# Competence-Based Analysis of Language Models

**Adam Davies** **Jize Jiang** **ChengXiang Zhai**
Siebel School of Computing and Data Science
The Grainger College of Engineering
University of Illinois Urbana-Champaign
adavies4@illinois.edu

## Abstract

Despite the recent successes of large, pretrained neural language models (LLMs), comparatively little is known about the representations of linguistic structure they learn during pretraining, which can lead to unexpected behaviors in response to prompt variation or distribution shift. To better understand these models and behaviors, we introduce a general model analysis framework to study LLMs with respect to their representation and use of human-interpretable linguistic properties. Our framework, CALM (Competence-based Analysis of Language Models), is designed to investigate LLM competence in the context of specific tasks by intervening on models' internal representations of different linguistic properties using causal probing, and measuring models' alignment under these interventions with a given ground-truth causal model of the task. We also develop a new approach for performing causal probing interventions using gradient-based adversarial attacks, which can target a broader range of properties and representations than prior techniques. Finally, we carry out a case study of CALM using these interventions to analyze and compare LLM competence across a variety of lexical inference tasks, showing that CALM can be used to explain behaviors across these tasks.

## 1 Introduction

The rise of large, pretrained neural language models (LLMs) has led to rapid progress in a wide variety of natural language processing tasks [11, 17, 22]. However, these models can also be quite sensitive to minor changes in input prompts [24, 49, 48] and fail to generalize outside their training or fine-tuning distribution [77, 82]. It is usually unclear where these limitations come from, as LLM task performance is typically studied using only "black box" behavioral analysis where limitations can only be detected if they are adequately represented in evaluation datasets, which cannot cover every potentially relevant limitation using a finite dataset [58, 69]. Thus, a deeper understanding of how these models can perform as well as they usually do while exhibiting unexpected limitations is critical for ensuring robust, trustworthy, and socially-responsible LLM-enabled applications [68, 39, 87, 6], and constitutes a key question in the basic science of LLM interpretation and analysis [6, 2].

We approach this question in terms of *competence*, drawing on the traditional competence-performance distinction in linguistic theory to motivate the study of LLMs in terms of their underlying representation of language. We define LLM competence in the context a given linguistic task as the alignment between the ground-truth causal structure of the task and the LLM's latent representation of the task's structure, measured by intervening on the LLM's representation of task-causal versus spurious properties and observing how its behavior changes in response. Models leveraging causal representations to perform a task generalize better under distribution shift than those that do not [55, 3, 13], meaning that more competent LLMs are also expected to exhibit greater robustness to distribution shift.

While the representations of causal or spurious properties are not directly observable, we take inspiration from *causal probing*, which intervenes on LLMs' representations of latent properties using causal interventions to study how these representations contribute to their behavior [23, 38].

NeurIPS 2024 Workshop on Causality and Large Models (CaLM).

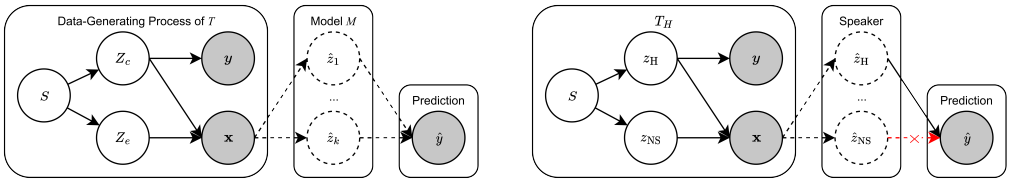


Figure 1: **Structural causal models (SCMs)** of task $\mathcal{T}$'s data-generating process and how it may be performed by model $M$ (left), or hypernym prediction task ($\mathcal{T}_H$) and how it is performed by a competent English speaker (right). Shaded and white nodes denote observed and unobserved variables, respectively. In CALM, the goal is to determine which representations $Z_j = z_j$ are causally implicated in $M$'s predictions $\hat{\mathbf{y}}$.

We introduce a general model interpretation and analysis framework, CALM (for *Competence-based Analysis of Language Models*), to study and measure LLM competence using causal probing. While CALM can be instantiated using a variety of existing causal probing interventions (e.g., [61, 59, 60, 67, 5]), we develop a new methodology for intervening on LLM representations, *Gradient-Based Interventions* (GBIs), which use white-box adversarial attacks against supervised probes to modify LLM embedding representations. GBIs are the first causal probing technique that allow one to study models' use of arbitrarily-encoded feature representations, enabling the investigation of new questions in language model interpretation. We carry out a case study of CALM using GBIs to intervene on two well-studied LLMs in order to measure and compare their competence across 14 lexical inference tasks, showing that CALM can indeed explain important patterns in behavior across these tasks by distinguishing between models' use of causal versus spurious properties.

## 2 Competence-based Analysis of Language Models

**Linguistic Competence** Linguistic competence is generally understood as the ability to utilize one's knowledge of a language in producing and understanding utterances in that language, and is typically defined in contrast with linguistic performance, which is speakers' actual use of their language in practice, considered independently of the underlying knowledge that supports it [47]. Given a linguistic task, we may understand competence in terms of the underlying linguistic knowledge that one draws upon to perform the task. If fluent human speakers rely on (implicit or explicit) knowledge of the same set of linguistic properties to perform a given task, then we may understand their performance of this task as being causally determined by these properties, and invariant to other properties. (See Appendix A.1 for further discussion of linguistic competence.)

While the study of human competence has a rich history in linguistics [15, 43, 50, 66, 46, 47], there is currently no generally accepted framework for studying LLM competence [45, 53]. In order to make the study of competence tractable in the context of LLMs, we introduce the CALM (Competence-based Analysis of Language Models) framework, which describes an LLM's competence with respect to a given linguistic task in terms of its latent representation of the causal structure of the task.

**Task Structure** Formally, given supervised task $\mathcal{T} \sim P(\mathcal{X}, \mathcal{Y})$ where the goal is to correctly predict $\mathbf{y} \in \mathcal{Y}$ given $\mathbf{x} \in \mathcal{X}$, and a collection of latent properties $\mathbf{Z} = \{Z_j\}_{j=1}^{m}$ that are (potentially) involved in generating $\mathbf{x}$, we formulate the causal structure of $\mathcal{T}$ in terms of the data-generating process

$$\mathbf{x} \sim \Pr(\mathbf{x}|\mathbf{Z}_c, \mathbf{Z}_e), \quad \mathbf{y} \sim P(\mathbf{y}|\mathbf{Z}_c) \tag{1}$$

where $\mathbf{Z}$ may be decomposed into $\mathbf{Z} = \mathbf{Z}_c \cup \mathbf{Z}_e, \mathbf{Z}_c \cap \mathbf{Z}_e = \varnothing$, where $\mathbf{Z}_c$ contains all properties that causally determine $\mathbf{y}$, and $\mathbf{Z}_e$ are the remaining properties that may be involved in generating $\mathbf{x}$ (cf. [34]). However, there may be an unobserved confounder $S$ that produces spurious correlations between $\mathbf{y}$ and $\mathbf{Z}_e$, which, if leveraged by language model $M$ in the course of predicting $\hat{\mathbf{y}}$, can lead to unexpected failures on $\mathcal{T}$ when the spurious association is broken [54]. The structural causal model (SCM)[1] of this data-generating process is visualized on the left side of both diagrams in Figure 1.

**Internal Representation** Our main concern is measuring how attributable an LLM $M$'s behavior in a given task $\mathcal{T}$ is to its representation of various properties $\mathbf{Z} = \{Z_1, ..., Z_m\}$, and how these properties correspond to the causal structure of the task. If $M$ respects the data-generating process of $\mathcal{T}$, then its behavior should be attributable only to causal properties $Z \in \mathbf{Z}_c$ (and not to environmental properties $Z \in \mathbf{Z}_e$), in which case we say that $M$ is *competent* with respect to $\mathcal{T}$. We study model

[1] An SCM is a directed acyclic graph where each node represents a variable and directed edges indicate causal dependencies (see [9] for an introduction to SCMs).

$M$'s use of each property $Z_j \in \mathbf{Z}$ by performing causal interventions $\text{do}(Z_j)$ on its representation of $Z_j$ in the course of performing task $\mathcal{T}$, and measure the impact that these interventions have on its predictions.

**Measuring Competence** We evaluate the competence of $M$ with respect to task $\mathcal{T} \sim P(\mathcal{X}, \mathcal{Y})$ by measuring its causal alignment with a *competence graph* $\mathcal{G}_\mathcal{T}$, which we define as a structural causal model (SCM) of $\mathcal{T}$ with nodes corresponding to each latent variables $Z_j \in \mathbf{Z}$ and an additional node for outputs $\mathbf{y} \in \mathcal{Y}$ and directed edges denoting causal dependencies between these variables. That is, the set of causal properties $\mathbf{Z}_c$ defined by $\mathcal{G}_\mathcal{T}$ is the set of all properties $Z_j \in \mathbf{Z}$ such that there is an edge or path from $Z_j$ to $\mathbf{y}$. To determine the extent to which $M$'s behavior is correctly explained by the causal dependencies (and lack thereof) in $\mathcal{G}_\mathcal{T}$, we measure their consistency under interventions $\text{do}(\mathbf{z})$, where setting $\mathbf{z} = \{z_j\}_{j=1}^m \sim \text{val}(\mathbf{Z})$ is a combination of values $Z_j = z_j \in \text{val}(Z_j)$ taken by each corresponding latent variable $Z_j \in \mathbf{Z}$.

The alignment of $M$ with $\mathcal{G}_\mathcal{T}$ is measured in terms of the similarity $S$ of their predictions under interventions $\text{do}(\mathbf{z})$ given input $\mathbf{x} \sim P(\mathcal{X})$, and can be computed using a given similarity metric $S : \mathcal{Y}, \mathcal{Y} \to [0, 1]$ (e.g., equality, n-gram overlap, cosine similarity, etc.) depending on the SCM $\mathcal{G}_\mathcal{T}$ and output space $\mathcal{Y}$. That is, we define $\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T})$ as $M$'s competence with respect to task $\mathcal{T}$ as a function of its alignment with corresponding task SCM $\mathcal{G}_\mathcal{T}$ under interventions $\text{do}(\mathbf{z})$ measured by similarity metric $S$, as follows:

$$\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T}) = \mathbb{E}_{\mathbf{x},\mathbf{z}\sim P(\mathcal{X},\text{val}(\mathbf{Z}))} S\big(M(\mathbf{x}|\,\text{do}(\mathbf{z})), \mathcal{G}_\mathcal{T}(\mathbf{x}|\,\text{do}(\mathbf{z}))\big) \tag{2}$$

This $\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T})$ metric (bounded by $[0, 1]$) is an adaptation of the Interchange Intervention Accuracy (IIA) metric [29, 27] to the context of causal probing, where instance-level interventions are replaced with concept-level interventions enabled by the gradient-based intervention methodology we introduce in Section 3. (See Appendix D.1 for a detailed comparison of our competence metric with IIA.)

**Causal Probing** A key technical challenge in implementing CALM (and causal probing more generally) is designing an algorithm to perform causal interventions $\text{do}(Z)$ that maximally damage the representation of a property $Z$ while otherwise minimally damaging representations of other properties $Z'$ [59]. For example, *amnesic probing* [23] uses the INLP algorithm [61] to produce interventions $g_Z$ that remove all information that is linearly predictive of property $Z$ from a set of embedding representations $\mathbf{H}$. However, when such information removal methods are used to remove representation of these properties in early LLM layers, models are often able to "recover" this representation in later layers [23, 59], which is likely due to models encoding these properties nonlinearly. Beyond recoverability, linear information removal methods like INLP also cannot account for relational properties between multiple input embeddings (see Appendix A.1). Thus, it is important to develop interventions that do not require restrictive assumptions about the structure of LLMs' representations such as linearity [76], a problem which we aim to solve in the following section.

## 3 Gradient-based Interventions

Our goal in developing gradient-based interventions (GBIs) as a causal probing technique is to enable interventions over arbitrarily-encoded LLM representations. GBIs allow users to flexibly specify the class of representations they wish to target, expanding the scope of causal probing to arbitrarily-encoded properties. We take inspiration from Kos, Fischer, and Song [35], who developed a technique to perturb latent representations using gradient-based adversarial attacks.[2] They begin by training probe $g_Z : \mathbf{h} \mapsto z$ to predict image class $z \in Z$ from latent representations $\mathbf{h} = f_{\text{enc}}(\mathbf{x})$ of images $\mathbf{x}$, where $f_{\text{enc}}$ is the encoder of a VAE-GAN [37] trained on an unsupervised image reconstruction task (i.e., $f_{\text{dec}}(f_{\text{enc}}(\mathbf{x})) = \hat{\mathbf{x}} \approx \mathbf{x}$, for decoder $f_{\text{dec}}$ and reconstructed image $\hat{\mathbf{x}}$ approximating $\mathbf{x}$). Next, gradient-based attacks like FGSM [30] and PGD [44] are performed against $g_Z$ in order to minimally manipulate $\mathbf{h}$ such that it resembles encoded representations of target image class $Z = z'$ (where $z' \neq z$, the original image class), yielding perturbed representation $\mathbf{h}'$. Finally, $\mathbf{h}$ and $\mathbf{h}'$ are each fed into the VAE decoder to reconstruct corresponding output images $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}'$ (respectively), where $\hat{\mathbf{x}}$ resembles input image class $Z = z$ and $\hat{\mathbf{x}}'$ resembles target class $Z = z'$.

We reformulate this approach in the context of causal probing as visualized in Figure 2, treating layers $L = 1, ..., l$ as the encoder and layers $L = l + 1, ..., |L|$ (composed with language modeling head

[2] Notably, Tucker, Qian, and Levy [75] developed a similar methodology without explicit use of such attacks (see Section 7).

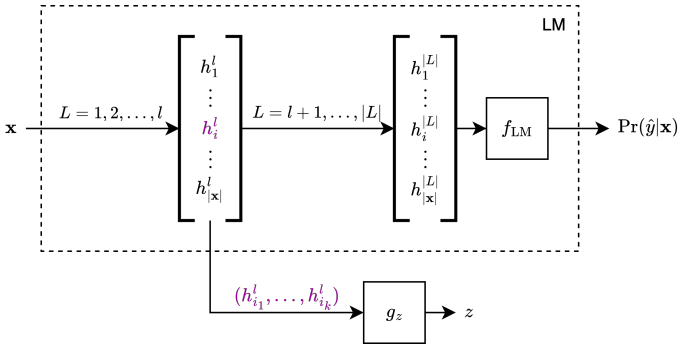


Figure 2: **Gradient-based Interventions.** Input tokens $\mathbf{x} = (x_1, ..., x_{|\mathbf{x}|})$ are passed through layers $L = 1, ..., l$, where embedding $\mathbf{h}_i^l$ (encoding the value $Z = z$) is extracted from layer $l$ and given to $g_Z$ as input. Next, the embedding is modified by gradient-based attacks on $g_Z$ to encode the counterfactual value $Z = z'$, then fed back into subsequent layers $L = l + 1, ..., |L|$ and language modeling head $f_{\text{LM}}$ to obtain the intervened predictions $M(\mathbf{x}|\operatorname{do}(Z = z'))$.

$f_{\text{LM}}$) as the decoder, allowing us to target representations of property $Z$ across embeddings $\mathbf{h}_i^l$ of token $x_i \in \mathbf{x}$ in layer $l$. We train $g_Z$ to predict $Z$ from a set of such $\mathbf{h}_i^l$, then attack $g_Z$ using FGSM and PGD to intervene on $\mathbf{h}_i^l$ (representing the original value $Z = z$), producing $\mathbf{h}_i^{l'}$ (representing the counterfactual value $Z = z'$). Finally, we replace $\mathbf{h}_i^l$ with $\mathbf{h}_i^{l'}$ in the LLMs' forward pass from layers $L = l + 1, ..., |L|$, simulating the intervention $\operatorname{do}(Z = z')$, and observe the impact on its word predictions $M(\mathbf{x}|\operatorname{do}(Z = z'))$.

There are several key benefits associated with GBIs relative to existing causal probing interventions (e.g., they can be applied to any differentiable probe), as well as some important limitations (e.g., lack of theoretical guarantees). as we discuss in detail in Appendices A.2 and B.1 (respectively).

## 4 Experiments

In this work, we begin by examining BERT [21] and RoBERTa [41],[3] two language models which have been extensively studied in the context of probing [63, 61, 42, 23, 38]. Our primary goal in the following experiments is to develop and test an experimental implementation of CALM using GBIs in the context of comparatively small, well-studied models and tasks in order to validate whether CALM can explain behavioral findings of earlier work in this simplified environment. (We motivate this choice in greater detail in Appendix B.2.)

**Tasks** We use the collection of 14 lexical inference tasks included in the ConceptNet [70] subset of LAMA [56], each of which are formulated as a collection of cloze prompts [40]. For example, the LAMA "IsA" task contains ∼2K hypernym prompts corresponding to the "IsA" ConceptNet relation (including, e.g., "A laser is a `[MASK]` which creates coherent light.", where the task is to predict that the `[MASK]` token should be replaced with "device", a hypernym of "laser"), with the remaining 13 LAMA ConceptNet tasks corresponding to other lexical relations such as "PartOf", "HasProperty", and "CapableOf". (See Appendix C.1 for additional details.)

Using these task datasets allow us to test how the representation of each relation is used across all other tasks. In the context of a single task $\mathcal{T}_j$, intervening on a model's representation of the task-causal relation $Z_j$ allows us to measure the extent to which its predictions are attributable to its representation of the causal property $\mathbf{Z}_c = \{Z_j\}$ (where a large impact indicates competence). On the other hand, intervening on the representations of the other 13 lexical relations $Z_k \in \mathbf{Z}_e$ allows us (in the aggregate) to measure how much the model is performing task $\mathcal{T}_j$ by leveraging representations of general, non-causal lexical information (where a large impact indicates incompetence).[4]

**Experimentally Measuring Competence** Given LLM $M$ and task $\mathcal{T}$, measuring the competence $\mathcal{C}_{\mathcal{T}}(M|\mathcal{G}_{\mathcal{T}})$ of $M$ given $\mathcal{G}_{\mathcal{T}}$ requires us to specify an experimental model $E = (\mathbf{Z}, \mathcal{G}_{\mathcal{T}}, S)$, where $\mathbf{Z}$ is a set of properties, $\mathcal{G}_{\mathcal{T}}$ is a competence graph for task $\mathcal{T}$, and $S$ is a scoring function that compares

[3] Specifically, `BERT-base-uncased` and `RoBERTa-base` [79].

[4] Note that the strictest interpretation of this formulation of competence makes the simplifying assumption that each non-causal property is equally (un)related to the target property, which is not always true; see Appendix B.3.

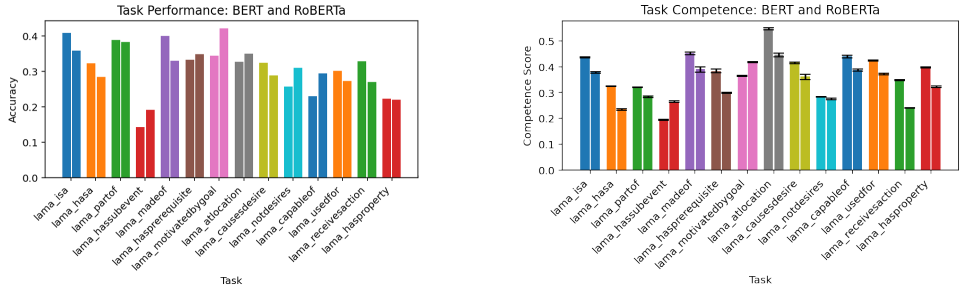


Figure 3: Performance (left) and competence (right) of BERT (left bars) and RoBERTa (right bars) for all tasks, using FGSM with $\epsilon = 0.1$. In the competence plot, y-values are the average competence score and error bars are the maximum and minimum competence score, as measured over 10 experimental iterations (each with a different randomly-initialized probe $g_Z$).

the predictions of $M$ and $\mathcal{G}_\mathcal{T}$ in order to compute the approximated $\hat{\mathcal{C}_\mathcal{T}}$. Given that each task $\mathcal{T}_i$ is defined by a single causal lexical relation $Z_i$ (i.e., $\mathbf{Z}_{c_i} = \{Z_i\}$), we model settings $\mathbf{z}$ as a collection of values $Z_j = z_j$ taken by each property $Z_j$ in the context of a specific task instance $(\mathbf{x}, \mathbf{y}) \sim \mathcal{T}_i$, where $Z_j = 1$ if $i = j$ (i.e., where the property $Z_j$ is the causal property for the task $\mathcal{T}_i$) or $Z_j = 0$ otherwise. That is, for each instance $(\mathbf{x}, \mathbf{y}) \sim \mathcal{T}_i$, the corresponding setting $\mathbf{z}$ is a one-hot vector whose $i$-th element $\mathbf{z}_i = 1$. We may specify $\mathcal{G}_{\mathcal{T}_i}$ in a similar manner: for task $\mathcal{T}_i \sim P(\mathcal{X}, \mathcal{Y})$, outputs $\mathbf{y} \in \mathcal{Y}$ are causally dependent on the property $Z_i$, and invariant to other concepts $Z_j, j \neq i.$, meaning that the only direct parent node of $\mathbf{y}$ in $\mathcal{G}_{\mathcal{T}_i}$ is $Z_i$. Finally, as we are dealing with masked language models whose output space $\mathcal{Y}$ for each task consists only of single tokens in $M$'s vocabulary $V_M$, our experimental model can define the scoring function $S$ as the overlap $\text{overlap}(\mathbf{y}_i, \mathbf{y}_j)$ for top-$k$ token predictions $\mathbf{y}_i = \{y_1, ..., y_k\} \subset V_M$, where $\text{overlap}(\cdot, \cdot)$ is the size of the intersection of each set of predictions divided by the total number of predictions $\text{overlap}(\mathbf{y}_i, \mathbf{y}_j) = \frac{|\mathbf{y}_i \cap \mathbf{y}_j|}{k}$, and $\hat{\mathcal{C}_\mathcal{T}}_k$ denotes $\hat{\mathcal{C}_\mathcal{T}}$ as measured using the top-$k$ token predictions $\mathbf{y}_i$. (See Appendix D.2 for additional details on how we compute competence in each experiment.)

**Probes** We implement probes $g_Z$ as a 2-layer MLP over each language model's final hidden layer, and train the probe on the task of classifying whether there is a particular relation $Z$ between a final-layer `[MASK]` token in the context of a cloze prompt and the final-layer object token from the "unmasked" version of the same prompt. All reported figures are the average of 10 runs of our experiment, using different randomly-initialized $g_Z$ each time. (See Appendix C.2 for further details.)

**Interventions** We implement GBIs against $g_Z$ using two gradient attack strategies, FGSM [30] and PGD [44]. We bound the magnitude of each intervention as follows: where $h$ is the input to $g_Z$ and $h'$ is the intervened representation following a GBI, $||h - h'||_\infty \leq \epsilon$. For all experiments reported in our main paper, we use FGSM with $\epsilon = 0.1$. (See Appendix C.3 for more details and PGD results.)

## 5 Results

In Figure 3, we visualize the performance and competence of BERT and RoBERTa across the test set of each LAMA ConceptNet task. Performance is measured using $(0, 1)$-accuracy, competence is measured using the experimental competence metric in Equation (3), and both metrics are averaged across the top-$k$ predictions of each model for $k \in [1, 10]$. That is, for ground truth $(\mathbf{x}, y)$ and $n = 10$, we compute accuracy and competence as follows:

$$\text{acc}(M) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}[y \in \underset{\hat{y}}{\text{top-}k} \Pr_M(\hat{y}|\mathbf{x})] \qquad \text{and} \qquad \hat{\mathcal{C}_\mathcal{T}}(M|\mathcal{G}_\mathcal{T}) = \frac{1}{n}\sum_{k=1}^{n} \hat{\mathcal{C}_\mathcal{T}}_k(M|\mathcal{G}_\mathcal{T})$$

To account for stochasticity in initializing and training probes $g_Z$, scores are also averaged over 10 randomized experiments for each target task where the probe is randomly re-initialized each time (resulting in different GBIs).

**Performance** While their accuracies on individual tasks vary, BERT and RoBERTa have quite similar aggregate performance: BERT outperforms RoBERTa on just over half (8/14) of the tasks, achieving essentially equivalent performance when averaged across all tasks (0.3099 versus 0.3094).

**Competence** Given our experimental model $E$ with $m = 14$ tasks, consider a random baseline language model $R$ whose predictions always change in response to each intervention, making equal

use of all properties in each task. $R$ would yield a competence score of $\mathcal{C}(R|\mathcal{G}_{\mathcal{T}}) = \frac{1}{m} \approx 0.0714$ for each task. Both BERT and RoBERTa score above this threshold for all tasks, meaning that their competence is consistently greater than that of a model ($R$) that does not distinguish between causal and environmental properties. However, RoBERTa is consistently less competent than BERT (on 12/14 tasks), and also has lower competence scores averaged across all tasks (0.381 vs. 0.334). Furthermore, relative performance and competence are correlated: the Spearman's Rank correlation coefficient between the average difference in accuracy and average difference in performance is a moderately strong positive correlation $\rho = 0.508$ with significance $p = 0.064$.

## 6 Discussion

**Explananda** The performance of BERT and RoBERTa on lexical inference tasks such as hypernym prediction has been shown to be highly variable under small changes to prompts [33, 62, 26, 24]. Our findings offer one possible explanation for such brittle performance: BERT and RoBERTa's partial competence in hypernym prediction indicates that it should be possible to prompt these models in a way that will yield high performance, but that its reliance on spurious lexical associations may lead it to fail when these correlations are broken – e.g., by substituting singular terms for plurals [62] or paraphrasing a prompt [24].

**Future Work** While the simplified experimental context considered in this work is a necessary first step in empirically validating our theoretical CALM framework, competence metric, and GBI methodology, we anticipate a much broader range of future research directions and potential applications for CALM. We elaborate several such directions in Appendix E.

## 7 Related Work

**Causal Probing** Most related to our work is amnesic probing [23], as discussed in Section 2. Lasri et al. [38] applied amnesic probing to study the use of grammatical number representations in performing an English verb conjugation prompt task. As this experiment involves intervening on the representation of a property which is causal with respect to the prompt task, it may be understood as an informal instantiation of CALM (albeit without considering environmental properties or measuring competence).

**Gradient-based Interventions** Tucker, Qian, and Levy [75] developed an approach similar to GBIs without explicit use of gradient-based adversarial attacks. Their methodology is equivalent to performing a targeted, unconstrained attack, where gradient updates are continually applied to embeddings until the target probe loss saturates (irrespective of perturbation magnitude). In such attacks, it is standard practice to constrain the magnitude of resulting perturbations [30, 44, 35], which we do here in order to minimize the effect of "collateral damage" done by such attacks (see Appendix C.3). Failing to impose such constraints may result in indiscriminate damage to representations [14] (see Appendix B.1 for further discussion).

**Unsupervised Probing** Instead of training supervised probes to predict a pre-specified property of interest (as we do here), an alternative approach is to train *unsupervised* probes such as Sparse Auto-Encoders (SAEs; [71, 83, 19]), which learn a "dictionary" of features that can be used to sparsely represent embeddings, and can also be used to control models' use of these learned features [10, 72]. Unlike supervised probing, unsupervised dictionary features must be retroactively interpreted in order to determine their relationship to a given task [20]; but given a suitable approach to interpreting such features (see, e.g., [8, 52]) and a sufficiently reliable method for intervening on them (cf. [14]), it is also possible to implement CALM using unsupervised probes like SAEs.

## 8 Conclusion

In this work, we introduced CALM, a general analysis framework that enables the study of LLMs' linguistic competence using causal probing, including the first quantitative measure of linguistic competence. We developed the gradient-based intervention (GBI) methodology, a novel causal probing method that can target a far greater range of representations than previous techniques, expanding the scope of causal probing to new questions in LLM interpretability and analysis. Finally, we carried out a case study of CALM using GBIs, analyzing BERT and RoBERTa's competence across a collection of lexical inference tasks, finding that even a simple experimental model is sufficient to explain their behavior across a variety of lexical inference tasks.

## Acknowledgments

This work is supported in part by the National Science Foundation and the Institute of Education Sciences, U.S. Department of Education, through Award #2229612 (National AI Institute for Inclusive Intelligent Technologies for Education). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or the U.S. Department of Education.

We thank Julia Hockenmaier, Marc E. Canby, Francesco Pinto, Bo Li, and Arindam Banerjee for helpful discussions regarding our framework and empirical analysis, as well as their valuable feedback on earlier drafts of this manuscript.

## References

[1] Eldar D Abraham et al. "Cebab: Estimating the causal effects of real-world concepts on nlp model behavior". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17582–17596.

[2] Usman Anwar et al. "Foundational challenges in assuring alignment and safety of large language models". In: *arXiv preprint arXiv:2404.09932* (2024).

[3] Martin Arjovsky et al. "Invariant risk minimization". In: *arXiv preprint arXiv:1907.02893* (2019).

[4] Yonatan Belinkov. "Probing Classifiers: Promises, Shortcomings, and Advances". In: *Computational Linguistics* 48.1 (Mar. 2022), pp. 207–219. DOI: 10.1162/coli_a_00422. URL: https://aclanthology.org/2022.cl-1.7.

[5] Nora Belrose et al. "Leace: Perfect linear concept erasure in closed form". In: *Advances in Neural Information Processing Systems* 36 (2024).

[6] Leonard Bereska and Efstratios Gavves. "Mechanistic Interpretability for AI Safety–A Review". In: *arXiv preprint arXiv:2404.14082* (2024).

[7] BigScience et al. "Bloom: A 176b-parameter open-access multilingual language model". In: *arXiv preprint arXiv:2211.05100* (2022).

[8] Steven Bills et al. "Language models can explain neurons in language models". In: *URL https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html* 2 (2023). URL: https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

[9] Stephan Bongers et al. "Foundations of structural causal models with cycles and latent variables". In: *The Annals of Statistics* 49.5 (2021), pp. 2885–2915.

[10] Trenton Bricken et al. "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning". In: *Transformer Circuits Thread* (2023). https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[11] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[12] Sébastien Bubeck et al. "Convex optimization: Algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[13] Peter Bühlmann. "Invariance, causality and robustness". In: *Statistical Science* 35.3 (2020), pp. 404–426.

[14] Marc Canby et al. "Measuring the reliability of causal probing methods: Tradeoffs, limitations, and the plight of nullifying interventions". In: *arXiv preprint arXiv:2408.15510* (2024).

[15] Noam Chomsky. "Aspects of the Theory of Syntax". In: *MIT Press* (1965).

[16] Leshem Choshen et al. "Where to start? Analyzing the potential value of intermediate models". In: *arXiv preprint arXiv:2211.00107* (2022).

[17] Aakanksha Chowdhery et al. "Palm: Scaling language modeling with pathways". In: *arXiv preprint arXiv:2204.02311* (2022).

[18] Arthur Conmy et al. "Towards Automated Circuit Discovery for Mechanistic Interpretability". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. arXiv: 2304.14997 [cs.LG].

[19] Hoagy Cunningham et al. "Sparse autoencoders find highly interpretable features in language models". In: *arXiv preprint arXiv:2309.08600* (2023).

[20] Adam Davies and Ashkan Khakzar. "The Cognitive Revolution in Interpretability: From Explaining Behavior to Interpreting Representations and Algorithms". In: *arXiv preprint arXiv:2408.05859* (2024).

[21] Jacob Devlin et al. "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[22] Abhimanyu Dubey et al. "The llama 3 herd of models". In: *arXiv preprint arXiv:2407.21783* (2024).

[23] Yanai Elazar et al. "Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 160–175. DOI: 10.1162/tacl_a_00359. URL: https://aclanthology.org/2021.tacl-1.10.

[24] Yanai Elazar et al. "Measuring and Improving Consistency in Pretrained Language Models". In: *Transactions of the Association for Computational Linguistics* 9 (Dec. 2021), pp. 1012–1031. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00410. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00410/1975957/tacl_a_00410.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00410.

[25] Nelson Elhage et al. "A mathematical framework for transformer circuits". In: *Transformer Circuits Thread* 1 (2021).

[26] Allyson Ettinger. "What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models". In: *Transactions of the Association for Computational Linguistics* 8 (Jan. 2020), pp. 34–48. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00298. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00298/1923116/tacl_a_00298.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00298.

[27] Atticus Geiger, Chris Potts, and Thomas Icard. "Causal Abstraction for Faithful Model Interpretation". In: *arXiv preprint arXiv:2301.04709* (2023).

[28] Atticus Geiger, Kyle Richardson, and Christopher Potts. "Neural Natural Language Inference Models Partially Embed Theories of Lexical Entailment and Negation". In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Ed. by Afra Alishahi et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 163–173. DOI: 10.18653/v1/2020.blackboxnlp-1.16. URL: https://aclanthology.org/2020.blackboxnlp-1.16.

[29] Atticus Geiger et al. "Inducing Causal Structure for Interpretable Neural Networks". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022, pp. 7324–7338. URL: https://proceedings.mlr.press/v162/geiger22a.html.

[30] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *International Conference on Learning Representations*. 2015. URL: http://arxiv.org/abs/1412.6572.

[31] Sven Gowal et al. "Scalable verified training for provably robust image classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4842–4851.

[32] Dirk Groeneveld et al. "OLMo: Accelerating the Science of Language Models". In: *arXiv preprint arXiv:2402.00838* (2024).

[33] Michael Hanna and David Mareček. "Analyzing BERT's Knowledge of Hypernymy via Prompting". In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 275–282. DOI: 10.18653/v1/2021.blackboxnlp-1.20. URL: https://aclanthology.org/2021.blackboxnlp-1.20.

[34] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. "Selecting data augmentation for simulating interventions". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4555–4562.

[35] Jernej Kos, Ian Fischer, and Dawn Song. "Adversarial examples for generative models". In: *2018 ieee security and privacy workshops (spw)*. IEEE. 2018, pp. 36–42.

[36] Abhinav Kumar, Chenhao Tan, and Amit Sharma. "Probing classifiers are unreliable for concept removal and detection". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17994–18008.

[37] Anders Boesen Lindbo Larsen et al. "Autoencoding beyond pixels using a learned similarity metric". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1558–1566. URL: https://proceedings.mlr.press/v48/larsen16.html.

[38] Karim Lasri et al. "Probing for the Usage of Grammatical Number". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8818–8831. DOI: 10.18653/v1/2022.acl-long.603. URL: https://aclanthology.org/2022.acl-long.603.

[39] Q Vera Liao and Jennifer Wortman Vaughan. "AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap". In: *arXiv preprint arXiv:2306.01941* (2023).

[40] Pengfei Liu et al. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing". In: *ACM Comput. Surv.* 55.9 (Jan. 2023). ISSN: 0360-0300. DOI: 10.1145/3560815. URL: https://doi.org/10.1145/3560815.

[41] Yinhan Liu et al. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[42] Zeyu Liu et al. "Probing Across Time: What Does RoBERTa Know and When?" In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 820–842. DOI: 10.18653/v1/2021.findings-emnlp.71. URL: https://aclanthology.org/2021.findings-emnlp.71.

[43] John Lyons. *Semantics: Volume 2*. Vol. 2. Cambridge university press, 1977.

[44] Aleksander Madry et al. "Towards deep learning models resistant to adversarial attacks". In: *arXiv preprint arXiv:1706.06083* (2017).

[45] Kyle Mahowald et al. "Dissociating language and thought in large language models: a cognitive perspective". In: *arXiv preprint arXiv:2301.06627* (2023). URL: https://doi.org/10.48550/arXiv.2301.06627.

[46] D. Marconi. *Lexical Competence*. A Bradford book. Bradford Book, 1997. ISBN: 9780262133333. URL: https://books.google.com/books?id=lcrEq%5C_7o5m0C.

[47] Diego Marconi. "Semantic competence". In: *The Routledge Handbook of Philosophy of Skill And Expertise*. Routledge, 2020, pp. 409–418.

[48] Moran Mizrahi et al. "State of what art? a call for multi-prompt llm evaluation". In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 933–949.

[49] Milad Moradi and Matthias Samwald. "Evaluating the robustness of neural language models to input perturbations". In: *arXiv preprint arXiv:2108.12237* (2021).

[50] Frederick J Newmeyer. "The Prague School and North American functionalist approaches to syntax". In: *Journal of Linguistics* 37.1 (2001), pp. 101–126.

[51] Catherine Olsson et al. "In-context learning and induction heads". In: *arXiv preprint arXiv:2209.11895* (2022).

[52] Gonçalo Paulo et al. "Automatically Interpreting Millions of Features in Large Language Models". In: *arXiv preprint arXiv:2410.13928* (2024).

[53] Ellie Pavlick. "Symbols and grounding in large language models". In: *Philosophical Transactions of the Royal Society A* 381.2251 (2023), p. 20220041.

[54] Judea Pearl. "Causal inference in statistics: An overview". In: *Statistics Surveys* 3 (2009).

[55] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. "Causal inference by using invariant prediction: identification and confidence intervals". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78.5 (2016), pp. 947–1012.

[56] Fabio Petroni et al. "Language Models as Knowledge Bases?" In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2463–2473. DOI: 10.18653/v1/D19-1250. URL: https://aclanthology.org/D19-1250.

[57] Tiago Pimentel et al. "The Architectural Bottleneck Principle". In: *arXiv preprint arXiv:2211.06420* (2022). URL: https://doi.org/10.48550/arXiv.2211.06420.

[58] Inioluwa Deborah Raji et al. "AI and the everything in the whole wide world benchmark". In: *arXiv preprint arXiv:2111.15366* (2021). URL: https://doi.org/10.48550/arXiv.2111.15366.

[59] Shauli Ravfogel et al. "Adversarial Concept Erasure in Kernel Space". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6034–6055. URL: https://aclanthology.org/2022.emnlp-main.405.

[60] Shauli Ravfogel et al. "Linear adversarial concept erasure". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 18400–18421.

[61] Shauli Ravfogel et al. "Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7237–7256. DOI: 10.18653/v1/2020.acl-main.647. URL: https://aclanthology.org/2020.acl-main.647.

[62] Abhilasha Ravichander et al. "On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT". In: *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*. Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 88–102. URL: https://aclanthology.org/2020.starsem-1.10.

[63] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. "A Primer in BERTology: What We Know About How BERT Works". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 842–866. DOI: 10.1162/tacl_a_00349. URL: https://aclanthology.org/2020.tacl-1.54.

[64] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. "The risks of invariant risk minimization". In: *arXiv preprint arXiv:2010.05761* (2020).

[65] Sebastian Ruder. "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (2017).

[66] Ivan A Sag and Thomas Wasow. "Performance-Compatible Competence Grammar". In: *Non-Transformational Syntax: Formal and Explicit Models of Grammar* (2011), pp. 359–377.

[67] Shun Shao, Yftah Ziser, and Shay B. Cohen. *Gold Doesn't Always Glitter: Spectral Removal of Linear and Nonlinear Guarded Attribute Information*. arXiv:2203.07893 [cs]. Mar. 2022. DOI: 10.48550/arXiv.2203.07893. URL: http://arxiv.org/abs/2203.07893 (visited on 09/22/2022).

[68] Donghee Shin. "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI". In: *International Journal of Human-Computer Studies* 146 (2021), p. 102551.

[69] Charlotte Siska et al. "Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 10406–10421.

[70] Robyn Speer, Joshua Chin, and Catherine Havasi. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 2017.

[71] Anant Subramanian et al. "Spine: Sparse interpretable neural embeddings". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 2018.

[72] Adly Templeton et al. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet". In: *Transformer Circuits Thread* (2024). URL: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

[73] Hugo Touvron et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[74] Hugo Touvron et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[75] Mycal Tucker, Peng Qian, and Roger Levy. "What if This Modified That? Syntactic Interventions with Counterfactual Embeddings". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 862–875. DOI: 10.18653/v1/2021.findings-acl.76. URL: https://aclanthology.org/2021.findings-acl.76.

[76] Francisco Vargas and Ryan Cotterell. "Exploring the Linear Subspace Hypothesis in Gender Bias Mitigation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2902–2913. DOI: 10.18653/v1/2020.emnlp-main.232. URL: https://aclanthology.org/2020.emnlp-main.232.

[77] Jindong Wang et al. "On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective". In: *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*. 2023. URL: https://openreview.net/forum?id=uw6HSkgoM29.

[78] Kevin Ro Wang et al. "Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=NpsVSN6o4ul.

[79] Thomas Wolf et al. "Huggingface's transformers: State-of-the-art natural language processing". In: *arXiv preprint arXiv:1910.03771* (2019).

[80] Zhengxuan Wu et al. "Interpretability at Scale: Identifying Causal Mechanisms in Alpaca". In: *arXiv preprint arXiv:2305.08809* (2023). arXiv: 2305.08809 [cs.LG].

[81] Karren Yang, Abigail Katcoff, and Caroline Uhler. "Characterizing and learning equivalence classes of causal DAGs under interventions". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5541–5550.

[82] Linyi Yang et al. "GLUE-X: Evaluating Natural Language Understanding Models from an Out-of-Distribution Generalization Perspective". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 12731–12750. DOI: 10.18653/v1/2023.findings-acl.806. URL: https://aclanthology.org/2023.findings-acl.806.

[83] Zeyu Yun et al. "Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors". In: *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Ed. by Eneko Agirre, Marianna Apidianaki, and Ivan Vulić. Online: Association for Computational Linguistics, June 2021, pp. 1–10. DOI: 10.18653/v1/2021.deelio-1.1. URL: https://aclanthology.org/2021.deelio-1.1.

[84] Susan Zhang et al. "Opt: Open pre-trained transformer language models". In: *arXiv preprint arXiv:2205.01068* (2022).

[85] Han Zhao et al. "Fundamental limits and tradeoffs in invariant representation learning". In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 15356–15404.

[86] Ziqian Zhong et al. "The clock and the pizza: Two stories in mechanistic explanation of neural networks". In: *arXiv preprint arXiv:2306.17844* (2023).

[87] Andy Zou et al. "Representation engineering: A top-down approach to ai transparency". In: *arXiv preprint arXiv:2310.01405* (2023).

# A Additional Context

## A.1 Background and Related Work

**Linguistic Competence** There has been significant debate in linguistics and the philosophy of language regarding the precise definition and nature of competence [43, 50, 66, 47]. However, the formalization of competence provided by CALM is sufficiently general to incorporate most notions of competence, which may be flexibly specified by instantiating CALM in different ways. In this work, we focus on *lexicosemantic competence*, the ability to utilize knowledge of word meaning relationships in performing tasks such as lexical inference [46, 47].

**Relational Properties** Why is it not possible for linear information removal methods such as INLP [61] to remove relational properties between multiple input embeddings? Consider a binary relational property $Z$ denoting whether a relation $Z(i, j)$ holds between multiple embeddings $\mathbf{h}_i, \mathbf{h}_j \in \mathbb{R}^d$. In

INLP, we begin by training linear classifier $W \in \mathbb{R}^{2d}$ to predict $Z$ from concatenated embeddings $(\mathbf{h}_i; \mathbf{h}_j)$, where the goal is to train $W : (\mathbf{h}_i; \mathbf{h}_j) \mapsto Z$. We may decompose $W(\mathbf{h}_i; \mathbf{h}_j) = W_{[0:d]}\mathbf{h}_i + W_{[d+1:2d]}\mathbf{h}_j$, where $W_{[0:d]}$ is the first $d$ dimensions of $W$ and $W_{[d+1:2d]}$ are the second $d$ dimensions. In this case, there is no interaction between the two inputs $\mathbf{h}_i, \mathbf{h}_j$, meaning there is no way for $W$ to take into account any relationship $Z$ between them.

### A.2 Benefits of GBIs

The key advantage of gradient-based interventions (GBIs) as a causal probing methodology is that they may be applied to any differentiable probe. For example, if we are investigating the hypothesis that $M$'s representation of $Z$ is captured by a linear subspace of representations in a given layer (see [76]), then we may train a linear probe and various nonlinear probes on representations and observe whether GBIs against the linear probe have a comparable impact to those against the nonlinear probes. Alternatively, if we believe that a probe's architecture should mirror the architecture of the model it is probing (as argued by [57]), we may implement probes as such. Finally, where previous intervention methodologies for causal probing have focused on *nullifying* interventions that remove the representation of the target property $Z$ [61, 59, 60, 67, 5], GBIs allow one to perform targeted interventions that set LLMs' representations to counterfactual values $do(Z = z')$, effectively simulating the model's behavior under counterfactual inputs, which may be useful for predicting behaviors under various distribution shifts (see Appendix D.1). However, the benefits associated with GBIs do come with some important limitations, as discussed below in Appendix B.1.

## B Limitations

### B.1 Gradient-Based Interventions

For causal probing to operate successfully – as is required to reliably deploy CALM in practice – it is important that probes leverage the underlying model's representation of the target property to make predictions, rather than relying on spurious information. However, there is some evidence that probes often leverage such spurious information [4, 36, 14]. For instance, in followup work studying our GBI methodology alongside other causal probing methods, Canby et al. [14] find that each method they studied (including GBIs) shows a tradeoff between its ability to manipulate the targeted property (*completeness*) and the extent to which it also modifies the representation of other, non-targeted property (*selectivity*). Notably, they also found that the flexibility of GBIs allows for precise control of this tradeoff by modulating the magnitude of perturbations ($\epsilon$), an advantage that is not shared by most other causal probing methods.

Furthermore, while GBIs are applicable to a more general range of model representations than most prior intervention methods (see Section 3), this generality comes with a lack of constraints on probes ($g_Z$); and as a result, GBIs cannot provide the strong theoretical constraints on collateral damage as can methods like, e.g., INLP [61], which provably preserves distances between embeddings as well as possible while completely removing the linear representation of the target property (which also generally leads to higher selectivity in practice [14]). To minimize collateral damage to representations, the magnitude of perturbations should be modulated via constraints on gradient attacks against $g_Z$ (see Section 4) and experimentally validated to control the damage done to representations (see Appendix C.3); and in the ideal case, should be calibrated to achieve the desired tradeoff between *selectivity* and *completeness* (a novel procedure introduced in [14], a followup work building on GBIs as initially developed in this work). Alternatively, in cases where the structure of representations is believed to satisfy strong assumptions (e.g., being restricted to a linear subspace; [76]) or strong upper bounds on collateral damage are required, CALM interventions can be implemented with methods like INLP rather than GBIs.[5]

### B.2 Simple Experimental Setting

As noted in Section 4, our primary goal in our experiments is to validate CALM by testing it in a simplified experimental setting consisting of comparatively small, well-studied models and tasks. As

[5] It may also be possible to control for collateral damage by developing GBI strategies that offer more principled protection against damage to non-targeted properties, such as adding a loss term to penalize damage to non-targeted probes or leveraging interval bound propagation [31] to place intervened embeddings inside the adversarial polytope for non-targeted properties. We leave such possibilities to future work.

such, we need models that are *just complex enough* for CALM to be applicable (i.e., neural language models that are capable of performing the tasks we consider at a nontrivial level of performance), making BERT and RoBERTa ideal candidates; and in future work plan to scale CALM to more complex contexts covering larger, more powerful models as they perform more difficult tasks (see Appendix E). This is a common setting in the context of substantial recent interpretability work: first, a theoretical framework is developed for interpreting an internal representation or mechanism and initially tested in the context of "toy" models or tasks [25, 51, 86, 27], and subsequent work scales these frameworks to the context of larger models "in the wild" [78, 18, 80]. Analogously, all of our major contributions (the CALM framework, competence metric, and GBI causal probing method) are directly scalable to much larger, more recent LLMs (e.g., [84, 7, 74, 73, 32], etc.).

### B.3 Task Independence

In our experiments, we modeled the 14 LAMA ConceptNet tasks as representing fully independent properties, which is not necessarily true – e.g., knowing that a tree is made of bark or contains leaves tells us something about whether it is a type of plant. However, in the aggregate (with impacts summed across 14 widely-varying lexical relation types in computing the final competence score for each task; see Appendix D.2), it may nonetheless be appropriate to treat the relations which are not causal with respect to a given task as collectively capturing spurious lexical associations.

## C Experimental Details

### C.1 Tasks

The full set of LAMA ConceptNet tasks is as follows: IsA, HasA, PartOf, HasSubEvent, MadeOf, HasPrerequisite, MotivatedByGoal, AtLocation, CausesDesire, NotDesires, CapableOf, UsedFor, ReceivesAction, and HasProperty. We split each task dataset into train, validation, and test sets with a random $80\%/10\%/10\%$ split. Train and validation instances are fed to each model to produce embeddings used to train $g_Z$ and select hyperparameters, respectively; and test instances are used to measure LLMs' competence with respect to each task by observing how predictions change under various interventions. In all experiments, we restrict each model $M$'s output space for each task $\mathcal{T}$ to the subset of vocabulary $V_M$ that occurs as a ground-truth answer $y^*$ for at least one instance $(\mathbf{x}, y^*) \sim \mathcal{T}$ in the respective task dataset. This lowers the probability of false negatives in evaluation (e.g., penalizing the model for predicting $\hat{y} =$ "mammal" for "a dog is a type of $y$" instead of $y^* =$ "animal").

### C.2 Probes

We use BERT's final layer $L$ to encode $h_i^l$ embeddings for each such example, where $i$ is the index of the `[MASK]` token or target word in the input prompt $x_i$. To encode the `[MASK]` token, we issue BERT masked prompts (as discussed above) to extract $h_{\texttt{[MASK]}}$, then repeat with the `[MASK]` token filled-in with the target word to encode it as $h_+$ (e.g., "device" in "A laser is a device which creates coherent light."), and concatenate matching embeddings $h = (h_{\texttt{[MASK]}}; h_+)$ to produce positive $(y = 1)$ training instances. We also construct one negative $(y = 0)$ instance, $h = (h_{\texttt{[MASK]}}; h_-)$, for each $h_{\texttt{[MASK]}}$ by sampling an incorrect target word $x_i$ corresponding to an answer to a random prompt from the same task, feeding it into the cloze prompt in the place of the correct answer, and obtaining BERT's contextualized final-layer embedding of this token $(h_-)$. Finally, we train $g_Z$ on the set of all such $(h, y)$.

We implement $g_Z$ as a multi-layer perceptron with 2 hidden layers, each with a width of 768 (which is one half the concatenated input dimension of 1536), using ReLU activations and dropout with $p = 0.1$, training it for 32 epochs using Binary Cross Entropy with Logits Loss[6] and the Adam optimizer, saving the model from the epoch with the highest validation-set accuracy for use in all experiments.

For all competence results reported in Section 5, we run the same experiment 10 times – each with a different random initialization of $g_Z$ and shuffled training data – and report each figure as the average among all 10 runs.

[6] `https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html`

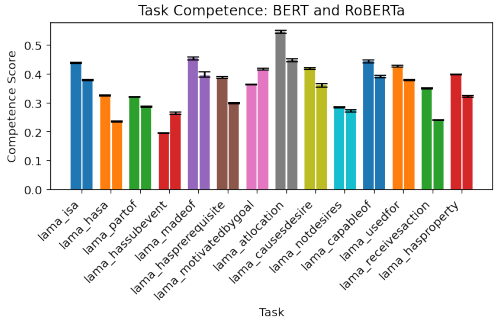


Figure 4: Competence of BERT (left bars) and RoBERTa (right bars) for all tasks, using PGD with $\epsilon = 0.1$. Y-values are the average competence score and error bars are the maximum and minimum competence score, as measured over 10 experimental iterations (each with a different randomly-initialized probe $g_Z$).

### C.3 Interventions

For embedding $h$, target (counterfactual) class $y'$, probe $g_Z$, loss function $\mathcal{L}$, and $L_\infty$-bound $\epsilon \in \{0.01, 0.03, 0.1, 0.3\}$[7], each intervention (gradient attack) $g_z$ may be used to produce perturbed representations $h' = g_z(h, y', f_{\text{cls}}, \mathcal{L}, \epsilon)$ where $||h - h'||_\infty \leq \epsilon$. In particular, given $h = (h_{\texttt{[MASK]}}; h_{\pm}) \in \mathbb{R}^{2d}$, let $h'_{\texttt{[MASK]}}$ be the first $d$ dimensions of $h'$ (which also satisfies the $L_\infty$-bound with respect to $h_{\texttt{[MASK]}}$, $||h_{\texttt{[MASK]}} - h'_{\texttt{[MASK]}}||_\infty \leq \epsilon$). To measure BERT's use of internal representations of $Z$ on each prompt task, we evaluate its performance when perturbed $h'_{\texttt{[MASK]}}$ is used to compute masked-word predictions, compared to unperturbed $h_{\texttt{[MASK]}}$.

Our intent in intervening only on the final-layer mask embedding $h_{\texttt{[MASK]}}$ in our experiments is that, in the final layer of a masked language model such as BERT or RoBERTa, the only embedding which is used to compute masked-word probabilities is that of the `[MASK]` token. Thus, any representation of the property that is *used* by the model in its final layer must be a part of its representation of the `[MASK]` token, preventing "recoverability" phenomena such as those observed by Elazar et al. [23].

**FGSM** FGSM [30] takes one gradient step of magnitude $\epsilon$ in the direction that minimizes the loss of a classifier (here, the probe $f_{\text{cls}}$) with respect to target class $y'$. We implement FGSM interventions as

$$h' = h + \epsilon \cdot \text{sgn}(\nabla_h \mathcal{L}(f_{\text{cls}}, x, y'))$$

where $\mathcal{L}$ is the same loss function used to train $f_{\text{cls}}$ (here, binary cross entropy).

**PGD** PGD [12, 44] iteratively minimizes the loss of a classifier (here, the probe $f_{\text{cls}}$) with respect to target class $y'$ by performing gradient descent within a $L_\infty$ ball of radius $\epsilon$. We implement PGD interventions as $h' = h^T$, where

$$h^{t+1} = \Pi_{N(h)}\big(h^t + \alpha \cdot \text{sgn}(\nabla_h \mathcal{L}(f_{\text{cls}}, x, y))\big)$$

for iterations $t = 0, 1, ..., T$, projection operator $\Pi$, $L_\infty$-neighborhood $N(h) = \{h' : ||h - h'|| \leq \epsilon\}$, and $\mathcal{L}$ is the same loss function used to train $f_{\text{cls}}$ (here, binary cross entropy). This method also introduces two hyperparameters: the number of PGD iterations $T$ and step size $\alpha$. We use hyperparameter grid search over $\alpha \in \{0.001, 0.003, 0.01, 0.03\}$ and $T \in \{20, 40, 60, 80, 100\}$, finding that setting $\alpha = \frac{\epsilon}{10}$ and $T = 40$ produces the most consistent impact on $g_Z$ accuracy across all tasks; so we use these values for the results visualized in Figure 4.

### C.4 Compute Budget

BERT-base-uncased has 110 million parameters, and RoBERTa-base has 125M parameters. As our goal is to study the internal representation and use of linguistic properties in existing pre-trained models, and we are not directly concerned with training or fine-tuning such models, we use these models only for inference (including encoding text inputs, using embeddings to train probes, and

[7] All reported results use $\epsilon = 0.1$, as greater $\epsilon$ resulted in unacceptably high "collateral damage" across target tasks (e.g., even random perturbations of magnitude $\epsilon = 0.3$ do considerable damage), and lesser values meant that predictions changed on target tasks consisted of only a few test instances.

feeding intervened embeddings back into the language models). The only models we trained were probes $g_Z$, which each had 1.77M parameters.

Each experimental iteration (including encoding text inputs, training probes on all 14 tasks, and performing all GBIs) for either BERT or RoBERTa took less than one hour on a single NVIDIA GeForce GTX 1080 GPU, meaning that running all 10 iterations across both language models took less than 20 hours on a single GPU. Each iteration, probe, and GBI can easily be parallelized across GPUs: in our case, running all iterations across both models took less than 3 hours total across 8 GTX 1080 GPUs.

# D Competence Metric

## D.1 Comparison With IIA

As noted in Section 2, the $\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T})$ metric defined in Equation (2) is an adaptation of the Interchange Intervention Accuracy (IIA) metric [29, 27], which evaluates the faithfulness of a causal abstraction like $\mathcal{G}_\mathcal{T}$ as a (potential) explanation of the behavior of a "black box" system like $M$. In our case, this is equivalent to evaluating the competence of $M$ on task $\mathcal{T}$, provided that $\mathcal{G}_\mathcal{T}$ is the appropriate SCM for $\mathcal{T}$, as an LLM is competent only to the extent that its behavior is determined by a causally invariant representation of the task.[8] IIA requires performing *interchange interventions* $\mathrm{do}_{II}(\mathbf{z}_j)$, where the part of $M$'s intermediate representation of input $\mathbf{x}_i$ hypothesized to encode latent variables $\mathbf{Z}$ (taking the values $\mathbf{z}_i$ when provided input $\mathbf{x}_i$) is replaced with that of $\mathbf{x}_j$ (which, in principle, means that the modified representation encodes the values $\mathbf{z}_j$ instead of $\mathbf{z}_i$), at which point these interchange interventions are used to compute predictions $M(\mathbf{x}_i|\,\mathrm{do}_{II}(\mathbf{z}_j))$, and the output is compared with $\mathcal{G}_\mathcal{T}(\mathbf{x}_i|\,\mathrm{do}(\mathbf{z}_j))$ to measure how faithfully $\mathcal{G}_\mathcal{T}$ predicts $M$'s behavior under these interventions. Thus, given access to high-quality interchange interventions over $M$, IIA measures the extent to which $\mathcal{G}_\mathcal{T}$ correctly models $M$'s behavior under counterfactuals, and thus its faithfulness as a causal abstraction of $M$.

To adapt IIA to the context of causal probing and define $\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T})$, we replace instance-level interchange interventions $\mathrm{do}_{II}$ with concept-level interventions $\mathrm{do}_Z$ for any given property $Z$. That is, instead of swapping $M$'s representation of variables $\mathbf{Z} = Z_1, ..., Z_k$ given input $\mathbf{x}_i$ with that of $\mathbf{x}_j$, we intervene on the representation of each property $Z \in \mathbf{Z}$ at the level of arbitrary values $\mathbf{z} : Z_1 = z_1, ..., Z_k = z_k$ that need not correspond to previously observed $\mathbf{x}$, allowing us to simulate the behavior of $M$ under previously-unseen distribution shifts (i.e., settings $\mathbf{z}$ representing previously-unseen combinations of values) and in doing so predict $M$'s consistency with a given causal model $\mathcal{G}_\mathcal{T}$ under these new conditions. As one of our primary motivations in studying LLM competence is to provide a framework useful for predicting and explaining behavior under distribution shifts, $\mathcal{C}_\mathcal{T}$ is more appropriate than IIA in this setting. However, this also introduces greater room for error: where interchange interventions only requires modifying representations to match the values taken by another input – as counterfactual representations can be obtained simply by "plugging in" representations from a different input – computing $\mathcal{C}_\mathcal{T}$ instead requires one to perform open-ended interventions that may not correspond to any ground-truth input, in which case there may be no region of the embedding space that corresponds to the intended setting $\mathbf{z}$ [28, 1, 14].

## D.2 Experimental Competence Metric

To compute the expectation in Equation (2) for test set $\{\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i\}_{i=1}^n \sim \mathcal{T} \times \mathbf{Z}$, we sum the competence score over all samples $\mathbf{x}_i$ and perform one intervention $\mathrm{do}(Z_j = 0)$ corresponding to each concept $Z_j \in \mathbf{Z}$.[9] As our goal is to measure the extent to which $M$'s behavior is attributable to an underlying representation of the causal property $Z_c$ or environmental property $Z \in \mathbf{Z}_e$, our experimental model defines $\mathcal{G}_\mathcal{T}$'s predictions with reference to $M$'s original predictions $M(\mathbf{x}_i) = \hat{\mathbf{y}}_i$, according to the following principle: if $M$ is competent, then its prediction $M(\mathbf{x}_i) = \hat{\mathbf{y}}_i$ is wholly

[8] For many tasks, there is more than one valid $\mathcal{G}_\mathcal{T}$ (see, e.g., the "price tagging game" constructed by Wu et al. [80]). In such cases, $\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T})$ should be computed with respect to each valid $\mathcal{G}_\mathcal{T}$ and the highest result should be selected, as conforming to any such $\mathcal{G}_\mathcal{T}$ carries the same implications.

[9] Note that this intervention changes the prediction $\mathcal{G}_\mathcal{T}(\mathbf{x}_i) \neq \mathcal{G}_\mathcal{T}(\mathbf{x}_i|\,\mathrm{do}(Z_j = 0))$ if and only if $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{T}_j$ – i.e., where the corresponding $(\mathbf{z}_i)_j = 1$ – otherwise, $(\mathbf{z}_i)_j$ is already 0, so the intervention has no effect. Thus, as $\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T})$ measures $M$'s consistency with $\mathcal{G}_\mathcal{T}$, then to the extent that $M$ is competent, its prediction should change under all and only the same interventions as $\mathcal{G}_\mathcal{T}$.

attributable to its representation of causal property $Z_c$, so its predictions $M(\mathbf{x}_i|\,\mathrm{do}(Z_c)) = \hat{\mathbf{y}}_i{}'$ will not overlap with its original predictions $\hat{\mathbf{y}}_i$ (i.e., $\mathrm{overlap}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_i{}') = 0$); and conversely, a competent $M$ will make the *same* predictions $M(\mathbf{x}_i|\,\mathrm{do}(Z_j)) = \hat{\mathbf{y}}_i{}''$ for any $Z_j \in \mathbf{Z}_e$, because its prediction is not caused by its representation of these environmental properties (i.e., $\mathrm{overlap}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_i{}'') = 1$). Motivated by this reasoning, our experimental model defines $\mathcal{G}_\mathcal{T}(\mathbf{x}_i|\,\mathrm{do}(Z_j = 0)) = M(\mathbf{x}_i)$ for environmental $Z_j \in \mathbf{Z}_e$; and for causal property $Z_c$, defines $\mathcal{G}_\mathcal{T}(\mathbf{x}_i|\,\mathrm{do}(Z_c = 0)) = \{y' \in V_M : y' \notin M(\mathbf{x}_i)\}$ (i.e., the set of all tokens $y'$ in $M$'s vocabulary that were not in its original prediction $M(\mathbf{x}_i)$). Thus, under experimental model $E$, we approximate $\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T})$ by computing it as follows:

$$\hat{\mathcal{C}_\mathcal{T}}(M|\mathcal{G}_\mathcal{T}) = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{overlap}\left(M\big(\mathbf{x}_i|\,\mathrm{do}(Z_j = 0)\big), \mathcal{G}_\mathcal{T}\big(\mathbf{x}_i|\,\mathrm{do}(Z_j = 0)\big)\right) \tag{3}$$

Notably, our experimental model $E$ only accounts for the relationship between $M$'s intervened and non-intervened predictions, independently of ground truth labels – instead, what is being measured is $M$'s consistency under meaning-preserving interventions $\mathrm{do}(Z_{j'})$ and its mutability under meaning-altering interventions $\mathrm{do}(Z_j)$. However, as we find in Section 5, the resulting competence metric $\mathcal{C}_\mathcal{T}(M|\mathcal{G}_\mathcal{T})$ is nonetheless useful for predicting $M$'s accuracy.

# E Future Work

## E.1 Representation Learning

The CALM framework, competence measure, and GBI methodology developed in Sections 2 and 3 are sufficiently general to be directly applied to analyze arbitrary LLMs on any language modeling task whose causal structure is already well understood (or, for tasks where this is not the case, we may apply the causal graph discovery approach described in Appendix E.4), allowing us to study the impact of various model architectures, pre-training regimes, and fine-tuning strategies on the representations LLMs learn and use for arbitrary tasks of interest.

## E.2 Multitask Learning

Are high competence scores on task $\mathcal{T}$ correlated with an LLMs' robustness to meaning-preserving transformations (see, e.g., [24]) on tasks $\mathcal{T}'$ that share several causal properties $\mathbf{Z}_c$ with task $\mathcal{T}$. Through the lens of causally invariant prediction [55, 3, 13], this hypothesis is likely true (however, see [64] for appropriate caveats) – if so, this would make it possible to use clusters of related tasks to predict LLMs' robustness (and other behavioral patterns, such as brittleness in the face of distribution shifts introduced by spurious dependencies) between related tasks using CALM, given an appropriate experimental model. Furthermore, the ability to characterize tasks based on mutual (learned) dependency structures could be valuable in transfer learning applications such as guiding the selection of auxiliary tasks in multi-task learning [65] or predicting the impact of intermediate task fine-tuning on downstream target tasks [16].

## E.3 Task Dependencies

Another possible application of CALM concerns causal invariance under multi-task applications. Existing approaches in invariant representation learning generally require task-specific training [85], as the notion of invariance is inherently task-centric (i.e., the properties which are invariant predictors of output values vary by task, and different tasks may have opposite notions of which properties are causal versus environmental), so applying such approaches to train models to be causally invariant with respect to a specific downstream task $\mathcal{T}$ is expected to come at the cost of performance on other downstream tasks $\mathcal{T}'$. Therefore, considering the recent rise of open-ended, task-general LLMs [84, 7, 74, 73, 32], it is important to understand the relationship between different task dependencies learned when fine-tuning task-general models on specific downstream tasks to account for applications involving tasks with different (and perhaps contradictory) causal structures, such as CALM.

### E.4 Causal Competence Graph Discovery

A key feature of CALM is that, instead of simply measuring consistency with respect to a known, static task description $\mathcal{G}_{\mathcal{T}}$, the competence metric in Equation (2) can also be used to dynamically discover a competence graph $\mathcal{G}$ which most faithfully explains a model $M$'s behavior in a given task or context (see Section 2) by computing $\mathcal{C}(M|\mathcal{G})$ "in-the-loop" of existing causal graph discovery algorithms like IGSP [81]. Such algorithms could be used to suggest likely competence graphs based on interventional data collected by running CALM experiments, to recommend the experiments that would yield the most useful interventional data for the graph discovery algorithm, or to evaluate candidate graphs $\mathcal{G}$ according to $\mathcal{C}(M|\mathcal{G})$, terminating the graph discovery algorithm once a competence graph $\mathcal{G}$ that offers sufficiently faithful explanations of $M$'s behavior has been found (e.g., where $\mathcal{C}(M|\mathcal{G}) > \tau$ for some threshold $\tau$). In this case, it is still necessary to define the set of properties $\mathbf{Z}$ being probed and the scoring function $S$ used to compare the predictions of $M$ and $\mathcal{G}$; but no knowledge of the causal dependencies (or structural functions $F : \mathbf{pa}(Z_j) \mapsto Z_j$ mapping from causal parents $\mathbf{pa}(Z_j)$ to causal dependents $Z_j$; see [9]) is required.