Parasite Features: Causal Abstraction Through Spurious Pathways

Denis Sutter,[*,1] **Qing Yao,**[*,2] **Sasha Boguraev,**[*,2] **Julian Minder,**[3] **Antonio Di Maio,**[1]
Thomas Hofmann,[1] **Kyle Mahowald,**[2] **Tiago Pimentel**[1]
[1] ETH Zurich, [2] The University of Texas at Austin, [3] EPFL
densutter@ethz.ch, sasha.boguraev@utexas.edu,
qyao@utexas.edu, tiago.pimentel@inf.ethz.ch

dounick/parasite-features

Abstract

Causal Abstraction defines a notion of *equivalence* between a neural network and an algorithm, thus serving as a powerful framework for mechanistic interpretability. In this paper, however, we show the existence of a class of *spurious features* which straightforwardly pass such equivalence tests. Suppose there are two directions in a model's representation space where: (i) one discriminates the input based on some feature, and (ii) the other steers the model towards a certain output. Interpretability methods based on Causal Abstraction (e.g., Distributed Alignment Search) can *spuriously* simulate these directions' combined behavior by leveraging pathways passing through them, giving the impression that the model implements this feature. In such cases, we say the method discovered a **parasite feature**. Experimentally, we show that Distributed Alignment Search finds parasite features using linear alignment maps in two tasks: syntactic agreement and two-digit addition. Our paper highlights an important consideration for Causal Abstraction and contributes to the ongoing debate about best practices for mechanistic interpretability.

1 Introduction

The field of mechanistic interpretability aims to uncover the mechanisms underlying a model's inner workings (Olah et al., 2020), encompassing many methods and frameworks (e.g. Elhage et al., 2021; Mueller et al., 2026; Ferrando et al., 2024; Sharkey et al., 2025). Among these, **Causal Abstraction** is particularly promising, formally defining notions of equivalence between causal graphs (Beckers and Halpern, 2019). Under this framework, a deep neural network (DNN) is equivalent to a (conceptually simpler) algorithm if they behave similarly under interventions (Geiger et al., 2025b); whenever this is the case, we informally say that the DNN *implements* the algorithm. Causal Abstraction thus allows us to interpret a DNN in terms of the algorithms it implements.

In recent years, however, a number of works have raised potential pitfalls in using mechanistic interpretability approaches, including Causal Abstraction, to draw firm conclusions about a model (e.g., Mueller, 2024; Sutter et al., 2025; Méloux et al., 2025a, *inter alia*). Most relevant to us are two concerns. First, mechanistic interpretability approaches may be *non-identifiable* and a DNN might be found to implement many algorithms (Méloux et al., 2025b). Second, Causal Abstraction methods may find *interpretability illusions*, i.e., interventions that leverage a DNN's null space to imitate genuine causal behavior (Makelov et al., 2024). As noted by Wu et al. (2024b) in their response to Makelov et al., though, interpretability illusions may not pose an actual obstacle for Causal Abstraction, as it is not inherently problematic for an intervention to leverage a model's null space.

[*]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

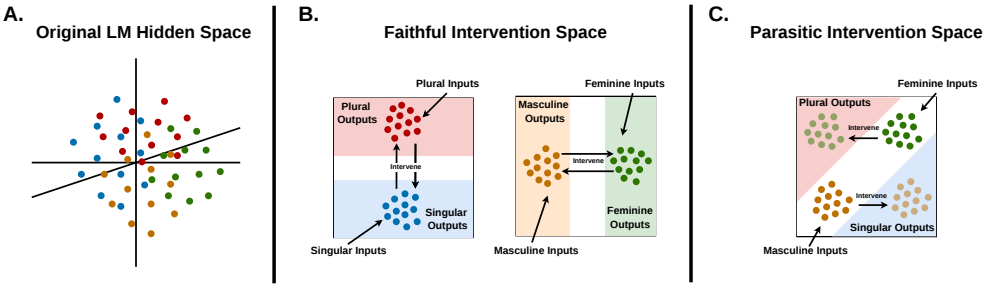


Figure 1: A schematic of parasite features for a hypothesized case of syntactic agreement (described in §3). **A:** Prompts in the original high-dimensional DNN hidden space are not basis-aligned by feature. **B:** By projecting onto subspaces, inputs can be separated by features. Interventions in this space move inputs to the region with corresponding output behavior. **C:** However, a linear alignment map may connect causally unrelated directions. Interventions in this space create the illusion of causality, despite there being no such relationship in the model. This depicts a *parasite feature.*

In this paper, we identify a related set of interpretability illusions which masquerade as causal features by manipulating spurious pathways in a DNN. Under mild assumptions, we show that alignment search methods can discover **parasite features**, a mechanistic interpretation of model behavior which *spuriously* mimics a target feature by leveraging two directions in the model where: (i) one discriminates the inputs in terms of this feature, and (ii) another steers the model's output towards an intended output. With the right alignment map and hidden state geometry, these directions can be linked together, leading to seemingly causal behavior under intervention, despite the leveraged directions being (under normal operation) causally unrelated in the model (see Fig. 1 for a visual example).

We present a theoretical account of parasite features (§4 and §5), followed by two empirical case-studies showing such parasite features are not merely theoretical concerns in toy models, but a practical concern (§6 and §7). Concretely, we find parasite features in two tasks, syntactic agreement and two-digit addition, by training interventions with a constrained class of linear alignment maps (which are minimally more expressive than rotation maps). We end with a discussion on the prevalence of parasite features and what these results mean for interpretability research more broadly (§8).

Our work demonstrates that while matching an algorithm's behavior under intervention is necessary for a feature to be faithful, it is not sufficient, even when alignments are restricted to linear or rotation maps. Claims that rest on high intervention accuracy alone should thus be treated with care, with our results further highlighting the importance of sound scientific practice in interpreting the results of mechanistic interpretability experiments. In doing so, we join a growing chorus emphasizing careful interpretation of mechanistic results, with a focus on robustness and generalizability (Geiger et al., 2025a; Sutter et al., 2025; Méloux et al., 2025a; Potts, 2025).

2 Causal Abstraction

In this section, we introduce relevant aspects of the Causal Abstraction framework (Rubenstein et al., 2017; Beckers and Halpern, 2019) used throughout our work. Causal Abstraction defines notions of equivalence between two structural causal models. In our case, these causal models are a deep neural network (DNN) $\mathtt{N}$ and an algorithm $\mathtt{A}$ which we might hypothesize the DNN implements.

First, we represent the **algorithm** $\mathtt{A}$ as an acyclic directed causal graph, which contains: (i) input nodes $\eta_{\mathbf{x}}$ representing input elements $\mathbf{x} \in \mathcal{X}$, (ii) output nodes $\eta_{\mathbf{y}}$ representing elements $\mathbf{y} \in \mathcal{Y}$, and (iii) inner nodes $\eta_{\mathtt{int}}$ representing intermediate variables whose value is computed as a function of its parents' values.[2] For simplicity of exposition, we will focus on analyses with a single feature, and we thus consider this graph to have a single inner node $\eta_{\mathtt{int}}$. Our results, however, can be easily extended to more complex graphs. Algorithm $\mathtt{A}$'s computation can be summarized as: $v = f_{\mathtt{A}}^{\eta_{\mathtt{int}}}(\mathbf{x})$ and $\mathbf{y} = f_{\mathtt{A}}^{\eta_{\mathbf{y}}}(\mathbf{x}, v)$. Now, consider this algorithm's behavior $\mathtt{A}$ under two different inputs, which we will call the **base** ($\mathbf{x}^{\mathtt{b}}$) and the **source** ($\mathbf{x}^{\mathtt{s}}$). We define a **source-to-base intervention** on inner node $\eta_{\mathtt{int}}$ as replacing this node's value in the base forward pass by its value under the source input, and continuing

[2]For more general formalisms covering, e.g., cyclic causal models, see Beckers and Halpern (2019) or Geiger et al. (2025b).

its forward pass as usual. Effectively, the intervention replaces just the feature value represented by $\eta_{\texttt{int}}$ in the base with a counterfactual value in its computation, leaving everything else untouched:

$$v^{\texttt{s}} = f_{\texttt{A}}^{\eta_{\texttt{int}}}(\mathbf{x}^{\texttt{s}}), \qquad v^{\texttt{s}\to\texttt{b}} = v^{\texttt{s}}, \qquad \mathbf{y}_{\texttt{A}}^{\texttt{s}\to\texttt{b}} = f_{\texttt{A}}^{\boldsymbol{\eta}_{\mathbf{y}}}(\mathbf{x}^{\texttt{b}}, v^{\texttt{s}\to\texttt{b}}) \tag{1}$$

Many **DNNs** $\texttt{N}$ can be similarly represented as acyclic directed causal graphs. In the case of transformer LMs, we can focus on a single layer ℓ and position, and write: $\mathbf{h} = f_{\texttt{N}}^{\ell}(\mathbf{x})$ and $\mathbf{y} = f_{\texttt{N}}^{\mathbf{y}}(\mathbf{h})$, where $\mathbf{h} \in \mathbb{R}^d$ is the hidden state. Representing each neuron in this layer as a distinct node, $\texttt{N}$ is thus a causal graph with nodes $\boldsymbol{\psi}_{\mathbf{x}} \cup \boldsymbol{\psi}_{\mathbf{y}} \cup \boldsymbol{\psi}_{\texttt{int}}$, where $\boldsymbol{\psi}_{\mathbf{x}}$, $\boldsymbol{\psi}_{\mathbf{y}}$ are defined as for the algorithm, and $\boldsymbol{\psi}_{\texttt{int}}$ contains d nodes corresponding to neurons in $\mathbf{h}$. We can now define source-to-base interventions on a DNN as before: given a chosen neuron i, compute hidden states for both the base and source, replace the base value of a neuron i with the source's, and continue the forward pass. Formally:

$$\mathbf{h}^{\texttt{b}} = f_{\texttt{N}}^{\ell}(\mathbf{x}^{\texttt{b}}), \quad \mathbf{h}^{\texttt{s}} = f_{\texttt{N}}^{\ell}(\mathbf{x}^{\texttt{s}}), \quad \mathbf{h}^{\texttt{s}\to\texttt{b}} = \left(\mathbf{h}_1^{\texttt{b}}, \mathbf{h}_2^{\texttt{b}}, \ldots, \mathbf{h}_i^{\texttt{s}}, \ldots, \mathbf{h}_d^{\texttt{b}}\right), \quad \mathbf{y}_{\texttt{N}}^{\texttt{s}\to\texttt{b}} = f_{\texttt{N}}^{\mathbf{y}}(\mathbf{h}^{\texttt{s}\to\texttt{b}}) \tag{2}$$

Several definitions of equivalence between causal graphs fall under Causal Abstraction. **Constructive abstraction** (Beckers and Halpern, 2019), defines an equivalence between $\texttt{N}$ and $\texttt{A}$ iff there exists a set of neurons such that $\mathbf{y}_{\texttt{N}}^{\texttt{s}\to\texttt{b}} = \mathbf{y}_{\texttt{A}}^{\texttt{s}\to\texttt{b}}$ holds for any $\mathbf{x}^{\texttt{s}}$, $\mathbf{x}^{\texttt{b}}$ pair.[3] This definition may be too restrictive, as it assumes the high-level conceptual information represented by $\eta_{\texttt{int}}$ is encoded in neuron-aligned bases of the representation space—something typically deemed too strong in DNNs (see Sutter et al., 2025 for a longer discussion). In turn, other notions of Causal Abstraction (Geiger et al., 2024, 2025b; Sutter et al., 2025) define an equivalence between $\texttt{N}$ and $\texttt{A}$ allowing for information to be latently encoded in $\mathbf{h}$. Formally, let $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^d$ be an **alignment map**: a learned bijective function which exposes a set of latent variables the representations $\mathbf{h}$ encode, with $\mathbf{h}_{\boldsymbol{\phi}} = \boldsymbol{\phi}(\mathbf{h})$. We can use such an alignment map to perform source-to-base interventions on $\mathbf{h}_{\boldsymbol{\phi}}$ as in eq. (2), producing vector $\mathbf{h}_{\boldsymbol{\phi}}^{\texttt{s}\to\texttt{b}}$. The notion of $\boldsymbol{\mathcal{V}}$**-abstraction** then defines an equivalence between $\texttt{N}$ and $\texttt{A}$ if there exists a $\boldsymbol{\phi} \in \mathcal{V}$ (for a variational family of bijective functions $\mathcal{V}$) such that $\mathbf{y}_{\texttt{N}}^{\texttt{s}\to\texttt{b}} = \mathbf{y}_{\texttt{A}}^{\texttt{s}\to\texttt{b}}$ holds for any $\mathbf{x}^{\texttt{s}}$, $\mathbf{x}^{\texttt{b}}$ pair.

The chosen class of alignment maps $\mathcal{V}$ reflects assumptions about how DNNs encode conceptual information. Under the linear representation hypothesis (Alain and Bengio, 2016), this class would be restricted to rotations (Geiger et al., 2024), or the more expressive class of invertible linear maps (Grant et al., 2025a,b). However, there is growing evidence that neural networks encode relevant feature information nonlinearly (White et al., 2021; Pimentel et al., 2022; Csordás et al., 2024; Levy and Geva, 2025; Kantamneni and Tegmark, 2025; Mueller et al., 2025), which warrants the use of nonlinear maps. At the same time, increasing $\mathcal{V}$'s expressivity has risks: if $\mathcal{V}$ encompasses arbitrary nonlinear bijections, *any* neural network can be mapped to *any* algorithm (Sutter et al., 2025). In this work, we show even rotation or linear alignment maps are not safe from fitting spurious features.

3 Motivating Example: Parasitic Agreement

We start with a thought experiment to motivate parasite features focusing on two forms of syntactic agreement: gender-agreement and number-agreement. An example of gender-agreement can be seen in Example 1. Here, the grammatical gender of the sentence's subject (as in *man* vs. *woman*) is responsible for licensing the correct anaphoric pronoun (*he* vs. *she*).

(1) a. The **man** ran down the street because ⇒ **he**
b. The **woman** ran down the street because ⇒ **she**

Similarly, in the case of number-agreement (Example 2), the plurality of the sentence's subject (as in *key* vs. *keys*) is responsible for licensing the correct verb form (*is* vs. *are*).

(2) a. The **key** next to the counter ⇒ **is**
b. The **keys** next to the counter ⇒ **are**

A competent LM should recognize these dependencies and predict the correct output given the upstream item. One should thus be able to learn an alignment map such that manipulating the

[3] More precisely, constructive abstraction (Beckers and Halpern, 2019) requires two stronger conditions to hold: (i) $\mathbf{y}_{\texttt{N}}^{\texttt{s}\to\texttt{b}} = \mathbf{y}_{\texttt{A}}^{\texttt{s}\to\texttt{b}}$ must hold for any "*compatible*" intervention-pair ($\mathbf{h}^{\texttt{s}\to\texttt{b}}$, $v^{\texttt{s}\to\texttt{b}}$) not just pairs generated as above; (ii) the intervened value $\mathbf{h}_i^{\texttt{s}}$ must encode sufficient information to predict $v^{\texttt{s}}$. Given transformer's injectivity property (Sutter et al., 2025; Nikolaou et al., 2025), we assume (ii) holds here. Regarding (i), we will work here with the input-restricted variants of causal abstraction (presented by Sutter et al., 2025), thus softening requirement (i) to intervention-pairs ($\mathbf{h}^{\texttt{s}\to\texttt{b}}$, $v^{\texttt{s}\to\texttt{b}}$) defined as above. See Sutter et al. (2025) for a longer discussion on these definitions, and a more precise definition of "compatible" interventions.

upstream feature in an LM leads to a corresponding change in the LM's output behavior. However, let us now consider a manufactured example in which the gender of a sentence's subject is responsible for licensing the plurality of a downstream verb (Example 3).

(3) a. The **man** next to the counter ⇒ **is**
b. The **woman** next to the counter ⇒ **are**

We would expect our linguistically-competent LM to not behave accordingly. As such, there should not be an alignment map under which manipulating subject gender controls downstream verb plurality. But this is not guaranteed, and below we construct a simple counterexample.

Consider a normal English LM, $\mathtt{N}$, that performs number-agreement. Assume that in some layer ℓ, there exists a direction that influences downstream verb number. We can construct another LM, $\mathtt{N}'$, which is an exact copy of $\mathtt{N}$, except for an additional neuron in layer ℓ. This neuron is computationally inert, having no outgoing weights, and simply encodes gender (e.g., with activation -1 for masculine, $+1$ for feminine). By construction, $\mathtt{N}'$ is behaviorally indistinguishable from $\mathtt{N}$ on all inputs.

However, despite the additional neuron playing no causal role in the model, it can be exploited by alignment maps. Consider a class of maps $\mathcal{V}$ and an alignment map $\phi \in \mathcal{V}$ that projects the model's activations onto a direction at a $45°$ angle with: (i) the direction of the extra inert neuron and (ii) the direction causally associated with **number**—virtually combining them into a latent feature. By linking these directions, ϕ can induce changes in downstream verb plurality, despite the signal originating from a feature (**gender**) which is never used by our model (as is exemplified in Fig. 1C).

This is an archetypal parasite feature — a counterfactual gender-to-number feature is assumed to be implemented in $\mathtt{N}$ by an alignment map which combines unrelated directions in a manner which is irrelevant to the model's computation. In real LMs, there are no artificial neurons that encode input features, but such directions often natively exist (e.g., a direction which spuriously separates inputs by gender in a model that performs number-agreement; Ravichander et al., 2021). While these directions are not as explicitly causally inert as the added neuron in $\mathtt{N}'$, there is likewise no guarantee that interventions leveraging them are faithful to the model's computation.[4]

4 Defining Features, Faithful Features, and Parasite Features

Before formalizing *parasite features*, we will first formalize the concept of a *feature*. We take the perspective that a **feature** is a node in a causal graph. In this sense, our **ground truth features** are elements in $\eta_{\mathtt{int}}$. This has at least one important implication: a feature has not only an **input-side component**, which defines the set of values it takes given the input (i.e., $v = f_{\mathtt{A}}^{\eta_{\mathtt{int}}}(\mathbf{x})$), but also an **output-side component**, which defines its impact on an algorithm's outputs (i.e., $\mathbf{y} = f_{\mathtt{A}}^{\eta_{\mathbf{y}}}(\mathbf{x}, v)$). We also have **candidate features** in a DNN: each of its explicit $\mathbf{h}_i$ or latent $\mathbf{h}_{\phi,i}$ nodes is a potential feature used for computations. Our goal is then to find features in a DNN which are faithful to a target feature.

Given a target feature ($\eta_{\mathtt{int}}$), we can use different techniques to find it in a model (i.e., to find an equivalent $\mathbf{h}_{\phi,i}$), including diagnostic probing (Alain and Bengio, 2016), sparse-autoencoders (Bricken et al., 2023; Huben et al., 2024), and distributed alignment search (Geiger et al., 2024). However, how do we know the feature we find in a model is faithful? Jacovi and Goldberg (2020) intuitively define a **faithful feature** to be one that *accurately represents the reasoning process behind the model's prediction* (although see Ribeiro et al., 2016; Wiegreffe and Pinter, 2019; Bastings and Filippova, 2020; Hanna et al., 2024, *inter alia*, for more definitions).[5] Under this definition, a clear condition for a feature's faithfulness is being **causally meaningful**: i.e., changes to the candidate feature in $\mathtt{N}$ must match the ground truth feature's behavior under $\mathtt{A}$. Importantly, causal abstraction gives us—by construction—features which are causally meaningful in this sense, as interventions on the model $\mathtt{N}$ must match the evaluated algorithm $\mathtt{A}$.

Is causal meaningfulness a sufficient condition for a feature to be faithful, however? We argue that it is not. Sutter et al. (2025) showed that, given a strong enough class of alignment maps $\mathcal{V}$, we can find *any* feature in *any* model using causal abstraction. In this paper, we document another class of features which we argue are not faithful, but still pass the causally meaningful test: parasite features.

[4] We note that such an example is very similar to the one provided by Makelov et al. (2024). Our example differs, however, in a couple regards. First, our steering direction is the LM's *natively active, on-manifold* direction, whereas Makelov's "dormant" steer direction is zero on-data and only fires off-distribution; second, we link *two distinct concepts* rather than a disconnected correlate of the *same* feature. For a larger discussion on the similarities and differences between the phenomena see §8.

[5] Faithfulness is thus a property of a candidate feature as compared to a ground truth, being a property of a triple: $\langle \mathtt{A}, \mathtt{N}, \phi \rangle$.

Parasite features are, informally, a spurious connection between two directions of a DNN's hidden space: a discriminative direction $\psi^{\texttt{DD}}$, which separates hidden states into two distinct clusters; and a steering direction $\psi^{\texttt{SD}}$, which predictably affects the model's output behavior when perturbed. Formally, let $f : \mathcal{X} \to \{0, 1\}$ be an input-feature function and $g : \mathcal{X} \times \{0, 1\} \to \mathcal{Y}$ be an output-behavior function. Now, let an input $\mathbf{x}$ be **measurably true in input-feature** f if $f(\mathbf{x}) = 1$, and measurably false otherwise; we can use this definition to cluster $\texttt{N}$'s hidden states (i.e., $\mathcal{H} = \{f^{\ell}_{\texttt{N}}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$) into sets $\mathcal{H}_0$ and $\mathcal{H}_1$ produced by measurably false and true inputs, respectively. Further, let an input–output pair $\mathbf{x}, \mathbf{y}$ be **measurably true in output-behavior** g if $\mathbf{y} = g(\mathbf{x}, 1)$, measurably false if $\mathbf{y} = g(\mathbf{x}, 0)$, and undefined otherwise.[6] We can now define directions $\psi^{\texttt{DD}}$ and $\psi^{\texttt{SD}}$ formally.

Definition 1. *Let $\texttt{N}$ be a DNN and f an input-feature function. A unit vector $\psi^{\texttt{DD}} \in \mathbb{R}^d$ is a **discriminative direction** for $(\texttt{N}, f)$ if there exists a threshold $\delta_{\texttt{disc}} > 0$ which separates the prompts into clusters such that: $\langle \psi^{\texttt{DD}}, \mathbf{h} \rangle > \delta_{\texttt{disc}}$ for all $\mathbf{h} \in \mathcal{H}_1$ and $\langle \psi^{\texttt{DD}}, \mathbf{h} \rangle < -\delta_{\texttt{disc}}$ for all $\mathbf{h} \in \mathcal{H}_0$.*

Definition 2. *Let $\texttt{N}$ be a DNN and g an output-behavior function. A unit vector $\psi^{\texttt{SD}} \in \mathbb{R}^d$ is a **steering direction** for $(\texttt{N}, g)$ if there exists a threshold $\delta_{\texttt{steer}} > 0$ such that whenever $\langle \mathbf{h}, \psi^{\texttt{SD}} \rangle > \delta_{\texttt{steer}}$, the output of $\mathbf{h}$ under $\texttt{N}$ is **measurably true** in behavior g, and **measurably false** whenever $\langle \mathbf{h}, \psi^{\texttt{SD}} \rangle < -\delta_{\texttt{steer}}$.*

Our definition of steering directions is intentionally broad. Some steering directions are intuitive: a model can implement a direction that *discriminates* inputs by grammatical number and that also *steers* its outputs to change a verb's number. However, some steering directions can be more abstract: e.g., a model can have a direction that systematically adds a constant to arithmetic results, or that steers an output verb's number without simultaneously discriminating inputs. Strange though we know these directions may seem, we find that they exist in pretrained models (in §7; with further evidence of their prevalence in §A). These directions allow parasite features to emerge.

Definition 3. *Consider an $\texttt{N}$, f, g with two* causally independent *directions: a discriminative direction, $\psi^{\texttt{DD}}$, and steering direction, $\psi^{\texttt{SD}}$. Finally, let $\texttt{A}$ have $f^{\eta_{\texttt{int}}}_{\texttt{A}} = f$ and $f^{\boldsymbol{\eta}_{\mathbf{y}}}_{\texttt{A}} = g$. We say an alignment map discovers a **parasite feature** if it leverages a combination of $\psi^{\texttt{DD}}$ and $\psi^{\texttt{SD}}$ to show $\texttt{N}$ implements $\texttt{A}$, using the discriminative direction $\psi^{\texttt{DD}}$ to move representations in the steering direction $\psi^{\texttt{SD}}$.*[7]

Importantly, we rely informally on the term 'causally independent' above to mean the input-side feature f and output-side feature g 'operate independently' within the model. Defining what it means for these directions to 'operate independently' in a deep neural network, however, is non-trivial. Instead, we offer two conditions which we believe should (in general) be sufficient for such independence to hold: **(i) data-level independence**, in which directions operate on conceptually distinct and causally unrelated features of the training data (e.g., subject gender and verb number), so a competent model should process them independently, and **(ii) subspace-level independence**, in which the two directions do not interact in the model's computation, meaning that moving along the discriminative direction does not affect the decision boundary set by the steering direction. We argue only that these conditions are sufficient — but not strictly necessary — for two directions to operate independently within a competent model. Two directions ($\psi^{\texttt{DD}}$ and $\psi^{\texttt{SD}}$) which operate on conceptually related features (data-level dependent) may still be processed independently by the model. Similarly, two directions may influence one another (subspace-level dependence), but the model may have mechanisms in place to offset those interactions under its normal operation.

5 Sufficient Conditions for Discovering Parasite Features

We now state formal conditions under which there exists an invertible linear map that can leverage a discriminative direction for f to influence the model's behavior towards g via a steering direction, regardless of any causal relationship between them. The implications of the theorem are particularly worrisome when f and g are causally disconnected in a target domain, making a feature $\eta_{\texttt{int}}$ combining their behavior implausible. Now, let $\mathcal{H} = \{f^{\ell}_{\texttt{N}}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$ be a set of d-dimensional hidden states at layer ℓ, with $d \geq 2$, partitioned by f into $\mathcal{H}_0, \mathcal{H}_1$. We denote elements of $\mathcal{H}$ by $\mathbf{h}$.

[6] Note that, given a binary feature $\eta_{\texttt{int}}$, our definitions of f and g precisely match an algorithm's structure $v = f^{\eta_{\texttt{int}}}_{\texttt{A}}(\mathbf{x})$ and $\mathbf{y} = f^{\boldsymbol{\eta}_{\mathbf{y}}}_{\texttt{A}}(\mathbf{x}, v)$. We make this connection more precise later. Further note that the definitions of 'measurably true' and 'measurably false' in g can be trivially extended to operate on $\texttt{N}$'s logits directly (e.g., the probability of a True class should exceed a False class), instead of on the produced output $\mathbf{y}$. This fact allows us to treat the logit-based metrics utilized in §7 as "measurably true/false", allowing experimental validation of our theoretical proof.

[7] Parasitism is a property of a candidate feature as compared to a ground truth, and thus a property of $\langle \texttt{A}, \texttt{N}, \boldsymbol{\phi} \rangle$.

We define our theorem's conditions as follows.

Assumption 1. ***(Steerability.)*** *A steering direction $\psi^{\texttt{SD}}$ exists for g with threshold $\delta_{\texttt{steer}} > 0$.*

Assumption 2. ***(Discrimination.)*** *A discriminative direction $\psi^{\texttt{DD}}$ exists for f with threshold $\delta_{\texttt{disc}} > 0$.*

Assumption 3. ***(Sufficient separation.)*** *Let (i) $\varepsilon_{\texttt{gap}} > 0$ be the minimum gap between hidden states and the $\pm\delta_{\texttt{steer}}$ steering threshold; (ii) $\varepsilon_{\texttt{max}} > 0$ be the maximum activation of the steering direction by hidden states; and (iii) $\varepsilon_{\texttt{spread}} > 0$ be the maximum within-class spread along $\psi^{\texttt{DD}}$:*

$$\varepsilon_{\texttt{gap}} = \min_{\mathbf{h}\in\mathcal{H}} \Big| \, |\langle \mathbf{h}, \psi^{\texttt{SD}} \rangle| - \delta_{\texttt{steer}} \, \Big|, \quad \varepsilon_{\texttt{max}} = \max_{\mathbf{h}\in\mathcal{H}} |\langle \mathbf{h}, \psi^{\texttt{SD}} \rangle|, \quad \varepsilon_{\texttt{spread}} = \max_{\mathbf{h}_s, \mathbf{h}_b \in \mathcal{H}_i} |\langle \psi^{\texttt{DD}}, \mathbf{h}_s - \mathbf{h}_b \rangle| \tag{3}$$

We require $|\langle \psi^{\texttt{DD}}, \psi^{\texttt{SD}} \rangle| < 1$ and a minimum gap-to-spread ratio satisfying $\frac{\varepsilon_{\texttt{gap}}}{\varepsilon_{\texttt{spread}}} > \frac{\varepsilon_{\texttt{max}} + \delta_{\texttt{steer}}}{2\,\delta_{\texttt{disc}}}$.

In words, assumptions 1 and 2 state the functions are associated with discriminative and steering directions in the model, while assumption 3 defines a minimum ratio between the gap between representations and the steering direction; and the spread of representations per cluster $\mathcal{H}_i$ along the discriminative direction. As representations naturally cluster according to input features, we expect the mere existence of steering directions to make these conditions reasonable and to hold for a number of f, g function pairs at some position and layer within a model.

We can now present our theorem.

Theorem 1. *Let $\texttt{N}$ be a DNN, and $\texttt{A}$ be an algorithm with $f_{\texttt{A}}^{\eta_{\texttt{int}}} = f$ and $f_{\texttt{A}}^{\eta_{\texttt{Y}}} = g$. Further, let $\mathcal{V}^{\texttt{lin}}$ be the class of invertible linear maps. Under assumptions 1, 2, and 3, $\texttt{A}$ is a $\mathcal{V}^{\texttt{lin}}$-abstraction of $\texttt{N}$.*

Proof sketch. See §B for a formal proof. We prove this by construction, building an alignment map $\phi \in \mathcal{V}^{\texttt{lin}}$ with the following intuition: given two points in the separate sets $\mathcal{H}_0, \mathcal{H}_1$, provided the inputs are cleanly separated and clustered along the discriminative direction, an invertible linear matrix can always rotate the difference between the two points to align with the steering direction, and amplify it appropriately to "bridge the gap" in the steering direction. Assumption 3 further ensures that perturbations for within-class interventions remain in their output region. □

Intuitively, Theorem 1 shows that, if assumptions 1–3 are satisfied, there exists an alignment map $\phi \in \mathcal{V}^{\texttt{lin}}$ which aligns $\texttt{A}$ to $\texttt{N}$, such that $\texttt{N}$'s output aligns with the behavior of a combined (f, g). Notably, this theorem holds even if directions $(\psi^{\texttt{DD}}, \psi^{\texttt{SD}})$ are orthogonal (suggesting subspace-level independence) or operate on conceptually unrelated features (suggesting data-level independence). We prove a similar theorem, with stricter assumptions, for the class of rotation alignment maps in §C. This highlights that parasite features are a general problem for Causal Abstraction, and that even less expressive alignment maps are not necessarily immune to discovering spurious features.

6 Experimental Setup

We find parasite features in two tasks: first, a nonsensical syntactic task (PARASITIC AGREEMENT) involving unrelated syntactic features; second, a task localizing a carry-one feature in two-digit addition (CARRY-ONE). These tasks differ in that we have strong reason to believe the model should *not* be competent in the syntax task, as it is constructed to be parasitic. In contrast, the carry-one feature in two-digit addition appeared in the mechanistic interpretability benchmark (MIB; Mueller et al., 2025) as a plausible faithful feature. We discuss additional parasite features and tasks in §A.

6.1 Tasks and Causal Algorithms

PARASITIC AGREEMENT. We create pairs of single-clause sentences labeled for subject gender. Sentences with masculine subjects are then given a singular verb, whereas feminine subjects have plural verbs (see Example 3). A candidate causal-algorithm, $\texttt{A}_{\texttt{agr}}$, for solving this task can be informally stated as: if the subject of the sentence is masculine, output a singular verb, otherwise if the subject of the sentence is feminine, output a plural verb.

We use `pythia-1b` (Biderman et al., 2023) because it is competent at number and gender agreement, and rotation alignment maps can align these features with strong causal effects (Arora et al., 2024). In relation to Theorem 1, the output behavior is whether singular continuations are favored over

plurals, with the steering direction controlling logits of singular and plural verbs. All prompts begin in the singular region (subjects are always singular by design), so the alignment map must transport feminine-subject hidden states to the plural region under intervention.

CARRY-ONE. Models receive string inputs of a two-digit addition problem and provide the output string of the correct sum (see Example 4). The mechanistic interpretability benchmark (Mueller et al., 2025) hypothesizes a model may use the following algorithm, $\texttt{A}_{\texttt{carry}}$: compute the sum of the 'ones' digits and the sum of the 'tens' digits, along with a carry-one indicator that is true when the ones-digit sum exceeds 9. The final output concatenates the ones-digit sum (modulo 10) with the tens-digit sum, incrementing the latter by one when the carry-one indicator is true.

Successfully localizing the carry-one variable permits source-to-base interventions that adjust the tens digit accordingly. An example of this intervention is in Example 4. Here, the carry-one value is $\eta^b_{\text{carry}} = 1$ (Example 4a) and $\eta^s_{\text{carry}} = 0$ (Example 4b). A successful intervention from Example 4b to Example 4a should make $\eta^b_{\text{carry}} = 0$, causing the tens digit to decrement by 1.

(4)
a. The sum of 27 and 64 is $\Rightarrow$ 91 ($\eta^b_{\text{carry}} = 1$)
b. The sum of 18 and 11 is $\Rightarrow$ 29 ($\eta^s_{\text{carry}} = 0$)
c. **Intervention** ($\eta^{s\to b}_{\textbf{carry}}$)**:** The sum of 27 and 64 is $\Rightarrow$ 81

For this task, we use `Llama3.1-8b` (Grattafiori et al., 2024), which despite achieving over 99% accuracy on these addition problems, cannot be aligned with the carry-one variable using rotation-based alignment maps (i.e., maps of class $\mathcal{V}^{\texttt{rot}}$ defined below; Mueller et al., 2025).

6.2 Alignment Maps

Theorem 1 defines conditions in which parasitic features could arise for linear alignment maps $\mathcal{V}^{\texttt{lin}}$. In practice, though, we find parasite features with maps much more constrained than the full class of alignment maps defined by $\mathcal{V}^{\texttt{lin}}$ and only marginally more expressive than that of rotation maps $\mathcal{V}^{\texttt{rot}}$.

Classically, DAS uses rotation alignment maps, mapping hidden states $\mathbf{h}$ to a latent space under $\mathbf{h} \mapsto \boldsymbol{\phi}(\mathbf{h})$, with $\boldsymbol{\phi} \in \mathcal{V}^{\texttt{rot}}$ a learned orthogonal transformation. In our case studies, however, we apply alignment maps from the class **rotation+linear**, consisting of maps which first apply an orthogonal map to the full hidden state, then an invertible linear transformation to the first i coordinates ($1 \leq i \leq d$), leaving the remaining coordinates unchanged. Formally, we define this class as

$$\mathcal{V}^{\texttt{r+l}} \;=\; \left\{ (\mathbf{L} \oplus \mathbf{I}_{d-i})\, \boldsymbol{\phi} \mid \mathbf{L} \in \mathcal{V}^{\texttt{lin}}_{i\times i}, \boldsymbol{\phi} \in \mathcal{V}^{\texttt{rot}} \right\}, \tag{4}$$

with I_{d-i} being an identity matrix of size $d-i$. In practice, we choose $i = 2$.

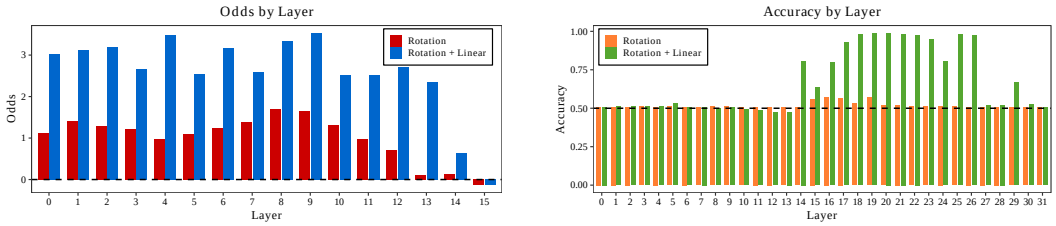
Our choice of map and i is motivated by two main factors. First, we find that rotation+linear with $i = 2$ can align CARRY-ONE and PARASITIC AGREEMENT. This shows parasite features can be discovered with only minimal increase to expressivity from rotation. Second, the $i = 2$ dimensions processed by $\mathcal{V}^{\texttt{lin}}$ can be visualized to assess the geometry associated with parasite features.

We are not aware of any other work using these alignment maps, but more expressive linear maps have already been employed (Grant et al., 2025a,b). The effectiveness of rotation+linear at finding parasite features suggests that over-expressivity is not restricted to nonlinear maps (Sutter et al., 2025).

6.3 Training and Evaluation Details

Training. We train interventions in the residual stream at the last token of the subject for PARASITIC AGREEMENT (a site shown to be highly causally efficacious in syntax; Arora et al., 2024; Boguraev et al., 2025) and the last token of the input for CARRY-ONE (as per Mueller et al., 2025). We train only on pairs of inputs that the models solve correctly. We train alignments of both $\mathcal{V}^{\texttt{rot}}$ and $\mathcal{V}^{\texttt{r+l}}$, for contrast between different maps. For $\mathcal{V}^{\texttt{rot}}$, we learn 1D alignments, and for $\mathcal{V}^{\texttt{r+l}}$, we use a 2D linear map ($i = 2$), and intervene on the first dimension. See §D for details.

Evaluation. For CARRY-ONE it is straightforward to measure an intervention's accuracy, as there is only one correct answer. However, it is more complicated for PARASITIC AGREEMENT since there are multiple grammatical continuations rather than a single correct continuation. Instead, we measure



(a) **ODDS** by layer for PARASITIC AGREEMENT. (b) Intervention accuracy by layer for CARRY-ONE.

Figure 2: **Left: ODDS** of the learned $\mathcal{V}^{\texttt{rot}}$ and $\mathcal{V}^{\texttt{r+l}}$ for PARASITIC AGREEMENT on a held-out test set. **Right:** Accuracy of the learned $\mathcal{V}^{\texttt{rot}}$ and $\mathcal{V}^{\texttt{r+l}}$ for CARRY-ONE when evaluated similarly. Dotted line indicates chance performance of intervention. In both cases, we see $\mathcal{V}^{\texttt{r+l}}$ display strong causal efficacy, more so than the less expressive $\mathcal{V}^{\texttt{rot}}$.

the **ODDS** of a given intervention. This metric, introduced by Arora et al. (2024), intuitively measures how much more probable the intervened label gets after intervention, relative to the original label.

7 Results

Our results can be seen in Fig. 2: Fig. 2a displays the **ODDS** of the learned interventions in each layer of `pythia-1b` when evaluated on our held-out test-set, and Fig. 2b the accuracy of the learned interventions in `Llama3.1-8b` when evaluated similarly. These plots demonstrate that we are able to learn alignment maps of the class $\mathcal{V}^{\texttt{r+l}}$ which show strong causal efficacy across layers. This is especially notable in comparison to the alignment maps we learn of class $\mathcal{V}^{\texttt{rot}}$. While $\mathcal{V}^{\texttt{rot}}$ achieves non-trivial odds, they lack the level of expressivity to consistently align the features.

The discrepancy between the efficacy of $\mathcal{V}^{\texttt{rot}}$ and $\mathcal{V}^{\texttt{r+l}}$ is not itself diagnostic of parasitic features. After all, it is not out of the question that DNNs may use representations more complicated than those found by $\mathcal{V}^{\texttt{rot}}$. We must perform further analysis to understand the nature of these alignments.

7.1 Confirming Parasitism

To investigate whether alignments are parasitic, we further analyze the most efficacious layers (layers 4 and 20 for PARASITIC AGREEMENT and CARRY-ONE respectively).

We uniformly sweep through points in the 2D latent space of our learned alignments, and for each point, we randomly select a prompt from the test set as our base value and intervene on it to relocate it to that point in space. We then measure the degree to which the output of this intervention deviates from the original prompt's output. For PARASITIC AGREEMENT our measure of deviation is

$$\Delta_{\text{PARASITIC AGREEMENT}} = (P_{\text{after}}(\text{plural}) - P_{\text{before}}(\text{plural})) + (P_{\text{before}}(\text{singular}) - P_{\text{after}}(\text{singular}))$$

where P_{before} and P_{after} are the LM's next-token log-probability before and after intervention. This is a gradable measure of how much more likely a singular (negative) or plural (positive) next token is.

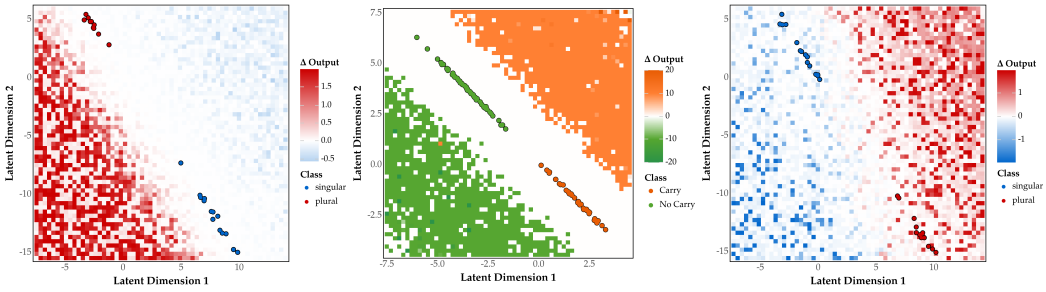
Similarly, for CARRY-ONE, our measure of deviation is

$$\Delta_{\text{CARRY-ONE}} = n_{\text{after}} - n_{\text{before}}$$

where n_{before} and n_{after} are the values of the `argmax` numerical token before and after intervention.

The resulting landscapes in Fig. 3 show remarkably consistent structure. The two classes of minimally paired data are cleanly separated along one diagonal — the discriminative direction — with the two output regions cleanly separated along the orthogonal diagonal — the steering direction. These landscapes, in matching our description of parasite features, serve as diagnosis for parasitism.

To further show the dimensions discovered by rotation+linear for CARRY-ONE do not participate in the computation, we ablate them — projecting the residual stream at the given layer onto the orthogonal complement — before continuing the forward pass. As a baseline, `Llama3.1-8b` achieves an accuracy of 99.1% on two-digit addition tasks. Ablating the 2D subspace discovered by the alignment map gives an accuracy of 98.8%. As a control, ablating random 2D subspaces at that layer over 10 trials yields $99.12\% \pm 0.13$ accuracy. These results suggest that while the parasite



(a) PARASITIC AGREEMENT (b) CARRY-ONE (c) Subject-Verb Agreement

Figure 3: Intervention landscapes for three tasks. Each point is colored by the class the model outputs when we intervene at that location. In the parasitic cases (left and middle), original prompts are cleanly separated in a region of no deviation. The remaining intervention space is partitioned into regions corresponding to plural/singular distinctions (left) and ± 10 (middle). The discriminative and steering directions are along the anti-diagonal and main-diagonal, respectively. In the faithful landscape (right), input points are positioned within their respective class's intervention region and the discriminative and steering directions can be perfectly aligned.

feature is sufficient for controlling a desired intervention behavior, it is not *necessary* for addition.[8] We note that a similar ablation for parasitic agreement is not possible, as the model already does not solve this parasitic task under normal operation.

7.2 Aligning on Faithful Features

Fig. 3c presents the intervention landscapes of a rotation+linear alignment trained on faithful subject-verb number agreement in the layer of best efficacy (layer 2, **ODDS** of 4.3). The prompts lie in the corresponding output regions, with discriminative and steering directions non-orthogonal. This confirms our orthogonality diagnostic reflects meaningful geometry rather than visualization artifact.

8 Discussion

Prevalence of Parasite-Features. The ubiquity of our conditions suggests parasite features are likely pervasive, and indeed we easily constructed the PARASITIC AGREEMENT task to find them. Furthermore, it is not the case that these parasite features are model specific: in §E, we show that a parasitic solution to PARASITIC AGREEMENT can be found in other models. However, such a constructed feature is not one a researcher would hypothesize a model implements, so its existence is benign. Parasite features only become problematic when they masquerade as faithful features in a proposed causal algorithm; we provide one such example in the CARRY-ONE task.

Still, the pervasiveness of discovering parasite features requires additional work. Beyond PARASITIC AGREEMENT and CARRY-ONE, we study two other parasitic tasks in §A. First, we show that rotation+linear can leverage the carry-one discriminative feature to steer for $\pm n$ for $n \neq 10$. Second, we devise a sentiment analysis task, where we use whether the subject is entertainment-related or food-related to steer for sentiment (rather than using the positive or negative adjectives modifying those nouns). We take these additional tasks to show that the geometries that enable discovering parasite features may not be hard to come by.

Parasite-Like Behavior in Other Frameworks. We speculate analogous issues could arise in other interpretability methods. For instance, if a linear probe for feature f happens to identify a direction that coincidentally steers an unrelated output g, steering along this direction could create spurious causal effects. Similarly, unsupervised methods like Sparse Autoencoders (SAEs; e.g. Huben et al., 2024; Bricken et al., 2023; Elhage et al., 2022; Yun et al., 2021) could learn to compress discriminative and steering directions into a single feature, leading to similar behavior where a feature combines unrelated concepts — akin to feature absorption (Chanin et al., 2025).

[8] Extant interpretability work suggests LMs can compensate for ablated components (McGrath et al., 2023; Rushing and Nanda, 2024), so a small accuracy drop alone does not fully establish the subspace is unused, merely providing strong evidence.

9 Relation to Prior Work

Makelov et al. (2024) define the closely related notion of *interpretability illusions* in subspace interventions. Similarly to parasite features, these illusions rely on two orthogonal directions: $\psi_{\text{disc.}}$ which separates the inputs but is causally disconnected (and thus in the null-space) from the outputs, and $\psi_{\text{dorm.}}$ which does not separate the inputs but steers the outputs. By patching along an intermediate direction, one can induce a desired downstream effect. Makelov et al. argue these illusions are unfaithful *because* they explicitly leverage the model's null space, however Wu et al. (2024b) counter that faithful upstream representations may lie non-orthogonally to downstream null-spaces, so such 'illusions' are not inherently unfaithful.

Unlike $\psi_{\text{disc.}}$, parasite features do not require discriminative directions (ψ^{DD}) to lie in a strict null space of the model; this direction may thus encode a real feature merely not used in the subtask at hand. Further, ψ^{DD} and ψ^{SD} need not be orthogonal. Parasite features simply require these to be two causally unrelated directions, which we argue makes their combination inherently unfaithful. Orthogonality between ψ^{DD} and ψ^{SD} is one such signal that the directions are causally unrelated, however others exist, such as the domain-specific independence between number and gender in PARASITIC AGREEMENT. In this sense, Makelov et al.'s illusions are a special case of parasite features. Theorem 1 acts as an auxiliary argument for parasite feature's spuriosity: any $(\psi^{\text{SD}}, \psi^{\text{DD}})$ pair satisfying assumptions 1–3 will pass the causal abstraction check (and vice-versa). We believe such indiscriminate acceptance of "implemented" features is characteristic of an unfaithful behavior.

Separately, Grant et al. (2025b) introduce *divergent representations* — directions inactive in a model's native computation which become active under intervention. They further distinguish *harmless divergence*, interventions harmlessly leveraging the null-space, and *pernicious divergence*, interventions activating off-manifold representations. They note interpretability illusions can be understood as a case of divergent representations due to their exploitation of $\psi_{\text{dorm.}}$, an off-distribution representation. Parasite features can be understood as a specific, theoretically-grounded instance of pernicious divergence where alignment maps discover hidden pathways by linking discriminative and steering directions. However, since ψ^{SD} may be natively active for some inputs, the degree of distributional divergence need not be as stark as in other cases of pernicious divergence — the intervened representation may land somewhere the model visits naturally, just not from these inputs.

Finally, Méloux et al. (2025b) demonstrate that Causal Abstraction is *non-identifiable*, that is, one can align multiple different maps with the same LM. Our work presents a complementary issue with Causal Abstraction: namely that a given alignment map can exploit spurious geometric relationships to give the 'illusion' of a feature. We can have a perfectly identifiable alignment between an algorithm–DNN pair, which is nevertheless parasitic. However, we do believe that parasitism can contribute to the non-identifiability issue, as it can, e.g., connect any directions which follow our Theorem 1's assumptions, thus permitting more alignments than a faithful method should.

10 Conclusion

In this paper, we introduce intuitive and formal notions of parasite features in Causal Abstraction. We provide theoretical conditions for when linear alignment maps can discover parasite features, and empirically show that a subclass of linear alignment maps discover parasite features.

Overall, our work demonstrates features found with linear alignment maps may be parasitic, linking causally unrelated features to imitate a hypothesized causal variable. In doing so, our results align with the growing body of work on the expressivity-identifiability trade-off in alignment maps. In tasks that require more expressivity, rotation maps fail to recover meaningful representations. Conversely, with more expressive maps, such as rotation+linear, we increase the possibility of finding spurious causal pathways, such as having subject gender control verb number in PARASITIC AGREEMENT. Notably, finding parasitic features does not necessarily mean the hypothesized feature does not exist—this feature could, for instance, still be represented non-linearly within the model.

How should we proceed? With growing evidence that models utilize complex internal representations that may exceed the Linear Representation Hypothesis (White et al., 2021; Pimentel et al., 2022; Csordás et al., 2024; Levy and Geva, 2025; Kantamneni and Tegmark, 2025; Engels et al., 2025; Wurgaft et al., 2026; Feucht et al., 2026; Bhalla et al., 2026), we should not hesitate to design challenging tasks and deploy more expressive alignment maps when simple methods fail. However, doing so requires careful science that rigorously checks whether apparently causal pathways are parasitic.

Acknowledgements

Denis Sutter and Antonio Di Maio acknowledge funding from the Swiss National Science Foundation (SNSF) under the project "Efficient Distributed Intelligent Applications in Mobile-Network Dynamics (eDIAMOND)" (grant number 223443). Kyle Mahowald acknowledges funding from NSF CAREER grant 2339729 and Coefficient Giving. Qing Yao acknowledges funding from the Harrington Fellowship at UT Austin. We thank Atticus Geiger for insightful discussions throughout the project.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Aryaman Arora, 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, Bangkok, Thailand. Association for Computational Linguistics.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Sander Beckers and Joseph Y. Halpern. 2019. Abstracting causal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2678–2685.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, Tal Haklay, Daniel Wurgaft, Siddharth Boppana, Matthew Kowal, Vasudev Shyam, Jack Merullo, Atticus Geiger, and Ekdeep Singh Lubana. 2026. Do sparse autoencoders capture concept manifolds? *Preprint*, arXiv:2604.28119.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430.

Sasha Boguraev, Christopher Potts, and Kyle Mahowald. 2025. Causal interventions reveal shared structure across English filler–gap constructions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25021–25042, Suzhou, China. Association for Computational Linguistics.

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

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Isaac Bloom. 2025. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Róbert Csordás, 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, Miami, Florida, US. Association for Computational Linguistics.

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

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Joshua Engels, 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*.

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

Sheridan Feucht, Tal Haklay, Usha Bhalla, Daniel Wurgaft, Can Rager, Raphaël Sarfati, Jack Merullo, Thomas McGrath, Owen Lewis, Ekdeep Singh Lubana, Thomas Fel, and Atticus Geiger. 2026. Arithmetic in the wild: Llama uses base-10 addition to reason about cyclic concepts. *Preprint*, arXiv:2605.01148.

Atticus Geiger, Jacqueline Harding, and Thomas Icard. 2025a. How causal abstraction underpins computational explanation. *Preprint*, arXiv:2508.11214.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. 2025b. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning, 1-3 April 2024, Los Angeles, California, USA*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.

Satchel Grant, Noah Goodman, and James Lloyd McClelland. 2025a. Emergent symbol-like number variables in artificial neural networks. *Transactions on Machine Learning Research*.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. 2025b. Addressing divergent representations from causal interventions on neural networks. *Preprint*, arXiv:2511.04638.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic,

Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best,

Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Michael Hanna, 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*.

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

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Subhash Kantamneni and Max Tegmark. 2025. Language models use trigonometry to do addition. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Amit Arnold Levy and Mor Geva. 2025. Language models encode numbers using digit representations in base 10. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 385–395, Albuquerque, New Mexico. Association for Computational Linguistics.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. 2024. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*.

Thomas McGrath, 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*.

Maxime Méloux, Giada Dirupo, François Portet, and Maxime Peyrard. 2025a. The dead salmons of AI interpretability. *Preprint*, arXiv:2512.18792.

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

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

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2026. The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis. *Computational Linguistics*, 52(1):331–378.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. 2025. MIB: A mechanistic interpretability benchmark. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*.

Giorgos Nikolaou, Tommaso Mencattini, Donato Crisostomi, Andrea Santilli, Yannis Panagakis, and Emanuele Rodolà. 2025. Language models are injective and hence invertible. *Preprint*, arXiv:2510.15511.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Tiago Pimentel, Josef Valvoda, Niklas Stoehr, and Ryan Cotterell. 2022. The architectural bottleneck principle. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11459–11472, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christopher Potts. 2025. Assessing skeptical views of interpretability research. Blog post.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Abhilasha Ravichander, 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, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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, New York, NY, USA. Association for Computing Machinery.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. 2017. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence*.

Cody Rushing and Neel Nanda. 2024. Explorations of self-repair in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42836–42855. PMLR.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. 2025. Open problems in mechanistic interpretability. *Preprint*, arXiv:2501.16496.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. 2025. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 132–138, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah Goodman, Christopher Manning, and Christopher Potts. 2024a. 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, Mexico City, Mexico. Association for Computational Linguistics.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. 2024b. A reply to Makelov et al. (2023)'s "interpretability illusion" arguments. *Preprint*, arXiv:2401.12631.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphaël Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. 2026. Manifold steering reveals the shared geometry of neural network representation and behavior. *Preprint*, arXiv:2605.05115.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zeyu Yun, Yubei Chen, Bruno Olshausen, and Yann LeCun. 2021. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10, Online.

A Other Parasite Features

To demonstrate that parasite features are not limited to the specific scenarios which we discuss in the main text, we design two further tasks for which rotation+linear identifies parasite features.

Variant of CARRY-ONE: $\pm n$ Steering. One of these tasks focuses on the same CARRY-ONE task as the main text, but instead of steering by ± 10, we steer by $\pm n$ for $n \in \{1, \ldots, 9\}$. If rotation+linear can align such features for all n's, we ought to be cautious about its over-expressivity in aligning CARRY-ONE. We use the exact setup as CARRY-ONE, fitting at layer 20, except changing the desired steering behavior to $\pm n$. The results are shown in Fig. 4 – every choice of n can be aligned with at least 96.6%.

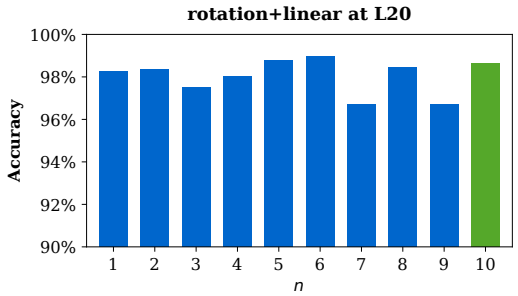


Figure 4: rotation+linear can leverage carry-one to steer for all $\pm n$, where $n = 1, \ldots, 10$. $n = 10$ results from main text.

Evidently, there's nothing special about the feature geometry of carry-one controlling ± 10 – rotation+linear is expressive enough to find $\pm n$ more generally. The $\pm n$ features are clearly parasitic, and further suggests that ± 10 is likewise parasitic.

Sentiment Analysis: Subject Topic Controls Sentiment. We also consider a sentiment analysis task, which is significantly different from existing tasks. We input the sentence "The X was Y and Z. Overall, the sentiment is " to `Llama3.1-8b` and measure the next-token probabilities of the tokens "positive" and "negative". X is an entertainment-related noun or a food-related noun, and Y and Z are either both positive adjectives or negative adjectives. The training and held-out vocabulary sets are in Table 1. The expected mechanism of the model is to use Y and Z to determine the sentiment, but we find that rotation+linear can leverage the discriminative feature of the subject being entertainment or food to steer sentiment directly.

Category	Training vocabulary	Held-out test vocabulary
X: entertainment	movie, film, show, play, musical, opera, ballet, documentary	concert, exhibition, festival, production
X: food	meal, dinner, lunch, supper, brunch, snack, dessert, feast	breakfast, banquet, buffet, appetizer
Y, Z: positive	excellent, pleasant, satisfying, wonderful, delightful, enjoyable, fantastic, superb	impressive, rewarding, amazing, splendid
Y, Z: negative	awful, disappointing, unpleasant, bad, horrible, dreadful, unsatisfying, poor	terrible, frustrating, boring, revolting

Table 1: Disjoint training and test vocabularies for the sentiment analysis task. Each sentence uses two distinct adjectives from the same sentiment category.

We train rotation+linear to associate the noun category of entertainment with positive and food with negative. For example, intervening from "The breakfast was impressive and rewarding" into "The concert was impressive and rewarding" should increase the probability of "negative", even though both sentences express positive sentiment. We learn the alignment map at the period position of the first sentence rather than the final token. We use 600 training pairs and 400 held-out test pairs drawn from the vocabulary, each balanced across the two noun categories and within- vs. across-category interventions.

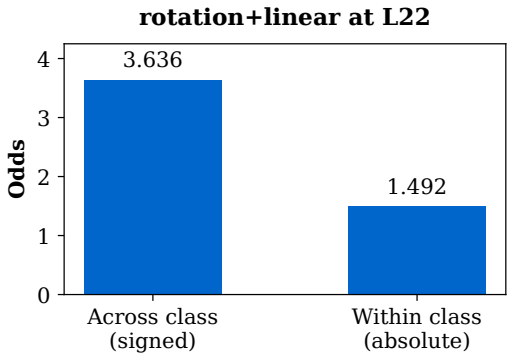


Figure 5: L22 rotation+linear **ODDS** metric on held-out sentiment analysis sentences. The alignment map successfully finds the parasite feature, leveraging subject category (entertainment/food) to steer for sentiment (positive/negative).

The model has a bias toward "positive" rather than "negative". But since we optimize for the odds metric, which is the difference between their probabilities, this bias is canceled out. As seen in Fig. 5, layer 22, the learned rotation+linear intervention achieves **ODDS** of 3.6 on held-out across-class interventions – making the target label 37 times more probable relative to the other label on average. Within-class intervention, however, is not neutral, as the intervention is optimized to still increase the probability of the shared sentiment label.

These additional examples of parasite features provide further evidence that the geometric conditions laid out in Theorem 1 are not too stringent in practice.

B Proof of Theorem 1

We restate the conditions and Theorem 1 here and provide a proof.

At a fixed layer ℓ in a DNN $\mathtt{N}$:

Assumption 1. ***(Steerability.)*** *A steering direction $\psi^{\mathtt{SD}}$ exists for g with threshold $\delta_{\mathtt{steer}} > 0$.*

Assumption 2. ***(Discrimination.)*** *A discriminative direction $\psi^{\mathtt{DD}}$ exists for f with threshold $\delta_{\mathtt{disc}} > 0$.*

Assumption 3. ***(Sufficient separation.)*** *Let (i) $\varepsilon_{\mathtt{gap}} > 0$ be the minimum gap between hidden states and the $\pm\delta_{\mathtt{steer}}$ steering threshold; (ii) $\varepsilon_{\mathtt{max}} > 0$ be the maximum activation of the steering direction by hidden states; and (iii) $\varepsilon_{\mathtt{spread}} > 0$ be the maximum within-class spread along $\psi^{\mathtt{DD}}$:*

$$\varepsilon_{\mathtt{gap}} = \min_{\mathbf{h}\in\mathcal{H}} \big|\, |\langle \mathbf{h}, \psi^{\mathtt{SD}}\rangle| - \delta_{\mathtt{steer}} \,\big|, \quad \varepsilon_{\mathtt{max}} = \max_{\mathbf{h}\in\mathcal{H}} |\langle \mathbf{h}, \psi^{\mathtt{SD}}\rangle|, \quad \varepsilon_{\mathtt{spread}} = \max_{\mathbf{h}_s,\mathbf{h}_b\in\mathcal{H}_i} |\langle \psi^{\mathtt{DD}}, \mathbf{h}_s - \mathbf{h}_b\rangle| \tag{3}$$

We require $|\langle \psi^{\mathtt{DD}}, \psi^{\mathtt{SD}}\rangle| < 1$ and a minimum gap-to-spread ratio satisfying $\frac{\varepsilon_{\mathtt{gap}}}{\varepsilon_{\mathtt{spread}}} > \frac{\varepsilon_{\mathtt{max}} + \delta_{\mathtt{steer}}}{2\,\delta_{\mathtt{disc}}}$.

Theorem 1. *Let $\mathtt{N}$ be a DNN, and $\mathtt{A}$ be an algorithm with $f_{\mathtt{A}}^{\eta_{\mathtt{int}}} = f$ and $f_{\mathtt{A}}^{\boldsymbol{\eta}_{\mathtt{y}}} = g$. Further, let $\mathcal{V}^{\mathtt{lin}}$ be the class of invertible linear maps. Under assumptions 1, 2, and 3, $\mathtt{A}$ is a $\mathcal{V}^{\mathtt{lin}}$-abstraction of $\mathtt{N}$.*

Proof. We will prove this theorem by construction. More precisely, we will prove it in 4 steps:

1. First, we will construct a linear alignment map $\phi \in \mathcal{V}^{\mathtt{lin}}$—expanded as $\phi(\mathbf{h}) = \boldsymbol{\Phi}\,\mathbf{h}$ for $\boldsymbol{\Phi} \in \mathbb{R}^{d\times d}$—such that:

$$\Phi_{\mathtt{r}} = \psi^{\mathtt{DD}} \tag{5a}$$

$$\Phi_{\mathtt{w}} = (1 - \lambda_{\mathtt{mix}}\,\rho)\psi^{\mathtt{DD}} + \lambda_{\mathtt{mix}}\,\psi^{\mathtt{SD}} \tag{5b}$$

where $\rho = \langle \psi^{\mathtt{DD}}, \psi^{\mathtt{SD}}\rangle$, $\Phi_{\mathtt{r}}$ is the first row in $\boldsymbol{\Phi}$, $\Phi_{\mathtt{w}}$ is the first column in $\boldsymbol{\Phi}^{-1}$, and we choose $\lambda_{\mathtt{mix}}$ such that $\frac{\varepsilon_{\mathtt{max}}+\delta_{\mathtt{steer}}}{2\delta_{\mathtt{disc}}} < \lambda_{\mathtt{mix}}(1-\rho^2) + \rho < \frac{\varepsilon_{\mathtt{gap}}}{\varepsilon_{\mathtt{spread}}}$.

2. Second, we show that, given a matrix constructed as above, our interventions have form:
$$\mathbf{h}^{\mathtt{s}\to\mathtt{b}} = \mathbf{h}^{\mathtt{b}} + \langle \Phi_{\mathtt{r}}, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}} \rangle \, \Phi_{\mathtt{w}} \tag{6}$$
where $\mathbf{h}^{\mathtt{s}\to\mathtt{b}}$ is typical source-to-base intervention: $\mathbf{h}^{\mathtt{s}\to\mathtt{b}} = \mathbf{\Phi}^{-1} \, [\mathbf{h}^{\mathtt{s}}_{\phi,1}, \mathbf{h}^{\mathtt{b}}_{\phi,2}, ..., \mathbf{h}^{\mathtt{b}}_{\phi,d}]^\top$ and $\mathbf{h}_\phi = \mathbf{\Phi} \, \mathbf{h}$.
3. Third, we prove that, given Assumptions 1–3, within class interventions preserve the DNN's output. Formally, this shows that for these interventions $\mathbf{y}^{\mathtt{s}\to\mathtt{b}}_{\mathtt{N}} = \mathbf{y}^{\mathtt{s}\to\mathtt{b}}_{\mathtt{A}} = g(\mathbf{x}^{\mathtt{b}}, f(\mathbf{x}^{\mathtt{s}}))$;
4. Fourth, we prove that, given Assumptions 1–3, across class interventions change the DNN's output. Formally, this shows that for these interventions $\mathbf{y}^{\mathtt{s}\to\mathtt{b}}_{\mathtt{N}} = \mathbf{y}^{\mathtt{s}\to\mathtt{b}}_{\mathtt{A}} = g(\mathbf{x}^{\mathtt{b}}, f(\mathbf{x}^{\mathtt{s}}))$.

Together, steps 3 and 4 cover all pairs of inputs $\mathbf{x}^{\mathtt{s}}, \mathbf{x}^{\mathtt{b}}$, showing thus that for every such pair $\mathbf{y}^{\mathtt{s}\to\mathtt{b}}_{\mathtt{N}} = \mathbf{y}^{\mathtt{s}\to\mathtt{b}}_{\mathtt{A}}$. The proof for each of these steps is given below, and together they prove our theorem. Notably, our proof assumes that the data is centered around the origin along $\psi^{\mathtt{DD}}$ and $\psi^{\mathtt{SD}}$; we do this without loss of generality, though, since the argument only depends on relative differences along $\psi^{\mathtt{DD}}$ and distance away from the steering threshold along $\psi^{\mathtt{SD}}$. □

TheoremProofStep 1. *For any $\lambda_{\mathtt{mix}} \in \mathbb{R}$, there exists an invertible matrix $\mathbf{\Phi} \in \mathbb{R}^{d\times d}$ such that:*
$$\Phi_{\mathtt{r}} = \psi^{\mathtt{DD}}, \qquad \Phi_{\mathtt{w}} = (1 - \lambda_{\mathtt{mix}} \, \rho)\psi^{\mathtt{DD}} + \lambda_{\mathtt{mix}} \, \psi^{\mathtt{SD}} \tag{7}$$
where $\rho = \langle \psi^{\mathtt{DD}}, \psi^{\mathtt{SD}} \rangle$, and $\Phi_{\mathtt{r}}$ is the first row in $\mathbf{\Phi}$ and $\Phi_{\mathtt{w}}$ is the first column in $\mathbf{\Phi}^{-1}$.

Proof. We prove this by giving a precise construction for $\mathbf{\Phi}$. First, let:
$$\boldsymbol{\theta} = \Phi_{\mathtt{w}} - ||\Phi_{\mathtt{w}}|| \, \mathbf{e}_1, \qquad \mathbf{\Theta} = \mathbf{I}_d - 2 \, \frac{\boldsymbol{\theta}\boldsymbol{\theta}^\top}{\boldsymbol{\theta}^\top \boldsymbol{\theta}} \tag{8}$$
where $\mathbf{\Theta}$ is known as a Householder reflector. Now, define the matrix $\mathbf{\Phi}$ as:
$$\mathbf{\Phi} = \begin{bmatrix} (\psi^{\mathtt{DD}})^\top \\ \text{2nd row of } \mathbf{\Theta} \\ \cdots \\ d\text{-th row of } \mathbf{\Theta} \end{bmatrix} \tag{9}$$
By construction, it is clear that $\Phi_{\mathtt{r}}$ is the first row in $\mathbf{\Phi}$. We now verify that $\Phi_{\mathtt{w}}$ is indeed the first column of $\mathbf{\Phi}^{-1}$, i.e., that $\mathbf{\Phi} \, \Phi_{\mathtt{w}} = \mathbf{e}_1$ (as that implies $\Phi_{\mathtt{w}} = \mathbf{\Phi}^{-1} \, \mathbf{e}_1$). Write $\nu = ||\Phi_{\mathtt{w}}||$. By construction:
$$\begin{aligned}
2 \, \langle \boldsymbol{\theta}, \Phi_{\mathtt{w}} \rangle &= 2 \left(\langle \Phi_{\mathtt{w}}, \Phi_{\mathtt{w}} \rangle - ||\Phi_{\mathtt{w}}|| \, \langle \mathbf{e}_1, \Phi_{\mathtt{w}} \rangle \right) && \text{(10a)} \\
&= 2 \, (\nu^2 - \nu \, \langle \mathbf{e}_1, \Phi_{\mathtt{w}} \rangle) && \text{(10b)} \\
&= \nu^2 - 2 \, \nu \, \langle \mathbf{e}_1, \Phi_{\mathtt{w}} \rangle + \nu^2 && \text{(10c)} \\
&= \langle \Phi_{\mathtt{w}}, \Phi_{\mathtt{w}} \rangle - 2 \, ||\Phi_{\mathtt{w}}|| \, \langle \mathbf{e}_1, \Phi_{\mathtt{w}} \rangle + ||\Phi_{\mathtt{w}}||^2 \, \langle \mathbf{e}_1, \mathbf{e}_1 \rangle && \text{(10d)} \\
&= \langle \Phi_{\mathtt{w}} - ||\Phi_{\mathtt{w}}|| \, \mathbf{e}_1, \; \Phi_{\mathtt{w}} - ||\Phi_{\mathtt{w}}|| \, \mathbf{e}_1 \rangle && \text{(10e)} \\
&= \langle \boldsymbol{\theta}, \boldsymbol{\theta} \rangle && \text{(10f)}
\end{aligned}$$
The Householder reflector $\mathbf{\Theta}$ thus sends $\Phi_{\mathtt{w}}$ onto the first axis:
$$\begin{aligned}
\mathbf{\Theta} \, \Phi_{\mathtt{w}} &= \Phi_{\mathtt{w}} - 2 \, \boldsymbol{\theta} \, \frac{\langle \boldsymbol{\theta}, \Phi_{\mathtt{w}} \rangle}{\langle \boldsymbol{\theta}, \boldsymbol{\theta} \rangle} && \text{(11a)} \\
&= \Phi_{\mathtt{w}} - \boldsymbol{\theta} && \text{(11b)} \\
&= \Phi_{\mathtt{w}} - (\Phi_{\mathtt{w}} - ||\Phi_{\mathtt{w}}|| \, \mathbf{e}_1) && \text{(11c)} \\
&= \nu \, \mathbf{e}_1 && \text{(11d)}
\end{aligned}$$
We now show that the same holds for $\mathbf{\Phi}$, applying it to $\Phi_{\mathtt{w}}$ row by row. Its first row is $(\psi^{\mathtt{DD}})^\top$, so—using $||\psi^{\mathtt{DD}}|| = 1$ and $\rho = \langle \psi^{\mathtt{DD}}, \psi^{\mathtt{SD}} \rangle$—the first entry is
$$\langle \psi^{\mathtt{DD}}, \Phi_{\mathtt{w}} \rangle = (1 - \lambda_{\mathtt{mix}} \, \rho) \, \langle \psi^{\mathtt{DD}}, \psi^{\mathtt{DD}} \rangle + \lambda_{\mathtt{mix}} \, \langle \psi^{\mathtt{DD}}, \psi^{\mathtt{SD}} \rangle = (1 - \lambda_{\mathtt{mix}} \, \rho) + \lambda_{\mathtt{mix}} \, \rho = 1. \tag{12}$$

Its remaining rows are rows $2, \dots, d$ of $\mathbf{\Theta}$, so for $k \geq 2$ the k-th entry is $\mathbf{e}_k^\top \mathbf{\Theta}\, \Phi_{\mathtt{w}} = \nu \langle \mathbf{e}_k, \mathbf{e}_1 \rangle = 0$. Hence $\mathbf{\Phi}\, \Phi_{\mathtt{w}} = \mathbf{e}_1$.

Finally, $\mathbf{\Phi}$ is invertible: rows $2, \dots, d$ of $\mathbf{\Theta}$ are orthonormal and, by the identity above, orthogonal to $\Phi_{\mathtt{w}}$, hence span $\Phi_{\mathtt{w}}^\perp$; and since $\langle \psi^{\mathtt{DD}}, \Phi_{\mathtt{w}} \rangle = 1 \neq 0$, the first row $\psi^{\mathtt{DD}} \notin \Phi_{\mathtt{w}}^\perp$, so the d rows are linearly independent. Therefore $\mathbf{\Phi}^{-1} \mathbf{e}_1 = \Phi_{\mathtt{w}}$, i.e., $\Phi_{\mathtt{w}}$ is the first column of $\mathbf{\Phi}^{-1}$, while $\Phi_{\mathtt{r}} = \psi^{\mathtt{DD}}$ is its first row by construction. □

TheoremProofStep 2. *Consider a linear alignment map $\phi \in \mathcal{V}^{\mathtt{lin}}$ where: $\phi(\mathbf{h}) = \mathbf{\Phi}\, \mathbf{h}$ for some invertible matrix $\mathbf{\Phi} \in \mathbb{R}^{d \times d}$. Further, let $\Phi_{\mathtt{r}}$ be the first row in $\mathbf{\Phi}$ and $\Phi_{\mathtt{w}}$ be the first column in $\mathbf{\Phi}^{-1}$. Given this alignment map, our interventions have form:*

$$\mathbf{h}^{\mathtt{s} \to \mathtt{b}} = \mathbf{h}^{\mathtt{b}} + \langle \Phi_{\mathtt{r}}, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}} \rangle\, \Phi_{\mathtt{w}} \tag{13}$$

Proof. $\mathbf{h}^{\mathtt{s} \to \mathtt{b}}$ is a typical source-to-base intervention. It is thus (without loss of generality) defined as:

$$\mathbf{h}^{\mathtt{s} \to \mathtt{b}} = \mathbf{\Phi}^{-1} \underbrace{[\mathbf{h}^{\mathtt{s}}_{\phi,1}, \mathbf{h}^{\mathtt{b}}_{\phi,2}, ..., \mathbf{h}^{\mathtt{b}}_{\phi,d}]^\top}_{\mathbf{h}_\phi^{\mathtt{s} \to \mathtt{b}}}, \qquad \mathbf{h}_\phi = \mathbf{\Phi}\, \mathbf{h} \tag{14}$$

where we apply our intervention on the first position of our latent variables. We can rewrite $\mathbf{h}_\phi^{\mathtt{s} \to \mathtt{b}}$ as:

$$\mathbf{h}_\phi^{\mathtt{s} \to \mathtt{b}} = \begin{bmatrix} \mathbf{e}_1^\top \mathbf{\Phi}\, \mathbf{h}^{\mathtt{s}} \\ \mathbf{e}_2^\top \mathbf{\Phi}\, \mathbf{h}^{\mathtt{b}} \\ \dots \\ \mathbf{e}_d^\top \mathbf{\Phi}\, \mathbf{h}^{\mathtt{b}} \end{bmatrix} = \mathbf{\Phi}\, \mathbf{h}^{\mathtt{b}} + \begin{bmatrix} \mathbf{e}_1^\top \mathbf{\Phi}\, \mathbf{h}^{\mathtt{s}} - \mathbf{e}_1^\top \mathbf{\Phi}\, \mathbf{h}^{\mathtt{b}} \\ 0 \\ \dots \\ 0 \end{bmatrix} \tag{15a}$$

$$= \mathbf{\Phi}\, \mathbf{h}^{\mathtt{b}} + \mathbf{e}_1^\top \mathbf{\Phi}\, (\mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}})\, \mathbf{e}_1 \tag{15b}$$

$$= \mathbf{\Phi}\, \mathbf{h}^{\mathtt{b}} + \langle \Phi_{\mathtt{r}}, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}} \rangle\, \mathbf{e}_1 \tag{15c}$$

Now, applying $\mathbf{\Phi}^{-1}$ on both sides gives us:

$$\mathbf{h}^{\mathtt{s} \to \mathtt{b}} = \mathbf{\Phi}^{-1}\, \mathbf{h}_\phi^{\mathtt{s} \to \mathtt{b}} = \mathbf{h}^{\mathtt{b}} + \langle \Phi_{\mathtt{r}}, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}} \rangle \mathbf{\Phi}^{-1} \mathbf{e}_1 \tag{16a}$$

$$= \mathbf{h}^{\mathtt{b}} + \langle \Phi_{\mathtt{r}}, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}} \rangle \Phi_{\mathtt{w}} \tag{16b}$$

which completes the proof. □

TheoremProofStep 3. *Let $\mathbf{x}^{\mathtt{b}}, \mathbf{x}^{\mathtt{s}} \in \mathcal{X}$ be two inputs sharing a class, i.e., $f(\mathbf{x}^{\mathtt{b}}) = f(\mathbf{x}^{\mathtt{s}})$. Further, let us have an alignment map ϕ such that $\Phi_{\mathtt{r}} = \psi^{\mathtt{DD}}$ and $\Phi_{\mathtt{w}} = (1 - \lambda_{\mathtt{mix}}\, \rho) \psi^{\mathtt{DD}} + \lambda_{\mathtt{mix}}\, \psi^{\mathtt{SD}}$. The DNN $\mathtt{N}$'s output is preserved under this intervention, meaning that $\mathbf{y}_{\mathtt{N}}^{\mathtt{s} \to \mathtt{b}} = \mathbf{y}_{\mathtt{A}}^{\mathtt{s} \to \mathtt{b}} = g(\mathbf{x}^{\mathtt{b}}, f(\mathbf{x}^{\mathtt{s}}))$.*

Proof. Recall that $\varepsilon_{\mathtt{gap}}$ (Assump. 3) is the minimum gap between hidden states and the $\pm\delta_{\mathtt{steer}}$ steering threshold, i.e., $\varepsilon_{\mathtt{gap}} = \min_{\mathbf{h} \in \mathcal{H}} \big| |\langle \mathbf{h}, \psi^{\mathtt{SD}} \rangle| - \delta_{\mathtt{steer}} \big|$. We must now show that $\mathbf{y}_{\mathtt{N}}^{\mathtt{s} \to \mathtt{b}} = \mathbf{y}_{\mathtt{A}}^{\mathtt{s} \to \mathtt{b}} = g(\mathbf{x}^{\mathtt{b}}, f(\mathbf{x}^{\mathtt{s}}))$. As $\mathbf{x}^{\mathtt{b}}, \mathbf{x}^{\mathtt{s}}$ share a class (i.e., $f(\mathbf{x}^{\mathtt{b}}) = f(\mathbf{x}^{\mathtt{s}})$), if we show the intervention does not make hidden states cross the $\delta_{\mathtt{steer}}$ boundary, they will preserve their class and the condition will hold. In other words, we need to show that the hidden states displacement along the steering direction is smaller than $\varepsilon_{\mathtt{gap}}$, i.e., we must prove that:

$$|\langle \mathbf{h}^{\mathtt{s} \to \mathtt{b}} - \mathbf{h}^{\mathtt{b}},\, \psi^{\mathtt{SD}} \rangle| < \varepsilon_{\mathtt{gap}} \tag{17}$$

By Step 2, this displacement is

$$|\langle \mathbf{h}^{\mathtt{s} \to \mathtt{b}} - \mathbf{h}^{\mathtt{b}},\, \psi^{\mathtt{SD}} \rangle| = |\langle \langle \Phi_{\mathtt{r}}, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}} \rangle\, \Phi_{\mathtt{w}},\, \psi^{\mathtt{SD}} \rangle| \tag{18a}$$

$$= |\langle \Phi_{\mathtt{r}}, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}} \rangle\, \langle \Phi_{\mathtt{w}},\, \psi^{\mathtt{SD}} \rangle| \tag{18b}$$

$$\leq \varepsilon_{\mathtt{spread}}\, |\langle \Phi_{\mathtt{w}},\, \psi^{\mathtt{SD}} \rangle| \qquad \text{definition of } \varepsilon_{\mathtt{spread}} \tag{18c}$$

$$= \varepsilon_{\mathtt{spread}}\, \langle (1 - \lambda_{\mathtt{mix}}\, \rho) \psi^{\mathtt{DD}} + \lambda_{\mathtt{mix}}\, \psi^{\mathtt{SD}},\, \psi^{\mathtt{SD}} \rangle \qquad \text{definition of } \Phi_{\mathtt{w}} \tag{18d}$$

$$= \varepsilon_{\mathtt{spread}}\, (\rho - \lambda_{\mathtt{mix}}\, \rho^2 + \lambda_{\mathtt{mix}}) \qquad \rho = \langle \psi^{\mathtt{DD}}, \psi^{\mathtt{SD}} \rangle \tag{18e}$$

$$= \varepsilon_{\mathtt{spread}}\, (\rho + \lambda_{\mathtt{mix}}\, (1 - \rho^2)) \tag{18f}$$

$$< \varepsilon_{\mathtt{spread}}\, (\rho + \frac{\varepsilon_{\mathtt{gap}}}{\varepsilon_{\mathtt{spread}}} - \rho) \qquad \lambda_{\mathtt{mix}}(1 - \rho^2) + \rho < \frac{\varepsilon_{\mathtt{gap}}}{\varepsilon_{\mathtt{spread}}} \tag{18g}$$

$$= \varepsilon_{\mathtt{gap}} \tag{18h}$$

which completes the proof. □

TheoremProofStep 4. *Let $\mathbf{x}^{\mathtt{b}}, \mathbf{x}^{\mathtt{s}} \in \mathcal{X}$ be two inputs with opposite classes, i.e., $f(\mathbf{x}^{\mathtt{b}}) \neq f(\mathbf{x}^{\mathtt{s}})$. Further, let us have an alignment map ϕ such that $\Phi_{\mathtt{r}} = \boldsymbol{\psi}^{\mathtt{DD}}$ and $\Phi_{\mathtt{w}} = (1 - \lambda_{\mathtt{mix}}\,\rho)\boldsymbol{\psi}^{\mathtt{DD}} + \lambda_{\mathtt{mix}}\,\boldsymbol{\psi}^{\mathtt{SD}}$. The DNN $\mathtt{N}$'s output flips to the source's class under this intervention, meaning that $\mathbf{y}_{\mathtt{N}}^{\mathtt{s}\to\mathtt{b}} = \mathbf{y}_{\mathtt{A}}^{\mathtt{s}\to\mathtt{b}} = g(\mathbf{x}^{\mathtt{b}}, f(\mathbf{x}^{\mathtt{s}}))$.*

Proof. Recall that $\varepsilon_{\mathtt{max}}$ (Assump. 3) is the maximum activation of the steering direction by hidden states, i.e., $\varepsilon_{\mathtt{max}} = \max_{\mathbf{h}\in\mathcal{H}} |\langle \mathbf{h}, \boldsymbol{\psi}^{\mathtt{SD}}\rangle|$. We must now show that $\mathbf{y}_{\mathtt{N}}^{\mathtt{s}\to\mathtt{b}} = \mathbf{y}_{\mathtt{A}}^{\mathtt{s}\to\mathtt{b}} = g(\mathbf{x}^{\mathtt{b}}, f(\mathbf{x}^{\mathtt{s}}))$. As $\mathbf{x}^{\mathtt{b}}, \mathbf{x}^{\mathtt{s}}$ have opposite classes (i.e., $f(\mathbf{x}^{\mathtt{b}}) \neq f(\mathbf{x}^{\mathtt{s}})$), if we show the intervention makes hidden states cross the $\delta_{\mathtt{steer}}$ boundary in the correct direction, they will force the output class to change correctly. In other words, we need to show that the hidden states displacement along the steering direction is greater than $\varepsilon_{\mathtt{max}} + \delta_{\mathtt{steer}}$, i.e., we must prove that:

$$\underbrace{\langle \mathbf{h}^{\mathtt{s}\to\mathtt{b}} - \mathbf{h}^{\mathtt{b}},\, \boldsymbol{\psi}^{\mathtt{SD}}\rangle > \varepsilon_{\mathtt{max}} + \delta_{\mathtt{steer}}}_{\text{if } f(\mathbf{x}^{\mathtt{b}})=0 \text{ and } f(\mathbf{x}^{\mathtt{s}})=1}, \qquad \underbrace{\langle \mathbf{h}^{\mathtt{s}\to\mathtt{b}} - \mathbf{h}^{\mathtt{b}},\, \boldsymbol{\psi}^{\mathtt{SD}}\rangle < -(\varepsilon_{\mathtt{max}} + \delta_{\mathtt{steer}})}_{\text{if } f(\mathbf{x}^{\mathtt{b}})=1 \text{ and } f(\mathbf{x}^{\mathtt{s}})=0} \tag{19}$$

Now, assume $f(\mathbf{x}^{\mathtt{b}}) = 0$ and $f(\mathbf{x}^{\mathtt{s}}) = 1$. By Step 2, this displacement is:

$$\begin{aligned}
\langle \mathbf{h}^{\mathtt{s}\to\mathtt{b}} - \mathbf{h}^{\mathtt{b}},\, \boldsymbol{\psi}^{\mathtt{SD}}\rangle &= \langle \langle \Phi_{\mathtt{r}}, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}}\rangle\, \Phi_{\mathtt{w}},\, \boldsymbol{\psi}^{\mathtt{SD}}\rangle && \text{(20a)}\\
&= \langle \Phi_{\mathtt{r}}, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}}\rangle\, \langle \Phi_{\mathtt{w}},\, \boldsymbol{\psi}^{\mathtt{SD}}\rangle && \text{(20b)}\\
&> 2\delta_{\mathtt{disc}}\, \langle \Phi_{\mathtt{w}},\, \boldsymbol{\psi}^{\mathtt{SD}}\rangle && \text{by Assump. 2 and } \langle \Phi_{\mathtt{w}},\, \boldsymbol{\psi}^{\mathtt{SD}}\rangle > 0 \quad \text{(20c)}\\
&= 2\delta_{\mathtt{disc}}\, \langle (1 - \lambda_{\mathtt{mix}}\,\rho)\boldsymbol{\psi}^{\mathtt{DD}} + \lambda_{\mathtt{mix}}\,\boldsymbol{\psi}^{\mathtt{SD}},\, \boldsymbol{\psi}^{\mathtt{SD}}\rangle && \text{definition of } \Phi_{\mathtt{w}} \quad \text{(20d)}\\
&= 2\delta_{\mathtt{disc}}\, (\rho - \lambda_{\mathtt{mix}}\,\rho^2 + \lambda_{\mathtt{mix}}) && \rho = \langle \boldsymbol{\psi}^{\mathtt{DD}}, \boldsymbol{\psi}^{\mathtt{SD}}\rangle \quad \text{(20e)}\\
&= 2\delta_{\mathtt{disc}}\, (\rho + \lambda_{\mathtt{mix}}\,(1 - \rho^2)) && \text{(20f)}\\
&> 2\delta_{\mathtt{disc}}\, (\rho + \frac{\varepsilon_{\mathtt{max}} + \delta_{\mathtt{steer}}}{2\delta_{\mathtt{disc}}} - \rho) && \lambda_{\mathtt{mix}}(1 - \rho^2) + \rho > \frac{\varepsilon_{\mathtt{max}} + \delta_{\mathtt{steer}}}{2\delta_{\mathtt{disc}}} \quad \text{(20g)}\\
&= \varepsilon_{\mathtt{max}} + \delta_{\mathtt{steer}} && \text{(20h)}
\end{aligned}$$

Notably, for $f(\mathbf{x}^{\mathtt{b}}) = 1$ and $f(\mathbf{x}^{\mathtt{s}}) = 0$, $\langle \Phi_{\mathtt{r}}, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}}\rangle < -2\delta_{\mathtt{disc}}$ instead and the inequalities change sign. This completes the proof. □

C Conditions for Rotation Alignment Maps Discovering Parasite Features

We characterize the conditions under which a rotation alignment map can discover parasite features. First, we give a general characterization of the conditions before giving a precise set of definitions.

At a fixed layer ℓ in a $\mathtt{N}$:

Assumption 4. ***(Steerability.)*** *A steering direction $\boldsymbol{\psi}^{\mathtt{SD}}$ exists for g with threshold $\delta_{\mathtt{steer}} > 0$.*

Assumption 5. ***(Discrimination.)*** *A discriminative direction $\boldsymbol{\psi}^{\mathtt{DD}}$ exists for f with threshold $\delta_{\mathtt{disc}} > 0$.*

Assumption 6. ***(Sufficient separation.)*** *Let $\varepsilon_{\mathtt{gap}}$, $\varepsilon_{\mathtt{max}}$ and $\varepsilon_{\mathtt{spread}}$ be defined as in Assump. 3. We require the discriminative gap to exceed the dynamic range of the steering projection: $\delta_{\mathtt{disc}} > \varepsilon_{\mathtt{max}} + \delta_{\mathtt{steer}}$.*

Assumption 7. ***(Within-class stability.)*** *Let $\varepsilon_{\mathtt{spread}}^{\mathtt{SD}}$ denote the within-class variation along $\boldsymbol{\psi}^{\mathtt{SD}}$:*

$$\varepsilon_{\mathtt{spread}}^{\mathtt{SD}} = \max_{i\in\{0,1\}} \max_{\mathbf{h}_b, \mathbf{h}_s \in \mathcal{H}_i} \left|\langle \mathbf{h}_b - \mathbf{h}_s, \boldsymbol{\psi}^{\mathtt{SD}}\rangle\right|. \tag{21}$$

We require $\varepsilon_{\mathtt{spread}} + \varepsilon_{\mathtt{spread}}^{\mathtt{SD}} < 2\varepsilon_{\mathtt{gap}}$.

Theorem 2. *Let $\mathtt{N}$ be a DNN, and $\mathtt{A}$ be an algorithm with $f_{\mathtt{A}}^{\eta_{\mathtt{int}}} = f$ and $f_{\mathtt{A}}^{\boldsymbol{\eta}_{\mathtt{y}}} = g$. Further, let $\mathcal{V}^{\mathtt{rot}}$ be the class of orthogonal maps. Under assumptions 4–7, $\mathtt{A}$ is a $\mathcal{V}^{\mathtt{rot}}$-abstraction of $\mathtt{N}$.*

Proof. We will prove this theorem by construction, following the same steps as the proof of Theorem 1. More precisely, we will prove it in 4 steps:

1. First, we will construct an orthogonal alignment map $\phi \in \mathcal{V}^{\texttt{rot}}$—expanded as $\phi(\mathbf{h}) = \mathbf{P}\,\mathbf{h}$ for an orthogonal $\mathbf{P} \in \mathbb{R}^{d\times d}$—whose first row is
$$\mathbf{v}_1 = \frac{\psi^{\texttt{DD}} - \psi^{\texttt{SD}}}{\|\psi^{\texttt{DD}} - \psi^{\texttt{SD}}\|}. \tag{22}$$
2. Second, we show that, given such an orthogonal matrix, our interventions have form:
$$\mathbf{h}^{\texttt{s}\to\texttt{b}} = \mathbf{h}^{\texttt{b}} + \langle \mathbf{v}_1, \mathbf{h}^{\texttt{s}} - \mathbf{h}^{\texttt{b}}\rangle\, \mathbf{v}_1, \tag{23}$$
Unlike the general case, orthogonality forces the read and write directions to coincide—both equal to $\mathbf{v}_1$.
3. Third, we prove that, given Assumptions 4–7, within class interventions preserve the DNN's output. Formally, this shows that for these interventions $\mathbf{y}_{\texttt{N}}^{\texttt{s}\to\texttt{b}} = \mathbf{y}_{\texttt{A}}^{\texttt{s}\to\texttt{b}} = g(\mathbf{x}^{\texttt{b}}, f(\mathbf{x}^{\texttt{s}}))$;
4. Fourth, we prove that, given Assumptions 4–7, across class interventions change the DNN's output. Formally, this shows that for these interventions $\mathbf{y}_{\texttt{N}}^{\texttt{s}\to\texttt{b}} = \mathbf{y}_{\texttt{A}}^{\texttt{s}\to\texttt{b}} = g(\mathbf{x}^{\texttt{b}}, f(\mathbf{x}^{\texttt{s}}))$.

Together, steps 3 and 4 cover all pairs of inputs $\mathbf{x}^{\texttt{s}}, \mathbf{x}^{\texttt{b}}$, showing that for every such pair $\mathbf{y}_{\texttt{N}}^{\texttt{s}\to\texttt{b}} = \mathbf{y}_{\texttt{A}}^{\texttt{s}\to\texttt{b}}$. The proof for each of these steps is given below, and together they prove our theorem. As in Theorem 1, our proof assumes that the data is centered around the origin along $\psi^{\texttt{DD}}$ and $\psi^{\texttt{SD}}$. We do this without loss of generality, since the argument only depends on relative differences along $\psi^{\texttt{DD}}$ and distance away from the steering threshold along $\psi^{\texttt{SD}}$. □

TheoremProofStep 1. *There exists an orthogonal matrix $\mathbf{P} \in \mathbb{R}^{d\times d}$ whose first row is $\mathbf{v}_1 = (\psi^{\texttt{DD}} - \psi^{\texttt{SD}})/\|\psi^{\texttt{DD}} - \psi^{\texttt{SD}}\|$.*

Proof. Assump. 6 implies $|\langle \psi^{\texttt{DD}}, \psi^{\texttt{SD}}\rangle| < 1$: if $\psi^{\texttt{DD}} = \pm\psi^{\texttt{SD}}$, then every $\mathbf{h} \in \mathcal{H}$ would have $|\langle \mathbf{h}, \psi^{\texttt{SD}}\rangle| = |\langle \mathbf{h}, \psi^{\texttt{DD}}\rangle| > \delta_{\texttt{disc}}$, giving $\varepsilon_{\texttt{max}} > \delta_{\texttt{disc}}$, a contradiction. Hence $\psi^{\texttt{DD}}$ and $\psi^{\texttt{SD}}$ are not collinear, $\psi^{\texttt{DD}} - \psi^{\texttt{SD}} \neq \mathbf{0}$, and $\mathbf{v}_1$ is well defined. We can extend to an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_d\}$ of $\mathbb{R}^d$, and define $\mathbf{P}$ as the change-of-basis matrix with rows $\mathbf{v}_1^\top, \mathbf{v}_2^\top, \ldots, \mathbf{v}_d^\top$. As its rows form an orthonormal basis, $\mathbf{P}$ is orthogonal, with its first row being $\mathbf{v}_1^\top$ by construction. □

TheoremProofStep 2. *Consider an orthogonal alignment map $\phi \in \mathcal{V}^{\texttt{rot}}$ where $\phi(\mathbf{h}) = \mathbf{P}\,\mathbf{h}$ for an orthogonal matrix $\mathbf{P} \in \mathbb{R}^{d\times d}$ with first row $\mathbf{v}_1^\top$. Given this alignment map, our interventions have form:*
$$\mathbf{h}^{\texttt{s}\to\texttt{b}} = \mathbf{h}^{\texttt{b}} + \langle \mathbf{v}_1, \mathbf{h}^{\texttt{s}} - \mathbf{h}^{\texttt{b}}\rangle\, \mathbf{v}_1. \tag{24}$$
Moreover, writing $d_b = \langle \mathbf{h}^{\texttt{b}}, \psi^{\texttt{DD}}\rangle$, $d_s = \langle \mathbf{h}^{\texttt{s}}, \psi^{\texttt{DD}}\rangle$, $\texttt{s}_b = \langle \mathbf{h}^{\texttt{b}}, \psi^{\texttt{SD}}\rangle$, and $\texttt{s}_s = \langle \mathbf{h}^{\texttt{s}}, \psi^{\texttt{SD}}\rangle$, the induced displacement along the steering direction is
$$\langle \mathbf{h}^{\texttt{s}\to\texttt{b}} - \mathbf{h}^{\texttt{b}},\, \psi^{\texttt{SD}}\rangle = \tfrac{1}{2}(d_b - d_s) + \tfrac{1}{2}(\texttt{s}_s - \texttt{s}_b). \tag{25}$$

Proof. As $\mathbf{P}$ is orthogonal it is invertible, so the intervention factorization from Step 2 of the proof of Theorem 1 applies: $\mathbf{h}^{\texttt{s}\to\texttt{b}} = \mathbf{h}^{\texttt{b}} + \langle \mathbf{r}, \mathbf{h}^{\texttt{s}} - \mathbf{h}^{\texttt{b}}\rangle\, \mathbf{w}$, where $\mathbf{r}$ is the first row of $\mathbf{P}$ and $\mathbf{w}$ is the first column of $\mathbf{P}^{-1}$. For an orthogonal map, $\mathbf{P}^{-1} = \mathbf{P}^\top$, so the first column of $\mathbf{P}^{-1}$ equals the first row of $\mathbf{P}$: $\mathbf{w} = \mathbf{P}^\top \mathbf{e}_1 = \mathbf{v}_1 = \mathbf{r}$. The read and write directions thus coincide, giving $\mathbf{h}^{\texttt{s}\to\texttt{b}} = \mathbf{h}^{\texttt{b}} + \langle \mathbf{v}_1, \mathbf{h}^{\texttt{s}} - \mathbf{h}^{\texttt{b}}\rangle\, \mathbf{v}_1$.

For the displacement along the steering direction, we take the inner product with $\psi^{\texttt{SD}}$:
$$\langle \mathbf{h}^{\texttt{s}\to\texttt{b}} - \mathbf{h}^{\texttt{b}},\, \psi^{\texttt{SD}}\rangle = \langle \mathbf{v}_1, \mathbf{h}^{\texttt{s}} - \mathbf{h}^{\texttt{b}}\rangle\, \langle \mathbf{v}_1, \psi^{\texttt{SD}}\rangle. \tag{26}$$
Using $\|\psi^{\texttt{DD}} - \psi^{\texttt{SD}}\|^2 = \langle \psi^{\texttt{DD}}, \psi^{\texttt{DD}}\rangle - 2\,\langle \psi^{\texttt{DD}}, \psi^{\texttt{SD}}\rangle + \langle \psi^{\texttt{SD}}, \psi^{\texttt{SD}}\rangle = 2(1 - \langle \psi^{\texttt{DD}}, \psi^{\texttt{SD}}\rangle)$, we have:
$$\langle \mathbf{v}_1, \psi^{\texttt{SD}}\rangle = \left\langle \frac{\psi^{\texttt{DD}} - \psi^{\texttt{SD}}}{\|\psi^{\texttt{DD}} - \psi^{\texttt{SD}}\|}, \psi^{\texttt{SD}} \right\rangle \qquad \text{definition of } \mathbf{v}_1 \tag{27a}$$
$$= \frac{\langle \psi^{\texttt{DD}}, \psi^{\texttt{SD}}\rangle - 1}{\|\psi^{\texttt{DD}} - \psi^{\texttt{SD}}\|} \tag{27b}$$
$$= -\frac{1}{2}\|\psi^{\texttt{DD}} - \psi^{\texttt{SD}}\| \qquad \langle \psi^{\texttt{DD}}, \psi^{\texttt{SD}}\rangle - 1 = -\frac{1}{2}\|\psi^{\texttt{DD}} - \psi^{\texttt{SD}}\|^2 \tag{27c}$$

Similarly, we have:

$$\langle \mathbf{v}_1, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}} \rangle = \left\langle \frac{\boldsymbol{\psi}^{\mathtt{DD}} - \boldsymbol{\psi}^{\mathtt{SD}}}{\|\boldsymbol{\psi}^{\mathtt{DD}} - \boldsymbol{\psi}^{\mathtt{SD}}\|}, \mathbf{h}^{\mathtt{s}} - \mathbf{h}^{\mathtt{b}} \right\rangle \qquad \text{definition of } \mathbf{v}_1 \tag{28a}$$

$$= \frac{(d_s - \mathtt{s}_s) - (d_b - \mathtt{s}_b)}{\|\boldsymbol{\psi}^{\mathtt{DD}} - \boldsymbol{\psi}^{\mathtt{SD}}\|} \qquad \text{definitions} \tag{28b}$$

Replacing both and multiplying them in eq. (26) leads to eq. (25). □

TheoremProofStep 3. *Let $\mathbf{x}^{\mathtt{b}}, \mathbf{x}^{\mathtt{s}} \in \mathcal{X}$ be two inputs sharing a class, i.e., $f(\mathbf{x}^{\mathtt{b}}) = f(\mathbf{x}^{\mathtt{s}})$. Further, let ϕ be any orthogonal alignment map with first row $\mathbf{v}_1^\top$ (which exists by Step 1). The DNN $\mathtt{N}$'s output is preserved under this intervention, meaning that $\mathbf{y}_{\mathtt{N}}^{\mathtt{s}\to\mathtt{b}} = \mathbf{y}_{\mathtt{A}}^{\mathtt{s}\to\mathtt{b}} = g(\mathbf{x}^{\mathtt{b}}, f(\mathbf{x}^{\mathtt{s}}))$.*

Proof. Recall that $\varepsilon_{\mathtt{gap}}$ (Assump. 6) is the minimum gap between hidden states and the $\pm\delta_{\mathtt{steer}}$ steering threshold. As $\mathbf{x}^{\mathtt{b}}, \mathbf{x}^{\mathtt{s}}$ share a class, if we show the intervention does not make hidden states cross the $\delta_{\mathtt{steer}}$ boundary, they will preserve their class and the condition will hold. In other words, we need to show that the displacement along the steering direction is smaller than $\varepsilon_{\mathtt{gap}}$:

$$|\langle \mathbf{h}^{\mathtt{s}\to\mathtt{b}} - \mathbf{h}^{\mathtt{b}}, \boldsymbol{\psi}^{\mathtt{SD}} \rangle| < \varepsilon_{\mathtt{gap}}. \tag{29}$$

By eq. (25) and the triangle inequality,

$$|\langle \mathbf{h}^{\mathtt{s}\to\mathtt{b}} - \mathbf{h}^{\mathtt{b}}, \boldsymbol{\psi}^{\mathtt{SD}} \rangle| = \left|\tfrac{1}{2}(d_b - d_s) + \tfrac{1}{2}(\mathtt{s}_s - \mathtt{s}_b)\right| \tag{30a}$$

$$\leq \tfrac{1}{2}|d_s - d_b| + \tfrac{1}{2}|\mathtt{s}_s - \mathtt{s}_b| \tag{30b}$$

$$\leq \tfrac{1}{2}\varepsilon_{\mathtt{spread}} + \tfrac{1}{2}\varepsilon^{\mathtt{SD}}_{\mathtt{spread}} \qquad \text{definitions of } \varepsilon_{\mathtt{spread}} \text{ and } \varepsilon^{\mathtt{SD}}_{\mathtt{spread}} \tag{30c}$$

$$< \varepsilon_{\mathtt{gap}} \qquad \varepsilon_{\mathtt{spread}} + \varepsilon^{\mathtt{SD}}_{\mathtt{spread}} < 2\varepsilon_{\mathtt{gap}} \tag{30d}$$

which completes this step. □

TheoremProofStep 4. *Let $\mathbf{x}^{\mathtt{b}}, \mathbf{x}^{\mathtt{s}} \in \mathcal{X}$ be two inputs with opposite classes, i.e., $f(\mathbf{x}^{\mathtt{b}}) \neq f(\mathbf{x}^{\mathtt{s}})$. Further, let ϕ be any orthogonal alignment map with first row $\mathbf{v}_1^\top$ (which exists by Step 1). The DNN $\mathtt{N}$'s output flips to the source's class under this intervention, meaning that $\mathbf{y}_{\mathtt{N}}^{\mathtt{s}\to\mathtt{b}} = \mathbf{y}_{\mathtt{A}}^{\mathtt{s}\to\mathtt{b}} = g(\mathbf{x}^{\mathtt{b}}, f(\mathbf{x}^{\mathtt{s}}))$.*

Proof. Recall that $\varepsilon_{\mathtt{max}}$ (Assump. 6) is the maximum activation of the steering direction by hidden states. As $\mathbf{x}^{\mathtt{b}}, \mathbf{x}^{\mathtt{s}}$ have opposite classes, if we show the intervention moves the hidden state across the $\delta_{\mathtt{steer}}$ boundary in the correct direction, they will force the output class to change correctly. Unlike the general case, an orthogonal map cannot amplify the displacement, so instead we bound the intervened activation $\langle \mathbf{h}^{\mathtt{s}\to\mathtt{b}}, \boldsymbol{\psi}^{\mathtt{SD}} \rangle$ directly, using the absolute separation $\delta_{\mathtt{disc}} > \varepsilon_{\mathtt{max}} + \delta_{\mathtt{steer}}$ of Assump. 6.

Assume $f(\mathbf{x}^{\mathtt{b}}) = 0$ and $f(\mathbf{x}^{\mathtt{s}}) = 1$, so $\mathbf{h}^{\mathtt{b}} \in \mathcal{H}_0$ and $\mathbf{h}^{\mathtt{s}} \in \mathcal{H}_1$. By Assump. 5, $d_b < -\delta_{\mathtt{disc}}$ and $d_s > \delta_{\mathtt{disc}}$, so $d_b - d_s < -2\delta_{\mathtt{disc}}$. Adding $\mathtt{s}_b = \langle \mathbf{h}^{\mathtt{b}}, \boldsymbol{\psi}^{\mathtt{SD}} \rangle$ to eq. (25) and using $\mathtt{s}_b, \mathtt{s}_s \leq \varepsilon_{\mathtt{max}}$:

$$\langle \mathbf{h}^{\mathtt{s}\to\mathtt{b}}, \boldsymbol{\psi}^{\mathtt{SD}} \rangle = \tfrac{1}{2}(d_b - d_s) + \tfrac{1}{2}(\mathtt{s}_b + \mathtt{s}_s) \tag{31a}$$

$$< -\delta_{\mathtt{disc}} + \varepsilon_{\mathtt{max}} \qquad d_b - d_s < -2\delta_{\mathtt{disc}} \text{ and } \mathtt{s}_b, \mathtt{s}_s \leq \varepsilon_{\mathtt{max}} \tag{31b}$$

$$< -\delta_{\mathtt{steer}} \qquad \delta_{\mathtt{disc}} > \varepsilon_{\mathtt{max}} + \delta_{\mathtt{steer}} \tag{31c}$$

Thus $\mathbf{h}^{\mathtt{s}\to\mathtt{b}}$ is moved past the $-\delta_{\mathtt{steer}}$ boundary, flipping the output regardless of where $\mathbf{h}^{\mathtt{b}}$ originally resided along $\boldsymbol{\psi}^{\mathtt{SD}}$. For $f(\mathbf{x}^{\mathtt{b}}) = 1$ and $f(\mathbf{x}^{\mathtt{s}}) = 0$, the same argument with signs reversed gives $\langle \mathbf{h}^{\mathtt{s}\to\mathtt{b}}, \boldsymbol{\psi}^{\mathtt{SD}} \rangle > \delta_{\mathtt{steer}}$. This completes the proof. □

D Training Details

We use the `pyvene` (Wu et al., 2024a) package to implement and train alignment maps for both PARASITIC AGREEMENT and CARRY-ONE.

We parameterize our alignment maps of class $\mathcal{V}^{\mathtt{r+l}}$ as compositions of a $\mathcal{V}^{\mathtt{rot}}$ map with a $\mathcal{V}^{\mathtt{lin}}$ map. In practice, the linear part of our map is parameterized as $UDU^\top$, where $U \in \mathbb{R}^{i\times i}$ is orthogonal

and D is diagonal with strictly positive entries. This convenient choice of alignment map, which is strictly contained in the broader $\mathcal{V}^{\mathtt{r}+1}$, turns out to be sufficient for discovering parasite features.

Our alignment maps are randomly initialized and optimized through backpropagation using a standard cross-entropy loss between the 'correct' label under intervention and the actual output of the model under intervention. The model's weights are frozen during this process.

In the case of PARASITIC AGREEMENT, we train DAS on 400 unique, balanced, minimal pairs, and evaluate on a held-out test set of 50 minimal pairs (as per Arora et al., 2024). We do so with 5 epochs with a batch size of 4 with a learning rate of 5e–3. Different from Arora et al. (2024), our training set includes pairs which do not differ in input feature, with the intended intervention behavior as no change in **ODDS**.

For CARRY-ONE, we train DAS on 34,400 unique minimal pairs for 20 epochs with a batch size of 64 with a learning rate of 1e–4, and evaluate the accuracy of our interventions on a held-out test set of 1000 items (as per Mueller et al., 2025).

All training is done on A40 GPUs. For PARASITIC AGREEMENT, training time for 5 epochs on 1 layer of `pythia-1b` (Biderman et al., 2023) is 2 minutes on an A40. For CARRY-ONE, training time for 20 epochs on 1 layer of `Llama3.1-8b` (Grattafiori et al., 2024) is 95 minutes.

E Parasite Features in Other Models

To determine whether $\mathcal{V}^{\mathtt{r}+1}$'s ability to find parasite features is model-specific, we also compare its performance on PARASITIC AGREEMENT against $\mathcal{V}^{\mathtt{rot}}$ in two other models. Specifically, we perform the same experiments as in §6, but with two models of different families and different sizes: `GPT2-Medium` (Radford et al., 2019) and `Qwen2-1.5b` (Yang et al., 2024).

Our results can be seen in Table 2. Notably, we see that, in line with our previous results, maps of the class rotation+linear can find parasite features for PARASITIC AGREEMENT more so than those of rotation in `Qwen2-1.5b`. That is, both the maximal **ODDS** across layers is higher for rotation+linear than for rotation, and rotation+linear shows higher causal efficacy across all layers. Interestingly, we find that each intervention is roughly equal in `GPT2-Medium`. This could be due to many causes, such as rotations finding parasites in this model or parasites being a phenomenon that emerges with scale. However, we do not attempt to adjudicate on such questions here, instead leaving it up to future work.

	GPT2-medium		*Qwen2-1.5b*	
Layer	$\mathcal{V}^{\texttt{rot}}$	$\mathcal{V}^{\texttt{r+l}}$	$\mathcal{V}^{\texttt{rot}}$	$\mathcal{V}^{\texttt{r+l}}$
0	2.43	2.38	0.11	0.24
1	**2.53**	**2.46**	1.59	2.03
2	2.46	2.33	1.59	1.99
3	2.35	2.25	1.85	2.04
4	1.89	1.80	**2.03**	2.05
5	1.64	1.47	1.81	2.22
6	1.67	1.30	0.98	1.89
7	1.72	1.37	1.53	1.98
8	1.65	1.28	1.29	1.82
9	1.71	1.39	1.64	2.07
10	1.69	1.31	1.45	2.21
11	1.56	1.43	1.51	**2.83**
12	2.03	2.06	1.53	2.55
13	2.13	0.92	1.85	2.43
14	2.31	1.08	1.80	2.55
15	1.96	0.94	1.40	2.23
16	1.67	0.77	1.67	2.06
17	1.44	0.70	1.69	2.19
18	0.76	0.29	1.92	2.39
19	0.52	0.50	1.93	2.26
20	0.18	0.21	1.66	2.06
21	0.18	0.16	1.74	2.09
22	0.10	0.11	1.52	1.92
23	0.00	0.00	1.30	1.70
24	–	–	0.15	0.37
25	–	–	0.16	0.13
26	–	–	0.04	0.09
27	–	–	0.00	0.00

Table 2: PARASITIC AGREEMENT with 1D interventions using rotation and rotation+linear maps. Bold marks the best layer for each map. GPT2-medium has 24 layers.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction introduce the notion of parasite features (our main contribution). We then explain appropriate background for these before providing theoretical justification for their existence (including a proof), and empirical experiments in the following sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In §B and §C we provide a long discussion on the limitations of the theoretical framework in which we work, and specifically in the assumptions we use in constructing our proof.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We have two main proofs in this work. For the main one, Theorem 1 in the main text, we provide assumptions and a sketch of the proof in the main text, before providing a larger discussion of conditions and the full proof in the appendix. For the auxiliary proof, Theorem 2, we provide a full proof and conditions in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide a clear methodology and rationale for our experiments, including models, training conditions and evaluation conditions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not yet provide our code or experimental data, but we plan to release all relevant artifacts before publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all such information distributed across §6 and §D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: For our experiments, due to computational requirements we only ran one seed per configuration. This makes calculating error bars and significance not possible. However, the consistent trends in these results, however, still enable us to have confidence in their robustness.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide such information in §D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the code of ethics and confirm no such violations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: We do not see broader societal impacts for our work. We point out additional considerations for using already released, known methods.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We discuss the models used in §6 and §E and discuss libraries we use to implement our research in §D.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not do crowdsourcing not research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We do not run human experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We do not utilize LLMs for any important, original, or non-standard component of the core methods in this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.