Probing for the Usage of Grammatical Number

Karim Lasri[ə,β] **Tiago Pimentel**[ð] **Alessandro Lenci**[β] **Thierry Poibeau**[ə] **Ryan Cotterell**[ſ]

[ə]Lattice (École Normale Supérieure-PSL, CNRS, U. Sorbonne Nouvelle)
[β]University of Pisa   [ð]University of Cambridge   [ſ]ETH Zürich
karim.lasri@ens.psl.eu   tp472@cam.ac.uk
alessandro.lenci@unipi.it   thierry.poibeau@ens.psl.eu
ryan.cotterell@inf.ethz.ch

Abstract

A central quest of probing is to uncover *how* pre-trained models encode a linguistic property within their representations. An encoding, however, might be spurious—i.e., the model might not rely on it when making predictions. In this paper, we try to find an encoding that the model actually uses, introducing a usage-based probing setup. We first choose a behavioral task which cannot be solved without using the linguistic property. Then, we attempt to remove the property by intervening on the model's representations. We contend that, if an encoding is used by the model, its removal should harm the performance on the chosen behavioral task. As a case study, we focus on how BERT encodes grammatical number, and on how it uses this encoding to solve the number agreement task. Experimentally, we find that BERT relies on a linear encoding of grammatical number to produce the correct behavioral output. We also find that BERT uses a separate encoding of grammatical number for nouns and verbs. Finally, we identify in which layers information about grammatical number is transferred from a noun to its head verb.

1 Introduction

Pre-trained language models have enabled researchers to build models that achieve impressive performance on a wide array of natural language processing (NLP) tasks (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020). How these models encode and use the linguistic information necessary to perform these tasks, however, remains a mystery. Over recent years, a number of works have tried to demystify the inner workings of various pre-trained language models (Alain and Bengio, 2016; Adi et al., 2017; Elazar et al., 2021), but no comprehensive understanding of how the models work has emerged. Such analysis methods are typically termed *probing*, and are methodologically diverse. In our assessment, most research in probing can

be taxonomized into three distinct paradigms. In the first paradigm, **diagnostic probing**, researchers typically train a supervised classifier to predict a linguistic property from the models' representations. High accuracy is then interpreted as an indication that the representations encode information about the property (Alain and Bengio, 2016; Adi et al., 2017; Hupkes et al., 2018; Conneau et al., 2018). A second family of methods, **behavioral probing**, consists in observing a model's behavior directly, typically studying the model's predictions on hand-picked evaluation datasets (Linzen et al., 2016; Goldberg, 2019; Warstadt et al., 2020; Ettinger, 2020). Finally, **causal probing** methods rely on interventions to evaluate how specific components impact a model's predictions (Giulianelli et al., 2018; Vig et al., 2020b; Elazar et al., 2021).

In this paper, we will investigate how linguistic properties are encoded in a model's representations, where we use the term encoding to mean the subspace on which a model relies to extract—or decode—the information. While probing has been extensively used to investigate whether a linguistic property is encoded in a set of representations, it still cannot definitively answer whether a model actually uses a certain encoding. Diagnostic probes, for instance, may pick up on a **spurious encoding** of a linguistic property, i.e., an encoding that allows us to extract our target property from the representation, but which the model being probed may not actually use to make a prediction.

Combining the three paradigms above, we instead seek to find encodings that are actually used by a pre-trained model, which we term **functional encodings**. To that end, we take a usage-based perspective on probing. Under this perspective, a researcher first identifies a linguistic property to investigate (e.g., grammatical number), and selects a behavioral task which requires knowledge of this property (e.g., selecting a verb's inflection which agrees in number with its subject). The researcher

then performs a causal intervention with the goal of removing a specific encoding (of the linguistic property under consideration) from the model's representations. If the encoding is a functional encoding, i.e., an encoding that the model indeed uses to make a prediction, then the intervention should prevent the model from solving the task.[1] Finally, once a functional encoding is discovered, we can use it to track how the property's information flows through the model under investigation.

As a case study, we examine how BERT (Devlin et al., 2019) uses grammatical number to solve a number agreement task. In English, grammatical number is a binary morpho-syntactic property: A word is plural or singular. In turn, subject–verb number agreement is a behavioral task; it inspects whether a model can predict the correct verbal inflection given its subject's number. For a model to solve the task, it thus requires information about the grammatical number of the subject and the verb. Our goal is to find how the model encodes this information when using it to make predictions. In other words, we want to find the structure from which the model *decodes* number information when solving the task.

In our experiments, we make three findings. First, our experiments provide us with strong evidence that BERT relies on a linear functional encoding of grammatical number to solve the number agreement task. Second, we find that nouns and verbs do *not* have a shared functional encoding of number; in fact, BERT relies on disjoint sub-spaces to extract their information. Third, our usage-based perspective allows us to identify where number information (again, as used by our model to make predictions) is transferred from a noun to its head verb. Specifically, we find that this transfer occurs between BERT's 3rd and 8th layers, and that most of this information is passed indirectly through other tokens in the sentence.

2   Paradigms in Probing

A variety of approaches to probing have been proposed in the literature. In this paper, we taxonomize them into three paradigms: (i) diagnostic probing, (ii) behavioral probing, and (iii) causal probing.

Diagnostic Probing. Traditionally, probing papers focus on training supervised models on top of frozen pre-trained representations (Adi et al., 2017;

---

[1]See Ravfogel et al. (2021) for a similar pipeline.

Hall Maudslay et al., 2020). The general assumption behind the work is that, if a probe achieves high accuracy, then the property of interest is encoded in the representations. Many researchers have expressed a preference for linear classifiers in probing (Alain and Bengio, 2016; Ettinger et al., 2016; Hewitt and Manning, 2019), suggesting that a less complex classifier gives us more insight into the model. Others, however, called this criterion into question (Tenney et al., 2019a,b; Voita and Titov, 2020; Papadimitriou et al., 2021; Sinha et al., 2021; Pimentel et al., 2020a; Pimentel and Cotterell, 2021). Notably, Hewitt and Liang (2019) proposed that complex classifiers may learn to extract a property by themselves, and may thus not reflect any true pattern in the representations. Further, Pimentel et al. (2020b) showed that, under a weak assumption, contextual representations encode as much information as the original sentences. Ergo, it is not clear what we can conclude from diagnostic probing alone.

Behavioral Probing. Another probing paradigm analyzes the behavior of pre-trained models on carefully curated datasets. By avoiding the use of diagnostic probes, they do not fall prey to the criticism above—tasks are directly performed by the model, and thus must reflect the pre-trained models' acuity. One notable example is Linzen et al. (2016), who evaluate a language model's syntactic ability via a careful analysis of a number agreement task. By controlling the evaluation, Linzen et al. could disentangle the model's syntactic knowledge from a heuristic based on linear ordering. In a similar vein, a host of recent work makes use of carefully designed test sets to perform behavioral analysis (Ribeiro et al., 2020; Warstadt et al., 2020; Warstadt and Bowman, 2020; Lovering et al., 2021; Newman et al., 2021). While behavioral probing often yields useful insights, the paradigm typically treats the model itself as a blackbox, thus failing to explain how individual components of the model work.

Causal Probing. Finally, the third probing paradigm relies on causal interventions (Vig et al., 2020b; Tucker et al., 2021; Ravfogel et al., 2021). In short, the researcher performs causal interventions that modify parts of the network during a forward pass (e.g., a layer's hidden representations) to determine their function. For example, Vig et al. (2020a) fix a neuron's value while manipulating the model's input to evaluate this neuron's role

in mediating gender bias. Relatedly, Elazar et al. (2021) propose a method to erase a target property from a model's intermediate layers. They then analyze the effect of such interventions on a masked language model's outputs.

3 Probing for Usage

Under our usage-based perspective, our goal is to find a functional encoding—i.e., an encoding that the model actually uses when making predictions. We achieve this by relying on a combination of the paradigms discussed in §2. To this end, we first need a behavioral task that requires the model to use information about the target property. We then perform a causal intervention to try to remove this property's encoding. We explain both these components in more detail now.

Behavioral Task. We first require a behavioral task which can *only* be solved with information about the target property. The choice of task and target property are thus co-dependent. Further, we require our model to perform well on this task. On one hand, if the model cannot achieve high performance on the behavioral task, we cannot be sure the model encodes the target property, e.g., grammatical number, at all. On the other hand, if the model can perform the task, it must make use of the property.

Causal Intervention. Our goal in this work is to answer a causal question: Can we identify a property's functional encoding? We thus require a way to intervene in the model's representations. If a model relies on an encoding to make predictions, removing it should harm the model's performance on the behavioral task. It follows that, by measuring the impact of our interventions on the model's behavioral output, we can assess whether our model was indeed decoding information from our targeted encoding.

4 Grammatical Number and its Usage

The empirical portion of this paper focuses on a study of how BERT encodes grammatical number in English. We choose number as our object of study because it is a well understood morpho-syntactic property in English. Thus, we are able to formulate simple hypotheses about how BERT passes information about number when performing number agreement. We use Linzen et al.'s (2016) number agreement task as our behavioral task.

4.1 The Number Agreement Task

In English, a verb and its subject agree in grammatical number (Corbett, 2006). Consider, for instance, the sentences:

(1)   a.   The boy goes to the movies.
      b.   *The boy go to the movies.
      c.   The boy that holds the *keys* goes to the movies.
      d.   *The boy that holds the *keys* go to the movies.

In the above sentences, both (1-a) and (1-c) are grammatical, but (1-b) and (1-d) are not; this is because, in the latter two sentences, the highlighted verb does not agree in number with its subject.

The subject–verb number agreement task evaluates a model's ability to predict the correct verbal inflection, measuring its preference for the grammatical sentence. In this task, the probed model is typically asked to predict the verb's number given its context. The model is then considered successful if it assigns a larger probability to the correct verb inflection:

context: The boy that holds the keys [MASK] to the movies.

success: $p(goes \mid context) > p(go \mid context)$

failure: $p(go \mid context) > p(goes \mid context)$

In this setting, the subject is usually called the **cue** of the agreement, and the verb is called the **target**.

Examples similar to the above are often designed to study the impact of distractors (the word *keys* in (1-c) and (1-d)) on the model's ability to predict the correct verb form. Success on the task is usually taken as evidence that a model is able to track syntactic dependencies. In this regard, this phenomenon has been studied in a variety of settings to investigate the syntactic abilities of neural language models (Gulordava et al., 2018; Marvin and Linzen, 2018; Newman et al., 2021; Lasri et al., 2022). In this work, however, we do not use this task to make claims about the syntactic abilities of the model, as done by Linzen et al. (2016). Instead, we employ it as a case study to investigate how BERT encodes and uses grammatical number.

4.2 Related Work on Grammatical Number

A number of studies have investigated how grammatical number is encoded in neural language models.[2] Most of this work, however, focuses on diagnostic probes (Klafka and Ettinger, 2020; Torroba Hennigen et al., 2020). These studies are thus

---

[2]We focus on grammatical number here. There is, however, also a vast literature investigating how BERT encodes number from a numeracy point of view (Wallace et al., 2019; Geva et al., 2020; Spithourakis and Riedel, 2018).

agnostic about whether the probed models actually use the encodings of number they discover. Some authors, however, do consider the relationship between how the model encodes grammatical number and its predictions. Notably, Giulianelli et al. (2018) use a diagnostic probe to investigate how an LSTM encodes number in a subject–verb number agreement setting. Other approaches (Lakretz et al., 2019; Finlayson et al., 2021) have been proposed to apply interventions at the neuron level and track their effect on number agreement. In this work, we look for functional encodings of grammatical number—encodings which are in fact used by our probed model when solving the task.

5 From Encoding to Usage

We discuss how to identify and remove an encoding from a set of contextual representations using diagnostic probing. Our use of diagnostic probing is thus twofold. For a model to rely on an encoding of our property when making predictions, the property must be encoded in its representations. We thus first use diagnostic probing to measure the amount of information a representation contains about the target linguistic property. In this sense, diagnostic probing serves to sanity-check our experiments—if we cannot extract information from the representations, there is no point in going forward with our analysis. Second, we make use of diagnostic probing in the context of amnesic probing (Elazar et al., 2021), which allows us to determine whether this probe finds a functional or a spurious encoding of the target property.

5.1 Estimating Extractable Information

In this section, we discuss how to estimate the amount of extractable number information in our probed model's representations. This is the probing perspective taken by Pimentel et al. (2020b) and Hewitt et al. (2021) in their diagnostic probing analyses. The crux of our analysis relies on the fact that the encoding extracted by diagnostic probes is not necessarily the functional encoding used by our probed model. Nevertheless, for a model to use a property in its predictions, this property should at least be extractable, which is true due to the data processing inequality. In other words, extractability is a necessary, but not sufficient, condition for a property to be used by the model.

We quantify the amount of extractable information in a set of representations in terms of a di-

agnostic probe's $\mathcal{V}$-information (Xu et al., 2020), where the $\mathcal{V}$-information is a direct measure of the amount of extractable information in a random variable. We compute the $\mathcal{V}$-information as:[3]

$$\mathrm{I}_{\mathcal{V}}(R \to N) = \mathrm{H}_{\mathcal{V}}(N) - \mathrm{H}_{\mathcal{V}}(N \mid R) \quad (1)$$

where R and N are, respectively, a representation-valued and a number-valued random variables, $\mathcal{V}$ is a variational family determined by our diagnostic probe, and the $\mathcal{V}$-entropies are defined as:

$$\mathrm{H}_{\mathcal{V}}(N) = \inf_{q \in \mathcal{V}} \mathbb{E}_{n \sim p(n)} \log \frac{1}{q(n)} \quad (2)$$

$$\mathrm{H}_{\mathcal{V}}(N \mid R) = \inf_{q \in \mathcal{V}} \mathbb{E}_{n, \boldsymbol{r} \sim p(n, \boldsymbol{r})} \log \frac{1}{q(n \mid \boldsymbol{r})} \quad (3)$$

where p is the true joint distribution over number and representations.[4]

Further, if we denote our analyzed model's (i.e., BERT's) hidden representations as:

$$\boldsymbol{r}_{t,l} = \mathrm{model}(\texttt{sentence})_{t,l} \quad (4)$$

we define our linear diagnostic probe as:

$$p_{\boldsymbol{\theta}}(N = \mathrm{SG} \mid R = \boldsymbol{r}_{t,l}) = \sigma(\boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{r}_{t,l} + b) \quad (5)$$

where $\boldsymbol{r}_{t,l} \in \mathbb{R}^{768}$ is a column vector, t is a sentence position and l is a layer, N is the binary number label associated with the word at position t, σ is the sigmoid function, $\boldsymbol{\theta}$ is a real-valued column parameter vector and b is a bias term. In this case, we can define our variational family as $\mathcal{V} = \{p_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathbb{R}^{768}\}$.

5.2 Intervening on the Representations

We now discuss how we perform a causal intervention to prevent the analyzed model from using a given encoding. The goal is to damage the model and make it "forget" a property's information. This allows us to analyze whether that encoding actually influences the probed model's predictions—i.e., whether this encoding is indeed functional. To this end, we employ amnesic probing (Elazar et al., 2021).[5] In short, we first learn a linear diagnostic classifier, following eq. (5). We then compute

---

[3]See App. B for a detailed description of $\mathcal{V}$-information.

[4]Representations are model-specific. However, the true distributions over representations here refers to the distribution over representations that is induced by the true distribution over natural language sentences.

[5]In particular, this intervention consists in applying iterative null-space projection to the representations, originally proposed by Ravfogel et al. (2020). We note that Ravfogel et al. (2022a,b) recently proposed two new methods to remove information from a set of representations.

the projector onto the kernel (or null) space of this linear transform $\boldsymbol{\theta}$, shown below:

$$\mathbf{P}_{\text{null}} = \mathbf{I} - \frac{\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}}{||\boldsymbol{\theta}||_2^2} \tag{6}$$

By iterating this process, we store a set of parameter vectors $\boldsymbol{\theta}^{(k)}$ and their associated projectors $\mathbf{P}_{\text{null}}^{(k)}$ until we are unable to extract the property. The composition of these projectors makes it possible to remove all linearly extractable number information from the analyzed representations. We can then apply the resulting composition to the said representations to get a new set of vectors:

$$\boldsymbol{r}'_{t,l} = \mathbf{P}_{\text{null}}^{(k)} \cdots \mathbf{P}_{\text{null}}^{(2)} \mathbf{P}_{\text{null}}^{(1)} \boldsymbol{r}_{t,l} \tag{7}$$

After learning the projectors, we can measure how erasing a layer's encoding impacts: (i) the subsequent layers, and (ii) our model's performance on the number agreement task. Removing a functional encoding of grammatical number should cause a performance drop on the number agreement task. Further, looking at both (i) and (ii) allows us to make a connection between the amount of information we can extract from our probed model's layers and its behavior. We are thus able to determine whether the encodings revealed by our diagnostic probes are valid from a usage-based perspective— are they actually used by the probed model on a task that requires them?

6 Experimental Setup

Data. We perform our analysis on Linzen et al.'s (2016) number agreement dataset, which consists of sentences extracted from Wikipedia. In this dataset, each sentence has been labeled with the position of the cue and target, along with their grammatical number. We assume here that this dataset is representative of the number agreement task; this may not be true in general, however.

Model. In our experiments, we probe BERT (Devlin et al., 2019).[6] Specifically, BERT is a bidirectional transformer model with 12 layers, trained using a masked language modeling objective. As BERT has been shown to perform well on this dataset (Goldberg, 2019), we already know that our probed model passes our first requirement; BERT does use number information in its predictions.

---

[6] We focus on `bert-base-uncased`, as implemented in the transformers library (Wolf et al., 2020).

Distinguishing Nouns and Verbs. While number is a morpho-syntactic property common to nouns and verbs, we do not know *a priori* if BERT relies on a single subspace to encode number in their representations. Though it is possible for BERT to use the same encoding, it is equally plausible that each part of speech would get its own number encoding. This leads us to perform our analyses using independent sets of representations for nouns and verbs; as well as a mixed set which merges both of them. Further, verbs are masked when performing the number agreement task, so their representations differ from those of unmasked verbs. Ergo, we analyze both unmasked and masked tokens at the target verb's position—which for simplicity we call verbs and masked verbs, respectively. This leaves us with four probed categories: nouns, verbs, masked-verbs, and mixed.

7 Experiments and Results

In our experiments, we focus on answering two questions: (i) How is number information encoded in BERT's representations? and (ii) How is number information transferred from a noun to its head verb for the model to use it on the behavioral task? We answer question (i) under both extractability and usage-based perspectives. In §7.1, we present our sanity-check experiments that demonstrate that grammatical number is indeed linearly extractable from BERT's representations. In §7.2 and §7.3, we use our causal interventions: we identify BERT's *functional encodings* of number; and analyze whether these functional encodings are shared across parts of speech. Finally, in §7.4 and §7.5 we investigate question (ii), taking a closer look at the layers in which information is passed.

7.1 What do diagnostic probes say about number?

Fig. 2 presents diagnostic probing results in all four of our analyzed settings.[7] *A priori*, we expect that verbs' and nouns' representations should already contain a large amount of $\mathcal{V}$-information about their grammatical number at the type level. As expected, we see that the $\mathcal{V}$-information is near its maximum for both verbs and nouns in all layers; this means that nearly 100% of the uncertainty about grammatical number is eliminated given BERT's representations. Further, the mixed category results also reach a maximal $\mathcal{V}$-information,

---

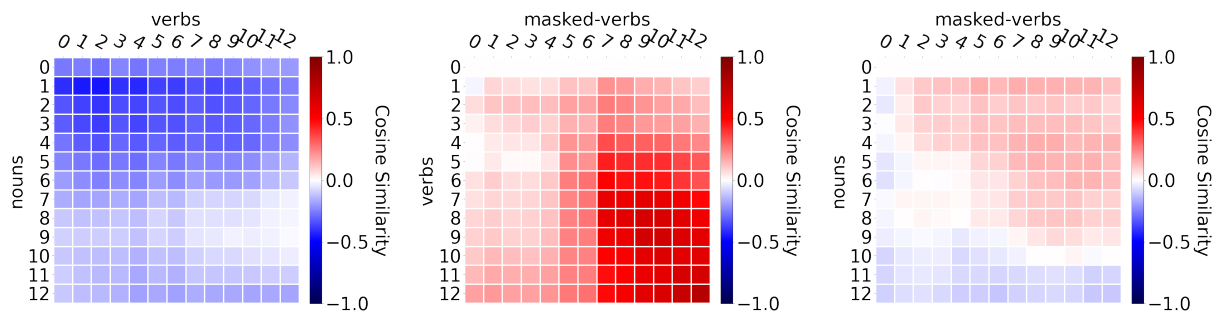[7] We further present accuracy results in App. A.

Figure 1: Cosine similarities between the learned parameter vectors of our diagnostic probes. The matrices display similarities between different layers, and across categories.
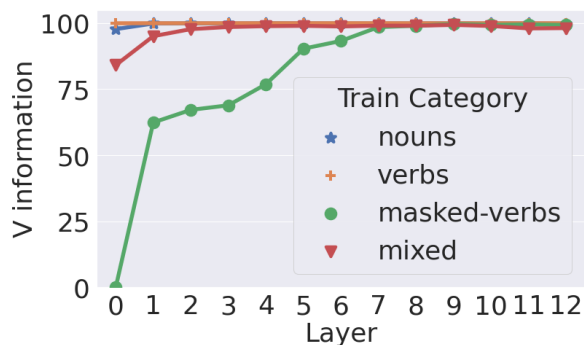

Figure 2: The amount of $\mathcal{V}$-information BERT representations hold about grammatical number, as estimated with linear diagnostic probes.

which indicates that it is possible to extract information linearly about both categories at the same time. On the other hand, the $\mathcal{V}$-information of masked verbs is 0 at the non-contextual layer and it progressively grows as we get to the upper layers.[8] As we go to BERT's deeper layers, the $\mathcal{V}$-information steadily rises, with nearly all of the original uncertainty eliminated in the mid layers. This suggests that masked verbs' representations acquire number information in the first 7 layers.

However, from these results alone we cannot confirm whether the encoding that nouns and verbs use for number is shared or disjoint. We thus inspect the encoding found by our diagnostic probes, evaluating the cosine similarity between their learned parameters $\boldsymbol{\theta}$ (ignoring the probes' bias terms b here). If there is a single shared encoding across categories, these cosine similarities should be high. If not, they should be roughly 0. Fig. 1 (left) shows that nouns and verbs might encode number along different directions. Specifically, noun represen-

tations on the first 6 layers seem to have a rather opposite encoding from verbs, while the later layers are mostly orthogonal. Further, while masked verbs and verbs do not seem to share an encoding in the first few layers, they are strongly aligned from layer 6 on (Fig. 1; center).

We now know that there are encodings from which we can extract number from nouns and verbs, and that these encodings are disjoint. However, we still do not know whether the encoding is spurious or functional.

7.2 Does the model use these encodings?

The patterns previously observed suggest there is a linear encoding, from which grammatical number can be extracted from BERT's representations. We, however, cannot determine whether these encodings are actually those used by the model to make predictions. We now answer this question taking our proposed usage-based perspective, studying the impact of linearly removing number information at both the cue and target positions.[9] We evaluate the model's change in behavior, as evaluated by its performance on the number agreement (NA) task.

Fig. 3a and Fig. 3c show the decrease in how much information is extractable at the target position after the interventions are applied. Fig. 3b and Fig. 3d show BERT's accuracy drops on the NA task (as measured at the output level). By comparing these results, we find a strong alignment between the information lost across layers and the damage caused to the performance on the task— irreversible information losses resulting from our intervention are mirrored by a performance decrease on the NA task. This alignment confirms that the model indeed uses the linear information erased by our probes. In other words, we have found the probed property's functional encoding.

---

[8] We note that, in Fig. 2, layer 0 corresponds to the non-contextual representations (i.e. the word embeddings before being summed to BERT's position embeddings). Non-contextual layers thus contain no information about the number of a masked verb, as the mask token contains no information about its replaced verb's number.

[9] The number of dimensions removed by our amnesic projectors in each layer and category is presented in Tab. 1.
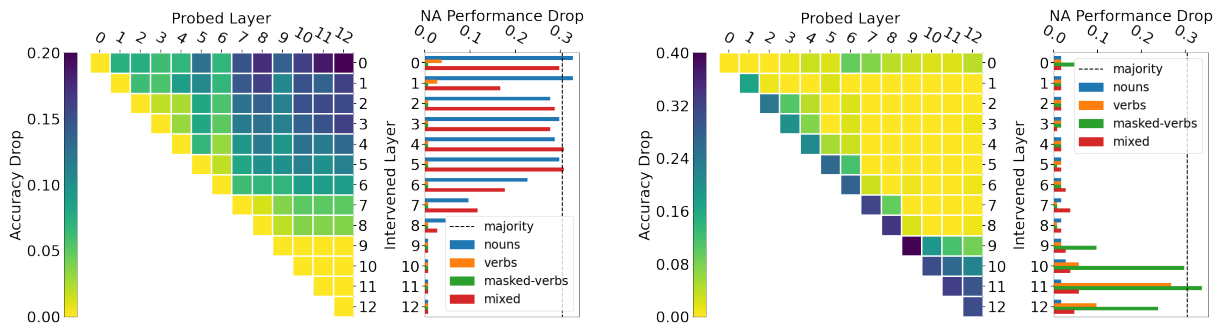
(a) Information loss (measured at the target) after erasing nouns' number information at the cue position.

(b) NA performance drop after erasing number at the cue position.

(c) Information loss (measured at the target) after erasing masked verbs' number at the target position.

(d) NA performance drop after erasing number at the target position.

Figure 3: Effect of our causal interventions on information recovery in subsequent layers (triangular matrices) and on the number agreement task (bar charts). Information loss is measured at the target position by a diagnostic probe; we display the probing accuracy drop compared to when no intervention was performed. The legend in the bar charts indicates what category the amnesic projectors have been trained on. Majority represents the difference in performance between BERT and a trivial baseline which always guesses the majority label.

7.3 Does BERT use the same encoding for verbs and nouns?

We now return to the question of whether nouns and verbs share a functional encoding of number, or whether BERT encodes number differently for them. To answer this question, we investigate the impact of removing a category's encoding from another category, e.g., applying an amnesic projector learned on verbs to a noun. In particular, we measure how these interventions decrease BERT's performance in our behavioral task. Figs. 3b and 3d presents these results.

We observe that each category's projector has a different impact on performance depending on whether it is applied to the cue or the target. Fig. 3b, for instance, shows that using the verb's, or masked verb's, projector to erase information at the cue's (i.e., the noun's) position does not hurt the model. It is similarly unimpactful (as shown in Fig. 3d) to use the noun's projectors to erase a target's (i.e., the masked verb's) number information. Further, the projector learned on the mixed set of representations does affect the cue, but has little effect on the target. Together, these results confirm that BERT relies on rather distinct encodings of number information for nouns and verbs.[10]

These experiments allow us to make stronger claims about BERT's encoding of number information. First, the fact that our interventions have a direct impact on BERT's behavioral output confirms that the encoding we erase actually bears number information *as used by the model when making predictions*. Second, the observation from Fig. 1—that number information could be encoded orthogonally for nouns and verbs—is confirmed from a usage-based perspective. Indeed, using amnesic probes trained on nouns has no impact when applied to masked verbs, and amnesic probes trained on verbs have no impact when applied to nouns. These fine-grained differences in encoding may affect larger-scale probing studies if one's goal is to understand the inner functioning of a model. Together, these results invite us to employ diagnostic probes more carefully, as the encoding found may not be actually used by the model.

7.4 Where does number erasure affect the model?

Once we have found which encoding the model uses, we can pinpoint at which layers the information is passed from the cue to the target. To that end, we observe how interventions applied in each layer affect performance. We know number information must be passed from the cue to the target's representations—otherwise the model cannot solve the task. Therefore, applying causal interventions to remove number information should harm the model's behavioral performance when applied

---

[10]A potential criticism of amnesic probing is that it may remove more information than necessary. Cross-testing our amnesic probes, however, results in little effect on BERT's behavior. It is thus likely that they are not overly harming our model. Further, we also run a control experiment proposed by Elazar et al., removing random directions at each layer (instead of the ones found by our amnesic probes). These results are displayed in the appendix in Tab. 1.

to: (i) the cue's representations before the transfer occurs; (ii) the target's representations after the transfer occurred.

Interestingly, we observe that target interventions are only harmful after the 9th layer; while noun interventions only hurt up to the 8th layer (again, shown in Fig. 3). This suggests that the cue passes its number information in the first 8 layers, and that the target stops acquiring number information in the last three layers. While we see a clear stop in the transfer of information after layer 8, Fig. 3a shows that the previous layers' contribution decreases slowly up to that layer. We thus conclude that information is passed in the layers before layer 8; however, we concede that our analysis alone makes it difficult to pinpoint exactly which layers.

7.5 Where does attention pruning affect number transfer?

Finally, in our last experiments, we complement our analysis by performing attention removal to investigate how and where information is transmitted from the cue to the target position. This causal intervention first serves the purpose of identifying the layers where information is transmitted. Further, we wish to understand whether information is passed directly, or through intermediary tokens. To this end, we look at the effect on NA performance after: (i) cutting direct attention from the target to the cue at specific layers, (ii) cutting attention from all tokens to the cue (as information could be first passed to intermediate tokens, which the target could attend to in subsequent layers).[11] Specifically, we perform these interventions in ranges of layers (from layer i up to j). We report number agreement accuracy drops in Fig. 4.[12]

The diagonals from this figure show that removing attention from a single layer has basically no effect. Further, cutting attention from layers 6 to 10 suffices to observe near-maximal effect for direct attention. Interestingly, it is at those layers where we see a transition from it being more harmful to apply amnesic projectors to the cue or to the target (in §7.4). However, while those layers play a role in carrying number information to the target position, the drop is relatively modest when cutting only direct attention ($\approx 10\%$). Cutting attention from all tokens to the cue, in turn, has a significant effect
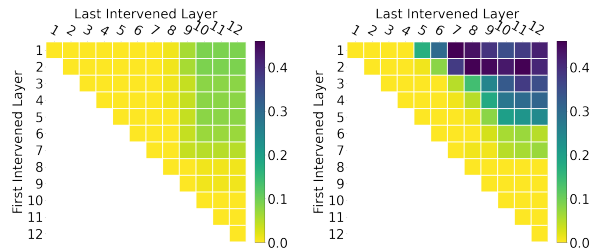
(a) Removing attention from the target to the cue only

(b) Removing attention from all tokens to the cue

Figure 4: Number agreement task performance drops after performing attention removal. The attention cut is performed on a range of layers. Rows and columns, respectively, represent the first and last intervened layer.

on performance (up to $\approx 40\%$), and is maximal for layers from 2 to 8. This first suggests that, while other clues in the sentence could indicate the target verb's number (such as a noun's determiner), the noun itself is the core source of number information. Further, this shows the target can get information from intermediate tokens, instead of number being passed exclusively through direct attention.[13]

8 Conclusion

Our analysis of grammatical number allows us to track how a simple morpho-syntactic property, grammatical number, is encoded across BERT's layers and where it is transferred between them before being used on the model's predictions. Using carefully chosen causal interventions, we demonstrate that forgetting number information impacts both: (i) BERT's behavior and (ii) how much information is extractable from BERT's inner layers. Further, the effects of our interventions on these two, i.e., behavior and information extractability, line up satisfyingly, and reveal the encoding of number to be orthogonal for nouns and verbs. Finally, we are also able to identify the layers in which the transfer of information occurs, and find that the information is not passed directly but through intermediate tokens. Our ability to concretely evaluate our interventions' impact is due to our focus on grammatical number and the number agreement task which directly aligns probed information and behavioral performance.

Ethics Statement

The authors foresee no ethical concerns with the work presented in this paper.

---

[11] Klafka and Ettinger (2020), for instance, showed that number information of a given token was distributed to neighboring tokens in the upper layers.

[12] We detail these interventions in App. C.

[13] See App. E for further experiments.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Greville G. Corbett. 2006. *Agreement*. Cambridge University Press, Cambridge.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *CoRR*, abs/1901.05287.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A tale of a probe and a parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7389–7395, Online. Association for Computational Linguistics.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. 2021. Conditional probing: Measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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), pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61(1):907–926.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Karim Lasri, Alessandro Lenci, and Thierry Poibeau. 2022. Does BERT really agree ? Fine-grained analysis of lexical dependence on a syntactic task. In *Findings of the Association for Computational Linguistics: ACL 2022*, Online. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. Predicting inductive biases of pre-trained models. In *International Conference on Learning Representations*.

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.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Virtual. Association for Computational Linguistics.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Tiago Pimentel and Ryan Cotterell. 2021. A Bayesian framework for information-theoretic probing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2869–2887, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020a. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3138–3153, Online. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020b. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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, 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. arXiv preprint arXiv:2201.12091.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022b. Adversarial concept erasure in kernel space. arXiv preprint arXiv:2201.12191.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online. Association for Computational Linguistics.

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

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for language models: Evaluating and improving their ability to predict numbers. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2104–2115, Melbourne, Australia. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In International Conference on Learning Representations.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 197–216, Online. Association for Computational Linguistics.

Mycal Tucker, Peng Qian, and Roger Levy. 2021. What if this modified that? Syntactic interventions with counterfactual embeddings. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 862–875, Online. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020a. Investigating gender bias in language models using causal mediation analysis. In Advances in Neural Information Processing Systems, volume 33, pages 12388–12401.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020b. Causal mediation analysis for interpreting neural NLP: The case of gender bias. CoRR, abs/2004.12265.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 183–196, Online. Association for Computational Linguistics.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? Probing numeracy in embeddings. 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), pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt and Samuel R. Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? In Proceedings of the Cognitive Science Society (CogSci), pages 1737–1743.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 217–235, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In International Conference on Learning Representations.

A   Diagnostic Probing Cross-Evaluation

In addition to comparing the angles of our diagnostic probes trained on different categories, we performed cross-evaluation of our trained diagnostic probes. In this setting, we trained probes on one category and tested them on the others. Fig. 5 presents our cross-evaluation results. The performance of probes evaluated in one category, but trained on another, again suggests that BERT encodes number differently across lexical categories. Interestingly, in the lower layer, the probe tested on nouns (top-left) guesses the wrong number systematically when trained on verbs, and vice-versa (top-right). This can be due to token ambiguity, as some singular nouns (e.g. "hit") are also plural verbs. This is further evidence that the encoding might be different for nouns and verbs, though this analysis still cannot tell us whether this is true from our usage-based perspective. Additionally, the mixed results (Fig. 5; bottom-right), show it is possible to linearly separate both nouns and verbs with a single linear classifier trained on both categories, reaching perfect performance on all other categories, including masked-verbs (bottom-left).
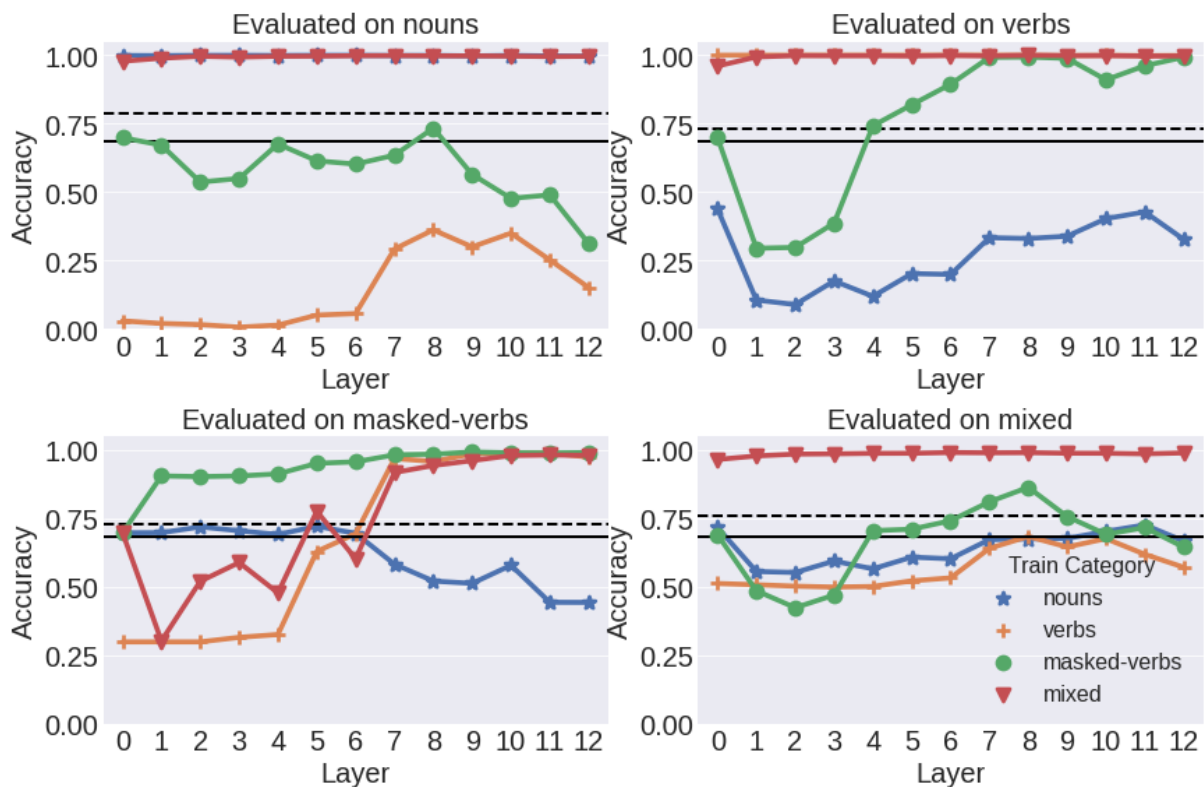

Figure 5: Probes cross-evaluation. Each plot corresponds to a test category, and colors correspond to the category used for training. Solid lines represent the percentage of majority-class (plural vs singular) tokens; dashed lines represent the percentage of majority-class tokens per lemma, averaged across lemmas.

B $\mathcal{V}$-information and mutual information

While a probing classifier's performance is often measured with accuracy metrics, in their analysis, Pimentel et al. (2020b) defined probing as extracting a mutual information. Formally, we write

$$I(R; N) = H(N) - H(N \mid R) \tag{8}$$

where R and N are, respectively, a representation-valued and a number-valued random variables. The mutual information, however, is a mostly theoretical value—hard to approximate in practice.

To compute this, we must first define a variational family $\mathcal{V}$ of interest; which we define as the set of linear transformations representable by eq. (5). We can then compute the $\mathcal{V}$-information as:

$$I_{\mathcal{V}}(R \to N) = H_{\mathcal{V}}(N) - H_{\mathcal{V}}(N \mid R) \tag{9}$$

where $\mathcal{V}$-entropies are defined as:

$$\mathrm{H}_\mathcal{V}(N) = \inf_{\boldsymbol{\theta} \in \mathcal{V}} \mathbb{E}_{n \sim p(n)} \log \frac{1}{p_{\boldsymbol{\theta}}(n)} \tag{10}$$

$$\mathrm{H}_\mathcal{V}(N \mid R) = \inf_{\boldsymbol{\theta} \in \mathcal{V}} \mathbb{E}_{n, \boldsymbol{r} \sim p(n, \boldsymbol{r})} \log \frac{1}{p_{\boldsymbol{\theta}}(n \mid \boldsymbol{r})} \tag{11}$$

This $\mathcal{V}$-information can vary in the range $[0; \mathrm{H}_\mathcal{V}(N)]$; thus a more interpretable value is the $\mathcal{V}$-**uncertainty**, which we define here as:

$$\mathrm{U}_\mathcal{V}(R \to N) = \frac{\mathrm{I}_\mathcal{V}(R \to N)}{\mathrm{H}_\mathcal{V}(N)} \tag{12}$$

We note that the $\mathcal{V}$-information lower-bounds the mutual information: $\mathrm{I}_\mathcal{V}(R \to N) \leq \mathrm{I}(R; N)$. It follows that, if we can extract some $\mathcal{V}$-information from a set of representations, they contain at least the same amount of information in Shannon's (1948) more classic sense.

C Attention Intervention

Formally, let $\mathbf{A}^{l,h} \in \mathbb{R}^{T \times T}$ be a model's attention weights for a given layer $1 \leq l \leq 12$, a head $1 \leq h \leq 12$, and a sentence with length T.[14] Further, we define a binary mask matrix $\mathbf{M}^l \in \{0, 1\}^{T \times T}$. We can now perform an intervention by masking the attention weights of all heads in a layer. Given a layer l:

$$\widehat{\mathbf{A}}^{l,h} = \mathbf{A}^{l,h} \circ \mathbf{M}^l, \quad 1 \leq h \leq 12 \tag{13}$$

where $\circ$ represents an elementwise product between two matrices. Now assume a given sentence with cue position p_c, and with target position p_t. In our intervention (i), matrix $\mathbf{M}^l$ is set to all 1's except for $\mathbf{M}^l_{p_t, p_c} = 0$; the target's attention to the cue is thus set to 0. In intervention (ii), we set $\mathbf{M}^l_{:, p_c} = 0$ and other positions to 1, which removes all attention to the cue.

D Removing random directions from representations

Removing directions from intermediate spaces could harm the model's normal functioning independently from removing our targeted property. We thus run a control experiment proposed by Elazar et al. (2021), removing random directions at each layer (as opposed to the specific directions found by our amnesic probes). This experiment allows us to verify that the observed information loss and decrease in performance do not only result from removing too many directions. To do so, we remove an equal number of random directions at each layer. The results are displayed in Tab. 1 and show that removing randomly chosen directions has little to no effect compared to our targeted causal interventions.

E The effect of linear distance

Here, we test whether the linear distance between the cue and the target influences the effect of attention removal. Fig. 6a shows that cutting attention from one layer has negligible effect over performance regardless of distance, which is in line with results from the diagonals of Fig. 4. When cutting attention from several subsequent layers (Fig. 6b), we observe that performance drop depends on the linear position, and decreases when the model is not faced with short-range agreement. This is not surprising as many of the attention maps attend to surrounding tokens (Kovaleva et al., 2019). Extensive analysis targeting individual attention heads (instead of cutting all attention from a given layer) is necessary to examine both their contribution to the model's successes, and their dependence on linear distance.

---

[14]Our analyzed model, BERT base, has 12 layers, and 12 attention heads in each layer.

Layer	0	1	2	3	4	5	6	7	8	9	10	11	12
Masked Verbs													
Number of Directions	1	13	15	26	30	17	21	44	24	22	22	26	33
Loss in Layers	0.0	0.33	0.3	0.34	0.34	0.34	0.37	0.38	0.42	0.39	0.41	0.41	0.41
Loss in Layers (Random)	0.08	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NA Performance Drop	0.04	0.01	0.01	0.01	0.01	0.0	0.01	0.0	0.0	0.09	0.29	0.33	0.23
NA Performance Drop (Random)	0.03	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.01
Nouns													
Number of Directions	17	51	33	70	22	37	48	52	64	39	22	39	26
Loss in Layers	0.49	0.37	0.39	0.38	0.37	0.38	0.43	0.4	0.43	0.4	0.37	0.41	0.4
Loss in Layers (Random)	0.0	0.0	0.0	0.02	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0
NA Performance Drop	0.32	0.32	0.27	0.29	0.28	0.29	0.22	0.09	0.04	0.0	0.0	0.0	0.0
NA Performance Drop (Random)	0.06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 1: Causal intervention results using both the default or random directions. For each category, we display the number of directions removed in each layer, the information loss resulting from amnesic interventions in each layer and the effect on the NA task. We also display the loss in layers and performance decrease on NA resulting from the removal of random directions as a control experiment.

(a) Cutting attention from the target to the cue only

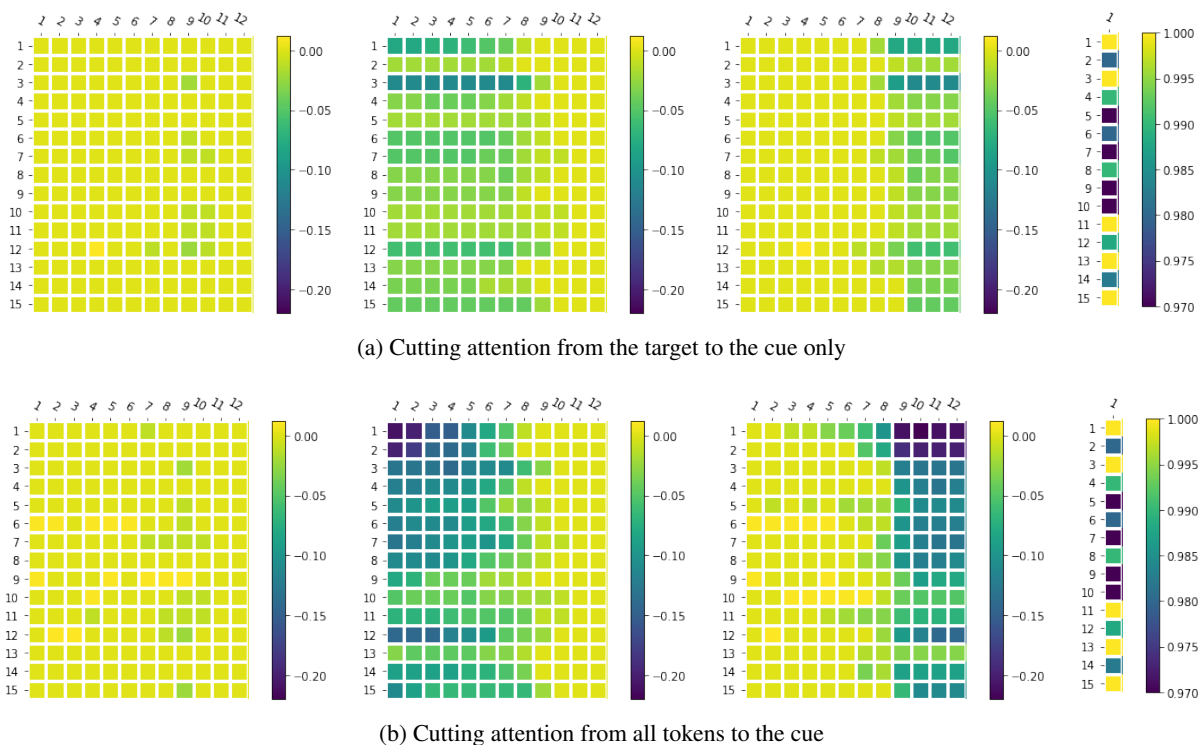
(b) Cutting attention from all tokens to the cue

Figure 6: Agreement task performance drops resulting from attention interventions, as a function of linear distance between the cue and the target. The rows represent distances (from 1 to 15) and columns represent the intervened layers. Three conditions are tested: cutting attention only at current layer (left), cutting attention starting from current layer up to the last one (middle) and from the first layer to current layer (right). The color map on the far right represent agreement scores without intervention for each linear distance.