

A REPLY TO MAKELOV ET AL. (2023)’S “INTERPRETABILITY ILLUSION” ARGUMENTS

Zhengxuan Wu* Atticus Geiger[◊] Jing Huang* Aryaman Arora*

Thomas Icard* Christopher Potts* Noah D. Goodman*

*Stanford University [◊]Pr(Ai)²R Group

{wuzhengx, atticusg, hij, aryamana, icard, cgpotts, ngoodman}@stanford.edu

ABSTRACT

We respond to the recent paper by Makelov et al. (2023), which reviews subspace interchange intervention methods like distributed alignment search (DAS; Geiger et al. 2023) and claims that these methods potentially cause “interpretability illusions”. We first review Makelov et al. (2023)’s technical notion of what an “interpretability illusion” is, and then we show that even intuitive and desirable explanations can qualify as illusions in this sense. As a result, their method of discovering “illusions” can reject explanations they consider “non-illusory”. We then argue that the illusions Makelov et al. (2023) see in practice are artifacts of their training and evaluation paradigms. We close by emphasizing that, though we disagree with their core characterization, Makelov et al. (2023)’s examples and discussion have undoubtedly pushed the field of interpretability forward.

1 INTRODUCTION

It has long been known that individual neurons in trained neural networks often play multiple roles (Smolensky, 1986; Rumelhart et al., 1986; McClelland et al., 1986), giving rise to what Smolensky (1986) calls “patterns”: distributed representations defined by linear combinations of unit vectors. In the recent wave of interpretability research, such distributed representations have once again entered the limelight under the guise of “polysemantic neurons” (Olah et al., 2020; Bricken et al., 2023). Furthermore, the widely discussed “linear representation hypothesis” (Mikolov et al., 2013; Elhage et al., 2022; Nanda et al., 2023; Park et al., 2023) takes the fundamental unit of analysis to be linear subspaces of neural activations.

It is vital that our model interpretability methods have the capacity to find this kind of structure if it does exist. Not all recent methods achieve this, however. In particular, the classical interchange interventions of Geiger et al. (2020) (also known as activation patching) presuppose that neurons will play unique causal roles (Vig et al., 2020; Csordás et al., 2021; Feder et al., 2021; Ravfogel et al., 2020; Elazar et al., 2020; De Cao et al., 2022; Abraham et al., 2023; Chan et al., 2022; Wang et al., 2023) Thus, these methods will inevitably fail to uncover the more abstract representations that McClelland et al. (1986), Smolensky (1986), and others sought to characterize.

In response to this limitation of interchange interventions, we recently proposed *distributed interchange interventions*, which intervene on subspaces and can be thought of as remapping sets of neurons in the original network using a learned change-of-basis matrix, and we developed Distributed Alignment Search (DAS) as a method for finding this matrix (Geiger et al., 2023). Our experiments with DAS showed that these interventions can reveal previously overlooked aspects of the causal structure of networks trained to solve hard problems. For instance, a simple network solving a hierarchical equality task was found to carry out an intuitive symbolic computation exactly, but only in distributed representations (Geiger et al., 2023).

In a recent paper, Makelov et al. (2023) argue that distributed interchange interventions can give rise to what they call “interpretability illusions”: situations in which neurons that are causally disconnected in a normal run of the model come to play causally efficacious roles when distributed interchange interventions are performed. An earlier version of the argument is given in a blog post on LessWrong (Lange et al., 2023).

Here we respond to [Makelov et al.](#)'s argument. Fundamentally, we reject the label “illusions” for these phenomena. These are simply discoveries about distributed representations. The crux of our response is a technical observation: it follows from [Makelov et al.](#)'s definition that a so-called illusion arises where (and essentially only where) the distributed interchange intervention leads the model to induce representations that are not orthogonal to the nullspace of the model's subcomponents (e.g., nullspace of the down-projection weights in the MLP layer of GPT-2). There is nothing problematic about such situations, though, and indeed we show that these situations arise with standard interchange interventions or activation patches. Overall, then, we should not dismiss these phenomena as illusions. They are simply facts about how networks work, and our interpretability methods should be able to discover them. Indeed, as we describe later, the prevalence of such situations reflects the fact that representations will reflect the natural variation in activations driven by inputs, and this variation need not vanish in null directions.

In the present paper, we first review [Makelov et al.](#)'s technical notion of what an “interpretability illusion” is, and then we show that even intuitive and desirable explanations termed “non-illusory” by [Makelov et al.](#) can qualify as illusions in this sense. The fundamental insight here is the general one we identified above: unless the inputs and all subsequent representations are orthogonal to the nullspace of the relevant model components (unlikely in practice, and not something we would impose), so-called illusions are inevitable. We then argue that the illusions [Makelov et al.](#) see in practice are artifacts of their training and evaluation paradigms. Furthermore, we conduct additional analyses on the indirect object identification (IOI) task, aiming to reproduce the findings of [Makelov et al.](#) and to offer further insights. We close by emphasizing that, though we disagree with their core characterization, [Makelov et al.](#)'s examples and discussion have undoubtedly pushed the field of mechanistic interpretability forward.

2 BACKGROUND: DEFINING [MAKELOV ET AL. \(2023\)](#)'S “ILLUSION”

We first formally distill [Makelov et al. \(2023\)](#)'s definition of “illusion”.

2.1 SET-UP

The main objective of DAS is to intervene on activations in a subspace represented by a set of orthonormal vectors $v \in \mathbb{R}^{n \times d}$ where each vector has d elements (corresponding to the dimension of the locations where we've chosen to look for a representation). Let's take two examples from our data distribution, **A** and **B**. Intervening on activations created by **A** with **B** in the subspace spanned by v can be expressed as,

$$u_{\mathbf{A}}^{v \leftarrow \mathbf{B}} = u_{\mathbf{A}} + (u_{\mathbf{B}} v^T - u_{\mathbf{A}} v^T)v \quad (1)$$

where $u_{\mathbf{A}}, u_{\mathbf{B}} \in \mathbb{R}^d$ are the original activations in the intervention location, using **A** (resp. **B**) as the input to the model.

2.2 NULLSPACE DECOMPOSITION

Specializing to the case where the locations of interest are just prior to the down-projection layer of an MLP, [Makelov et al.](#) decompose v into two orthogonal parts as,

$$v = v_{\text{nullspace}} + v_{\text{rowspace}} \quad (2)$$

where the definitions of these two sub-components are,

- $v_{\text{nullspace}}$: v 's projection onto the nullspace of $W_{\text{out}} \in \mathbb{R}^{d \times o}$, the weights of the downward projection layer in the MLP with an input dimension of d and an output dimension of o .
- $v_{\text{rowspace}} = v - v_{\text{nullspace}}$: the complement.

2.3 THE “ILLUSION”

The *intervened output* of the MLP layer as the result of intervention along v in its input can be written out based off Eqn. 1,

$$u_{\mathbf{A}}^{v \leftarrow \mathbf{B}} W_{\text{out}} = u_{\mathbf{A}} W_{\text{out}} + (u_{\mathbf{B}} v^T - u_{\mathbf{A}} v^T)v W_{\text{out}} \quad (3)$$

Now, let's expand the above equation with the decomposition provided in Eqn. 2,

$$u_{\mathbf{A}}^{v \leftarrow \mathbf{B}} W_{\text{out}} = u_{\mathbf{A}} W_{\text{out}} + (u_{\mathbf{B}} v_n^T + u_{\mathbf{B}} v_r^T - u_{\mathbf{A}} v_n^T - u_{\mathbf{A}} v_r^T)(v_n + v_r) W_{\text{out}} \quad (4)$$

where we denote $v_{\text{nullspace}}$ as v_n , and v_{rowspace} as v_r . Now, let's reorganize things a little and rewrite the right-hand side (RHS) as,

$$u_A^{v \leftarrow \mathbf{B}} W_{\text{out}} = u_A W_{\text{out}} + ((u_B - u_A)v_n^\top + (u_B - u_A)v_r^\top)(v_n + v_r)W_{\text{out}} \quad (5)$$

and expand the RHS to five different parts,

$$\begin{aligned} u_A^{v \leftarrow \mathbf{B}} W_{\text{out}} &= u_A W_{\text{out}} + (u_B - u_A)v_n^\top v_n W_{\text{out}} \\ &\quad + (u_B - u_A)v_r^\top v_n W_{\text{out}} \\ &\quad + (u_B - u_A)v_n^\top v_r W_{\text{out}} \\ &\quad + (u_B - u_A)v_r^\top v_r W_{\text{out}} \end{aligned} \quad (6)$$

We know that $xv_n W_{\text{out}} = 0$, for any vector x . Thus, we can cross out the first two terms after $u_A W_{\text{out}}$ as they are always 0. The RHS thus can be simplified as,

$$\begin{aligned} u_A^{v \leftarrow \mathbf{B}} W_{\text{out}} &= u_A W_{\text{out}} + (u_B - u_A)v_n^\top v_r W_{\text{out}} \\ &\quad + (u_B - u_A)v_r^\top v_r W_{\text{out}} \end{aligned} \quad (7)$$

The equation above shows that the causal effects of the subspace interventions can be written in two parts: (1) the projected difference between two examples onto the *nullspace* of W_{out} , and (2) the projected difference between two examples onto the *rowspace* of W_{out} .

[Makelov et al.](#)'s intuition about “interpretability illusion” is that variations of activations along v_n (i.e., the *nullspace* of W_{out}) should not have any causal effect on model predictions. Thus, the causal effects of intervening along v and v_r are approximately the same for any non-illusory v . Following this intuition, the main experiments that [Makelov et al.](#) designed is to compare the causal effect on model predictions of intervening on v with intervening on v_r alone (as noted in Section 3.4 and the experimental sections of the paper). Formally, the “interpretability illusion” of intervening on the targeted MLP layer can be defined as,

$$\text{ILLUSIONEFFECT}(\Phi, W_{\text{out}}, v, \mathbf{A}, \mathbf{B}) = \Phi(u_A^{v \leftarrow \mathbf{B}} W_{\text{out}}) - \Phi(u_A^{v_r \leftarrow \mathbf{B}} W_{\text{out}}) \quad (8)$$

where the model’s downstream components are folded into $\Phi(\cdot)$. Assuming intervening on v produces the desired counterfactual behavior in general, an “illusion” arises if the causal effect of intervening on v_r is *much smaller* than v towards the desired goal (e.g., flip the model’s prediction from the correct name to the incorrect name in their indirect object identification (IOI) experiment with GPT-2).

3 REVISITING [MAKELOV ET AL. \(2023\)](#)’S TOY EXAMPLE

[Makelov et al. \(2023\)](#) introduce an illustrative toy example. We revisit this example in some detail before returning to more complex models.

3.1 SET-UP

The toy example involves a simple linear neural network $f(x) = (xW_1^\top)W_2$ where $W_1 = [1, 0, 1]^\top$ and $W_2 = [0, 2, 1]^\top$. As a result, this network implements a “copy” function as,

$$f(x) = 0 \times (1 \times x)_{H1} + 2 \times (0 \times x)_{H2} + 1 \times (1 \times x)_{H3} = x \quad (9)$$

where we use $(*)_{H1}$, $(*)_{H2}$ and $(*)_{H3}$ to index the activations of three hidden representations. We can then write out the nullspace of W_2 as a plane linearly spanned by two vectors, $[1, 0, 0]$ and $[0, -\frac{1}{\sqrt{3}}, \frac{2}{\sqrt{3}}]$ (i.e., a trivial vector where the weighted sum of last two dimensions nullify the causal effect). The other relevant geometric structure is what we will call the **data-induced submanifold**, which is the subspace of activations that can be achieved by some input to the network. It is clear in this simple example that the data-induced submanifold is the line extended by the unit vector $[\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}]$, the image of W_1 .

In the toy example, the ILLUSIONEFFECT definition in Eqn. 8 can be further simplified given Eqn. 7 and the fact that there are no downstream computations (i.e., Φ_{toy} in Eqn. 8 is a no-op) as,

$$\text{ILLUSIONEFFECT}(\Phi_{\text{no-op}}, W_2, v, \mathbf{A}, \mathbf{B}) = u_A^{v \leftarrow \mathbf{B}} W_2 - u_A^{v_r \leftarrow \mathbf{B}} W_2 = (u_B - u_A)v_n^\top v_r W_2 \quad (10)$$

3.2 AN OBVIOUS NON-ILLUSORY DIRECTION

In the toy example, activations on H3 are taken to mediate the signal through the network. [Makelov et al.](#) are worried that interchange interventions along directions like $[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0]$ would nullify H3 and create a new causal pathway by intervening on the sum of H1 and H2. By contrast, [Makelov et al.](#) identify $v^{\text{non-illusory}} = [0, 0, 1]$ as an obvious non-illusory direction in this case since it identifies H3 as the only causally relevant activation.

Intriguingly, we can see that the direction $v^{\text{non-illusory}}$ is **not** orthogonal to the nullspace of W_2 by simply checking the dot-product between $v^{\text{non-illusory}}$ and any vector on the nullspace of W_2 such as $[0, -\frac{1}{\sqrt{3}}, \frac{2}{\sqrt{3}}]$. This entails that the non-illusory representation will have an illusion effect as defined in Eqn. 10. To see this in detail we work through the calculations. We can decompose $v^{\text{non-illusory}}$ into two non-zero orthogonal parts following [Makelov et al.](#)'s paradigm as,

$$v^{\text{non-illusory}} = v_n^{\text{non-illusory}} + v_r^{\text{non-illusory}} \quad (11)$$

where we denote $v_{\text{nullspace}}^{\text{non-illusory}}$ as $v_n^{\text{non-illusory}}$, and $v_{\text{rowspace}}^{\text{non-illusory}}$ as $v_r^{\text{non-illusory}}$. To find $v_n^{\text{non-illusory}}$ as the orthogonal projection of $v^{\text{non-illusory}}$ onto the nullspace of W_2 , we create a matrix $\mathbf{M} \in \mathbb{R}^{3 \times 2}$ representing the span of two vectors in the nullspace of W_2 as,

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

We can then get the orthogonal projection of $v^{\text{non-illusory}}$ onto this plane as,

$$v_n^{\text{non-illusory}} = \text{proj}_{\mathbf{M}} v^{\text{non-illusory}} = \mathbf{M}(\mathbf{M}^\top \mathbf{M})^{-1} \mathbf{M}^\top v^{\text{non-illusory}} \quad (12)$$

As a result, we get,

$$v_n^{\text{non-illusory}} = [0, -0.4, 0.8] \quad (13)$$

$$v_r^{\text{non-illusory}} = v - v_n^{\text{non-illusory}} = [0, 0, 1] - [0, -0.4, 0.8] = [0, 0.4, 0.2] \quad (14)$$

To derive the exact ILLUSIONEFFECT term, we can sample two random inputs \mathbf{A} (as example x) and \mathbf{B} (as example x') to first obtain hidden representations realized by our simple network as,

$$u_{\mathbf{A}} = [x, 0, x]; u_{\mathbf{B}} = [x', 0, x']; u_{\mathbf{B}} - u_{\mathbf{A}} = [x' - x, 0, x' - x]$$

Now, ILLUSIONEFFECT as in Eqn. 10 can be calculated step-by-step as,

$$\begin{aligned} \text{ILLUSIONEFFECT}(\Phi_{\text{no-op}}, W_{\text{out}}, v, \mathbf{A}, \mathbf{B}) &= (u_{\mathbf{B}} - u_{\mathbf{A}}) v_n^\top v_r W_2 \\ &= [x' - x, 0, x' - x] [0, -0.4, 0.8]^\top [0, 0.4, 0.2] [0, 2, 1]^\top \\ &= 0.8(x' - x) [0, 0.4, 0.2] [0, 2, 1]^\top \\ &= (x' - x) [0, 0.32, 0.16] [0, 2, 1]^\top \\ &= 0.8(x' - x) \end{aligned} \quad (15)$$

Recall that ILLUSIONEFFECT represents the difference in causal effect between intervening on v and v_r alone. This shows the causal effect by intervening with v_r alone *greatly diminishes* by only retaining 20% of the causal effect when intervening on v . For instance, when $x = 1$ and $x' = 5$, intervening along v will result in an output of 5, and intervening along v_r will result in an output of $5 - 0.8 \times (5 - 1) = 1.8$ which is *much lower*.¹ [Makelov et al.](#) identify $v_n^{\text{non-illusory}}$ as an obvious non-illusory direction in the toy example. However, we are able to show that $v_n^{\text{non-illusory}}$ would be classified as an “illusory” direction according to their own definitions (outlined in Section 3.4).

¹[Makelov et al.](#)'s original implementation also normalizes $v_n^{\text{non-illusory}}$ and $v_r^{\text{non-illusory}}$, which will lead to even larger “illusory” effect, given that the norms of both vectors are smaller than 1, thereby further invalidating this non-illusory direction.

3.3 A BROADER LESSON FROM THE EXAMPLE ABOVE

We have shown that [Makelov et al. \(2023\)](#)’s paradigm can reject a non-illusory direction. More importantly, we also have shown that a faithful distributed representation may **not** be orthogonal to the nullspace of downstream computation. **Why?** Simply because the image of upstream computations need not be (and indeed generically will not be). To expand that: the space of possible inputs to the network leads to variation within a given set of neurons that covers some data-induced submanifold. This submanifold need not, and generally will not, span the whole activation space. It should be clear that, if our goal is understanding how the network computes, we should be interested in variation along this submanifold; indeed interchange interventions will operate on this submanifold by construction. However, there is no reason to insist that the nullspace of downstream computations must be orthogonal to this data-induced submanifold. Hence we agree with [Makelov et al.](#) that, for example, their MLP-in-the-middle “illusion” will be common (see Section 7 in the paper). Except it is not an “illusion” in any useful sense. It is a simple fact about the geometry of representations. We believe the work of [Makelov et al. \(2023\)](#) highlights the need to think carefully about these geometric relationships and can lead to important research directions.

Beyond questions of “illusory” causation, the toy example provides a lovely case of multiple abstraction. In brief, a computation such as $h(x) = x - x + x$ can be seen as implementing the identity by copying the input in *two different ways*. As we describe in Appendix A, the toy example is an instance of this pattern with one copy of a distributed representation. The possibility of such multiple abstractions is connected with the fundamentals of the causal abstraction framework and the lack of preferred bases in linear algebra.

4 REMARKS ON [MAKELOV ET AL. \(2023\)](#)’S EXPERIMENTAL EVIDENCE FOR DISCOVERING “ILLUSIONS” IN THE WILD

We close by reviewing the experimental evidence that [Makelov et al. \(2023\)](#) offer for the claim that their “illusions” arise in real pretrained LMs. Our view is that the evidence provided in favor of this claim is not strong. However, for the reasons given above, we would find substantiated examples of this effect to be interesting evidence of latent causal structure rather than a worrisome illusion.

4.1 BACKGROUND: THE INDIRECT OBJECTIVE IDENTIFICATION (IOI) AND FACTUAL RECALL TASKS

[Makelov et al. \(2023\)](#) conduct two experiments to find “illusions” in the wild. We first describe their two tasks in detail.

Indirect Objective Identification (IOI) asks the model to complete sentences of the form “When Mary and John went to the store, John gave a bottle of milk to” with the correct answer being “Mary”. “John” is the repeated name as S (i.e., the subject), and “Mary” is the non-repeated name as IO (i.e., the indirect object). GPT-2 is effective at solving this task in a zero-shot fashion ([Wang et al., 2023](#)). Figure 1 shows one high-level causal model that can solve this task, and the goal is to find alignments between activations and high-level causal variables (e.g., the name position variable or the correct IO name variable).

Factual Recall involves subject-relation-object triple (s, r, o) (e.g., $s =$ “Eiffel Tower”, $r =$ “is in”, $o =$ “Paris”). We test whether a model can recall the fact associated with a triple by constructing a prompt using a subject-relation tuple (e.g., “The Eiffel Tower is in”) and checking whether the model can generate the correct output (e.g., “Paris”) given the prompt. Given any pair of triples that a model can recall correctly, $\mathbf{A} = (s, r, o)$; $\mathbf{B} = (s', r, o')$, the goal is to intervene on the model’s activations induced using \mathbf{A} with activations using \mathbf{B} and have the output changed from o to o' .

4.2 INTERCHANGE INTERVENTION ACCURACY (IIA)

In our work on causal abstraction (starting from the initial work of [Geiger et al. 2020](#)), we evaluate the degree to which a high-level model is a faithful description of a neural network based on our ability to control model behavior through interventions on neural activations aligned with high-level variables. Interchange intervention accuracy (IIA) is the percentage of counterfactual predictions matched under interchange interventions between the high-level causal model and the neural model. Although IIA has limitations (as outlined in [Geiger et al. 2023; Wu et al. 2023](#)), it offers an intuitive

measure for gauging the interpretability of a neural model in the following sense: IIA is bounded by 100%, and when IIA is 100%, the causal model is a faithful explanation of how the neural network behaves under the set of interventions we perform, which provides transparent insights into the neural network’s causal mechanisms (Geiger et al., 2022). In practice, achieving 100% IIA is uncommon, but the higher the value of IIA, the more confidence we have in reasoning about the high-level causal model as a stand-in for the low-level neural model.

In Makelov et al. (2023)’s IOI experiment, they try to learn a single DAS direction that abstracts the name position information. For instance, if the IO (i.e., correct output token) name appears as the second name in the base, and the first name in the source example, the counterfactual label for the base would switch to the first name mentioned in the base after the intervention. When finding such directions, Makelov et al. (2023) end up with an IIA ranging from 0.0% to 4.2% for intervention locations in MLP layers, as shown in their Table 1. This is far from a strong signal given the fact that GPT-2 can achieve a task performance of 91.6%.² In our work, we routinely see significantly higher IIA values. For example, in the study of how Alpaca solves a simple reasoning task in context (Wu et al., 2023), we see IIA values around 85% even in ambitious, out-of-domain evaluations. Thus, we interpret 4%, which is much lower than the task performance, as a failure to find any relevant structure.

Instead of relying on IIA, Makelov et al. (2023) use a new metric, which they call the fractional logit difference decrease (FLDD) to quantify the actual “illusion” effect. FLDD measures a fractional logit difference between the IO and S (i.e., incorrect output token) tokens before and after the intervention, as described in Equation 4 in the paper. This measure is interesting because it encodes a qualitative pattern that very roughly tracks IIA. Assuming near-perfect performance of the model on a given task, when FLDD is above 100% this means the intervention, on average, produces the right counterfactual prediction (e.g., switching from IO to S in the IOI experiment). As their Table 1 shows, having FLDD greater than 100% does correlate somewhat with relatively high IIA scores.

However, the measure may also mislead one into thinking that the magnitudes themselves are closely reflective of counterfactual model behavior. Because FLDD is based on a ratio of log-likelihoods, it can be brought arbitrarily close to 100% **with no effect whatsoever on counterfactual behavior**. For instance, if the model predicts IO with a probability of 0.505, while the intervention leads to a slightly lower probability of 0.5025, this would mean approximately 50% FLDD, even though model behavior does not change at all. This is why their Table 1 reveals a generally poor correlation with IIA (e.g., 46.7% FLDD results in only 4.2% IIA; see also their discussion on p. 14).

Despite the limitations of FLDD, we are enthusiastic about considering other metrics than IIA for evaluating model explanations, including those that compare logit differences (e.g., KL divergence).

4.3 CHECKING DORMANT OR DISCONNECTED COMPONENTS VIA CORRELATIONAL ANALYSIS OF ACTIVATIONS

Makelov et al. (2023) extensively study the distributions of projected activations when projecting onto nullspace and rowspace to further validate their hypothesis around dormant or disconnected pathways in GPT-2. Furthermore, they often draw conclusions by analyzing the activation distributions. For instance in their Figure 5, they observe that activation projections onto the nullspace separate more saliently across two groups compared to activation projection onto the rowspace. On this basis, they claim that the rowspace is predominantly dormant and the nullspace projection drives the causal effect.

While it is interesting to observe how activations are distributed differently, we want to point out that they may carry less causally interesting information than one might expect. For example, Huang et al. (2023) assess the proposal of Bills et al. (2023) to use GPT-4 to explain neurons in GPT-2-XL. One of their core observations is that activations that are found to be strongly correlated with a certain high-level concept can have little to no causal effect on model predictions in downstream tasks where the concept is being used. We thus argue that analysis solely based on activations cannot support strong causal claims about model structures.

²Makelov et al. (2023) reported 91.6% accuracy in their publicly released codebase. We reproduced this number and provided detailed scripts in https://github.com/frankaging/pyvne/blob/main/tutorials/advanced_tutorials/IOI_with_DAS.ipynb.

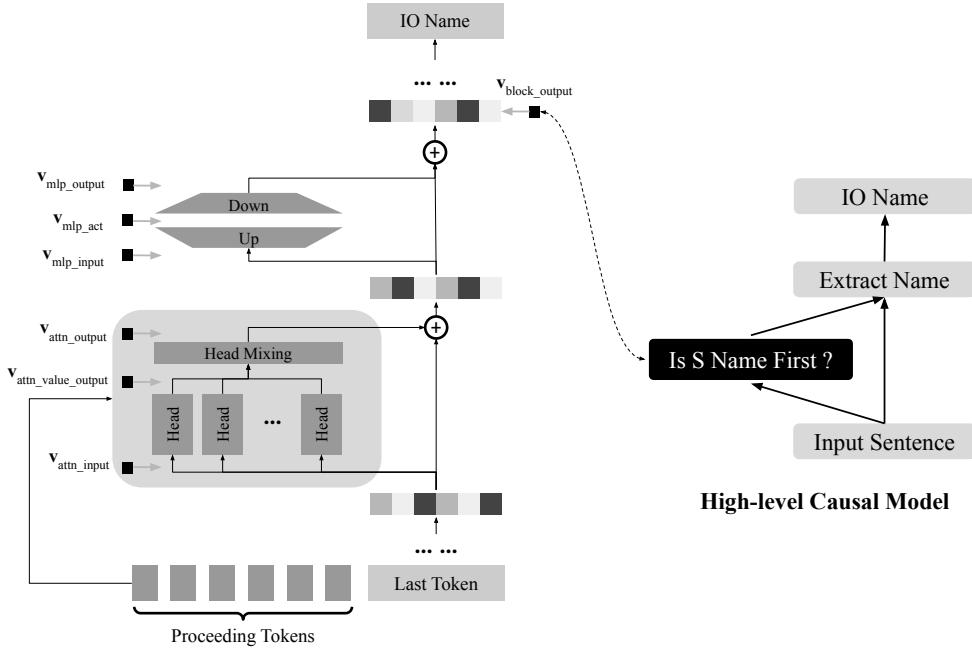


Figure 1: An illustration of aligning a high-level causal model with key intervention locations of the streams on top of the last input token in the GPT-2 model. Besides aligning the main residual streams and the MLP activations, we align other streams to study how the name position information emerges in the GPT-2 model. The head mixing layer is a linear layer.

4.4 POTENTIAL PITFALLS IN THE FACTUAL RECALL EXPERIMENT

Makelov et al. (2023) also try to replicate their finding of “illusion” in the factual recall task as in their Section 6.1 (p. 20). Their core experimental pipeline for finding “illusions” relies on learning a single DAS direction for a single pair of facts with MLP activations at different layers of GPT2-XL.³ In other words, Makelov et al. (2023) find a DAS direction that would change the label of the base example to the source example when intervening on that single DAS direction. This is **not** an experimental design that we would endorse. Two potential pitfalls are worth calling out for future research when using DAS to find alignments between neural networks and high-level causal models.

First, the learned DAS direction aligns the neural network with an uninteresting high-level causal model: a dummy model that always returns the answer of the seen source example. It should be radically unsurprising to find the “return the answer” high-level causal model is indeed a causal abstraction of any model that returns the answer. Consequently, we suspect there could be many DAS directions that could achieve this alignment with a seemingly perfect alignment score (as shown in Figure 7 on p. 20), but such results are not meaningful explanations. Second, the current experimental setup also entails that the learned DAS direction overfits the single pair of facts seen during training. DAS is a model with learned parameters. As such (as with essentially all ML models), it should be trained on numerous examples and tested on examples that are disjoint from the train set. Makelov et al. (2023)’s factual recall experiment violates both of these conditions. As a result, their directions would likely not generalize to any other facts except the ones it is trained on. Additionally, the latter point is related to the former: generalization would show evidence of alignment with a more interesting high-level model.

³They compare the causal effect between intervening on learned DAS direction v and $v_{rowspace}$ (i.e., the rowspace decomposition as shown in Section 2) for the seen pair of facts to trace “illusion”. Details can be found in Makelov et al. (2023)’s original implementation at https://github.com/amakelov/activation-patching-illusion/blob/main/fact_patching.ipynb.

5 ADDITIONAL INDIRECT OBJECT IDENTIFICATION (IOI) RESULTS WITH GPT-2

In this section, we show that DAS can be productively used to reproduce some of the key findings outlined in previous mechanistic interpretability results on the IOI task (Wang et al., 2023; Makelov et al., 2023). Furthermore, we show how to use DAS to gain new insights into causal mechanisms induced by GPT-2 on IOI.⁴

5.1 SET-UP

We use the same model and dataset generation script as Wang et al. (2023) and Makelov et al. (2023). Following Makelov et al. (2023), we learn a single DAS direction to localize the name position in the IOI task. Since there are only two names (i.e., the S or IO name) mentioned in the input sentence, the name position information can be binarized (e.g., a binary variable indicating whether S is the first name or not) and thus represented in a single dimension. As shown in Figure 1, one possible high-level causal model of solving the IOI task is to first determine whether the S name is the first name mentioned in the sentence, and then extract the name based on the intermediate result of the first step before returning the final answer. In our experiments as well as Makelov et al. (2023)'s experiments, DAS is used to find causal variables representing the name position information (i.e., a causal variable representing whether S name is the first name or not as shown in Figure 1). We try seven intervention locations to find the causal variable as annotated in Figure 1. We use IIA as our evaluation metric. Other details about our experimental set-up can be found in Appendix B.

5.2 THE NAME POSITION INFORMATION IN THE MAIN RESIDUAL STREAM AND THE MLP ACTIVATIONS

As shown in Figure 2, our results are consistent with Makelov et al. (2023)'s findings, where the name position information best aligns with the 8th⁵ layer in GPT-2. The learned DAS direction achieves an IIA of 70% with the main residual stream (v_{block_out}) and an IIA of 4% with the MLP activations (v_{mlp_act}), which is approximately the same as Makelov et al.'s results (see their Table 1). Our IIA scores suggest that MLP activations carry hardly any information about name position information.

Our results also suggest that single token representations before the last token rarely carry any name position information.⁶ Furthermore, we use DAS to check whether name position information is distributed across tokens by learning a single DAS direction on the concatenated token representation of all the tokens before the last token (16 tokens in total) at the 7th layer (a layer before we find name position information). We find an IIA of 3%, which means name position information is not distributed across tokens either. This gives another piece of evidence that name position information emerges at the 8th layer only.

5.3 DIFFERENT STREAMS SHOW WHERE THE NAME POSITION INFORMATION EMERGES

As shown in Figure 1, we align with seven different locations across layers above the last token to trace the name information across components. As shown in Figure 3, consistent with previous work (Wang et al., 2023; Makelov et al., 2023), the MLP layers carry little to no information about the name position. For instance, the MLP input and output streams (v_{mlp_input} and v_{mlp_output}) both flatten out at 0% IIA across all layers.

Our results also suggest that the name position information emerges most saliently through the self-attention block at the 8th layer. As shown on the top panel of Figure 3, the block input stream (v_{block_input}) of the 8th layer (i.e., the block output stream of the previous layer) reaches about 11% IIA, the attention output stream (v_{attn_out}) reaches about 48% IIA, and the resulting block output stream reaches about 69% IIA. These results suggest that the self-attention block is largely responsible for adding the name position information into the residual stream. Furthermore, a similar

⁴To foster reproducibility, we release our code and data at https://github.com/frankaging/pyvene/blob/main/tutorials/advanced_tutorials/IOI_with_DAS.ipynb.

⁵Author correction: we use 0-indexing that “0th” layer means the first layer.

⁶We additionally align with each token representation other than the four proceeding tokens before the last token and find IIAs are approximately zero.

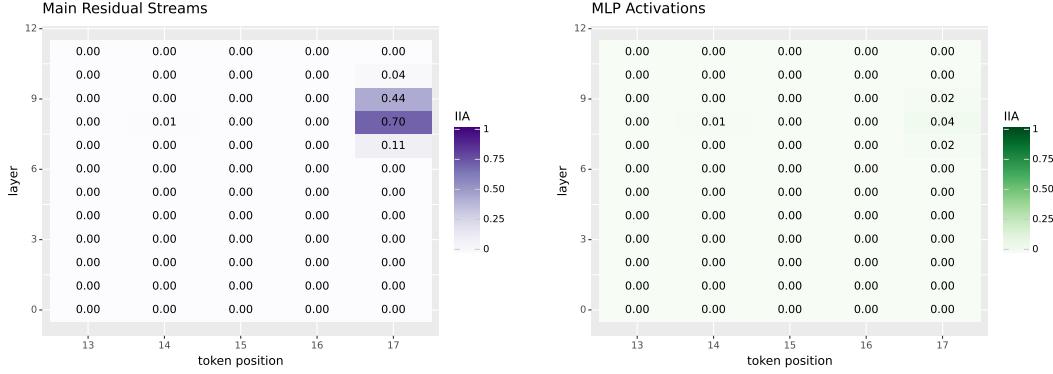


Figure 2: Interchange Intervention Accuracy (IIA) when aligning the name position variable with different intervention locations in the main residual streams ($v_{\text{block.out}}$) as well as the MLP activations ($v_{\text{mlp.act}}$) above the last token and the four tokens proceeding it. The GPT-2 model achieves 96% task accuracy. Higher IIA means better alignment. Overall our results are consistent with [Makelov et al. \(2023\)](#)'s findings where name position information mainly resides above the last token at the 8th layer.

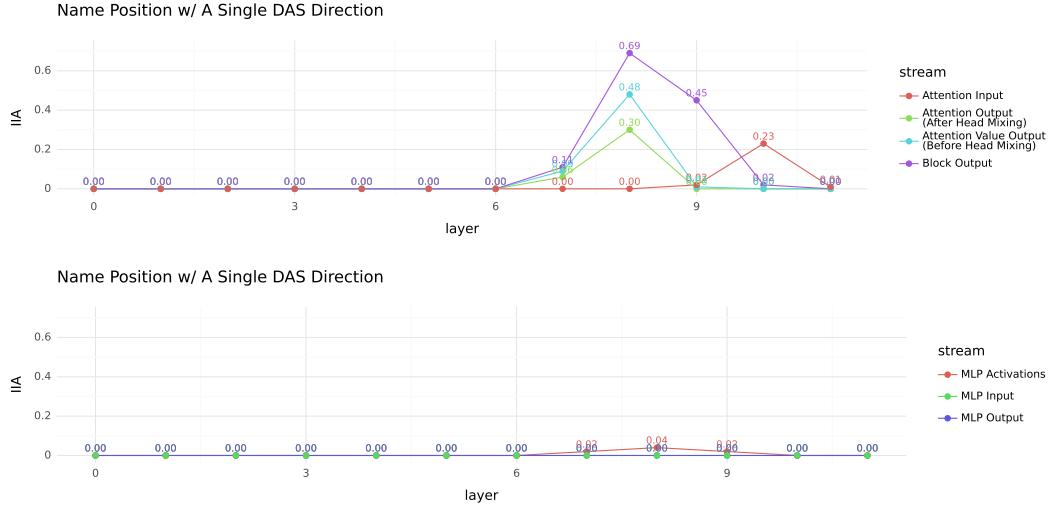


Figure 3: Interchange Intervention Accuracy (IIA) when aligning the name position variable with different intervention locations above the last token.

trend also happens in the 7th layer, suggesting the name information emerges across layers in a distributed way as well. Strikingly, the self-attention streams after the 8th layer carry nothing about the name position information suggesting these self-attention heads can be pruned, as shown in previous work ([Wang et al., 2023](#)).

DAS assumes a fixed-size linear subspace; however, a binary variable may not always fit within a linear subspace of a single dimension. Thus, we also experimented with Boundless DAS ([Wu et al., 2023](#)), a variant of DAS that does not assume a fixed dimensionality for the DAS subspace. With Boundless DAS, the boundaries of the distributed intervention are learned along with the change-of-basis matrix. Supporting results using this method can be found in Appendix D.

5.4 NAME POSITION INFORMATION IS DISTRIBUTED ACROSS ATTENTION HEADS

[Wang et al. \(2023\)](#) show that there are certain heads (e.g., heads 6 and 10) in the 8th layer that are responsible for deriving the name position information. To verify whether a single head can carry the name position information, we train a single DAS direction on the head's representation individually

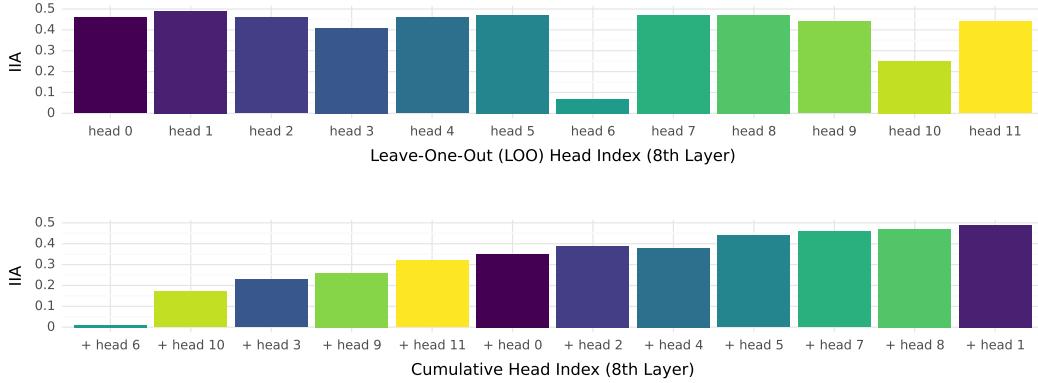


Figure 4: Interchange Intervention Accuracy (IIA) when aligning the name position variable with head representations. The **top panel** shows IIA when aligning with a concatenated representation of all heads in the 8th layer by leaving one head out at a time. The **bottom panel** shows IIA when aligning with cumulated head representations by starting from the head resulting in the largest drop in the top panel and concatenating with one additional head at a time based on the sorted order of IIA drops from the top panel.

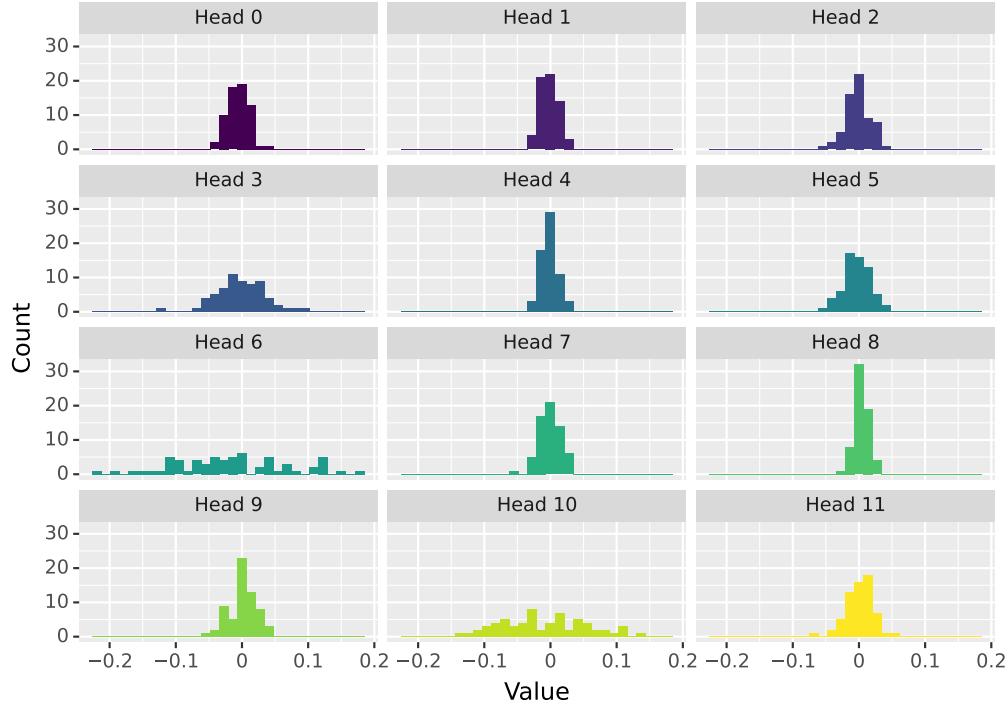


Figure 5: Distributions of learned DAS weights (a single dimension DAS) when aligning with the name position information at attention value output stream of the 8th layer. The number of non-zero entries maps well to the head importance discovered through our ablation studies in Section 5.4 as well as findings from previous works (Wang et al., 2023).

at the 8th layer. Surprisingly, when aligning with each head, IIA flattens out at 0%, which suggests the name position information may be distributed across multiple heads instead.

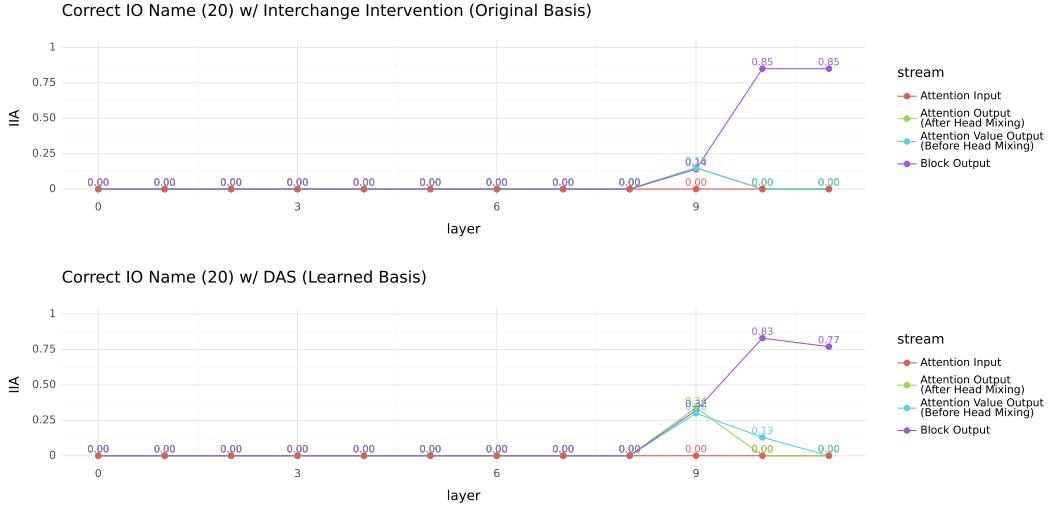


Figure 6: Interchange Intervention Accuracy (IIA) when aligning the correct IO name variable with interchange intervention (i.e., no training) and DAS at different intervention locations above the last token.

Instead of aligning with individual heads, we perform Leave-One-Out (LOO) alignment: we align with concatenated head representations by leaving one head out at a time. As shown in the top panel of Figure 4, leaving heads 6 and 10 out results in the largest drops in IIA, which aligns with previous findings by Wang et al. (2023) claiming these two heads are responsible for carrying out the name position information. Nevertheless, we find that the name position information is fairly distributed across heads, or the information emerges only when multiple heads are considered together and heads need to act together to reach good IIA.

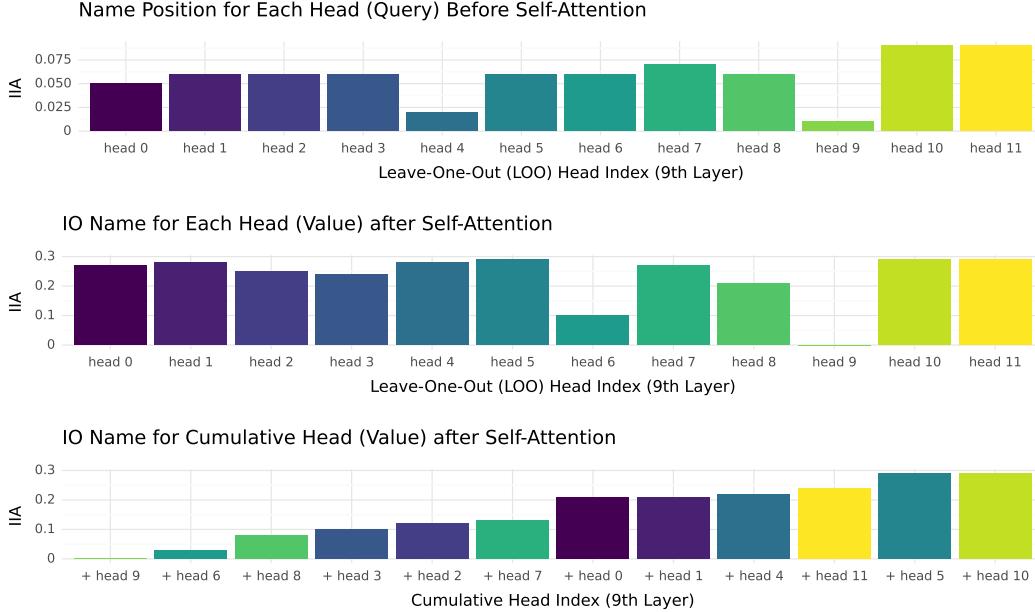


Figure 7: Interchange Intervention Accuracy (IIA) when aligning the name position variable or the correct IO name with head representations at the 9th layer.

To verify this, we then rank the IIA drops across heads from the highest (i.e., leaving this head out results in the biggest loss in IIA) to the lowest based on LOO experiments. We re-align with

cumulated head representations by starting from the lowest head itself and concatenating with one additional head at a time based on the sorted order. Our results show that head 6 alone reaches 0% IIA, and it reaches less than 20% when considered together with head 10 as shown in the bottom panel of Figure 4. Our findings suggest that all heads are crucial to carrying out all the name position information. Figure 5 shows the breakdown of the learned DAS weights across heads in the attention value output stream ($v_{\text{attn_value_output}}$), without head mixing. Our results indicate that the importance of each head correlates well with the number of non-zero entries in the learned DAS weight matrix. Additional analyses at 7th layer can be found in Appendix C.

5.5 NAME POSITION INFORMATION IS PROCESSED IN A DISTRIBUTED WAY

We further check where the correct IO token emerges above the last token. To find the IO token representation, we use vanilla interchange interventions (Geiger et al., 2021) as well as DAS. As shown in Figure 6, the correct IO token seems to emerge starting from the 9th layer. Comparing results from DAS with vanilla interchange intervention, we find that DAS finds new alignments in the attention value output stream with an IIA of 13% where the vanilla interchange intervention flats out with a 0% IIA.

One hypothesis is that self-attention heads at the 9th and 10th layers process the name position information passed in and fetch the correct IO name by adding it to the main residual stream. Indeed, Wang et al. (2023) shows head 9 at the 9th layer is identified as the strongest “name mover head” that copies the IO name from the sequence into the last token. To verify this, we align the name position information in the query output stream ($v_{\text{attn_input}}$) at each head as well as the IO name in the attention value output at each head. Instead of aligning each head individually, we first perform the LOO experiment as in Section 5.4. Figure 7 shows our results. We find that head 9 indeed absorbs the name position information the most compared to other heads (i.e., removing head 9 results in the largest IIA drops when aligning the name position before self-attention) and contributes to generating the IO name in the attention value output stream (i.e., removing head 9 results in the largest IIA drops when aligning the IO name after self-attention). However, our results also suggest that head 9 alone cannot produce the correct IO name result in 0% IIA when aligning the IO name by itself. We need up to 11 heads to reach approximately the ceiling IIA of aligning the IO name from the attention value output stream. Our findings suggest that attention heads process information in a very distributed way. Interestingly, the correct IO name alignments disappear after head mixing at 10th layer, suggesting another causal variable may emerge to add the correct IO name information into the main residual stream. Appendix C includes additional analyses at 10th layer which shows similar results.

6 CONCLUSIONS

We disagree with Makelov et al. (2023)’s characterization of their examples and results, but we find their contribution to be extremely valuable, and we applaud their effort to probe methods like DAS very deeply. Overall, their work has helped us to more deeply understand DAS as well as how neural networks work in general, and thus we feel the discussion has furthered the goals of explainable AI in general.

ACKNOWLEDGMENTS

We thank Aleksandar Makelov, George Lange, and Neel Nanda for providing feedback on an earlier version of this paper.

REFERENCES

- Eldar David Abraham, Karel D’Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. CEBaB: Estimating the causal effects of real-world concepts on NLP model behavior. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL <https://arxiv.org/abs/2205.14140>.
- Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. 2023. URL <https://openaipublic.blob.core.windows.net/neuron-explainer/>

[paper/index.html](#).

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.

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

Róbert Csordás, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Are neural nets modular? inspecting functional modularity through differentiable weight masks. In *International Conference on Learning Representations (ICLR)*, 2021. URL <https://openreview.net/forum?id=7uVcpu-gMD>.

Nicola De Cao, Leon Schmid, Dieuwke Hupkes, and Ivan Titov. Sparse interventions in language models with differentiable masking. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2022. URL <https://aclanthology.org/2022.blackboxnlp-1.2>.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. In *Proceedings of the 2020 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2020. URL <https://arxiv.org/abs/2006.00995>.

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, 2022. URL <https://arxiv.org/abs/2209.10652>.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. CausaLM: Causal Model Explanation Through Counterfactual Language Models. In *Computational Linguistics*, 2021. URL https://doi.org/10.1162/coli_a.00404.

Atticus Geiger, Kyle Richardson, and Chris Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the 2020 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2020. URL <https://arxiv.org/abs/2004.14623>.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL <https://papers.nips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html>.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning (ICML)*, 2022. URL <https://proceedings.mlr.press/v162/geiger22a.html>.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations, 2023. URL <https://arxiv.org/abs/2303.02536>.

Jing Huang, Atticus Geiger, Karel D’Oosterlinck, Zhengxuan Wu, and Christopher Potts. Rigorously assessing natural language explanations of neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2023. URL <https://arxiv.org/abs/2309.10312>.

- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL <https://arxiv.org/abs/1412.6980>.
- Georg Lange, Alex Makelov, and Neel Nanda. An interpretability illusion for activation patching of arbitrary subspaces. *LessWrong*, 2023. URL <https://www.lesswrong.com/posts/RFtkRXHebkwygDe2/an-interpretability-illusion-for-activation-patching-of#comments>.
- Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. *arXiv preprint arXiv:2311.17030*, 2023. URL <https://arxiv.org/abs/2311.17030>.
- J. L. McClelland, D. E. Rumelhart, and PDP Research Group (eds.). *Parallel Distributed Processing. Volume 2: Psychological and Biological Models*. MIT Press, Cambridge, MA, 1986.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Association for Computational Linguistics (ACL)*, 2013. URL <https://aclanthology.org/N13-1090/>.
- Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2023. URL <https://aclanthology.org/2023.blackboxnlp-1.2>.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. 2020. URL <https://distill.pub/2020/circuits-zoom-in>.
- Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2023. URL <https://arxiv.org/abs/2311.03658>.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by Iterative Nullspace Projection. In *Association for Computational Linguistics (ACL)*, 2020. URL <https://doi.org/10.18653/v1/2020.acl-main.647>.
- D. E. Rumelhart, J. L. McClelland, and PDP Research Group (eds.). *Parallel Distributed Processing. Volume 1: Foundations*. MIT Press, Cambridge, MA, 1986.
- P. Smolensky. Neural and conceptual interpretation of PDP models. In *Parallel Distributed Processing: Explorations in the Microstructure, Vol. 2: Psychological and Biological Models*. MIT Press, Cambridge, MA, USA, 1986.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL <https://arxiv.org/abs/2004.12265>.
- Kevin 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 *International Conference on Learning Representations (ICLR)*, 2023. URL <https://arxiv.org/abs/2211.00593>.
- Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL <https://arxiv.org/abs/2305.08809>.

APPENDIX

A MULTIPLE CAUSAL ABSTRACTIONS EXIST IN THE TOY EXAMPLE

Besides revisiting the “illusion” in the toy example, we also find there is an interesting connection between the disconnected or dormant direction and the existence of multiple causal abstractions.

A.1 SET-UP

In the original basis, the neural model is implemented as,

$$f(x) = 0 \times (1 \times x)_{H1} + 2 \times (0 \times x)_{H2} + 1 \times (1 \times x)_{H3} = x \quad (16)$$

Now, let's say we have a rotation matrix \mathbf{R} (change-of-basis matrix) as,

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

Importantly, we use the subspace identified as an “illusion” by [Makelov et al. \(2023\)](#) as the first direction in this rotation matrix. We can then rewrite $f(x) = g(x) = W_2^T R^T (RxW_1)$ given $R^{-1} = R^T$ for orthonormal matrix. Thus, $g(x)$ can be expressed as,

$$g(x) = \frac{1}{\sqrt{2}} \times (\sqrt{2} \times x)_{H1} - \frac{1}{\sqrt{2}} \times (\sqrt{2} \times x)_{H2} + 1 \times (1 \times x)_{H3} = x \quad (17)$$

A.2 MULTIPLE CAUSAL ABSTRACTIONS

$f(x)$ and $g(x)$ are behaviorally identical yet structurally different due to the basis change. Likewise, $g(x)$ gives us additional intervention signals when intervening on H1 in the rotated basis. In other words, if we want to align the representation of x (i.e., the input identity) with some representations among $\{H1, H2, H3\}$, it could be H3 in the original basis by $f(x)$, or H1 in the rotated basis by $g(x)$. This resembles the case where there are multiple causal abstractions when aligning causal variables. We discussed this extensively in [Geiger et al. 2023](#) and [Wu et al. 2023](#). In essence, the causal abstraction framework cannot determine structural equivalence between objects. Instead, it says that, under a set of known interventions drawn from the data distribution, we cannot distinguish two objects (e.g., a Python program and a neural network) behaviorally under any combinations of these interventions. One could try to align components between two objects more exhaustively to distinguish multiple causal abstractions. For instance, we could further align all hidden representations (e.g., intervention on H2 in $f(x)$ and $g(x)$ will give different counterfactual behavior) in the toy example. However, for sufficiently complex models, there are still likely to be multiple distinct fully correct abstractions. Additionally, there is a correspondence with the complexity of the high-level model. This means that more accurate abstractions are likely when the low-level model is complex, and the high-level model is simple.

B EXPERIMENTAL SETUP

We use the GPT-2 model Small checkpoint from HuggingFace.⁷ We use the dataset generation scripts from the publicly released code of [Makelov et al. \(2023\)](#).⁸ To train a single DAS direction, we sample 200 pairs of base and source example pairs. We use the Adam optimizer ([Kingma & Ba, 2015](#)) with an initial learning rate of 0.01 for an epoch number of 10 with a batch size of 20. As in [Makelov et al. \(2023\)](#), examples in the dataset have three different templates, yet the base and the source examples in each pair share the same template. Two templates are used during training while an unseen template is used during evaluation. Each input example has the same sequence length of 18. In case the predicting name is tokenized into multiple tokens, the accuracy is calculated using only the first token. Training of a single DAS direction takes less than a minute using a single 12G Nvidia GPU.

To find the IO name representations, we sample 200 pairs of base and source randomly and have the correct counterfactual label to be the source IO name. Since DAS is likely to overfit to seen IO names, we have 20 distinct IO names as the possible output label during training and testing. We train DAS with a dimension size of 20. We keep other training settings the same as our name position experiments above. For our cumulative head experiments, we sort heads based on their original indices for stability.⁹ For Boundless DAS, we initialize the dimensionality to occupy half of

⁷GPT-2 can be accessed publicly at <https://huggingface.co/gpt2>.

⁸The code is publicly released at <https://github.com/amakelov/activation-patching-illusion>.

⁹Ideally, the order should not matter. In practice, given the scale of the dataset and the magnitude of IIAs, we observe variance in results as the order changes.

the total hidden representation size (e.g., a size of 384 when aligned with the block output with a size of 768). The learning rate for boundary learning is set at 0.05. Furthermore, we apply a weighted loss for Boundless DAS, assigning a weight of 1.0 to the DAS loss and 2.0 to the boundary loss. We set the initial temperature for boundary learning to be 50.0, and the end temperature to be 0.1 with a linear temperature annealing throughout the training.

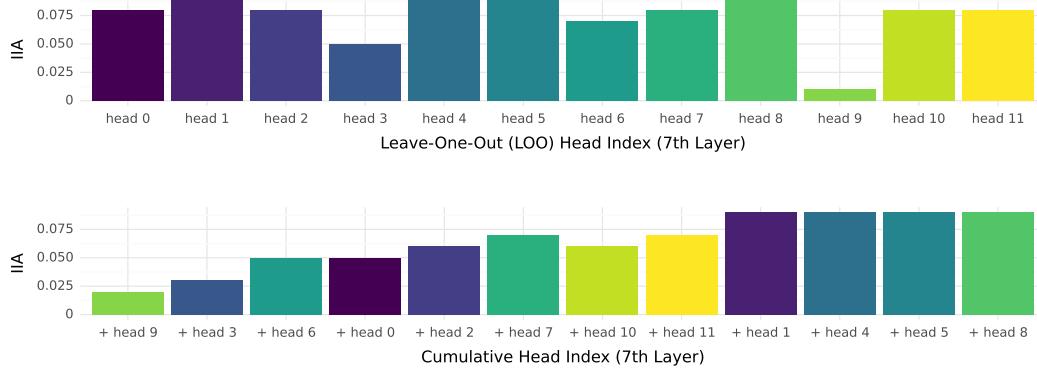


Figure 8: Interchange Intervention Accuracy (IIA) when aligning the name position variable with head representations. The **top panel** shows IIA when aligning with a concatenated representation of all heads in the 7th layer by leaving one head out at a time. The **bottom panel** shows IIA when aligning with cumulated head representations by starting from the head resulting in the largest drop in the top panel and concatenating with one additional head at a time based on the sorted order of IIA drops from the top panel.

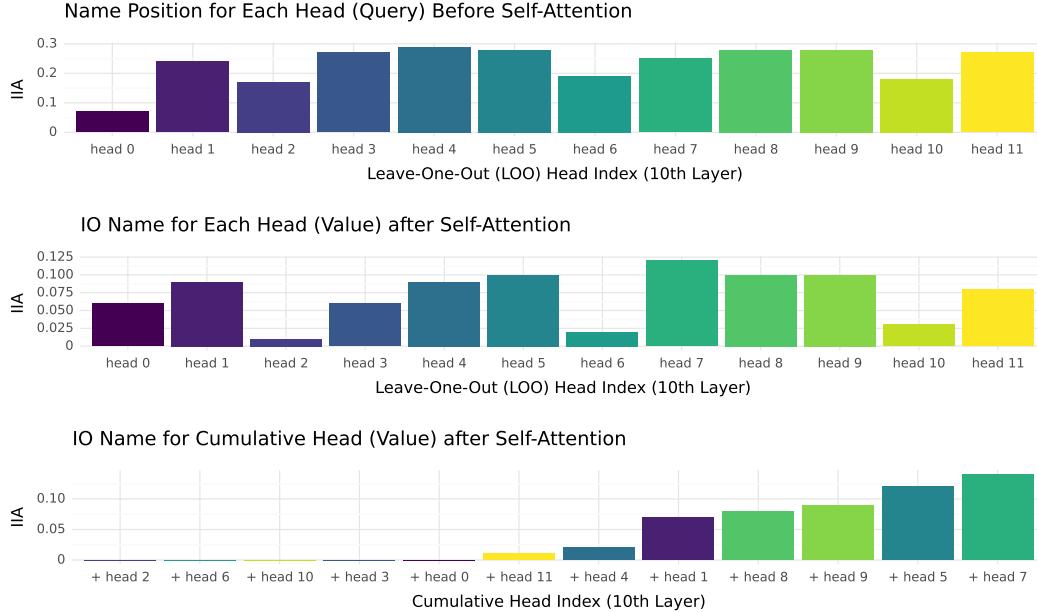


Figure 9: Interchange Intervention Accuracy (IIA) when aligning the name position variable or the correct IO name with head representations at the 10th layer.

C ADDITIONAL HEAD ALIGNMENT RESULTS

We conduct the same set of analyses outlined in Section 5.4 at the 7th layer as well since our results in Figure 3 show that attention value representations in the 7th layer also carry some name position

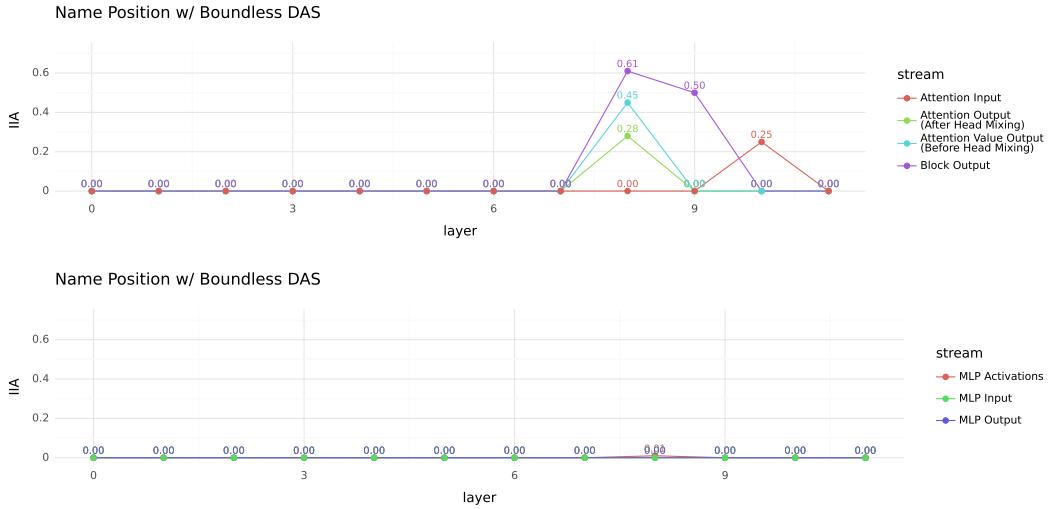


Figure 10: Interchange Intervention Accuracy (IIA) when aligning the name position variable with Boundless DAS and at different intervention locations above the last token.

information. As shown in Figure 8, there are particular heads (e.g., heads 9 and 3) that result in larger IIA drops compared to others when removed. Similarly, the name position information is fairly distributed across heads. Notably, removing heads in $\{4, 5, 8\}$ results in almost no drop in IIA suggesting information is less distributed compared to the 8th layer. Figure 9 shows additional results when aligning the name position variable or the correct IO name with head representations at the 10th layer. Figure 13 shows the distributions of the learned DAS weights for each head when aligning the correct IO name with the attention value output of the 9th layer. Consistent with our findings in Section 5.4, the importance of each head corresponds well with the number of non-zero entries in the learned weight matrix.

D ADDITIONAL BOUNDLESS DAS RESULTS

In Figure 10 and Figure 11, additional alignment results are presented using Boundless DAS (Wu et al., 2023). Boundless DAS achieves comparable outcomes to DAS when aligning the name position variable. However, it demonstrates superior performance in alignment with the correct IO name. Figure 12 shows the distributions of learned DAS weight for each head when aligning the name position information with the attention value output of the 8th layer. Compared with DAS results in Figure 5, contrasts between heads are much less salient, suggesting that Boundless DAS learns a much more distributed subspace. Figure 14 shows the distributions of learned DAS weight for each head when aligning the correct IO name with the attention value output of the 9th layer, which suggests a similar finding.

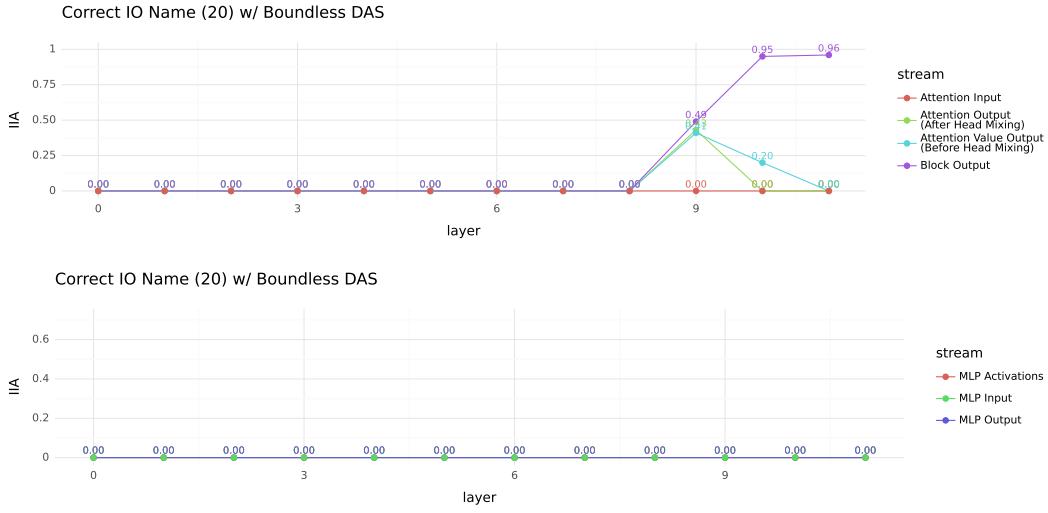


Figure 11: Interchange Intervention Accuracy (IIA) when aligning the correct IO name variable with Boundless DAS and at different intervention locations above the last token.

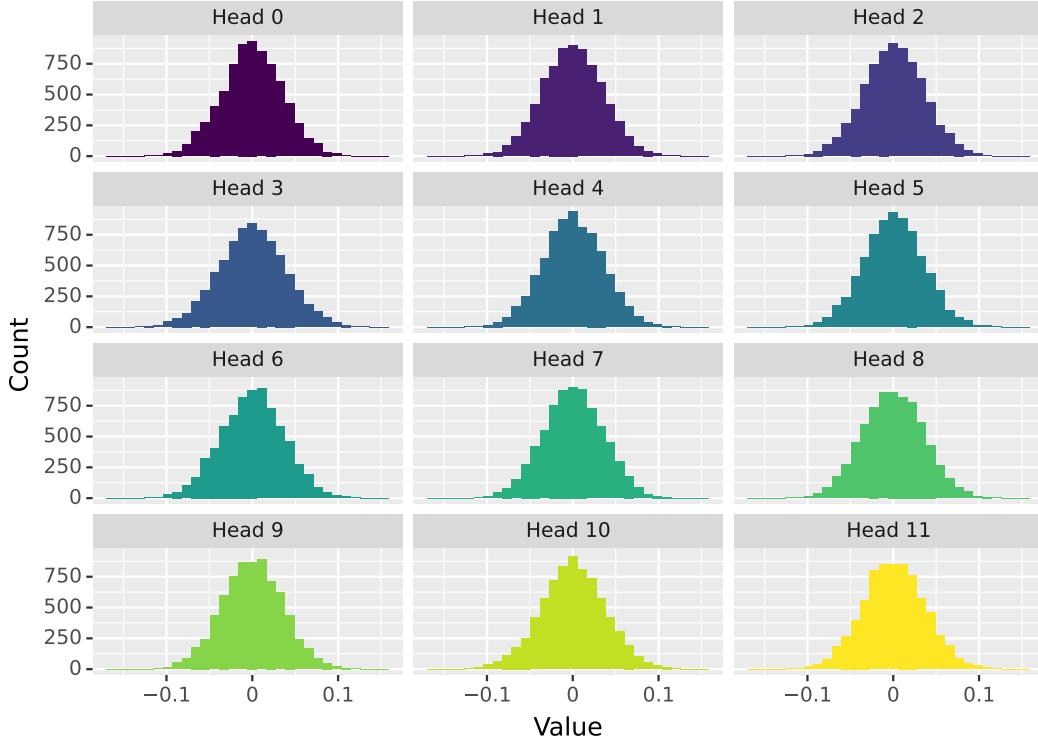


Figure 12: Distributions of learned Boundless DAS weights (learned boundary occupied 14.88% of the original dimension size) when aligning with the name position information at attention value output stream of the 8th layer.

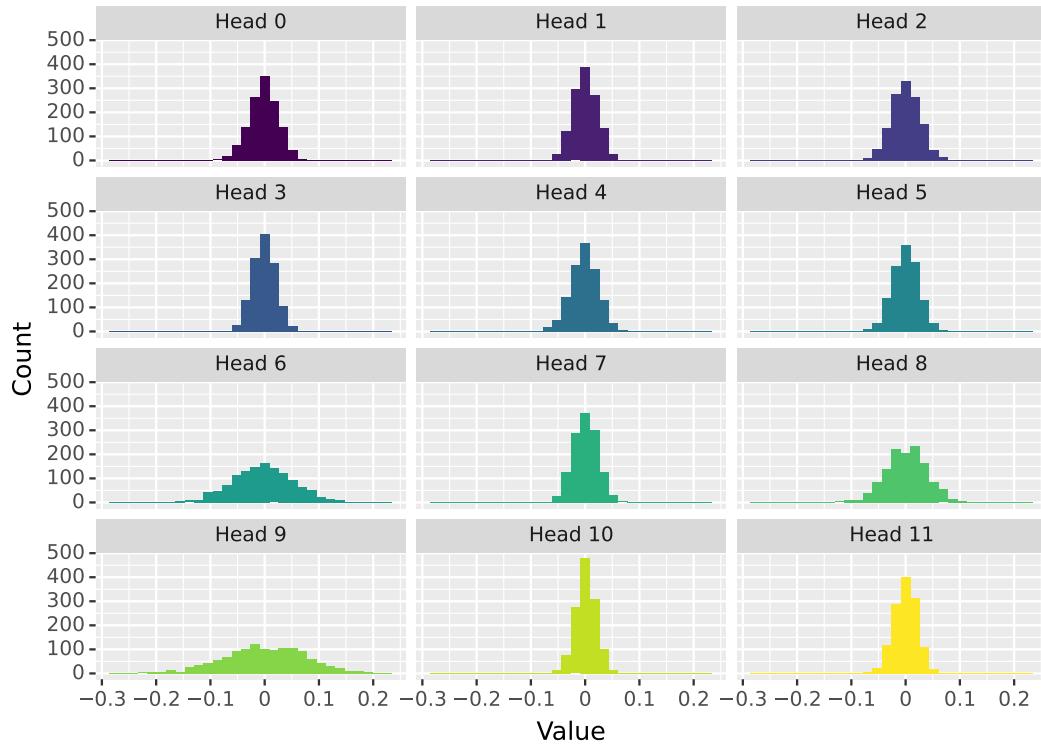


Figure 13: Distributions of learned DAS weights (a fixed dimensionality of 20) when aligning with the correct IO name at attention value output stream of the 9th layer.

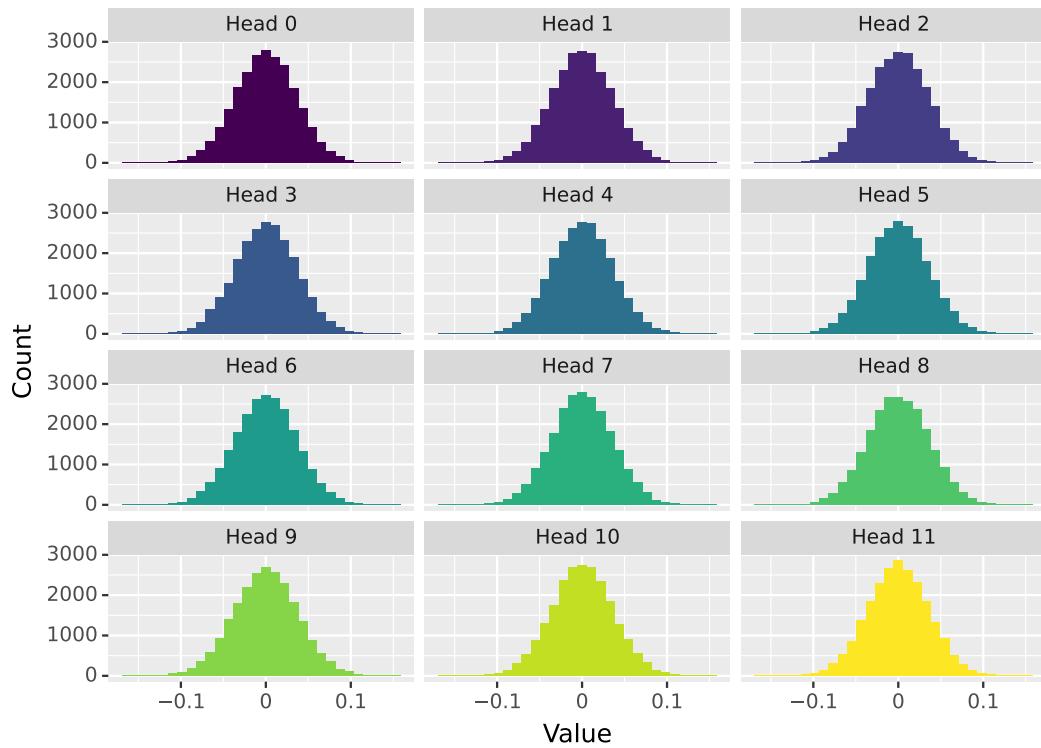


Figure 14: Distributions of learned Boundless DAS weights (learned boundary occupied 46.87% of the original dimension size) when aligning with the correct IO name at attention value output stream of the 9th layer.