On the Relationship between Skill Neurons and Robustness in Prompt Tuning

Leon Ackermann
Institute for Cognitive Science
Osnabrück University
Osnabrück, Germany
lackermann@uni-osnabrueck.de

Xenia Ohmer
Institute for Cognitive Science
Osnabrück University
Osnabrück, Germany
xenia.ohmer@uni-osnabrueck.de

Abstract

Prompt Tuning is a popular parameter-efficient finetuning method for pre-trained large language models (PLMs). Recently, 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, with higher robustness for T5 than RoBERTa. At the same time, we replicate the existence of skill neurons in RoBERTa and further show that skill neurons also seem to 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 activate the relevant skill neurons on adversarial data.

1 Introduction

Pretrained large language models (PLMs) are ever-increasing in size. Finetuning such models for downstream tasks is extremely expensive both in terms of computation and storage. As a solution to this problem, parameter-efficient finetuning methods have been developed. 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 methods for parameter-efficient finetuning of 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 achieves comparable performance to conventional finetuning when applied to larger models. The authors further demonstrated that—next to reducing computational and storage requirements—Prompt Tuning has the additional advantage of being more robust to domain shifts, as adapting fewer parameters reduces the risk of overfitting.

To understand how Prompt Tuning actually 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 FNN neurons activated by different prompts is predictive of the prompt transferability [Su et al., 2022]. More recently, Wang et al. [2022] showed that the activations of some FFN neurons are highly

R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023.

predictive of the task labels after Prompt Tuning. Further analyses indicated 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 in Prompt Tuning, 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 models. We tune several prompts (different seeds) for various tasks for both models and identify the models’ skill neurons for each task. While DNNs are not robust to adversarial examples in various contexts [Zhang and Li, 2020], we would like to investigate whether Prompt Tuning might constitute an exception. If skill neurons really encode task-specific skills they should also function on adversarial data. Otherwise, they encode skills that correlate with the task but are not fully aligned. We test this by exposing the prompt-tuned PLMs to adversarial data. Our main contributions are:

1. Like previous work, we find that tuned prompts are transferable to other datasets, including domain shifts, when these datasets belong to the same type of task. However, using `Adversarial GLUE` [Wang et al., 2021], we show that Prompt Tuning is not robust to adversarial data.
2. Wang et al. [2022] run their skill neuron analysis only for RoBERTa. We replicate their findings and additionally identify skill neurons in (the encoder of) T5.
3. We establish a connection between adversarial robustness and skill neurons. T5 is more robust to adversarial data than RoBERTa. At the same time, while T5 seems to have skill neurons on adversarial data, which are relatively consistent with its skill neurons on the corresponding non-adversarial data, this is not the case for RoBERTa.

In sum, we provide further evidence for the existence of skill neurons in PLMs. While Prompt Tuning is not robust to adversarial data, our findings suggest that robustness may be increased by supporting the model in consistently activating the same skill neurons on adversarial and non-adversarial data.

2 Methods

2.1 Prompt Tuning

The model embeds input sequence $X_{orig} = [\text{token } 1, \text{token } 2, \ldots, \text{token } s]$ into $\mathbf{X} \in \mathbb{R}^{s \times h}$, where h is the embedding dimension. Prompt Tuning prepends additional prompt tokens $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_p], \mathbf{p}_i \in \mathbb{R}^h$ to that input in the embedding space, such that the new model input is $(\mathbf{P}, \mathbf{X}) = [\mathbf{p}_1, \ldots, \mathbf{p}_p, \mathbf{x}_1, \ldots, \mathbf{x}_s]$, with $(\mathbf{P}, \mathbf{X}) \in \mathbb{R}^{(p+s) \times h}$. The continuous prompt tokens in the embedding space are treated as free parameters of the model and their values are learned via backpropagation during the training phase while 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.

2.2 Neuron predictivity and 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 the following 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, but it can also be applied to multi-class problems [see Wang et al., 2022].

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

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

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 token $\mathbf{x}$.

Baseline activations. Let the training set be defined as $D_{\text{train}} = \left\{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), \ldots, \left(\mathbf{X}_{|D|}, y_{|D|}\right)\right\}$, 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 $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_p], \mathbf{p}_i \in \mathbb{R}^h$. The baseline activation $a(\mathcal{N}, \mathbf{p}_i) \in \mathbb{R}$ is defined as the average activation of neuron $\mathcal{N}$ for a 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

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

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

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

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 correspond to 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

$$\text{Pred}(\mathcal{N}, \mathbf{p}_i) = \max\big(\text{Acc}(\mathcal{N}, \mathbf{p}_i), 1 - \text{Acc}(\mathcal{N}, \mathbf{p}_i)\big) \; . \tag{4}$$

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

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

When sorting the neurons in the model based on their predictivity, the most predictive neurons are considered to be the "skill neurons" of the model for the given task.

3 Experiments

Models and tasks. We run our experiments with RoBERTa-base (125 million parameters) and T5-base (223 million parameters). 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` [Talat et al., 2022], and (4) natural language inference (NLI), including `QNLI` [Wang et al., 2018]. To test adversarial robustness we use `Adversarial QQP`, `Adversarial QNLI`, and `Adversarial SST2` from `Adversarial GLUE` [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[1] includes one tuned prompt for each of the (non-adversarial) datasets that we use. We train four additional prompts per dataset, giving us a total of five prompts per dataset. We analyze the models' performance on the non-adversarial data and test their robustness to adversarial data (using the prompts from the corresponding non-adversarial tasks).

Skill neurons. We calculate the neuron predictivities (Equation 5) for all non-adversarial datasets following the method described in Section 2.2. For calculating the neuron predictivities on the adversarial datasets, we use the baseline activations from the corresponding non-adversarial tasks. All of our analyses that involve neuron predictivities are done for each layer in the model—or each layer in the encoder model in the case of T5[2]—simultaneously.

[1] `https://github.com/thunlp/Prompt-Transferability/`

[2] The skill neuron calculation depends on neuron activations for specific prompt tokens, which only exist for the encoder, not the decoder.

4 Results

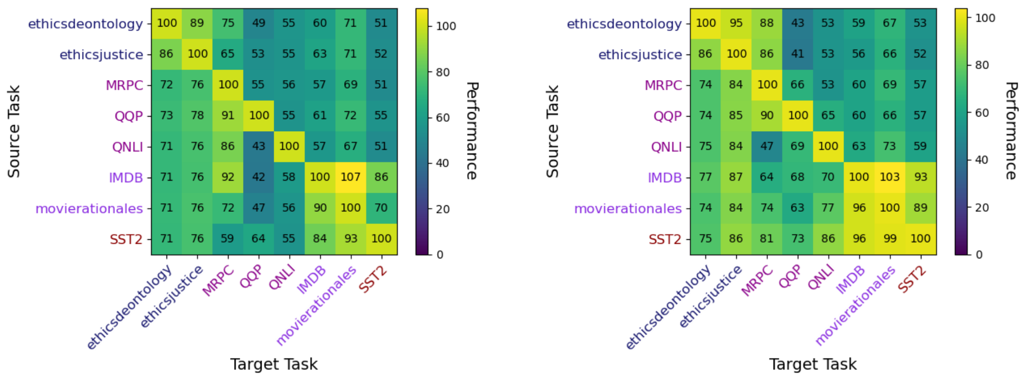
4.1 Prompt Tuning and robustness

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 correspond to those observed in other studies using Prompt Tuning [e.g. Lester et al., 2021, Su et al., 2022]. 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.

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

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 ± 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

Robustness. We analyze two different kinds of robustness: adversarial robustness and transferability. Table 1 shows the models' accuracies on the three adversarial datasets when evaluated with the continuous prompts trained on their non-adversarial counterparts. The accuracies drop significantly. For RoBERTa, they are consistently below chance performance. T5 is somewhat more robust, with below chance performance on `Adversarial SST-2` but around 60% accuracy on the other two adversarial datasets. Figure 1 shows the relative 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 transferrable between 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.



(a) RoBERTa

(b) T5

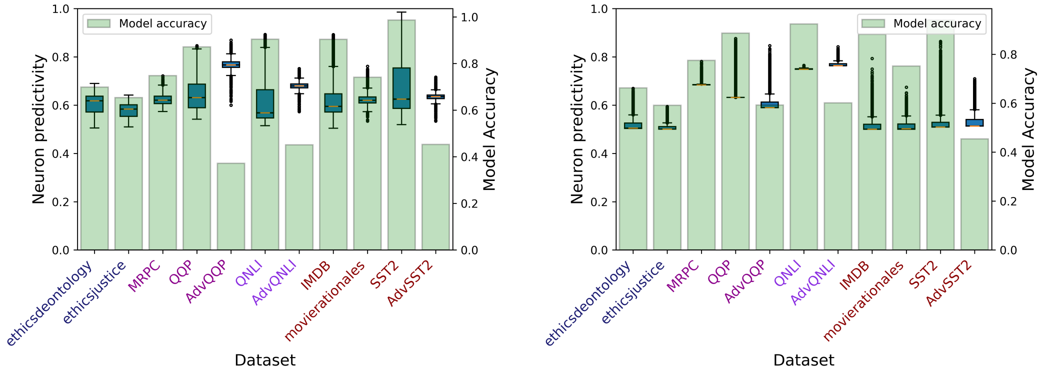
Figure 1: Prompt transferability. We calculate the accuracy when using the prompt for the source task on the target task divided by the accuracy when using the prompt for the target task on the target task for each seed, and report the average across seeds.

4.2 Skill neurons

Following the procedure by 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 indeed *important* for solving the task.

High predictivity. 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 sometimes reach and sometimes fall (slightly) short of the model's accuracy (see Figure

2b). Regarding the adversarial datasets, the predictivities of almost all neurons exceed the models' accuracy. Possible reasons are discussed in section 4.3.

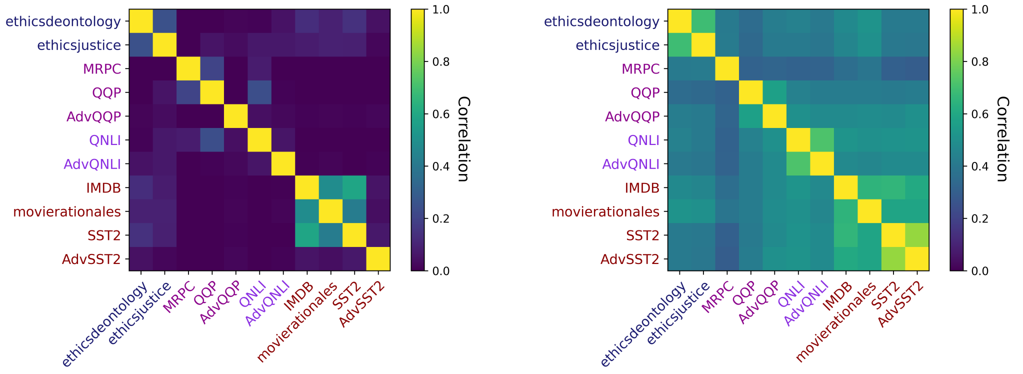


(a) RoBERTa

(b) T5

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

Task-specificity. We calculate Spearman's 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 both 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]. Appendix B shows the normalized correlation values.



(a) RoBERTa

(b) T5

Figure 3: Spearman's rank correlation between the neuron predictivities for different tasks.

Importance. To make sure 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 model's most predictive neurons versus the same number of random neurons.[3] Neurons are suppressed by setting their activations to zero. For both models the accuracy drops much more strongly when suppressing skill neurons compared to random neurons, supporting the importance of the skill neurons for the models' task performance (see `IMDB` example in Figure 4 and results for all tasks in Appendix D). Suppressing random neurons has a larger impact on RoBERTa than T5, which we again attribute to T5's sparse activations: When selecting neurons at random, many of them would not have been active anyway.

[3] Wang et al. [2022] perturb the neurons with random noise instead of suppressing them completely. Since T5's activations are very large they remain relatively unaffected by such perturbations. Therefore, we decided to do a suppression analysis, which has also been used in other work [e.g. Dai et al., 2022].

Additionally, we study what happens when skill neurons for adversarial datasets are suppressed (see Appendix D), ignoring the datasets where model performance is below chance to begin with—leaving us with T5: `Adversarial QQP` and `Adversarial QNLI`. In both cases, suppressing the skill neurons leads to a decrease in performance, with a stronger decrease when more neurons are suppressed.

4.3 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. In most cases, also the neuron accuracies (Equation 3) are higher than the model accuracies (see Appendix C), which means that the high predictivities are not caused by "inhibitory" neurons. These findings suggest that highly predictive neurons may exist that do not function as skill neurons, either because they do not really encode the necessary skill (e.g. do not correlate with the skill neurons determined on similar tasks) or because their activations do not contribute to the model's prediction.

To investigate these possibilities, we look at Spearman's 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—with the exception of `(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.

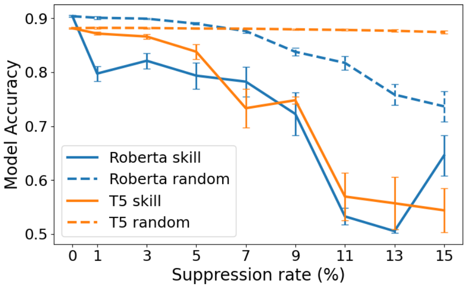


Figure 4: Suppression of skill neurons versus random neurons on `IMDB` for Roberta and T5.

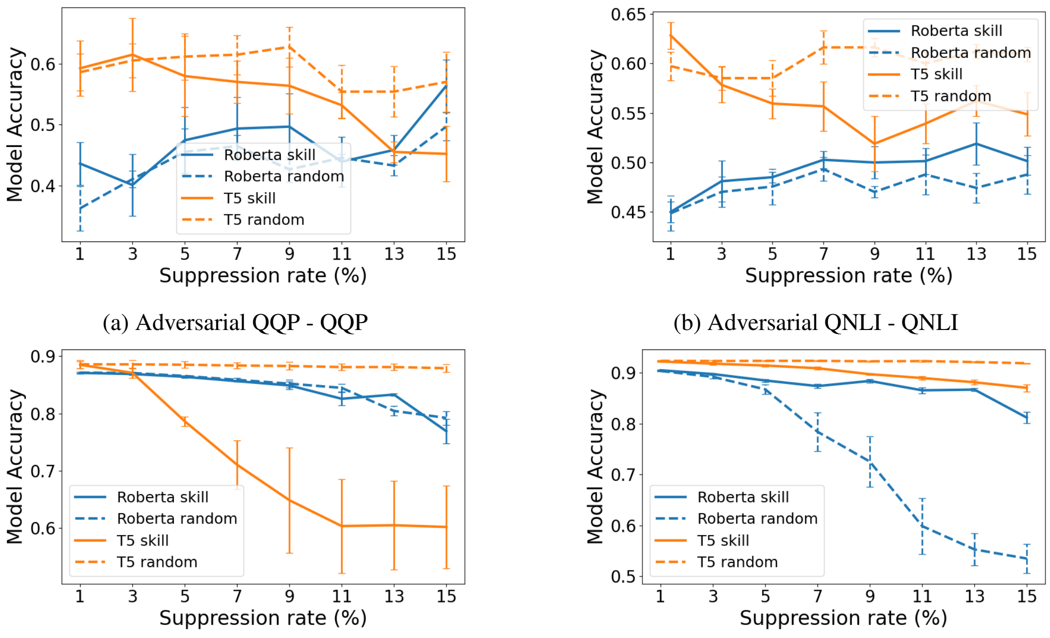
To further test whether T5 uses the same set of skill neurons on both adversarial and non-adversarial data, we run an ablation experiment: We evaluate the model's performance on the adversarial datasets when suppressing the skill neurons identified on the corresponding non-adversarial datasets and vice versa (see Figure 5). 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 T5, but not RoBERTa, consistently activates at least some of the same skill neurons on adversarial and non-adversarial data of the same task.

Taken together, these findings suggest that T5's higher robustness to adversarial data might be related to the fact that it can activate the skill neurons for the corresponding non-adversarial dataset, and therefore—given the high prompt transferability—neurons that generally encode knowledge about the relevant type of task.

5 Discussion and conclusion

In this paper, we studied the robustness of Prompt Tuning in relation to model activations. Firstly, we demonstrated that Prompt Tuning leads to a high prompt transferability between similar tasks but is not robust to adversarial data. Regarding adversarial robustness, T5 seems to be more robust than RoBERTa, probably because the examples in `AdversarialGLUE` were generated against models based on BERT Devlin et al. [2019] and RoBERTa.

Secondly, we identified skill neurons in both RoBERTa and T5 (for non-adversarial tasks). The perturbation analysis revealed that while skill neurons are crucial for performing the task, suppressing them affects RoBERTa more than T5. It might be that T5 encodes more redundant information. In particular, it is known that the encoder output of a transformer can be significantly compressed



(a) Adversarial QQP - QQP

(b) Adversarial QNLI - QNLI

(c) QQP - Adversarial QQP

(d) QNLI - Adversarial QNLI

Figure 5: 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 `Adversarial QQP` when the most predictive neurons of `QQP` (solid lines) or randomly selected neurons (dashed lines) are suppressed.

before being passed to the decoder without negatively impacting performance [Zhang et al., 2021]. Future work should extend the skill neuron analysis method to encompass both encoder and decoder and study whether neurons in the decoder are potentially more predictive and more essential for performing the task.

Finally, we established a link between adversarial robustness and skill neurons. Computer Vision studies suggest that the lack of adversarial robustness is likely caused by the existence of non-robust features [e.g., Dong et al., 2017, Ilyas et al., 2019, Ortiz-Jiménez et al., 2021]. Adversarial examples take advantage of spurious correlations, which act as discriminative features. Even if some model activations are highly predictive, they do not necessarily align with the task at hand. The skill neurons determined for RoBERTa and T5 are indeed not perfectly aligned with the task, which is reflected in a lack of adversarial robustness in both models. On the other hand, we observe that T5 (but not RoBERTa) activates and uses the same skill neurons on adversarial and non-adversarial data. At the same time, T5 is also more robust than RoBERTa. This interaction between skill neuron activation and robustness suggests that at least *some* skill neurons encode *some* task-relevant properties beyond spurious correlations. Building on this insight, future research on adversarial robustness for continuous prompts could develop methods to consistently activate the relevant skill neurons for a given task.

Acknowledgements

We are indebted to Yusheng Su for answering many questions about their paper on prompt transferability, and to Xiaozhi Wang for answering several questions about the skill neuron method. We would further like to thank Elia Bruni and Michael Rau for helpful discussions. Finally, we would like to thank the Universitätsgesellschaft Osnabrueck for sponsoring our participation at the NeurIPS R0-FoMo Workshop.

References

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 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, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL `https://aclanthology.org/2022.acl-long.581`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL `https://aclanthology.org/I05-5002`.

Yinpeng Dong, Hang Su, Jun Zhu, and Fan Bao. Towards interpretable deep neural networks by leveraging adversarial examples. *ArXiv Preprint*, arXiv:1708.05493, 2017. URL `http://arxiv.org/abs/1708.05493`.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf`.

Brian Lester, Rami Al-Rfou, and Noah Constant. 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, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL `https://aclanthology.org/2021.emnlp-main.243`.

Xiang Lisa Li and Percy Liang. 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, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL `https://aclanthology.org/2021.acl-long.353`.

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. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=TJ2nxciYCk-`.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *ArXiv Preprint*, arXiv:2103.10385, 2021. URL `https://arxiv.org/abs/2103.10385`.

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

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 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, 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1015`.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. Peft: State-of-the-art parameter-efficient fine-tuning methods, 2022. URL `https://github.com/huggingface/peft`.

Guillermo Ortiz-Jiménez, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness. *Proceedings of the IEEE*, 109(5):635–659, 2021. doi: 10.1109/JPROC.2021.3050042.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(1):1–67, 2020. URL https://dl.acm.org/doi/abs/10.5555/3455716.3455856.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 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, 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

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. 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, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.290. URL https://aclanthology.org/2022.naacl-main.290.

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

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 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, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.56. URL https://aclanthology.org/2022.naacl-main.56.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 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, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://openreview.net/pdf?id=GF9cSKI3A_q.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 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, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.765. URL https://aclanthology.org/2022.emnlp-main.765.

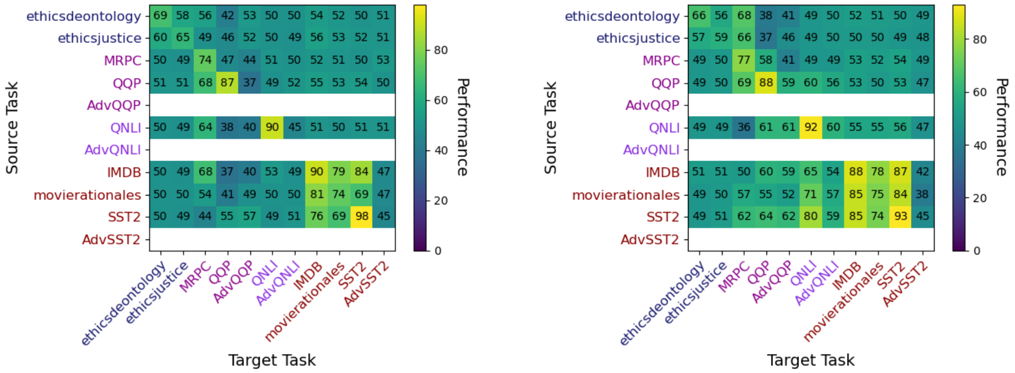
Omar F. Zaidan, Jason Eisner, and Christine Piatko. Machine learning with annotator rationales to reduce annotation cost. In *Proceedings of the NeurIPS 2008 Workshop on Cost Sensitive Learning*, 2008.

Biao Zhang, Ivan Titov, and Rico Sennrich. On sparsifying encoder outputs in sequence-to-sequence models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2888–2900, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.255. URL https://aclanthology.org/2021.findings-acl.255.

Jiliang Zhang and Chen Li. Adversarial examples: Opportunities and challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2578–2593, 2020. doi: 10.1109/TNNLS.2019.2933524.

A Transferability

Figure 6 shows the performance of each prompt on each dataset. Rows of adversarial datasets are left blank since no prompt was trained on these datasets. When T5 is evaluated on adversarial datasets, it achieves higher performance with prompts from the corresponding non-adversarial datasets compared to prompts from other non-adversarial datsets. RoBERTa exhibits the opposite pattern. It achieves higher performance on adversarial datasets when using prompts other than the ones trained on the corresponding non-adversarial data. These differences are in line with our observation that RoBERTa does not activate the relevant skill neurons (those determined on the non-adversarial datasets) when facing adversarial data. It is unclear, though, why the original prompt performs worse than all other prompts.



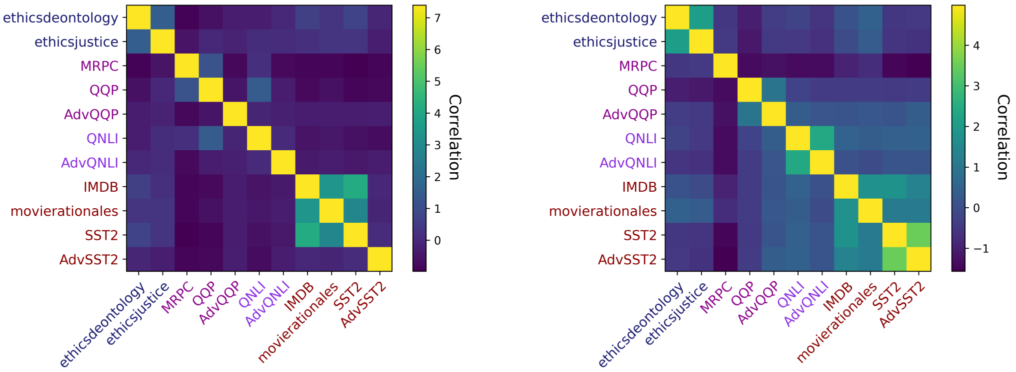
(a) RoBERTa

(b) T5

Figure 6: Zero-shot accuracy when transferring the prompt tuned on the source task to the target task.

B Task-specificity (normalized)

The correlations between neuron predictivities for different tasks are generally higher for T5 than RoBERTa (see Figure 3). To account for this fact, we applied a Z-score normalization to the correlation values, as illustrated in Figure 7. Normalizing the correlation values does not change the results. It still holds that skill neurons in both T5 and RoBERTa are task-specific and further that there is a strong correlation between neuron predictivities on adversarial and corresponding non-adversarial data for T5 but not RoBERTa.



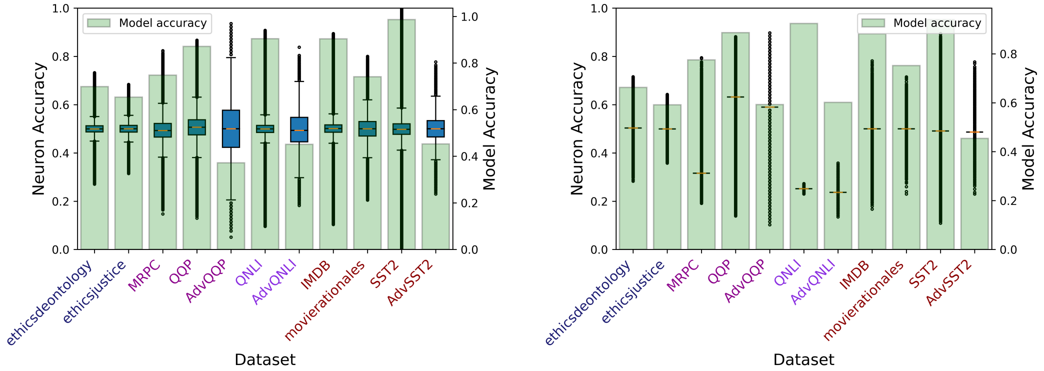
(a) RoBERTa

(b) T5

Figure 7: Z-score normalized Spearman rank correlations of the neuron predictivities for different tasks.

C Neuron accuracies

Figure 8 shows the distributions of neuron accuracies, as calculated by Equation 3, for both models and each task. For both 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

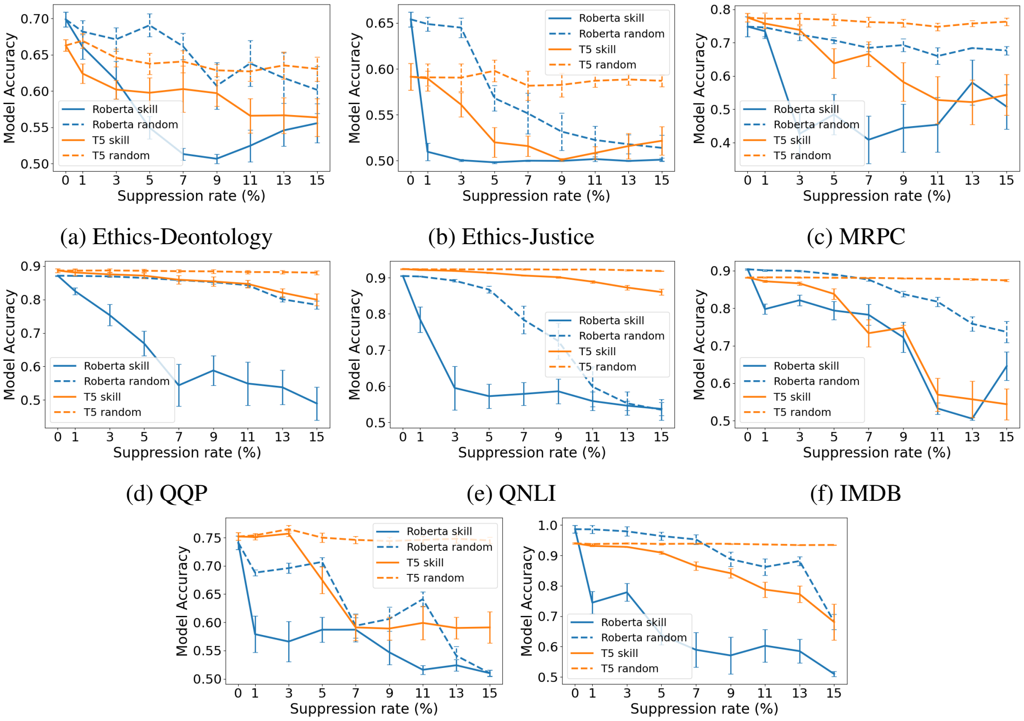
(b) T5

Figure 8: Distribution of neuron accuracies (box plots) on top of model accuracy (bar plots).

D Suppression analysis

Figure 9 shows the results of the suppression analysis for all non-adversarial datasets. Suppression of skill neurons is consistently more detrimental to performance than suppression of random neurons. Furthermore, the more skill neurons are suppressed, the more performance decreases. Suppressing random neurons hardly affects T5, while it leads to a—sometimes strong—decrease in performance for RoBERTa.

Figure 10 shows the results for the suppression analysis on all adversarial datasets. Suppressing skill neurons has a stronger negative impact on performance than suppressing random neurons, at least in those cases where the base performance (at 0% suppression rate) is above chance level.



(a) Ethics-Deontology

(b) Ethics-Justice

(c) MRPC

(d) QQP

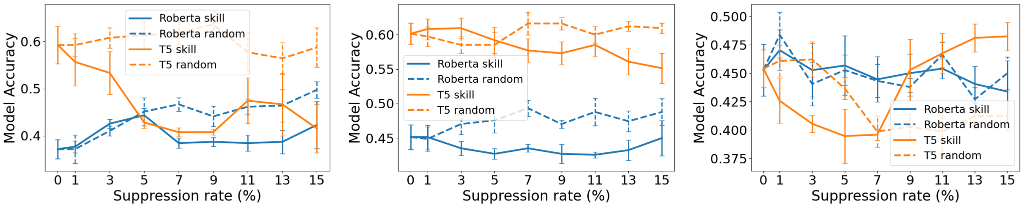
(e) QNLI

(f) IMDB

(g) Movie Rationales

(h) SST2

Figure 9: Model accuracies on each non-adversarial task when neurons are suppressed. For each dataset, the activations of 0-15% of the most predictive neurons (solid lines) or the same amount of randomly selected neurons (dashed lines) are set to zero.



(a) Adversarial QQP

(b) Adversarial QNLI

(c) Adversarial SST2

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