Towards Agile Text Classifiers for Everyone
Maximilian Mozes\textsuperscript{1,2}\textsuperscript{*} Jessica Hoffmann\textsuperscript{1}\textsuperscript{*} Katrin Tomanek\textsuperscript{1} Muhamed Kouate\textsuperscript{1,}\textsuperscript{†}
Nithum Thain\textsuperscript{1} Ann Yuan\textsuperscript{1} Tolga Bolukbasi\textsuperscript{1} Lucas Dixon\textsuperscript{1}
\textsuperscript{1}Google Research
\textsuperscript{2}University College London
\{jhoffmann,katrintomanek,kouate,nthain,annyuan,tolgab,ldixon\}@google.com
maximilian.mozes@ucl.ac.uk

Abstract

Text-based safety classifiers are widely used for content moderation and increasingly to tune generative language model behavior—a topic of growing concern for the safety of digital assistants and chatbots. However, different policies require different classifiers, and safety policies themselves improve from iteration and adaptation. This paper introduces and evaluates methods for agile text classification, whereby classifiers are trained using small, targeted datasets that can be quickly developed for a particular policy. Experimenting with 7 datasets from three safety-related domains, comprising 15 annotation schemes, led to our key finding: prompt-tuning large language models, like PaLM 62B, with a labeled dataset of as few as 80 examples can achieve state-of-the-art performance. We argue that this enables a paradigm shift for text classification, especially for models supporting safer online discourse. Instead of collecting millions of examples to attempt to create universal safety classifiers over months or years, classifiers could be tuned using small datasets, created by individuals or small organizations, tailored for specific use cases, and iterated on and adapted in the time-span of a day.

1 Introduction

Conversation moderation has changed rapidly over the past decade as platforms have evolved new tools. In the last few years, general purpose classifiers supporting online discourse, like Perspective API, have seen broad adoption; they are used to assist moderation (Rieder and Skop, 2021a), give feedback to authors (Simon, 2020), and advance research in online safety. The Perspective API’s most widely used model is for toxicity detection and was trained on hundreds of millions of annotations (Borkan et al., 2019). This results in a useful model for its domain distribution. However, the way people use language is continuously changing, many forums have different policies, and policies themselves change frequently (e.g., to deal with new topics, such as COVID-19). In practice, one faces the challenge of either using the model with the quality degradation caused by the distribution and policy shift, or of training a custom model. Training a new high-quality neural text classification model (i.e., fine-tuning models such as BERT (Devlin et al., 2019)) typically requires collecting thousands or millions of textual annotations, a process which is both time-consuming and cost-intensive.

In parallel to these developments, there has been rapid progress in chatbots, with ChatGPT representing a particular turning point in public awareness of the capabilities of large language models (LLMs).\textsuperscript{1} An important emerging strand of this research explores the role of human feedback and reinforcement learning in mitigating safety concerns.

---

\textsuperscript{*} Equal contribution.
\textsuperscript{†} Work done during an internship at Google Research.
\textsuperscript{1}https://openai.com/blog/chatgpt/
cerns with their outputs (Bai, 2022; Glaese, 2022). This field has not yet arrived at consensus on policies, nor developed high-quality large datasets for the proposed policies. Classifiers are already used to scale human feedback for tuning models, but being able to quickly iterate on high-quality text classifiers could play a particularly important role for the safety of modern chatbots.

This paper explores alternative approaches to text classification that leverage modern generative large language models like T5 (Raffel et al., 2020) and GPT-3 (Brown et al., 2020). Not only can LLMs generate comments and conversations, they can also themselves act as safety classifiers, detecting which comments may need moderation (Chaudhary et al., 2021; Rieder and Skop, 2021b). We explore prompting these models in two ways: using few-shot examples—often referred to as in-context learning (ICL), which includes task demonstrations in the input prompt and does not require model parameter updates —and parameter-efficient tuning (PET), a branch of transfer learning that adapts only a small number of parameters for new tasks and has been shown to provide high performance (up to 500x more annotated data, while also requiring only a fraction of the parameters (up to 1300x less). We also found our methods significantly surpass widely used classifiers trained on tens of millions of annotations on closely related domains. We argue this represents a new paradigm for training agile, domain-adapted safety classifiers using PET for LLMs.

2 Related work

2.1 Safe online dialogue

Over the past decade, a significant branch of research has been motivated by the goal of improving the safety of online conversations. A popular approach is developing classifiers to detect whether individual comments contain toxicity, personal attacks, hate speech, and other unhealthy attributes of online conversations (Wulczyn et al., 2017; Davidson et al., 2017). Classifiers for positive attributes such as constructiveness (Kolhatkar et al., 2020a) have also been explored. The underlying models for detecting these attributes have also grown in sophistication, from linear models such as support vector machines (MacAvaney et al., 2019) to more modern approaches that rely on deep learning architectures like CNNs (Gambäck and Sikdar, 2017) and, more recently, transformers (Caselli et al., 2020; Zhou et al., 2021).

With the rapid progress of chatbots, online chat support, and digital assistants, there has been a growing focus on multi-turn dialogue, with the goal of improving conversational agents and making them more robust to adversarial users (Dinan et al., 2019; Xu et al., 2021). Another recent area of research has focused on reducing the toxicity generated by large language models, although this risks diminished performance for the language of marginalized groups (Welbl et al., 2021).
Our contribution explores 9 ratings schemes across 7 datasets related to classifications of negative and positive characteristics of multi-turn dialogue and online commentary.

2.2 Large language models

Large language models, also known as foundation models, use the transformer architecture (Vaswani et al., 2017), typically have tens to hundreds of billions of parameters and are pretrained on datasets consisting of hundreds of billions of tokens. These pretrained models, when fine-tuned, have demonstrated state-of-the-art performance on a multitude of tasks (Devlin et al., 2019; Chowdhery et al., 2022; Tran et al., 2022). One of the most important characteristics of LLMs is that they can be prompted with input texts to configure them to perform new tasks (Radford et al., 2019; Brown et al., 2020). This ability to use text in the input to solve new tasks is called in-context learning (ICL).

The prompting method that has the highest performance and which can directly use a small fragment of an existing task dataset is called few-shot prompting. This involves inserting a few examples of the task (typically between 2 and 20 examples) into the LLM prompt before the new example input. There is now a significant branch of research on improving ICL performance using, for example, self-consistency (Wang et al., 2022) or chain-of-thought reasoning (Wei et al., 2022).

Our contribution focuses on evaluating the performance of few-shot prompts, leaving further comparisons to the quickly growing catalog of prompting tricks as further work.

2.3 Parameter-efficient tuning

Domain adaptation for pretrained language models has typically achieved state-of-the-art performance by fine-tuning all model parameters on relatively large datasets (tens of thousands to millions of annotations; Peters et al., 2018; Devlin et al., 2019). However, a recent branch of research has found that one can keep the majority of the parameters in a large pretrained LLM fixed, instead updating only a small fraction of the parameters, with comparable downstream task performance (Li and Liang, 2021b; Lester et al., 2021). This branch of research has largely focused on the trade-offs between how many parameters to tune, which parameters to tune (including the addition of new parameters), how to initialize them (Vu et al., 2022b), and the resulting performance. However, a few new studies have indicated that these methods can also perform better than ICL when given the same tiny set of examples (Liu et al., 2022; Agrawal et al., 2022).

We specifically focus on prompt-tuning, which prepends a set of learnable token embeddings (also called soft prompts) to an LLM’s input via concatenation. The soft prompt vector is optimized during training using the cross-entropy loss, while the LLM parameters remain fixed. At inference time, the trained token embeddings are then fed into the model along with the input prompt. Prompt-tuning is one of the most parameter-efficient approaches to PET; this allows the LLM to remain fixed and the soft prompt to be provided at query time with the input. In particular, it avoids having to manage multiple copies of the LLM (Houlsby et al., 2019), or swap state for a significant number of parameters during inference.

A novel characteristic of our contribution to the PET field is that we explore the trade-offs when one has more data than can fit into the LLM’s input context for ICL, but less than is effective for fine-tuning: in the range of tens of examples to 2,000. We argue that this is an important scale of data to consider as it represents what a small organization is likely to meaningfully create for a specialized text classification.

3 Datasets

Our datasets cover three broad domains related to the quality of online discourse: classification of offensiveness to support safer chatbot dialogue (Section 6.1), attributes of online comments related to toxicity that result in unhealthy conversations (Section 6.2), and expressing responses in a neutral way, inspired by Wikipedia’s policies (Section 6.3).

3.1 Dialogue safety

For the dialogue safety domain, we consider 5 datasets for our experiments based on the ParlAI (Dinan et al., 2019) and Bot Adversarial Dialogue (BAD; Xu et al., 2021) data collection efforts.

ParlAI. We consider three independent datasets from ParlAI, namely ParlAI Single Standard, ParlAI Single Adversarial, and ParlAI Multi. All three datasets come with a pre-defined split as introduced by Dinan et al. (2019), containing 24,000 training examples, and 3,000 each for validation and testing.

ParlAI Single Standard and ParlAI Single Adversarial are both single-turn conversational datasets.
For the former, crowdworkers were simply asked to construct sentences that they would consider offensive. The latter, in contrast, was built by asking crowdworkers to submit sentences that are offensive, but are predicted to be safe by a classifier.

ParlAI Multi is a multi-turn conversational dataset, consisting of sequences representing multiple turns of human conversations in which the last utterance is meant to be offensive.

**Bot Adversarial Dialogue.** The Bot Adversarial Dialogue (BAD) dataset, presented in Xu et al. (2021), contains a collection of dialogues related to conversational toxicity. The dataset was collected by asking humans to converse with a bot, with the intention to lead the bot into generating offensive output. The dataset comes with a predefined split of 5,080 conversations for training, 513 for validation, and 191 for testing.

Since the conversations in the dataset can be lengthy (up to 14 turns), we follow Xu et al. (2021) and truncate conversations. Specifically, we experiment with versions of the dataset where we only consider the last four (BAD-4) and two (BAD-2) utterances of the conversation.

### 3.2 Unhealthy Comment Corpus

The Unhealthy Comment Corpus (UCC; Price et al., 2020) consists of 44,355 comments from the Globe and Mail news site (randomly sampled from the Simon Fraser University Opinion and Comment Corpus dataset (Kolhatkar et al., 2020b)), labeled by crowdworkers for 7 labels. Each comment is labeled as either healthy or unhealthy, in addition to binary labels for the presence of six unhealthy sub-attributes: (1) hostile; (2) antagonistic, insulting, provocative or trolling; (3) dismissive; (4) condescending; (5) sarcastic; and (6) generalization. It is worth mentioning that this dataset is highly imbalanced, with positive examples (the unhealthy attributes) comprising less than 10% across splits and attributes (details can be found in Table 3 in the Appendix). This is typical of datasets sourced from online discourse labeled with forms of toxicity.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Total (pos. percentage)</th>
<th>Human AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple perspectives</td>
<td>113 (75.33 %)</td>
<td>0.965 ± 0.023</td>
</tr>
<tr>
<td>Neutral</td>
<td>76 (50.67%)</td>
<td>0.960 ± 0.025</td>
</tr>
<tr>
<td>Well-explained</td>
<td>91 (60.67%)</td>
<td>0.876 ± 0.040</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the Neutral Responses dataset, including the absolute number as well as percentage of positive examples per attribute, as well as the Human AUC baseline. Each example was labeled by three expert annotators.

To experiment with the effectiveness of prompt-tuning on a novel task, we built a new Neutral Responses dataset comprised of human-written texts annotated according to three attributes that one might want from a chatbot’s response to a difficult or polarizing question. The dataset contains 150 examples, each composed of a topic, a question regarding the topic, an answer to the question, and yes/no annotations labeling the answer along three attributes: whether it covers multiple perspectives on the topic, whether it is written from a neutral perspective (without expression of opinion or judgement), and whether it is well-explained (see Appendix A for additional dataset details and examples). The 150 examples span 75 different topics which were selected to be controversial. Each question and answer was written by an expert annotator. This expert and two more experts then annotated the data along each attribute. The final label is obtained by majority vote.

The distribution of attributes is shown in Table 1. The answers were crafted to limit the class imbalance so there are sufficiently many positive and negative examples to train a classifier. Additionally, each consistent triplet of values of attributes (multiple perspectives, not neutral, not well-explained) is also represented in a relatively balanced proportion: every triplet of values comprises between 12% and 34% of the dataset. In our experiments, when sampling the splits for train and test sets, we use stratified sampling to approximately match the label distribution of the overall training set.

The overall inter-annotator agreement measures for both Krippendorff’s alpha and Fleiss’s kappa are 0.72. This is a strong agreement for subjective tasks of this nature; for reference, inter-annotator agreement for crowdsourced attributes related to online toxicity typically falls into the 0.4–0.6 range (Wulczyn et al., 2017). We also compute these measures by attributes: the Multiple perspectives attribute achieves a Krippendorff’s agreement: every triplet of values comprises between 12% and 34% of the dataset. In our experiments, when sampling the splits for train and test sets, we use stratified sampling to approximately match the label distribution of the overall training set.

The overall inter-annotator agreement measures for both Krippendorff’s alpha and Fleiss’s kappa are 0.72. This is a strong agreement for subjective tasks of this nature; for reference, inter-annotator agreement for crowdsourced attributes related to online toxicity typically falls into the 0.4–0.6 range (Wulczyn et al., 2017). We also compute these measures by attributes: the Multiple perspectives attribute achieves a Krippendorff’s agreement:

Note that some attribute labels are incompatible.
alpha of 0.76 (resp. Fleiss’s kappa of 0.76), Neutral achieves 0.79 (resp. 0.80), and Well-Explained 0.59 (resp. 0.60).

To provide a strong baseline for classifier quality for each attribute, we also compute the AUC for each annotator against the majority vote, and report the average as the human mean AUC; this is artificially high as the annotator contributes to the majority baseline. We also see that as the mean AUC for these tasks decreases, the standard deviation increases. This is an aspect of agreement measured by AUC, and correlates with subjective feedback from the annotators on task difficulty.

### 4 Models and prompt-tuning

For all three domains, we train soft prompts based on T5 XXL (Raffel et al., 2020) and PaLM 62B (Chowdhery et al., 2022). Soft prompt tokens have embedding dimensionality according to the language model, which results in 4,096- and 8,192-dimensional embeddings for T5 XXL and PaLM 62B, respectively. We follow Lester et al. (2021) and use an adapted version of T5 XXL, which has been trained on a prefix language modeling objective. For all prompt-tuning experiments we follow Lester et al. (2021) and initialize each prompt with a random sample of vocabulary token embeddings from the respective model’s 5,000 most frequent tokens.

We train soft prompts on 10, 20, 50, 80, 100, 200, 500, 1,000, and 2,000 randomly sampled training examples. For the Neutral Responses dataset, we use only up to 100 examples due to the dataset’s more limited size. To account for variability in the sampling process, we repeat each experiment three times with different seeds and report average scores.

**Dialogue safety and UCC.** Soft prompts consist of 10 tokens, resulting in a total of 40,960 (T5 XXL) and 81,920 (PaLM 62B) learnable parameters per task, constituting only a small fraction of the 11 and 62 billion total parameters of each LLM. We train each prompt for 20 epochs, validate the loss after each epoch on a sampled subset of 500 validation examples, and select the best-performing checkpoint for testing. Each experiment uses the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.1 and weight decay 0.00001. For PaLM 62B we use a batch size of 4, and for T5 XXL one of 32.

**Neutral Responses dataset.** Soft prompts of 5 tokens are trained for 1,000 steps for both PaLM 62B and T5 XXL. We use the AdaFactor optimizer (Shazeer and Stern, 2018) with a constant learning rate of 0.1, a batch size of 16 and a prompt of length 5. For T5 XXL, we also set the dropout rate to 0.1.

**Evaluation metrics.** For the dialogue safety experiments we report model performance in terms of binary $F_1$ on the positive (toxic) class, in line with experiments in Dinan et al. (2019). For both the Unhealthy Comments and Neutral Responses datasets, we report ROC-AUC scores. We obtain classification scores in the zero to one range from the LLMs by scoring specific tokens corresponding to the output class labels (e.g., the tokens yes and no), applying softmax, and then taking the score value of the positive class. This provides a threshold-agnostic and dataset-imbalance-agnostic metric, and allows comparison to the previous reported performance results.

### 5 Baselines

#### 5.1 In-context learning

We compare prompt-tuning to in-context learning (ICL) baselines, which include training data directly in the prompt-template sent to the model. We conduct experiments with zero-shot, 6-shot and 12-shot prompt-templates. For the latter two, for each seed we sample a fixed set of few-shot examples to be used for the inference prompt. For both few-shot learning and prompt-tuning, half of the training set is sampled from the positive class, half from the negative class. We repeat the few-shot experiments three times with different seeds to account for variation due to random sampling.

#### 5.2 Perspective API

For the dialogue safety and online comments tasks, we further compare prompt-tuning with the Perspective API baseline: an off-the-shelf toxicity classifier which computes a toxicity confidence value for a given input text.\(^7\) In our experiments, we consider the eight attributes TOXICITY, SEVERE_TOXICITY, IDENTITY_ATTACK, INSULT, PROFANITY, THREAT, FLIRTATION and SEXUALLY_EXPLICIT. For each test set example, we compute the Perspective score individually for each category, and use a threshold-based approach.

\(^7\)https://perspectiveapi.com/
with threshold values of 0.0, 0.1, 0.2, \ldots, 0.9 to predict whether a given piece of text is toxic. In the results, we only report the highest achieved performance across thresholds and attributes.

6 Results

6.1 Dialogue safety

Performance results on the dialogue safety datasets can be found in Table 2, while Figure 1 shows the quality of prompt-tuning as the number of training examples increases on the ParlAI Single Adversarial dataset. The plots for the ParlAI Single Standard, ParlAI Multi, BAD-2 and BAD-4 datasets are similar, and can be found in Appendix B. For each dataset, we show the best-performing few-shot baselines. For the three ParlAI datasets we also show the test set scores reported in Dinan et al. (2019), obtained by fine-tuning BERT-Base on the entire training set containing 24,000 samples in each respective dataset.

We observe that across the dialogue safety datasets, prompt-tuning PaLM 62B outperforms T5 XXL when trained on both 80 and 2,000 examples. Furthermore, we can see that for T5 XXL, there is a critical change in behavior as the training data increases (e.g., 2,000 instead of 80): in two cases, T5 XXL jumps from being significantly worse than random to outperforming the previous state-of-the-art (e.g., increasing from 0.18 to 0.91 $F_1$ on ParlAI Single Adversarial). Such differences are much less clear for PaLM 62B, since the model performs well with as little as 80 examples across datasets, where it already outperforms the previous state-of-the-art (Dinan et al., 2019).

Taking a closer look at the differences between few-shot learning and prompt-tuning, we observe that utilizing PaLM 62B for few-shot learning also suffices to perform on par with (ParlAI Single Adversarial) or outperform (ParlAI Single Standard) the previous state-of-the-art, indicating that few-shot learning represents a competitive baseline in this setting. However, prompt-tuning on 80 training examples suffices to outperform the few-shot ICL in four out of the five tasks. Note that a comparison to few-shot learning with 80 examples is not possible due to the context window restrictions for T5 XXL and PaLM 62B.

Both LLMs perform substantially better on single-turn datasets (ParlAI Single Standard, ParlAI Single Adversarial) compared to multi-turn datasets (ParlAI Multi, BAD-2, BAD-4). This demonstrates the difficulty of detecting safety concerns in multi-turn conversations.

6.2 Unhealthy Comment Corpus

For UCC, like previous experiments, we ran few-shot and prompt-tuning experiments with PaLM 62B and T5 XXL to classify comments for each class by prompting the models with a target text and the question Is the text above {class}? This is inspired by the Question and Answering template presented in Rae et al. (2021), which adapts the verbalization methods proposed by Schick et al. (2021).

To compare with the previous state-of-the-art, we report the performance in terms of ROC-AUC, and we use the same test set as Price et al. (2020). We contrast our results on the UCC dataset with their Human and BERT baselines, which evaluate performance according to assessments from human crowdworkers as well as a BERT model fine-tuned on the entire training dataset.

In Figure 2, we show a typical example of how AUC varies as we use more training examples for prompt-tuning (results for the remaining attributes can be found in Appendix C). The trends are comparable to Figure 1.

Quantitative results can be found in Table 2 (Unhealthy Comment Corpus). As can be seen, with only 80 training examples, prompt-tuning...
### Table 2: Summary of results for the dialogue safety ($F_1$ score), Neutral Responses (ROC-AUC), and Unhealthy Comments datasets (ROC-AUC), averaged over three seeds. We compare ICL for PaLM 62B and T5 XXL across 0, 6, and 12 shots on a validation set, and report the results of the best model (PaLM 62B, either 6 or 12 shots). We add the human agreement baseline and the previous state-of-the-art (Dinan et al., 2019) for dialogue safety, and results with BERT for the Unhealthy Comment Corpus (Price et al., 2020). T5 XXL prompt-tuned on 2,000 examples outperforms the human agreement. PaLM 62B prompt-tuned on 80 examples also outperforms human agreement, and achieves SOTA. PaLM 62B prompt-tuned on 2,000 examples shows that the quality of results keeps improving as the dataset size increases.

PaLM 62B outperforms the human baseline, and is comparable to the BERT baselines across all experiments. Performance scores for PaLM 62B increase further when training on more examples up to our upper bound of 2,000. However, as in the results discussed in Section 6.1, the performance differences between 80 and 2,000 examples are much more drastic for T5 XXL, showing absolute AUC improvements of around 0.3 for Generalization and 0.2 for Antagonistic, Condescending, Dismissive, Hostile, and Sarcastic. Additionally, T5 XXL outperforms the BERT baselines in two out of the five cases (Generalization and Sarcastic). While this shows that prompt-tuning T5 XXL can be competitive with full model fine-tuning, it also demonstrates the benefit of using a larger model (i.e., PaLM 62B) for prompt-tuning on small datasets.

We also observe that the PaLM 62B few-shot baseline is competitive on many of the attributes in this dataset, nearly reaching or outperforming the BERT baseline in terms of AUC. However, few-shot ICL is also constrained by the context window that creates a hard limit on the number of examples that can be provided, and has high variance (depending significantly on the specific examples in the few-shot prompt).

Overall, these results indicate that with fewer than 100 examples, prompt-tuning PaLM 62B achieves competitive performance across attributes, demonstrating its ability to serve as a method for efficiently building safety classifiers in data-scarce settings.

### 6.3 Neutral Responses
Quantitative results are reported in Table 2 (Neutral Responses), and a typical plot is shown in Figure 3 (the plots for the remaining attributes can be found in Appendix D). From left to right, the tasks are shown by increasing order of complexity for humans. Unsurprisingly, the number of examples needed to achieve (or surpass) the human baseline increases with the complexity of the task. Like in our results with the other datasets, we observe that prompt-tuning with 80 examples is enough to get human-level accuracy (and to exceed few-shot prompting).
In contrast to the previous two domains, the difference between T5 XXL and PaLM 62B appears to be smaller for the Neutral Responses dataset. This may be because it is an easier task, or because the training data quality is higher.

7 Further Work

We encourage researchers working towards healthier online discourse to further explore the new paradigm of small datasets and PET. While our results show that prompt-tuning is a strong method for our tasks, there are many other PET methods, e.g., prefix-tuning (Liu et al., 2022), that may have different properties for text classification at the scale of data in between fine-tuning and ICL few-shot prompts.

The limits of PET for text classifiers are also important areas to explore. For example, one would expect poor results when prompt-tuning an LLM pretrained on one language in order to obtain an effective classifier in a different language. A deeper understanding of these limits will be important for the maturation of the field. In line with this, another interesting direction for future work could focus on the potential of instruction-tuned models for PET in this context, since such models have shown promising performances on unseen tasks (Wei et al., 2021).

Understanding unintended biases in small datasets will also become critical if PET becomes more widely adopted as a tool for agile classifier development. This is because datasets written by an individual may induce more unintended biases in the resulting classifiers since a smaller number of people review the policy. On the other hand, if the datasets are much smaller it may also become simpler to review and correct them in the underlying dataset.

Finally, another important branch of future work is to investigate augmenting, scaling, and replacing aspects of current human annotation with synthetic generation. Early work in this direction for Question and Answer tasks has recently proved successful (Agrawal et al., 2022). By combining synthetic generation with prompt-tuning, we speculate that a rich methodology can be created for the agile development of text classifiers.

8 Conclusions

In this paper, we demonstrated that we can use LLM-based parameter-efficient tuning to build high-performance classifiers with small datasets (e.g., of as few as 80 examples) across three domains: 5 datasets related to offensive dialogue (Section 6.1), 7 annotation schemes related to toxicity in online comments (Section 6.2), and three attributes related to neutral responses to questions on sensitive topics (Section 6.3).

We focused on the prompt-tuning approach to PET; it is one of the most parameter-efficient PET methods, allowing a single model to be served and the task-specialization to be provided in a soft prompt vector at query time, much like an ICL prompt-template. In contrast, fine-tuning requires changing many more model parameters and serving a separate model per task.

Our results show that prompt-tuning on small datasets consistently achieves performance that is competitive with the previous state-of-the-art (e.g., BERT-based fine-tuning approaches that use much larger datasets of human-annotated examples). Prompt-tuning performance also appears to be equal to or better than human-annotation quality. We found that ICL with few-shot templates is sometimes very effective, but also has much more variable performance.

When prompt-tuning T5 XXL, we observed that much more data is needed for effective performance on most datasets; the model only becomes competitive with fine-tuned baselines when trained on thousands of examples. In contrast, prompt-tuning on the much larger PaLM 62B model requires less than a hundred examples. This sug-
gests that there is a form of scaling law for prompt-tuning, and that its utility dramatically improves as the model size grows.

We remark that the scale of data needed for prompt-tuning to achieve useful and even state-of-the-art performance in many tasks is now sufficiently small that a single individual could create the needed dataset. We argue that this represents a paradigm shift for text classification, especially in the domain of online safety: ICL and prompt-tuning large LLMs could enable even small forums to develop customized classifiers for their own evolving policies, rather than depend on centralized classifiers, trained on millions of annotations representing a single common policy.

References


Bernhard Rieder and Yarden Skop. 2021b. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of perspective api.


Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou’, and Daniel Cer. 2022a. SPoT: Better frozen model


A Neutral Responses Dataset

A.1 Detailed generative process of the dataset

The first expert gathered sources on each controversial topic, and was sometimes aided by two generative models: at first, they used a PaLM 540B model (Chowdhery et al., 2022) using examples from the subreddit r/eli5 as few-shot. Once a small subset of the dataset was created, they used that subset as few-shot for a FLAN-PaLMChilla 62B model (Chung et al., 2022). Most answers were written without that aid. Each answer written with that aid was heavily edited.

A.2 Examples

Examples from the neutral responses dataset can be seen in Table 4.

B Prompt-tuning results for dialogue safety

Prompt-tuning results on the ParlAI Single Standard, ParlAI Multi, BAD-2, and BAD-4 datasets can be found in Figure 4. In line with the results reported in Table 2, we observe that prompt-tuning PaLM 62B with 100 (ParlAI Single Standard) and 80 (ParlAI Multi) examples suffices to outperform the performance of the fine-tuned BERT as reported in Dinan et al. (2019). Similarly, for both BAD-2 and BAD-4, we can see that PaLM 62B outperforms both the Perspective and few-shot baselines when trained on 50 examples.

In contrast, T5 XXL requires a larger amount of training data for prompt-tuning (2,000 examples) on ParlAI Single Standard to outperform the BERT baseline. On the remaining three datasets, prompt-tuning T5 XXL does not achieve a performance better than the few-shot baseline.

C Prompt-tuning results for UCC

The prompt-tuning performances for the six remaining attributes of UCC can be found in Figure 5. Across attributes, we can see that prompt-tuning PaLM 62B achieves test set scores superior to the ones obtained with BERT fine-tuning. Interestingly, the model outperforms BERT with as little as 10 training examples (Generalization and Sarcastic). With 500 training examples, PaLM 62B outperforms BERT for all six attributes.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Train</th>
<th>Test</th>
<th>Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antagonistic</td>
<td>1689 (4.8%)</td>
<td>203 (4.6%)</td>
<td>174 (3.9%)</td>
</tr>
<tr>
<td>Condescending</td>
<td>1927 (5.4%)</td>
<td>269 (6.1%)</td>
<td>238 (5.4%)</td>
</tr>
<tr>
<td>Dismissive</td>
<td>1071 (3.0%)</td>
<td>150 (3.4%)</td>
<td>143 (3.2%)</td>
</tr>
<tr>
<td>Generalization</td>
<td>752 (2.1%)</td>
<td>96 (2.2%)</td>
<td>96 (2.2%)</td>
</tr>
<tr>
<td>Hostile</td>
<td>923 (2.6%)</td>
<td>108 (2.4%)</td>
<td>99 (2.2%)</td>
</tr>
<tr>
<td>Sarcastic</td>
<td>1501 (4.2%)</td>
<td>201 (4.5%)</td>
<td>195 (4.4%)</td>
</tr>
<tr>
<td>Unhealthy</td>
<td>2655 (7.5%)</td>
<td>320 (7.2%)</td>
<td>366 (8.3%)</td>
</tr>
</tbody>
</table>

Table 3: The absolute number as well as percentage of positive examples per split and attribute of the UCC dataset splits.
Table 4: Three examples from the neutral responses dataset.

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>Multiple perspectives</th>
<th>Neutral</th>
<th>Well-explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone believes we should legalize marijuana, right?</td>
<td>Some people believe legalizing marijuana would reduce crime and create jobs, while others oppose legalization because they believe cannabis is a gateway drug and legalizing it will increase hard drug use.</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>What is this about taxes and churches?</td>
<td>Churches are currently tax exempt according to US law. There is a part of the population who wants to keep the status quo, and another who is awaiting a change.</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Can you tell me about the water ban?</td>
<td>I’m so tired of hearing about the water ban. Nobody cares.</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Figure 4: Prompt-tuning results (test set $F_1$ averaged across three seeds with standard deviations visualized) for the ParlAI Single Standard, ParlAI Multi, and the two BAD datasets across models (T5 XXL, PaLM 62B). Prompt-tuning performance is compared to few-shot learning and the Perspective API baseline (only the best scores across categories and thresholds are shown). For few-shot, we show PaLM 62B’s best performing score across 0, 6, and 12 shots.

The results for T5 XXL reveal a different picture. While prompt-tuning achieves competitive performances across attributes, only for the attributes Generalization and Sarcastic does it manage to perform better than the BERT baseline.

Taking a closer look at the few-shot baseline, we observe that it performs close to or outperforms BERT in five out of the six categories (only for Hostile do we see a notable performance gap between the two). These results demonstrate the strength of the few-shot baseline, which is nevertheless outperformed by prompt-tuning PaLM 62B across all...
attributes.

D Prompt-tuning results for the Neutral Responses dataset

The prompt-tuning results on the attributes multiple perspectives and neutral can be found in Figure 6. As can be seen, for PaLM 62B, 20 training examples are enough to outperform the Human baseline for the multiple perspectives attribute, whereas for the neutral one around 80 examples are needed. Furthermore, prompt-tuning PaLM 62B performs better than the few-shot baselines for the majority of comparisons.

Looking at T5 XXL, we observe that the model does on average perform worse than PaLM 62B and needs 80 training examples to perform on par with the Human baseline. Nevertheless, we can see that prompt-tuning T5 XXL on 20 examples suffices to outperform the PaLM 62B few-shot baseline for multiple perspectives (80 examples are needed for neutral).

Overall, these results underline the effectiveness of prompt-tuning on small datasets.

E Computational budget

We estimate that it took around 1075200 GPU hours in total to create this research paper. The cost of reproducing our final results would be around 76800 GPU hours. Our model training times were in the range of 1 to 4 hours.
Figure 5: Prompt-tuning results (ROC-AUC) on six attributes of the UCC dataset using T5 XXL and PaLM 62B. HUMAN and BERT denote the human and BERT-based baselines from Price et al. (2020), respectively. For few-shot, we show PaLM 62B’s best performing score across 0, 6, and 12 shots.
Figure 6: Prompt-tuning results (ROC-AUC) on the two remaining attributes of the Neutral Responses dataset, using T5 XXL and PaLM 62B. For few-shot, we show PaLM 62B’s best performing score across 0, 6, and 12 shots.