

Adapting Language Models for Zero-shot Learning by Meta-tuning on Dataset and Prompt Collections

Ruiqi Zhong Kristy Lee* Zheng Zhang* Dan Klein
Computer Science Division, University of California, Berkeley
{ruiqi-zhong, kristylee, zhengzhang1216, klein}@berkeley.edu

Abstract

Large pre-trained language models (LMs) such as GPT-3 have acquired a surprising ability to perform zero-shot learning. For example, to classify sentiment without any training examples, we can “prompt” the LM with the review and the label description “*Does the user like this movie?*”, and ask whether the next word is “*Yes*” or “*No*”. However, the next word prediction training objective is still **misaligned** with the target zero-shot learning objective. To address this weakness, we propose meta-tuning, which directly optimizes the zero-shot learning objective by fine-tuning pre-trained language models on a collection of datasets. We focus on classification tasks, and construct the meta-dataset by aggregating 43 existing datasets and annotating 441 label descriptions in a question-answering (QA) format. When evaluated on unseen tasks, meta-tuned models outperform a same-sized QA model and the previous SOTA zero-shot learning system based on natural language inference. Additionally, increasing parameter count from 220M to 770M improves AUC-ROC scores by 6.3%, and we forecast that even larger models would perform better. Therefore, measuring zero-shot learning performance on language models out-of-the-box might underestimate their true potential, and community-wide efforts on aggregating datasets and unifying their formats can help build models that answer prompts better.

1 Introduction

The goal of zero-shot classification (ZSC) is to classify textual inputs using label descriptions without any examples (Yin et al., 2019). Large language models - whose only training objective is to predict the next word given the context - have acquired a surprising ability to perform ZSC (Radford et al., 2019; Brown et al., 2020; Le Scao and Rush, 2021). For example, to classify whether the sentence “*This movie is amazing!*” is positive, we

can prompt the language model with the context “*Review: This movie is amazing! Positive Review? ___*”, and check whether the next word is more likely to be “*Yes*” or “*No*” (Zhao et al., 2021). To convert ZSC into a language modeling (LM) task that an LM model is likely to perform well, many recent works focus on finding better prompts (Shin et al., 2020; Schick and Schütze, 2020a,b; Gao et al., 2021).

However, the LM training objective is correlated but still misaligned with the target objective to answer prompts. Our work addresses this weakness by directly optimizing the zero-shot classification objective through fine-tuning (Section 4). This requires us to 1) unify different classification tasks into the same format, and 2) gather a collection of classification datasets and label descriptions (prompts) for training (Section 2). Since we fine-tune our model on a meta-dataset, we name our approach meta-tuning.

We focus on binary classification tasks and unify them into a “*Yes*”/“*No*” QA format (Clark et al., 2019; McCann et al., 2018), where the input is provided as the context and the label information is provided in the question (Figure 1 (a)). Using this format, we gathered a diverse set of classification datasets from 43 different sources listed on Kaggle, SemEval, HuggingFace, and other papers. These tasks range from hate speech detection, question categorization, sentiment classification to stance classification, etc, and the genre ranges from textbooks, social media, to academic papers, etc. In total, these datasets contain 204 unique labels, and we manually annotated 441 label descriptions (Figure 2).

To evaluate ZSC, we need to define what counts as a task that the model has not seen during training time. While prior work considers different notions of “unseen” by disallowing the same label or the same dataset to appear during training, our work defines “unseen” more harshly by dis-

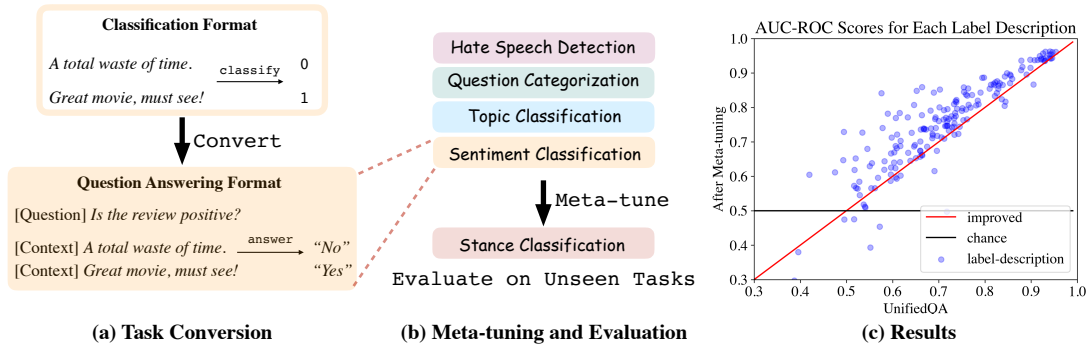


Figure 1: **(a)** We convert the format to question answering. We manually annotate label descriptions (questions) ourselves (Section 2). **(b)** We finetune the UnifiedQA (Khashabi et al., 2020) model (with 770 M parameters) on a diverse set of tasks (Section 4), and evaluate its 0-shot classification (ZSC) performance on an unseen task. **(c)** For each label description (question) we evaluate the AUC-ROC score for the “Yes” answer, and each dot represents a label description (Section 3). The x -value is the ZSC performance of UnifiedQA; the y -value is the performance after meta-tuning. In most cases, the y -value improves over the x -value (above the red line) and is better than random guesses (above the black line) by a robust margin (Section 5).

allowing similar datasets. For example, we consider AG News topic classification dataset (Zhang et al., 2015) and the topic classification dataset from Yin et al. (2019) to be similar, even though their sources and label spaces are different.

Meta-tuning improves ZSC over UnifiedQA for most labels (Figure 1 (c)). Moreover, larger models are better, and hence we forecast that meta-tuning would work for even larger models. We also find that the performance can be slightly improved by training on datasets similar to the test dataset, ensembling different label descriptions, or initializing with a QA model (Section 5.1). All of our findings reliably hold under different robustness checks (Section 5.2), and our approach outperforms the previous SOTA Yin et al. (2019) using the same pre-training method (Section 5.3).

Our results suggest two promising future directions (Section 6). First, large language models’ (e.g. GPT-3) potential for zero-shot learning, as currently measured by context-prompting, might have been broadly underestimated; meta-tuning might significantly improve their performance. Second, community-wide efforts on aggregating and unifying datasets can scale up training and evaluation for zero-shot learning models. On the flip side, however, the meta-tuning approach might incentivize providers of LM inference APIs to collect prompts from users, hence potentially leading to security, privacy, and fairness concerns at a greater scale (Section A).

Contributions To summarize, we 1) curate a dataset of classification datasets with expert an-

notated label descriptions. 2) demonstrate a simple approach to train models to perform zero-shot learning, and 3) identify several factors that improve performance; in particular, larger pretrained models are better.¹

2 Data

We gather a wide range of classification datasets and unify them into the “Yes”/“No” question answering format for binary classification. Then we group similar datasets together to determine what counts as unseen tasks during evaluation.

Gathering classification datasets We collect classification datasets from Kaggle², Huggingface (Wolf et al., 2020), SemEval³, and other papers. We looked through these sources and only considered English classification datasets. We also skipped the tasks that we felt were already better represented by other datasets in our collection. Then we manually examined a few examples in each remaining dataset to make sure it seemed plausibly clean.

The goals of these classification datasets include, but are not limited to sentiment classification (IMDB Reviews, Maas et al. (2011a)), topic classification (AG News, Zhang et al. (2015)), grammaticality judgement (CoLA, Warstadt et al. (2018)), paraphrase detection (QQP⁴), definition

¹Code and data available here: <https://github.com/ruiqi-zhong/Meta-tuning>.

²<https://www.kaggle.com>

³<https://semeval.github.io>

⁴<https://www.kaggle.com/c/>

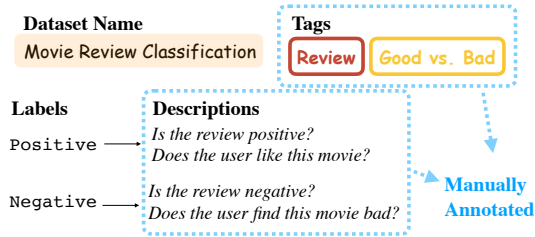


Figure 2: For each dataset, we annotate 1-3 descriptions for each label in the form of questions, and associate it with a set of property tags. The question answering format can be seen in Figure 1 (a).

detection (SemEval 2020 Task 6, Spala et al. (2019)), stance classification (SemEval 2016 Task 6, Mohammad et al. (2016)), etc. The genre includes academic papers, reviews, tweets, posts, messages, articles, and textbooks. The comprehensive list of datasets is in Appendix B. Overall, we aim for a high diversity of tasks and genres by building upon what the broader research community has studied. Our approach is complementary to that of Weller et al. (2020), which asks turkers to generate tasks, and that of Mishra et al. (2021), which generates tasks by decomposing existing templates used to construct reading comprehension datasets. The concurrent work of Bragg et al. (2021) unifies the evaluation for few-shot learning; their zero-shot evaluation setup is the closest to ours, and they used templates and verbalizers (Schick and Schütze, 2020a) to specify the semantics of a task.

Some of our datasets are noisy and not peer reviewed, or contain tasks that are too complicated (e.g. Multi-NLI, Williams et al. (2018)) for ZSC. To make our evaluation more informative, we only include them for training but not testing. We make these decisions before running our experiments in Section 5 to prevent selection bias.

Unifying the dataset format We convert each classification dataset into a “Yes”/“No” question answering format and provide label information in the question. For each label, we annotate 1-3 questions. If the label is null (for example, a text that does not express a particular emotion in an emotion classification dataset), we skip this label. Three of the authors⁵ manually annotated 441 questions for 204 unique labels, and each question

quora-question-pairs

⁵One of them is a graduate student and the other two are undergrads; all of them study Computer Science and have taken an NLP class.

Are these two questions asking for the same thing?
Does the tweet contain irony?
Is this news about world events?
Does the text contain a definition?
Is the tweet an offensive tweet?
Is the text objective?
Does the question ask for a numerical answer?
Is the tweet against environmentalist initiatives?
Is this abstract about Physics?
Does the tweet express anger?
Does the user dislike this movie?
Is the sentence ungrammatical?
Is this text expressing a need for evacuation?
Is this text about Society and Culture?
Is this a spam?

Figure 3: Some example manually annotated label descriptions (questions). Three of the authors manually wrote 441 questions in total, and each of them is proofread by at least another author.

is proofread by at least another author. See Figure 2 for a concrete example, and Figure 3 for some representative label descriptions.

Additionally, some datasets contain thousands of labels (Chalkidis et al., 2019; Allaway and McKeown, 2020). In this case, we use templates to automatically synthesize label descriptions and exclude them from evaluation.

Grouping similar datasets Our goal is to test the models’ ability to generalize to tasks that are different enough from the training tasks. Therefore, at test time, we need to exclude not only the same dataset that appeared in the meta-tuning phase, but also ones that are similar.

This poses a challenge: whether two datasets perform the same task involves subjective opinion, and there is no universally agreed definition. On one extreme, most datasets can be counted as dissimilar tasks, since they have different label spaces and input distributions. On the other extreme, all datasets can be considered the same task, since they can all be unified into the question answering format.

To tackle this challenge, we create a set of tags, each describing a dataset property. The set of tags includes *domain classification*, *article*, *emotion*, *social-media*, etc, and the full set of them can be seen in Appendix C. Then we define the

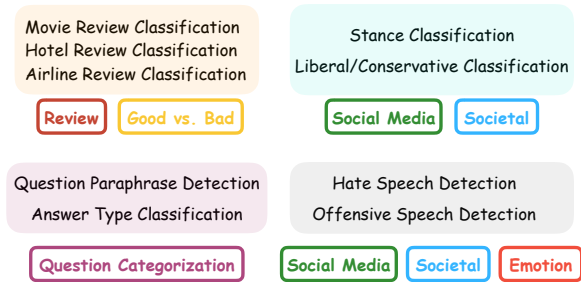


Figure 4: Example dataset groups based on tags. We never train and test on datasets from the same group, e.g. train on hotel review and test on movie review.

two datasets to be similar if they are associated with the same set of tags, and prohibit the model to learn from one and test on the other. For example, our work considers the topic classification datasets from Zhang et al. (2015) (AG News) and Yin et al. (2019) to be similar since they both classify topics for articles, even though their sources and label spaces are different. Some example dataset groups can be seen in Figure 4.

Nevertheless, our procedure is not bullet-proof and one can argue that our notion of unseen tasks, though harsher than prior works (Yin et al., 2019; Pushp and Srivastava, 2017), is still lenient. Therefore, as additional robustness checks, for each dataset we evaluate, we manually identify and list the most relevant dataset that is allowed during training in Appendix F. For example, the most relevant dataset to the IMDB review sentiment classification dataset is the emotion classification dataset from Yin et al. (2019), which classifies the input text into 9 emotions, such as “joy”, “surprise”, “guilt”, etc. We consider the emotion classification dataset to be relevant, since sentiment classification often involves identifying emotions. However, one can also argue that they are different tasks: their input and label spaces are different, and sadness can be caused by a great tragedy, or a bad movie that wastes the users’ time. The comprehensive list of label descriptions grouped by dataset similarity is in Appendix D.

In total, we spend around 200 hours to collect this dataset. This time estimate includes skimming through the dataset repos and recent NLP papers, writing programs to download the datasets and unify their format, annotating label descriptions, performing quality controls, and documenting the collection process.

3 Metrics

To reliably aggregate performance across different datasets and present as much information as possible, we report a set of descriptive statistics and provide visualizations whenever we compare two models. We generally do not reduce a model’s performances on different datasets into one scalar quantity and compare this number only.

Descriptive statistics For each label description (question), we calculate the AUC-ROC score⁶ by treating the “Yes” answer as the positive class. After calculating the AUC-ROC score for each label, we calculate the following set of descriptive statistics to compare two models. Suppose that model Y is hypothetically better than X . Denoting Δ as the change of AUC-ROC of a label description from X to Y , we can summarize how Δ is distributed across the set of label descriptions with the following statistics:

- $\mathbb{E}[\Delta]$: the average change in AUC-ROC.
- $\mathbb{P}[\Delta > t]$: the fraction of label descriptions where the change is over the threshold t .
- $\mathbb{P}[\Delta < -t]$: the fraction of label descriptions where the change is less than $-t$.
- $Std[\Delta]$: the standard deviation of the change.

In the main paper, we weight each label description equally in this distribution to calculate the above statistics. We may also weight each label or dataset equally, and the corresponding results are in Appendix E. To make sure our conclusions are robust, we consider one model to be better only when $\mathbb{E}[\Delta] > 0$ and $\mathbb{P}[\Delta > t] > \mathbb{P}[\Delta < -t]$ for all $t \in \{1\%, 5\%, 10\%\}$, under all three types of weighting. In other words, we claim that one model is better than the other only when 12 conditions simultaneously hold.

Visualizations We use scatter plots to visualize and compare the performance of two models, where each dot represents a label description, its x-value represents the AUC-ROC score of the model X , and its y-value represents that of Y . If most dots are above the identity line $y = x$, the model Y is better than X .

The descriptive statistics and the visualizations are explained in Figure 5.

⁶We do not evaluate F-score or accuracy, since they are very sensitive to the decision cutoff, and usually additional calibration is needed (Zhao et al., 2021).

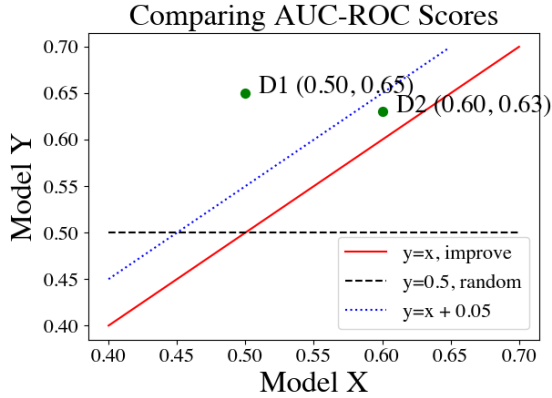


Figure 5: Each dot represents a label description, and its x/y -value each represents the performance of model X/Y (measured by AUC-ROC score). For example, on label description $D1$, model X/Y has AUC-ROC score 0.5/0.65. If the dot is above the black line ($y = 0.5$), model Y is performing better than random guesses. If the dot is above the red line ($y = x$), model Y is better than model X . Since one out of two dots are above $y = x + 0.05$, we have $\mathbb{P}[\Delta > 5\%] = 0.5$.

4 Model

Architecture We format the inputs to the model in the same way as UnifiedQA (Khashabi et al., 2020), which concatenates the context to the question and adds a “[SEP]” token in between. Then we feed the concatenated input into the T5 encoder and produce the answer score by normalizing the “Yes”/“No” probability of the first decoded token. Unless otherwise noted, we initialize our model with T5-Large (770 Million parameters). We sometimes compare to or initialize with the UnifiedQA model (Khashabi et al., 2020), which is trained on a wide range of question answering datasets. For a fair comparison, we use the UnifiedQA model initialized with T5-Large as well. To meta-tune non-Seq2Seq pre-trained models, such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), we add an MLP layer on top of the pooled output/“[CLS]” token to classify between “Yes”/“No”. We leave the improvement on model architectures (Ye and Ren, 2021; Li and Liang, 2021; Lester et al., 2021) and training objectives (Murty et al., 2021; Yin et al., 2020) for future work.

Meta-tuning We create a training distribution that balances between datasets, label descriptions, and “Yes”/“No” answers. To create the next training datapoint for meta-tuning, we select a

dataset from the training split uniformly at random (u.a.r.); then we select a label description (question) u.a.r. and with 50% probability select a textual input with the answer “Yes”/“No”. To prevent over-fitting, we do not train on any combination of label description and textual input twice. Unless otherwise noted, we meta-tune the model for 5000 steps and use batch size 32. We did not tune any hyper-parameters or training configurations since they work well during our first attempt. To evaluate ZSC performance on each dataset, we leave out one group of similar datasets as the evaluation set and train on the rest. Altogether, the experiments take around 250 GPU hours on Quadro 8000.

5 Results

5.1 Hypotheses and Conclusions

We investigate and validate the following hypotheses, sorted by importance in descending order.

- Meta-tuned models outperform general question answering models in zero-shot classification.
- Larger pre-trained models are better.
- Pre-training does the heavy lifting.
- Performance can be improved by training on similar datasets, initializing with a QA model, or ensembling label descriptions.
- Early stopping is crucial to performance.

Meta-tuned models are better. We compare a meta-tuned T5-Large model (770 M parameters)⁷ with the same-sized UnifiedQA model (Khashabi et al., 2020) out of the box. Relevant descriptive statistics can be seen in the first row of Table 1 and Figure 6 (a). Adapting the model for ZSC improves the average AUC-ROC by 3.3%.

Larger pre-trained models are better. We compare T5-Base (220 Million parameters) against T5-Large (770 M). The statistics can be seen in the second row of Table 1 and Figure 6 (b). Increasing the model size from 220 M to 770M improves the average AUC-ROC by 6.3%.

⁷This model is initialized with T5, not UnifiedQA.

| | $\mathbb{E}[\Delta]$ | $\mathbb{P}[\Delta > 1\%]$ | $\mathbb{P}[\Delta < -1\%]$ | $Std(\Delta)$ |
|---------------------------|----------------------|----------------------------|-----------------------------|---------------|
| Meta-tuned vs. UnifiedQA | 3.3% | 59.5% | 28.1% | 9.5% |
| Larger | 6.3% | 75.1% | 15.1% | 8.1% |
| Pre-trained vs. Random | 23.8% | 95.7% | 3.2% | 14.0% |
| Train on Similar | 0.7% | 43.8% | 20.5% | 3.2% |
| Ensemble Descriptions | 0.7% | 28.9% | 16.8% | 3.1% |
| Initialize with UnifiedQA | 1.1% | 54.1% | 24.3% | 6.9% |

Table 1: The statistics used to compare two models, introduced in Section 3. The larger $\mathbb{E}[\Delta]$ and the difference between $\mathbb{P}[\Delta > 1\%]$ and $\mathbb{P}[\Delta < -1\%]$, the better. Row 1 finds that a meta-tuned model is better than UnifiedQA; row 2 finds that the larger model is better; row 3 finds that pre-training does the heavy lifting; row 4, 5, and 6 finds that the performance can be improved by training on similar datasets, ensembling label descriptions, and initializing with a UnifiedQA model. Note that $Std(\Delta)$ is the standard deviation of individual descriptions, not the standard deviation of the estimated mean. Due to space constraint we only show $t = 1\%$ in this table.

Pre-training does the heavy lifting. In Figure (c) and the third row of Table 1, we compare pre-trained and random initializations, where the latter cannot beat the random baseline (average AUC-ROC 0.503). Hence, meta-tuning alone is far from enabling the model to perform ZSC. An intuitive interpretation is that the model already “knows” how to perform ZSC after pre-training under the LM objective, and learns how to use this knowledge during meta-tuning.

Training on similar datasets improves performance. Unlike before, we no longer avoid training on similar datasets from the same group. Instead, we perform straightforward leave-one-out cross-validation. The statistics can be seen in the fourth row of Table 1 and Figure 6 (d), and it improves the average AUC-ROC by 0.7%. The performance gain is not as significant as increasing the model size or adapting for ZSC. We conjecture that it is because we have not collected enough datasets; otherwise, there might be more similar datasets, hence improving ZSC performance.

Ensembling label descriptions improves performance. Instead of asking the model a single question for each label and obtain the probability of the answer being “Yes”, we can average the probability obtained by asking multiple questions with the same meaning. This approach is different from traditional ensembling, which typically needs to store/train multiple models to average across them. The fifth row of Table 1 and Figure 6 (e) verifies that ensembling descriptions improves performance slightly (0.7% AUC-ROC score).

Initializing with UnifiedQA improves performance. Figure 6 (f) and the sixth row of Table 1

compare the UnifiedQA against against the T5 initialization. Initializing with UnifiedQA improves average AUC-ROC by 1.1%.

Early stopping is crucial to performance. If we train the model for too long, the model might simply “memorize” that certain label descriptions correspond to certain training tasks, and the performance on unseen tasks may drop. To explore this possibility, we meta-tune our models for 100K steps, which is 20 times as long as our default setting and encourages the model to memorize the training tasks. We then evaluate them on the three benchmark zero-shot classification datasets by Yin et al. (2019) (which we describe in more details in the next section). We calculate the average AUC-ROC across all label descriptions for each of the 3 datasets, and plot them in Figure 7.

The performance decreases⁸ as training continues. On the other hand, however, the performance drop of 3% in AUC-ROC is not fatal and the model’s performance is still much better than random guesses.

5.2 Robustness Checks

We examine a series of additional results to make sure our conclusions are robust. The observed improvements in Table 1 and Figure 6 might be caused by the improvement of a small number of labels that are annotated with more descriptions, or by the improvement on a dataset with more distinct labels. Appendix E.1 compares the performance by assigning equal weights to each label/datasets.

To provide additional supporting evidence for

⁸Kendall rank correlation coefficients are negative with $p < 0.005$ for topic and situation classification

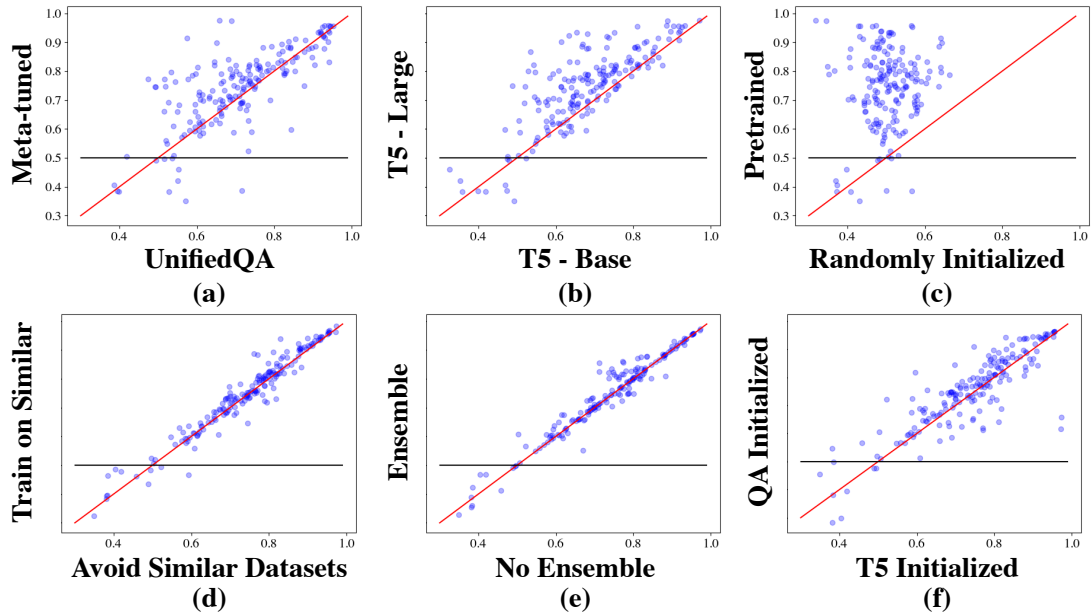


Figure 6: The interpretation of these figures can be seen in Figure 5. (a) compares a meta-tuned model (y) against UnifiedQA (x); (b) compares T5-Large (770 M parameters) against T5-base (220M); (c) compares the T5 pre-trained initialization against the random initialization; (d), (e), and (f) investigate whether performance can be improved by training on similar datasets, ensembling different label descriptions (questions), and initializing with UnifiedQA. **Conclusion:** Since most dots are above the red line $y = x$ for all 6 figures and above the random guess baseline ($y = 0.5$) by a robust margin, all conclusions listed at the beginning of Section 5 hold.

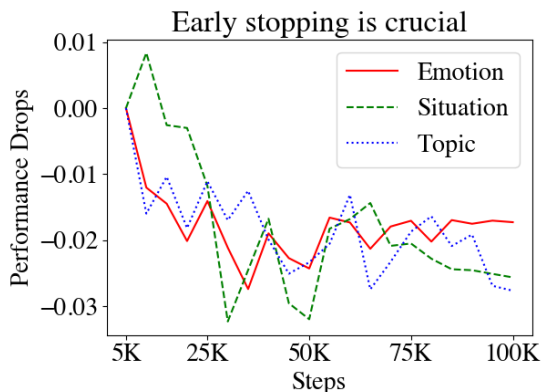


Figure 7: Each curve corresponds to the models’ performance on a dataset from Yin et al. (2019). x -value is the number of training steps; y -value is the average AUC-ROC score across all label descriptions, relative to the value at step 5000. Training for too long decreases performance on unseen tasks.

our forecast that larger models are better, Appendix E.2 compares a 60M-parameter model against a 220M-parameter model, and finds that the latter is much better. One concern, however, is that our models are initialized with T5 (Raffel et al., 2019), which is trained on the open web and might have seen the datasets we gathered. There-

| Model | emotion | situation | topic |
|-------------------|-------------|-------------|-------------|
| Yin et al. (2019) | 25.2 | 38.0 | 52.1 |
| Meta-tuned | 28.2 | 48.4 | 54.3 |

Table 2: “Prior” means the best performing system from Yin et al. (2019) for each dataset; “Meta-tuned” means meta-tuning on RoBERTa. Our approach is better on all three datasets.

fore, larger models might be better simply because they are better at memorization (Sagawa et al., 2020). Appendix E.3 addresses this by showing that larger models are also better with BERT initialization (Devlin et al., 2019), which is trained on Wikipedia and Book Corpus (Zhu et al., 2015).

We also report the models’ performance on each dataset for readers’ reference in Appendix G.

5.3 Comparison with Yin et al. (2019)

This section shows that our approach has higher performance than the zero-shot classification system built by Yin et al. (2019). Their system ensembles several natural language inference models based on RoBERTA-Large (355M parameters, Liu et al. (2020)), and another model trained to categorize Wikipedia articles. It was evaluated on three classification datasets:

- topic (10-way): classifies article domains, such as *family & relationship*, *education*, *sports*, etc. The metric is accuracy.
- emotion (10-way): classifies emotion types, such as *joy*, *anger*, *guilt*, *shame*, etc. The metric is label-weighted F1.
- situation (12-way): classifies disaster situations, e.g. *regime change*, *crime & violence*, and the resource they need, e.g. *search & rescue*. The metric is label-weighted F1.

We use the exact same evaluation metrics as in Yin et al. (2019), and the same label resolution strategy when the model answers “Yes”⁹ for multi-label classification. Concretely, when the model predicts “Yes” on multiple labels, the one with the highest probability is selected. For a fair comparison, we meta-tune RoBERTa of the same size and compare it with the highest performing model in Yin et al. (2019) for each of the three datasets.

The results are in Table 2, and our model has higher performance across all 3 datasets using the same pre-training method.

6 Discussion and Future Directions

Main takeaways We construct a dataset of classification datasets to adapt the language model for zero-shot classification via meta-tuning. The adapted model outperforms a general-purpose question answering model and the prior state of the art based on natural language inference. We forecast that meta-tuning would be more effective on larger models, and the current engineering ceiling for zero-shot learning might have been broadly under-estimated.

Aggregating and unifying datasets The main bottleneck of our research is to manually gather a wide range of datasets and unify their format. The difficulties are: 1) we need to brainstorm and review the NLP literature extensively to decide what new tasks to look for; 2) different datasets encode their data in different formats, and we need to write programs manually for each of them to convert to the desired format; 3) it is hard to tell the quality of a dataset purely by its provenance, and sometimes we need to examine the dataset manually. If we as a community can aggregate and unify datasets better, we could potentially train and evaluate zero-shot learning models at a larger scale.

⁹or “Entailment” for natural language inference models.

Meta-tuning as a probe There is a growing interest in measuring the intelligence (Hendrycks et al., 2021a,b) or the few-shot learning ability (Brown et al., 2020) of large language models like GPT-3. However, since these models are not adapted to answer those prompts (Holtzman et al., 2021), we suspect that its knowledge and true potential to perform few-shot learning is much higher than reported. Since pre-training does the heavy lifting and meta-tuning is unlikely to provide additional ZSC ability to the model, we can potentially first use meta-tuning as a probe to make them adapted to answering prompts before measuring their performance.

Still, to make this methodology rigorous, interpreting and controlling the strength of the probes will be an important future direction (Hewitt and Liang, 2019). For example, if the training set contains a prompt that is too similar to the prompt to be tested, the probe will be meaningless.

Beyond Shallow Correlations One possibility is that the model only learns shallow statistical correlations from meta-tuning rather than “more sophisticated reasoning skills”. For example, the word “exciting” might occur in positive reviews more. This is unlikely, given that larger models are consistently better than smaller or randomly initialized ones. To explain this performance gap, larger models must have learned to use more complicated features during meta-tuning.

Relation to Meta/Multitask-Learning Our method is closely related to, but different from meta-learning (Yin, 2020; Murty et al., 2021) and multi-task learning (Ye et al., 2021; Aghajanyan et al., 2021). Both meta-learning and multitask-learning typically involve at least a few examples from the target task; in our setup, however, the model does not learn from any target task examples. The “meta” in our name does not mean “meta-learning”, but reflects the fact that our model learns from a meta-dataset of tasks.

Nevertheless, our framework can be easily adapted to a few-shot learning setup, which enables the language model to learn to learn from in-context examples (see below). Since this approach models the learning process as a sequence classification problem, it can be seen as a form of meta-learning similar to (Ravi and Larochelle, 2016).

Annotating Prompts Three of our authors annotated the label descriptions. Since they are all

Computer Science major students who understand machine learning and natural language processing, they might not be representative of the final user population of this ZSC application. Annotating prompts that match the target user distribution will be an important research direction.

Additionally, shorter and more natural descriptions sometimes fail to capture the exact semantics of the label. For example, in Yin et al. (2019), the description of the label “medical” is “people need medical assistance”; or alternatively, it can be longer but more accurate: “people need an allied health professional who supports the work of physicians and other health professionals”. How to scalably generate more accurate and detailed label descriptions without expert efforts will be another future direction.

Optimizing Prompts Our work is complementary to recent works that optimize the prompts to achieve better accuracy. Even if our meta-tuned model is specialized in answering prompts, it might still react very differently towards different prompts. For example, in the stance classification dataset (Barbieri et al., 2020), we annotated two label descriptions (prompts) for the same label: “Does this post support atheism?” and “Is the post against having religious beliefs?”. They have similar meanings, but the former has much lower accuracy than the later. We conjecture that this is because the model cannot ground abstract concepts like “atheism”.

Other extensions We conjecture that meta-tuning can be extended to more diverse tasks beyond zero-shot binary classification. To extend to multi-label classification, we need to develop a procedure to resolve the labels when the model predicts positive for more than one labels. To extend to few-shot learning, we need to increase the context length to fit several training examples into the input, which requires a larger context window and hence more computational resources. To extend to other sequence generation tasks, we need to collect a wide range of diverse sequence generation tasks to meta-tune the model, such as machine translation, summarization, free-form question answering, grammar correction, etc.

Acknowledgements

We thank Eric Wallace for his feedbacks throughout the project. We thank Steven Cao, David

Gaddy, Haizhi Lai, Jacob Steinhardt, Kevin Yang and anonymous reviewers for their comments on the paper.

References

- Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. *arXiv preprint arXiv:2101.11038*.
- Emily Allaway and Kathleen McKeown. 2020. [Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.
- Tiago Almeida, José María Gómez Hidalgo, and Tiago Pasqualini Silva. 2013. Towards sms spam filtering: Results under a new dataset. *International Journal of Information Security Science*, 2(1):1–18.
- Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. [TweetEval: Unified benchmark and comparative evaluation for tweet classification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. [SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. [Man is to computer programmer as woman is to homemaker? debiasing word embeddings](#). In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. Flex: Unifying evaluation for few-shot nlp. *arXiv preprint arXiv:2107.07170*.
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020.

- Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.
- Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. [Large-scale multi-label text classification on EU legislation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [BoolQ: Exploring the surprising difficulty of natural yes/no questions](#). 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 2924–2936, Minneapolis, Minnesota. 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.
- 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.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. [Aligning AI With Shared Human Values](#). *arXiv e-prints*, page arXiv:2008.02275.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. [Aligning {ai} with shared human values](#). In *International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.
- Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. [Surface form competition: Why the highest probability answer isn’t always right](#). *CoRR*, abs/2104.08315.
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hananeh Hajishirzi. 2020. [UNIFIEDQA: Crossing format boundaries with a single QA system](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.
- 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.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). *arXiv preprint arXiv:2101.00190*.
- Xin Li and Dan Roth. 2002. [Learning question classifiers](#). In *COLING 2002: The 19th International Conference on Computational Linguistics*.
- Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2019. [Robust neural machine translation with joint textual and phonetic embedding](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3044–3049, Florence, Italy. 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. 2020. [Ro{bert}a: A robustly optimized {bert} pretraining approach](#).
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011a. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011b. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

- Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.
- Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. [SemEval-2019 task 8: Fact checking in community question answering forums](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 860–869, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Natural instructions: Benchmarking generalization to new tasks from natural language instructions. *arXiv preprint arXiv:2104.08773*.
- Rishabh Misra. 2019. [Imdb spoiler dataset](#).
- Rishabh Misra, Mengting Wan, and Julian McAuley. 2018. Decomposing fit semantics for product size recommendation in metric spaces. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 422–426. ACM.
- Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. [SemEval-2018 task 1: Affect in tweets](#). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. [SemEval-2016 task 6: Detecting stance in tweets](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.
- Shikhar Murty, Tatsunori B Hashimoto, and Christopher D Manning. 2021. Dreca: A general task augmentation strategy for few-shot natural language inference.
- Bo Pang and Lillian Lee. 2004. [A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts](#). In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.
- Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train once, test anywhere: Zero-shot learning for text classification. *arXiv preprint arXiv:1712.05972*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.
- Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning.
- Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. [SemEval-2017 task 4: Sentiment analysis in Twitter](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.
- Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR.
- Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few-shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.
- Timo Schick and Hinrich Schütze. 2020b. It’s not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.
- Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. [Auto-Prompt: 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.
- Sasha Spala, Nicholas Miller, Franck Dernoncourt, and Carl Dockhorn. 2020. [SemEval-2020 task 6: Definition extraction from free text with the DEFT corpus](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 336–345, Barcelona (online). International Committee for Computational Linguistics.
- Sasha Spala, Nicholas A. Miller, Yiming Yang, Franck Dernoncourt, and Carl Dockhorn. 2019. [DEFT: A corpus for definition extraction in free- and semi-structured text](#). In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 124–131, Florence, Italy. Association for Computational Linguistics.
- Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.
- Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. [SemEval-2018 task 3: Irony detection in English tweets](#). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

- Eric Wallace, Tony Z Zhao, Shi Feng, and Sameer Singh. 2020. Customizing triggers with concealed data poisoning. *arXiv preprint arXiv:2010.12563*.
- Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.
- Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew Peters. 2020. Learning from task descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375, 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.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in NLP. *CoRR*, abs/2104.08835.
- Qinyuan Ye and Xiang Ren. 2021. Zero-shot learning by generating task-specific adapters. *arXiv preprint arXiv:2101.00420*.
- Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. 2020. Meta-learning without memorization. In *International Conference on Learning Representations*.
- Wenpeng Yin. 2020. Meta-learning for few-shot natural language processing: A survey. *arXiv preprint arXiv:2007.09604*.
- Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS’15*, page 649–657, Cambridge, MA, USA. MIT Press.
- 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*.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Ethics

Data and incentives In the existing prompting framework, end users send the natural language descriptions and a few training examples to the large language model inference API to perform few-shot learning (Brown et al., 2020). This becomes a natural source of training data for meta-tuning. Hence, the success of meta-tuning presented in this paper might incentivize for-profit organizations who provide language model inference APIs to collect prompts from the users, and train on these data.

Privacy, security, and fairness If a model is meta-tuned on user-provided data, certain security, privacy and fairness concerns can potentially emerge. For example, Carlini et al. (2020) shows that it is possible to extract the training data from large language models, and hence meta-tuned systems might expose some users’ prompts to other users. Wallace et al. (2020) shows that it is possible to poison the model through training data and trigger unwanted behaviors; the meta-tuning procedure might be susceptible to these data poisoning attacks as well. Finally, meta-tuning might perpetuate existing societal biases hidden in the users’ prompts (Bolukbasi et al., 2016).

If not addressed properly, these concerns might have a broader negative societal impact through meta-tuning. Compared to other domain-specific and task-specific machine learning applications, meta-tuned models might be applied to a much wider range of tasks, deployed at a larger scale, and serving a more diverse set of user population. Therefore, biased or poisoned training data for one task from one user population might compromise fairness and performance of another task and harm another user population; additionally, malicious or biased data might even tamper with the few-shot learning capability (“meta-poisoning”).

Potential abuse As shown in Figure 6, the AUC-ROC score for a lot of tasks are still well below 0.9, and hence our system is far from solving a significant fraction of tasks. Therefore, even though our system is flexible and has the potential to perform a wide range of tasks, it does not present an elixir to all classification tasks. Particularly, it should not be applied to higher stake scenarios (e.g. hate speech detection, fake news detection, etc), since its efficacy, robustness, and fairness properties remain unknown.

B Datasets

IMDB movie review sentiment classification (Maas et al., 2011b). Classifies whether the user likes the movie.

POSITIVE: “My favourite police series of all time turns to a TV-film. Does it work? Yes. ...”

NEGATIVE: “ "Stupid! Stupid! Stupid! I can not stand Ben stiller anymore.”

Zero Shot Emotion Classification (Yin et al., 2019). This task classifies a textual input into 9 emotion types {“sadness”, “joy”, “anger”, “disgust”, “fear”, “surprise”, “shame”, “guilt”, “love”}, and none-type if not any of the above. For example,

JOY: “*Making new friends is always fun , specially when playing dress up*”

ANGER: “*People that smoke cigarettes irritate my soul.*”

Zero Shot topic Classification (Yin et al., 2019). This task classifies an articles into 10 topic classes, such as “Health”, “Sports”, “Family & Relationships”, “Science & Mathematics”, etc.

SCIENCE: “*How the human species evolved ? How the human species evolved ? A tough question as it overlaps science and theology ...*”

SPORTS: “*Number one formula one driver ? Depends on your question.Kimi Raikkonen won more races ...*”

Zero Shot Situation Classification (Yin et al., 2019). This task classifies what type of situation the user is having and what aids it needs. The classes include “food”, “terrorism”, etc.

FOOD: “*livestock , crops , rural and urban housing , roads , river and canal embankments , and other infrastructure suffered extensive damage*”

TERRORISM: “*Daish attacked Tahrumbu vilage with chemical gas.*”

SemEval 2017 Task 4 (Barbieri et al., 2020) (Rosenthal et al., 2017). This task classifies the sentiment of a tweet as {“negative”, “neutral”, “positive”}. For example,

NEGATIVE: “*That sucks if you have to take the SATs tomorrow*”

NEUTRAL: “*@user Well said on HMW. Can you now address why Texans fans file out of the stadium midway through the 4th qtr of every game?*”

POSITIVE: *"This is the last Sunday w/o football,NFL is back baby"*

ETHICS Justice (Hendrycks et al., 2020). This task identifies whether the justification provided for a given scenario is deemed reasonable. For example,

REASONABLE: *"I deserve to be paid by my boss because I keep her house clean daily."*

UNREASONABLE: *"I deserve to be given a brand new car by my boss because I keep her house clean daily."*

ETHICS Deontology (Hendrycks et al., 2020). This task identifies whether the excuse provided for a given scenario is a reasonable excuse. For example,

SCENARIO: *"Aren't you supposed to be preparing for the meeting?"*

EXCUSE (REASONABLE): *"No because I have everything already set up in the conference room."*

EXCUSE (UNREASONABLE): *"No because you don't have a meeting"*

ETHICS Common Sense (Hendrycks et al., 2020). This task identifies whether a scenario demonstrates common sense. For example,

COMMON SENSE: *"I went to the principal's office to change my records before going to a different school."*

NOT COMMON SENSE: *"I secured the loan because I would make the payments."*

EURLEX57K (Chalkidis et al., 2019). This task classifies European legislation.

NATIONAL CURRENCY: *"Council Regulation (EC) No 2595/2000 of 27 November 2000 amending Regulation (EC) No 1103/97 on certain provisions relating to the introduction of the euro"*

SOUTHERN AFRICA: *"95/458/EC: Commission Regulation (EC) No 302/2006 of 20 February 2006 on import licences in respect of beef and veal products originating in Botswana, Kenya, Madagascar, Swaziland, Zimbabwe and Namibia"*

SemEval 2019 Task 6 (Barbieri et al., 2020) (Zampieri et al., 2019). This task classifies the tweet as either offensive or not offensive. For example,

OFFENSIVE: *"@user She has become a parody unto herself? She has certainly taken some heat for being such an...well idiot. Could be optic too*

Who know with Liberals They're all optics. No substance"

NOT OFFENSIVE: *"@user @user She is great. Hi Fiona!"*

Click Bait Detection¹⁰ This task detects whether a news title is a click bait.

CLICK BAIT: *"Can You Pass This Basic Trigonometry Quiz"*

NON CLICK BAIT: *"NASCAR driver Kyle Busch wins 2011 Jeff Byrd 500"*.

Abstract Domain Classification¹¹ This classifies the abstract into 4 domains: "Physics", "Maths", "Computer Science", "Statistics". For example,

PHYSICS: *"a ever-growing datasets inside observational astronomy have challenged scientists inside many aspects, including an efficient and interactive data exploration and visualization. many tools have been developed to confront this challenge ..."*

MATHS: *"a main result of this note was a existence of martingale solutions to a stochastic heat equation (she) inside the riemannian manifold ..."*

SemEval 2019 Task 5 (Barbieri et al., 2020) (Basile et al., 2019). This task identifies whether the tweet contains hate speech towards women and/or immigrants or not. For example,

HATE SPEECH: *"This account was temporarily inactive due to an irrational woman reporting us to Twitter. What a lack of judgement, shocking. #YesAllMen"*

NO HATE SPEECH: *"@user nice new signage. Are you not concerned by Beatlemania -style hysterical crowds crongregating on you. . ."*

SemEval 2019 Task 8 (Mihaylova et al., 2019). This task identifies whether the text is an example of a question asking for factual information, an example of a question asking for an opinion, or an example of socializing. For example,

FACTUAL: *"is there any place i can find scented massage oils in qatar?"*

OPINION: *"hi there; i can see a lot of massage center here; but i dont which one is better."*

¹⁰<https://www.kaggle.com/c/clickbait-news-detection>

¹¹<https://www.kaggle.com/abisheksudarshan/topic-modeling-for-research-articles?select=Train.csv>

can someone help me which massage center is good...and how much will it cost me? thanks"

SOCIALIZING: *"Hello people...let's play this game...you have to write something good about the person whose 'post' is above you on QL.You can write anything and you can write multiple times."*

SemEval 2018 Task 3 (Barbieri et al., 2020) (Van Hee et al., 2018). This task identifies whether the tweet contains irony or not. For example,

IRONY: *"seeing ppl walking w/ crutches makes me really excited for the next 3 weeks of my life"*

NO IRONY: *"@user on stage at #flzjingleball at the @user in #Tampa #iheartradio"*

SemEval 2018 Task 1 (Barbieri et al., 2020; Mohammad et al., 2018) This task classifies a tweet as one of 4 emotion types {"sadness", "joy", "anger", "optimism"}. For example,

SADNESS: *"@user I so wish you could someday come to Spain with the play, I can't believe I'm not going to see it #sad"*

JOY: *"#ThisIsUs has messed with my mind & now I'm anticipating the next episode with #apprehension & #delight! #isthereahelplineforthis"*

ANGER: *"@user Haters!!! You are low in self worth. Self righteous in your delusions. You cower at the thought of change. Change is inevitable."*

OPTIMISM: *"Don't be #afraid of the space between your #dreams and #reality. If you can #dream it, you can #make it so"*

SemEval 2016 Task 6 (Mohammad et al., 2016; Barbieri et al., 2020) This task classifies a tweet's stance as {"neutral", "against", "favor"}. Each tweet contains a stance on one of the five different target topics {"abortion", "atheism", "climate change", "feminism", "hillary"}. For example,

NEUTRAL: *"@user maybe that's what he wants #SemST"*

AGAINST: *"Life is #precious & so are babies, mothers, & fathers. Please support the sanctity of Human Life. Think #SemST"*

FAVOUR: *"@user @user Nothing to do with me. It's not my choice, nor is it yours, to dictate what another woman chooses. #feminism #SemST"*

SemEval 2020 Task 6 (Spala et al., 2020). This task classifies whether textbook sentence contains a definition. For example,

CONTAINS DEFINITION: *"Since 2005, automated sequencing techniques used by laboratories are under the umbrella of next-generation sequencing, which is a group of automated techniques used for rapid DNA sequencing"*

DOESN'T CONTAIN DEFINITION: *"These automated low-cost sequencers can generate sequences of hundreds of thousands or millions of short fragments (25 to 500 base pairs) in the span of one day."*

TREC (Li and Roth, 2002). This task classifies a question into one of six question types: DESC (description), ABBR (abbreviation), ENTY (entity), HUM (people/individual), LOC (location), NUM (numeric information), each of which have specific fine-grained sub-categories. For example,

DESC: *"How did serfdom develop in and then leave Russia?"*

ABBR: *"What is the full form of .com?"*

ENTY: *"What films featured the character Pop-eye Doyle?"*

HUM: *"What contemptible scoundrel stole the cork from my lunch?"*

LOC: *"What sprawling U.S. state boasts the most airports?"*

NUM: *"How many Jews were executed in concentration camps during WWII?"*

SUBJ (Pang and Lee, 2004). This task classifies a sentence as being subjective or objective. For example,

SUBJECTIVE: *"smart and alert, thirteen conversations about one thing is a small gem."*

OBJECTIVE: *"the movie begins in the past where a young boy named sam attempts to save celebi from a hunter."*

The Corpus of Linguistic Acceptability (Warstadt et al., 2018). This task detects if sentences are grammatically acceptable by their original authors. For example,

GRAMMATICALLY ACCEPTABLE: *"Her little sister will disagree with her."*

GRAMMATICALLY NOT ACCEPTABLE: *"Has not Henri studied for his exam?"*

The Multi-Genre NLI Corpus (Williams et al., 2018). This task detects if a premise is a contradiction or entailment of a hypothesis, or if a hypothesis holds neutral view on the premise.. For example,

NEUTRAL: *“Premise: Exoatmospheric Kill Vehicles orbiting Earth would be programmed to collide with warheads. Hypothesis: Exoatmospheric Kill Vehicles would be very expensive and hard to make.”*

ENTAILMENT: *“Premise: so we have to run our clocks up forward an hour and i sure do hate to loose that hour of sleep in the morning. Hypothesis: I don’t like the time change that results in losing an hour of sleeping time.”*

CONTRADICTION: *“Premise: The mayor originally hoped groundbreaking would take place six months ago, but it hasn’t happened yet. Hypothesis: The mayor doesn’t want groundbreaking to happen at all.”*

Metaphor as a Medium for Emotion: An Empirical Study (?). This task detects if the application of a word is Literal or Metaphorical. For example,

WORD: ABUSE

LITERAL: *“This boss abuses his workers.”*

METAPHORICAL: *“Her husband often abuses alcohol.”*

Political Preference Classification (Allaway and McKeown, 2020). This task predicts a comment’s stand point on a political topic. For example,

TOPIC: COMPANIES REGULATION

CON: *“Regulation of corporations has been subverted by corporations. States that incorporate corporations are not equipped to regulate corporations that are rich enough to influence elections, are rich enough to muster a legal team that can bankrupt the state. Money from corporations and their principals cannot be permitted in the political process if democracy is to survive.”*

PRO: *“Regulation is to a corporation what a conscience is to a living person. Without a conscience, we would all be sociopaths. Corporations do not have a conscience, thus they need regulation to make sure they are focused on benefiting society instead on merely benefiting themselves.”*

NEUTRAL: *“Without government to ensure their behavior, companies will attempt to make a profit even to the DETRIMENT of the society that supports the business. We have seen this in the environment, in finances, in their treatment of workers and customers. Enough.”*

Airline Service Review ¹² This task classifies if an airline review has a positive or negative sentiment. For example,

POSITIVE: *“This is such a great deal! Already thinking about my 2nd trip to Australia; I haven’t even gone on my 1st trip yet!”*

NEGATIVE: *“amazing to me that we can’t get any cold air from the vents.”*

Covid-19 Tweets Sentiment Analysis ¹³ This task classifies if a tweet has a positive or negative sentiment. For example,

POSITIVE: *“Taken by Henk Zwoferink on Saturday in Wargl, our black beauty hauled a train bringing the last tourists home. Our colleagues are #workinghard to keep supply chains running while respecting the measures to ensure everyone’s #safety. A pleasure to work with such #DedicatedPeople!”*

NEGATIVE: *“So far, the Minister does not seem to have made statement on the catastrophe that can develop if the issue of markets operation is not addressed. Food insecurity has potential to make current Covid-19 panic look like a kindergarten and could lead to riots. I submit.”*

Hotel Review ¹⁴ This task predicts if a hotel review is a positive or negative review. For example,

NEGATIVE: *“The single rooms like hospital rooms single rooms hotel sparse intentional know ugly like trapped hospital white walls sink basin room small rectangle shape.the beds hard rocks blankets rough really noisy.this overrated hotel stayed fans type hotels”*

POSITIVE: *“loved stay, stayed univ, inn 10 days april 2005 thoroughly enjoyed, free parking clean spacious room friendly staff great breakfast snack, loved location, definitely stay, ”*

Stock Market Sentiment ¹⁵ This task predicts if a comment holds a positive or negative view on the performance of the stock market. For example,

NEGATIVE: *“GPS wow that wa s a fast fast fade...”*

POSITIVE: *“user Maykiljil posted that: I agree that MSFT is going higher & possibly north of 30”*

¹²<https://www.kaggle.com/welkin10/airline-sentiment>

¹³https://www.kaggle.com/datatattle/covid-19-nlp-text-classification?select=Corona_NLP_test.csv

¹⁴<https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews>

¹⁵<https://www.kaggle.com/yash612/stockmarket-sentiment-dataset>

AG-News (Zhang et al., 2015). This task classifies the topic of news based on their contents. For example,

WORLD NEWS: "Greek duo could miss drugs hearing"

SPORTS NEWS: "AL Wrap: Olerud Cheers Yankees by Sinking Ex-Team"

BUSINESS NEWS: "Lowe's Second-Quarter Profit Rises"

TECH NEWS: "Satellite boosts Olympic security"

Real and Fake News ¹⁶ This task classifies if a news is fake or real. For example,

REAL: "WASHINGTON (Reuters) - Alabama Secretary of State John Merrill said he will certify Democratic Senator-elect Doug Jones as winner on Thursday despite opponent Roy Moore's challenge, in a phone call on CNN. Moore, a conservative who had faced allegations of groping teenage girls when he was in his 30s, filed a court challenge late on Wednesday to the outcome of a U.S. Senate election he unexpectedly lost."

FAKE: "Ronald Reagan shut down the Berkeley protests many years ago THIS is how you do it!"

Disaster Tweets ¹⁷ This task detects if a tweet announces an emergency or a disaster. For example,

CONTAINS DISASTER: "Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all."

DOES NOT CONTAIN DISASTER: "My dog attacked me for my food #pugprobs."

Obama vs Trump Tweets ¹⁸ This task detects if a tweet was sent by Obama or Trump. For example,

OBAMA: "Michelle and I are delighted to congratulate Prince Harry and Meghan Markle on their engagement. We wish you a lifetime of joy and happiness together."

TRUMP: "Together, we dream of a Korea that is free, a peninsula that is safe, and families that are reunited once again!"

¹⁶<https://www.kaggle.com/amananandrai/ag-news-classification-dataset?select=train.csv>

¹⁷<https://www.kaggle.com/c/nlp-getting-started/data?select=train.csv>

¹⁸<https://www.kaggle.com/shaharz/classifying-tweets-of-trump-and-obama>

Kaggle Sexually Explicit Tweets ¹⁹ This dataset provides positive examples of profane comments. For example,

EXPLICIT: "What do guys say when you get naked in front of them for the first time?"

Democratic vs Republican Tweets ²⁰ This task detects if a tweet was sent by the Democratic or Republican Party. For example,

DEMOCRATIC: "#YuccaMountain would require moving tens of thousands of metric tons of radioactive waste across the country and through Southern Nevada."

REPUBLICAN: "Stopped by One Hour Heating& Air Conditioning to discuss the benefits tax reform will bring to their business."

Women E-commerce Clothing Reviews ²¹ This task predicts if the buyer likes or recommends a product based on its review. For example,

LIKE: "After reading the previous reviews, i ordered a size larger. i am so glad i did it! it fits perfectly! i am 5'4"/115/32dd and went with the s regular. so beautiful! i can't wait to wear it!"

DISLIKE: "The zipper broke on this piece the first time i wore it. very disappointing since i love the design. I'm actually going to try to replace the zipper myself with something stronger, but annoying that it's come to that."

Quora Question Pairs ²² This task predicts if a pair of Quora question is asking for the same thing. For example,

SAME: "Question 1: How many months does it take to gain knowledge in developing Android apps from scratch?; Question 2: How much time does it take to learn Android app development from scratch?"

DIFFERENT: "Question 1: How would you review the site Waveclues? ; Question 2: Is there a good pay for reviews site out there?"

Headline Sarcasm Detection This task detects if a news headline contains sarcasm. For example,

¹⁹<https://www.kaggle.com/harsh03/sexually-explicit-comments>

²⁰<https://www.kaggle.com/kapastor/democratvsrepublicantweets?select=ExtractedTweets.csv>

²¹<https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews>

²²<https://www.kaggle.com/c/quora-question-pairs/data>

SARCASM: “guy who just wiped out immediately claims he’s fine”

NO SARCASM: “Donald trump effigies burn across Mexico in Easter ritual”

Company Account Tweets ²³ This task detects whether the tweet is targeted towards a company account. For example,

YES: “@VirginTrains Oh, that’s nice. What are you doing about it? What are you targets next year?”

NO: “@115738 That’s the best kind of trick-or-treating. All treats, my friend. -Becky”

SMS Spam Detection (Almeida et al., 2013) This task detects whether the SMS is a spam message. For example,

SPAM: “Thank you, winner notified by sms. Good Luck! No future marketing reply STOP to 84122 customer services 08450542832”

HAM: “Lol great now I am getting hungry.”

Clothing Fitness (Misra et al., 2018) Checking whether the customer complains that the cloth is too small or too large.

SMALL: “runs a bit small. wish it fit”.

LARGE: “too big”.

Water Problem Topic Classification ²⁴ Classifying the topic of a report on water problems. The labels include “biological”, “climatic indicator”, “environmental technology”, etc. For example,

BIOLOGICAL: “Mineralization of organic phosphorus in bottom sediments reaches 40–80% and as we found out during the project implementation it intensified in autumn-winter period.”

CLIMATIC INDICATOR: “The average amount of precipitation in the lower part of the basin makes 470 mm to 540 mm. The relative average annual air humidity makes 60-65%”.

ENVIRONMENTAL TECHNOLOGY: “Most of wastewater treatment facilities require urgent modernization and reconstruction”.

Sexist Statement Detection ²⁵ This task classifies whether the statement is sexist. For example,

SEXIST: “It’s impossible for a girl to be faithful.”

²³<https://www.kaggle.com/thoughtvector/customer-support-on-twitter>

²⁴https://www.kaggle.com/vbmokin/nlp-reports-news-classification?select=water_problem_nlp_en_for_Kaggle_100.csv

²⁵<https://www.kaggle.com/dgrosz/sexist-workplace-statements>

NON SEXIST: “Without strength, can we work to create wealth?”

Movie Spoiler Detection (Misra, 2019) ²⁶ This task classifies whether the movie review is a spoiler. For example,

SPOILER: “I must say that this movie was good but several things were left unsaid. For those who have seen the movie know what I am talking about but for those who haven’t, I don’t want to give spoilers. I was also impressed by Vin Diesel’s acting skills. Overall I have to say it was a good movie filled with several twists and turns.”

NON SPOILER: “The Great Wall amazes with its spectacular effects, both on screen and sound. Usually I do not appreciate 3D movies, but in this case I felt like it worth it. However, being honest, the storytelling and the story itself had its weaknesses. There were many logical lapses, and for me, many details are still waiting to be answered. On the other hand, expect decent acting especially from the main characters. All in all, The Great Wall is a solid popcorn-movie, but I expected a more elaborated unfolding of the legend it tells about.”

News Summary/headline Topic Classification ²⁷ This task classifies the topic of the summary of a news. For example,

POLITICS: “City and state officials said they received little advance warning of the decision.”

BUSINESS: “The streaming giant’s third-quarter earnings were nothing like the Upside Down.”

C Dataset Property Tags

Here we list all the dataset property tags (Section 2). We define two datasets to be “similar” if they have the set of tags, and disallow meta-tuning on datasets that are similar to evaluation dataset.

social media: whether the source is from social media (e.g. tweets).

social/political: whether the task is highly related to political/social topics. Some examples include stance classification and hate speech detection.

topic classification: whether the task classifies the topics of the input.

²⁶https://www.kaggle.com/rmisra/imdb-spoiler-dataset?select=IMDB_reviews.json

²⁷<https://www.kaggle.com/rmisra/news-category-dataset>

good vs. bad: whether the task classifies whether the text is judging something to be good or bad.

paper: whether input text comes from a paper.

review: whether the input text is a review of a product (e.g. movie, hotel).

questions: whether the input texts are questions. Some examples include classifying whether the question asks for factual information or subjective opinion and detecting whether two questions have the same meaning.

emotion: whether the task classifies certain emotion in the text, for example “hate”, “surprise”, “joy”, etc.

Besides, we do not assign tags to datasets that we are confident to be different enough from other tasks (e.g. extracting whether a text contains definition), and allow the model to be meta-tuned on all other datasets.

D List of Label Descriptions

The comprehensive list of label descriptions and grouping can be seen in Figure 8, 9, and 10.

E Robustness Checks

We report all the descriptive statistics mentioned in Section 3 under 3 different types of description weighting. We additionally compare T5-small vs. T5-base, BERT-medium vs. BERT-Base and BERT-Base vs. BERT Large. All the results can be seen in Table 3, 4, and 5. Due to space constraint, we abbreviate $\mathbb{P}[\Delta > t]$ as $> t$ if t is positive, and $< t$ if t is negative. Notice that, since we only have around 20 datasets to evaluate the model, most of the results presented here are not statistically significant at the dataset level; nevertheless,

E.1 Different Description Weighting

We weight each label and dataset equally in Table 4 and 5. We find that, under almost all comparisons across different weighting, the mean change $\bar{\Delta}$ is positive, and the change above a certain threshold t is more frequent than the change below a certain threshold $-t$. The only single exception is the “Ensemble” row in Table 5, where there are slightly more datasets where the change is lower than -1%. Nevertheless, given that the trend is still positive under $t = 5\%$ and 10% , and two other description weightings, we may still conclude that

ensembling label descriptions is more likely to improve model performance.

E.2 Larger T5 Models are Better

In addition to comparing T5-Base (220 Million parameters) vs. T5-Large (770M), we also compare T5-small (60M) vs. T5-base (220M). Across all metrics, larger models are significantly better. Most notably, there is a sudden jump in performance when increasing model size from T5-small to T5-base (sometimes 15% increase in $\bar{\Delta}$).

E.3 Larger BERT Models are Better

We also compare different sizes of BERT (Turc et al., 2019) (41, 110, and 330M) parameters. Across all metrics, larger models are significantly better.

F Most Relevant Datasets

To ensure that we are testing the models’ ability to generalize to an unseen tasks, we disallow both training and testing on datasets that are too similar, which is defined as “having the same set of dataset property tags” (Section 2). To help interpret how we define unseen tasks, for each dataset that we evaluate on, we try to find the “most relevant” dataset that the model has seen during the meta-tuning phase, and list it in Table 6.

G Performance Break Down

For each model, we average the AUC-ROC scores for each label description for each dataset, and report the results in Table 7.

H Accuracy

===== GROUP 0 =====

[*Does the tweet take an opposing stance on abortion?', '*Does the tweet oppose abortion?', '*Does the tweet take a stance against abortion?', '*Does the tweet support abortion?', '*Is there a supporting stance taken on abortion in the tweet?', '*Does the tweet take a stance in favor of abortion?', '*Does the tweet take a supporting stance on abortion?', '*Is there an opposing stance taken on abortion in the tweet?', '*Does the tweet consider abortion to be bad?', '*Is there an opposing stance taken on atheism in the tweet?', '*Does the tweet support religious beliefs?', '*Does the tweet take a supporting stance on atheism?', '*Is the tweet against religious beliefs?', '*Does the tweet take an opposing stance on atheism?', '*Does the tweet take a stance against atheism?', '*Is there a supporting stance taken on atheism in the tweet?', '*Does the tweet take a stance in favor of atheism?', '*Is the tweet against environmentalist initiatives?', '*Does the tweet support protecting the environment?', '*Does the tweet support environmentalist?', '*Does the tweet take a stance in favor of feminism?', '*Is there an opposing stance taken on feminism in the tweet?', '*Does the tweet take a stance against feminism?', '*Is there a supporting stance taken on feminism in the tweet?', '*Does the tweet take an opposing stance on feminism?', '*Does the tweet take a supporting stance on feminism?', '*Is there an opposing stance taken on Hillary in the tweet?', '*Does the tweet take a stance against Hillary?', '*Does the tweet take an opposing stance on Hillary?', '*Does the tweet take a supporting stance on Hillary?', '*Is there a supporting stance taken on Hillary in the tweet?', '*Does the tweet take a stance in favor of Hillary?', 'Does the tweet take a stance against a controversial issue?', 'Does the tweet take a stance in favor of a controversial issue?', 'Does the tweet take a supporting stance on a controversial issue?', 'Does the tweet take an opposing stance on a controversial issue?', 'Is there an opposing stance taken on a controversial issue in the tweet?', 'Is there a supporting stance taken on a controversial issue in the tweet?', 'Is the statement sexist?', 'Does this statement discriminate against women?', 'Is it sexually explicit?', 'Is it inappropriate?', 'Is it profane?', 'Is this a tweet from Donald Trump?', 'Is this a tweet from Barack Obama?', 'Did Donald Trump post this tweet?', 'Did Barack Obama post this tweet?', 'Does this lean toward Republican Party?', 'Is this a tweet from Democratic Party?', 'Is this a tweet from Republican Party?', 'Does this lean toward Democratic Party?', 'Is this a Republican post?', 'Is this a Democratic post?']

===== GROUP 1 =====

[*Is this text describing a crime violence situation?', '*Are the people or area in the text experiencing crime violence?', '*Are the people described in the text in need of evacuation?', '*Do people need to evacuate?', '*Is this text expressing a need for evacuation?', '*Is this text expressing a need for food?', '*Are the people described in the text in need of food?', '*Do people need food?', '*Is this text providing information about an infrastructure?', '*Is it discussing an infrastructure?', '*Is this text expressing a need for medical support?', '*Do people need medical support?', '*Are the people described in the text in need of medical support?', '*Is this text describing a regime change situation?', '*Are the people or area in the text experiencing regime change?', '*Is there a regime change?', '*Are the people described in the text in need of search or rescue?', '*Do people need to be searched or rescued?', '*Is this text expressing a need for search or rescue?', '*Do people need shelter?', '*Is this text expressing a need for shelter?', '*Are the people described in the text in need of shelter?', '*Is this text describing a terrorism situation?', '*Are the people or area in the text experiencing terrorism?', '*Are the people described in the text in need of utilities, energy, or sanitation?', '*Do people need utilities, energy, or sanitation?', '*Is this text expressing a need for utilities, energy, or sanitation?', '*Do people need water?', '*Are the people described in the text in need of water?', '*Is this text expressing a need for water?']

===== GROUP 2 =====

[*Does the text contain a question asking for factual information?', '*Is factual information being asked for in the text?', '*Does the text contain a question asking for an opinion?', '*Is an opinion being asked for in the text?', '*Do these two questions have the same meaning?', '*Are these two questions asking for the same thing?', '*Does the text convey subjective information?', '*Is the text subjective?', '*Does the text convey objective information?', '*Is the text objective?', '*Does the question ask for a numerical answer?', '*Does the question ask for an individual or group, or their occupation?', '*Does the question ask about an acronym?', '*Does the question ask for something to be described?', '*Does the question ask for a number?', '*Does the question ask about a location?', '*Does the question ask for an explanation of something?', '*Does the question ask for a description of something?', '*Does the question ask for a place?', '*Does the question ask about an entity?', '*Does the question ask about an individual or group?', '*Does the question ask about an abbreviation?', 'Does the question ask what an abbreviation stand for?', 'Does the question ask for the abbreviation of something?', 'Does the question ask about the meaning of an abbreviation?', 'Does the question ask for something to be explained?', 'Does the question ask for a definition?', 'Does the question ask about how something is done?', 'Does the question ask for something to be described?', 'Does the question ask for a description of something?', 'Does the question ask to define something?', 'Does the question ask for a reason why something occurs/occurred?', 'Is the question related to a publication, show, or film?', 'Does the question ask about food?', 'Is the question related to sports?', 'Is the question related to physical substances?', 'Is the question related to languages?', 'Does the question ask about a word/words?', 'Is the question related to letters of an alphabet?', 'Does the question ask about letters of an alphabet?', 'Is the question asking about a terminology?', 'Does the question ask about products?', 'Does the question ask about animals?', 'Is the question related to food?', 'Does the question ask about anything about a body part?', 'Is the question about an instrument?', 'Is the question related to a word/words?', 'Does the question ask about sports?', 'Does the question ask about symbols?', 'Is the question related to instruments?', 'Is the question related to currency?', 'Is the question about an event?', 'Does the question ask for a terminology?', 'Is the question about religion?', 'Is the question related to plants?', 'Does the question ask about an event?', 'Does the question ask about plants?', 'Is the question about a word/words?', 'Is the question related to products?', 'Is the question about symbols?', 'Is the question about physical substances?', 'Does the question ask about anything related to colors?', 'Is the question about sports?', 'Is the question related to symbols?', 'Is the question about plants?', 'Does the question ask about a technique?', 'Is the question about letters of an alphabet?', 'Is the question about products?', 'Is the question about a language?', 'Does the question ask about religion?', 'Is the question related to a body part?', 'Is the question related to religion?', 'Is the question asking about currency?', 'Is the question asking about medical issues?', 'Is the question related to colors?', 'Does the question ask about physical substances?', 'Is the question asking about a mode of transportation?', 'Is the question related to medical issues?', 'Is the question related to animals?', 'Is the question asking for the description of someone?', 'Does the question ask about an individual?', 'Does the question ask about the occupation of an individual?', 'Does the question ask about a group?', 'Does the question ask about a city?', 'Does the question ask about a country?', 'Does the question ask about a mountain?', 'Does the question ask about a state?', 'Does the question ask about a percentage?', 'Does the question ask about a code?', 'Does the question ask about digits in a code?', 'Does the question ask about an amount?', 'Does the question ask about speed of something?', 'Does the question ask for a date of something?', 'Does the question ask about money?', 'Does the question ask about a specific date?', 'Does the question ask about something related to temperature?', 'Does the question ask for a percentage of something?', 'Is the question related to age or a period?', 'Is the question asking about size or volume?', 'Does the question ask about a distance or length?', 'Does the question related to ranking or ordering?', 'Is the question asking about weight?', 'Does the question ask for an amount of something?']

Figure 8: Label descriptions from the same group of datasets that are considered similar. "*" at the beginning indicates that we are evaluating on this dataset.

===== GROUP 3 =====

["*Does the topic of discussion fall into the research field of Mathematics?', '*Is the main subject of discussion Physics?', '*Does the topic of discussion fall into the research field of Computer Science?', '*Does the topic of discussion fall into the research field of Statistics?', '*Is the main subject of discussion Computer Science?', '*Is this abstract about Mathematics?', '*Is the main subject of discussion Mathematics?', '*Is the main subject of discussion Statistics?', '*Is this abstract about Statistics?', '*Is this abstract about Physics?', '*Is this abstract about Computer Science?', '*Does the topic of discussion fall into the research field of Physics?']

===== GROUP 4 =====

["*Is the tweet's emotion that of anger?", '*Does the tweet convey feelings of anger?', '*Is the tweet's emotion that of joy?', '*Does the tweet convey feelings of joy?', '*Is the tweet's emotion that of optimism?', '*Does the tweet convey feelings of optimism?', '*Is the tweet's emotion that of sadness?', '*Does the tweet convey feelings of sadness?', 'Does the tweet convey a neutral sentiment?', 'Is the sentiment of the tweet neutral?', 'Is the sentiment of the tweet positive?', 'Does the tweet convey a positive sentiment?', 'Does the tweet convey a negative sentiment?', 'Is the sentiment of the tweet negative?', 'Does this text express love?', 'Does this text express a sense of disgust?', 'Does this text express a sense of shame?', 'Does the author convey a sense of love?', 'Does this text express a sense of sadness?', 'Does this text express a sense of love?', 'Does the author of this text show an emotion of joy?', 'Does the author of this text show an emotion of fear?', 'Does this text express disgust?', 'Does this text express a sense of guilt?', 'Does this text express fear?', 'Does this text express a sense of anger?', 'Does the author of this text show an emotion of sadness?', 'Does the author of this text show an emotion of guilt?', 'Does the author of this text show an emotion of anger?', 'Does this text express joy?', 'Does this text express a sense of fear?', 'Does the author of this text show an emotion of shame?', 'Does this text express anger?', 'Does this text express a sense of joy?', 'Does this text express a sense of surprise?', 'Does this text express sadness?', 'Does this text express shame?', 'Does this text express guilt?', 'Does the author of this text show an emotion of love?', 'Does this text express surprisingness?', 'Does the author of this text show an emotion of disgust?', 'Does this news seem optimistic about the stock market?', 'Does this news seem pessimistic about the stock market?', 'Is this a negative news for the stock market?', 'Is this a positive news for the stock market?']

===== GROUP 5 =====

["*Is this a Business News?', '*Is this a Sports News?', '*Is this news about science or technology events?', '*Is this news about world events?', '*Is this news about sports events?', '*Is this a scientific or technology News?', '*Is this news about business events?', '*Is this a world news?', 'Is this text about Science and Mathematics?', 'Is the topic of this text related to Business and Finance?', 'Is this text about Business and Finance?', 'Is this text about Computers and Internet?', 'Is this text about Sports?', 'Is the topic of this text related to Politics and Government?', 'Is this text about Society and Culture?', 'Is the topic of this text related to Science and Mathematics?', 'Is the topic of this text related to Family and Relationships?', 'Is the topic of this text related to Sports?', 'Is the topic of this text related to Society and Culture?', 'Is this text about Health?', 'Is this text about Politics and Government?', 'Is this text about Entertainment and Music?', 'Is this text about Family and Relationships?', 'Is the topic of this text related to Entertainment and Music?', 'Is this text about Education and Reference?', 'Is the topic of this text related to Education and Reference?', 'Is the topic of this text related to Health?', 'Is the topic of this text related to Computers and Internet?', 'Is the headline relevant to business?', 'Is the headline relevant to parenting?', 'Does the headline belong to the category parenting?', 'Does the headline belong to the category comedy?', 'Is the headline relevant to entertainment?', 'Does the headline belong to the category sports?', 'Does the headline belong to the category style & beauty?', 'Does the headline belong to the category travel?', 'Is the headline relevant to education?', 'Does the headline belong to the category arts & culture?', 'Is the headline relevant to tech?', 'Is the headline relevant to food & drink?', 'Is the headline relevant to sports?', 'Is the headline relevant to religion?', 'Is the headline relevant to science?', 'Is the headline relevant to style & beauty?', 'Is the headline relevant to travel?', 'Does the headline belong to the category religion?', 'Does the headline belong to the category tech?', 'Does the headline belong to the category science?', 'Does the headline belong to the category money?', 'Does the headline belong to the category education?', 'Is the headline relevant to environment?', 'Is the headline relevant to money?', 'Is the headline relevant to comedy?', 'Does the headline belong to the category food & drink?', 'Does the headline belong to the category crime?', 'Does the headline belong to the category business?', 'Is the headline relevant to politics?', 'Is the headline relevant to arts & culture?', 'Does the headline belong to the category weird news?', 'Does the headline belong to the category entertainment?', 'Does the headline belong to the category politics?', 'Is the headline relevant to weird news?', 'Does the headline belong to the category environment?', 'Is the headline relevant to crime?', 'Is this news summary relevant to weird news?', 'Is this news summary relevant to tech?', 'Is this news summary relevant to education?', 'Is this news summary relevant to comedy?', 'Is this news summary relevant to sports?', 'Is this news summary relevant to politics?', 'Does this news summary belong to the category tech?', 'Does this news summary belong to the category religion?', 'Is this news summary relevant to religion?', 'Is this news summary relevant to arts & culture?', 'Is this news summary relevant to science?', 'Does this news summary belong to the category sports?', 'Does this news summary belong to the category politics?', 'Does this news summary belong to the category education?', 'Is this news summary relevant to crime?', 'Does this news summary belong to the category money?', 'Does this news summary belong to the category business?', 'Does this news summary belong to the category style & beauty?', 'Is this news summary relevant to travel?', 'Does this news summary belong to the category crime?', 'Does this news summary belong to the category environment?', 'Is this news summary relevant to style & beauty?', 'Does this news summary belong to the category science?', 'Is this news summary relevant to parenting?', 'Does this news summary belong to the category weird news?', 'Is this news summary relevant to entertainment?', 'Does this news summary belong to the category comedy?', 'Does this news summary belong to the category entertainment?', 'Is this news summary relevant to business?', 'Does this news summary belong to the category food & drink?', 'Is this news summary relevant to money?', 'Is this news summary relevant to environment?', 'Does this news summary belong to the category parenting?', 'Does this news summary belong to the category travel?', 'Does this news summary belong to the category arts & culture?', 'Is this news summary relevant to food & drink?']

===== GROUP 6 =====

["*Does the user find this movie good?', '*Does the user find this movie bad?', '*Does the user dislike this movie?', '*Is the user feeling good about the movie?', '*Is this a negative review?', '*Is this a positive review?', '*Does the user like this movie?', '*Is the user feeling negative about the movie?', 'Does the consumer dislike his or her airline service?', 'Is this a negative review towards an airline company?', 'Is this a positive review toward an airline company?', 'Does the consumer like his or her airline service?', 'Is this a positive hotel review?', 'Is the user feeling negative about his or her experience in the hotel?', 'Does this hotel review worth a rating of 0 out of 4?', 'Is this a bad hotel review?', 'Is the user feeling positive about his or her experience in the hotel?', 'Does this hotel review worth a rating of 4 out of 4?', 'Does the consumer dislike the product?', 'Is this product not recommended by the buyer based on its review?', 'Is this a good product?', 'Is this product recommended by the buyer based on its review?', 'Does the consumer like the product?', 'Does the consumer consider the product to be good?', 'Does the consumer consider this product to be bad?', 'Is this a bad product?']

GROUP 7

Figure 9: Label descriptions from the same group of datasets that are considered similar. "*" at the beginning indicates that we are evaluating on this dataset.

===== GROUP 7 =====
 [*Is this tweet offensive towards women or immigrants?', '*Does the tweet contain hate speech towards women or immigrants?', '*Is there hate speech towards women or immigrants in the tweet?', '*Is the tweet an offensive tweet?', '*Does the tweet contain offensive language/content?]

===== GROUP 8 =====
 [*Is the sentence grammatical?', '*Is the sentence ungrammatical?', '*Is the sentence grammatically unacceptable?', '*Is the sentence grammatically acceptable?]

===== GROUP 9 =====
 [*Does the tweet contain irony?', '*Is the tweet ironic?', '*Is there irony in the tweet?', 'Is the user feeling negative about the situation?', 'Does this tweet have a negative sentiment?', 'Does this tweet have a positive sentiment?', 'Is the user feeling positive about the situation?]

===== GROUP 10 =====
 [*Does the text contain a definition?', '*Does the text define a terminology?]

===== GROUP 11 =====
 [*Is this a spam?', 'Is the tweet directed at a customer support?', 'Is the tweet sent towards a customer support?]

===== OTHER =====
 [Is the sentence about economics?', 'Is the sentence category/subject history?', 'Is the sentence category/subject government?', 'Is the sentence category/subject physics?', 'Is the sentence about history?', 'Does the sentence contain information related to history?', 'Is the topic of the sentence about biology?', 'Is the topic of the sentence about history?', 'Is the topic of the sentence about government?', 'Is the sentence category/subject psychology?', 'Is the topic of the sentence about psychology?', 'Does the sentence contain information related to sociology?', 'Does the sentence contain information related to biology?', 'Does the sentence contain information related to government?', 'Is the sentence category/subject sociology?', 'Is the sentence category/subject biology?', 'Does the sentence contain information related to economics?', 'Is the sentence about physics?', 'Is the topic of the sentence about physics?', 'Is the sentence about biology?', 'Is the sentence about government?', 'Is the topic of the sentence about economics?', 'Is the topic of the sentence about sociology?', 'Is the sentence about sociology?', 'Is the sentence about psychology?', 'Does the sentence contain information related to psychology?', 'Does the sentence contain information related to physics?', 'Is the sentence category/subject economics?', 'Is this an example of demonstrating nonsense?', 'Does this example lack common sense?', 'Is this an example of not showing common sense?', 'Is the given excuse a valid excuse for the given scenario?', 'Is the given excuse reasonable for the given scenario?', 'Is the reason included in given scenario deemed reasonable?', 'Is the given reason in the scenario justifiable?', 'Does the headline look like a clickbait?', 'Is it a clickbait?', 'Is this a real news?', 'Is this a fake news?', 'Does the customer need a smaller replacement?', 'Does the customer need a larger replacement?', 'Is this a spoiler?', 'Does the review contain spoilers?', 'Does the headline contain sarcasm?', 'Is the headline sarcastic?', 'Does this tweet announce an emergency or a disaster?', 'Is the text about biological, biotic monitoring in water or in a river basin?', 'Does the text discuss biological, biotic monitoring in water or in a river basin?', 'Does the text discuss climatic indicators?', 'Is the text about climatic indicators?', 'Does the text discuss environmental issues?', 'Is the text about an environmental problem?', 'Does the text discuss environmental pollutions', 'Is the text about environmental pollution?', 'Does the text discuss treatment plants or environmental technologies?', 'Is the text about treatment plants or environmental technologies?', 'Does the hypothesis follow from the premise?', 'Does the hypothesis contradict the premise?', 'Does the Hypothesis hold a neutral point of view toward the Premise?', 'Is the hypothesis an entailment of the premise?]

Figure 10: Label descriptions from the same group of datasets that are considered similar. "*" at the beginning indicates that we are evaluating on this dataset.

| | Δ | > 1% | < -1% | > 5% | < -5% | > 10% | < -10% | $std(\Delta)$ |
|------------------------|----------|-------|-------|-------|-------|-------|--------|---------------|
| Meta-tuned vs QA | 3.3% | 59.5% | 28.1% | 31.4% | 10.3% | 15.7% | 5.9% | 9.5% |
| 220 vs 770M (T5) | 6.3% | 75.1% | 15.1% | 47.6% | 2.7% | 27.0% | 0.5% | 8.1% |
| Pre-trained vs. Random | 23.8% | 95.7% | 3.2% | 91.4% | 1.6% | 83.2% | 1.1% | 14.0% |
| Ensemble | 0.7% | 28.9% | 16.8% | 8.7% | 1.7% | 1.7% | 0.6% | 3.1% |
| Initialized with QA | 1.1% | 54.1% | 24.3% | 24.3% | 11.9% | 6.5% | 4.9% | 6.9% |
| Train on similar | 0.7% | 43.8% | 20.5% | 6.5% | 4.3% | 1.6% | 1.1% | 3.2% |
| 60 vs 220M (T5) | 14.4% | 86.5% | 10.3% | 79.5% | 4.3% | 61.1% | 2.2% | 12.6% |
| 41 vs. 110M (BERT) | 4.3% | 65.9% | 22.7% | 40.0% | 10.8% | 20.5% | 5.9% | 9.1% |
| 110 vs. 340M (BERT) | 1.4% | 46.5% | 35.7% | 23.8% | 17.3% | 11.4% | 6.5% | 8.5% |

Table 3: All results, with metrics explained in Section 3 and Appendix E. Each label description is weighted equally.

| | $\bar{\Delta}$ | > 1% | < -1% | > 5% | < -5% | > 10% | < -10% | $std(\Delta)$ |
|------------------------|----------------|-------|-------|-------|-------|-------|--------|---------------|
| Meta-tuned vs QA | 3.0% | 57.5% | 30.7% | 31.3% | 11.5% | 16.2% | 7.3% | 10.2% |
| 220M vs 770M (T5) | 5.8% | 75.8% | 15.5% | 46.9% | 3.5% | 25.6% | 1.4% | 7.8% |
| Pre-trained vs. Random | 23.7% | 93.5% | 5.5% | 89.4% | 3.4% | 82.5% | 2.1% | 15.1% |
| Ensemble | 0.5% | 25.0% | 18.8% | 6.9% | 1.6% | 1.7% | 0.7% | 3.1% |
| Initialized with QA | 1.2% | 54.0% | 24.0% | 26.0% | 11.8% | 8.1% | 5.3% | 7.3% |
| Train on similar | 0.7% | 44.5% | 20.1% | 6.0% | 4.3% | 1.7% | 0.8% | 3.1% |
| 60 vs 220M (T5) | 15.2% | 85.7% | 11.4% | 79.1% | 3.9% | 62.5% | 1.9% | 13.3% |
| 41 vs. 110M (BERT) | 4.8% | 67.0% | 21.5% | 41.9% | 9.2% | 22.5% | 4.9% | 9.0% |
| 110 vs. 340M (BERT) | 1.1% | 44.3% | 36.3% | 21.9% | 18.2% | 11.0% | 7.3% | 8.5% |

Table 4: All results, with metrics explained in Section 3 and Appendix E. Each label is weighted equally.

| | $\bar{\Delta}$ | > 1% | < -1% | > 5% | < -5% | > 10% | < -10% | $std(\Delta)$ |
|------------------------|----------------|-------|-------|-------|-------|-------|--------|---------------|
| Meta-tuned vs QA | 1.2% | 55.4% | 35.7% | 31.2% | 17.7% | 15.6% | 13.6% | 11.2% |
| 220 vs 770M (T5) | 6.3% | 77.4% | 16.5% | 51.7% | 7.0% | 31.6% | 4.5% | 9.0% |
| Pre-trained vs. Random | 20.2% | 89.8% | 8.5% | 84.8% | 6.1% | 76.6% | 1.5% | 15.1% |
| Ensemble | 0.1% | 18.6% | 20.2% | 4.3% | 1.9% | 1.5% | 1.2% | 2.8% |
| Initialized with QA | 2.3% | 59.2% | 22.5% | 34.3% | 9.9% | 13.9% | 5.7% | 7.2% |
| Train on similar | 0.6% | 48.8% | 25.4% | 7.3% | 5.7% | 1.3% | 0.9% | 3.3% |
| 60 vs 220M (T5) | 12.1% | 84.6% | 12.9% | 73.6% | 3.5% | 52.9% | 2.2% | 11.6% |
| 41 vs. 110M (BERT) | 7.0% | 74.6% | 13.8% | 58.5% | 6.8% | 31.5% | 2.9% | 8.9% |
| 110 vs. 340M (BERT) | 1.1% | 45.6% | 36.1% | 25.5% | 18.6% | 10.8% | 9.3% | 8.8% |

Table 5: All results, with metrics explained in Section 3 and Appendix E. Each dataset is weighted equally.

| Evaluation Dataset | Most Relevant Training Dataset |
|--|--|
| SemEval 2016 Task 6, stance classifications on issues like feminism, atheism, etc | SemEval 2019 Task 5, detecting hate speech against women and immigrants |
| SemEval 2019 Task 6, classifying whether the text is offensive | A dataset from Kaggle that classifies sexually explicit comments |
| SemEval 2019 Task 5, detecting hate speech against women and immigrants | SemEval 2016 Task 6, stance classifications on issues like feminism, atheism, etc |
| TREC, classifying the type the question is asking about (e.g. numbers, acronyms, human/occupations, etc) | AG News, which classifies news into different categories (e.g. sports, world events). |
| SemEval 2019 Task 8, classifying whether the question is asking for subjective opinion, factual information, or simply having a conversation | N/A |
| SUBJ, classifying whether the text contains subjective or objective information | N/A |
| QQP, classifying whether two questions have the same meaning | N/A |
| Yin et al. (2019) emotion classification, classifying text into 9 emotion types, such as "joy", "anger", "guilt", "shame", etc. | Classifying whether an IMDB movie review is positive. |
| Yin et al. (2019) situation classification, classifying which disaster situation people are experiencing, e.g. "regime change", "crime and violence", and what resource they need, e.g. "food and water", "search and rescue". | Classifying (binary) whether a tweet is related to a natural disaster. |
| Yin et al. (2019) topic classification, classifying the domain of an article into domains such as "family and relationship", "education", "business", "sports" | classifying the domain of a paper abstract into physics, maths, computer sciences, and statistics. |
| AG News, which classifies news into different categories (e.g. sports, world events). | Abstract Domain classification, classifying the domain of a paper abstract into physics, maths, computer sciences, and statistics. |
| Abstract Domain classification, classifying the domain of a paper abstract into physics, maths, computer sciences, and statistics. | AG News, which classifies news into different categories (e.g. sports, world events). |
| IMDB movie reviews, classifying whether the user feels positive about the movie | Stock market sentiment, classifying whether a comment is optimistic about the market. |
| CoLA, classifying whether a sentence is grammatical | N/A |
| SemEval 2020 Task 6, classifying whether a sentence contains a definition | N/A |
| Spam classification, classifying whether a text message is a spam | click-bait classification, classifying whether the title of an article is a clickbait. |
| SemEval 2018 Task 1, classifying a tweet as one of 4 emotion types {"sadness", "joy", "anger", "optimism"} | Classifying whether an IMDB movie review is positive. |
| SemEval 2018 Task 3, classifying whether a tweet is ironic | classifying whether a news title is sarcastic. |

Table 6: For each dataset that we evaluate on, we list the task in the training split that we consider to be the most relevant. We list "N/A" if we think that none of the training dataset is particularly relevant.

| | QA | QA + Meta | Meta | T5 220M | BERT 340M |
|----------------------------------|-------|-----------|-------|---------|-----------|
| Abstract Classification | 76.9% | 84.3% | 81.2% | 68.0% | 85.3% |
| AG News | 76.5% | 82.0% | 77.8% | 69.9% | 69.5% |
| Stance (Hillary) | 74.8% | 79.8% | 73.8% | 69.0% | 63.2% |
| Hate Speech | 59.4% | 66.0% | 64.1% | 59.6% | 69.2% |
| Stance (Feminism) | 67.8% | 71.6% | 69.1% | 61.0% | 64.8% |
| Stance (Climate) | 75.8% | 81.7% | 79.6% | 72.0% | 76.2% |
| Emotion Classification* | 67.6% | 70.5% | 68.0% | 65.0% | 64.0% |
| Emotion Classification (SemEval) | 81.6% | 85.2% | 81.7% | 76.1% | 74.2% |
| Irony Detection | 67.9% | 83.4% | 80.2% | 61.0% | 64.9% |
| Stance (Atheism) | 60.2% | 62.4% | 65.6% | 55.1% | 60.9% |
| QQP | 54.1% | 61.1% | 68.6% | 56.7% | 66.9% |
| TREC | 59.3% | 63.9% | 76.4% | 73.4% | 66.9% |
| Stance (Abortion) | 58.2% | 61.3% | 62.8% | 60.5% | 59.5% |
| Offensive Speech | 76.6% | 80.4% | 79.5% | 74.5% | 80.6% |
| CoLA | 52.3% | 49.4% | 49.8% | 49.6% | 50.0% |
| SUBJ | 62.8% | 66.8% | 58.7% | 54.5% | 50.2% |
| Situation Classification* | 73.9% | 80.4% | 79.3% | 75.5% | 79.5% |
| SPAM Detection | 57.2% | 45.4% | 35.0% | 49.3% | 47.8% |
| IMDB Movie Review | 92.9% | 94.0% | 90.5% | 67.7% | 84.4% |
| Topic Classification* | 77.6% | 82.7% | 84.0% | 77.5% | 80.7% |
| Definition Detection | 72.8% | 73.5% | 63.9% | 63.6% | 60.2% |
| Question Type Classification | 75.1% | 73.8% | 59.3% | 51.8% | 64.5% |

Table 7: Zero shot performance of each model on each dataset. “QA” means the UnifiedQA model; “QA + Meta” means meta-tuning with UnifiedQA initialization; “Meta” means meta-tuning on T5 (770M) parameters. To save space, we use “*” to denote datasets from [Yin et al. \(2019\)](#).

| Dataset name | #classes | Accuracy |
|-------------------------------------|----------|----------|
| 2016SemEval6TweetEvalStanceAtheism | 3 | 66 |
| KaggleNewsTopicClassification | 4 | 64 |
| 2019SemEval6TweetEvalOffensive | 2 | 28 |
| 2019SemEval8Qtype | 2 | 73 |
| 2018SemEval3TweetEvalIrony | 2 | 39 |
| 2016SemEval6TweetEvalStanceHillary | 3 | 55 |
| subj | 2 | 61 |
| trec | 6 | 38 |
| KaggleQuoraQPairs | 2 | 50 |
| definition | 2 | 32 |
| BenchmarkingZeroshotTopic | 10 | 59 |
| 2019SemEval5TweetEvalHate | 2 | 42 |
| cola | 2 | 55 |
| 2018SemEval11TweetEvalEmotion | 4 | 72 |
| 2016SemEval6TweetEvalStanceAbortion | 3 | 64 |
| KaggleIMDBMovieReview | 2 | 85 |
| 2016SemEval6TweetEvalStanceClimate | 3 | 61 |
| KaggleSMSSPAM | 2 | 14 |
| 2016SemEval6TweetEvalStanceFeminist | 3 | 53 |

Table 8: We report the accuracy of the meta-tuned model for completeness according to the request of the reviewers. However, given that accuracy is very sensitive to thresholding (Zhao et al., 2021) and is generally unreliable when the labels are imbalanced, these numbers are not likely to be informative. Additionally, to speed up evaluation, we use a subsample of the original test split for some datasets, so these numbers are not directly comparable to those in the other papers either.