Linear Representation Hypothesis & Geometry of LLMs

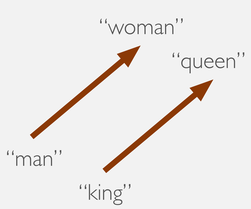
Kiho Park, Yo Joong Choe, Victor Veitch

University of Chicago

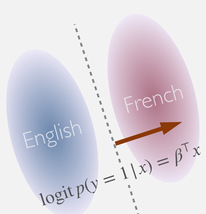
What Does "Linear" Even Mean?

There are three natural ways to define the notion of linear representation.

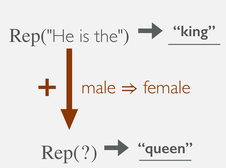
Subspace
(word2vec)



Measurement
(linear probe)



Intervention
(steering vector)

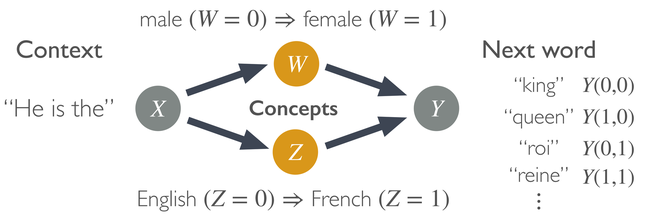


Problem: It is not clear how these ideas relate to each other, nor which is the right notion of linear representation.

Concepts in LLMs

LLMs generate the next word using the softmax distribution.

Embedding	Softmax	Unembedding
$\lambda(x) \in \mathbb{R}^d$	$\mathbb{P}(y \mid x) \propto \exp(\lambda(x)^\top \gamma(y))$	$\gamma(y) \in \mathbb{R}^d$



A concept W is defined by counterfactual outputs $Y(W=0), Y(W=1)$.
Concepts W and Z are ***causally separable*** if $Y(w,z)$ is well-defined, that is, they can be manipulated freely and in isolation.

Formalizing Linear Representation Hypothesis

We first formalize the subspace notions of linear representations, then use softmax structure to connect them to measurement and intervention.

Unembedding Representation $\bar{\gamma}_W$

$\gamma(Y(1)) - \gamma(Y(0)) = \alpha \bar{\gamma}_W \quad (\alpha > 0)$

for any counterfactual pairs $(Y(0), Y(1))$

Theorem

Measurement Representation

$\text{logit}\, \mathbb{P}(W=1 \mid \lambda) = \alpha \lambda^\top \bar{\gamma}_W$

Embedding Representation $\bar{\lambda}_W$

$\lambda_1 - \lambda_0 = \alpha' \bar{\lambda}_W \quad (\alpha' > 0)$

for any counterfactual pairs (λ_0, λ_1)

s.t. $\frac{\mathbb{P}(W=1 \mid \lambda_1)}{\mathbb{P}(W=1 \mid \lambda_0)} > 1$ and $\frac{\mathbb{P}(Z \mid W, \lambda_1)}{\mathbb{P}(Z \mid W, \lambda_0)} = 1$

Theorem

Intervention Representation

$\mathbb{P}(W=1 \mid Z, \lambda + c\bar{\lambda}_W)$ increasing in $c \in \mathbb{R}$

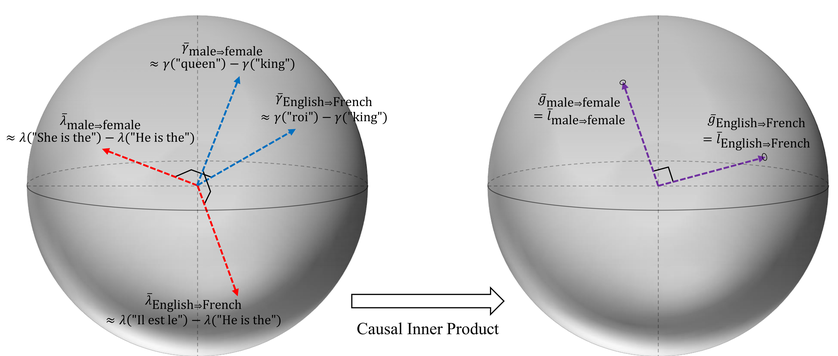
$\mathbb{P}(Z=1 \mid W, \lambda + c\bar{\lambda}_W)$ constant in $c \in \mathbb{R}$

Causal Inner Product

Problems:

How do the unembedding and embedding representations relate?

What is the right inner product for the representation space?



We introduce the causal inner product $\langle \cdot, \cdot \rangle_C$ such that $\langle \bar{\gamma}_W, \bar{\gamma}_Z \rangle_C = 0$ whenever W and Z are causally separable.

Theorem

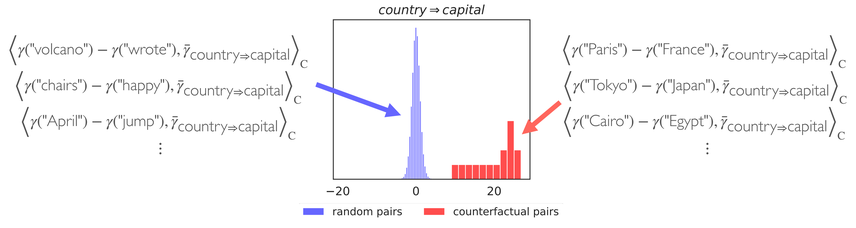
This unifies the unembedding and embedding representations via $\langle \bar{\gamma}_W, \cdot \rangle_C = (\bar{\lambda}_W)^\top$. (This is the Riesz isomorphism.)
We can estimate the causal inner product as $\langle \bar{\gamma}, \bar{\gamma}' \rangle_C = \bar{\gamma}^\top \text{Cov}(\gamma)^{-1} \bar{\gamma}'$.

Experiments with LLaMA-2

We estimate the unembedding representations for various concepts by using the counterfactual pairs from a word analogy dataset.

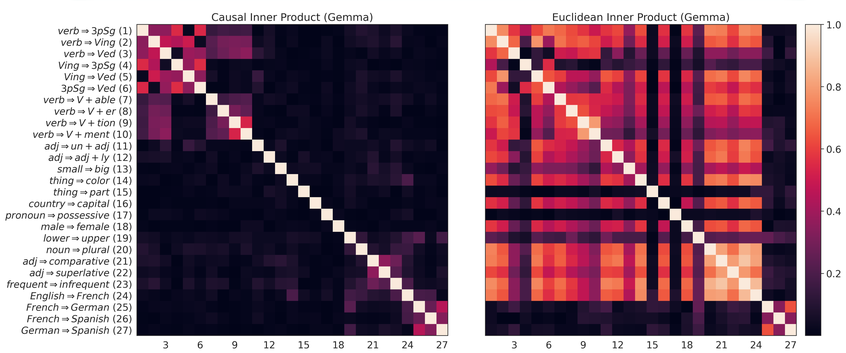
$$\bar{\gamma}_W := \frac{\tilde{\gamma}_W}{\langle \tilde{\gamma}_W, \tilde{\gamma}_W \rangle_C} \text{ where } \tilde{\gamma}_W = \frac{1}{n_W} \sum_{i=1}^{n_W} \gamma(y_i(1)) - \gamma(y_i(0))$$

Linear Representations Exist



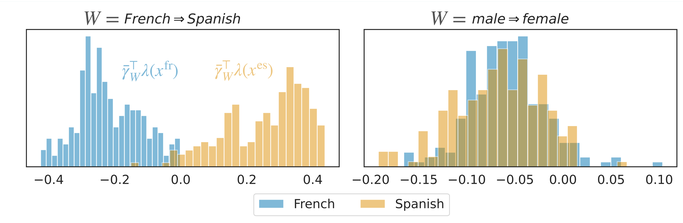
Differences between counterfactual pairs are more parallel to $\bar{\gamma}_W$ than those between random pairs, supporting the linear representation hypothesis.

Causally Separable Concepts are Represented Orthogonally under the Causal Inner Product



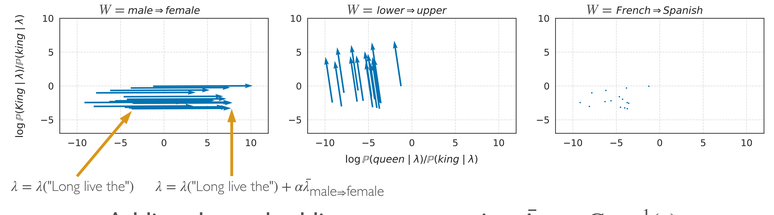
Heatmaps of $\langle \bar{\gamma}_W, \bar{\gamma}_{W'} \rangle$ show that the causal inner product (left) between the unembedding representations of causally separable concepts is close to zero. It clearly improves on the naive Euclidean inner product (right).

Unembedding Representation Yields Linear Probe



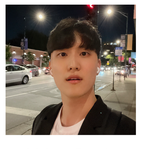
$\bar{\gamma}_{\text{French}\Rightarrow\text{Spanish}}$ (left) separates the embeddings of French and Spanish contexts, while $\bar{\gamma}_{\text{male}\Rightarrow\text{female}}$ (right) does not.

Embedding Representation Yields Steering Vector



Adding the embedding representation $\bar{\lambda}_W := \text{Cov}^{-1}(\gamma) \bar{\gamma}_W$ to context embeddings changes the target concept, without changing other causally separable concepts.

Find us online! Email: parkkiho@uchicago.edu


Kiho Park


YJ Choe


Victor Veitch

arXiv:2311.03658