Table 2. Concept names, one example of the counterfactual pairs, and the number of the used pairs

#	Concept	Example	Count
1	verb ⇒ 3pSg	(accept, accepts)	32
2	verb ⇒ Ving	(add, adding)	31
3	verb ⇒ Ved	(accept, accepted)	47
4	Ving ⇒ 3pSg	(adding, adds)	27
5	Ving ⇒ Ved	(adding, added)	34
6	3pSg ⇒ Ved	(adds, added)	29
7	verb ⇒ V + able	(accept, acceptable)	6
8	verb ⇒ V + er	(begin, beginner)	14
9	verb ⇒ V + tion	(compile, compilation)	8
10	verb ⇒ V + ment	(agree, agreement)	11
11	adj ⇒ un + adj	(able, unable)	5
12	adj ⇒ adj + ly	(according, accordingly)	18
13	small ⇒ big	(brief, long)	20
14	thing ⇒ color	(ant, black)	21
15	thing ⇒ part	(bus, seats)	13
16	country ⇒ capital	(Austria, Vienna)	15
17	pronoun ⇒ possessive	(he, his)	4
18	male ⇒ female	(actor, actress)	11
19	lower ⇒ upper	(always, Always)	34
20	noun ⇒ plural	(album, albums)	63
21	adj ⇒ comparative	(bad, worse)	19
22	adj ⇒ superlative	(bad, worst)	9
23	frequent ⇒ infrequent	(bad, terrible)	32
24	English ⇒ French	(April, avril)	46
25	French ⇒ German	(ami, Freund)	35
26	French ⇒ Spanish	(année, año)	35
27	German ⇒ Spanish	(Arbeit, trabajo)	22

C. Experiment Details

The LLaMA-2 model We utilize the `llama-2-7b` variant of the LLaMA-2 model (Touvron et al., 2023), which is accessible online (with permission) via the `huggingface` library.[7] Its seven billion parameters are pre-trained on two trillion `sentencepiece` (Kudo & Richardson, 2018) tokens, 90% of which is in English. This model uses 32,000 tokens and 4,096 dimensions for its token embeddings.

Counterfactual pairs Tokenization poses a challenge in using certain words. First, a word can be tokenized to more than one token. For example, a word "princess" is tokenized to "prin" + "cess", and γ("princess") does not exist. Thus, we cannot obtain the meaning of the exact word "princess". Second, a word can be used as one of the tokens for another word. For example, the French words "bas" and "est" ("down" and "east" in English) are in the tokens for the words "basalt", "baseline", "basil", "basilica", "basin", "estuary", "estrange", "estoppel", "estival", "esthetics", and "estrogen". Therefore, a word can have another meaning other than the meaning of the exact word.

When we collect the counterfactual pairs to identify $\bar{\gamma}_W$, the first issue in the pair can be handled by not using it. However, the second issue cannot be handled, and it gives a lot of noise to our results. Table 2 presents the number of the counterfactual pairs for each concept and one example of the pairs. The pairs for 13, 17, 19, 23-27th concepts are generated by ChatGPT-4 (OpenAI, 2023), and those for 16th concept are based on the csv file[8]). The other concepts are based on The Bigger Analogy Test Set (BATS) (Gladkova et al., 2016), version 3.0[9], which is used for evaluation of the word analogy task.

Context samples In Section 4, for a concept W (e.g., `English⇒French`), we choose several counterfactual pairs $(Y(0), Y(1))$ (e.g., (house, maison)), then sample context $\{x_j^0\}$ and $\{x_j^1\}$ that the next token is $Y(0)$ and $Y(1)$, respectively, from Wikipedia. These next token pairs are collected from the `word2word` bilingual lexicon (Choe et al., 2020), which is a

[7] `https://huggingface.co/meta-llama/Llama-2-7b-hf`

[8] `https://github.com/jmerullo/lm_vector_arithmetic/blob/main/world_capitals.csv`

[9] `https://vecto.space/projects/BATS/`

Table 3. Concepts used to investigate measurement notion

Concept	Example	Count
English ⇒ French	(house, maison)	(209, 231)
French ⇒ German	(déjà, bereits)	(278, 205)
French ⇒ Spanish	(musique, música)	(218, 214)
German ⇒ Spanish	(Krieg, guerra)	(214, 213)

Table 4. Contexts used to investigate intervention notion

j	x_j
1	Long live the
2	The lion is the
3	In the hierarchy of medieval society, the highest rank was the
4	Arthur was a legendary
5	He was known as the warrior
6	In a monarchy, the ruler is usually a
7	He sat on the throne, the
8	A sovereign ruler in a monarchy is often a
9	His domain was vast, for he was a
10	The lion, in many cultures, is considered the
11	He wore a crown, signifying he was the
12	A male sovereign who reigns over a kingdom is a
13	Every kingdom has its ruler, typically a
14	The prince matured and eventually became the
15	In the deck of cards, alongside the queen is the

publicly available word translation dictionary. We take all word pairs between languages that are the top-1 correspondences to each other in the bilingual lexicon and filter out pairs that are single tokens in the LLaMA-2 model's vocabulary.

Table 3 presents the number of the contexts $\{x_j^0\}$ and $\{x_j^1\}$ for each concept and one example of the pairs $(Y(0), Y(1))$.

In the experiment for intervention notion, for a concept W, Z, we sample texts which $Y(0,0)$ (e.g., "king") should follow, via ChatGPT-4. We discard the contexts such that $Y(0,0)$ is not the top 1 next word. Table 4 present the contexts we use.

D. Additional Results

D.1. Histograms of random and counterfactual pairs for all concepts

In Figure 7, we include an analog of Figure 2 where we check the causal inner product of the differences between the counterfactual pairs and an LOO estimated unembedding representation for each of the 27 concepts. While the most of the concepts are encoded in the unembedding representation, some concepts, such as `thing⇒part`, are not encoded in the unembedding space Γ.

D.2. Comparison with the Euclidean inner products

In Figure 8, we also plot the cosine similarities induced by the Euclidean inner product between the unembedding representations. Surprisingly, the Euclidean inner product somewhat works in the LLaMA-2 model as most of the causally separable concepts are orthogonal! This may due to some initialization or implicit regularizing effect that favors learning unembeddings with approximately isotropic covariance. Nevertheless, the estimated causal inner product clearly improves on the Euclidean inner product. For example, `frequent⇒infrequent` (concept 23) has high Euclidean inner product with many separable concepts, and these are much smaller for the causal inner product. Conversely, `English⇒French` (24) has low Euclidean inner product with the other language concepts (25-27), but high causal inner product with `French⇒German` and `French⇒Spanish` (while being nearly orthogonal to `German⇒Spanish`, which does not share French).

Table 5. Context: "The prince matured and eventually became the "

Rank	$\alpha = 0$	0.1	0.2	0.3	0.4
1	*king*	*king*	em	**queen**	**queen**
2	em	em	r	em	woman
3	leader	r	leader	r	lady
4	r	leader	*king*	leader	wife
5	King	head	**queen**	woman	em

Table 6. Context: "In a monarchy, the ruler is usually a "

Rank	$\alpha = 0$	0.1	0.2	0.3	0.4
1	*king*	*king*	her	woman	woman
2	monarch	monarch	monarch	**queen**	**queen**
3	member	her	member	her	female
4	her	member	woman	monarch	her
5	person	person	**queen**	member	member

Interestingly, the same heatmaps for a more recent Gemma-2B model (Mesnard et al., 2024) in Figure 9 illustrate that the Euclidean inner product doesn't capture semantics, while the causal inner product still works. One possible reason is that the origin of the unembeddings is meaningful as the Gemma model ties the unembeddings to the token embeddings used before the transformer layers.

D.3. Additional results from the measurement experiment

We include analogs of Figure 4, specifically where we use each of the 27 concepts as a linear probe on either `French⇒Spanish` (Figure 10) or `English⇒French` (Figure 11) contexts.

D.4. Additional results from the intervention experiment

In Figure 12, we include an analog of Figure 5 where we add the embedding representation $\alpha\bar{\lambda}_C$ (4.1) for each of the 27 concepts to $\lambda(x_j)$ and see the change in logits.

D.5. Additional tables of top-5 words after intervention

Table 5 and Table 6 are analogs of Table 1 where we use different contexts $x =$ "In a monarchy, the ruler usually is a " and $x =$ "The prince matured and eventually became the ". For the first example, note that "r" and "em" are the prefix tokens for words related to royalty, such as "ruler", "royal", and "emperor". For the second example, even when the target word "**queen**" does not become the most likely one, the most likely words still reflect the concept direction ("woman", "**queen**", "her", "female").

D.6. A sanity check for the estimated causal inner product

In earlier experiments, we found that the choice $M = \mathrm{Cov}(\gamma)^{-1}$ from (3.3) yields a causal inner product and induces an embedding representation $\bar{\lambda}_W$ in the form of (4.1). Here, we run a sanity check experiment where we verify that the induced embedding representation satisfies the uncorrelatedness condition in Assumption D.6. In Figure 13, we empirically show that $\bar{\lambda}_W^\top\gamma$ and $\bar{\lambda}_Z^\top\gamma$ are uncorrelated for the causally separable concepts (left plot), while they are correlated for the non-causally separable concepts (right plot). In these plots, each dot corresponds to the point $(\bar{\lambda}_W^\top\gamma, \bar{\lambda}_Z^\top\gamma)$, where γ is an unembedding vector γ corresponding to each token in the LLaMA-2 vocabulary (32K total).