Mallory, F. (2026). Formats of representation in large language models. *Philosophy and the Mind Sciences*, 7. https://doi.org/10.33735/phimisci.2026.12091

Formats of Representation in Large Language Models

Abstract: This paper argues for a pluralist approach to representation in large language models. There are two parts to this pluralism, the first is that we should recognise more than one *vehicle* of representation in transformer models. Call this *vehicle pluralism*. Rather than identifying the vehicles of representation with a single component of a system, e.g. individual neurons, patterns of activation, regions in the activation space, we should acknowledge multiple systems of representation within a network operating with different vehicles. The second claim is that we should recognise that there are different *formats* of representation in transformer models. Transformer models do not operate with a purely analogue, structural, or symbolic architecture but are a hybrid system of representation. Finally, I will discuss how this relates to several working hypotheses about representation that have become adopted in the field of mechanistic interpretability.

Keywords: Representational format, LLMs, linear representation hypothesis, analogue representation, structural representation, vehicle

1. Introduction

In this paper, I will argue for a pluralist approach to representation in large language models. There are two parts to this pluralism, the first is that we should recognise more than one *vehicle* of representation in neural networks. Call this *vehicle pluralism*. Rather than identifying the vehicles of representation with one component of a system, e.g. individual neurons, patterns of activation, regions in the activation space and so forth, we should acknowledge multiple systems of representation within a network operating with different vehicles. The second claim is that we should recognise that there are different *formats* of representation in deep neural networks. Specifically, I will argue that neural language models exhibit nominal, analogue, and structural formats of representation at different stages in processing. An advantage of adopting pluralism will be that we can see how different vehicles emerge depending on the properties of network architecture as well as the data to which the network has been exposed in training.

I'll argue for this by presenting several cases where it makes sense to think of a system as representing with different types of vehicles and different formats of representation. This will involve a simple case study contrasting the analogue format of simple word embeddings with the structural representation facilitated by transformer models. In short, structural representation arises when relations between vehicles of representation represent relations between contents (Shea, 2018, Swoyer, 1991). In NLP, this is most commonly facilitated by attention heads. I will then situate more recent work on linear representation and superposition within the context of the philosophical literature on representation and suggest that this indicates a new, more cartographic format of representation in LLMs.

Before starting, it's worth noting that it shouldn't be surprising if language models exploit multiple formats and vehicles of representation. Few would claim that brains have a single unit of representation with contemporary work ascribing contents not just to individual neural activations (Quian Quiroga et al. 2005), and clusters of neural activation (Nagelhus et al. 2023), but also to oscillatory dynamics of neural activations (Martinez & Artiga, 2024). Similarly with formats, while some cortical regions appear to have analogue correspondences with the domains they represent (Shagrir, 2010, Chang et al. 2025), other regions allow for richer structural representations (e.g. hippocampal maps, Shea, 2018) and in other cases, human brains may exploit symbolic representations (Pulvermüller, 2013). It may be a feature of complex information processing systems that they utilise multiple formats for organising and processing information. In any case, pluralism is a modest hypothesis.

The structure of this paper is as follows. Section 2 sets out some basic assumptions that the paper will make about representation. Section 3 argues that there are multiple vehicles of representation in language models. Section 4 argues that there are multiple formats of representation in language models. To give an example of this, I will argue that the analogue format of simple word embeddings becomes a form of structural representation when the relations between embeddings become an object of computations with attention heads. The upshot of this will be that even relatively simple models utilise several different kinds of representation and that a theory of representation for LLMs should be sensitive to the interaction between these. Section 5 will introduce some of the recent representation hypotheses within the mechanistic interpretability literature and situation them within this discussion.

Before discussing systems of representation, we need to disambiguate two kinds of representation in DNNs. The combination of memory and computation within the same architecture is one of the major differences between artificial neural networks and traditional computers (Shea, 2021). Despite the unity of memory and computation, we can nevertheless distinguish information that has been stored from information that is being processed. On the one hand, we have the representational content encoded in the network's parameters. This *model-representation* (or 'm-representation', m for memory, matrix, mechanism) is the object of enquiry in research that probes model weights. It corresponds to the representational content a network acquires through training. It is in virtue of this information that layers of the network output the particular values they compute. The information encoded in these parameters determines how the mechanism will respond to an input; geometrically put, how it will scale, project, rotate, translate, and compress its inputs.

In contrast, when a vector or higher-rank tensor is output from one layer and passed to another, we can view this as a signal sent from one layer of the network to the next. This is *signal representation* (s, 2010, LaCroix, 2020). These signals are output by producer-mechanisms and consumed by downstream consumer-mechanisms although most layers are both producer and consumer. Signals can also be understood as keys that function to unlock values, distributions over data in a second layer (Geva et a. 2020). The content of the signal determines how the consumer mechanism will respond to it. At different layers, different contents may be more important, e.g., syntactic content plays a greater role in early layers, (Peters et al. 2018). It may be that mechanisms *add* content to a signal as it passes through the network (Geva et al. 2022, Ferrando et al. 2024) while other layers may compress information so that the signal is more

efficiently encoded. Activation patching can be used to identify signal representations (Wang et al. 2022).

In many cases, it will be plausible to say that signals represent the inputs being processed, i.e. tokens of text, while their location within an embedding space represents *that* those input tokens possess a particular feature. This is similar to how the name of a town may appear on a map in text, while its location in the world is given by where that name is written on the map and its relations to other names. Hybridity is a common feature of systems of representation.

2. Systems of representation

The following discussion will require some minimal, overly simplistic, commitments about representation. In what follows, a system of representation is formed of a set of vehicles and a mapping function mapping these vehicles and/or the relations between them onto a domain of contents. The system S and domain D may be structured by any number of relations $\{R_1,R_2,R_3 \ldots R_n\}$ and $\{R'_1,R'_2,R'_3 \ldots R'_n\}$. Put roughly, we can think of a system as an ordered triple of an interpretation function, a set of vehicles, S, and a set of relations **R** and we can think of them mapping onto a set of contents D with a structure given by a set of relations **R'**:

$$\langle f, v \in S, R_1,R_2 \ldots R_n\rangle \rightarrow \langle c \in D, R'_1,R'_2 \ldots R'_n\rangle$$

This kind of quasi-formalisation is a massive oversimplification but it provides a useful tool for distinguishing different systems of representation. Structured systems of vehicles can be points on a map, words in a sentence, notes on a stave, points on a thermometer, place cells in a rat's hippocampus and so on. Relations might be physical distances, differences in size, semantic structures, co-activation structures, and countless other structures identified by researchers. The content of a representation is what it represents and the vehicle of representation is the syntactically individuable particular within the system that bears this content. A *format of representation* (e.g. analogue, symbolic, iconic, distributed etc.) is a property of the system of representation as a whole and, in this paper, formats will be individuated by both intensional properties of the interpretation function and the set of relations over which the function is defined (this isn't the only way to do this, e.g. Lee et al. 2023, Maley, 2011). I'll say more about each of these below but first I should clarify the scope of this discussion.

Work on the philosophy of representation primarily focusses on the conditions under which a mapping function *f* holds. In this paper, I want to avoid the metasemantic question of what *grounds* representation and focus instead on the intensional properties of the interpretation function. These properties can help explain why the system is good or bad for its task, without themselves explaining why it is that the system represents. A map might have structural properties that make it good for navigating space without these structural properties explaining why it is a map. Likewise, a signalling system might be compositional but that might not be *why*

it corresponds to the world.[1] I won't take a stance on *why* there is representation here. At the very least, talking about representation saves a lot of time. Unlike human brains, we can give completely causal or mathematical accounts of the operations of LLMs. The model parameters can be identified (though there may be many billions) and the core operations are typically simple matrix multiplications and ReLUs. But speaking at this level of description is not just practically infeasible but seems to miss out on *real patterns* in the networks' organisation (Dennett, 1991). Talk of representation is extremely useful.

The discussion that follows will rely on some core, widely shared assumptions. First, the relation of representation is neither causation nor correlation but some third thing. The states of a system may correlate with a variety of variables without representing them (just as the rings of a tree correlate with the tree's age but don't represent it). Any adequate theory of representation should be able to distinguish between information that is unexploited yet recoverable from the system and information that is represented by the system. There are also countless possible trivial *fs* that a complete theory of representation must exclude somehow (the 'liberality concern'). Accordingly, I will also assume that a vehicle of representation must in some sense be *used* ('consumed', 'exploited') by a system to count as a representation (Rai et al. 2024, Harding, 2023, Geiger et al. 2025). So any reference to systems, vehicles, or relations in what follows is relative to some consumer. Saying that representation is relative to a consumer is not saying that it is relative to a *probe.* One might hold that there is a fact of the matter about what a system represents to some other system without saying that this representation is relative to the scientist (or other third party) trying to study representation.

I'll similarly avoid commitments regarding the contents of representation. Different variants of the standard model take different contents as primitives: entities (Gallistel, 1994), states-of-affairs (Millikan, 1984), possible worlds (Stalnaker, 1999), conditions (Shea, 2018), facts, situations, etc. Within machine learning, different contents are typically subsumed under the category of 'features' and while there is an active debate about whether 'features' must be interpretable by humans the discussion here won't take a stance on this issue.[2] Some features are taken to be continuous variables (e.g. how large is the creature described) while others are taken to be discrete (e.g. is the text in French or not). In the next section, I will set out some general principles for distinguishing formats of representation and argue that language models can exhibit nominal, analogue and structural representation.

---

[1] For example, a recent debate concerns whether or not structural correspondence (cast typically as second-order isomorphism) is necessary for representation (Facchin, 2024, Artiga, 2025). Shepard & Chipman originally developed their theory of second order isomorphism as a response to Wittgenstein's private language argument which was taken to discredit the idea of a first order isomorphism between private objects, 'meanings' in the head and entities in the world (Shepard & Chipman, 1970). Inspired by Hebb, they focus on functional relations of co-activation aligning with similarity relations between entities (much as Shea presents structural representation in the hippocampal place area). Second order isomorphisms have been identified in color and letter spaces (Hanson et al. 2020). The Shephard & Chipman approach has recently been championed by Ned Block (Block, 2023) and in classic work (Swoyer, 1991).

[2] While some stipulate that features are human-interpretable (Rai, et al. 2024, Ferrando et al, 2024), others hold that "it seems important to allow for features we might not understand" (Elhage et al, 2022). Both sides have valid concerns. Shea articulates the concern with appealing to features that are more fine-grained, 'microfeatural', than human concepts clearly: "The microfeatural approach encourages the view that, if connectionist systems represent, they do so in a way that is highly complex, with contents quite unlike those in everyday explanations. Their contents may even be ineffable. From here it is only a short step, via viewing connectionist networks as some kind of model of human brains, to eliminativism about the contents ascribed in everyday psychological explanations, which some are happy to embrace" (Shea, 2007: 249). Yet if we are to be realist about representation, then this realism should presumably make us willing to accept that some complex systems represent the world differently to us.

3. Vehicles of Representation in LLMs

A vehicle of representation must be a non-semantically individuable unit that plays a causal role in the operation of a system. In the following, we will largely be dealing with vehicle types rather than tokens so that different particular systems of representation may share the same vehicle types.

The particular systems of representation we will be considering here are transformer-based language models (for a readable introduction to the architecture see Elhage et al. 2021). Transformer models can be broken down into several subsystems; tokenisation and embedding layers, a series of transformer blocks, each containing attention heads (i.e. key, query, value matrices) along with MLPs, a residual stream from which information is copied to the transformer blocks and to which the output of the transformers is written, some more MLPs, and a final unembedding layer.

Various candidates have been identified as the syntactically-individuated vehicles of representation in artificial neural networks, from individual neurons (Eliasmith, 2003, Clark, 2001), activations (Harding, 2023), clusters within the activation space (Shea, 2007, Azhar, 2016) to the whole network (Fodor, 2000, though he seemed to regard this as a *reductio*). The two dominant positions in the philosophical literature are the activations view and the cluster view.

> **Activation Hypothesis:** The vehicles of representation in an ANN are the activations at a given layer (i.e. the vector values output at a layer) (Churchland, 1998).

According to the activation hypothesis, vehicle types are individuated at the level of activation vectors. Each vector output at a given layer within an ANN is the vehicle for the particular representational content that vector bears. This does not entail that that every vector picks out a different content but it does lead to a very fine-grained delineation of vehicle types. Particular activations, often a list of floating-point values, are the bearers of content.

Nevertheless, the most developed account of representation in ANNs is the polytope hypothesis according to which vehicle types correspond to *regions* within an embedding space (Azhar, 2016, Shea, 2007, Shea, 2023).

> **Polytope Hypothesis:** The vehicles of representation in ANNs are clusters or polytopes within the state space of the neural network (Hinton, 1986, Shea, 2007, Azhar, 2016, Shea, 2023)

The polytope hypothesis treats regions within the activation space of the network as the vehicles of representation. There are several motivations for ascribing content to regions rather than individual embedding vectors. Abstracting away from individual activations allows for comparisons between different models which may represent the same contents but with different

individual vector values.[3] Furthermore, by definition, differences within a cluster are computationally irrelevant. The non-linear activations at each layer may not distinguish between neighbouring inputs whose values are at a distance from the activation threshold.[4] In effect, "downstream processing has effectively carved boundaries in the state space" (Shea, 2023: 175). The status of regions as vehicles therefore depends upon the activation functions used in downstream layers. This approach to model interpretation is most obvious seen in the case of clustering algorithms but the final layer of a logistic classification network can be understood in these terms. At the final layer of a classifier performing logistic regression, what matters is the side of the regression curve the output lands on, not its particular values. In a sense, according to the model, any vector in that region counts as the same *answer* to the classification *question* parameterised by the model.

However, there are also reasons to think that the polytope hypothesis, while appropriate for some architectures (i.e. feed-forward classificatory networks) and layers, may not be the best way to think about vehicles of representation at all points of processing in all networks, and in particular, transformers. To give a quick overview of transformer architecture, a typical transformer will have a tokeniser mapping text to tokens, and an embedding layer which maps tokens and their positions in a sequence onto embedding vectors. This is followed by transformer layers composed of attention heads which compute the distances between embeddings and then use these distances to alter each embedding which is then returned to the residual stream of the model before being passed to the next layer. Eventually, an unembedding layer maps the embeddings you've been working on to logit values for each word in the vocabulary. If the model is performing autoregressive language modelling (i.e. guessing the next word), the a softmax function is applied to the final set of logits which ensures the values sum to 1 enabling the model to treat its output as a probability distribution of possible next words.

First, standard tokenisation methods like byte-pair encoding map tokens to particular embedding vectors. This constitutes a discrete division of the activation space. If, due to some electrical interference, the tokeniser maps a token of 'th' to a slightly different embedding, it is reasonable to claim that the tokenizer is *misrepresenting* its input (though some philosophers may disagree with this analysis). We also find this in simple language models in which embedding layers are simple affine transformations which do not carve up the activation space.[5]

---

[3] There are theoretically interesting questions we can ask about the nature of these clusters, for example, whether clusters are all convex or how they can be identified. Shea, 2023 uses the example of clusters identified with T-SNE which is a non-linear dimensionality reduction algorithm. It is an interesting question whether clusters must correspond to linearly connected convex regions within a model or whether they can correspond to curved submanifolds. As with all theories of representation, there may be challenges arising from pre-training and model-stitching. If the regions are a product of downstream computations and these are changed, i.e. early layers are plugged into a different model, then these vehicles will change. If we get different contents when a pre-trained model has additional layers added, how do we explain the success of pre-training?

[4] The example networks the authors of these papers discuss use sigmoid and tanh activation functions which are more conducive to the polytope account than networks with ReLUs. Azhar is explicit about the focus on 'feedforward classificatory neural networks' (Azhar, 2016: 697). Networks using softmax at the final layer can be understood as outputting probability distributions and even if performing classification tasks, it is a further question how the credences should be interpreted, that is, whether the credence is part of the content of the representation. More recently, Park et al. have applied the polytope approach to categorical representation LLMs: "The *polytope representation* of a categorical concept $W = \{w_0, \ldots, w_{k-1}\}$ is the convex hull of the vector representations of the elements of the concept" (Park et al. 2024).

[5] For example, the first layer of Word2vec is linear (in effect, a one-hot encoding vector selects a vector from the column space of the embedding layer). This selection function is not indifferent to slight differences between vectors, instead, its whole function is to select exactly one of the stored column vectors for projection to the next layer. In this case, it makes sense to identify the vehicle of representation with the word embedding vector itself and not the region in which the vector sits. This is, after all, the point of generating word embeddings.

Second, the residual stream of a transformer does not perform non-linear operations on embeddings. Rather, it is a place where the outputs of attention heads are added to existing embeddings. This process of incremental addition can be understood as refining the model's prediction but, because the process is additive and non-linear, it does not itself provide a means for clustering the activation space into regions and while there is evidence that the final layer of an LLM clusters contents into polytopes (Park et al. 2024), it is still open whether this is the best way to view contents at earlier layers in the network. The cluster approach may also not be appropriate for attention heads themselves where weights are stored as scalar values. Consider the attention weights QK^T in an attention head:

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

A standard interpretation of these weights is that the scalar values represent the relatedness (in some sense) between embeddings. This may be viewed as a 'non-content specific computation' (Shea, 2023) but the scaled dot-product values themselves may still be understood as vehicles that represent the relatedness of embeddings (according some attention-head specific feature). In this case, the scaled dot products are scalar values and so not regions of the embedding space, and yet it may nonetheless be useful to view them as vehicles of content. The information stored in these scalars is used to scale value embeddings (in the V matrix) and so plays a causal role in downstream operations. Furthermore, these scalar values correspond to real features of the text being processed. A variety of studies have found that attention weights may represent whether a token is the subject or object of a sentence, when co-reference occurs, or if the same topic is being discussed (Olsson et al. 2022, Htut et al. 2019, Ravishankar et al. 2021). The contents of attention heads are a core focus of work on mechanistic interpretability (Sharkey et al. 2025).

In the case of the transformer architecture, there appear to be several distinct vehicles of representation. Individual token embeddings have contents determined by the tokeniser, attention heads represent tokens as activations while key-query matrices represent the relations between these tokens as scalar values. Each of these provides a case of a syntactically individuable unit playing a causal role in computations that can be assigned a unique content explaining its role in that computation. Nevertheless, at feed forward layers, it's likely that clusters of embeddings are the right level of description, in other words, the set of vehicles which map to contents are regions within the activation space of the network.

Rejecting the polytope hypothesis does not mean that we can't perform cross-model comparisons. When we treat clusters as the vehicles of content, the reason a new sample belongs to a cluster is that it has 'some property in virtue of which they fall into existing hidden layer clusters' (Shea, 2007: 255). In this case, clusters are vehicles of content but individual inputs are mapped to certain clusters because of their individual, content-determining properties. To use the terminology from earlier, signal representation carries content by virtue of model representation. After all, it would be circular to say that an input bears some content because it occupies some region of the activation space and also that the region forms a region because all the points in it have the same content. What is needed is an account of representation that explains why particular regions within an activation space have the content they do. Fortunately, recent work in mechanistic interpretability is beginning to indicate how this can be the case and in section 5, I'll give a quick tour of some representation hypotheses that have been formulated

about LLMs. In the next section though, I will argue that LLMs utilise multiple formats of representation.

4. Formats of Representation

Talk about representational formats concerns the structure of a system of representation. This paper will draw on both the structure on S and the definition of the interpretation function to individuate formats of representation. The set of relations R_1,R_2R_3 …R_n on S gives the *structure* of the system of representation. This can range from the one-dimensional structure of a mercury thermometer, the two-dimensional structure of a map, the five-dimensional structure of the LAB colour space, to the n-dimensional structures within the latent spaces of deep neural networks. Furthermore, the relations on S can be characterised by different properties; they may be partial or total, and they can be ordinal, metric, or dense etc.[6] To clarify things, we need some definitions:

> **Isomorphism (first-order):** For two structured sets, X, Y, a function f(X) = Y is an isomorphism iff it is a bijection and for R on X and R' on Y, $fR(x_1, x_2)$ if and only if R'($f(y_1), f(y_2)$). If we drop the bijection requirement, f is a homomorphism.

Following recent work, I will call a system whose representation relation is a first order isomorphism to its domain an *analogue* system of representation. To be clear, there will be countless first (and second) order isomorphisms between a system and its domain. We are concerned here with some already identified representation relation. When a system is analogue, vehicles that are similar, i.e. closer according to relations R on the domain, map onto contents that are similar, i.e. closer according to relations R' on the domain.

> **Isomorphism (second-order):** For two structured sets, X, Y, a function f(X) = Y is a second-order isomorphism iff it is a bijection from the relations R…R^n on X to the relations R'…. R'^{n} on Y. If we drop the bijection requirement, f is a second-order homomorphism.[7]

A system's representation relation may be a first-order isomorphism without being a second-order isomorphism. In other words, a system can have an analogue format (first-order isomorphism) without performing structural representation (second-order isomorphism). In either case though, we will typically want to weaken the claim that there is an isomorphism between the system of representation and the domain it represents. Whether a complete isomorphism holds is an all-or-notion property whereas many systems exhibit may be only *partially isomorphic* to the domain they represent.

---

[6] A structure is metric if it can support a notion of distance satisfying usual constraints (e.g. the triangle inequality) whereas it is dense if, for every pair of vehicles v_m v_n, R$v_m v_n$ there is some vehicle v_{m+1} between them. A set can have a metric without being dense (e.g. graphs) and dense without supporting a metric. A set can also be dense without being continuous, e.g. the system for representing the Rationals $\mathbb{Q}$ (Maley, 2011).

[7] Second-order isomorphism is defined here as a mapping from relations to relations. This is the definition in Shepard & Chipman, (1970) and underlying the idea of 'structural correspondence' (Shea, 2018). It may be strengthened though by the requirement that this mapping preserve higher level properties such as the order or distances between relations. As with first-order isomorphism, the choice of constraints will depend on the set of properties one decides to include when specifying the system of representation.

Format	Semantic Mapping Function	S-Relations
Nominal	Arbitrary	N/A
Analogue-ordinal	First-order isomorphism	Ordinal
Analogue-metric	First-order isomorphism	Metric
Analogue-dense	First-order isomorphism	Dense
Structural-ordinal	Second-order isomorphism	Ordinal
Structural-metric	Second-order isomorphism	Metric
Structural-dense	Second-order isomorphism	Dense

Systems of Representation

> **Partial isomorphism**: a function *f* from two structured sets is a partial isomorphism iff it *f* a partial function taking elements from V and mapping them onto elements of W satisfying the isomorphism conditions.[8]

While analogue representation involves a (partial) first-order isomorphism from vehicles in a system to contents in a domain, structural representation involves a second-order mapping between the relations between these vehicles and relations between those contents.

> **Structural Representation:** "A complex representation in which a relation on representational vehicles v_1,… v_n represents a relation on the entities represented by v_1, ..., v_n" (Shea, 2018, Shagrir, 2012, Swoyer, 1991).[9]

In other words, systems perform structural representation when their structure represents the structure of the domain of representation. There is evidence of nominal, analogue and structural formats within natural language processing.

The simplest format of representation is a *nominal sign system*. In a nominal sign system, the mapping *f* is arbitrary in that it need not respect any structures within the domain of vehicles or contents. The paradigm example of such a system is the North Church lantern code by which the sexton Robert Newman signalled to Paul Revere that the British army was approaching

[8] A partial isomorphism is distinct from homomorphism which is defined over the whole system. Lee el al. can speak of analogue mirroring as a gradient notion determined by the ratio between the cardinality of an isomorphic subsystem and the overall system. For example, in the original Mikolov paper, around 39% of vector offsets were correct, giving a mirroring-score of .39 (Mikolov et al., 2013, Lee et al. 2023). This gives us a measure of 'approximate isomorphism'(Shea, 2014). The task of rendering isomorphism a gradient notion has generated fascinating work both in philosophy and machine learning. Søgaard et al. 2018, for example, propose *eigenvector similarity* as a measure (Vulić et al. 2020, and a similar idea in Gutzen et al. 2023), Gromov-Hausdorff distance, and relational similarity have also been proposed. Unlike Lee et al, 2023, these metrics are more directly concerned with the structure of the embedding space rather than the cardinality of a strictly isomorphic subsystem. I don't assume that the concept of a format of representation can be wholly exhausted by the structure of the system (see Coelho Mollo & Vernazzani, 2024 for a distinction between inner and outer constraints).

[9] As there is considerable overlap between discussion of structural representation and iconic representation (Carey, 2009, Burge, 2018, Block, 2023). This literature focusses on the iconicity of perceptual representation. "An emerging consensus that the best way to understand representation in the context of cognitive explanation is structural" (Piccinini, 2018). Millikan's original account of intentional icons may also fit in here though she is concerned with mappings from sets of transformations on the vehicles and transformations on the domain represented (Millikan, 1984).

('one if by land, two if by sea') (Godfrey-Smith, 2017, Shea, 2023). In this system, all that matters is that the signals are distinguishable and neither the relations between the parts of the representation nor relations between properties shared by both signs have any semantic significance.[10] Newman could have chosen to invert the encoding and it would have made no difference. Further examples of nominal systems are the alarm calls of vervet monkeys (Camp, 2022) and the symbol-sound correspondence of the Latin alphabet. Many of the traditional concerns about the 'symbol grounding problem', e.g. Chinese rooms, Blockhead etc., seem to arise from the arbitrariness of nominal sign systems (Harnad, 1990).

Tokenisation is the most obvious case of a nominal format in a language model. Tokenisation is the mapping of input text to a numerical representation that facilitates computation. The clearest example of this is when a one-hot encoding represents a word in the model's vocabulary as a vector of 0s with a single dimension set to 1, e.g. 'cat' = [0,0,0,0,….0,1,0,0….].[11] This encoding is a bijection, each vocabulary token has an embedding and each embedding corresponds to exactly one token, but this bijection does not *necessarily* depend on a particular structure of the dictionary, you can sort the dictionary however you want. The standard is to use more efficient tokenisations (e.g. Byte-pair or WordPiece) in which the length of the code corresponds inversely with the frequency of the content. In these cases, there may be an unexploited partial isomorphism between the length of the encoding and frequency of the token encoded (matters are complicated by the reuse of tokens as substrings of other tokens). The partial isomorphism is unexploited as there are no computations that compute the length of the token encodings and respond based upon this and so it would be implausible to argue that the tokenisation *represents* the word's frequency. Natural languages exhibit a similar property in which more frequently use words like 'a' and 'the' are typically shorter than less common words like 'equidistant' and 'declination' but this doesn't entail that languages *represent* the frequency of a word by its length even if this fact can be recovered from a corpus (Zipf, 1935).

As mentioned, a system is analogue if its interpretation function is a partial isomorphism and it may be more or less analogue depending on the extent to which it is isomorphic (Lee et al 2023, Shea, 2023, and Godfrey-Smith, 2017, call this 'structural organisation'). One of the most obvious forms this takes is when similar vehicles correspond to similar contents. For example, the spatial proximity of points on a thermometer or notes on a stave corresponds to the thermal proximity of temperatures and pitches with closer frequencies. To consider an example from natural language processing, Word2Vec was originally developed to produce embeddings that mirrored their contents, where cosine similarity would correspond to semantic similarity

[10] LeCun calls such systems 'symbolic'. Kulviki defends a conception of analogicity according to which analogue systems support open ended search of content across levels (Kulviki, 2015). According to this view, properties of analogue systems (continuity, similarity structures etc.) contribute to those systems being analogue because they facilitate this. This corresponds well to properties like the translation invariance of convolutions in CNNs. The pattern-matching like properties of CNNs implement a kind of open-ended search.

[11] If the vocabulary is sorted prior to the generation of one-hot encodings, then the unit dimension may carry information about the structure of this datatype. A reasonable way to generate one-hot encodings is to generate a vocabulary data structure, generate a list of zero vectors, and then use the index of a vocabulary item within the vocabulary to select a dimension to be set to 1. While this does transfer the structure of the vocabulary onto the one-hot encoding, this information cannot be reconstructed at later layers. Byte-pair encodings are more complicated (e.g. the letter 'a', for example, may not be encoded as a single value).

(Mikolov, et al. 2013).[12] In other words, Word2Vec was developed to produce an analogue representation of semantic similarity. Similarity is always similarity with respect to an aspect, a given relation R', and it has been a non-trivial challenge to work out the various feature subspaces that give rise to these general similarity relations (Mikolov et al., 2013, Grand et al., 2022).

However, these relations are not *exploited* in downstream computations. The Word2Vec algorithm produces embeddings in an analogue format but the relations between these embeddings are not exploited by the algorithm. The algorithm may compute the distance between word embeddings and separate context vectors, but it doesn't directly compute the distance between word embeddings. No measure of the distance between embeddings is taken by the model, the value of the distance is neither the output of a computation nor the input to a computation, and so it is not accurate to say the system represents these relations (i.e. that the model 'represents' gender as a relation between embeddings). To make this claim is to conflate the information we can extract from the model with the information the system represents. Even if structural information is not exploited by the system, analogue organisation has advantages (Shea, 2023). It is efficient to implement, generalise to new cases, and is error-tolerant. The analogue organisation of Word2Vec can emerge as a byproduct of simple learning tasks (Goldberg & Levy, 2014). Analogue systems can also generalise, a point long discussed in machine learning: "The ability to capture the similarity between concepts by the similarity of their representations, and the consequent ability to generalise new information in sensible ways" (Hinton, 1986: 3). If a vehicle is subject to some change due to error, it will still correspond to a content that is *similar* to what it was supposed to represent thus minimising the effects of the error. This fact, however, depends on the landscape of the loss function and thus indirectly on the structure of the domain being represented.

While the simple word embeddings produced by Word2Vec have an analogue correspondence with word senses, structural representation requires an operation like attention (in the following, I will focus on scaled dot product attention).[13] Attention enables the embeddings in sequential data to inform downstream representations of the other embeddings in the sequence. The attention mechanism uses three learned linear weight matrices to create three copies of an individual token embedding called the 'query', 'key', and 'value' vectors. Each of these vectors 'stands for' the original embedding but is put to slightly different use by the attention head. For each embedding in the sequence, an attention head computes the difference, i.e. the dot product, between that embedding's query vector and the key vectors of the other embeddings in the sequence. This gives a measure of how 'related' the embeddings in a sequence are to each other (Vaswani et al. 2017). The dot product, having been run through a softmax activation, is then used to scale the value embedding thus making sure this information is carried forward. In other words, an attention head ensures that the relation between two representational vehicles is

---

12 "We use recently proposed techniques for measuring the quality of the resulting vector representations, with the expectation that not only will similar words tend to be close to each other, but that words can have multiple degrees of similarity" (Mikolov et al. 2013). This idea assumes that what is represented by embeddings is word meanings, *senses*, and not what those meanings pick out *referents*. There's no guarantee that sense similarity and reference similarity align; Hesperus and Phosphorus have identical referents but dissimilar senses.

13 It is an interesting question whether adding an input vector to a hidden state in a RNN counts as computation over relations between vehicles by proponents of structural representation. Likewise, GLoVE computes the dot product of embeddings (Pennington, 2014). I suspect both cases would count as structural representation but illustrate that the category may require some more granular analysis. I present the attention mechanism here as it is an unambiguous example.

exploited directly in a downstream computation thereby enabling any content implicit in the analogue structure of a language model to become exploitable.[14] In other words, attention enables structural representation as it allows relations between embeddings to play a role in computations.

This may not be the only case of structural representation in a transformer. There have also been several hypotheses formulated about how *structure* is represented.

> **Graph Embedding Hypothesis:** Graphs are embedded in vector spaces such that distances and norms within the embedding space approximate the geometry of the graph (Hamilton et al, 2017).

We see a special form of this in the:

> **Syntax Embedding Hypothesis:** Syntactic structures are embedded in subspaces where euclidean distance corresponds to the distance between embedding's within a parse tree (Hewitt & Manning, 2019).[15]

The idea here is that the distance between two token embeddings corresponds to the distance between the words in a parse tree the embeddings encode. This has made it possible to extract dependent parse trees for sentences from early layers of transformer models. It is currently less clear though how this information is exploited and the extent to which syntactic information may be separable from semantic information (Agarwal et al. 2025). However, we have reason to believe that the distances between embeddings are exploited to capture both semantic and syntactic relations in the data. Further research is required on the relation of syntax to semantics in LLMs.

In this section, I have claimed that a transformer model uses at least three distinct formats of representation. At the tokenisation stage, embeddings are nominal representations of the tokens they represent. The residual stream has an analogue structure according to which similar embeddings correspond to similar words (along various dimensions). When this structure is exploited in an attention head, the relations between embeddings may be used to correspond to the relations between entities in the world. However, these distinctions only concern relations between signal representations (i.e. activations or clusters of activations) rather than what and how those signal representations represent. The next section will give an overview of one prominent account of how this happens.

5. Linear Representation and Superposition

---

[14] The scaling here is carried out with the softmax function. While the softmax is sometimes said to output a probability distribution, it is clear that in this case, it serves a purely norming function (and can actually be replaced and improved with other functions, Saratchandran et al. 2024). For simplicity I will speak as though operations are operations on individual vector embeddings rather than matrices of embeddings with the understanding that multiplication of these matrices really is just the computation of the dot products of the individual vectors.

[15] While the authors avoid unwarranted conjectures, it's worth noting that a representation of composable, hierarchical structure within an embedding space could be evidence of language-like structures within the model, providing an implementation of something like a Language of Thought (Fodor, 1989, Quilty-Dunn et al. 2023).

In recent years, several hypotheses have been formulated about how features are encoded in deep neural networks. One of the most influential is the idea that features are encoded as directions in the latent space of a network (Elhage et al. 2024). These hypotheses can be seen as generalisations from the discovery that properties like gender appear to be encoded as vectors in simple word embeddings, famously, v(uncle) - v(man) + v(woman) ≈ v(aunt), (Mikolov et al., 2013). The claim that features are encoded as linear directions in a model is called the Linear Representation Hypothesis, and it has taken several forms.

> **The Weak Linear Representation Hypothesis (wLRH):** Neural networks learn to encode 'high level' features as linear directions in activation space (Elhage et al. 2022, Park et al. 2024).[16]
>
> **The Strong Linear Representation Hypothesis (sLRH):** All features are encoded as linear directions

The strong-weak distinction concerns the scope of the hypothesis. While it is not yet clear what it means for a feature to be 'high level', typical examples are grammatical or linguistic properties such as whether a word is singular or plural, in English or French, an adjective or superlative, as well as general categorial relations such as the relationship between a country and capital, thing and colour, whether the sentiment is positive or negative (Tigges et al. 2023, Park et al. 2024).[17] The weak linear representation hypothesis is supported by counterfactual intervention studies in which features are altered by shifting vectors in different directions within a network (Hao & Linzen, 2023, Park et al. 2023).

Systems of linear representation can be classified with respect to several properties. Two important properties are whether there is a *privileged basis* and whether vectors are *orthogonal.* A system has a privileged basis if the coordinates of the system have a unique interpretation. One way to think about this is in terms of the invariance properties of the system, e.g. if a system has a privileged basis, you can't just rotate it and get the same result. The concept of translation invariance is used in several ways in the literature to include geometric translations ('shift invariance', LeCun, 1989) as well as rotation invariance, scaling invariance (Buckner, 2024). In a transformer model, token embeddings, attention patterns and MLP layers have privileged bases while keys, queries, values, and the residual stream don't. A more important consideration is whether the directions representing distinct features are orthogonal to each other or not. The concept of orthogonality is a generalisation of the notion of lines being perpendicular, i.e. at a 90° angle. On a map of two dimensional space, we might represent the east-west dimension along the left-right axis and the north-south feature on the top-bottom axis.

---

16 Some versions of the LRH take a core feature to be the claim that the composition of multiple concepts can be understood as the addition of their corresponding feature vectors (Olah & Jermyn, 2024, Sharkey et al. 2025). An earlier and stronger form of the hypothesis was that all concepts are encoded in one-dimensional vectors (Elhage et al. 2022). This has been rejected (Engles et al. 2024). Some formulations of the LRH add the further claim that the 'intensity' of a concept corresponds to the magnitude of its corresponding vector (Olah & Jermyn, 2024). This claim has two possible readings that are worth disentangling. One reading takes this as a claim about analogue representation according to which vectors of greater magnitude correspond to greater presence of the feature in a sample while according to a second reading, vector magnitude corresponds to the intensity of the model's confidence about the presence of a feature (this is a more Bayesian reading). Being *very* sure that a dog is small is not the same as being slightly sure that a dog is *very* small. We will set aside these issues in the following discussion.

17 One way the strong thesis can be refuted by the discovery of features than are not encoded linearly. There is considerable evidence for this (Engels et al. 2024, Csordás et al. 2024). For example, recent work suggests that numbers are encoded as helices (Kantamneni & Tegnmark, 2025).

Format	Function	Vehicle	Example
Nominal Representation	Arbitrary mapping	Activations/embeddings	Simple tokenisation
Analogue Format	First-order isomorphism	Activations/embeddings	Simple word embeddings (e.g. Word2Vec)
Structural Representation	Second-order isomorphism	Relations in residual stream, attention heads	Transformer LMs
Virtual Representation	Yes (but it's complicated)	Higher dimensions of the residual stream	Transformer LMs

Formats and Vehicles of Representation

These axes provide an interpretable basis for our encoding system. If this basis is orthogonal, moving left will only take one west, it won't take one north or south as well. In other words, features that are encoded orthogonally are semantically independent of each other.

A strictly orthogonal embedding space can only represent as many distinct features as its dimensionality.[18] This has given rise to the suggestion that the embedding spaces of transformer networks aren't strictly orthogonal but *almost orthogonal*. This, in turn, motivates a particular variant of the linear representation hypothesis that holds that features in a neural network are not just encoded as linear directions within the embedding space but that these directions are almost orthogonal to each other.

> **The Superposition Hypothesis**: Features are represented *almost orthogonally* in embedding spaces. As a result, an embedding space can represent many more features than dimensions by noisily simulating a higher dimensional embedding space. As a consequence, neurons in a DNN are polysemantic, i.e. they may correspond to multiple features (Elhage et al. 2022, Engels et al., 2024).

The advantage of using an almost orthogonal embedding space is that, instead of having n features orthogonally encoded in n-dimensions, an embedding space can represent e^n almost orthogonal features (Elhage et al, 2022). One way to view this is that an almost orthogonal space noisily simulates a higher dimensional, orthogonal embedding space — a kind of *virtual* representation. To get an intuition for this claim, it may help to imagine a drawing of a 3D map on a 2D sheet of paper (Figure 1). Angles which in a 'true map' would be 90° do not appear as 90° exactly in the 2D drawing. This is how the drawing depicts more dimensions than its medium strictly affords. The cost of this compression (or 'perspective') is that moving right within the image no longer corresponds exclusively to moving right in the world as it also takes one deeper into the third dimension. In the case of the

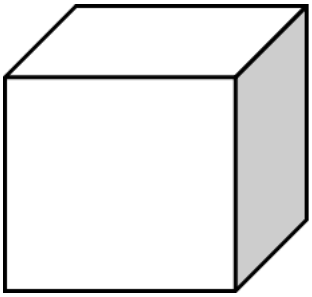

Fig 1. A 3D shape can be projected to 2D space at the expense of right angles. x-wards direction is also z-wards (i.e. superposition).

[18] In some cases it has been possible to find linearly separable orthonormal subspaces corresponding to high-level concepts like grammatical number (Hao & Linzen, 2023). A 'linearly separable orthonormal subspace' is a region of our embedding space with an orthogonal basis across which you could plot a line separating it into two sides, on one side, we have singular forms of nouns, on the other, their plural forms. You can change a nouns number value by shifting it across this line.

neural network, this simulation is noisy as features encoded by distinct dimensions may interact with each other and this noise manifests as neurons displaying polysemanticity — the phenomenon by which a single neuron encodes multiple features. In this way, a model can represent more features than its dimensionality. With virtual representations, vehicles in a system simulate vehicles in a higher dimensional system which map onto contents (Haugeland, 2012).

Granted that non-orthogonally encoded features semantically overlap to some extent, there is some evidence that orthogonality between vectors may correspond to semantically meaningful relations between the corresponding features represented, specifically the relationship of hierarchical inclusion (Park et al. 2024) This work also proposes that vectors may be the appropriate vehicles for the representation of binary concepts (features like 'is an animal') while polytopes are the vehicles of subordinate contrastive categories (e.g. features like mammal, bird) (Park et al. 2024). For example, the feature ANIMAL stands in a hierarchical relationship to the features CAT and DOG. Assuming that neither a CAT nor a DOG is more of an ANIMAL, we should expect these corresponding vectors to have the same dot product and we should not want a modification of the vector for ANIMAL to alter the probabilities for CAT or DOG features in a computation. If this is the case, and these relationships are in some way exploited, this may be another case of structural representation, where relations between vehicles (i.e. the orthogonality of embeddings) correspond to relations between properties in the world.

Superposition raises concerns for mechanistic interpretability: "Such tangled representations undermine our ability to decompose networks into independently meaningful and composable components" (Gurnee et al. 2023) but it is important to note that polysemanticity does not entail that the mapping from vehicles to contents, *f,* is not a function. What it requires is that we rethink what a content is. If directions correspond to multiple features then the function *f* may be best thought of as a mapping to distributions over contents. This view of representation aligns with the idea that LLMs perform *iterative inference* according to which each layer of the residual stream of a transformer performs an update on the prediction of the next token in the data (Jastrzębski et al. 2017, Belrose, 2023). To be clear, the claim isn't that polysemanticity supports iterative inference but that the view that the content of representation is a probability distribution aligns with it.

It is possible that some of the evidential support for the LRH is a result of the use of linear probes when searching for representations.[19] By linear probe, I mean one which either linearly separates an embedding space or learns a linear transformation to some subspace. Typical causal analyses of neural networks use interventions on linear representations (Csordás et al. 2024). There is a good reason for the focus on linear methods. Non-linear probes are extremely powerful, that is, they can identify 'find' any pattern in the data and so are more likely to provide false-positives in the search for internal representations. The constraint of linearity serves to prevent this. As one recent paper puts it, "the field is rapidly converging around methods that can only find solutions consistent with the LRH" (Csordás et al. 2024). To take the example from above, Hewitt & Manning, discovered latent syntactic structure within the early layers of BERT by training a network to learn a linear transformation from the embedding space

[19] Similar questions can be raised about the similarity metrics used in this research. Research into the LRH relies on methods and metrics for comparing different embedding spaces and there are a range to choose from; CKA, Eigenvector similarity, Linear Procrustes etc., each involving trade-offs, e.g. CKA is invariant to orthogonal transforms but can destroy spatial structure.

of a network to a subspace in which the distances between embedding vectors corresponds to distances within a parse tree. This allowed the authors to reconstruct dependency parse trees for input sentences based on the representation of syntactic structure internalised within the model. However, researchers have shown that this can be improved upon by rejecting the linearity constraint and availing of non-linear methods (White et al. 2021). This raises the question of whether this work has captured an essential fact about the representation of syntactic information in large language models or whether the structure uncovered is an artefact of the probing method.[20] This concern is not confined to ML and has been raised in neuroscience as well. Since transformations are typically non-linear, it has been objected that linear probes may not give us a 'faithful representation' of downstream processes (Ivanova et al. 2021). The point of this digression is not to dismiss the LRH but to acknowledge that everything that has been said in this section is provisional.

Before finishing, it is worth highlighting one final hypothesis about representation.

> **The Platonic Representation Hypothesis (PRH)**: Neural networks trained with different objectives on different data and modalities are converging to a shared statistical model of reality in their representation spaces (Bansal et al, 2021, Huh, Cheung et al. 2024).

The Platonic Representation Hypothesis is a version of the idea that good networks learn similar representations.[21] Evidence for the PRH comes from several sources. Models that have been trained on different datasets can be *stitched* together without much loss in their accuracy (Lenc & Vedaldi 2015, Moschella et al., 2023). For example, the early layers of a model trained on ImageNet can be combined with the later layers of a model trained on Places-365 by means of a learned stitching layer to produce a third, high-performing model. The success of stitched models is taken to indicate the *compatibility of representations* up to a given layer. Importantly, stitching can show the compatibility of representations across models trained with different initializations, subsets of the same dataset, and training tasks.[22] If the PRH is correct, then this will likely raise more issues for our understanding of the formats of representation in ANNs and will help shed light on the ways in which the structure of data constrains and incentivises the emergence of the systems that represent it.

---

[20] The fact that we can formulate these questions may be a reason not to relativise the concept of representation to a probe. Some researchers have argued that. "We should always think of the representation-and-its-probe *together*...This may sound like a radical suggestion: to stop worrying and love the fact that patterns will *always* depend on our means of finding them. But it is in fact a commitment we most likely already have once we have accepted the impossibility of direct access to things-in-themselves" (Cao, 2022). Cao acknowledges that controls should be run to distinguish artefacts from real contents but it is the notion of dependence I am resisting here. Realism about patterns, including the pattern of linearity, assumes truth conditions are independent of our particular probes.

[21] This is also described as the Anna Karenina scenario; c.f. "All happy families are alike; each unhappy family is unhappy in its own way". It is also related to the Isomorphic Embedding Hypothesis: All sufficiently resourced embedding spaces are isomorphic (Vulíc, Rudder & Søgaard, 2020). If true, this would be a remarkable discovery about the nature of representation in artificial neural networks. Behind much of this work likes a final major hypothesis, the manifold hypothesis which holds that real-world data lies on a low-dimensional non-linear manifold embedded in a high-dimensional space (Fefferman et al., 2013, Ghojogh et al. 2023).

[22] Stitching also counts against a skeptical, Kuhnian concern sometimes levelled at DNNs implementing SGD. According to this concern, different local minimal correspond to different and incommensurable conceptualizations of the same data — to inhabit a local minimum is to inhabit a unique conceptual scheme. Work on *mode connectivity* and stitching gives us reasons to reject this skeptical view. Mode connectivity shows that there are low-loss paths between minima (indicating commensurability) (Draxler et al. 2018). Stitching research indicates that alternative minima can be directly stitched with linear layers (translatability between alternative schemes).

6. General Remarks

This paper has made two claims. The first is that we should recognise multiple vehicles of representation in transformer language models. The outputs of a tokeniser are discrete embeddings, attention weights are scalar values, and features in a residual stream are encoded as directions within the embeddings space. In each case, these vehicles are exploited in computations and can be assigned contents by researchers. The second claim is that we should recognise multiple formats of representation. The process of tokenisation generates a nominal representation of individual tokens. There may be unexploited analogue correspondences between embeddings and features at later layers and attention heads can support structural representation by enabling the relations between embeddings to be exploited in downstream computations. The phenomenon of superposition may give rise to a further representational format, though more evidence is needed to say anything certain about this. Much of what has been claimed here is dependent on the accuracy of current probing methods and the discussion has skimmed over ignored many important issues. What is clear is that the science of representation is beginning to find its feet as a field that allows for precise theorisation and novel experimental paradigms.

References

Agarwal, A., Jian, J., Manning, C. D., & Murty, S. (2025). Mechanisms vs. Outcomes: Probing for Syntax Fails to Explain Performance on Targeted Syntactic Evaluations. *arXiv preprint arXiv:2506.16678*.

Artiga, M. (2023). Understanding Structural Representations. *The British Society for the Philosophy of Science* https://doi.org/10.1086/728714

Azhar, F. (2016). Polytopes as vehicles of informational content in feedforward neural networks. Philosophical Psychology. https://doi.org/10.1080/09515089.2016.1142070

Bansal, Y., Nakkiran, P., & Barak, B. (2021). Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, *34*, 225-236.

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., ... & Steinhardt, J. (2023). Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Bengio, Y., Goodfellow, I., & Courville, A. (2017). Deep learning (Vol. 1). Cambridge, MA, USA: MIT press.

Block, N. (2023). The border between seeing and thinking. Oxford University Press.

Brandom, R. (1994). Making It Explicit: Reasoning, Representing, and Discursive Commitment. Harvard University Press.

Brody, S., Alon, U., & Yahav, E. (2023). On the Expressivity Role of LayerNorm in Transformers' Attention. *arXiv preprint arXiv:2305.02582*.

Buckner, C. J. (2024). From deep learning to rational machines: What the history of philosophy can teach us about the future of artificial intelligence. Oxford University Press.

Burge, T. (2018). Iconic representation: Maps, pictures, and perception. The map and the territory: Exploring the foundations of science, thought and reality, 79-100.

Camp, E. (2007). Thinking with maps. Philosophical perspectives, 21, 145-182.

Camp, E., Grzankowski, A., & Montague, M. (2018). Why maps are not propositional. Non-propositional intentionality, 19-45.

Carey, S. (2009). The origin of concepts. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195367638.001.0001

Cao, R. (2022). Putting representations to use. Synthese, 200(2), 151.

Chang, A., Poeppel, D., & Teng, X. (2025). Temporally dissociable neural representations of pitch height and chroma. *Journal of Neuroscience*.

Churchland, P. M. (1998). Conceptual Similarity Across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered. Journal of Philosophy, 95(1), 5-32.

Clark, A. (2001) *Mindware*. Oxford: O.U.P.

Csordás, R., Potts, C., Manning, C. D., & Geiger, A. (2024). Recurrent neural networks learn to store and generate sequences using non-linear representations. arXiv preprint arXiv:2408.1

Dennett, D. C. (1991). Real Patterns. *The Journal of Philosophy*, *88*(1), 27-51.

Draxler, F., Veschgini, K., Salmhofer, M., & Hamprecht, F. (2018). Essentially no barriers in neural network energy landscape. In *International conference on machine learning* (pp. 1309-1318). PMLR.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... & Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*, *1*(1), 12.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., ... & Olah, C. (2022). Toy models of superposition. arXiv preprint arXiv:2209.10652.

Eliasmith, C. (2003). Moving beyond metaphors: understanding the mind for what it is. Journal of Philosophy, 493 – 520.

Engels, J., Michaud, E. J., Liao, I., Gurnee, W., & Tegmark, M. (2024). Not all language model features are linear. arXiv preprint arXiv:2405.14860.

Facchin, M. (2024). Maps, simulations, spaces and dynamics: On distinguishing types of structural representations. *Erkenntnis*, 1-22. https://doi.org/10.1007/s10670-024-00831-6

Fefferman, C., Mitter, S., & Narayanan, H. (2016). Testing the manifold hypothesis. *Journal of the American Mathematical Society*, *29*(4), 983-1049.

Ferrando, J., Sarti, G., Bisazza, A., & Costa-jussà, M. R. (2024). A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*.

Fodor, J. A. (1989). Why there still has to be a language of thought. In Computers, brains and minds: Essays in cognitive science (pp. 23-46). Dordrecht: Springer Netherlands.

Fodor, J. (2000). *The Mind Doesn't Work That Way.* Cambridge, MA, MIT Press.

Fodor, J. & E. Lepore (1999). All at sea in semantic space: Churchland on meaning similarity. Journal of Philosophy, 96(8), 381-403.

Gallistel, C. R. (1994). Foraging for brain stimulation: toward a neurobiology of computation. Cognition, 50(1-3), 151-170.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34, 9574-9586.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., ... & Icard, T. (2025). Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, *26*(83), 1-64.

Geva, M., Schuster, R., Berant, J., & Levy, O. (2020). Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Geva, M., Caciularu, A., Wang, K. R., & Goldberg, Y. (2022). Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*.

Godfrey-Smith, P. (2017). Senders, receivers, and symbolic artifacts. *Biological Theory*, 12(4), 275–286.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour,* 6(7), 975-987.

Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., & Bertsimas, D. (2023). Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*.

Gurnee, W., & Tegmark, M. (2023). Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*.

Harding, J. (2023). Operationalising Representation in Natural Language Processing. *British Journal for the Philosophy of Science*. https://arxiv.org/abs/2306.08193

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335-346.

Hao, S., & Linzen, T. (2023). Verb conjugation in transformers is determined by linear encodings of subject number. arXiv preprint arXiv:2310.15151.

Hewitt, J., & Manning, C. D. (2019, June). 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 Pape*rs) (pp. 4129-4138).

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society* (Vol. 1, p. 12).

Htut, P. M., Phang, J., Bordia, S., & Bowman, S. R. (2019). Do attention heads in BERT track syntactic dependencies?. *arXiv preprint arXiv:1911.12246*.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. arXiv preprint arXiv:2405.07987.

Ivanova, A. A., Schrimpf, M., Anzellotti, S., Zaslavsky, N., Fedorenko, E., & Isik, L. (2021). Is it that simple? Linear mapping models in cognitive neuroscience. *bioRxiv*, 438248.

Jastrzębski, S., Arpit, D., Ballas, N., Verma, V., Che, T., & Bengio, Y. (2017). Residual connections encourage iterative inference. *arXiv preprint arXiv:1710.04773*.

Jermyn, A. S., Schiefer, N., & Hubinger, E. (2022). Engineering monosemanticity in toy models. *arXiv preprint arXiv:2211.09169.*

José Hanson, S., Caglar, L. R., & Hanson, C. (2020). Decoding Second Order Isomorphisms in the Brain: The case of colors and letters. *bioRxiv*, 2020-05.

Kantamneni, S., & Tegmark, M. (2025). Language Models Use Trigonometry to Do Addition. *arXiv preprint arXiv:2502.00873.*

Kulvicki, J. (2015). Analog representation and the parts principle. *Review of Philosophy and Psychology,* 6, 165-180.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation,* 1(4), 541-551.

Lee, A. Y., Myers, J., & Rabin, G. O. (2023). The structure of analog representation. *Noûs*, 57(1), 209-237.

Lenc, K., & Vedaldi, A. (2015). Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 991-999).

Maley, C. J. (2011). Analog and digital, continuous and discrete. *Philosophical Studies*, *155*, 117-131.

Martínez, M., & Artiga, M. (2023). Neural oscillations as representations. *The British Journal for the Philosophy of Science,* 74(3), 619-648.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Millikan, R. G. (1987). *Language, thought, and other biological categories: New foundations for realism.* MIT press.

Moschella, L., Maiorca, V., Fumero, M., Norelli, A., Locatello, F., & Rodolà, E. (2022). Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430.*

Mu, J., Bhat, S., & Viswanath, P. (2016). Geometry of polysemy. *arXiv preprint arXiv:1610.07569*.

Narayanan, H., & Mitter, S. (2010). Sample complexity of testing the manifold hypothesis. *Advances in neural information processing systems*, *23*.

Haugeland, J. (2012). Semantic engines: An introduction to mind design. In *Machine Intelligence* (pp. 29-63). Routledge.

Olah, C., & Jermyn, A. (2024). What is a linear representation? what is a multidimensional feature. *Transformer Circuits Thread*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. (2022) In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*

Park, K., Choe, Y. J., & Veitch, V. (2023). The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658.*

Park, K., Choe, Y. J., Jiang, Y., & Veitch, V. (2024). The geometry of categorical and hierarchical concepts in large language models. *arXiv preprint arXiv:2406.01506*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Pulvermüller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in cognitive sciences*, *17*(9), 458-470.

Piccinini, G. (2018). Computation and representation in cognitive neuroscience. Minds and Machines, 28(1), 1-6.

Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. *Brain and language,* 127(1), 86-103.

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 46, e261.

Quiroga, R. Quian, et al. "Invariant visual representation by single neurons in the human brain." *Nature* 435.7045 (2005): 1102-1107.

Radford, A., Józefowicz, R., Sutskever, I.: Learning to generate reviews and discovering sentiment (2017), http://arxiv.org/abs/1704.01444

Ravishankar, V., Kulmizev, A., Abdou, M., Søgaard, A., & Nivre, J. (2021). Attention can reflect syntactic structure (if you let it). *arXiv preprint arXiv:2101.10927*.

Saratchandran, H., Ramasinghe, S., Shevchenko, V., Long, A., & Lucey, S. (2024). A sampling theory perspective on activations for implicit neural representations. *arXiv preprint arXiv:2402.05427*.

Shagrir, O. (2010). Brains as analog-model computers. Studies In History and Philosophy of Science Part A, 41(3), 271-279.

Sharkey, Lee, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill et al. (2025). Open Problems in Mechanistic Interpretability. *arXiv preprint arXiv:2501.16496*

Shea, N. (2007). Content and its vehicles in connectionist systems. *Mind & Language*, *22*(3), 246-269.

Shea, N. (2014). Exploitable isomorphism and structural representation. In *Proceedings of the Aristotelian Society* (Vol. 114, No. 2_pt_2, pp. 123-144). Oxford, UK: Oxford University Press.

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Shea, N. (2023). Moving beyond content-specific computation in artificial neural networks. *Mind & Language*, *38*(1), 156-177.

Shea, N. (2023). Organized representations forming a computationally useful processing structure. *Synthese*, 202(6), 175.

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, 1(1), 1-17.

Skyrms, B. (2010). Signals: Evolution, learning, and information. Oxford University Press.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, 87, 449-508.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., ... & Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread.*

Hollinsworth, O., Tigges, C., Geiger, A., & Nanda, N. (2024, November). Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP* (pp. 58-87).

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30*

Vulić, I., Ruder, S., & Søgaard, A. (2020). Are all good word vector spaces isomorphic?. *arXiv preprint arXiv:2004.04070.*

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

White, J. C., Pimentel, T., Saphra, N., & Cotterell, R. (2021). A non-linear structural probe. *arXiv preprint arXiv:2105.10185*.