The Functional Relevance of Probed Information: A Case Study

Michael Hanna[*], **Roberto Zamparelli**[†], and **David Mareček**[§]
[*]ILLC, University of Amsterdam, The Netherlands: m.w.hanna@uva.nl
[†]CIMeC, University of Trento, Italy: roberto.zamparelli@unitn.it
[§]MFF ÚFAL, Charles University, Czechia: marecek@ufal.mff.cuni.cz

Abstract

Recent studies have shown that transformer models like BERT rely on number information encoded in their representations of sentences' subjects and head verbs when performing subject-verb agreement. However, probing experiments suggest that subject number is also encoded in the representations of all words in such sentences. In this paper, we use causal interventions to show that BERT only uses the subject plurality information encoded in its representations of the subject and words that agree with it in number. We also demonstrate that current probing metrics are unable to determine which words' representations contain functionally relevant information. This both provides a revised view of subject-verb agreement in language models, and suggests potential pitfalls for current probe usage and evaluation.

1 Introduction

The phenomenon of subject-verb agreement has received significant attention from the NLP community. In English, this phenomenon is very simple: present tense verbs must agree in number with their subject noun, which is either singular or plural. In the present tense, verbs that agree with 3rd-person singular nouns receive one verb conjugation, generally ending in "-s"; in all other cases, the bare form of the verb is used.

The simplicity of this phenomenon, combined with the potential for long-distance subject and verb agreement across intervening adverbs or relative clauses (e.g. "The friend [that probably called my parents] is…") has made it the object of intense study in humans (Vigliocco et al. (1995) and Franck et al. (2010), *inter alia*). It also prompted early investigations on the ability of language models to capture it (left-to-right LSTMs in Linzen et al. (2016); Gulordava et al. (2018)). More recently, the popular pre-trained model BERT (Devlin et al., 2019) has been shown to be relatively proficient at subject-verb agreement (Goldberg, 2019), although these abilities depend somewhat on verb frequency and lexical patterns (Newman et al., 2021; Lasri et al., 2022a).

Other studies ask not how models behave (e.g. with respect to subject-verb agreement), but how models' representations support this behavior. Such studies often use probing, a technique in which an auxiliary classifier (probe) is trained to extract some property of words or sentences from models' internal representations thereof (Belinkov and Glass, 2019; Belinkov, 2022). If the probe can extract the property with high accuracy, one concludes that the model has encoded the property in the word's representation. Klafka and Ettinger (2020) discover that probes can extract the plurality of a sentence's subject from last-layer BERT representations of any word in the sentence.

However, probing has received criticism because the information discovered by probes is not always used by models (Ravichander et al., 2021). Causal interventions have been proposed as a means of connecting models' internal representations to their external behavior. Such techniques make targeted changes to models' representations, and observe how model behavior changes, in order to establish a causal connection between the two (Ravfogel et al., 2021; Geiger et al., 2021). For example, Ravfogel et al. (2020) remove gender information from models' noun representations, and observe that models behave as if they do not know the nouns' gender.

Recently, Lasri et al. (2022b) unified these lines of work by using causal probing interventions to investigate subject-verb agreement. They did so by first training probes to predict subject number information from representations of verbs and their subjects. Then, they removed subject number information from the representations of subjects and verbs. This caused BERT to make errors on a subject-verb agreement task, indicating that the probes discovered functionally relevant informa-

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 835–848
May 2-6, 2023 ©2023 Association for Computational Linguistics

tion about subject number in subject and verb representations. However, this leaves open the question of whether models use the subject number information in the representations of other words of the sentence, found by Klafka and Ettinger (2020).

In this paper, we focus on these questions: do large language models (LLMs) like BERT rely on subject number information stored outside of subject and verb representations when performing subject-verb agreement? Moreover, do current probing evaluation metrics allow us to determine which information is used and which is not? Our goal is thus twofold: first, to clarify how subject-verb agreement occurs in LLMs, and second, to determine if any current non-causal probing metrics can determine which probes have found functionally relevant information.

We achieve these goals as follows. First, we adopt the setup of Klafka and Ettinger (2020), examining simple sentences of a fixed structure and length. We use probing to demonstrate that subject number information is extractable from representations of any word in the sentence, at most layers. Next, we use causal interventions to show that although subject number information exists in the representations of all words of the sentence, it is not always used; rather, it seems that BERT uses information stored in words that agree with the subject in number. Finally, we evaluate our probes using modern probe evaluation metrics, to ascertain if any metric can determine whether subject number information found by a probe is used by the model or not. We find that these metrics are unable to do this.[1]

2 Probing for Plurality

To determine if the subject number information that is contained in all words of a sentence is used, we first verify that all words' representations contain said information. We adopt the setup of Klafka and Ettinger (2020), and thus investigate simple sentences with a fixed structure, using a synthetic English-language dataset. Their original dataset contains sentences of the form "The [subject] [verb-past] the [object]". However, we create our own dataset, which contains not only subject-verb agreement but also article-subject agreement. It consists of sentences of the form "[This / these] [adjective] [subject] [adverb] [verb-present] the [object]", e.g. "This short boy definitely admires the firefighter." The dataset contains 6000 sentences, with roughly equal numbers of singular and plural subjects. We split our dataset into train, valid, and test splits containing 4000/1000/1000 examples.

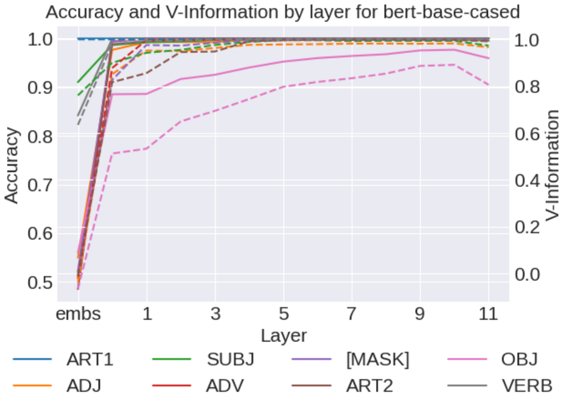

Figure 1: Test accuracy (left, solid) and $\mathcal{V}$-information (right, dashed) of probes of a given word and layer

Having generated a dataset, we must then define the model representations to be probed for subject number information. Klafka and Ettinger (2020) find said information in all words' final-layer BERT representations; in contrast, we examine the representations generated as the output of the entire transformer block at *each* layer, of a variety of LLMs. We also consider representations from the embedding, which have by definition no contextual information (except for positional information). Concretely, we analyze the base, large, and distilled variants of BERT and RoBERTa (Devlin et al., 2019; Liu et al., 2019; Sanh et al., 2019). However, we present only the BERT-base representations unless discussing another model, as results are similar across all models analyzed.

Having extracted model representations, we train probes on the subject number prediction task. Each probe is specific to a word in the sentence and to a model layer; a given probe might predict the number of a sentence's subject given the 5th-layer representation of the sentence's verb. Each probe is a linear layer with a sigmoid activation. We use HuggingFace (Wolf et al., 2020) implementations of these models; for precise model names, and dataset and training details, see Appendices A and B.

For each probe, we record 3 metrics, averaged across the dataset: (i) test accuracy; (ii) $\mathcal{V}$-information (Xu et al., 2020), and (iii) codelength, as measured by online minimum description length (MDL) probing (Voita and Titov, 2020). We record

[1] The code for all experiments is available at `https://github.com/hannamw/probed-information`.

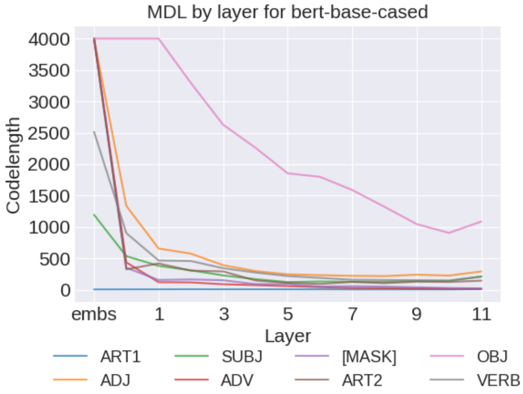

Figure 2: Probe codelength for a given word and layer

the latter two metrics to see if recently-proposed metrics can distinguish functionally relevant from irrelevant information in representations. We report metrics averaged over 10 probe training runs.

$\mathcal{V}$-information measures, given a family $\mathcal{V}$ of probes used to extract a label from a representation, how much those probes benefit from receiving a representation as input, as opposed to no input at all. In contrast, online MDL probing measures codelength by retraining probes on subsets of the dataset of increasing size. If probes assign low entropy to a dataset when training data is limited, the information being probed for is easily accessible from the input representations. A detailed description of these metrics can be found in Appendix C.

Results Figure 1 shows the accuracy and $\mathcal{V}$-information of our probes (higher is better). Note that for all figures, 'ART1' refers to the subject's article, and 'ART2' to the object's. We also probe the representation of masked verbs ('[MASK]'), discussed in the following section. As in Klafka and Ettinger (2020), accuracy is high for all words in later BERT layers. Unsurprisingly, accuracy is also high for subjects, their articles, and unmasked verbs using just BERT's embeddings. But starting in layer 0, when information is first able to move between positions, accuracy is high for most other words in the sentence as well. $\mathcal{V}$-information tracks accuracy; this is unsurprising, as it combines entropy, which tracks with accuracy, and a model-family-specific baseline, which is the same for all words / layers. This indicates that subject number information is present in the representations of all words of the sentence, and could be used for subject-verb agreement.

Figure 2 shows the codelength in bits for each probe, by layer; lower codelength means probes could learn to extract subject number information given less data. Broad trends are similar to those from accuracy and $\mathcal{V}$-information. Before layer 0, codelength given by the unmasked verb, subject, and subject article probes is notably lower than in others; this makes sense, because the very form of these words indicates the number of the subject. However, the plurality of the subject quickly becomes available, and the codelength for other words' probes drops. By the last layer, the disparity between words that agree in number with the subject, and those that do not, has mostly disappeared; the exception is the object probe, whose codelength remains high.

3 Causal Interventions

We now apply causal interventions to test if the information found by probes in the prior experiment is actually used by BERT[2]. In our case, this means we will alter BERT's internal representations with respect to subject number information, and observe BERT's performance on a subject-verb agreement task. In order to perform such interventions, we alter our dataset to accommodate this task: we mask out the verb of the sentence, and task BERT with predicting this masked word.

Then, we apply causal interventions, specifically reflection (Ravfogel et al., 2021) and interchange (Geiger et al., 2021) interventions. To define these, consider the case where BERT's input $\mathbf{s}$ is "This short boy definitely [MASK] the firefighter.". In the middle of computation, we extract $\mathbf{z}$, BERT's nth-layer representation of our word of interest; for example, the sentence's object ("firefighter"). $\mathbf{z}$ serves as input to a linear probe trained on the full training set, defined as $h(\mathbf{x}) = \sigma\left(\mathbf{W}^\top\mathbf{x} + \mathbf{b}\right)$. Let $h(\mathbf{z}) < 0.5$, indicating that the probe predicts the subject of this sentence is singular. In the **reflection** intervention, we reflect $\mathbf{z}$ over the decision boundary of h (a hyperplane defined by $(\mathbf{W}, \mathbf{b})$), creating $\mathbf{z}_r$. By definition, $h(\mathbf{z}_r) > 0.5$; the probe now predicts that the sentence's subject is plural. We replace the original representation $\mathbf{z}$ in the BERT with $\mathbf{z}_r$, and observe how BERT's output changes. In contrast, in the **interchange** intervention, we do not use the probe. Define $\mathbf{z}_i$ as the same representation as $\mathbf{z}$, but taken from the opposite-plurality context, e.g. the n-th layer representation of "firefighter" in "These short boys definitely [MASK]

[2] In this section, we discuss only BERT, but the procedure is identical for all models tested.

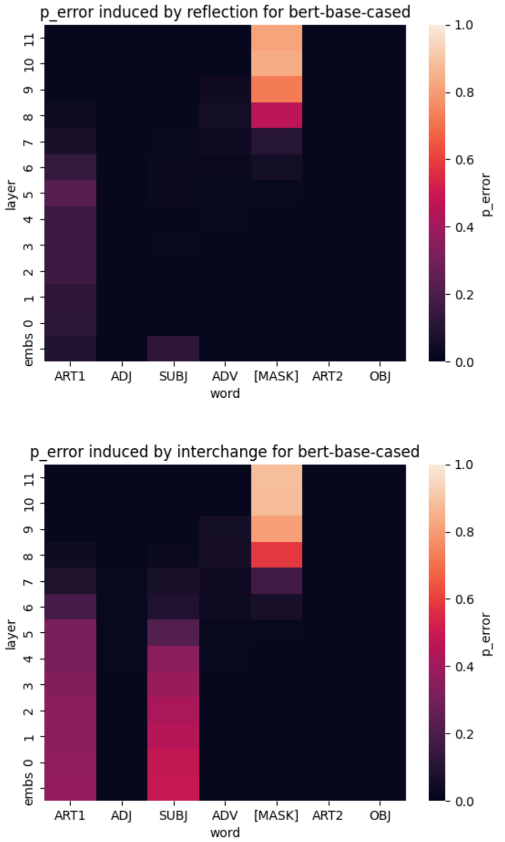

Figure 3: p_{error} induced by intervening on a representation, by word (x-axis), and layer (y-axis). Top: reflection intervention; bottom: interchange intervention

the firefighter." We then run BERT on $\mathbf{s}$, but replace $\mathbf{z}$ with $\mathbf{z}_i$, and observe how its behavior changes.

In both cases, we attempt to reverse the plurality of the subject number information in $\mathbf{z}$ by replacing it with $\mathbf{z}'$. Thus, if the intervention was effective (i.e. BERT was using said information in $\mathbf{z}$), BERT should predict verbs that do not agree in number with the original subject of the sentence.

We perform interventions on examples from the test split. Each intervention targets one word at a specific layer. To measure the effectiveness of a given intervention, we record p_{error}, the probability mass assigned by BERT to all 3rd-person present-tense verbs in its vocabulary that do not agree in number with the sentence's original subject. Higher p_{error} indicates a more successful intervention, i.e. that BERT relied on information in the targeted word representation to perform subject-verb agreement.

Results Figure 3 shows, for each word and layer of BERT, the error induced by performing either a reflection or interchange intervention on that word's representation at the given layer, averaged across the test split. For reflection interventions, we previously trained 10 sets of probes, and thus also average over the results obtained for each set. Note that with no interventions, $p_{error} \approx 0$; BERT assigns 80% of its probability mass to verbs that agree with the subject in number (BERT sometimes predicts conjunctions or other words that neither agree nor disagree with the sentence's subject).

The results show clearly that not all subject number information in a sentence is used. Information stored in representations of the adjective, object, and object's article, is not used at all, producing no effects when either intervention is performed. On the other hand, considering only the interchange intervention, information from the subject (in early layers) and masked verb (in later layers) is heavily used, as reported by Lasri et al. (2022b). Moreover, there is minor usage of number information ($p_{error} \approx 0.05$) in the adverb in layers 9-10, the transition layers between subject and verb. Although the error induced is very small, this could hint at instances where BERT's subject-verb agreement processing diverge from our expectations.

Our setup also reveals a new phenomenon: BERT uses the number information in the representation of the subject's article (a demonstrative, which agrees with the subject). The information in the subject article representation is used in the same layers as the subject information, with slightly weaker intervention effects. On this basis, we conclude that BERT uses subject number information in the subject and words that agree with it (not just the masked verb); however, further study, with additional agreement effects, would strengthen this.

These results are strengthened by the reflection intervention, which yields only slightly different results. It produces the same effects as the interchange at the masked verb and subject's article, albeit at a lower magnitude, indicating that the information found by probes is indeed the information the model is using to perform subject-verb agreement. Moreover, the strong effects at the masked verb position indicate that the model may be encoding subject number linearly. While linearity in BERT's representations of subject number was also found by (Lasri et al., 2022b), they used a much more complex approach, iteratively projecting representations into classifiers' null space; in contrast, our reflection approach is non-iterative, and suggests the model may be encoding this information in just one dimension.

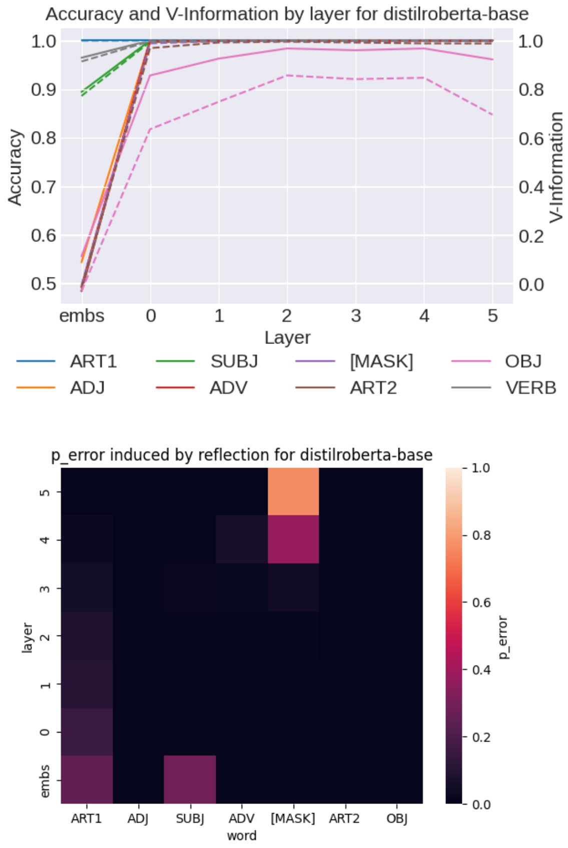

Figure 4: Top: Accuracy and $\mathcal{V}$-information for distilroberta-base. Bottom: effects of reflection interventions on distilroberta-base

It is notable that the reflection intervention produces almost no effect when performed on the subject. As this effect does appear under the interchange intervention, this result likely does not reflect a non-reliance on subject information there encoded; rather, it is due to specific properties of this intervention. First, the information encoded in the subject may well be used, but simply not encoded as the probe found it. Second, the subject is often split into multiple tokens; while we apply the reflection to each token, reflecting multiple tokens representations might not be as effective, as multiple tokens of the same word might not share the same plurality boundary as that of single tokens.

Although we primarily discuss results for BERT, results generalize to the RoBERTa models and their large / distilled variants: interchange interventions show use of information in representations of the subject, its article, and the verb, at expected locations (see Appendix D for plots and heatmaps). This could be surprising, especially in the case of the distilled models. If one views the unused subject number information present in model representations as extraneous or irrelevant encoded information, one might expect that these smaller distilled models do not have the space to encode irrelevant information. However, this is not the case, as seen in Figure 4; probe achieve high accuracies even when the their information is not used.

4 Discussion and Conclusions

In our prior two experiments, we showed that subject number information is extractable from the representation of any word in our simple sentences; however, it is only used if it comes from a representation of the subject or a word that agrees with it. Could any existing metrics have warned us of this, without requiring causal interventions?

Accuracy is insufficient to distinguish functional relevant information; probes extract functionally irrelevant subject number information with high accuracy. The same is true for $\mathcal{V}$-information, for the same reasons; high $\mathcal{V}$-information is necessary but not sufficient for a property to be used by models.

This leaves codelength, which measures ease of extracting the property from the data. It is an appealing hypothesis that models might encode functionally relevant information more accessibly in their representations; if this were the case, MDL probing could detect functionally relevant information. Indeed in early layers, it distinguishes between words that directly reflect subject number (the verb, subject, its article) and those that do not. However, when we mask the verb, such that subject number cannot be determined from the verb's form, but its representation's subject number information is still functionally relevant to subject-verb agreement, this distinction is lost. Thus, MDL probing seems unable to determine functional relevance.

So, we conclude the following. First, we show via probing that subject number information is present in representations of all words of our simple sentences. Then, using causal interventions, we show that only in the subject and words that agree in number with it, is said information functionally relevant. This indicates, as previously shown, that probing is not a reliable method for understanding how models function. Moreover, a way of distinguishing between functionally relevant and irrelevant information in model representations remains elusive. For now, causal interventions remain the most promising way to make this determination.

Limitations

This study is limited to a single phenomenon (number agreement), in a morphologically poor language (English). Our conclusions should be checked against different domains (person/gender agreement) and in morphologically complex languages. In particular, the study of sentences in which agreement extends beyond a demonstrative, the subject, and the verb, could help determine the extent to which models rely on agreement information in each word whose form expresses it. This study is also limited by its use of small, simple, synthetic data; expanding to real-world data, or data that follows a less rigid template, would strengthen our conclusions. From a modeling point of view, these results do generalize to various different masked language models, but this study does not investigate larger, more modern language models; autoregressive language models are also excluded. Thus, it is unclear to what extent these popular models exhibit the phenomenon studied.

Acknowledgments

MH received travel support from ELISE - the European Network of AI Excellence Centers - to present this work at the 2022 ELLIS Doctoral Symposium. This work has been supported by the grant 23-06912S of the Czech Science Foundation.

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

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.

Julie Franck, Gabriela Soare, Ulrich H. Frauenfelder, and Luigi Rizzi. 2010. Object interference in subject–verb agreement: The role of intermediate traces of movement. *Journal of memory and language*, 62(2):166–182. ID: unige:31707.

Atticus Geiger, Hanson Lu, Thomas F Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *ArXiv*, 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.

Sarah M Harmon. 2017. *Narrative Encoding for Computational Reasoning and Adaptation*. Ph.D. thesis, University of California, Santa Cruz.

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.

Karim Lasri, Alessandro Lenci, and Thierry Poibeau. 2022a. Does BERT really agree ? fine-grained analysis of lexical dependence on a syntactic task. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2309–2315, Dublin, Ireland. Association for Computational Linguistics.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022b. Probing for the usage of grammatical number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

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.

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*, pages 3710–3723, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Gabriella Vigliocco, Brian Butterworth, and Carlo Semenza. 1995. Constructing subject-verb agreement in speech: The role of semantic and morphological factors. *Journal of Memory and Language*, 34(2):186–215.

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.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

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 Dataset Details

Our dataset is a tightly-controlled synthetic dataset created using the BLiMP sentence generator (Warstadt et al., 2020), which we altered and updated to fit our use-case. A benefit of using this specific generator is that it forms (relatively) coherent sentences, such that e.g. all subjects and objects are valid (if not especially plausible) arguments of their verb. Moreover, it also allows us to specify additional constraints, and sample an arbitrary number of sentences fulfilling them.

Each example in the dataset consists of a sentence of the form "[This / these / that / those] [adjective] [subject] [adverb] [verb-present] the [object]"; we also automatically generate the same sentence, but with an opposite-number subject (and corresponding verb / article) for use with interchange interventions. At intervention time, we append (but do not probe / investigate) the word "nowadays". This encourages BERT to output a present-tense verb for the masked token; otherwise, BERT often predicts past-tense verbs or conjunctions. Subjects have distinct singular and plural forms, where the plural form ends in "-s".

Each of the word types contained in brackets is always one word long (so there are no multiword subjects, or phrasal verbs). If a word is split into multiple tokens, we handle it in the following way. During the probe training phase, we consider each token as a separate training example (so the number of training examples for each probe might differ, and be slightly greater than the number of sentences in the dataset). Then, during the intervention phase, we simply apply the intervention to every token composing the word. For the interchange intervention, both the original word and its opposite-number counterpart must have the same number of tokens. Thus, we discard from our dataset any examples where the two have different token lengths. In order to determine whether the model's prediction is a present-tense verb, and whether it was conjugated in the singular or plural, we use the `nodebox_linguistics_extended` package (Harmon, 2017).

B Experimental Details

Probes are scikit-learn (Pedregosa et al., 2011) `LogisticRegression` models, trained with L2 normalization to predict the number (singular / plural) of the subject. For some experiments, we use the weights from these models as the weights for identical PyTorch (Paszke et al., 2019) models, consisting of linear layer with bias and sigmoid activation. The `bert-base-cased` model has approximately 110 million parameters (Devlin et al., 2019). Each probe has 1538 parameters, for a total of 140000 parameters summed over all probes. All experiments were performed using an NVIDIA A100 GPU, and take no more than 24 GPU hours to run in total. We study `bert-base-cased`, `bert-large-cased`, `distilbert-base-cased`, `roberta-base`, `roberta-large`, and `distilroberta-base`.

C Metric Details

We consider two metrics in addition to standard accuracy. The first, predictive $\mathcal{V}$-information (Xu et al., 2020), measures the degree to which information in representations is made available to a chosen model family $\mathcal{V}$. We study linear probes f, where $f(x)$ is the probability that the subject of the sentence from which the representation x comes, is plural. For the purposes of $\mathcal{V}$-information, however, it is useful to define $f'(y|x) = (1-y)(1-f(x)) + yf(x)$, and $\mathcal{V}$ as the family of such functions.

$\mathcal{V}$-information relies on conditional $\mathcal{V}$-entropy:

$$H_{\mathcal{V}}(Y|X) = \inf_{f' \in \mathcal{V}} \mathbb{E}_{x,y \sim \mathcal{D}} - \log f'(y|x)$$

where $\mathcal{D}$ is our dataset of representation, number label pairs. We can define baseline $\mathcal{V}$-entropy as

$$H_{\mathcal{V}}(Y|\emptyset) = \inf_{f' \in \mathcal{V}} \mathbb{E}_{y \sim Y\mathcal{D}} - \log f'(y|\emptyset)$$

where $\emptyset$ represents the absence of representational information, e.g. a constant vector. Now we can define predictive $\mathcal{V}$-information as

$$I(\emptyset \to X) = H_{\mathcal{V}}(Y|\emptyset) - H_{\mathcal{V}}(Y|X)$$

i.e. the predictive advantage given to models of the family $\mathcal{V}$' by representational information. Put simply, this metric asks how much better our probes can extract subject number given representational information, as opposed to if they had none.

The conditional $\mathcal{V}$-information can be estimated via probing; it is the entropy achieved by a probe trained on our dataset. We compute the baseline as the maximum likelihood estimate of the label distribution, i.e. $f'(y|\emptyset) = c(y)/(c(0) + c(1))$, where $y \in \{0, 1\}$ is a label, and $c(y)$ is its count.

We also use minimum description length (MDL) probing, described by Voita and Titov (2020). Their codelength metric rewards probes that are both high-performing and simple. We measure it using their online MDL method, which measures the cost of encoding a dataset's labels with probes trained on limited data. In this setup, we have a dataset of data X and binary labels Y; we begin by partitioning this dataset into groups $(x_{1:t_1}, y_{1:t_1}); (x_{t_1+1:t_2}, y_{t_1+1:t_2}); \ldots; (x_{t_{S-1}+1:t_S}, y_{t_{S-1}+1:t_S})$ of increasing size. We encode the first group of labels $y_{1:t_1}$ at full cost, t_1 bits. For each timestep $i = 1, \ldots, S-1$, we train a probe p_{θ_i} on $\{(x_j, y_j)\}_{j=1}^{i}$, and encode the next set of labels using the updated model, for a cost of $-\log p_{\theta_i}(y_{t_i+1:t_{i+1}}|x_{t_i+1:t_{i+1}})$. The online codelength is, in bits

$$L = t_1 + \sum_{i=1}^{S-1} - \log p_{\theta_i}(y_{t_i+1:t_{i+1}}|x_{t_i+1:t_{i+1}}).$$

In this setup, probes that quickly learn how to extract subject plurality from BERT representations will have a shorter codelength. So, comparing codelength across representations from different words will tell us the words that are easier or more difficult to extract plurality information from.

We follow this procedure in our analysis: we split our data into partitions; then, we repeatedly train probes on increasing portions of our dataset, until we have trained on all data. Probes are trained using scikit-learn's (Pedregosa et al., 2011) `LogisticRegression`, with L2 normalization. After each round of training, we compute the cost of encoding the next partition. We sum over all partitions, and repeat this process 10 times, reporting average codelength. The partition sizes we use are identical to those in Voita and Titov (2020), i.e. 0.1%, 0.2%, 0.4%, 0.8%, 1.6%, 3.2%, 6.25%, 12.5%, 25%, 50%, and 100% of the dataset (cumulatively). In some cases (e.g. when training on embeddings that contain no subject plurality information), MDL is very high; in this case we cap it at 4000 (the cost of transmitting the labels uncompressed, i.e. 4000 examples, at 1 bit / example).

D Results for All Models

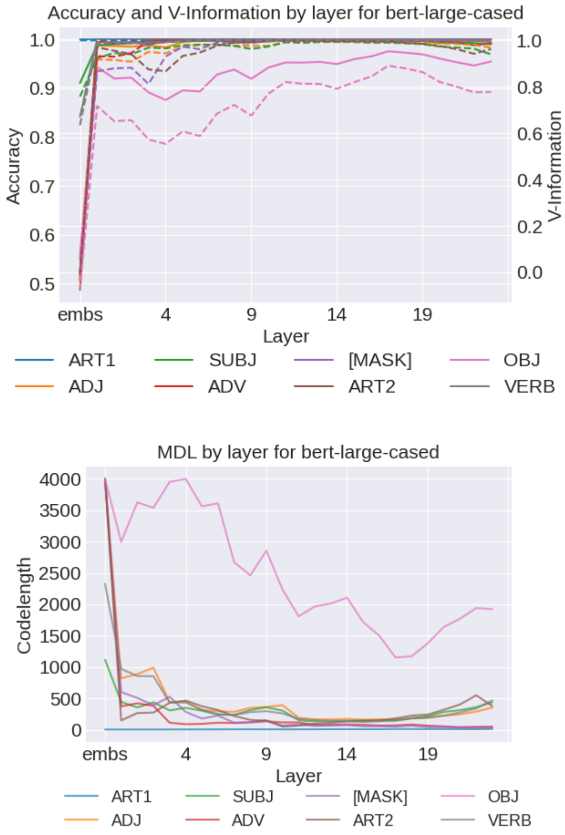

Figure 5: Top: Accuracy and V-information; bottom: MDL codelength

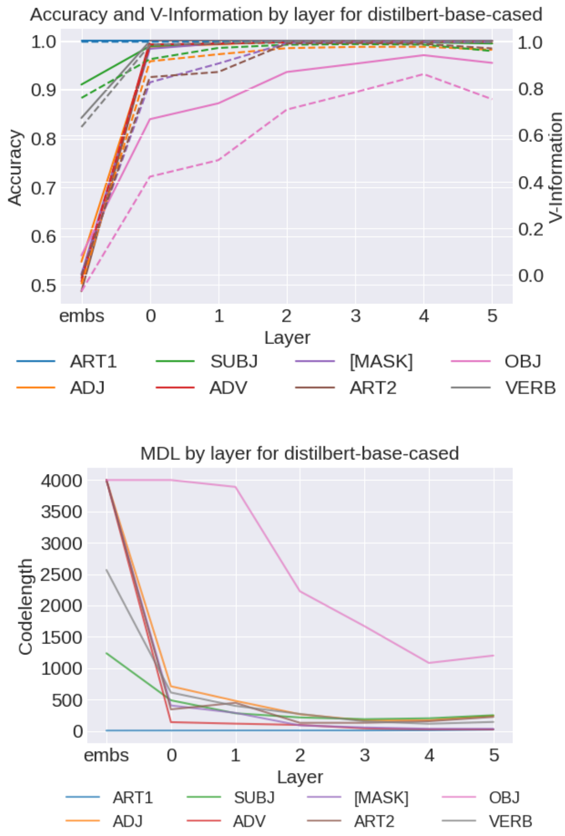

Figure 6: Top: Accuracy and V-information; bottom: MDL codelength

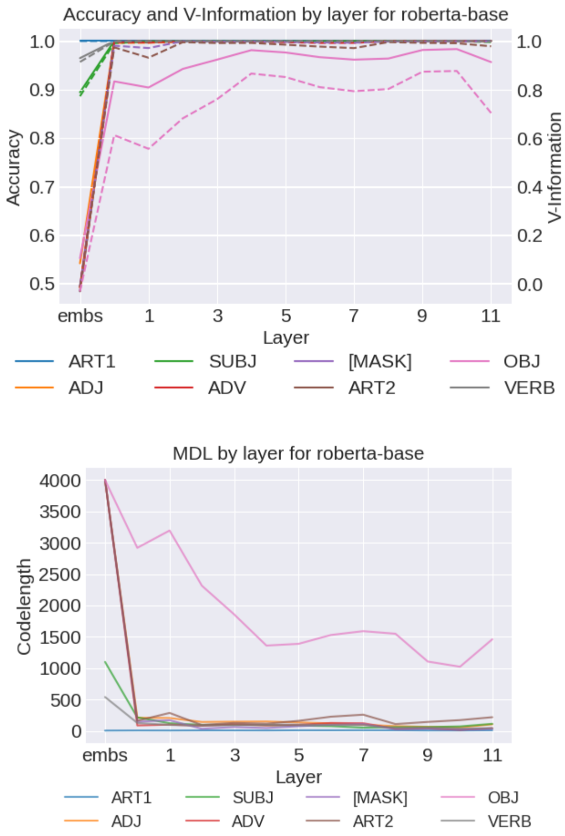

Figure 7: Top: Accuracy and V-information; bottom: MDL codelength

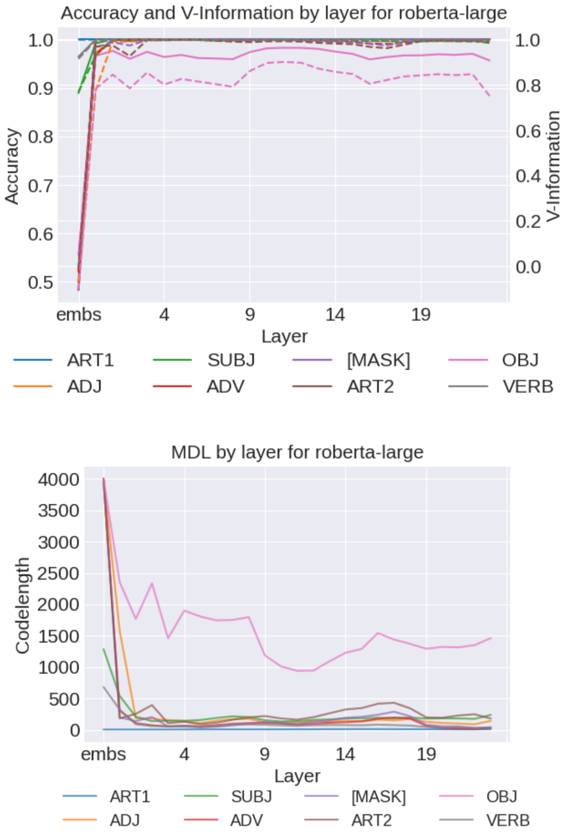

Figure 8: Top: Accuracy and V-information; bottom: MDL codelength

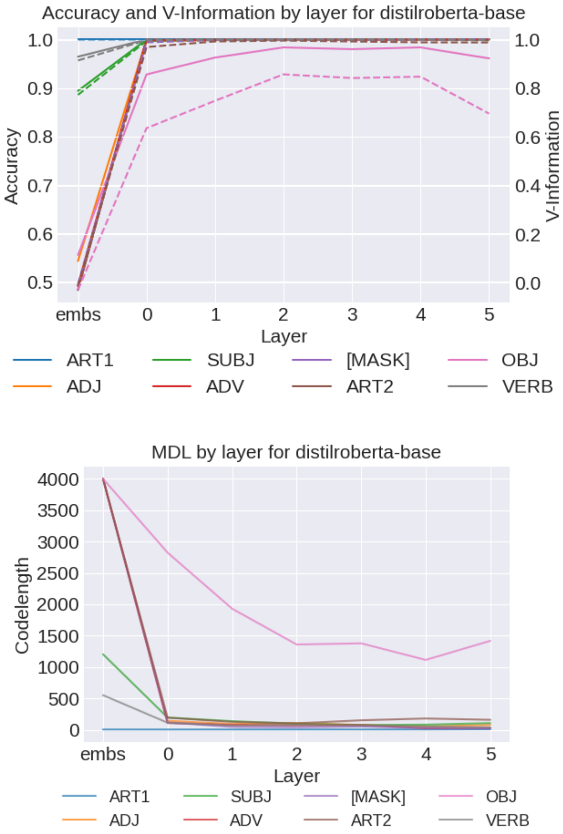

Figure 9: Top: Accuracy and V-information; bottom: MDL codelength

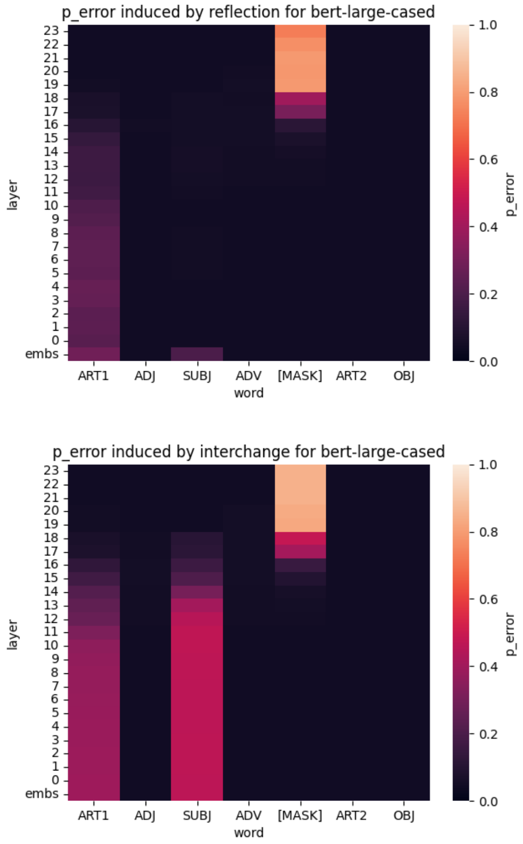

Figure 10: p_{error} induced by intervening on a representation, by word (x-axis), and layer (y-axis). Top: reflection intervention; bottom: interchange intervention

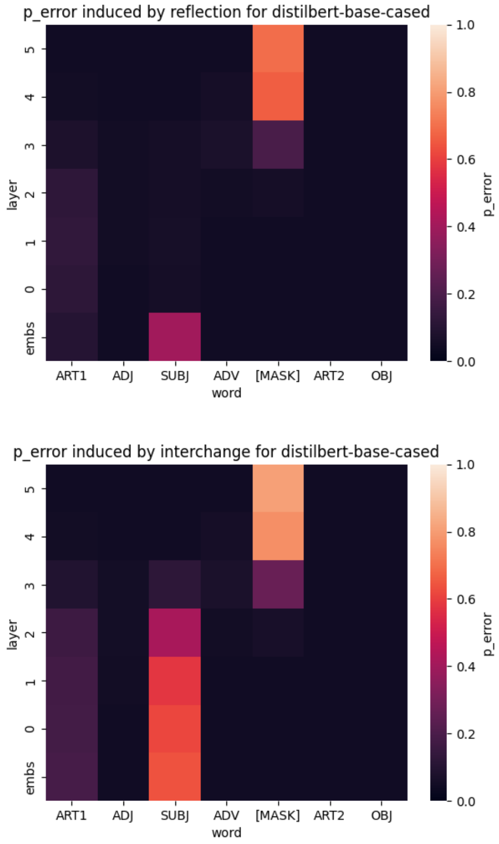

Figure 11: p_{error} induced by intervening on a representation, by word (x-axis), and layer (y-axis). Top: reflection intervention; bottom: interchange intervention

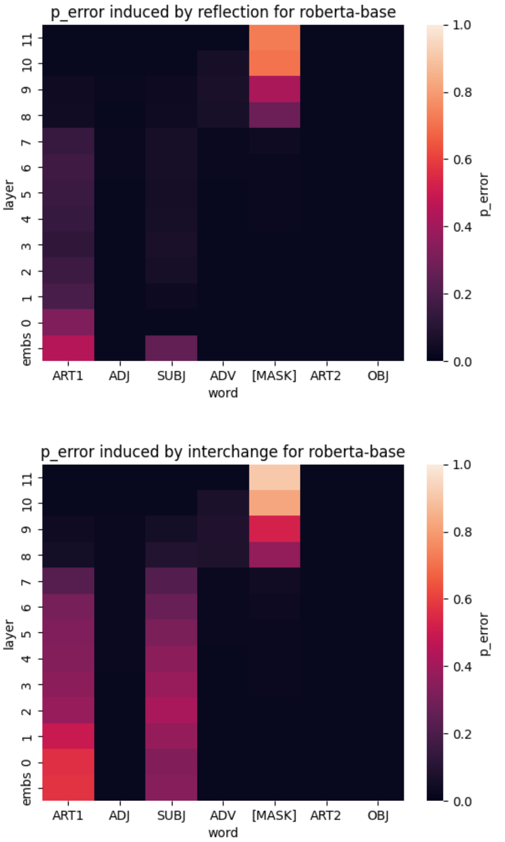

Figure 12: p_{error} induced by intervening on a representation, by word (x-axis), and layer (y-axis). Top: reflection intervention; bottom: interchange intervention

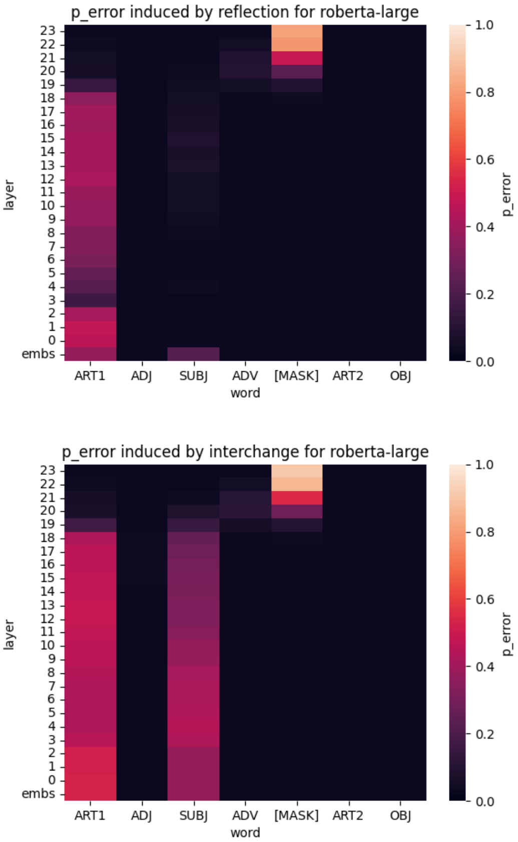

Figure 13: p_{error} induced by intervening on a representation, by word (x-axis), and layer (y-axis). Top: reflection intervention; bottom: interchange intervention

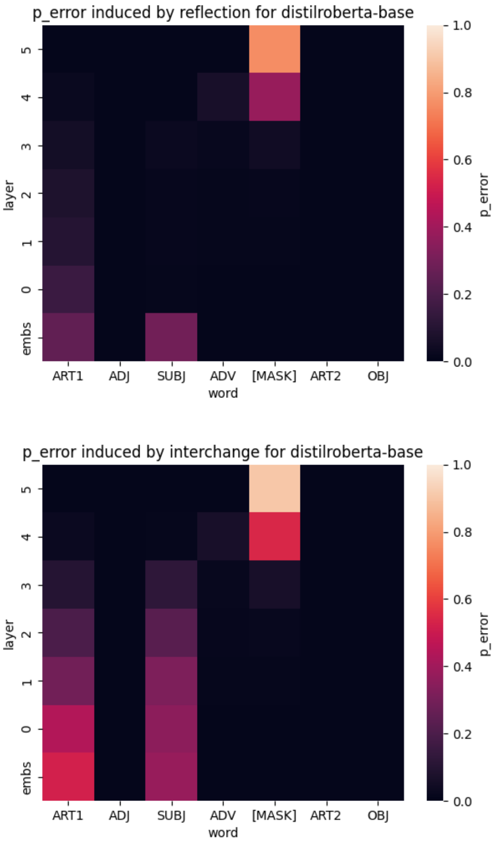

Figure 14: p_{error} induced by intervening on a representation, by word (x-axis), and layer (y-axis). Top: reflection intervention; bottom: interchange intervention