Survey

The Quest for the Right Mediator: Surveying Mechanistic Interpretability for NLP Through the Lens of Causal Mediation Analysis

Aaron Mueller[1*], Jannik Brinkmann[2], Millicent Li[3], Samuel Marks[4], Koyena Pal[3], Nikhil Prakash[3], Can Rager[3], Aruna Sankaranarayanan[5], Arnab Sen Sharma[3], Jiuding Sun[6], Eric Todd[3], David Bau[3], and Yonatan Belinkov[7]

[1]Boston University
amueller@bu.edu
[2]University of Mannheim
[3]Northeastern University
[4]Anthropic
[5]Massachusetts Institute of Technology
[6]Stanford University
[7]Technion – Israel Institute of Technology

Interpretability provides a toolset for understanding how and why language models behave in certain ways. However, there is little unity in the field: Most studies use ad-hoc evaluations and do not share theoretical foundations, making it difficult to measure progress and compare the pros and cons of different techniques. Furthermore, while mechanistic understanding is frequently discussed, the basic causal units underlying these mechanisms are often not explicitly defined. In this article, we propose a perspective on interpretability research grounded in causal mediation analysis. Specifically, we describe the history and current state of interpretability taxonomized according to the types of causal units (mediators) utilized, as well as methods used to search over mediators. We discuss the pros and cons of each mediator, providing insights as to when particular kinds of mediators and search methods are most appropriate. We argue that this framing yields a more cohesive narrative of the field and helps researchers select appropriate methods based on their research objective. Our analysis yields actionable recommendations for future work, including the discovery of new mediators and the development of standardized evaluations tailored to these goals.

* Corresponding author.

Action Editor: Afra Alishahi. Submission received: 20 May 2025; revised version received: 9 August 2025; accepted for publication: 8 September 2025.

https://doi.org/10.1162/COLI.a.572

1. Introduction

To understand how language models (LMs) generalize, we must understand the causes of their behavior. These causes include inputs, but also the intermediate computations of the network; this survey is concerned with understanding these intermediate computations. How can we understand what an LM's computations represent, such that we can obtain a deeper algorithmic understanding of *how* and *why* models behave the way they do? For example, if a model decides to refuse a user's request, was the refusal mediated by an underlying concept of toxicity, by the presence of superficial correlates of toxicity (such as the mention of particular demographic groups), or some other unexpected variable? The former would be significantly more likely to robustly and safely generalize. These questions motivate the field of mechanistic interpretability (MI), which aims to understand how LMs arrive at particular behaviors by understanding the functional roles of their components.

We view mechanistic interpretability[1] as equivalent to extracting **causal graphs** explaining how intermediate LM computations mediate model outputs. This framing based in causality enables a new perspective of the field: Prior surveys have organized the field according to methodological differences, whereas we taxonomize work in the field according to the kinds of causal mediators—or types of nodes in the causal graphs—that a study uses (e.g., neurons, non-basis-aligned directions, attention heads, and the like). We start by describing causal mediation analysis and its role in MI (§2); we also define the goals of MI, and goal-specific criteria by which the success of a mediator can be measured (§2.1). We then contrast this survey with other MI surveys (§3), noting in particular the lack of surveys that center the mediator type. Following this, we present a history of mechanistic interpretability for neural networks more broadly (§4), from backpropagation to the beginning of the current wave of mechanistic interpretability research.

We survey common mediators (units of causal analysis) used in MI studies (§5), discussing the pros and cons of each mediator type. Should one analyze individual neurons? Combinations of neurons? Full activation vectors? More broadly, *what is the right unit of abstraction for analyzing and discussing neural network behaviors?* Any model component has pros and cons related to its level of granularity, whether it is a causal bottleneck, and whether it is natively part of the model (as opposed to whether it is learned via a separate module). The mediator type determines the kinds of methods that may be used to search over them; these search methods have their own pros and cons, which we use to organize the field in §6.

Finally, after surveying the field, we discuss practical considerations and implications for future work (§7). We point out mediators that have been underexplored, but have significant potential to yield new insights; propose future mediators that are likely to satisfy the criteria laid out in §2.1; and suggest ways to measure progress in mechanistic interpretability moving forward. Figure 1 summarizes the content of this survey.[2]

1 The meaning of "mechanistic interpretability" is debated; see Saphra and Wiegreffe (2024). We define the term as any study that aims to understand, explain, or modify a neural network's behavior by studying the model's internal components, such as its representations or weights.

2 Note that the methods discussed in this survey generalize beyond language models: Most can, in theory, apply to any neural network. Our focus on language models is motivated by (i) the centrality of mechanistic interpretability in natural language processing research in recent years, (ii) our primary expertise lying in this field, and (iii) a desire to keep the survey size tractable.

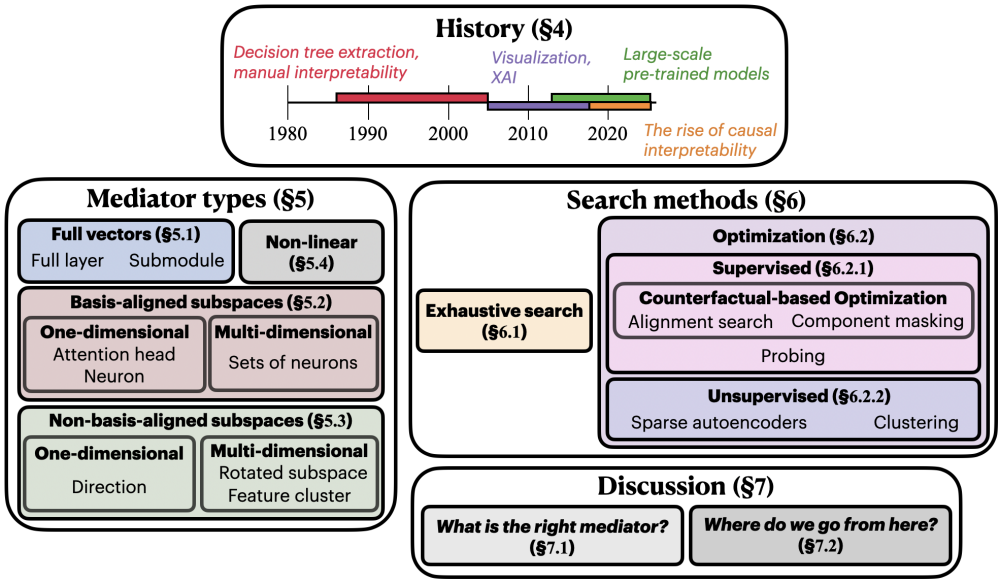

Figure 1
Outline of survey. We first define necessary causal terminology (§2) and contextualize our perspective with others' (§3). We subsequently give an overview of the history of mechanistic interpretability centered on units of causal analysis (§4). We then survey and categorize commonly used units of analysis and describe their strengths and weaknesses (§5), as well as methods for searching over them (§6). Finally, we discuss (§7) what we consider to be among the most important questions in mechanistic interpretability: What are the right causal abstractions for understanding and discussing the inner workings of LMs (§7.1)? What kinds of mediators and research will be needed to advance the field (§7.2)?

2. Preliminaries

The Counterfactual Theory of Causality. Lewis (1973) poses that a **causal dependence** holds iff the following condition holds:

> "An event E *causally depends* on C [iff] (i) if C had occurred, then E would have occurred, and (ii) if C had not occurred, then E would not have occurred."

Lewis (1986) extends the definition of causal dependence to be whether there is a **causal chain** linking C to E; a causal chain is a connected series of causes and effects that proceeds from an initial event to a final one, with potentially many intermediate events between them. This idea was later extended from a binary notion of whether the effect happens at all to a more nuanced notion of causes having influence on *how* or *when* events occur (Lewis 2000). Other work defines notions of cause and effect as measurable quantities (Pearl 2000); this includes direct and **indirect effects** (Robins and Greenland 1992; Pearl 2001), which are common metrics in causal interpretability studies.

Causal Abstractions in Mechanistic Interpretability. **Causal graphs** are fundamental abstractions in the causality literature (Pearl 2000). A causal graph $\mathcal{H}$ is a directed acyclic graph consisting of nodes V and directed edges E. A **node** $V_i \in V$ corresponds to an action or event; in neural networks, it can correspond to any component (or combination

Table 1
Table of notation grouped by whether the terms refer to concepts in neural networks, causal graphs, or both.

X	The input to a neural network or an exogenous variable in a causal graph. A specific value of X is denoted x.
Y	The output of a neural network or outcome node in a causal graph. A specific value of Y is denoted y.
$\mathcal{C}$	The computation graph of a neural network. Also used to refer to the neural network itself.
Z	A generic placeholder referring to any possible representation in a neural network between X and Y.
ℓ	A layer of $\mathcal{C}$.
$\mathbf{h}^\ell$	The representation vector at the output of layer ℓ.
$\mathbf{h}^{\ell\text{-MLP}}$, $\mathbf{h}^{\ell\text{-Attn}}$	The vector output of the MLP or attention block, respectively, at layer ℓ.
$\mathbf{h}_i^\ell$	A neuron in $\mathbf{h}^\ell$.
h_i^ℓ	A scalar activation of neuron $\mathbf{h}_i^\ell$.
A_i^ℓ	An attention head in layer ℓ.
$\mathbf{a}_i^\ell$	The vector output (attention score, equivalent to $Q \cdot K$ before the softmax) of A_i^ℓ.
d	The size of $\mathbf{h}^\ell$.
$\mathcal{H}$	A causal graph.
V	The set of nodes in causal graph $\mathcal{H}$ between X and Y.
V_i	A node in V. A specific value of V_i is denoted as v_i.
E	The set of edges in causal graph $\mathcal{H}$.
$E_{i,j}$	An edge in E drawn from V_i to V_j.

thereof), as described below and in §5. An **edge** $E_{i,j} \in E$ encodes a causal relationship between nodes, where the source is the **cause** and destination is the **effect**.[3] For example, if edge $E_{i,j}$ is drawn from one neuron V_i to another V_j, this indicates that V_i has significant counterfactual influence over V_j. Note that in a causal chain (a connected path in a causal graph), any node V_j can simultaneously function as both a cause of some downstream node V_k and an effect of a prior node V_i.

The abstraction of causal graphs extends naturally to neural networks: The computation graph of a model $\mathcal{C}$ is, by definition, the full causal graph that explains how inputs X (an exogenous variable in the causal graph) are transformed into a probability distribution over outputs Y (an outcome or leaf node in the causal graph). A causal node can correspond to any unit or intermediate representation Z produced by the network—for example, a neuron, a full layer, an attention head, or even some grouping of these. An edge encodes a causal relationship between any two nodes in the network, where the only restriction is that the source of the edge come before the destination in the computation graph. We summarize the notation we use for describing (components of) causal graphs and computation graphs in Table 1.

Each node can be viewed as a causal mediator that has some functional role in explaining how X is transformed into Y. The main challenge of MI studies, then, is to define a mapping from components Z in computation graph $\mathcal{C}$ to a high-level causal graph $\mathcal{H}$ consisting of nodes and edges (V, E) that explains how the model performs

3 In interpretability studies, it is also frequently required that edge E_i have strong influence on the final target behavior or output Y, rather than just the downstream intermediate component V_j.

some specific behavior. This entails deciding which of the components $\{Z_i\}_{i=1}^N$ should be filtered out from the nodes of the causal graph V, and optionally filtering out edges between these nodes.[4] This survey focuses on how the type of component $Z \in \mathcal{C}$ will affect one's findings; this is discussed in detail in §5.

In the causality literature, a **mechanism** is defined as a causal chain from cause X to effect Y (Salmon 1984; Pearl 2009). The mechanistic interpretability literature, while closely related to causal interpretability,[5] does not enforce this causally grounded definition of mechanism (cf. Miller, Chughtai, and Saunders 2024; Nanda, Lee, and Wattenberg 2023). The overlap between mechanistic and causal interpretability is significant, but not total: For example, sparse autoencoders (Bricken et al. 2023; Cunningham et al. 2024) are correlational, but are common in mechanistic interpretability, as they can reveal the concepts encoded in a model component without requiring the researcher to hypothesize what these concepts are ahead of time. Meanwhile, methods like LIME (Ribeiro, Singh, and Guestrin 2016b) involve interventions to input variables, but not the internals of a model. We believe that the causal definition of "mechanism" is an actionable one that makes the main challenge of mechanistic interpretability more precise—to reverse-engineer algorithmic understanding or control of model behaviors, where "algorithm" is equivalent to a task-specific causal graph $\mathcal{H}$ explaining how the model $\mathcal{C}$ performs a given task.

Counterfactual Interventions. In interpretability, "causal method" generally refers to a method that uses **counterfactual interventions** (Lewis 1973) to some part of the model or its inputs. Early interpretability work focused on interpreting model decision boundaries by intervening on the inputs X to the network (e.g., Ribeiro, Singh, and Guestrin 2016b), but contemporary work is primarily concerned with understanding which intermediary model components Z are responsible for some behavior Y—i.e., finding the model components $Z \in \mathcal{C}$ from the low-level computation graph to keep as nodes $V \in \mathcal{H}$ in the high-level causal graph (e.g., Geiger et al. 2021; Hanna, Liu, and Variengien 2023).

Causal mediation analysis (Pearl 2001) provides a framework for performing counterfactual interventions and interpreting their results. Given input X, output Y, and a causal graph consisting of many intermediate nodes V between X and Y, the causal influence of an intermediate node (mediator) $V_i \in V$ on the output Y is quantified as V_i's **indirect effect** (IE; Pearl 2001; Robins and Greenland 1992). This metric is based on the notion of counterfactual dependence, where one measures the difference in some target metric m before and after intervening on the mediator. In practice, m is typically (but not always) a function of the model's output Y.[6] More precisely, one starts by measuring m given a normal run of the model on input $X = x$, where V_i takes its natural value(s) v_i.

4 For some studies, it is sufficient to discover unordered sets of causally relevant components. In such cases, we assume that the causal graph is fully connected (where it is possible given the computation graph's directionality).

5 To be more precise, we will classify any study that aims to understand a model via understanding the roles of its components or its inner representations as mechanistic interpretability. We will classify any study that uses counterfactual interventions to a model's inputs and/or representations (resulting in states the model would not naturally have taken) as causal interpretability.

6 m can be any scalar value, including the value of an intermediate causal variable $\in V$, or even the aggregated values of multiple intermediate variables. For example, ACDC (Conmy et al. 2023) recursively computes a circuit by first finding components with high IE on Y (where m is derived from Y), and then finding components with high IE on those components (where m is now derived from intermediate components Z), and so on.

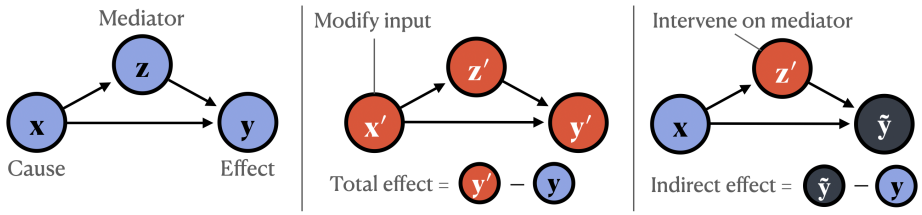

Figure 2
Visual summary of causal mediation analysis. Given input $X = x$ and the resulting output (model prediction) $Y = y$, and another input $X = x'$ that results in different output $Y = y'$, we can compute the total effect of changing x to x' as $y' - y$. In language models, there exist components Z that mediate the influence of X on Y. A common way to quantify the importance of Z is by measuring its **indirect effect** (Equation (1)), where, given $X = x$, one sets Z to some counterfactual value z'. In this figure, we set Z to what it would have been given x'; this results in $Y = \tilde{y}$. One can then measure the indirect effect as $\tilde{y} - y$.

We then compare this to m given $X = x$ where we intervene to set mediator V_i to some alternate (counterfactual) value v_i':[7]

$$\text{IE}(m; X, x; V_i, v_i, v_i') = m(V_i = v_i \mid X = x) - m(\text{do}(V_i = v_i') \mid X = x), \tag{1}$$

where "do" is an operation where the experimenter intervenes, setting a value in the experiment to one that it would not naturally have taken without the experimenter's involvement. See Figure 2 for an illustration of the indirect effect given a language model component Z.[8]

From Intervention to Evaluation. The tools of causal mediation analysis allow us to quantify the causal influence of a mediator on model behavior. However, identifying influential components is not sufficient for mechanistic understanding on its own. Depending on the goal of a study, we may define additional criteria by which the appropriateness or quality of a mediator may be judged. In the following subsection, we outline an evaluation framework that will help frame the following survey of mediator types used in existing research.

2.1 Evaluation Criteria

The goals of a mechanistic interpretability study determine how success should be defined. We identify three primary goals: (1) explaining model behavior, (2) verifying a mechanistic hypothesis, and (3) localization and editing. Below, we define criteria by which we can compare the utility of mediator types.

7 Appendix A surveys methods for sourcing v_i'. This value can come from alternate inputs where the answer is flipped, means over many inputs, or arbitrary constants (typically 0).

8 The choice of counterfactual v_i' significantly affects which components will be identified as influential. Common approaches for selecting v_i' include using values from alternative inputs x', applying a constant (e.g., 0), or adding noise ϵ, where ϵ is typically scaled according to the norm or variance of v_i. These are conceptually distinct operations that will uncover different kinds of components; we provide a more detailed discussion in Appendix A. We leave this to the appendix, as this survey is concerned with how results are affected by the choice of mediator Z—a topic distinct from the choice of intervention value v_i'.

First, we note four criteria that apply to each goal. At a high level, a good causal explanation explains the most general effects from the fewest non-trivial causes.[9] Thus, the ideal causal mediator is one that achieves a Pareto optimum between maximal **sparsity** (using as little of the computation graph as possible), **generality** (explaining the widest breadth of data), **selectivity** (explaining only the phenomenon of interest and as little else as possible), and **faithfulness** (retaining fidelity to the original model).

Explaining Model Behavior. Often, our goal is to produce qualitative and human-understandable insights about how models perform a certain behavior or task. What makes a good explanation? In mechanistic interpretability, this can be defined as one that maximizes the above metrics of sparsity, generality, selectivity, and faithfulness, as well as **human understandability**. Faithfulness is already a common metric in mechanistic explanations (Hanna, Pezzelle, and Belinkov 2024; Marks et al. 2025; Wang et al. 2023); measuring it is straightforward, as it is a generalization of causal mediation analysis to evaluation.[10] Sparsity is often called "minimality" (Mueller et al. 2025; Wang et al. 2023); this metric is not as common as faithfulness, but is becoming increasingly common in circuit analyses. It is typically either defined as the inclusion of as few redundant dependencies as possible (Wang et al. 2023), or as the inclusion of as few components as is necessary to achieve high faithfulness (Mueller et al. 2025). Generality is a rarer metric, as it requires evaluating on out-of-distribution examples that may need to be carefully curated. Recent work has argued that generality is a crucial metric, and has begun to define ways to measure it (Huang et al. 2025; Li et al. 2025). Selectivity is not often explicitly discussed nor measured, as it is assumed that maximizing faithfulness and sparsity should implicitly optimize selectivity; more work is needed to verify this assumption. The most difficult of these criteria to quantify is human understandability; well-trained sparse autoencoder features are typically easier to interpret than neurons, but it is not yet clear whether this metric can be measured in a reliable manner.

Verifying a Mechanistic Hypothesis. Sometimes, we already have a well-defined guess as to how a model accomplishes a task (a *mechanistic hypothesis*), and we would like to verify to what extent our hypothesis is accurate. Here, the criteria are similar as when explaining model behaviors: We would like to locate the smallest set of components or lowest-rank subspace that aligns best with the hypothesized explanation on the broadest possible data distribution. The primary difference is that the human-understandability of the mediator is of lesser importance for this goal, as understandability is a function of the human's mechanistic hypothesis and its accuracy, rather than the method used to align the model with the hypothesized causal variables. Criteria for success thus include sparsity, generality, and selectivity, as well as a modified form of faithfulness that we term **counterfactual faithfulness**. Counterfactual faithfulness measures whether the model's behavior changes in the expected manner when we perform counterfactual interventions to a specific part of the hypothesized mechanism; this is typically quantified as the interchange intervention accuracy (Geiger et al. 2021).

Localization and Editing. We often simply want to know where in a model some ability is implemented without necessarily understanding the components that implement it. This enables applications like model editing, steering, and parameter-efficient model

9 We add the "non-trivial" qualifier because one could trivially maximize the scope of an explanation by taking the entire computation graph $\mathcal{C}$ as a single mediator.

10 We refer readers to Hanna, Pezzelle, and Belinkov (2024) for further detail.

adaptation (e.g., with LoRAs applied to specific layers). Here, we want to maximize sparsity, generality, and selectivity, but we do not assign as great an importance to human understandability, and do not necessarily require a hypothesis for this to work well. If using localized components for an application like steering or model editing, then one should maximize primarily for *downstream task performance*, with sparsity and generality ideally being integrated into the downstream evaluation metrics. Note that high-quality evaluation is essential: One can maximize the efficacy of model editing via fine-tuning all model parameters, but this may produce unintended side effects if sparsity is ignored, and may not generalize well out-of-distribution if generality is ignored. Put simply, if there exist multiple mediator types that achieve similar test performance, then the best solution is likely that which then also attains the greatest sparsity and generality. Note that faithfulness is also important, but is not explicitly considered in model editing because it is trivially satisfied if the model's behavior changes after the editing operation.

3. Related Work

Causal interpretability surveys do not always focus on model internals, and mechanistic interpretability surveys do not necessarily require causal grounding. We give a brief overview of both types of survey here, contrasting them with ours. We also discuss recent tooling efforts that have accompanied the growing interest in mechanistic interpretability.

Mechanistic/Model-Internal Interpretability Surveys. Some surveys catalogue studies that aim to understand the latent representations of neural networks (Belinkov and Glass 2019; Belinkov 2022; Sajjad, Durrani, and Dalvi 2022); these have often called for more causal validations of correlational observations. More recent surveys tend to focus increasingly on giving practical overviews of how to use common methods for intervening on model internals (Ferrando et al. 2024; Rai et al. 2024). Others provide perspectives for understanding the trajectory of the mechanistic interpretability field (Räuker et al. 2023), and/or cataloguing the impacts of the field (Bereska and Gavves 2024). Notably, each of these surveys taxonomizes the field based on methodological differences; for example, a common contrast might be circuit analysis (discovering causal graphs of task-specific causal dependencies from model components, as in Wang et al. 2023; Conmy et al. 2023) vs. causal variable alignment methods (aligning model representations to human-provided concepts, as in Geiger et al. 2021; Wu et al. 2023), even if they both operate over the same kinds of components.

There is a gap here: Many of these methods implicitly deploy the same units of analysis, and thus benefit/suffer from the same fundamental pros/cons as a result. For example, it is not clear whether one would be better off discovering circuits over neurons, or attention heads, or abstractions over groupings of these. The same issue applies to any model-internals method. Thus, in this survey, we instead foreground the units of causal analysis that a study uses, as well as the way in which the study searches over those units, as primary factors in categorizing the study. We also ground the field in the language of causality, which grounds the goals of mechanistic interpretability: to discover causal subgraphs explaining how inputs are transformed into outputs.

Causal Interpretability Surveys. The survey by Moraffah et al. (2020) is a causal interpretability survey that categorizes various streams of causal interpretability research

according to the methods they utilize, though the studies they summarize are not necessarily based in the ideas of causal mediation analysis. The units of analysis were also not foundational to their organization, nor directly compared to each other. Other interpretability surveys (Subhash et al. 2022; Gilpin et al. 2018; Singh et al. 2024a) focus on methods for explaining the decisions of neural networks without causally grounding the explanation methods or focusing on model internals. Many causality-focused surveys are domain-specific, including areas such as cybersecurity (Rawal et al. 2024) and healthcare (Wu et al. 2024b). Some focus on particular domains; for example, in natural language processing (NLP), some focus on how causal inference can improve interpretability (Feder et al. 2022), or ways to explain (Danilevsky et al. 2020; Lyu, Apidianaki, and Callison-Burch 2024) or interpret (Madsen, Reddy, and Chandar 2022) neural NLP systems. Our survey is more specifically focused on studies that aim to understand NLP systems via their internal components—and even more specifically, those that do so via causal techniques such as interventions to those components.

Tools. Several libraries have recently been released to facilitate causal interpretability methods that involve interventions to model components. These tools can implicitly prioritize certain types of mediators over others. For instance, `pyvene` (Wu et al. 2024d) is designed specifically to aid in locating non-basis-aligned multidimensional subspaces via alignment search methods such as distributed alignment search and its successors (Geiger et al. 2024; Wu et al. 2023; Huang et al. 2024; Wu et al. 2024c). While it can also be used for other kinds of model interventions, this library could be particularly useful for those wishing to verify existing causal hypotheses. `TransformerLens` (Nanda and Bloom 2022) and libraries based on it (`Prisma`; Joseph 2023) are interpretability tools for examining Transformer-based neural networks. In these libraries, the interface is standardized across model architectures. This tends to encourage a focus on basis-aligned components such as neurons and attention heads, subspaces, and layers, as interventions to these mediators are natively supported. `NeuroX` (Dalvi, Sajjad, and Durrani 2023) similarly incentivizes neuron-level interpretability in particular. `NNsight` (Fiotto-Kaufman et al. 2025) and `Baukit` (Bau 2022) are more transparent interfaces that provide access to the underlying PyTorch model architecture, which allows for flexible modifications of the model's computation graph. Due to different naming conventions across model developers, this more transparent access may make it harder to generalize basis-aligned intervention code across architectures at first, but research on both basis-aligned and non-basis-aligned mediators is more accessible under this paradigm.

Note that this survey is intended more as a scientific review of the field rather than a practical guide to using these tools, so we have mainly discussed these toolkits with respect to the mediators that they enable working with. Some surveys such as those of Ferrando et al. (2024) and Rai et al. (2024) or code tutorials (Nanda and Bloom 2022; Wu et al. 2024d; Fiotto-Kaufman et al. 2025; Mohebbi et al. 2024) make hands-on practical introduction to particular methods their explicit purpose, without necessarily assuming (nor discussing the benefits nor drawbacks of) a particular categorization of methods, nor their units of analysis. See these surveys for more hands-on guides to implementing interpretability methods.

4. Lessons from the History of Interpretability

Causal interpretability techniques have existed since the beginning of deep learning. What distinguishes the current wave of mechanistic interpretability studies from past causal interpretability work? What actionable lessons can past work (which often used

very different methods and mediators from contemporary studies) teach us about analyzing intermediate model computations? We claim that the lens of causal mediation analysis (1) enables a novel and clear narrative of the trajectory of interpretability research; (2) links current issues in the field to longstanding issues that have existed since at least the 1980s; and (3) highlights actionable research directions. We focus primarily on (1) and (2) in this section, and return to (3) in §7.

Interpretability at the Beginning of Deep Learning. In 1986, Rumelhart, Hinton, and Williams published an algorithm for neural network backpropagation and an analysis of this algorithm. This enabled and massively popularized research into multi-layer perceptrons (MLPs)—now often called feedforward layers. That work arguably represents the first mechanistic interpretability study: The authors evaluated their method by inspecting each activation and weight in the neural network, and observing whether the learned algorithm corresponded to the human intuition of how the task should be performed. In other words, they reverse-engineered the algorithm of the network by labeling the rules encoded by each neuron and weight!

From the 1980s through the early 2000s, rule extraction via neuron-level activation and weight analysis remained popular. At first, this was a manual process: Networks were either small enough to be manually interpreted (Rumelhart, Hinton, and Williams 1986; McClelland and Rumelhart 1985) or interpreted with the aid of carefully crafted datasets (Elman 1989, 1990, 1991). For example, Elman (1990, 1991) found that recurrent neural networks were capable of capturing hierarchical semantic relationships, and were sensitive to syntactic context. Alternatively, researchers could prune the network (Mozer and Smolensky 1988; Karnin 1990) to a sufficiently small size to be manually interpretable. Later, researchers proposed techniques for automatically extracting rules (Hayashi 1990) or decision trees from neural networks (NNs) (Craven and Shavlik 1994, 1995; Krishnan, Sivakumar, and Bhattacharya 1999; Boz 2002)—often after the network had been pruned. At this point, interest in automated causal methods based on interventions had not yet been established, as networks were often small and simple enough to directly understand without significant abstraction.

Nonetheless, as the size of neural networks increased, the number of rules that could be encoded in a network increased. Thus, rule/decision tree extraction techniques could not generate easily human-interpretable explanations nor algorithmic abstractions of model behaviors beyond a certain size. This led to the rise of **visualization methods** in the 2000s, which became a popular way to demonstrate the complexity of phenomena that models had learned to encode. Visualizations of network inputs and outputs (Tzeng and Ma 2005) and interactive visualizations of model activations (Erhan et al. 2009) were valuable initial tools for generating hypotheses as to what kinds of concepts models could represent. While visualization research was generally *not* causal, this subfield would remain influential for interpretability research as neural networks scaled in size in the following decade.

Large-Scale Pre-trained Models. The 2010s were a time of rapid change in machine learning. In 2012, the first large-scale and widely adopted pre-trained neural network, AlexNet (Krizhevsky, Sutskever, and Hinton 2012), was released. Not long after, pre-trained word embeddings (Mikolov et al. 2013a, b; Pennington, Socher, and Manning 2014) became common in NLP, and further pre-trained deep networks followed (He et al. 2016). These were based on ideas from *deep learning*. This represented a significant paradigm shift: Formerly, each study would build ad-hoc models which

were not shared across studies, but which were generally more transparent.[11] After 2012, there was a transition toward using a shared collection of significantly larger and more capable—but also more opaque—models. This raised new questions on what was encoded in the representations of these shared scientific artifacts. The rapid scaling of these models rendered old neuron-level rule extraction methods either intractable or made their results difficult to interpret. Thus, interpretability methods in the early 2010s deployed scalable and relatively fast *correlational* methods, including visualizations (Zeiler and Fergus 2014) and saliency maps (Simonyan, Vedaldi, and Zisserman 2014). This trend continued into 2014–2015, when recurrent neural network–based (Elman 1990) language models (Mikolov et al. 2010) began to overtake non-neural statistical models in performance (Bahdanau, Cho, and Bengio 2015); for example, visualizing RNN and LSTM (Hochreiter and Schmidhuber 1997) hidden states was proposed as a way to better understand their incremental processing (Karpathy, Johnson, and Fei-Fei 2016; Strobelt et al. 2017).

At the same time, interpretability methods started to focus more on *explaining* model predictions.[12] The explainable AI field was and is extensive. One line of work designed supervised auxiliary (correlational) models to explain particular model predictions, such as LIME (Ribeiro, Singh, and Guestrin 2016a, b), Anchors (Ribeiro, Singh, and Guestrin 2018), and extensions like CLEAR that explicitly integrate notions of counterfactual fidelity to the output explanations (White and d'Avila Garcez 2020). These models learn local decision boundaries, or some human-interpretable simplified representation of a model's behavior. Other works interpreted predictions via feature importance measures like SHAP (Lundberg and Lee 2017). Influence functions (Koh and Liang 2017) traced the model's behavior back to specific instances from the training data. Another line of work sought to directly manipulate intermediate concepts to control model behavior at test time (Koh et al. 2020), or to decompose distributed representations into interpretable symbolic representations post hoc (Odense and Garcez 2020). The primary difference between these visualization-/correlation-/input-based methods and current methods lies in whether they prioritize black-box explanations or white-box explanations—that is, whether they explain model behaviors in terms of input/output relationships or require analysis of model internals, respectively. Black-box explanations allow us to generate hypotheses as to the types of *input concepts* that explain particular model predictions. In contrast, current work prioritizes *white-box explanations*—i.e., highly localized and causal explanations of *how* and in *which components of the computation graph* models perform a given behavior.

The years 2017–2019 featured perhaps the largest architectural shift (among many) in machine learning methods at this time: Transformers (Vaswani et al. 2017) were released and quickly became popular due to scalability and high performance. This led directly to the first successful large-scale pretrained language models, such as (Ro)BERT(a) (Devlin et al. 2019; Liu et al. 2019b) and GPT-2 (Radford et al. 2019). These significantly outperformed prior models, but it was unclear *why*—and at this

11 Many systems built before deep learning were based on feature engineering, and so the information they relied on was more transparent than in current systems.

12 There has classically been a distinction between *local* and *global* interpretability; local interpretability is concerned with explaining specific model predictions, whereas global interpretability is concerned with explaining a given model behavior in general across examples (Lipton 2018; Guidotti et al. 2018). Both styles of interpretability can be valuable, depending on one's research question. A benefit of causal mediation analysis is that it can encompass both styles. As recent work has tended to focus more on global interpretability, we devote more attention to this style of work, though we cite and acknowledge examples of local interpretability methods in this section.

scale, analyzing neural networks at the neuron level using past techniques had become intractable. This combination of high performance and little mechanistic understanding created demand for interpretability techniques that allowed us to see *how* language models had learned to perform so well.

Hence, correlational probing methods rose to meet this demand. In this approach, classifiers are trained on intermediate activations to extract some target phenomenon. Probing classifiers have been used to investigate the latent morphosyntactic structures encoded in static word embeddings (Köhn 2015; Gupta et al. 2015) or intermediate hidden representations in pre-trained language models—for example, in neural machine translation systems (Shi, Padhi, and Knight 2016; Belinkov et al. 2017; Conneau et al. 2018) and pre-trained language models (Hewitt and Manning 2019; Hewitt et al. 2021; Lakretz et al. 2019, 2021). However, probing classifiers lack consistent baselines, and the claims made in these studies were not often causally verified (Belinkov 2022). For instance, although an intervention may target a causal property of the task $V_i \rightarrow Y$, an alternative spurious property V_j may be picked up by the probe, which impedes causal claims about $V_i \rightarrow Y$ (Ravichander, Belinkov, and Hovy 2021). This encouraged researchers to search for more causally efficacious methods.

The Rise of Causal and Mechanistic Interpretability. Subsequently, 2017–2018 featured the first hints of our current wave of mechanistic interpretability, primarily based on interventions to neurons or full layers. Giulianelli et al. (2018) trained a probing classifier, but then used gradients from the probe to modify the activations of the network. Other studies analyzed the functional role of individual neurons in static word embeddings (Li, Monroe, and Jurafsky 2017) by forcing certain neurons on or off. Parallel developments in computer vision were influential: Bau et al. (2019b) found that interpretable concepts in the outputs of generative adversarial networks (Goodfellow et al. 2014) could be modified via interventions to specific neurons. The idea of manipulating neurons to steer behaviors was then applied to downstream task settings, such as machine translation (Bau et al. 2019a). These techniques were popularized in 2020 when Vig et al. (2020) proposed a method for assigning task-specific causal importance scores to specific neurons and attention heads by systematically computing each component's indirect effect (Equation (1)) on the model's output. It was an application of the counterfactual theory of causality (Lewis 1973, 1986), as well as Judea Pearl's causal mediation analysis framework (Pearl 2001, 2000). This enabled a new line of interpretability research that aimed to faithfully localize model behaviors to specific components—an idea that would become foundational to contemporary causal and mechanistic interpretability.

At the same time, however, researchers began to realize the significant performance improvements that could be gained by massively increasing the number of parameters and training corpus sizes of neural networks (Brown et al. 2020; Kaplan et al. 2020). Increasing model sizes resulted in more interesting subjects of study, but also rendered causal interpretability significantly more difficult. Thus, a primary challenge in interpretability has been to balance the often contradictory goals of (i) obtaining a causal understanding of how and why models behave in a given manner, while also (ii) designing methods that are efficient and scalable.

Presently, there exist many subfields of interpretability that propose and apply causal methods to understand which model components contribute to an observed model behavior (e.g., Elhage et al. 2021; Geiger et al. 2021; Conmy et al. 2023). There have also been efforts to discover more human-interpretable mediators by moving toward latent-space structures aside from (collections of) neurons (Cunningham et al.

2024; Bricken et al. 2023; Wu et al. 2023). These methods and the units they are based on form the focus of this survey.

The Lens of Causal Mediation Analysis. For much of the history of deep learning, layers (§5.1) and neurons (§5.2) were the basic unit of study in mechanistic interpretability. They are natural units of the model (i.e., require no external modules to discover), thus making them faithful to the model's computations by definition. In toy models, they can sometimes be human-interpretable, and their simplicity and small quantity enable researchers to exhaustively search over all of them (§6.1). Large-scale pre-trained models, however, contain far too many neurons for such methods to be tractable. Furthermore, their neurons are typically *not* straightforwardly interpretable because representations in neural networks are generally **distributed** (Hinton, McClelland, and Rumelhart 1986)—in other words, there is a many-to-many relationship between neurons and concepts. Thus, the field has turned to more sophisticated abstractions like *sets of neurons*, *attention heads* (§5.2), or even *non-basis-aligned subspaces* (§5.3) that require external modules (such as probes or sparse autoencoders) to locate (§6.2). Each of these mediator types has strengths and weaknesses. A mediator type also determines, to a large extent, the kinds of concepts that can be found, and the class of methods that can be used to find them. In the following section, we more precisely define these units of analysis, and compare their strengths and weaknesses (§5). Then, we give practical context to each mediator type by describing the methods that can be used to find them (§6), and the strengths and weaknesses thereof.

5. Mediator Types

In this section, we discuss different types of causal mediators in neural networks, and the pros and cons of each. Figure 3 visualizes a computation graph of a Transformer block in a typical language model, and units in the graph that are often used as mediators in mechanistic interpretability studies. In mechanistic interpretability, we

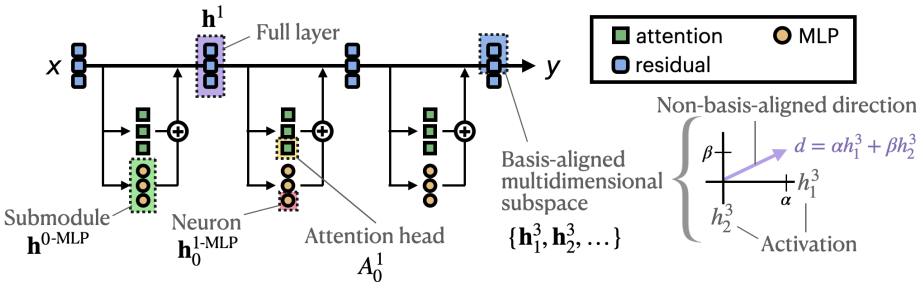

Figure 3
Visualization of common mediator types in neural networks. *Neurons* or *attention heads* are common units of analysis. *Full layer* and *submodule* vectors are more coarse-grained, but more easily enumerable. One can also implicate a *multidimensional subspace*, which could be neuron-basis-aligned (as in a group of neurons, pictured here) or non-basis-aligned. Non-basis-aligned mediators—e.g., arbitrary *directions* in activation space—have recently become a popular mediator type due to their monosemanticity. However, discovering non-basis-aligned mediators requires external modules such as classifiers, autoencoders, or other modifications to the original computation graph. Note that while this figure depicts a Transformer, many of the mediator types generalize to other architectures (the primary exception being attention heads).

Table 2
Summary of mediator types, the pros and cons of each, the search methods that are typically used to search over them, and examples of studies that use them.

Mediator type	Strengths	Weaknesses	Common search methods	Example studies
Full layers and submodules (§5.1)	Small search space. Some useful applications.	Difficult to interpret. Not well-suited to explaining model behavior.	Exhaustive search (§6.1), supervised probing (§6.2.1)	Hupkes, Zuidema et al. (2017), Giulianelli et al. (2018), Hewitt and Manning (2019), Geva et al. (2023), Meng et al. (2022)
Neurons and attention heads (§5.2)	Discrete and enumerable. Relatively fine-grained; sometimes enables model control and editing.	Search space not always tractable. Often not interpretable.	Exhaustive search (§6.1), optimization (§6.2)	Vig et al. (2020), Bau et al. (2020), Finlayson et al. (2021), Lakretz et al. (2019), Cao, Sanh, and Rush (2021)
Non-basis-aligned spaces (§5.3)	Fine-grained. Interpretable. Enables precise control of NN behaviors.	Non-enumerable. Typically requires optimization; sensitive to training setup and random variance. May not be faithful to original model.	Optimization (§6.2): supervised probing (§6.2.1), unsupervised methods such as sparse autoencoders (§6.2.2)	Wu et al. (2023), Bricken et al. (2023), Cunningham et al. (2024), Marks and Tegmark (2023), Marks et al. (2025), Ravfogel et al. (2021)

often do not want to treat the full computation graph as the final causal graph, as it is large and difficult to directly interpret. Thus, we typically want to build higher-level causal abstractions that capture only the most important mediators, and/or where each causal node is human-interpretable.

In this section, our primary questions are: What kinds of model components can be used as mediators? What are the strengths and weaknesses of using particular kinds of components as mediators? In this subsection, we define each mediator type. Then, in the following mediator-specific subsections, we discuss how they have been used in interpretability, and their pros and cons. Table 2 summarizes mediator types, their strengths and weaknesses, and search methods commonly used to identify each.

One possible mediator type is a **full layer**—typically the output activations or *hidden state* $\mathbf{h}^\ell$ of a specific layer ℓ (§5.1). Each index $\mathbf{h}_i^\ell$ is a **neuron** that can take some activation $\mathbf{h}_i^\ell = h_i^\ell$.[13] One can also use the output vector of an intermediate **submodule** within the layer (e.g., an MLP), rather than the output of the whole layer. For example, in Transformers (Vaswani et al. 2017),[14] a layer typically consists of two submodules: an MLP and an attention block, which can be arranged either sequentially or in parallel. The outputs of these submodules $\mathbf{h}^{\ell\text{-MLP}}$ and $\mathbf{h}^{\ell\text{-Attn}}$ are also activation vectors, so we will refer to their individual dimensions as neurons as well.[15]

One can also group neurons into sets, and use a neuron set as a single mediator (§5.2). A set of neurons (possibly of size 1) $\{\mathbf{h}_i^\ell, \mathbf{h}_j^\ell, \ldots\}$ from a vector $\mathbf{h}^\ell$ is referred to as a **basis-aligned subspace** of $\mathbf{h}^\ell$. A one-dimensional basis-aligned subspace is equivalent to a neuron; for clarity, we will use basis-aligned subspace primarily to refer to multidimensional spaces (sets of neurons of size > 1).

Basis alignment refers to whether concept representations are aligned with specific dimensions of $\mathbf{h}^\ell$ (as contrasted with weighted combinations of dimensions). Basis alignment is a key concept: If a mediator V is aligned with the latent space basis vectors defined by a (group of) neuron(s) $\{\mathbf{h}_i^\ell, \mathbf{h}_j^\ell, \ldots\}$, then we can discover it via

13 In other words, we use "neuron" to refer to any basis-aligned direction in activation space.

14 Transformers are currently the dominant architecture for language models; as such, most work in this space focuses on this architecture. However, our ideas are presented in a general way that will also apply (with minor modifications) to other neural network–based architectures, such as recurrent neural networks (Mikolov et al. 2010) and state space models (Gu, Goel, and Re 2022; Gu and Dao 2024).

15 Using the same notation emphasizes that these are mediators of the same level of granularity. However, we acknowledge that this obscures that neurons in different locations often encode different types of features.

non-parametric methods. For example, it is straightforward to exhaustively search over and intervene on individual neurons; it is less tractable, but still theoretically possible, to enumerate all $2^n - 1$ possible combinations of neurons without using any additional parameters. However, causally relevant mediators are not guaranteed to be aligned with neurons in activation space; indeed, recent work has found human-interpretable features in arbitrary directions that are *not* aligned to neuron bases (Elhage et al. 2022b; Bricken et al. 2023).

Thus, in recent studies, it is common to study **non-basis-aligned spaces** (§5.3). Each dimension in a non-basis-aligned subspace can be defined as a weighted linear combination of neuron activations. For example, to obtain a non-basis-aligned **direction**,[16] we could learn coefficients α and β to weight the activations of neurons $\mathbf{h}_i^\ell$ and $\mathbf{h}_j^\ell$ (optionally with a bias term $\mathbf{b}$):

$$\mathbf{d} = \alpha \cdot \mathbf{h}_i^\ell + \beta \cdot \mathbf{h}_j^\ell + \ldots + \mathbf{b}, \tag{2}$$

where neurons $\mathbf{h}_i^\ell$ and $\mathbf{h}_j^\ell$ are allowed to take their natural activations given some input x, but α and β remain fixed across inputs. Note that α and β are *not* part of the original computation graph $\mathcal{C}$. This means that discovering non-basis-aligned directions often requires external modules that weight components from the computation graph in some way—e.g., classifiers or autoencoders.

The primary trade-off between these mediator types is their granularity and quantity. This section proceeds from coarser to finer granularity and in increasing quantity. Broadly speaking, finer-grained mediators are more likely to optimize sparsity and selectivity, but are more difficult to search over, and may not be faithful to the original model if optimization is required to locate them. Coarser-grained mediators are more likely to optimize generality and are generally easier to search over, but tend to sacrifice selectivity and sparsity.

5.1 Full Layers and Submodules

Full layers $\mathbf{h}^\ell$ and submodules $\mathbf{h}^{\ell\text{-MLP}}$, $\mathbf{h}^{\ell\text{-Attn}}$ are relatively coarse-grained mediators; there exist only ℓ of them in a model $\mathcal{C}$.[17] Early probing classifiers studied the information encoded in full layers (Shi, Padhi, and Knight 2016; Hupkes, Veldhoen, and Zuidema 2018; Belinkov et al. 2017; Conneau et al. 2018; Liu et al. 2019a; Hewitt and Manning 2019; Giulianelli et al. 2018), and recent studies that leverage classifiers as part of causal techniques still frequently do the same (e.g., Elazar et al. 2021; Marks and Tegmark 2023; Li et al. 2023). This makes layers a natural mediator for exploratory interventions where using more fine-grained mediators is infeasible, as in Conmy et al. (2023), or where broad characterizations of information flow are sufficient, as in Geva et al. (2023) or Sharma, Atkinson, and Bau (2024).

Full layers are rare mediators in mechanistic interpretability. This is because $\mathbf{h}^\ell$ is a **causal bottleneck**, such that all information in the model must pass through it. For example, if we run input x_i through a model, and intervene on $\mathbf{h}^\ell$ to set it to what it would have been given another input x_i', the model's output would be identical to the case where we simply change the input to x_i' and perform no interventions. That

16 We use "direction" to refer to one-dimensional spaces.

17 Henceforth, we will refer to each of these with $\mathbf{h}^\ell$ for concision, as they are equivalent with respect to what kinds of search methods can be applied to them.

said, interventions to full layers were used in a pruning study where the motivation was not interpretability (Sajjad et al. 2023); we believe that a more refined intervention technique has the potential to inform our understanding of which model *regions* (as opposed to components) are more responsible for certain behaviors (e.g., Lad, Gurnee, and Tegmark 2024).[18] Past work has used coarse-grained methods based on full layers to investigate factual recall in language models (Geva et al. 2023), and to update these factual associations (Meng et al. 2022, 2023).

The primary advantage of using full layers and submodules as mediators is their small quantity and broad scope of information (high generality). This means that even slow or resource-intensive methods will generally be easy to apply to all layers. In some cases, this is sufficient. However, an obvious disadvantage is that full layers are generally difficult to understand, as they are not particularly sparse nor selective: Even if we know that a concept is encoded in a given layer, it is unclear precisely how this information is encoded, composed, or used, and how we might intervene on it without affecting other concepts (Conmy et al. 2023). Thus, layers and submodules will generally be too coarse to explain model behavior or verify mechanistic hypotheses; under these goals, one is better off using full layers as tools to narrow the search space for harder-to-localize mediators (e.g., Brinkmann et al. 2024; Geva et al. 2023). However, if one's goal is localization or model editing, then full layers may be sufficient (Meng et al. 2022, 2023; Sharma, Atkinson, and Bau 2024; Gandikota et al. 2023, 2024).

5.2 Basis-Aligned Subspaces

Neurons. Compared with a full layer or submodule $\mathbf{h}^\ell$, a neuron $\mathbf{h}_i^\ell$ represents a more fine-grained component that could feasibly represent an individual concept (though we discuss below that this is not often the case due to polysemanticity). A neuron can be considered the smallest meaningful unit in the computation graph $\mathcal{C}$; the neuron's activation h_i^ℓ is a scalar corresponding to a single dimension (1-dimensional subspace) of a hidden representation vector. Each neuron can sometimes be distinguished from others based on its functional role in the network; for instance, Bau et al. (2019a) locate neurons in a machine translation model responsible for detecting or generating items of a particular tense, gender, or number, and causally verify their roles by intervening on their activations in a targeted manner.

Neurons are a natural choice of mediator, as they are both fine-grained (sparse) and easy to exhaustively iterate over (see §6.1); there are $O(\ell \cdot d)$ of them, where d is size of the activation vector $|\mathbf{h}^\ell|$.[19]

However, a major disadvantage of neuron-based interpretability methods is *polysemanticity*. Individual neurons are often polysemantic—i.e., they respond to multiple seemingly unrelated concepts simultaneously (Arora et al. 2018), such that a neuron may be *sparse* but not *selective* nor *human-understandable*. This gives them relatively low utility for explaining model behaviors, but they can be useful for verifying mechanistic hypotheses (Geiger et al. 2021) or localizing model behaviors (Vig et al. 2020). For example, if the same neuron were sensitive to capitalized words, animal names, one-digit

18 Relatedly, one could intervene on submodules, such as $\mathbf{h}^{\ell\text{-MLP}}$, and observe how this impacts the model, or the features present in the final layer output $\mathbf{h}^\ell$.

19 We use big-O notation because the exact number will depend on whether one includes just the activation vectors at the end of a layer, or additionally includes the vectors output by the MLP and/or attention blocks (among other possible vectors). Also, if looking at intermediate MLP neurons, the number is typically $e \cdot |\mathbf{h}^\ell|$, where e is some constant expansion factor, typically in $[1.5, 4.0]$.

numbers, among other phenomena, and a researcher were to inspect the activations of that neuron, it would be difficult to disentangle each of these individual roles. Elhage et al. (2022b) investigate polysemanticity and suggest that neural networks represent features through linear superposition, where they represent features along non-basis-aligned linear subspaces, resulting in interpretable units being smeared across multiple neurons. In other words, in an activation vector $\mathbf{h}^\ell$ of size $|\mathbf{h}^\ell| = d$, a model can encode $k \gg d$ concepts as directions (Park, Choe, and Veitch 2023), such that only a sparse subset of concepts are active given a particular input.

Basis-Aligned Multidimensional Subspaces. The computations of a neuron $\mathbf{h}_i^\ell$ are often not independent: *Sets of* neurons can compose to encode some concept. For example, in language models, subsets of neurons $\{\mathbf{h}_i^\ell, \mathbf{h}_j^\ell, \ldots\} \in \mathbf{h}^\ell$ can be implicated in encoding gender bias (Vig et al. 2020), and implementing latent linguistic phenomena (Finlayson et al. 2021; Mueller, Xia, and Linzen 2022; Bau et al. 2019a; Lakretz et al. 2019). Thus, some early mechanistic interpretability work used heuristic-based searches over sets of neurons responsible for some behavior (e.g., Bau et al. 2019b; Vig et al. 2020; Cao, Sanh, and Rush 2021; Antverg and Belinkov 2022). This is a generalization of individual neurons as mediators, where multiple dimensions in activation space are intervened upon simultaneously.

Sets of neurons have strictly more expressive power than individual neurons, and thus have the potential to explain model behavior more broadly than finer-grained mediators. In other words, one can conceptualize neuron groups as trading off some sparsity for increased generality. If a concept is encoded across multiple neurons, then neuron groups may also enable more human-interpretable interventions than one-dimensional subspaces. Despite this, basis-aligned multidimensional subspaces are not commonly studied. This is for two main reasons: (1) There is a combinatorial explosion when we are allowed to search over arbitrarily-sized sets of neurons; if using exhaustive search, this increases the number of required forward passes from $O(\ell \cdot d)$ to $O(2^{\ell \cdot d})$, which makes this intractable. (2) Furthermore, interpretable concepts are not guaranteed to be aligned to neuron bases, meaning that groups of neurons still do not directly address the problem of polysemanticity (Morcos et al. 2018; Chughtai, Chan, and Nanda 2023; Wang et al. 2023).

Attention Heads. Similar to neurons, attention heads are fundamental components of Transformer-based neural networks. They mediate the flow of information between token positions (Vaswani et al. 2017); thus, using attention heads as units of causal analysis can help us understand how models synthesize contextual information (Ma et al. 2021; Neo, Cohen, and Barez 2024) to predict subsequent tokens (Wang et al. 2023; Hanna, Liu, and Variengien 2023; Prakash et al. 2024; García-Carrasco, Maté, and Trujillo 2024; Brinkmann et al. 2024).

Each head A_i^ℓ in layer ℓ can be understood as an independent operation contributing a vector output $\mathbf{a}_i^\ell$; the outputs of all heads are concatenated and projected to form the output of the attention block $\mathbf{h}^{\ell\text{-Attn}}$, which is then added into the residual stream $\mathbf{h}^\ell$.[20] For example, some heads specialize in syntactic relationships (Chen et al. 2024a),

20 This is the *residual stream perspective* (Elhage et al. 2021) of Transformers, which has been adopted in recent interpretability research (Ferrando et al. 2024). The *residual stream perspective* suggests that the residual stream, which comprises the sum of the outputs of all the previous layers and the original input embedding, acts as a passive communication channel through which the MLP and attention submodules route the information they add.

others in semantic relationships such as co-reference (Vig et al. 2020), and others still in maintaining long-range dependencies in text (Wu et al. 2024a). Attention heads have also been directly implicated in acquiring the ability to perform in-context learning (Olsson et al. 2022; Brown et al. 2020), or to detect and encode functions in latent space (Todd et al. 2024; Feucht et al. 2025).

Attention heads are attractive mediators because they are easily enumerable: There are generally far fewer attention heads $O(\ell \cdot |A^\ell|)$ than neurons $O(\ell \cdot d)$ in a model, as $|A^\ell| \ll d$ in typical language models. Attention heads also track multi-token relationships. However, in contrast to the activation h_i^ℓ of a neuron, the output of an attention head $\mathbf{a}_i^\ell$ is multidimensional. Thus, it is difficult to directly interpret the full set of functional roles a single head might have: Attention heads are almost always polysemantic, so one cannot typically determine the function(s) of an attention head solely by observing its outputs (Janiak, cmathw, and Heimersheim 2023).[21] It has additionally been observed that intervening on an attention head can cause other attention heads to compensate, which further complicates causal analyses (Jermyn, Olah, and Henighan 2023; Wang et al. 2023; McGrath et al. 2023).[22] To summarize, attention heads are easier to exhaustively search over than (sets of) neurons, but have the same issues of low selectivity and human-interpretability.

5.3 Non-basis-aligned Spaces

Non-basis-aligned Multidimensional Subspaces. Due to their polysemanticity, neurons, attention heads, and sets thereof do not necessarily correspond to cleanly interpretable concepts. For example, individual neurons typically activate on many seemingly unrelated inputs (Elhage et al. 2022b), and this issue cannot be cleanly resolved by adding more dimensions. This is because the features may actually be encoded in directions or subspaces that are *not aligned to neuron bases* (Mikolov et al. 2013a; Arora et al. 2016); §6.2 defines and visualizes this concept in more detail.

To overcome this disadvantage, one can generalize causal mediators to include arbitrary *non-basis-aligned* subspaces of $\mathbf{h}^\ell$. This allows us to capture more sophisticated causal abstractions encoded in latent space, such as causal nodes corresponding to greater-than relationships (Wu et al. 2023), or equality relationships (Geiger et al. 2024). Common methods for locating these are discussed in §6.2.

The primary advantage of considering an arbitrary subspace as a mediator is its expressivity *and* precision: Subspaces often capture distributed abstractions that are not cleanly aligned to specific neurons. However, they are generally more difficult to locate than basis-aligned components or lower-dimensional directions, as we are required to have specific hypotheses as to how a model accomplishes a task and access to minimal pairs that isolate the target concept. Some optimization-based procedure is also usually required. Thus, non-basis-aligned multidimensional subspaces are generally more selective and human-interpretable than basis-aligned subspaces, and can maintain similar

21 However, there is initial evidence that some dimensions of an attention head's output can be meaningfully explained (Merullo, Eickhoff, and Pavlick 2024a, b; Hu et al. 2025). Thus, by decomposing the vector output of a head into smaller subspaces or even individual neurons, it may be easier to explain the set of functional roles of a given head.

22 This phenomenon where downstream components only have causal relevance after an upstream component has been removed is sometimes called **preemption** in the causality literature (Mueller 2024), or the "Hydra effect" in mechanistic interpretability (McGrath et al. 2023). Preemption is not limited to attention heads; future work should analyze how common preemption is between other types of components, such as MLP submodules.

sparsity and generality; however, they may sacrifice faithfulness, as the optimized component can in theory learn concepts that were not in the model itself. This, combined with the strict data requirements, make them useful primarily for verifying mechanistic hypotheses, but less useful for explaining model behaviors or localization/editing.

Directions. A recent line of work aims to automatically identify specific directions (one-dimensional non-basis-aligned objects) $\mathbf{d}$ that correspond to monosemantic concept representations; see Equation (2) for a definition of $\mathbf{d}$. Identifying and labeling these monosemantic model abstractions (often called **features**; Bricken et al. 2023; Cunningham et al. 2024; Huang et al. 2024) can reveal units of computation the model uses to solve tasks in a way that is often easier for humans to interpret.[23]

There is also initial evidence that these directions may enable fine-grained model control (Panickssery et al. 2024; Marks et al. 2025; Tigges et al. 2023). Past work has found initial signs that basis-aligned directions could be leveraged to edit (Meng et al. 2022) or steer (Subramani, Suresh, and Peters 2022; Turner et al. 2023) model behavior, whereas more recent work has found that steering using non-basis aligned directions is both more effective and more precise (Marks et al. 2025; Arora, Jurafsky, and Potts 2024; Wu et al. 2025). For example, there is work that uses linear probes to understand the effects of a direction on the model behavior (Chen et al. 2024b; Ravfogel et al. 2020; Elazar et al. 2021; Ravfogel et al. 2021; Lasri et al. 2022; Marks and Tegmark 2023), as well as work that uses these directions to steer model behaviors—e.g., by minimizing (Marks et al. 2025; Cunningham et al. 2024) or amplifying (Templeton et al. 2024) directions corresponding to fine-grained concepts, such as typically female names or the Golden Gate Bridge.

Nonetheless, directions have key disadvantages. The search space over possible non-basis-aligned directions is infinite, making it impossible to exhaustively search over them. To discover them, we are generally *required* to modify the computation graph in some way to obtain some discrete search space—for example, by learning coefficients on each neuron, as in Equation (2), requires learning new parameters. Regardless of the method, optimization algorithms will introduce confounds due to their stochastic nature. In short, non-basis-aligned directions have significant advantages in human-interpretability, selectivity, and sparsity over basis-aligned mediator types, and their data requirements are less strict than non-basis-aligned multidimensional subspaces. This makes them a good starting point if one's goal is to explain model behaviors. Nonetheless, faithfulness is likely to be worse than that of basis-aligned components, given that optimization is required; generality may also suffer because of reconstruction error, and because the mediators are so fine-grained.

5.4 Non-linear Mediators

Non-basis-aligned directions and subspaces can be the most sparse and selective *linear* mediator types. However, recent work has demonstrated that some features in language models can be represented non-linearly. For example, there exist features that are encoded as vector *magnitudes* in any direction (Csordás et al. 2024). Past work has similarly found that many concepts can be more easily extracted using non-linear

23 Note that these directions are not necessarily subspaces of activation space: There are often non-linearities used in computing them, even though the vectors in activation space are involved in computing the directions. Therefore, we will refer to any one-dimensional space as a **direction**, but do not require it to be a subspace of activation space.

probes (Liu et al. 2019a), and that non-linear concept erasure techniques tend to outperform strictly linear techniques (Iskander, Radinsky, and Belinkov 2023; Ravfogel et al. 2022). However, in causal and mechanistic interpretability, most work has thus far tended toward using linear representations as units of causal analysis. Thus, there is significant potential in future work for systematically locating non-linearly-represented features—e.g., using group sparse autoencoders (Theodosis and Ba 2023), which could isolate multiple directions simultaneously, and/or probing and clustering techniques to identify multidimensional features (Engels et al. 2025). Non-linear features have not been extensively studied, despite their expressivity; we therefore advocate investigating these mediators in §7.

6. Searching for Task-Relevant Mediators

Once one has selected a task and a type of mediator, how does one identify task-relevant mediators of that type? The answer depends largely on the type of mediator chosen. If a given mediator type is finite in number—as is the case for sets of model components $\mathbf{h}^\ell$ such as neurons, layers, and submodules—one could perform an exhaustive search over all possible mediators, choosing which to keep according to some metric; §6.1 discusses this approach. However, other mediator types, including non-basis-aligned directions and subspaces, are continuous, rendering an exhaustive search impossible. There are two common solutions to the problem of continuous mediator search spaces: (i) use optimization to search this space, or (ii) narrow the space into an enumerable discrete set. Both are discussed in §6.2. Table 2 (§5) summarizes the kinds of mediators that tend to be paired with particular search methods. We do not discuss runtimes in this section, but see Appendix B for an overview of the relative computational requirements for each mediator/search method combination.

6.1 Exhaustive Search

Suppose we are given a neural network with a finite set of candidate mediators $\{Z_i\}_{i=1}^N$, such as the set of all neurons $\{\mathbf{h}_0^0, \ldots, \mathbf{h}_d^\ell\}$. One way to identify task-relevant mediators from this set is to assign each mediator Z_i a task-relevancy score $S(Z_i)$ and then select the mediators with the top scores. S is typically the indirect effect (IE; Pearl 2001; Robins and Greenland 1992), as defined in Equation (1).[24] Computing this generally entails iterating over each mediator Z_i, setting its activation to some counterfactual value (either from a different input where the answer is flipped, or a value that destroys the information in the neuron, such as its mean value), and measuring how much this intervention changes the output. For example, Vig et al. (2020) and Finlayson et al. (2021) perform counterfactual interventions to the activation of each neuron in an LM, measuring how much each intervention changes the probability of correct completions. This metric is based on the notion of counterfactual dependence, where we measure the difference in some output metric m before and after intervening on a given component Z_i.

24 Other causal metrics include the **direct effect**, which measures the direct influence of the input on the output behavior except via the mediator. While more rarely used, it can be a helpful metric in tandem with indirect effects, as in Vig et al. (2020). There is also the **total effect**, which is the impact of changing the input on the model's output behavior. Note that the total effect does not directly implicate any particular component in model behavior, as it depends only on the input.

Exhaustive searches have many advantages: Their results are comprehensive, causally efficacious, and relatively conceptually precise if our mediators are fine-grained units like neurons. We are also not required to have a pre-existing causal hypothesis as to how or where in its computations a model performs a task: We may simply observe how interventions to a model component changes the model's behavior or the probability of some continuation. Because of these advantages, this method is common when we have a finite set of mediators—for example, in neuron-based analyses (Vig et al. 2020; Geiger et al. 2021; Finlayson et al. 2021) or attention-head-based analyses (Vig et al. 2020; Conmy et al. 2023; Syed, Rager, and Conmy 2023).

However, exhaustive searches also have two significant disadvantages. The most obvious is that, in exact form, an exhaustive search requires $O(N)$ forward passes, where N is the number of mediators. This does not scale efficiently as models scale, both because the number of components increases and because the computational cost of inference scales with model size. This may be why exhaustive searches have not often been extended to *sets of* neurons or heads, as this results in a combinatorial explosion in the size of the search space such that the number of required forward passes increases to $O(2^N)$. Searches over sets of components can be approximated using greedy or top-k approaches, as in Vig et al. (2020), but this is not guaranteed to find the best solution to the problem of assigning causal credit to groups of components.

To overcome these challenges, gradient attributions have become common. These can be conceptualized as fast linear approximations to causal mediation analysis. As they are local linear approximations, gradient attributions are technically not causal[25] and not always accurate, but they have far better asymptotic runtime than causal mediation analysis. Example methods include attribution patching (Kramár et al. 2024; Syed, Rager, and Conmy 2023), and improved versions thereof inspired by integrated gradients (Sundararajan, Taly, and Yan 2017; Marks et al. 2025; Hanna, Pezzelle, and Belinkov 2024). These techniques usually entail backpropagating from some target metric m; this is typically the probability of the correct next token or correct label y, or the probability difference between y and some minimally differing incorrect output y'. This yields $\frac{\partial m}{\partial h_i^\ell}\Big|_x$ (the gradient of m with respect to the activation of neuron h_i^ℓ given input x), which can be conceptualized as a local estimate of the slope of m with respect to h_i^ℓ. If we multiply this slope by the difference in h_i^ℓ and a counterfactual value $h_i^{\ell\prime}$,[26] then we can obtain a linear approximation of how much changing h_i^ℓ would have changed m—in other words, a linear approximation of the indirect effect (Equation (1)). We can perform this attribution for all h_i^ℓ in parallel using $O(1)$ forward and backward passes.

The second and more difficult disadvantage to overcome is that exhaustive search constrains us to finite sets of mediators. Thus, this approach will not be possible if the search space is continuous (infinitely large). This is a key motivation behind the methods in the following subsection.

25 Gradient attributions provide a scalar value whose magnitude can be interpreted as a local linear approximation of the model output's sensitivity to the component. This is conceptually related but distinct from causal analysis—e.g., because it is sensitive to confounding, does not capture non-linear effects, and does not directly test a causal model nor counterfactual hypothesis.

26 A common counterfactual value includes $\mathbb{E}_{x \in X}[h_i^\ell]$; setting h_i^ℓ to this mean is known as a mean ablation. Setting h_i^ℓ to 0 is known as a zero ablation. See Appendix A for more details.

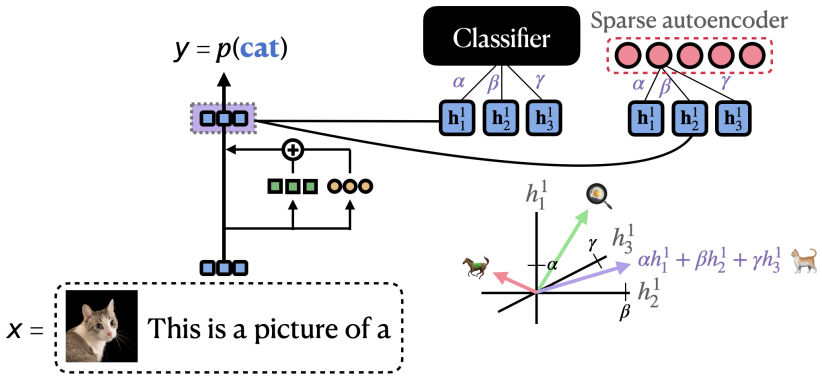

Figure 4
Neurons are not guaranteed to encode interpretable features. If non-basis-aligned directions encode the features of interest, then a neuron may activate on many different features that are non-orthogonal to its basis. Locating non-basis-aligned mediators requires components in addition to the model's computation graph that encode the coefficients on each activation. One can, for example, obtain these coefficients via supervised optimization with probing classifiers (§6.2.1) or unsupervised optimization with sparse autoencoders (§6.2.2). Note that optimization-based techniques sometimes introduce non-linearities, meaning that the discovered directions will not necessarily be a subspace of activation space.

6.2 Optimizing Over Large Spaces of Mediators

For some types of mediators, the collection of candidate mediators is continuous or far too large to exhaustively search over; this precludes using methods described in §6.1. To search over large but enumerable sets, some researchers utilize modified versions of exhaustive search, including greedy search methods (Vig et al. 2020) or manual searches (Wang et al. 2023). For continuous spaces, however, interpretability researchers generally use optimization. We taxonomize these optimization problems based on whether they require the interpretability researcher to manually select and incorporate task-specific information into the loss function (supervised methods; §6.2.1) or not (unsupervised methods; §6.2.2). We illustrate the intuition behind optimization-based search in Figure 4.

6.2.1 Supervised Mediator Searches. By **supervised mediator searches**, we mean parametric approaches that require labeled task data $\mathcal{D}$ and/or human-provided hypothesized high-level causal graphs $\mathcal{H}$. For example, these methods might require the researcher to propose candidate intermediate concepts $V_i \in V$ that they expect the model to use in performing some task, or a candidate mechanism (a subset of $\mathcal{H}$) by which the model might complete the task. Others might simply require labeled data $(x, y) \in \mathcal{D}$ for training classifiers.

Supervised Probing. In supervised probing approaches, the researcher hypothesizes that the model represents some concept $V_i \in V$, designs some labeled dataset consisting of $\{(x_i, y_i)\}_{i=1}^{n} \in \mathcal{D}$ to isolate the concept, and trains a probe to locate some direction(s) in Z that *correlate(s)* with the concept. A probe is generally formulated as a classifier Π that maps from model representations Z given input x_i to probabilities over class predictions: $p_\Pi(y_i \mid x_i, Z)$. Generally, Z is a representation vector $\mathbf{h}^\ell$ at the end of a particular

layer ℓ, and the probe updates its weights over all possible subspaces therein to attend to signals that are predictive of the labels. Many papers use probing classifiers (Belinkov and Glass 2019), though most do not validate the causal efficacy of the probe (Belinkov 2022). A drawback of this is that certain activation subspaces are often *correlated* with a concept without *causally mediating* the concept. This means that probing can return many false positives (Hewitt and Liang 2019; Elazar et al. 2021; Ravichander, Belinkov, and Hovy 2021; Amini et al. 2023; Belinkov 2022).

Thus, much recent work complements supervised probing approaches with additional checks of causality. For example, one can apply causal mediation analysis to the directions $W_{\Pi}^{\top}\mathbf{h}^{\ell}$ identified by supervised probing (Marks and Tegmark 2023; Nanda, Lee, and Wattenberg 2023)—i.e., one can measure whether the IE (Equation (1)) of the direction used by the probe is significantly different from zero. This would fall under non-basis-aligned mediators. One can also backpropagate from the classifier to modify the behavior of the model (Giulianelli et al. 2018) or to generate counterfactual representations (Tucker, Qian, and Levy 2021), which is more akin to a full-layer intervention. One can also directly compare the probe's predictions to a causally grounded probe (Amini et al. 2023).

Another line of work uses the directions discovered by probes to *guard* or *erase* information about a particular concept from the model's representations. For example, a direction in a model's activation space that is most predictive of the target concept can be nullified via orthogonal projections, such that the model can no longer use the information (Ravfogel et al. 2020); this process can be repeated until linear guarding is achieved. The aim of this procedure is to make the indirect effect of the concept on the model's behavior equal to zero for all inputs.

Concept erasure and guarding can be used to measure the causal importance of particular concepts, as in Elazar et al. (2021), though studies that use methods like these tend to focus on single layers. More recently, techniques such as LEACE (Belrose et al. 2023) and follow-ups (Singh et al. 2024b) have generalized this idea to provably prevent any linear classifier from using a concept; this moves beyond orthogonal projections and projects out the information at *every* layer. One could use such methods to causally understand the *set of* directions, or non-basis-aligned multidimensional space, that encode some concept. Note that many of these methods are still susceptible to the problems entailed by using linear mediators; thus, future work could follow Iskander, Radinsky, and Belinkov (2023); Ravfogel et al. (2022) in generalizing these analyses to non-linear mediator types.

Counterfactual-Based Optimization. A related line of methods involves learning binary masks over sets of components (e.g., sets of neurons, attention heads, layers) to determine which are relevant mediators for a task. Examples include subnetwork probing (Cao, Sanh, and Rush 2021) and desiderata-based component masking (DCM) (Davies et al. 2023; Prakash et al. 2024). These allow us to find basis-aligned multidimensional subspaces (§5.2), but they require triplets (x_i, x_i', y_i) consisting of contrastive input pairs (x_i, x_i') and a target y_i.

A more expressive class of approaches involves using the result of causal mediation analysis as a metric to directly optimize. One such line of methods includes distributed alignment search (DAS) and follow-up methods such as Boundless DAS (Geiger et al. 2024; Wu et al. 2023; Huang et al. 2024). These methods are powerful in that they allow us to locate non-basis-aligned multidimensional subspaces (§5.3) that correspond to a particular concept, though they have relatively strict data requirements compared with probing: We need not only labeled examples, but also contrastive input pairs *and* a

human-provided high-level causal graph. More precisely, given an activation vector $\mathbf{h}^\ell$ and hypothesized causal graph $\mathcal{H}$ containing nodes and edges (V, E), the goal is to learn an invertible linear transformation $\mathbf{R}$ such that a concept of interest $V_i \in V$ is aligned to the bases of the transformed space $\mathbf{R}(\mathbf{h}^\ell)$. To locate the concept, one iterates over contrastive input pairs (x_i, x_i') in training dataset $\mathcal{D}$ that vary only with respect to V_i and (ideally) no other concepts. For each pair, interventions are performed in the transformed space such that v_i (the value of V_i given x) is set to a counterfactual value v_i' that it would have taken given x_i'. Finally, the space is transformed back to the original space via $\mathbf{R}^{-1}$. This should result in a predictable change in model behavior, as defined by $\mathcal{H}$. See Appendix C for an illustration and more detailed description.

These methods provide time-efficient ways to search for human-interpretable variables encoded in intractably large or innumerable mediator sets. However, their key limitation is that they require either labeled data, contrastive pairs of inputs, and/or a pre-existing causal hypothesis as to how a model accomplishes some behavior. These methods can be evaluated with respect to accuracy in capturing model behavior, but they do not directly indicate a priori what those hypotheses should be. When one obtains negative results, these methods also do not indicate in what specific ways the hypotheses are wrong. They also require sufficient training data to recover the concept of interest during training. As with all parametric methods, the above approaches are subject to overfitting or underfitting.

6.2.2 Unsupervised Mediator Searches. Supervised search methods (see §6.2.1) require specific hypotheses about the internal representations of neural networks. However, neural networks implement various behaviors, many of which may be counterintuitive to humans and therefore more likely to be missed in supervised settings. For example, while Li et al. (2023) hypothesized a constant board state representation in a Transformer learning to play Othello, Nanda, Lee, and Wattenberg (2023) later found that the model actually switches the board state representation with every turn, taking the view of "my pieces vs. opponent's pieces" rather than "black pieces vs. white pieces". This example demonstrates that it can be desirable to search for mediators without specifying beliefs as to what those mediators do ahead of time.

Hence, some studies employ *unsupervised* methods. Typically, unsupervised methods return large—but finite—collections of mediators. Unsupervised methods are largely *correlative*, meaning that the discovered mediators may not necessarily capture causally relevant or faithful aspects of a model's computation. However, the discovered mediators can then be implicated in a model's computation post hoc by utilizing additional techniques, such as exhaustive searches for the highest-indirect-effect mediators (§6.1), to select task-relevant mediators from this collection.

Dictionary Learning Using Sparse Autoencoders. Exhaustive search for meaningful non-basis-aligned directions is impossible due to the infinite search space. The *dictionary learning* literature tackles this problem by performing an unsupervised search for directions in neuron activations which both (1) capture the information encoded in the internal representations and (2) are *disentangled* from other meaningful directions. Bengio, Courville, and Vincent (2013) characterize disentangled representations as *factors of variation* in the training dataset.

To identify these factors of variation, Sharkey, Braun, and Millidge (2023) used sparse autoencoders (SAEs) to perform dictionary learning on a one-layer transformer, identifying a large (overcomplete) basis of features. Given activation vector $\mathbf{h}^\ell$, SAEs

are trained to reconstruct $\mathbf{h}^\ell$ as $\hat{\mathbf{h}}^\ell$ while only activating a sparse subset of dictionary features. Concretely, SAEs typically consist of an encoder and decoder:

$$\mathbf{f} = \text{ReLU}(W_e(\mathbf{h}^\ell - \mathbf{b}_d) + \mathbf{b}_e), \tag{3}$$

$$\hat{\mathbf{h}}^\ell = W_d\mathbf{f} + \mathbf{b}_d, \tag{4}$$

where $\mathbf{f}$ is the feature vector (the encoded space), W denotes a learned weight matrix, and $\mathbf{b}$ denotes a learned bias vector.[27] It is common to refer to a single dimension of $\mathbf{f}$, or $\mathbf{f}_i$, as a *feature*. Cunningham et al. (2024) applied SAEs to language models and demonstrated that the observed dictionary features are highly interpretable and can be used to localize and edit model behavior. Since then, numerous researchers have found promising results in identifying funtionally relevant and human-interpretable features (Templeton et al. 2024; Rajamanoharan et al. 2024; Braun et al. 2024; Bricken et al. 2023; Fel et al. 2025, inter alia), many of them having predictable effects on the model's behavior under interventions. That said, SAEs are not able to perfectly reconstruct the model's activations, and may not be optimal for counterfactual operations such as steering (Wu et al. 2025). Most importantly, however, we do not know a priori what the ground truth features are in the model's computation, and can only use the reconstruction performance as a proxy measure of performance.

Correlation-Based Clustering. Another unsupervised way of discovering meaningful units is clustering mediators by the similarity of their behavior over some dataset $\mathcal{D}$. This idea is not new (cf. Elman 1990), but running causal verifications of the qualitative insights from clustering studies is relatively rare. Bau et al. (2019a) showed that neurons sharing a similar behavior are causally important for various functionalities in recurrent neural machine translation models. Dalvi et al. (2020) cluster neurons in language models, and are able to maintain performance after ablating a significant portion of them. The goal of Dalvi et al.'s study was not interpretability, but their results nonetheless causally verify that redundancy is very common in neural networks.

There has recently been renewed interest in mediator search via clustering. Michaud et al. (2023) propose a method to identify interpretable behaviors within neural networks by clustering parameters. Because the identified behaviors tend to be coherent, the units implicated in each cluster can be viewed as a set of components that have a functionally coherent role in the network. Marks et al. (2025) and Engels et al. (2025) generalize this from gradients to neuron or sparse autoencoder activations. The activations that compose the clusters are then labeled according to the dataset samples on which they activate most highly. When based on neurons, clusters are basis-aligned subspaces; when based on sparse autoencoder features (as in Marks et al. 2025), they are non-basis-aligned subspaces. To compute the indirect effect of a cluster, one can intervene on each component in the cluster simultaneously and compute the resulting indirect effect. Then, to find the most causally relevant clusters, one can exhaustively search by taking the top cluster by indirect effect. This method has not yet been extensively employed or explored. However, intervening on elements within these clusters could be a useful way to establish the functional role of *groups of* components in future work,

27 This is a basic formulation of SAEs. There are more sophisticated architectures that have empirically demonstrated better reconstruction performance and/or more interpretable features (e.g., Rajamanoharan et al. 2024; Braun et al. 2024).

or assess whether a subset of a model's behavior is implicated in a more complex task. We discuss this in §7.

In summary, unsupervised mediator searches enable us to locate human-interpretable concepts in language models without any labeled data. They allow us to do so relatively quickly, and these concepts can then be causally implicated in model behavior post hoc. However, the recovered concepts are not guaranteed to be faithful to the model's true concept space, and we are never guaranteed to recover a complete or non-redundant concept set. Another pressing issue in unsupervised search is evaluation: While it is possible to compare the relative quality of two different unsupervised interpretability searches, it is difficult to devise absolute metrics that meaningfully capture closeness to some ideal solution; indeed, an ideal or identifiable solution may not exist (Méloux et al. 2025). For SAEs, some have tried to devise thorough evaluations (Karvonen et al. 2025), but robust evaluations that enable comparisons between supervised and unsupervised methods have only just started to become common (e.g., Wu et al. 2025; Mueller et al. 2025).

7. Discussion

What is the right unit of analysis for describing the inner workings of language models? Thus far, we have categorized past work by mediator type and search method. Now, we turn our attention to practical questions: What kinds of studies can more easily be done with certain mediator types? Where is there still room to deploy less-common mediators in useful ways? Our main goal is to encourage researchers to conceptualize the field in this way, in the hope that we can encourage more work on discovering better causal abstractions and better terms for discussing the inner workings of language models—and, more ambitiously, more rigorous theoretical foundations for interpretability.

7.1 What Is the Right Mediator?

There are pros and cons to any mediator, and the best mediator will therefore depend on one's goals. In this section, we ask: When is it appropriate to deploy particular kinds of mediators and search methods? To answer this question, we revisit the three goals laid out in §2.1 and the pros and cons discussed in §5 and §6. We first hypothesize which mediators and search methods are likely to be best given recent evidence. Then, in the following subsections, we describe directions for future work that can further improve on these criteria, and call for work formalizing these criteria into concrete benchmarks. In Appendix D, we provide an even more practical guide by giving concrete examples in specific research scenarios.

Explaining Model Behavior. If we wish to understand at an algorithmic level how a model performs some behavior, then in the absence of compute restrictions and with no strong prior hypotheses, *unsupervised optimization-based methods* (§6.2.2) over fine-grained mediators (such as *non-basis-aligned directions*, §5.3) provide a strong starting point. For example, unsupervised methods like sparse autoencoders provide a fine-grained (selective) and human-interpretable interface to a model's computation, making explanations easier to derive in the absence of any pre-existing mechanistic hypotheses. They have also been found to yield higher faithfulness with fewer components compared with components like neurons (Marks et al. 2025), meaning that non-basis-aligned directions generally achieve a better trade-off between sparsity and faithfulness than basis-aligned directions. That said, autoencoder features are not guaranteed to be faithful to a neural

network's behavior in next-token prediction, and they require either a human or an LLM to label or interpret the features, which is laborious and expensive (Bills et al. 2023; Paulo et al. 2024). Moreover, natural language explanations of model components have inherent flaws (Huang et al. 2023): They may often exhibit both low precision and recall. Finally, non-basis-aligned directions, while more interpretable than basis-aligned components, require more human effort and/or compute than basis-aligned components to locate, and one may need to rediscover these directions if model fine-tuning, adaptation, or editing is part of the study.[28]

Verifying a Mechanistic Hypothesis. If we already have a hypothesis as to how a model performs some behavior and wish to measure the accuracy of our hypothesis, then *multidimensional non-basis-aligned subspaces* (§5.3) may be the right mediators, and a reasonable corresponding search method would be *counterfactual-based optimization* (§6.2.1). One can automatically search for the subspaces that correspond to a particular node in one's hypothesized causal graph using alignment search methods (Geiger et al. 2024; Wu et al. 2023; Sun et al. 2025; Geiger et al. 2021). Alignment search entails learning a linear transformation to isolate some target concept; this allows us to locate distributed representations that act as independent causal variables in non-basis-aligned spaces. This is relatively scalable, and enables us to qualitatively understand intermediate model computations. The primary downside is that we must anticipate the mechanisms that models employ to perform a task, as demonstrated in Appendix C; if we cannot anticipate them, then curating data and refining one's causal hypotheses may require significant human effort. If these are significant concerns, then probing may be a better (though correlative and less precise) option. These methods are subject to the same confounds as any other optimization-based technique: The module we are optimizing could directly learn the phenomenon, rather than extracting it from the language model (Hewitt and Liang 2019; Sutter et al. 2025).

Localization and Editing. If one's goal is to localize some phenomenon in a model, then *exhaustive searches* (§6.1) over *basis-aligned subspaces* (§5.2) or *full layers* (§5.1) may be sufficient. There are many comprehensive causal techniques for locating these, including causal tracing (Meng et al. 2022) and activation patching (Vig et al. 2020), as well as techniques for locating graphs of basis-aligned mediators, such as circuit discovery algorithms (Goldowsky-Dill et al. 2023; Wang et al. 2023; Conmy et al. 2023). Some of these methods are slow in their exact form, but fast approximations exist to these causal metrics, including attribution patching (Syed, Rager, and Conmy 2023) and improved versions thereof (Kramár et al. 2024; Hanna, Pezzelle, and Belinkov 2024; Marks et al. 2025). Even in the absence of a deep understanding of the role of these mediators, localization can be useful for downstream applications like model editing (Meng et al. 2022, 2023)[29] and model steering (Todd et al. 2024; Goyal et al. 2020). That said, if meaningful features are not actually aligned with neurons/heads, then we are not guaranteed to get the best performance unless we use more fine-grained and selective mediators. Future work should analyze the performance of model editing and steering methods when using different kinds of mediators. For example, Marks et al. (2025) compare the efficacy of debiasing approaches based on ablating neurons versus non-basis-aligned

28 Though Prakash et al. (2024) find that the same model components are implicated in an entity tracking task before and after fine-tuning.

29 Though Hase et al. (2023) find that causal localizations do not always reflect the optimal locations for editing models.

directions discovered via sparse autoencoders; they find that ablating non-basis-aligned directions is significantly more effective. Wu et al. (2025), however, find that supervised approaches for locating non-basis-aligned directions (e.g., difference-in-means, Marks and Tegmark 2023) are significantly more effective than sparse autoencoders. These studies represent promising initial steps toward the kind of principled comparison between mediators that we advocate.

7.2 Suggestions for Future Work

By centering the mediator type and the criteria defined in §2.1, we can gain new insights into the kinds of research that will be necessary to advance mechanistic interpretability. Here, we discuss lines of work we believe will be fruitful.

7.2.1 Finding Better Causal Mediators. There are almost certainly better causal mediators that have not yet been explored. By "better", we mean achieving a better Pareto optimum between the criteria described in §2.1 for at least one of the listed goals. Current work on improving mediators tends to focus on non-basis-aligned directions, such as sparse features or directions discovered from supervised probing on the activations of a single layer/submodule $\mathbf{h}^\ell$. One could consider pursuing coarser-grained mediators by discovering multi-layer *model regions* or *component sets* that accomplish a single behavior. Because these regions can cross layers, they would include non-linearities that allow them to represent more complex functions or concepts. We believe this would improve the generality of our findings, but not necessarily sparsity nor human-interpretability. Thus, we believe coarser-grained mediators will be better-suited to verifying mechanistic hypotheses or localization and editing.

Non-linear and Multidimensional Feature Discovery. As discussed in §5.4, there is recent work demonstrating the existence of human-interpretable *multidimensional* features. For example, days of the week are encoded circularly as a set of 7 directions in a two-dimensional subspace (Engels et al. 2025), and current methods cannot easily capture these multidimensional features. The ability to discover these features could greatly improve our understanding of the feature space of language models, and thus our ability to systematically explain more of their behavior (i.e., increase generality) in a faithful and human-interpretable way. Group sparse autoencoders (Theodosis and Ba 2023) or clusters of autoencoder features could be a way to capture multidimensional non-basis-aligned features in an unsupervised manner, but despite promising initial evidence, empirical work has not yet demonstrated to what degree this will be effective for interpreting or controlling language models. Additionally, many current mechanistic interpretability methods require us to define binary distinctions between correct and incorrect answers, whereas causal mediation analysis does not have any theoretical linearity, dimensionality, or Boolean restrictions; thus, interpretability methods will need to be extended to handle new kinds of variables.

There may also exist higher-order non-linear concepts in latent space that we have not yet located due to the linear focus of contemporary methods. For example, a subgraph or subcircuit can encode a coherent variable representation or functional role, as in Lepori, Serre, and Pavlick (2024) or Li, Davies, and Nadeau (2024). How can we discover these subgraphs? Path patching (Goldowsky-Dill et al. 2023; Wang et al. 2023) provides a manual approach to implicating subgraphs as causal mediators, but we do not yet have automatic methods that can scalably search over subgraphs of a computation graph. Even if we were able to locate them, how might non-linear and/or

coarse-grained mediators like these be useful in practice? As an example, we might expect fundamental phenomena like syntax processing to be spread across many layers of a model. Syntax-sensitive components should be implicated in downstream tasks like question answering (QA) if we expect that language models are robustly parsing the meaning of the inputs. Thus, one could use causal mediation analysis to locate all components in the model with some high indirect effect on syntactic processing (e.g., using a subject–verb agreement task); this can be conceptualized as the syntax processing region of the model. Then, one could implicate the region(s) in the model's QA performance by intervening on each component in the region and observing whether performance changes. If the syntax processing region is not strongly implicated in QA performance, then we have a strong hint that the model may not be parsing the meaning of the inputs, but instead relying on a mixture of surface-level spurious heuristics—for example, memorized bigram associations, or giving answers with high prior probabilities. These examples illustrate how these more coarse-grained mediators could help us verify new kinds of mechanistic hypotheses.

7.2.2 Inherently Interpretable Models. More ambitiously, one could consider building models with inherently interpretable components—i.e., whose fundamental units of computation (or some subset thereof) are designed to be sparse, monosemantic, and/or human-interpretable, but ideally still expressive enough to attain good performance on downstream tasks. Examples based in neural networks include differentiable masks (De Cao et al. 2020; Bastings, Aziz, and Titov 2019), transcoders (Dunefsky, Chlenski, and Nanda 2024), codebook features (Tamkin, Taufeeque, and Goodman 2023), and softmax linear units (Elhage et al. 2022a). These are primarily post hoc methods that decompose model components into interpretable units, but they could potentially be integrated into the network itself during pre-training alongside a loss term (in addition to the language modeling loss) that enables fine-grained interpretability at all stages of pre-training.

Alternatively, more focus could be devoted to building models that are designed from the ground up to be interpretable, such as backpack language models (Hewitt et al. 2023), concept bottleneck models (Koh et al. 2020; Oikarinen et al. 2023), or decision trees (Hu, Rudin, and Seltzer 2019). A related idea is to train the model using loss terms that encourage success on intermediate tasks, or induce particular kinds of feature representations (Hupkes, Zuidema et al. 2017). Perhaps least invasively, we could consider pre-training methods that softly encourage interpretable features to be aligned to neuron bases; this would remove the need for optimization to find non-basis-aligned components, and therefore make interpreting model decisions significantly easier and less confounded. However, this would reduce the number of features that could be encoded per neuron, so it would likely require significantly more parameters, or accepting degradations in performance. Regardless, we believe that this line of work will improve our ability to explain the behaviors of language models via directly improving the human-understandability of its intermediate computations; this will make it far easier to explain model behaviors, verify mechanistic hypotheses, and localize/edit particular computations—but at the potential expense of performance on downstream tasks.

7.2.3 Scalable Search. As the size of neural networks increases, the number of potential mediators to search over will also increase. The situation worsens as we start searching over continuous sets of fine-grained mediators such as non-basis-aligned directions. Although a few gradient-based or optimization-based approximations to causal influence have been proposed to improve time efficiency, such as attribution patching (Syed, Rager, and Conmy 2023) and DCM (Davies et al. 2023), more work is still needed to

evaluate the efficacy of these techniques in identifying the correct causal mediators. Additionally, better techniques beyond greedy search methods should be devised to identify causally important *groups of* mediators; these should aim to produce Pareto improvements over time complexity and causal efficacy.

As discussed in §6.2.1, optimization-based mediator search methods can be effective, but often require pre-existing hypotheses as to how a model implements a particular behavior. Thus, one could investigate using large language model (LLM) agents to automate the process of hypothesis generation. Qiu et al. (2024) showed that current LLMs can generate hypotheses, and Shaham et al. (2024) showed that hypothesis refinement via LLMs can aid humans in interpreting the causal role of neurons in multimodal models. Similary, LLMs could be used to automate and scale hypothesis generation regarding the role of particular mediators across a wider variety of tasks and models. Optimization-based methods such as DAS or DCM could then be used to causally verify the automatically generated hypotheses, potentially in an iterative loop of hypothesis refinement and empirical testing.

7.2.4 Benchmarking Progress in Mechanistic Interpretability. How will we know when we have made genuine improvements along any of the criteria that we have proposed? There exist few standardized methods or datasets for measuring general progress. To address this, research is needed on *standard benchmarks for measuring progress in mechanistic interpretability*. Currently, most studies develop ad hoc evaluations, and only compare to similar methods that employ the same mediators. Thus, to measure whether new mediators or search methods are truly giving us improvements over previous ones, we need to develop methods for performing principled direct comparisons. In circuit discovery, it is theoretically possible to use the same metrics to compare any circuit discovered for a particular model and task, regardless of whether sparse autoencoders are used, whether the circuit is based on nodes or edges, among other variations. Some recent work has begun to perform direct comparisons across mediator types, such as Miller, Chughtai, and Saunders (2024). Huang et al. (2024) propose to directly evaluate interpretability methods according to the generality of the abstractions they recover, and directly compare across different mediator types given the same model and task. Arora, Jurafsky, and Potts (2024) and Makelov, Lange, and Nanda (2024) also propose standardized interpretability benchmarks that allow us to compare across mediator search methods, though they do not directly compare across mediator types.

Direct comparisons require defining criteria for success, but in mechanistic interpretability, there is little agreement about the kinds of phenomena we should be measuring and precisely how they should be measured (with the exception of faithfulness, which is very common but still not standardized; Hanna, Pezzelle, and Belinkov 2024; Wang et al. 2023). Some tasks have started to integrate human/user evaluations, which will be especially useful for building interpretability tools that are grounded in real-world use cases and settings (Saphra et al. 2024). One benchmark that aims to enable direct comparison across mediator types and search methods is the Mechanistic Interpretability Benchmark (Mueller et al. 2025), which consists of two tracks: one for comparing circuits based on basis-aligned mediators, and another for comparing across mediator types/search methods. The fusion of these two tracks—i.e., comparing full causal graphs based on non-basis-aligned mediators, or even non-basis-aligned subspaces—could add value beyond the sum of these two separate tracks, and enable us to more directly benchmark progress on each of the goals of MI. Furthermore, Mueller et al. only measure (counterfactual) faithfulness and sparsity, as these are the most tractable metrics at the relatively large scale needed for a benchmark. Future

benchmarks should develop more scalable measures of generality and selectivity (e.g., via out-of-distribution evaluations).

For model editing and localization, more task-specific downstream metrics and datasets will be needed. For example, Cohen et al. (2024) and Zhong et al. (2023) propose benchmarks to evaluate model editing methods on out-of-distribution examples, and Karvonen et al. (2024) propose to measure progress in feature disentanglement using board game models. Wu et al. (2025) define a measure for the quality of steering methods based on different mediator types and search methods. While not the main focus of this survey, we believe that building standardized benchmarks will be a key means to the end of assessing whether advancements in causal mediators are producing real improvements in applications of interpretability. More broadly, robust evaluation metrics and methods will lead to a more accurate science of the inner workings of language models, which will allow us to assess whether new causal abstractions are fundamentally more useful—for explaining the computations of a neural network, for verifying hypotheses, *and* for practical applications.

8. Conclusion

In any study analyzing model behaviors via analyzing model components, the type of component(s) analyzed will determine what kinds of findings are possible. Some units are more closely aligned to the target concepts, while others are more faithful to the model's computation. Some units are easier to search over, but more difficult to understand (or vice versa). We have proposed a narrative and taxonomy of mechanistic interpretability research grounded in these units of analysis, or causal mediators. We have discussed the strengths and weaknesses of each mediator type, as well as what kinds of search methods are commonly employed for each. We have also discussed open problems in the field, focusing on those where this perspective reveals actionable and impactful research opportunities.

Appendix A: Types of Interventions

To compute the indirect effect (Equation (1)), we must replace the value v_i of causal node V_i with some counterfactual value v_i'. Say we are using mediator type Z, and that we are performing an exhaustive search by computing the IE for all $z_i \in Z$, and then taking the top components by IE. How should we compute z_i'? There are many ways to derive z_i: some of these depend on x, others depend on whether x is a member of some class (e.g., inputs about dogs or inputs not about dogs), and others still depend on neither (e.g., are constant values or involve adding a noise term). Here, we briefly describe each of these classes of interventions, and describe how they will affect the kinds of components one will uncover.

Broadly, constant interventions will tell one which components have *any* impact on model behavior, regardless of how. Input-dependent interventions are more precise, but tend to have lower recall: They isolate components whose impact changes when the input changes in the specific way defined by the intervention. Class-dependent interventions are a sort of medium between these.

A.1 Input-Dependent Interventions

Deterministic Interventions. If one cares about neurons that are sensitive to a specific contrast, then one can use input-dependent interventions (i.e., interventions where z_i'

depends on x). For example, assume our target task is subject–verb agreement. Given an input $x =$ "The **key**", we want to locate neurons that increase the probability difference $m = p(\textbf{is}) - p(\textbf{are})$. We obtain z_i by running x through model $\mathcal{C}$ in a forward pass (denoted $\mathcal{C}(x)$) and storing the activation z_i of component(s) Z_i (which could be a neuron, for example). We then obtain z_i' by running $\mathcal{C}(x')$, where x' is a minimally different input that swaps the answer; here, x' would be "The **keys**". This type of intervention preserves example-specific information, and varies only the grammatical number of the subject. This will only reveal neurons for which swapping grammatical number *and nothing else* will significantly affect the model's output.

Input-dependent interventions are precise: They reveal components targeted to a specific contrast between two prompts. However, humans must carefully curate controlled input pairs in which only one phenomenon is varied across x and x'. Input-dependent interventions work best for *binary* contrasts, where one defines two minimally different inputs that isolate neurons sensitive only to the difference between items in the pair. This yields counterfactuals that are semantically meaningful, making the results of an intervention easier to interpret. When working with categorical or ordinal variables, it is not immediately clear how to construct x' to recover all relevant components. Additionally, it does not recover all task-relevant components; it only recovers those sensitive to the contrast between x and x'. In other words, this is a low-recall method. For instance, Vig et al. (2020) enumerates through all possible gender pronouns and nouns related to a specific gender to measure gender bias; they note that full generalizability to all grammatical gender pronouns is difficult. Furthermore, such interventions can be privy to unreliable explanations. This was shown in Srivastava, Oikarinen, and Weng (2023), where the input data was corrupted to manipulate the concept assigned to a neuron. Hence, input-dependent interventions may require additional safeguards to ensure safety and fairness in critical real-life applications.

Stochastic Interventions. Another common intervention type entails adding noise to z_i', without defining some specific x_i' from which to derive it. For example, Meng et al. (2022, 2023) derive the counterfactual as $z_i' = z_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 3\sigma_{Z_i})$. σ_{Z_i} is the standard deviation of Z_i on some dataset. This intervention depends on z_i by definition, but does not isolate a semantically meaningful contrast as input-dependent interventions do. This is conceptually closer to class- and independent-independent interventions (§A.3), in that it will isolate components with *any* impact on the model's behavior, regardless of the semantics of that impact. However, its stochasticity introduces variability in results and can be harder to interpret causally.

A.2 Class-Dependent Interventions

In contrast to input-dependent interventions, class-dependent interventions define a single intervention across a class of inputs. For example, Li, Davies, and Nadeau (2024) learn a mask over the computational graph of a language model to prevent the model from producing toxic content; here, the two classes are *toxic* and *not toxic*, and the intervention within one of those classes is the same as for all other inputs in that class.

Class-dependent interventions provide a single flexible intervention that works for any given input. However, they require a dataset of input-label pairs that can be used to learn the interventions. This requires labeled data, and is sensitive to spurious correlations. Furthermore, many labels we care about are hard to definitively label without ambiguity (e.g., bias or toxicity).

A.3 Class- and Input-Independent Interventions

This type of intervention does not rely on the input nor a class label. The goal of these interventions is generally to fully remove (*ablate*) the information encoded by a mediator, regardless of whether the information is task-relevant.[30] A common ablation type is **zero ablations** (Dabkowski and Gal 2017; Lakretz et al. 2019; Geva et al. 2023), where the activation of a component is set to 0. This is not entirely principled, since 0 has no inherent meaning in an activation—for example, a neuron's default activation may be non-zero, whereas 0 itself is out of distribution relative to what the model expects. A more principled ablation type is a **mean ablation** (Zeiler and Fergus 2014; Ghorbani and Zou 2020; McDougall et al. 2024), where the neuron's activation is set to its mean value over some distribution—either task-specific data or general text data. A **resampling ablation** (Robnik-Šikonja and Kononenko 2008; Chan et al. 2022) is typically defined as a special case of mean ablations where the sample size is 1, and where the counterfactual input is randomly sampled.

Class- and input-independent interventions are a more general type of intervention that can be run without access to contrastive input/output pairs, and without labeled inputs. They allow us to tell whether *any* of the information in a mediator is necessary for a model to perform the task, but they may also affect other information in unanticipated ways; in other words, they have high recall and low precision relative to the previous intervention types. They may also cause performance on a task to drop in a way that reveals spurious mediators, rather than mediators that are conceptually relevant. For example, in subject–verb agreement, ablating a neuron that detects the word "dog" may reduce the probability of the correct verb form "is" over the incorrect verb form "are", but this is a highly input-specific neuron that does not, in isolation, reveal general information about how models perform syntactic agreement.

Appendix B: Computational Considerations

We have briefly touched on the computational considerations inherent to each mediator type and search method throughout the survey. Here, we expand this discussion by more directly comparing their computational costs.

Table B.1 contains estimates of the number of forward and backward passes through $\mathcal{C}$ needed to locate the most causally relevant mediators, assuming we compute causal relevance using the indirect effect as in Equation (1). Where applicable, ℓ refers to the number of layers, d to the size of an activation vector $\mathbf{h}^{\ell}$, f to the size of an SAE's latent vector (i.e., the size of the output of the encoder), and K to the number of clusters (a hyperparameter used in clustering algorithms). We see that gradient attributions are always fastest where applicable, but as they are linear estimates, we expect them to be less accurate than more exact methods like exhaustive search.

Note that training times are excluded from these estimates. In general, training a single probe should have similar amortized runtime compared to a single training run of alignment search, though the difference lies in how many runs would be needed to find the correct features, and in how many examples would be needed to properly train

30 An ablation is a type of intervention. The goal of an ablation is to *remove* the information stored in a component. "Intervention" is a broader term that refers to setting some v_i to any value it would not naturally have taken.

Table B.1
Summary of number of forward passes (and backward passes, when applicable) needed to locate the most causally relevant mediators of a given type (columns) using a particular method (rows). When a method involves training, we do not include training time in these estimates.

Method	Layers/submodules	Neurons	Basis-aligned spaces	Non-basis-aligned spaces
Exhaustive search	$O(\ell)$	$O(\ell \cdot d)$	$O(2^{\ell \cdot d})$	N/A
Gradient attribution	$O(1)$	$O(1)$	N/A*	N/A
Probing	$O(\ell)$	N/A	$O(\ell)$	$O(\ell)$
Alignment search	$O(\ell)$	$O(\ell \cdot d)^{\dagger}$	$O(\ell)$	$O(\ell)$
Sparse autoencoders	N/A	N/A	N/A	$O(\ell \cdot f)^{\ddagger}$
Clustering	$O(K)$	$O(K)$	$O(K)$	$O(K)$

*One could operationalize this as the sum of neurons' gradient attributions (in which case it would be $O(1)$, though finding the best combination could still be exponential), but this is not recommended for three reasons: interaction effects, redundancy, and potential nonlinear compositions.

†This estimate is based on the method of Geiger et al. (2021), but this is not common; with more recent methods like Boundless DAS (Wu et al. 2023), it could in theory be reduced to $O(\ell)$.

‡This assumes exhaustive search; the time becomes $O(1)$ if using gradient attributions.

them. If using a typical linear classification probe, one only needs to train $O(\ell)$ times maximum to obtain the best probe. If using Boundless DAS (Wu et al. 2023), one needs to train $O(\ell \cdot |x|)$ times, where $|x|$ is the length of the input sequence. For unsupervised methods, training time can vary significantly; training a sparse autoencoder can take over a day given even a relatively small model of $<$ 1B parameters (assuming access to one A100 GPU), but one only needs to train $O(\ell)$ of them. For clustering, the training time depends on the clustering method, but one can, in theory, cluster essentially any set of scalars relatively quickly. Moreover, clustering can be performed over all possible components in a model simultaneously without needing to iterate over layers (though it may sometimes be beneficial to perform clustering iteratively).

Appendix C: Illustration of Alignment Search

Here, we provide an illustration of an alignment search example (Figure C.1), as described in §6.2, and use this example to illustrate the goals of mechanistic interpretability more broadly. We have a hypothesis as to how the model $\mathcal{C}$ performs addition; in other words, we have a guess as to what $\mathcal{H}$ looks like. To isolate a variable in the hypothesized $\mathcal{H}$, we must design a dataset of contrastive pairs that vary only with respect to the variable. For example, if we believe the model contains a carry-the-one feature, we can design a dataset of inputs that vary with respect to whether the model must carry the one while leaving all other variables unchanged. We can use these pairs to isolate the variable during a training procedure; see Wu et al. (2023), Geiger et al. (2024) for more details.

Generally, one does not search for every variable in the hypothesized $\mathcal{H}$; one might select a few variables of interest. To quantify whether one has found them, one performs interventions to the discovered variable, as depicted in Figure C.1. Success is measured by whether the predicted change to the model's output under the intervention given $\mathcal{H}$ is what is actually observed.

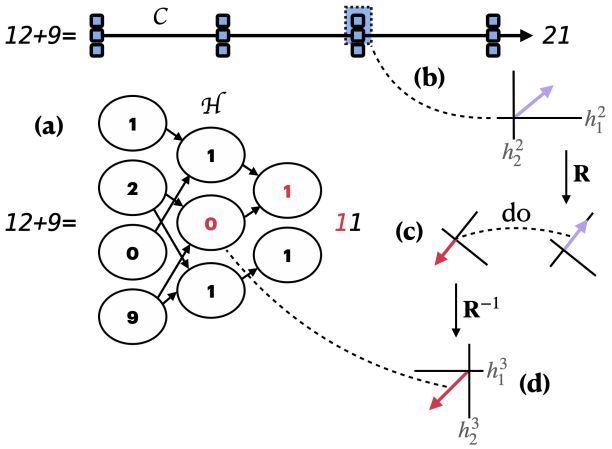

Figure C.1
Example of alignment search, based on an example from Mueller et al. (2025). (a) We start with the computation graph $\mathcal{C}$, and a hypothesized high-level causal graph $\mathcal{H}$. The hypothesis is that the model accomplishes addition using a tens-place addition, a ones-place addition, and a carry-the-one variable. (b) We hypothesize that the carry-the-one variable exists in layer two ($\mathbf{h}^2$). This variable may exist between multiple neurons, so interventions to neurons will not suffice. (c) We learn a rotation $\mathbf{R}$ into a new space where the target variable is aligned to the basis. This allows us to perform an intervention (the do-operation) to change the carry-the-one variable to some counterfactual value. (d) After intervening, we rotate back out using $\mathbf{R}^{-1}$. If the hypothesized causal graph is correct, the new output should be 11 instead of 21 after changing the carry-the-one variable's value.

Appendix D: Concrete Examples

The discussion in §7.1 is abstract. Here, we aim to give more concrete examples as to when certain kinds of mediators may be more appropriate.

Assume we are interested in understanding how a language model performs multiple-choice QA. Also assume that we can afford to rerun fine-tuning and adaptation if this model is particularly bad, so we do not intend to perform precise model editing based on the results of benchmarking evaluations or interpretability experiments. Instead, we care mainly about predicting success and failure modes on future examples so that we know whether this model could be deployed in production, and in what cases we should double-check the model's outputs. In this case, the goal is to *explain model behavior*, and we do not have a specific mechanistic hypothesis. Thus, we should deploy an *unsupervised method* to locate and search over meaningful features, such as non-basis-aligned directions. This will help one find unantipicated mechanisms.

Assume instead that we want to precisely edit the knowledge of the model on cases where it gets the answer wrong. Now, we do not necessarily care as much about interpreting the model's general answering process (as helpful as this would be), but rather, debugging and fixing specific mistakes. Thus, this would fall under localization and editing: We would like to locate the source of incorrect answers, and patch them to improve performance. Thus, as a first step, we should deploy an exhaustive search over a relatively coarse-grained mediator, such as submodules or full layers. We can use model editing techniques like ROME (Meng et al. 2022) or MEMIT (Meng et al. 2023), which perform targeted updates to basis-aligned components, to edit facts in cases

where the model answered incorrectly. It is theoretically possible that localizing editing over non-basis-aligned mediators could result in even better performance. For example, AlphaEdit (Fang et al. 2025) performs a targeted update to the *null space* of MLP layers, outperforming both ROME and MEMIT on several benchmarks. Future work should investigate whether this is possible, and whether the expected time complexity increase is worth the performance improvements.

Now assume that we are running a different kind of study: The task is still multiple-choice QA, but we are testing a specific hypothesis as to how the model accomplishes the task. We want to know whether it represents "truthfulness" as an independent concept, and the study is only concerned with to what extent this holds—not accuracy on the task per se. Here, we want to verify a specific mechanistic hypothesis, so we should design a dataset of labeled examples, where the label is based on truthfulness, and then deploy a supervised method such as probing over layers. If we can find a way to design counterfactual input pairs that vary only with respect to truthfulness, then we could instead deploy a more precise supervised method such as counterfactual-based optimization over non-basis-aligned subspaces. This would yield a set of scores that indicate to what extent the hypothesis causally explains the model's output behavior.

Note that none of these examples have recommended the use of basis-aligned subspaces, such as (sets of) neurons or attention heads. This is not to say that they are not useful, but it does indicate that when compute is not a significant limitation, they are often not the best place to start when working with realistic neural networks trained on large-scale data. Basis-aligned units are often difficult to interpret, and there are many of them; other mediator types are generally either more interpretable or easier to search over. That said, basis-aligned subspaces may be useful when we expect that they may have interpretable meanings (e.g., in toy task settings), or when we expect that unsupervised methods like sparse autoencoders are likely to yield bad results, or are simply not effectively trainable given one's resources.

References

Amini, Afra, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. Naturalistic causal probing for morpho-syntax. *Transactions of the Association for Computational Linguistics*, 11:384–403. https://doi.org/10.1162/tacl_a_00554

Antverg, Omer and Yonatan Belinkov. 2022. On the pitfalls of analyzing individual neurons in language models. In *International Conference on Learning Representations*.

Arora, Aryaman, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663. https://doi.org/10.18653/v1/2024.acl-long.785

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399. https://doi.org/10.1162/tacl_a_00106

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495. https://doi.org/10.1162/tacl_a_00034

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Bastings, Jasmijn, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977. https://doi.org/10.18653/v1/P19-1284

Bau, Anthony, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James R. Glass. 2019a. Identifying and

controlling important neurons in neural machine translation. In *7th International Conference on Learning Representations, ICLR 2019*.

Bau, David. 2022. Baukit. https://github.com/davidbau/baukit

Bau, David, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078. https://doi.org/10.1073/pnas.1907375117, PubMed: 32873639

Bau, David, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. 2019b. Visualizing and understanding generative adversarial networks. In *International Conference on Learning Representations*.

Belinkov, Yonatan. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219. https://doi.org/10.1162/coli_a_00422

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872. https://doi.org/10.18653/v1/P17-1080

Belinkov, Yonatan and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72. https://doi.org/10.1162/tacl_a_00254

Belrose, Nora, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828. https://doi.org/10.1109/TPAMI.2013.50, PubMed: 23787338

Bereska, Leonard and Stratis Gavves. 2024. Mechanistic interpretability for AI safety - A review. *Transactions on Machine Learning Research*. *arXiv:2404.14082*.

Bills, Steven, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. OpenAI Blog.

Boz, Olcay. 2002. Extracting decision trees from trained neural networks. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 456–461. https://doi.org/10.1145/775047.775113

Braun, Dan, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. 2024. Identifying functionally important features with end-to-end sparse dictionary learning. *arXiv preprint arXiv:2405.12241*. https://doi.org/10.52202/079017-3408

Bricken, Trenton, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Brinkmann, Jannik, Abhay Sheshadri, Victor Levoso, Paul Swoboda, and Christian Bartelt. 2024. A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task. *arXiv preprint arXiv:2402.11917*. https://doi.org/10.18653/v1/2024.findings-acl.242

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Cao, Steven, Victor Sanh, and Alexander Rush. 2021. Low-complexity probing via finding subnetworks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–966. https://doi.org/10.18653/v1/2021.naacl-main.74

Chan, Lawrence, Adrià Garriga-Alonso, Nicholas Goldwosky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal scrubbing, a method for rigorously testing interpretability hypotheses. *AI Alignment Forum*.

Chen, Angelica, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. 2024a. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

Chen, Yida, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, et al. 2024b. Designing a dashboard for transparency and control of conversational AI. *arXiv preprint arXiv:2406.07882*.

Chughtai, Bilal, Lawrence Chan, and Neel Nanda. 2023. A toy model of universality: Reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning*.

Cohen, Roi, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298. https://doi.org/10.1162/tacl_a_00644

Conmy, Arthur, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352.

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. 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)*, pages 2126–2136. https://doi.org/10.18653/v1/P18-1198

Craven, Mark and Jude Shavlik. 1995. Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, 8.

Craven, Mark W. and Jude W. Shavlik. 1994. Using sampling and queries to extract rules from trained neural networks. In *Machine Learning Proceedings 1994*, pages 37–45. https://doi.org/10.1016/B978-1-55860-335-6.50013-1

Csordás, Róbert, Christopher Potts, Christopher D. Manning, and Atticus Geiger. 2024. Recurrent neural networks learn to store and generate sequences using non-linear representations. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 248–262. https://doi.org/10.18653/v1/2024.blackboxnlp-1.17

Cunningham, Hoagy, Logan Riggs Smith, Aidan Ewart, Robert Huben, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Dabkowski, Piotr and Yarin Gal. 2017. Real time image saliency for black box classifiers. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6970–6979.

Dalvi, Fahim, Hassan Sajjad, and Nadir Durrani. 2023. NeuroX library for neuron analysis of deep NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 226–234. https://doi.org/10.18653/v1/2023.acl-demo.21

Dalvi, Fahim, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing redundancy in pretrained transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4908–4926. https://doi.org/10.18653/v1/2020.emnlp-main.398

Danilevsky, Marina, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459. https://doi.org/10.18653/v1/2020.aacl-main.46

Davies, Xander, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. 2023. Discovering variable binding circuitry with desiderata. *arXiv preprint arXiv:2307.03637*.

De Cao, Nicola, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? Interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255. https://doi.org/10.18653/v1/2020.emnlp-main.262

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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), pages 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dunefsky, Jacob, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable LLM feature circuits. *arXiv preprint 2406.11944*.

Elazar, Yanai, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175. https://doi.org/10.1162/tacl_a_00359

Elhage, Nelson, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, et al. 2022a. Softmax linear units. *Transformer Circuits Thread*.

Elhage, Nelson, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022b. Toy models of superposition. *Transformer Circuits Thread*.

Elhage, Nelson, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Elman, Jeffrey L. 1989. Representation and structure in connectionist models. University of California, San Diego, Center for Research in Language. https://doi.org/10.21236/ADA259504

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science*, 14:179–211. https://doi.org/10.1207/s15516709cog1402_1

Elman, Jeffrey L. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2–3):195–225. https://doi.org/10.1023/A:1022699029236

Engels, Joshua, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*.

Erhan, Dumitru, Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2009. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal.

Fang, Junfeng, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. AlphaEdit: Null-space constrained model editing for language models. In *The Thirteenth International Conference on Learning Representations*.

Feder, Amir, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158. https://doi.org/10.1162/tacl_a_00511

Fel, Thomas, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba Ba, and Talia Konkle. 2025. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. *arXiv:2502.12892*.

Ferrando, Javier, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *arXiv:2405.00208*.

Feucht, Sheridan, Eric Todd, Byron Wallace, and David Bau. 2025. The dual-route model of induction. In *Second Conference on Language Modeling*.

Finlayson, Matthew, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843. https://doi.org/10.18653/v1/2021.acl-long.144

Fiotto-Kaufman, Jaden Fried, Alexander Russell Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, et al. 2025. NNsight and NDIF: Democratizing access to open-weight foundation model internals. In *The Thirteenth International Conference on Learning Representations*.

Gandikota, Rohit, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*. https://doi.org/10.1109/ICCV51070.2023.00230

Gandikota, Rohit, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David

Bau. 2024. Unified concept editing in diffusion models. *IEEE/CVF Winter Conference on Applications of Computer Vision*. https://doi.org/10.1109/WACV57701.2024.00503

García-Carrasco, Jorge, Alejandro Maté, and Juan Trujillo. 2024. How does GPT-2 predict acronyms? Extracting and understanding a circuit via mechanistic interpretability. *arXiv:2405.04156*.

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

Geiger, Atticus, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187.

Geva, Mor, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.751

Ghorbani, Amirata and James Zou. 2020. Neuron Shapley: Discovering the responsible neurons. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89. https://doi.org/10.1109/DSAA.2018.00018

Giulianelli, Mario, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248. https://doi.org/10.18653/v1/W18-5426

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

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 1–9.

Goyal, Yash, Amir Feder, Uri Shalit, and Been Kim. 2020. Explaining classifiers with causal concept effect (CaCE). *arXiv preprint arXiv:1907.07165*.

Gu, Albert and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*.

Gu, Albert, Karan Goel, and Christopher Re. 2022. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5). https://doi.org/10.1145/3236009

Gupta, Abhijeet, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21. https://doi.org/10.18653/v1/D15-1002

Hanna, Michael, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than? Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 76033–76060.

Hanna, Michael, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Hase, Peter, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hayashi, Yoichi. 1990. A neural expert system with automated extraction of fuzzy if-then rules and its application to medical diagnosis. In *Advances in Neural Information Processing Systems*, volume 3.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and*

Pattern Recognition, pages 770–778. https://doi.org/10.1109/CVPR.2016.90

Hewitt, John, Kawin Ethayarajh, Percy Liang, and Christopher Manning. 2021. Conditional probing: Measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639. https://doi.org/10.18653/v1/2021.emnlp-main.122

Hewitt, John and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743. https://doi.org/10.18653/v1/D19-1275

Hewitt, John and Christopher D. Manning. 2019. 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)*, pages 4129–4138.

Hewitt, John, John Thickstun, Christopher Manning, and Percy Liang. 2023. Backpack language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9103–9125. https://doi.org/10.18653/v1/2023.acl-long.506

Hinton, G. E., J. L. McClelland, and D. E. Rumelhart. 1986. Distributed representations. In *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. MIT Press.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computing*, 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, PubMed: 9377276

Hu, Xiyang, Cynthia Rudin, and Margo Seltzer. 2019. Optimal sparse decision trees. In *Advances in Neural Information Processing Systems*, volume 32, pages 7265–7273.

Hu, Xinyan, Kayo Yin, Michael I. Jordan, Jacob Steinhardt, and Lijie Chen. 2025. Understanding in-context learning of addition via activation subspaces. *arXiv:2505.05145*.

Huang, Jing, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. 2023. Rigorously assessing natural language explanations of neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 317–331. https://doi.org/10.18653/v1/2023.blackboxnlp-1.24

Huang, Jing, Junyi Tao, Thomas Icard, Diyi Yang, and Christopher Potts. 2025. Internal causal mechanisms robustly predict language model out-of-distribution behaviors. In *Forty-second International Conference on Machine Learning*.

Huang, Jing, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. 2024. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8669–8687. https://doi.org/10.21665/2318-3888.v12n24p13

Hupkes, Dieuwke, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926. https://doi.org/10.1613/jair.1.11196

Hupkes, Dieuwke, Willem Zuidema, et al. 2017. Diagnostic classification and symbolic guidance to understand and improve recurrent neural networks. *Interpreting, Explaining and Visualizing Deep Learning. Workshop at Neural Information Processing Systems 2017*.

Iskander, Shadi, Kira Radinsky, and Yonatan Belinkov. 2023. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5961–5977. https://doi.org/10.18653/v1/2023.findings-acl.369

Janiak, Jett, cmathw, and Stefan Heimersheim. 2023. Polysemantic attention head in a 4-layer transformer. LessWrong post.

Jermyn, Adam, Chris Olah, and Tom Henighan. 2023. Attention head superposition. *Transformer Circuits Thread*.

Joseph, Sonia. 2023. ViT Prisma: A mechanistic interpretability library for vision transformers. https://github.com/soniajoseph/vit-prisma

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020.

Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
Karnin, E. D. 1990. A simple procedure for pruning back-propagation trained neural networks. *IEEE Transactions on Neural Networks*, 1(2):239–242. https://doi.org/10.1109/72.80236, PubMed: 18282841
Karpathy, Andrej, Justin Johnson, and Li Fei-Fei. 2016. Visualizing and understanding recurrent networks. In *The Fourth International Conference on Learning Representations*.
Karvonen, Adam, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, et al. 2025. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. *arXiv:2503.09532*.
Karvonen, Adam, Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Riggs Smith, Claudio Mayrink Verdun, David Bau, and Samuel Marks. 2024. Measuring progress in dictionary learning for language model interpretability with board game models. In *ICML 2024 Workshop on Mechanistic Interpretability*.
Koh, Pang Wei and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894.
Koh, Pang Wei, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348.
Köhn, Arne. 2015. What's in an embedding? Analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073. https://doi.org/10.18653/v1/D15-1246
Kramár, János, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*.
Krishnan, R., G. Sivakumar, and P. Bhattacharya. 1999. Extracting decision trees from trained neural networks. *Pattern Recognition*, 32(12):1999–2009. https://doi.org/10.1016/S0031-3203(98)00181-2
Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25.
Lad, Vedang, Wes Gurnee, and Max Tegmark. 2024. The remarkable robustness of LLMs: Stages of inference? In *ICML 2024 Workshop on Mechanistic Interpretability*.
Lakretz, Yair, Théo Desbordes, Jean-Rémi King, Benoît Crabbé, Maxime Oquab, and Stanislas Dehaene. 2021. Can RNNs learn recursive nested subject-verb agreements? *arXiv preprint arXiv:2101.02258*.
Lakretz, Yair, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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)*, pages 11–20. https://doi.org/10.18653/v1/N19-1002
Lasri, Karim, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. Probing for the usage of grammatical number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831. https://doi.org/10.18653/v1/2022.acl-long.603
Lepori, Michael A., Thomas Serre, and Ellie Pavlick. 2024. Uncovering intermediate variables in transformers using circuit probing. In *First Conference on Language Modeling*.
Lewis, David. 1986. Causation. In *Philosophical Papers II*. Oxford University Press, pages 159–213. https://doi.org/10.1093/0195036468.003.0006
Lewis, David. 2000. Causation as influence. *Journal of Philosophy*, 97(4):182–197. https://doi.org/10.2307/2678389
Lewis, David K. 1973. *Counterfactuals*. Blackwell.
Li, Jiwei, Will Monroe, and Dan Jurafsky. 2017. Understanding neural networks through representation erasure. *arXiv:1612.08220*.
Li, Kenneth, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. In the

Eleventh International Conference on Learning Representations.

Li, Maximilian, Xander Davies, and Max Nadeau. 2024. Circuit breaking: Removing model behaviors with targeted ablation. *arXiv preprint arXiv:2309.05973*.

Li, Victoria R., Jenny Kaufmann, Martin Wattenberg, David Alvarez-Melis, and Naomi Saphra. 2025. Can interpretation predict behavior on unseen data? *arXiv preprint 2507.06445*.

Lipton, Zachary C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57. https://doi.org/10.1145/3236386.3241340

Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual 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)*, pages 1073–1094. https://doi.org/10.18653/v1/N19-1112

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lundberg, Scott M. and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777.

Lyu, Qing, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723. https://doi.org/10.1162/coli_a_00511

Ma, Weicheng, Kai Zhang, Renze Lou, Lili Wang, and Soroush Vosoughi. 2021. Contributions of transformer attention heads in multi- and cross-lingual tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1956–1966. https://doi.org/10.18653/v1/2021.acl-long.152

Madsen, Andreas, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural NLP: A survey. *ACM Computing Surveys*, 55(8). https://doi.org/10.1145/3546577

Makelov, Aleksandar, Georg Lange, and Neel Nanda. 2024. Towards principled evaluations of sparse autoencoders for interpretability and control. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Marks, Samuel, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*.

Marks, Samuel and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

McClelland, James L. and David E. Rumelhart. 1985. Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology. General*, 114(2):159–97. https://doi.org/10.1037/0096-3445.114.2.159, PubMed: 3159828

McDougall, Callum Stuart, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. 2024. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 337–363. https://doi.org/10.18653/v1/2024.blackboxnlp-1.22

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

Méloux, Maxime, Silviu Maniu, François Portet, and Maxime Peyrard. 2025. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *The Thirteenth International Conference on Learning Representations*.

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

Meng, Kevin, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Merullo, Jack, Carsten Eickhoff, and Ellie Pavlick. 2024a. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Merullo, Jack, Carsten Eickhoff, and Ellie Pavlick. 2024b. Talking heads: Understanding inter-layer communication in transformer language models. *arXiv preprint arXiv:2406.09519*.

Michaud, Eric J., Ziming Liu, Uzay Girit, and Max Tegmark. 2023. The quantization model of neural scaling. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Martin Karafiát, Lukáš Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, pages 1045–1048. https://doi.org/10.21437/Interspeech.2010-343

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26.

Miller, Joseph, Bilal Chughtai, and William Saunders. 2024. Transformer circuit evaluation metrics are not robust. In *First Conference on Language Modeling*.

Mohebbi, Hosein, Jaap Jumelet, Michael Hanna, Afra Alishahi, and Willem Zuidema. 2024. Transformer-specific interpretability. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 21–26. https://doi.org/10.18653/v1/2024.eacl-tutorials.4

Moraffah, Raha, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33. https://doi.org/10.1145/3400051.3400058

Morcos, Ari S., David G. T. Barrett, Neil C. Rabinowitz, and Matthew Botvinick. 2018. On the importance of single directions for generalization. In *International Conference on Learning Representations*.

Mozer, Michael C. and Paul Smolensky. 1988. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In *Advances in Neural Information Processing Systems*, volume 1.

Mueller, Aaron. 2024. Missed causes and ambiguous effects: Counterfactuals pose challenges for interpreting neural networks. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Mueller, Aaron, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, et al. 2025. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*.

Mueller, Aaron, Yu Xia, and Tal Linzen. 2022. Causal analysis of syntactic agreement neurons in multilingual language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 95–109. https://doi.org/10.18653/v1/2022.conll-1.8

Nanda, Neel and Joseph Bloom. 2022. TransformerLens. https://github.com/TransformerLensOrg/TransformerLens

Nanda, Neel, Andrew Lee, and Martin Wattenberg. 2023. 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*, pages 16–30. https://doi.org/10.18653/v1/2023.blackboxnlp-1.2

Neo, Clement, Shay B. Cohen, and Fazl Barez. 2024. Interpreting context look-ups in transformers: Investigating attention-MLP interactions. *arXiv preprint arXiv:2402.15055*. https://doi.org/10.18653/v1/2024.emnlp-main.930

Odense, Simon and Artur d'Avila Garcez. 2020. Layerwise knowledge extraction from deep convolutional networks. *arXiv preprint arXiv:2003.09000*.

Oikarinen, Tuomas, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. 2023. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*.

Olsson, Catherine, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Panickssery, Nina, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Park, Kiho, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*.

Paulo, Gonçalo, Alex Mallen, Caden Juang, and Nora Belrose. 2024. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, Judea. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420.

Pearl, Judea. 2009. *Causality*, 2nd edition. Cambridge University Press.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162

Prakash, Nikhil, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*.

Qiu, Linlu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. 2024. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. In *The Twelfth International Conference on Learning Representations*.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Rai, Daking, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Rajamanoharan, Senthooran, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. 2024. Improving dictionary learning with gated sparse autoencoders. *arXiv:2404.16014*.

Ravfogel, Shauli, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256. https://doi.org/10.18653/v1/2020.acl-main.647

Ravfogel, Shauli, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209. https://doi.org/10.18653/v1/2021.conll-1.15

Ravfogel, Shauli, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055. https://doi.org/10.18653/v1/2022.emnlp-main.405

Ravichander, Abhilasha, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377. https://doi.org/10.18653/v1/2021.eacl-main.295

Rawal, Atul, Adrienne Raglin, Danda B. Rawat, Brian M. Sadler, and James McCoy. 2024. Causality for trustworthy artificial intelligence: Status, challenges and perspectives. *ACM Computing Surveys*. https://doi.org/10.1145/3665494

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016a. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016b. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144. https://doi.org/10.1145/2939672.2939778

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. https://doi.org/10.1609/aaai.v32i1.11491

Robins, James M. and Sander Greenland. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155. https://doi.org/10.1097/00001648-199203000-00013, PubMed: 1576220

Robnik-Šikonja, Marko and Igor Kononenko. 2008. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600. https://doi.org/10.1109/TKDE.2007.190734

Räuker, Tilman, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. *arXiv:2207.13243*. https://doi.org/10.1109/SaTML54575.2023.00039

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536. https://doi.org/10.1038/323533a0

Sajjad, Hassan, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2023. On the effect of dropping layers of pre-trained transformer models. *Computer Speech & Language*, 77:101429. https://doi.org/10.1016/j.csl.2022.101429

Sajjad, Hassan, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level interpretation of deep NLP models: A survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303. https://doi.org/10.1162/tacl_a_00519

Salmon, Wesley C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press. https://doi.org/10.1515/9780691221489

Saphra, Naomi, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. 2024. First tragedy, then parse: History repeats itself in the new era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2310–2326. https://doi.org/10.18653/v1/2024.naacl-long.128

Saphra, Naomi and Sarah Wiegreffe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498. https://doi.org/10.18653/v1/2024.blackboxnlp-1.30

Shaham, Tamar Rott, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. 2024. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*.

Sharkey, Lee, Dan Braun, and Beren Millidge. 2023. Taking features out of superposition with sparse autoencoders. Accessed: 2023-05-10.

Sharma, Arnab Sen, David Atkinson, and David Bau. 2024. Locating and editing factual associations in Mamba. *arXiv preprint arXiv:2404.03646*.

Shi, Xing, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534. https://doi.org/10.18653/v1/D16-1159

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Singh, Chandan, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024a. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Singh, Shashwat, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. 2024b. Representation surgery: Theory and practice of affine steering. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Srivastava, Divyansh, Tuomas Oikarinen, and Tsui-Wei Weng. 2023. Corrupting neuron explanations of deep visual features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1877–1886. https://doi.org/10.1109/ICCV51070.2023.00180

Strobelt, Hendrik, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M. Rush. 2017. LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE Transactions on Visualization and Computer Graphics*,

24(1):667–676. https://doi.org/10.1109/TVCG.2017.2744158, PubMed: 28866526

Subhash, Varshini, Zixi Chen, Marton Havasi, Weiwei Pan, and Finale Doshi-Velez. 2022. What makes a good explanation?: A harmonized view of properties of explanations. In *Progress and Challenges in Building Trustworthy Embodied AI*.

Subramani, Nishant, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581. https://doi.org/10.18653/v1/2022.findings-acl.48

Sun, Jiuding, Jing Huang, Sidharth Baskaran, Karel D'Oosterlinck, Christopher Potts, Michael Sklar, and Atticus Geiger. 2025. HyperDAS: Towards automating mechanistic interpretability with hypernetworks. In *The Thirteenth International Conference on Learning Representations*.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3319–3328.

Sutter, Denis, Julian Minder, Thomas Hofmann, and Tiago Pimentel. 2025. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? *arXiv:2507.08802*.

Syed, Aaquib, Can Rager, and Arthur Conmy. 2023. Attribution patching outperforms automated circuit discovery. In *NeurIPS Workshop on Attributing Model Behavior at Scale*. https://doi.org/10.18653/v1/2024.blackboxnlp-1.25

Tamkin, Alex, Mohammad Taufeeque, and Noah D. Goodman. 2023. Codebook features: Sparse and discrete interpretability for neural networks. *arXiv preprint arXiv:2310.17230*.

Templeton, Adly, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 sonnet. *Transformer Circuits Thread*.

Theodosis, Emmanouil and Demba Ba. 2023. Learning silhouettes with group sparse autoencoders. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095958

Tigges, Curt, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*.

Todd, Eric, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*.

Tucker, Mycal, Peng Qian, and Roger Levy. 2021. What if this modified that? Syntactic interventions with counterfactual embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 862–875. https://doi.org/10.18653/v1/2021.findings-acl.76

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

Tzeng, F. Y. and K.-L. Ma. 2005. Opening the black box: Data driven visualization of neural networks. In *VIS 05. IEEE Visualization, 2005*, pages 383–390.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Vig, Jesse, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401.

Wang, Kevin Ro, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

White, Adam and Artur d'Avila Garcez. 2020. Measurable counterfactual local explanations for any classifier. *European Conference on Artificial Intelligence*, pages 2529–2535. https://doi.org/10.3233/FAIA200387

Wu, Wenhao, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024a. Retrieval head mechanistically explains

long-context factuality. *arXiv preprint arXiv:2404.15574*.

Wu, Xing, Shaoqi Peng, Jingwen Li, Jian Zhang, Qun Sun, Weimin Li, Quan Qian, Yue Liu, and Yike Guo. 2024b. Causal inference in the medical domain: A survey. *Applied Intelligence*, pages 1–24.

Wu, Zhengxuan, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. *arXiv:2501.17148*.

Wu, Zhengxuan, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024c. ReFT: Representation finetuning for language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wu, Zhengxuan, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah Goodman, Christopher Manning, and Christopher Potts. 2024d. pyvene: A library for understanding and improving PyTorch models via interventions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 158–165. https://doi.org/10.18653/v1/2024.naacl-demo.16

Wu, Zhengxuan, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. 2023. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zeiler, Matthew D. and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833. https://doi.org/10.1007/978-3-319-10590-1_53

Zhong, Zexuan, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702. https://doi.org/10.18653/v1/2023.emnlp-main.971