Towards Agile Text Classifiers for Everyone

Maximilian Mozes[1,2*†] **Jessica Hoffmann**[1*] **Katrin Tomanek**[1] **Muhamed Kouate**[1†]
Nithum Thain[1] **Ann Yuan**[1] **Tolga Bolukbasi**[1] **Lucas Dixon**[1]
[1]Google Research
[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.

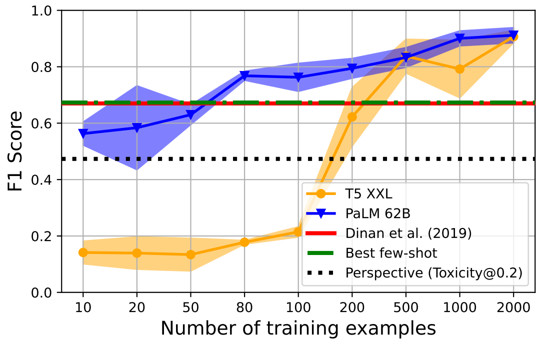


Figure 1: Prompt-tuning PaLM 62B and T5 XXL with as few as 80 and 500 examples, respectively, outperforms both in-context learning (12-shot) on PaLM 62B and a BERT model fine-tuned on 24,000 training examples as reported in Dinan et al. (2019) for the ParlAI Single Adversarial dataset.

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).[1] An important emerging strand of this research explores the role of human feedback and reinforcement learning in mitigating safety con-

[*] Equal contribution.

[†] Work done during an internship at Google Research.

[1] https://openai.com/blog/chatgpt/

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 400–414
December 6-10, 2023 ©2023 Association for Computational Linguistics

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 obtain performance comparable to fine-tuning all parameters (Li and Liang, 2021a; Lester et al., 2021; Vu et al., 2022a). Recent work suggests that PET may also enable more data-efficient training of models (Liu et al., 2022; Agrawal et al., 2022).

This paper's key contribution is to show the surprising effectiveness of PET for text classification on small datasets related to having safe and productive online discourse (Figure 1). A novelty of our study is the scale of data we experiment with: in addition to few-shot templates that are widely studied in the LLM literature, we explore datasets too large to fit into an LLM's context window for prompting, but too small for traditional fine-tuning.[2] Our experiments use two recent LLMs, T5 XXL (Raffel et al., 2020) and PaLM 62B (Chowdhery et al., 2022), on three related domains. The first domain concerns offensiveness in dialogue, for which we study 5 datasets: three from ParlAI (Dinan et al., 2019) and two from Bot Adversarial Dialogue (Xu et al., 2021). The second domain concerns 7 different annotation schemes for attributes related to toxicity in online comments that were introduced in the Unhealthy Comment Corpus (Price et al., 2020). For the third domain, we evaluate three new annotation schemes inspired by Wikipedia's neutral point of view (NPOV) on a new dataset we created.[3] We focus specifically on such moderation tasks to demonstrate the effectiveness of PET in the context of small datasets, however the methods provided in this work are adaptable to any text classification task.

Our primary result is that prompt-tuning provides high performance (> 0.9 ROC-AUC) for text classifiers across all domains we studied, matching or exceeding the quality of neural models on the same domain that were trained on 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).

[2] We noticed that fine-tuning LLMs on small-scale datasets can lead to overfitting.

[3] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

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.[4] Each comment is labeled as either *healthy* or *unhealthy*, in addition to binary labels for the presence of six unhealthy subattributes: (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.

3.3 Neutral Responses

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 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.[5] This expert and two more experts then annotated the data along each attribute. The final label is obtained by majority vote.

Attribute	Total (pos. percentage)	Human AUC
Multiple perspectives	113 (75.33 %)	0.965 ± 0.023
Neutral	76 (50.67%)	0.960 ± 0.025
Well-explained	91 (60.67%)	0.876 ± 0.040

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.

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.[6] 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

[4] For our experiments, we use the version of UCC available at `https://github.com/conversationai/unhealthy-conversations`.

[5] The questions and answers on each topic can be opinionated and do not reflect the views of the authors.

[6] 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, …, 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).[8] 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

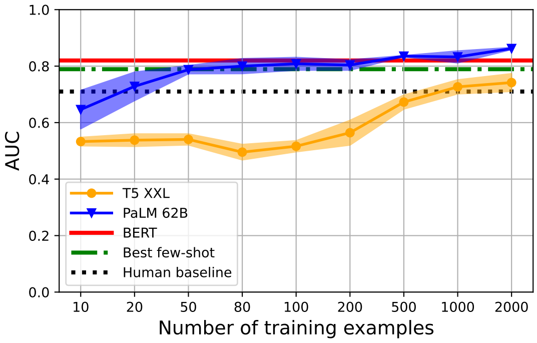


Figure 2: Prompt-tuning AUC results for the *Antagonistic* attribute of the UCC dataset, across three seeds. PaLM 62B outperforms both 12-shot prompt on PaLM 62B and a BERT model fine-tuned on 44,335 training examples. The plot for the other attributes look similar and can be found in Appendix C.

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

[8] Note that F_1 scores below 0.50, that of a random classifier, happen when the data is imbalanced, and for smaller datasets T5 XXL biases towards predicting the majority class (which leads to high accuracy, but low F_1 as recall is low on the minority class).

	Dialogue Safety					Neutral Responses		
Model	ParlAI Single Standard	ParlAI Single Adversarial	ParlAI Multi	BAD-2	BAD-4	Multiple Perspectives	Neutral	Well-Explained
PaLM 62B best few-shot	0.89	0.67	0.56	0.54	0.54	0.84	0.87	0.87
T5 XXL - 80	0.18	0.18	0.19	0.29	0.48	0.94	0.96	0.76
T5 XXL - 2,000	0.90	0.91	0.48	0.20	0.44	—	—	—
Human Agreement	—	—	—	—	—	**0.94**	0.95	**0.90**
Previous SOTA	0.88	0.67	0.66	—	—	—	—	—
PaLM 62B - 80	0.87	0.77	0.71	0.60	0.65	0.94	**0.96**	0.88
PaLM 62B - 2,000	**0.95**	**0.91**	**0.81**	**0.68**	**0.70**	—	—	—

	Unhealthy Comment Corpus						
Model	Antagonistic	Condescending	Dismissive	Generalization	Hostile	Sarcastic	Unhealthy
PaLM 62B best few-shot	0.79	0.78	0.81	0.76	0.79	0.76	0.70
T5 XXL - 80	0.50	0.55	0.56	0.49	0.57	0.54	0.51
T5 XXL - 2,000	0.74	0.74	0.75	0.80	0.80	0.74	0.66
Human Agreement	0.71	0.72	0.68	0.73	0.76	0.72	0.62
Previous SOTA	0.82	0.78	0.82	0.74	0.84	0.64	0.69
PaLM 62B - 80	0.80	0.80	0.74	0.81	0.84	0.81	0.63
PaLM 62B - 2,000	**0.86**	**0.84**	**0.87**	**0.90**	**0.89**	**0.85**	**0.77**

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).

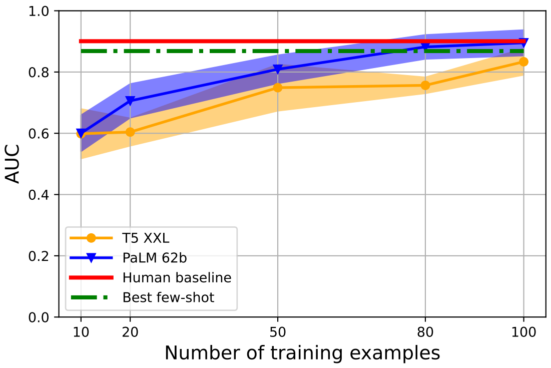


Figure 3: Prompt-tuning AUC results for the *Well-Explained* attribute of the Neutral Responses dataset, across six seeds. PaLM 62B outperforms both 12-shot prompts and the human agreement baseline. Prompt-tuning exhibits similar behavior on the other attributes—results can be found in Appendix D.

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 suggests 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

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2022. Qameleon: Multilingual qa with only 5 examples. *arXiv preprint arXiv:2211.08264*.

Yuntao et al Bai. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Mudit Chaudhary, Chandni Saxena, and Helen Meng. 2021. Countering online hate speech: An nlp perspective. *arXiv preprint arXiv:2109.02941*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

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.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Amelia et al Glaese. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. 2020a. Classifying constructive comments. *arXiv preprint arXiv:2004.05476*.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2020b. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021a. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021b. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124, Online. Association for Computational Linguistics.

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

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Bernhard Rieder and Yarden Skop. 2021a. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of perspective api. *Big Data & Society*, 8(2):20539517211046181.

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

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Guy Simon. 2020. Openweb tests the impact of "nudges" in online discussions. *OpenWeb Blog*.

Dustin Tran, Jeremiah Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, et al. 2022. Plex: Towards reliability using pretrained large model extensions. *arXiv preprint arXiv:2207.07411*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. 2022a. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, Dublin, Ireland. Association for Computational Linguistics.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2022b. Spot: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A Smith. 2021. Challenges in automated debiasing for toxic language detection. In *EACL*.

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.

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

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

A.2 Examples

Examples from the neutral responses dataset can be seen in 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.

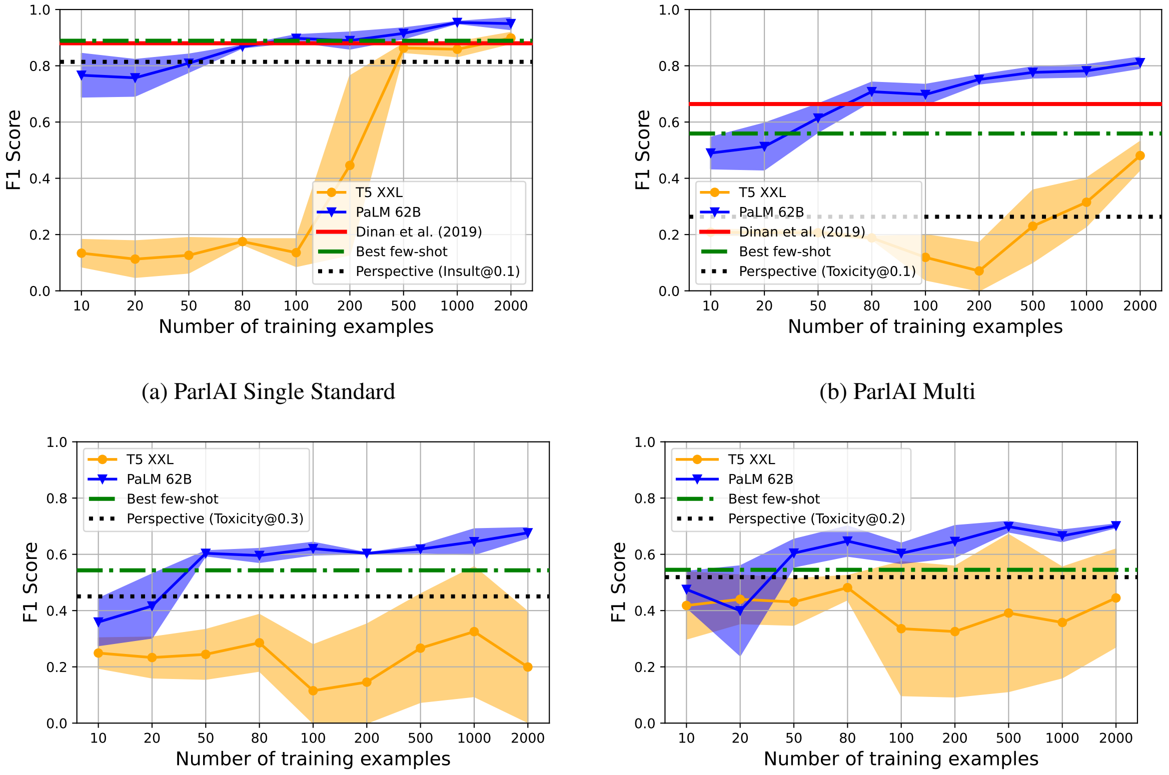
In contrast, T5 XXL requires a larger amount of training data for prompt-tuning (2,000 examples) on ParlAI Single Standard to outperform the BERT baseline. On the remaining three datasets, prompt-tuning T5 XXL does not achieve a performance better than the few-shot baseline.

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.

Question	Answer	Multiple perspectives	Neutral	Well-explained
Everyone believes we should legalize marijuana, right?	Some people believe legalizing marijuana would reduce crime and create jobs, while others oppose legalization because they believe canabis is a gateway drug and legalizing it will increase hard drug use.	Yes	Yes	Yes
What is this about taxes and churches?	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.	Yes	Yes	No
Can you tell me about the water ban?	I'm so tired of hearing about the water ban. Nobody cares.	No	No	No

Table 4: Three examples from the neutral responses dataset.



(a) ParlAI Single Standard

(b) ParlAI Multi

(c) BAD-2

(d) BAD-4

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

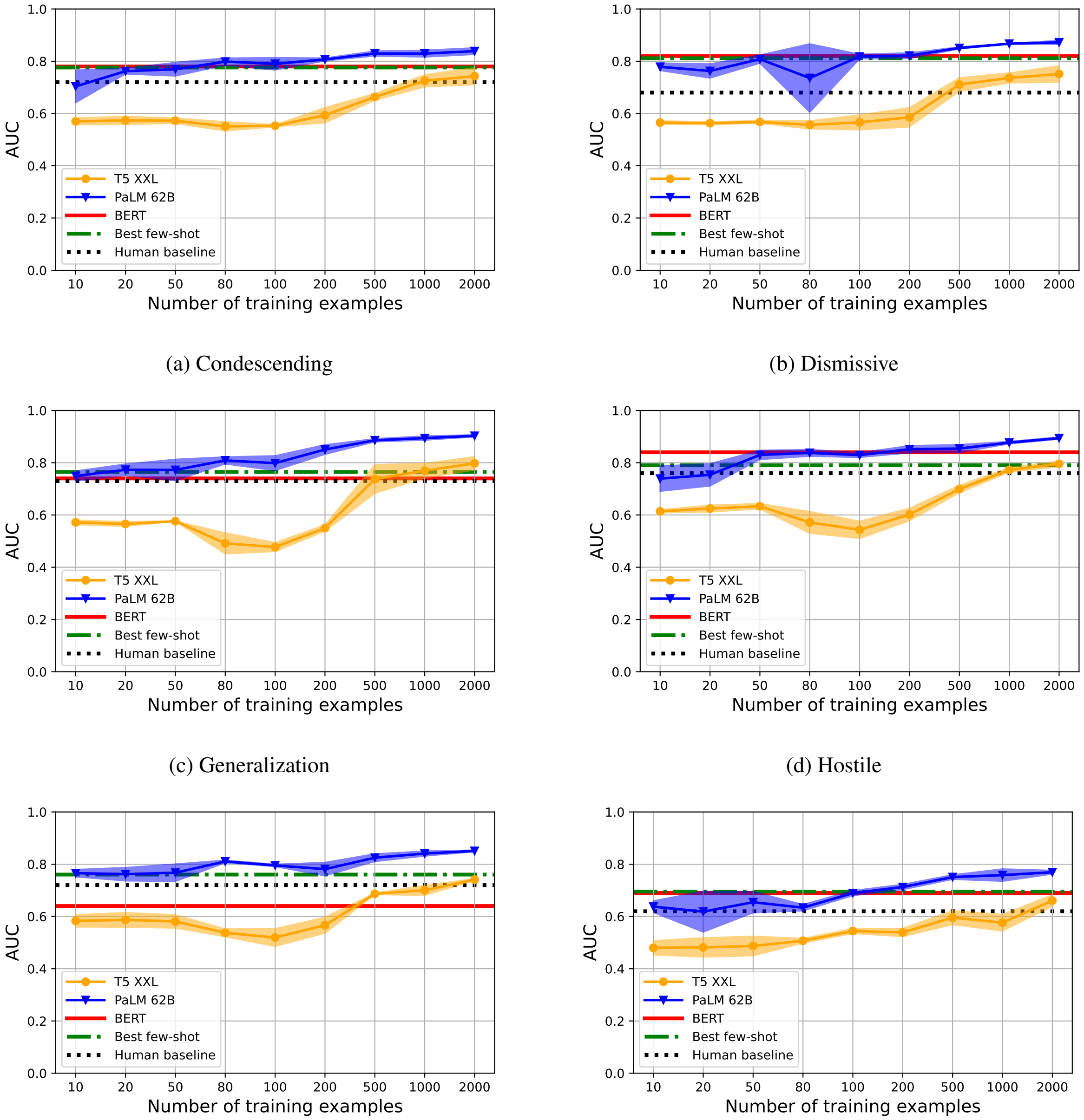
The prompt-tuning results on the attributes *multiple perspectives* and *neutral* can be found in Figure 6. As can be seen, for PaLM 62B, 20 training examples are enough to outperform the Human baseline for the *multiple perspectives* attribute, whereas for the *neutral* one around 80 examples are needed. 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.



(a) Condescending

(b) Dismissive

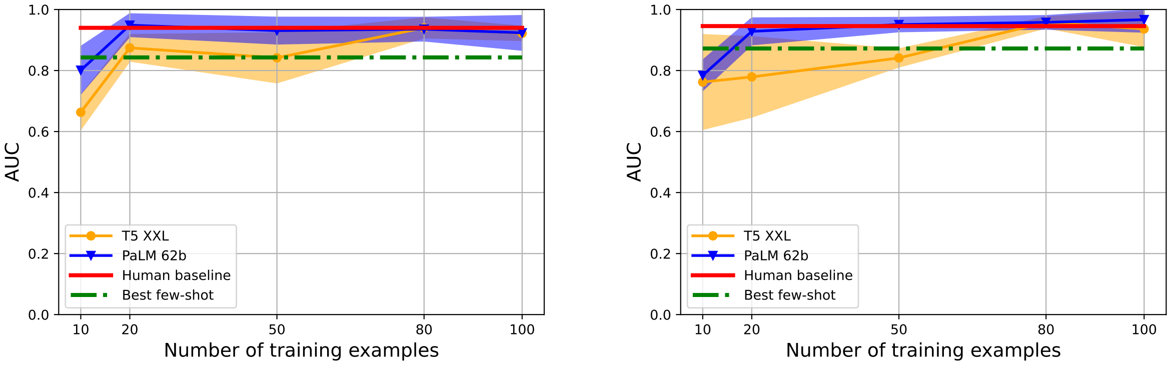
(c) Generalization

(d) Hostile

(e) Sarcastic

(f) Unhealthy

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.



(a) Multiple perspectives

(b) Neutral

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.