What's your Use Case? A Taxonomy of Causal Evaluations of Post-hoc Interpretability

David Reber[1], Cristina Gârbacea[1,2], and Victor Veitch[1,2]

[1]*Department of Statistics, University of Chicago*
[2]*Data Science Institute, University of Chicago*

Abstract

Post-hoc interpretability of neural network models, including Large Language Models (LLMs), often aims for mechanistic interpretations — detailed, causal descriptions of model behavior. However, human interpreters may lack the capacity or willingness to formulate intricate mechanistic models, let alone evaluate them. This paper addresses this challenge by introducing a taxonomy which dissects the overarching goal of mechanistic interpretability into constituent claims, each requiring distinct evaluation methods. By doing so, we transform these evaluation criteria into actionable learning objectives, providing a data-driven pathway to interpretability. This framework enables a methodologically rigorous yet pragmatic approach to evaluating the strengths and limitations of various interpretability tools.

1 Introduction

Despite the impressive performance of large language models (LLMs) in many complex tasks and domains, interpreting the inner workings of these black-box models remains an open challenge. Ideally, we would like to identify high-level variables that encode relevant concepts, as well as relationships between these concepts, with the goal of understanding internal model representations, manipulating these in a disentangled manner, assessing model behaviour and enhancing trust in these models. For example, consider a movie recommendation scenario: is the algorithm suggesting specific movies based on user preferences or because it has an inherent bias towards Tom Hanks movies? This example illustrates the need for a nuanced understanding of the model's internal variables and their causal relationships.

In order to make these neural models transparent and understand their inner workings, mechanistic interpretability [Räu+23] proposes translating model computations into human-readable code for gaining an algorithmic-level understanding of the neural network model's computations, such as for example a python program that captures the pseudocode of the LLM. While it is debatable to what extent programming code is human-interpretable, assuming this condition is met, it would offer us the possibility to understand and control model behavior. Nevertheless, this goal may not be readily attainable in practice due to the inherent difficulty of the objective: there is no guarantee that the surrogate python program is a reliable and consistent approximation of the true LLM behaviour. This holds particularly true when the program has not been evaluated on the entirety of the data the LLM has seen at training time. In addition, the principles of causal hierarchy [Bar+22] demonstrate that while two python programs may present indistinguishable behavior when evaluated on a given dataset, they may behave completely different from each other on other datasets in the presence of distributional shift. Besides, if the python program is billion lines of code in length, it is not going to satisfy the human-understandable criteria.

Given the above mentioned challenges, we follow the approach proposed in [GPI23] to narrow down the search space of LLM-compatible surrogate programs to those that emulate the LLM behavior on a *particular task of interest* only with a given data distribution. While this does not

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

guarantee consistency between the surrogate program and LLM behaviour on a different task other than the one of immediate interest, focusing the interpretability analysis on a narrow task (as opposed to the entire LLM-compatible programs space) allows us to substantially reduce the complexity of compatible programs for generating and validating hypotheses and the intricacy of their internal computations. The main contribution of our work consists in proposing partial evaluations of existing interpretability methods for the given task of interest.

Therefore, in this work we are interested in identifying a subset of high-level variables in the LLM that encode relevant concepts, along with partially specified relationships between these concepts. By leveraging techniques from mechanistic intepretability and program synthesis literature, we aim to find valid human-understandable interpretations that are consistent with the LLM behaviour. Unsurprisingly though, there are many ways in which any given interpretation can be falsified, even in complex domains with limited human understanding To overcome these challenges, we introduce a taxonomy rooted in the principles of causal hierarchy [Bar+22] which helps with the selection of interpretability tools, evaluation methods and mechanistic uncertainty given multiple compatible programs.

More specifically, our framework consists in identifying high-level variables, verifying constraints, probing for intermediate concepts, and assessing the causal sufficiency and generalizability of the proposed model interpretations. We consider the case where the high-level model is only partially specified, i.e we only know the direction of causation between intermediate variables, or even less, merely their joint distribution. We assume the high level model defines a set of high-level concepts that are interpretable to humans; in case of a data-driven approach, the data itself determines the high-level model. Our goal is to evaluate whether a given interpretation in terms of this ontology is *faithful* to the LLM, that is whether the interpretation can be used as a proxy for reasoning about the LLM. By focusing on these aspects, we aim to enhance the rigor of post-hoc interpretability in LLMs.

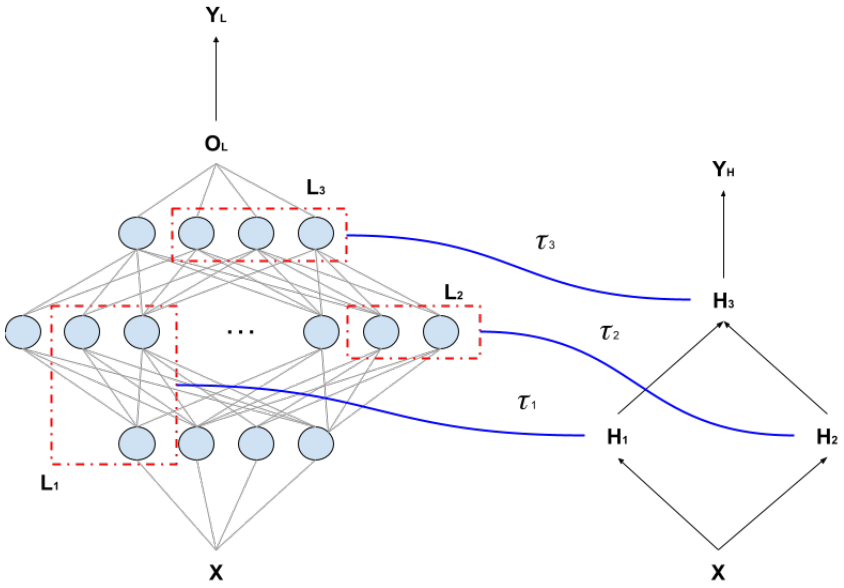

Figure 1: An *interpretation* (H, τ) consists of a human-understandable model H of how the output feature of interest, Y_H, is related to high-level concepts $H_1, H_2 \ldots, H_n$, along with mappings τ_k relating these high-level concepts to parts of the neural network. Here, we depict the case where H is a directed acyclic causal graph. In general, H could be cyclic or involve non-causal relations, so long as it induces a joint probability distribution over concepts.

2 Related Work

Our approach builds upon the framework of causal abstractions [GPI23; Gei+23; Wu+23] which evaluates the faithfulness of an interpretation by comparing model behavior to a given high-level algorithm. This algorithm is mechanistically specified and encapsulates human knowledge about

a domain at multiple levels, be it in the choice of variables, the choice of abstraction, the direction of causal influence, or the structural causal equations themselves.

Related to our work, [MPT23] presents a causal framework for evaluating interpretations focusing on the question of *alignment*, i.e. the extent to which an interpretation preserves human-semantics, rather than *faithfulness* as we do in our work. Since interpretations need to be both faithful and human-interpretable, these are complementary investigations. [Rä+23] offers a taxonomy of interpretability methods which is broader than ours, covering *intrinsic* interpretability (enforcing interpretability during training) in addition to the *post-hoc* interpretability methods we cover, and does not employ the same causal taxonomy for post-hoc interpretability methods.

We formulate the prompt/inputs to be exogenous variables outside the scope of intervention, and then consider how causal evaluations of the internals of the neural network can provide guarantees against distributional shift on the inputs. Alternative, one can directly model this distributional shift by intervening directly on the inputs as in [HL22].

3 Notation

Consider L, a feed-forward neural network, generating output $O_L(x)$ for a prompt $x \sim X$, where X represents the prompt distribution. The pair (X, L) forms a large structural causal model whose neurons deterministically map inputs x to activations l at each layer. The use-case and evaluation distributions of prompts are denoted as X^U and X^E, respectively.

For interpretations focused on a single task, we analyze Y_L, a summary statistic of O_L. Examples of Y_L include the sum of numbers mentioned in the prompt, and evaluations of whether O_L is harmful, biased, factual. Distributions induced by model L with X^U and X^E are Y_L^U and Y_L^E.

To relate the low-level computations of L to a high-level model, we introduce a structural causal model $H = (H_1, H_2, \ldots, H_n)$. H_k is thus a measurable function over the inputs x, representing a high-level concept. As with the neural network L, each H_k deterministically maps from a subset of $\{x, H_1, ..., H_{k-1}\}$ to values h_k. Even if H_k does not depend on x directly, we use $H_k(x, h_1, ..., h_{k-1})$ to refer to the value H_k takes on when $H_1 = h_1, ..., H_{k-1} = h_{k-1}$ and input x. Domain$(H)$is the set of x where H is well-defined: we assume it is supported by X^U and X^E.

The mapping $\tau : L \to H$ translates activations in L to high-level concepts in H. This is achieved through functions $\tau_k : \text{domain}(L_k) \to \text{domain}(H_k)$, linking pairwise disjoint sets of activation vectors L_k to their corresponding H_k; collective L_k may not encompass the entire activation space of L, allowing for parts of L that do not map to any high-level concept for this task.

The interpretation (H, τ) enables both passive monitoring and active control of L's behavior. Mechanistic faithfulness involves ensuring that the causal relations between the H_k are a perfect match for the causal relations between the L_k, so that H can serve as a proxy for reasoning about L even under distributional shift. As we will show, partial evaluations of mechanistic faithfulness are possible, allowing analysis even without a fully-structural specification of H.

4 Mechanistic Faithfulness as an Ideal

In order to reason about the behaviour of black-box model L , one would ideally specify a proxy model H that is entirely expressed in human-understandable language. Ideally, H would be *mechanistically faithful* to L, in the sense that all correlations between inputs, outputs, and intermediate concepts are preserved even when intermediate concepts are intervened on.

Definition 1 (Mechanistic Faithfulness). Suppose we are given low-level model L, an interpretation (H, τ), a use case consisting of prompt distribution X^U and target concepts Y. We say the high-level model H is **mechanistically faithful** to the low-level model L if, $\forall x \in \text{domain}(H)$, $\forall l_i \in \text{domain}(\tau_i)$,

$$\tau_k(L_k(x, l_1, ..., l_{k-1})) = H_k(x, \tau_1(l_i), ..., \tau_{k-1}(l_{k-1})) \tag{4.1}$$

$$Y_L(x, l_1, ..., l_n) = Y_H(x, \tau_1(l_1), ..., \tau_n(l_n)) \tag{4.2}$$

where $L_k(x, l_1, ..., l_{k-1})$ is the value L_k would take on if $X = x, L_1 = l_1, ..., L_k = l_k$, and $H_k(x, \tau_1(l_i), ..., \tau_{k-1}(l_{k-1}))$ is defined similarly.

Intuitively, the definition of mechanistic faithfulness states that at least for inputs $x \in \text{domain}(H)$, each $\tau(L_k)$ and H_k have the exact same functional relationship to preceding concepts $L_{<k}$ and $H_{<k}$ and the input x.

If we pass in the same distribution of inputs X to L and a mechanistically faithful H, all their counterfactual, interventional, and observational distributions will agree. Hence, this is an extremely strong faithfulness criteria. Therefore, an important question to ask is what evaluations can we perform using L, H, τ, and X^E to falsify Equation 4.1?

This is challenging, as we may have limited understanding of:

- how the use-case prompt distribution X^U differs from the evaluation distribution X^E;
- the joint distribution $P(X, H_1, ..., H_n, Y_H)$ of high-level intermediate concepts (for eg., we may only have estimates of some conditional distributions with sampling bias);
- which H is consistent with the data-generating process of L's training data.

Given these uncertainties, we can at least check whether certain weaker claims hold. In what follows we present implications of Equation 4.1 in increasing order of their difficulty; these implications are the cornerstone for the hierarchy of evaluations we propose in Section 5. Note that if the platonic objective 1 holds true, then each of the following statements is also true (although the converse is not always true) [1].

Implication 1 (Input/Output Observational Equivalence)**.** H should preserve the correlations between the input prompts and the target concepts:

$$P_X(Y_L) = P_X(Y_H) \tag{4.3}$$

Implication 2 (Input/Intermediate Observational Equivalence)**.** τ should preserve the correlations which an input induces on intermediate concepts of H:

$$P_X(\tau_1(L_1), \ldots, \tau_k(L_k)) = P_X(H_1, \ldots, H_k) \tag{4.4}$$

Probing is a common approach to search for a model's representation of a concept [Bel21], especially when a concept is believed to be represented by a single neuron [Gur+23]. Note that if one trains probes for high-level concepts independently, one may fail to capture the information about how the concepts relate to each other: this would correspond to just enforcing the marginal equivalence $P(\tau(L_k)|X) = P(H_k|X)$ for each k individually. It is more informative to leverage how the high-level concepts relate to each other as in 4.4, as in [Bur+22; PH22].

However, just as one would not assume that just because two Python programs are perfectly identical simply because they pass the same unit tests on a limited dataset, we would like to evaluate how 1. H might fail to predict L's behavior out of distribution, or 2. whether H responds the same to model steering as L in-distribution.

Implication 3 (Input/Output Interventional Equivalence)**.** L and H should produce the same target concepts when intermediate concepts are intervened on.

$$P_X(Y_L; \text{do}(L_i = l)) = P_X(Y_H; \text{do}(H_i = \tau_i(l))) \tag{4.5}$$

There are many recent examples of evaluating the faithfulness of a concept by its effect on the output, although the nature of the interventions vary widely, from synthetic interventions optimized a probe value [Li+23], ablations [Bel+23], *interchange interventions* [Wu+23] (setting activations to values they would take on other inputs), *concept algebra* [Wan+23] (interventions based on linear projections in a score-based representation).

Equation 4.5 is only a statement about the relation between intermediate concepts and the output, not about the relationships *between* intermediate concepts themselves. Of course, if we truly

[1] Since these ought to hold for both X^E and X^U, we just write X for simplicity. We will add the superscripts back in when discussing the implications of evaluations.

have recovered the true Python program, then the high-level model should predict the effect of an intervention of one intermediate variable on other intermediate variables as well.

Implication 4 (Input/Intermediate Interventional Equivalence)**.** Post-interventional probing correlations match real data:

$$P_X(\tau(L_{i+1}),\ldots,\tau(L_n);\text{do}(L_i=l)) = P_X(H_{i+1},\ldots H_n;\text{do}(H_i=\tau_i(l))) \tag{4.6}$$

Here, we seek to ensure that probes and the high-level model react similarly to interventions. In the literature, causal scrubbing [Cha+22] uses ablations to enforce the invariance of non-descendant concepts to interventions. Another approach involves activation patching [Con+23], which is commonly used in mechanistic interpretability to test specific network functions. Lastly, we seek to evaluate counterfactuals [2], which provide even stronger guarantees that H will continue to serve as a proxy for L under distributional shift.

Implication 5 (Input/Output Counterfactual Equivalence)**.** L and H should produce the same target concepts when intermediate concepts are intervened on.

$$P_X(Y_L|L_i=l;\text{do}(L_i=l')) = P_X(Y_H|H_i=\tau_i(l);\text{do}(H_i=\tau_i(l'))) \tag{4.7}$$

Furthermore, ensuring that intermediate concepts H_k are counterfactually equivalent is important if we want to ensure that model steering will continue to behave as expected out of distribution. Intermediate counterfactual equivalence may also prove necessary for 1) concepts involving a degree of adaptiveness, such as agency, situational awareness, cooperativeness, and consciousness, and 2) notions of blame and attribution, such as counterfactual fairness with regard to an intermediate sensitive concept.

Implication 6 (Input/Intermediate Counterfactual Equivalence)**.** For any given prompt $x \in X$, L and H should produce the same target concepts when intermediate concepts are intervened on (i.e. intermediate concepts H_k are counterfactually equivalent).

$$P_X(\tau(L_{i+1}),\ldots,\tau(L_n)|L_i=l;\text{do}(L_i=l')) = P_X(H_{i+1},\ldots H_n|H_i=\tau_i(l);\text{do}(H_i=\tau_i(l'))) \tag{4.8}$$

No current interpretability methods directly test for counterfactual equivalence of L and H.

5 Partial Evaluations of Interpretability

We propose algorithmic evaluations for each causal hierarchy level (associational, interventional, counterfactual) based on Section 4's implications. Ideally, L and H should behave identically within the use case domain. If X^E and X^U are very similar, weak evaluations are sufficient, but stronger faithfulness evaluations are required when X^E and X^U differ substantially.

5.1 Associational/Observational Evaluation

If X^U and X^E are identically distributed, then it's enough to show that L and H are *observationally equivalent* over the evaluation distribution, $\forall k$:

Evaluation 1 (Associational)**.** Assume X^U and X^E are identically distributed.

$$\textbf{Evaluating Implication 1: } \text{Accuracy}[Y_L(X^E),\quad Y_H(X^E)] \tag{5.1}$$

$$\textbf{Evaluating Implication 2: } \text{Accuracy}[\tau_k(L_k(X^E)),\quad H_k(X^E)] \tag{5.2}$$

Relevant interpretability methods: Equation 5.2 is **probing** [AB16] - τ_k is the probe trained to predict concept H_k from the activations L_k. Algorithm discovery in neural networks is complex: a wide diversity of solutions are plausible even for a simple learning problem [Zho+23].

[2] We only describe simple counterfactuals, but our evaluations hold for any counterfactual.

5.2 Interventional Evaluation

In the section 5.1 we assume that X^U and X^E are identically distributed, now we break this assumption, i.e. prompt distribution of the use case X^U differs from that of the evaluation distribution X^E. If we have enough high-level concepts H_k that Y_L can be fully derived from the values of $L_1, ..., L_n$ alone (without even needing the prompt), then this implies our high-level concepts H_k include all possible confounders.

Definition 2 (Causal Sufficiency)**.** We say that $L_1, ...L_n$ are causal sufficient for the target concept Y_L if there exists some f : domain$(X) \to$ domain(Y_L) such that $\forall x \in$ domain(H), $\forall l_i \in$ domain(τ_i),

$$Y_L(x, l_1, ..., l_n) = f(l_1, ..., l_n) \tag{5.3}$$

Causal sufficiency states H, τ account for all the variability of the target concept Y_L in L. Causal sufficiency fails, for instance, when there are computation paths from X to Y_L which bypass the $L_1, ..., L_n$, as in Figure 1. Therefore, known intermediates L_k may not capture all relevant computation in L. This will occur if H is causally insufficient for Y_H, i.e. there are unobserved confounders affecting both the input and target concepts. A possible example of this is [McG+23], demonstrates the existence of adaptive computation through ablations and causal analysis. Even if a computation is ablated, later redundancy ensures the computation is still completed. As long as we have causal sufficiency (Definition 5.3), then we can evaluate Implication 3 and Implication 4 by intervening on the low and high level model.

Evaluation 2 (Interventional)**.** Assume causal sufficiency holds.

Evaluating Implication 3:

$$\text{Accuracy}[Y_L(X^E; \text{edit}(L_k = l)), \quad Y_H(X^E; \text{edit}(H_k = \tau_k(l)))] \qquad \forall k \tag{5.4}$$

Evaluating Implication 4:

$$\text{Accuracy}[\tau_k(L_k(X^E; \text{edit}(L_j = l))), \quad H_k(X^E; \text{edit}(H_j = \tau_j(l)))] \qquad \forall k, \forall j < k \tag{5.5}$$

where $L_k(x; \text{edit}(L_j = l))$ is the value L_k takes on input x when L_j is intervened to value l.

Relevant interpretability methods: **Causal abstractions** [GPI23] and **Othello-GPT** [NLW23] employ a version of Equation 5.4, but differ in the choice of interventions l: causal abstractions selects values which occur naturally on the data manifold (via Equation 5.6), while causal scrubbing [Men+22] samples random interventions.

5.3 Counterfactual Evaluation

Now, assume that X^E and X^U are not identically distributed, and that the L_k's are not causally sufficient for the target concepts Y_L (i.e., they do not fully screen off the effect of the prompt). In this case, we can still monitor and control concepts if we can 'recreate' the values H_k takes over X^U, by stitching together the values each H_k takes on from X^E individually:

$$\forall u \sim X^U, \quad \forall H_k \sim H, \quad \exists e \sim X^E \quad \text{such that} \quad H_k(u) = H_k(e) \tag{5.6}$$

Evaluation 3 (Counterfactual)**.** Given Equation 5.6, evaluate for each k, j, and $j < k$:
Evaluating Implication 5:

$$\text{Accuracy}(Y_L(X^E | L_j = l; \text{edit}(L_j = l')), Y_H(X^E | H_j = \tau(l); \text{edit}(H_j = \tau(l')))) \tag{5.7}$$

Evaluating Implication 6:

$$\text{Accuracy}(\tau_k(L_k(X^E | L_j = l; \text{edit}(L_j = l'))), H_k(X^E | H_j = \tau(l); \text{edit}(H_j = \tau(l'))))) \tag{5.8}$$

$X | L_j = l$ indicates that we only pass in the prompts which induce $L_j = l$; we then pass these filtered prompts through the intervened model $L_k(X | L_j = l; \text{edit}(L_j = l'))$.

6 Towards Data-Driven Interpretability

Our approach to data-driven interpretability begins by fragmenting mechanistic faithfulness into partial evaluations. We aim to use the findings from 'weaker' evaluations as a stepping stone towards passing 'stronger' evaluations iteratively. In cases where a complete counterfactual specification of H is unattainable, possibly due to constraints in human understanding of intermediate concepts, we limit our evaluation to the strictest level feasible. Building on this framework, we adopt three key strategies: focusing on partial evaluations, prioritizing simpler ones to narrow the search space, and combining these results for a more in-depth interpretation.

Algorithm 1 Data-Driven Interpretability Algorithm

1: **Ensure H is well-defined:** $P(Y_L \mid x) = P(Y_H \mid x)$
2: **Probe Intermediate Concepts:** Verify $P(\tau(L_i) \mid x) = P(H_i \mid x)$
3: **Intervene on Probe:** Confirm $P(Y_L \mid x; \text{do}(L_i = l)) = P(Y_H \mid x; \text{do}(H_i = \tau(l)))$
4: **Refine using Intermediate Concepts:**
Confirm $P(\tau(L_{i+1}), \ldots, \tau(L_n)|x; \text{do}(L_i = l)) = P(H_{i+1}, ...H_n|x; \text{do}(H_i = \tau_i(l)))$
5: **Check for Causal Sufficiency:** Validate $Y_L(x; \text{do}(L_i = l)) = Y_L(x'; \text{do}(L_i = l))$
6: **Check for Input/Output Counterfactual Equivalence:**
$P_X(Y_L|L_i = l; \text{do}(L_i = l')) = P_X(Y_H|H_i = \tau_i(l); \text{do}(H_i = \tau_i(l')))$
7: **Check for Input/Intermediate Counterfactual Equivalence:**
$P_X(\tau(L_{i+1}), \ldots, \tau(L_n)|L_i = l; \text{do}(L_i = l')) = P_X(H_{i+1}, ...H_n|H_i = \tau_i(l); \text{do}(H_i = \tau_i(l')))$

The algorithm proceeds as follows: Step 1 - Confirms H has the same input-output behavior as L for a sanity check; Step 2 - Identifies relevant L_i subspaces by probing intermediate concepts; Step 3 - Refines L_i, conditioning on specific inputs x for probe validity; Step 4: Further refines L_i for matching interventional and predicted distributions; Step 5 - Validates causal sufficiency of probes, impacting final evaluation sufficiency; Step 6 - Confirms input-output counterfactual equivalence between L and H; Step 7 - Checks counterfactual equivalence of intermediate concepts.

Collectively, these steps provide a structured approach to approximating mechanistic faithfulness. Steps 2 and 3 guide the interpretation from inputs to intermediate variables and from intermediate variables to the target concept, respectively. The validity of combining these steps hinges on the causal sufficiency verified in Step 4. Importantly, if the high-level model lacks full mechanstic precision, the algorithm will make the most of easier early steps, and terminate once we reach a condition we cannot validate (e.g. if we don't know what the causal effect of a certain high-level variable is). Of particular note, the check for causal sufficiency fails only if there are unobserbed confounders of high-level intermediate concepts.

7 Conclusion

Achieving a mechanistic, human-understandable interpretation of a neural network/LLM is challenging, especially when the relationships among relevant high-level variables are not fully understood — either in the real world or within the model's internal representation.

In this work, we present a methodologically rigorous and pragmatic approach grounded in the causal hierarchy to navigate the complexities inherent in neural networks/LLMs interpretability. We dissect mechanistic faithfulness into constituent claims, revealing that existing interpretability methods are, in essence, tackling different facets of the same complex problem. Each facet warrants a distinct evaluation method and offers varying degrees of certification. In this way, we provide a data-driven pathway to post-hoc interpretability evaluations, and leverage any (limited) existing understanding of a low-level model L via the lens of a mechanistically faithful high-level model H.

Acknowledgments and Disclosure of Funding

We would like to thank Tom Everitt and Yibo Jiang for valuable discussions and feedback. DR is supported by the Long-Term Future Fund.

References

[AB16] G. Alain and Y. Bengio. "Understanding intermediate layers using linear classifier probes". *arXiv preprint arXiv:1610.01644* (2016) (cit. on p. 5).

[Bar+22] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. "On pearl's hierarchy and the foundations of causal inference". In: *Probabilistic and causal inference: the works of judea pearl*. 2022 (cit. on pp. 1, 2).

[Bel21] Y. Belinkov. *Probing classifiers: promises, shortcomings, and advances*. 2021. arXiv: 2102.12452 [cs.CL] (cit. on p. 4).

[Bel+23] N. Belrose, Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, S. Biderman, and J. Steinhardt. *Eliciting latent predictions from transformers with the tuned lens*. 2023. arXiv: 2303.08112 [cs.LG] (cit. on p. 4).

[Bur+22] C. Burns, H. Ye, D. Klein, and J. Steinhardt. *Discovering latent knowledge in language models without supervision*. 2022. arXiv: 2212.03827 [cs.CL] (cit. on p. 4).

[Cha+22] L. Chan, A. Garriga-Alonso, N. Goldwosky-Dill, R. Greenblatt, J. Nitishinskaya, A. Radhakrishnan, B. Shlegeris, and N. Thomas. "Causal scrubbing, a method for rigorously testing interpretability hypotheses". *AI Alignment Forum* (2022). https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing (cit. on p. 5).

[Con+23] A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. *Towards automated circuit discovery for mechanistic interpretability*. 2023. arXiv: 2304.14997 [cs.LG] (cit. on p. 5).

[GPI23] A. Geiger, C. Potts, and T. Icard. *Causal abstraction for faithful model interpretation*. 2023. arXiv: 2301.04709 [cs.AI] (cit. on pp. 1, 2, 6).

[Gei+23] A. Geiger, Z. Wu, C. Potts, T. Icard, and N. D. Goodman. *Finding alignments between interpretable causal variables and distributed neural representations*. 2023. arXiv: 2303.02536 [cs.AI] (cit. on p. 2).

[Gur+23] W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii, and D. Bertsimas. *Finding neurons in a haystack: case studies with sparse probing*. 2023. arXiv: 2305.01610 [cs.LG] (cit. on p. 4).

[HL22] Z. Hu and L. E. Li. *A causal lens for controllable text generation*. 2022. arXiv: 2201.09119 [cs.CL] (cit. on p. 3).

[Li+23] K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg. *Emergent world representations: exploring a sequence model trained on a synthetic task*. 2023. arXiv: 2210.13382 [cs.LG] (cit. on p. 4).

[MPT23] E. Marconato, A. Passerini, and S. Teso. *Interpretability is in the mind of the beholder: a causal framework for human-interpretable representation learning*. 2023. arXiv: 2309.07742 [cs.LG] (cit. on p. 3).

[McG+23] T. McGrath, M. Rahtz, J. Kramar, V. Mikulik, and S. Legg. *The hydra effect: emergent self-repair in language model computations*. 2023. arXiv: 2307.15771 [cs.LG] (cit. on p. 6).

[Men+22] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. "Locating and editing factual associations in gpt". *Advances in Neural Information Processing Systems* (2022) (cit. on p. 6).

[NLW23] N. Nanda, A. Lee, and M. Wattenberg. "Emergent linear representations in world models of self-supervised sequence models". *arXiv preprint arXiv:2309.00941* (2023) (cit. on p. 6).

[PH22] S. T. Piantadosi and F. Hill. *Meaning without reference in large language models*. 2022. arXiv: 2208.02957 [cs.CL] (cit. on p. 4).

[Räu+23] T. Räuker, A. Ho, S. Casper, and D. Hadfield-Menell. "Toward transparent ai: a survey on interpreting the inner structures of deep neural networks". In: *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE. 2023 (cit. on p. 1).

[Rä+23] T. Räuker, A. Ho, S. Casper, and D. Hadfield-Menell. *Toward transparent ai: a survey on interpreting the inner structures of deep neural networks*. 2023. arXiv: 2207.13243 [cs.LG] (cit. on p. 3).

[Wan+23] Z. Wang, L. Gui, J. Negrea, and V. Veitch. *Concept algebra for score-based conditional models*. 2023. arXiv: 2302.03693 [cs.CL] (cit. on p. 4).

[Wu+23] Z. Wu, A. Geiger, C. Potts, and N. D. Goodman. *Interpretability at scale: identifying causal mechanisms in alpaca*. 2023. arXiv: 2305.08809 [cs.CL] (cit. on pp. 2, 4).

[Zho+23] Z. Zhong, Z. Liu, M. Tegmark, and J. Andreas. "The clock and the pizza: two stories in mechanistic explanation of neural networks". *arXiv preprint arXiv:2306.17844* (2023) (cit. on p. 5).