining whether they attend to **IO** over **S**. This motivates us to consider the gradient \mathbf{v}_{grad} of the difference of attention scores of these heads on the **S** and **IO** names with respect to the residual stream at the last token, right after layer 8. We choose this layer because it right after the **S**-Inhibition heads (in layers 7 and 8) and before the name mover heads (in layers 9 and 10); see Figure 3. This gradient is the direction in the space of residual stream activations at this location that maximally shifts attention between the two names (per unit ℓ_2 norm), so we expect it to be a strong mediator of the position signal. Implementation details are given in Appendix B.2.

Importantly, the transformation from residual stream activations to attention scores is an approximately linear map: it consists of layer normalization followed by matrix multiplication. Layer normalization is a linear operation modulo the scaling step, and empirically, the scales of different examples in a trained model at inference time are tightly concentrated (see also ‘Handling Layer Normalization’ in Elhage et al. (2021)). This justifies the use of the gradient – which is in general only locally meaningful – as a direction in the residual stream globally meaningful for the attention scores of the name mover heads.

Distributed alignment search. We can also directly optimize for a direction that mediates the position signal. This is the approach taken by DAS (Geiger et al., 2023b). In our context, DAS optimizes for an activation subspace which, when activation patched from prompt *B* into prompt *A*, makes the model behave as if the relative position of the **IO** and **S** names in the sentence is as in prompt *B*. Specifically, if we patch between examples where the positions of the two names are the same, we optimize for a patch that *maximizes* the difference in predicted logits for the **IO** and **S** names. Conversely, if we patch between examples where the positions of the two names are switched, we optimize to *minimize* this difference. This approach is based purely on the model’s predictions, and does not make any assumptions about its internal computations.

We let \mathbf{v}_{MLP} and $\mathbf{v}_{\text{resid}}$ be 1-dimensional subspaces found by DAS in the layer 8 MLP activations and layer 8 residual stream output at the last token, respectively (see Figure 3). Both of these locations are between the **S**-Inhibition and Name Mover heads; however, Wang et al. (2023) did not

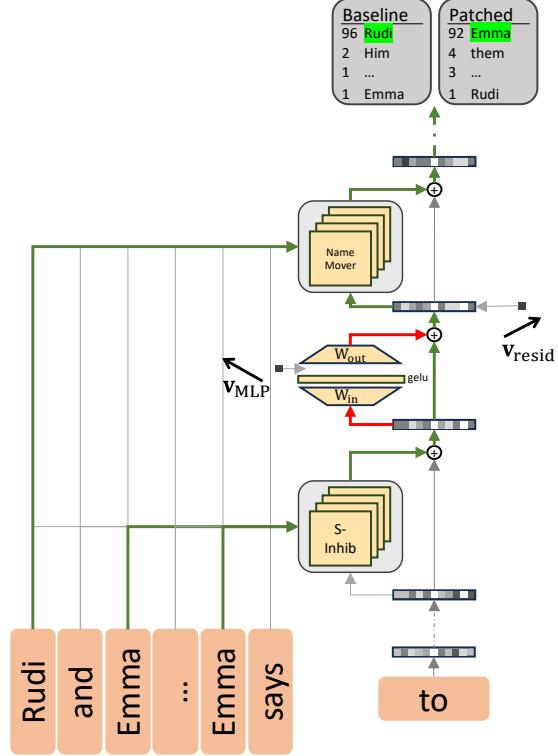


Figure 3: Schematic of the IOI circuit and locations of key interventions. As argued in Wang et al. (2023), GPT2-Small predicts the correct name by **S**-inhibition heads writing positional information to the residual stream, which is used by the name movers to copy the non-duplicated name (green arrows). Location of subspace interventions \mathbf{v}_{MLP} (analyzed in Subsection 4.4) and $\mathbf{v}_{\text{resid}}$ (analyzed in Section 5) are marked. Patching the illusory subspace \mathbf{v}_{MLP} adds a new path (red) along the established one that is used to flip positional information when patched.

find any significant contribution from the MLP layer, making it a potential location for our illusion. Implementation details are given in Appendix B.3.

4.3 Measuring Patching Success via the Logit Difference Metric

In our experiments, we perform all patches between examples that only differ in the variable we want to localize in the model, i.e. the position of the **S** and **IO** names in the first sentence. In other words, we patch from an example of the form ‘Then, Mary and John went to the store. John gave a book to’ (an ABB example) into the corresponding example ‘Then, John and Mary went to the store. John gave a book to’ (a BAB example), and vice-versa. Our activation patches have the goal of making the model output the **S** name instead of the **IO** name.

Accordingly, we use the *logit difference* between the logits assigned to the **IO** and **S** names as our main measure of how well a patch performs. We note that the logit difference is a meaningful quantifier of the model’s confidence for one name over the other (it is equal to the log-odds between the two names assigned by the model), and has been extensively used in the original IOI circuit work Wang et al. (2023) to measure success on the IOI task.

Given a prompt x from the IOI distribution, let $\text{logit}_{\text{IO}}(x)$, $\text{logit}_{\text{S}}(x)$ denote the last-token logits output by the model on input x , for the **IO** and **S** names in the prompt x respectively (note that in our IOI distribution, all names are single tokens in the vocabulary of the model). The logit difference

$$\text{logitdiff}(x) := \text{logit}_{\text{IO}}(x) - \text{logit}_{\text{S}}(x)$$

when x is sampled from the IOI distribution is > 0 for almost all examples (99 + %), and is on average ≈ 3.5 (for this average value, the probability ratio in favor of the **IO** name is $e^{3.5} \approx 33$).

Similarly, for an activation patching intervention ι , let $\text{logit}_{\text{IO}}^{\iota(x \leftarrow x')}(x)$, $\text{logit}_{\text{S}}^{\iota(x \leftarrow x')}(x)$ denote the last-token logits output by the model when run on input x but patching from x' using ι . The logit difference after intervening via ι is thus

$$\text{logitdiff}_{\iota(x \leftarrow x')}(x) := \text{logit}_{\text{IO}}^{\iota(x \leftarrow x')}(x) - \text{logit}_{\text{S}}^{\iota(x \leftarrow x')}(x)$$

Our main metric is the average **fractional logit difference decrease (FLDD)** due to the intervention ι , where

$$\text{FLDD}_{\iota(x \leftarrow x')}(x) := \frac{\text{logitdiff}(x) - \text{logitdiff}_{\iota(x \leftarrow x')}(x)}{\text{logitdiff}(x)} = 1 - \frac{\text{logitdiff}_{\iota(x \leftarrow x')}(x)}{\text{logitdiff}(x)} \quad (4)$$

The average FLDD is 0 when the patch does not, on average, change the model’s log-odds. The more positive FLDD is, the more successful the patch, with values above 100% indicating that the patch more often than not makes the model prefer the **S** name over the **IO** name. Finally, an average FLDD below 0% means that the patch on average helps the model do the task (and thus the patch has failed).

We also measure the **interchange accuracy** of the intervention: the fraction of patches for which the model predicts the **S** (i.e., wrong) name for the patched run. This is a ‘hard’ 0-1 counterpart to the FLDD metric.

Why prefer the FLDD metric to interchange accuracy? We argue that our main metric, which is based on logit difference (Equation 4), is a better reflection of the success of a patch than the accuracy-based interchange accuracy. Specifically, there are practical cases (e.g. the results in

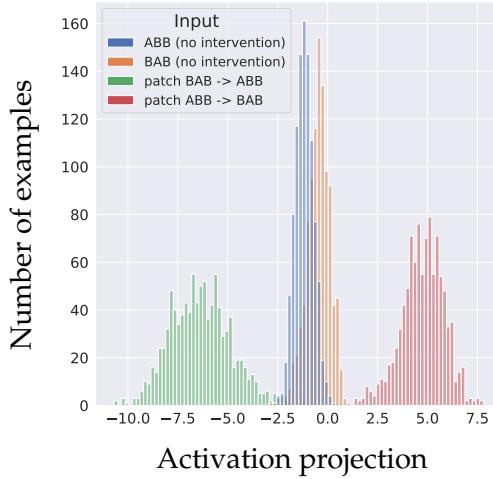


Figure 4: Projections of the output of the MLP layer on the gradient direction \mathbf{v}_{grad} before (blue/orange) and after (green/red) the activation patch along \mathbf{v}_{MLP} . In the legend, ‘ABB’ denotes examples where the **IO** name comes before the **S** name, and ‘BAB’ the other kind of examples.

While before the patch the contribution of the MLP layer to the causally relevant direction \mathbf{v}_{grad} distinguishes between values of the **IO** position in the prompt, after the patch there is a strong distinction (in the opposite direction). This shows that the patch activates a potential mediator of this feature that is normally dormant, taking model activations off-distribution.

Patching subspace	FLDD	Interchange accuracy
full MLP	-8%	0.0%
\mathbf{v}_{MLP}	46.7%	4.2%
\mathbf{v}_{MLP} rowspace	13.5%	0.2%
\mathbf{v}_{MLP} nullspace	0%	0.0%
full residual stream	123.6%	54.8%
$\mathbf{v}_{\text{resid}}$	140.7%	74.8%
$\mathbf{v}_{\text{resid}}$ rowspace	127.5%	63.1%
$\mathbf{v}_{\text{resid}}$ nullspace	13.9%	0.4%
\mathbf{v}_{grad}	111.5%	45.1%
\mathbf{v}_{grad} rowspace	106.47%	40.6%
\mathbf{v}_{grad} nullspace	2.2%	0.0%

Table 1: Effects of activation patching of full components and 1-dimensional subspaces on the IOI task: fractional logit difference decrease (FLDD, higher means more successful patch; 0% means no change) and interchange accuracy (fraction of predictions flipped; higher means more successful patch).

The first 5 interventions are described in more detail in Section 4, and the next 6 in Section 5.

An FLDD metric of $> 100\%$ indicates that the patch is more successful than not on average; however, an FLDD of $\approx 50\%$ is still significant, even if the associated interchange accuracy may be $\approx 0\%$. See Subsection 4.3 for more on interpreting the FLDD metric.

Subsection 4.4) in which an intervention consistently achieves $\text{FLDD} \approx 50\%$, even though the interchange accuracy is $\approx 0\%$. In practice, circuits often have multiple components contributing to the same signal (including the IOI circuit found in Wang et al. (2023)). So a single non-residual-stream component consistently responsible for shifting 50% of the model’s log-odds towards another prediction is significant (even more so for a low-dimensional subspace of the component). Indeed, even if this component’s contribution alone is insufficient to cause the predicted token to change, three such components would robustly change the prediction.

4.4 Results: Demonstrating the Illusion for the \mathbf{v}_{MLP} Direction

We now show that patching the \mathbf{v}_{MLP} direction exhibits the illusion from Section 3. By contrast, we revisit \mathbf{v}_{grad} and $\mathbf{v}_{\text{resid}}$ in Section 5, where we show that both are representations of the name position information that are highly faithful to the model’s computation.

Methodology and interventions considered To contextualize the effect of the \mathbf{v}_{MLP} patch, we compare it to several additional subspace- and component-level activation patching interventions:

- **full MLP:** patching the full value of the hidden activation of the 8-th MLP layer at the last token.
- **\mathbf{v}_{MLP} :** patching along the 1-dimensional subspace spanned by the direction \mathbf{v}_{MLP} found in Subsection 4.2.
- **$\mathbf{v}_{\text{MLP}} \text{ nullspace}$:** patching along the 1-dimensional subspace spanned by the causally disconnected component of \mathbf{v}_{MLP} . This is the orthogonal projection $\mathbf{v}_{\text{MLP}}^{\text{nullspace}}$ of \mathbf{v}_{MLP} on the nullspace $\ker W_{\text{out}}$ of the down-projection W_{out} of the MLP layer. Note that $W_{\text{out}} \in \mathbb{R}^{768 \times 3072}$, so its kernel occupies at least 2304 dimensions, or 3/4 of the total dimension of the space of MLP activations.
- **$\mathbf{v}_{\text{MLP}} \text{ rowspace}$:** patching along the 1-dimensional subspace spanned by causally relevant component of \mathbf{v}_{MLP} . This is the orthogonal projection $\mathbf{v}_{\text{MLP}}^{\text{rowspace}}$ of \mathbf{v}_{MLP} on the rowspace of W_{out} . Note that we have the orthogonal decomposition

$$\mathbf{v}_{\text{MLP}} = \mathbf{v}_{\text{MLP}}^{\text{nullspace}} + \mathbf{v}_{\text{MLP}}^{\text{rowspace}}.$$

- **full residual stream:** patching the entire activation of the residual stream at the last token after layer 8 of the model. This is indicated as the location of $\mathbf{v}_{\text{resid}}$ in Figure 3.

Results. Metrics are shown in Table 1. In particular, we confirm the mechanics of the illusion are at play through the following observations.

The causally disconnected component of \mathbf{v}_{MLP} drives the effect. While patching the \mathbf{v}_{MLP} direction has a significant effect on the FLDD metric (46.7%), this effect is greatly diminished when we remove the component of \mathbf{v}_{MLP} in $\ker W_{\text{out}}$ whose activations are (provably) causally disconnected from model predictions (13.5%), or when we patch the entire MLP activation (−8%, actually increasing confidence). By contrast, performing analogous ablations on $\mathbf{v}_{\text{resid}}$ leads to roughly the same numbers for the three analogous interventions (140.7%/127.5%/123.6%; we refer the reader to Section 5 for details on the $\mathbf{v}_{\text{resid}}$ experiments).

Patching \mathbf{v}_{MLP} activates a dormant pathway through the MLP. To corroborate these findings, in Figure 4, we plot the projection of the MLP layer’s contribution to the residual stream on the gradient direction \mathbf{v}_{grad} before and after patching, in order to see how it contributes to the attention of name mover heads. We observe that in the absence of intervention, the MLP output is weakly sensitive to the name position information, whereas after the patch this changes significantly.

Further validations of the illusion. We observe that the disconnected-dormant decomposition from the illusion approximately holds: the causally disconnected component of \mathbf{v}_{MLP} (the one in $\ker W_{\text{out}}$) is significantly more discriminating between ABB and BAB examples than the component in $(\ker W_{\text{out}})^{\top}$, which is the one driving the causal effect (Figure 5); in this sense, the causally relevant component is ‘dormant’ relative to the causally disconnected one⁶.

⁶The projection of \mathbf{v}_{MLP} onto $\ker W_{\text{out}}$ is substantial: it has norm ≈ 0.65 , and the orthogonal projection onto $(\ker W_{\text{out}})^{\top}$ has norm ≈ 0.75 (as predicted by our model, the two components are approximately equal in norm; see Appendix A.2).

While the contribution of the \mathbf{v}_{MLP} patch to the FLDD metric may appear relatively small, and the interchange accuracy of this intervention is very low, in Appendix B.4 we argue that this is significant for a single component.

A potential concern when evaluating these results is overfitting by DAS. In our experiments, we always evaluate trained subspaces on a held-out test dataset which uses different names, objects, places and templates for the sentences; this makes sure that we learn a general (relative to our IOI distribution) subspace representing position information and not a subspace that only works for particular names or other details of the sentence. We investigated overfitting in DAS further in Appendix B.7, and found that when DAS is trained on a dataset with a small number of names, overfitting is a real concern. However, the extent of overfitting is not such that DAS works in layers of the model where a generalizing DAS solution can also be found.

Another potential concern is that the model could be somehow representing the position information in the MLP layer in a higher-dimensional subspace, and that our 1-dimensional intervention is perhaps not fit to illuminate the properties of that larger representation. In Appendix B.6, we show that the illusion occurs when patching 100-dimensional subspaces as well, and the quantitative effect of the illusion is just a little stronger than that for 1-dimensional subspaces (as measured by the FLDD metric).

Finally, in Appendix B.5, we show that we can find a direction within the post-gelu activations that has an even stronger effect on the model’s behavior, *even when* we replace the MLP weights with random matrices.

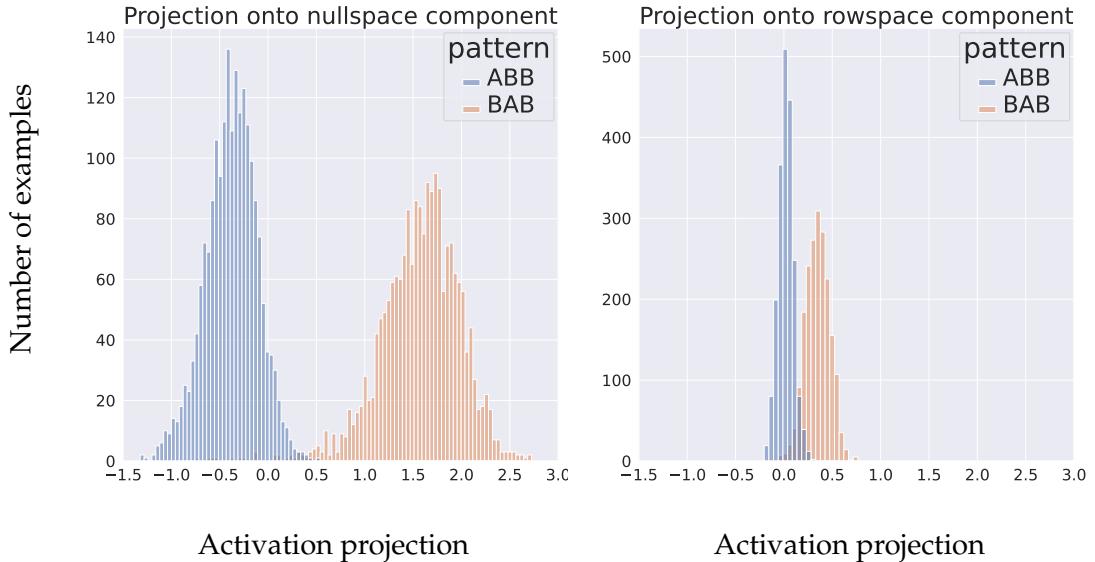


Figure 5: Projections of dataset examples onto the two (normalized to unit ℓ_2 norm) components of the illusory patching direction found in MLP8: the nullspace (irrelevant) component (left), and the rowspace (dormant) component (right).

5 Finding and Validating a Faithful Direction Mediating Name Position in the IOI Task

As a counterpoint to the illusion, in this section we demonstrate a success case for subspace activation patching, as well as for DAS as a method for finding meaningful subspaces, by revisiting the directions \mathbf{v}_{grad} and $\mathbf{v}_{\text{resid}}$ defined in Subsection 4.2, and arguing they are faithful to the model’s computation to a high degree.

Specifically, we subject these directions to the same tests we used for the illusory direction \mathbf{v}_{MLP} , and arrive at significantly different results. Through these and additional validations, we demonstrate that these directions possess the necessary and sufficient properties of a successful activation patch – being both correlated with input variation and causal for the targeted behavior – in a way that does not rely on a large causally disconnected component for the effect.

5.1 Ruling Out the Illusion

Intuitively, the main property of $\mathbf{v}_{\text{resid}}$ we want to establish in order to rule out the illusion is that it is simultaneously (1) strongly discriminating between ABB and BAB examples (i.e., projections of activations on $\mathbf{v}_{\text{resid}}$ separate these two classes well), and (2) is highly aligned with the direction \mathbf{v}_{grad} that downstream model components read this information from in order to put attention on the **IO** name and not the **S** name. To this end, we define a notion of causally disconnected component for $\mathbf{v}_{\text{resid}}$, and we show that removing it does not diminish the effect of the patch; we further show that $\mathbf{v}_{\text{resid}}$ and \mathbf{v}_{grad} are quite similar, and that \mathbf{v}_{grad} is also strongly activated by position information.

What is the causally (dis)connected subspace of the residual stream? While for an MLP layer it is clear that $\ker W_{\text{out}}$ is the subspace of post-GELU activations which is causally disconnected from model outputs, the residual stream after layer 8 has no subspace which is simultaneously in the kernel of all downstream model components, or even of all the query matrices of downstream attention heads (we checked this empirically).

To overcome this, recall from Section 4 that Wang et al. (2023) argued that the three Name Mover heads in layers 9 and 10 are mostly responsible for the IOI task specifically. Let $W_Q^{\text{NM}} \in \mathbb{R}^{(3 \times 64) \times 768} = \mathbb{R}^{192 \times 768}$ be the stacked query matrices of the three name mover heads (which are full-rank). We use the 192-dimensional subspace $(\ker W_Q^{\text{NM}})^\top$ as a proxy for the causally relevant subspace⁷ of the residual stream at the last token position after layer 8.

To further narrow down the precise subspace read by the Name Mover heads, we also compare $\mathbf{v}_{\text{resid}}$ with the gradient \mathbf{v}_{grad} , which is the direction that the Name Mover’s attention on the **IO** vs. **S** name is most sensitive to.

Results. In Table 1, we report the fractional logit difference decrease (FLDD, recall Subsection 4.3) and interchange accuracy when patching $\mathbf{v}_{\text{resid}}$ and \mathbf{v}_{grad} , as well as their components along

⁷Note that, while technically all attention heads in layers 9, 10 and 11 read information from the residual stream after layer 8, using their collective query matrices instead of just the name movers would lead to a vacuous concept of a causally relevant subspace, because their collective query matrices’ rowspaces span the entire residual stream. As a rough baseline, a random isotropic unit vector would have on average $\sqrt{\frac{192}{768}} = \frac{1}{2}$ of its ℓ_2 -norm in $(\ker W_Q^{\text{NM}})^\top$. We also note that this is on par with the decomposition of \mathbf{v}_{MLP} we considered in Section 4, where $\ker W_{\text{out}}$ occupied 3/4 of the dimension of the full space of activations.

$\ker W_Q^{NM}$ (denoted ‘nullspace’) and its orthogonal complement $(\ker W_Q^{NM})^\top$ (denoted ‘rowspace’). We observe that the non-nullspace metrics are broadly similar⁸; in particular, removing the causally disconnected component of $\mathbf{v}_{\text{resid}}$ does not significantly diminish the effect of the patch in terms of the logit difference metrics (as it does for \mathbf{v}_{MLP}).

Furthermore, we find that the cosine similarity between $\mathbf{v}_{\text{resid}}$ and \mathbf{v}_{grad} is ≈ 0.78 , which is significant (the baseline for random vectors in the residual stream is on the order of $\frac{1}{\sqrt{768}} \approx 0.03$).

Both $\mathbf{v}_{\text{resid}}$ and \mathbf{v}_{grad} have a significant fraction of their norm in the $(\ker W_Q^{NM})^\top$ subspace (91% and 98%, respectively). These results suggest that this $\mathbf{v}_{\text{resid}}$ and \mathbf{v}_{grad} are highly similar directions, and that they’re both strongly causally connected to the model’s output.

In Figure 6, we also find that both directions are strongly discriminating between ABB and BAB examples.

Discussion. A key observation about the residual stream at the last token is that it is a full bottleneck for the model’s computation over the last token position: all updates to that position are added to it. This provides another viewpoint on why the successful patches we find don’t rely on a dormant subspace: there can be no earlier model component that activates the directions we find in a way that skips over the patch via a residual connection (unlike for \mathbf{v}_{MLP}). Indeed, in Figure 20 in Appendix C we show that the $\mathbf{v}_{\text{resid}}$ direction gets written to by the S-Inhibition heads.

5.2 Additional Validations

In Appendix C, we further validate these directions’ faithfulness to the computation of the IOI circuit from Wang et al. (2020) by finding the model components that write to them and studying how they generalize on the pre-training distribution (OpenWebText); representative samples annotated with attention scores are shown in Figures 23, 21, 22 in Appendix C.

6 Factual Recall

This section has two major goals. One is to show that the interpretability illusion can also be exhibited for the factual recall capability of language models, a much broader setting than the IOI task. The other is to exhibit in practice an approximate equivalence between two seemingly different interventions: rank-1 weight editing and interventions on 1-dimensional subspaces of activations. We do this in several complementary ways:

1. we show that DAS (Geiger et al., 2023b) finds illusory 1-dimensional subspace patches that change factual recall (e.g., to make a model complete ‘The Eiffel Tower is in’ with ‘Rome’ instead of ‘Paris’). The patches found strongly update the model’s confidence towards the false completion, but the effect disappears when the causally disconnected component of the subspace is removed, or when the entire MLP activation containing the subspace is patched instead.
2. we show that for a wide range of layers in the middle of the model (GPT2-XL Radford et al. (2019)), rank-1 fact editing using the ROME method (Meng et al., 2022a) is approximately

⁸We observe that the $\mathbf{v}_{\text{resid}}$ patch is more successful than the \mathbf{v}_{grad} patch; we conjecture that this is due to $\mathbf{v}_{\text{resid}}$ being able to contribute to all downstream attention heads, not just the three name-mover heads. In particular, the original IOI paper Wang et al. (2023) found that there is another class of heads, Backup Name Movers, which act somewhat like Name Mover heads.

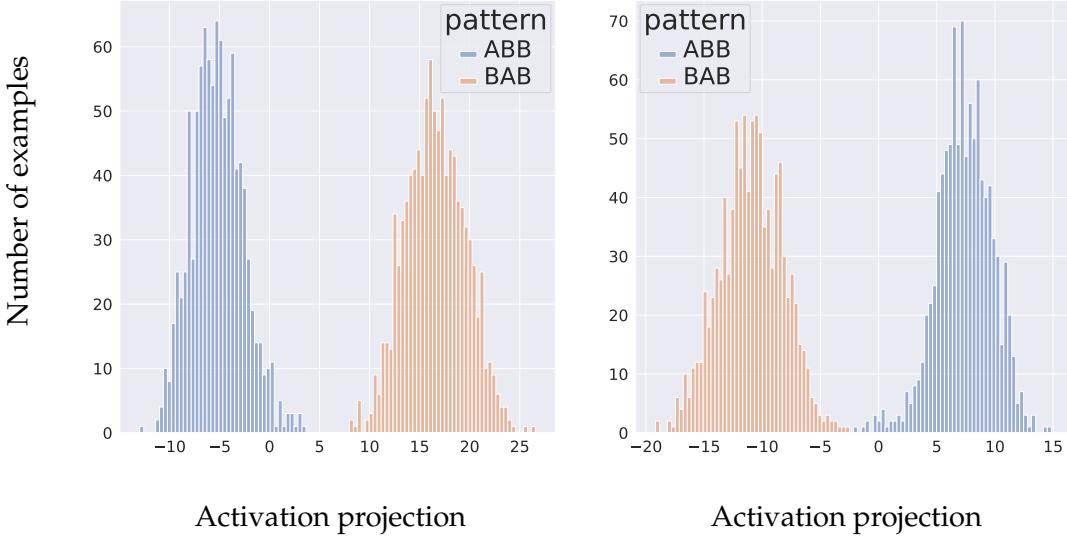


Figure 6: Projections of dataset examples’ activations in the residual stream after layer 8 onto the v_{resid} direction found by DAS (left) and the v_{grad} direction (right) which is the gradient for difference in attention of the name mover heads to the two names.

equivalent to a 1-dimensional subspace intervention that generalizes activation patching. The same arguments from Sections 3 and 8 apply to this intervention, suggesting that it is likely to work successfully in a wide range of MLP layers, regardless of the role of these MLP layers for factual recall.

- Finally, we show that the existence of the illusory patches from 1. implies the existence of rank-1 weight edits which have identical effect at the token being patched, and comparable overall effect on the model. This provides the other direction of an approximate equivalence between 1-dimensional subspace interventions and rank-1 editing, which may be of independent interest.

In particular, these findings provide a mechanistic explanation for the observation of prior work (Hase et al., 2023) that ROME works even in layers where the fact is supposedly not stored. As we discuss in Section 8, we expect that in practice all MLP layers between two model components communicating some feature are likely to contain an illusory subspace – and, by virtue of the approximate equivalence we demonstrate, rank-one fact edits will exist in these MLP layers, regardless of whether they are responsible for recalling the fact being edited.

6.1 Finding Illusory 1-Dimensional Patches for Factual Recall

Given a fact expressed as a subject-relation-object triple (s, r, o) (e.g., $s = \text{'Eiffel Tower'}$, $r = \text{'is in'}$, $o = \text{'Paris'}$), we say that a model M *recalls* the fact (s, r, o) if M completes a prompt expressing just the subject-relation pair (s, r) (e.g., ‘The Eiffel Tower is in’) with the object o (‘Paris’).

Let us be given two facts (s, r, o) and (s', r, o') for the same relation that a model recalls correctly, with corresponding factual prompts A expressing (s, r) and B expressing (s', r) (e.g., $r = \text{'is in'}$, $A = \text{'The Eiffel Tower is in'}$, $B = \text{'The Colosseum is in'}$). In this subsection, we patch from B into A , with

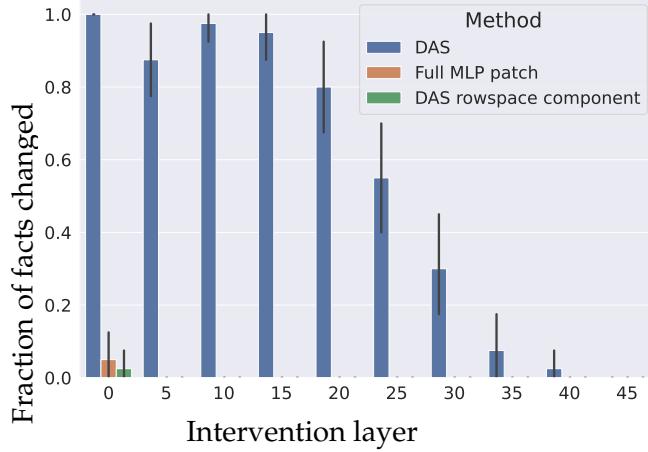


Figure 7: Fraction of successful fact patches under three interventions: patching along the direction found by DAS (blue), patching the component of the DAS direction in the rowspace of W_{out} (green), and patching the entire hidden MLP activation (orange).

the goal of changing the model’s output from o to o' . Implementation details are given in Appendix D.1.

Results are shown in figure 7. We find a stronger version of the same qualitative phenomena as for the IOI illusory direction from Section 4: (i) the directions we find have a strong causal effect (successfully changing o to o'), but (ii) this effect disappears when we instead patch along the subspace spanned by the component orthogonal to $\ker W_{out}$, and (iii) patching the entire MLP activation instead has a negligible effect on the difference in logits between the correct and incorrect objects. Further experiments confirming the illusion are in Appendix D.2.

We conclude that it is possible to make a model output a different object for a given fact by exploiting a 1-dimensional subspace patch that activates a dormant circuit in the model; in particular, using such a patch to localize the fact in the model is prone to interpretability illusions. Next, we turn to a more sophisticated intervention that has been used to edit a fact in a more holistic way, so that related facts update accordingly while the model otherwise stays mostly the same.

6.2 Background on ROME

Meng et al. (2022a) propose an intervention to overwrite a fact (s, r, o) with another fact (s, r, o') while minimally changing the model otherwise. The intervention is a *rank-1 weight edit*, which updates the down-projection W_{out} of a chosen MLP layer to become $W'_{out} = W_{out} + \mathbf{a}\mathbf{b}^\top$ for some $\mathbf{a} \in \mathbb{R}^{d_{\text{resid}}}, \mathbf{b} \in \mathbb{R}^{d_{\text{MLP}}}$. The edit takes a ‘key’ vector $\mathbf{k} \in \mathbb{R}^{d_{\text{MLP}}}$ representing the subject (e.g., an average of its last-token MLP post-gelu activations over many prompts containing it) and a ‘value’ vector $\mathbf{v} \in \mathbb{R}^{d_{\text{resid}}}$ which, when output by the MLP layer, will cause the model to predict the new object o' for the factual prompt (together with some other conditions incentivizing the model to not change much otherwise).

Importantly, we demonstrate that ROME can be formulated as an optimization problem with a natural objective, and this objective allows us to compare it to related interventions. Namely, the vectors \mathbf{a}, \mathbf{b} are the solution to

$$\min_{\mathbf{a}, \mathbf{b}} \text{trace} \left(\text{Cov}_{\mathbf{x} \sim \mathcal{N}(0, \Sigma)} [\mathbf{a}\mathbf{b}^\top \mathbf{x}] \right) \quad \text{subject to} \quad W'_{out} \mathbf{k} = \mathbf{v}. \quad (5)$$

where $\text{Cov}[\mathbf{r}] = \mathbb{E}[(\mathbf{r} - \boldsymbol{\mu})(\mathbf{r} - \boldsymbol{\mu})^\top]$ denotes the covariance matrix of a random vector \mathbf{r} with mean $\boldsymbol{\mu}$, and $\Sigma \succeq 0$ is an empirical (uncentered) covariance matrix for the MLP activations (proof in Appendix D.4). In words, the ROME update is the update that alters W_{out} so it outputs \mathbf{v} on input \mathbf{k} , and minimizes the total variance of the extra contribution of the update in the output of the MLP layer under the assumption that the pre- W_{out} activations are normally distributed with mean zero and covariance $\Sigma \succeq 0$.

6.3 Rank-1 Fact Edits Imply Approximately Equivalent 1-Dimensional Subspace Interventions

Comparing the effect of a rank-1 edit to the MLP layer’s output with equation 3 expressing the effect of patching on the MLP’s outputs, we see that the two are quite similar. This leads to a natural question: given a rank-1 weight edit $W'_{out} = W_{out} + \mathbf{a}\mathbf{b}^\top$ such as ROME, can we find a 1-dimensional activation patch that has the same contribution to the MLP’s output for any MLP activation \mathbf{x} ?

Motivation and details. As it turns out, finding a patch that has the same effect as a rank-1 edit is not feasible in practice. For an activation \mathbf{x} , the extra contribution to the MLP’s output due to a rank-1 edit is $(\mathbf{b}^\top \mathbf{x}) \mathbf{a}$, whereas the extra contribution of a 1-dimensional patch from activation \mathbf{x}' is $(\mathbf{v}^\top \mathbf{x}' - \mathbf{v}^\top \mathbf{x}) W \mathbf{v}$, where crucially $\|\mathbf{v}\|_2 = 1$. In particular, the vectors \mathbf{a}, \mathbf{b} are not norm-constrained, unlike \mathbf{v} . This restricts the magnitude of the contribution of a patch, and we find this matters in practice.

To overcome this, we consider a closely related subspace intervention,

$$\mathbf{x}_{\text{intervention}} = \mathbf{x} + (\mathbf{v}^\top \mathbf{0} - \mathbf{v}^\top \mathbf{x}) \mathbf{v} = \mathbf{x} - (\mathbf{v}^\top \mathbf{x}) \mathbf{v}$$

where \mathbf{v} is no longer restricted to be unit norm, and \mathbf{x}' is chosen to be $\mathbf{0}$ to match the expectation of the rank-1 edit’s contribution (see Appendix D.7 for details). This intervention bears many similarities to subspace patching; in particular, this intervention leaves the projections on all directions orthogonal to \mathbf{v} the same, and the intuitions about the illusion from Sections 3 and 7 also apply to this intervention. We also remark that, at the same time, this intervention is exactly equivalent to the rank-1 edit $W''_{out} = W_{out} + W_{out}\mathbf{v}(-\mathbf{v})^\top$ in terms of contribution to the MLP output.

It turns out that this more general intervention can often approximate the ROME intervention well. Specifically, given a rank-1 edit $W_{out} + \mathbf{a}\mathbf{b}^\top$, we can treat the problem probabilistically over activations $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ like done in Meng et al. (2022a), and ask for a direction \mathbf{v} with the following properties:

- the expected value of both interventions is the same;
- the resulting extra contribution $(-\mathbf{v}^\top \mathbf{x}) W_{out}\mathbf{v}$ to the MLP’s output points in the same direction as the extra contribution $(\mathbf{b}^\top \mathbf{x}) \mathbf{a}$ of the rank-1 edit;
- the total variance trace $(\text{Cov}[(-\mathbf{v}^\top \mathbf{x}) W_{out}\mathbf{v} - (\mathbf{b}^\top \mathbf{x}) \mathbf{a}])$ of the difference of these two contributions is minimized over $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$.

Details are given in Appendix D.7. The important takeaway is that the solution \mathbf{v} has the form

$$\mathbf{v} = \alpha W_{out}^+ \mathbf{a} + \mathbf{u} \quad \text{where} \quad \mathbf{u} \in \ker W_{out}$$

for a constant $\alpha \geq 0$ that is optimized. In particular, $W_{out}\mathbf{v}$ points in the direction \mathbf{a} , and the component \mathbf{u} (which is causally disconnected) is a ‘free variable’ that is essentially optimized to bring $\mathbf{v}^\top \mathbf{x}$ close to $-\mathbf{b}^\top \mathbf{x}/\alpha$ (subject to accounting for Σ).

Metrics and evaluation. We apply this method to find subspace interventions corresponding to edits extracted from the COUNTERFACT dataset (Meng et al., 2022a); see Appendix D.1 for details. Specifically, we run ROME for all the edits, and we find the approximate subspace intervention (defined by a vector \mathbf{v}) corresponding to each ROME edit. To compare the interventions, we consider the following metrics:

Rewrite score. Defined in Hase et al. (2023) (and a closely related metric is optimized by Meng et al. (2022a)), the rewrite score is a relative measure of how well a change to the model (ROME or our subspace intervention) increases the probability of the false target o' we are trying to substitute for o . Specifically, if $p_{\text{clean}}(o)$ is the probability assigned by the model to output o under normal operation, and $p_{\text{intervention}}(o)$ is the probability assigned when the intervention is applied, the rewrite score is

$$\frac{p_{\text{intervention}}(o') - p_{\text{clean}}(o')}{1 - p_{\text{clean}}(o')} \in (-\infty, 1].$$

with a value of 1 indicating the model assigns probability 1 to o' after the intervention. We measure the rewrite score for the ROME intervention, our approximation of it, and also the corresponding subspace intervention with the $\ker W_{out}$ component of \mathbf{v} removed, in analogy with how we examine the subspace patches in Sections 4 and 6.1. That is, if \mathbf{v}_{null} is the orthogonal projection of \mathbf{v} on $\ker W_{out}$ and $\mathbf{v}_{\text{rowsp}} = \mathbf{v} - \mathbf{v}_{\text{null}}$, we apply the intervention

$$\mathbf{x}_{\text{rowspace intervention}} = \mathbf{x} - (\mathbf{v}_{\text{rowsp}}^\top \mathbf{x}) \mathbf{v}_{\text{rowsp}}.$$

Results comparing ROME and the subspace intervention we use to approximate it are shown in Figure 8. When using the rowspace intervention, all rewrite scores are less than 10^{-3} , indicating a strong reliance on the nullspace component.

Cosine similarity of \mathbf{v} and \mathbf{b} . Our intervention contributes $-(\mathbf{v}^\top \mathbf{x}) W \mathbf{v}$, and the ROME edit contributes $(\mathbf{b}^\top \mathbf{x}) \mathbf{a}$. Note that, by construction, the cosine similarity of $W \mathbf{v}$ with \mathbf{a} is 1. So, the cosine similarity of \mathbf{v} and \mathbf{b} measures how well the direction we are projecting the activation \mathbf{x} on matches that from the ROME edit. Results are shown in Figure 9 (left); in a range of layers we observe cosine similarity significantly close to 1.

Overall change to the model relative to ROME. This is the total variance introduced by this intervention as a fraction of the total variance introduced by the corresponding ROME edit. It measures the extent to which the subspace intervention damages the model overall, following our formulation of ROME as an optimization problem (see Appendix D.4). Results are shown in Figure 9 (right). Note that this metric is a ratio of variances; a ratio of standard deviations can be obtained by taking the square root.

In conclusion, we observe that in layers 20-35 inclusive, the two interventions are very similar according to all metrics considered.

What is the interpretability illusion implied by this? An important difference between the IOI case study from Section 4 and the factual recall results from the current section is that, while activating a dormant circuit is contrary to activation patching’s interpretability fact editing is, by definition, allowed to alter the model. In this sense, activating a dormant circuit via a rank-1 edit should no longer be considered a sign of spuriousity.

Instead, we argue that the interpretability illusion is to assume that the success of ROME means that the fact is stored in the layer being edited. This was already observed in Hase et al. (2023). Our work provides a mechanistic explanation for this observation.

We also note that we have evaluated the success of ROME and our approximately-equivalent subspace intervention only using the rewrite score metric and the measure of total variance implicit in the ROME algorithm. Ideally, there would be other validations of a fact edit that test the behavior of the intervened-upon model on other facts that should be changed by the edit.

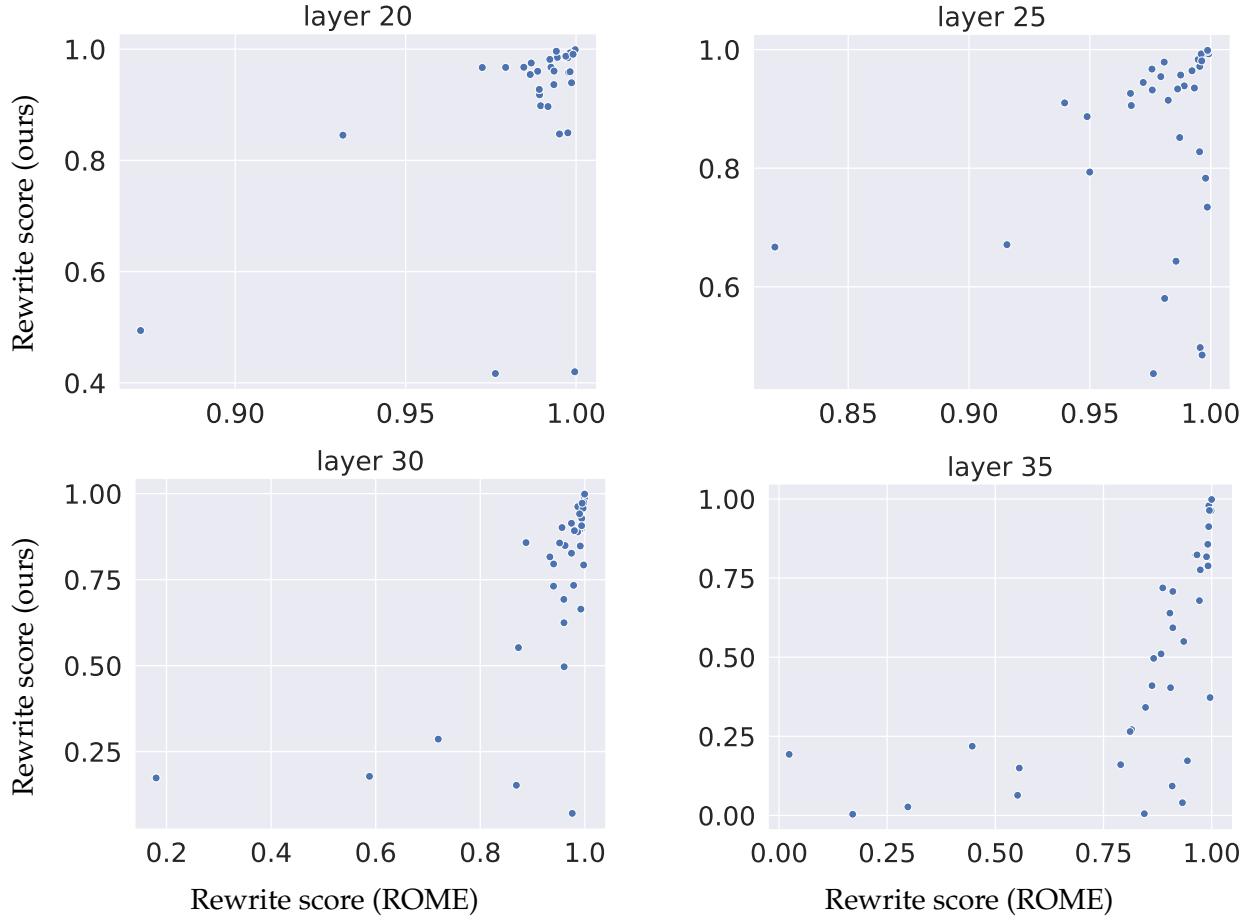


Figure 8: Rewrite score comparison between ROME (x-axis) and our approximation to it (y-axis) via a subspace intervention for layers 20, 25, 30, 35.

6.4 1-Dimensional Fact Patches Imply Equivalent Rank-1 Fact Edits

Finally, we show that the existence of an activation patch as in Subsection 6.1 implies the existence of a rank-1 weight edit which has the same contribution to the MLP’s output at the token being patched, and otherwise results in very similar model outputs as the activation patch.

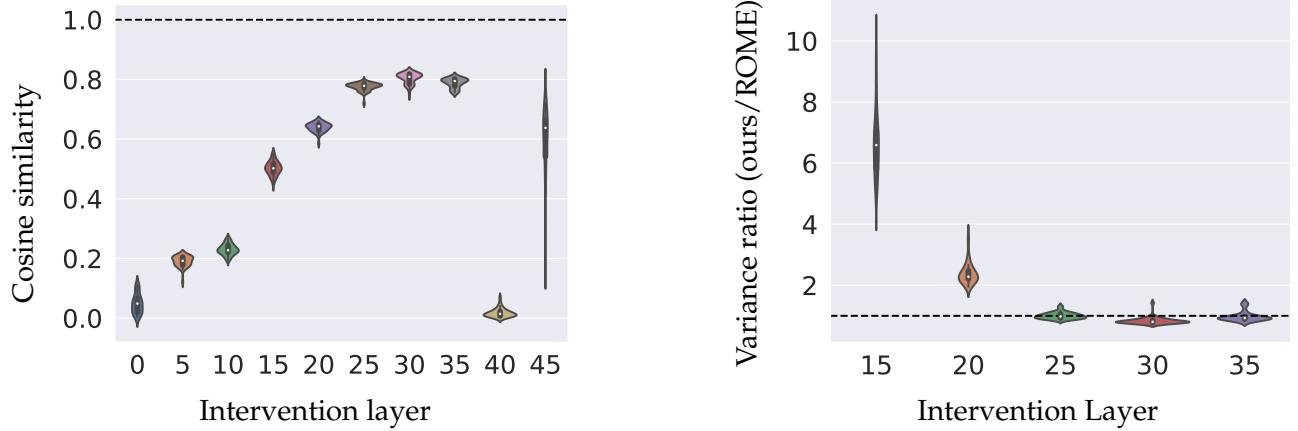


Figure 9: Comparisons between ROME rank-1 edits and our approximation via a subspace intervention. Left: cosine similarity between the vector \mathbf{v} defining the subspace we intervene on and the vector \mathbf{b} from the ROME edit (dashed horizontal line is at $y = 1$). Right: ratio of the total variance introduced by the subspace intervention to the total variance of the ROME intervention (x-axis scale is restricted to make the plot readable; dashed horizontal line is at $y = 1$).

Intuitively, a ‘fact patch’ as in Subsection 6.1 should have a corresponding rank-1 edit with the same effect: the last subject token MLP activation \mathbf{u}_A for prompt A takes the role of \mathbf{k} , and the patch modifies the MLP’s output (making it \mathbf{v}) to change the model’s output to o' . We make this intuition formal in Appendix D.5, where we show that for each 1-dimensional activation patch between a pair of examples in the post-GELU activations of an MLP layer, there is a rank-1 model edit to W_{out} that results in the same change to the MLP layer’s output at the token where we do the patching, and minimizes the variance of the extra contribution in the sense of Equation 5.

While this shows that the patch implies a rank-1 edit with the same behavior *at the token where we perform the patch*, the rank-1 edit is applied *permanently* to the model, which means that it (unlike the activation patch) applies to *every token*. Thus, it is not a priori obvious whether the rank-1 edit will still succeed in making the model predict o' instead of o . To this end, in Appendix D.6, we evaluate empirically how using the rank-1 edit derived in Appendix D.5 instead of the activation patch changes model predictions, and we find negligible differences.

7 Reasons to Expect the MLP-in-the-middle Illusion to be Prevalent

We only exhibit our illusion empirically in two settings: the IOI task and factual recall. However, we believe it is likely prevalent in practice. In this section, we provide several theoretical, heuristic and empirical arguments in support of this.

Specifically, we expect the illusion to be likely occur whenever we have an MLP M which is not used in the model’s computation on a given task, but is between two components A and B which are used, and communicate by writing to / reading from the direction \mathbf{v} via the skip connections of the model. This structure has been frequently observed in the mechanistic interpretability literature (Lieberum et al., 2023; Wang et al., 2023; Olsson et al., 2022; Geva et al., 2021): circuits contain components composing with each other separated by multiple layers, and circuits have often been observed to be sparse, with most components (including most MLP layers) not playing a significant role.

7.1 Assumptions: a Simple Model of Linear Features in the Residual Stream

The linear representation hypothesis suggests a natural way to formalize this intuition. Namely, let's assume for simplicity that there is a binary feature C in the data, and the value of C influences the model's behavior on a task, by e.g. making some next-token predictions more likely than others. Concretely, there is a residual stream direction $\mathbf{v} \in \mathbb{R}^{d_{\text{resid}}}$ that mediates this effect: projections on \mathbf{v} (at an appropriate token position) linearly separate examples according to the value of C , and intervening on this projection by setting it to e.g. the mode of the opposite class has the same effect on model outputs as changing the value of C in the input itself. Furthermore, we assume that this direction has this property in all residual stream layers between some two layers $a < b$.

We note that these assumptions can be realized ‘in the wild’: the highly similar directions \mathbf{v}_{grad} and $\mathbf{v}_{\text{resid}}$ discussed in Section 5 are both examples of such directions \mathbf{v} for the binary concept of whether the IO name comes first or second in the sentence, as we argued empirically.

7.2 Overview of Argument

The key hypothesis is that, given the setup from the previous Subsection 7.1, the post-nonlinearity activations of every MLP layer between layers a and b are likely to contain a 1-dimensional subspace whose patching will have the same effect (possibly with a smaller magnitude) on model outputs as changing the value of C in the input. For this, it is sufficient to have two kinds of directions in the MLP's activation space:

- a ‘causal’ direction, such that changing the projection of an activation along this direction results in the expected change of model outputs. Such a direction will exist simply because W_{out} is a full-rank matrix, so we can simply pick $W_{\text{out}}^+ \mathbf{v}$. We give an empirically-supported theoretical argument for this in Appendix E.1.
- a ‘correlated’ direction that linearly discriminates between the values of C : such a direction will exist because the pre-nonlinearity activations (which are an approximately linear image of the residual stream) will linearly discriminate C , and the transformation $\mathbf{x} \mapsto \text{gelu}(\mathbf{x}) \mapsto \text{proj}_{\ker W_{\text{out}}} \text{gelu}(\mathbf{x})$ approximately preserves linear separability. We give an empirically-supported theoretical argument for this in Appendix E.2.

8 Discussion, Limitations, and Recommendations

Throughout this paper, we have seen that interventions on arbitrary linear subspaces of model activations, such as subspace activation patching, can have counterintuitive properties. In this section, we take a step back and provide a more conceptual point of view on these findings, as well as concrete advice for interpretability practitioners.

Why should this phenomenon be considered an illusion? One argument for the illusory nature of the subspaces we find is the reliance on a large causally disconnected component (in all our examples, this component is in the kernel of the down-projection W_{out} of an MLP layer). In particular, patching along only the causally-relevant component of the subspace (the one in $(\ker W_{\text{out}})^\perp$) destroys the effect of the subspace patch; we find this a convincing reason to be suspicious of the explanatory faithfulness of these subspaces.

Beyond this argument, there are several more subtle considerations. For an explanation to be ‘illusory’, there has to be some notion of what the ‘true’ explanation is. We admit that a

definition of a ‘ground truth’ mechanistic explanation is conceptually challenging. In the absence of such a definition, our claims rest on various observations about model’s inner workings that we now collect in one place and make more explicit. We believe these findings collectively point to meaningful constraints on mechanistic explanations.

For example, the IOI circuit work of Wang et al. (2023) finds through various component-level interventions that the layer 8 MLP *as a whole* does not contribute significantly to the model’s ability to do the IOI task. However, does this imply that there aren’t individual subspaces of the MLP layer that mediate the model’s behavior on the IOI task? Not necessarily: there could be, for example, two subspaces mediating the position signal, but which have opposite contributions to the MLP’s output that cancel out. This is compatible with our model of the illusion from Section 3: for example, we can form two ‘cancelling’ 1-dimensional subspaces by taking the sum and difference of the causally disconnected and dormant directions in our model. From this point of view, our subspace intervention decouples these (ordinarily coupled) subspaces by changing the activation only along one of them. This is impossible for an intervention that operates on entire model components.

Should we prefer the view under which the MLP layer simply does not participate in any meaningful way in the IOI task, or the view under which it contains subspaces that mediate information about the IOI task, but whose contributions cancel out? Note that meaningful cancellation behavior between entire model components has been observed to some extent in the mechanistic interpretability literature, such as with negative heads (Wang et al., 2023), anti-induction heads (Olsson et al., 2022) and copy suppression heads (McDougall et al., 2023). Furthermore, it is not clear that, in general, a component-level explanation should take precedence over subspace-level explanations. So, *a priori*, we have a conundrum: two different kinds of interventions arrive at conflicting interpretations.

Nevertheless, based on our experiments, we suggest that the view under which the MLP layer contains meaningful subspaces that cancel out is the less likely mechanistic explanation for several reasons. A first argument is that, as we argue in Section 7, the existence of the illusory subspace only relies on the existence of certain directions in the residual stream; the MLP weights themselves don’t play a role. In some sense, the illusory subspace is a ‘necessity of linear algebra’. This is further reinforced by the fact that we find illusory directions even when the MLP weights are replaced by random matrices (see Appendix B.5). A second argument is that features that are individually strong, but whose contributions almost exactly cancel out, seem unlikely to be prevalent.

Finally, we again remark that circuits for specific tasks have been observed to be sparse (recall Section 7). Our model of the illusion from Section 3 and the evidence from Section 7 suggest that any MLP layer between two circuit components using the residual stream as a communication bottleneck for some feature will contain a subspace that appears to mediate this feature. Thus, even if we cannot conclusively rule out any given MLP layer on the path as not being meaningfully involved in the computation, it would be quite surprising if always all of them are involved. So we expect that at least some of these subspaces will be illusory.

The importance of correct model units. A further implicit assumption in our work is that model components are meaningful boundaries for mechanistic explanation. As we illustrate in Appendix A.3, our toy example of the illusion can be considered in a rotated basis, in which the ‘illusory’ direction appears ‘meaningful’. In a similar way, if we allow ourselves to arbitrarily reparametrize spaces of activations by crossing the boundaries between e.g. attention heads and MLP layers, calling the MLP subspace ‘illusory’ is much more tenuous.

To respond to this criticism, we point to the many observations in the mechanistic literature

that different components (like heads and MLP layers) perform qualitatively different functions in a model. For example, tasks involving algorithmic manipulations of in-context text, such as the IOI task, often rely predominantly on attention heads (Wang et al., 2023; Heimersheim & Janiak). On the other hand, MLP layers have so far been implicated in tasks having to do with recalling bigrams and facts (Meng et al., 2022a; Gurnee et al., 2023). On these grounds, mixing activations between them is likely to lead to less parsimonious and less principled mechanistic explanations.

Takeaways and recommendations. As we have seen, optimization-based methods using subspace activation patching, such as DAS (Geiger et al., 2023b) can find both faithful (Section 5) and illusory (Section 4) features with respect to the model’s computation. We recommend running such methods in activation bottlenecks, especially the residual stream, as well as using validations beyond end-to-end evaluations to ascertain the precise role of such features.

9 Acknowledgments

We are deeply indebted to Atticus Geiger for many useful discussions and helpful feedback, as well as help writing parts of the paper. We particularly appreciate his thoughtful pushback on framing and narrative, and commitment to rigour, without which this manuscript would be far poorer. We would also like to thank Christopher Potts, Curt Tigges, Oskar Hollingsworth, Tom Lieberum, Senthooran Rajamanoharan and Peli Grietzer for valuable feedback and discussions. The authors used the open-source library `transformerlens` (Nanda & Bloom, 2022) to carry out the experiments related to the IOI task. AM and GL did this work as part of the SERI MATS independent research program, with support from AI Safety Support.

10 Author Contributions

NN was the main supervisor for this project, and provided high level feedback on experiments, prioritisation, and writing throughout. NN came up with the original idea of the illusion and the conceptual example. AM came up with the correspondence with factual recall, developed the factual recall results, and ran the experiments for Sections 6, 7 and part of 5 (with the exception of experiments from Appendix E.6 ran by GL), and wrote the majority of the paper and appendices, as well as the public version of the code for the paper. GL also ran the experiments for Sections 4 and 5 with some guidance from AM, wrote and ran the experiments for Appendices B.5, B.6, B.7, C and E.6, and contributed to Section 5 and various other sections.

References

- Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. *arXiv preprint arXiv:2109.06129*, 2021.
- Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolo Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*,

December 3-8, 2018, Montreal, Canada, pp. 9525–9536, 2018. URL <https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html>.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

Mariette Awad, Rahul Khanna, Mariette Awad, and Rahul Khanna. Support vector machines for classification. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, pp. 39–66, 2015.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for bert. *arXiv preprint arXiv:2104.07143*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Lawrence Chan, Adria 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, 2022. URL <https://www.alignmentforum.org/posts/JvZhhzyCHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing>.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*, 2023.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*, 2023.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrau, and Marco Baroni. What you can cram into a single \$&#!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL <https://aclanthology.org/P18-1198>.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL <https://aclanthology.org/2022.acl-long.581>.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.

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>.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 163–173, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.16. URL <https://www.aclweb.org/anthology/2020.blackboxnlp-1.16>.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Atticus Geiger, Christopher Potts, and Thomas Icard. Causal abstraction for faithful model interpretation. Ms., Stanford University, 2023a. URL <https://arxiv.org/abs/2301.04709>.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D Goodman. Finding alignments between interpretable causal variables and distributed neural representations. *arXiv preprint arXiv:2303.02536*, 2023b.

Robert Geirhos, Roland S Zimmermann, Blair Bilodeau, Wieland Brendel, and Been Kim. Don’t trust your eyes: on the (un)reliability of feature visualizations. *arXiv preprint arXiv:2306.04719*, 2023.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL <https://aclanthology.org/2021.emnlp-main.446>.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*, 2023.

Aaron Gokaslan and Vanya Cohen. Openwebtextcorpus, 2019. URL <http://Skylion007.github.io/OpenWebTextCorpus>.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

G Grand, I Blank, F Pereira, and E Fedorenko. Semantic projection: Recovering human knowledge of multiple, distinct object features from word embeddings. arxiv. *arXiv preprint arXiv:1802.01241*, 2018.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *arXiv preprint arXiv:2301.04213*, 2023.

Stefan Heimersheim and Jett Janiak. 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>.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models, 2023.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. Implicit representations of meaning in neural language models. *arXiv preprint arXiv:2106.00737*, 2021.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023a.

Maximilian Li, Xander Davies, and Max Nadeau. Circuit breaking: Removing model behaviors with targeted ablation. *arXiv preprint arXiv:2309.05973*, 2023b.

Tom Lieberum, Matthew Rahtz, János Kramár, Geoffrey Irving, Rohin Shah, and Vladimir Mikulík. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

Zachary C. Lipton. The mythos of model interpretability, 2016. URL <https://arxiv.org/abs/1606.03490>.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head. *arXiv preprint arXiv:2310.04625*, 2023.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

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

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022b.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 3111–3119, 2013. URL <https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html>.

Neel Nanda. Actually, othello-gpt has a linear emergent world representation. 2023a.

Neel Nanda. Attribution patching: Activation patching at industrial scale, 2023b. URL <https://www.neelnanda.io/mechanistic-interpretability/attribution-patching>.

Neel Nanda and Joseph Bloom. Transformerlens. <https://github.com/neelnanda-io/TransformerLens>, 2022.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. <https://www.transformer-circuits.pub>, 2022. URL <https://www.transformer-circuits.pub/2022/mech-interp-essay>.

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.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.

OpenAI. Gpt-4 technical report, 2023.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, UAI'01*, pp. 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

P. Smolensky. Neural and conceptual interpretation of PDP models. In *Parallel Distributed Processing: Explorations in the Microstructure, Vol. 2: Psychological and Biological Models*, pp. 390–431. MIT Press, Cambridge, MA, USA, 1986. ISBN 0262631105.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. Understanding arithmetic reasoning in language models using causal mediation analysis. *arXiv preprint arXiv:2305.15054*, 2023.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601,

Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL <https://aclanthology.org/P19-1452>.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fdb053c1c4a845aa-Abstract.html>.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias. URL <https://arxiv.org/abs/2004.12265>.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Jonas Wallat, Jaspreet Singh, and Avishek Anand. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 174–183, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.17. URL <https://aclanthology.org/2020.blackboxnlp-1.17>.

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>.

Ziheng Wang, Jeremy Wohlwend, and Tao Lei. Structured pruning of large language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6151–6162, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.496. URL <https://aclanthology.org/2020.emnlp-main.496>.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Zhengxuan Wu, Atticus Geiger, Christopher Potts, and Noah D Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. *arXiv preprint arXiv:2305.08809*, 2023.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions, 2023.

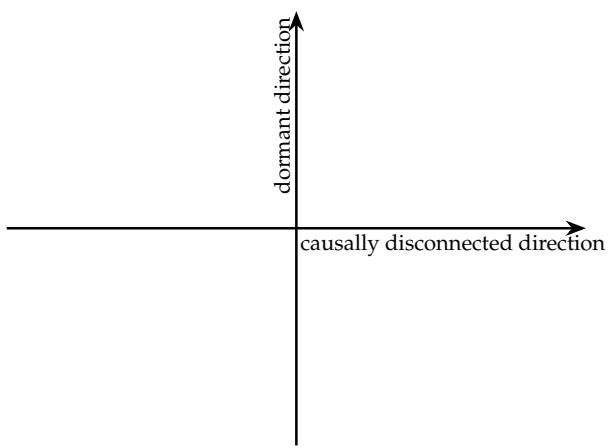


Figure 10: Consider a 2-dimensional subspace of model activations, with an orthogonal basis where the x -axis is *causally disconnected* (changing the activation along it makes no difference to model outputs) and values on the y axis are always zero for examples in the data distribution (a special case of a *dormant* direction).

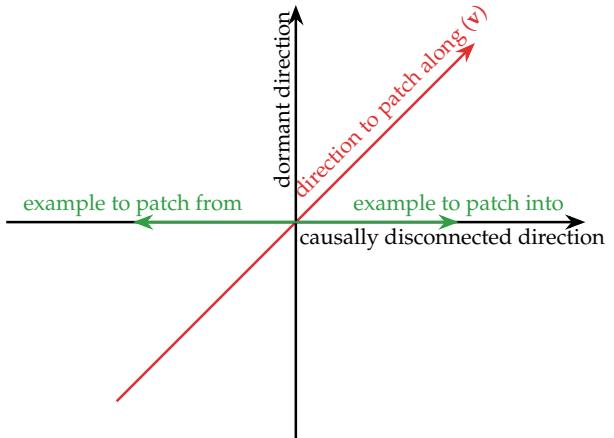


Figure 11: Suppose we have two examples (green) which differ in their projection on the causally disconnected direction (and have zero projection on the dormant direction, by definition). Let's consider what happens when we patch from the example on the left into the example on the right along the 1-dimensional subspace v spanned by the vector $(1, 1)$ (red)

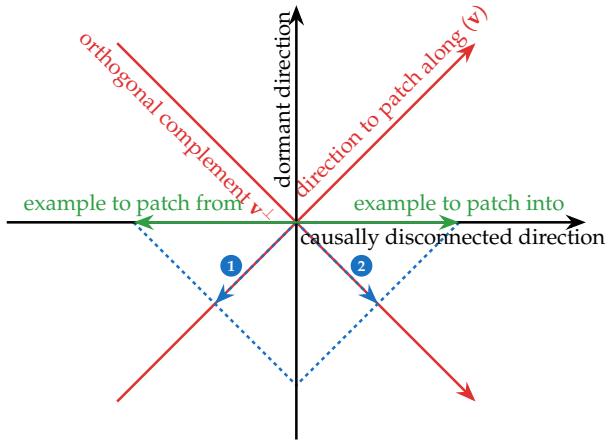


Figure 12: To patch along v from the left into the right example, we match the projection on v from the left one, and leave the projection on v^\perp unchanged. In other words, we take the component of the left example along v (①) and sum it with the v^\perp component (②) of the original activation.

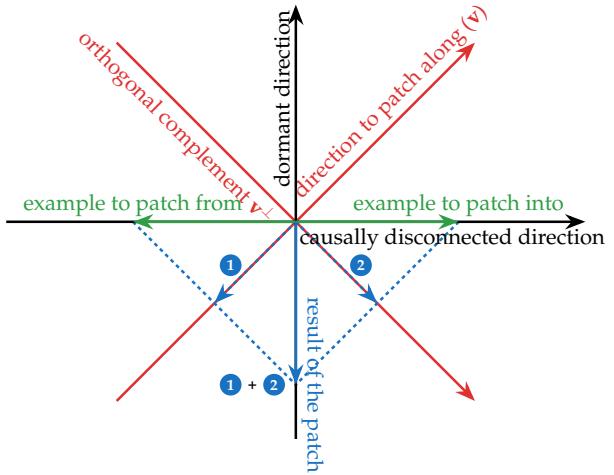


Figure 13: This results in the patched activation ① + ②, which points completely along the dormant direction. In this way, activation patching makes the variation of activations along the causally disconnected x -axis result in activations along the previously dormant y -axis.

Figure 14: A step-by-step illustration of the phenomenon shown in Figure 1.

A Additional Notes on Section 3

A.1 The Illusion for Higher-Dimensional Subspaces

In the main text, we mostly discuss the illusion for activation patching of 1-dimensional subspaces for ease of exposition. Here, we develop a more complete picture of the mechanics of the illusion for higher-dimensional subspaces.

Let \mathcal{C} be a model component taking values in \mathbb{R}^d , and let $U \subset \mathbb{R}^d$ be a linear subspace. Let V be a matrix whose columns form an orthonormal basis for U . If the \mathcal{C} activations for examples A and B are $\mathbf{act}_A, \mathbf{act}_B \in \mathbb{R}^d$ respectively, patching U from A into B gives the patched activation

$$\mathbf{act}_B^{\text{patched}} = \mathbf{act}_B + VV^\top(\mathbf{act}_A - \mathbf{act}_B) = (I - VV^\top)\mathbf{act}_B + VV^\top\mathbf{act}_A$$

For intuition, note that VV^\top is the orthogonal projection on U , so this formula says to replace the orthogonal projection of \mathbf{act}_B on U with that of \mathbf{act}_A , and keep the rest of \mathbf{act}_B the same.

Generalizing the discussion from Section 3, for the illusion to occur for subspace S , we need S to be sufficiently aligned with a causally disconnected subspace $V_{\text{disconnected}}$ that is correlated with the feature being patched, and a dormant but causal subspace V_{dormant} which, when set to out of distribution values, can achieve the wanted causal effect.

For example, a particularly simple way in which this could happen is if we let $V_{\text{disconnected}}, V_{\text{dormant}}$ be 1-dimensional subspaces (like in the setup for the 1-dimensional illusion), and we form S by combining $V_{\text{disconnected}} \oplus V_{\text{dormant}}$ with a number of orthogonal directions that are approximately constant on the data with respect to the feature we are patching. These extra directions effectively don't matter for the patch (because they are cancelled by the $\mathbf{act}_A - \mathbf{act}_B$ term). Given a specific feature, it is likely that such weakly-activating directions will exist in a high-dimensional activation space. Thus, if the 1-dimensional illusion exist, so will higher-dimensional ones.

A.2 Optimal Illusory Patches are Equal Parts Causally Disconnected and Dormant

In this subsection, we prove a quantitative corollary of the model of our illusion that suggests that we should expect optimal illusory patching directions to be of the form

$$\mathbf{v}_{\text{illusory}} = \frac{1}{\sqrt{2}} (\mathbf{v}_{\text{disconnected}} + \mathbf{v}_{\text{dormant}})$$

for unit vectors $\mathbf{v}_{\text{disconnected}} \perp \mathbf{v}_{\text{dormant}}$. In other words, we expect the strongest illusory patches to be formed by combining a disconnected and illusory direction with *equal* coefficients, like depicted in Figure 1:

Lemma A.1. *Suppose we have two distributions of input prompts $\mathcal{D}_A, \mathcal{D}_B$. In the terminology of Section 3, let $\mathbf{v}_{\text{disconnected}} \perp \mathbf{v}_{\text{dormant}}$ be unit vectors such that the subspace spanned by $\mathbf{v}_{\text{disconnected}}$ is a causally disconnected subspace, and the subspace spanned by $\mathbf{v}_{\text{dormant}}$ is **strongly** dormant, in the sense that the projections of the activations of all examples $\mathcal{D}_{\text{source}} \cup \mathcal{D}_{\text{base}}$ onto $\mathbf{v}_{\text{dormant}}$ are equal to some constant c .*

Let $\mathbf{v} = \mathbf{v}_{\text{disconnected}} \cos \alpha + \mathbf{v}_{\text{dormant}} \sin \alpha$ be a unit-norm linear combination of the two directions parametrized by an angle α . Then the magnitude of the expected change in projection along $\mathbf{v}_{\text{dormant}}$ when patching from $x_A \sim \mathcal{D}_A$ into $x_B \sim \mathcal{D}_B$ is maximized when $\alpha = \frac{\pi}{4}$, i.e. $\cos \alpha = \sin \alpha = \frac{1}{\sqrt{2}}$.

Proof. Recall that the patched activation from x_A into x_B along \mathbf{v} is

$$\mathbf{act}_B^{\text{patched}} = \mathbf{act}_B + (p_A - p_B)\mathbf{v}$$

where $p_A = \mathbf{v}^\top \mathbf{act}_A$, $p_B = \mathbf{v}^\top \mathbf{act}_B$ are the projections of the two examples' activations on v . The change along $\mathbf{v}_{\text{dormant}}$ is thus

$$\begin{aligned}\mathbf{v}_{\text{dormant}}^\top (\mathbf{act}_B^{\text{patched}} - \mathbf{act}_B) &= (p_A - p_B) \sin \alpha = (\mathbf{v}^\top \mathbf{act}_A - \mathbf{v}^\top \mathbf{act}_B) \sin \alpha \\ &= \mathbf{v}_{\text{disconnected}}^\top (\mathbf{act}_A - \mathbf{act}_B) \cos \alpha \sin \alpha\end{aligned}$$

where we used the assumption that $\mathbf{v}_{\text{dormant}}^\top \mathbf{act}_A = \mathbf{v}_{\text{disconnected}}^\top \mathbf{act}_B$. Hence, the expected change is

$$\cos \alpha \sin \alpha \mathbf{v}_{\text{disconnected}}^\top \mathbb{E}_{A \sim \mathcal{D}_A, B \sim \mathcal{D}_B} [\mathbf{act}_A - \mathbf{act}_B].$$

The function $f(\alpha) = \cos \alpha \sin \alpha$ for $\alpha \in [0, \pi/2]$ is maximized for $\alpha = \pi/4$, concluding the proof. \square

A.3 The Toy Illusion in a Rotated Basis

There is a subtlety in the toy example of the illusion from 3.3. Suppose we reparametrized the hidden layer of the network so that, instead of the standard basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, we use a rotated basis where one of the directions is $\mathbf{e}_1 + \mathbf{e}_2$, the other direction is orthogonal to it and to \mathbf{w}_2 (hence will be causally disconnected), and the last direction is the unique direction orthogonal to the first two.

The unit basis vectors for this new basis are given by

$$\begin{aligned}\mathbf{d}_1 &= (\mathbf{e}_1 + \mathbf{e}_2) / \sqrt{2}, \\ \mathbf{d}_2 &= (-\mathbf{e}_1 + \mathbf{e}_2 - 2\mathbf{e}_3) / \sqrt{6}, \\ \mathbf{d}_3 &= (-\mathbf{e}_1 + \mathbf{e}_2 + \mathbf{e}_3) / \sqrt{3}.\end{aligned}$$

If we assemble these into the rows of a rotation matrix

$$R = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\ \frac{\sqrt{6}}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

the re-parametrized network is then given by

$$x \mapsto \mathbf{h}' = R\mathbf{w}_1 x \mapsto y = (R\mathbf{w}_2)^\top \mathbf{h}'$$

and a diagram of this new network is shown in Figure 15. From this point of view, \mathbf{d}_1 takes the role that \mathbf{e}_3 had before: the input is essentially copied to it (modulo scalar multiplication), and then read from it at the output. By contrast, \mathbf{d}_2 is now a causally disconnected direction, and \mathbf{d}_3 is a dormant direction.

B Additional details for Section 4

B.1 Dataset, Model and Evaluation Details for the IOI Task

We use GPT2-Small for the IOI task, with a dataset that spans 216 single-token names, 144 single-token objects and 75 single-token places, which are split 1 : 1 across a training and test set. Every example in the data distribution includes (i) an initial clause introducing the indirect object (**IO**, here 'Mary') and the subject (**S**, here 'John'), and (ii) a main clause that refers to the subject a second time. Beyond that, the dataset varies in the two names, the initial clause content, and the main clause content. Specifically, use three templates as shown below:

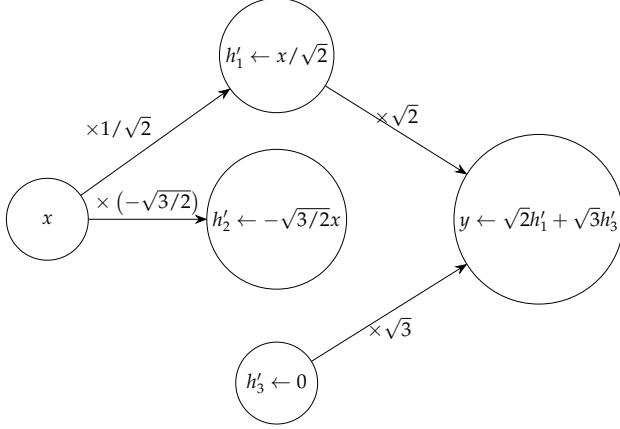


Figure 15

Then, [] and [] had a long and really crazy argument. Afterwards, [] said to Then, [] and [] had lots of fun at the [place]. Afterwards, [] gave a [object] to Then, [] and [] were working at the [place]. [] decided to give a [object] to

and we use the first two in training and the last in the test set. Thus, the test set relies on unseen templates, names, objects and places. We used fewer templates than the IOI paper Wang et al. (2020) in order to simplify tokenization (so that the token positions of our names always align), but our results also hold with shifted templates like in the IOI paper.

On the test partition of this dataset, GPT2-Small achieves an accuracy of $\approx 91\%$. The average difference of logits between the correct and incorrect name is ≈ 3.3 , and the logit of the correct name is greater than that of the incorrect name in $\approx 99\%$ of examples. Note that, while the logit difference is closely related to the model’s correctness, it being > 0 does not imply that the model makes the correct prediction, because there could be a third token with a greater logit than both names.

B.2 Details for Computing the Gradient Direction \mathbf{v}_{grad}

For a given example from the test distribution and a given name mover head, we compute the gradient of the difference of attention scores from the final token position to the 3rd and 5th token in the sentence (where the two name tokens always are in our data). We then average these gradients over a large sample of the full test distribution and over the three name mover heads, and finally normalize the resulting vector to have unit ℓ_2 norm.

We note that there is a ‘closed form’ way to compute approximately the same quantity that requires no optimization. Namely, for a single example we can collect the keys k_S, k_{IO} to the name mover heads at the first two names in the sentence (the **S** and **IO** name). Then, for a single name mover head with query matrix W_Q , a maximally causal direction v in the residual stream at the last token position after layer 8 will be one such that $W_Q v$ is in the direction of $k_S - k_{IO}$, because the attention score is simply the dot product between the keys and queries. We can use this to ‘backpropagate’ to v by multiplying with the pseudoinverse W_Q^+ . This is slightly complicated by the fact that we have been ignoring layer normalization, which can be approximately accounted for by estimating the scaling parameters (which tend to concentrate well) from the IOI data distribution. We note that this approach leads to broadly similar results.

B.3 Training Details for DAS

To train DAS, we always sample examples from the training IOI distribution. We sample equal amounts of pairs of base (which will be patched into) and source (where we take the activation to patch in from) prompts where the two names are the same between the prompts, and pairs of prompts where all four names are distinct. We optimize DAS to maximize the logit difference between the name that should be predicted if the position information from the source example is correct and the other name.

For training, we use a learned rotation matrix as in the original DAS paper ([Geiger et al., 2023b](#)), parametrized with `torch.nn.utils.parametrizations.orthogonal`. We use the Adam optimizer and minibatch training over a training set of several hundred patching pairs. We note that results remain essentially the same when using a higher number of training examples.

B.4 Discussion of the Magnitude of the Illusion

While the contribution of the \mathbf{v}_{MLP} patch to logit difference may appear relatively small, we note that this is the result of patching a direction in a single model component at a single token position. Typical circuits found in real models (including the IOI circuit from [Wang et al. \(2023\)](#)) are often composed of multiple model components, each of which contribute. In particular, the position signal itself is written to by 4 heads, and chiefly read by 3 other heads. As computation tends to be distributed, when patching an individual component accuracy may be a misleading metric (eg patching 1 out of 3 heads is likely insufficient to change the output), and a fractional logit diff indicates a significant contribution. By contrast, patching in the residual stream is a more potent intervention, because it can affect *all* information accumulated in the model that is communicated to downstream components.

B.5 Random ablation of MLP weights

How certain are we that MLP8 doesn't actually matter for the IOI task? While we find the IOI paper analysis convincing, to make our results more robust to the possibility that it does matter, we also design a further experiment.

Given our conceptual picture of the illusion, the computation performed by the MLP layer where we find the illusory subspace does not matter as long as it propagates the correlational information about the position feature from the residual stream to the hidden activations, and as long as the output matrix W_{out} is full rank (also, see the discussion in [8](#)). Thus, we expect that if we replace the MLP weights by randomly chosen ones with the same statistics, we should still be able to exhibit the illusion.

Specifically, we randomly sampled MLP weights and biases such that the norm of the output activations matches those of MLP8. As random MLPs might lead to nonsensical text generation, we don't replace the layer with the random weights, but rather train a subspace using DAS on the MLP activations, and add the difference between the patched and unpatched output of the random MLP to the real output of MLP8. This setup finds a subspace that reduces logit difference even more than the \mathbf{v}_{MLP} direction.

This suggests that the existence of the \mathbf{v}_{MLP} subspace is less about *what* information MLP8 contains, and more about *where* MLP8 is in the network.

B.6 Generalization to high-dimensional Subspaces

In the main text, we focus on activation patching in one-dimensional subspaces for clarity. Here, we extend the discussion to higher-dimensional subspaces and show that the interpretability illusion generalizes to high-dimensional linear subspaces.

We investigate two different 100-dimensional subspaces U_{MLP8} in MLP8 and U_{resid8} in the output of layer 8. Specifically, we used DAS to find orthonormal bases V_{MLP} and V_{resid} that align the position information in these two locations, as explained in A.1. We found that these subspaces performed slightly better compared to their 1-dimensional counterparts (for V_{resid} : 190% FLDD and 89% interchange accuracy; for V_{MLP} : 62% FLDD and 13% interchange accuracy).

We hypothesize that the subspace trained on MLP8 is pathological while the subspace in the residual stream is not. To test this, we decompose every basis vector $\mathbf{v}_{\{\text{MLP8}, \text{resid8}\}}^{(d)}$ into its projection $\mathbf{v}_{\{\text{MLP8}, \text{resid8}\}}^{\text{nullspace}}$ on the nullspace $\ker\{W_{\text{out}}, W_Q\}$ and its orthogonal complement $\mathbf{v}_{\{\text{MLP8}, \text{resid8}\}}^{\text{rowspace}}$ such that

$$\mathbf{v}_{\{\text{MLP8}, \text{resid8}\}}^{(d)} = \mathbf{v}_{\{\text{MLP8}, \text{resid8}\}}^{\text{nullspace}} + \mathbf{v}_{\{\text{MLP8}, \text{resid8}\}}^{\text{rowspace}}.$$

Note that W_Q denotes the query weight of name mover head 9.9. We then patched the 200-dimensional subspace spanned by \hat{V} with

$$\hat{V} = \mathcal{QR}([\mathbf{v}_1^{\text{nullspace}}, \dots, \mathbf{v}_d^{\text{nullspace}}, \mathbf{v}_1^{\text{rowspace}}, \dots, \mathbf{v}_d^{\text{rowspace}}])$$

composed out of the decomposed subspace vectors and orthonormalized using QR-decomposition (see Figure 16). For patching the output of layer 8, FLDD and interchange accuracy remained similarly high. (FLDD: 200%, interchange accuracy: 86%). However, patching on MLP8 mostly removes the effect of the patch (FLDD: 17%, interchange accuracy: 2%). Thus, the causally disconnected subspace is required for patching MLP8 which suggests that the interpretability illusion generalizes to higher-dimensional subspaces.

B.7 Overfitting on Small Datasets

How important is a large and diverse dataset for training DAS? We initially hypothesized that for very small datasets, it is possible to find working subspaces in all layers as there are only a few fixed activation vectors in each layer and we might be able to find subspaces that utilize this noise to overfit.

To test this, we created a small IOI dataset containing only two names from a fixed template. We fitted a one-dimensional subspace using DAS for every layer on that dataset and the full dataset as a control (Figure 17). We repeated the experiment for subspaces in the MLP and residual stream and evaluated the subspaces on their train distribution and a test distribution containing all names and templates. FLDD was highest in layer 8, the component between S-inhibition heads and name movers, and also high in neighboring layers that still contain IOI information (e.g. some of the S-inhibition heads are in layer 7 and some of the name movers are in layer 10). Moreover, train FLDD was significantly higher than test FLDD when trained on only 2 names.

Importantly, we also observe that subspaces optimized on the small dataset reached a FLDD bigger than zero in some of the other layers but contrary to our expectation, this was neither high in absolute terms nor compared to subspaces trained on the full distribution (see Figure 17).

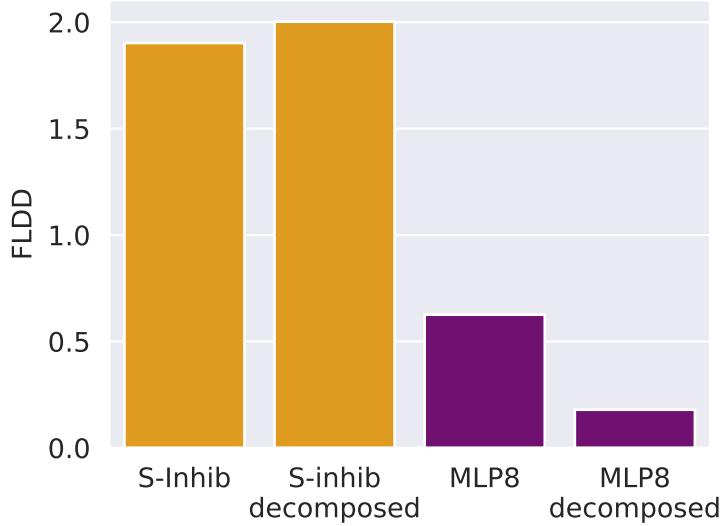


Figure 16: Fractional logit difference decrease (FLDD) for patching a 100-dimensional subspace on the S-inhibition heads or on MLP8; "decomposed" patches the 200-dimensional subspace made out of the nullspace projection vectors into W_Q of name mover or W_{out} of MLP8, respectively, and their orthogonal complements

C Additional Details for Section 5

Which model components write to the $\mathbf{v}_{\text{resid}}$ direction? To test how every attention head and MLP contributes to the value of projections on \mathbf{v}_{MLP} , we sampled activations from head and MLP outputs at the last token position of IOI prompts, and calculated their dot product with $\mathbf{v}_{\text{resid}}$ (Figure 20). We found that the dot products of most heads and MLPs was low, and that the S-inhibition heads were the only heads whose dot product differed between different patterns ABB and BAB. This shows that only the S-inhibition heads write to the $\mathbf{v}_{\text{resid}}$ direction (as one would hope). Importantly, this test separates $\mathbf{v}_{\text{resid}}$ from the interpretability illusion \mathbf{v}_{MLP} . While patching \mathbf{v}_{MLP8} also writes to $\mathbf{v}_{\text{resid8}}$ (i.e. $\mathbf{v}_{\text{MLP8}}W_{out} \approx \mathbf{v}_{\text{resid8}}$), the MLP layer does not write this subspace on the IOI task (see Figure 4). This further supports the observation that the \mathbf{v}_{MLP} patch activates a dormant pathway in the model.

Generalization beyond the IOI distribution. We also investigate how the subspace generalizes. We sample prompts from OpenWebText-10k and look at those with particularly high and low activations in $\mathbf{v}_{\text{sinhib}}$. Representative examples are shown in Figure 21 together with the name movers attention at the position of interest, how the probability changes after subspace ablation, and how the name movers attention changes.

Stability of found solution. Finally, we note that solutions found by DAS in the residual stream are stable, including when trained on a subset of S-inhibition heads (see Figure 18).

D Additional details for Section 6

D.1 Dataset construction and training details

We use the first 1000 examples from the COUNTERFACT dataset (Meng et al., 2022a). We filter the facts which GPT2-XL correctly recalls. Out of the remaining facts, for each relation we form all pairs

of distinct facts, and we sample 5 such pairs from each relation with at least 5 facts. This results in a collection of 40 fact pairs spanning 8 different relations. We then use these facts as follows:

- for the ROME experiments in Subsection 6.3, we define edits by requesting one of the facts in each pair to be rewritten with the object of the other fact;
- for the activation patching experiments in Subsection 6.1, we patch from the last token of s' in B to the last token of s in A (prior work has shown that the fact is retrieved on s (Geva et al., 2023)), and we again use DAS Geiger et al. (2023b) to optimize for a direction that maximizes the logit difference between o' and o .

D.2 Additional fact patching experiments

In figure 24, we show the distribution of the fractional logit difference metric (see Subsection 4.2 for a definition) when patching between facts as described in Subsection 6.1. Like in the related Figure 7, we observe that, while patching along the directions found by DAS achieves strongly negative values (indicating that the facts are very often successfully changed by the patch), the interventions that replace the entire MLP layer or only the causally relevant component of the DAS directions have no such effect.

Next, we observe that the nullspace component of the patching direction is the one similar to the variation in the inputs (difference of last-token activations at the two subjects). Specifically, in Figure 25, we plot the (absolute value of the) cosine similarity between the difference in activations for the two last subject tokens, and the nullspace component of the DAS direction. We note that this similarity is consistently significantly high (note that it can be at most 1, which would indicate perfect alignment).

Finally, we observe that the nullspace component of the patching direction is a non-trivial part of the direction in Figure 26, where we plot the distribution of the ℓ_2 norm of this component.

D.3 ROME implementation details

ROME takes as input a vector $\mathbf{k} \in \mathbb{R}^{d_{\text{MLP}}}$ representing the subject (e.g. an average of last-token representations of the subject) and a vector $\mathbf{v} \in \mathbb{R}^{d_{\text{resid}}}$ which, when output by the MLP layer, will cause the model to predict a new object for the factual prompt, but at the same time won't change other facts about the subject. ROME modifies the MLP weight by setting $W'_{\text{out}} = W_{\text{out}} + \mathbf{a}\mathbf{b}^\top$, where $\mathbf{a} \in \mathbb{R}^{d_{\text{resid}}}$, $\mathbf{b} \in \mathbb{R}^{d_{\text{MLP}}}$ are chosen so that $W'_{\text{out}}\mathbf{k} = \mathbf{v}$, and the MLP's output is otherwise minimally changed. Without loss of generality, the first condition implies that $\mathbf{a} = \mathbf{v} - W_{\text{out}}\mathbf{k}$ and $\mathbf{b}^\top \mathbf{k} = 1$; the second condition is then modeled by minimizing the variance of $\mathbf{b}^\top \mathbf{x}$ when $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ for an empirical estimate $\Sigma \in \mathbb{R}^{d_{\text{MLP}} \times d_{\text{MLP}}}$ of the covariance of MLP activations (see Lemma D.1 in Appendix D for details and a proof). In all our experiments involving ROME, we use GPT2-XL (Radford et al., 2019), and we use the precomputed values of Σ from Meng et al. (2022a) accessible online [here](#).

D.4 ROME as an Optimization Problem

We now review the ROME method from Meng et al. (2022a) and show how it can be characterized as the solution of a simple optimization problem. Following the terminology of 6.4, let us have an MLP layer with an output projection W_{out} , a key vector $\mathbf{k} \in \mathbb{R}^{d_{\text{MLP}}}$ and a value vector $\mathbf{v} \in \mathbb{R}^{d_{\text{resid}}}$.

In Meng et al. (2022a), equation 2, the formula for the rank-1 update to W_{out} is given by

$$W'_{out} = W_{out} + (\mathbf{v} - W_{out}\mathbf{k}) \frac{\mathbf{k}^\top \Sigma^{-1}}{\mathbf{k}^\top \Sigma^{-1} \mathbf{k}} \quad (6)$$

where Σ is an empirical estimate of the uncentered covariance of the pre- W_{out} activations. We derive the following equivalent characterization of this solution (which may be of independent interest):

Lemma D.1. *Given a matrix $W_{out} \in \mathbb{R}^{d_{resid} \times d_{MLP}}$, a key vector $\mathbf{k} \in \mathbb{R}^{d_{MLP}}$ and a value vector $\mathbf{v} \in \mathbb{R}^{d_{resid}}$, let $\Sigma \succ 0, \Sigma \in \mathbb{R}^{d_{MLP} \times d_{MLP}}$ be a positive definite matrix (specifically, the uncentered empirical covariance), and let $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ be a normally distributed random vector with mean 0 and covariance Σ . Then, the ROME weight update is $W'_{out} = W_{out} + \mathbf{ab}^\top$ where $\mathbf{a} \in \mathbb{R}^{d_{resid}}, \mathbf{b} \in \mathbb{R}^{d_{MLP}}$ solve the optimization problem*

$$\min_{\mathbf{a}, \mathbf{b}} \text{trace}(\text{Cov}_{\mathbf{x}} [W'_{out}\mathbf{x} - W_{out}\mathbf{x}]) \quad \text{subject to} \quad W'_{out}\mathbf{k} = \mathbf{v}.$$

In other words, the ROME update is the update that causes W_{out} to output \mathbf{v} on input \mathbf{k} , and minimizes the total variance of the extra contribution of the update in the output of the MLP layer under the assumption that the pre- W_{out} activations are normally distributed with covariance Σ .

Proof. We have $W'_{out}\mathbf{x} - W_{out}\mathbf{x} = \mathbf{ab}^\top \mathbf{x}$. Next, Using $\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top] = \Sigma$ and the cyclic property of the trace, we see that

$$\text{trace}(\text{Cov}_{\mathbf{x}} [W'_{out}\mathbf{x} - W_{out}\mathbf{x}]) = \|\mathbf{a}\|_2^2 \mathbf{b}^\top \Sigma \mathbf{b}$$

We must have $\mathbf{ab}^\top \mathbf{k} = \mathbf{v} - W_{out}\mathbf{k}$, so without loss of generality we can rescale \mathbf{a}, \mathbf{b} so that $\mathbf{a} = \mathbf{v} - W_{out}\mathbf{k}$. Then, we want to solve the problem

$$\min_{\mathbf{b}} \mathbf{b}^\top \Sigma \mathbf{b} \quad \text{subject to} \quad \mathbf{b}^\top \mathbf{k} = 1$$

which we can solve using Lagrange multipliers. The Lagrangian is

$$\mathcal{L}(\mathbf{b}, \lambda) = \frac{1}{2} \mathbf{b}^\top \Sigma \mathbf{b} - \lambda \mathbf{b}^\top \mathbf{k}$$

and the derivative w.r.t. \mathbf{b} is $\Sigma \mathbf{b} - \lambda \mathbf{k} = 0$, which tells us that \mathbf{b} is in the direction of $\Sigma^{-1} \mathbf{k}$. Then the constraint $\mathbf{b}^\top \mathbf{k} = 1$ forces the constant of proportionality, and we arrive at $\mathbf{b} = \frac{\mathbf{k}^\top \Sigma^{-1}}{\mathbf{k}^\top \Sigma^{-1} \mathbf{k}}$. \square

D.5 Connection between 1-dimensional activation patching and model editing

Lemma D.2. *Given prompts A and B, two token positions t_A, t_B , and an MLP layer with output projection weight $W_{out} \in \mathbb{R}^{d_{resid} \times d_{MLP}}$, let $u_A, u_B \in \mathbb{R}^{d_{MLP}}$ be the respective (post-nonlinearity) activations at these token positions in this layer. If v is a direction in the activation space of the MLP layer, then there exists a ROME edit $W'_{out} = W_{out} + ab^\top$ such that the activation patch from u_B into u_A along v and the edit result in equal outputs of the MLP layer at token t_A when run on prompt A. Moreover, the ROME edit is given by*

$$a = \left((u_B - u_A)^\top v \right) W_{out} v \quad \text{and any } b \text{ that satisfies } b^\top u_A = 1.$$

Choosing $b = \frac{\Sigma^{-1} u_A}{u_A^\top \Sigma^{-1} u_A}$ minimizes the change to the model (in the sense of Meng et al. (2022a)) over all such rank-1 edits.

Proof. The activation after patching from B into A along v is $u'_A = u_A + ((u_B - u_A)^\top v)v$, which means that the change in the output of the MLP layer at this token will be

$$W_{out}u'_A - W_{out}u_A = ((u_B - u_A)^\top v)W_{out}v$$

The change introduced by a fact edit at this token is

$$W'_{out}u_A - W_{out}u_A = ab^\top u_A = \left(b^\top u_A\right) \left((u_B - u_A)^\top v\right) W_{out}v$$

and the two are equal because $b^\top u_A = 1$.

To find the b that minimizes the change to the model, we minimize the variance of $b^\top x$ when $x \sim \mathcal{N}(0, \Sigma)$ subject to $b^\top u_A = 1$. The variance is equal to $b^\top \Sigma b$, so we have a constrained (convex) minimization problem

$$\min \frac{1}{2} b^\top \Sigma b \quad \text{subject to} \quad b^\top u_A = 1$$

The rest of the proof is the same as in Lemma D.1. Namely, we can solve this optimization problem using Lagrange multiplies. The Lagrangian is

$$\mathcal{L}(b, \lambda) = \frac{1}{2} b^\top \Sigma b - \lambda b^\top u_A$$

and the derivative w.r.t. b is $\Sigma b - \lambda u_A = 0$, which tells us that b is in the direction of $\Sigma^{-1}u_A$. Then the constraint $b^\top u_A = 1$ forces the constant of proportionality. \square

D.6 Additional experiments comparing fact patching and rank-1 editing

In Figure 27, we plot the distributions of the logit difference between the correct object for a fact and the object we are trying to substitute when patching the 1-dimensional subspaces found by DAS, and performing the equivalent rank-1 weight edit according to Lemma D.2. We observe that the two metrics quite closely track each other, indicating that the additional effects of using a weight edit (as opposed to only intervening at a single token) are negligible.

Similarly, in Figure 28, we show the success rate of the the two methods in terms of making the model output the object of the fact we are patching from. Again, we observe that they quite closely track each other.

D.7 From Rank-1 Model Edits to Subspace Interventions

In this section, we describe how, given a rank-1 edit $W'_{out} = W_{out} + ab^\top$, to obtain a direction $v \in \mathbb{R}^{d_{MLP}}$ such that intervening on the model by setting the projection on v to some constant $c \in \mathbb{R}$ (at each token) is approximately equivalent to intervening via the rank-1 edit.

Specifically, given an activation $x \in \mathbb{R}^{d_{MLP}}$, the patched activation is $x' = x + (c - v^\top x)v$ and the extra contribution of the subspace intervention to the output of the MLP layer will be

$$\text{contrib}_{\text{subspace}}(x) = W_{out}x' - W_{out}x = \left(c - v^\top x\right) Wv.$$

Similarly, the extra contribution of the rank-1 edit to the output of the MLP layer is

$$\text{contrib}_{\text{rank-1}}(x) = W'_{out}x - W_{out}x = \left(b^\top x\right) a.$$

Recall (see Appendix D.4) that the ROME method (Meng et al., 2022a) implicitly treats the activation x as a random vector sampled from $\mathcal{N}(0, \Sigma)$ where Σ is an empirical estimate of the covariance. In particular, this distribution is used to quantify the amount to which a rank-1 edit changes the model.

Motivated by this, we formalize approximating the rank-1 edit by the subspace intervention using the following criteria analogous to the ROME method:

- $\mathbb{E}_{x \sim \mathcal{N}(0, \Sigma)} [\text{contrib}_{\text{subspace}}(x)] = \mathbb{E}_{x \sim \mathcal{N}(0, \Sigma)} [\text{contrib}_{\text{rank-1}}(x)]$, i.e. the interventions have the same expectation;
- $W_{out}v \parallel a$, i.e. the interventions point in the same direction;
- $\text{trace}(\text{Cov}_x [\text{contrib}_{\text{subspace}}(x) - \text{contrib}_{\text{rank-1}}(x)])$ is minimized, i.e. the two interventions are maximally similar with respect to the activation distribution modeled as $x \sim \mathcal{N}(0, \Sigma)$ (this is the criterion used by ROME; recall D.4).

The expectation of $\text{contrib}_{\text{rank-1}}(x)$ is zero, while the expectation of $\text{contrib}_{\text{subspace}}(x)$ is $cW_{out}v$, and since $W_{out}v = 0$ would lead to a trivial intervention, we must have

$$c = 0.$$

Next, to ensure $W_{out}v \parallel a$, we have to pick $v = \alpha W_{out}^+ a + u$ where $u \in \ker W_{out}$. With this, the covariance minimization can then be written as

$$\min_{\alpha, v} \|a\|_2^2 (b + \alpha v)^T \Sigma (b + \alpha v)$$

(this is a similar derivation to the one in Appendix D.4). After removing constant terms and setting $w = u/\alpha$, we are left with

$$\min_{\alpha, w} \left[\alpha^4 (W_{out}^+ a + w)^T \Sigma (W_{out}^+ a + w) + 2\alpha^2 b^T \Sigma (W_{out}^+ a + w) \right].$$

subject to $W_{out}w = 0$. The Lagrangian is

$$\mathcal{L}(\alpha, w, \lambda) = \alpha^4 (W_{out}^+ a + w)^T \Sigma (W_{out}^+ a + w) + 2\alpha^2 b^T \Sigma (W_{out}^+ a + w) + \lambda^T W_{out}w$$

with the first-order conditions

$$\frac{\partial \mathcal{L}}{\partial w} = 2\alpha^4 \Sigma (W_{out}^+ a + w) + 2\alpha^2 \Sigma b + W_{out}^T \lambda = 0$$

and $\partial \mathcal{L}/\partial \lambda = W_{out}w = 0$. Multiplying the w derivative with $W_{out}\Sigma^{-1}$ on the left gives us a linear system for λ :

$$W_{out}\Sigma^{-1}W_{out}^T \lambda = -2\alpha^2 W_{out}b - 2\alpha^4 a,$$

which can be solved assuming we know α , and then substituting λ in $\frac{\partial \mathcal{L}}{\partial w} = 0$ gives us w . In practice, we guess several values for α (typically, $\alpha^2 = 0.05$ performs best) and pick the one resulting in the best value for the objective.

E Additional Details for Section 7

E.1 Prevalence of Causal Directions in MLP Layers

Given an MLP activation \mathbf{x} and a vector $\mathbf{u} \in \mathbb{R}^{d_{MLP}}$, changing the projection of \mathbf{x} on \mathbf{u} means replacing \mathbf{x} with the new activation $\mathbf{x}' = \mathbf{x} + \alpha\mathbf{u}$ for some $\alpha \in \mathbb{R}$. This translates to the new output of the MLP layer being

$$W_{out}(\mathbf{x} + \alpha\mathbf{u}) = W_{out}\mathbf{x} + \alpha W_{out}\mathbf{u}.$$

Under our assumptions, the direction \mathbf{u} will be causally relevant if the extra contribution to the residual stream $\alpha W_{out}\mathbf{u}$ points along \mathbf{v} ; thus it suffices to find a \mathbf{u} such that $W_{out}\mathbf{u} \parallel \mathbf{v}$.

As it turns out, we can simply choose $\mathbf{u} = W_{out}^+\mathbf{v}$. Indeed, we empirically observe that $W_{out} \in \mathbb{R}^{d_{resid} \times d_{MLP}}$ is a full-rank matrix⁹, with almost all singular values bounded well away from 0 (see Appendix E.5). Since $d_{MLP} > d_{resid}$, it follows that $W_{out}W_{out}^+\mathbf{v} = \mathbf{v}$. This establishes that \mathbf{u} is a causal direction.

E.2 Prevalence of Directions Discriminating for C in MLP layers

For a feature $\mathbf{u} \in \mathbb{R}^{d_{MLP}}$ to discriminate between values of C , we need projections of the post-nonlinearity activations on \mathbf{u} to linearly separate examples according to the values of C . By assumption, \mathbf{v} is a good linear separator for the values of C in the residual stream. We can thus frame our goal as a more general question:

If two sets of activations are linearly separable in the residual stream, are their images after the non-linearity also (approximately) linearly separable?

The transformation from residual vectors $\mathbf{x} \in \mathbb{R}^{d_{resid}}$ to post-nonlinearity activations is given by the steps

$$\mathbf{x} \mapsto \text{LayerNorm}(W_{in}\mathbf{x}) \mapsto \text{gelu}(\text{LayerNorm}(W_{in}\mathbf{x}))$$

The composition of LayerNorm and W_{in} is approximately a linear operation (Elhage et al., 2021), so the values of the concept C are also linearly separated in the pre-gelu activations. However, it is not a priori clear if the gelu operation (approximately) preserves linear separability.

We show ample empirical evidence in Appendix E.4 that this transformation approximately preserves the Euclidean geometry of activations in a certain restricted sense; then, we prove that this preservation implies that points remain approximately linearly separable after this transformation. We further argue this empirically in Appendix E.6, where we show that linear separability is approximately preserved in MLP activations for random directions \mathbf{v} in the residual stream.

E.3 Empirical Analysis of Distortion Introduced by the Non-linearity

Methodology. We use the first 10K texts of OpenWebText dataset (Gokaslan & Cohen, 2019). Each of these texts contains 1,024 tokens; we pass each text through GPT-2 Small, and for each layer

⁹This is also heuristically plausible: models want to maximize their expressive capacity, and pre-training datasets are very complex, so making W_{out} low-rank would not be preferred by optimization.

collect the pre-gelu activations \mathbf{x}_i of the MLP layer, as well as the values $\mathbf{z}_i = \text{proj}_{\ker W_{out}}(\text{gelu}(\mathbf{x}_i))$. We sample 250 quadruples of distinct i, j, k, l per text, and compute the values

$$\begin{aligned} a_{ijkl} &= (\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_k - \mathbf{x}_l) \\ b_{ijkl} &= (\mathbf{z}_i - \mathbf{z}_j)^\top (\mathbf{z}_k - \mathbf{z}_l) \end{aligned}$$

We collect these numbers across the first 1000 texts out of the first 10K, resulting in 250K datapoints per layer, and perform linear regression of b_i against a_i .

We note that there is some inherent linearity in the quantities a_{ijkl}, b_{ijkl} that could in principle skew the results of the linear regression towards a higher r^2 statistic in the presence of enough samples. In particular, there are linear dependencies of the form

$$a_{ijkl} = a_{ijkp} + a_{ijpl}$$

for any p , and similarly for the b_{ijkl} quantities. This makes these quantities potentially misleading targets for linear regression. However, in our regime, we sample 250 4-element subsets from the set $\{1, \dots, 1024\}$, and the probability of sampling quadruples that are linearly related is quite small.

Results. We find that the coefficients of determination r^2 for the linear regression are consistently high (≈ 0.8 or higher) for all layers except for layer 0, indicating a high degree of fit. The r^2 values are given in Table TODO, and regression plots are shown in Figure TODO. We also remark that the intercept coefficients are ≈ 0 relative to the standard deviation in the dependent variable.

E.4 Details for Subsection E.2

We will show the stronger property that activations remain approximately linearly separable even after projecting on the kernel of W_{out} . Define the function

$$f : \mathbf{x}' \mapsto \text{gelu}(\mathbf{x}') \mapsto \text{proj}_{\ker W_{out}}(\text{gelu}(\mathbf{x}'))$$

where $\text{proj}_{\ker W_{out}}$ is orthogonal projection on the kernel of W_{out} .

To overcome the non-linearity of f , we establish an empirical property of f on activations from the model’s pre-training distribution. Specifically, we show that f approximately preserves the Euclidean geometry of activations in a certain restricted sense:

$$(f(\mathbf{x}_i) - f(\mathbf{x}_j))^\top (f(\mathbf{x}_k) - f(\mathbf{x}_l)) \approx \lambda (\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_k - \mathbf{x}_l) + \eta \quad (7)$$

for $\lambda > 0$ and $\eta \approx 0$ (relative to the standard deviation of the expression on the left-hand side of the approximation). Specifically, we perform linear regression of $(f(\mathbf{x}_i) - f(\mathbf{x}_j))^\top (f(\mathbf{x}_k) - f(\mathbf{x}_l))$ using $(\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_k - \mathbf{x}_l)$ as the predictor variable in Appendix E.3, and find very high coefficients of determination ($r^2 \approx 0.8$) in all layers except for layer 0. To avoid relying solely on the coefficient of determination, we also generate regression plots for the data; see Figure 29 for regression lines over samples of 10^4 points from each layer of GPT2-Small.

Finally, we prove that f maintains linear separability if we assume that Equation 7 holds exactly with $\eta = 0$:

Lemma E.1. *Let $x_i \in \mathbb{R}^d, 1 \leq i \leq n$ be linearly separable with respect to binary labels $y_i \in \{-1, 1\}$. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a transformation with the property that*

$$(f(x_i) - f(x_j))^\top (f(x_k) - f(x_l)) = \lambda (x_i - x_j)^\top (x_k - x_l)$$

for all distinct i, j, k, l and some $\lambda > 0$. Then, $f(x_i)$ are linearly separable with respect to the labels y_i as well.

Proof. Consider the hard SVM objective for the points (x_i, y_i) ,

$$\min_{w,b} \frac{1}{2} \|w\|_2^2 \quad \text{subject to} \quad y_i (w^\top x_i + b) \geq 1.$$

Since the examples are linearly separable, we know that the minimizer (w^*, b^*) exists and satisfies all constraints. Furthermore, from the optimality conditions of the dual formulation of the objective we know that we can write $w^* = \sum_i \alpha_i x_i$ where $\sum_i \alpha_i = 0$ (see for example Awad et al. (2015)). Let $S = \{s_1, \dots, s_t\} \subset \{1, \dots, n\}$ be the support of α , i.e. the indices s_j such that $\alpha_{s_j} \neq 0$. Since $\sum_i \alpha_i = 0$, we can rewrite w^* as

$$w^* = \sum_j \beta_j (x_{s_j} - x_{s_{j+1}})$$

with indices modulo $|S|$. Since (w^*, b^*) is a separating hyperplane for (x_i, y_i) , we have

$$\begin{aligned} (w^*)^\top x_i &\geq 1 - b & \text{when } y_i = 1 \\ (w^*)^\top x_i &\leq -1 - b & \text{when } y_i = -1 \end{aligned}$$

and thus

$$(w^*)^\top (x_i - x_j) \geq 2 \quad \text{when } y_i = 1, y_j = -1.$$

Using the expansion of w^* as a linear combination of differences between examples, this says

$$\sum_j \beta_j (x_{s_j} - x_{s_{j+1}})^\top (x_i - x_j) \geq 2 \quad \text{when } y_i = 1, y_j = -1.$$

and thus

$$\sum_j \beta_j (f(x_{s_j}) - f(x_{s_{j+1}}))^\top (f(x_i) - f(x_j)) \geq 2\lambda > 0 \quad \text{when } y_i = 1, y_j = -1.$$

Now we claim that

$$\hat{w} = \sum_j \beta_j (f(x_{s_j}) - f(x_{s_{j+1}}))$$

is a linear separator for $(f(x_i), y_i)$ for some bias to be determined later. Indeed, let $M = \min_{y_i=1} \hat{w}^\top f(x_i)$ and $m = \max_{y_i=-1} \hat{w}^\top f(x_i)$. Then we have $M - m \geq 2\lambda > 0$. Choosing any $\hat{b} \in (m, M)$, we have

$$\begin{aligned} \hat{w}^\top f(x_i) - \hat{b} &\geq M - \hat{b} > 0 & \text{when } y_i = 1 \\ \hat{w}^\top f(x_i) - \hat{b} &\leq m - \hat{b} < 0 & \text{when } y_i = -1 \end{aligned}$$

which shows that (\hat{w}, \hat{b}) linearly separates the points $(f(x_i), y_i)$. □

E.5 MLP weights are full-rank matrices

In figure 30, we plot the 100 smallest singular values of the MLP weights in GPT2-Small for all 12 layers. We observe that they the vast majority are bounded well away from 0. This confirms that both MLP weights are full-rank transformations.

E.6 Features in the residual stream propagate to hidden MLP activations

Intuition. Suppose we have two classes of examples that are linearly separable in the residual stream. The transformation from the residual stream to the hidden MLP activations is a linear map followed by a nonlinearity, specifically $x \mapsto \text{gelu}(W_{in}x)$. As we observed in E.5, the W_{in} matrix is full-rank, meaning that all the information linearly present in x will also be so in $W_{in}x$. Even better, since W_{in} maps x from a d_{resid} -dimensional space to a $d_{\text{MLP}} = 4d_{\text{resid}}$ -dimensional space, this should intuitively make it much easier to linearly separate the points, because in a higher-dimensional space there are many more linear separators. On the other hand, the non-linearity has an opposite effect: by compressing the space of activations, it makes it harder for points to be separable. So it is a priori unclear which intuition is decisive.

Empirical validation. However, it turns out that empirically this is not such a problem. To test this, we run the model GPT2-Small on random samples from its data distribution (we used OpenWebText-10k), and extract 2000 activations of an MLP-layer after the non-linearity. We train a linear regression with ℓ_2 -regularization to recover the dot product of the residual stream immediately before the MLP-layer of interest and a randomly chosen direction. We repeat this experiment with different random vectors and for each layer. We observe that all regressions are better than chance and explain a significant amount of variance on the held-out test set ($R^2 = 0.71 \pm 0.17$, $\text{MSE} = 0.31 \pm 0.18$, $p < 0.005$). Results are shown in Figure 31 (right) (every marker corresponds to one regression model using a different random direction).

The position information in the IOI task is really a binary feature, so we are also interested in whether *binary* information in general is linearly recoverable from the MLP activations. To test this, we sample activations from the model run on randomly-sampled prompts. This time however, we add or subtract a multiple of a random direction v to the residual stream activation u , and calculate the MLP activations using this new residual stream vector u' :

$$u' = u + y \times z \times \|u\|_2 \times v$$

where $y \in \{-1, 1\}$ is uniformly random, z is a scaling factor we manipulate, and v is a randomly chosen direction of unit norm. For each classifier, we randomly sample a direction v that we either add or subtract (using y) from the residual stream. The classifier is trained to predict y . We rescale v to match the average norm of a residual vector and then scale it with a small scalar z .

Then, a logistic classifier is trained on 1600 samples. Again, we repeat this experiment for different v and z , and for each layer. We observe that the classifier works quite well across layers even with very small values of z (still, accuracy drops for $z = 0.0001$). Results are shown in Figure 31 (right), and Table 2.

Table 2: Mean Accuracy for Different Values of z

z	Mean Accuracy
0.0001	0.69
0.001	0.83
0.01	0.87
0.1	0.996

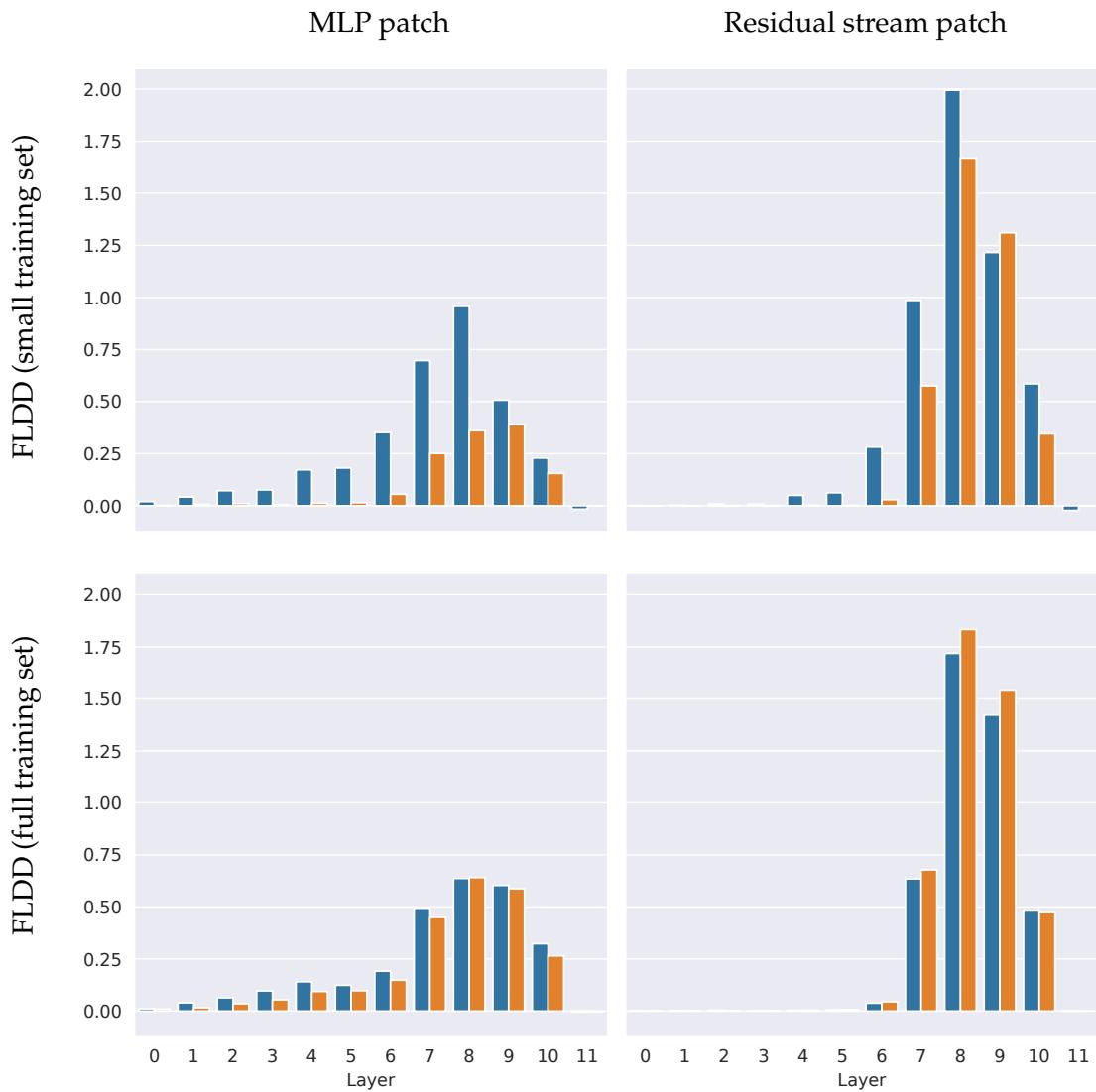


Figure 17: FLDD for different IOI-position subspaces: Subspaces were fitted to either a small version of the IOI dataset that only contained 2 names (first row) or on the full dataset (second row) using activations from the MLP (first column) or the residual stream (second column). Subspace performance on the IOI task was evaluated on the training (blue) distribution and the full test dataset containing all names (orange)



Figure 18: Cosine Similarity between learned position subspaces in the S-inhibition heads is high even when using only a subset of S-inhibition heads for training

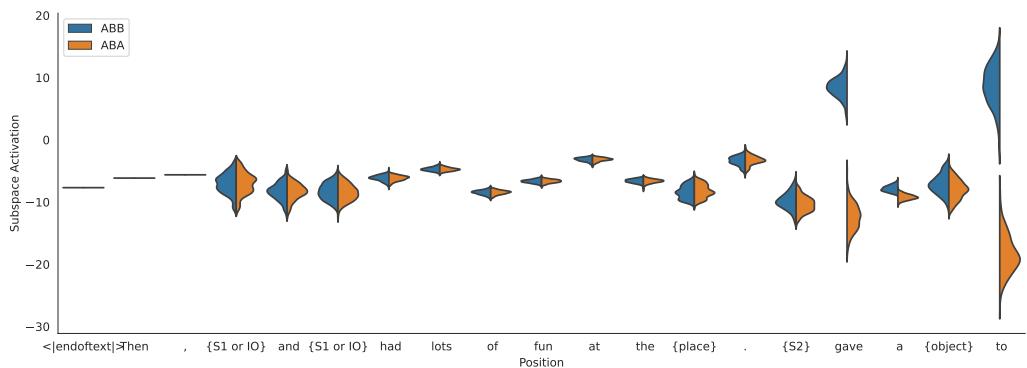


Figure 19: The IOI position subspace activates at words that predict a repeated name. S-inhibition subspace activations for different IOI prompts per position

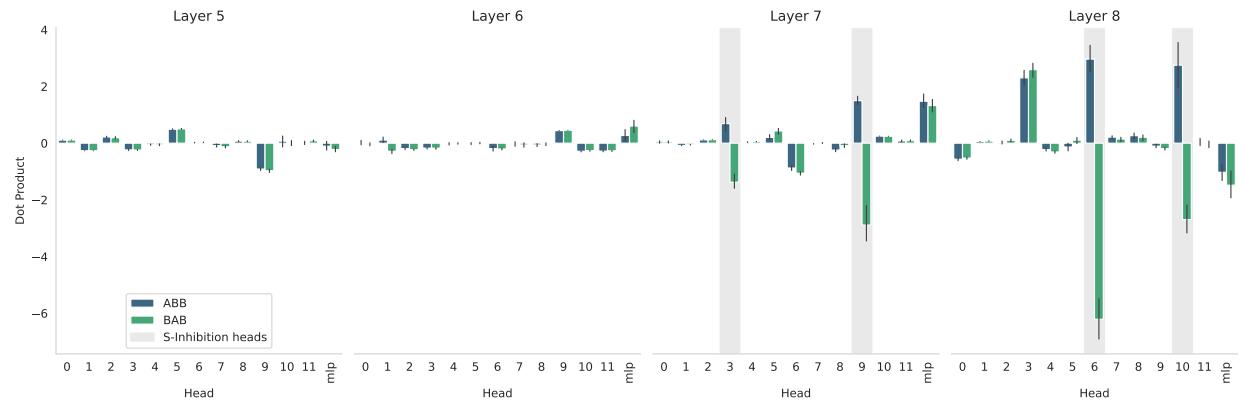
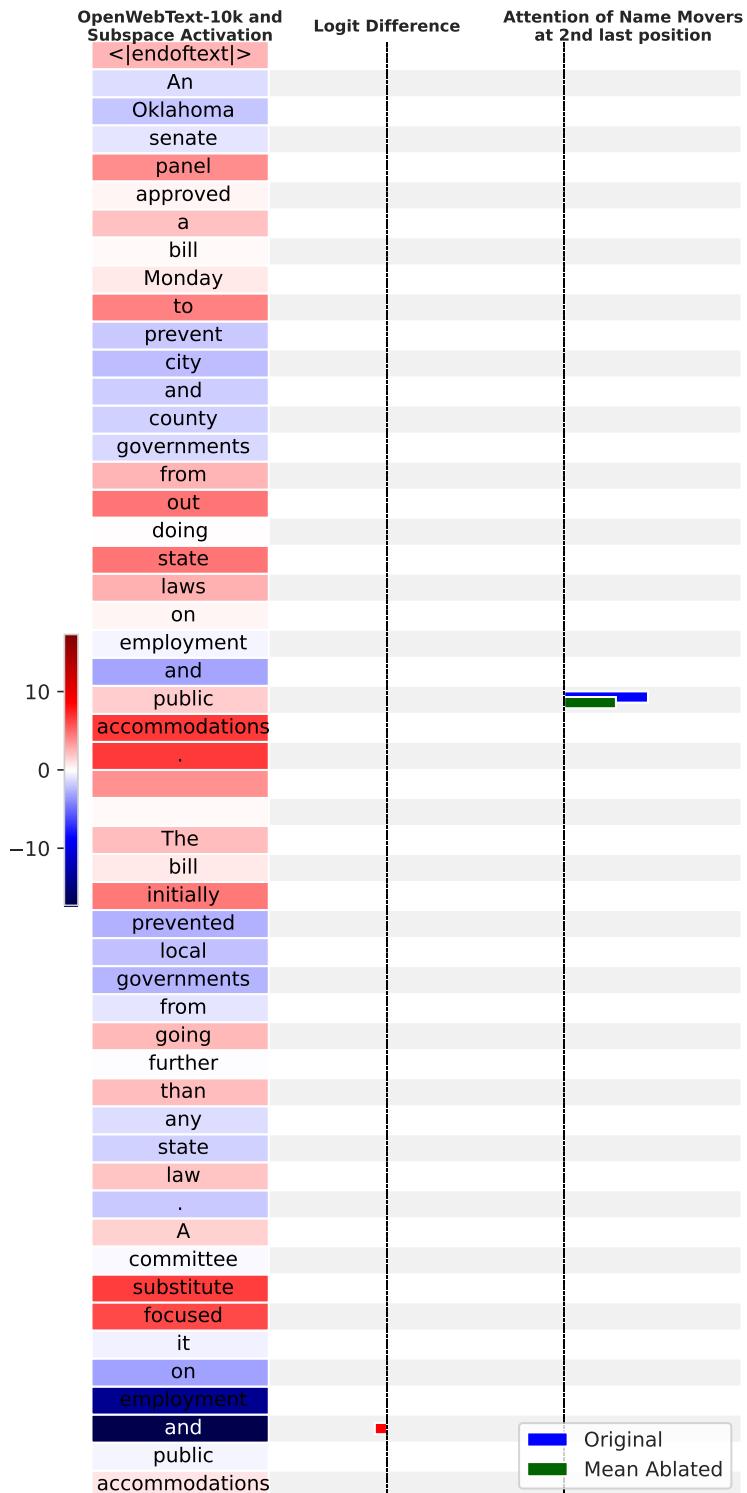
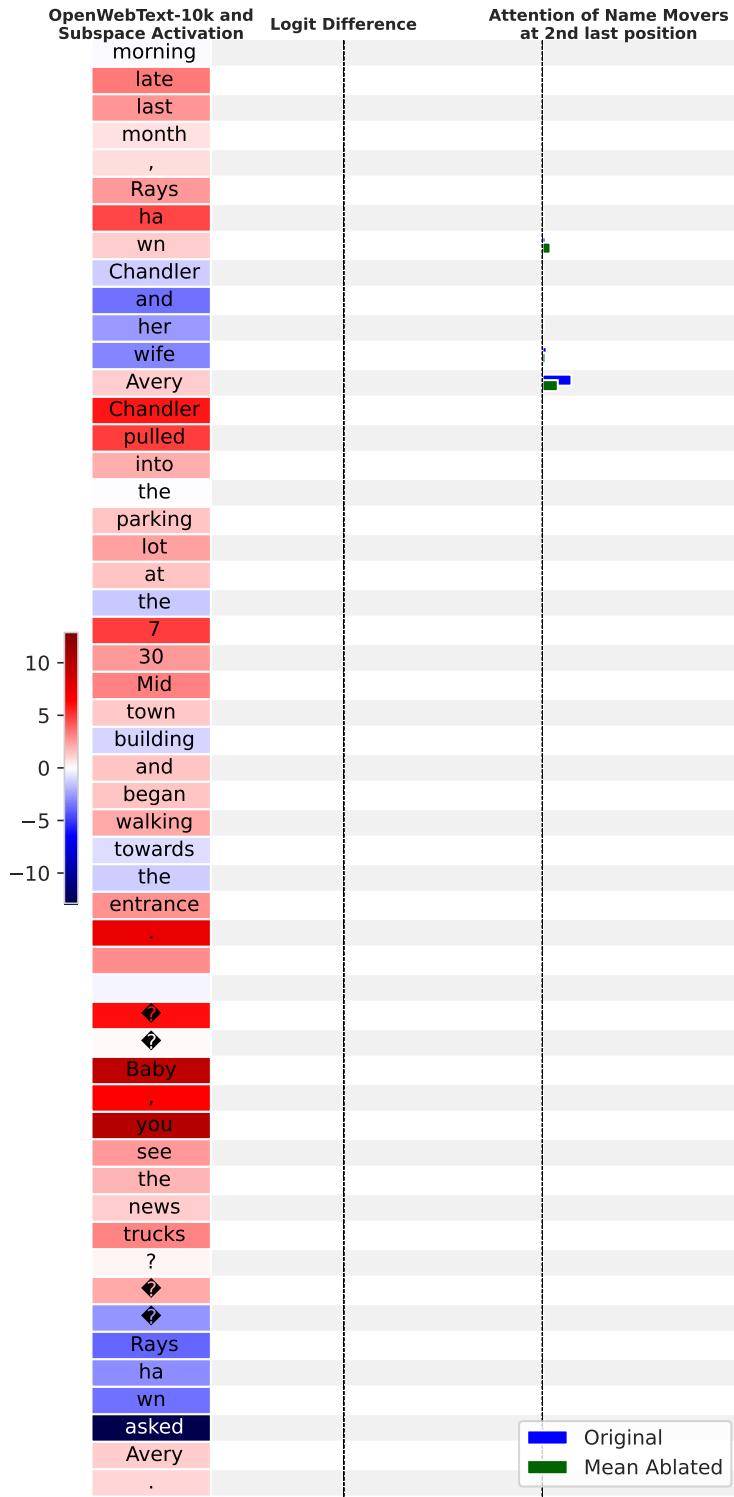
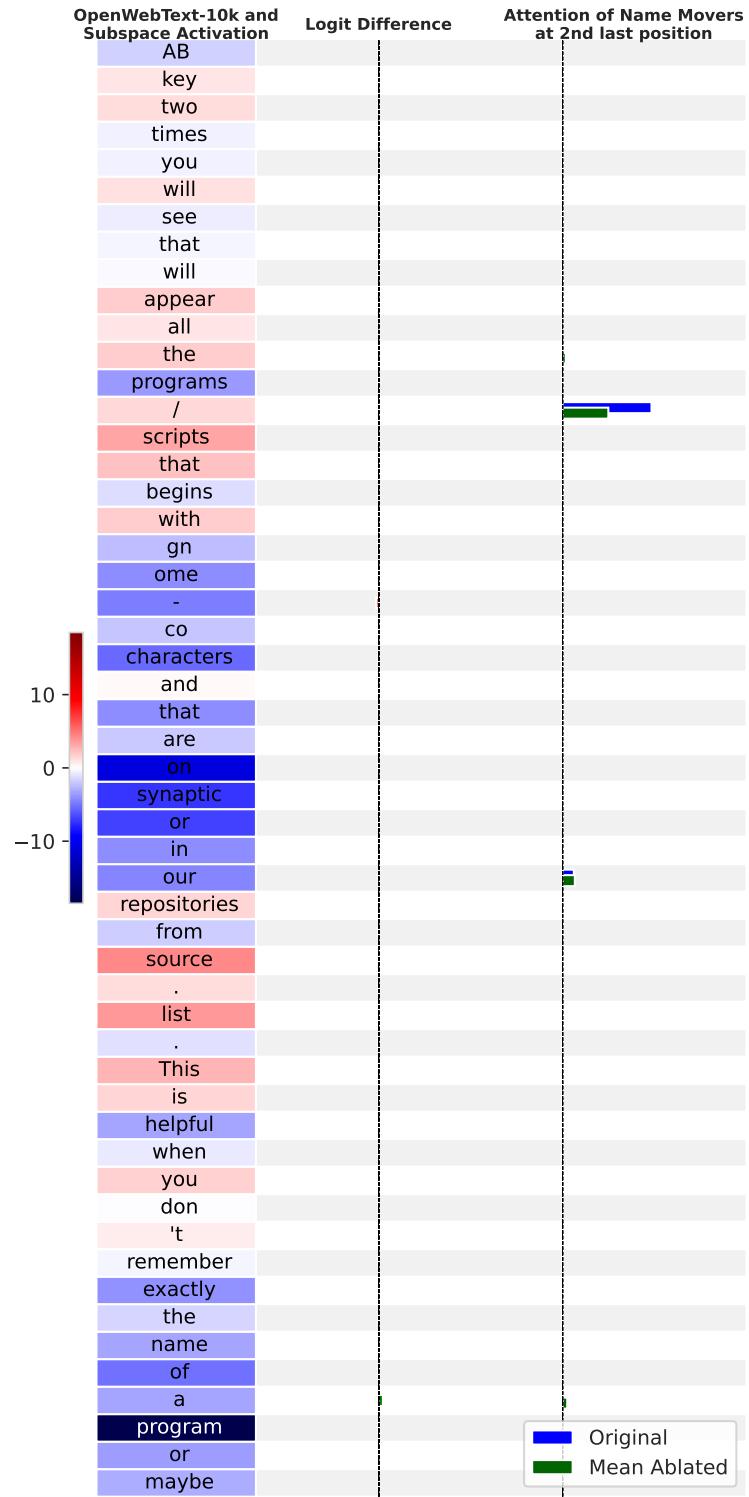


Figure 20: S-Inhibition heads but not MLP8 write to the position subspace in the residual stream that is causally connected to the name movers on the IOI task







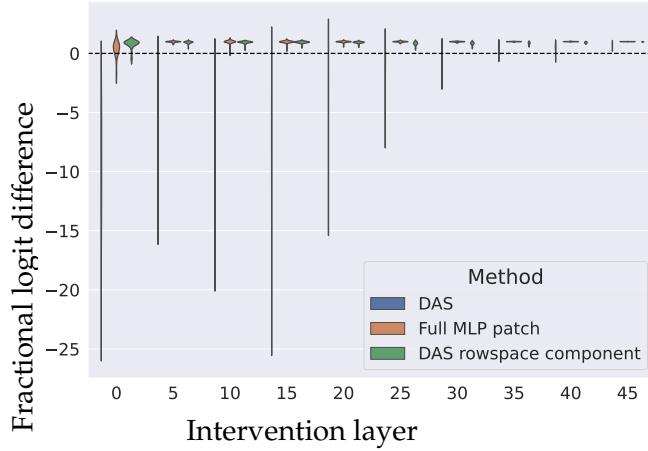


Figure 24: Fractional logit difference distributions under three interventions: patching along the direction found by DAS (blue), patching the component of the DAS direction in the rowspace of W_{out} (green), and patching the entire hidden MLP activation (orange).

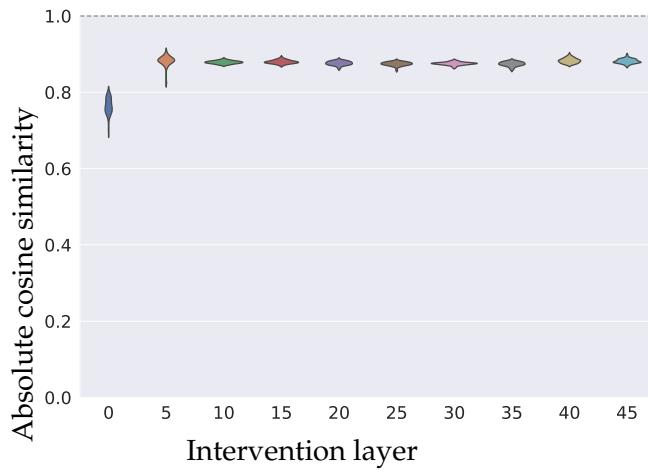


Figure 25: Distribution of the absolute value of the cosine similarity between the nullspace component of the DAS fact patching directions and the difference in activations of the last tokens of the two subjects.

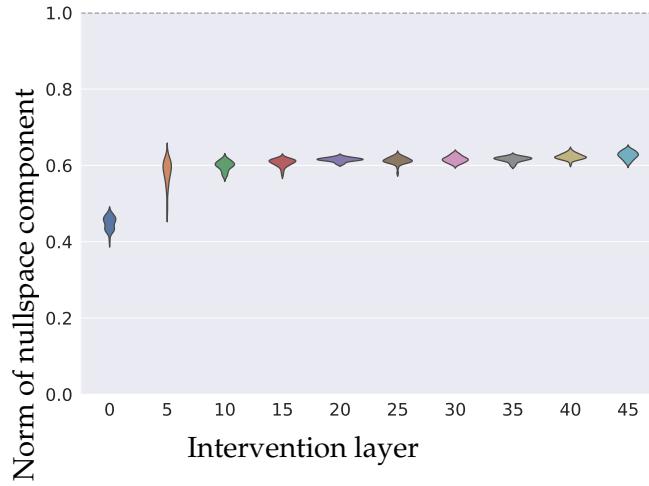


Figure 26: Distribution of the norm of the nullspace component of the DAS direction across intervention layers.

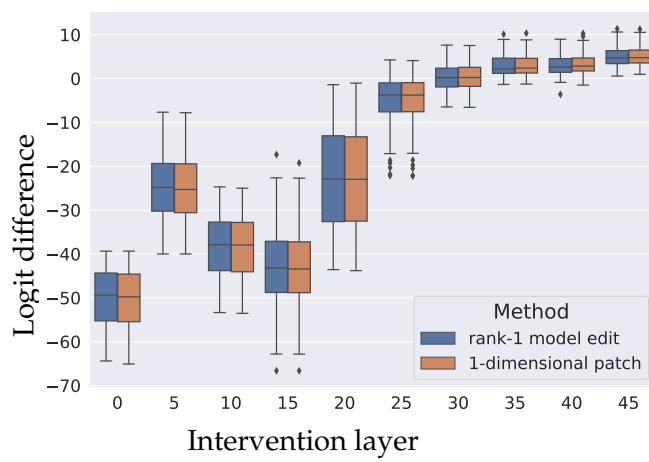


Figure 27: Comparison of logit difference between 1-dimensional fact patches and their derived rank-1 model edits

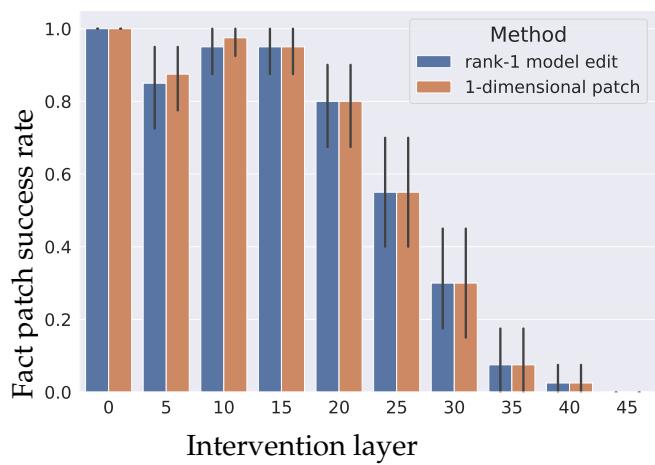


Figure 28: Comparison of fact editing success rate between 1-dimensional fact patches and their derived rank-1 model edits

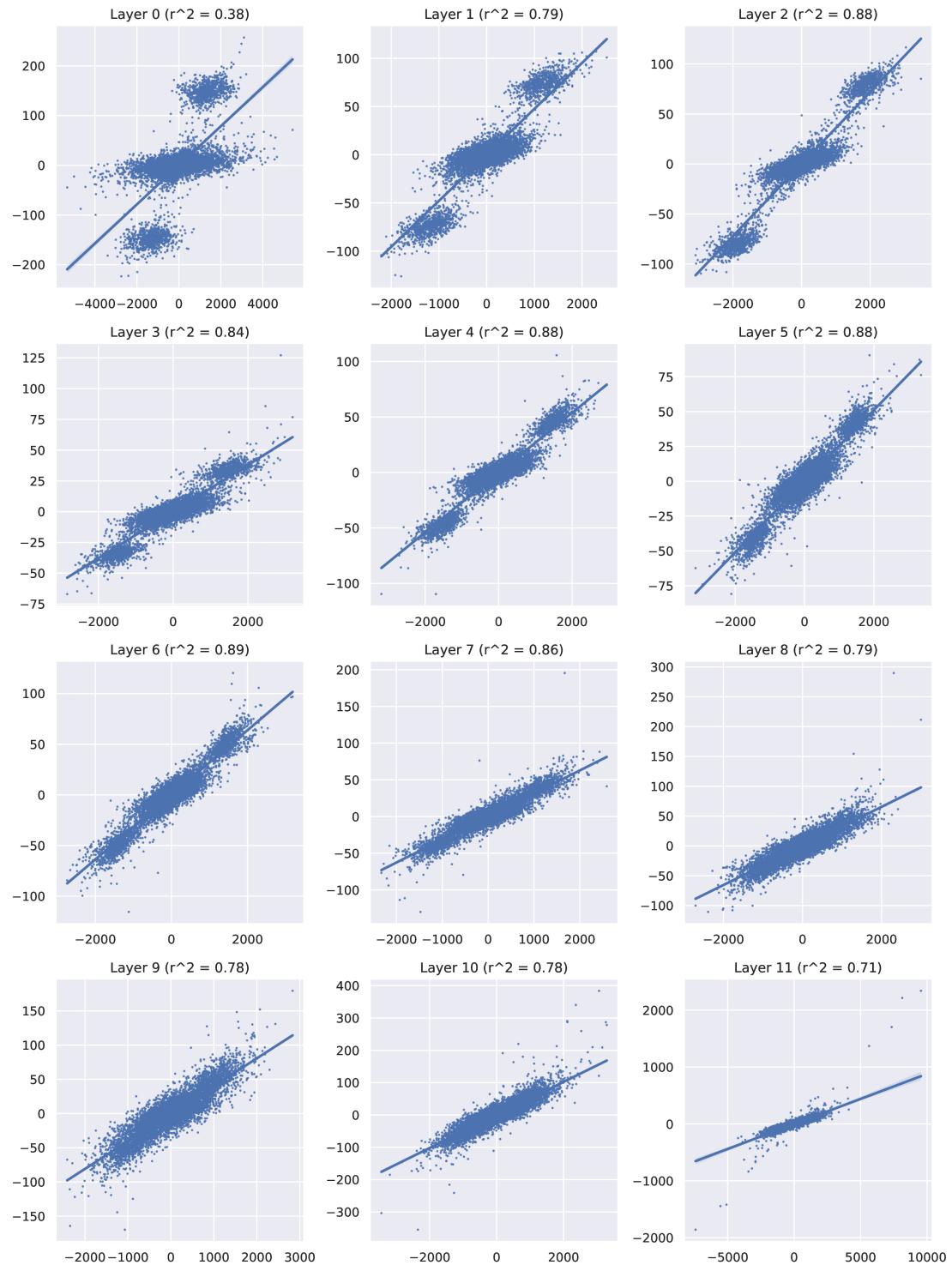


Figure 29: Regression plots accompanying the experiments in Appendix E.4.

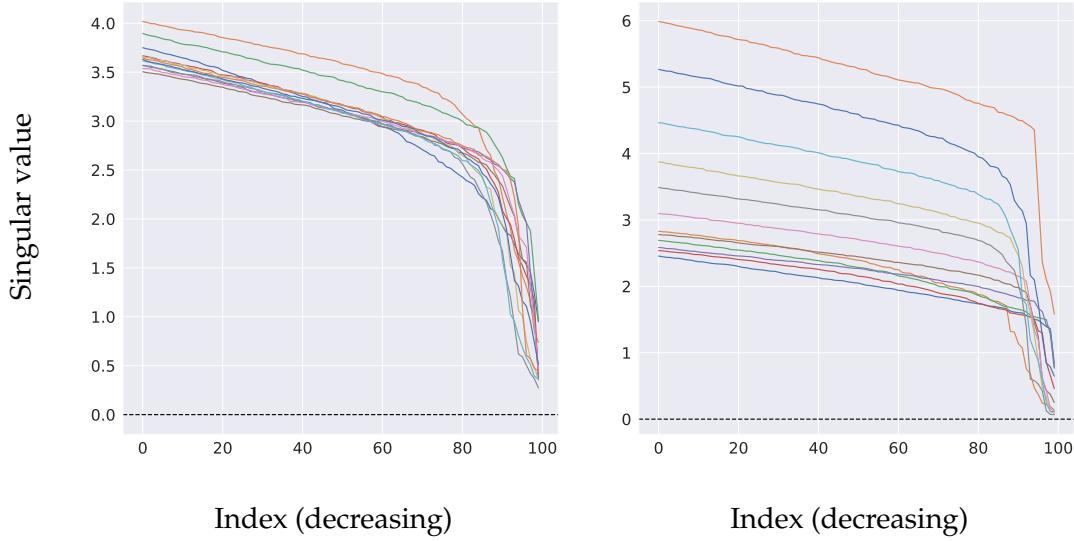


Figure 30: Smallest 100 singular values of the W_{in} (left) and W_{out} (right) MLP weights by layer in GPT2-Small

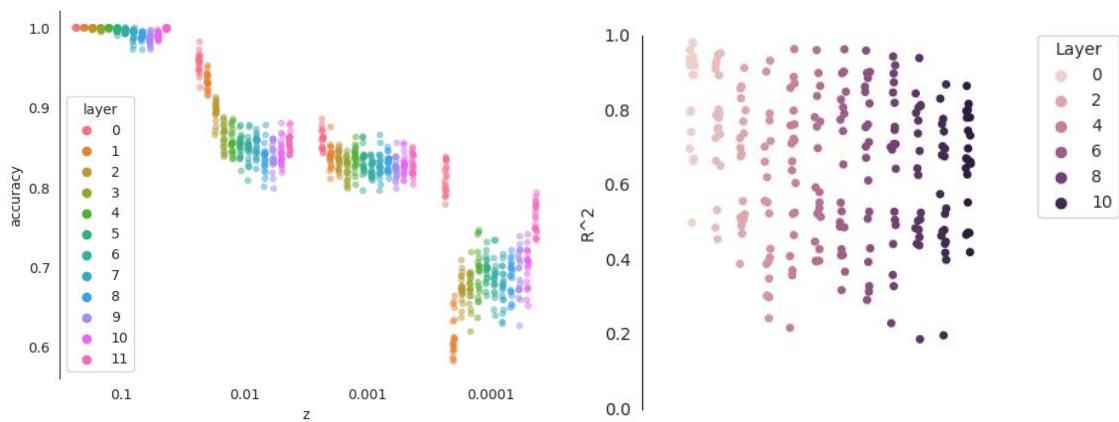


Figure 31: Recovering residual stream features linearly from hidden MLP activations: classification (left) and regression (right).