

A Geometric Notion of Causal Probing

Clément Guerner Anej Svete Tianyu Liu Alexander Warstadt Ryan Cotterell

{cguerner, anej.svete, tianyu.liu, awarstadt, ryan.cotterell}@inf.ethz.ch

ETH zürich

Abstract

The linear subspace hypothesis (Bolukbasi et al., 2016) states that, in a language model’s representation space, all information about a concept such as verbal number is encoded in a linear subspace. Prior work has relied on auxiliary classification tasks to identify and evaluate candidate subspaces that might give support for this hypothesis. We instead give a set of intrinsic criteria which characterize an ideal linear concept subspace and enable us to identify the subspace using only the language model distribution. Our information-theoretic framework accounts for spuriously correlated features in the representation space (Kumar et al., 2022a). As a byproduct of this analysis, we hypothesize a causal process for how a language model might leverage concepts during generation. Empirically, we find that LEACE (Belrose et al., 2023) returns a one-dimensional subspace containing roughly half of total concept information under our framework for verbal-number. Our causal intervention for controlled generation shows that, for at least one concept, the subspace returned by LEACE can be used to manipulate the concept value of the generated word with precision.

1 Introduction

The reliance of language models (LMs) on concepts to make predictions—especially linguistic concepts such as *verbal-number*¹—is a well-studied phenomenon (Ravfogel et al., 2021; Lasri et al., 2022; Amini et al., 2023). Earlier studies on this topic test whether an LM uses the concept of *verbal-number* by giving it forced choice between a grammatical and an ungrammatical variant of a sentence (Linzen et al., 2016; Marvin and Linzen, 2018; Goldberg, 2019; Lasri et al., 2022). Consider, for example, the sentences:

- (1) a. *The kids walk the dog.*
the kid.PL walk.3PL.PRES the dog.SG

¹Throughout the text, we will use a distinguished typesetting to refer to concepts. For instance, the concept of a bird is written as *bird*.

- b. **The kids walks the dog.*
the kid.PL walk.3SG.PRES the dog.SG

Goldberg (2019) shows that LMs can achieve near perfect accuracy when forced to choose between two such variants. Such results strongly suggest that LMs make use of *verbal-number* and other concepts to perform next-word prediction, but tell us little about how the representation spaces of these models encode such concepts.

Our primary contribution is to construct a novel geometric notion of what it means for a neural LM’s representation space² to have information about a concept. Following Bolukbasi et al. (2016) and Ravfogel et al. (2022a), we argue that concepts are naturally operationalized by *linear* subspaces. Linear subspaces lend themselves to tractable algorithms, and they have a simple geometric interpretation which makes it possible to erase a concept from a representation. Existing work (Lasri et al., 2022; Ravfogel et al., 2023) has relied on \mathcal{V} -information (Xu et al., 2020) to quantify the amount of information in the representation space of a language model, before and after concept erasure. This measure is *extrinsic* to the language model, in the sense that it relies on a variational family \mathcal{V} of auxiliary classifiers to measure concept information. In contrast, we propose an *intrinsic*, information-theoretic (Shannon, 1948) definition of information, by which we mean that information is quantified using distributions induced from the language model, i.e., without relying on an additional classifier.

We show, via an example inspired by Kumar et al. (2022a), that a naïve approach to measuring intrinsic information in a subspace falls victim to spurious correlations. Specifically, while a ground truth, *causal* concept subspace may exist in the representation space, correlated non-concept features can also contain information about the concept, complicating the task of estimating concept information in either subspace. Our frame-

²For now, we define a representation space simply as the d -dimensional vector space that a language model relies on to encode text. We propose a more formal definition in §2.

work breaks the dependence between the concept subspace and its orthogonal complement, allowing us to *correctly* compute information contained in either subspace while marginalizing out the other. This approach is counterfactual in the sense that it creates representations that would not otherwise occur under the language model. Crucially, it allows us to talk about the mutual information between linear subspaces and concepts.

We derive four geometric properties within our counterfactual framework that characterize a precise geometric encoding of a concept. First, **erasure** is the condition that the orthogonal complement of the concept subspace should contain *no* information about the concept. Second, **encapsulation** states that projecting a representation onto our concept subspace should preserve *all* the information about the concept. Third, **stability** quantifies the requirement that projection onto the orthogonal complement of our concept subspace should preserve non-concept information. Finally, **containment** ensures that the concept subspace does not contain additional information beyond the concept.

Empirically, we study *verbal-number* in English and *grammatical-gender* in French. We find, for *verbal-number*, that LEACE (Belrose et al., 2023) yields a one-dimensional concept subspace which, according to our novel counterfactual metrics, contains a large share of concept information while leaving non-concept information untouched. We then leverage our intrinsic measure of information to posit a causal graphical model by which a latent concept may govern LM text generation. This model enables us to derive a causal controlled generation method by manipulating the concept component of a representation. And, indeed, we find evidence that it is possible to use a one-dimensional subspace to control the generation behavior of the language model with respect to *verbal-number*, but not for *grammatical-gender*.³

2 Concepts and Information

In this section, we build towards a definition of mutual information between representations and the concept of interest.

2.1 Language Modeling Basics

A language model is a probability distribution p_{LM} over Σ^* , the Kleene closure over an alphabet Σ .

³Code will be made available in camera ready version.

We parameterize p_{LM} in an autoregressive manner (Du et al., 2023) as follows:

$$p_{\text{LM}}(\mathbf{x}) = p_{\text{LM}}(\text{EOS} \mid \mathbf{x}) \prod_{t=1}^T p_{\text{LM}}(x_t \mid \mathbf{x}_{<t}) \quad (1)$$

where $x_t \in \Sigma$ refers to t^{th} word⁴ in a string $\mathbf{x} \in \Sigma^*$, where $\mathbf{x}_{<t}$ represents the first $(t - 1)$ words of \mathbf{x} , and $\text{EOS} \notin \Sigma$ being a distinguished end-of-string symbol.

Many language models make use of contextual representations, i.e., they encode a textual context $\mathbf{x}_{<t}$ as a real-valued column vector $h(\mathbf{x}_{<t}) \in \mathbb{R}^d$. Generally, $h(\mathbf{x}_{<t})$ is deterministically computed from the context string $\mathbf{x}_{<t}$ ⁵, such that the representation space of Eq. (1) is defined as

$$\mathbb{H} \stackrel{\text{def}}{=} \left\{ h(\mathbf{x}) \mid \mathbf{x} \in \Sigma^* \right\} \subset \mathbb{R}^d. \quad (2)$$

2.2 Language Models and Concepts

We now discuss an exact sense in which a language model can be said to encode a concept. First, we define a concept based on the possible values it can take. We formalize this with a **concept set**, a finite, non-empty set \mathcal{C} whose elements are those values. For example, we take the concept set for *verbal-number* to include three values: **sg** (e.g., *walks*), **pl** (e.g., *walk*), and **n/a** (e.g., *consternation*). For various reasons, including syncretism (Baerman, 2007), some verbs in English can have ambiguous concept value depending on context. For example, the *You* in the sentence *You walked to the store* can be **sg** or **pl**. We find similar facts for other concepts in different languages. For instance, for *grammatical-gender* in French, the adjective *marron* can be both **fem** and **msc**.

To relate language models to concept sets, we introduce a deterministic probability distribution $\iota(c \mid \mathbf{x}_{<t}, \mathbf{x})$. ι tells us the probability that, in the sequential context $\mathbf{x}_{<t} \in \Sigma^*$, word $x \in \Sigma$ is annotated with the concept value $c \in \mathcal{C}$. For now, we make the simplifying assumption that ι is

⁴We refer to $x \in \Sigma$ as words for simplicity, even though in the context of neural language modeling, these are often called subwords, tokens, or symbols.

⁵We relax this assumption later on, such that $h(\mathbf{x}_{<t})$ can be stochastic given $\mathbf{x}_{<t}$. One example of a language model with stochastic contextual embeddings is Bowman et al. (2016).

⁶Despite consisting of real vectors, the cardinality of \mathbb{H} is *countably* infinite, because it contains exactly one element for every string in the countably infinite set Σ^* . Thus, summing over \mathbb{H} is discrete and does not require integration.

deterministic, i.e., we have $\iota(c | \mathbf{x}_{<t}, x) \in \{0, 1\}$ for all $c \in \mathcal{C}$, $x \in \Sigma$, and $\mathbf{x}_{<t} \in \Sigma^*$.⁷ We later relax this assumption in §4 by proposing a stochastic operationalization of concepts.

2.3 Unigram Information

To construct a mutual information between the model’s notion of a concept and its contextual representations, we require a joint distribution between a concept-valued random variable and a representation-valued random variable. In order for this estimate to be intrinsic, we obtain this distribution from the language model itself.

We begin in Eq. (3) by defining the **joint induced unigram** distribution of the language model over words and representations. In words, this distribution tells how frequently each word $x \in \Sigma$ co-occurs with a representation $h \in \mathbb{H}$, on average, in a string $\mathbf{x} \sim p_{\text{LM}}$.

$$p_u(x, h) \stackrel{\text{def}}{=} \sum_{\mathbf{x} \in \Sigma^*} p_{\text{LM}}(\mathbf{x}) \frac{\sum_{t=1}^T \mathbb{1}\{x = x_t \wedge h = h(\mathbf{x}_{<t})\}}{T} \quad (3)$$

Next, using ι , we can define a concept–representation induced unigram distribution as

$$p_u(c, h) \stackrel{\text{def}}{=} \sum_{\mathbf{x} \in \Sigma^*} p_{\text{LM}}(\mathbf{x}) \frac{\sum_{t=1}^T \iota(c | \mathbf{x}_{<t}, x_t) \mathbb{1}\{h = h(\mathbf{x}_{<t})\}}{T} \quad (4)$$

The only difference, relative to Eq. (3), is that we count instances of concept values mapped from context strings and words via our distribution ι .

We can now use Eq. (4) to compute our intrinsic measure of concept information in representations:

$$I(\mathcal{C}; \mathbb{H}) = \sum_{c \in \mathcal{C}} \sum_{h \in \mathbb{H}} p_u(c, h) \log \frac{p_u(c, h)}{p_u(c)p_u(h)} \quad (5)$$

where \mathcal{C} is a \mathcal{C} -valued random variable and \mathbb{H} is a \mathbb{H} -valued random variable. Eq. (5) tells us how much information on average a representation $h \in \mathbb{H}$ tells us about the identity of a concept $c \in \mathcal{C}$.

⁷To illustrate this formalism, consider the concept *verbal-number* and sentences (1-a) and (1-b). The concept set for *verbal-number* is $\mathcal{C} = \{\text{sg, pl, n/a}\}$, and ι maps as follows, e.g., $\iota(\text{sg} | \text{The kids, walk}) = 0$, $\iota(\text{pl} | \text{The kids, walk}) = 1$.

Next, we define the following conditional mutual information:

$$I(X; H | \mathcal{C}) = \sum_{c \in \mathcal{C}} \sum_{x \in \Sigma} \sum_{h \in \mathbb{H}} p_u(x, h, c) \log \frac{p_u(x, h | c)}{p_u(x | c)p_u(h | c)} \quad (6)$$

where $p_u(x, h, c)$ is trivially obtained by combining approaches used to derive Eq. (3) and Eq. (4). This quantity measures, given a particular concept value $c \in \mathcal{C}$, how much additional information about a word $x \in \Sigma$ is encoded in the model’s representations.

Our information-theoretic framework can be generalized to handle different language-generating processes, i.e., different decoding algorithms for language models. Let \tilde{p} be a distribution over Σ^* , which we assume we can easily draw samples from, e.g., a language model decoded with nucleus sampling (Holtzman et al., 2020). We obtain the joint induced unigram distribution $\tilde{p}_u(c, h)$ with respect to this distribution by replacing p_{LM} with \tilde{p} in Eq. (4).

3 A Geometric Encoding of Concepts

The **linear subspace hypothesis** (Bolukbasi et al., 2016) makes a prediction about how the concept information we quantify in the previous section is represented geometrically in the LM’s representation space. Specifically, it postulates that there exists a *linear subspace* $S_{\mathcal{C}} \subseteq \mathbb{H}$ that contains all of the information about a concept with values \mathcal{C} . This hypothesis has been tested on various linguistic concepts, including *verbal-number* (Ravfogel et al., 2021; Lasri et al., 2022; Amini et al., 2023) and *grammatical-gender* (Amini et al., 2023). We follow in this vein, and decompose the representation space \mathbb{H} into a concept linear subspace and a non-concept, orthogonal subspace. We provide four definitions, using our information-theoretic framework, that characterize these subspaces in terms of the information that they contain.

3.1 Concept Partition

Given a concept set \mathcal{C} , we define a partition of a language model’s representation space \mathbb{H} into a **concept subspace** $S_{\mathcal{C}}$ and its orthogonal complement, the **non-concept subspace** $S_{\mathcal{C}}^{\perp}$. We refer to $\mathbf{P} \in \mathbb{R}^{d \times d}$ as the orthogonal projection matrix that projects onto $S_{\mathcal{C}}^{\perp}$, i.e., $\mathbf{P}h = \text{proj}_{S_{\mathcal{C}}^{\perp}}(h)$. In turn, $\mathbf{I}_d - \mathbf{P}$ projects onto our concept

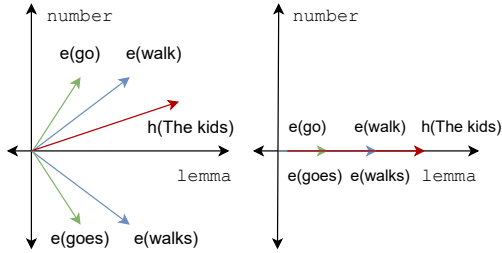


Figure 1: Example of erasure of a **verbal-number** subspace, when predicting the next word given *The kids*. The representation space is two-dimensional with the y -axis representing the correct subspace encoding the concept **verbal-number**, while the x -axis encodes the lemma. Word representations are denoted with \mathbf{v} and contextual representation with \mathbf{h} . On the left, we have the original representation space, and on the right, we have the space resulting from erasing information in our concept subspace, i.e., setting the y -coordinates of all vectors in the space to 0.

	walks	walk	goes	go
sg	0	0	0.7	0
pl	0	0.3	0	0

Table 1: Hypothetical joint unigram distribution $p_u(x, c)$ of **verbal-number** and word. The lemma *walk* is only used as **pl** and *go* only as **sg**.

subspace S_C with dimensionality $|C| - 1$, such that $(\mathbf{I}_d - \mathbf{P})\mathbf{h} = \text{proj}_{S_C}(h)$. The partition of \mathbb{H} into S_C and S_C^\perp is an *information* partition.

We use Eq. (5) to define the information about the concept encoded in both. Consider, for example, information in S_C^\perp about C on average over textual contexts

$$I(C; \mathbf{P}\mathbf{H}) = \sum_{c \in C} \sum_{h \in \mathbb{H}} p_u(c, \mathbf{P}h) \log \frac{p_u(c, \mathbf{P}h)}{p_u(c)p_u(\mathbf{P}h)} \quad (7)$$

where the language model’s representations are orthogonally projected onto S_C^\perp using \mathbf{P} . Eq. (7) relates the *geometric* notion of a linear subspace with the *information-theoretic* notion of information. Thus, if $I(C; \mathbf{P}\mathbf{H})$ is low, we can say that \mathbf{P} erases a lot of concept information in \mathbb{H} by projecting onto the subspace S_C^\perp . We refer to $H_\parallel \stackrel{\text{def}}{=} \{(\mathbf{I}_d - \mathbf{P})\mathbf{h} \mid \mathbf{h} \in \mathbb{H}\}$, $H_\perp \stackrel{\text{def}}{=} \{\mathbf{P}h \mid \mathbf{h} \in \mathbb{H}\}$ as random variables corresponding to contextual representations projected onto concept and non-concept subspaces, respectively.

3.2 The Perils of Correlation

Eq. (7) suggests an attractive property we might ask from \mathbf{P} : It should satisfy $I(C; \mathbf{P}\mathbf{H}) = 0$, i.e., completely erase the information about the concept by projecting onto S_C^\perp . However, as we show next, this naïve characterization is flawed. We illustrate this point with a counterexample inspired by Kumar et al. (2022a), as shown in Fig. 1. Intuitively, such a transformation constitutes successful erasure.⁸ To the extent that such a subspace exists in reality, finding the \mathbf{P} that erases this subspace seems like the correct objective.

Now, consider the hypothetical joint word–concept unigram distribution $p_u(x, c)$ in Table 1. Under this distribution, a projection matrix \mathbf{P} that erases the correct y -axis as shown in Fig. 1 is *not* the minimizer of Eq. (7). Knowledge of the lemma alone reveals the **verbal-number**, because H_\perp (x -axis) and H_\parallel (y -axis) are heavily correlated. This means that $I(C; H_\perp) = 0.88 > 0$ in our toy example in Fig. 1. In order to have $I(C; H_\perp) = 0$, we would need to let $\mathbf{P} = \mathbf{0}$, thereby erasing all lemma information as well. Thus, requiring \mathbf{P} to satisfy $I(C; H_\perp) = 0$ does not characterize successful erasure because it requires removing all spuriously correlated features.

3.3 A Counterfactual Unigram Distribution

The underlying problem with the example given in §3.2 is that H_\parallel and H_\perp have a common cause that introduces a spurious correlation—the Σ^* -valued context random variable $\mathbf{X}_{<t}$. This means $I(H_\perp; H_\parallel) > 0$, i.e., these variables are *not* statistically independent. We resolve this issue by building a variant of our information-theoretic objective in Eq. (7) that *assumes* these two variables are statistically *independent*, i.e., $I(H_\perp; H_\parallel) = 0$. Under this assumption, H_\perp would contain no information about the concept, and identification of H_\parallel would be possible via mutual information. While this assumption likely never holds for a concept in practice, this does not matter here—we are crafting a metric under which the correct subspace will be optimal.

We denote with $h_\parallel \stackrel{\text{def}}{=} (\mathbf{I}_d - \mathbf{P})\mathbf{h}$ and $h_\perp \stackrel{\text{def}}{=} \mathbf{P}h$ the projections onto the concept and non-concept subspace for $\mathbf{h} \in \mathbb{H}$. Marginalizing with respect to

⁸One might refer to the y -axis as the *causal* subspace, in the sense that manipulating the values of that subspace would result in changing precisely the concept encoded by the representation while leaving other aspects intact.

the induced unigram distribution defined in §2, we arrive at the following unigram distributions:

$$p_u(\mathbf{h}_\perp) \stackrel{\text{def}}{=} \sum_{\mathbf{h} \in \mathbb{H}} \mathbb{1}\{\mathbf{h}_\perp = \mathbf{P}\mathbf{h}\} p_u(\mathbf{h}) \quad (8)$$

$$p_u(\mathbf{h}_\parallel) \stackrel{\text{def}}{=} \sum_{\mathbf{h} \in \mathbb{H}} \mathbb{1}\{\mathbf{h}_\parallel = (\mathbf{I}_d - \mathbf{P})\mathbf{h}\} p_u(\mathbf{h}) \quad (9)$$

We now construct a variant of our induced unigram $p_u(\mathbf{x}, \mathbf{c}, \mathbf{h})$ that assumes independence between \mathbf{h}_\perp and \mathbf{h}_\parallel , i.e., $q_u(\mathbf{h}) = q_u(\mathbf{h}_\perp, \mathbf{h}_\parallel) \stackrel{\text{def}}{=} p_u(\mathbf{h}_\perp) p_u(\mathbf{h}_\parallel)$. This **counterfactual unigram distribution** q_u assigns probability mass to $(\mathbf{h}_\perp, \mathbf{h}_\parallel)$ pairs which, under $p_u(\mathbf{h})$, would have zero probability.

$$q_u(\mathbf{x}, \mathbf{c}, \mathbf{h}_\parallel, \mathbf{h}_\perp) \stackrel{\text{def}}{=} \sum_{\mathbf{x}_{<t} \in \Sigma^*} \iota(\mathbf{c} | \mathbf{x}, \mathbf{x}_{<t}) \quad (10)$$

$$p_{\text{LM}}(\mathbf{x} | \mathbf{h}_\parallel, \mathbf{h}_\perp) p(\mathbf{x}_{<t}) p_u(\mathbf{h}_\parallel) p_u(\mathbf{h}_\perp)$$

The choice of the name counterfactual, as well as the implications of this decoupling, will be made precise in §4 when we introduce the causal interpretation of the word–concept model.

We define the **counterfactual mutual information** between the concept and the projection onto the non-concept subspace as

$$I_q(\mathcal{C}; \mathbf{H}_\perp) \stackrel{\text{def}}{=} \sum_{\mathbf{c} \in \mathcal{C}} \sum_{\mathbf{h}_\perp \in \mathbb{H}_\perp} q_u(\mathbf{c}, \mathbf{h}_\perp) \log \frac{q_u(\mathbf{c}, \mathbf{h}_\perp)}{q_u(\mathbf{c}) q_u(\mathbf{h}_\perp)} \quad (11)$$

Importantly, Eq. (11) is minimized by the correct subspace in our example in §3.2. Note that $I_q(\mathcal{C}; \mathbf{H}_\parallel)$ can also be obtained by marginalizing out \mathbf{h}_\perp instead. Finally, we define $I_q(\mathcal{X}; \mathbf{H}_\perp | \mathcal{C})$ by using q_u instead of p_u in Eq. (6).

3.4 Erasure and Encapsulation

We now give formal definitions of erasure and encapsulation based on Eq. (11). These two notions, combined, determine the extent to which a projection matrix \mathbf{P} has decomposed the representation space into concept and non-concept subspaces.

Definition 3.1 (Counterfactual Erasure). *Let $\mathbf{H}_\perp \stackrel{\text{def}}{=} \mathbf{P}\mathbf{H}$ be an \mathbb{R}^d -valued random variable. An orthogonal projection matrix $\mathbf{P} \in \mathbb{R}^{d \times d}$ is an ε -eraser of \mathcal{C} if $I_q(\mathcal{C}; \mathbf{H}_\perp) < \varepsilon$.*

As $\varepsilon \rightarrow 0$, the subspace $S_{\mathcal{C}}^\perp$ characterized by an ε -eraser \mathbf{P} for concept set \mathcal{C} with respect to \mathbb{H} encodes very little information about the concept.

This means that the language model is no longer able to determine the concept value required by the textual context when generating the next word. We now show that given an ε -eraser \mathbf{P} , projecting onto its orthogonal complement with $\mathbf{I}_d - \mathbf{P}$ preserves nearly all of the information.

Definition 3.2 (Counterfactual Encapsulation). *Let $\mathbf{H}_\parallel \stackrel{\text{def}}{=} (\mathbf{I}_d - \mathbf{P})\mathbf{H}$ be an \mathbb{R}^d -valued random variable. An orthogonal projection matrix $\mathbf{I}_d - \mathbf{P} \in \mathbb{R}^{d \times d}$ is an ε -encapsulator of \mathcal{C} if $I_q(\mathcal{C}; \mathbf{H}) - I_q(\mathcal{C}; \mathbf{H}_\parallel) < \varepsilon$.*

The quantity $I_q(\mathcal{C}; \mathbf{H}) - I_q(\mathcal{C}; \mathbf{H}_\parallel)$ is always non-negative due to the data-processing inequality (Cover and Thomas, 2006, §2.8). Encapsulation operationalizes the idea that a subspace gives us all the information needed to correctly identify the concept value required by textual context. Combining erasure and encapsulation, we show that the mutual information decomposes additively in the following sense.

Proposition 3.3. *Suppose \mathbf{P} is a ε -eraser and $(\mathbf{I}_d - \mathbf{P})$ is a ε -encapsulator of \mathcal{C} with respect to \mathbb{H} . Then, as $\varepsilon \rightarrow 0$, the following holds*

$$I_q(\mathcal{C}; \mathbf{H}) = I_q(\mathcal{C}; \mathbf{H}_\perp) + I_q(\mathcal{C}; \mathbf{H}_\parallel) \quad (12)$$

Proof. See App. A. ■

3.5 Containment and Stability

Erasure and encapsulation do not consider the information content of the representation aside from the concept. With perfect erasure and encapsulation, the learned orthogonal projection matrix \mathbf{P} could erase much of the non-concept related information from $S_{\mathcal{C}}^\perp$. Specifically, if \mathcal{C} is encoded non-linearly (Ravfogel et al., 2022b), then erasure via a linear orthogonal projection could require the removal of additional dimensions that also contain non-concept information. Therefore, in the concept erasure literature, tests of successful erasure are paired with a verification that the representations are not otherwise damaged (Kumar et al., 2022a; Ravfogel et al., 2020, 2022a,b; Elazar et al., 2021). We, too, need an information-theoretic notion of preservation of non-concept information in \mathbf{H}_\perp .

Preserving information about non-concept aspects of $\mathbf{x}_{<t}$ in \mathbf{H}_\perp requires that \mathbf{H}_\parallel only capture information about the concept, i.e. that it should be the *minimal* subspace that captures \mathcal{C} . Containment formalizes this notion by requiring that, conditioned on \mathcal{C} , \mathbf{H}_\parallel contains little information about the next word \mathcal{X} .

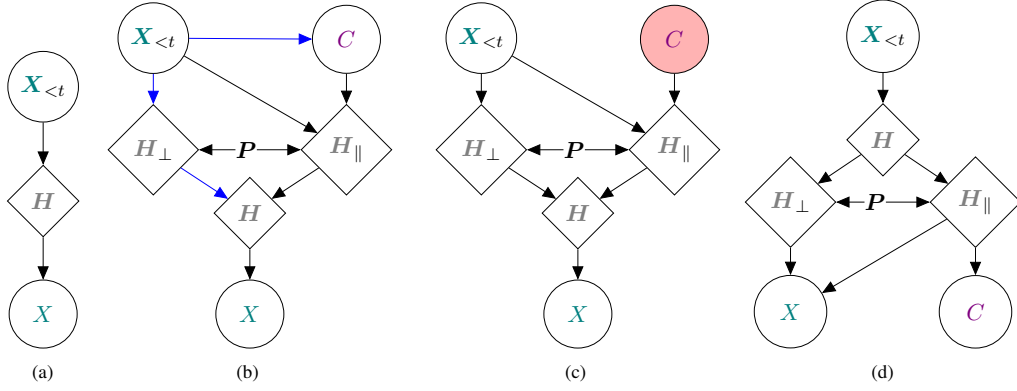


Figure 2: Causal graphical models that demonstrate how a concept may have a causal effect on word generation. Circles represent random variables and diamonds represent deterministic variables. $\mathbf{X}_{<t}, \mathbf{C}, \mathbf{X}$ represent the random variables for the textual context, the underlying concept, and the next word, respectively. $\mathbf{H}, \mathbf{H}_{\parallel}, \mathbf{H}_{\perp}$ are the representation at step t , its concept-related component, and its component whose concept-related information is erased by orthogonal projection matrix \mathbf{P} . Fig. 2a shows the traditional autoregressive causal structure for generation. Fig. 2b is our proposed causal structure for generation with a \mathbf{C} -valued latent variable \mathbf{C} . Fig. 2c is the causal structure induced by a do-intervention on \mathbf{C} . Finally, Fig. 2d is the causal structure implied by Yang and Klein’s (2021) concept-controlled generation approach.

Definition 3.4 (Counterfactual Containment). Let \mathbf{P} be an eraser for concept set \mathbf{C} with respect to \mathbf{H} . Let $\mathbf{H}_{\parallel} \stackrel{\text{def}}{=} (\mathbf{I}_d - \mathbf{P})\mathbf{H}$ be an \mathbb{R}^d -valued random variable. Then, we say that \mathbf{P} is ε -contained with respect to \mathbf{H} and \mathbf{C} if $\mathbb{I}_q(\mathbf{X}; \mathbf{H}_{\parallel} \mid \mathbf{C}) < \varepsilon$.

Lastly, we define stability to measure how much non-concept information about the next word is preserved in the non-concept subspace \mathbf{H}_{\perp} . Ideally, this should be as close as possible to the information present in the entire representation space, ignoring the information about the concept.

Definition 3.5 (Counterfactual Stability). Let \mathbf{P} be an eraser for concept set \mathbf{C} with respect to \mathbf{H} . Let $\mathbf{H}_{\perp} \stackrel{\text{def}}{=} \mathbf{P}\mathbf{H}$ be an \mathbb{R}^d -valued random variable. Then, we say that \mathbf{P} is an ε -stabilizer with respect to \mathbf{H} and \mathbf{C} if $\mathbb{I}_q(\mathbf{X}; \mathbf{H} \mid \mathbf{C}) - \mathbb{I}_q(\mathbf{X}; \mathbf{H}_{\perp} \mid \mathbf{C}) < \varepsilon$.

The data processing inequality once again ensures that $\mathbb{I}_q(\mathbf{X}; \mathbf{H} \mid \mathbf{C}) - \mathbb{I}_q(\mathbf{X}; \mathbf{H}_{\perp} \mid \mathbf{C}) \geq 0$. Containment and stability together characterize the preservation of information not related to concepts.

4 A Causal Graphical Model

We now propose a causal structure by which language models leverage concepts, in the form of a latent variable, in the generation process. We relate this causal structure to the information partition definitions given in §3. This enables causal controlled generation via a do-intervention (Pearl, 2009) on the concept random variable \mathbf{C} . We finish with a discussion of how our causal controlled generation approach improves upon existing approaches.

4.1 Concept as a Latent Variable

We illustrate the traditional autoregressive causal structure, based on the model definition put forth in §2.1, in Fig. 2a. In it, the Σ^* -valued random variable $\mathbf{X}_{<t}$ represents the textual context that was previously sampled from the model, \mathbf{H} is the deterministic contextual representation, and \mathbf{X} the word which is sampled using \mathbf{H} .

To enable controlled generation with respect to the concept, we introduce a \mathbf{C} -valued latent variable \mathbf{C} in the generation process, as shown in Fig. 2b. We make two assumptions about \mathbf{C} . First, we assume that the distribution of \mathbf{C} is influenced by the textual context $\mathbf{X}_{<t}$, and, moreover, that \mathbf{C} is not fully determined by the context $\mathbf{x}_{<t}$, i.e., \mathbf{C} is stochastic. This assumption is justified by the fact that the concept value of the next word may not be fully determined by the preceding context, as discussed in §2. Second, we assume that the concept is determined before the word is sampled. This enables controlled generation, as the concept can directly influence the sampled word x . In doing so, we break away from ι , which deterministically assigned a concept value to a word based on the preceding context.

Our two assumptions on \mathbf{C} have an important implication: $\mathbf{X}_{<t}$ is no longer the only source of stochasticity in \mathbf{H} , as in Fig. 2a. Rather, we assume that both $\mathbf{X}_{<t}$ as well as \mathbf{C} influence the representation \mathbf{H} , i.e., $h = h(\mathbf{x}_{<t}, c)$. Although this construction is not the norm in neural language models, it is a minor departure from reality that

greatly enables our model.

4.2 Causal Controlled Generation

We now derive a formal relationship between erasure, encapsulation, stability, containment, and the assumed causal graph in Fig. 2b. First, inspecting Fig. 2b, we see that if we wish to intervene on C to influence X , there is a single backdoor path from C to H . As shown in Fig. 2c, *intervening* on C directly (denoted by $\text{do}(C = c)$) removes the edge $X_{<t} \rightarrow C$, which lets us easily compute the distribution over the next word after intervention as follows

$$\begin{aligned} p(x \mid H_{\perp} = h_{\perp}, \text{do}(C = c)) & \quad (13) \\ &= \sum_{g \in \mathbb{H}} p(x \mid H = h_{\perp} + (\mathbf{I}_d - \mathbf{P})g) p(g \mid c) \end{aligned}$$

where, as shown in Fig. 2b, we assume that h_{\perp} is deterministic given the context $\mathbf{x}_{<t}$. g is an \mathbb{R}^d -valued contextual representation that encodes a textual context $\mathbf{x}'_{<t}$ with concept value c . With high probability, $h(\mathbf{x}_{<t})$ and $g(\mathbf{x}'_{<t})$ will be different. This is the logical conclusion of our decision to treat h_{\perp} and h_{\parallel} as statistically independent—we can intervene on the generation process by setting the value of the concept component independently.

We now make good on our decision to name the counterfactual unigram distribution from Eq. (10) as such. Assuming the model Fig. 2b, a do-intervention on C —as depicted in Fig. 2c—implies erasure, encapsulation, stability, and containment. We make this idea formal in the following theorem.

Theorem 4.1. *Consider a joint distribution p that factors as in Fig. 2b parameterized by orthogonal projection matrix \mathbf{P} . Under the distribution*

$$\begin{aligned} p_{\text{do}}(x, h_{\perp}, h_{\parallel}, c) &= p(x \mid h_{\perp}, h_{\parallel}) & (14) \\ p(h_{\perp} \mid \text{do}(C = c)) & p(h_{\parallel} \mid \text{do}(C = c)) p(c) \end{aligned}$$

we have that \mathbf{P} is an ε -eraser, $\mathbf{I}_d - \mathbf{P}$ is an ε -encapsulator, $\mathbf{I}_d - \mathbf{P}$ is an ε -container and \mathbf{P} is an ε -stabilizer for every $\varepsilon > 0$.

Proof. See App. B. ■

What Theorem 4.1 tells us is that the graph given in Fig. 2b is consistent with the technical elaboration in §3. Specifically, it means that erasure, encapsulation, stability, and containment are all properties that we expect a causal distribution resulting

from an intervention on a concept to have. The interventional distributions, hence, motivate our discussion on independent $p(h_{\parallel})$ and $p(h_{\perp})$ in §3.3.

4.3 Non-causal Controlled Generation

Controlled generation involving the manipulation of concepts is not a new problem. We contextualize our approach relative to Yang and Klein’s (2021) method. They perform controlled generation as follows. First, they train a classifier to predict a concept value $c \in \mathcal{C}$ from the contextual representation h of a language model. Then, they perform controlled generation by conditioning on a concept value $C = c$ and applying Bayes’ rule as follows:

$$\begin{aligned} p(x \mid \mathbf{x}_{<t}, C = c) & \quad (15) \\ \propto p(C = c \mid (\mathbf{I}_d - \mathbf{P})h(\mathbf{x}_{<t})) p(x \mid \mathbf{x}_{<t}) \end{aligned}$$

We illustrate the causal structure implied by this approach in Fig. 2d. We use \mathbf{P} to relate this approach to our subspace formulation,⁹ but Yang and Klein (2021) do not make use of concept subspaces.

A do-intervention on C has no effect on X with this causal structure, because there is no causal path from C to X in Fig. 2d. This is why the authors *condition* on C instead. In this sense, Yang and Klein’s (2021) and similar methods are not causal and cannot easily be extended to be so. As discussed in §4.2, our approach *is* causal, but such an analysis may come at the price of a number of restricting assumptions that are not fully met in practice. In the next section, we explain how we go about testing these assumptions with data.

5 Experiments and Results

In the remainder of the paper, we test our framework empirically. Specifically, we answer two questions. First, are we able to find a projection matrix \mathbf{P} that meets our definitions in §3? Second, can we use the resulting concept subspace to successfully control the model’s generation behavior, as theorized in §4?

5.1 Experimental Setup

Concepts and Models. We perform our analysis on two concepts, `verbal-number` in English with $\mathcal{C} = \{\text{sg, pl, n/a}\}$ and `grammatical-gender` in French with $\mathcal{C} = \{\text{fem, msc, n/a}\}$. For each of these concepts, we study the representation spaces

⁹Thus, we assume that the classifier is restricted to looking at H_{\parallel} to make its prediction.

	Number (gpt2-large)		Gender (gpt2-base-french)	
	Ancestral	Nucleus	Ancestral	Nucleus
Total Info, $I_q(\mathcal{C}; H)$	0.27±0.01	0.37±0.01	0.42±0.01	0.47±0.02
Erasure, $I_q(\mathcal{C}; H_{\perp})$	0.11±0.01	0.19±0.01	0.27±0.01	0.30±0.02
Subspace Info, $I_q(\mathcal{C}; H_{\parallel})$	0.16±0.01	0.19±0.01	0.10±0.00	0.11±0.01
Encapsulation, $I_q(\mathcal{C}; H) - I_q(\mathcal{C}; H_{\parallel})$	0.11±0.02	0.18±0.02	0.32±0.01	0.36±0.02
Reconstructed Info, $I_q(\mathcal{C}; H_{\perp}) + I_q(\mathcal{C}; H_{\parallel})$	0.27±0.02	0.39±0.01	0.37±0.01	0.41±0.02
Erasure Ratio	0.42±0.04	0.53±0.04	0.64±0.03	0.64±0.04
Subspace Info Ratio	0.58±0.06	0.52±0.04	0.23±0.02	0.23±0.02
Encapsulation Ratio	0.42±0.06	0.48±0.04	0.77±0.02	0.77±0.02
Reconstructed Info Ratio	1.00±0.10	1.05±0.06	0.87±0.04	0.87±0.04
Baseline, $I_q(X; H \mathcal{C})$	1.08±0.12	1.34±0.10	1.81±0.02	2.06±0.04
Containment, $I_q(X; H_{\parallel} \mathcal{C})$	0.12±0.09	0.40±0.11	0.27±0.12	0.36±0.11
Stability, $I_q(X; H \mathcal{C}) - I_q(X; H_{\perp} \mathcal{C})$	0.06±0.10	0.08±0.09	0.02±0.11	-0.02±0.08

Table 2: Counterfactual Information-Theoretic Results. In separate columns, we report two sets of results for each concept, estimated over samples generated using ancestral and nucleus sampling. Each entry shows mean \pm standard deviation over random restarts (see §5.1). The first set of rows shows concept-related information metrics, namely *Erasure* and *Encapsulation*, along with previously undefined metrics. *Total Info* is the amount of concept information in the original representations, *Subspace Info* is concept information in the concept subspace, and *Reconstructed Info* tests whether the partition is lossy relative to total information. The second set of rows shows the ratio of each of these quantities relative to *Total Info*. The third set of rows shows *Containment* and *Stability* metrics, i.e., information about the next word X , conditioned on \mathcal{C} . We include *Baseline* as a reference point.

of an autoregressive language model, namely GPT2 (Radford et al., 2019).¹⁰

Data. For *verbal-number* in English, we use Linzen et al.’s (2016) number agreement dataset. This dataset consists of sentences from Wikipedia that contain a *sg* or *pl* verb with the **fact** (ground truth verb) and the **foil** (inflected form of the fact to have opposite concept value). For *grammatical-gender* in French, we rely on three treebanks from Universal Dependencies (Nivre et al., 2020): French GSD (Guillaume et al., 2019), ParTUT (Sanguinetti and Bosco, 2015, 2014; Bosco and Sanguinetti, 2014), and Rhapsodie (Lacheret et al., 2014). We replicate the pre-processing steps of Linzen et al. (2016) on each of these datasets, i.e., we filter sentences to those containing *fem* or *msc* nouns with an associated adjective, and we obtain the foil by inflecting the *grammatical-gender* of this adjective.

Vocabulary Partition. In §2.2, we defined our context-dependent distribution ι as a means of relating language models and concepts. In practice, we drop the context-dependent aspect and consider a single partition of Σ as our definition of a concept. We start constructing the partition

¹⁰We rely on the implementations in the transformers library (Wolf et al., 2020), namely: gpt2-large for *verbal-number* and gpt2-base-french for *grammatical-gender*.

in a model-agnostic manner: we use SpaCy (Montani et al., 2022) to tag the French and English Wikipedia corpora (Foundation, 2023), respectively. For *verbal-number*, we use the tagged English words to obtain lists of third person present *sg* and *pl* verbs, which we then align to obtain matching pairs, e.g., (*walks*, *walk*). The process is the same for *grammatical-gender* in French, leading to gendered pairs of adjectives, e.g., (*français*, *française*). For each model, we then partition the vocabulary according to the appropriate list, with tokens not included in either list classified as *n/a*. One limitation of our work is that we do not consider concept words that are tokenized into more than one subword in our analysis—for example, if *disambiguate*s tokenizes to [*disambiguate*, "#s"], then the pair (*disambiguate*s, *disambiguate*) is assigned to *n/a*.

The concept value *n/a*. In practice, we exclude *n/a* from our concept set when computing our metrics. The reason for this is that in both text sampled from the model and natural text, the vast majority of words do not invoke the concept, meaning the concept marginals $p_u(c)$ and $\tilde{p}_u(c)$ assign very low probabilities to our values of interest. Resulting information metrics would therefore be dominated by *n/a*, so we choose to exclude it.

Finding the Concept Subspace. We find \mathcal{P} using LEACE (Belrose et al., 2023), the state-of-the-art method for linear concept erasure. LEACE maximizes a cross-entropy loss on samples from \tilde{p}_u with respect to \mathcal{P} , which constitutes a lower bound on our *correlational* $I(\mathcal{C}; \mathbf{H}_\perp)$. Results are reported for three \mathcal{P} estimates obtained from randomized train, test splits for each concept, and three random restarts of the experiment for each \mathcal{P} .

5.2 Partitioning of Concept Information

In this section, we test empirically whether LEACE (Belrose et al., 2023) yields a \mathcal{P} that performs well according to our counterfactual information-theoretic framework defined in §3. Although LEACE is the state of the art for concept erasure, we can anticipate several reasons why it might perform poorly. First, LEACE does not optimize counterfactual erasure, i.e., it does not distinguish between causal vs. spuriously correlated components of \mathbf{H} (see §3.2). Second, as discussed in §3.5, the concept may be non-linearly encoded, such that the removal of a $(|\mathcal{C}| - 1)$ -dimensional subspace would not be sufficient to significantly reduce information. Third, the models under study are not the state of the art, meaning their representation spaces are not necessarily of high quality, and we are limited by available data.

Results in Table 2 for *verbal-number* show that LEACE finds a one-dimensional concept subspace that partitions only about 50% of concept information according to our counterfactual metrics (*Erase Ratio* and *Subspace Info Ratio*). Total amount of information in \mathbf{H}_\perp and \mathbf{H}_\parallel is slightly greater than *Total Info*, but no information is lost in the partitioning (*Reconstructed Info Ratio*). *Stability* values near 0 show that this partitioning preserves non-concept information in S_C^\perp . *Containment* is relatively high compared to the *Baseline*, meaning that the concept subspace found by LEACE is not minimal. For *grammatical-gender* however, LEACE does not find a good projection matrix. Approximately only 30% of concept information is erased, and information is lost in the case of ancestral sampling (*Reconstructed Info Ratio* less than 1).

Looking back on the three failure modes outlined at the start of this section, these results show that LEACE can sometimes return an adequate partitioning according to our framework, but does not drive erasure to 0. Empirically, it is difficult to determine whether this is due to the LEACE objec-

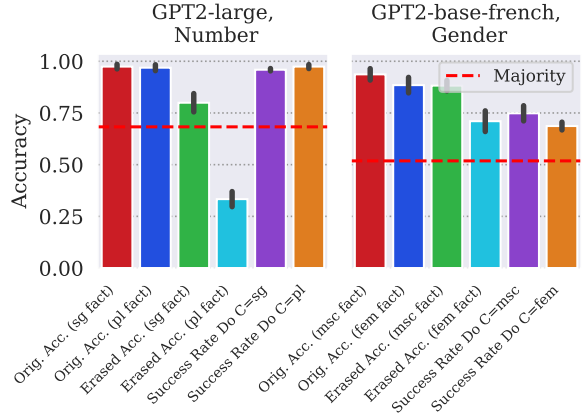


Figure 3: Controlled Generation Experiment. Reported values are computed on (context, fact, foil) samples from the test split of our curated datasets of natural text used to train LEACE. We report results separately depending on whether the fact for the given context is *sg* or *pl*. *Orig. Acc.* refers to the accuracy with which the model chooses fact over foil using original representations. *Erased Acc.* is the accuracy after erasure. *Success Rate Do* measures, for example for Do $\mathbf{C}=\text{sg}$ (see Eq. (13)), the rate at which the intervention induces the model to assign higher probability to the *sg* element of the (fact, foil) pair over its *pl* counterpart, reported on average over *sg* and *pl* contexts in the test set.

tive falling victim to spurious correlations or to the concept encoding being non-linear. If the former is true, one solution might be to learn a \mathcal{P} that optimizes our counterfactual framework. However, this is computationally intractable due to nested sums over the infinite representations space \mathbf{H} , and we leave the development of a tractable approximation to future work. We attribute the failure to learn a concept partition for *grammatical gender* to limitations of the model itself. Compared to English, the best available French gpt2 model is trained on less data and has fewer parameters. In our preliminary experiments, we noticed that smaller English gpt2 models for *verbal-number* were also notably worse than gpt2-large.

5.3 Causal Controlled Generation

In §4, we argued for a causal structure for language generation that allows us to intervene on the concept-valued random variable \mathcal{C} . We now test this causal model empirically by computing the do-intervention in Eq. (13). We define success for the intervention using the forced-choice setup shown in sentences (1-a) and (1-b). For example, given a context with a *sg* fact, we consider $\text{do}(\mathbf{C} = \text{pl})$ successful if $p(x | \mathbf{H}_\perp = h_\perp, \text{do}(\mathbf{C} = \text{pl}))$ assigns

higher probability to the **pl** foil over **sg** fact.

Results for this experiment are shown in Fig. 3. For context, we report the model’s accuracy in the forced-choice setup before (*Orig. Acc.*) and after (*Erased Acc.*) erasure. We note the consistency between information-theoretic metrics in Table 2 and post-erasure accuracy in Fig. 3—the erasure intervention successfully lowers the accuracy of the minority class **pl** for **verbal-number**, however the intervention fails to significantly reduce accuracy for **grammatical-gender**.

With this context in mind, the do-intervention is remarkably successful for **verbal-number**. In particular, $\text{do}(C = \text{pl})$ succeeds for roughly 90% of textual contexts in getting the model to assign higher probability to the **pl** form of the (fact, foil) pair. This result is notable because the low accuracy on predicting the **pl** fact after erasure means that, without the do-intervention, erasure strongly biases the model against generating the **pl** form. By acting solely in our concept subspace via **pl** values of h_{\parallel} , we are able to instead get the model to almost always predict **pl**.

Results for `gpt2-base-french` are much worse— $\text{do}(C = \text{msc})$ actually reduces the accuracy relative to after erasure, while, with $\text{do}(C = \text{fem})$, we see no significant difference. Viewed together with results in §5.2, this confirms that our causal structure only holds given an adequate P under our counterfactual framework. Nonetheless, the success of the do-intervention on **verbal-number** despite concept information not being perfectly isolated in H_{\parallel} suggests that identifying the causal concept direction is not a necessary requirement for causal concept-based controlled generation, so long as H_{\parallel} contains a significant share of concept information.

6 Related Work

In terms of our stated goal of developing a geometrically oriented causal probing framework, our work is most closely related to Elazar et al. (2021) and Lasri et al. (2022). Elazar et al. (2021) pose the problem of identifying a subspace used by a model to perform a task via an erasure intervention, on the assumption that a reduction in word prediction accuracy after intervention certifies the usage of the subspace. Lasri et al. (2022) applied Linzen et al.’s (2016) forced choice approach to the problem of evaluating the impact of concept erasure. Our work relies on a behavioral (forced choice) dataset for

learning the linear projection matrix, but importantly, not for evaluating usage. Our definition of erasure, for example, formalizes the expectation that *all* concept-related word pairs should be indistinguishable under the language model after erasure given any contextual representation.

Previous work is also interested in measuring the degree to which erasure preserves non-concept related features. Ravfogel et al. (2020, 2022a,b) perform various tests, e.g., evaluating whether the model’s understanding of word similarity is affected by erasure using SimLex-999 (Hill et al., 2015), which have little to do with language modeling. Elazar et al. (2021) assess damage to p_{LM} via two different tests. First, they attempt to recover task performance after concept erasure by finetuning the language model on gold annotations for the concept. Fine-tuning results in the paper show an increase in task performance, which suggests that further training can improve the model overall, casting doubt on the validity of performance recovery as an evaluation criterion. Second, the authors also report the overall KL divergence in the LM’s output distribution, over the entire vocabulary. This last approach was a source of inspiration for our work, which delves much deeper into this distributional distance idea via our stability and containment tests.

7 Conclusion

In this paper, we set out to define an *intrinsic* measure of information in a subspace of a language model’s representation space. In light of the correlational failure mode of linear concept erasure methods (Kumar et al., 2022a), doing so requires a counterfactual approach: By assuming statistical independence between the components of a representation in the concept subspace and its orthogonal complement, we are able to correctly measure information in a subspace by marginalizing out the remainder of the space. To the extent that a causal concept subspace exists for a particular concept and model, erasure under this metric is optimized by that subspace. In practice, we did not actually optimize this metric. Our theoretical analysis, combined with the efficacy of linear erasure methods using a correlational objective, suggests a tantalizing prospect: That a counterfactual objective could identify a one-dimensional causal subspace containing *all* information about the concept empirically.

References

- Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. [Naturalistic causal probing for morpho-syntax](#). *Transactions of the Association for Computational Linguistics*, 11:384–403.
- Matthew Baerman. 2007. [Syncretism](#). *Language and Linguistics Compass*, 1(5):539–551.
- Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. [LEACE: Perfect linear concept erasure in closed form](#). *arXiv preprint arXiv:2306.03819*.
- Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. [Man is to computer programmer as woman is to homemaker? Debiasing word embeddings](#). In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Cristina Bosco and Manuela Sanguinetti. 2014. [Towards a Universal Stanford Dependencies parallel treebank](#). In *Proceedings of the 13th Workshop on Treebanks and Linguistic Theories (TLT-13)*, pages 14–25, Tubingen, Germany.
- Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. [Generating sentences from a continuous space](#). In *Proceedings of the 20th SIGLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.
- Thomas M. Cover and Joy M. Thomas. 2006. *Elements of Information Theory*, second edition. Wiley-Interscience.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Li Du, Lucas Torroba Hennigen, Tiago Pimentel, Clara Meister, Jason Eisner, and Ryan Cotterell. 2023. [A measure-theoretic characterization of tight language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9744–9770, Toronto, Canada. Association for Computational Linguistics.
- Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. [Amnesic probing: Behavioral explanation with amnesic counterfactuals](#). *Transactions of the Association for Computational Linguistics*, 9:160–175.
- Wikimedia Foundation. 2023. [Wikimedia downloads](#).
- Yoav Goldberg. 2019. [Assessing BERT’s syntactic abilities](#). *ArXiv*.
- Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. [Conversion et améliorations de corpus du français annotés en Universal Dependencies](#). *Revue TAL*, 60(2):71–95.
- Felix Hill, Roi Reichart, and Anna Korhonen. 2015. [SimLex-999: Evaluating Semantic Models With \(Genuine\) Similarity Estimation](#). *Computational Linguistics*, 41(4):665–695.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *International Conference on Learning Representations*.
- Abhinav Kumar, Chenhao Tan, and Amit Sharma. 2022a. [Probing classifiers are unreliable for concept removal and detection](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17994–18008. Curran Associates, Inc.
- Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. 2022b. [Gradient-based constrained sampling from language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2251–2277, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. 2014. [Rhapsodie: a prosodic-syntactic treebank for spoken French](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*

- (LREC'14), pages 295–301, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. [Probing for the usage of grammatical number](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland. Association for Computational Linguistics.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of LSTMs to learn syntax-sensitive dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Rebecca Marvin and Tal Linzen. 2018. [Targeted syntactic evaluation of language models](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.
- Ines Montani, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O’Leary McCann, Maxim Samsonov, Jim Geovedi, Jim O’Regan, Duygu Altinok, György Orosz, Søren Lind Kristiansen, , Roman, Explosion Bot, Lj Miranda, Leander Fiedler, Daniël De Kok, Grégory Howard, , Edward, Wannaphong Phatthiyaphaibun, Yohei Tamura, Sam Bozek, , Murat, Mark Amery, Ryn Daniels, Björn Böing, Pradeep Kumar Tippa, and Peter Baumgartner. 2022. [spaCy: Industrial-strength natural language processing in Python](#).
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal Dependencies v2: An evergrowing multilingual treebank collection](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Judea Pearl. 2009. [Causal inference in statistics: An overview](#). *Statistics Surveys*, 3:96–146.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. [Information-theoretic probing for linguistic structure](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. [Null it out: Guarding protected attributes by iterative nullspace projection](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.
- Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. 2023. [Log-linear guardedness and its implications](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9413–9431, Toronto, Canada. Association for Computational Linguistics.
- Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. [Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.
- Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022a. [Linear adversarial concept erasure](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.
- Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022b. [Kernelized concept erasure](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Manuela Sanguinetti and Cristina Bosco. 2014. [Converting the parallel treebank ParTUT in Universal Stanford Dependencies](#). In *Proceedings of the 1st Conference for Italian Computational Linguistics (CLiC-it 2014)*, Pisa, Italy.

- Manuela Sanguinetti and Cristina Bosco. 2015. *PartTUT: The Turin University Parallel Tree-bank*, pages 51–69. Springer International Publishing, Cham.
- Claude E. Shannon. 1948. *A mathematical theory of communication*. *The Bell System Technical Journal*, 27(3):379–423.
- Elena Voita and Ivan Titov. 2020. *Information-theoretic probing with minimum description length*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. *Transformers: State-of-the-art natural language processing*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. *A theory of usable information under computational constraints*. In *International Conference on Learning Representations*.
- Kevin Yang and Dan Klein. 2021. *FUDGE: Controlled text generation with future discriminators*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

A Proof of Proposition 3.3

Proposition 3.3. *Suppose \mathbf{P} is a ε -eraser and $(\mathbf{I}_d - \mathbf{P})$ is a ε -encapsulator of \mathcal{C} with respect to \mathbb{H} . Then, as $\varepsilon \rightarrow 0$, the following holds*

$$I_q(\mathcal{C}; \mathbf{H}) = I_q(\mathcal{C}; \mathbf{H}_\perp) + I_q(\mathcal{C}; \mathbf{H}_\parallel) \quad (12)$$

Proof. On the left-hand side,

$$I_q(\mathcal{C}; \mathbf{H}) + \varepsilon \geq I_q(\mathcal{C}; \mathbf{H}_\parallel) + \varepsilon \quad (\text{data-processing inequality}) \quad (16a)$$

$$\geq I_q(\mathcal{C}; \mathbf{H}_\parallel) + I_q(\mathcal{C}; \mathbf{H}_\perp) \quad (\mathbf{P} \text{ is an } \varepsilon\text{-eraser}) \quad (16b)$$

On the right-hand side,

$$I_q(\mathcal{C}; \mathbf{H}) + \varepsilon \leq I_q(\mathcal{C}; \mathbf{H}_\parallel) + 2\varepsilon \quad ((\mathbf{I}_d - \mathbf{P}) \text{ is an } \varepsilon\text{-encapsulator}) \quad (17a)$$

$$\leq I_q(\mathcal{C}; \mathbf{H}_\parallel) + I_q(\mathcal{C}; \mathbf{H}_\perp) + 2\varepsilon \quad (\text{non-negativity of MI}) \quad (17b)$$

Combining Eq. (16b) and Eq. (17b), we have

$$I_q(\mathcal{C}; \mathbf{H}_\parallel) + I_q(\mathcal{C}; \mathbf{H}_\perp) \leq I_q(\mathcal{C}; \mathbf{H}) + \varepsilon \quad (18a)$$

$$\leq I_q(\mathcal{C}; \mathbf{H}_\parallel) + I_q(\mathcal{C}; \mathbf{H}_\perp) + 2\varepsilon \quad (18b)$$

Taking $\varepsilon \rightarrow 0$ in Eq. (18), we have Eq. (12)

$$I_q(\mathcal{C}; \mathbf{H}) = I_q(\mathcal{C}; \mathbf{H}_\parallel) + I_q(\mathcal{C}; \mathbf{H}_\perp)$$

■

B Proof of Theorem 4.1

Theorem 4.1. *Consider a joint distribution p that factors as in Fig. 2b parameterized by orthogonal projection matrix \mathbf{P} . Under the distribution*

$$p_{\text{do}}(\mathbf{x}, \mathbf{h}_\perp, \mathbf{h}_\parallel, \mathbf{c}) = p(\mathbf{x} \mid \mathbf{h}_\perp, \mathbf{h}_\parallel) \quad (14)$$

$$p(\mathbf{h}_\perp \mid \text{do}(\mathcal{C} = \mathbf{c})) p(\mathbf{h}_\parallel \mid \text{do}(\mathcal{C} = \mathbf{c})) p(\mathbf{c})$$

we have that \mathbf{P} is an ε -eraser, $\mathbf{I}_d - \mathbf{P}$ is an ε -encapsulator, $\mathbf{I}_d - \mathbf{P}$ is an ε -container and \mathbf{P} is an ε -stabilizer for every $\varepsilon > 0$.

Proof. Given the factorization in Fig. 2b, we derive the following equation using the independence assumptions given in Fig. 2b:

$$p_{\text{do}}(\mathbf{x}, \mathbf{h}_\perp, \mathbf{h}_\parallel, \mathbf{c}) = p(\mathbf{x} \mid \mathbf{h}_\perp, \mathbf{h}_\parallel) p(\mathbf{h}_\perp \mid \text{do}(\mathcal{C} = \mathbf{c})) p_{\text{do}}(\mathbf{h}_\parallel \mid \text{do}(\mathcal{C} = \mathbf{c})) p(\mathbf{c}) \quad (20a)$$

$$= p(\mathbf{x} \mid \mathbf{h}_\perp, \mathbf{h}_\parallel) p(\mathbf{h}_\perp) p_{\text{do}}(\mathbf{h}_\parallel \mid \mathbf{c}) p(\mathbf{c}) \quad (20b)$$

Erasure. Given Eq. (20b), we have the following joint distribution

$$p_{\text{do}}(\mathbf{c}, \mathbf{h}_\perp) = \sum_{\mathbf{h}_\parallel \in \mathbb{H}_\parallel} \sum_{\mathbf{x} \in \Sigma} p_{\text{do}}(\mathbf{x}, \mathbf{h}_\perp, \mathbf{h}_\parallel, \mathbf{c}) \quad (21a)$$

$$= \sum_{\mathbf{h}_\parallel \in \mathbb{H}_\parallel} \sum_{\mathbf{x} \in \Sigma} p(\mathbf{x} \mid \mathbf{h}_\perp, \mathbf{h}_\parallel) p(\mathbf{h}_\perp) p_{\text{do}}(\mathbf{h}_\parallel \mid \mathbf{c}) p(\mathbf{c}) \quad (21b)$$

$$= \sum_{\mathbf{h}_\parallel \in \mathbb{H}_\parallel} \underbrace{\left(\sum_{\mathbf{x} \in \Sigma} p(\mathbf{x} \mid \mathbf{h}_\perp, \mathbf{h}_\parallel) \right)}_{=1} p(\mathbf{h}_\perp) p_{\text{do}}(\mathbf{h}_\parallel \mid \mathbf{c}) p(\mathbf{c}) \quad (21c)$$

$$= \underbrace{\left(\sum_{\mathbf{h}_\parallel \in \mathbb{H}_\parallel} p_{\text{do}}(\mathbf{h}_\parallel \mid \mathbf{c}) \right)}_{=1} p(\mathbf{h}_\perp) p(\mathbf{c}) \quad (21d)$$

$$= p(\mathbf{h}_\perp) p(\mathbf{c}) \quad (21e)$$

The mutual information $I(\mathbf{C}; \mathbf{H}_\perp)$ can be computed as follows

$$I(\mathbf{C}; \mathbf{H}_\perp) = \sum_{\mathbf{c} \in \mathcal{C}} \sum_{\mathbf{h}_\perp \in \mathbf{H}_\perp} p_{\text{do}}(\mathbf{c}, \mathbf{h}_\perp) \log \frac{p_{\text{do}}(\mathbf{c}, \mathbf{h}_\perp)}{p(\mathbf{c})p(\mathbf{h}_\perp)} \quad (22a)$$

$$= \sum_{\mathbf{c} \in \mathcal{C}} \sum_{\mathbf{h}_\perp \in \mathbf{H}_\perp} p_{\text{do}}(\mathbf{c}, \mathbf{h}_\perp) \log \frac{p(\mathbf{h}_\perp)p(\mathbf{c})}{p(\mathbf{c})p(\mathbf{h}_\perp)} \quad (\text{applying Eq. (21e)}) \quad (22b)$$

$$= 0 < \varepsilon \quad (22c)$$

for every $\varepsilon > 0$.

Encapsulation. The following equation holds given Eq. (20b)

$$I(\mathbf{C}; \mathbf{H}) - I(\mathbf{C}; \mathbf{H}_\parallel) = I(\mathbf{C}; \mathbf{H}_\parallel, \mathbf{H}_\perp) - I(\mathbf{C}; \mathbf{H}_\parallel) \quad (\mathbf{H} = \mathbf{H}_\perp, \mathbf{H}_\parallel) \quad (23a)$$

$$= I(\mathbf{C}; \mathbf{H}_\perp | \mathbf{H}_\parallel) \quad (23b)$$

$$= I(\mathbf{C}; \mathbf{H}_\perp) \quad (\mathbf{H}_\perp, \mathbf{H}_\parallel \text{ are independent (\S 3.3)}) \quad (23c)$$

$$= 0 < \varepsilon \quad (\text{applying Eq. (22c)}) \quad (23d)$$

$$(23e)$$

Containment. The following joint distribution can be derived from Eq. (20b)

$$p_{\text{do}}(\mathbf{x}, \mathbf{h}_\parallel, \mathbf{c} = \mathbf{c}) = \sum_{\mathbf{h}_\perp \in \mathbf{H}_\perp} p_{\text{do}}(\mathbf{x}, \mathbf{h}_\perp, \mathbf{h}_\parallel, \mathbf{c} = \mathbf{c}) \quad (24a)$$

$$= \sum_{\mathbf{h}_\perp \in \mathbf{H}_\perp} p(\mathbf{x} | \mathbf{h}_\perp, \mathbf{h}_\parallel) p(\mathbf{h}_\perp) p_{\text{do}}(\mathbf{h}_\parallel | \mathbf{c} = \mathbf{c}) p(\mathbf{c} = \mathbf{c}) \quad (24b)$$

$$= \sum_{\mathbf{h}_\perp \in \mathbf{H}_\perp} \frac{p(\mathbf{x}, \mathbf{h}_\perp, \mathbf{h}_\parallel)}{p(\mathbf{h}_\perp, \mathbf{h}_\parallel)} p(\mathbf{h}_\perp) p_{\text{do}}(\mathbf{h}_\parallel | \mathbf{c} = \mathbf{c}) p(\mathbf{c} = \mathbf{c}) \quad (24c)$$

$$= \sum_{\mathbf{h}_\perp \in \mathbf{H}_\perp} \frac{p(\mathbf{x}, \mathbf{h}_\perp, \mathbf{h}_\parallel)}{p(\mathbf{h}_\perp) p(\mathbf{h}_\parallel)} p(\mathbf{h}_\perp) p_{\text{do}}(\mathbf{h}_\parallel | \mathbf{c} = \mathbf{c}) p(\mathbf{c} = \mathbf{c}) \quad (\mathbf{H}_\perp, \mathbf{H}_\parallel \text{ are independent (\S 3.3)}) \quad (24d)$$

$$= \underbrace{\sum_{\mathbf{h}_\perp \in \mathbf{H}_\perp} p(\mathbf{x}, \mathbf{h}_\perp | \mathbf{h}_\parallel)}_{=p(\mathbf{x}|\mathbf{h}_\parallel)} p_{\text{do}}(\mathbf{h}_\parallel | \mathbf{c} = \mathbf{c}) p(\mathbf{c} = \mathbf{c}) \quad (24e)$$

$$= p(\mathbf{x} | \mathbf{h}_\parallel) p_{\text{do}}(\mathbf{h}_\parallel | \mathbf{c} = \mathbf{c}) p(\mathbf{c} = \mathbf{c}) \quad (24f)$$

$$= p(\mathbf{x} | \mathbf{h}_\parallel, \mathbf{c} = \mathbf{c}) p(\mathbf{h}_\parallel | \mathbf{c} = \mathbf{c}) \quad (\mathbf{H}_\parallel \text{ is deterministic given } \mathbf{C}) \quad (24g)$$

The mutual information $I(\mathbf{X}; \mathbf{H}_\parallel | \mathbf{C} = \mathbf{c})$ can be computed as follows

$$I(\mathbf{X}; \mathbf{H}_\parallel | \mathbf{C} = \mathbf{c}) \quad (25a)$$

$$= \sum_{\mathbf{x} \in \Sigma} \sum_{\mathbf{h}_\parallel \in \mathbf{H}_\parallel} p_{\text{do}}(\mathbf{x}, \mathbf{h}_\parallel, \mathbf{c} = \mathbf{c}) \log \frac{p_{\text{do}}(\mathbf{x}, \mathbf{h}_\parallel, \mathbf{c} = \mathbf{c})}{p(\mathbf{x} | \mathbf{c} = \mathbf{c}) p(\mathbf{h}_\parallel | \mathbf{c} = \mathbf{c})} \quad (25b)$$

$$= \sum_{\mathbf{x} \in \Sigma} \sum_{\mathbf{h}_\parallel \in \mathbf{H}_\parallel} p_{\text{do}}(\mathbf{x}, \mathbf{h}_\parallel, \mathbf{c} = \mathbf{c}) \log \frac{p(\mathbf{x} | \mathbf{h}_\parallel, \mathbf{c} = \mathbf{c}) p(\mathbf{h}_\parallel | \mathbf{c} = \mathbf{c})}{p(\mathbf{x} | \mathbf{c} = \mathbf{c}) p(\mathbf{h}_\parallel | \mathbf{c} = \mathbf{c})} \quad (\text{applying Eq. (24g)}) \quad (25c)$$

$$= \sum_{\mathbf{x} \in \Sigma} \sum_{\mathbf{h}_\parallel \in \mathbf{H}_\parallel} p_{\text{do}}(\mathbf{x}, \mathbf{h}_\parallel, \mathbf{c} = \mathbf{c}) \log \frac{p(\mathbf{x} | \mathbf{h}_\parallel, \mathbf{c} = \mathbf{c}) p(\mathbf{h}_\parallel | \mathbf{c} = \mathbf{c})}{p(\mathbf{x} | \mathbf{h}_\parallel, \mathbf{c} = \mathbf{c}) p(\mathbf{h}_\parallel | \mathbf{c} = \mathbf{c})} \quad (\mathbf{H}_\parallel \text{ is deterministic given } \mathbf{C}) \quad (25d)$$

$$= 0 < \varepsilon \quad (25e)$$

Stability. The following equation holds given Eq. (20b)

$$\begin{aligned} & I(X; H \mid C = c) - I(X; H_{\perp} \mid C = c) && (26a) \\ = & I(X; H_{\perp}, H_{\parallel} \mid C = c) - I(X; H_{\perp} \mid C = c) && (H = (H_{\perp}, H_{\parallel})) \quad (26b) \\ = & I(X; H_{\parallel} \mid H_{\perp}, C = c) && \text{(conditional mutual information)} \quad (26c) \\ = & I(X; H_{\parallel} \mid C = c) && (H_{\perp}, H_{\parallel} \text{ are independent (\S 3.3)}) \quad (26d) \\ = & 0 < \varepsilon && \text{(applying Eq. (25e))} \quad (26e) \end{aligned}$$

■