How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope

Yiyun Zhao
Department of Linguistics
University of Arizona
yiyunzhao@arizona.edu

Steven Bethard
School of Information
University of Arizona
bethard@arizona.edu

Abstract

Large pretrained language models like BERT, after fine-tuning to a downstream task, have achieved high performance on a variety of NLP problems. Yet explaining their decisions is difficult despite recent work probing their internal representations. We propose a procedure and analysis methods that take a hypothesis of how a transformer-based model might encode a linguistic phenomenon, and test the validity of that hypothesis based on a comparison between knowledge-related downstream tasks with downstream control tasks, and measurement of cross-dataset consistency. We apply this methodology to test BERT and RoBERTa on a hypothesis that some attention heads will consistently attend from a word in negation scope to the negation cue. We find that after fine-tuning BERT and RoBERTa on a negation scope task, the average attention head improves its sensitivity to negation and its attention consistency across negation datasets compared to the pre-trained models. However, only the base models (not the large models) improve compared to a control task, indicating there is evidence for a shallow encoding of negation only in the base models.

1 Introduction

As large-scale pre-trained language models such as BERT and ELMo have achieved high performance in a variety of natural language processing tasks (Peters et al., 2018a; Radford et al., 2018; Devlin et al., 2019), a growing body of research is devoted to understanding what linguistic properties these language models have acquired. Recent work uses *probes*, which are supervised models trained to predict linguistic properties including morphology (Belinkov et al., 2017), syntax (Hewitt and Manning, 2019) and semantics (Peters et al., 2018b), etc. (See Belinkov and Glass (2019) for a complete survey.) A good probing performance is considered as evidence that the language models have learned the linguistic knowledge.

What is not yet well understood is how this encoded linguistic knowledge changes when a pretrained language model is fine-tuned for a downstream task. Peters et al. (2019) applies a supervised probe both before and after fine-tuning BERT, and suggests that fine-tuning makes the internal representation task-sensitive. But with supervised probes it can be difficult to disentangle what was learned by the probe from what was present in the internal representation (Hewitt and Liang, 2019).

Recent studies have thus turned to unsupervised probes that require no additional training of the model and instead look directly at the attention mechanism, i.e., how much to care about other words when computing the next version of the current word. Clark et al. (2019) inspected pretrained transformers and found several syntactic properties encoded in an intuitive way, where the maximum attention from a dependent is on its syntactic head. But only the pretrained models were considered, not what happened to these intuitive encodings after fine-tuning to a downstream task.

We argue that if some interpretable encoding of linguistic knowledge is a good explanation of a model, rather than showing it in the pretrained model, it is more important to show it will be enhanced by fine-tuning on a task where that linguistic knowledge is necessary. If the encoding is not enhanced by such fine-tuning, then the model must be using some other mechanism to encode that linguistic knowledge. We therefore propose the following methodology for testing whether a hypothesized encoding of a linguistic phenomenon is a good explanation for a transformer's predictions.

1. Hypothesize an attention representation of the knowledge of interest and design an unsupervised probe, such that each attention head can

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4729–4747
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

make its own prediction.

2. Identify a downstream task related to the knowledge of interest, and design a control task that is learnable and has a similar input and output space but is not related to the knowdge of interest.
3. Fine-tune on both the downstream and control tasks, and measure the unsupervised probe performance of each attention head before and after fine-tuning.

Applying this methodology and a variety of analyses that it enables, and focusing on the phenomenon of linguistic negation scope in a intuitive encoding (the maximal attention from a word in negation scope will be on the negation cue), we find that:

1. Before fine-tuning, several attention heads are sensitive to negation scope. The best heads are better than a fixed-offset baseline, with the best BERT-base head achieving an F_1 of 53.8 in a fully unsupervised setting.
2. There is consistency in which heads are negation-sensitive across different datasets.
3. After fine-tuning on a negation scope task, the average sensitivity of attention heads improved over the pretrained model for all four models (BERT-base, BERT-large, RoBERTa-base, RoBERTa-large) but only the two base models improved more than the control task.
4. The rich do not get richer: attention heads that had the top F_1s in the pretrained model do not have the top-ranked improvements after fine-tuning on negation scope.
5. The behavior of individual attention heads becomes more consistent across datasets after fine-tuning on the negation task, compared to the pretrained model and the control task, except for RoBERTa-large.

Items 1 and 2 suggest that in the pretrained models negation scope may be encoded via attention to negation cues. Items 3 to 5 indicate that during fine-tuning, this encoding continues to play a role in BERT-base and RoBERTa-base, but RoBERTa-large and BERT-large may rely on other mechanisms to represent negation scope. The analysis code is available at https://github.com/yiyunzhao/negation-scope-probing

Though our findings are specific to the linguistic phenomenon of negation scope and the specific attention encoding we hypothesized, our proposed methodology and analyses are general, and can easily be applied to other linguistic phenomena or other encoding hypotheses to discover the role they play in modern pre-trained neural network models.

2 Background

2.1 BERT and attention heads

We performed our analysis on the attention mechanism of uncased BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), large Transformer models (Vaswani et al., 2017). In the following text, we primarily focus on BERT-base and refer the reader to the appendix for detailed results on the other models. BERT-base contains 12 layers and each layer contains 12 attention heads. Each attention head takes a sequence of input vectors $h = [h_1, .., h_n]$ that correspond to the n tokens. An attention head transforms each h_i into query (q_i), key (k_i) and value (v_i) vectors and computes an output vector (o_i) via a weighted sum of value vectors based on attention weights (a_i) :

$$a_{ij} = \frac{\exp(q_i^T k_j)}{\sum_{l=1}^{n} \exp(q_i^T k_l)} \tag{1}$$

$$o_i = \sum_{j=1}^{n} a_{ij} v_j \tag{2}$$

Attention weights can be viewed as the amount of contribution from other tokens to the new representation of the current token.

2.2 Negation scope

Negation is a grammatical structure that reverses the truth value of a proposition. The tokens that express the presence of negation are the *negation cue* and the tokens that are affected by the negation cue belong to the *negation scope*. For example, in the following sentence, *not* is the negation cue and the underlined tokens are the negation scope.

> *Holmes was sitting with his back to me, and <u>I had given him</u> {no} <u>sign of my occupation</u>.*

Knowledge about negation and its scope is important for tasks such as sentiment anlaysis and logical inference. And as a linguistic phenomenon that bridges between syntax and semantics, it is a good candidate for exploring BERT's attention, as related phenomena have already been found in BERT (Tenney et al., 2019; Clark et al., 2019).

3 Methodology and Analyses

In this section, we explain our proposed methodology and analyses, and illustrate their application to the linguistic phenomenon of negation scope.

Step 1: hypothesize an interpretable representation of the phenomenon of interest. Transformer models could represent linguistic knowledge in many ways: attention, contextualized embeddings, etc. To apply our methodology, one must first hypothesize a specific encoding of the phenomenon of interest. For negation scope, we hypothesize that for some subset of attention heads, words in negation scope will attend primarily to the negation cue, while words out of negation scope will attend primarily to other words (see Section 4.1). Under this hypothesis, each attention head is an unsupervised negation scope classifier.

Step 2: Identify a downstream task that requires the phenomenon of interest. To infer that a transformer model is explainable in terms of the hypothesized encoding, we must see evidence that the encoding is strengthened when fine-tuning on a task that requires the phenomenon of interest. If the encoding is visible in the pre-trained model but disappears during fine-tuning, then the model is handling the phenomenon through some other mechanism. For negation scope, our downstream tasks are supervised negation scope prediction problems (see Section 5.1).

Step 3: Design a control task where the phenomenon of interest is irrelevant. The control task should have input and output spaces that match those of the downstream task but should be learnable without any knowledge of the phenomenon. For negation scope, we arbitrarily assign word types to binary labels (see Section 5.1).

Step 4: Analyze differences between models fine-tuned on the downstream and control tasks. If the hypothesized encoding explains the model predictions, changes observed when fine-tuning on the downstream task must be greater than changes observed when fine-tuning on the control task. For negation scope, we analyze changes in performance of individual attention heads as unsupervised negation classifiers.

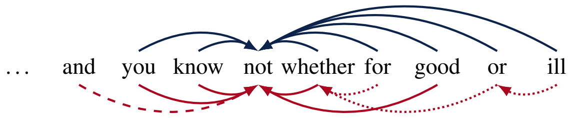


Figure 1: Example text with true negation scope on top and layer 8 head 4's maximally-attended word for each input on the bottom. Dashed lines are precision errors and dotted lines are recall errors.

4 Does BERT pay 'attention' to negation scope before fine-tuning?

We start by hypothesizing a way that negation scope could be encoded in transformer models. This hypothesis must not rely on any negation-specific training data, as we want to be able to measure evidence of the encoding equally well both before and after fine-tuning. Our hypothesized encoding treats each attention head as an unsupervised negation scope classifier.

4.1 Attention as a negation classifier

Our goal is to see if any individual attention head is good at detecting negation scope. Because attention heads by definition compare two tokens to each other, we formulate negation scope detection as a pair-wise task. We treat each attention head as an unsupervised classifier that considers each token in the sentence, and if the maximum attention from that token is to the negation cue, we classify the token as within the negation scope. Formally, the prediction of an attention head for token i is:

$$attendneg(i) = \begin{cases} 1 & \text{if } j_{neg} = \operatorname*{argmax}_{j=1}^{n} a_{ij} \\ 0 & otherwise \end{cases} \quad (3)$$

where j_{neg} is the index of the negation cue, and a_{ij} is attention as defined in Equation (1).

The quality of each attention head as such a negation classifier can be evaluated based on how often it agrees with the true negation scope, as shown in Figure 1. We use the standard measures of precision, recall, and F_1:

$$precision = \frac{\sum_{i=1}^{n} attendneg(i) \wedge negscope(i)}{\sum_{i=1}^{n} attendneg(i)}$$

$$recall = \frac{\sum_{i=1}^{n} attendneg(i) \wedge inscope(i)}{\sum_{i=1}^{n} negscope(i)}$$

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

where $attendneg(i)$ is the unsupervised classifier of Equation (3) and $negscope(i)$ is 1 if i is within the annotated negation scope and 0 otherwise.

4.2 Checking for confounds

If we find an attention head that achieves a high F_1 for negation detection, are we sure that BERT has learned negation? Or could the head be doing something simpler to achieve that F_1? If most negation scopes were just one word after the negation cue, simply attending to the previous word would achieve high performance on the negation task.

To build confidence that attention heads that achieve high F_1 in negation detection aren't somehow cheating, we (1) look at several baselines to establish the difficulty of the task, (2) use a regression to see which factors explain the attention, and (3) look for consistency in attention head performance across different datasets. We use the baselines:

all in-scope: Always attend to the negation token, regardless of the input word. This guarantees 100% recall, but is somewhat unrealistic, since the attention mechanism doesn't know where the negation word is[1].

fixed offset: Always attend to a fixed position relative to the input word. For example, a fixed offset of +1 would mean to always attend to the next word in the sentence, and therefore, according to Equation (3), to only predict a token is in the negation scope if it is immediately followed by the negation cue. Clark et al. (2019) observed several of BERT's attention heads displaying such behavior. We considered fixed offsets from -3 to +3.

Predictors of attention If an attention head has truly learned something about negation, its attention should not be easily explainable by something simpler like the proximity in the text. We thus build a simple regression model using the token's negation scope label (in-scope or out-of-scope) and the distance to the negation cue as predictors, and the attention of the token to the negation cue as the dependent variable. If an attention head is truly detecting negation scope, we expect that scope label will be a significant predictor in this model, and token distance will be much less important.

Consistency across domains If an attention head has truly learned something about negation, we would expect it to perform reasonably well regardless of changes in text genre or style of negation annotation. Several studies show that generalization ability to a different dataset is not always guaranteed despite a good test performance on the same dataset (Weber et al., 2018; McCoy et al., 2019). We thus consider two different corpora annotated for negation: ConanDoyle-neg (Morante and Daelemans, 2012) and SFU Review (Konstantinova et al., 2012)[2]. These datasets differ in genre (Sherlock Holmes stories vs. movie, book, and consumer product reviews) and in annotation schema (e.g., they have different rules for what sentences are considered to contain negation, and how to deal with coordination structure).

Models		**P**	**R**	**F_1**
baseline	all in scope	34.0	**100.0**	50.7
baseline	average fixed offset	66.1	8.6	15.2
baseline	best fixed offset (-1)	**83.5**	11.6	20.4
attention	average head	49.5	5.2	9.0
attention	best head (8-4)	76.2	41.5	**53.8**

Table 1: Performance of unsupervised BERT-base attention-based classifiers and baselines on the negation scope detection task in terms of precision (P), recall (R) and F_1. The best fixed offset and attention head according to their F_1 score are reported.

To see whether the same attention heads are performing well at negation scope detection across the two corpora, we measure kendall rank correlation:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j)$$

where x_i is the performance of attention head i on the Conan Dolye dataset and y_i is the performance of head i on the SFU-review dataset.

4.2.1 Results

Table 1 shows the performance of BERT-base's attention heads and the baselines. Table A1 in the Appendix shows the results for other models. BERT-base attention heads on average are not good predictors of negation scope (49.5% in precision, 5.2% in recall, 9.0% in F_1) but the 4th attention head in layer 8 stands out (76.2% in precision, 41.5% in recall, 53.8% in F_1). This performance is unlike either the best fixed offset baseline (-1) or the

[1] Note that our classifier in Equation (3) *does* know where the negation word is, since it is given j_{neg} as an input. But a standalone transformer model is not given such information.

[2] We exclude cases in these datasets where the negation cue is part of a word (e.g., *im* in *impossible*) because such subword segmentation does not always align to BERT's tokenization.

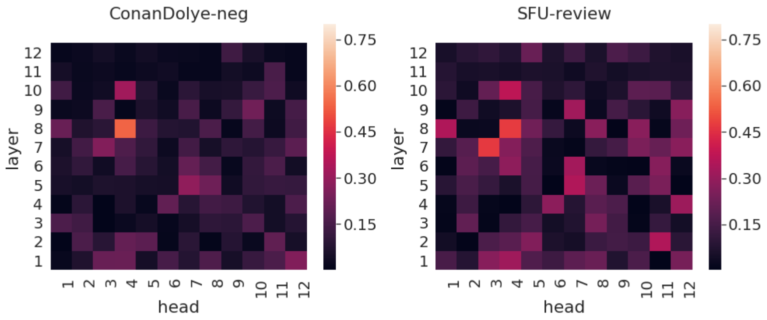


Figure 2: The heatmap of unsupervised negation-scope classification F_1 for BERT-base's 12 layers x 12 heads across two different datasets. The consistency (measure by kendall rank correlation) between the two datasets for precision, recall and F_1 are 0.440, 0.418 and 0.415 respectively. See fig. A1 for precision and recall.

all-in-scope baseline, exceeding both of these in F_1, and with very different precision/recall trade-offs. When we fit a regression model to predict layer 8 head 4's attention based on token distance and the true negation scope label, we found that both distance ($\beta = 0.043, p < 2 \times 10^{-16}$) and label ($\beta = 0.310, p < 2 \times 10^{-16}$) were significant predictors for the attention, but the true negation scope label had a much larger coefficient. Anova tests comparing the full model with a model leaving out distance or label found that true negation scope explains more variance (207.7) than distance (1.5). This suggests that a large part of what the best attention head is doing can be best explained as detecting negation.

Figure 2 shows that there is consistency in the F_1 of BERT-base's attention heads across the two negation scope datasets, e.g., BERT-base's layer 8 head 4 has the best F_1 in both. Kendall correlation tests confirm that the similarities across attention heads of BERT-base are significant: 0.440 tau coefficient ($p = 5.24 \times 10^{-15}$) in precision, 0.418 tau coefficient ($p = 1.20 \times 10^{-13}$) in recall and 0.415 tau coefficient ($p = 1.56 \times 10^{-13}$) in F_1. Figures A1 to A4 in the Appendix show plots for precision and recall, and that similar results hold for the other models. Seeing that attention heads that are predictive of negation in one dataset continue to be predictive in another differently annotated dataset from a different text genre suggests that these most successful heads are indeed learning some form of linguistic negation during the BERT pre-training.

5 What happens to negation-sensitive attention heads when you fine-tune?

We have seen that without any explicit training on a negation task, some attention heads are sensitive to negation scope in an intuitive way (in-scope words attend primarily to the negation cue). What happens to the attention when we fine-tune (i.e., continue training the pre-trained model) on a downstream task that requires an understanding of negation scope? Will this attention-based encoding of negation scope be strengthened? Or will the model choose to represent negation-scope knowledge in some other way during fine-tuning? What about for a downstream task that is unrelated to negation? We answer these questions and others in the following sections by fine-tuning models on downstream tasks, and measuring how this changes the negation-sensitivity of different attention heads.

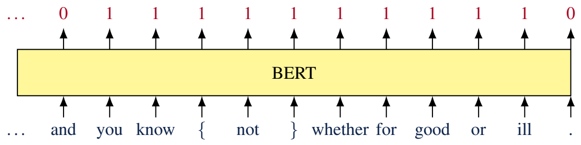


Figure 3: Negation scope detection as a word-piece-by-word-piece binary classification task.

5.1 Downstream Tasks

Downstream negation task We construct a downstream negation scope detection task from the ConanDoyle-neg dataset. As shown in Figure 3, we formulate the problem as a word-piece-by-word-piece binary classification problem, where a word-piece should be labeled 1 if it is in a negation scope and 0 otherwise. To provide the location of the negation cue as an input to the classifier, we add two tokens to the input, surrounding the cue with "{" and "}". As is standard for BERT token classification models, a fully-connected layer with sigmoid activation connects BERT's contextual embedding for each token with the binary outputs that must be predicted. This model can then be trained with BERT's standard back-propagation procedure.

Downstream control task Inspired by the control tasks of Hewitt and Liang (2019), we construct a downstream control task on the ConanDoyle-neg dataset that has the same input space and output space as the downstream negation task, but is constructed to be irrelevant to negation and most other linguistic phenomena. We arbitrarily assign each unique token in the training vocabulary to be always in-scope or always out-of-scope, with a distribution close to the empirical in-scope and out-of-scope distribution. To succeed in this control task, the model must memorize the category (in-scope or out-of-scope) for each token type. Since the assignment is arbitrary, there is no way for the model to generalize to unseen tokens, and thus when we evaluate performance on this task, we consider performance only on the tokens seen during training.

5.2 Fine-tuning classifiers

We split the data into 662 negation frames for training and 200 negation frames for testing. We use the same data split for both the downstream negation scope task and the downstream control task. For each task, we take pre-trained BERT base as our starting point. We fine-tune this model for 50 epochs with a learning rate of 4×10^{-5} using the transformers libary (Wolf et al., 2019), and pick the best epoch based upon its performance on the testing data. For the negation scope task, performance is measured in F_1. For the control task, performance is measured in accuracy on the testing data tokens that have been seen in the training data. We repeat this process 10 times, generating 10 different fine-tuned BERT models for each task, to allow us to quantify variance due to the inherent randomness in neural network training[3].

5.3 Results

Table 2 and Table A2 in the Appendix show that after fine-tuning all models achieve very high performance in both downstream tasks. BERT-base achieves on average 92.8% F_1 for the negation scope task and on average 95.9% accuracy for the control task. The BERT-base model trained on the control task has learned essentially nothing about negation scope relationship, achieving an average 35.4% F_1. These results show that both tasks are learnable from their data, and that the control task is irrelevant to negation scope.

How does fine-tuning change attention? Fine-tuning changes many parameters to make a model better at a downstream task. Will the change be reflected in our hypothesized encoding, i.e., will in-scope words increase their attention to negation cues? And what will the patterns of such a change be? Will sensitivity to negation be spread throughout the attention heads of the model? Will just the attention heads that were already sensitive to negation improve? Or maybe no individual attention heads will get better at negation; the model will only becomes sensitive to negation in aggregate?

We first look at overall changes. Table 3 shows the average performance change across all 144 heads of BERT-base, and for just the best head (layer 8, head 4). Table A3 shows average performance changes for the other models. When BERT-base is fine-tuned on the control task, the F_1 for most heads is similar to what it was before fine-tuning. When BERT is fine-tuned on the negation task, both the average F_1 and the F_1 of the best attention head increase. The Wilcoxon test shows that both the average F_1 ($p = 7.578 \times 10^{-5}$) and the F_1 of the best head ($p = 0.002089$) fine-tuned on the negation task are significantly higher than when fine-tuned on the control task. Table A3 shows that all negation-finetuned models improve over the pretrained models, but only BERT-base and RoBERTa-base improve over the controls.

We next look at changes at the level of individual attention heads.

Figure 4 plots the average F_1 performance gain for each of BERT-base's 144 attention heads after fine-tuning on either the negation or control task. Figure A5 in the Appendix plots the same for the other models. These plots show that in negation-finetuned models the mid-to-late layers of attention heads improve their sensitivity to negation scope, while in control-finetuned models the changes are less positive and spread more broadly. Figure 4 shows that when BERT-base is fine-tuned on the negation task, the biggest gains in F_1 are on attention heads in layers 6 through 10, while no such pattern is visible when BERT-base is fine-tuned on the control task.

Do the rich heads get richer? Are attention heads that are already good predictors of negation

[3] Random restarts with the exact same hyperparameters can induce a surprising amount of instability in performance (Reimers and Gurevych, 2017; Devlin et al., 2019).

	Testing Task			
	Negation			Control
Training Task	P ± sd	R ± sd	F_1± sd	A ± sd
Negation	96.1± 1.3	89.7 ± 1.3	92.8 ± 1.1	-
Control	34.8 ± 0.4	36.1 ± 2.2	35.4 ± 1.2	95.9 ± 3.0

Table 2: Performance of fine-tuned BERT-base models on the supervised negation scope detection and control tasks in terms of precision (P), recall (R) and F_1 for negation scope and accuracy (A) for the control task. We report the average performance of 10 runs and 1 standard deviation.

Attention Head	Fine-Tune	P ± sd	R ± sd	F_1 ± sd
Average	None	49.5	5.2	9.0
Average	Control	48.6 ± 1.7	5.3 ± 0.2	9.0 ± 0.4
Average	Negation	52.2 ± 2.2	6.6 ± 0.8	11.1 ± 1.2
Best (8-4)	None	76.2	41.5	53.8
Best (8-4)	Control	65.0 ± 8.9	47.5 ± 11.7	53.1 ± 6.7
Best (8-4)	Negation	82.3 ± 4.1	58.6 ± 10.8	67.7 ± 7.9

Table 3: Performance of unsupervised BERT-base attention-based classifiers on the scope detection task in terms of precision (P), recall (R) and F_1 after the BERT model has been fine-tuned on different downstream tasks.

scope improve more after fine-tuning? That is, if an attention head has a high negation-scope prediction performance before fine-tuning, will it increase in performance more than other attention heads that had lower performance before fine-tuning? To test this, we measure the kendall rank correlation between an attention head's performance before fine-tuning on the downstream negation task, and its change in performance after fine-tuning. For the BERT-base model, most coefficients are very small and many of the runs show no significant correlation: the average τ coefficient for precision is -0.07 and only 3 out of 10 runs show a significant correlation, the average τ coefficient for recall is 0.10 and only 5 out of 10 runs show a significant correlation, and the τ coefficient for F_1 is 0.08 and only 5 out of 10 runs show a significant correlation. Table A4 in the Appendix shows that in other models the rich on average get poorer: we find weak negative correlations. This suggests fine-tuning, even on a relevant downstream task, does not focus on improving the attention heads that are already good at the problem.

Which layers improve the most? Are attention heads at certain layers more sensitive to fine-tuning than other layers? We measure the average performance gain for attention heads in each layer of BERT-base, and plot how these vary across the 10 runs in Figure 5. Figure A6 in the Appendix plot the same for the other models. After the model is fine-tuned on the negation task, we see that attention heads in mid-to-later layers (e.g., layers 6 through 10 in BERT-base) become more sensitive to negation scope. The models fine-tuned on the control task generally show smaller changes. The exception is BERT-large, whose pattern is very different, perhaps because it is the only model to have perfectly memorized the control task.

Is the change consistent across datasets? We have seen that fine-tuning on a downstream negation task increases the negation sensitivity broadly across the many attention heads. Do these changes truly represent a better understanding of the linguistic phenomenon of negation, or are they simply a form of better fitting the training data? If a more general understanding is being learned, when looking across several different types of negation problems, there should be greater consistency in which attention heads are paying attention to negation than in the pretrained model or control task.

We thus take models after fine-tuning on the ConanDoyle-neg downstream negation scope task, treat each of the attention heads as unsupervised negation-scope classifiers as in Section 4.1, and calculate performance on both the ConanDoyle-neg data (the same type of data as was used for

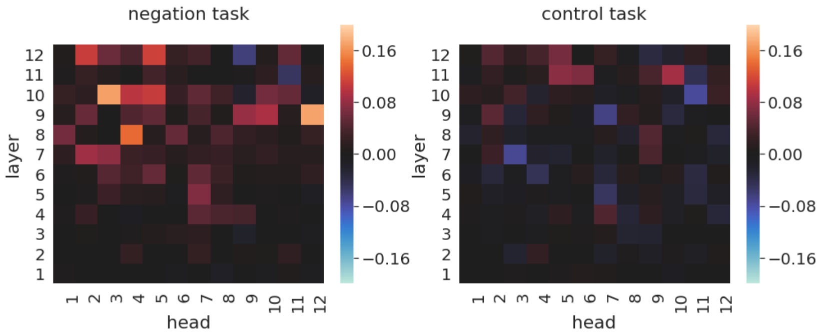


Figure 4: Change in F_1 for each attention head in BERT-base (averaged across 10 runs) before and after fine-tuning.

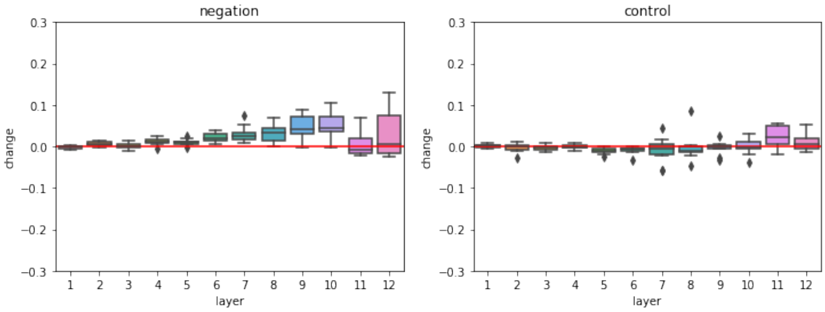


Figure 5: Average change in F_1 for the attention heads in each layer in BERT-base, repeated for 10 runs.

fine-tuning) and the SFU-review data (a different text genre and annotation scheme). We then run kendall rank correlation tests between the two sets of attention-head performances and report them in Table 4 for BERT-base and Table A5 in the Appendix for the other models. Fine-tuning BERT-base on the downstream negation task indeed yields more similar performance across datasets (0.516 F_1) than for the original model before fine-tuning (0.415 F_1) or the model fine-tuned on the downstream control task (0.409 F_1). A Wilcoxon test shows that the τ coefficients fine-tuned on the negation task are significantly higher compared to those fine-tuned on the control task ($p = 1.083 \times 10^{-5}$). RoBERTa-base patterns similarly. For BERT-large the negation-tuned models show a marginal consistency improvement over the pretrain and the attention head consistency in the negation-tuned RoBERTa-large models does not exceed that of the control-tuned ones.

6 Discussion

We have presented a methodology for looking for explanations of transformer models, where a hypothesized encoding of knowledge within the transformer is measured before and after fine-tuning and the changes are compared to those seen when fine-tuning on a control task. We considered a specific linguistic phenomenon, negation scope detection, proposed an intuitive way that attention may encode negation-scope (in-scope words pay attention to the negation cue), and applied our methodology to test whether the hypothesized encoding was indeed an explanation of the behavior of BERT and/or RoBERTa models. We found evidence that BERT-base and RoBERTa-base encode some negation knowledge in the proposed way as both average negation sensitivity and cross-dataset consistency improved over the pretrained model and the control task. Evidence for the large versions of the models was weaker, suggesting that they may be representing negation knowledge in other ways.

Other works have explored the effects of fine-tuning on attention without testing for specific linguistic knowledge. Serrano and Smith (2019), Jain and Wallace (2019) and Wiegreffe and Pinter (2019) found many redundancies in the attention of sequence-to-sequence models, suggesting that attention may encode knowledge in many ways. Kovaleva et al. (2019) found that removal of attention heads in transformers does not necessarily damage downstream performance. Our results suggest an explanation for this finding: knowledge sensitivity spreads broadly, so recovering from a small number of missing heads should be easy.

Htut et al. (2019) investigated the role of gram-

Fine-Tune	Precision		Recall		$\mathbf{F_1}$	
	mean τ ± sd	sig	mean τ ± sd	sig	mean τ ± sd	sig
Pretrain	0.440		0.418		0.415	
Control	0.438 ± 0.020	10/10	0.406 ± 0.034	10/10	0.409 ± 0.026	10/10
Negation	0.469 ± 0.025	10/10	0.519 ± 0.020	10/10	0.516 ± 0.020	10/10

Table 4: Kendall rank correlation (τ) between an attention head's performance on the ConanDoyle-neg dataset and its performance in the SFU-review dataset. For the fine-tuning settings, we report the average τ across 10 runs with 1 standard deviation, and the number of runs where there was a significant correlation.

matical relations in BERT's changes before and after fine-tuning. They found that long distance grammatical relations such as *advcl* and *csubj* improved greatly after finetuning on a semantically related task, but other relations did not. They included no control task and did not report changes for individual attention heads (only changes in the maximum performance) so their work inspires some questions: Do *advcl* and *csubj* improve more than expected by chance? For the other relations, does performance not improve because they are irrelevant? Or maybe performance of one of the non-maximal heads improved quite a bit, but not enough to exceed the maximal head? Applying our methodology for comparing against a control task and examining changes in individual heads could address these questions.

Other work has tested for specific linguistic knowledge in pretrained models, but not explored how the encoding of that knowledge changes during fine-tuning. For instance, Clark et al. (2019) identified several syntactic relationships that are encoded in an intuitive way: the dependent's primary attention is on its grammtical head. We argue that testing whether this hypothesized encoding of grammatical relations survives fine-tuning is critical if this is to be an explanation of how transformer models make predictions.

We found no past work that considered the cross-dataset consistency of attention. We believe measuring such consistency is important for differentiating between an attention head that learned to encode a linguistic phenomenon for a single dataset vs. an attention head that learned an encoding of the true linguistic phenomenon. For example, it could have been the case that fine-tuning improves sensitivity to negation in both datasets, but the improvements happen at different heads. We see this for example in BERT-large on the control task, where there is essentially zero consistency in which attention heads are active across the two datasets.

Some limitations of our current work suggest future research directions. First, we have focused on one interpretable way of encoding of negation scope knowledge but one can hypothesize many other ways. For instance, instead of assuming that all in-scope words directly pay attention to negation cue, it is possible that the head of in-token words are organized in a tree of attention that leads to the negation cue. We use a single nonlinguistic control task, but one could imagine exploring attention head changes in the face of a gradient of fine-tuning tasks that are more or less relevant to the linguistic phenomenon of interest. We also focus primarily on the attention mechanism, but it would be useful to explore the value vectors that transformers apply the attention to, since these form the outputs and are thus more directly tied to classification decisions.

7 Conclusion

In this paper, we propose a basic procedure and analysis methods that take a hypothesis of how a transformer-based model might encode a linguistic phenomenon, and test the validity of that hypothesis based on unsupervised probes, downstream control tasks, and measurement of cross-dataset consistency. We hypothesize an interpretable encoding of negation scope, where in-scope words attend to the negation cue, and find evidence of such an encoding in BERT-base and RoBERTa-base.

Acknowledgements

Thanks to the anonymous reviewers for their helpful suggestions. This work was supported in part by National Institutes of Health grant R01LM012918 from the National Library of Medicine (NLM). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

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

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. 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.

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.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do attention heads in bert track syntactic dependencies? *ArXiv*, abs/1911.12246.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3190–3195, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

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*.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2019. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *CoRR*, abs/1911.02969.

Roser Morante and Walter Daelemans. 2012. Conandoyle-neg: Annotation of negation in conan doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL* *https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Noah Weber, Leena Shekhar, and Niranjan Balasubramanian. 2018. The fine line between linguistic generalization and failure in Seq2Seq-attention models. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*, pages 24–27, New Orleans, Louisiana. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

A Appendix

The main text of the paper focused on the results for BERT-base. This appendix contains detailed results for all four models: BERT-base, RoBERTa-base, BERT-large, and RoBERTa-large.

Models		**P**	**R**	**F_1**
baseline	all in scope	34.0	**100.0**	50.7
baseline	average fixed offset	66.1	8.6	15.2
baseline	best fixed offset (-1)	83.5	11.6	20.4
BERT-base attention	average head	49.5	5.2	9.0
BERT-base attention	best head (8-4)	76.2	41.5	**53.8**
BERT-large attention	average head	45.4	3.3	5.9
BERT-large attention	best head (14-4)	74.9	28.3	41.0
RoBERTa-base attention	average head	56.0	6.9	12.1
RoBERTa-base attention	best head (9-12)	**92.9**	19.1	31.1
RoBERTa-large attention	average head	50.2	5.3	9.4
RoBERTa-large attention	best head (15-15)	66.7	21.3	32.3

Table A1: Performance of unsupervised attention-based classifiers and baselines on the negation scope detection task in terms of precision (P), recall (R) and F_1. The best fixed offset and attention head according to their F_1 score are reported. Finding: *all models have attention heads that know more about negation than the simple baselines.*

	Testing Task			
	Negation			**Control**
Training Task	**P ± sd**	**R ± sd**	**F_1± sd**	**A ± sd**
BERT-base Negation	96.1± 1.3	89.7 ± 1.3	92.8 ± 1.1	-
BERT-base Control	34.8 ± 0.4	36.1 ± 2.2	35.4 ± 1.2	95.9 ± 3.0
BERT-large Negation	97.3± 0.9	93.0 ± 1.1	95.1 ± 0.6	-
BERT-large Control	39.2 ± 0.9	33.1 ± 1.0	35.9 ± 0.6	100.0 ± 0.0
RoBERTa-base Negation	97.2± 0.9	92.9 ± 1.0	95.9 ± 0.3	-
RoBERTa-base Control	43.4 ± 0.7	45.4 ± 1.2	44.4 ± 0.7	98.3 ± 0.4
RoBERTa-large Negation	97.9± 0.9	93.5 ± 1.2	95.7 ± 0.9	-
RoBERTa-large Control	44.1 ± 0.6	45.2 ± 1.8	44.6 ± 1.0	97.9 ± 2.2

Table A2: Performance of fine-tuned models on the supervised negation scope detection and control tasks in terms of precision (P), recall (R) and F_1 for negation scope and accuracy (A) for the control task. We report the average performance of 10 runs and 1 standard deviation. Finding: *All models successfully learned both supervised tasks.*

Attention Head	**Fine-Tune**	**P ± sd**	**R ± sd**	**F_1 ± sd**
BERT-base Average	None	49.5	5.2	9.0
BERT-base Average	Control	48.6 ± 1.7	5.3 ± 0.2	9.0 ± 0.4
BERT-base Average	Negation	52.2 ± 2.2	6.6 ± 0.8	**11.1 ± 1.2**
BERT-large Average	None	45.4	3.3	5.9
BERT-large Average	Control	44.8 ± 0.3	4.6 ± 0.1	8.3 ± 0.2
BERT-large Average	Negation	46.0 ± 3.7	4.8 ± 1.5	8.0 ± 2.3
RoBERTa-base Average	None	56.0	6.9	12.1
RoBERTa-base Average	Control	53.7 ± 1.7	7.0 ± 0.3	12.0 ± 0.5
RoBERTa-base Average	Negation	55.5 ± 1.9	7.9 ± 0.9	**13.4 ± 1.4**
RoBERTa-large Average	None	50.2	5.3	9.4
RoBERTa-large Average	Control	48.2 ± 2.2	7.0 ± 1.0	11.5 ± 1.3
RoBERTa-large Average	Negation	54.2 ± 3.4	8.0 ± 1.8	13.2 ± 2.7

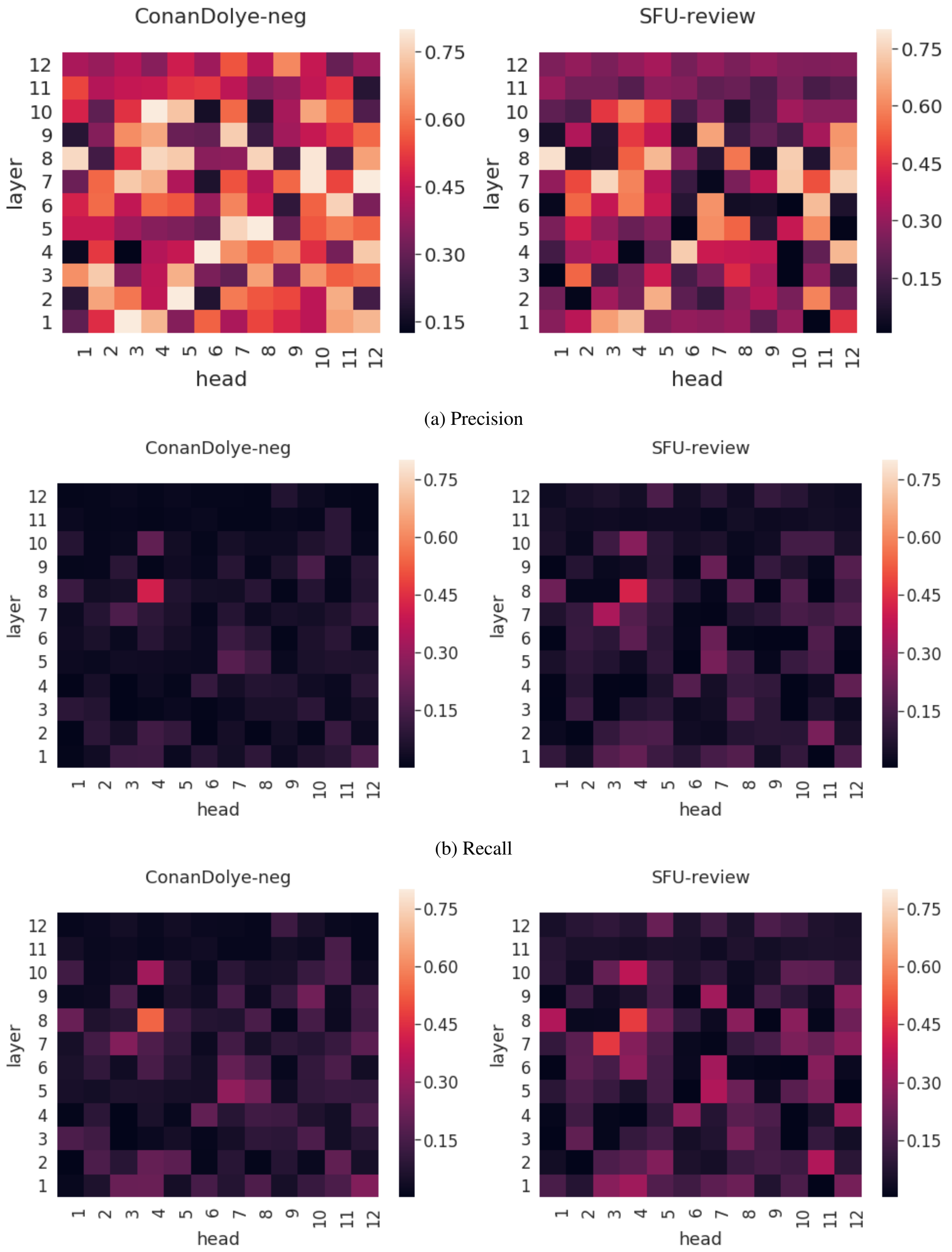
Table A3: Performance of unsupervised attention-based classifiers on the scope detection task in terms of precision (P), recall (R) and F_1 after models have been fine-tuned on different downstream tasks. All models fine-tuned on negation-scope significantly outperformed their pretrained counterparts in F_1, but only two (in bold) significantly outperformed the controls. Finding: *In BERT-base and RoBERTa-base, attention can be a explanation of negation.*

Negation change	**Precision**			**Recall**			**$\mathbf{F_1}$**		
	τ	pos/neg	sig	τ	pos/neg	sig	τ	pos/neg	sig
BERT-base	-0.065	0/3	3/10	0.096	5/0	5/10	0.085	5/0	5/10
BERT-large	-0.098	2/5	7/10	-0.132	0/7	7/10	-0.132	0/8	8/10
RoBERTa-base	-0.134	0/7	7/10	-0.107	0/5	5/10	-0.113	0/6	6/10
RoBERTa-large	-0.155	0/8	8/10	-0.142	0/8	8/10	-0.144	0/8	8/10

Table A4: Kendall rank correlation (τ) between the change of an attention head after fine-tuning on the negation task and its performance in the pretrained model. We report the average τ across 10 runs, the number of runs where there was a significant correlation, and the direction (positive or negative) of the significant correlations. Finding: *The rich do not get richer: attention heads that had the top F_1s in the pretrained model do not have the top-ranked improvements after fine-tuning on negation scope.*

Consistency	**Precision**		**Recall**		**$\mathbf{F_1}$**	
	mean τ ± sd	sig	mean τ ± sd	sig	mean τ ± sd	sig
BERT-base Pretrain	0.440		0.418		0.415	
BERT-base Control	0.438 ± 0.020	10/10	0.406 ± 0.034	10/10	0.409 ± 0.026	10/10
BERT-base Negation	0.469 ± 0.025	10/10	0.519 ± 0.020	10/10	**0.516 ± 0.020**	10/10
BERT-large Pretrain	0.295		0.487		0.482	
BERT-large Control	0.0005 ± 0.057	3/10	0.007 ± 0.039	1/10	0.006 ± 0.039	1/10
BERT-large Negation	0.474 ± 0.038	10/10	0.523 ± 0.082	10/10	0.530 ± 0.066	10/10
RoBERTa-base Pretrain	0.438		0.472		0.471	
RoBERTa-base Control	0.456 ± 0.022	10/10	0.502 ± 0.023	10/10	0.487 ± 0.021	10/10
RoBERTa-base Negation	0.521 ± 0.024	10/10	0.538 ± 0.033	10/10	**0.531 ± 0.033**	10/10
RoBERTa-large Pretrain	0.377		0.504		0.493	
RoBERTa-large Control	0.389 ± 0.031	10/10	0.579 ± 0.029	10/10	0.561 ± 0.026	10/10
RoBERTa-large Negation	0.516 ± 0.037	10/10	0.593 ± 0.056	10/10	0.584 ± 0.054	10/10

Table A5: Kendall rank correlation (τ) between an attention head's performance on the ConanDoyle-neg dataset and its performance in the SFU-review dataset. For the fine-tuning settings, we report the average τ across 10 runs with 1 standard deviation, and the number of runs where there was a significant correlation. Only in two models (in bold) was the correlation for the negation-trained model significantly higher than the correlation for both the pretrained model and the control model. Finding: *In BERT-base and RoBERTa-base, attention performance finetuned on a negation task is more consistent scope across different domains and annotation schemes.*

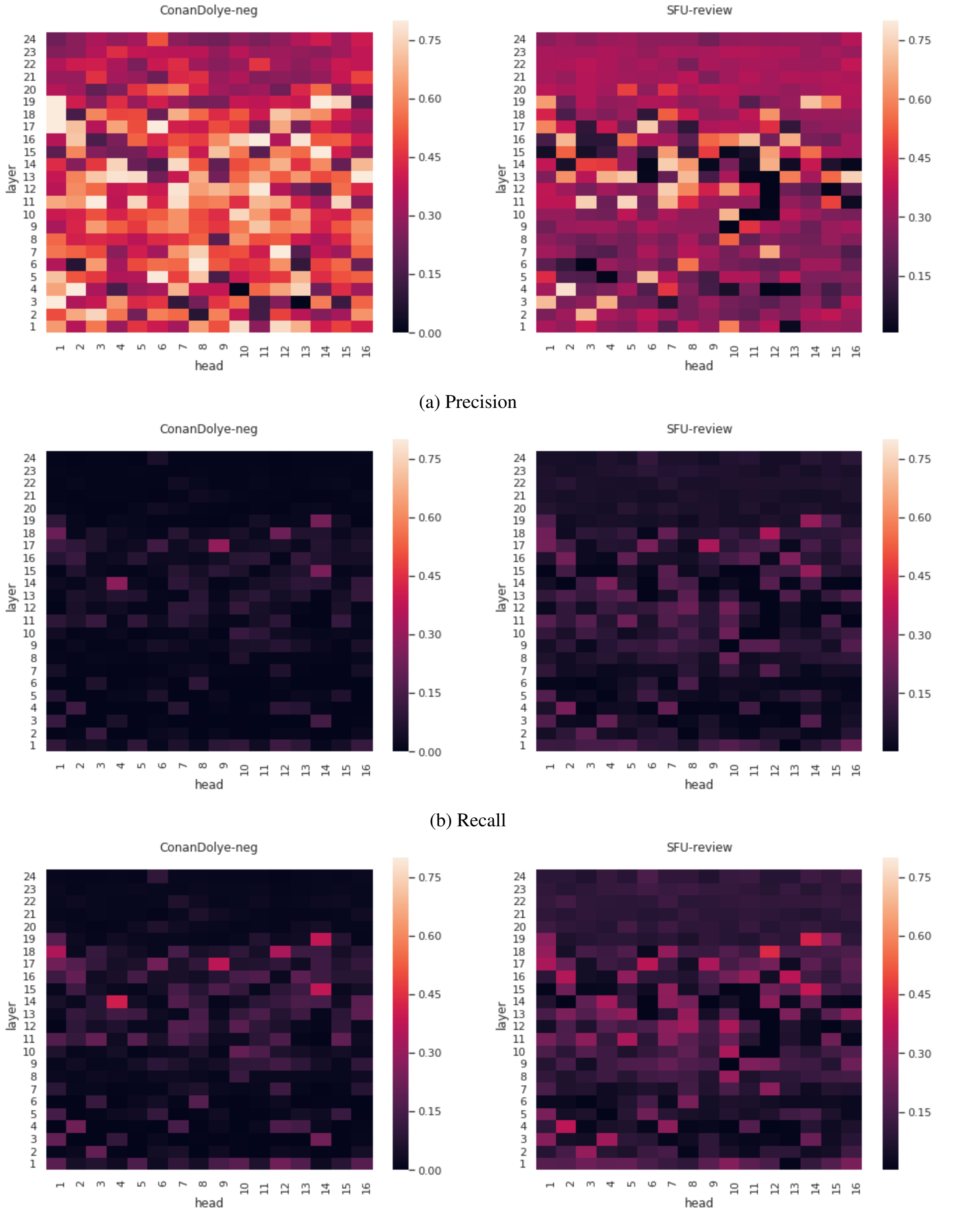


(a) Precision

(b) Recall

(c) F_1

Figure A1: The heatmap of unsupervised negation-scope classification performance for BERT-base's 12 layers x 12 heads across two different datasets. The consistency (measure by kendall rank correlation) between the two datasets for precision, recall and F_1 are 0.440, 0.418 and 0.415 respectively.

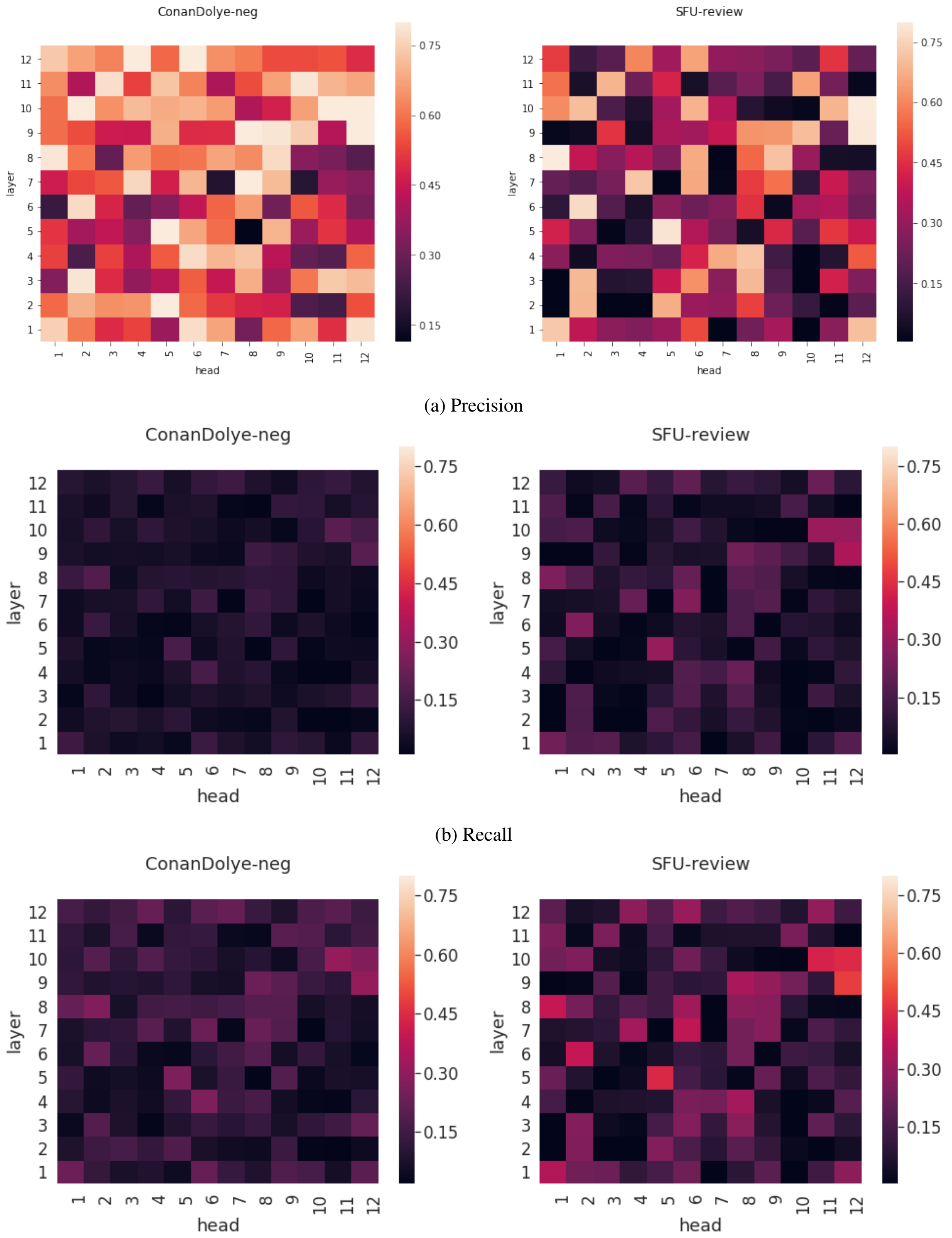


(a) Precision

(b) Recall

(c) F_1

Figure A2: The heatmap of unsupervised negation-scope classification performance for BERT-large's 24 layers x 16 heads across two different datasets. The consistency (measure by kendall rank correlation) between the two datasets for precision, recall and F_1 are 0.295, 0.487 and 0.482 respectively.

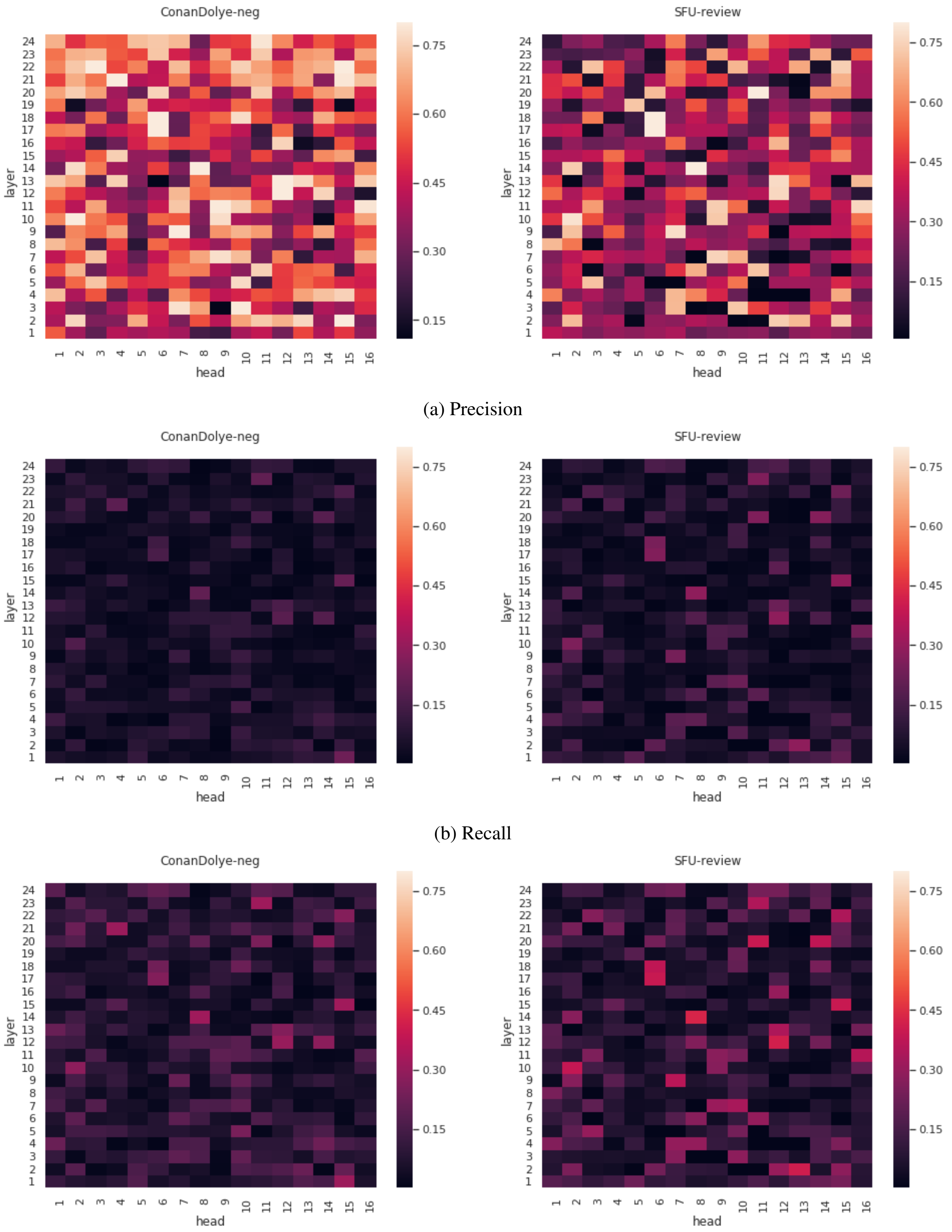


(a) Precision

(b) Recall

(c) F_1

Figure A3: The heatmap of unsupervised negation-scope classification performance for RoBERTa-base's 12 layers x 12 heads across two different datasets. The consistency (measure by kendall rank correlation) between the two datasets for precision, recall and F_1 are 0.438, 0.472 and 0.471 respectively.



(a) Precision

(b) Recall

(c) F_1

Figure A4: The heatmap of unsupervised negation-scope classification performance for RoBERTa-large's 24 layers x 16 heads across two different datasets. The consistency (measure by kendall rank correlation) between the two datasets for precision, recall and F_1 are 0.377, 0.504 and 0.493 respectively.

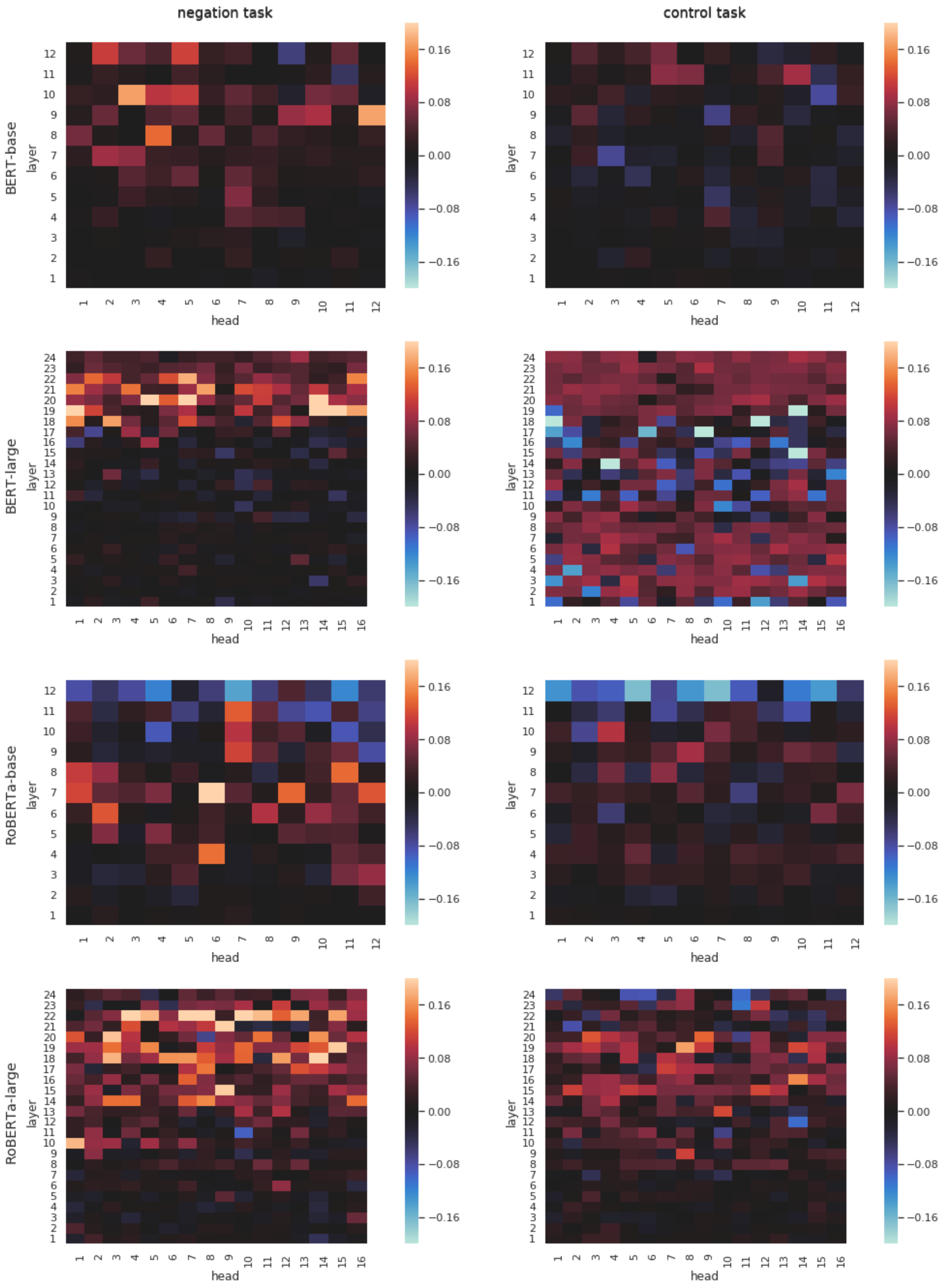


Figure A5: Change in F_1 for each attention head (averaged across 10 runs) before and after fine-tuning.

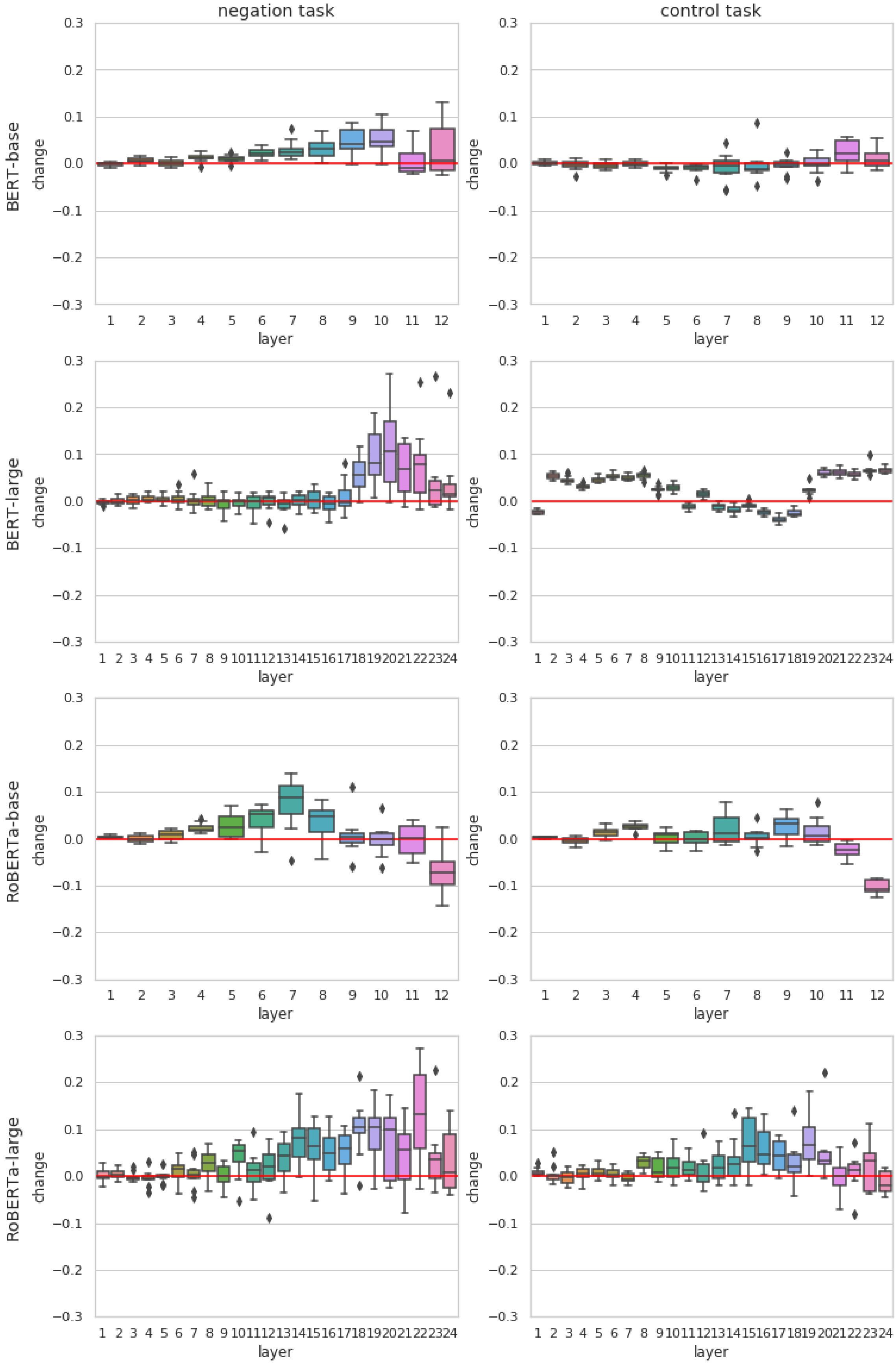


Figure A6: Change in F_1 for each attention head (averaged across 10 runs) before and after fine-tuning.