On the Relationship between Skill Neurons and Robustness in Prompt Tuning

Leon Ackermann, Xenia Ohmer
University of Osnabrueck
Osnabrueck, Germany
{lackermann, xenia.ohmer}@uni-osnabrueck.de

Abstract

Prompt Tuning is a popular parameter-efficient finetuning method for pre-trained large language models (PLMs). Based on experiments with RoBERTa, it has been suggested that Prompt Tuning activates specific neurons in the transformer's feed-forward networks, that are highly predictive and selective for the given task. In this paper, we study the robustness of Prompt Tuning in relation to these "skill neurons", using RoBERTa and T5. We show that prompts tuned for a specific task are transferable to tasks of the same type but are not very robust to adversarial data. While prompts tuned for RoBERTa yield below-chance performance on adversarial data, prompts tuned for T5 are slightly more robust and retain above-chance performance in two out of three cases. At the same time, we replicate the finding that skill neurons exist in RoBERTa and further show that skill neurons also exist in T5. Interestingly, the skill neurons of T5 determined on non-adversarial data are also among the most predictive neurons on the adversarial data, which is not the case for RoBERTa. We conclude that higher adversarial robustness may be related to a model's ability to consistently activate the relevant skill neurons on adversarial data.

1. Introduction

Pretrained large language models (PLMs) comprise increasingly large numbers of parameters. For example, while Roberta-Large has "only" 355 million parameters (Liu et al., 2019), T5-XXL has 11 billion parameters (Raffel et al., 2020), and LLama-2 up to 70 billion (Touvron et al., 2023). Finetuning such models for downstream tasks is extremely expensive both in terms of computation and storage. Parameter-efficient finetuning (PEFT) methods have been developed as a solution to this problem. These methods adapt PLMs to downstream tasks by finetuning only a small set of (additional) parameters.

Next to Low Rank Adaptation (LoRA) (Hu et al., 2022), Prefix Tuning (Li and Liang, 2021), and P-Tuning (Liu et al., 2021), Prompt Tuning (Lester et al., 2021) is one of the state-of-the-art PEFT methods for PLMs (see e.g., Mangrulkar et al., 2022). In Prompt Tuning, prompt tokens are prepended to the model input *in the embedding space*, and only these prepended tokens are learned during finetuning while the actual model parameters are frozen. In experiments with various T5 model sizes, Lester et al. (2021) showed that Prompt Tuning performance is on par with conventional finetuning for larger models. The authors further demonstrated that—next to reducing computational and storage requirements—Prompt Tuning has the advantage of being more robust to domain shifts, as adapting fewer parameters reduces the risk of overfitting.

To understand how Prompt Tuning works, researchers have started looking at its effects on PLM activations. In general, it is known that activations in the feed-forward networks (FFNs) of transformers (Vaswani et al., 2017) can specialize to encode specific knowledge (Dai et al., 2022) or concepts (Suau et al., 2020). For Prompt Tuning, it has been shown that the overlap between the FFN neurons activated by different prompts can predict prompt transferability (Su et al., 2022). More recently, Wang et al. (2022) showed that the activations of some FFN neurons are highly predictive of the task labels after Prompt Tuning. Analyses by the authors indicate that these "skill neurons" are task-specific, essential for task performance, and likely already generated during pretraining.

Our work extends ongoing research on robustness and skill neurons and establishes a connection between these two aspects. We run experiments with RoBERTa (Liu et al., 2019) and T5 (Raffel et al., 2020) to capture differences between encoder-only and encoder-decoder architectures and utilize various benchmark datasets that cover a broad spectrum of NLP tasks. For each dataset and model, we tune several prompts (using different seeds) and identify the associated skill neurons. Our main contributions are:

1. Consistent with previous research, we find that tuned prompts can be transferred to other datasets, including datasets involving domain shifts, provided these datasets pertain to the same type of task.

2. Using `AdversarialGLUE` (Wang et al., 2021), we show that Prompt Tuning is not robust to adversarial data.

3. Wang et al.'s (2022) skill neuron analysis was limited to RoBERTa. We replicate their findings and further identify skill neurons in (the encoder of) T5.

4. We establish a potential link between adversarial robustness and skill neurons. T5 exhibits greater robustness to adversarial data than RoBERTa. While T5's skill neurons on adversarial data are relatively consistent with its skill neurons on the corresponding non-adversarial data, this is not the case for RoBERTa.

Our code is publicly available at https://github.com/LeonAckermann/robust-neurons.

In conclusion, our study offers additional evidence supporting the existence of skill neurons in PLMs. Although Prompt Tuning typically lacks adversarial robustness, our findings indicate that a model's robustness against adversarial attacks may depend on its ability to maintain task-relevant skill neurons on adversarial data. Since skill neurons already emerge during pretraining and are thus independent of Prompt Tuning, our results are of general relevance for PLMs.

2. Related Work

Prompt Tuning is a PEFT method. In line with other work, we examine to what extent tuned prompts generalize to other (non-adversarial) datasets of the same task as well as to adversarial datasets. Furthermore, we use the tuned prompts to identify whether specific neurons in the models encode specific skills, by analyzing the models' FFN activations.

Parameter-efficient Finetuning. All PEFT methods train only a few (additional) parameters to adapt PLMs to a certain task. They can be divided into *adapter-based* and *prompt-based* methods. Adapter-based methods insert small neural modules (adapters) into the transformer layers, which are tuned to the task. They were first used in Computer Vision (Rebuffi et al., 2017) and then also became popular in NLP (Houlsby et al., 2019), especially in the form of low-rank adapter tuning (Hu et al., 2022). Instead of inserting additional modules, prompt-based methods extend the original inputs with additional parameters. An important example is Prefix-Tuning (Li and Liang, 2021), which prepends virtual tokens to each layer in the encoder stack, including the input (embedding) layer. Prompt Tuning (Lester et al., 2021) further simplifies that method by only adding tokens to the input layer. Compared to adapter-based methods, prompt-based methods tend to converge more slowly, and perform worse with smaller datasets

and models (e.g. Ding et al., 2023). On the other hand, they are easy to implement and require even fewer changes to the model.

Prompt Transferability. An additional advantage of Prompt Tuning is that tuned prompts may be reusable. Lester et al. (2021) showed that Prompt Tuning is robust to domain shifts by evaluating prompts, tuned on specific question answering (QA) and paraphrase identification datasets, on other QA and paraphrase detection datasets. Prompt Tuning was more robust than traditional finetuning, with an increasing advantage for larger domain shifts. Using a larger set of tasks, Su et al. (2022) confirmed that prompts learned with Prompt Tuning can be transferred effectively to similar tasks (within models), and further showed that they can also be transferred between models (within tasks). The authors examined various prompt similarity metrics as transferability indicators. Especially the overlapping rate of activated neurons in the transformer FFN layers was indicative of prompt transferability—more so than metrics based on prompt similarity in the embedding space.

Adversarial Robustness. We are interested in whether Prompt Tuning is also robust to adversarial examples. Adversarial examples are datapoints that are misclassified even though they are only slightly—often imperceptibly—different from correctly classified examples. Szegedy et al. (2014) discovered that several machine learning models, including neural networks, are vulnerable to such examples and that the same examples tend to be adversarial for different models. Several studies have shown that PLMs are affected as well (e.g., Garg and Ramakrishnan, 2020; Wang et al., 2021; Jin et al., 2020; Zhang et al., 2021; Li et al., 2020). In the text domain, adversarial examples are generated through perturbations that preserve semantic meaning. Perturbations can be applied at the word level (e.g. synonym replacement, typos) or the sentence level (e.g. paraphrasing, adding distracting text). They can be generated automatically or crafted by humans (e.g. Naik et al., 2018; Ribeiro et al., 2020; Nie et al., 2020; Jia and Liang, 2017). A total of 14 perturbation methods were applied to the multitask benchmark GLUE (Wang et al., 2018) to generate AdvGLUE (Wang et al., 2021).

Analyzing FFN Activations in PLMs. There is a wide interest in understanding the inner workings of PLMs and transformer FFN layers are increasingly studied in this context. In particular, evidence is growing that FFN layers serve as memories that store factual and linguistic knowledge. For example, Geva et al. (2021) found that FFNs function

similarly to key-value memories, in that they detect certain text input patterns and map them to an output distribution over tokens; and that the detected patterns contain increasingly semantic information when progressing through the transformer layers. Furthermore, knowledge encoded in the FFNs seems to be highly localized. Specific neurons seem to encode specific concepts or pieces of information, and by modifying the activations of these neurons, the model's expression of the corresponding knowledge can be regulated (Dai et al., 2022; Suau et al., 2020; Yao et al., 2022). Wang et al. (2022) analyzed the activations in FFN layers when prepending task-specific continuous prompts to the input (learned through Prompt Tuning). Their findings suggest that certain neurons encode task-specific skills and that these skills, like factual knowledge, are already acquired during model pretraining.

3. Methods

This section introduces Prompt Tuning more formally (Section 3.1) and describes how the predictivities of individual neurons can be determined using tuned prompts (Section 3.2). The neurons with the highest predictivity for a specific task are considered the model's skill neurons for that task.

3.1. Prompt Tuning

The model embeds input sequence $X_{orig} = [\text{token } 1, \text{token } 2, \dots, \text{token } s]$ into $\mathbf{X} \in \mathbb{R}^{s \times h}$, where h is the embedding dimension. Prompt Tuning prepends additional prompt tokens $\mathbf{P} = [\mathbf{p}_1, \dots, \mathbf{p}_p], \mathbf{p}_i \in \mathbb{R}^h$ to that input in the embedding space, such that the new model input is $(\mathbf{P}, \mathbf{X}) = [\mathbf{p}_1, \dots, \mathbf{p}_p, \mathbf{x}_1, \dots, \mathbf{x}_s]$, with $(\mathbf{P}, \mathbf{X}) \in \mathbb{R}^{(p+s) \times h}$. The continuous prompt tokens in the embedding space are treated as free model parameters and their values are learned via backpropagation during the training phase. All other model parameters are frozen. Thus, Prompt Tuning does not change any of the model's original weights, and only a few new parameters ($p \times h$) are learned per task.

3.2. Skill Neurons

Based on the method by Wang et al. (2022), skill neurons are identified as neurons in the FFNs of a transformer model whose activations are highly predictive of the task labels. Skill neurons are defined in relation to task-specific prompts, such as the ones generated through Prompt Tuning. They are calculated in three steps: 1) The *baseline activation* for each neuron is calculated. 2) The *predictivity* of each neuron is calculated, and 3) The consistently most predictive neurons are identified as *skill neurons*. In the following, we describe how the skill

neurons of one FFN (one layer) are determined using Prompt Tuning. The method is described for binary classification tasks, which we use in our analyses.

Notation. An FFN with activation function f can formally be defined as

$$\text{FFN}(\mathbf{x}) = f\left(\mathbf{x}\mathbf{K}^\top + \mathbf{b}_1\right)\mathbf{V} + \mathbf{b}_2 , \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^h$ is the embedding of an input token, $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{f \times h}$ are weight matrices, and $\mathbf{b}_1, \mathbf{b}_2$ are biases. Given that the first linear transformation produces the activations $\mathbf{a} = f\left(\mathbf{x}\mathbf{K}^\top + \mathbf{b}_1\right)$, a_i is considered the activation of the i-th neuron on input $\mathbf{x}$.

Baseline Activations. Let the training set be defined as $D_{\text{train}} = \left\{\left(\mathbf{X}_1, y_1\right), \left(\mathbf{X}_2, y_2\right), \dots, \left(\mathbf{X}_{|D|}, y_{|D|}\right)\right\}$, with $\mathbf{X}_i \in \mathbb{R}^{s \times h}$ (where s is the input sequence length), and $y_i \in \{0, 1\}$. Let $\mathbf{P}$ be the task prompt with $\mathbf{P} = [\mathbf{p}_1, \dots, \mathbf{p}_p], \mathbf{p}_i \in \mathbb{R}^h$. The baseline activation $a_{\text{bsl}}(\mathcal{N}, \mathbf{p}_i) \in \mathbb{R}$ is defined as the average activation of neuron $\mathcal{N}$ for prompt token $\mathbf{p}_i$ across the training data. Let $a(\mathcal{N}, \mathbf{t}, \mathbf{X}_i)$ be the activation of neuron $\mathcal{N}$ for token embedding $\mathbf{t}$ given input $\mathbf{X}_i$. Then

$$a_{\text{bsl}}(\mathcal{N}, \mathbf{p}_i) = \frac{1}{|D_{\text{train}}|} \sum_{\mathbf{X}_i \in D_{\text{train}}} a\left(\mathcal{N}, \mathbf{p}_i, (\mathbf{P}, \mathbf{X}_i)\right) . \quad (2)$$

Predictivities. The accuracy of neuron $\mathcal{N}$ is calculated over the validation set D_{dev} with respect to the baseline activations calculated on the training set as

$$\text{Acc}(\mathcal{N}, \mathbf{p}_i) =$$
$$\frac{\sum_{(\mathbf{X}_i, y_i) \in D_{\text{dev}}} \mathbf{1}_{[\mathbf{1}_{[a(\mathcal{N}, \mathbf{p}_i, (\mathbf{P}, \mathbf{X}_i)) > a_{\text{bsl}}(\mathcal{N}, \mathbf{p}_i)]} = y_i]}}{|D_{\text{dev}}|} , \quad (3)$$

where $\mathbf{1}_{[\text{condition}]} \in \{0, 1\}$ is the indicator function. In other words, the neuron's accuracy describes how often (on average) activations above or below the baseline activation co-occur with a positive or a zero label, respectively. Finally, to account for the fact that inhibitory neurons may also encode skills, the predictivity per neuron and prompt token is calculated as

$$\text{Pred}(\mathcal{N}, \mathbf{p}_i) = \max\left(\text{Acc}(\mathcal{N}, \mathbf{p}_i), 1 - \text{Acc}(\mathcal{N}, \mathbf{p}_i)\right) . \quad (4)$$

Skill Neurons. Given that a set of k continuous prompts are trained $\mathcal{P} = \{\mathbf{P}_1, \dots, \mathbf{P}_k\}$ (with different seeds), the final predictivity of each neuron is

given by

$$\text{Pred}(\mathcal{N}) = \frac{1}{k} \sum_{\mathbf{P}_i \in \mathcal{P}} \max_{\mathbf{p}_j \in \mathbf{P}_i} \text{Pred}(\mathcal{N}, \mathbf{p}_j) . \quad (5)$$

When sorting the neurons in the model based on their predictivity, the most predictive neurons are considered the model's "skill neurons" for the given task.[1]

4. Experiments

Models. We run our experiments with RoBERTa-base (125M parameters) (Liu et al., 2019) and T5-base (223M parameters) (Raffel et al., 2020).

Tasks. We tune prompts for various types of binary classification tasks: (1) paraphrase detection, including QQP (Wang et al., 2018) and MRPC (Dolan and Brockett, 2005); (2) sentiment analysis, including Movie Rationales (Zaidan et al., 2008), SST2 (Socher et al., 2013), and IMDB (Maas et al., 2011); (3) ethical judgment, including Ethics-Deontology and Ethics-Justice (Hendrycks et al., 2020), and (4) natural language inference (NLI), including QNLI (Wang et al., 2018). We had also planned to include RTE (and AdvRTE) but performance after Prompt Tuning was poor, with accuracies between 55%–60% at a 50% chance level.

Importantly, datasets belonging to the same task cover domain and format shifts. The paraphrase detection tasks consist of question pairs from *Quora* (QQP) and sentence pairs from Newswire articles (MRPC). The sentiment analysis tasks are based on movie reviews from different websites and sometimes contain the full review (IMDB) and sometimes single sentences (SST2). The two ethics datasets are crowdsourced and focus on different types of ethical judgments, related to justice (Ethics-Justice) or deontology (Ethics-Deontology).

To test adversarial robustness we use AdvQQP, AdvQNLI, and AdvSST2 from AdvGLUE (Wang et al., 2021). We work with the validation sets of the adversarial tasks since the submission format for evaluation on the test sets does not allow for a skill-neuron analysis.

Prompt Tuning. We build on the code by Su et al. (2022) and use the same parameters for Prompt Tuning. In particular, the learned prompts consist of 100 (continuous) tokens. Their repository (https://github.com/thunlp/Prompt-Transferability/) includes one tuned prompt for each of the (non-adversarial) datasets that we use. We train four additional prompts (with different seeds) per dataset, resulting in a total of five prompts per dataset.

Skill Neurons. We calculate the neuron predictivities (Equation 5) for all non-adversarial datasets following the method described in Section 3.2. We use the baseline activations from the non-adversarial datasets to calculate the neuron predictivities on the corresponding adversarial datasets. All analyses involving neuron predictivities are conducted simultaneously for each layer in the model, or each layer in the encoder model in the case of T5. The calculation of skill neurons relies on neuron activations for specific prompt tokens, which can be extracted from the encoder but not the decoder.

5. Results

5.1. Prompt Tuning and Robustness

Prompt Tuning. We report mean accuracies and standard deviations across the five random seeds in Table 1. Both the accuracies and the observed variations between seeds are in line with the results from other Prompt Tuning experiments (e.g. Lester et al., 2021; Su et al., 2022). Performance is lowest on the ethical judgment tasks, with high accuracies on at least one dataset from the other tasks. Overall, the performance of the two models is similar, with a slight advantage for RoBERTa on ethical judgment and sentiment analysis, and a slight advantage for T5 on paraphrase detection and NLI.

Dataset	RoBERTa	T5
ethicsdeontology	69.9 ± 2.0	66.3 ± 1.6
ethicsjustice	65.4 ± 1.6	59.1 ± 2.9
MRPC	74.8 ± 5.9	77.5 ± 2.6
QQP	$87.1 \pm 0,2$	88.7 ± 1.1
AdvQQP	37.2 ± 4.1	59.2 ± 8.0
QNLI	90.4 ± 0.2	92.4 ± 0.2
AdvQNLI	45.1 ± 3.5	60.1 ± 3.1
IMDB	90.4 ± 0.3	88.2 ± 0.2
movierationales	74.1 ± 2.4	75.2 ± 1.4
SST2	98.7 ± 2.6	94.0 ± 0.4
AdvSST2	45.3 ± 4.5	45.4 ± 3.3

Table 1: Mean and standard deviation of the models' accuracy after Prompt Tuning.

Robustness. We analyze two different kinds of robustness: adversarial robustness and transfer-

---

[1]We never determine a fixed set of skill neurons. Our analyses either involve *all* predictivities, or we modify the activations of the top $k\%$ predictive neurons for different values of k.

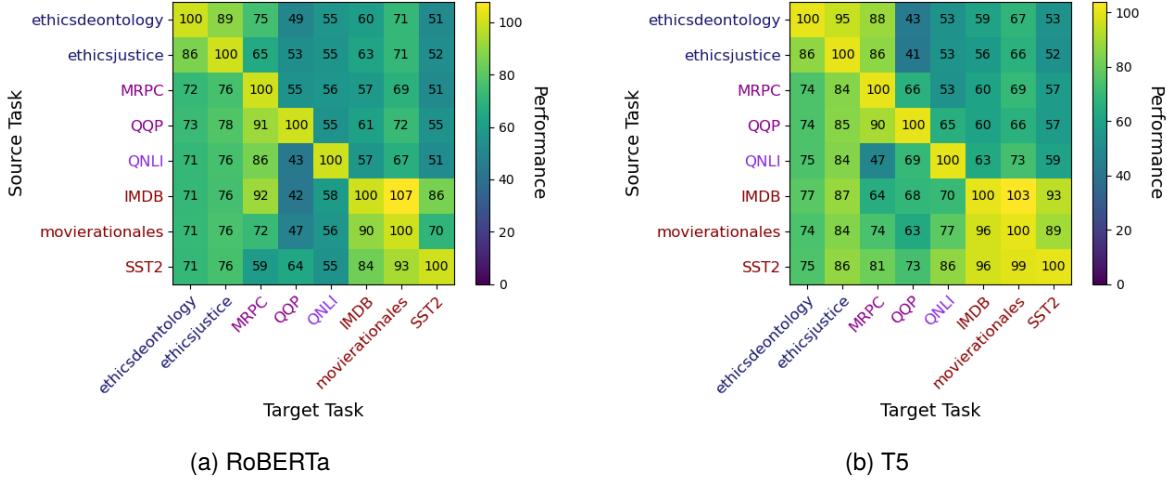|	(a) RoBERTa	(b) T5

Figure 1: Prompt transferability. We calculate the accuracy when using the prompt for the source task on the target task divided by the accuracy when using the prompt for the target task on the target task for each seed, and report the average across seeds.
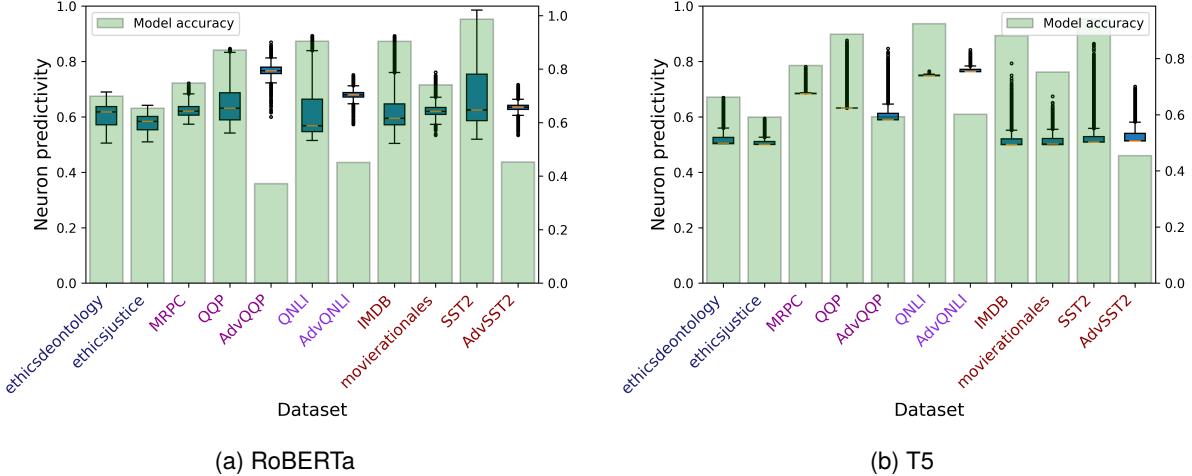

	(a) RoBERTa	(b) T5

Figure 2: Distribution of neuron predictivities (box plots) on top of model accuracy (bar plots).

ability. Table 1 shows the models' accuracy on the adversarial datasets, evaluated with the prompts of their non-adversarial counterparts. The accuracies drop significantly. For RoBERTa, they are consistently below chance performance. The score is especially low for AdvQQP, which is unbalanced (41% versus 59%)—unlike AdvQNLI and AdvSST2. T5 is somewhat more robust, with below chance performance on AdvSST2 but around 60% accuracy on the other two adversarial datasets.[2] Figure 1 shows the relative task accuracies when transferring a continuous prompt from a source task to a target task (see Appendix A for absolute values). In line with earlier findings, the prompts tend to be highly transferable to datasets belonging to the same type of task (Lester et al., 2021; Su et al.,

2022). In conclusion, Prompt Tuning is robust to data changes, including domain shifts (within the same type of task), but not to adversarial data.

5.2. Skill Neurons

Following a similar procedure to Wang et al. (2022), we test for the existence of skill neurons by calculating the neuron predictivities (Equation 5) and making sure that the most predictive neurons are *highly predictive*, *task-specific*, and *important* for solving the task.

High Predictivity. RoBERTa and T5 have only a few highly predictive neurons for each dataset. This observation is in line with the initial findings presented by Wang et al. (2022). Consistent with other studies that analyzed FFN activations across layers (Geva et al., 2021; Dai et al., 2022), neurons

---

[2]F1 scores on AdvQQP range between 36.4–44.4% for RoBERTa, with an average of 39.4%; and between 38.9–59.7% for T5, with an average of 48.2%.
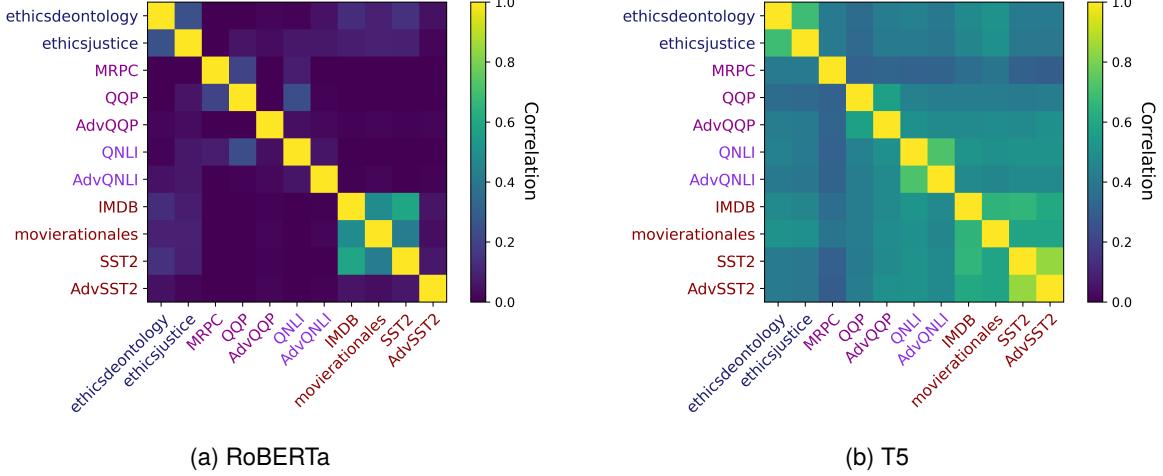
(a) RoBERTa   (b) T5

Figure 3: Spearman rank correlation between the neuron predictivities for different datasets.

with high predictivities tend to be located in the upper layers. The predictivities of the most predictive neurons of RoBERTa largely correspond to the model's accuracy for the non-adversarial datasets (see Figure 2a). The most predictive neurons of T5 reach or fall (slightly) short of the model's accuracy (see Figure 2b). For T5, especially in the cases where neuron predictivity is lower than model accuracy, more predictive neurons can probably be found in the decoder. Regarding the adversarial datasets, the predictivities of almost all neurons exceed the models' accuracy. Possible reasons are discussed in section 5.3.

Task-specificity. We calculate the Spearman rank correlation between the neuron predictivities for all datasets (see Figure 3). The correlations are calculated per layer, based on the neuron predictivities when evaluated on the corresponding dataset, and then averaged across layers. High values within but not between different types of tasks for RoBERTa and T5 indicate a high task-specificity of the models' skill neurons. Notably, the correlations are generally higher for T5 which might be due to its sparse activations (Li et al., 2022). If there is a large number of consistently inactive neurons, they will always rank lower than active neurons and thus lead to a net positive correlation.

Importance. To ensure that the most predictive neurons are in fact essential for performing the task, we compare the decrease in accuracy when suppressing 1-15% of the models' most predictive neurons versus the same number of random neurons. Wang et al. (2022) perturb the neurons with Gaussian noise instead of suppressing them completely. Given that the activations in T5 are much higher than those in RoBERTa, using the same amount of noise (same standard deviation) will have differ-

ent effects on the models. Instead, we decided to suppress the neurons, which has also been done in other work (e.g. Dai et al., 2022). Neurons are suppressed by setting their activations to zero. For both models, the accuracy drops much more when suppressing skill neurons compared to random neurons, highlighting the importance of skill neurons for the models' task performance (see the `IMBD` example in Figure 4 and results for all datasets in Appendix D). In general, suppressing random neurons has a larger impact on RoBERTa than T5, which we again attribute to T5's sparse activations: A significant proportion of the randomly selected neurons would not have been active anyway.
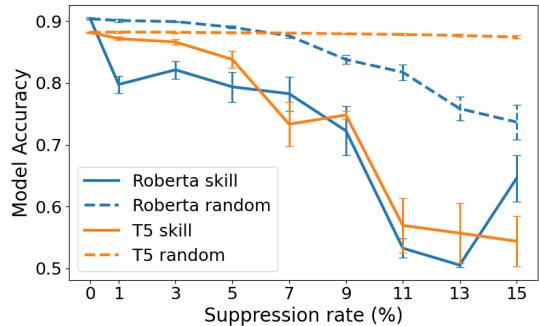

Figure 4: Model accuracies on `IMDB` when suppressing skill neurons (solid lines) versus randomly selected neurons (dashed lines).

In sum, both models have neurons that are predictive and selective for specific tasks, and their performance declines when suppressing these neurons, especially in comparison to suppressing random neurons.
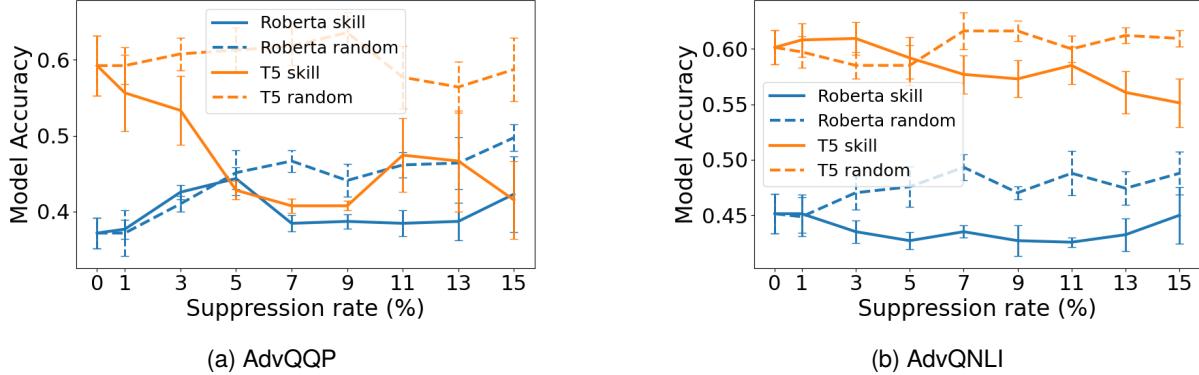
Figure 5: Model accuracies on each adversarial dataset when suppressing the skill neurons determined for these tasks (solid lines) and when suppressing randomly selected neurons (dashed lines).

5.3. The Relationship between Robustness and Skill Neurons

Our analyses above (Figure 2) show that the most predictive neurons on the adversarial datasets are, in fact, more predictive than the model itself. Notably, this is also true for the neuron accuracies (Equation 3) and can therefore not be attributed to inhibitory activations (see Appendix C). These findings suggest that highly predictive neurons may exist that do not function as skill neurons because they do not encode the necessary skill (e.g. do not correlate with the skill neurons determined on similar tasks) or because the model does not rely on their activations in making a prediction.

To investigate these possibilities, we look at the Spearman rank correlation between the neuron predictivities on the adversarial datasets and the corresponding non-adversarial datasets (see Figure 3). There are important differences between RoBERTa and T5. T5 exhibits strong (ρ: $0.57-0.84$) and significant ($p < 0.01$) correlations between the predictivities. For RoBERTa, in contrast, correlations are close to zero (ρ: -0.01–0.07), and largely non-significant—except for `(Adversarial) QNLI` ($p = 0.02$). Even when accounting for the generally higher correlations for T5 (by normalizing the scores, see Appendix B), T5 still exhibits a much stronger correspondence between adversarial and non-adversarial predictivities than RoBERTa.

Additionally, we study what happens when skill neurons for adversarial datasets are suppressed, ignoring the datasets where model performance is below chance to begin with, leaving us with T5: `AdvQQP` and `AdvQNLI`. Figure 5 shows the model accuracies on these tasks. The results for RoBERTa are included for comparison. In both cases, suppressing the skill neurons leads to a decrease in performance, with a stronger decrease when more neurons are suppressed. Thus, the analyses so far establish that T5's "adversarial" skill neurons are

important for the model's performance and further that they correlate with the "non-adversarial" skill neurons of the corresponding task.

To further test whether T5 uses the same set of skill neurons on both adversarial and non-adversarial data, we run an ablation experiment: We evaluate the model's performance on the adversarial datasets when suppressing the skill neurons identified on the corresponding non-adversarial datasets and vice versa (see Figure 6). Indeed, in both cases, performance is negatively affected, and suppressing the alternative skill neurons decreases performance more strongly than suppressing random neurons. For RoBERTa, in contrast, suppressing the alternative skill neurons is not more (and sometimes even less) harmful to performance than suppressing random neurons. In line with the correlation analysis, these results further support that at least some of T5's (but not RoBERTa's) skill neurons continue to be predictive and influential on adversarial data. Taken together, these findings suggest that T5's higher robustness to adversarial data might be related to the fact that it recruits some of the skill neurons determined on the corresponding non-adversarial dataset, and therefore—given the high prompt transferability—neurons that generally encode knowledge about the relevant type of task.

6. Discussion

This paper investigated the robustness of Prompt Tuning with respect to model activations.

Firstly, we demonstrated that Prompt Tuning leads to a high prompt transferability between datasets of the same type of task but is not robust to adversarial data. Regarding adversarial robustness, T5 is more robust than RoBERTa, probably because the examples in `AdvGLUE` were generated against surrogate models based on BERT and RoBERTa (Devlin et al., 2019). Comparing

(a) AdvQQP–QQP
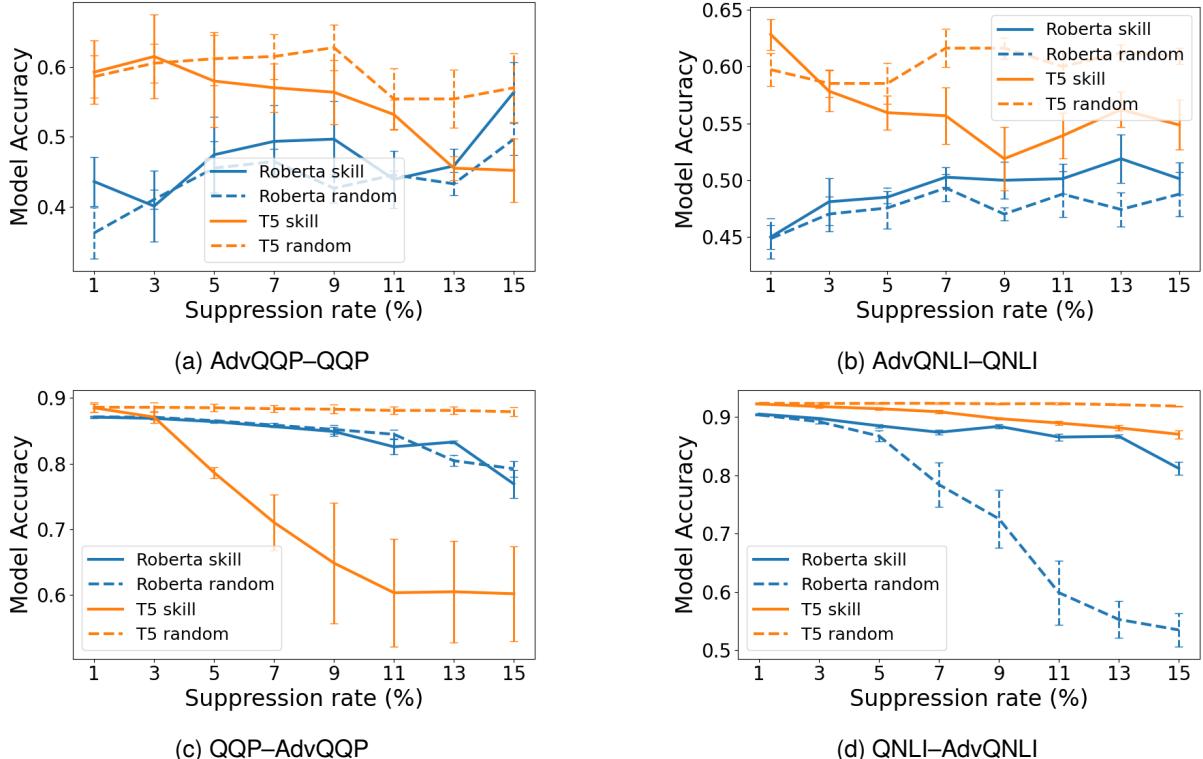
(b) AdvQNLI–QNLI

(c) QQP–AdvQQP

(d) QNLI–AdvQNLI

Figure 6: Model accuracies on the adversarial datasets when suppressing the skill neurons identified on the corresponding non-adversarial datasets, and vice versa. For example, (a) shows the accuracies on `AdvQQP` when the most predictive neurons of `QQP` (solid lines) or randomly selected neurons (dashed lines) are suppressed.

our results on `AdvGLUE` to those of finetuned (and larger model versions of) T5 and RoBERTa (Wang et al., 2021) suggests that there is no advantage of Prompt Tuning over model finetuning in terms of adversarial robustness. In other words, susceptibility to adversarial attacks likely arises during pretraining and persists across different task-adaptation methods.

Secondly, we identified skill neurons in both RoBERTa and T5. The suppression analysis revealed that, even though both models rely on these skill neurons when performing a task, suppressing them affects RoBERTa more strongly than T5. It might be that T5 encodes more redundant information. For example, it is known that a transformer's encoder output can be significantly compressed before being passed to the decoder without negatively impacting performance (Zhang et al., 2021). Besides, the skill neurons we identified for T5 tend to be slightly less predictive than those we identified for RoBERTa. More predictive neurons possibly reside in the T5 decoder and future work should extend the skill neuron analysis method to encompass both the transformer encoder and decoder.

Finally, we identified a potential link between robustness and skill neurons. Our results suggest that the activation (as measured by the correlation analysis) and use (as measured by the suppression analysis) of the same skill neurons on non-adversarial and the corresponding adversarial data may be related to model robustness. We discussed above that the adversarial attacks in `AdvGLUE` were generated against BERT- and RoBERTa-based models, and that T5 is slightly more robust against these adversarial attacks. Given that T5, but not RoBERTa, use similar skill neurons on corresponding pairs of adversarial and non-adversarial datasets, it could be that adversarial attacks work because they modify relevant skill neuron activations. Unlike RoBERTa, T5 has sparse activations. While sparsity might explain some of our results, such as the task specificity and importance of the skill neurons (see section 5.3), it is still an open question whether there is a connection between sparsity and adversarial robustness.

Building on our insight that adversarial robustness may be regulated by individual neuron activations, future research on model robustness could aim to develop methods for consistently activating the relevant skill neurons for a given task. For example, given that prompts are transferable between similar tasks, one could search for a prompt that activates skill neurons both on non-adversarial and adversarial data for a specific task, and transfer this

prompt to other tasks of that type. Similar to methods that enforce the expression of certain concepts or facts by activating the corresponding FFN neurons (see Section 2), these methods could enforce the use of specific skills.

7. Conclusion

In this paper, we investigated the robustness of Prompt Tuning in relation to model activations, specifically focusing on the existence of skill neurons and their connection to adversarial robustness. Our findings revealed that Prompt Tuning yields prompts that are transferable between similar tasks but not robust to adversarial attacks. We identified skill neurons in both RoBERTa and T5, which were highly predictive and task-specific. Suppressing these skill neurons significantly impacted task performance, highlighting their importance. Interestingly, T5 demonstrated higher adversarial robustness than RoBERTa, and skill neurons in T5 exhibited stronger correlations between adversarial and non-adversarial data. This suggests a potential link between the activation of skill neurons on adversarial data and model robustness. Future research could explore methods to enhance model robustness by consistently activating relevant skill neurons.

8. Acknowledgements

9. Bibliographical References

Jeff Da, Ronan Le Bras, Ximing Lu, Yejin Choi, and Antoine Bosselut. 2021. Analyzing commonsense emergence in few-shot knowledge models. In *3rd Conference on Automated Knowledge Base Construction*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. 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.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *ArXiv Preprint*, arxiv:2008.02275.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8018–8025.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J. Reddi, Ke Ye, Felix Chern, Felix Yu, Ruiqi Guo, and Sanjiv Kumar. 2022. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. In *International Conference on Learning Representations (ICLR)*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *ArXiv Preprint*, arXiv:2103.10385.

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.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

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

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 506—516. Curran Associates Inc.

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013*

Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. 2022. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969, Seattle, United States. Association for Computational Linguistics.

Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2020. Finding experts in transformer models. *ArXiv Preprint*, arxiv:2005.07647.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, ..., and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv Preprint*, arXiv:2307.09288.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Boxin Wang, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang, and Bo Li. 2020. T3: Tree-autoencoder constrained adversarial text generation for targeted attack. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6134–6150, Online. Association for Computational Linguistics.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yunzhi Yao, Shaohan Huang, Li Dong, Furu Wei, Huajun Chen, and Ningyu Zhang. 2022. Kformer: Knowledge injection in transformer feed-forward layers. In *Natural Language Processing and Chinese Computing (NLPCC)*, pages 131–143. Springer International Publishing.

Omar F. Zaidan, Jason Eisner, and Christine Piatko. 2008. Machine learning with annotator rationales to reduce annotation cost. In *Proceedings of the NeurIPS 2008 Workshop on Cost Sensitive Learning*.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2021. On sparsifying encoder outputs in sequence-to-sequence models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2888–2900, Online. Association for Computational Linguistics.

A. Transferability

Figure 7 shows the average absolute accuracy when evaluating the prompts tuned on a specific source dataset on a specific target dataset. Adversarial datasets are not included because we did not tune any prompts for these. Note that the matrices correspond to the matrices in Figure 1 except that we report absolute instead of relative accuracies here.
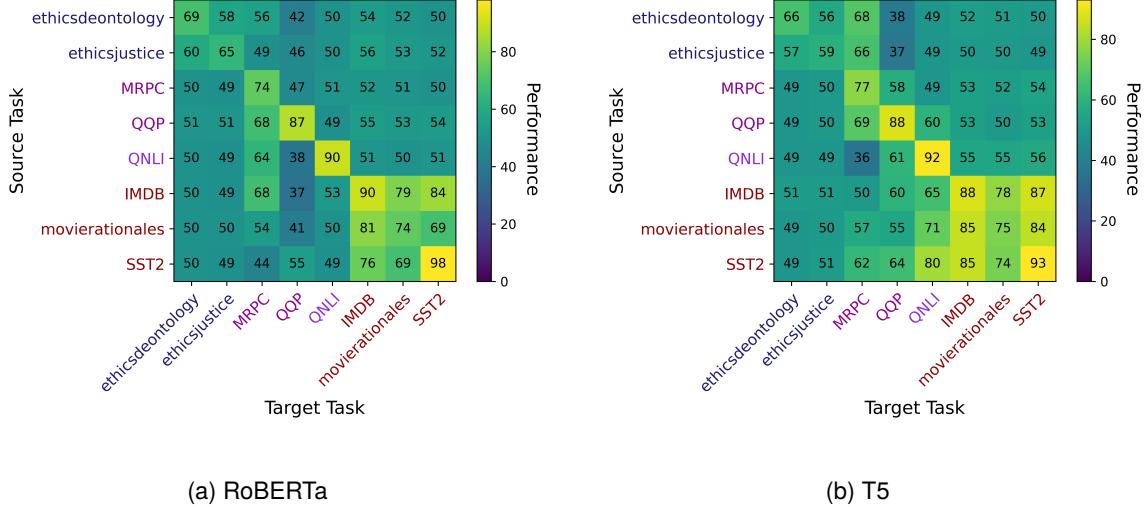

(a) RoBERTa

(b) T5

Figure 7: Prompt transferability. We calculate the accuracy when using the prompt for the source task on the target task.

B. Task-specificity (normalized)

The correlations between neuron predictivities for different datasets are generally higher for T5 than RoBERTa (see Figure 3). We applied a Z-score normalization to the correlation values to account for this difference (see Figure 8). The normalization does not affect our conclusions: Skill neurons in both T5 and RoBERTa are task-specific, and there is a strong correlation between neuron predictivities on adversarial and corresponding non-adversarial data for T5 but not RoBERTa.
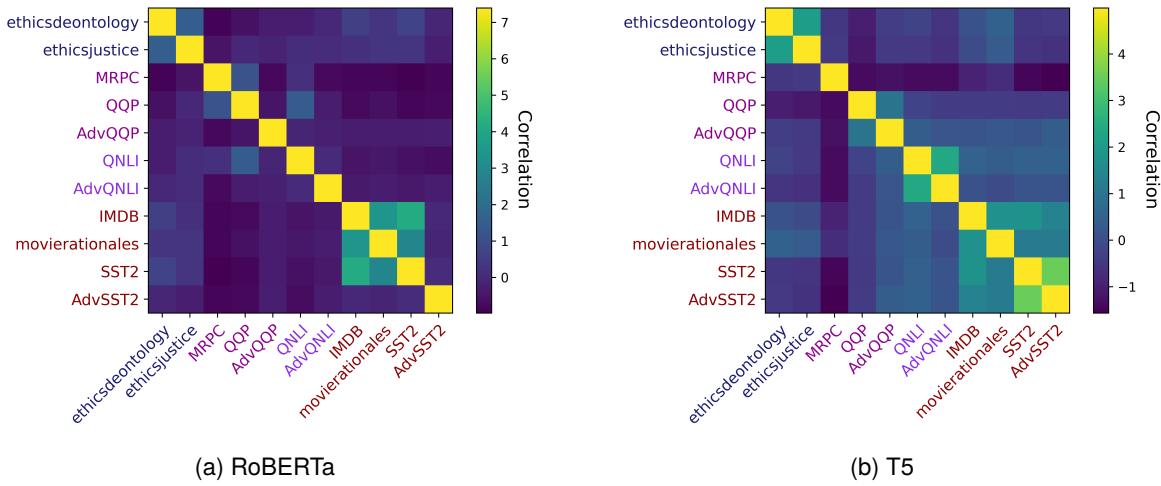

(a) RoBERTa

(b) T5

Figure 8: Z-score normalized Spearman rank correlations of the neuron predictivities for different datasets.

C. Neuron accuracies

Figure 9 shows the distributions of neuron accuracies (Equation 3) for both models and each dataset. For non-adversarial and adversarial datasets, neuron accuracies are excitatory and inhibitory. `(Adv)QNLI` poses an exception in that neuron activations are exclusively inhibitory.
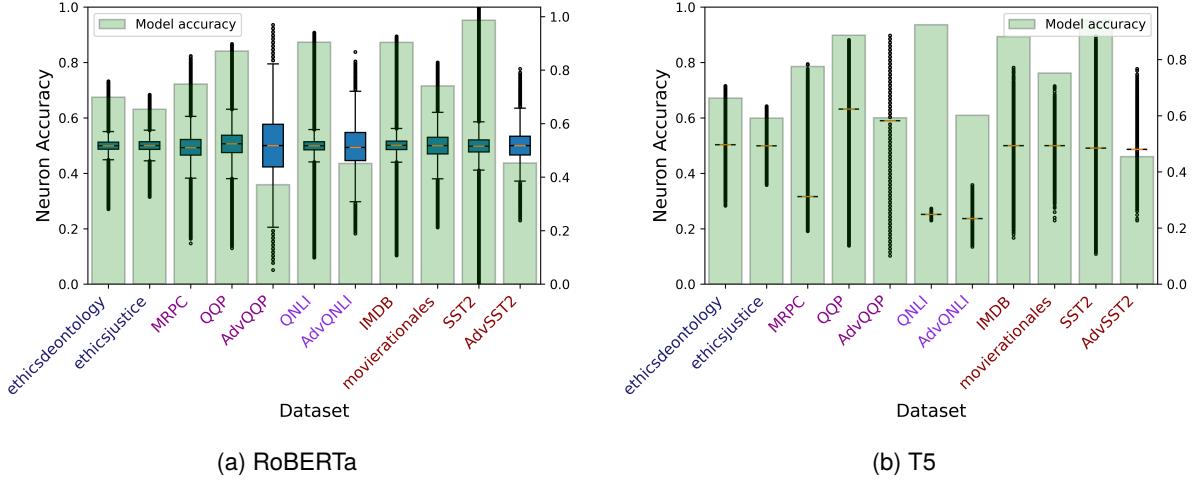
(a) RoBERTa

(b) T5

Figure 9: Distribution of neuron accuracies (box plots) on top of model accuracy (bar plots).

D.   Suppression analysis

Figure 10 shows the results of our suppression analysis for all non-adversarial datasets. Suppressing skill neurons consistently leads to a stronger decrease in accuracy than suppressing random neurons. In addition, accuracy tends to decrease with increasing suppression rates. Suppressing random neurons hardly affects T5 but leads to a—sometimes strong—decrease in performance for RoBERTa.

(a) Ethics-Deontology

(b) Ethics-Justice

(c) MRPC

(d) QQP

(e) QNLI

(f) IMDB
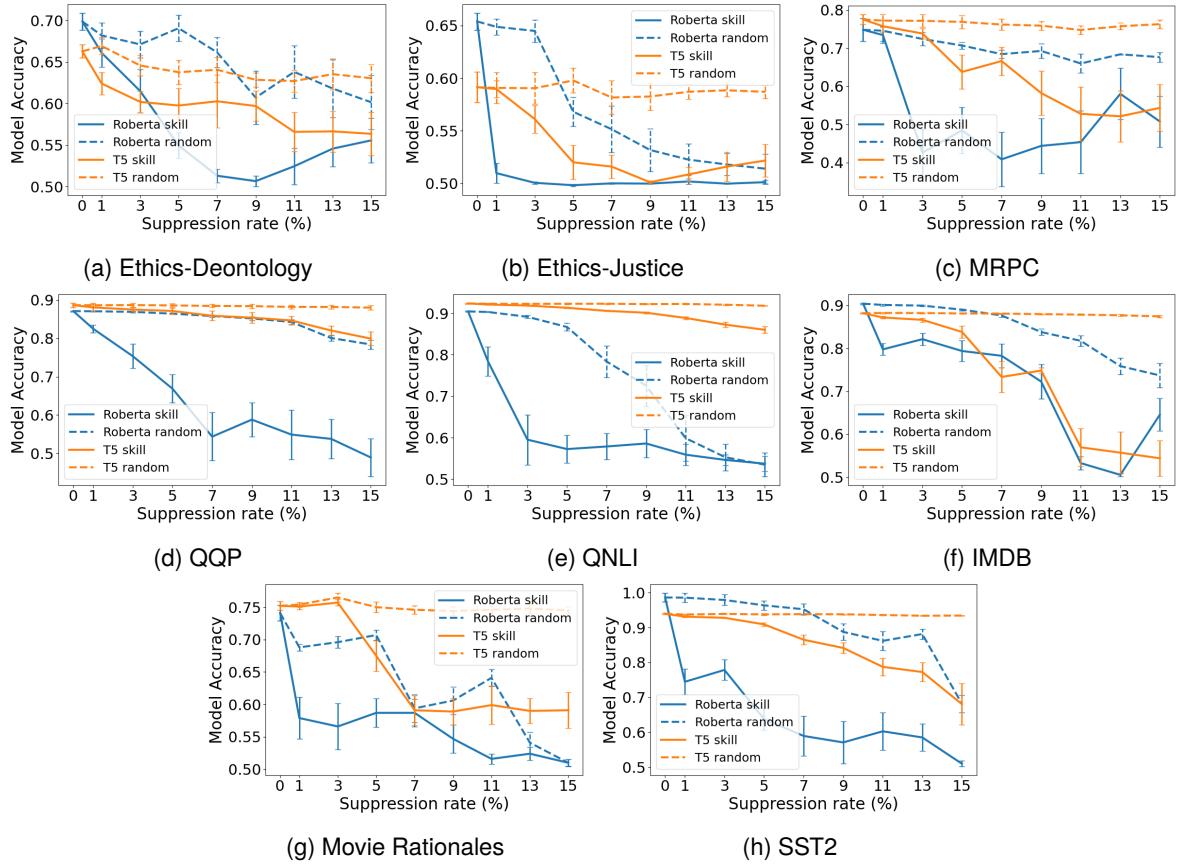
(g) Movie Rationales

(h) SST2

Figure 10: Model accuracies when neurons are suppressed. For each dataset, activations of the 0-15% most predictive neurons (solid lines) or the same amount of randomly selected neurons (dashed lines) are set to zero.