A Survey on the Linear Representation *Hypothesis*

Anonymous ACL submission

Abstract

The term "the linear representation hypothesis" has appeared across diverse subfields of artificial intelligence, neuroscience, and cognitive science. Yet previous works have not consistently treated this so-called hypothesis as a falsifiable scientific hypothesis; we thus analyze the inconsistencies of the usage of this term and point out what this means for existing literature. Based on this analysis, we argue that claims about linear representations are only well-defined for a specific model, representation location, feature definition, and evaluation dataset. We therefore propose a more rigorous formalization of the LRH that makes these dependencies explicit and allows the hypothesis to be evaluated as a falsifiable scientific claim. Finally, we identify the non-trivial open problems that warrant further attention from the research community.

1 Introduction

Neural activations enable a neural system to do remarkably complex things. Yet it is not obvious how such capabilities arise, which is why their underlying mechanisms have long been a focus of inquiry. One natural way to understand such complexity is to look for a linear structure. In an artificial neural network, the activation corresponds to a d-dimensional hidden embedding vector $\text{emb}(x) \in \mathbb{R}^d$ for an input x. The discovery of linear regularities in these embeddings, such as $\text{emb}(\text{"king"}) - \text{emb}(\text{"queen"}) \approx \text{emb}(\text{"man"}) - \text{emb}(\text{"woman"})$, in recurrent neural network language models (Mikolov et al., 2013) later influenced the concept of linear representations. If such a relation holds, it suggests the existence of a linear direction in the embedding space, such that:

$$\text{emb}(\text{"queen"}) \approx \text{emb}(\text{"king"}) + v_{\text{male}\to\text{female}}, \tag{1}$$

where $v_{\text{male}\to\text{female}} \in \mathbb{R}^d$ is a gender transition vector (or, more formally, a *counterfactual intervention vector*), providing a more interpretable picture of previously human-incomprehensible hidden embedding vectors.

Year	New papers	Cumulative
1976	1	1
2001	1	2
2007	1	3
2022	2	5
2023	11	16
2024	149	165
2025	447	612

Table 1: Number of papers explicitly employing the term "linear representation hypothesis" over time, based on Google Scholar search results (counts as of December 2025). The early history is discussed in Section 2.1. The sharp increase after 2022 was triggered by Elhage et al. (2022).

There has also been considerable debate over whether the relations generalize. For example, GloVe word embeddings (Pennington et al., 2014) were reported to achieve 75% accuracy on analogy tests of the form "*a* is to *b* as *c* is to _?" where the answer is predicted by finding the embedding with the highest cosine similarity to $\text{emb}(\text{b}) - \text{emb}(\text{a}) + \text{emb}(\text{c})$. However, Gladkova et al. (2016) introduced the bigger analogy test set (BATS) and reported that GloVe's accuracy on this dataset drops below 30%. Despite these debates, it is uncontroversial that whether linear representations exist in machine-learned embeddings depends on the choice of embedding model, dataset, and the linguistic features under investigation.

More recently, however, there has been a noticeable shift in terminology: as shown in Table 1, the term "linear representation hypothesis (LRH)" has come into widespread use. This naturally leads to a deeper question that has received surprisingly little discussion: *is "linear representation" truly something that merits being called a hypothesis?*

And if so, what would it mean for this field if this hypothesis is ever falsified? Would large parts of the research that rely on it need to be reconsidered?

We first categorize the different ways in which the LRH is used in the literature. Some works treat it as a theoretical premise about the structure of neural representations, from which other results are derived. In others, it is examined empirically, with studies presenting supporting evidence, counterexamples, or analyses of when linear representations appear to hold. Still other work relies on the LRH more implicitly, as a methodological assumption that motivates techniques such as representation steering and related interventions. Our examination leads us to consider whether the LRH, as commonly formulated, satisfies the criteria of a falsifiable scientific hypothesis. Our analysis suggests that claims about linear representations are only well-defined relative to a specific model, representation location, feature definition, and evaluation dataset; we therefore propose a formalization of the LRH that makes these requirements explicit and places prior empirical findings within the scope of a well-defined hypothesis.

2 Formulation of the LRH

2.1 Historical Origins of the LRH

If we exclude cases that merely mention "linear representation" and instead trace the origin of the exact term "linear representation *hypothesis*," the story can be summarized as follows.

The earliest use of this term appears in Pitz et al. (1976) in the field of cognitive science. There, it was introduced as a hypothesis in behavioral experiments on humans: the numbers that participants saw as samples were hypothesized to be encoded as points on an imaginary real number line (linearly) in their minds. Later, Zentall (2001) used the term in animal cognition experiments, framing it as the hypothesis that animals, such as pigeons, can construct different stimuli (A, B, C) in their brain as a linear ordering structure, such as (A > B > C).

In machine learning, the term appears only later, first in González et al. (2007), and this is where it becomes closer to the modern sense. They introduced the term "non-linear representation hypothesis" to refer to cases in which a classification problem is not separable by a linear decision boundary in the input space (e.g., with a linear support vector machine). After that, the kind of hype shown in Table 1 began when Elhage et al. (2022) pointed out that the following claim is non-trivial and explicitly named it the linear representation hypothesis (LRH):

Definition 1 (LRH; Elhage et al. (2022)'s)**.** *Neural networks represent input features*[1] *as directions in activation space.*

2.2 Early Empirical Findings

Building on Elhage et al. (2022)'s definition, Nanda et al. (2023) provides early empirical evidence for the LRH by showing that the board state of an OthelloGPT model trained on Othello move sequences can be recovered using supervised linear probes. Unsupervised methodological progress has also been made, beginning with the findings by Huben et al. (2024) that sparse autoencoders (SAEs) with a single-layer (linear) encoder can find interpretable features, such as words beginning with "w", from large language models like Pythia-70M (Biderman et al., 2023). Notably, Templeton et al. (2024) identifies the well-known "Golden Gate Bridge" feature in Claude 3 Sonnet (Anthropic, 2024) and shows that clamping it causally steers the model's outputs toward that concept.

2.3 The Strong, the Weak and the Causal

The Strong LRH Later, Smith (2024) points out that the prior definition (Definition 1) is vague, and shows that it can be understood in two distinct ways:

Definition 2 (LRH; Smith (2024)'s)**.** *In a given representation space,*

*(i) (**Strong LRH**) **All** features used by neural networks are represented by linear directions.*

*(ii) (**Weak LRH**) **Some** features used by neural networks are represented as linear directions.*

This classification is meaningful in that it makes at least the strong LRH more easily falsifiable. Under such constraints, several recent studies provide concrete counterexamples to the strong LRH by identifying features whose representations are inherently non-linear. Csordás et al. (2024) demonstrate that GRU models encode certain concepts, such as token position, using concentric representations. Similarly, Engels et al. (2024) argue that

[1] While defining what constitutes a *feature* is also a non-trivial problem (Olah, 2024), a detailed discussion is beyond the scope of this paper. We therefore refer the reader to Table 2 for examples of features considered in this study.

Linear Features	Model	Layer	LRH	Source
Board state of Othello	OthelloGPT	∘ • ∘	LRH holds	Nanda et al. (2023)
Negative-to-positive sentiment shift, detoxification	GPT2-XL, GPT-J, OPT, LLaMA3 8B	∘ • ∘	LRH holds	Turner et al. (2023)
The Golden Gate Bridge, implementing addition functions in code, security vulnerability, gender bias, sycophancy, deception, harmful behavior	Claude 3 Sonnet	○ • ○	LRH holds	Templeton et al. (2024)
Movie review sentiment	GPT2-small, Pythia 1.4B, Pythia 2.8B	∘ • ∘	LRH holds	Tigges et al. (2024)
Multilingualism, morphology, gender, hypernym/hyponym, holonym/meronym	LLaMA2 7B	○ ○ •	LRH holds	Park et al. (2024)
Multilingualism, morphology, gender, holonym/meronym, capitalization, plurality	LLaMA2 7B	○ ○ •	LRH holds	Jiang et al. (2024)
Latitude/longitude of places, release date of books	LLaMA2, Pythia	∘ • ∘	LRH holds	Gurnee and Tegmark (2024)
Truthfulness, honesty	LLaMA2 7B, LLaMA2 13B, LLaMA2 70B	• • •	LRH holds	Nguyen and Leng (2025)
Hypernym/hyponym	LLaMA3 8B	○ ○ •	LRH holds	Park et al. (2025)
Political perspective (liberal/conservative)	LLaMA2 7b, Mistral 7b, Vicuna 7b	○ • ○	LRH holds	Kim et al. (2025)
Hallucination, persuasion, pessimism, refusal, sycophancy, truthfulness	Qwen 2.5 7B and LLaMA2 13B	∘ • ∘	LRH holds	Agarwal et al. (2025)
Relations with high co-occurrence frequency (e.g., country-capital)	GPT-J, OLMO 7B	○ • •	LRH holds	Merullo et al. (2025)

Non-linear Features	Model	Layer	LRH	Source
Token position	GRU (RNN)	N/A	LRH does not hold (Onion/concentric representation)	Csordás et al. (2024)
Days of the week, months of the year	GPT2-small, Mistral 7B	○ • ○	LRH does not hold (Circular representation)	Engels et al. (2024); Modell et al. (2025)
Numbers used in two digit addition	GPT-J, Pythia 6.9B, LLaMA3 8B	○ ○ •	LRH does not hold (Helix/spiral representation)	Kantamneni and Tegmark (2025)
Periodic table	LLaMA3 70B	• • •	LRH does not hold (Helix/spiral representation)	Lei and Cooper (2025a)
Numbers in MNIST, "an" token in language	DINOv2, GPT2-Large	○ ○ •	LRH does not hold (Centroid affinity representation)	Walker et al. (2025)
Relations with low co-occurrence frequency (e.g., company-CEO)	GPT-J, OLMO 7B	○ • •	LRH does not hold (Representation not found)	Merullo et al. (2025)

Table 2: **Summary of studies providing empirical evidence for, or counterexamples to, the linear representation hypothesis (LRH).** In the **Layer** column, • ○ ○ denotes the initial (non-contextualized) embedding layer, ○ • ○ intermediate layers, and ○ ○ • the final layer before unembedding. Gray circles (∘) indicate that the corresponding layer was also examined but exhibited weaker linear representations. The **LRH** column indicates whether the LRH holds in the study; when it does not hold, the observed non-linear representation is specified.

transformer-based language models represent periodic concepts, such as days of the week, using circular representation. Together, these results suggest that *not* all features used by neural networks are linearly represented.

The Weak LRH In Definition 2, the weak LRH asserts the existence of some features that are represented as linear directions. As summarized in Table 2, several studies report empirical evidence for linear representations of particular features. However, the existence of supporting examples does not entail that the weak LRH is generally true. This leads researchers to habitually refer to it as a hypothesis, while leaving unclear how long we should continue to call it a hypothesis. More concretely, the difficulty of generalizing evidence for the weak LRH can be understood along two main dimensions.

First, empirical evidence supporting linear representation for one feature is insufficient to justify the claim that another feature is also linearly represented. Moreover, because each study defines a feature with respect to a particular dataset or input, features that appear identical on the surface may in fact differ. For example, for the linguistic feature of gender, Park et al. (2024) examine whether masculine nouns such as "king" can be transformed into feminine counterparts like "queen", whereas Templeton et al. (2024) focus on gender bias using sentence-level contexts such as "I asked the nurse a question, and ...". Under this approach, researchers are forced to carry out an endless series of validations for the infinitely many fine-grained features that may exist.

Second, a more serious challenge is that the class of embedding spaces under investigation is not fixed: they change as new LLMs are developed. This makes hypothesis formulation in this field particularly challenging. For instance, in physics, while researchers gather evidence in support of a hypothesis, the universe has not suddenly changed into a different one overnight. By contrast, an LLM under study may eventually become outdated, and researchers will then encounter the embedding spaces of new LLMs. In other words, experiments that support the LRH for one model do not guarantee generalization to another model trained with different data or architectures.

In sum, although the weak LRH may sound weak by name, it is in fact the most difficult to handle; unlike the strong LRH, it is not easily falsified. Because its loose formulation can cover nearly all LRH-related work, references to the LRH in practice typically fall into this category, and we follow this categorization in this paper unless otherwise specified. At the same time, because of its broad applicability, the merit of calling it a hypothesis is questionable; since this issue lies at the core of this paper, Section 5 will focus on this question in greater depth.

The Causal LRH There is also another line of work investigating linear representations of causality (Park et al., 2024; Jiang et al., 2024; Nguyen and Leng, 2025). Prior work has often cited Park et al. (2024) as a seminal contribution. Their study starts from the premise that causality can be defined by constructing counterfactual pairs between two opposing factual states, as illustrated in Eq. (1). Given such pairs, they characterize the geometry of linear representations through the following three questions:

Definition 3 (LRH; Park et al. (2024)'s)**.** *For all* ***counterfactual pairs*** *of a concept,*

(i) ***(Subspace LRH)*** *All pairs belong to a shared subspace.*

(ii) ***(Measurement LRH)*** *The concept can be measured using a linear probe.*

(iii) ***(Intervention LRH)*** *The concept can be modified by adding a steering vector.*

A central theoretical contribution of Park et al. (2024) is to show that these three definitions can be unified under an appropriate choice of inner product, which they term the *causal inner product*. Empirically, they further demonstrate that a range of linguistic features, such as male → female (gender), English → French (multilingualism), and country → capital (holonym/meronym relations), have linear representations that are mutually orthogonal when the concepts are causally separable (see Table 2).

However, upon closer inspection, this formulation is not fully aligned with the notion of the LRH as discussed in prior literature. For instance, what would constitute a causal and counterfactual pair for the *Golden Gate Bridge* in the sense of Templeton et al. (2024) (Section 2.2)? The set of "everything that is not the Golden Gate Bridge" does not satisfy a one-to-one correspondence, and thus

Context of LRH Usage	**Papers** (may appear in multiple rows)
Type 1. Theoretical analyses supporting LRH	Park et al. (2024); Jiang et al. (2024); Park et al. (2025); Liu et al. (2025)
Type 2. Empirical evidence: Supporting examples	Nanda et al. (2023); Turner et al. (2023); Templeton et al. (2024); Tigges et al. (2024); Park et al. (2024); Jiang et al. (2024); Gurnee and Tegmark (2024); Nguyen and Leng (2025); Park et al. (2025); Kim et al. (2025); Agarwal et al. (2025)
Type 3. Empirical evidence: Counterexamples	Csordás et al. (2024); Engels et al. (2024); Modell et al. (2025); Kantamneni and Tegmark (2025); Lei and Cooper (2025a); Walker et al. (2025)
Type 4. Empirical analysis of when LRH holds	Merullo et al. (2025)
Type 5. Statistical hypothesis testing of LRH	Reizinger et al. (2025); Guo et al. (2025)
Type 6. Critical discussions of LRH limitations	Marshall et al. (2024); Wattenberg and Viégas (2024); Tan et al. (2024); Canby et al. (2025a); Muraleedharan (2025); Geiger et al. (2025); Tiblias et al. (2025)
Type 7. Theoretical assumption†	Boix-Adsera (2024); Lee et al. (2025); Hübotter et al. (2025); Zhu et al. (2025); Stevinson et al. (2025)
Type 8. Methodological assumption‡	Templeton et al. (2024); Chanin et al. (2024); Ayonrinde et al. (2024); Heinzerling and Inui (2024); Skean et al. (2024); Bhalla et al. (2024); Tan et al. (2024); Makelov et al. (2024); Wu et al. (2024); Ayonrinde et al. (2024); Karvonen et al. (2024); Song et al. (2025); Li et al. (2025a,b); Cho and Hockenmaier (2025); Agarwal et al. (2025); Valois et al. (2025); Engels et al. (2025); Mencattini et al. (2025); Shen et al. (2025); Chen et al. (2024); Sutter et al. (2025); Lee et al. (2025); Yao and Du (2025); Makelov et al. (2025); Qiu et al. (2025); Tas and Wagner (2024); Zhang et al. (2025); Wang and Xu (2025); Vompa et al. (2025); Baek and Tegmark (2025); Tolooshams et al. (2025); Lei and Cooper (2025b); Costa et al. (2025); Kalajdzievski (2025); Chen et al. (2025); Paulo et al. (2025); Dalva et al. (2025)

Table 3: **Contexts in which the term "linear representation hypothesis (LRH)" is used across the literature.** A substantial fraction of prior work employs the LRH primarily as a methodological assumption (**Type 8**) to justify linear analysis tools, such as probing, sparse autoencoders, and steering vectors. As argued in Section 5.3, when such assumptions are invoked across different layers or feature definitions, the resulting methodological justification may not necessarily transfer across studies.

† Theoretical findings derived under the assumption that the LRH is true.
‡ LRH used to justify linear methods (e.g., linear probing, SAE, steering vectors).

cannot be formulated using pairs of the form:

$$\text{male} \to \text{female} = \{(\text{man}, \text{woman}), (\text{king}, \text{queen}), (\text{roi}, \text{reine}), \ldots\}.$$

These limitations are partially addressed by Park et al. (2025), which extends the framework to general is-a relations by relaxing the reliance on counterfactual pairs, while retaining the constraint that features should be causally separable.[2] Nevertheless, Park et al. (2024), Jiang et al. (2024), and Park et al. (2025) all share the common limitation that their theoretical analyses apply exclusively to the last layer of the model. We discuss this layer-specific limitation in more detail in Section 3.

[2]For example, gender (e.g., king→queen) and multilingualism (e.g., king→roi) are causally separable because they can be intervened on independently.

3 Target Embedding Layers

Last layer Park et al. (2024) provides a formalization of the linear representation hypothesis for binary concepts defined with counterfactual word pairs (e.g., male → female), showing that such concepts are well-defined linear representations as directions in the representation space. Further, Jiang et al. (2024) provides a theoretical analysis demonstrating that architectural choices such as softmax and cross-entropy promote a linear structure in the representations used for next-token prediction.

However, it is important to note that, in both works (Park et al., 2024; Jiang et al., 2024), the model is defined as

$$P(y \mid x) = \text{softmax}(\text{emb}(x)^\top \text{unemb}(y)),$$

where x denotes the input sequence (e.g., "He is the") and y denotes the next token to predict (e.g.,

"king"). By construction, this formulation applies only to the final layer representation immediately preceding the unembedding linear transformation. Accordingly, the theoretical guarantees established in these works are restricted to the last layer. Therefore, using these results to justify claims about intermediate layers $1, \ldots, L-1$ in a network with a total of L extends their conclusions beyond the scope of what is formally established.

Intermediate layers As an early attempt to support the LRH, Nanda et al. (2023) present that, when linear probes are trained on the hidden states of a GPT model trained on Othello game records to distinguish between white and black stones, performance is relatively poor (62–75%). This might naively be interpreted as evidence for a non-linear representation. However, when the same probes are trained to distinguish *my* stone versus *your* stone, linear probing achieves extremely high accuracy (99%) at the 4th layer of the 7-layer model, highlighting that the choice of features is crucial for determining whether a representation is linearly encoded.

Studies measuring linear probing performance and steering effectiveness for sentiment and detoxification tasks similarly find that these effects peak in the middle layers of the network and then decline toward the final layers (Turner et al., 2023; Tigges et al., 2024). Additionally, information about time and position is also linearly represented, with layer-wise performance plateauing around the halfway point of the network (Gurnee and Tegmark, 2024). Linear probing and intervention for political perspectives in LLMs are also observed most effective in middle layers (Kim et al., 2025), and a broad investigation spanning hallucination, persuasion, pessimism, refusal, sycophancy, and truthfulness reports that, in a 28-layer model, linear probing and steering are most effective at the 15th layer Agarwal et al. (2025).

Recently, Merullo et al. (2025) provide empirical evidence about when linear relations emerge for (subject, relation, object) triplets by using the subject-side hidden state $h^{(l)}_{\text{subject}}$ extracted from the l-th intermediate layer, and modeling the relation as:

$$h^{(L)}_{\text{object}} = W h^{(l)}_{\text{subject}} + b, \text{ where } l \in \{1, \ldots, L-1\}.$$

Building on the findings of Hernandez et al. (2024)[3], Merullo et al. (2025) further demonstrates that linear relational embeddings tend to emerge when the corresponding triplets have high co-occurrence frequency in the training data, and fail to do so when the frequency is low. This work is distinguished from prior studies in that it does not merely add supporting or counterexamples to the LRH, but instead seeks to identify the underlying boundary conditions that determine *when* linear representations arise.

4 Gap between Theory and Experiment

The existing literature on the LRH can be organized along several dimensions, including the perspective taken on LRH, the models studied, the layers examined, and the features considered, as summarized in Table 2. In addition, the extent to which each work provides theoretical versus empirical support for LRH is summarized in Table 3. Having organized the literature by target embedding type, a natural question arises: If the **final layer** is where linear structure is theoretically formalized and promoted (Park et al., 2024; Jiang et al., 2024; Park et al., 2025; Liu et al., 2025), why do empirical studies consistently report the clearest linear separability (Marks and Tegmark, 2024), the lowest probing error (Nanda et al., 2023; Tigges et al., 2024; Gurnee and Tegmark, 2024), and the most effective linear intervention (Li et al., 2023, 2025a; Shen et al., 2025; Kim et al., 2025; Agarwal et al., 2025) at **intermediate layers**? Furthermore, as shown in Table 1, the layers that experimentally exhibited non-linear representations were most often the last layer (Kantamneni and Tegmark, 2025; Lei and Cooper, 2025a; Walker et al., 2025).

Skean et al. (2025) propose an explanation why the last layer in LLMs is not always the best for embeddings, and why intermediate layers often provide better representations, based on information theory, geometry, and stability analysis. However, why *linear* representations should emerge effectively in the intermediate layers remains an open problem, which we summarize in Section 6.

5 What exactly is hypothetical in LRH?

The term *hypothesis* can be used in at least two distinct senses in the literature: scientific hypotheses, which propose explanatory claims about the world, and statistical hypotheses, which are formal

[3]Earlier work by Hernandez et al. (2024) observed that some relations (e.g., country–capital) are well captured by faithful linear relational embeddings, whereas others (e.g., company–CEO) are not, despite both being well learned.

statements used within statistical testing procedures (Alger, 2022). However, regarding the LRH, the literature is inconsistent about whether it should be understood as a statistical or a scientific hypothesis. In this section, we review both possibilities and, in particular, discuss the modifications required for the LRH to function as a falsifiable scientific hypothesis.

5.1 LRH as a Statistical Hypothesis

There is a vein of work that approaches the linear representation hypothesis from the perspective of statistical hypothesis testing. For instance, Guo et al. (2025) applied statistical hypothesis testing, finding that embeddings significantly satisfy a linear representation hypothesis, exhibiting stronger linear alignment and additive compositional generalization with respect to interpretable attributes (e.g., gender, age, and occupation) than expected under randomly permuted attribute–embedding pairings. Another example is Reizinger et al. (2025), who treat binary semantic attributes in ImageNet-X (Idrissi et al., 2022) defined by human annotations (e.g., object size or position) as proxies for latent factors, and test whether these attributes can be predicted from the second-to-last layer representations of ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2021) using linear decoders at levels significantly above chance.

Despite their methodological clarity, formulating the LRH as a statistical hypothesis faces two major limitations. First, specifying an appropriate null hypothesis requires strong assumptions about how embeddings would be structured *in the absence of linear representations*. In practice, such tests do not allow us to accept the hypothesis that an embedding space is structured by linear representations; for example, they only allow us to reject the hypothesis that the embeddings follow a multivariate Gaussian distribution (Li et al., 2025c), offering limited insight into whether the LRH itself holds. Second, these approaches rely on assumptions about the underlying data-generating process, including the distributions of latent features or embeddings, such as von Mises–Fisher models (Reizinger et al., 2025) or Monte Carlo simulation (Guo et al., 2025), which are difficult to justify or not clearly generalizable beyond the given dataset.

Nevertheless, viewing the LRH through the lens of statistical hypothesis testing has the advantage of yielding concrete and well-scoped contributions under clearly defined assumptions. However, work that formulates and evaluates the LRH in hypothesis testing remains relatively rare in the literature.

5.2 LRH as a Scientific Hypothesis

Arguably, one of the fundamental requirements of a scientific hypothesis is falsifiability (Popper, 1959); an idea that is formulated in a way that does not allow for falsification cannot be regarded as a scientific hypothesis. In Section 2.3, we discussed that the definition implicitly used in the majority of the literature corresponds to the weak LRH (Definition 2), and that this formulation is inherently challenging to falsify. Although a number of counterexamples are shown to be represented in nonlinear ways (Table 2), these results do not directly refute the weak LRH, since they can always escape refutation through post hoc redefinition of features—"well, that wasn't one of the ***some*** features!" Can the LRH, as it is currently formulated and being *increasingly* used (Table 1), be regarded as a scientific hypothesis?

We address that much of the recent discussion of the LRH has overlooked an early yet fundamental lesson pointed out by Gladkova et al. (2016): linguistic regularities must be carefully examined in conjunction with the data under investigation. In contemporary LLM-based LRH studies, the definition of a feature determines which subsets of the data are evaluated for linearity. We therefore propose that the LRH must be framed not as a hypothesis about 'some' feature, but as a hypothesis about 'one specific' feature f (along with model M and layer l as in Table 2), if it is to function as a falsifiable scientific claim.

More specifically, the LRH must be stated relative to a particular model M, representation location l, feature f, and evaluation domain $\mathcal{D}$. Accordingly, we treat the **LRH not as a single global statement, but as a family of local claims parameterized by** $(M, l, f, \mathcal{D})$:

Definition 4 (LRH; proposed falsifiable statement)**.** *Let M be a neural network model. Let $\mathrm{emb}^{(l)}(x) \in \mathbb{R}^d$ denote the embedding of an input x at a location l (e.g., layer index) in the model. Let $\mathcal{D}$ be an evaluation dataset with a feature labeling function $f : \mathcal{D} \to \{1, \ldots, K\}$.*

We say that the feature f has a ***linear representation*** *with respect to $M, l,$ and $\mathcal{D}$, if there exist $W \in \mathbb{R}^{K \times d}, b \in \mathbb{R}^K$ such that*

$$\Pr_{x \sim \mathcal{D}} \left[\operatorname*{argmax}_{k \in \{1,\ldots,K\}} \left(W_k^\top \mathrm{emb}^{(l)}(x) + b_k \right) = f(x) \right]$$

$$\geq 1 - p \quad \textit{for a small } p \geq 0.$$[4]

5.3 From Slogan to Statement

Would there be merit in formulating the hypothesis as in Definition 4? One reason this question matters becomes clear from Table 3: the LRH is frequently used as experimental background for employing specific methods (e.g., linear probing or sparse autoencoder with a single-layer encoder). This practice is a double-edged sword. On the one hand, building methods on prior work is essential for cumulative scientific progress. On the other hand, grounding methodological justification in different layers or unrelated linguistic features can create an illusion that a choice of methods is theoretically justified, even when no shared justification actually exists. In particular, when it is unclear whether different studies are referring to the same feature and layer, using the LRH as a slogan to motivate linear methodologies leads to claims that are unfalsifiable and thus, borrowing Wolfgang Pauli's famous phrase, *not even wrong*.

Then what is currently built on top of the LRH? Table 3 also sheds light on this issue by indicating which lines of work would be affected if the LRH is falsified for a particular combination of model, feature, and layer. For studies that invoke the LRH as a theoretical assumption (**Type 7** in Table 3), the implications are straightforward: once the LRH is falsified for a given model, feature, or layer, the corresponding theoretical claims no longer hold in that setting. For studies that rely on the LRH as a methodological justification (**Type 8** in Table 3), while their empirical findings remain valid as observations about specific models and datasets, the methodological justification becomes weaker if the LRH is assumed without being tested. In particular, when non-linear representations are not explored as alternatives, one cannot rule out the possibility that more appropriate non-linear methods would yield stronger or more informative results.

[4]The restriction to *small* positive p avoids a trivial regime: if p is too large, the condition becomes vacuous, while $p = 0$ corresponds to the exact/perfect case. This formulation is also consistent with prior work (Table 3, Type 2), which typically allows a nonzero tolerance. As discussed further in Appendix A, this formulation is aligned with Definitions 1, 2, and 3. It also provides a precise language for describing the empirical claims organized in Table 2: M corresponds to the **Model** column, l to the **Layer** column, and $(f, \mathcal{D})$ refines the **Feature** column. Under this view, the **LRH** column in Table 2 can be interpreted as validating the statement $\mathrm{LRH}(M, l, f, \mathcal{D}) \to \{\texttt{holds, does not holds}\}$.

6 Open problems

Building on the preceding review of the literature, we highlight three open problems that are well-formulated and warrant further investigation.

Beyond Transformers For binary-labeled features, linear representations are commonly observed in generative pre-trained transformers (Table 2). However, whether a similar linear structure arises in other model groups, such as language diffusion models, remains largely unexplored.

The relationship between layer depth and linear representation Existing research has theoretically examined linear representations of various features and has shown that cross-entropy loss and softmax can promote a linear structure in the final layer of neural networks (Park et al., 2024; Jiang et al., 2024; Park et al., 2025). However, it remains unclear why empirical studies consistently find stronger linear representations in intermediate layers than at the final layer (Section 4).

Why do linear representations arise in the first place? Despite extensive empirical evidence, the mechanisms underlying the emergence of linear representations remain poorly understood. One explanation is that neural networks prefer to retrieve information "cheaply" (Henighan and Olah, 2023). Another possibility is that linear representations arise from the data distribution itself, such as co-occurrence in pre-training data (Merullo et al., 2025). However, whether this is the only factor, and whether the relationship extends beyond correlation to causality, remains an open question. More fundamentally, using Definition 4, we currently lack an explanation of why a feature f exhibits a linear representation for some tuples $(M, l, f, \mathcal{D})$ but not for others.

7 Conclusion

In this work, we examined the Linear Representation Hypothesis (LRH). We showed that definitions of LRH have evolved across prior work and are not always consistent. In particular, existing studies often draw on results obtained across different layers or feature types. We proposed that LRH must be stated together with an explicit specification of the data (or the feature defined over it), layer, and model under consideration. Based on a systematic categorization of prior work, we identified several open problems for future research.

Limitations

We deliberately focus on literature that explicitly frames its contributions around the linear representation *hypothesis*. While this choice provides coherence when organizing an increasingly large body of work, it necessarily excludes many earlier and influential studies that investigate linear structure but do not connect their findings to the LRH. A notable example is Hewitt and Manning (2019), which demonstrates that distances in syntactic parse trees can be recovered by a linear probe. In that work, linearity is not introduced to support a hypothesis, but rather as a methodological constraint intended to prevent the probe from learning the syntactic parsing task itself. Such studies fall outside the scope of our survey, although they have inspired later research in this literature.

Another limitation concerns the scope of the causal claims considered in this survey, particularly those that are central to the mechanistic interpretability literature (Saphra and Wiegreffe, 2024). In this paper, we focus on causal structure within the representation space, corresponding to the mapping $x \to \mathrm{emb}(x)$. By contrast, a different line of work asks whether such representations are causally used by the model when producing predictions, corresponding to $\mathrm{emb}(x) \to y$. Prior work has emphasized that correlations between embeddings and linguistic properties do not necessarily imply that models causally rely on those properties in their predictions (Elazar et al., 2021; Belinkov, 2022; Davies et al., 2023; Canby et al., 2025b). Understanding whether representations that appear linear are also causally used by the model remains an important direction for future research.

References

Isha Agarwal, Saharsha Navani, and Fazl Barez. 2025. Context matters: Analyzing the generalizability of linear probing and steering across diverse scenarios. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Bradley E. Alger. 2022. Neuroscience needs to test both statistical and scientific hypotheses. *Journal of Neuroscience*, 42(45):8432–8438.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4.

Kola Ayonrinde, Michael T Pearce, and Lee Sharkey. 2024. Interpretability as compression: Reconsidering sae explanations of neural activations. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*.

David D Baek and Max Tegmark. 2025. Towards understanding distilled reasoning models: A representational approach. *Building Trust Workshop at ICLR 2025*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. 2024. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Enric Boix-Adsera. 2024. Towards a theory of model distillation. *arXiv preprint arXiv:2403.09053*.

Marc Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. 2025a. How reliable are causal probing interventions? *AACL*.

Marc E Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. 2025b. How reliable are causal probing interventions? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 857–878.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Isaac Bloom. 2024. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Interpretable AI: Past, Present and Future*.

Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. 2025. Transferring linear features across language models with model stitching. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xi Chen, Zhihui Zhu, and Andrew Perrault. 2024. Understanding learned representations and action collapse in visual reinforcement learning. In *Reinforcement Learning Journal*.

Ikhyun Cho and Julia Hockenmaier. 2025. Toward efficient sparse autoencoder-guided steering for improved in-context learning in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28961–28973, Suzhou, China. Association for Computational Linguistics.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba Ba. 2025. From flat to hierarchical: Extracting sparse representations with matching pursuit. *39th 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.

Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. 2025. Fluxspace: Disentangled semantic editing in rectified flow models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13083–13092.

Adam Davies, Jize Jiang, and ChengXiang Zhai. 2023. Competence-based analysis of language models. *arXiv preprint arXiv:2303.00333*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

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*.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2024. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*.

Joshua Engels, Logan Riggs Smith, and Max Tegmark. 2025. Decomposing the dark matter of sparse autoencoders. *Transactions on Machine Learning Research*.

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

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Adrián González, Graham Russel, Astrid Márquez, Jose Ali Moreno, Cristina Garcia, Carlos Dominguez, Omar Colmenares, and Juan José Machado. 2007. Supervised farm classification from remote sensing images based on kernel adatron algorithm. In *2007 IEEE International Geoscience and Remote Sensing Symposium*, pages 3345–3348. IEEE.

Zhijin Guo, Chenhao Xue, Zhaozhen Xu, Hongbo Bo, Yuxuan Ye, Janet B Pierrehumbert, and Martha Lewis. 2025. Quantifying compositionality of classic and state-of-the-art embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22130–22146.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Benjamin Heinzerling and Kentaro Inui. 2024. Monotonic representation of numeric attributes in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Bangkok, Thailand. Association for Computational Linguistics.

Tom Henighan and Chris Olah. 2023. Dictionary learning worries.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*.

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

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*.

Jonas Hübotter, Patrik Wolf, Alexander Shevchenko, Dennis Jüni, Andreas Krause, and Gil Kur. 2025.

Specialization after generalization: Towards understanding test-time training in foundation models. *39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop on Continual and Compatible Foundation Model Updates (CCFM)*.

Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. 2022. Imagenet-x: Understanding model mistakes with factor of variation annotations. *arXiv preprint arXiv:2211.01866*.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. 2024. On the origins of linear representations in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21879–21911. PMLR.

Damjan Kalajdzievski. 2025. The logical implication steering method for conditional interventions on transformer generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 28689–28720. PMLR.

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*.

Adam Karvonen, Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Smith, Claudio Mayrink Verdun, David Bau, and Samuel Marks. 2024. Measuring progress in dictionary learning for language model interpretability with board game models. *Advances in Neural Information Processing Systems*, 37:83091–83118.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. *ICLR*.

János Kramár, Joshua Engels, Zheng Wang, Bilal Chughtai, Rohin Shah, Neel Nanda, and Arthur Conmy. 2026. Building production-ready probes for gemini. *arXiv preprint arXiv:2601.11516*.

Sewoong Lee, Adam Davies, Marc E. Canby, and Julia Hockenmaier. 2025. Evaluating and designing sparse autoencoders by approximating quasi-orthogonality. In *Second Conference on Language Modeling*.

Ge Lei and Samuel J Cooper. 2025a. Do llamas understand the periodic table? *Digital Discovery*, 4(12):3455–3465.

Ge Lei and Samuel J Cooper. 2025b. The representation and recall of interwoven structured knowledge in llms: A geometric and layered analysis. *arXiv preprint arXiv:2502.10871*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. 2025a. Fairsteer: Inference time debiasing for llms with dynamic activation steering. *arXiv preprint arXiv:2504.14492*.

Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. 2025b. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4).

Yuxiao Li, Eric J Michaud, David D Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. 2025c. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344.

Yuhang Liu, Dong Gong, Yichao Cai, Erdun Gao, Zhen Zhang, Biwei Huang, Mingming Gong, Anton van den Hengel, and Javen Qinfeng Shi. 2025. I predict therefore i am: Is next token prediction enough to learn human-interpretable concepts from data? *arXiv preprint arXiv:2503.08980*.

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*.

Aleksandar Makelov, Georg Lange, and Neel Nanda. 2025. Towards principled evaluations of sparse autoencoders for interpretability and control. In *The Thirteenth International Conference on Learning Representations*.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Thomas Marshall, Adam Scherlis, and Nora Belrose. 2024. Refusal in llms is an affine function. *arXiv preprint arXiv:2411.09003*.

Tommaso Mencattini, Riccardo Cadei, and Francesco Locatello. 2025. Exploratory causal inference in saence. *arXiv preprint arXiv:2510.14073*.

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. 2025. On linear representations and pretraining data frequency in language models. In *The Thirteenth International Conference on Learning Representations*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. 2025. The origins of representation manifolds in large language models. *arXiv preprint arXiv:2505.18235*.

Abhinav Muraleedharan. 2025. On the limits of linear representation hypotheses in large language models: A dynamical systems analysis. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore. Association for Computational Linguistics.

Trung Nguyen and Yan Leng. 2025. Toward a flexible framework for linear representation hypothesis using maximum likelihood estimation. *arXiv preprint arXiv:2502.16385*.

Chris Olah. 2024. The next five hurdles. `https://transformer-circuits.pub/2024/july-update/index.html#hurdles`. Circuits Updates — July 2024, Transformer Circuits Thread.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2025. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. 2025. Automatically interpreting millions of features in large language models. In *Forty-second International Conference on Machine Learning*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Gordon F Pitz, Lock Sing Leung, Christopher Hamilos, and Wilbann Terpening. 1976. The use of probabilistic information in making predictions. *Organizational Behavior and Human Performance*, 17(1):1–18.

Karl Popper. 1959. *The Logic of Scientific Discovery*. Julius Springer, Hutchinson & Co.

Haiquan Qiu, You Wu, Dong Li, Jianmin Guo, and Quanming Yao. 2025. Superpose task-specific features for model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4200–4214, Suzhou, China. Association for Computational Linguistics.

Patrik Reizinger, Alice Bizeul, Attila Juhos, Julia E Vogt, Randall Balestriero, Wieland Brendel, and David Klindt. 2025. Cross-entropy is all you need to invert the data generating process. In *The Thirteenth International Conference on Learning Representations*.

Naomi Saphra and Sarah Wiegreffe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498.

William F. Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D. Lane. 2025. LLM unlearning via neural activation redirection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Oscar Skean, Md Rifat Arefin, and Ravid Shwartz-Ziv. 2024. Does representation matter? exploring intermediate layers in large language models. In *Workshop on Machine Learning and Compression, NeurIPS 2024*.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by Layer: Uncovering hidden representations in language models. *ICML*.

Lewis Smith. 2024. The 'strong' feature hypothesis could be wrong.

Jiajun Song, Zhuoyan Xu, and Yiqiao Zhong. 2025. Out-of-distribution generalization via composition: A lens through induction heads in transformers. *Proceedings of the National Academy of Sciences*, 122(6):e2417182122.

Edward Stevinson, Lucas Prieto, Melih Barsbey, and Tolga Birdal. 2025. Adversarial attacks leverage interference between features in superposition. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

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*.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37:139179–139212.

Omer Sahin Tas and Royden Wagner. 2024. Words in motion: Extracting interpretable control vectors for motion transformers. In *Interpretable AI: Past, Present and Future*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Federico Tiblias, Irina Bigoulaeva, Jingcheng Niu, Simone Balloccu, and Iryna Gurevych. 2025. Shape happens: Automatic feature manifold discovery in llms via supervised multi-dimensional scaling. *arXiv preprint arXiv:2510.01025*.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. 2024. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US. Association for Computational Linguistics.

Bahareh Tolooshams, Ailsa Shen, and Anima Anandkumar. 2025. Sparse autoencoder neural operators: Model recovery in function spaces. In *UniReps: 3rd Edition of the Workshop on Unifying Representations in Neural Models*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Pedro HV Valois, Dipesh Satav, Rodrigo AP de Campos, Gulpi QO Pratamasunu, and Kazuhiro Fukui. 2025. Vision language model interpretability with concept guided decoding. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 397–402. IEEE.

Enrico Vompa, Tanel Tammet, and Mohit Vaishnav. 2025. Beyond the linear separability ceiling: Aligning representations in vlms. *arXiv preprint arXiv:2507.07574*.

Thomas Walker, Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. 2025. Centroid affinity: How deep networks represent features. In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Zijian Wang and Chang Xu. 2025. Thoughtprobe: Classifier-guided llm thought space exploration via probing representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6029–6050.

Martin Wattenberg and Fernanda B Viégas. 2024. Relational composition in neural networks: A survey and call to action. *Workshop on Mechanistic Interpretability*.

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

Yifei Yao and Mengnan Du. 2025. Adaptivek sparse autoencoders: Dynamic sparsity allocation for interpretable llm representations. *arXiv preprint arXiv:2508.17320*.

Thomas R Zentall. 2001. The case for a cognitive approach to animal learning and behavior. *Behavioural processes*, 54(1-3):65–78.

Shenyi Zhang, Yuchen Zhai, Keyan Guo, Hongxin Hu, Shengnan Guo, Zheng Fang, Lingchen Zhao, Chao Shen, Cong Wang, and Qian Wang. 2025. Jbshield: defending large language models from jailbreak attacks through activated concept analysis and manipulation. In *Proceedings of the 34th USENIX Conference on Security Symposium*, SEC '25, USA. USENIX Association.

Jianing Zhu, Zongze Li, Chandler Squires, Qizhou Wang, Bo Han, and Pradeep Ravikumar. 2025. On the fragility of latent knowledge: Layer-wise influence under unlearning in large language model. In *ICML 2025 Workshop on Machine Unlearning for Generative AI*.

A Connections Between Definitions

A.1 Park et al. (2024)'s Approach

First, the three different notions of linear representation summarized by Park et al. (2024) can be expressed mathematically as follows using the example described in Eq. 1:

Subspace:

$$v_{\text{male}\to\text{female}} \approx \text{emb}(\text{"queen"}) - \text{emb}(\text{"king"}).$$

Intervention:

$$\text{emb}(\text{"king"}) + \alpha v_{\text{male}\to\text{female}} \approx \text{emb}(\text{"queen"}).$$

Measurement:[5].

$$\text{feature}(x) = \text{emb}(x)^\top v_{\text{feature}}.$$

[5]This can be viewed as a formalization of Definition 1, which states that "neural networks represent input features as directions in activation space." Under this formulation, the feature vector can be defined more flexibly than under the other two definitions, since it does not require the feature vector to correspond to a counterfactual concept. For example, in Section 2.2, when using a single-layer sparse autoencoder, each feature is measured as $\text{emb}(x)^\top w_{\text{enc}}$, which implicitly adopts this definition as the underlying assumption for feature detection. Importantly, "being the Golden Gate Bridge" is not itself a counterfactual concept, yet it can still be represented by a measurement of the form $\text{emb}(x)^\top v_{\text{GoldenGateBridge}}$.

Here, Park et al. (2024) restrict $\mathrm{emb}(x)$ to the final layer where the softmax is applied in order to derive the following relationship:

$$P(Y = y \mid x) = \mathrm{softmax}\Big(\mathrm{emb}(x)^\top \mathrm{unemb}(y)\Big) = \frac{\exp\big(\mathrm{emb}(x)^\top \mathrm{unemb}(y)\big)}{\sum_{y'} \exp(\mathrm{emb}(x)^\top \mathrm{unemb}(y'))},$$

where Y is a random variable representing the next token and x denotes the input text. For example, x = "He is the ", $Y(0)$ = "king", and $Y(1)$ = "queen". To measure how much more likely the outcome $Y(1)$ is compared to $Y(0)$, consider the log-odds:

$$\log \frac{P(Y(1) \mid x)}{P(Y(0) \mid x)} = \log \frac{\exp(\mathrm{emb}(x)^\top \mathrm{unemb}(Y(1)))}{\exp(\mathrm{emb}(x)^\top \mathrm{unemb}(Y(0)))}$$

$$= \mathrm{emb}(x)^\top \mathrm{unemb}(Y(1)) - \mathrm{emb}(x)^\top \mathrm{unemb}(Y(0))$$

$$= \mathrm{emb}(x)^\top \left(\mathrm{unemb}(Y(1)) - \mathrm{unemb}(Y(0))\right)$$

$$= \mathrm{emb}(x)^\top v_{\text{feature}}.$$

While Park et al. (2024)'s derivation is advantageous in that v_{feature} can be easily obtained from next-token unembedding vectors and is more theoretically tractable, the three definitions can be unified without restricting the analysis to the final layer, as will be described in the next section.

A.2 Beyond the Final Layer

Formally, if the subspace exists, the definition of intervention follows trivially with $\alpha = 1$. To see the connection with the measurement definition, suppose we apply an intervention to the embedding of an input text x, producing a modified embedding $\mathrm{emb}(x)' = \mathrm{emb}(x) + \alpha v_{\text{feature}}$. Then,

$$\mathrm{feature}(\mathrm{emb}(x)') = (\mathrm{emb}(x) + \alpha v_{\text{feature}})^\top v_{\text{feature}}$$

$$= \mathrm{emb}(x)^\top v_{\text{feature}} + \alpha v_{\text{feature}}^\top v_{\text{feature}}$$

$$= \mathrm{feature}(\mathrm{emb}(x)) + \alpha \|v_{\text{feature}}\|^2.$$

Thus, an intervention defined as above leads to a linear change in the feature value measured by the measurement definition.

A.3 The Fourth Definition

Notably, relying solely on the measurement score can be misleading: a high inner product with a feature direction is not sufficient to establish a linear representation unless representations that lack the feature also yield consistently lower scores. In addition to the aforementioned three notions, there exists another related notion that is not expressed in exactly the same form, namely linear separability.

Linear separability:

$$\exists\, w, b \quad \text{s.t.:} \quad w^\top \mathrm{emb}(\text{"queen"}) + b > 0$$
$$w^\top \mathrm{emb}(\text{"king"}) + b \leq 0.$$

Given the measurement definition

$$\mathrm{feature}(\mathrm{emb}(x)) = \mathrm{emb}(x)^\top v_{\text{feature}},$$

introducing a threshold b yields the linear decision rule

$$\mathrm{emb}(x)^\top v_{\text{feature}} + b > 0$$
$$\text{or} \quad \mathrm{emb}(x)^\top v_{\text{feature}} + b \leq 0.$$

Thus, linear separability can be viewed as a thresholded version of the measurement definition, with $w = v_{\text{feature}}$. As discussed in Appendix A.2, this notion captures a core property shared across these formulations. It also aligns with the widespread use of linear probing for testing whether a feature is linearly represented. This observation explains why the proposed Definition 4 takes the same form as a linear probe classifier, which learns the form $\hat{y} = \sigma(w^\top \mathrm{emb}(x) + b)$, where w and b are learned parameters and $\sigma(\cdot)$ denotes the sigmoid function. When the learned probe direction w aligns with v_{feature}, the probe measures the same feature and applies a threshold to determine its presence.

This connection yields a simple corollary: linear probes can be used to find linear representations. To encompass the diverse probing methodologies in existing literature, Definition 4 explicitly introduces the (M, l, f, D) tuple. For instance, the exact conceptualization of the target feature f determines whether a linear representation is identifiable, as shown in recent studies where representations emerge only under specific feature definitions (Nanda et al., 2023). Furthermore, modern probes frequently aggregate activations across multiple token positions rather than relying on a single token (Kramár et al., 2026). Definition 4 accounts for these variations by requiring the probing location l to be explicitly specified. By formalizing these elements, our definition establishes a mapping between empirical linear probes and the theoretical framework of the linear representation hypothesis.