

Avoiding Inference Heuristics in Few-shot Prompt-based Finetuning

Prasetya Ajie Utama^{†‡}, Nafise Sadat Moosavi[‡], Victor Sanh[♣], Iryna Gurevych[‡]

[†]Research Training Group AIPHES

[‡]UKP Lab, Technische Universität Darmstadt

[♣]Hugging Face, Brooklyn, USA

[‡]<https://www.ukp.tu-darmstadt.de>

utama@ukp.tu-darmstadt.de

Abstract

Recent *prompt-based* approaches allow pre-trained language models to achieve strong performances on *few-shot finetuning* by reformulating downstream tasks as a language modeling problem. In this work, we demonstrate that, despite its advantages on low data regimes, finetuned prompt-based models for sentence pair classification tasks still suffer from a common pitfall of adopting inference heuristics based on lexical overlap, e.g., models incorrectly assuming a sentence pair is of the same meaning because they consist of the same set of words. Interestingly, we find that this particular inference heuristic is significantly less present in the zero-shot evaluation of the prompt-based model, indicating how finetuning can be *destructive* to useful knowledge learned during the pretraining. We then show that adding a regularization that preserves pretraining weights is effective in mitigating this destructive tendency of few-shot finetuning. Our evaluation on three datasets demonstrates promising improvements on the three corresponding challenge datasets used to diagnose the inference heuristics.¹

1 Introduction

Prompt-based finetuning has emerged as a promising paradigm to adapt Pretrained Language Models (PLM) for downstream tasks with limited number of labeled examples (Schick and Schütze, 2021a; Radford et al., 2019). This approach reformulates downstream task instances as a language modeling input,² allowing PLMs to make non-trivial task-specific predictions even in zero-shot settings. This in turn, provides a good initialization point for data efficient finetuning (Gao et al., 2021), resulting in

¹The code is available at <https://github.com/UKPLab/emnlp2021-prompt-ft-heuristics>

²E.g., appending a cloze prompt “It was [MASK]” to a sentiment prediction input sentence “Delicious food!”, and obtaining the sentiment label by comparing the probabilities assigned to the words “great” and “terrible”.

a strong advantage on low data regimes where the standard finetuning paradigm struggles. However, the success of this prompting approach has only been shown using common held-out evaluations, which often conceal certain undesirable behaviors of models (Niven and Kao, 2019).

One such behavior commonly reported in downstream models is characterized by their preference to use surface features over general linguistic information (Warstadt et al., 2020). In the Natural Language Inference (NLI) task, McCoy et al. (2019) documented that models preferentially use the lexical overlap feature between sentence pairs to blindly predict that one sentence *entails* the other. Despite models’ high in-distribution performance, they often fail on counterexamples of this *inference heuristic*, e.g., they predict that “*the cat chased the mouse*” entails “*the mouse chased the cat*”.

At the same time, there is a mounting evidence that pre-training on large text corpora extracts rich linguistic information (Hewitt and Manning, 2019; Tenney et al., 2019). However, based on recent studies, standard finetuned models often overlook this information in the presence of lexical overlap (Nie et al., 2019; Dasgupta et al., 2018). We therefore question whether direct adaptation of PLMs using prompts can better transfer the use of this information during finetuning. We investigate this question by systematically studying the heuristics in a prompt-based model finetuned across three datasets with varying data regimes. Our intriguing results reveal that: (i) zero-shot prompt-based models are more robust to using the lexical overlap heuristic during inference, indicated by their high performance on the corresponding challenge datasets; (ii) however, prompt-based finetuned models quickly adopt this heuristic as they learn from more labeled data, which is indicated by gradual degradation of the performance in challenge datasets.

We then show that regularizing prompt-based finetuning, by penalizing the learning from up-

dating the weights too far from their original pre-trained values, is an effective approach to improve the in-distribution performance on target datasets, while mitigating the adoption of inference heuristics. Overall, our work suggests that while prompt-based finetuning has gained impressive results on standard benchmarks, it can have a negative impact regarding inference heuristics, which in turn suggests the importance of a more thorough evaluation setup to ensure meaningful progress.

2 Inference Heuristics in Prompt-based Finetuning

Prompt-based PLM Finetuning In this work, we focus on sentence pairs classification tasks, where the goal is to predict semantic relation y of an input pair $x = (s_1, s_2)$. In a standard finetuning setting, s_1 and s_2 are concatenated along with a special token [CLS], whose embedding is used as an input to a newly initialized *classifier head*.

The *prompt-based* approach, on the other hand, reformulates pair x as a masked language model input using a pre-defined template and word-to-label mapping. For instance, Schick and Schütze (2021a) formulate a natural language inference instance (s_1, s_2, y) as:

$$[\text{CLS}] s_1 ? [\text{MASK}], s_2 [\text{SEP}]$$

with the following mapping for the masked token: “yes” → “entailment”, “maybe” → “neutral”, and “no” → “contradiction”. The probabilities assigned by the PLM to the label words at the [MASK] token can then be directly used to make task-specific predictions, allowing PLM to perform in a zero-shot setting. Following Gao et al. (2021), we further finetune the prompt-based model on the available labeled examples for each task. Note that this procedure finetunes only the existing pre-trained weights, and does not introduce new parameters.

Task and Datasets We evaluate on three English language datasets included in the GLUE benchmark (Wang et al., 2018) for which there are challenge datasets to evaluate the lexical overlap heuristic: MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), and Quora Question Pairs (QQP). In MNLI and SNLI, the task is to determine whether premise sentence s_1 *entails*, *contradicts*, or is *neutral* to the hypothesis sentence s_2 . In QQP, s_1 and s_2 are a pair of questions that are labeled as either *duplicate* or *non-duplicate*.

Original Input	
Premise	The actors that danced saw the author.
Hypothesis	The actors saw the author.
Label	entailment (support)
Premise	The managers near the scientist resigned.
Hypothesis	The scientist resigned.
Label	non-entailment (against)
Reformulated Input	
Premise	The actors that danced saw the author? [MASK], the actors saw the author.
Label word	Yes
Premise	The managers near the scientist resigned? [MASK], the scientist resigned.
Label word	No / Maybe

Table 1: **Top:** input examples of the NLI task that **support** or are **against** the lexical overlap heuristics. **Bottom:** reformulated NLI instances as masked language model inputs with the expected label words.

Researchers constructed corresponding *challenge* sets for the above datasets, which are designed to contain examples that are *against* the heuristics, i.e., the examples exhibit word overlap between the two input sentences but are labeled as non-entailment for NLI or non-duplicate for QQP. We evaluate each few-shot model against its corresponding challenge dataset. Namely, we evaluate models trained on MNLI against entailment and non-entailment subsets of the HANS dataset (McCoy et al., 2019), which are further categorized into lexical overlap (lex.), subsequence (subseq.), and constituent (const.) subsets; SNLI models against the long and short subsets of the Scramble Test challenge set (Dasgupta et al., 2018); and QQP models against the PAWS dataset (Zhang et al., 2019).³ We illustrate challenge datasets examples and their reformulation as prompts in Table 1.

Model and Finetuning Our training and standard evaluation setup closely follow Gao et al. (2021), which measure finetuning performances across five different randomly sampled training data of size K to account for finetuning instability on small datasets (Dodge et al., 2020; Mosbach et al., 2021). We perform five data subsampling for each dataset and each data size K , where $K \in \{16, 32, 64, 128, 256, 512\}$. Note that K indicates the number of examples *per label*. We use the original development sets of each training dataset for testing the *in-distribution* performance. We per-

³See appendix A for details of HANS, PAWS, and Scramble Test test sets.

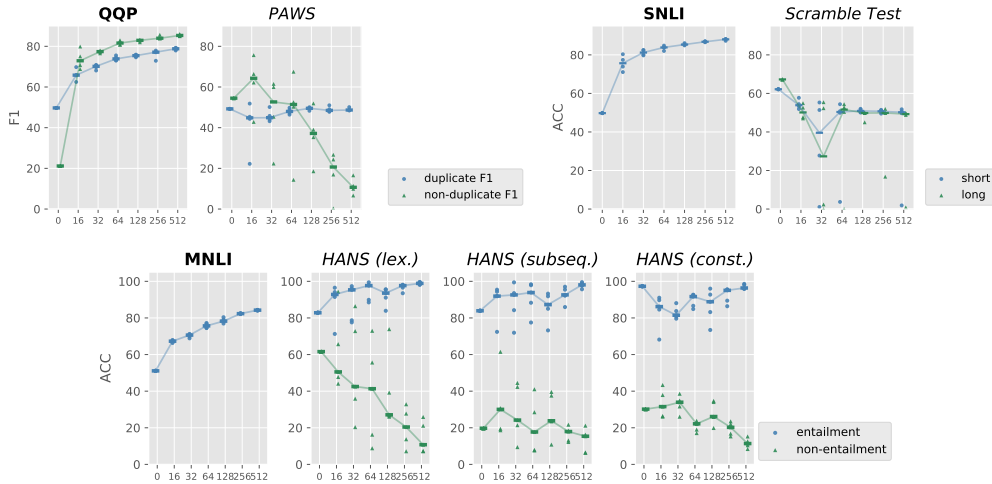


Figure 1: In-distribution (**bold**) vs. challenge datasets (*italic*) evaluation results of prompt-based finetuning across different data size \mathbf{K} (x axis), where $\mathbf{K} = 0$ indicates zero-shot evaluation. In all challenge sets, the overall zero-shot performance (both blue and green plots) degrades as the model is finetuned using more data.

form all experiments using the RoBERTa-large model (Liu et al., 2019b).

Inference heuristics across data regimes We show the results of the prompt-based finetuning across different \mathbf{K} in Figure 1. For the in-distribution evaluations (leftmost of each plot), the prompt-based models finetuned on MNLI, SNLI, and QQP improve rapidly with more training data before saturating at $\mathbf{K} = 512$. In contrast to the in-distribution results, we observe a different trajectory of performance on the three challenge datasets. On the Scramble and HANS sets, prompt-based models show non-trivial zero-shot performance ($\mathbf{K} = 0$) that is above its in-distribution counterpart. However, as more data is available, the models exhibit stronger indication of adopting heuristics. Namely, the performance on examples subset that *support* the heuristics increases, while the performance on cases that are *against* heuristics decreases. This pattern is most pronounced on the lexical overlap subset of HANS, where the median accuracy on non-entailment subset drops to below 10% while the entailment performance reaches 100%. The results suggest that few-shot finetuning can be destructive against the initial ability of prompt-based classifier to ignore surface features like lexical overlap. Finetuning appears to over-adjust model parameters to the small target data, which contain very few to no counter-examples to the heuristics (Min et al., 2020; Lovering et al., 2021).

3 Avoiding Inference Heuristics

Here we look to mitigate the adverse impact of finetuning by viewing the issue as an instance of catastrophic forgetting (French, 1999), which is characterized by the loss of performance on the original dataset after subsequent finetuning on new data. We then propose a regularized prompt-based finetuning based on the Elastic Weight Consolidation (EWC) method (Kirkpatrick et al., 2017), which penalizes updates on weights crucial for the original zero-shot performance. EWC identifies these weights using empirical Fisher matrix (Martens, 2020), which requires samples of the original dataset. To omit the need of accessing the pretraining data, we follow Chen et al. (2020) that assume stronger independence between the Fisher information and the corresponding weights. The penalty term is now akin to the L2 loss between updated weights θ_i and the original weights θ_i^* , resulting in the following overall loss:

$$\mathbf{L}_{rFT} = \alpha \mathbf{L}_{FT} + (1 - \alpha) \frac{\lambda}{2} \sum_i (\theta_i - \theta_i^*)^2$$

where \mathbf{L}_{FT} is a standard cross entropy, λ is a quadratic penalty coefficient, and α is a coefficient to linearly combine the two terms. We use the RecAdam implementation (Chen et al., 2020) for this loss, which also applies an annealing mechanism to gradually upweight the standard loss \mathbf{L}_{FT} toward the end of training.⁴

⁴See Appendix A for implementation details.

	MNLi (acc.)			QQP (F1)			SNLI (acc.)		
	In-dist.	HANS	avg.	In-dist.	PAWS	avg.	In-dist.	Scramble	avg.
Prompt-based									
<i>zero-shot</i> #0	51.1	62.6	56.8	35.4	51.8	43.6	49.7	64.7	57.2
FT #512	84.3	54.8	69.5	82.1	29.6	55.8	88.1	50.1	69.1
rFT #512	82.7	60.2	71.5	81.5	37.1	59.3	87.6	55.4	71.5
FT-fix18 #512	76.5	61.6	69.1	78.6	35.6	57.1	84.5	45.3	64.9
FT-fix12 #512	83.5	54.3	68.9	81.9	35.3	57.1	87.1	50.5	68.8
FT-fix6 #512	84.2	52.9	68.5	82.1	32.7	57.4	87.9	50.1	68.9
Classifier head									
FT #512	81.4	52.6	67.0	80.9	26.8	53.8	86.5	49.8	68.1

Table 2: Results of different strategies for finetuning prompt-based model (using $\#k$ examples). Models are evaluated against the in-distribution set and corresponding challenge sets. The zero-shot row indicates prompting results before finetuning. The *avg* columns report the average score on in-distribution and challenge datasets.

Baselines We compare regularized finetuning with another method that also minimally update the pretraining weights. We consider simple weight fixing of the first n layers of the pretrained model, where the n layers are frozen (including the token embeddings) and only the weights of upper layers and LM head are updated throughout the finetuning. In the evaluation, we use $n \in \{6, 12, 18\}$. We refer to these baselines as FT-fix n .

Results We evaluate all the considered finetuning strategies by taking their median performance after finetuning on 512 examples (for each label) and compare them with the original zero-shot performance. We report the results on Table 2, which also include the results of standard classifier head finetuning (last row). We observe the following: (1) Freezing the layers has mixed challenge set results, e.g., FT-fix18 improves over vanilla prompt-based finetuning on HANS and PAWS, but degrades Scramble and all in-distribution performances; (2) The L2 regularization strategy, rFT, achieves consistent improvements on the challenge sets while only costs small drop on the corresponding in-distribution performance, e.g., +6pp, +8pp, and +5pp on HANS, PAWS, and Scramble, respectively; (3) Although vanilla *prompt-based* finetuning performs relatively poorly, it still has an advantage over standard *classifier head* finetuning by +2.5pp, +2.0pp, and +1.0pp on the average scores of each in-distribution and challenge dataset pair.

Additionally, Figure 2 shows rFT’s improvement over vanilla prompt-based finetuning across data regimes on MNLI and HANS. We observe that the advantage of rFT is the strongest on the lexical overlap subset, which initially shows the highest

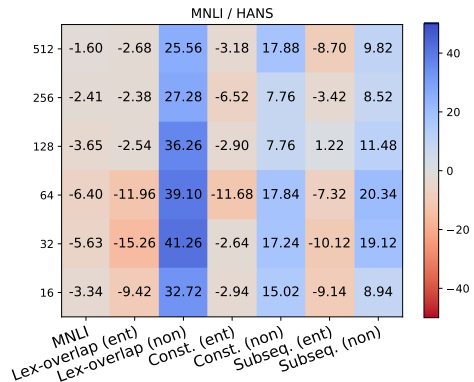


Figure 2: Relative difference between median accuracy of prompt-based finetuning across data regimes (y axis) with and without regularization on MNLI and HANS.

zero-shot performance. The results also suggest that the benefit of rFT peaks at mid data regimes (e.g., $K = 32$), before saturating when training data size is increased further. We also note that our results are consistent when we evaluate alternative prompt templates, or finetune for varying number of epochs.⁵ The latter indicates that the adoption of inference heuristics is more likely attributed to the amount of training examples rather than the number of learning steps.

4 Related Work

Inference Heuristics Our work relates to a large body of literature on the problem of “bias” in the training datasets and the ramifications to the resulting models across various language understanding tasks (Niven and Kao, 2019; Poliak et al., 2018; Tsuchiya, 2018; Gururangan et al., 2020). Previ-

⁵See Appendix B for the detailed results.

ous work shows that the artifacts of data annotations result in spurious surface cues, which gives away the labels, allowing models to perform well without properly learning the intended task. For instance, models are shown to adopt heuristics based on the presence of certain indicative words or phrases in tasks such as reading comprehension (Kaushik and Lipton, 2018), story cloze completion (Schwartz et al., 2017; Cai et al., 2017), fact verification (Schuster et al., 2019), argumentation mining (Niven and Kao, 2019), and natural language inference (Gururangan et al., 2020). Heuristics in models are often investigated using constructed “challenge datasets” consisting of counter-examples to the spurious cues, which mostly result in incorrect predictions (Jia and Liang, 2017; Glockner et al., 2018; Naik et al., 2018; McCoy et al., 2019). Although the problem has been extensively studied, most works focus on models that are trained in standard settings where larger training datasets are available. Our work provides new insights in inference heuristics in models that are trained in zero- and few-shot settings.

Heuristics Mitigation Significant prior work attempt to mitigate the heuristics in models by improving the training dataset. Zellers et al. (2019); Sakaguchi et al. (2020) propose to reduce artifacts in the training data by using adversarial filtering methods; Nie et al. (2020); Kaushik et al. (2020) aim at a similar improvement via iterative data collection using human-in-the-loop; Min et al. (2020); Schuster et al. (2021); Liu et al. (2019a); Rozen et al. (2019) augment the training dataset with adversarial instances; and Moosavi et al. (2020a) augment each training instances with their semantic roles information. Complementary to this, recent work introduces various learning algorithms to avoid adopting heuristics including by re-weighting (He et al., 2019; Karimi Mahabadi et al., 2020; Clark et al., 2020) or regularizing the confidence (Utama et al., 2020a; Du et al., 2021) on the training instances which exhibit certain biases. The type of bias can be identified automatically (Yaghoobzadeh et al., 2021; Utama et al., 2020b; Sanh et al., 2021; Clark et al., 2020) or by hand-crafted models designed based on prior knowledge about the bias. Our finding suggests that prompted zero-shot models are less reliant on heuristics when tested against examples containing the cues, and preserving this learned behavior is crucial to obtain more robust finetuned models.

Efficiency and Robustness Prompting formulation enables language models to learn efficiently from a small number of training examples, which in turn reduces the computational cost for training (Le Scao and Rush, 2021). The efficiency benefit from prompting is very relevant to the larger efforts towards sustainable and green NLP models (Moosavi et al., 2020b; Schwartz et al., 2020a) which encompass a flurry of techniques including knowledge distillation (Hinton et al., 2015; Sanh et al., 2019), pruning (Han et al., 2015), quantization (Jacob et al., 2018), and early exiting (Schwartz et al., 2020b; Xin et al., 2020). Recently, Hooker et al. (2020) show that methods improving compute and memory efficiency using pruning and quantization may be at odds with robustness and fairness. They report that while performance on standard test sets is largely unchanged, the performance of efficient models on certain underrepresented subsets of the data is disproportionately reduced, suggesting the importance of a more comprehensive evaluation to estimate overall changes in performance.

5 Conclusion

Our experiments shed light on the negative impact of low resource finetuning to the models’ overall performance that is previously obscured by standard evaluation setup. The results indicate that while finetuning helps prompt-based models to rapidly gain the *in-distribution* improvement as more labeled data are available, it also gradually increases models’ reliance on *surface heuristics*, which we show to be less present in the zero-shot evaluation. We further demonstrate that applying regularization that preserves pretrained weights during finetuning mitigates the adoption of heuristics while also maintains high in-distribution performances.

Acknowledgement

We thank Michael Bugert, Tim Baumgärtner, Jan Buchman, and the anonymous reviewers for their constructive feedback. This work is funded by the German Research Foundation through the research training group AIPHES (GRK 1994/1) and by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

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.
- Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. [Pay attention to the ending: strong neural baselines for the ROC story cloze task](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 616–622, Vancouver, Canada. Association for Computational Linguistics.
- Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. [Recall and learn: Fine-tuning deep pretrained language models with less forgetting](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online. Association for Computational Linguistics.
- Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2020. [Learning to model and ignore dataset bias with mixed capacity ensembles](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3031–3045, Online. Association for Computational Linguistics.
- Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel Gershman, and Noah D. Goodman. 2018. [Evaluating compositionality in sentence embeddings](#). In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018, Madison, WI, USA, July 25-28, 2018*. cognitivesciencesociety.org.
- Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.
- Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. [Towards interpreting and mitigating shortcut learning behavior of NLU models](#). *arXiv preprint arXiv:2103.06922*.
- R. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3:128–135.
- 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.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. [Breaking NLI systems with sentences that require simple lexical inferences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS’15*, page 1135–1143, Cambridge, MA, USA. MIT Press.
- He He, Sheng Zha, and Haohan Wang. 2019. [Unlearn dataset bias in natural language inference by fitting the residual](#). In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. [Distilling the knowledge in a neural network](#). In *NeurIPS Deep Learning and Representation Learning Workshop*.
- Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. [Characterising bias in compressed models](#). *arXiv preprint arXiv:2010.03058*.
- Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- 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.

- Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. [End-to-end bias mitigation by modelling biases in corpora](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.
- Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. [Learning the difference that makes a difference with counterfactually-augmented data](#). In *8th International Conference on Learning Representations, ICLR 2020, Virtual Conference, 26 April - 1 May, 2019*. OpenReview.net.
- Divyansh Kaushik and Zachary C. Lipton. 2018. [How much reading does reading comprehension require? a critical investigation of popular benchmarks](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.
- J. Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, J. Veness, G. Desjardins, Andrei A. Rusu, K. Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. 2017. [Overcoming catastrophic forgetting in neural networks](#). *Proceedings of the National Academy of Sciences*, 114:3521 – 3526.
- Teven Le Scao and Alexander Rush. 2021. [How many data points is a prompt worth?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.
- Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019a. [Inoculation by fine-tuning: A method for analyzing challenge datasets](#). 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 2171–2179, Minneapolis, Minnesota. 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. 2019b. [RoBERTa: A robustly optimized BERT pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. [Predicting inductive biases of pre-trained models](#). In *International Conference on Learning Representations, ICLR 2021, Virtual Conference, 3 May - 8 May, 2021*. OpenReview.net.
- James Martens. 2020. [New insights and perspectives on the natural gradient method](#). *Journal of Machine Learning Research*, 21(146):1–76.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. [Syntactic data augmentation increases robustness to inference heuristics](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.
- Nafise Sadat Moosavi, Marcel de Boer, Prasetya Ajie Utama, and Iryna Gurevych. 2020a. [Improving robustness by augmenting training sentences with predicate-argument structures](#). *arXiv preprint arXiv:2010.12510*.
- Nafise Sadat Moosavi, Angela Fan, Vered Shwartz, Goran Glavaš, Shafiq Joty, Alex Wang, and Thomas Wolf, editors. 2020b. *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*. Association for Computational Linguistics, Online.
- Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. [On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- 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, Yicheng Wang, and Mohit Bansal. 2019. [Analyzing compositionality-sensitivity of nli models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6867–6874.
- 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.
- Timothy Niven and Hung-Yu Kao. 2019. [Probing neural network comprehension of natural language arguments](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*,

- pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.
- Ohad Rozen, Vered Shwartz, Roei Aharoni, and Ido Dagan. 2019. [Diversify your datasets: Analyzing generalization via controlled variance in adversarial datasets](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 196–205, Hong Kong, China. Association for Computational Linguistics.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. [Winogrande: An adversarial winograd schema challenge at scale](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 8732–8740. AAAI Press.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter](#). *arXiv preprint arXiv:1910.01108*.
- Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M. Rush. 2021. [Learning from others’ mistakes: Avoiding dataset biases without modeling them](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. [Automatically identifying words that can serve as labels for few-shot text classification](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- 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, Adam Fisch, and Regina Barzilay. 2021. [Get your vitamin C! robust fact verification with contrastive evidence](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, 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.
- Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020a. [Green AI](#). *Communications of the ACM*, 63(12):54–63.
- Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. [The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.
- Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. 2020b. [The right tool for the job: Matching model and instance complexities](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6640–6651, Online. Association for Computational Linguistics.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. [BERT rediscovers the classical NLP pipeline](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Masatoshi Tsuchiya. 2018. [Performance impact caused by hidden bias of training data for recognizing textual entailment](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. [Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.
- Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. [Towards debiasing NLU models from unknown biases](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.
- 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.
- Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. [Learning which features matter: RoBERTa acquires a preference for linguistic generalizations \(eventually\)](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.
- 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.
- Ji Xin, Rodrigo Nogueira, Yaoliang Yu, and Jimmy Lin. 2020. [Early exiting BERT for efficient document ranking](#). In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 83–88, Online. Association for Computational Linguistics.
- Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordani. 2021. [Increasing robustness to spurious correlations using forgettable examples](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Yuan Zhang, Jason Baldridge, and Luheng He. 2019. [PAWS: Paraphrase adversaries from word scrambling](#). 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Experimental Details

Manual templates and mapping We use the following prompt templates and word-to-label mapping for the three tasks we evaluate on:

Template	Label Words
MNLI (manual): entailment, neutral, contradiction	
$s_1 ? [\text{MASK}], s_2$	Yes, Maybe, No
SNLI (manual): entailment, neutral, contradiction	
$s_1 ? [\text{MASK}], s_2$	Yes, Maybe, No
QQP (manual): duplicate, non-duplicate	
$s_1 [\text{MASK}], s_2$	Yes, No
MNLI (auto): entailment, neutral, contradiction	
$s_1 . [\text{MASK}], \text{you are right}, s_2$	Fine, Plus, Otherwise
$s_1 . [\text{MASK}], \text{you're right}, s_2$	There, Plus, Otherwise
$s_1 . [\text{MASK}] ! s_2$	Meaning, Plus, Otherwise

Table 3: Templates and label words used to finetune and evaluate on MNLI, SNLI, and QQP.

The last 3 rows are automatically generated templates and label words that are shown by Gao et al. (2021) to improve the few-shot finetuning further. Note that we use the corresponding task’s template when evaluating on the challenge datasets.

Challenge datasets We provide examples from each challenge datasets considered in our evaluation to illustrate sentence pairs that support or are against the heuristics. Table 4 shows examples for HANS, PAWS, and Scramble Test. Following McCoy et al. (2019), we obtain the probability for the *non-entailment* label by summing the probabilities assigned by models trained on MNLI to the *neutral* and *contradiction* labels. We use the *same-type* subset of Scramble Test (Dasgupta et al., 2018) which contain examples of both entailment (*support*) and contradiction (*against*) relations.

HANS details HANS dataset is designed based on the insight that the word overlapping between premise and hypothesis in NLI datasets is spuriously correlated with the *entailment* label. HANS consists of examples in which relying to this correlation leads to incorrect label, i.e., hypotheses are *not entailed* by their word-overlapping premises. HANS is split into three test cases: (a) **Lexical overlap** (e.g., “The doctor was paid by the actor” → “The doctor paid the actor”), (b) **Subsequence** (e.g., “The doctor near the actor danced” → “The actor danced”), and (c) **Constituent** (e.g., “If the artist slept, the actor ran” → “The artist

HANS (McCoy et al., 2019)	
premise	The artists avoided the senators that thanked the tourists.
hypothesis label	The artists avoided the senators. entailment (support)
premise	The managers near the scientist resigned.
hypothesis label	The scientist resigned. non-entailment (against)
PAWS (Zhang et al., 2019)	
S1	What are the driving rules in Georgia versus Mississippi?
S2	What are the driving rules in Mississippi versus Georgia?
label	duplicate (support)
S1	Who pays for Hillary’s campaigning for Obama?
S2	Who pays for Obama’s campaigning for Hillary?
label	non-duplicate (against)
Scramble Test (Dasgupta et al., 2018)	
premise	The woman is more cheerful than the man.
hypothesis label	The woman is more cheerful than the man. entailment (support)
premise	The woman is more cheerful than the man.
hypothesis label	The man is more cheerful than the woman. contradiction (against)

Table 4: Sampled examples from each of the challenge datasets we used for evaluation.

slept”). Each subset contains both entailment and non-entailment examples that always exhibit word overlap.

Hyperparameters Following Schick and Schütze (2021b,a), we use a fixed set of hyperparameters for all finetuning: learning rate of $1e^{-5}$, batch size of 8, and maximum length size of 256.

Regularization implementation We use the RecAdam implementation by Chen et al. (2020) with the following hyperparameters. We set the quadratic penalty λ to 5000, and the linear combination factor α is set dynamically throughout the training according to a sigmoid function schedule, where α at step t is defined as:

$$\alpha = s(t) = \frac{1}{1 + \exp(-k \cdot (t - t_0))}$$

where parameter k regulates the rate of the sigmoid, and t_0 sets the point where $s(t)$ goes above 0.5. We set k to 0.01 and t_0 to 0.6 of the total training steps.

B Additional Results

Standard CLS finetuning Previously, Gao et al. (2021) reported that the performance of standard

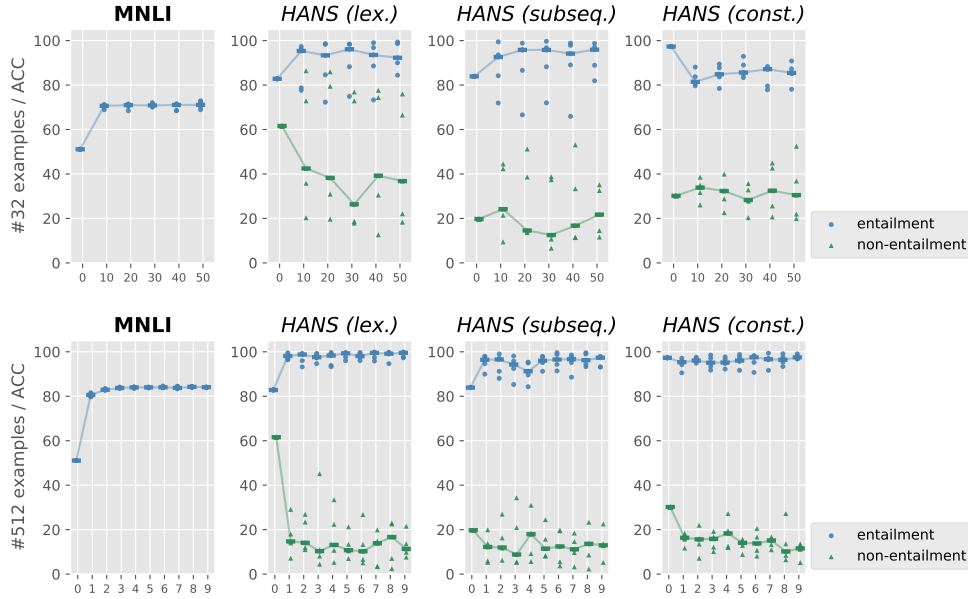


Figure 3: Results of prompt-based finetuning with varying number of epochs and fixed amount of training examples. Top: finetuning on 32 examples per label for epochs ranging from 10 to 50. Bottom: finetuning on 512 examples per label for 1 to 9 epochs. Both results show an immediate drop of non-entailment HANS performances which later stagnate even after more training steps.

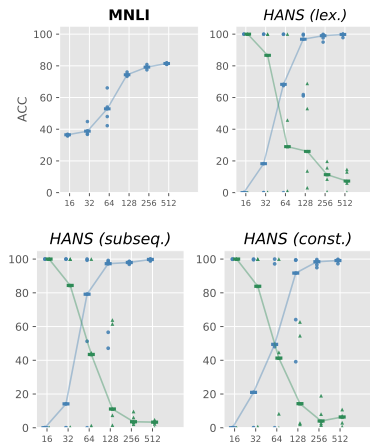


Figure 4: Results of non-prompt finetuning.

non-prompt finetuning with additional classifier head (CLS) can converge to that of prompt-based counterpart after certain amount of data, e.g., 512. It is then interesting to compare both finetuning paradigm in terms of their heuristics-related behavior. Figure 4 shows the results of standard finetuning using standard classifier head across varying data regimes on MNLI and the 3 subsets of HANS. We observe high instability of the results when only small amount of data is available (e.g., $K = 64$). The learning trajectories are consistent across the HANS subsets, i.e., they start making random predictions on lower data regime and im-

	MNLI (acc.)	
	IN	HANS
manual	51.1	62.6
manual Ft-#512	84.3	54.8
template-1	46.3	62.0
template-1 Ft-#512	84.2	53.2
template-2	49.9	61.3
template-2 Ft-#512	83.9	52.7
template-3	44.5	61.7
template-3 Ft-#512	84.4	56.0

Table 5: Evaluation results of different MNLI templates provided by Gao et al. (2021). Models are evaluated against both the in-distribution (IN) set and corresponding challenge set of MNLI.

mediately adopt heuristics by predicting almost all examples exhibiting lexical overlap as **entailment**. We observe that standard prompt-based finetuning still performs better than CLS finetuning, indicating that prompt-based approach provides good initialization to mitigate heuristics, and employing regularization during finetuning can improve the challenge datasets (out-of-distribution) performance further.

Impact of prompt templates A growing number of work propose varying prompt generation strategies to push the benefits of prompt-based predictions (Gao et al., 2021; Schick et al., 2020). We

	MNLi (acc.)			QQP (F1)			SNLI (acc.)		
	In.	HANS	avg.	In.	PAWS	avg.	In.	Scramble	avg.
zero-shot RoBERTa-large	51.1	62.6	56.8	35.4	51.8	43.6	49.7	64.7	57.2
FT #512 RoBERTa-large	84.3	54.8	69.5	82.1	29.6	55.8	88.1	50.1	69.1
zero-shot RoBERTa-base	48.2	58.1	53.15	37.3	41.5	39.4	48.8	56.4	52.6
FT #512 RoBERTa-base	74.4	49.9	62.15	79.0	26.9	52.9	83.7	48.5	66.1
zero-shot BERT-large-uncased	45.3	55.4	50.4	34.7	33.4	34.0	41.5	54.8	48.1
FT #512 BERT-large-uncased	70.9	50.0	60.4	77.3	26.3	51.8	79.9	49.5	64.7
zero-shot BERT-base-uncased	43.5	55.9	49.7	40.7	50.8	45.8	38.7	49.9	44.3
FT #512 BERT-base-uncased	63.2	50.1	56.65	73.9	29.1	51.5	74.5	42.6	58.5

Table 6: Evaluation results of different pretrained language models. Models are evaluated against both the in-distribution (**In.**) set and corresponding challenge set.

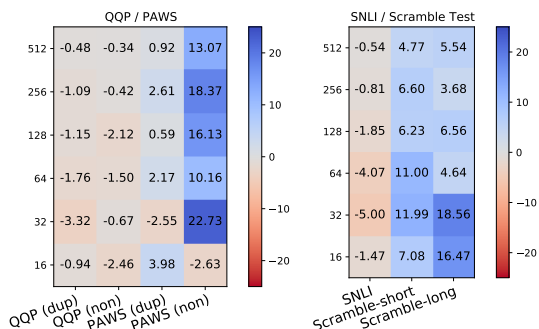


Figure 5: Relative difference between median accuracy of prompt-based finetuning across data regimes (y axis) with and without regularization on QQP / PAWS and SNLI / Scramble Test.

therefore questions whether different choices of templates would affect the model’s behavior related to lexical overlap. We evaluate the 3 top-performing templates for MNLi that are obtained automatically by Gao et al. (2021) and show the results in Table 5. We observe similar behavior from the resulting models over the manual prompt counterpart, achieving HANS average accuracy of around 62% and below 55% on zero-shot and finetuning with 512 examples.

Impact of learning steps We investigate the degradation of the challenge datasets performance as the function of the number of training data available during finetuning. However, adding more training examples while fixing the number of epochs introduces a confound factor to our finding, which is the number of learning steps to the model’s weights. To factor out the number of steps, we perform similar evaluation with a fixed amount of training data and varying number of training epochs. On 32 examples per label, we finetune for 10, 20, 30, 40, and 50 epochs. Additionally,

we finetune on 512 examples for 1 until 10 epochs to see if the difference in learning steps results in different behavior. We plot the results in Figure 3. We observe that both finetuning settings result in similar trajectories, i.e., models start to adopt heuristics immediately in early epochs and later stagnate even with increasing number of learning steps. For instance, finetuning on 32 examples for the same number of training steps as in 512 examples finetuning for 1 epoch still result in higher overall HANS performance. We conclude that the number of finetuning data plays a more significant role over the number of training steps. Intuitively, larger training data is more likely to contain more examples that disproportionately support the heuristics; e.g. NLI pairs with lexical overlap are rarely of non-entailment relation (McCoy et al., 2019).

Regularization across data regimes Figure 5 shows the results improvement of L2 weight regularization over vanilla prompt-based finetuning on QQP and SNLI. Similar to results in MNLi/HANS, the improvements are highest on mid data regimes, e.g., 32 examples per label.

Impact of pretrained model In addition to evaluating RoBERTa-large, we also evaluate on other commonly used pretrained language models based on transformers such as RoBERTa-base, BERT-base-uncased, and BERT-large-uncased. The results are shown in Table 6. We observe similar pattern across PLMs, i.e., improved in-distribution scores come at the cost of the degradation in the corresponding challenge datasets.