The Geometry of Categorical and Hierarchical Concepts in Large Language Models

ICLR 2025
April 25, Singapore


Kiho Park
UChicago
Graduating next year!


Yo Joong "YJ" Choe
UChicago / INSEAD

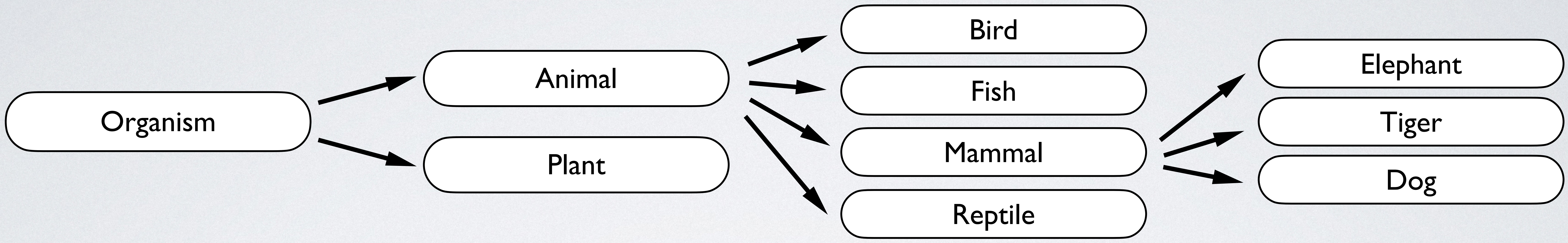

Yibo Jiang
UChicago


Victor Veitch
UChicago / Google

The "Big Picture" Question

How is semantic meaning encoded in the representation spaces of LLMs?

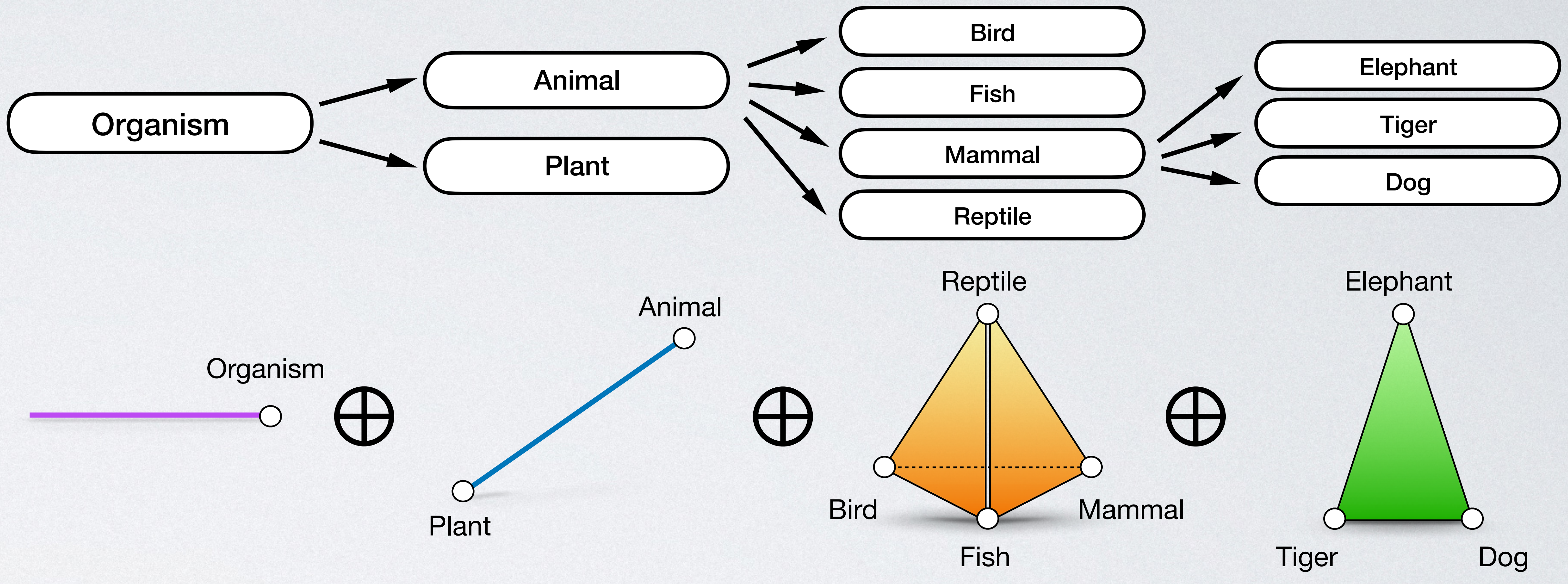
Key Questions in This Work



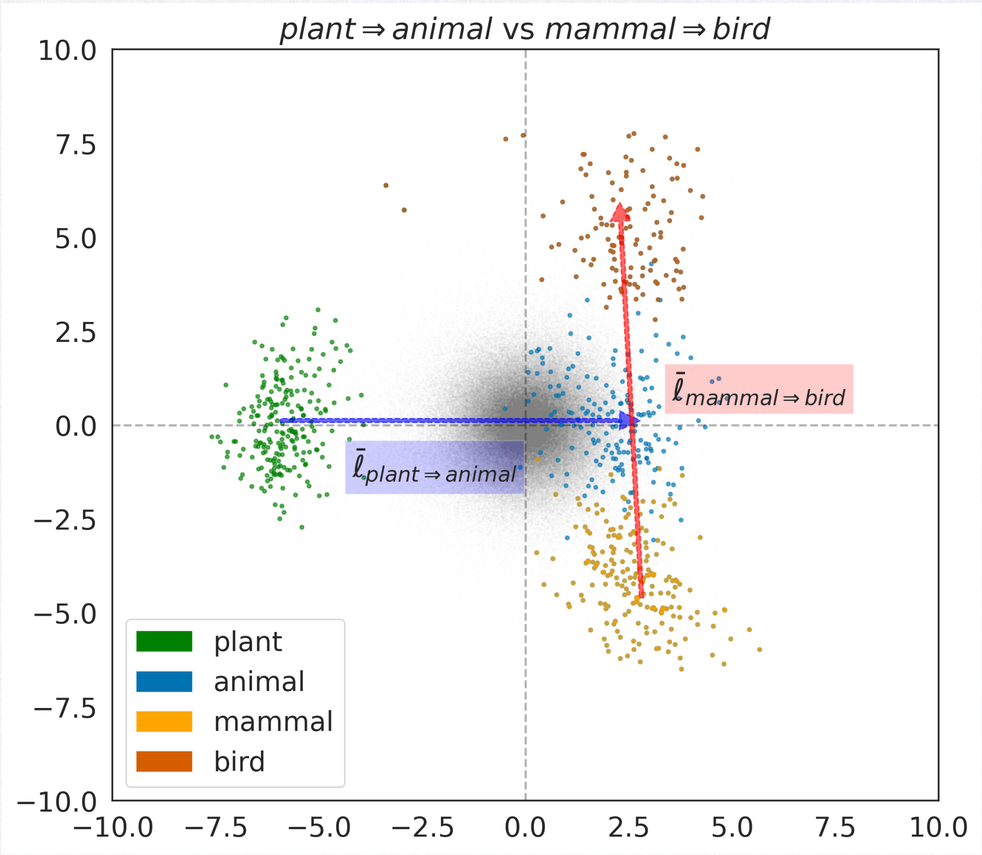
1. What is a representation of a single feature (e.g., *is_animal*)?
2. How are categorical concepts represented?
3. How are hierarchical relations between concepts represented?

Summary of Contributions

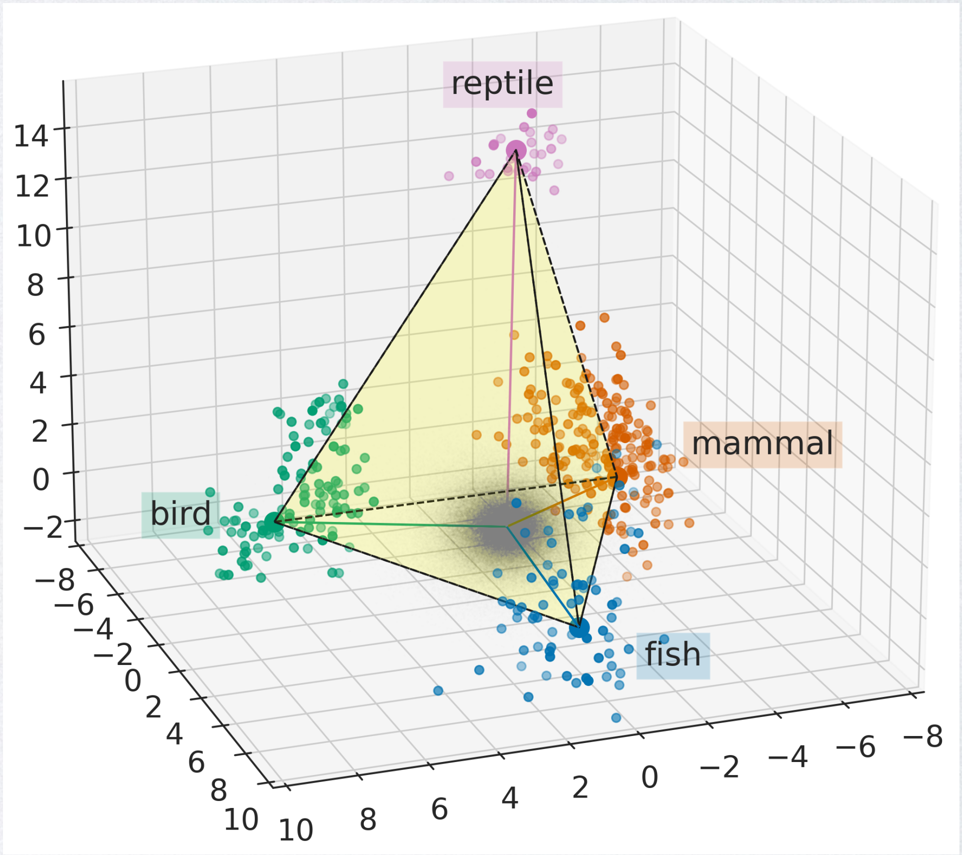
1. Single features are represented as *vectors*
2. Categorical concepts are represented as *polytopes*
3. Hierarchical relations are represented as *orthogonality*



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



(b) Hierarchy is encoded as orthogonality in Gemma.

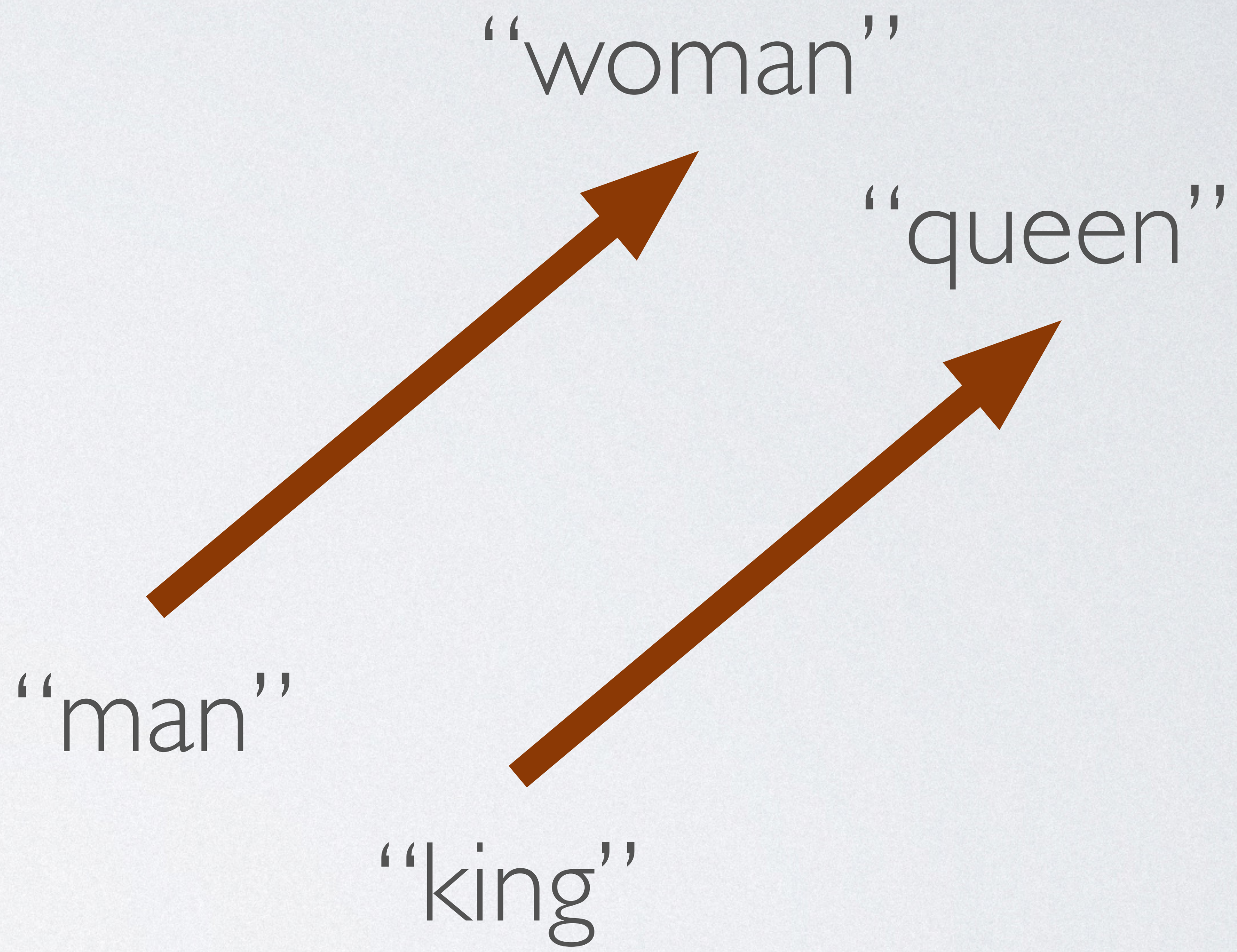


(c) Categorical concepts are represented as polytopes in Gemma.

Background

Background 1. Linear Representation Hypothesis

"High-level concepts are represented linearly as ***directions*** in the representation space"

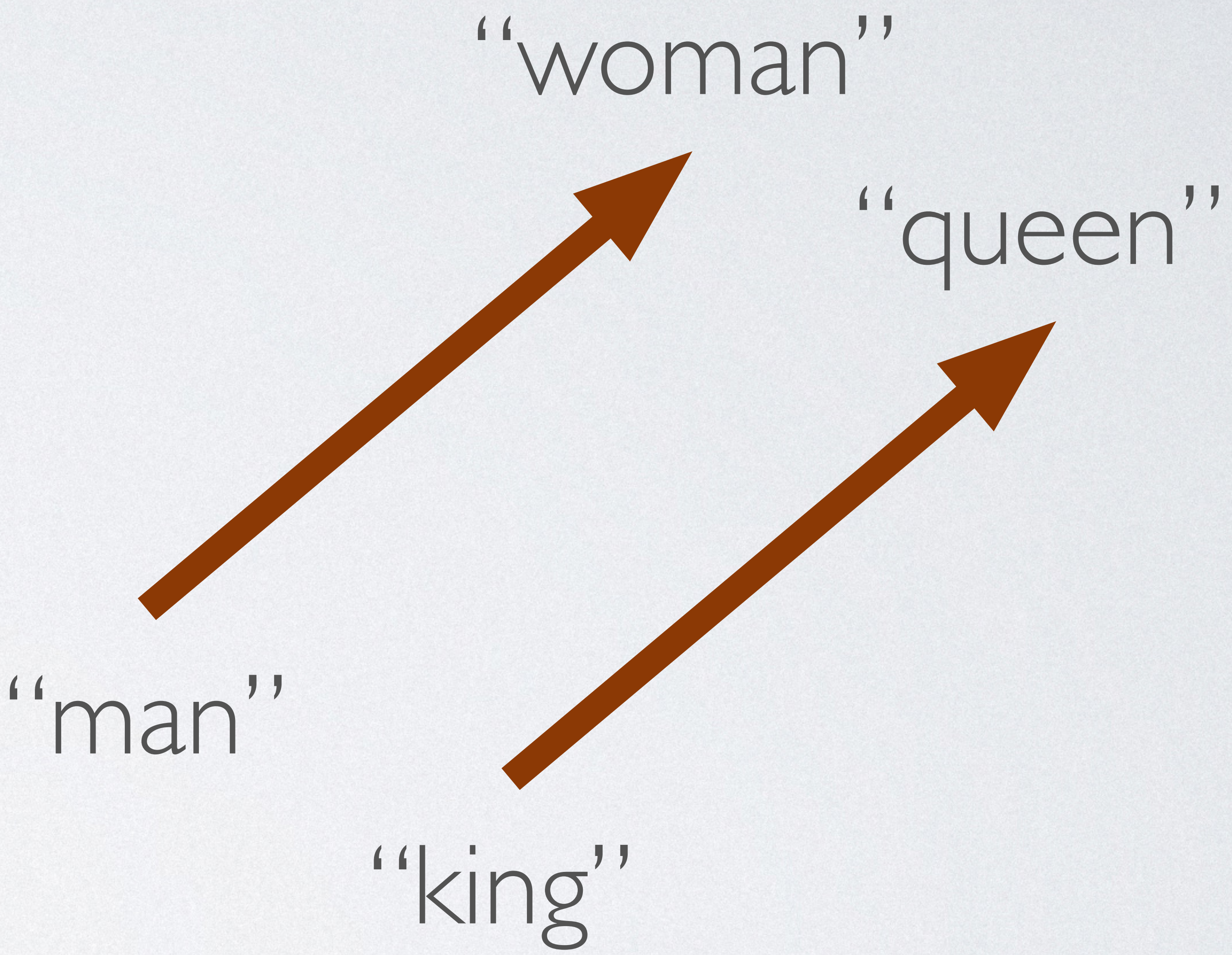


Mikolov et al., 2013; Arora et al., 2016; Elhage et al., 2022; …

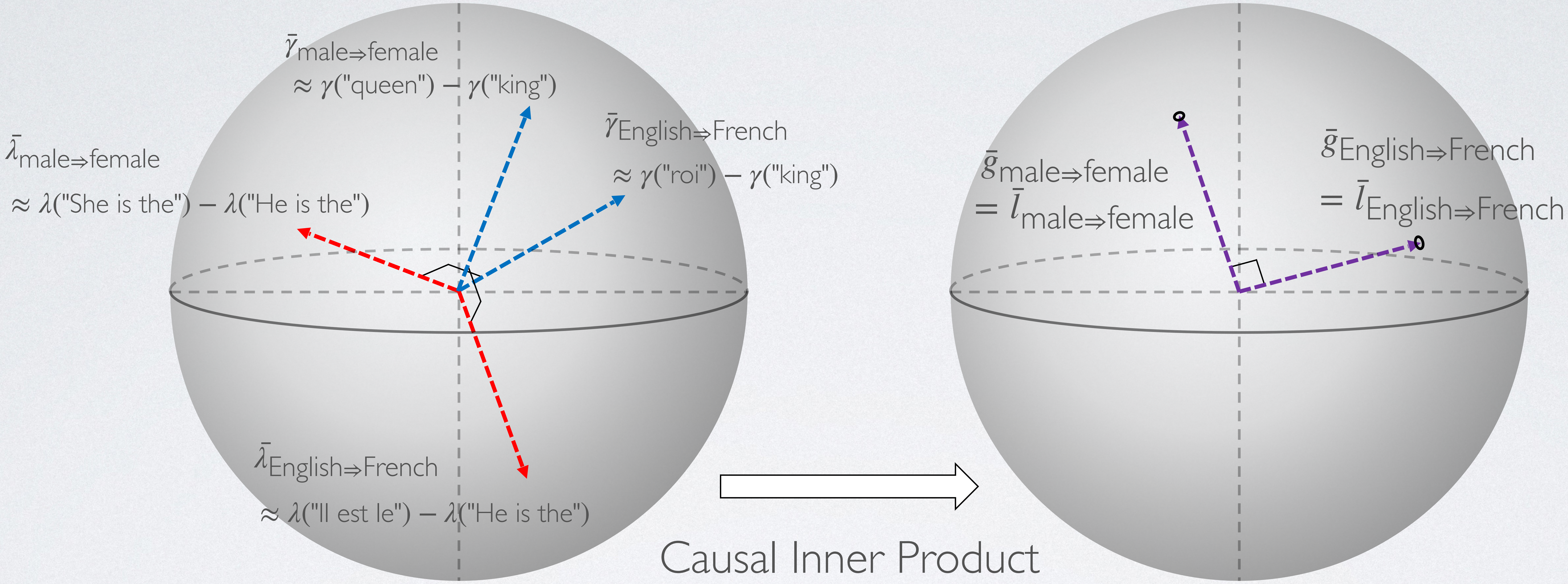
Background 1. Linear Representation Hypothesis

"High-level concepts are represented linearly as ***directions*** in the representation space"

But… a vector = direction + ***magnitude***



Background 2. Causal Inner Product



Embedding

$l(x) \in \mathbb{R}^d$

Softmax

$\mathbb{P}(y \mid x) \propto \exp(l(x)^\top g(y))$

Unembedding

$g(y) \in \mathbb{R}^d$

Result 1. Vector Representations of Binary Features

Linear Representations of Binary Concepts

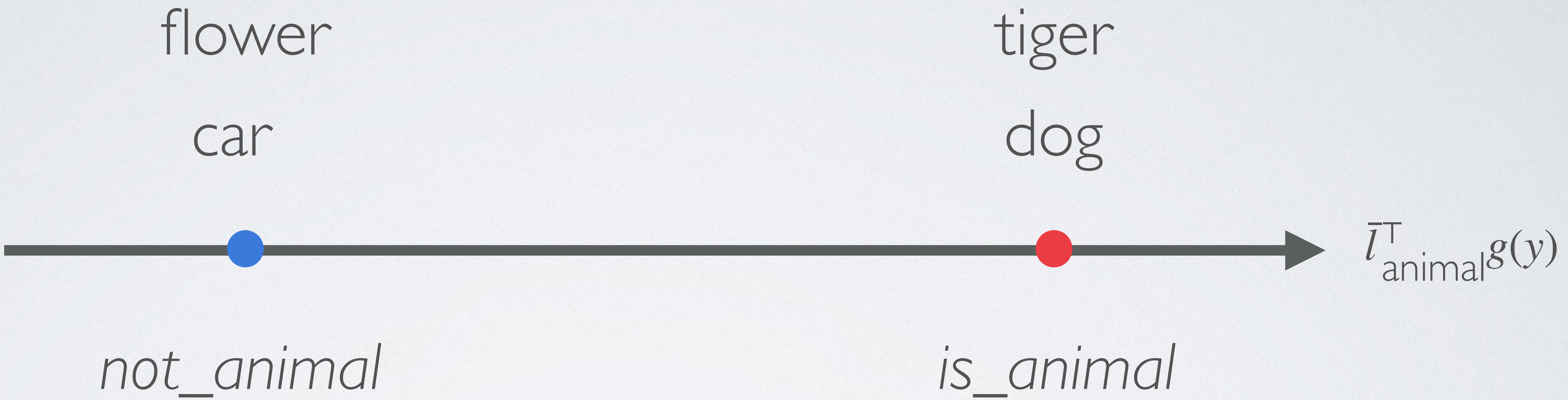
Desideratum: If a linear representation exists as a direction, moving an embedding vector in this direction should modify the probability of the target concept ***in isolation***

Result 1. Vector Representations of Binary Features

Logits: $l(\text{"I have a"})^\top g(y) \xrightarrow{?} \left(l(\text{"I have a"}) + \alpha \bar{l}_{\text{animal}}\right)^\top g(y)$

See Theorem 4 in the paper

Result 1. Vector Representations of Binary Features



See Theorem 4 in the paper

Result 1. Vector Representations of Binary Features

See Theorem 4 in the paper

Result 1. Vector Representations of Binary Features

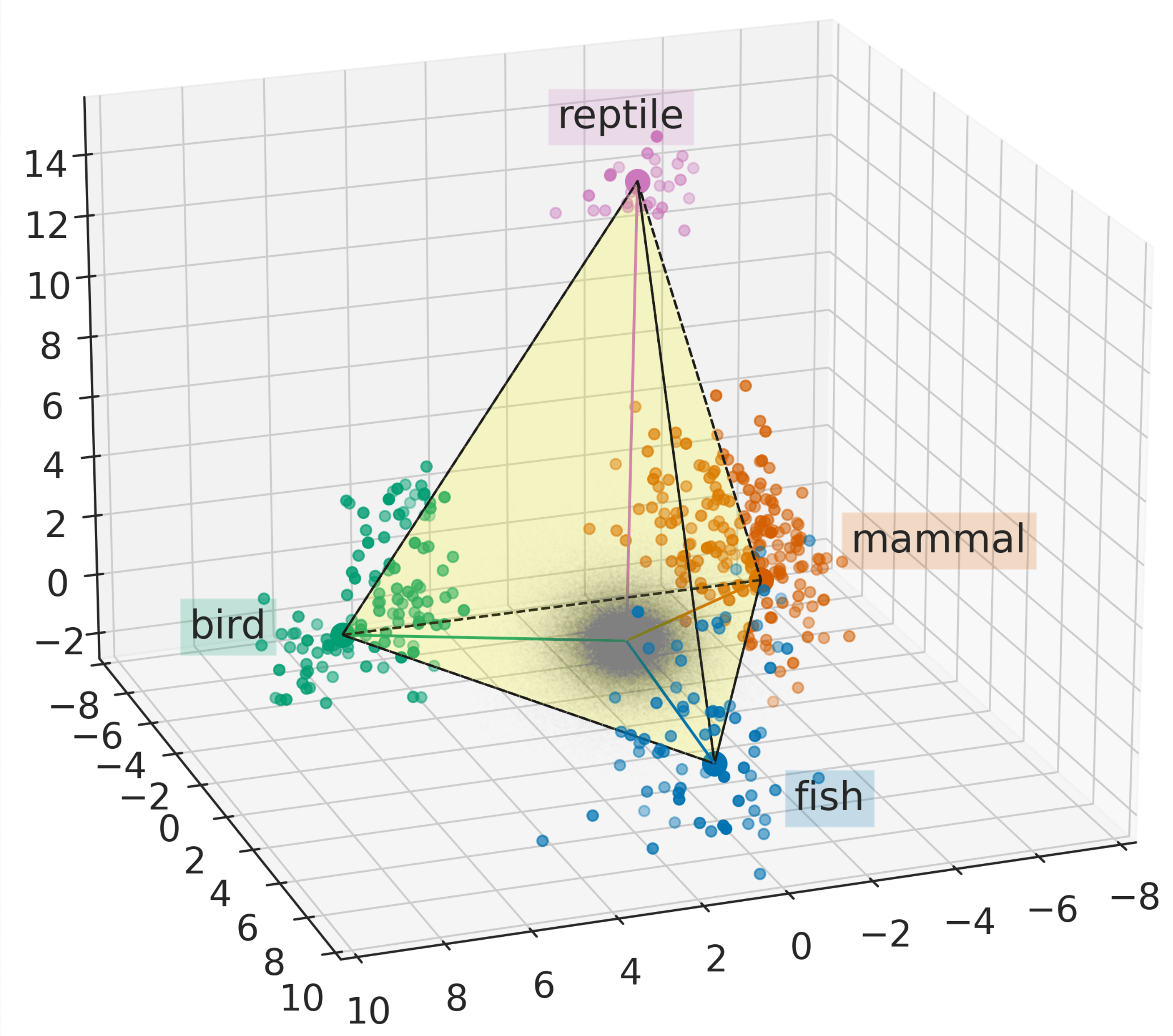
A binary feature $W = \{\text{not\_w}, \text{is\_w}\}$ has a vector representation $\bar{\ell}_w$ if it is a linear representation for W with an associated *magnitude*.

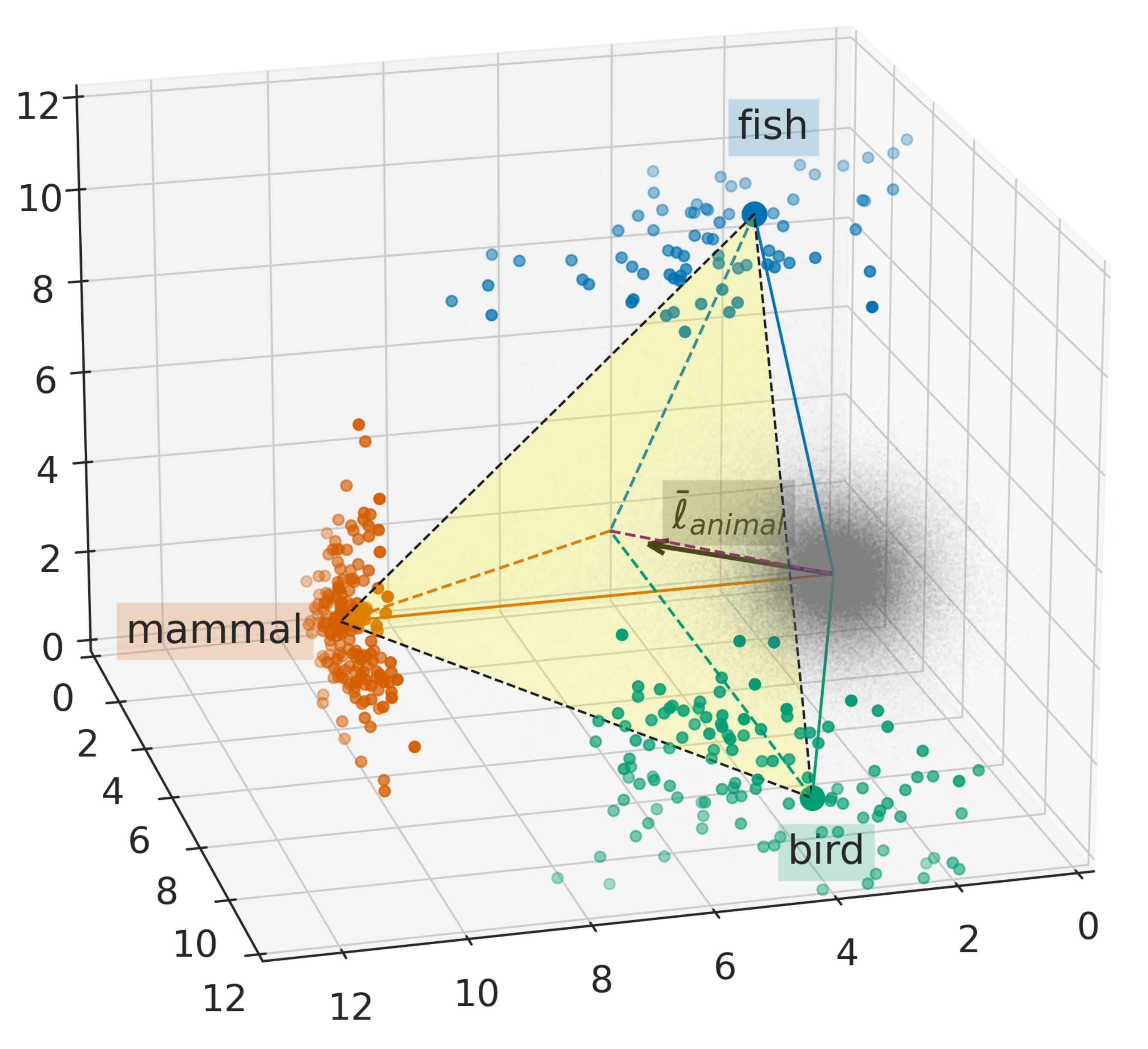
See Theorem 4 in the paper

Result 2. Polytope Representations of Categorical Concepts

Result 2. Polytope Representations of Categorical Concepts

The polytope representation of a categorical concept $\{w_0, \ldots, w_{k-1}\}$ is the convex hull of vector representations $\bar{\ell}_{w_0}, \ldots, \bar{\ell}_{w_{k-1}}$ for each element.





Experiment with Gemma-2B model

Result 3. Hierarchical Semantics Are Encoded as Orthogonality

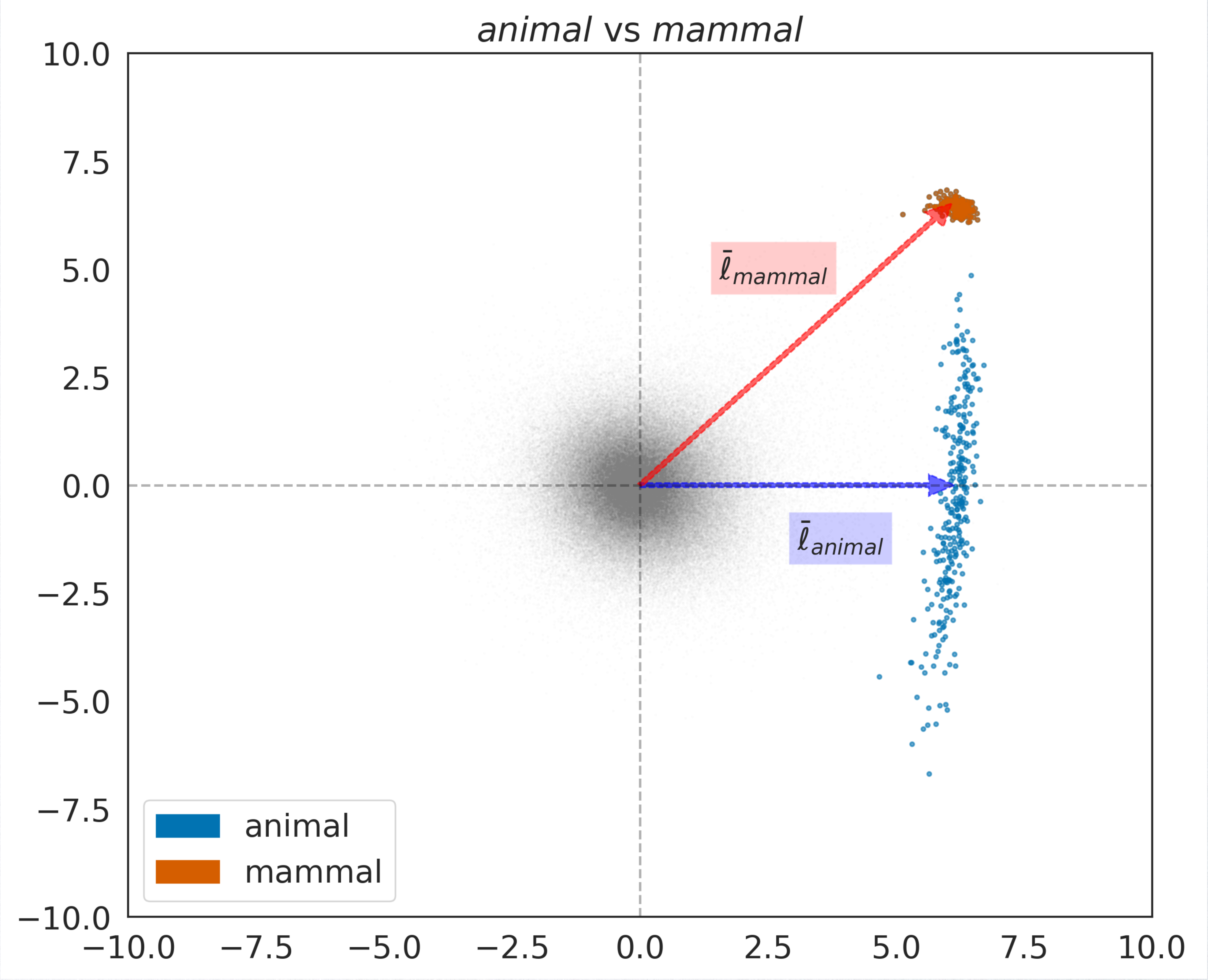
Result 3. Hierarchical Semantics Are Represented As Orthogonality

(a) $\bar{l}_w \perp \bar{l}_z - \bar{l}_w$ for $z \prec w$ (e.g., $\bar{l}_{animal} \perp \bar{l}_{mammal} - \bar{l}_{animal}$)

(b) $\bar{l}_w \perp \bar{l}_{z_1} - \bar{l}_{z_0}$ for $Z \in_R \{z_0, z_1\} \prec W \in_R \{\text{not\_w}, \text{is\_w}\}$

(c) $\bar{l}_{w_1} - \bar{l}_{w_0} \perp \bar{l}_{z_1} - \bar{l}_{z_0}$ for $Z \in_R \{z_0, z_1\} \prec W \in_R \{w_0, w_1\}$

(d) $\bar{l}_{w_1} - \bar{l}_{w_0} \perp \bar{l}_{w_2} - \bar{l}_{w_1}$ for $w_2 \prec w_1 \prec w_0$



See Theorem 8 in the paper; experiment with Gemma-2B model

Result 3’. Cosine Similarities Between Vector Representations Capture Their Semantic Relations

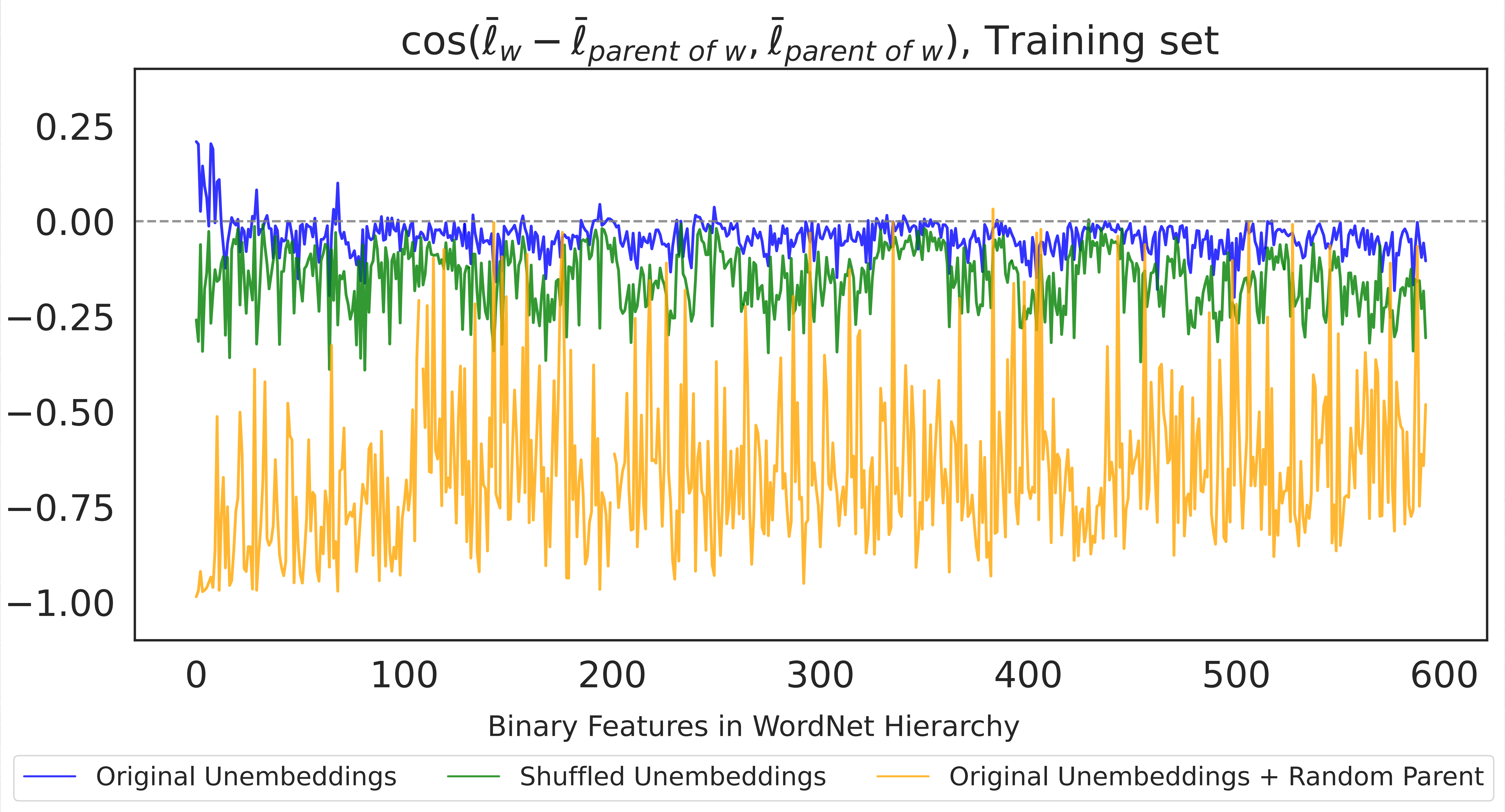
Shortest-path distances between features in WordNet hierarchy

Cosine similarities between vector representations

Cosine similarities between *shuffled* vector representations

Experiment with Gemma-2B model

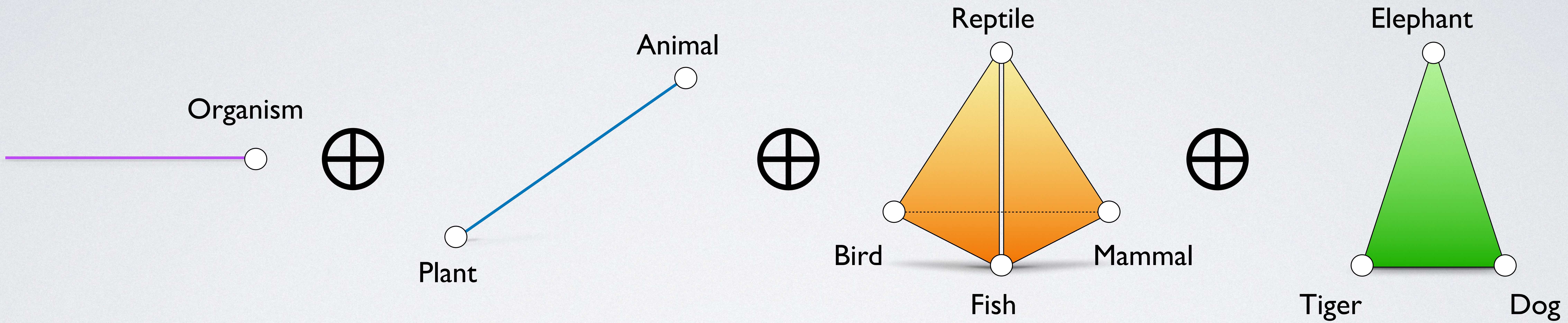
Result 3''. Validating Hierarchical Orthogonality on the Full WordNet Hierarchy



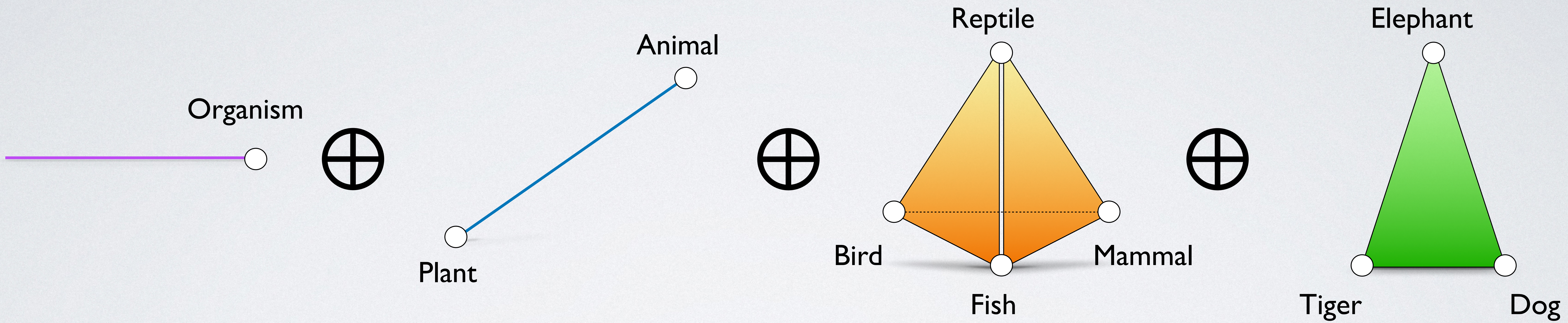
Experiment with Gemma-2B model and WordNet Hierarchy

How is semantic meaning encoded in the representation spaces of LLMs?

How is semantic meaning encoded in the representation spaces of LLMs?



How is semantic meaning encoded in the representation spaces of LLMs?



The Geometry of Categorical and Hierarchical Concepts in Large Language Models

Kiho Park, Yo Joong Choe, Yibo Jiang, Victor Veitch

arXiv:2406.01506

Appendix

Result 3. Hierarchical Semantics Are Represented As Orthogonality

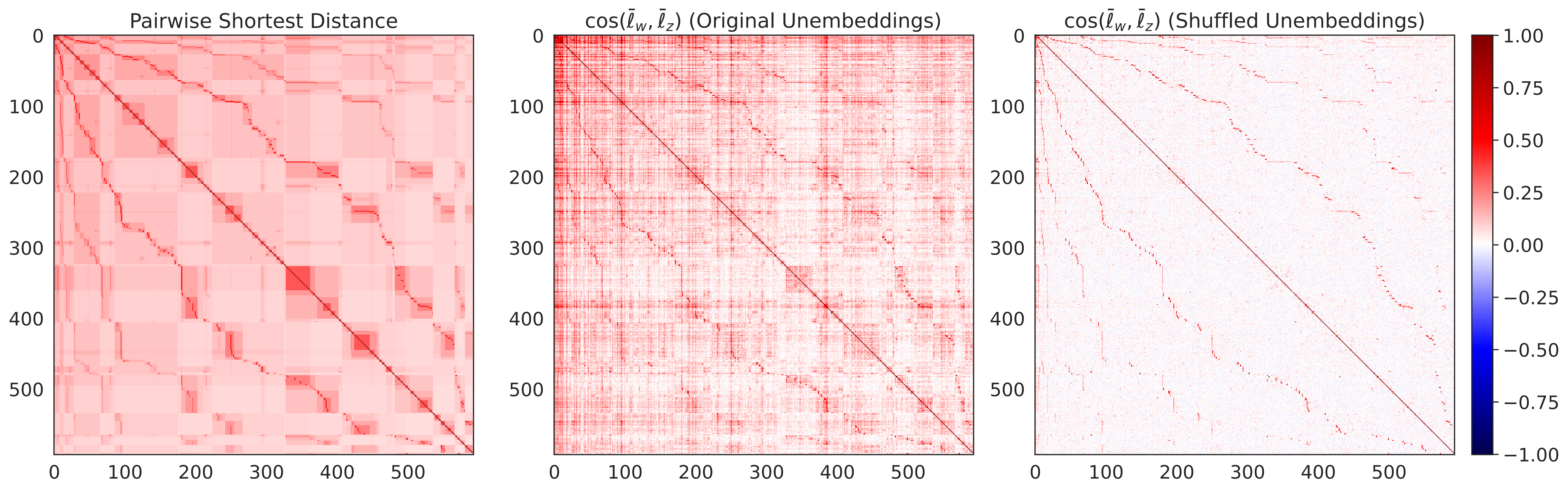
(a) $\bar{l}_w \perp \bar{l}_z - \bar{l}_w$ for $z \prec w$
(e.g., $\bar{l}_{animal} \perp \bar{l}_{mammal} - \bar{l}_{animal}$)

(b) $\bar{l}_w \perp \bar{l}_{z_1} - \bar{l}_{z_0}$ for
$Z \in_R \{z_0, z_1\} \prec W \in_R \{\text{not\_w}, \text{is\_w}\}$

(c) $\bar{l}_{w_1} - \bar{l}_{w_0} \perp \bar{l}_{z_1} - \bar{l}_{z_0}$ for
$Z \in_R \{z_0, z_1\} \prec W \in_R \{w_0, w_1\}$

See Theorem 8 in the paper; experiment using Gemma-2B model

Result 3’. Cosine Similarities Between Vector Representations Capture Their Semantic Relations



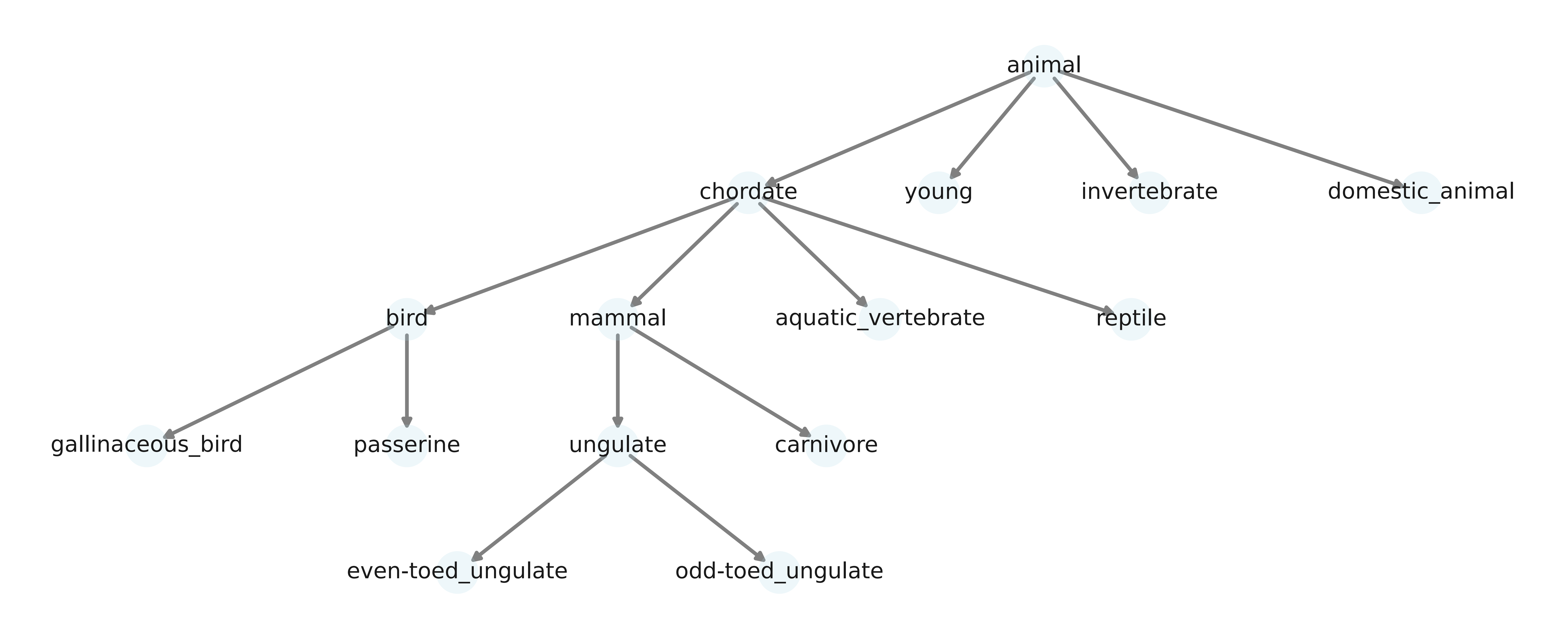
Shortest-path distances between features in WordNet hierarchy

Cosine similarities between vector representations

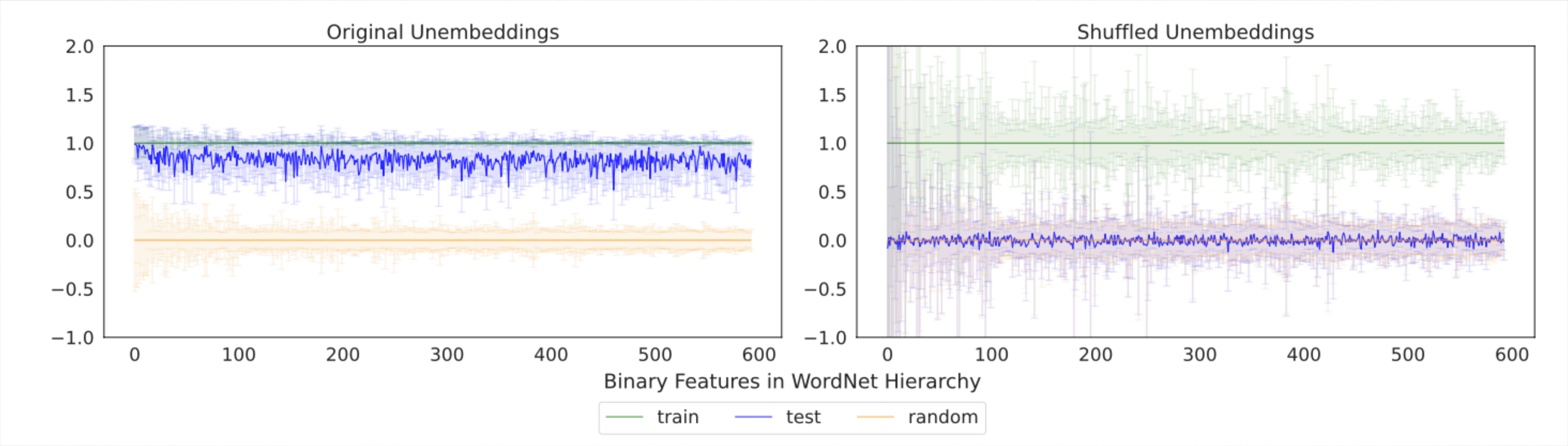
Cosine similarities between *shuffled* vector representations

Experiment with Gemma-2B model

A Subgraph of the WordNet Hierarchy



Existence of Vector Representations



Theoretical Prediction: $(\bar{\ell}_w)^\top g(y)/\|\bar{\ell}_w\|^2 = 1$ for any $y \in \mathcal{Y}(w)$.

Experiment with Gemma-2B model

Proof Sketch for the "Magnitude Theorem"

Theorem: Linear Representations have Magnitude

If ℓ_W is a linear representation of binary feature W then there is some $b_w > 0$ such that

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

Proof Sketch

Adding $\bar{\ell}_{\texttt{animal}}$ shouldn't change probability of "dog" vs "cat".

Softmax gives:

$$\log \frac{\mathrm{P}(\text{“dog”} \mid \ell(x) + \alpha \bar{\ell}_{\texttt{animal}})}{\mathrm{P}(\text{“cat”} \mid \ell(x) + \alpha \bar{\ell}_{\texttt{animal}}} = \ell(x)^\top (g_{\texttt{dog}} - g_{\texttt{cat}}) + \alpha \bar{\ell}_{\texttt{animal}}^\top (g_{\texttt{dog}} - g_{\texttt{cat}}).$$

This expression is free of α only if $\bar{\ell}_{\texttt{animal}}^\top g_{\texttt{dog}} = \bar{\ell}_{\texttt{animal}}^\top g_{\texttt{cat}}$.

The End