Improving Syntactic Probing Correctness and Robustness with Control Tasks

Weicheng Ma[1], **Brian Wang**[2], **Hefan Zhang**[2], **Lili Wang**[2], **Rolando Coto-Solano**[3], **Saeed Hassanpour**[4], and **Soroush Vosoughi**[5]

[1,2,5]Department of Computer Science, Dartmouth College
[3]Department of Linguistics, Dartmouth College
[4]Department of Biomedical Data Science, Dartmouth College
[1]weicheng.ma.gr@dartmouth.edu
[5]soroush.vosoughi@dartmouth.edu

Abstract

Syntactic probing methods have been used to examine whether and how pre-trained language models (PLMs) encode syntactic relations. However, the probing methods are usually biased by the PLMs' memorization of common word co-occurrences, even if they do not form syntactic relations. This paper presents a random-word-substitution and random-label-matching control task to reduce these biases and improve the robustness of syntactic probing methods. Our control tasks are also shown to notably improve the consistency of probing results between different probing methods and make the methods more robust with respect to the text attributes of the probing instances. Our control tasks make syntactic probing methods better at reconstructing syntactic relations and more generalizable to unseen text domains. Our experiments show that our proposed control tasks are effective on different PLMs, probing methods, and syntactic relations.

1 Introduction

To explain the high performance of PLMs on various natural language processing (NLP) tasks, efforts have been made to examine the syntactic relation-encoding ability of these models. For example, Manning et al. (2020) attempt to reconstruct syntactic relations from the attention heads of Transformer models (Vaswani et al., 2017) using raw attention scores. Leave-one-out probing methods (Brunner et al., 2020), instead, measure the influence of ablating parts of each syntactic relation on the hidden representations of the models.

However, the probing results may not faithfully reflect the encoding of syntactic relations as the memorization of common word co-occurrences in the training data of PLMs can lead to incorrect and non-generalizable probing results (Hewitt and Liang, 2019). We observe the same issues in our experiments, where many highly-ranked attention heads by the attention-as-classifier and leave-one-out probing methods highlight frequent word pairs regardless of whether there is a syntactic relation between them. This reduces the trustworthiness of the probing methods and any model interpretation that relies on them. To address this issue and improve the correctness, robustness, and generalizability of existing probing methods, we design two control tasks to reduce the adverse effects of the PLMs' memorization of word co-occurrences. The **random-word-substitution control task** substitutes one component word (i.e., the head or dependent words) of each syntactic relation with its other forms to make the text ungrammatical. The **random-label-matching control task** randomly

Positive:
Influential members of the House Ways and Means Committee introduced legislation that would restrict how the new savings-and-loan bailout agency can raise capital, creating another potential obstacle to the government's sale of sick thrifts.

Random-Word-Substitution:
Influential members of the House Ways and Means Committee introduction legislation that would restrict how the new savings-and-loan bailout agency can raise capital, creating another potential obstacle to the government's sale of sick thrifts.

Random-Label-Matching:
Influential members of the House Ways and Means Committee introduced legislation that would restrict how the new savings-and-loan bailout agency can raise capital, creating another potential obstacle to the government's sale of sick thrifts.

Figure 1: Top: An instance labeled with the correct "subject" dependency relation (Positive); Middle: the instance generated by Random-Word-Substitution where the instance is labeled with the correct pair of words but incorrect word form; Bottom: the instance generated by Random-Label-Matching where the instance is labeled with an incorrect pair of words. The head verb is in blue and the dependent is in red for all the examples.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 402–415

matches one component word of each syntactic relation with a random irrelevant word in the sentence to make the syntactic-relation labels incorrect. Figure 1 shows examples for each control task. The control instances (i.e., negative instances) are generated automatically by substituting words or labels of instances in the positive datasets.

By down-weighting the attention heads that are ranked highly by the probing methods on the control tasks, we observe notably more consistent probing results between the attention-as-classifier and leave-one-out methods on the BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models, with improvements above 0.1 for the Spearman's rank correlation coefficients (Spearman's ρ). [1]. The layer-wise distributions of top-ranked attention heads also become notably more consistent across different text attributes of the probing instances. The results demonstrate the effectiveness of our proposed control tasks for improving the quality and robustness of syntactic probing methods.

2 Syntactic Probing Methods

Different families of probing methods rely on different assumptions (Belinkov and Glass, 2019) and as such, probing results from different families cannot be meaningfully compared. Hence, we examine two probing methods that are both based on attention distributions: (1) Given a sentence and a headword for a syntactic relation, the **attention-as-classifier** method (Manning et al., 2020) predicts another word as the dependent if it puts the highest attention score on the headword; (2) As an attention-based version of the **leave-one-out** probing method used by Meister et al. (2021), we mask the headword of each syntactic relation for each sentence and predict the word whose attention distribution changes the most as the dependent word. Following Kobayashi et al. (2020), we additionally examine two variant methods, **norm-as-classifier** and **leave-one-out-norm** methods which predict the dependent words based on the distributions or changes of attention norms, respectively. We calculate the importance of each attention head for encoding each syntactic relation by evaluating the top-3 accuracy (ACC@3) of the predictions; defined as the percentage of instances where the dependent words from the ground truth are ranked among the top-3 in the predictions. We use ACC@3 since in many cases, the highest attention scores fall on separator tokens such as "[SEP]" and punctuation marks (Clark et al., 2019a).

3 Probing Datasets

We use the "subject" (subj), "object" (obj), "nominal modifier" (nmod), "adverbial modifier" (advmod), and "coreference" (coref) relations in our analyses. We use the English dataset for the CoNLL-2009 shared task (Hajič et al., 2009) to construct our positive and control probing datasets. Figure 1 shows an example instance from the positive dataset and each control dataset.

3.1 Positive Datasets

Our positive dataset for each syntactic relation contains the correct annotations of words that make up the syntactic relation, e.g., the subject words and the corresponding verbs for "subj". The gold-standard dependency annotations in the CoNLL-2009 dataset are used for the "subj", "obj", "nmod", and "advmod" relations and the SpanBERT model (Joshi et al., 2020) is used to annotate the "coref" relation [2].

3.2 Random-Word-Substitution Control

If an attention head in a Transformer model encodes a specific syntactic relation, it should not highlight the connections between words that do not form that syntactic relation. To measure and control for this effect, we construct the random-word-substitution control dataset by substituting one component word of the syntactic relation in each instance of the positive datasets with another part of speech of the same word (e.g., changing a verb to its noun form) to make the instance ungrammatical but not greatly change its semantics. We use the Language Tool [3], a grammar correction tool, to verify that the sentences become ungrammatical after word substitution.

3.3 Random-Label-Matching Control

We also extend the existing method of the random control task (Hewitt and Liang, 2019) to construct the random-label-matching control dataset. Specifically, for each instance in our positive datasets, we use our gold-standard labels and coreference labels generated by SpanBERT to remove word pairs that

[1] All the reported Spearman's ρ are statistically significant ($p < 0.01$), unless specified otherwise.

[2] SpanBERT achieves an F1 score of 79.60% on the Ontonotes v5.0 coreference dataset (Pradhan et al., 2012).

[3] https://languagetool.org/

are syntactically related, leaving us with words that are not syntactically related. These words are then used to create syntactically unrelated pairs by combining known head words with randomly selected dependent words. We then (intentionally) mislabel each pair as forming a specific syntactic relation, depending on the positive dataset from which the instance was taken. Attention heads that encode the relations between these syntactically unrelated word pairs are likely memorizing the co-occurrence of frequent word pairs without regard to syntactic correctness and thus should not be ranked highly by syntactic probing methods.

4 Experimental Results

We conduct three sets of experiments to examine our probing methods' sensitivity to "spurious" word correlations (Section 4.1), consistency (Section 4.2), and robustness to text attributes (Section 4.3). We run the experiments using the BERT-base and RoBERTa-base models for generality. All the experiments are run on an Nvidia RTX-6000 GPU.

4.1 Syntactic Relation Reconstruction

We follow Manning et al. (2020) to evaluate the correctness of attention-head rankings produced by the probing methods via syntactic relation reconstruction experiments. Specifically, for a given headword, we use the attention scores (for attention-as-classifier) or norms (for norm-as-classifier) between that headword and all other words in the instance to predict the dependent word. Similarly, We use the distribution changes of the attention scores (for leave-one-out) or norms (for leave-one-out-norm) when the headword is masked to predict the dependent word. Contributive attention heads for encoding a particular syntactic relation should achieve high syntactic-relation reconstruction performance (in ACC@3) given syntactically correct (positive) labels and low performance given incorrect (negative/control) labels.

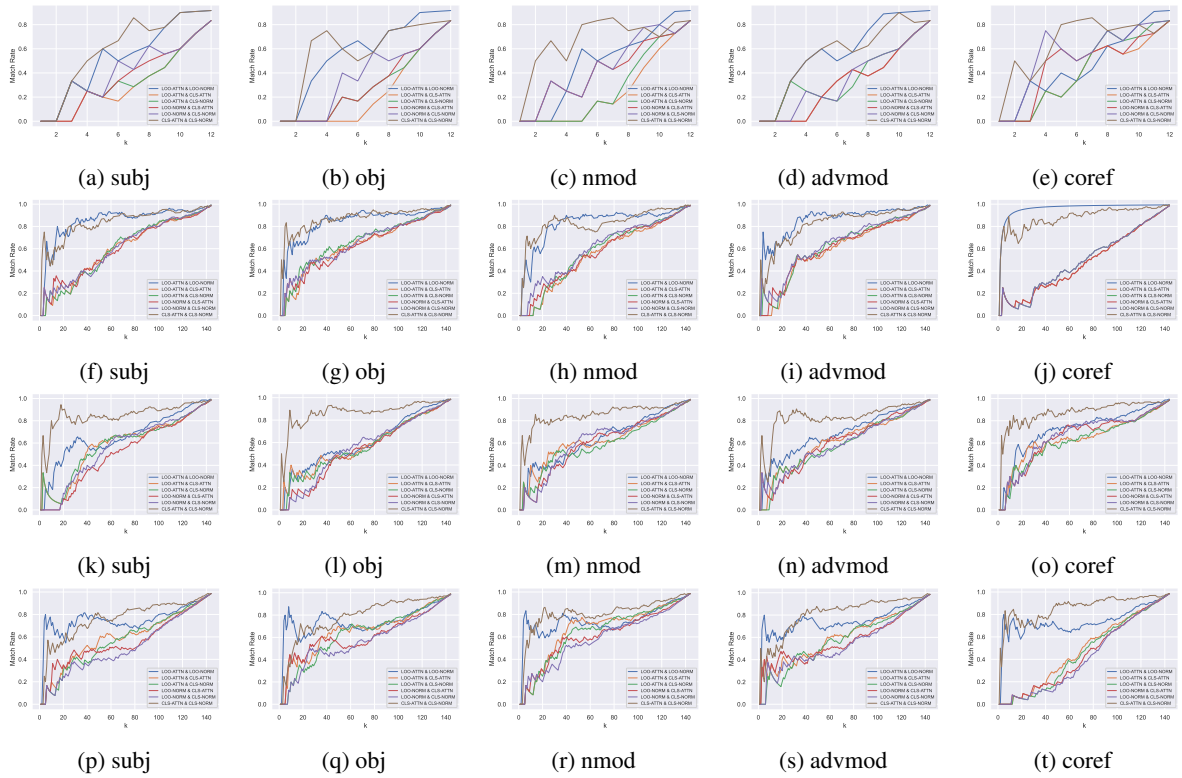
We use the left-out development set of the CoNLL-2009 dataset (labeled using the ground-truth annotations and SpanBERT) as one positive probing dataset (pos-main) and the corresponding random-word-substitution and random-label-matching control instances as two negative datasets. We construct an additional positive probing dataset (pos-uncommon) by substituting the dependent words with other words that have the same part of speech but rarely co-occur (<5 times) with the corresponding headwords in the English Wikipedia corpus [4]. This dataset enables us to study the effect of co-occurrence for syntactically related pairs of words on the syntactic relation reconstruction task. We use the English Wikipedia corpus as it is representative of the data used to pre-train BERT and RoBERTa. All the evaluations are conducted on the top-5 attention heads according to each probing method (with and without control tasks), and the scores are averaged across syntactic relations and heads.

Results show that applying our proposed control tasks does not harm the syntactic-relation reconstruction performance of the four probing methods on the pos-main dataset. In contrast, applying the random control task (Hewitt and Liang, 2019) occasionally leads to a performance drop of 1.32. This suggests that our proposed control tasks are more robust than the existing random control task. On the pos-uncommon dataset, our proposed control tasks lead to an average increase of 9.17 ± 0.13 (BERT) and 4.07 ± 0.15 (RoBERTa) in the syntactic-relation reconstruction performance. Additionally, the control tasks on average reduce the incorrect prediction of syntactic relations in our two negative datasets by 11.70 ± 0.09 (BERT) and 12.69 ± 0.06 (RoBERTa). These results suggest that our proposed control tasks can reduce the influence of the PLMs' memorization of syntactically-irrelevant word co-occurrences for encoding syntactic relations. The complete results of these experiments are shown in Appendix A.

4.2 Consistency of Attention-Head Rankings

We also observe that our control tasks lead to higher consistency between the two categories of probing methods. Without any control task, the Spearman's ρ between the head rankings produced by the four probing methods are always lower than 0.38 (for BERT) and 0.49 (for RoBERTa), while applying the control tasks improves the consistency from a minimum of 0.10 to 0.79 (for BERT) and 0.14 to 0.53 (for RoBERTa), in Spearman's ρ. Furthermore, the highest consistency improvements are achieved when applying both our random-word-substitution and random-label-matching control tasks. Applying the random control task independently or jointly with our two control tasks does not lead to higher consistency improvements. The complete results of these experiments are shown in

[4] https://dumps.wikimedia.org



(a) subj (b) obj (c) nmod (d) advmod (e) coref

(f) subj (g) obj (h) nmod (i) advmod (j) coref

(k) subj (l) obj (m) nmod (n) advmod (o) coref

(p) subj (q) obj (r) nmod (s) advmod (t) coref

Figure 2: Percentage of shared attention heads in the top-k ($1 \leq k \leq 144$) attention heads between each pair of probing methods on the positive data only ((a) to (e)), with the random-word-substitution control task ((f) to (j)), with the random-label-matching control task ((k) to (o)), and with both control tasks ((p) to (t)). Each line represents a pair of probing methods. The x-axes indicate k and the y-axes indicate the percentage of attention heads in common between the two probing methods.

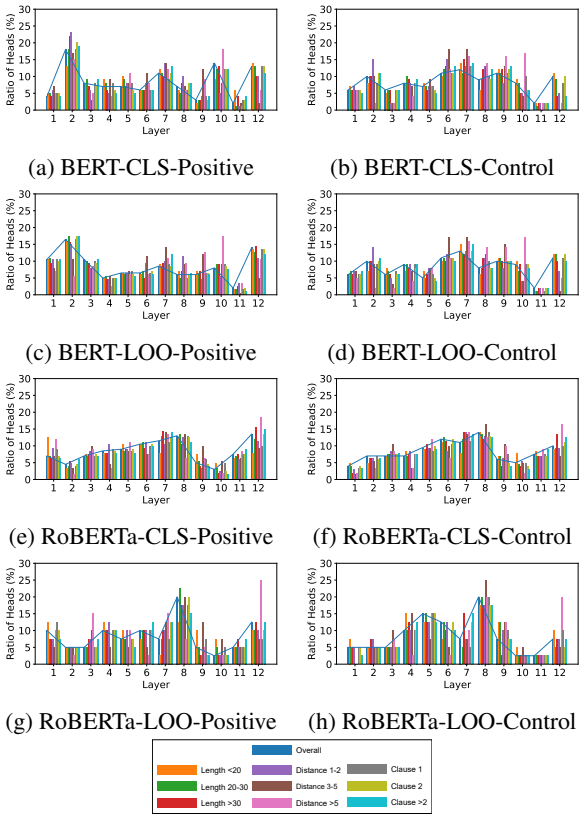
Appendix B.

Prior work has shown that only a small focused set of heads contributes to the encoding of each linguistic feature (Michel et al., 2019; Voita et al., 2019), and as such, a good probing method should highlight these select contributive heads. Figure 2 shows the percentage of attention heads in common among the top-k heads ($1 \leq k \leq 144$) between each pair of probing methods, either with or without control tasks. We find that applying the control tasks generally improves the agreement between attention-head rankings, with the effect being more pronounced for the top 15% of the heads, i.e., the attention heads that are deemed the most important for encoding each syntactic rule. These results show that our control tasks aid the probing methods in highlighting the small set of contributive heads.

4.3 Robustness to Text Attributes

The literature suggests that most contributive attention heads for encoding syntactic relations lie on the middle layers of Transformer models (Hewitt and Manning, 2019; Vig and Belinkov, 2019; Goldberg, 2019; Jawahar et al., 2019; Clark et al., 2019b). Consequently, the layer-wise distribution of the attention heads ranked highly by a robust syntactic probing method should follow a similar pattern and not be greatly affected by the variation in the text attributes.

We divide the pos-main dataset into nine subsets with different sentence lengths (< 20 tokens, $20 - 30$ tokens, and > 30 tokens), numbers of clauses (1, 2, and > 2 clauses), and distances between the head and dependent words ($1 - 2$ tokens, $3 - 5$ tokens, and > 5 tokens). The parameters for each of the attributes were selected to create a relatively uniform distribution of sentences for each of the datasets for a given attribute. We repeat all the experiments with the attention-as-classifier and leave-one-out probing methods on these nine datasets. The layer-wise distributions of top-5 attention heads for each probing method (aggregated for the five syntactic relations) are shown in Figure 3. We show the results for the two probing methods with both our combined control tasks and without any control.



(a) BERT-CLS-Positive (b) BERT-CLS-Control

(c) BERT-LOO-Positive (d) BERT-LOO-Control

(e) RoBERTa-CLS-Positive (f) RoBERTa-CLS-Control

(g) RoBERTa-LOO-Positive (h) RoBERTa-LOO-Control

(i) Figure Legend

Figure 3: Ratio of top-5 attention heads (aggregated across the five syntactic relations) falling on each layer, as ranked by the attention-as-classifier (CLS) and leave-one-out (LOO) probing methods. Positive and Control represent the settings with no control task and the combination of random-word-substitution and random-label-matching control tasks.

We note that the overall trend (represented by the blue line in each figure) shows that the top-ranked attention heads are over-represented on the middle layers, either with or without control tasks. This is well-aligned with the literature, suggesting that the most contributive attention heads for encoding syntactic relations (i.e., middle layers) are identified by the probing methods even without any control tasks (Hewitt and Manning, 2019; Vig and Belinkov, 2019; Goldberg, 2019; Jawahar et al., 2019). However, the probing methods without control tasks also put high weights on the low-level layers (below Layer 2) more frequently than those with control tasks. We speculate the cause to be the sensitivity of the probing methods (without control tasks) to the memorization of common word co-occurrences on each attention head; since the lower-layer attention heads are closer to the embedding layer, they usually encode richer lexical features (Limisiewicz and Mareček, 2021). Our claim is further supported by the observation that there is greater variation in the attention-head rankings between the individual probing results for each of the nine attributes when no control is used. This can be visually observed in Figure 3 by comparing the deviation between different colored bars (corresponding to different attributes) on the left and right figures, corresponding probing without and with controls, respectively. We additionally measure this difference in variation quantitatively by examining the consistency of the attention-head rankings over the entire 144 heads for individual probing results for each of the nine attributes. The Spearman's ρ of the rankings between all settings (i.e., using the entire development set or any of the nine subsets) range from 0.75 to 0.96 when using the combination of the random-word-substitution and random-label-matching control tasks. In comparison, Spearman's ρ of the rankings between the settings drops to 0.22 and 0.38 when no control task is applied and between 0.51 and 0.60 when the random control task is used. These experiments suggest that our proposed control tasks can improve syntactic probing methods' robustness and reduce syntactic probing methods' fragility to the models' memorization of common word co-occurrences.

5 Conclusion and Future Work

This paper proposes two control tasks to improve the syntactic probing of PLMs and reduce the noise in the probing results of the PLMs' memorization of common word co-occurrences. By applying these control tasks, we observe notable improvements in the correctness and consistency of the results produced by four attention-based probing methods across two categories of five diverse syntactic relations. The improvements are also robust to different PLMs' and attributes of the probing instances, suggesting the general applicability of our proposed control tasks.

Future work can expand the use of our proposed control tasks to other models or syntactic relations.

Acknowledgement

This work was partially funded by Dr. Vosoughi's 2022 Google Research Award.

Limitations

While our study provides promising results in reducing biases and improving the robustness of syn-

tactic probing methods, there are some limitations that must be discussed:

First, our experiments only utilized attention-based probing approaches, and it is unclear whether our results would generalize to other families of probing methods. Therefore, further investigation is needed to determine the effectiveness of our control tasks for other types of probing methods. Second, we only explored a subset of syntactic relations in English, including subject, object, nominal modifier, adverbial modifier, and coreference. Our results may not be generalizable to other syntactic relations or languages. Future studies could expand the exploration of other syntactic features and investigate the effectiveness of our control tasks in different languages. Third, our experiments only focused on two pre-trained language models, namely BERT and RoBERTa. It is unclear whether our control tasks would be effective for other types of PLMs, and further studies could investigate the effectiveness of our control tasks on other types of PLMs. Finally, our study only focused on syntactic probing methods and did not investigate probing methods for other types of NLP tasks, such as natural language inference, machine translation, and summarization. Therefore, further studies could explore the effectiveness of our control tasks on other types of NLP tasks.

Despite these limitations, our proposed control tasks have shown promising results in reducing biases and improving the robustness of syntactic probing methods, and we hope that our work will inspire further research in this direction.

Ethics Statement

This paper used publicly available pre-trained models (bert-base-cased and roberta-base models and the SpanBERT model) and a publicly available dataset (CoNLL-2009). No sensitive information is introduced to the data annotations or experiments. Also, we only examine the ways pre-trained language models encode general syntactic relations, which should not introduce stereotypes or biases into our results and analyses. We do not foresee any potential ethical concerns in our work. However, we should note that our work is limited to English syntactic relations and should not be generalized to other languages without additional experiments.

References

Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93.

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019a. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019b. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado. 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.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Tomasz Limisiewicz and David Mareček. 2021. Introducing orthogonal constraint in structural probes. 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 428–442, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Clara Meister, Stefan Lazov, Isabelle Augenstein, and Ryan Cotterell. 2021. Is sparse attention more interpretable? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 122–129, Online. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

A syntactic relation Reconstruction Results

We display in Tables A1 - A4 the average syntactic relation reconstruction performance on the top-5 attention heads produced by each probing method for the five syntactic relations ("subj", "obj", "nmod", "advmod", and "coref") on the pos-main, pos-uncommon, random-word-substitution, and random-label-matching datasets, respectively.

B The Inconsistency Across Probing Methods

The attention-head rankings produced by different probing methods are inconsistent when no control task is applied. As Figure B1 shows, the Spearman's ρ between each pair of probing methods are always lower than 0.38 for BERT and below 0.49 for RoBERTa, which falls under the "weak to moderate correlation" range given the interpretation of Akoglu (2018). As shown in Figures B2 and B3, by applying the random-word-substitution or the random-label-matching control tasks, Spearman's ρ across probing methods improve greatly, in some cases yielding Spearman's ρ above 0.7 (i.e., "very strong" correlations). Though not as effective as our proposed control tasks, applying the random control task also improves the consistencies of attention-head rankings across the probing methods.

As shown in the figures, combining our two control tasks generates the most consistent results for all four probing methods.

BERT				
Control	CLS	CLS-N	LOO	LOO-N
None	54.01 (0.10)	56.90 (0.11)	51.73 (0.13)	46.50 (0.47)
RAND	53.91 (0.10)	57.57 (0.12)	52.61 (0.14)	47.61 (0.45)
RWS	54.50 (0.11)	57.72 (0.10)	52.82 (0.11)	47.71 (0.12)
RLM	54.88 (0.10)	57.86 (0.11)	52.68 (0.12)	47.91 (0.15)
RWS+RAND	54.89 (0.11)	57.79 (0.10)	52.77 (0.10)	47.66 (0.17)
RLM+RAND	54.46 (0.11)	57.80 (0.10)	52.60 (0.09)	47.91 (0.18)
RWS+RLM	**54.99** (0.10)	**58.00** (0.10)	**52.98** (0.13)	**48.06** (0.14)
ALL	54.88 (0.11)	57.84 (0.11)	52.95 (0.14)	47.99 (0.17)
RoBERTa				
Control	CLS	CLS-N	LOO	LOO-N
None	55.03 (0.13)	58.33 (0.11)	56.69 (0.11)	57.79 (0.09)
RAND	55.50 (0.12)	59.15 (0.13)	57.93 (0.10)	56.47 (0.17)
RWS	56.34 (0.10)	60.17 (0.12)	58.17 (0.10)	58.19 (0.09)
RLM	56.37 (0.09)	60.18 (0.11)	58.13 (0.12)	58.34 (0.08)
RWS+RAND	55.65 (0.14)	60.02 (0.12)	58.48 (0.11)	58.03 (0.11)
RLM+RAND	55.70 (0.13)	60.51 (0.14)	58.44 (0.13)	58.28 (0.10)
RWS+RLM	**56.89** (0.10)	**60.83** (0.11)	**58.74** (0.10)	**58.95** (0.08)
ALL	56.39 (0.13)	**60.83** (0.14)	58.53 (0.12)	58.82 (0.10)

Table A1: syntactic relation reconstruction performance (ACC@3) on the pos-main dataset. The ACC@3 scores are averaged over all five syntactic relations and the top 5 attention heads as ranked by each probing method, both with and without control tasks. CLS, CLS-N, LOO, and LOO-N refer to the attention-as-classifier, norm-as-classifier, leave-one-out, and leave-one-out-norm probing methods, respectively. RAND, RWS, and RLM refer to the random, random-word-substitution, and random-label-matching control tasks, respectively. None and ALL indicate applying no or all three control tasks. The highest (best) scores are in bold, and the lowest (worst) scores are underlined.

BERT				
Control	CLS	CLS-N	LOO	LOO-N
None	41.37 (0.13)	40.27 (0.10)	54.40 (0.17)	58.68 (0.18)
RAND	37.95 (0.17)	45.26 (0.09)	54.43 (0.22)	59.05 (0.20)
RWS	42.68 (0.15)	46.97 (0.12)	56.26 (0.14)	61.51 (0.12)
RLM	42.06 (0.13)	46.67 (0.08)	55.75 (0.16)	60.08 (0.11)
RWS+RAND	41.89 (0.20)	47.15 (0.09)	56.60 (0.14)	64.99 (0.15)
RLM+RAND	40.41 (0.17)	45.61 (0.10)	55.89 (0.17)	61.03 (0.16)
RWS+RLM	**45.16** (0.15)	**48.75** (0.06)	**60.38** (0.10)	**77.13** (0.16)
ALL	43.74 (0.14)	47.04 (0.07)	60.15 (0.16)	75.36 (0.15)
RoBERTa				
Control	CLS	CLS-N	LOO	LOO-N
None	37.11 (0.09)	42.52 (0.16)	70.03 (0.22)	70.97 (0.18)
RAND	33.97 (0.12)	45.57 (0.16)	68.21 (0.17)	72.11 (0.18)
RWS	38.06 (0.12)	46.48 (0.14)	74.03 (0.20)	73.46 (0.15)
RLM	37.81 (0.13)	46.32 (0.10)	72.81 (0.18)	72.70 (0.20)
RWS+RAND	38.73 (0.10)	46.47 (0.15)	71.70 (0.15)	72.22 (0.19)
RLM+RAND	37.82 (0.11)	46.06 (0.13)	71.37 (0.17)	72.39 (0.20)
RWS+RLM	**39.26** (0.10)	**48.35** (0.10)	**74.84** (0.17)	**74.45** (0.21)
ALL	**39.26** (0.08)	47.81 (0.13)	72.69 (0.10)	72.39 (0.14)

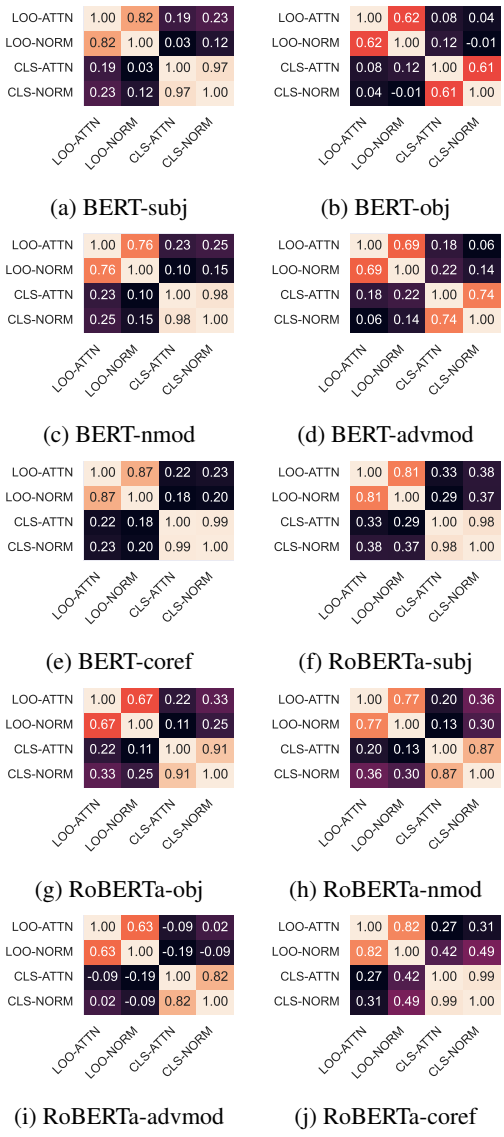
Table A2: syntactic relation reconstruction performance on the pos-uncommon dataset. The highest (best) scores are in bold, and the lowest (worst) scores are underlined.

BERT				
Control	CLS	CLS-N	LOO	LOO-N
None	67.75 (0.10)	70.28 (0.10)	54.59 (0.14)	51.09 (0.13)
RAND	64.08 (0.13)	66.13 (0.12)	51.63 (0.13)	46.74 (0.12)
RWS	56.37 (0.09)	58.56 (0.10)	42.02 (0.13)	39.19 (0.11)
RLM	55.99 (0.10)	58.83 (0.10)	42.11 (0.12)	38.64 (0.12)
RWS+RAND	56.57 (0.09)	58.95 (0.09)	42.72 (0.12)	39.30 (0.11)
RLM+RAND	56.08 (0.12)	59.15 (0.10)	42.92 (0.13)	39.16 (0.12)
RWS+RLM	**53.38** (0.09)	**56.72** (0.12)	**37.95** (0.13)	**34.55** (0.12)
ALL	54.79 (0.10)	57.13 (0.12)	39.20 (0.08)	36.07 (0.11)
RoBERTa				
Control	CLS	CLS-N	LOO	LOO-N
None	64.86 (0.10)	66.58 (0.17)	64.81 (0.07)	65.95 (0.08)
RAND	56.06 (0.17)	57.79 (0.10)	58.95 (0.09)	60.83 (0.11)
RWS	50.92 (0.09)	53.85 (0.10)	51.17 (0.08)	52.99 (0.08)
RLM	50.23 (0.07)	53.44 (0.09)	51.70 (0.06)	53.25 (0.08)
RWS+RAND	51.21 (0.11)	54.34 (0.09)	52.13 (0.07)	52.20 (0.11)
RLM+RAND	51.07 (0.09)	54.04 (0.10)	52.94 (0.10)	53.70 (0.08)
RWS+RLM	**46.97** (0.10)	**49.50** (0.12)	**47.45** (0.11)	**48.59** (0.10)
ALL	49.36 (0.12)	51.13 (0.09)	50.03 (0.09)	51.16 (0.09)

Table A3: syntactic relation reconstruction performance on the random-word-substitution negative dataset. The lowest (best) scores are in bold, and the highest (worst) scores are underlined.

BERT				
Control	CLS	CLS-N	LOO	LOO-N
None	18.22 (0.10)	17.45 (0.04)	17.88 (0.01)	18.56 (0.10)
RAND	14.13 (0.03)	13.21 (0.01)	13.29 (0.02)	14.24 (0.06)
RWS	10.82 (0.02)	11.50 (0.02)	11.31 (0.02)	11.36 (0.02)
RLM	10.83 (0.05)	12.29 (0.08)	11.29 (0.02)	11.86 (0.03)
RWS+RAND	12.03 (0.10)	11.81 (0.01)	11.86 (0.05)	11.33 (0.05)
RLM+RAND	12.29 (0.05)	12.74 (0.04)	11.95 (0.03)	12.17 (0.04)
RWS+RLM	**10.22** (0.02)	**9.65** (0.03)	**9.51** (0.02)	**10.16** (0.02)
ALL	10.57 (0.02)	10.13 (0.05)	10.15 (0.03)	11.19 (0.02)
RoBERTa				
Control	CLS	CLS-N	LOO	LOO-N
None	18.60 (0.01)	19.67 (0.02)	16.52 (0.04)	17.02 (0.02)
RAND	13.31 (0.09)	15.04 (0.10)	13.36 (0.02)	12.33 (0.04)
RWS	11.09 (0.01)	11.26 (0.01)	10.81 (0.01)	10.03 (0.02)
RLM	11.62 (0.01)	11.30 (0.03)	11.03 (0.01)	10.40 (0.01)
RWS+RAND	11.37 (0.04)	12.41 (0.08)	11.28 (0.02)	10.84 (0.02)
RLM+RAND	12.44 (0.03)	13.38 (0.05)	11.42 (0.01)	11.56 (0.02)
RWS+RLM	**10.85** (0.01)	**10.23** (0.02)	**9.37** (0.02)	**9.53** (0.02)
ALL	12.11 (0.02)	12.85 (0.05)	9.66 (0.03)	9.83 (0.02)

Table A4: syntactic relation reconstruction performance on the random-label-matching negative dataset. The lowest (best) scores are in bold, and the highest (worst) scores are underlined.



(a) BERT-subj (b) BERT-obj

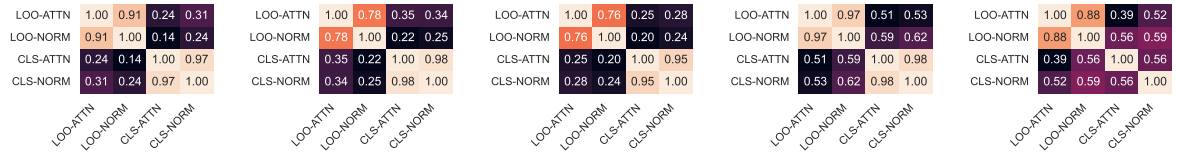
(c) BERT-nmod (d) BERT-advmod

(e) BERT-coref (f) RoBERTa-subj

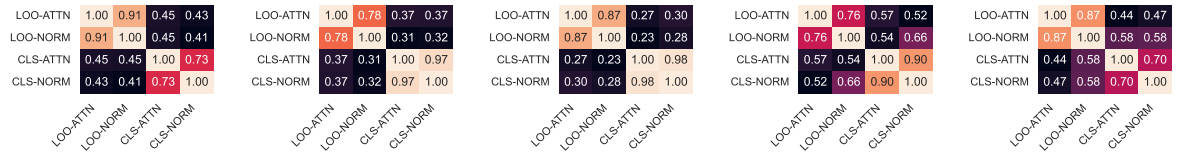
(g) RoBERTa-obj (h) RoBERTa-nmod

(i) RoBERTa-advmod (j) RoBERTa-coref

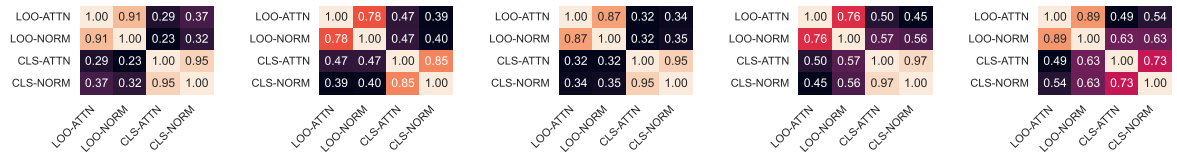
Figure B1: Spearman's ρ between BERT ((a) to (e)) and RoBERTa ((f) to (j)) head rankings produced by four probing methods on the positive dataset of five syntactic relations. LOO-ATTN, LOO-NORM, CLS-ATTN, and CLS-NORM refer to the leave-one-out-attention, leave-one-out-norm, attention-as-classifier, and norm-as-classifier probing methods, respectively.



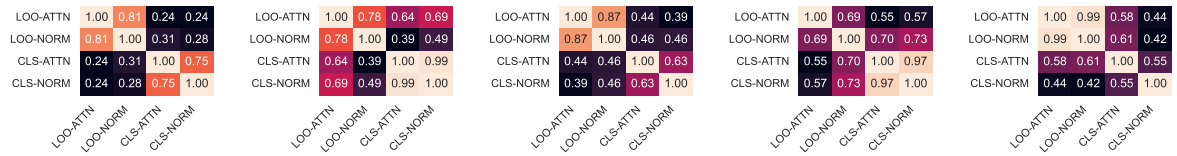
(a) RAND



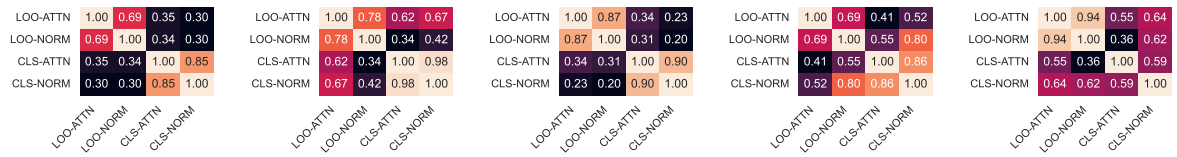
(b) RWS



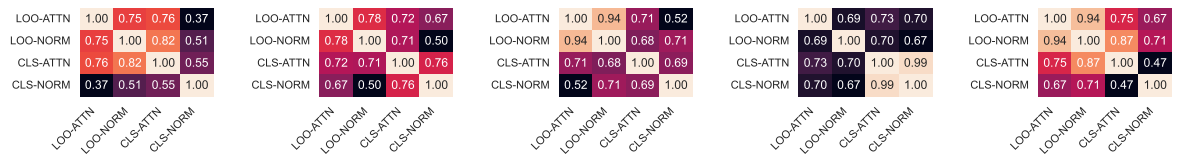
(c) RLM



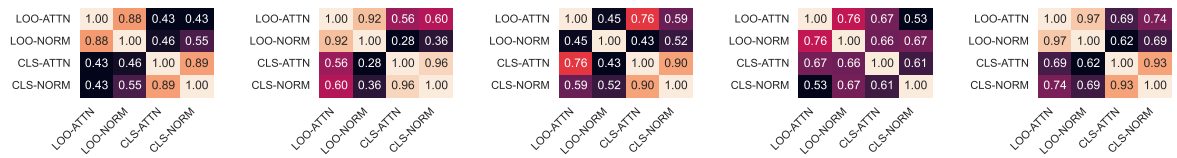
(d) RAND+RWS



(e) RAND+RLM

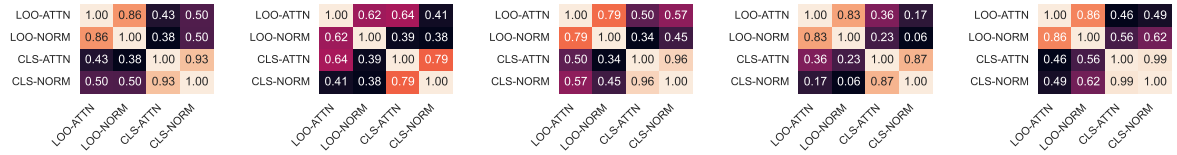


(f) RWS+RLM

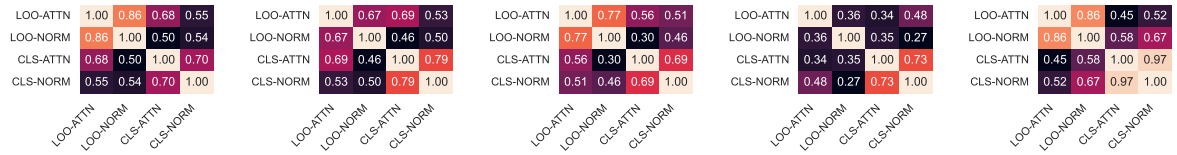


(g) RAND+RWS+RLM

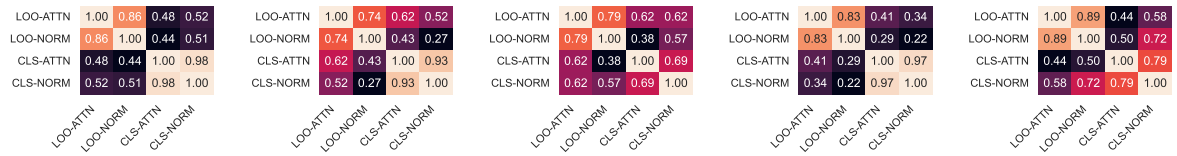
Figure B2: Spearman's ρ between BERT head rankings produced by four probing methods with different control tasks on five syntactic relations (in the column order of "subj", "obj", "nmod", "advmod", and "coref"). RAND, RWS, and RLM refer to the random, random-word-substitution, and random-label-matching control tasks. LOO-ATTN, LOO-NORM, CLS-ATTN, and CLS-NORM refer to the leave-one-out-attention, leave-one-out-norm, attention-as-classifier, and norm-as-classifier probing methods, respectively.



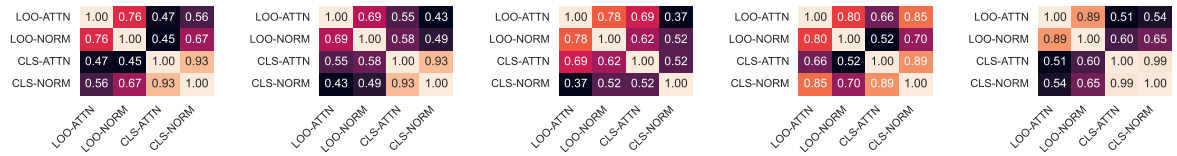
(a) RAND



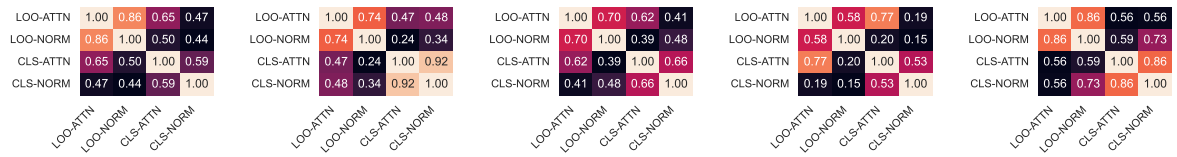
(b) RWS



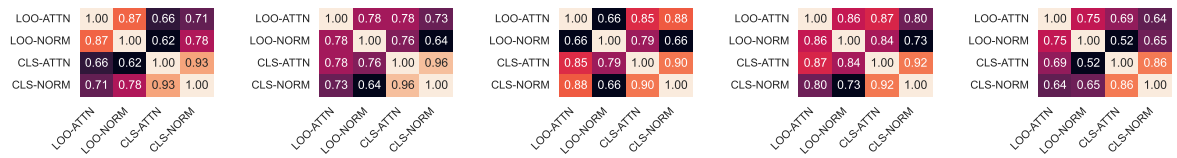
(c) RLM



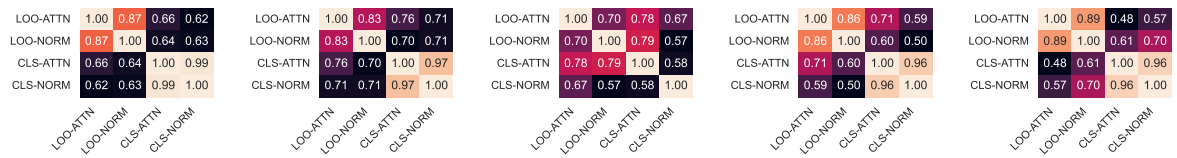
(d) RAND+RWS



(e) RAND+RLM



(f) RWS+RLM



(g) RAND+RWS+RLM

Figure B3: Spearman's ρ between RoBERTa head rankings produced by four probing methods with control tasks on five syntactic relations (in the order of "subj", "obj", "nmod", "advmod", and "coref" on each row). RAND, RWS, and RLM refer to the random, random-word-substitution, and random-label-matching control tasks. LOO-ATTN, LOO-NORM, CLS-ATTN, and CLS-NORM refer to the leave-one-out-attention, leave-one-out-norm, attention-as-classifier, and norm-as-classifier probing methods, respectively.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3.

☑ B1. Did you cite the creators of artifacts you used?
Section 3.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
All the datasets we use are publicly available, and they are cited in Section 3.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
sets we use are publicly available, and they are cited in Section 3.

C ☑ Did you run computational experiments?

Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. We conducted probing experiments that do not require training or hyperparameter search.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.