The Geometry of Categorical and Hierarchical Concepts in Large Language Models

Kiho Park [1] Yo Joong Choe [1] Yibo Jiang [1] Victor Veitch [1]

Abstract

Understanding how semantic meaning is encoded in the representation spaces of large language models is a fundamental problem in interpretability. In this paper, we study the two foundational questions in this area. First, how are categorical concepts, such as $\{\texttt{mammal}, \texttt{bird}, \texttt{reptile}, \texttt{fish}\}$, represented? Second, how are hierarchical relations between concepts encoded? For example, how is the fact that `dog` is a kind of `mammal` encoded? We show how to extend the linear representation hypothesis to answer these questions. We then find a remarkably simple structure: simple categorical concepts are represented as simplices, hierarchically related concepts are orthogonal in a sense we make precise, and (in consequence) complex concepts are represented as polytopes constructed from direct sums of simplices, reflecting the hierarchical structure. We validate the results on the Gemma large language model, estimating representations for 957 hierarchically related concepts using data from the WordNet hierarchy.

1. Introduction

This paper concerns how high-level semantic concepts are encoded in the representation spaces of large language models (LLMs). Understanding this is crucial for model interpretability and control. The ultimate aspiration is to monitor (and manipulate) the semantic behavior of LLMs (e.g., is the model's response truthful) by directly measuring (and editing) the internal vector representations of the model (e.g., Li et al., 2023a; Zou et al., 2023; Ghandeharioun et al., 2024). Achieving this requires understanding how the geometric structure of the representation spaces corresponds to the high-level semantic concepts that humans understand. In this paper, we are concerned with two fundamental questions in this direction:

1. How are categorical concepts represented? For example, what is the representation of the concept $\texttt{animal} = \{\texttt{mammal}, \texttt{bird}, \texttt{reptile}, \texttt{fish}\}$?
2. How are hierarchical relations between concepts represented? For example, what is the relationship between the representations of `animal`, `mammal`, `dog`, and `poodle`?

Our starting point is the *linear representation hypothesis*, the informal idea that high-level concepts are linearly encoded in the representation spaces of LLMs (e.g., Marks & Tegmark, 2023; Tigges et al., 2023; Gurnee & Tegmark, 2024). A main challenge for the linear representation hypothesis is that, in general, it's not clear what "linear" means, nor what constitutes a "high-level concept". Park et al. (2024) give a formalization in the limited setting of binary concepts that can be defined by counterfactual pairs of words. For example, the concept of $\texttt{male} \Rightarrow \texttt{female}$ is formalized using the counterfactual pairs $\{(\text{"man"}, \text{"woman"}), (\text{"king"}, \text{"queen"}), \dots\}$. They prove that such binary concepts have a well-defined linear representation as a direction in the representation space. They further connect semantic structure and representation geometry by showing that, under a suitably defined *causal inner product*, concepts that can be freely manipulated (e.g., $\texttt{male} \Rightarrow \texttt{female}$ and $\texttt{french} \Rightarrow \texttt{english}$) are represented by orthogonal directions.

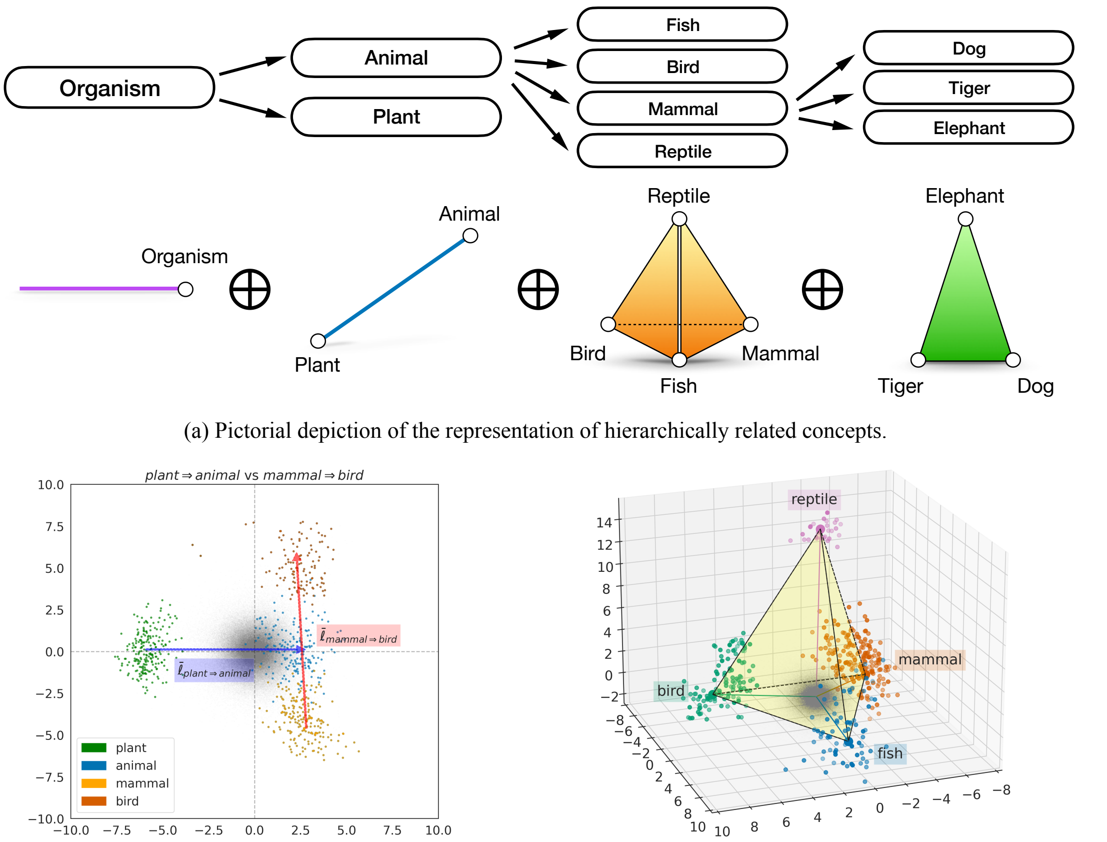
Our aim here is to extend this formalization beyond binary concepts represented as counterfactual word pairs. For example, the `animal` concept does not have a natural counterfactual definition, but such concepts are fundamental to human semantic understanding. Further, we aim to understand how the geometry of the representation space encodes semantic relationships between concepts that cannot be freely manipulated, such as `animal` and `mammal`.

To that end, we make the following contributions:

1. We show how to move from representations of binary concepts as *directions* to representations as *vectors*. That is, we show how to associate a magnitude to the representation of a binary concept. This allows us to study semantic composition using simple vector

[1] University of Chicago, Illinois, USA.

Work presented at TF2M workshop at ICML 2024, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(a) Pictorial depiction of the representation of hierarchically related concepts.

(b) Hierarchy is encoded as orthogonality in Gemma.

(c) Categorical concepts are represented as simplices in Gemma.

Figure 1. In large language models, categorical concepts are represented as simplices in the representation space. Further, hierarchically related concepts (such as animal and mammal ⇒ bird) live in orthogonal subspaces. The top panel illustrates the structure, the bottom panels show the measured representation structure in the Gemma LLM. See Appendix B for details.

addition.

2. Using this result, we show that semantic hierarchy between concepts is encoded geometrically as orthogonality between representations, in a manner we make precise.

3. Then, we construct the representation of categorical variables (e.g., animal) as the polytope where the vertices are the representations of the binary features that define the category (e.g., mammal, bird, . . .). We show that for "natural" concepts, the representation is a simplex.

4. Finally, we empirically validate these theoretical results on the Gemma large language model (Mesnard et al., 2024). To that end, we extract concepts from the WordNet hierarchy (Miller, 1995), estimate their representations, and show that the geometric structure of the representations align with the semantic hierarchy of WordNet.

The final structure is remarkably simple, and is summarized in Figure 1. In totality, these results provide a foundation for understanding how high-level semantic concepts are encoded in the representation spaces of LLMs.

Before we proceed, we introduce notations and definitions that we use throughout the paper in Appendix A, which cover large language models, concepts, the causal inner product, and linear representations.

2. Binary Concepts and Hierarchical Structure

Our high-level strategy will be to build up from binary concepts to more complex structure. We begin by defining the basic building blocks.

Binary and Categorical Concepts We consider two kinds of binary concept. A *binary feature* $W \in_{\mathcal{R}} \{\texttt{not\_w}, \texttt{is\_w}\}$ is an indicator of whether the output has the attribute w. For example, if the feature `is_animal` is true then the output will be about an animal. A *binary contrast* $\texttt{a} \Rightarrow \texttt{b} \in_R \{a, b\}$ is a binary variable that contrasts two specific attribute values. For example, the variable `mammal` ⇒ `bird` is a binary contrast. In the particular case where the concept may be represented as counterfactual pairs of words, we can identify the contrast with the notion of linear representation in Park et al. (2024).

We also define a categorical concept to be any concept corresponding to a categorical latent variable. This includes binary concepts as a special case.

Hierarchical Structure The next step is to define what we mean by a hierarchical relation between concepts. To that end, to each attribute w, we associate a set of tokens $\mathcal{Y}(w)$ that have the attribute. For example, $\mathcal{Y}(\texttt{mammal}) = \{\text{" dog"}, \text{" cats"}, \text{" Tiger"}, \dots\}$. Then,

Definition 2.1. A value z is *subordinate* to a value w (denoted by $z \prec w$) if $\mathcal{Y}(z) \subseteq \mathcal{Y}(w)$. We say a categorical concept $Z \in_R \{z_0, \dots, z_{k-1}\}$ is subordinate to a categorical concept $W \in_R \{w_0, \dots, w_{n-1}\}$ if there exists a value w_Z of W such that each value z_i of Z is subordinate to w_Z.

For example, the binary contrast `dog` ⇒ `cat` is subordinate to the binary feature {`is_mammal`, `not_mammal`}, and the binary contrast `parrot` ⇒ `eagle` is subordinate to the categorical concept {**`mammal`**, **`bird`**, **`fish`**}. On the other hand, `dog` ⇒ `eagle` is not subordinate to `bird` ⇒ `mammal`, and `bird` ⇒ `mammal` and `live_in_house` ⇒ `live_in_water` are not subordinate to each other.

Linear Representations of Binary Concepts Now we return to the question of how binary concepts are represented. A key desideratum is that if $\bar{\ell}_W$ is a linear representation then moving the representation in this direction should modify the probability of the target concept *in isolation*. If adding $\bar{\ell}_W$ also modified off-target concepts, it would not be natural to identify it with W. In Definition A.1, this idea is formalized by the requirement that the probability of causally separable concepts is unchanged when the representation is added to the context.

We now observe that, when there is hierarchical structure, this requirement is not strong enough to capture 'off-target' behavior. For example, if $\bar{\ell}_{\texttt{animal}}$ captures the concept of animal vs not-animal, then moving in this direction should not affect the relative probability of the output being about a mammal versus a bird. If it did, then the representation would actually capture some amalgamation of the animal and mammal concepts. Accordingly, we must strengthen our definition:

Definition 2.2. A vector $\bar{\ell}_W$ is a linear representation of a binary concept W if for all contexts ℓ,

$$\mathbb{P}(W = 1 \mid \ell + \alpha\bar{\ell}_W) > \mathbb{P}(W = 1 \mid \ell), \text{ and} \tag{2.1}$$

$$\mathbb{P}(Z \mid \ell + \alpha\bar{\ell}_W) = \mathbb{P}(Z \mid \ell), \tag{2.2}$$

for all $\alpha > 0$ and all concept variables Z that are either subordinate or causally separable with W. Here, if W is a binary feature for an attribute w, $W = 1$ denotes $W = \texttt{is\_w}$.

Notice that, in the case of binary concepts defined by counterfactual pairs, this definition is equivalent to Definition A.1, because such variables have no subordinate concepts.

3. Representations of Complex Concepts

Vector Representations of Binary Features To build up to complex concepts, we need to understand how to compose representations of binary concepts. At this stage, the representations are *directions* in the representation space—they do not have a natural notion of magnitude. In particular, this means we cannot use vector operations (such as addition) to compose representations. To overcome this, we now show how to associate a magnitude to the representation of a binary concept.

The key is the following result connecting binary feature representations and word unembeddings:

Theorem 3.1 (Magnitudes of Linear Representations)**.** *Suppose there exists a linear representation (normalized direction) $\bar{\ell}_W$ of a binary feature W for an attribute w. Then, there is a constant $b_w > 0$ and a choice of unembedding space origin $\bar{\gamma}_0^w$ in (A.2) such that*

$$\begin{cases} \bar{\ell}_W^\top g(y) = b_w & \text{if } y \in \mathcal{Y}(w) \\ \bar{\ell}_W^\top g(y) = 0 & \text{if } y \notin \mathcal{Y}(w). \end{cases} \tag{3.1}$$

Further, if there are d causally separable attributes $\{w_0, \dots, w_{d-1}\}$ with linear representations, we can choose a canonical origin $\bar{\gamma}_0$ in (A.2) as $\bar{\gamma}_0 = \sum_i \bar{\gamma}_0^{w_i}$.

All proofs are given in Appendix D.

In words, this theorem says that if a (perfect) linear representation of the `animal` feature exists, then every token that has the animal attribute has the *same* dot product with the representation vector; i.e., "cat" is exactly as much `animal` as "dog" is. If this weren't true, then increasing the probability that the output is about an animal would also increase the relative probability that the output is about a dog rather than a cat. In practice, such exact representations are unlikely to be found by gradient descent in LLM training. Rather, we expect $\bar{\ell}_W^\top g(y)$ to be isotropically distributed around b_w with variance that is small compared to b_w (so that animal and non-animal words are well-separated.)

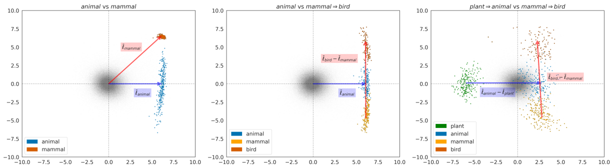


Figure 2. Hierarchical semantics are encoded as orthogonality in the representation space, as predicted in Theorem 3.3. The plots show the projection of the unembedding vectors on the 2D subspaces: $\text{span}\{\bar\ell_{\texttt{animal}}, \bar\ell_{\texttt{mammal}}\}$ (left; statement (b)), $\text{span}\{\bar\ell_{\texttt{animal}}, \bar\ell_{\texttt{bird}} - \bar\ell_{\texttt{mammal}}\}$ (middle; statement (c)), and $\text{span}\{\bar\ell_{\texttt{animal}} - \bar\ell_{\texttt{plant}}, \bar\ell_{\texttt{bird}} - \bar\ell_{\texttt{mammal}}\}$ (right; statement (d)). The gray points indicate all 256K tokens in the vocabulary, and the colored points are the tokens in $\mathcal{Y}(\texttt{w})$. The blue and red vectors are used to span the 2D subspaces.

With this result in hand, we can define a notion of vector representation for a binary feature:

Definition 3.2. We say that binary feature W for an attribute w has a *vector representation* $\bar\ell_w \in \mathbb{R}^d$ if $\bar\ell_w$ satisfies Definition 2.2 and $\|\bar\ell_w\|_2 = b_w$ in Theorem 3.1. If the vector representation of a binary feature is not unique, we say $\bar\ell_w$ is the vector representation that maximizes b_w.

Hierarchical Orthogonality We have now moved from representations as directions to representations as vectors. Using this result, we now establish how hierarchical relations between concepts are encoded in the vector space structure of the representation space. The structure is illustrated in Figure 2. Formally, we have the following connections between vector and semantic structure:

Theorem 3.3 (Hierarchical Orthogonality)**.** *Suppose there exist the vector representations for all the following binary features. Then, we have that*

(a) $\bar\ell_{w_1} - \bar\ell_{w_0}$ *is a linear representation* $\bar\ell_{w_0 \Rightarrow w_1}$ *defined in Definition 2.2;*

(b) $\bar\ell_w \perp \bar\ell_z - \bar\ell_w$ *for* $z \prec w$*;*

(c) $\bar\ell_w \perp \bar\ell_{z_1} - \bar\ell_{z_0}$ *for* $Z \in_R \{z_0, z_1\}$ *subordinate to* $W \in_R \{\texttt{not\_w}, \texttt{is\_w}\}$*;*

(d) $\bar\ell_{w_1} - \bar\ell_{w_0} \perp \bar\ell_{z_1} - \bar\ell_{z_0}$ *for* $Z \in_R \{z_0, z_1\}$ *subordinate to* $W \in_R \{w_0, w_1\}$*; and*

(e) $\bar\ell_{w_1} - \bar\ell_{w_0} \perp \bar\ell_{w_2} - \bar\ell_{w_1}$ *for* $w_2 \prec w_1 \prec w_0$*.*

We emphasize that these results—involving differences of representations—are only possible because we now have *vector* representations (mere direction would not suffice).

Categorical Concepts as Simplices The power of having a vector representation is that now we can use ordinary vector space operations to construct representation of other concepts. We now turn to the representation of categorical concepts, e.g., $\{\texttt{mammal}, \texttt{reptile}, \texttt{bird}, \texttt{fish}\}$. There is now a straightforward way to define the representation of such concepts:

Definition 3.4. The *polytope representation* of a categorical concept $Z = \{z_0, \ldots, z_{k-1}\}$ is the convex hull of the vector representations of the elements of the concept.

Polytopes are quite general objects. The definition here also includes representations of categorical variables that are semantically unnatural, e.g., $\{\texttt{dog}, \texttt{sandwich}, \texttt{running}\}$. We would like to make a more precise statement about the representation of "natural" concepts. One possible notion of a "natural" concept is one where the model can freely manipulate the output values. The next theorem shows that such concepts have a particularly simple structure:

Theorem 3.5 (Categorical Concepts are Represented as Simplices)**.** *Suppose that* $\{w_0, \ldots, w_{k-1}\}$ *is a collection of* k *mutually exclusive attributes such that for every joint distribution* $Q(w_0, \ldots w_{k-1})$ *there is some* ℓ_i *such that* $\mathbb{P}(W = w_i \mid \ell_i) = Q(W = w_i)$ *for every* i*. Then, the vector representations* $\bar\ell_{w_0}, \ldots, \bar\ell_{w_{k-1}}$ *form a* $(k-1)$*-simplex in the representation space. In this case, we take the simplex to be the representation of the categorical concept* $W = \{w_0, \ldots, w_{k-1}\}$*.*

Summary Together, Theorems 3.3 and 3.5 give the simple structure illustrated in Figure 1: hierarchical concepts are represented as direct sums of simplices. The direct sum structure is immediate from the orthogonality in Theorem 3.3.

Experiments In Appendix B, we empirically test the theoretical results in the representation space of the Gemma-2B large language model (Mesnard et al., 2024).

4. Discussion

We set out to understand how semantic structure is encoded in the geometry of representation space. We have arrived an astonishingly simple structure, summarized in Figure 1. The key contributions are moving from representing concepts as directions to representing them as vectors (and polytopes), and connecting semantic hierarchy to orthogonality. We include a discussion of related work and future work in Appendix C.

References

Allen, C. and Hospedales, T. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pp. 223–231. PMLR, 2019.

Allen, C., Balazevic, I., and Hospedales, T. What the vec? towards probabilistically grounded embeddings. *Advances in Neural Information Processing Systems*, 32, 2019.

Amini, A. A., Aragam, B., and Zhou, Q. A non-graphical representation of conditional independence via the neighbourhood lattice. *arXiv preprint arXiv:2206.05829*, 2022.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120, 2006.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Chang, T. A., Tu, Z., and Bergen, B. K. The geometry of multilingual language model representations. *arXiv preprint arXiv:2205.10964*, 2022.

Chen, B., Fu, Y., Xu, G., Xie, P., Tan, C., Chen, M., and Jing, L. Probing bert in hyperbolic spaces. *arXiv preprint arXiv:2104.03869*, 2021.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Ethayarajh, K., Duvenaud, D., and Hirst, G. Towards understanding linear word analogies. *arXiv preprint arXiv:1810.04882*, 2018.

Frandsen, A. and Ge, R. Understanding composition of word embeddings via tensor decomposition. *arXiv preprint arXiv:1902.00613*, 2019.

Ganea, O., Bécigneul, G., and Hofmann, T. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pp. 1646–1655. PMLR, 2018.

Ghandeharioun, A., Caciularu, A., Pearce, A., Dixon, L., and Geva, M. Patchscopes: A unifying framework for inspecting hidden representations of language models. *International Conference on Machine Learning (to appear)*, 2024.

Gittens, A., Achlioptas, D., and Mahoney, M. W. Skip-gram - zipf + uniform = vector additivity. In *Annual Meeting of the Association for Computational Linguistics*, 2017.

Gurnee, W. and Tegmark, M. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

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

He, Y., Yuan, Z., Chen, J., and Horrocks, I. Language models as hierarchy encoders. *arXiv preprint arXiv:2401.11374*, 2024.

Jiang, Y., Aragam, B., and Veitch, V. Uncovering meanings of embeddings via partial orthogonality. *arXiv preprint arXiv:2310.17611*, 2023.

Jiang, Y., Rajendran, G., Ravikumar, P., Aragam, B., and Veitch, V. On the origins of linear representations in large language models. In *International Conference on Machine Learning (to appear)*, 2024.

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*, 2020.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023a.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023b.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625, 2022.

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

Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Mikolov, T., Yih, W.-t., and Zweig, G. 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*, pp. 746–751, 2013.

Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Mimno, D. and Thompson, L. The strange geometry of skip-gram with negative sampling. In *Conference on Empirical Methods in Natural Language Processing*, 2017.

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

Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (to appear)*, 2024.

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

Reif, E., Yuan, A., Wattenberg, M., Viegas, F. B., Coenen, A., Pearce, A., and Kim, B. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32, 2019.

Ri, N., Lee, F.-T., and Verma, N. Contrastive loss is all you need to recover analogies as parallel lines. *arXiv preprint arXiv:2306.08221*, 2023.

Rudolph, M. and Blei, D. Dynamic bernoulli embeddings for language evolution. *arXiv preprint arXiv:1703.08052*, 2017.

Rudolph, M., Ruiz, F., Mandt, S., and Blei, D. Exponential family embeddings. *Advances in Neural Information Processing Systems*, 29, 2016.

Tigges, C., Hollinsworth, O. J., Geiger, A., and Nanda, N. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

Valeriani, L., Doimo, D., Cuturello, F., Laio, A., Ansuini, A., and Cazzaniga, A. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36, 2024.

Volpi, R. and Malagò, L. Evaluating natural alpha embeddings on intrinsic and extrinsic tasks. In *Workshop on Representation Learning for NLP*, 2020.

Volpi, R. and Malagò, L. Natural alpha embeddings. *Information Geometry*, 4(1):3–29, 2021.

Wang, Z., Gui, L., Negrea, J., and Veitch, V. Concept algebra for (score-based) text-controlled generative models. In *Advances in Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=SGlrCuwdsB`.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Preliminaries

Large Language Models For the purposes of this paper, we consider a large language model to consist of two parts. The first part is a function λ that maps input texts x to vectors $\lambda(x)$ in a representation space $\Lambda \simeq \mathbb{R}^d$. This is the function given by the stacked transformer blocks. We take $\lambda(x)$ to be the output of the final layer at the final token position. The second part is an unembedding layer that assigns a vector $\gamma(y)$ in an unembedding space $\Gamma \simeq \mathbb{R}^d$ to each token y in the vocabulary. Together, these define a sampling distribution over tokens via the softmax distribution:

$$\mathbb{P}(y \mid x) = \frac{\exp(\lambda(x)^\top \gamma(y))}{\sum_{y' \in \mathrm{Vocab}} \exp(\lambda(x)^\top \gamma(y'))}. \tag{A.1}$$

The broad goal is to understand how semantic structure is encoded in the geometry of the spaces Λ and Γ. (We do not address the "internal" structure of the LLMs in this paper, though we are optimistic that a clear understanding of the softmax geometry will shed light on this as well.)

Concepts We formalize a concept as a latent variable W that is caused by the context X and causes the output Y. That is, a concept is a thing that could—in principle—be manipulated to affect the output of the language model. In the particular case where a concept is a binary variable with a word-level counterfactual, we can identify the variable W with the counterfactual pair of outputs $(Y(0), Y(1))$. Concretely, we can identify $\texttt{male} \Rightarrow \texttt{female}$ with $(Y(0), Y(1)) \in_R \{(\text{"man"}, \text{"woman"}), (\text{"king"}, \text{"queen"}), (\text{"he"}, \text{"her"}), \dots\}$. We emphasize that the notion of a concept as a latent variable that affects the output is more general than the counterfactual binary case.

Given a pair of concept variables W and Z, we say that W is *causally separable* with Z if the potential outcome $Y(W = w, Z = z)$ is well-defined for all w, z. That is, two variables are causally separable if they can be freely manipulated—e.g., we can change the output language and the sex of the subject freely, so these concepts are causally separable.

Causal Inner Product and Linear Representations We are trying to understand how concepts are represented. At this stage, there are two distinct representation spaces: Λ and Γ. The former is the space of context embeddings, and the latter is the space of token unembeddings. We would like to unify these spaces so that there is just a single notion of representation.

Park et al. (2024) show how to achieve this unification via a "Causal Inner Product". This is a particular choice of inner product that respects the semantics of language in the sense that the linear representations of (binary, counterfactual) causally separable concepts are orthogonal under the inner product. Their result can be understood as saying that there is some invertible matrix A and constant vector $\bar{\gamma}_0$ such that, if we transform the embedding and unembedding spaces as

$$g(y) \leftarrow A(\gamma(y) - \bar{\gamma}_0), \quad \ell(x) \leftarrow A^{-\top}\lambda(x) \tag{A.2}$$

then the Euclidean inner product in the transformed spaces is the causal inner product, and the Riesz isomorphism between the embedding and unembedding spaces is simply the usual vector transpose operation. We can estimate A as the whitening operation for the unembedding matrix. Following this transformation, we can think of the embedding and unembedding spaces as the same space, equipped with the familiar Euclidean inner product.[1]

Notice that the softmax probabilities are unchanged for any A and $\bar{\gamma}_0$, so this transformation does not affect the model's behavior. The vector $\bar{\gamma}_0$ defines an origin for the unembedding space, and can be chosen arbitrarily. We give a particularly convenient choice below.

In this unified space, the linear representation of a binary concept $W \in_R \{0, 1\}$ is defined as:

Definition A.1. A vector $\bar{\ell}_W$ is a linear representation of a binary concept W if for all contexts ℓ, and all concept variables Z that are causally separable with W, we have, for all $\alpha > 0$,

$$\mathbb{P}(W = 1 \mid \ell + \alpha\bar{\ell}_W) > \mathbb{P}(W = 1 \mid \ell), \text{ and} \tag{A.3}$$

$$\mathbb{P}(Z \mid \ell + \alpha\bar{\ell}_W) = \mathbb{P}(Z \mid \ell). \tag{A.4}$$

[1] We are glossing over some technical details here; see (Park et al., 2024) for details.

That is, the linear representation is a direction in the representation space that, when added to the context, increases the probability of the concept, but does not affect the probability of any off-target concept. In the case of binary concepts that can be represented as counterfactual pairs of words, this direction can be shown to be proportional to the "linear probing" direction, and proportional to $g(Y(1)) - g(Y(0))$ for any counterfactual pair of words $Y(1), Y(0)$ that represent the concept W.

B. Experiments

Canonical representation space The results in this paper rely on transforming the representation space so that the Euclidean inner product is a causal inner product, aligning the embedding and unembedding representations. Following Park et al. (2024), we estimate the required transformation as:

$$g(y) = \mathrm{Cov}(\gamma)^{-1/2}(\gamma(y) - \mathbb{E}[\gamma])$$

where γ is the unembedding vector of a word sampled uniformly from the vocabulary. Centering by $\mathbb{E}[\gamma]$ is a reasonable approximation of centering by $\bar{\gamma}_0$ defined in Theorem 3.1 because this makes the projection of a random $g(y)$ on an arbitrary direction close to 0. This matches the requirement that the projection of a word onto a concept the words does not belong to should be close to 0.

WordNet We define a large collection of binary concepts using WordNet (Miller, 1995). Briefly, WordNet organizes English words into a hierarchy of synsets, where each synset is a set of synonyms. The WordNet hierarchy is based on word hyponym relations, and reflects the semantic hierarchy of interest in this paper. We take each synset as an attribute w and define $\mathcal{Y}(w)$ as the collection of all words belonging to any synset that is a descendant of w. For example, the synset `mammal.n.01` is a descendant of `animal.n.01`, so both $\mathcal{Y}(\texttt{mammal.n.01})$ and $\mathcal{Y}(\texttt{animal.n.01})$ contain the word "dog". We collect all noun and verb synsets, and augment the word collections by including plural forms of the nouns, multiple tenses of each verb, and capital and lower case versions of each word. We filter to include only those synsets with at least 50 words in the Gemma vocabulary. This leaves us with 593 noun and 364 verb synsets, each defining an attribute.

Estimation via Linear Discriminant Analysis Now, we want to estimate the vector representation $\bar{\ell}_w$ for each attribute w. To do this, we make use of vocabulary sets $\mathcal{Y}(w)$. Following Theorem 3.1, the vector associated to the concept w should have two properties. First, when the full vocabulary is projected onto this vector, the words in $\mathcal{Y}(w)$ should be well-separated from the rest of the vocabulary. Second, the projection of the unembedding vectors for $y \in \mathcal{Y}(w)$ should be approximately the same value. Equivalently, the variance of the projection of the unembedding vectors for $y \in \mathcal{Y}(w)$ should be small. To capture these requirements, we estimate the directions using a variant of Linear Discriminant Analysis (LDA), which finds a projection minimizing within-class variance and maximizing between-class variance. Formally, we estimate the vector representation of a binary feature W for an attribute w as

$$\bar{\ell}_w = \left(\tilde{g}_w^\top \mathbb{E}(g_w)\right) \tilde{g}_w, \quad \text{with} \quad \tilde{g}_w = \frac{\mathrm{Cov}(g_w)^\dagger \mathbb{E}(g_w)}{\|\mathrm{Cov}(g_w)^\dagger \mathbb{E}(g_w)\|_2},$$

where g_w is the unembedding vector of a word sampled uniformly from $\mathcal{Y}(w)$ and $\mathrm{Cov}(g_w)^\dagger$ is a pseudo-inverse of the covariance matrix. We estimate the covariance matrix $\mathrm{Cov}(g_w)$ using the Ledoit-Wolf shrinkage estimator (Ledoit & Wolf, 2004), because the dimension of the representation spaces is much higher than the number of samples.

B.1. Visualization of `animal`

As a concrete example, we check the theoretical predictions for the concept `animal`. For this, we generated two sets of tokens $\mathcal{Y}(\texttt{animal})$ and $\mathcal{Y}(\texttt{plant})$ using ChatGPT-4 (OpenAI, 2023) and manually inspected them. $\mathcal{Y}(\texttt{animal})$ is separated to six sets of tokens for each subcategory $\{\texttt{mammal}, \texttt{bird}, \texttt{fish}, \texttt{reptile}, \texttt{amphibian}, \texttt{insect}\}$.

Figure 2 illustrates the geometric relationships between various representation vectors. The main takeaway is that the semantic hierarchy is encoded as orthogonality in the manner predicted by Theorem 3.3. The figure also illustrates Theorem 3.1, showing that the projection of the unembedding vectors for $y \in \mathcal{Y}(w)$ is approximately constant, while the projection of $y \notin \mathcal{Y}(w)$ is zero.

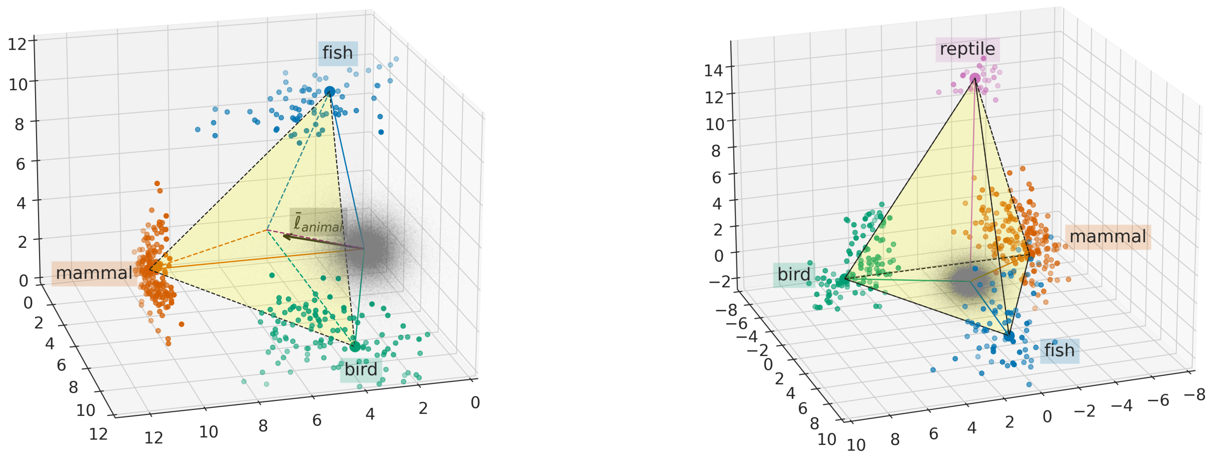


Figure 3. Categorical concepts are represented as simplices. The plots show the projection of the unembedding vectors on the 3D subspaces: $\mathrm{span}\{\bar{\ell}_{\texttt{mammal}}, \bar{\ell}_{\texttt{bird}}, \bar{\ell}_{\texttt{fish}}\}$ (left) and $\mathrm{span}\{\bar{\ell}_{\texttt{bird}} - \bar{\ell}_{\texttt{mammal}}, \bar{\ell}_{\texttt{fish}} - \bar{\ell}_{\texttt{mammal}}, \bar{\ell}_{\texttt{reptile}} - \bar{\ell}_{\texttt{mammal}}\}$ (right). The gray points indicate all 256K tokens in the vocabulary, and the colored points are the tokens in $\mathcal{Y}(\texttt{w})$. The left plot further shows the orthogonality between the triangle and the projection of $\bar{\ell}_{\texttt{animal}}$ (black arrow).

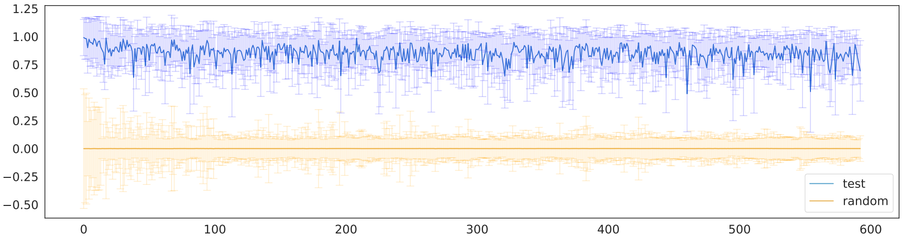


Figure 4. Linear representations exist for most binary features in the WordNet noun hierarchy. Comparison of projection of test and random words on estimated vector representations for each WordNet feature. The values are divided by the norm of the estimated vector representation. The x-axis indices denote all features in the noun hierarchy. The thick lines present the mean of the projections for each feature and the error bars indicate the 1.96 × standard error.

Figure 3 illustrates that the representation of a categorical concept is a simplex, as predicted in Theorem 3.5. It also shows that, as predicted, the simplex for $\texttt{fish}, \texttt{mammal}, \texttt{bird}$ is orthogonal to the vector representation of `animal`.

B.2. WordNet Hierarchy

We now turn to using the WordNet hierarchy to evaluate the theoretical predictions at scale. For space, we report the noun hierarchy here and defer the verb hierarchy to Appendix F.

Existence of Vector Representations for Binary Features To evaluate whether vector representations exist, for each synset w we split $\mathcal{Y}(w)$ into train words (80%) and test words (20%), fit the LDA estimator to the train words, and examine the projection of the unembedding vectors for the test words onto the estimated vector representation. Figure 4 shows the mean and standard error of the test projections, divided by the magnitude of each estimated $\bar{\ell}_w$. If a vector representation exists for an attribute, we would expect these values to be close to 1. We see that this is indeed the case, giving evidence that vector representations do indeed exist for these features.

Hierarchical Orthogonality It remains to evaluate the prediction that hierarchical relations are encoded as orthogonality in the representation space. Figure 5 shows the adjacency matrix of the WordNet noun hyponym inclusion graph (left), the cosine similarity between the vector representations $\bar{\ell}_w$ for each feature (middle), and the cosine similarity between child-parent vectors $\bar{\ell}_w - \bar{\ell}_{\text{parent of } w}$ for each feature (right). Strikingly, the cosine similarity clearly reflects the semantic hierarchy—

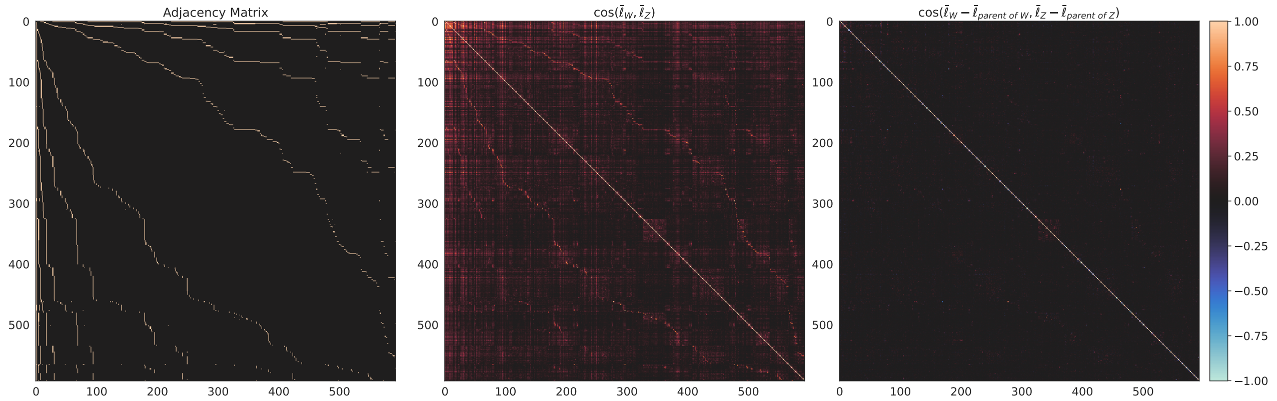


Figure 5. Hierarchical semantics in WordNet are encoded in Gemma representation space, with the orthogonal structure predicted in Theorem 3.3. The adjacency matrix of the hierarchical relations between features in the noun hierarchy (left), the cosine similarity between the vector representations $\bar{\ell}_w$ for each feature (middle), and the cosine similarity between child-parent vectors $\bar{\ell}_w - \bar{\ell}_{\text{parent of } w}$ for each feature (right). The features are ordered by the hierarchy.

the adjacency matrix is clearly visible in the middle heatmap. This is because, e.g., `mammal.n.01` and `animal.n.01` have high cosine similarity. By contrast, as predicted by Theorem 3.3, the child-parent and parent-grandparent vectors are orthogonal. This also straightforwardly implies all other theoretical connections between orthogonality and semantic hierarchy.

In Appendix F, we present zoomed-in heatmaps for the subtree of descendants of "animal", and the results for the verb hierarchy.

C. Related Work and Future Work

This result connects closely to the study of linear representations in language models (e.g., Mikolov et al., 2013; Pennington et al., 2014; Arora et al., 2016; Elhage et al., 2022; Burns et al., 2022; Tigges et al., 2023; Nanda et al., 2023; Moschella et al., 2022; Li et al., 2023b; Gurnee et al., 2023; Wang et al., 2023; Jiang et al., 2024; Park et al., 2024). In particular, Park et al. (2024) formalize the linear representation hypothesis by unifying three distinct notions of linearity: word2vec-like embedding differences, logistic probing, and steering vectors. Our work relies on this unification, and just focuses on the steering vector notion. Our work also connects to work aimed at theoretically understanding the existence of linear representations. Specifically, (Arora et al., 2016; 2018; Frandsen & Ge, 2019) use RAND-WALK model where the latent vectors are modeled to drift on the unit sphere. (Blei & Lafferty, 2006; Rudolph et al., 2016; Rudolph & Blei, 2017) consider a similar dynamic topic modeling. Gittens et al. (2017) and subsequent works (Allen & Hospedales, 2019; Allen et al., 2019) propose a paraphrasing model where a subset of words is semantically equivalent to a single word. Ethayarajh et al. (2018) try to explain linear representations by decomposing the pointwise mutual information matrix while Ri et al. (2023) connect it to contrastive loss. Jiang et al. (2024) connect the existence of linear representations to the implicit bias of gradient descent. In this paper, we do not seek to justify the *existence* of linear representations, but rather to understand their *structure* if they do exist. Though, by empirically estimating vector representations for thousands of concepts, we add to the body of evidence supporting the existence of linear representations. Elhage et al. (2022) also empirically observe the formation of polytopes in the representation space of a toy model, and the present work can be viewed as giving an explanation for this phenomenon.

There is also a growing literature studying the representation geometry of natural language (Mimno & Thompson, 2017; Reif et al., 2019; Volpi & Malagò, 2021; 2020; Li et al., 2020; Chen et al., 2021; Chang et al., 2022; Liang et al., 2022; Jiang et al., 2023; Park et al., 2024; Valeriani et al., 2024). Much of this work focuses on connections to hyperbolic geometry (Nickel & Kiela, 2017; Ganea et al., 2018; Chen et al., 2021; He et al., 2024). We do not find such a connection in existing LLMs, but it is an interesting direction for future work to determine if more efficient LLM representations could be constructed in hyperbolic space. Jiang et al. (2023) hypothesize that very general ”independence structures” are naturally represented by partial orthogonality in vector spaces (Amini et al., 2022). The results here confirm and expand on this hypothesis in the case of hierarchical structure in language models.

Future Work The results in this paper are foundational for understanding the structure of representation space in language models. Of course, the ultimate purpose of foundations is to build upon them. One immediate direction is to refine the attempts to interpret LLM structure to explicitly account for hierarchical semantics. As a concrete example, there is currently significant interest in using sparse autoencoders to extract interpretable features from LLMs (e.g., Cunningham et al., 2023; Bricken et al., 2023). This work searches for representations in terms of distinct binary features. Concretely, it hopes to find features for, e.g., `animal`, `mammal`, `bird`, etc. Based on the results here, these representations are strongly co-linear, and potentially difficult to disentangle. On the other hand, a representation in terms of $\bar{\ell}_{\texttt{animal}}$, $\bar{\ell}_{\texttt{mammal}} - \bar{\ell}_{\texttt{animal}}$, $\bar{\ell}_{\texttt{bird}} - \bar{\ell}_{\texttt{animal}}$, etc., will be cleanly separated and equally interpretable. Fundamentally, semantic meaning has hierarchical structure, so interpretability methods should respect this structure. Understanding the geometric representation makes it possible to design such methods.

In a separate, foundational, direction: the results in this paper rely on using the canonical representation space. We estimate this using the whitening transformation of the unembedding layer. However, this technique only works for the final layer representation. It is an important open question how to make sense of the geometry of internal layers.

D. Proofs

D.1. Proof of Theorem 3.1

Theorem 3.1 (Magnitudes of Linear Representations)**.** *Suppose there exists a linear representation (normalized direction) $\bar{\ell}_W$ of a binary feature W for an attribute w. Then, there is a constant $b_w > 0$ and a choice of unembedding space origin $\bar{\gamma}_0^w$ in (A.2) such that*

$$\begin{cases} \bar{\ell}_W^\top g(y) = b_w & \text{if } y \in \mathcal{Y}(w) \\ \bar{\ell}_W^\top g(y) = 0 & \text{if } y \notin \mathcal{Y}(w). \end{cases} \tag{3.1}$$

Further, if there are d causally separable attributes $\{w_0, \dots, w_{d-1}\}$ with linear representations, we can choose a canonical origin $\bar{\gamma}_0$ in (A.2) as $\bar{\gamma}_0 = \sum_i \bar{\gamma}_0^{w_i}$.

Proof. For any $y_1, y_0 \in \mathcal{Y}(w)$ or $y_1, y_0 \notin \mathcal{Y}(w)$, let Z be a binary concept where $\mathcal{Y}(Z = 0) = \{y_0\}$ and $\mathcal{Y}(Z = 1) = \{y_1\}$. Since Z is subordinate to W, (2.2) implies that

$$\operatorname{logit} \mathbb{P}(Y = y_1 \mid Y \in \{y_0, y_1\}, \ell + \bar{\ell}_W) = \operatorname{logit} \mathbb{P}(Y = y_1 \mid Y \in \{y_0, y_1\}, \ell) \tag{D.1}$$

$$\iff \bar{\ell}_W^\top (g(y_1) - g(y_0)) = \bar{\ell}_W^\top A(\gamma(y_1) - \gamma(y_0)) = 0 \tag{D.2}$$

where A is the invertible matrix in (A.2). This means that $\bar{\ell}_W^\top A\gamma(y)$ is the same for all $y \in \mathcal{Y}(w)$, and it is also the same for all $y \notin \mathcal{Y}(w)$.

Furthermore, for any $y_1 \in \mathcal{Y}(w)$ and $y_0 \notin \mathcal{Y}(w)$, (2.1) implies that

$$\operatorname{logit} \mathbb{P}(Y = y_1 \mid Y \in \{y_0, y_1\}, \ell + \bar{\ell}_W) > \operatorname{logit} \mathbb{P}(Y = y_1 \mid Y \in \{y_0, y_1\}, \ell) \tag{D.3}$$

$$\iff \bar{\ell}_W^\top (g(y_1) - g(y_0)) = \bar{\ell}_W^\top A(\gamma(y_1) - \gamma(y_0)) > 0. \tag{D.4}$$

Thus, by setting $b_w^0 = \bar{\ell}_W^\top A\gamma(y)$ for any $y \notin \mathcal{Y}(w)$, and $b_w = \bar{\ell}_W^\top A\gamma(y_1) - \bar{\ell}_W^\top A\gamma(y_0) > 0$ for any $y_1 \in \mathcal{Y}(w)$ and $y_0 \notin \mathcal{Y}(w)$, we get

$$\begin{cases} \bar{\ell}_W^\top A\gamma(y) = b_w^0 + b_w & \text{if } y \in \mathcal{Y}(w) \\ \bar{\ell}_W^\top A\gamma(y) = b_w^0 & \text{if } y \notin \mathcal{Y}(w). \end{cases} \tag{D.5}$$

Then, we can choose an origin as

$$\bar{\gamma}_0^w = b_w^0 A^{-1} \bar{\ell}_W \tag{D.6}$$

satisfying (3.1).

On the other hand, if there exist $\bar{\ell}_W$ and $\bar{\ell}_Z$ for causally separable attributes w and z, then $\bar{\ell}_W$ and $\bar{\ell}_Z$ are orthogonal by the property of the causal inner product. If they are not orthogonal, adding $\bar{\ell}_Z$ can change the other concept W, and it is a contradiction. Now if there exist the linear representation for d binary features for causally separable attributes $\{w_0, \dots, w_{d-1}\}$, we can choose a canonical $\bar{\gamma}_0$ in (A.2) as

$$\bar{\gamma}_0 = \sum_i \bar{\gamma}_0^{w_i}. \tag{D.7}$$

with (3.1) satisfied. □

D.2. Proof of Theorem 3.3

Theorem 3.3 (Hierarchical Orthogonality)**.** *Suppose there exist the vector representations for all the following binary features. Then, we have that*

(a) $\bar{\ell}_{w_1} - \bar{\ell}_{w_0}$ *is a linear representation* $\bar{\ell}_{w_0 \Rightarrow w_1}$ *defined in Definition 2.2;*

(b) $\bar{\ell}_w \perp \bar{\ell}_z - \bar{\ell}_w$ *for* $z \prec w$*;*

(c) $\bar{\ell}_w \perp \bar{\ell}_{z_1} - \bar{\ell}_{z_0}$ *for* $Z \in_R \{z_0, z_1\}$ *subordinate to* $W \in_R \{\texttt{not\_w}, \texttt{is\_w}\}$*;*

(d) $\bar{\ell}_{w_1} - \bar{\ell}_{w_0} \perp \bar{\ell}_{z_1} - \bar{\ell}_{z_0}$ *for* $Z \in_R \{z_0, z_1\}$ *subordinate to* $W \in_R \{w_0, w_1\}$*; and*

(e) $\bar{\ell}_{w_1} - \bar{\ell}_{w_0} \perp \bar{\ell}_{w_2} - \bar{\ell}_{w_1}$ *for* $w_2 \prec w_1 \prec w_0$*.*

Proof. (a) For $\bar{\ell}_{w_1}$ and $\bar{\ell}_{w_0}$, by Theorem 3.1, we have

$$\begin{cases} (\bar{\ell}_{w_1} - \bar{\ell}_{w_0})^\top g(y) = 0 - b_{w_0} = -b_{w_0} & \text{if } y \in \mathcal{Y}(w_0) \\ (\bar{\ell}_{w_1} - \bar{\ell}_{w_0})^\top g(y) = b_{w_1} - 0 = b_{w_1} & \text{if } y \in \mathcal{Y}(w_1) \\ (\bar{\ell}_{w_1} - \bar{\ell}_{w_0})^\top g(y) = 0 - 0 = 0 & \text{if } y \notin \mathcal{Y}(w_0) \cup \mathcal{Y}(w_1). \end{cases} \tag{D.8}$$

Since $\bar{\ell}_{w_1} - \bar{\ell}_{w_0}$ can change the target concept $w_0 \Rightarrow w_1$ without changing any other concept subordinate or causally separable to the target concept, $\bar{\ell}_{w_1} - \bar{\ell}_{w_0}$ is the linear representation $\bar{\ell}_{w_0 \Rightarrow w_1}$.

(b) For $\bar{\ell}_w$ and $\bar{\ell}_z$ where $z \prec w$, by Theorem 3.1, we have

$$\begin{cases} (\bar{\ell}_z - \bar{\ell}_w)^\top g(y) = b_z - b_w & \text{if } y \in \mathcal{Y}(z) \\ (\bar{\ell}_z - \bar{\ell}_w)^\top g(y) = 0 - b_w = -b_w & \text{if } y \in \mathcal{Y}(w) \setminus \mathcal{Y}(z) \\ (\bar{\ell}_z - \bar{\ell}_w)^\top g(y) = 0 - 0 = 0 & \text{if } y \notin \mathcal{Y}(w). \end{cases} \tag{D.9}$$

When $w \setminus z$ denotes an attribute defined by $\mathcal{Y}(w) \setminus \mathcal{Y}(z)$, $\bar{\ell}_z - \bar{\ell}_w$ can change the target concept $w \setminus z \Rightarrow z$ without changing any other concept subordinate or causally separable to the target concept. Thus, $\bar{\ell}_z - \bar{\ell}_w$ is the linear representation $\bar{\ell}_{w \setminus z \Rightarrow z}$. This concept means $\texttt{not\_z} \Rightarrow \texttt{is\_z}$ conditioned on w, and hence it is subordinate to w.

Therefore, $\bar{\ell}_w$ is orthogonal to the linear representation $\bar{\ell}_{w \setminus z \Rightarrow z} = \bar{\ell}_z - \bar{\ell}_w$ by the property of the causal inner product. If they are not orthogonal, adding $\bar{\ell}_w$ can change the other concept $w \setminus z \Rightarrow z$, and it is a contradiction.

(c) By the above result (b), $\bar{\ell}_w^\top(\bar{\ell}_{z_1} - \bar{\ell}_w) = \bar{\ell}_w^\top(\bar{\ell}_{z_0} - \bar{\ell}_w) = 0$. Therefore, $\bar{\ell}_w^\top(\bar{\ell}_{z_1} - \bar{\ell}_{z_0}) = 0$.

(d) Let's say that w_1 is w_Z defined in Definition 2.1. The binary contrast $z_0 \Rightarrow z_1$ is subordinate to the binary feature for the attribute w_0. By the property of the causal inner product, $\bar{\ell}_{w_0}$ is orthogonal to the linear representation $\bar{\ell}_{z_0 \Rightarrow z_1} = \bar{\ell}_{z_1} - \bar{\ell}_{z_0}$ (by (a)). Then, with the above result (c), we have $(\bar{\ell}_{w_1} - \bar{\ell}_{w_0})^\top(\bar{\ell}_{z_1} - \bar{\ell}_{z_0})$.

(e) By the above result (b), we have

$$\begin{cases} \|\bar{\ell}_{w_1} - \bar{\ell}_{w_0}\|_2^2 = \|\bar{\ell}_{w_1}\|_2^2 - \|\bar{\ell}_{w_0}\|_2^2 \\ \|\bar{\ell}_{w_2} - \bar{\ell}_{w_1}\|_2^2 = \|\bar{\ell}_{w_2}\|_2^2 - \|\bar{\ell}_{w_1}\|_2^2 \\ \|\bar{\ell}_{w_2} - \bar{\ell}_{w_0}\|_2^2 = \|\bar{\ell}_{w_2}\|_2^2 - \|\bar{\ell}_{w_0}\|_2^2. \end{cases} \tag{D.10}$$

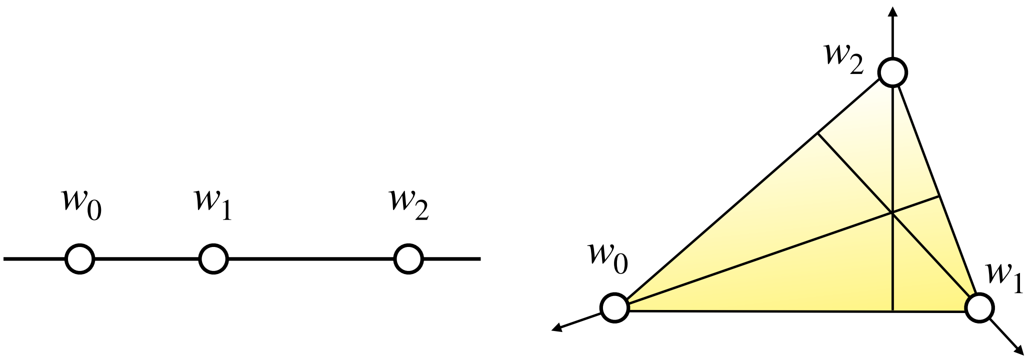


Figure 6. Illustration of the case $k = 3$ in the proof of Theorem 3.5.

Then,

$$\begin{aligned}
&\|\bar{\ell}_{w_1} - \bar{\ell}_{w_0}\|_2^2 + \|\bar{\ell}_{w_2} - \bar{\ell}_{w_1}\|_2^2 && \text{(D.11)}\\
&= \|\bar{\ell}_{w_1}\|_2^2 - \|\bar{\ell}_{w_0}\|_2^2 + \|\bar{\ell}_{w_2}\|_2^2 - \|\bar{\ell}_{w_1}\|_2^2 && \text{(D.12)}\\
&= \|\bar{\ell}_{w_2}\|_2^2 - \|\bar{\ell}_{w_0}\|_2^2 && \text{(D.13)}\\
&= \|\bar{\ell}_{w_2} - \bar{\ell}_{w_0}\|_2^2. && \text{(D.14)}
\end{aligned}$$

Therefore, $\bar{\ell}_{w_1} - \bar{\ell}_{w_0}$ is orthogonal to $\bar{\ell}_{w_2} - \bar{\ell}_{w_1}$.

□

D.3. Proof of Theorem 3.5

Theorem 3.5 (Categorical Concepts are Represented as Simplices)**.** *Suppose that $\{w_0, \dots, w_{k-1}\}$ is a collection of k mutually exclusive attributes such that for every joint distribution $Q(w_0, \dots w_{k-1})$ there is some ℓ_i such that $\mathbb{P}(W = w_i \mid \ell_i) = Q(W = w_i)$ for every i. Then, the vector representations $\bar{\ell}_{w_0}, \dots, \bar{\ell}_{w_{k-1}}$ form a $(k-1)$-simplex in the representation space. In this case, we take the simplex to be the representation of the categorical concept $W = \{w_0, \dots, w_{k-1}\}$.*

Proof. If we can represent arbitrary joint distributions, this means, in particular, that we can change the probability of one attribute without changing the relative probability between a pair of other attributes. Consider the case where $k = 3$, as illustrated in Figure 6. If $\bar{\ell}_{w_0}, \bar{\ell}_{w_1}, \bar{\ell}_{w_2}$ are on a line, then there is no direction in that line (to change the value in the categorical concept) such that adding the direction can change the probability of w_2 without changing the relative probabilities between w_0 and w_1. However, if $\bar{\ell}_{w_0}, \bar{\ell}_{w_1}, \bar{\ell}_{w_2}$ are not on a line, they form a triangle. Then, there exists a line that is toward $\bar{\ell}_{w_2}$ and perpendicular to the opposite side of the triangle. Now adding the direction $\tilde{\ell}$ can manipulate the probability of w_2 without changing the relative probabilities between w_0 and w_1. That is, for any $\alpha > 0$ and context embedding ℓ,

$$\begin{cases} \mathbb{P}(W = w_2 \mid \ell + \alpha\tilde{\ell}) > \mathbb{P}(W = w_2 \mid \ell), \text{ and} \\ \frac{\mathbb{P}(W = w_1 \mid \ell + \alpha\tilde{\ell})}{\mathbb{P}(W = w_0 \mid \ell + \alpha\tilde{\ell})} = \frac{\mathbb{P}(W = w_1 \mid \ell)}{\mathbb{P}(W = w_0 \mid \ell)}. \end{cases} \tag{D.15}$$

Therefore, the vectors $\bar{\ell}_{w_0}, \bar{\ell}_{w_1}, \bar{\ell}_{w_2}$ form a 2-simplex.

This argument extends immediately to higher k by induction. For each $i \in \{0, \dots, k-1\}$, there should exist a direction that is toward $\bar{\ell}_{w_i}$ and orthogonal to the opposite hyperplane ($(k-2)$-simplex) formed by the other $\bar{\ell}_{w_{i'}}$'s. Then, the vectors $\bar{\ell}_{w_0}, \dots, \bar{\ell}_{w_{k-1}}$ form a $(k-1)$-simplex. □

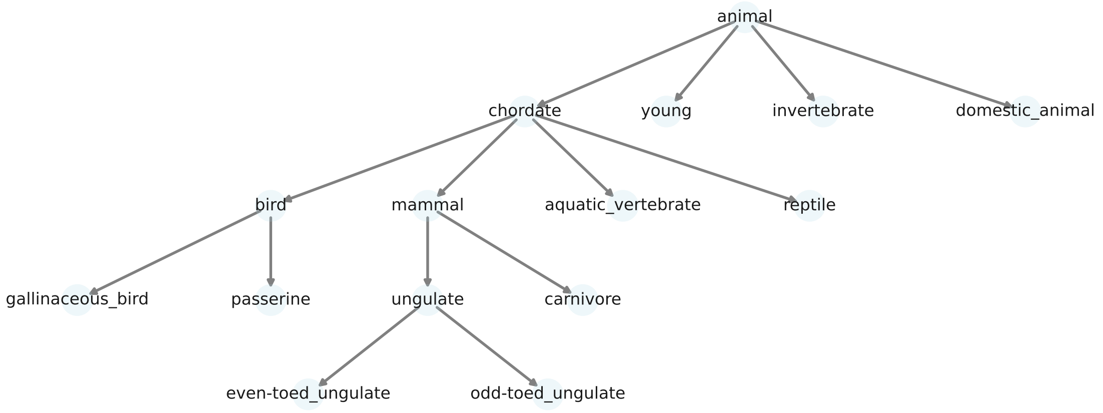


Figure 7. Subtree in WordNet noun hierarchy for descendants of `animal`.

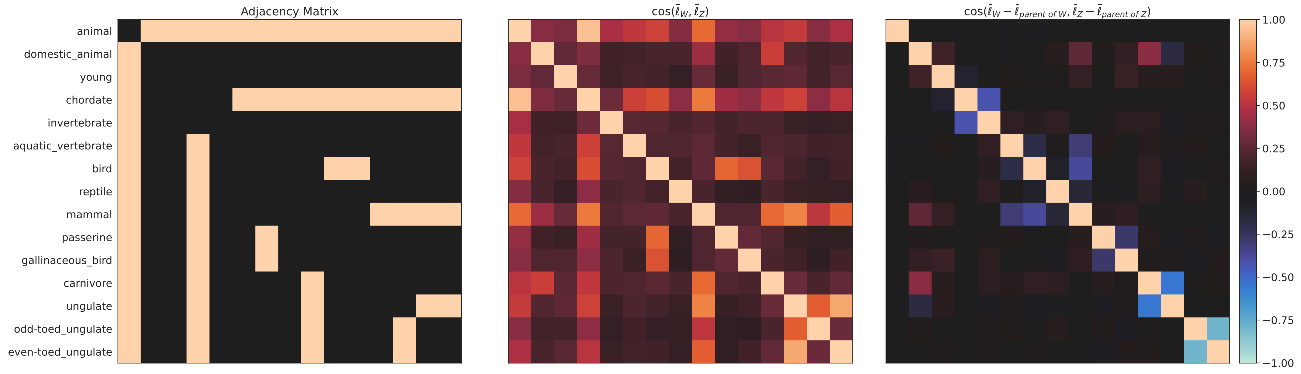


Figure 8. Zoomed-in Heatmaps of the subtree for `animal` in Figure 7.

E. Experiment Details

We employ the `Gemma-2B` version of the Gemma model (Mesnard et al., 2024), which is accessible online via the `huggingface` library. Its two billion parameters are pre-trained on three trillion tokens. This model utilizes 256K tokens and 2,048 dimensions for the representation space.

We always use tokens that start with a space ('\u2581') in front of the word, as they are used for next-word generation with full meaning. Additionally, like WordNet data we use, we include plural forms, and both capital and lowercase versions of the words in $\mathcal{Y}(\texttt{animal})$ and $\mathcal{Y}(\texttt{plant})$ for visualization in Appendix B.1.

In the WordNet synset data, each content of the synset `mammal.n.01` indicates that "mammal" is a word, "n" denotes "noun," and "01" signifies the first meaning of the word. In the WordNet hierarchy, if a parent has only one child, we combine the two features into one. Additionally, since the WordNet hierarchy is not a perfect tree, a child can have more than one parent. We use one of the parents when computing the $\bar{\ell}_w - \bar{\ell}_{\text{parent of } w}$.

Code is available at github.com/KihoPark/LLM_Categorical_Hierarchical_Representations.

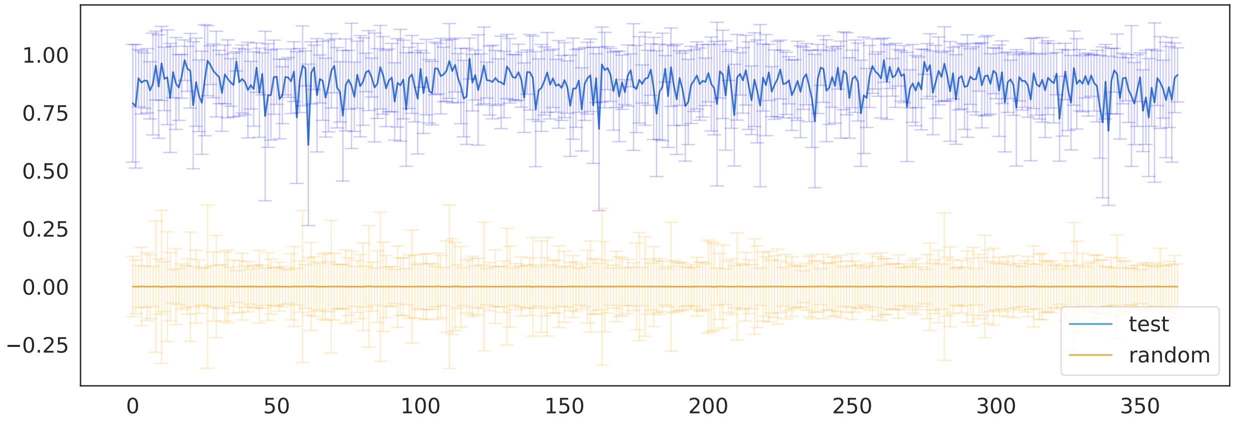


Figure 9. Linear representations exist for most binary features in the WordNet verb hierarchy. Comparison of projection of test and random words on estimated vector representations for each feature. The values are divided by the norm of the estimated vector representation. The x-axis indices denote all features in the verb hierarchy. The thick lines present the mean of the projections for each feature and the error bars indicate the 1.96 × standard error.

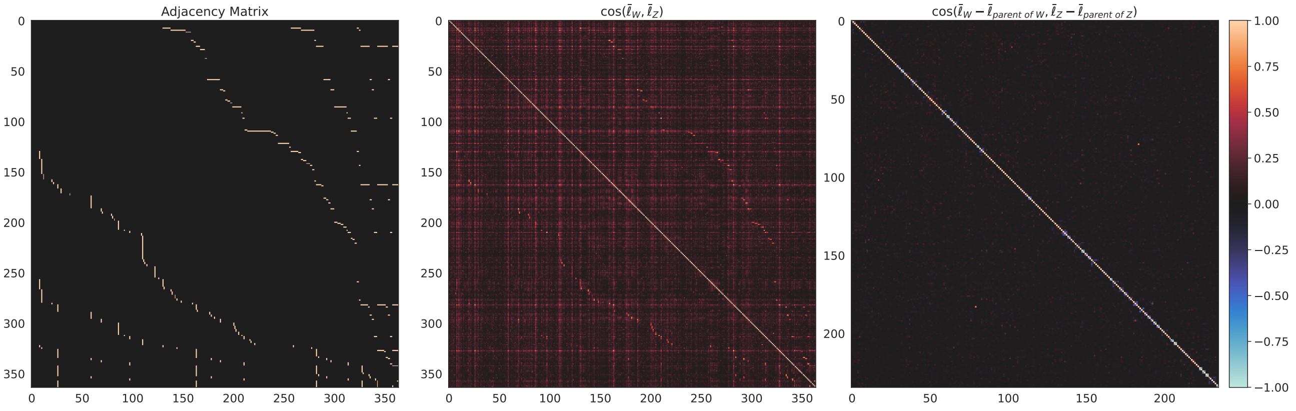


Figure 10. The adjacency matrix of the hierarchical relations between features in the WordNet verb hierarchy (left), the cosine similarity between the vector representations $\bar{\ell}_w$ for each feature (middle), and the cosine similarity between child-parent vectors $\bar{\ell}_w - \bar{\ell}_{\text{parent of } w}$ for each feature (right). The features are ordered by the hierarchy.

F. Additional Results

F.1. Zooming in on a Subtree of Noun Hierarchy

As it is difficult to understand the entire WordNet hierarchy at once from the heatmaps in Figure 5, we present a zoomed-in heatmap for the subtree (Figure 7) for the feature `animal` in Figure 8. The left heatmap displays the adjacency matrix of the hierarchical relations between features, aligned with the subtree in Figure 7. The middle heatmap shows that the cosine similarities between the vector representations $\bar{\ell}_w$ correspond to the adjacency matrix. The final heatmap demonstrates that the child-parent vector $\bar{\ell}_w - \bar{\ell}_{\text{parent of } w}$ and $\bar{\ell}_{\text{parent of } w} - \bar{\ell}_{\text{grandparent of } w}$ are orthogonal, as predicted in Theorem 3.3.

F.2. WordNet Verb Hierarchy

In the same way as for the noun hierarchy, we estimate the vector representations for the WordNet verb hierarchy. To evaluate whether vector representations exist, we split $\mathcal{Y}(w)$ for each synset w into train words (80%) and test words (20%), fit the LDA estimator to the train words, and examine the projection of the unembedding vectors for the test words onto the estimated vector representation. Figure 9 shows the mean and standard error of the test projections, divided by the magnitude of each estimated $\bar{\ell}_w$. If a vector representation exists for an attribute, we would expect these values to be close to 1. This is indeed the case, providing evidence that vector representations do indeed exist for these features.

Figure 10 displays the adjacency matrix of the WordNet verb hyponym inclusion graph (left), the cosine similarity between the vector representations $\bar{\ell}_w$ for each feature (middle), and the cosine similarity between child-parent vectors $\bar{\ell}_w - \bar{\ell}_{\text{parent of } w}$ for each feature (right). The cosine similarity clearly reflects the semantic hierarchy—the adjacency matrix is clearly visible in the middle heatmap. By contrast, as predicted by Theorem 3.3, the child-parent and parent-grandparent vectors are orthogonal. This straightforwardly implies all other theoretical connections between orthogonality and the semantic hierarchy.