Pathologies of Pre-trained Language Models in Few-shot Fine-tuning

Hanjie Chen[1], **Guoqing Zheng**[2], **Ahmed Hassan Awadallah**[2], **Yangfeng Ji**[1]
[1]Department of Computer Science, University of Virginia, Charlottesville, VA, USA
[2]Microsoft Research
{hc9mx, yangfeng}@virginia.edu
{zheng, hassanam}@microsoft.com

Abstract

Although adapting pre-trained language models with few examples has shown promising performance on text classification, there is a lack of understanding of where the performance gain comes from. In this work, we propose to answer this question by interpreting the adaptation behavior using post-hoc explanations from model predictions. By modeling feature statistics of explanations, we discover that (1) without fine-tuning, pre-trained models (e.g. BERT and RoBERTa) show strong prediction bias across labels; (2) although few-shot fine-tuning can mitigate the prediction bias and demonstrate promising prediction performance, our analysis shows models gain performance improvement by capturing non-task-related features (e.g. stop words) or shallow data patterns (e.g. lexical overlaps). These observations alert that pursuing model performance with fewer examples may incur pathological prediction behavior, which requires further sanity check on model predictions and careful design in model evaluations in few-shot fine-tuning.

1 Introduction

Pre-trained language models (Brown et al., 2020; Liu et al., 2019; Devlin et al., 2019) have shown impressive adaptation ability to dowstream tasks, achieving considerable performance even with scarce task-specific training data, i.e., few-shot adaptation (Radford et al., 2019; Schick and Schütze, 2021a; Gao et al., 2021). Existing few-shot adaptation techniques broadly fall in fine-tuning and few-shot learning (Shin et al., 2020; Schick and Schütze, 2021b; Chen et al., 2021b). Specifically, fine-tuning includes directly tuning pre-trained language models with few task-specific examples or utilizing a natural-language prompt to transform downstream tasks to masked language modeling task for better mining knowledge from pre-trained models (Petroni et al., 2019; Jiang et al., 2020; Wang et al., 2021a). Few-shot learning leverages unlabeled data or auxiliary tasks to provide additional information for facilitating model training (Zheng et al., 2021; Wang et al., 2021b; Du et al., 2021a).

Although much success has been made in adapting pre-trained language models to dowstream tasks with few-shot examples, some issues have been reported. Utama et al. (2021) found that models obtained from few-shot prompt-based fine-tuning utilize inference heuristics to make predictions on sentence pair classification tasks. Zhao et al. (2021) discovered the instability of model performance towards different prompts in few-shot learning. These works mainly look at prompt-based fine-tuning and discover some problems.

This paper looks into direct fine-tuning and provides a different perspective on understanding model adaptation behavior via post-hoc explanations (Strumbelj and Kononenko, 2010; Sundararajan et al., 2017). Specifically, post-hoc explanations identify the important features (tokens) contribute to the model prediction per example. We model the statistics of important features over prediction labels via local mutual information (LMI) (Schuster et al., 2019; Du et al., 2021b). We track the change of feature statistics with the model adapting from pre-trained to fine-tuned and compare it with the statistics of few-shot training examples. This provides insights on understanding model adaptation behavior and the effect of training data in few-shot settings.

We evaluate two pre-trained language models, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), on three tasks, including sentiment classification, natural language inference, and paraphrase identification. For each task, we test on both in-domain and out-of-domain datasets to evaluate the generalization of model adaptation performance. We discover some interesting observations, some of which may have been overlooked in prior work: (1) without fine-tuning, pre-trained mod-

Proceedings of the Third Workshop on Insights from Negative Results in NLP, pages 144 - 153
May 26, 2022 ©2022 Association for Computational Linguistics

els show strong prediction bias across labels; (2) fine-tuning with a few examples can mitigate the prediction bias, but the model prediction behavior may be pathological by focusing on non-task-related features (e.g. stop words); (3) models adjust their prediction behaviors on different labels asynchronously; (4) models can capture the shallow patterns of training data to make predictions. The insight drawn from the above observations is that pursuing model performance with fewer examples is dangerous and may cause pathologies in model prediction behavior. We argue that future research on few-shot fine-tuning or learning should do sanity check on model prediction behavior and ensure the performance gain is based on right reasons.

2 Setup

Tasks. We consider three tasks: sentiment classification, natural language inference, and paraphrase identification. Each task contains an in-domain/out-of-domain dataset pair: IMDB (Maas et al., 2011)/Yelp (Zhang et al., 2015) for sentiment classification, SNLI (Bowman et al., 2015)/MNLI (Williams et al., 2018) for natural language inference, and QQP (Iyer et al., 2017)/TwitterPPDB (TPPDB) (Lan et al., 2017) for paraphrase identification. The data statistics are in Table 4 in Appendix A.1.

Models. We evaluate two pre-trained language models, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). For each task, we train the models on the in-domain training set with different ratio ($r\%, r \in [0, 1]$) of clean examples and then test them on in-domain and out-of-domain test sets.

Explanations. We explain model prediction behavior via post-hoc explanations which identify important features (tokens) in input texts that contribute to model predictions. We test four explanation methods: sampling Shapley (Strumbelj and Kononenko, 2010), integrated gradients (Sundararajan et al., 2017), attentions (Mullenbach et al., 2018), and individual word masks (Chen et al., 2021a). For each dataset, we randomly select 1000 test examples to generate explanations due to computational costs. We evaluate the faithfulness of these explanation methods via the AOPC metric (Nguyen, 2018; Chen et al., 2020). Table 6 in Appendix A.2 shows that the sampling Shapley generates more faithful explanations than other methods. In the following experiments, we adopt it to explain model predictions.

More details about the models, datasets and explanations are in Appendix A.

3 Experiments

We report the prediction results (averaged across 5 runs) of BERT and RoBERTa trained with different ratio ($r\% : 0 \sim 1\%$) of in-domain training examples on both in-domain and out-of-domain test sets in Table 2. Overall, training with more examples, BERT and RoBERTa achieve better prediction accuracy on both in-domain and out-of-domain test sets.

We look into the predictions of models from pre-trained to fine-tuned and analyze model prediction behavior change during adaptation via post-hoc explanations. In subsection 3.1, we observe that pre-trained models without fine-tuning show strong prediction bias across labels. The models fine-tuned with a few examples can quickly mitigate the prediction bias by capturing non-task-related features, leading to a plausible performance gain. In subsection 3.2, we further quantify the prediction behavior change by comparing the feature statistics of model explanations and training data. We discover that the models adjust their prediction behavior on minority labels first rather than learning information from all classes synchronously and can capture the shallow patterns of training data, which may result in pathologies in predictions.

3.1 Prediction bias in pre-trained models

In our pilot experiments, we find the predictions of pre-trained models without fine-tuning are biased across labels (see an example of confusion matrix in Figure 2 in Appendix B). Original pre-trained models tend to predict all examples with a specific label on each dataset. We denote the specific label as the majority label and the rest labels as minority labels. The results of majority labels are in Table 1.

We propose a metric, prediction bias (PB), to quantify the bias of model predictions across labels,

$$\text{PB} = \left| \frac{T_{i_1} - T_{i_2}}{T_{i_1} + T_{i_2}} - \frac{D_{i_1} - D_{i_2}}{D_{i_1} + D_{i_2}} \right|, \quad (1)$$

$$i_1 = \underset{i\in\{1,\ldots,C\}}{\text{argmax}}\,(T_i), i_2 = \underset{i\in\{1,\ldots,C\}}{\text{argmin}}\,(T_i)$$

where i_1 and i_2 are the majority and most minority labels respectively. T_i and D_i denote the numbers of model predictions and test examples on label i respectively, and C is number of classes. The range

Models	IMDB	SNLI	QQP	Yelp	MNLI	TPPDB
BERT	Pos	Neu	Pa	Pos	Neu	Pa
RoBERTa	Pos	Con	Pa	Pos	Con	Pa

Table 1: The majority labels of original pre-trained models on different datasets. Pos: postive, Con: contradiction, Neu: neutral, Pa: paraphrases.

		In-domain						Out-of-domain					
Model	r	IMDB		SNLI		QQP		Yelp		MNLI		TPPDB	
		Acc	PB	Acc	PB	Acc	PB	Acc	PB	Acc	PB	Acc	PB
BERT	0	49.73	0.97	35.30	0.65	45.10	0.46	49.86	0.98	32.95	0.95	44.44	0.85
	0.01	-	-	48.45	0.20	65.33	0.45	-	-	34.77	0.92	80.25	0.35
	0.05	60.31	0.41	63.20	0.08	69.82	0.16	61.61	0.09	37.58	0.95	86.26	0.14
	0.1	70.76	0.13	69.13	0.12	73.65	0.04	67.11	0.41	38.27	0.93	86.69	0.07
	0.5	84.71	0.05	77.63	0.06	79.06	0.02	88.19	0.08	55.37	0.45	87.27	0.03
	1	85.46	0.05	80.33	0.06	80.16	0.05	89.09	0.03	58.81	0.34	85.22	0.07
RoBERTa	0	50.17	1.00	33.55	1.00	36.84	1.26	50.00	1.00	33.24	1.02	18.93	1.62
	0.01	-	-	36.27	0.61	66.26	0.54	-	-	32.48	1.00	81.07	0.38
	0.05	58.11	0.61	68.03	0.13	71.64	0.09	58.47	0.71	42.41	0.88	82.30	0.21
	0.1	78.58	0.10	77.04	0.07	76.82	0.04	76.59	0.37	54.72	0.75	83.54	0.21
	0.5	89.56	0.01	83.84	0.04	81.91	0.05	92.54	0.08	66.90	0.37	85.67	0.06
	1	90.34	0.01	85.43	0.03	83.19	0.05	93.76	0.01	70.47	0.20	85.78	0.08

Table 2: Prediction accuracy and bias of BERT and RoBERTa trained with different ratio ($r\%$) of in-domain training examples on both in-domain and out-of-domain test sets. Acc: accuracy (%), PB: prediction bias. For PB, darker pink color implies larger prediction bias. Note that we do not consider $r = 0.01$ for IMDB and Yelp datasets because the number of training examples is too small.

of PB is $[0, 2]$. PB takes 0 if the label distribtion of model predictions is consistent with that of data. For balanced dataset, the upper bound of PB is 1, that is all examples are predicted as one label. For imbalanced dataset, PB takes 2 in an extreme case, where the dataset only contains one label of examples, while the model wrongly predicts them as another label. We consider data bias because some datasets (e.g. QQP and TPPDB) have imbalanced label distributions.

The results in Table 2 show that both pre-trained BERT and RoBERTa have strong prediction bias on all of the datasets. The prediction bias decreases with models fine-tuned with more examples.

Models make biased predictions by focusing on non-task-related features. To understand which features are associated with model prediction labels, we follow Schuster et al. (2019); Du et al. (2021b) and analyze the statistics of model explanations via local mutual information (LMI). Specifically, we select top k important features in each explanation and get a set of important features ($E = \{e\}$) over all explanations. We empirically take $k = 10$ for the IMDB and Yelp datasets and $k = 6$ for other datasets based on their average sentence lengths. The LMI between a feature e and a particular label y is

$$\mathrm{LMI}(e, y) = p(e, y) \cdot \log\left(\frac{p(y \mid e)}{p(y)}\right), \qquad (2)$$

where $p(y \mid e) = \frac{count(e,y)}{count(e)}$, $p(y) = \frac{count(y)}{|E|}$, $p(e, y) = \frac{count(e,y)}{|E|}$, and $|E|$ is the number of occurrences of all features in E. Then we can get a distribution of LMI over all tokens in the vocabulary ($\{w\}$) built upon the dataset, i.e.

$$P_{\mathrm{LMI}}(w, y) = \begin{cases} \mathrm{LMI}(w, y) & \text{if token } w \in E \\ 0 & \text{else} \end{cases} \qquad (3)$$

We normalize the LMI distribution by dividing each value with the sum of all values.

Figure 1 shows LMI distributions of BERT on the IMDB dataset with different r, where top 5 tokens are pointed in each plot (see Table 7 in Appendix B for more results on other datasets). When $r = 0$, we can see that BERT makes biased predictions on the positive label (in Table 1) by focusing on some non-task-related high-frequency tokens. The top features associated with the negative label include some relatively low-frequency tokens (e.g.

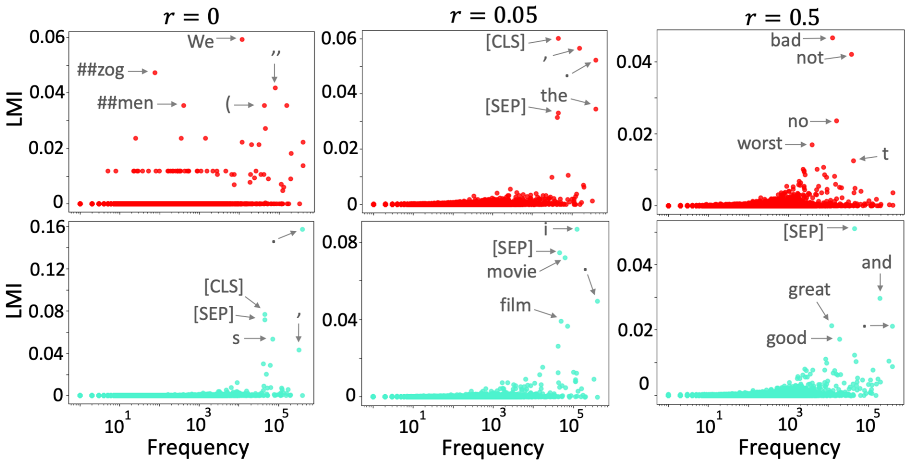

Figure 1: LMI distributions based on explanation statistics of BERT on the IMDB dataset with different r. The horizontal axis represents tokens in vocabulary in the ascending order of frequency. The upper and lower plots are on the negative and positive labels respectively. Top 5 tokens are pointed in each plot.

		In-domain														Out-of-domain															
		IMDB				SNLI						QQP				Yelp				MNLI						TPPDB					
Model	r	Ori		Data		Ori			Data			Ori		Data		Ori		Data		Ori			Data			Ori		Data			
		Neg	Pos	Neg	Pos	En	Con	Neu	En	Con	Neu	NPa	Pa	NPa	Pa	Neg	Pos	Neg	Pos	En	Con	Neu	En	Con	Neu	NPa	Pa	NPa	Pa		
BERT	0.01	-	-	-	-	0.71	0.43	0.33	0.70	0.42	0.51	0.67	0.32	0.93	0.45	-	-	-	-	0.35	0.09	0.29	0.40	0.33	0.76	0.74	0.16	1.55	0.18		
	0.05	2.26	0.45	0.90	0.63	0.58	0.60	0.47	0.31	0.17	0.16	0.49	0.14	0.23	0.22	2.20	0.69	0.66	0.43	0.43	0.45	0.41	0.76	0.27	0.63	0.87	0.02	0.58	0.03		
	0.1	2.00	0.76	0.80	0.54	0.56	0.82	0.45	0.30	0.42	0.46	0.46	0.53	0.19	0.37	2.06	0.79	0.37	0.45	0.46	0.49	0.41	0.61	0.40	1.25	0.67	0.21	0.53	0.00		
	0.5	1.39	0.80	1.16	0.52	0.70	1.51	0.94	0.14	0.54	0.46	0.31	0.67	0.08	0.21	1.61	0.93	0.73	0.52	0.92	1.70	0.78	0.82	0.91	1.02	0.93	0.09	0.37	0.04		
	1	1.21	1.60	0.68	0.86	0.80	1.02	0.65	0.14	0.48	0.52	0.21	1.01	0.00	0.42	0.73	1.94	0.46	0.83	0.73	1.31	0.55	0.76	0.69	1.14	0.46	0.54	0.33	0.11		
RoBERTa	0.01	-	-	-	-	-	0.96	-	0.76	0.52	0.56	-	0.08	0.54	0.36	-	-	-	-	-	0.95	-	0.33	0.84	0.95	-	0.00	1.55	0.00		
	0.05	-	0.66	0.17	0.72	-	0.62	-	0.50	0.32	0.67	-	0.43	0.22	0.35	-	0.38	0.14	0.62	-	0.26	-	0.89	0.22	1.07	-	0.26	1.43	0.39		
	0.1	-	1.03	0.69	0.71	-	1.05	-	0.22	0.57	0.45	-	1.27	0.17	0.59	-	0.96	0.30	0.47	-	0.18	-	1.05	0.10	0.62	-	0.39	0.72	0.36		
	0.5	-	1.33	0.81	0.42	-	2.07	-	0.21	0.60	0.55	-	1.01	0.15	0.69	-	1.70	0.66	0.43	-	0.70	-	0.87	0.70	0.79	-	0.59	0.79	0.48		
	1	-	1.41	0.86	0.62	-	0.30	-	0.17	0.32	0.23	-	0.42	0.27	0.23	-	1.91	0.65	0.78	-	0.18	-	0.72	0.66	0.51	-	0.64	0.95	0.47		

Table 3: The KL divergence between LMI distributions. The columns of "Ori" and "Data" show the results with original pre-trained models' explanations or few-shot training data as the reference respectively. Neg: negative, Pos: postive, En: entailment, Con: contradiction, Neu: neutral, NPa: nonparaphrases, Pa: paraphrases. Darker color indicates larger KL divergence.

##men, ##zog) which may have been seen by the model during pre-training.

Models adjust prediction bias by capturing non-task-related features on minority labels. Fine-tuning BERT with a few examples ($r = 0.05$, exactly 9 examples) from IMDB can quickly mitigate the prediction bias along with a plausible improvement on prediction accuracy (in Table 2). However, Figure 1 (the middle upper plot) shows that the model captures non-task-related high-frequency tokens to make predictions on the minority label (negative), implying the performance gain is not reasonable. Only when the model is fine-tuned with more examples ($r = 0.5$), it starts capturing task-specific informative tokens, such as "bad", "good".

3.2 Quantifying model adaptation behavior

To quantify the model prediction behavior change (in Figure 1) during adaptation, we compute the Kullback–Leibler divergence (KLD) between the LMI distributions of the model without/with fine-tuning, i.e. $KL_y(P^0_{LMI}(w, y), P^r_{LMI}(w, y))$. The superscripts ("0" or "r") indicate the ratio of training examples used in fine-tuning. Besides, we also evaluate how much the model prediction behavior is learned from the patterns of training data. Specifically, we compute the LMI distribution of few-shot training examples via Equation 2 and Equation 3, except that E represents the set of features appearing in those examples. Then we use the LMI distribution of data as the reference and compute the KLD between it and the LMI distribution of model explanations.

Table 3 records the results of KLD with the LMI distribution of original pre-trained model explanations as the reference (columns of "Ori") or that of training data as the reference (columns of "Data"). Note that we do not have the results of RoBERTa on some labels (e.g. "Neg") in "Ori" columns because the pre-trained RoBERTa does not make any predictions on those labels and we do not have the reference LMI distributions.

Models adjust their prediction behaviors on different labels asynchronously. In "Ori" columns, the KLDs on minority labels are larger than those on majority labels when r is small (e.g. 0.05). The changes of KLDs are discrepant across labels with r increasing. The results show that the models focus on adjusting their prediction behavior on minority labels first rather than learning from all classes synchronously in few-shot settings.

Models can capture the shallow patterns of training data. In "Data" columns, the KLDs on SNLI and QQP are overall smaller than those on IMDB, illustrating that it is easier for models to learn the patterns of datasets on sentence-pair classification tasks. With r increasing, the KLDs on the entailment label of SNLI are smaller than those on other labels, which validates the observations in previous work (Utama et al., 2021; Nie et al., 2019) that models can capture lexical overlaps to predict the entailment label. Another interesting observation is the KLDs on Yelp in "Data" columns are mostly smaller than those on IMDB. This indicates that models may rely on the shallow patterns of in-domain datasets to make predictions on out-of-domain datasets.

4 Conclusion

In this work, we take a closer look into the adaptation behavior of pre-trained language models in few-shot fine-tuning via post-hoc explanations. We discover many pathologies in model prediction behavior. The insight drawn from our observations is that promising model performance gain in few-shot learning could be misleading. Future research on few-shot fine-tuning or learning requires sanity check on model prediction behavior and some careful design in model evaluation and analysis.

Acknowledgments

We thank the anonymous reviewers for many valuable comments.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hanjie Chen, Song Feng, Jatin Ganhotra, Hui Wan, Chulaka Gunasekara, Sachindra Joshi, and Yangfeng Ji. 2021a. Explaining neural network predictions on sentence pairs via learning word-group masks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3917–3930, Online. Association for Computational Linguistics.

Hanjie Chen and Yangfeng Ji. 2020. Learning variational word masks to improve the interpretability of neural text classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4236–4251, Online. Association for Computational Linguistics.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5578–5593, Online. Association for Computational Linguistics.

Yiming Chen, Yan Zhang, Chen Zhang, Grandee Lee, Ran Cheng, and Haizhou Li. 2021b. Revisiting self-training for few-shot learning of language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9125–9135, Online and Punta Cana, Dominican Republic. 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.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021a. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online. Association for Computational Linguistics.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021b. Towards interpreting and mitigating shortcut learning behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929, Online. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. Quora question pairs.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. 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*.

Andrew 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.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. 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 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. 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 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of nli models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6867–6874.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games*, 2(28).

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Erik Strumbelj and Igor Kononenko. 2010. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Prasetya Ajie Utama, Nafise Sadat Moosavi, Victor Sanh, and Iryna Gurevych. 2021. Avoiding inference heuristics in few-shot prompt-based finetuning. *arXiv preprint arXiv:2109.04144*.

Chengyu Wang, Jianing Wang, Minghui Qiu, Jun Huang, and Ming Gao. 2021a. TransPrompt: Towards an automatic transferable prompting framework for few-shot text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2792–2802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2021b. Meta self-training for few-shot neural sequence labeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1737–1747.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.

Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. 2021. Meta label correction for noisy label learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

A Supplement of Setup

A.1 Models and Datasets

We adopt the pretrained BERT-base and RoBERTa-base models from Hugging Face[1]. For sentiment classification, we utilize movie reviews IMDB (Maas et al., 2011) as the in-domain dataset and Yelp reviews (Zhang et al., 2015) as the out-of-domain dataset. For natural language inference, the task is to predict the semantic relationship between a premise and a hypothesis as entailment, contradiction, or neutral. The Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) and Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018) are used as the in-domain and out-of-domain datasets respectively. The task of paraphrase identification is to judge whether two input texts are semantically equivalent or not. We adopt the Quora Question Pairs (QQP) (Iyer et al., 2017) as the in-domain dataset, while using the TwitterPPDB (TPPDB) (Lan et al., 2017) as the out-of-domain dataset. Table 4 shows the statistics of the datasets.

We implement the models in PyTorch 3.6. We set hyperparameters as: learning rate is $1e-5$, maximum sequence length is 256, maximum gradient norm is 1, and batch size is 8. All experiments were performed on a single NVidia GTX 1080 GPU. We report the time for training each model on each in-domain dataset (with full training examples) in Table 5.

A.2 Explanations

We adopt four explanation methods:

- sampling Shapley (SS) (Strumbelj and Kononenko, 2010): computing feature attributions via sampling-based Shapley value (Shapley, 1953);
- integrated gradients (IG) (Sundararajan et al., 2017): computing feature attributions by integrating gradients of points along a path from a baseline to the input;
- attentions (Attn) (Mullenbach et al., 2018): attention weights in the last hidden layer as feature attributions;
- individual word masks (IMASK) (Chen et al., 2021a): learning feature attributions via variational word masks (Chen and Ji, 2020).

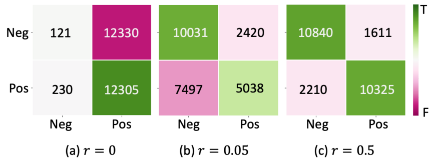

Figure 2: Confusion matrix of BERT (with different r) on the IMDB dataset. “Neg” and “Pos” represent negative and positive labels respectively. Vertical and horizontal dimensions show ground-truth and predicted labels respectively. Green and pink colors represent true or false predictions respectively. Darker color indicates larger number.

Explanation faithfulness. An important criterion for evaluating explanations is their faithfulness to model predictions (Jacovi and Goldberg, 2020). We evaluate the faithfulness of the four explanation methods via the AOPC metric (Nguyen, 2018; Chen et al., 2020). AOPC calculates the average change of prediction probability on the predicted class over all examples by removing top $1 \ldots u$ words identified by explanations.

$$\mathrm{AOPC} = \frac{1}{U+1} \langle \sum_{u=1}^{U} p(y|\boldsymbol{x}) - p(y|\boldsymbol{x}_{\backslash 1 \ldots u}) \rangle_{\boldsymbol{x}}, \tag{4}$$

where $p(y|\boldsymbol{x}_{\backslash 1 \ldots u})$ is the probability for the predicted class when words $1 \ldots u$ are removed and $\langle \cdot \rangle_{\boldsymbol{x}}$ denotes the average over all test examples. Higher AOPC score indicates better explanations.

We test the BERT and RoBERTa trained with 1% in-domain training examples on each task. For each dataset, we randomly select 1000 test examples to generate explanations due to computational costs. We report the results of AOPC scores when U = 10 in Table 6. Sampling Shapley consistently outperforms other three explanation methods in explaining different models on both in-domain and out-of-domain datasets.

B Supplement of Experiments

[1] https://github.com/huggingface/pytorch-transformers

Datasets	*C*	*L*	*#train*	*#dev*	*#test*	Label distribution
IMDB	2	268	19992	4997	24986	Positive: *train*(10036), *dev*(2414), *test*(12535) Negative: *train*(9956), *dev*(2583), *test*(12451)
Yelp	2	138	500000	60000	38000	Positive: *train*(250169), *dev*(29831), *test*(19000) Negative: *train*(249831), *dev*(30169), *test*(19000)
SNLI	3	14	549367	4921	4921	Entailment: *train*(183416), *dev*(1680), *test*(1649) Contradiction: *train*(183187), *dev*(1627), *test*(1651) Neutral: *train*(182764), *dev*(1614), *test*(1651)
MNLI	3	22	391176	4772	4907	Entailment: *train*(130416), *dev*(1736), *test*(1695) Contradiction: *train*(130381), *dev*(1535), *test*(1631) Neutral: *train*(130379), *dev*(1501), *test*(1581)
QQP	2	11	363178	20207	20215	Paraphrases: *train*(134141), *dev*(7435), *test*(7447) Nonparaphrases: *train*(229037), *dev*(12772), *test*(12768)
TPPDB	2	15	42200	4685	4649	Paraphrases: *train*(11167), *dev*(941), *test*(880) Nonparaphrases: *train*(31033), *dev*(3744), *test*(3769)

Table 4: Summary statistics of the datasets, where *C* is the number of classes, *L* is average sentence length, and *#* counts the number of examples in the *train/dev/test* sets. For label distribution, the number of examples with the same label in *train/dev/test* is noted in bracket.

Models	IMDB	SNLI	QQP
BERT	856.43	25402.52	17452.12
RoBERTa	912.47	256513.98	17514.80

Table 5: The average runtime (s/epoch) of each model on each in-domain dataset.

		In-domain			Out-of-domain		
Model	*r*	IMDB	SNLI	QQP	Yelp	MNLI	TPPDB
BERT	SS	0.41	0.82	0.61	0.53	0.77	0.40
	IG	0.08	0.34	0.19	0.12	0.31	0.10
	Attn	0.07	0.35	0.28	0.12	0.26	0.14
	IMASK	0.09	0.28	0.25	0.09	0.25	0.08
RoBERTa	SS	0.25	0.86	0.53	0.28	0.84	0.28
	IG	0.02	0.36	0.21	0.04	0.38	0.09
	Attn	0.02	0.33	0.26	0.03	0.23	0.09
	IMASK	0.02	0.18	0.18	0.03	0.17	0.05

Table 6: AOPC scores of different explanation methods in explaining different models.

Datasets	r	Labels	Top Features
IMDB	0	Neg	we ##zog " ##men (’ [SEP] capitalism lynch hell
		Pos	. [CLS] [SEP] s , t movie film plot)
	0.5	Neg	bad not no worst t off terrible nothing stupid boring
		Pos	[SEP] and great . good [CLS] love , film characters
Yelp	0	Neg	. they majestic adds state owners loud dirty priced thai
		Pos	. [CLS] [SEP] , s t for i you m
	0.5	Neg	not no bad t worst never off rude over nothing
		Pos	[SEP] great and good . [CLS] amazing love friendly experience
SNLI	0	En	a [SEP] man the woman dog sitting sits his fire
		Con	[SEP] [CLS] is the a , are in of there
		Neu	. people woman girl are playing looking [CLS] group boy
	0.5	En	[SEP] . [CLS] and is a man there woman people
		Con	the a in [SEP] at sitting with man on playing
		Neu	[SEP] are for . man [CLS] is the a girl
MNLI	0	En	the [SEP] ##ists israel ’ recession ata consultants discusses attacked
		Con	[SEP] [CLS] , s to of in . the not
		Neu	. [CLS] they we you people about it really i
	0.5	En	. [CLS] and is [SEP] there are , was of
		Con	the ’ . not no t [CLS] don to didn
		Neu	[SEP] [CLS] the for to all when . you it
QQP	0	NPa	? is the a ’ what india does quo why
		Pa	[SEP] [CLS] ? in i , of . best s
	0.5	NPa	? what [CLS] is how , why a the .
		Pa	[SEP] quo [CLS] best trump ##ra india life your sex
TPPDB	0	NPa	trump ’ the obama " we is russia a says
		Pa	[SEP] . [CLS] , s of in to ##t t
	0.5	NPa	. , [CLS] ? ’@ ; - a is
		Pa	[SEP] trump [CLS] inauguration obama russia repeal ##care cia senate

Table 7: Top 10 important tokens for BERT predictions on different labels. Neg: negative, Pos: postive, En: entailment, Con: contradiction, Neu: neutral, NPa: nonparaphrases, Pa: paraphrases.