A Geometric Notion of Causal Probing

Clément Guerner **Tianyu Liu** **Anej Svete** **Alexander Warstadt** **Ryan Cotterell**
{cguerner, tianyu.liu, anej.svete, 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., 2022) by reconciling the statistical notion of concept information and the geometric notion of how concepts are encoded in the representation space. 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 linear concept erasure is successful in erasing most concept information under our framework for verbal number as well as some complex aspect-level sentiment concepts from a restaurant review dataset. Our causal intervention for controlled generation shows that, for at least one concept across two languages models, the concept subspace can be used to manipulate the concept value of the generated word with precision.

https://github.com/rycolab/causalgeom

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; Arora et al., 2024). Earlier studies on this topic test whether an LM uses the concept of verbal-number by giving it a 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

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. This result suggests that LMs make use of verbal-number and other concepts to perform next-word prediction, but tells 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 space[2] 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

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

[2] For 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.

the language model, i.e., without relying on an additional classifier.

We show, via an example inspired by Kumar et al. (2022), 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 framework 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 linguistic concepts `verbal-number` in English and `grammatical-gender` in French. We find, for `verbal-number`, that the LEACE method for linear concept erasure (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 relatively 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 two language models with respect to `verbal-number`, but not for `grammatical-gender`. We test our causal intervention against the CEBaB (Abraham et al., 2022) benchmark, and find our do-intervention improves the performance of INLP (Ravfogel et al., 2020) and LEACE as causal effect estimators.

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

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

where $x_t \in \Sigma$ refers to t-th word[3] in a string $\boldsymbol{x} \in \Sigma^*$, $\boldsymbol{x}_{<t}$ represents the first $t-1$ words of $\boldsymbol{x}$, 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 $\boldsymbol{x}_{<t}$ as a real-valued column vector $\boldsymbol{h}(\boldsymbol{x}_{<t}) \in \mathbb{R}^d$. Generally, $\boldsymbol{h}(\boldsymbol{x}_{<t})$ is deterministically computed from the context string $\boldsymbol{x}_{<t}$[4], such that the representation space of Eq. (1) is defined as

$$\mathrm{H} \stackrel{\text{def}}{=} \left\{ \boldsymbol{h}(\boldsymbol{x}) \mid \boldsymbol{x} \in \Sigma^* \right\} \subseteq \mathbb{R}^d.^{5} \quad (2)$$

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.,

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

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

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

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 instance, in the sentence *You walked to the store*, *walked* can be sg or pl. We find similar facts for other concepts in different languages, e.g., for `grammatical-gender` in French, the adjective *marron* can be both fem and msc.

To relate language models to concept sets, we introduce a probability distribution $\iota(c \mid \boldsymbol{x}_{<t}, x)$. ι tells us the probability that, in the sequential context $\boldsymbol{x}_{<t} \in \Sigma^*$, word $x \in \Sigma$ is annotated with the concept value $c \in \mathcal{C}$. For now, we make a simplifying assumption that ι is deterministic, i.e., $\iota(c \mid \boldsymbol{x}_{<t}, x) \in \{0, 1\}$ for all $c \in \mathcal{C}$, $x \in \Sigma$, and $\boldsymbol{x}_{<t} \in \Sigma^*$.[6] This assumption will be relaxed in §4 with 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 by defining the **joint induced unigram** distribution of the language model over words, concepts, and representations in Eq. (3). In words, this distribution tells how frequently each word $x \in \Sigma$ co-occurs with a concept value $c \in \mathcal{C}$ and a representation $h \in \mathrm{H}$, on average, in a string $\boldsymbol{x} \sim p_{\mathrm{LM}}$ of length T:

$$p_u(x, c, h) \stackrel{\text{def}}{=} \sum_{\boldsymbol{x} \in \Sigma^*} p_{\mathrm{LM}}(\boldsymbol{x}) \frac{\sum_{t=1}^{T} \iota(c \mid \boldsymbol{x}_{<t}, x_t) \mathbb{1}\left\{x = x_t \wedge h = h(\boldsymbol{x}_{<t})\right\}}{T} \tag{3}$$

We can now use Eq. (3) to compute our intrinsic measure of concept information in representations:

$$\mathrm{I}(C; \boldsymbol{H}) = \sum_{c \in \mathcal{C}} \sum_{h \in \mathrm{H}} p_u(c, h) \log \frac{p_u(c, h)}{p_u(c) p_u(h)} \tag{4}$$

where C is a $\mathcal{C}$-valued random variable, $\boldsymbol{H}$ is a H-valued random variable, and $p_u(c, h)$ is obtained by marginalizing out x from Eq. (3). Eq. (4) tells us how much information on average a representation $h \in \mathrm{H}$ encodes about the identity of a concept $c \in \mathcal{C}$.

Next, we define the following conditional mutual information:

$$\mathrm{I}(X; \boldsymbol{H} \mid C) = \sum_{c \in \mathcal{C}} \sum_{x \in \Sigma} \sum_{h \in \mathrm{H}} p_u(x, h, c) \log \frac{p_u(x, h \mid c)}{p_u(x \mid c) p_u(h \mid c)} \tag{5}$$

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, e.g., different decoding algorithms for language models, or natural text generated by a process other than the language model under study.

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* $\mathrm{S}_{\mathcal{C}} \subseteq \mathbb{R}^d$ that contains all of the information about a concept with values $\mathcal{C}$.[7] 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 H into a concept linear subspace and an orthogonal, non-concept linear subspace. Then, we provide four information-theoretic metrics 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 H into a **concept subspace** $\mathrm{S}_{\mathcal{C}}$ and its orthogonal complement, the **non-concept subspace** $\mathrm{S}_{\mathcal{C}}^{\perp}$. We refer to $\boldsymbol{P} \in \mathbb{R}^{d \times d}$ as the orthogonal projection matrix that projects onto $\mathrm{S}_{\mathcal{C}}^{\perp}$, i.e., $\boldsymbol{P}h = \mathrm{proj}_{\mathrm{S}_{\mathcal{C}}^{\perp}}(h)$. In turn, $\boldsymbol{I}_d - \boldsymbol{P}$ projects onto the concept subspace $\mathrm{S}_{\mathcal{C}}$ with dimensionality $|\mathcal{C}| - 1$, such that

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

[7] Note that $\mathrm{S}_{\mathcal{C}}$ is *not* a linear subspace in the linear-algebraic sense because it is countable and thus not closed under scalar multiplication.

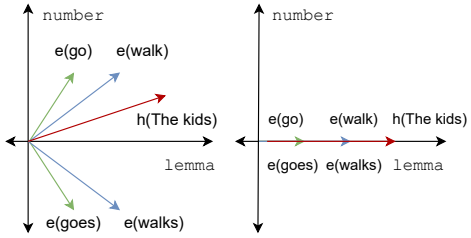

Figure 1: Example of erasure of a verbal-number subspace, when predicting the next word given *The kids*. The representation space is two-dimensional with the y-axis representing the correct subspace encoding the concept verbal-number, while the x-axis encodes the lemma. Word representations are denoted with $\mathbf{e}$ and contextual representation with 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.

$(\boldsymbol{I}_d - \boldsymbol{P})h = \text{proj}_{\text{S}_\mathcal{C}}(h)$. We refer to the partition of $\mathbb{R}^d$ into $\text{S}_\mathcal{C}$ and $\text{S}_\mathcal{C}^\perp$ as an **information partition**.

We use Eq. (4) to define the information about the concept encoded in both. Consider, for example, information in $\text{S}_\mathcal{C}^\perp$ about $\mathcal{C}$ on average over textual contexts:

$$\text{I}(C;\boldsymbol{P}\boldsymbol{H}) = \sum_{c\in\mathcal{C}}\sum_{h\in\text{H}} p_u(c, \boldsymbol{P}h)\log\frac{p_u(c, \boldsymbol{P}h)}{p_u(c)p_u(\boldsymbol{P}h)} \tag{6}$$

where the language model's representations are orthogonally projected onto $\text{S}_\mathcal{C}^\perp$ using $\boldsymbol{P}$. Eq. (6) relates the *geometric* notion of a linear subspace with the *information-theoretic* notion of information. Thus, if $\text{I}(C;\boldsymbol{P}\boldsymbol{H})$ is low, we can say that $\boldsymbol{P}$ erases a lot of concept information in H by projecting onto the subspace $\text{S}_\mathcal{C}^\perp$. We denote $\text{H}_\parallel \stackrel{\text{def}}{=} \{(\boldsymbol{I}_d - \boldsymbol{P})\,h \mid h \in \text{H}\}$, $\text{H}_\perp \stackrel{\text{def}}{=} \{\boldsymbol{P}h \mid h \in \text{H}\}$, and refer to $\boldsymbol{H}_\parallel, \boldsymbol{H}_\perp$ as random variables corresponding to contextual representations projected onto concept and non-concept subspaces, respectively.

3.2 The Perils of Correlation

Eq. (6) suggests an attractive property we might ask from $\boldsymbol{P}$: It should satisfy $\text{I}(C;\boldsymbol{P}\boldsymbol{H}) = 0$, i.e., completely erase the information about the concept by projecting onto $\text{S}_\mathcal{C}^\perp$. However, as we show next, this naïve characterization is flawed. We illustrate this point with a counterexample inspired by Kumar et al. (2022), shown in Fig. 1. Intuitively, such a transformation constitutes successful erasure.[8] To the extent that such a subspace exists in reality, finding the $\boldsymbol{P}$ 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 $\boldsymbol{P}$ that erases the correct y-axis as shown in Fig. 1 is *not* the minimizer of Eq. (6). Knowledge of the lemma alone reveals the verbal-number, because $\boldsymbol{H}_\perp$ (x-axis) and $\boldsymbol{H}_\parallel$ (y-axis) are heavily correlated. This means that $\text{I}(C;\boldsymbol{H}_\perp) = 0.88 > 0$ in our toy example in Fig. 1. In order to have $\text{I}(C;\boldsymbol{H}_\perp) = 0$, we would need to let $\boldsymbol{P} = \boldsymbol{0}$, thereby erasing all lemma information as well. Thus, requiring $\boldsymbol{P}$ to satisfy $\text{I}(C;\boldsymbol{H}_\perp) = 0$ 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 $\boldsymbol{H}_\parallel$ and $\boldsymbol{H}_\perp$ have a common cause that introduces a spurious correlation—the Σ^*-valued context random variable $\boldsymbol{X}_{<t}$. This means $\text{I}(\boldsymbol{H}_\perp;\boldsymbol{H}_\parallel) > 0$, i.e., $\boldsymbol{H}_\perp$ and $\boldsymbol{H}_\parallel$ are *not* statistically independent. We resolve this issue by building a variant of our information-theoretic objective in Eq. (6) that *assumes* these two variables are statistically *independent*, i.e., $\text{I}(\boldsymbol{H}_\perp;\boldsymbol{H}_\parallel) = 0$. Under this assumption, $\boldsymbol{H}_\perp$ would contain no information about the concept, and identification of $\boldsymbol{H}_\parallel$ 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 $h_\parallel \stackrel{\text{def}}{=} (\boldsymbol{I}_d - \boldsymbol{P})h$ and $h_\perp \stackrel{\text{def}}{=} \boldsymbol{P}h$ the projections onto the concept and non-concept subspace for $h \in \text{H}$. Marginalizing with respect to the induced unigram distribution defined in §2, we

[8] One might, but probably shouldn't, refer to the y-axis as the *causal* subspace, in the sense that manipulating the value in that subspace would result in changing precisely the concept encoded by the representation while leaving other aspects intact.

arrive at the following unigram distributions:

$$p_u(h_\perp) \stackrel{\text{def}}{=} \sum_{h \in \mathrm{H}} \mathbb{1}\{h_\perp = \boldsymbol{P}h\} p_u(h) \quad (7)$$

$$p_u(h_\parallel) \stackrel{\text{def}}{=} \sum_{h \in \mathrm{H}} \mathbb{1}\{h_\parallel = (\boldsymbol{I}_d - \boldsymbol{P})h\} p_u(h) \quad (8)$$

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

$$q_u(x, c, h_\parallel, h_\perp) \stackrel{\text{def}}{=} \sum_{\boldsymbol{x}_{<t} \in \Sigma^*} \iota(c \mid x, \boldsymbol{x}_{<t}) \quad (9)$$
$$p_{\mathrm{LM}}(x \mid h_\parallel, h_\perp)\, p_{\mathrm{LM}}(\boldsymbol{x}_{<t})\, p_u(h_\parallel)\, p_u(h_\perp)$$

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

$$\mathrm{I}_q(C; H_\perp) \stackrel{\text{def}}{=} \quad (10)$$
$$\sum_{c \in \mathcal{C}} \sum_{h_\perp \in \mathrm{H}_\perp} q_u(c, h_\perp) \log \frac{q_u(c, h_\perp)}{q_u(c) q_u(h_\perp)}$$

Importantly, Eq. (10) is minimized by the correct subspace in our example in §3.2. Note that $\mathrm{I}_q(C; H_\parallel)$ can also be obtained by marginalizing out $h_\perp$ instead. Finally, we define $\mathrm{I}_q(X; H_\perp \mid C)$ by using q_u instead of p_u in Eq. (5).

3.4 Erasure and Encapsulation

We now give formal definitions of erasure and encapsulation based on Eq. (10). These two notions, combined, determine the extent to which a projection matrix $\boldsymbol{P}$ has decomposed the representation space into concept and non-concept subspaces.

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

As $\varepsilon \to 0$, the subspace $\mathrm{S}^\perp_{\mathcal{C}}$ characterized by an ε-eraser $\boldsymbol{P}$ for concept set $\mathcal{C}$ with respect to 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 $\boldsymbol{P}$, projecting onto its orthogonal complement with $\boldsymbol{I}_d - \boldsymbol{P}$ preserves nearly all of the information.

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

The quantity $\mathrm{I}_q(C; H) - \mathrm{I}_q(C; H_\parallel)$ 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. In App. A, we show that the mutual information can be additively decomposed: $\mathrm{I}_q(C; H) = \mathrm{I}_q(C; H_\perp) + \mathrm{I}_q(C; H_\parallel)$.

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 $\boldsymbol{P}$ could erase much of the non-concept related information from $\mathrm{S}^\perp_{\mathcal{C}}$. 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., 2022; Ravfogel et al., 2020, 2022a,b; Elazar et al., 2021). We, too, need an information-theoretic notion of preservation of non-concept information in $H_\perp$.

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

Definition 3.3 (Counterfactual Containment)**.** *Let $\boldsymbol{P}$ be an eraser for concept set $\mathcal{C}$ with respect to* H*. Let $H_\parallel \stackrel{\text{def}}{=} (\boldsymbol{I}_d - \boldsymbol{P})\, H$ be an $\mathbb{R}^d$-valued random variable. Then, we say that $\boldsymbol{P}$ is ε-**contained** with respect to* H *and $\mathcal{C}$ if $\mathrm{I}_q(X; H_\parallel \mid C) < \varepsilon$.*

Lastly, we define stability to measure how much non-concept information about the next word

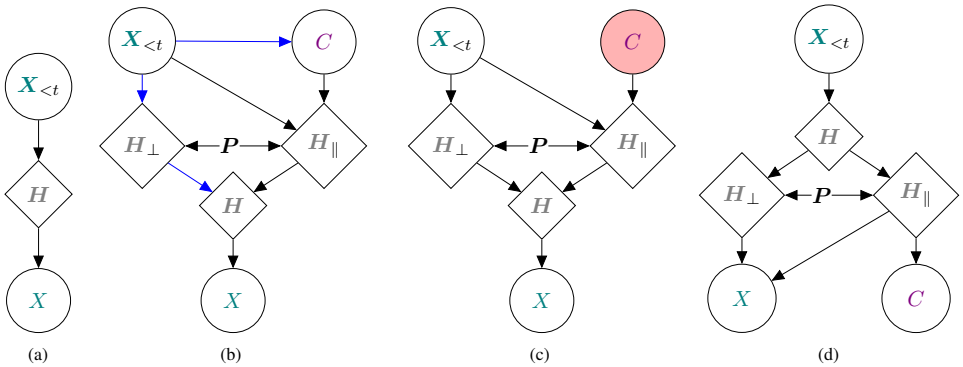

Figure 2: Causal graphical models that demonstrate how a concept may have a causal effect on word generation. Circles represent random variables and diamonds represent deterministic variables. $\boldsymbol{X}_{<t}, C, X$ represent the random variables for the textual context, the underlying concept,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

is *preserved* in the non-concept subspace $H_\perp$. Ideally, this should be as close as possible to the information present in the entire representation space, ignoring the information about the concept.

Definition 3.4 (Counterfactual Stability)**.** *Let $\boldsymbol{P}$ be an eraser for concept set $\mathcal{C}$ with respect to H. Let $H_\perp \stackrel{\text{def}}{=} \boldsymbol{P}H$ be an $\mathbb{R}^d$-valued random variable. Then, we say that $\boldsymbol{P}$ is an ε-**stabilizer** with respect to H and $\mathcal{C}$ if $\mathrm{I}_q(X; H \mid C) - \mathrm{I}_q(X; H_\perp \mid C) < \varepsilon$.*

The data processing inequality once again ensures that $\mathrm{I}_q(X; H \mid C) - \mathrm{I}_q(X; H_\perp \mid C) \geq 0$. 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 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. The Σ^*-valued random variable $\boldsymbol{X}_{<t}$ represents the textual context that was previously sampled from the model, H is the deterministic contextual representation, and X the word which is sampled using H.

To enable controlled generation with respect to the concept, we introduce a $\mathcal{C}$-valued latent variable C in the generation process, as shown in Fig. 2b. We make two assumptions about C. First, we assume that the distribution of C is influenced by the textual context $\boldsymbol{X}_{<t}$, and, moreover, that C is not *fully* determined by the context $\boldsymbol{x}_{<t}$, i.e., 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 C have an important implication: $\boldsymbol{X}_{<t}$ is no longer the only source of stochasticity in H, as in Fig. 2a. Rather, we assume that both $\boldsymbol{X}_{<t}$ as well as C influence the representation H, i.e., $h = h\left(\boldsymbol{x}_{<t}, c\right)$. Although this construction is not the norm in neural language models, it is a minor departure from reality that greatly enables our model.

We note that our causal structure in Fig. 2b differs from the high-level causal abstraction proposed by Geiger et al. (2023) for concept erasure via linear projection in the representation space. The authors don't include a concept-valued random variable in their causal abstraction. Relatedly, they argue that iterative null space projection (INLP, Ravfogel et al., 2020) attempts to determine

whether a concept is used by a model, not how it is used. In §1, we contend that language models rely on some notion of linguistic concepts like `verbal-number`, since they consistently predict the correct value. Our intrinsic information-theoretic framework helps us identify both whether *and* how a concept encoding is used by a language model because we measure changes in the model's predictions when projecting representations onto the concept and non-concept subspaces.

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 $\mathrm{do}(C = c)$) removes the edge $\boldsymbol{X}_{<t} \rightarrow C$, which lets us easily compute the distribution over the next word after intervention as follows

$$p(x \mid \boldsymbol{H}_\perp = \boldsymbol{h}_\perp, \mathrm{do}(C = c)) \tag{11}$$
$$= \sum_{g \in \mathrm{H}} p(x \mid \boldsymbol{H} = \boldsymbol{h}_\perp + (\boldsymbol{I}_d - \boldsymbol{P})g)\, p(g \mid c)$$

where, as shown in Fig. 2b, we assume that $\boldsymbol{h}_\perp$ is deterministic given the context $\boldsymbol{x}_{<t}$. g is an $\mathbb{R}^d$-valued contextual representation that encodes a textual context $\boldsymbol{x}'_{<t}$ with concept value c. With high probability, $\boldsymbol{h}(\boldsymbol{x}_{<t})$ and $g(\boldsymbol{x}'_{<t})$ will be different. This is the logical conclusion of our decision to treat $\boldsymbol{h}_\perp$ and $\boldsymbol{h}_\parallel$ 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. (9) 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 $\boldsymbol{P}$. Under the distribution*

$$p_{\mathrm{do}}(x, \boldsymbol{h}_\perp, \boldsymbol{h}_\parallel, c) = p(x \mid \boldsymbol{h}_\perp, \boldsymbol{h}_\parallel) \tag{12}$$
$$p(\boldsymbol{h}_\perp \mid \mathrm{do}\,(C = c))\, p(\boldsymbol{h}_\parallel \mid \mathrm{do}\,(C = c))\, p(c)$$

we have that $\boldsymbol{P}$ is an ε-eraser, $\boldsymbol{I}_d - \boldsymbol{P}$ is an ε-encapsulator, $\boldsymbol{I}_d - \boldsymbol{P}$ is an ε-container and $\boldsymbol{P}$ 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 $p(\boldsymbol{h}_\parallel)$ and $p(\boldsymbol{h}_\perp)$ 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 $\boldsymbol{h}$ of a language model. Then, they perform controlled generation by conditioning on a concept value $C = c$ and applying Bayes' rule as follows:

$$p(x \mid \boldsymbol{x}_{<t}, C = c) \tag{13}$$
$$\propto p(C = c \mid (\boldsymbol{I}_d - \boldsymbol{P})\boldsymbol{h}(\boldsymbol{x}_{<t}))\, p(x \mid \boldsymbol{x}_{<t})$$

We illustrate the causal structure implied by this approach in Fig. 2d. We use $\boldsymbol{P}$ 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 Experimental Setup

In the remainder of the paper, we test our framework empirically. Specifically, we answer three questions. First, are we able to find a projection matrix $\boldsymbol{P}$ that meets our definitions in §3, across multiple concepts and models? Second, can we use the resulting concept subspace to successfully control the model's generation behavior, as theorized in §4? Finally, how does our

[9] Thus, we assume that the classifier is restricted to looking at $\boldsymbol{H}_\parallel$ to make its prediction.

do-intervention compare to other concept-based explanation methods on the CEBaB benchmark?

5.1 Linguistic Concepts

Concepts and Models. We perform our analysis on two linguistic concepts, `verbal-number` in English with $\mathcal{C} = \{\text{sg}, \text{pl}, \text{n/a}\}$ and `grammatical-gender` in French with $\mathcal{C} = \{\text{fem}, \text{msc}, \text{n/a}\}$. For each of these concepts, we study the representation spaces of autoregressive language models, namely GPT2 (Radford et al., 2019) and Llama 2 (Touvron et al., 2023).[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 preprocessing 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.

5.2 CEBaB Benchmark

The CEBaB (Abraham et al., 2022) benchmark dataset consists of original restaurant reviews in English, annotated for their overall sentiment on a 5-star scale, as well as aspect-level sentiment labels on four concepts, `ambiance`, `food`, `noise`, and `service`, with concept values $\mathcal{C} = \{\text{pos}, \text{neg}, \text{n/a}\}$. The dataset includes human-written counterfactual reviews, where the text of the original review is altered to create pairs of reviews that differ according to a single aspect-level concept. We use this dataset in two experiments requiring different setups. In this section, we outline how we adapt the dataset to be able to learn $\boldsymbol{P}$ and compute our information-theoretic framework on the four CEBaB concepts. Later, in §6.3, we also test our do-intervention on the CEBaB benchmark. For that experiment, we refer the reader to Abraham et al. (2022) for experiment details, since we simply plugged our do-intervention into their pipeline.

Data and Models. For CEBaB aspect-level sentiment concepts, we use the pre-processing steps in the CEBaB repository[11], yielding train-exclusive, dev and test data splits. To adapt CEBaB to an autoregressive language modeling setting, we append a concept-specific prompt at the end of the restaurant review, e.g., *... The food was*. This prompt induces the model into predicting either a pos or neg adjective depending on the contents of the review with respect to that concept.[12] We randomly sample one of five concept-related prompts for each review and append it to the end of the review text (see Table 4 in App. C for prompts). Then, we take the last hidden state of the prompt to learn $\boldsymbol{P}$ and compute our counterfactual mutual information. This prompt-based approach enables the use of our framework for a wide range of complex concepts. Lastly, we study these four concepts using GPT2 (Radford et al., 2019) and Llama 2 (Touvron et al., 2023), using the same transformers implementations as for `verbal-number` (Wolf et al., 2020).

5.3 Shared Experimental Setup

Concept Definition. 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 define a concept via a list of words for each concept value. For linguistic concepts, we construct our lists of words by using 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 CEBaB concepts, we prompt ChatGPT (OpenAI, 2022) to create lists of pos and neg adjectives relating to each concept. In this case, the lists do not lend themselves to pairing, but include equal numbers of words for each

[10] We rely on the implementations in the transformers library (Wolf et al., 2020), namely: `gpt2-large` and `meta-llama/Llama-2-7b-hf` for `verbal-number` and `gpt2-base-french` for `grammatical-gender`.

[11] `https://github.com/CEBaBing/CEBaB`

[12] Initially, we learned $\boldsymbol{P}$ without the prompt, using the last hidden state of the review to replicate Abraham et al.'s (2022) training process for $\boldsymbol{P}$ using INLP (Ravfogel et al., 2020). Poor erasure performance led us to use prompts for both learning and evaluation.

Concept	Model	$\mathrm{I}(C;H)$	Concept Information				Non-Concept Information		
			Erasure (↑)	Corr. Erasure (↑)	Encaps. (↑)	Reconst. (↑)	$\mathrm{I}(X;H \mid C)$	Containment (↑)	Stability (↑)
number	gpt2-large	0.50±0.04	0.78±0.04	0.69±0.04	0.52±0.03	0.74±0.04	1.02±0.15	0.87±0.02	1.00±0.02
number	llama2	0.49±0.06	0.76±0.05	0.75±0.04	0.55±0.05	0.78±0.08	1.07±0.07	0.78±0.01	1.05±0.02
gender	$\text{gpt2-base}_{\text{fr}}$	0.46±0.04	0.57±0.04	0.49±0.08	0.34±0.03	0.77±0.03	2.02±0.13	0.93±0.01	0.95±0.01
ambiance	gpt2-large	0.25±0.02	0.84±0.04	0.97±0.01	0.09±0.02	0.25±0.04	0.41±0.10	0.77±0.07	0.76±0.13
ambiance	llama2	0.35±0.03	0.39±0.04	0.18±0.05	0.46±0.03	1.07±0.06	0.58±0.13	0.72±0.06	0.84±0.07
food	gpt2-large	0.28±0.02	0.74±0.04	0.27±0.12	0.59±0.04	0.85±0.06	0.44±0.09	0.81±0.03	0.60±0.09
food	llama2	0.41±0.03	0.76±0.03	0.68±0.03	0.97±0.07	1.21±0.06	0.66±0.07	0.73±0.03	0.81±0.05
noise	gpt2-large	0.21±0.01	0.78±0.01	0.60±0.03	0.35±0.05	0.57±0.05	0.31±0.11	0.71±0.14	0.82±0.10
noise	llama2	0.28±0.03	0.29±0.08	0.21±0.10	0.43±0.08	1.15±0.05	0.60±0.10	0.73±0.05	0.73±0.09
service	gpt2-large	0.35±0.03	0.65±0.03	0.40±0.11	0.44±0.05	0.79±0.05	0.39±0.11	0.74±0.08	0.74±0.14
service	llama2	0.39±0.02	0.51±0.04	0.25±0.05	0.86±0.07	1.35±0.09	0.53±0.13	0.68±0.08	0.86±0.05

Table 2: Information-theoretic evaluation results. The first set of columns quantifies concept information in concept and non-concept subspaces, with total concept information $\mathrm{I}(C;H)$ alongside the *Correlational Erasure Ratio*, and the *Counterfactual Erasure*, *Encapsulation* and *Reconstructed* ratios. The second set of columns describes the preservation of non-concept information, with total non-concept information $\mathrm{I}(X;H \mid C)$ alongside the *Containment Ratio* and *Stability Ratio*. We expect values of the ratios to fall within $[0,1]$, where higher is better. Each entry shows standard deviation over random restarts. The first set of rows shows linguistic concepts verbal-number and grammatical-gender, split by model. The second set of rows shows CEBaB aspect-level sentiment concepts ambiance, food, noise, and service, with $\boldsymbol{P}$ trained on the CEBaB dataset.

concept (see Table 5 for word lists).

Computing word probabilities. A word can be tokenized to one or more tokens. We assume that the tokenization for each word is unique. When computing the probability of a multi-token word while intervening on the representation space, e.g., $q_u(x|h_\perp)$, we apply concept erasure only when computing the probability of the first token in the word. Subsequent token probabilities are obtained by recomputing h without intervention. We then sum log probabilities of each successive token, as well as the log probability that the token after the word is either the beginning of a new word, punctuation, or the EOS token.

Finding the Concept Subspace. We find $\boldsymbol{P}$ 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 $\widetilde{p}_u$ with respect to $\boldsymbol{P}$, which constitutes a lower bound on the *correlational* $\mathrm{I}(C;H_\perp)$. Results are reported for three $\boldsymbol{P}$ estimates obtained from randomized train, test splits for each concept, and three random restarts of the experiment for each $\boldsymbol{P}$.

Estimating $p_u(h)$, $p_u(h_\perp)$, and $p_u(h_\parallel)$. For each model, we generate strings by repeatedly sampling from the language model using ancestral and nucleus sampling, starting with BOS, until the model's context size limit is reached or EOS is sampled. At each sampling step, we save the pair $(x_t, h(\boldsymbol{x}_{<t}), c_t)$, where c_t is the concept label of the sample. We then compute sums over H by randomly sampling from our dataset of generated $h(\boldsymbol{x}_{<t})$ assuming a uniform distribution. We apply $\boldsymbol{P}$ to get $h_\perp$ and $h_\parallel$.

The concept value n/a. We exclude n/a from our concept set when learning $\boldsymbol{P}$ and when measuring concept information. We are most interested in the behavior of the model when generating non-n/a words. In the overwhelming majority of textual contexts, n/a is far more likely and erasure is trivial. As such, our train sets for LEACE and test sets for computing our metrics contain only non-n/a context strings. This means $\boldsymbol{P}$ only learns to erase the distinction between non-n/a concept values, so we again exclude n/a when computing concept information metrics of erasure, encapsulation, and partition reconstruction defined in §3.4. We can, however, include n/a words for non-concept information metrics of containment and stability defined in §3.5. We estimate this by randomly sampling and computing the probabilities of 3k non-concept words in the concept language with each random restart of our evaluation pipeline.

6 Results

6.1 Partitioning of Concept Information

In Table 2, we test empirically whether LEACE (Belrose et al., 2023) yields a $\boldsymbol{P}$ that performs well according to our counterfactual information-

theoretic framework defined in §3. To facilitate the interpretation of our results, we reformulate our definitions into ratios with values between 0 and 1, such that a higher value is better:

Ratio	Definition
Erasure	$1 - \frac{\mathrm{I}_q(C;H_\perp)}{\mathrm{I}(C;H)}$
Corr. Erasure	$1 - \frac{\mathrm{I}(C;H_\perp)}{\mathrm{I}(C;H)}$
Encapsulation	$\frac{\mathrm{I}_q(C;H_\parallel)}{\mathrm{I}(C;H)}$
Reconstructed	$\frac{\mathrm{I}_q(C;H_\parallel)+\mathrm{I}_q(C;H_\perp)}{\mathrm{I}(C;H)}$
Containment	$1 - \frac{\mathrm{I}_q(X;H_\parallel\|C)}{\mathrm{I}(X;H\|C)}$
Stability	$\frac{\mathrm{I}_q(X;H_\perp\|C)}{\mathrm{I}(X;H\|C)}$

Table 2 shows that LEACE can yield an adequate representation space partitioning according to our framework, but does not drive the counterfactual erasure and encapsulation ratios to 1. The mean erasure ratio across all datasets and models is 0.64 and the encapsulation ratio is 0.50. This indicates that (a) the non-concept representation components $H_\perp$ still contain some information about the concept and (b) the concept representation components $H_\parallel$ cannot fully determine the concept c. Empirically, it is difficult to determine whether these findings are due to the LEACE objective falling victim to spurious correlations, as described in §3.2, or to the concept encoding being non-linear. For CEBaB concepts, the correlational erasure ratio is typically far lower than the counterfactual one. This suggests that while spuriously correlated information remains in $H_\perp$, the actual concept direction has been identified more accurately than a simple correlational analysis would imply.

A containment ratio less than 1 indicates that $H_\parallel$ generally contains some non-concept information, with a mean value of 0.77. Since all concepts are binary after dropping n/a, the concept subspaces are one-dimensional. The amount of non-concept information contained in a one-dimensional subspace suggests that the concept features are highly correlated with non-concept features. The stability ratio has a mean of 0.83, confirming that LEACE does some damage to non-concept information in $H_\perp$. We note, however, that for our simple linguistic concepts, this damage is minimal.

Lastly, 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`.

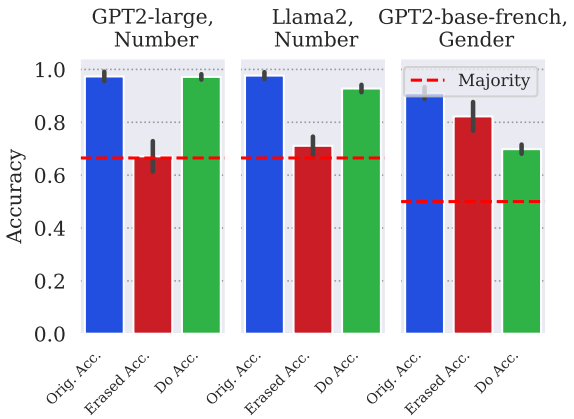

Figure 3: Controlled generation experiment. Reported values are computed on (context, fact, foil) samples from the test split of our curated datasets of natural text used to train LEACE. *Orig. Acc* refers to the accuracy with which the model chooses fact over foil using original representations. *Erased Acc.* is the accuracy after erasure, using our counterfactual $q_u(x|h_\perp)$ distribution. *Do Acc.* measures, for example for Do(C=sg) (see Eq. (11)), 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 aggregate over sg and pl contexts in the test set.

6.2 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 C. We now test this causal model empirically by computing the do-intervention in Eq. (11). 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 $\mathrm{do}(C = \mathrm{pl})$ successful if $p(x \mid H_\perp = h_\perp, \mathrm{do}(C = \mathrm{pl}))$ 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 of results 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`

Model	Method	Binary Overall Sentiment			3-way Overall Sentiment			5-way Overall Sentiment		
		ICaCE-cosine (↓)	ICaCE-L2 (↓)	ICaCE-normdiff(↓)	ICaCE-cosine (↓)	ICaCE-L2 (↓)	ICaCE-normdiff(↓)	ICaCE-cosine (↓)	ICaCE-L2 (↓)	ICaCE-normdiff(↓)
bert-base-uncased	Best CEBaB	**0.64±0.05**	**0.31±0.00**	**0.30±0.00**	**0.54±0.04**	**0.56±0.00**	**0.48±0.00**	0.63±0.01	**0.74±0.02**	**0.54±0.02**
bert-base-uncased	INLP	0.79±0.00	0.52±0.05	0.52±0.05	0.70±0.01	0.58±0.02	0.55±0.01	0.60±0.02	0.80±0.02	0.72±0.03
bert-base-uncased	Do-INLP	0.78±0.01	0.55±0.01	0.55±0.01	0.68±0.01	0.62±0.04	0.53±0.02	**0.54±0.01**	0.77±0.02	0.61±0.03
bert-base-uncased	LEACE	0.77±0.01	0.36±0.01	0.35±0.01	0.78±0.03	0.57±0.01	0.53±0.01	0.77±0.03	0.80±0.02	0.71±0.02
bert-base-uncased	Do-LEACE	0.75±0.01	0.37±0.01	0.37±0.01	0.69±0.02	0.57±0.01	0.49±0.01	0.58±0.02	0.75±0.02	0.61±0.03
gpt2	Best CEBaB	**0.58±0.01**	0.29±0.00	0.29±0.00	**0.50±0.01**	**0.52±0.01**	**0.42±0.01**	**0.59±0.01**	**0.60±0.02**	**0.40±0.01**
gpt2	INLP	1.00±0.00	0.49±0.04	0.44±0.03	1.00±0.00	0.65±0.04	0.52±0.01	1.00±0.00	0.72±0.02	0.58±0.02
gpt2	Do-INLP	0.98±0.02	0.30±0.01	0.30±0.00	0.99±0.02	0.53±0.01	0.51±0.01	0.99±0.01	0.68±0.02	0.66±0.02
gpt2	LEACE	1.00±0.00	0.31±0.01	0.30±0.00	1.00±0.00	0.53±0.01	0.51±0.01	1.00±0.00	0.68±0.02	0.66±0.02
gpt2	Do-LEACE	0.99±0.06	**0.28±0.00**	**0.28±0.00**	0.99±0.04	**0.52±0.01**	0.51±0.01	0.99±0.02	0.68±0.02	0.67±0.02
roberta-base	Best CEBaB	**0.70±0.03**	**0.29±0.01**	**0.29±0.01**	**0.62±0.02**	0.55±0.01	0.48±0.00	0.64±0.01	**0.78±0.01**	**0.59±0.01**
roberta-base	INLP	0.80±0.00	0.32±0.02	0.32±0.02	0.72±0.01	0.55±0.00	0.54±0.01	0.59±0.01	0.84±0.01	0.81±0.01
roberta-base	Do-INLP	0.79±0.00	0.30±0.06	0.30±0.06	0.70±0.01	**0.52±0.01**	0.47±0.02	**0.56±0.01**	0.80±0.00	0.72±0.01
roberta-base	LEACE	0.82±0.01	0.31±0.01	0.31±0.01	0.83±0.01	0.54±0.01	0.51±0.01	0.83±0.01	0.83±0.01	0.80±0.01
roberta-base	Do-LEACE	0.77±0.01	0.31±0.01	0.31±0.01	0.70±0.01	0.53±0.01	**0.47±0.01**	0.59±0.01	0.81±0.01	0.76±0.01

Table 3: CEBaB benchmark results for our causal do-intervention. In three sets of columns, we report empirical Individual Causal Concept Effect (ICaCE) losses for 2-, 3- and 5-class overall sentiment prediction. Three types of losses are reported, lower is better for all three. *ICaCE-cosine* indicates whether the estimated and observed effect have the same direction, without accounting for magnitude. *ICaCE-L2* compares the Euclidean norm of the difference of the estimated and observed effect, and therefore reflects both magnitude and direction differences. *ICaCE-normdiff* is the absolute difference between the Euclidean norms of the observed and estimated effects, thereby honing in on magnitude differences. Each entry shows mean ± standard deviation over random restarts. We report results for the best performing method from CEBaB (Abraham et al., 2022), INLP (Ravfogel et al., 2020), and LEACE (Belrose et al., 2023). We apply our causal do-intervention from §4.2, this time in a classification setting, using $\boldsymbol{P}$ returned by INLP and LEACE. See §6.3 for discussion of these results.

across two models. By acting solely in our concept subspace, we are able to almost match the original accuracy for both models. Results for `gpt2-base-french` are much worse, since the do-intervention actually has lower accuracy than erasure. Viewed together with results in §6.1, this confirms that our causal structure only holds given an adequate $\boldsymbol{P}$ under our counterfactual framework. Nonetheless, the success of the do-intervention on `verbal-number` despite an imperfect partitioning of concept information suggests that identifying the causal concept direction is not a necessary requirement for causal concept-based controlled generation, so long as $\boldsymbol{H}_{\|}$ contains a significant share of concept information.

6.3 CEBaB Benchmark Results

As explained in §7, our work joins an ongoing conversation about whether interpretability methods such as linear concept erasure, which operate on low-level neural representations, are good estimators for a high-level causal narrative, e.g., the process by which a language model generates grammatical text (Geiger et al., 2023). The CEBaB benchmark (Abraham et al., 2022) evaluates concept-based explanation methods as causal effect estimators by comparing the change in a sentiment classifier's predictions resulting, e.g., from concept erasure, against a ground truth. This ground truth is the change in the classifier's predictions from changing the concept value in the input text, without modifying other aspects of the restaurant review.

In Table 3, we report CEBaB results for INLP (Ravfogel et al., 2020) and LEACE (Belrose et al., 2023), our do-intervention with the projection matrices returned by these methods, and the best performing score from other methods in CEBaB. Results show that our do-intervention improves the performance of INLP and LEACE on CEBaB. For LEACE, our do-intervention improves the directional (cosine) loss, but not the magnitude losses. The inverse is true for INLP. Across models and tasks, our method occasionally beats other methods on CEBaB.

7 Related Work

The use of concept subspaces for estimating causal effects and performing causal interventions is a well-studied area. Feder et al. (2021) introduce a method for estimating the causal effect of a concept encoding on a language model's predictions, in a manner that accounts for confounders. They use adversarial training to obtain counterfactual representations with a different concept value, without

changing the textual context. The CEBaB benchmark (Abraham et al., 2022) builds on this work by providing a dataset of textual counterfactuals, enabling the comparison of different concept intervention methods. Arora et al. (2024) also provide textual counterfactuals to benchmark a much larger set of linguistics tasks. Elazar et al. (2021) and Lasri et al. (2022) measure changes in model accuracy on a concept-related task as a test of whether a model uses a linear concept encoding. Jacovi et al. (2021), Ravfogel et al. (2021), Park et al. (2023), Geiger et al. (2024), and Wu et al. (2024) create concept counterfactual representations using linear projection. Geiger et al. (2021, 2023) propose a theoretical framework relating low-level interpretability methods such as linear concept erasure to high level causal phenomena such as text generation.

Previous work in linear concept erasure 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 semantics of the representation space were affected by erasure using SimLex-999 (Hill et al., 2015), which is different from whether the language model's predictions have changed. Elazar et al. (2021) assess damage to p_{LM} via two tests: First, they test the model's ability to recover task performance after finetuning, and second, they 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.

8 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., 2022), 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, because it is computationally intractable due to nested sums over the infinite representations space $\mathbb{H}$. We leave the development of a tractable approximation to future work. 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

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. 2022. CEBaB: Estimating the causal effects of real-world concepts on NLP model behavior. In *Advances in Neural Information Processing Systems*, volume 35, pages 17582–17596. Curran Associates, Inc.

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.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. *arXiv:2402.12560*.

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 SIGNLL 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.

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.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Wikimedia Foundation. 2023. Wikimedia downloads.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc.

Atticus Geiger, Chris Potts, and Thomas Icard. 2023. Causal abstraction for faithful model interpretation.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.

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.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. 2022. Probing classifiers are unreliable for concept removal and detection. In *Advances in Neural Information Processing Systems*, volume 35, pages 17994–18008. Curran Associates, Inc.

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.

OpenAI. 2022. Introducing ChatGPT.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

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 1rst Conference for Italian Computational Linguistics (CLiC-it 2014)*, Pisa, Italy.

Manuela Sanguinetti and Cristina Bosco. 2015. *ParTUT: The Turin University Parallel Treebank*, chapter 1. Springer International Publishing, Cham.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou,

Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024. ReFT: Representation finetuning for language models. *arXiv:2404.03592*.

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 Decomposing $\mathrm{I}_q(C; H)$

Proposition A.1. *Suppose $\boldsymbol{P}$ is a ε-eraser and $(\boldsymbol{I}_d - \boldsymbol{P})$ is a ε-encapsulator of $\mathcal{C}$ with respect to H. Then, as $\varepsilon \to 0$, the following holds*

$$\mathrm{I}_q(C; H) = \mathrm{I}_q(C; H_\perp) + \mathrm{I}_q(C; H_\parallel) \tag{14}$$

Proof. On the left-hand side,

$$\begin{aligned} \mathrm{I}_q(C; H) + \varepsilon &\geq \mathrm{I}_q(C; H_\parallel) + \varepsilon && \text{(data-processing inequality)} && (15a)\\ &\geq \mathrm{I}_q(C; H_\parallel) + \mathrm{I}_q(C; H_\perp) && (\boldsymbol{P} \text{ is an } \varepsilon\text{-eraser}) && (15b) \end{aligned}$$

On the right-hand side,

$$\begin{aligned} \mathrm{I}_q(C; H) + \varepsilon &\leq \mathrm{I}_q(C; H_\parallel) + 2\varepsilon && ((\boldsymbol{I}_d - \boldsymbol{P}) \text{ is an } \varepsilon\text{-encapsulator}) && (16a)\\ &\leq \mathrm{I}_q(C; H_\parallel) + \mathrm{I}_q(C; H_\perp) + 2\varepsilon && \text{(non-negativity of MI)} && (16b) \end{aligned}$$

Combining Eq. (15b) and Eq. (16b), we have

$$\begin{aligned} \mathrm{I}_q(C; H_\parallel) + \mathrm{I}_q(C; H_\perp) &\leq \mathrm{I}_q(C; H) + \varepsilon && (17a)\\ &\leq \mathrm{I}_q(C; H_\parallel) + \mathrm{I}_q(C; H_\perp) + 2\varepsilon && (17b) \end{aligned}$$

Taking $\varepsilon \to 0$ in Eq. (17), we have Eq. (14)

$$\mathrm{I}_q(C; H) = \mathrm{I}_q(C; H_\parallel) + \mathrm{I}_q(C; H_\perp)$$

■

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 $\boldsymbol{P}$. Under the distribution*

$$\begin{aligned} p_{\mathrm{do}}(x, h_\perp, h_\parallel, c) = p(x \mid h_\perp, h_\parallel) \\ p(h_\perp \mid \mathrm{do}\,(C = c))\, p(h_\parallel \mid \mathrm{do}\,(C = c))\, p(c) \end{aligned} \tag{12}$$

we have that $\boldsymbol{P}$ is an ε-eraser, $\boldsymbol{I}_d - \boldsymbol{P}$ is an ε-encapsulator, $\boldsymbol{I}_d - \boldsymbol{P}$ is an ε-container and $\boldsymbol{P}$ 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:

$$\begin{aligned} p_{\mathrm{do}}(x, h_\perp, h_\parallel, c) &= p(x \mid h_\perp, h_\parallel)\, p(h_\perp \mid \mathrm{do}\,(C = c))\, p_{\mathrm{do}}(h_\parallel \mid \mathrm{do}\,(C = c))\, p(c) && (19a)\\ &= p(x \mid h_\perp, h_\parallel) p(h_\perp)\, p_{\mathrm{do}}(h_\parallel \mid c)\, p(c) && (19b) \end{aligned}$$

Erasure. Given Eq. (19b), we have the following joint distribution

$$\begin{aligned} p_{\mathrm{do}}(c, h_\perp) &= \sum_{h_\parallel \in \mathrm{H}_\parallel} \sum_{x \in \Sigma} p_{\mathrm{do}}(x, h_\perp, h_\parallel, c) && (20a)\\ &= \sum_{h_\parallel \in \mathrm{H}_\parallel} \sum_{x \in \Sigma} p(x \mid h_\perp, h_\parallel) p(h_\perp)\, p_{\mathrm{do}}(h_\parallel \mid c)\, p(c) && (20b)\\ &= \sum_{h_\parallel \in \mathrm{H}_\parallel} \underbrace{\left(\sum_{x \in \Sigma} p(x \mid h_\perp, h_\parallel) \right)}_{=1} p(h_\perp)\, p_{\mathrm{do}}(h_\parallel \mid c)\, p(c) && (20c)\\ &= \underbrace{\left(\sum_{h_\parallel \in \mathrm{H}_\parallel} p_{\mathrm{do}}(h_\parallel \mid c) \right)}_{=1} p(h_\perp)\, p(c) && (20d)\\ &= p(h_\perp)\, p(c) && (20e) \end{aligned}$$

The mutual information $\mathrm{I}(C; H_\perp)$ can be computed as follows

$$\mathrm{I}(C; H_\perp) = \sum_{c \in \mathcal{C}} \sum_{h_\perp \in \mathrm{H}_\perp} p_{\mathrm{do}}(c, h_\perp) \log \frac{p_{\mathrm{do}}(c, h_\perp)}{p(c)p(h_\perp)} \tag{21a}$$

$$= \sum_{c \in \mathcal{C}} \sum_{h_\perp \in \mathrm{H}_\perp} p_{\mathrm{do}}(c, h_\perp) \log \frac{p(h_\perp)\, p(c)}{p(c)p(h_\perp)} \qquad \text{(applying Eq. (20e))} \tag{21b}$$

$$= 0 < \varepsilon \tag{21c}$$

for every $\varepsilon > 0$.

Encapsulation. The following equation holds given Eq. (19b)

$$\mathrm{I}(C; H) - \mathrm{I}(C; H_\parallel) = \mathrm{I}(C; H_\parallel, H_\perp) - \mathrm{I}(C; H_\parallel) \qquad (H = H_\perp, H_\parallel) \tag{22a}$$

$$= \mathrm{I}(C; H_\perp \mid H_\parallel) \tag{22b}$$

$$= \mathrm{I}(C; H_\perp) \qquad (H_\perp, H_\parallel \text{ are independent (§3.3)}) \tag{22c}$$

$$= 0 < \varepsilon \qquad \text{(applying Eq. (21c))} \tag{22d}$$

$$\tag{22e}$$

Containment. The following joint distribution can be derived from Eq. (19b)

$$p_{\mathrm{do}}(x, h_\parallel, c = c) = \sum_{h_\perp \in \mathrm{H}_\perp} p_{\mathrm{do}}(x, h_\perp, h_\parallel, c = c) \tag{23a}$$

$$= \sum_{h_\perp \in \mathrm{H}_\perp} p(x \mid h_\perp, h_\parallel) p(h_\perp)\, p_{\mathrm{do}}(h_\parallel \mid c = c)\, p(c = c) \tag{23b}$$

$$= \sum_{h_\perp \in \mathrm{H}_\perp} \frac{p(x, h_\perp, h_\parallel)}{p(h_\perp, h_\parallel)} p(h_\perp)\, p_{\mathrm{do}}(h_\parallel \mid c = c)\, p(c = c) \tag{23c}$$

$$= \sum_{h_\perp \in \mathrm{H}_\perp} \frac{p(x, h_\perp, h_\parallel)}{\cancel{p(h_\perp)}\, p(h_\parallel)} \cancel{p(h_\perp)}\, p_{\mathrm{do}}(h_\parallel \mid c = c)\, p(c = c) \qquad (H_\perp, H_\parallel \text{ are independent (§3.3)}) \tag{23d}$$

$$= \underbrace{\sum_{h_\perp \in \mathrm{H}_\perp} p(x, h_\perp \mid h_\parallel)}_{= p(x \mid h_\parallel)}\, p_{\mathrm{do}}(h_\parallel \mid c = c)\, p(c = c) \tag{23e}$$

$$= p(x \mid h_\parallel)\, p_{\mathrm{do}}(h_\parallel \mid c = c)\, p(c = c) \tag{23f}$$

$$= p(x \mid h_\parallel, c = c)\, p(h_\parallel \mid c = c) \qquad (H_\parallel \text{ is deterministic given } C) \tag{23g}$$

The mutual information $\mathrm{I}(X; H_\parallel \mid C = c)$ can be computed as follows

$$\mathrm{I}(X; H_\parallel \mid C = c) \tag{24a}$$

$$= \sum_{x \in \Sigma} \sum_{h_\parallel \in \mathrm{H}_\parallel} p_{\mathrm{do}}(x, h_\parallel, c = c) \log \frac{p_{\mathrm{do}}(x, h_\parallel, c = c)}{p(x \mid c = c)p(h_\parallel \mid c = c)} \tag{24b}$$

$$= \sum_{x \in \Sigma} \sum_{h_\parallel \in \mathrm{H}_\parallel} p_{\mathrm{do}}(x, h_\parallel, c = c) \log \frac{p(x \mid h_\parallel, c = c)\, p(h_\parallel \mid c = c)}{p(x \mid c = c)\, p(h_\parallel \mid c = c)} \qquad \text{(applying Eq. (23g))} \tag{24c}$$

$$= \sum_{x \in \Sigma} \sum_{h_\parallel \in \mathrm{H}_\parallel} p_{\mathrm{do}}(x, h_\parallel, c = c) \log \frac{p(x \mid h_\parallel, c = c)\, p(h_\parallel \mid c = c)}{p(x \mid h_\parallel, c = c)\, p(h_\parallel \mid c = c)} \qquad (H_\parallel \text{ is deterministic given } C) \tag{24d}$$

$$= 0 < \varepsilon \tag{24e}$$

Stability. The following equation holds given Eq. (19b)

$$
\begin{aligned}
&\mathrm{I}(X; H \mid C = c) - \mathrm{I}(X; H_{\perp} \mid C = c) && && \text{(25a)}\\
=&\mathrm{I}(X; H_{\perp}, H_{\parallel} \mid C = c) - \mathrm{I}(X; H_{\perp} \mid C = c) && (H = (H_{\perp}, H_{\parallel})) && \text{(25b)}\\
=&\mathrm{I}(X; H_{\parallel} \mid H_{\perp}, C = c) && \text{(conditional mutual information)} && \text{(25c)}\\
=&\mathrm{I}(X; H_{\parallel} \mid C = c) && (H_{\perp}, H_{\parallel} \text{ are independent (§3.3)}) && \text{(25d)}\\
=&0 < \varepsilon && \text{(applying Eq. (24e))} && \text{(25e)}
\end{aligned}
$$

■

C Concept Word Lists and CEBaB Prompts

Ambiance	The ambiance was The atmosphere was The restaurant was The vibe was The setting was
Food	The cuisine was The dishes were The meal was The food was The flavors was
Noise	The ambient noise level was The background noise was The surrounding sound was The auditory atmosphere was The ambient soundscape was
Service	The service was The staff was The hospitality extended by the staff was The waiter was The host was

Table 4: List of prompts for CEBaB dataset.

verbal-number	sg	absorbs, accepts, accompanies, accounts, achieves, acknowledges, activates, adds, addresses, administers, admits, adopts, advises, advocates, affects, agrees, aims, allows, announces, appears, applies, appoints, ...
	pl	absorb, accept, accompany, account, achieve, acknowledge, activate, add, address, administer, admit, adopt, advise, advocate, affect, agree, aim, allow, announce, appear, apply, appoint, ...
grammatical-gender	msc	abbatial, absolu, actif, actuel, additionnel, administratif, afro-américain, agressif, aigu, algérien, allemand, alsacien, amer, américain, ancien, annuel, architectural, arménien, artificiel, artisanal, ...
	fem	abbatiale, absolue, active, actuelle, additionnelle, administrative, afro-américaine, agressive, aiguë, algérienne, allemande, alsacienne, amère, américaine, ancienne, annuelle, architecturale, arménienne, artificielle, artisanale, ...

Table 5: Subset of linguistic concept word lists.

ambiance	pos	cozy, elegant, inviting, charming, welcoming, intimate, sophisticated, tranquil, lively, romantic, chic, rustic, vibrant, serene, stylish, eclectic, enchanting, upscale, warm, bustling, idyllic, exquisite, radiant, harmonious, blissful, alluring, picturesque, opulent, sumptuous, dreamy, luxurious, polished, effervescent, enthralling
	neg	dingy, claustrophobic, dreary, uninviting, dull, sterile, disorganized, loud, cramped, stale, unpleasant, chaotic, gaudy, tacky, uncomfortable, grimy, stuffy, depressing, drab, cold, seedy, pretentious, overcrowded, gloomy, oppressive, grim, tense, repellent, muggy, sullen, bland, repugnant, dismal, shabby
food	pos	delicious, mouthwatering, flavorful, delectable, savory, scrumptious, tasty, heavenly, exquisite, succulent, aromatic, satisfying, appetizing, divine, gourmet, nutritious, fresh, yummy, fragrant, sumptuous, delightful, rich, zesty, indulgent, juicy, decadent, balanced, sweet, tangy, refreshing, warm, tender, crispy, herbacious, spiced
	neg	tasteless, bland, overcooked, undercooked, stale, soggy, greasy, unappetizing, flavorless, dry, tough, burnt, rancid, unpalatable, watery, disgusting, sour, bitter, mushy, unpleasant, unappealing, foul, off-putting, insipid, rubbery, dull, moldy, spoiled, repulsive, stinky, unbalanced, salty, fatty, stringy, oily
noise	pos	vibrant, lively, energetic, buoyant, festive, animated, cheerful, convivial, invigorating, buzzy, jubilant, pulsating, tranquil, serene, calm, peaceful, quiet, relaxed, soothing, gentle, mellow, harmonious, rejuvenating, vivacious, resonant
	neg	disruptive, deafening, chaotic, clamorous, unruly, boisterous, raucous, overwhelming, harsh, grating, jarring, unpleasant, discordant, intrusive, irritating, nerve-wracking, agitated, distracting, unbearable, silent, loud, obnoxious, booming, cacaphonous, blaring
service	pos	attentive, friendly, efficient, professional, courteous, prompt, welcoming, accommodating, hospitable, personable, polished, gracious, knowledgeable, warm, engaging, diligent, exemplary, seamless, outstanding, enthusiastic, meticulous, personalized, considerate, anticipatory, respectful, reliable, consistent, thoughtful, empathetic, genuine, polite, proactive, adaptable, detail-oriented, impeccable
	neg	inattentive, slow, rude, incompetent, dismissive, disorganized, impersonal, unprofessional, neglectful, indifferent, abrupt, inefficient, aloof, uncaring, clueless, forgetful, disrespectful, aggressive, insubordinate, inept, unfriendly, unaccommodating, inconsiderate, arrogant, sloppy, unknowledgeable, intrusive, negligent, careless, unresponsive, unreliable, distracted, impolite, disinterested, discourteous

Table 6: Full list of CEBaB concept words.