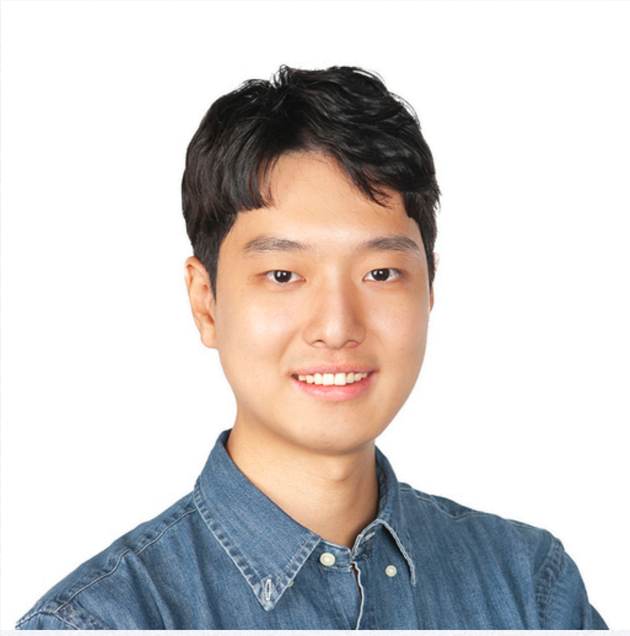
# The Geometry of Categorical and Hierarchical Concepts in Large Language Models

ICML 2024 Workshop on Mechanistic Interpretability
July 27, 2024, Vienna, Austria


Kiho Park
Stat @ UChicago

Yo Joong (YJ) Choe
DSI @ UChicago

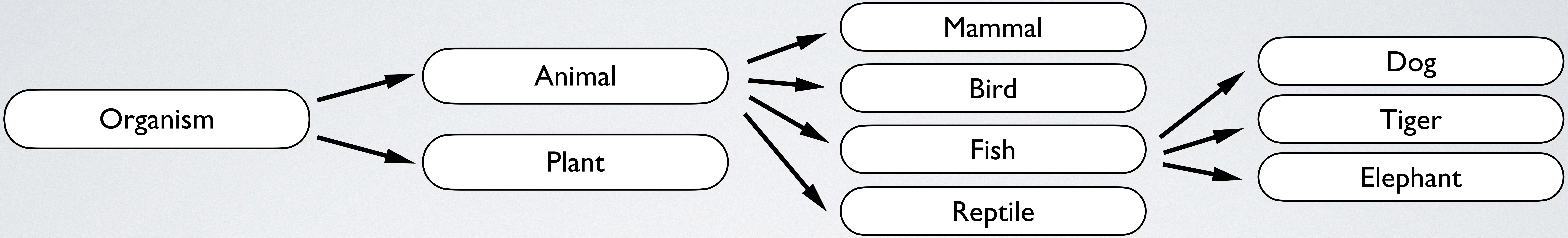
Yibo Jiang
CS @ UChicago

Victor Veitch
Stat & DSI @ UChicago

# Big Picture

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

# Extending the Linear Representation Hypothesis to Categorical and Hierarchical Concepts




How are categorical concepts represented?

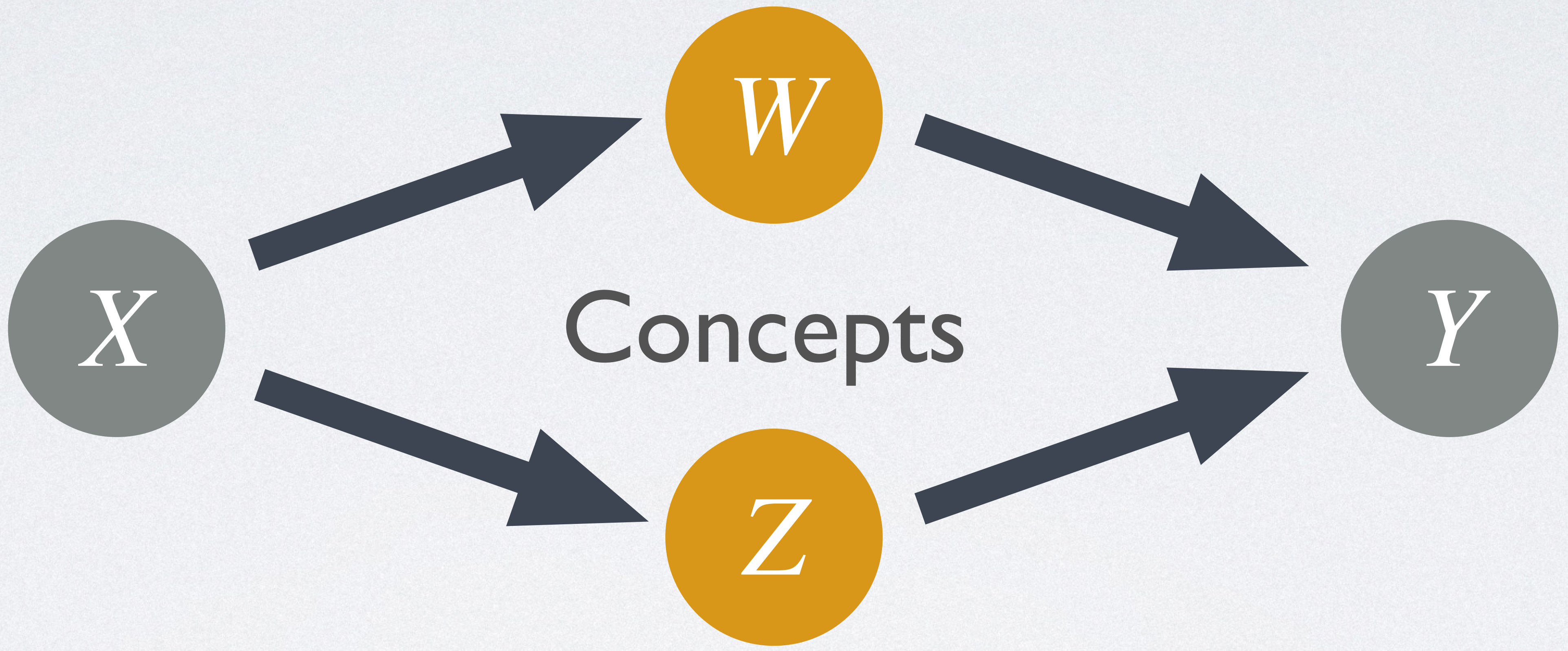
How are hierarchical relations between concepts represented?

*Challenge: a linear direction can only encode a binary concept*

# Background: Softmax Structure

Context

"He is the"




Next word

"king"

"man"

"PhD"

⋮

**Embedding**

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

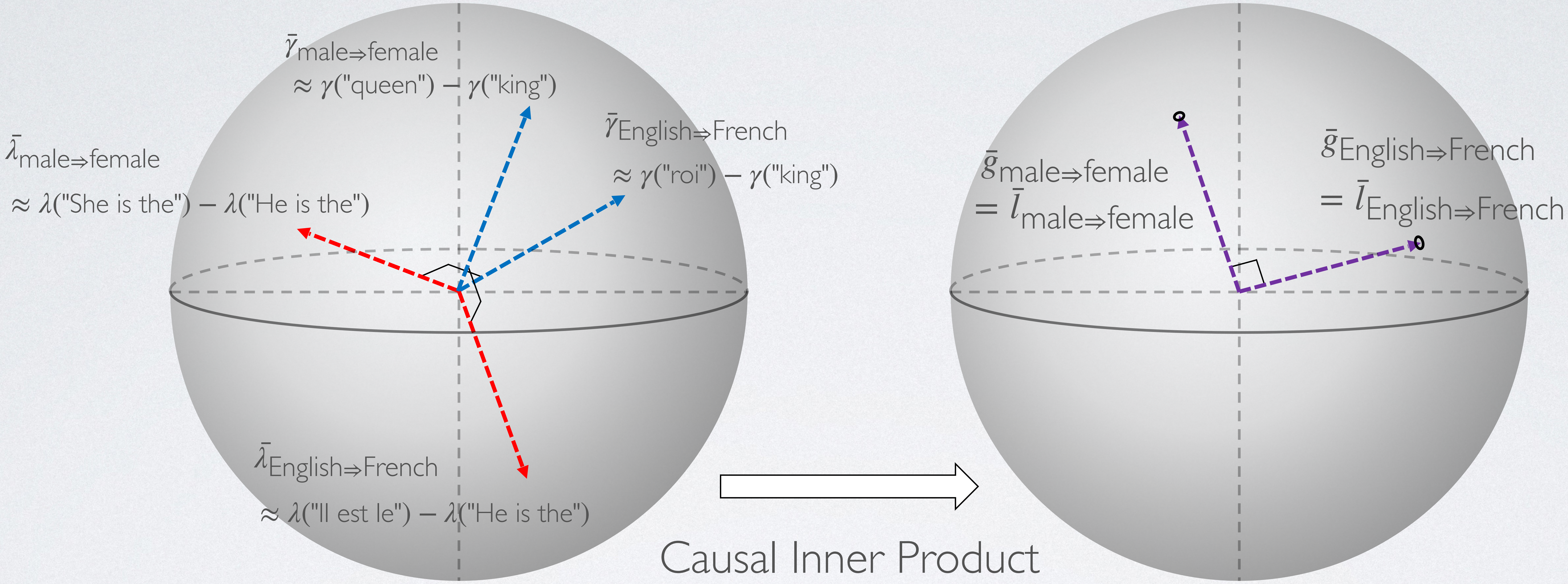
**Softmax**

$\mathbb{P}(y \mid x) \propto \exp(\lambda(x)^\top \gamma(y))$

**Unembedding**

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

# Background: Causal Inner Product (Park et al., 2024)




**Embedding**

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

**Softmax**

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

**Unembedding**

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

# How to Build up from Binary Concepts?

## Categorical Concepts

{mammal, bird, fish, reptile}

### Binary Concepts

#### Binary Contrast

male ⇒ female

mammal ⇒ bird

#### Binary Feature

{not_female, is_female}

{not_bird, is_bird}

# Hierarchical Structure

$Z$ is "subordinate" to $W$

$$Z = \text{dog} \Rightarrow \text{cat} \prec W = \{\text{not\_mammal}, \text{is\_mammal}\}$$

$$Z = \text{parrot} \Rightarrow \text{eagle} \prec W = \{\text{mammal}, \text{bird}, \text{fish}\}$$

# Linear Representation $\bar{l}_W$ of Binary Concept

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

$$\mathbb{P}(W = 1 \mid l + \alpha \bar{l}_W) > \mathbb{P}(W = 1 \mid l)$$

$$\mathbb{P}(Z \mid l + \alpha \bar{l}_W) = \mathbb{P}(Z \mid l)$$

$\forall l, \alpha > 0, Z$ subordinate to or causally separable with $W$

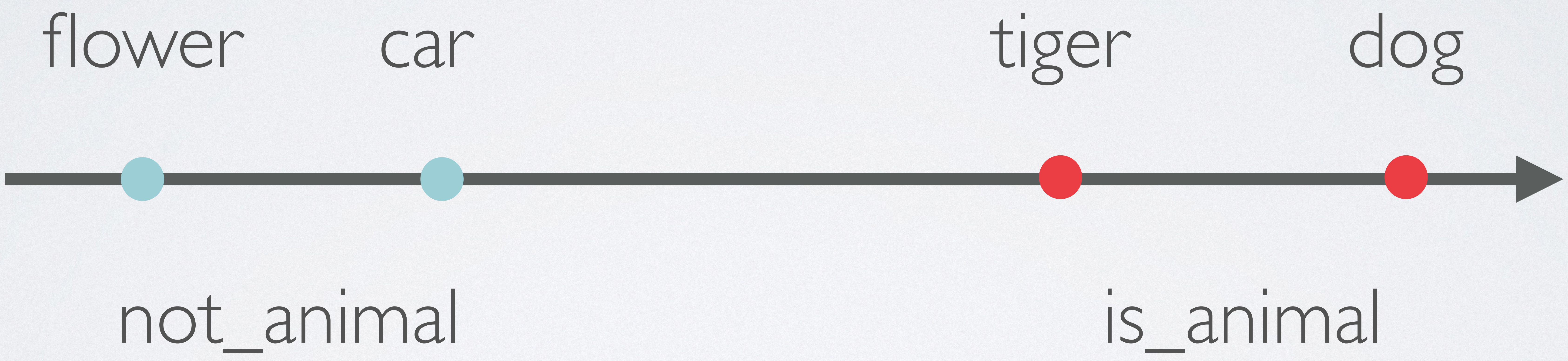
# Representations of Complex Concepts

How to compose representations of binary concepts?

*Challenge: linear representations are directions without magnitude*

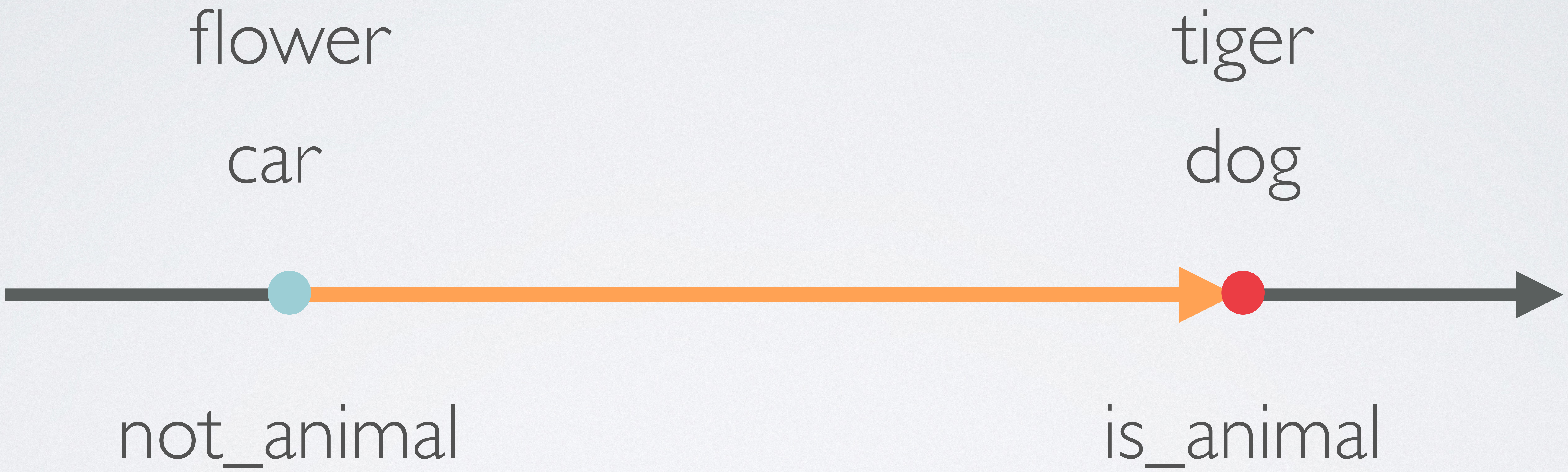
# Vector Representation $\bar{l}_w$ of Binary Feature

Theorem 4



# Vector Representation $\bar{l}_w$ of Binary Feature

Theorem 4



# Main Result 1: Semantic Hierarchy is Encoded as Orthogonality

Theorem 6

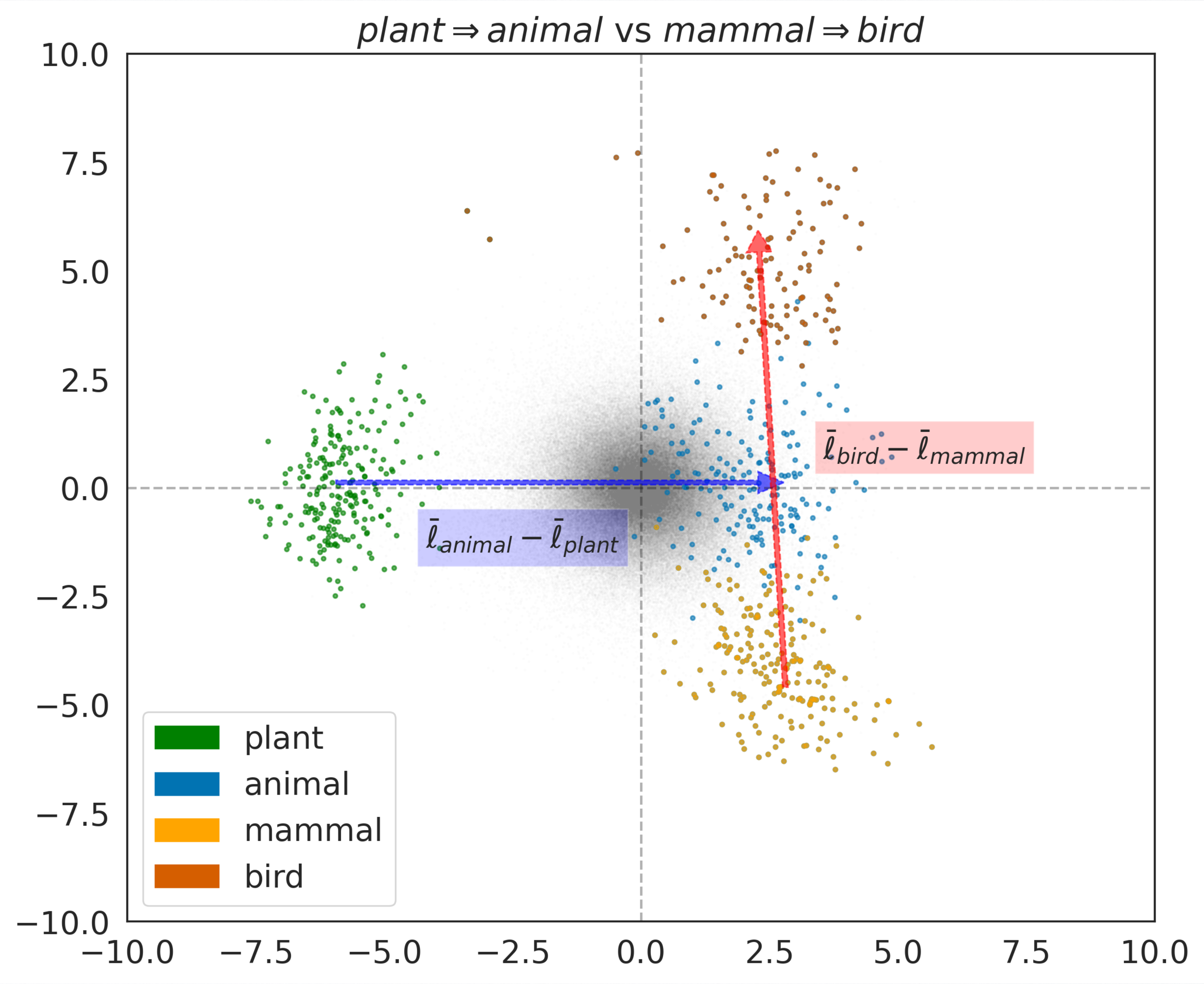
(a) $\bar{l}_{w_1} - \bar{l}_{w_0}$ is a linear representation $\bar{l}_{w_0 \Rightarrow w_1}$

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

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

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

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




Experiment with Gemma-2B model

# Main Result 1: Theoretical Predictions Hold on the Full WordNet Hierarchy

Theorem 6

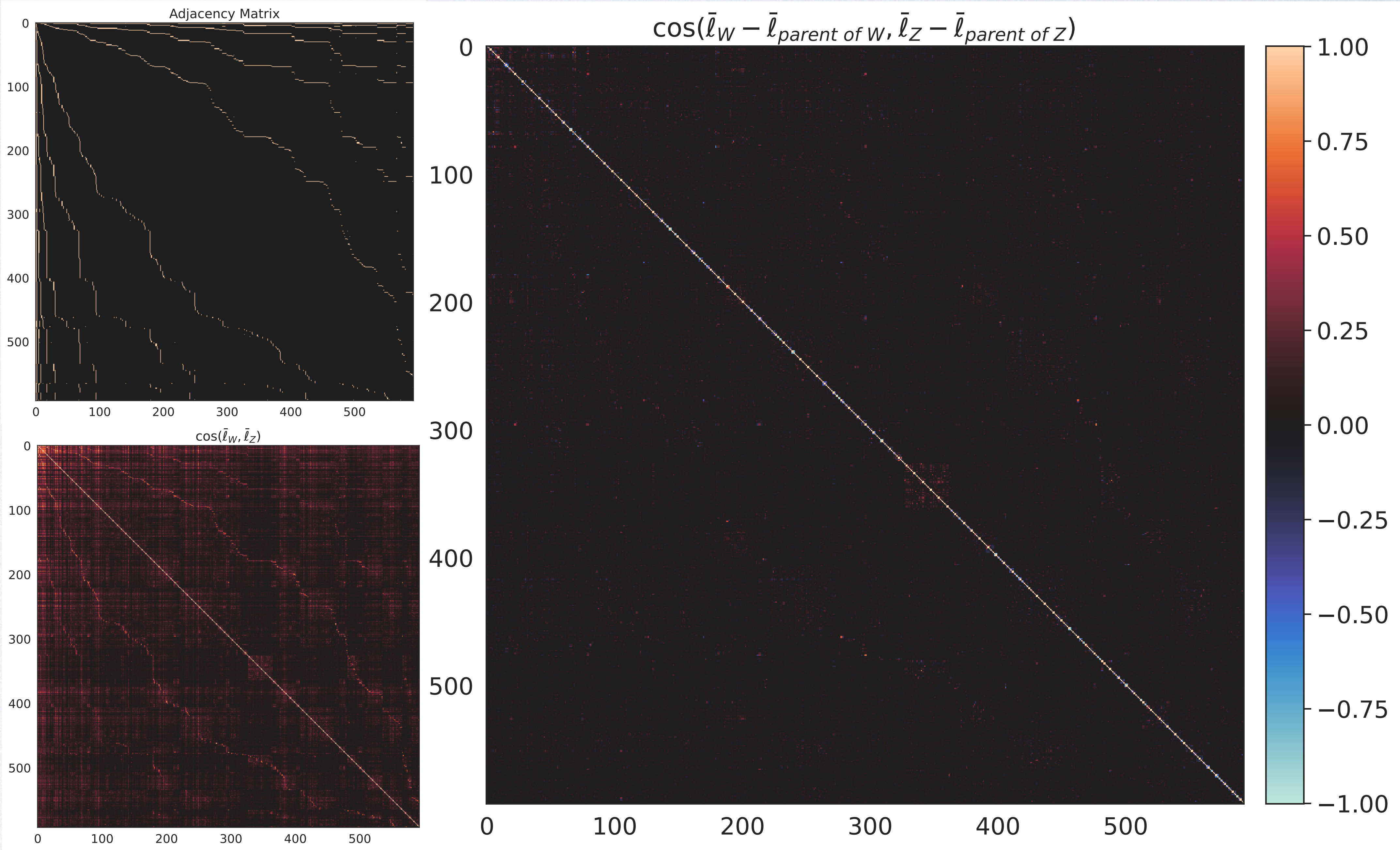
(a) $\bar{l}_{w_1} - \bar{l}_{w_0}$ is a linear representation $\bar{l}_{w_0 \Rightarrow w_1}$

(b) $\bar{l}_{w} \perp \bar{l}_{z} - \bar{l}_{w}$ for $z \prec w$

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

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

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




*Subtracting the parent of the feature gives orthogonality*

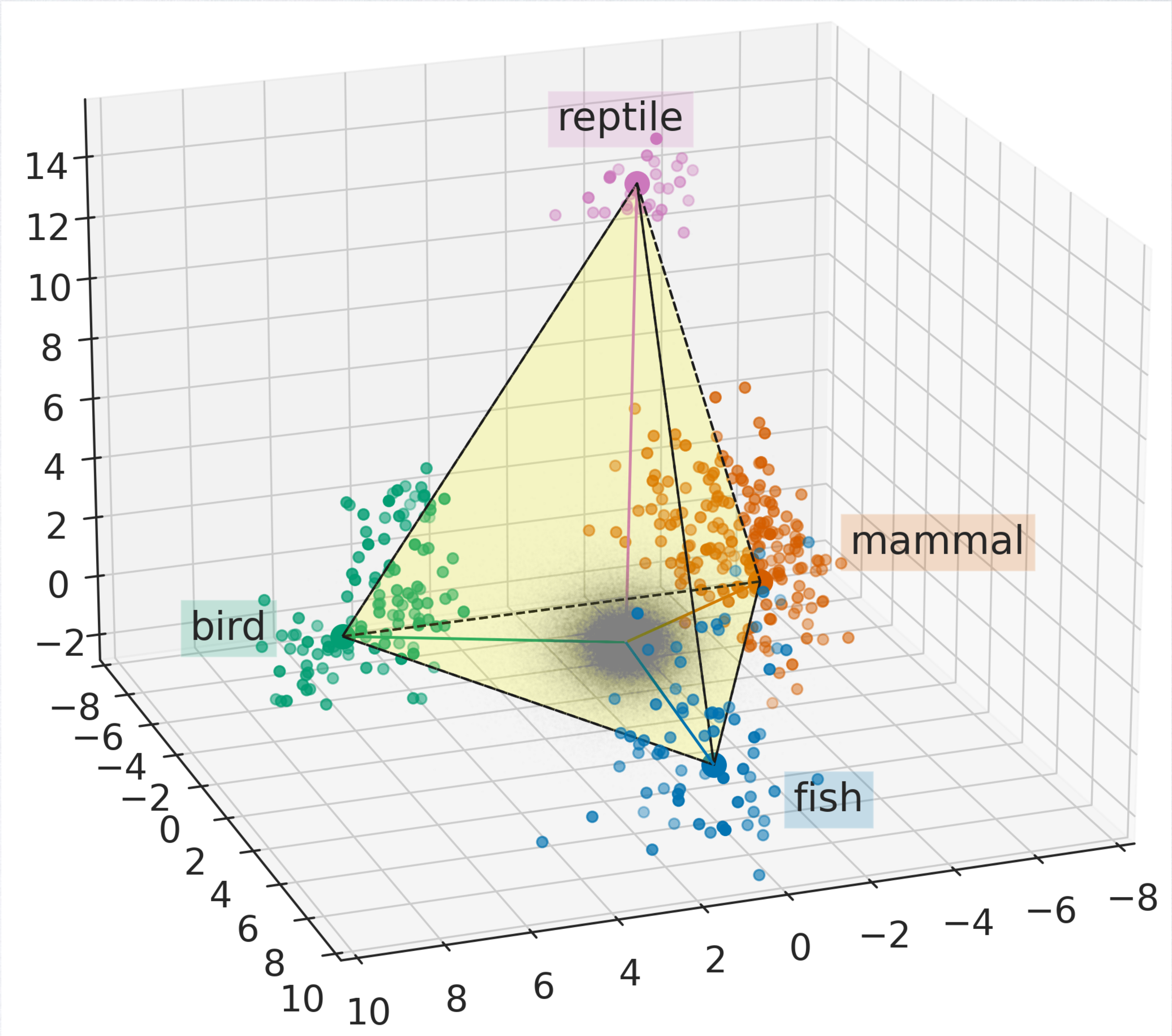


Experiment with Gemma-2B model

# Main Result 2: Natural Categorical Concepts are Encoded as Simplices

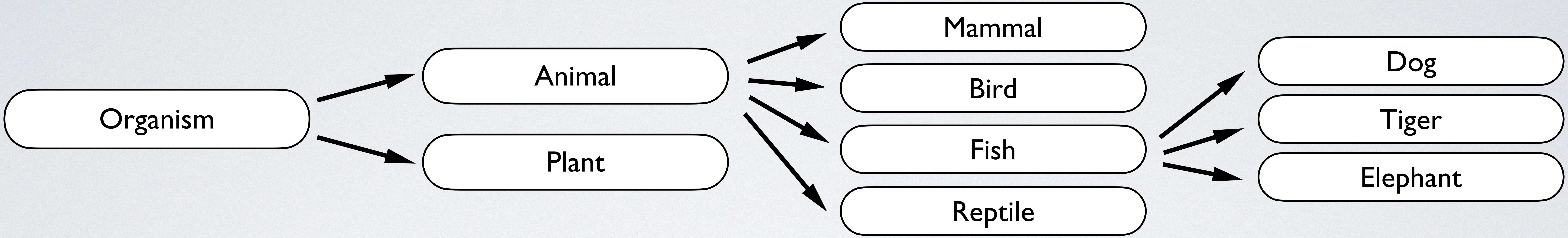
Theorem 8

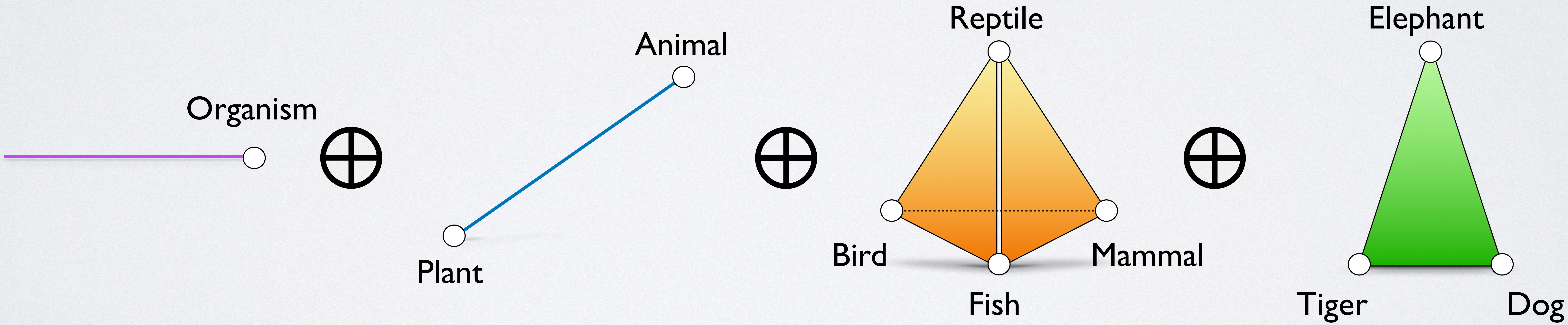
For every joint distribution $Q(w_0, \ldots, w_{k-1})$, if there exists some $l_i$ such that $\mathbb{P}(W = w_i \mid l_i) = Q(W = w_i)$ for every $i$, the vector representations $\bar{l}_{w_0}, \ldots, \bar{l}_{w_{k-1}}$ form a $(k-1)$-simplex in the representation space. In this case, we take the simplex to be the representation of the categorical concept $W = \{w_0, \ldots, w_{k-1}\}$.




Experiment with Gemma-2B model

# Overall Structure



# Summary

- Categorical Concepts are Represented as Simplices
- Hierarchical Relations are encoded as orthogonality

**The Geometry of Categorical and Hierarchical Concepts in Large Language Models**

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

arXiv:2406.01506