Learning to Initialize: Can Meta Learning Improve Cross-task Generalization in Prompt Tuning?

Anonymous ACL submission

Abstract

Prompt tuning (PT) which only tunes the embeddings of an additional sequence of tokens per task, keeping the pre-trained language model (PLM) frozen, has shown remarkable performance in few-shot learning. Despite this, PT has been shown to rely heavily on good initialization of the prompt embeddings. In this work, we study *meta prompt tuning* (MPT) to systematically explore how meta-learning can help improve (if it can) cross-task generalization in PT through learning to initialize the prompt embeddings from other relevant tasks. We empirically analyze a representative set of meta learning algorithms in a wide range of adaptation settings with different source/target task configurations on a large set of few-shot tasks. With extensive experiments and analysis, we demonstrate the effectiveness of MPT. We find the improvement to be significant particularly on classification tasks. For other kinds of tasks such as question answering, we observe that while MPT can outperform PT in most cases, it does not always outperform multi-task learning. We further provide an in-depth analysis from the perspective of task similarity.

1 Introduction

Humans can easily learn to perform new tasks with only few data by leveraging previously acquired knowledge from other relevant tasks. Such capability is a hallmark of human intelligence (Carey and Bartlett, 1978). However, when it comes to the models, they often face over-fitting issues when they are tasked to learn from a few labeled examples (Lake et al., 2017; Linzen, 2020), a problem commonly termed as *few-shot learning* (FSL).

With the recent advancements in developing large-scale pre-trained language models (PLMs), prompt-based methods have shown promising results in FSL. Brown et al. (2020) show that by virtue of in-context (meta) learning, a frozen GPT-3 model can achieve good results on a variety of few-shot tasks through manually designed *prompts*, which are task instructions along with a few examples expressed in natural language. However, the performance of in-context learning has been shown to be highly sensitive to the design of such "discrete" prompts (Zhao et al., 2021). It is also limited by the maximum sequence length supported by the PLMs (Li and Liang, 2021). Down this line, efforts have been made on automatically searching and optimizing for discrete prompts (Shin et al., 2020; Schick and Schütze, 2021; Gao et al., 2021).

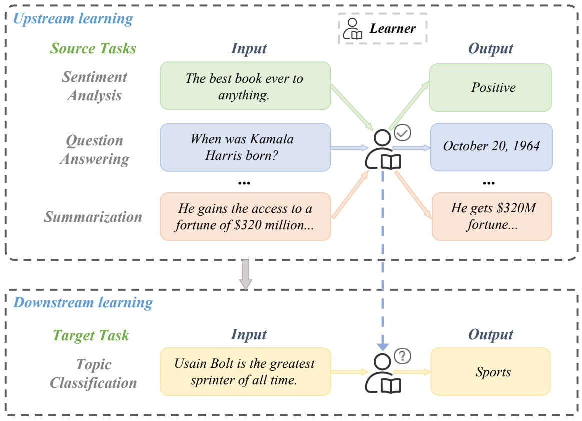

Figure 1: Illustration of cross-task generalization, where the model is expected to learn an unseen *target* task given the knowledge acquired from previously learned *source* tasks.

As an alternative to discrete prompts, recent efforts attempt to learn "soft" prompts that add additional trainable parameters (Liu et al., 2021b; Li and Liang, 2021; Lester et al., 2021), showing better results than discrete prompts (Liu et al., 2021a). Lester et al. (2021) introduce *prompt tuning* (PT) that prepends a sequence of *tunable* tokens to the input and optimize their embeddings keeping the PLM frozen. Despite its strong few-shot performance, PT has been shown to be sensitive to the initialization of the embeddings, which might limit its practical application (Qin and Joty, 2022b). To address this, Gu et al. (2022) propose *pre-trained* *prompt tuning* (PPT) to pre-train soft prompts using self-supervised tasks on unlabeled data. It relies on carefully designed pre-training tasks tailored to the downstream tasks, and the pre-training objectives are only applicable to classification tasks. Vu et al. (2022) introduce *soft prompt transfer* (SPoT), which uses the soft prompts learned from a set of source tasks through multi-task learning to initialize the prompt for a target task. Both PPT and SPoT demonstrate *cross-task generalization* (Fig. 1) – learning of a new task can benefit from learning of other related tasks (Ye et al., 2021).

In a recent survey, Lee et al. (2022) claim that *meta learning* (Schmidhuber, 1987) can play an important role for cross-task generalization in NLP.[1] Different from multi-task learning which considers the performance on the source tasks to learn the initial parameters, meta learning aims to find initial parameters suitable for adapting to a target few-shot task. Hence, it could outperform multi-task learning in several scenarios with *full-model* finetuning (Dou et al., 2019; Chen et al., 2020b). However, to our knowledge, there is no systematic study on the role of meta learning on PT. In a recent work, Huang et al. (2022) adopt MAML (Finn et al., 2017) for pre-training soft prompts. One major limitation of their study is that it is limited to only one type of meta learning algorithm and only sentiment classification tasks, lacking comprehensive understanding of cross-task generalization. Min et al. (2022) and Chen et al. (2022) show the effectiveness of in-context learning for PLMs, whereas we mainly focus on optimization-based meta learning.

To systematically study meta prompt tuning (MPT) for cross-task generalization, we conduct experiments on a large collection of few-shot tasks involving different types of datasets with a unified text-to-text format (Ye et al., 2021). We investigate a wide range of adaptation settings with different source/target task types, which helps better understand the capability and limitation of meta learning in PT. With extensive experiments, we aim to address the following research questions:

- **Q1.** Can MPT improve cross-task generalization in PT? Is it better than multi-task learning?
- **Q2.** What happens with more labelled data for source/target tasks (beyond few-shot settings)?
- **Q3.** Does it help with more diverse source tasks?
- **Q4.** Is the performance gain of MPT consistent across different backbone models?

To answer these questions, we empirically analyze MAML (Finn et al., 2017), FoMAML and Reptile (Nichol et al., 2018), which constitute a representative set of meta learning methods. Experimental results show that MPT can indeed help cross-task generalization, *e.g.,* MAML improves the performance of PT by more than 20% on classification tasks. However, we also notice that MPT does not always outperform multi-task learning, especially on non-classification tasks. We provide an in-depth analysis from the perspective of task similarity. As for Q2, we find that MPT does benefit cross-task generalization beyond few-shot settings. For Q3, we observe that increasing the diversity of source tasks does not necessarily improve cross-task generalization. Finally, the consistent gain of MPT across different models shows its robustness to model type and size. In summary, the two main contributions of this work are:

- To the best of our knowledge, we are the first to extensively explore how meta learning helps cross-task generalization in prompt tuning.
- With extensive experiments and analysis, we show the effectiveness and limitation of meta prompt tuning in various source/target settings. Our code base is available at `<redacted>`.

2 Related Work

Few-shot Learning (FSL) FSL aims to learn a task with only a few labeled examples, which often leads to the over-fitting problem. Existing methods to address this problem mainly focus on optimizing the hypothesis space of the few-shot tasks (Triantafillou et al., 2017; Finn et al., 2017; Hu et al., 2018) or augmenting the few-shot data (Gao et al., 2020; Qin and Joty, 2022a). Recently, large-scale pre-trained language models (PLMs) have demonstrated strong FSL ability through prompt-based methods, including both discrete (Brown et al., 2020) and soft prompts (Lester et al., 2021).

Prompt-based Learning (PL) PL is a new paradigm which prepends a task-specific template or prompt to the input for learning new tasks (Liu et al., 2021a). Initial PL methods mainly focus on designing, searching or optimizing discrete prompts (Brown et al., 2020; Shin et al., 2020; Gao et al., 2021). However, discrete prompts are hard to optimize. To solve this, recent PL methods

[1]Unless otherwise specified, by meta learning in this paepr we generally refer to the optimization-based meta learning algorithms, and use more specific names for the other kinds such as *in-context learning* for black-box meta learning and *metric learning* for non-parametric meta learning.

attempt to optimize prompts in a continuous space, *i.e.,* learn soft prompts (Li and Liang, 2021; Liu et al., 2021b; Lester et al., 2021), showing impressive FSL performance (Qin and Joty, 2022b). In addition to prompt design, several recent studies have explored the applications (Zhu et al., 2022; Li et al., 2022) and analysis (Zhong et al., 2021; Le Scao and Rush, 2021) of PL.

Meta Learning Meta Learning or learning to learn, has been applied to boost few-shot performance on various NLP tasks, *e.g.,* relation extraction (Han et al., 2018) and machine translation (Gu et al., 2018). Meta learning algorithms can be divided into three main categories. First, *black-box* methods adopt additional meta learners to help adaptation (Santoro et al., 2016; Garnelo et al., 2018; Mishra et al., 2018; Brown et al., 2020). Second, *non-parametric* methods explore how to learn metrics that can compare the distances between different samples, *i.e.,* learning to compare (Koch et al., 2015; Vinyals et al., 2016; Snell et al., 2017). Finally, *optimization-based* methods aim to learn better parameter initialization to effectively and efficiently adapt to unseen tasks, *i.e.,* learning to initialize (Finn et al., 2017; Nichol et al., 2018; Kedia et al., 2021). Lee et al. (2022) claim that meta learning can be effective for cross-task generalization, especially the optimization-based methods. They can be applied to various problems in a model-agnostic way to improve FSL on target tasks with model fine-tuning (Ye et al., 2021).

Summary. Existing work shows that meta learning can improve cross-task few-shot generalization with full model fine-tuning. However, there is no systematic study on whether (and how) meta learning can do so with prompt tuning of PLMs. To fill this research gap, our work provides a comprehensive understanding of the effectiveness and limitation of meta learning in prompt tuning.

3 Preliminaries

In this section, we revisit the basics about prompt tuning and optimization-based meta learning.

3.1 Prompt Tuning

Following Lester et al. (2021), we reframe all tasks into a text-to-text format. Given a training dataset $\mathcal{D}^{tr} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ for a task $\mathcal{T}$, different from traditional model fine-tuning, prompt tuning (PT) is a parameter-efficient learning method which freezes the PLM θ and prepends the input text X_i with a sequence of *tunable* soft tokens P, parameterized by prompt embeddings ϕ. The prompt embeddings ϕ are initialized from the vocabulary of the PLM and optimized through gradient descent with the following objective:

$$\mathcal{L}_\phi^{\mathcal{T}} = \mathcal{L}(\phi, \mathcal{D}^{tr}) = -\sum_{i=1}^{n} \log p(Y_i | [P, X_i], \phi, \theta) \quad (1)$$

3.2 Optimization-based Meta Learning

The main goal of optimization-based meta learning (or learning to initialize), is to learn better initial parameters that can effectively and efficiently adapt to a new task $\mathcal{T}^{\text{new}}$ with limited data. We denote the initial parameters (meta-parameters) as ϕ^*.

To obtain ϕ^*, the model needs to learn from a series of *meta-training* tasks $\mathcal{T}^{\text{meta}} = \{\mathcal{T}_1, ..., \mathcal{T}_n\}$. The dataset $\mathcal{D}_i$ of each task $\mathcal{T}_i$ is divided into two disjoint sets: a *support set* $\mathcal{S}_i$ and a *query set* $\mathcal{Q}_i$. The objective for learning ϕ^* is

$$\phi^* = \arg\min_{\phi} \sum_{\mathcal{T}_i \in \mathcal{T}^{\text{meta}}} \mathcal{L}\Big(\underbrace{\phi - \alpha \nabla_\phi \mathcal{L}(\phi, \mathcal{S}_i)}_{\text{inner update}}, \mathcal{Q}_i\Big) \quad (2)$$

where $\mathcal{L}$ is the objective function defined in Eq. (1), ϕ is the set of parameters to meta-learn and α is the inner learning rate. Denoting the overall loss as $\mathcal{L}_\phi^{\mathcal{T}^{\text{meta}}} = \sum_{\mathcal{T}_i \in \mathcal{T}^{\text{meta}}} \mathcal{L}(\phi', \mathcal{Q}_i)$ with ϕ' being the inner-updated value of ϕ, we use gradient descent to update ϕ further in the meta-training stage:

$$\phi = \phi - \beta \nabla_\phi \mathcal{L}_\phi^{\mathcal{T}^{\text{meta}}} \quad (3)$$

where β is the outer learning rate. This is actually the Model-Agnostic Meta-Learning or MAML (Finn et al., 2017). Notice that optimizing Eq. (3) requires calculating second-order gradients, which can be quite memory-consuming. To alleviate this, First-order MAML (FoMAML) and Reptile (Nichol et al., 2018) are proposed to use first-order approximations, allowing lower memory costs.

After the meta-training stage, ϕ^* serves as the initial parameters for learning an unseen *meta-testing* task $\mathcal{T}^{\text{new}}$ which is usually few-shot.

4 Approach

In this section, we first introduce the problem setting and evaluation metric. Then, we illustrate the key methods for meta prompt tuning (MPT).

4.1 Problem Setting

To evaluate cross-task generalization in prompt tuning, we select a large and diverse collection of few-shot tasks from Ye et al. (2021), covering various types including classification, question answering and generation. We partition the set of all tasks $\mathcal{T}^{\text{all}}$ into two disjoint parts: source tasks $\mathcal{T}^{\text{src}}$ and target tasks $\mathcal{T}^{\text{tgt}}$. Details of the tasks and partitions are provided later in our experiment setup (§5).

Following Min et al. (2022), we can divide the whole learning process into two stages (Fig. 1):

- **Upstream learning on source tasks** In this stage, the model has access to $\mathcal{T}^{\text{src}}$, which is regarded as *meta-training* tasks $\mathcal{T}^{\text{meta}}$ in Eq. (2). We divide the dataset $\mathcal{D}_i$ of every source task $\mathcal{T}_i$ into training (or support) and validation (or query) sets, and conduct optimization-based meta learning or multi-task learning on these sets to obtain meta-parameters ϕ^*. Note that we use both support and query sets for model training in multi-task learning to ensure fair data access for both methods.

- **Downstream learning on target tasks** After the upstream learning stage, we use the learned meta-parameters ϕ^* as the initial point for learning target tasks $\mathcal{T}^{\text{tgt}}$. Every target task $\mathcal{T}_k$ has its own training set $\mathcal{D}_k^{\text{tr}}$, validation set $\mathcal{D}_k^{\text{val}}$, and test set $\mathcal{D}_k^{\text{test}}$. The model is required to learn from $\mathcal{D}_k^{\text{tr}}$ via prompt tuning and will be evaluated on $\mathcal{D}_k^{\text{test}}$. The performance on $\mathcal{D}_k^{\text{val}}$ is used for hyper-parameters tuning and model selection.

This two-stage learning paradigm can naturally reflect cross-task generalization where the model needs to learn an unseen task given previously acquired knowledge from other tasks.

4.2 Evaluation Metric

We evaluate the model performance on a set of target tasks $\mathcal{T}^{\text{tgt}}$. As $\mathcal{T}^{\text{tgt}}$ may cover various task types, simply averaging the performance of different target tasks is unreasonable. Following Ye et al. (2021), we use *average relative gain* (ARG) as the main evaluation metric. We first calculate *relative gain* (RG) for each target task, *i.e.,* relative performance improvement before and after applying the upstream (meta or multi-task) learning on the source tasks. Then we average the relative gains of all target tasks to obtain the final result which indicates the overall performance improvement.

4.3 Meta Prompt Tuning (MPT)

As shown in Fig. 2, the key idea of MPT is to apply optimization-based meta-training as upstream learning to a set of source tasks in order to learn meta parameters, which in this case are prompt embeddings. The learned prompt embeddings serve as the initialization for learning unseen target tasks, referred to as meta-testing or downstream learning.

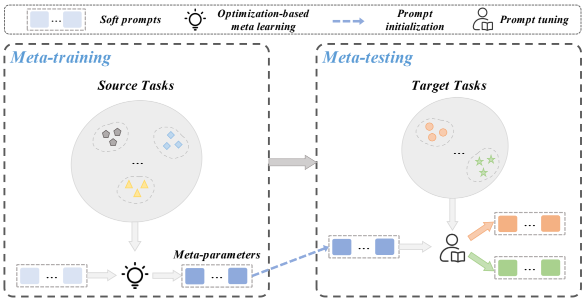

Figure 2: Overview of Meta Prompt Tuning (MPT). In the meta-training stage, we conduct optimization-based meta learning on source tasks to obtain meta-parameters (*i.e.,* soft prompts). The meta-parameters will then be used to initialize prompt embeddings for learning unseen target tasks in the meta-testing stage.

4.3.1 Meta-training

We meta-train the prompt embeddings on source tasks $\mathcal{T}^{\text{src}}$. Without loss of generality, we take MAML (Finn et al., 2017) as an example. For every iteration, we first sample one source task $\mathcal{T}_i$ which has a support set $\mathcal{S}_i$ and a query set $\mathcal{Q}_i$. Then we sample a support batch $\mathcal{B}_s$ from $\mathcal{S}_i$ and a query batch $\mathcal{B}_q$ from $\mathcal{Q}_i$. Denoting the trainable prompt embeddings as ϕ, $\mathcal{B}_s$ and $\mathcal{B}_q$ are used for one gradient update with the following objective:

$$\begin{aligned} \mathcal{L}_\phi^i &= \mathcal{L}(\phi - \alpha \nabla_\phi \mathcal{L}(\phi, \mathcal{B}_s), \mathcal{B}_q) \\ \phi &= \phi - \beta \nabla_\phi \mathcal{L}_\phi^i \end{aligned} \tag{4}$$

where $\mathcal{L}$ is the task loss defined in Eq. (1), and α and β are inner and outer learning rates, respectively. During the meta-training stage, we iterate over tasks in $\mathcal{T}^{\text{src}}$ to update prompt embeddings ϕ for a fixed number of steps. The learned meta-parameters ϕ^* is used in the meta-testing stage.

4.3.2 Meta-testing

In meta-testing, the model is expected to learn unseen target tasks $\mathcal{T}^{\text{tgt}}$. For each target task $\mathcal{T}_k$, we use the learned meta-parameters ϕ^* to initialize the prompt embeddings for the task. Denoting the training set of $\mathcal{T}_k$ as $\mathcal{D}_k^{\text{tr}}$, the learning objective during meta testing is defined as:

$$\mathcal{L}_{\phi^*}(\mathcal{D}_k^{\text{tr}}) = -\sum_{i=1}^{n} \log p(Y_i | [P^*, X_i], \phi^*, \theta) \tag{5}$$

where θ is the frozen PLM, $(X_i, Y_i) \sim \mathcal{D}_k^{\text{tr}}$ is a training sample and P^* are the prompt tokens.

Source		Target	
Setting	#tasks	Setting	#tasks
Random	114	Random	20
Classification (Cls) Both (Cls + Non-Cls) Non-Classification	45 23 + 22 45	Classification	10
Classification Both (Cls + Non-Cls) Non-Classification	45 23 + 22 45	Non-Classification	12
QA Non-QA	22 33	QA	15
Non-Paraphrase Cls	60	Paraphrase	4

Table 1: Statistics of ten distinct source/target task partitions. Appendix A.1 for details about each partition.

We evaluate the model with the best validation performance on the test set and calculate average relative gain on the test sets of $\mathcal{T}^{\text{tgt}}$.

5 Experimental Setup

We first describe the source/target task partitions, and then introduce methods compared in our work. Finally, we present the implementation details.

5.1 Task Partitions

We experiment with ten different source/target task partitions as shown in Table 1. Depending on the type of the target tasks, we can divide these ten settings into several groups:

- **R→R (Random→Random)**: We first experiment with the R→R setting where both source and target tasks are randomly selected, meaning that they can cover any task type. This setting mimics the learning paradigm of humans and reflects whether cross-task generalization can help obtain a general-purpose few-shot learner.
- **X→Cls (X=Cls, Both, Non-Cls)**: The target tasks involve classification, while the source tasks can be classification, non-classification tasks or both. This setting helps us better understand the influence of the source task distribution.
- **X→Non-Cls (X=Cls, Both, Non-Cls)**: The only difference between this and the previous setting is the type of target tasks. We investigate how meta learning improves cross-task generalization when target tasks are non-classification tasks.
- **X→QA (X=QA, Non-QA)**: Compared to the previous one, this group is more fine-grained. We only select target tasks from question answering (QA) instead of all non-classification tasks. We conduct experiment on different source task types, including QA and Non-QA tasks.
- **NP→P (Non-Paraphrase Cls→Paraphrase)**: This group has the finest granularity in our setting. We choose paraphrase identification which is a sub-category of classification as the target, and non-paraphrase classification as the source. The final two groups help understand how meta learning performs in more fine-grained scenarios.

Note that we ensure that there is no overlap between the source and target tasks. Following Ye et al. (2021), we use 16 samples per class in the training (or support) and validation (or query) sets for classification tasks, and 32 samples per set for non-classification tasks. For every task, we sample the training and validation sets 5 times with different random seeds to reduce variance in few-shot evaluation and cover more diverse samples in upstream learning. We provide full details of tasks and partitions in Appendix A.1.

5.2 Methods Compared

We mainly use T5-Large (Raffel et al., 2019) as the backbone language model and compare the following methods in our work.

- **Prompt Tuning (PT) on target tasks.** It is our baseline without the upstream learning. We directly apply PT (Lester et al., 2021) to target tasks and use its performance as the basis for computing average relative gain for other methods.
- **Model-Agnostic Meta-Learning (MAML).** We apply MAML (Finn et al., 2017) in the upstream learning (meta-training) stage. The learned meta-parameters are used to initialize prompt embeddings for learning target tasks.
- **First-order MAML (FoMAML) and Reptile.** We also investigate two first-order meta learning algorithms: FoMAML (Finn et al., 2017) and Reptile (Nichol et al., 2018). Compared to MAML, they are more memory-efficient.
- **Multi-task learning (MTL).** We conduct multi-task learning on source tasks instead of meta learning to obtain initial parameters. This is a straight-forward yet effective method as demonstrated by Vu et al. (2022).
- **Fine-tuning on target tasks.** Fine-tuning is the dominant paradigm where the whole language model is tuned for learning target tasks. We include it to verify whether cross-task generalization can help PT outperform fine-tuning.

In addition, we conduct experiments with different backbone models to verify MPT's robustness.

5.3 Implementation Details

All our methods are implemented with PyTorch/Transformers library (Wolf et al., 2020). We use higher library (Grefenstette et al., 2019) for higher-order optimization in meta learning methods. The prompt length in PT is set to 100 tokens following Lester et al. (2021). We provide details of other hyperparameters in Appendix A.4.

Since it is infeasible to search for optimal hyperparameters for each of the meta- and multi-task learning methods in each of the settings, we select them based on the R→R setting. We randomly select 5 tasks that are not in the source and target sets as validation tasks for hyperparameter search. The hyperparameters with best validation performance (ARG) are used for upstream learning. We select the inner learning rate, the outer learning rate and total training steps for MAML and adopt the same three hyperparameters for FoMAML and Reptile.

6 Results and Analysis

We now address the four research questions asked before in §1 with empirical results.

> **Q1.** Can meta prompt tuning improve cross-task generalization? Is it better than multi-task learning?

The ARG of different methods *w.r.t.* PT in various settings are shown in Table 2; more detailed results on every target task are in Appendix A.2.

• **MPT can indeed help cross-task generalization.** From the results in Table 2, we observe that MPT outperforms the baseline PT in most cases with +ve ARG scores. Out of 30 different runs for three meta learning methods in ten different settings (see the 1st block of results), MPT achieves better performance than PT in 23 runs, demonstrating its effectiveness in cross-task generalization.

For the R→R setting, MAML achieves the best performance, showing that it is a good general-purpose few-shot learner. For adapting to classification tasks, MAML outperforms PT by **20.16**% if the prompt embeddings are initialized from other classification tasks. The results in a more fine-grained setting (NP→P) also indicate the ability of MAML to learn classification tasks. While Reptile performs the best (20.44%) in this setting, MAML still outperforms PT by a large margin (**11.14**%).

However, as shown in Table 2, MAML falls behind FoMAML when adapting to non-classification tasks. Among the three meta learning methods, FoMAML achieves the best performance (**9.81**%) on non-classification target tasks in the Both→Non-Cls setting, showing effective knowledge transfer. We observe similar results in more fine-grained settings QA/Non-QA→QA, where FoMAML outperforms MAML and Reptile significantly. While Reptile is claimed empirically to be better than MAML/FoMAML (Lee et al., 2022), it falls short of MAML/FoMAML in many cases. This might be because MAML and FoMAML are more similar compared to Reptile from a gradient perspective (Nichol et al., 2018). And since the hyperparameter search is done based on MAML (§5.3), which means Reptile's method may be suboptimal.

In addition, we can see that meta learning helps PT outperform fine-tuning in several settings including Cls→Cls (MAML, FoMAML), Both→Cls (FoMAML) and NP→P (MAML, Reptile), which demonstrates the superiority of MPT.

• **MPT does not always outperform multi-task learning (MTL).** While meta learning is specifically designed for quickly adapting to unseen target tasks, it does not always outperform MTL in PT. From Table 2, we can observe that MTL achieves better performance than MPT in many cases, especially on non-classification target tasks. We analyze the reasons as follows:

- Meta learning methods have been shown to be highly sensitive to the hyperparameters (Antoniou et al., 2019), which we could not tune exhaustively due to memory/time constraints (see Appendix A.5 for hyperparameter sensitivity analysis). As mentioned in §5.3, we select the hyperparameters of MAML using the R→R setting, and then use the same hyperparameters for all meta learning methods in all settings, which might limit the performance of MPT.
- There might be less shared structure (or features) among non-classification tasks compared to classification. The classification tasks mostly involve sentence-level classification and in some cases the task labels correlate well (*e.g.,* AG News and DBpedia). Thus, they share some common semantics in both source and target tasks. The model can learn similar patterns (inferring the label of the entire input sentence) during both meta-training and meta-testing stages, enabling better knowledge transfer. The non-classification set on the other hand can include different types of tasks such as QA and summarization; modeling

Method	R→R	Cls →Cls	Both →Cls	Non-Cls →Cls	Cls →Non-Cls	Both →Non-Cls	Non-Cls →Non-Cls	QA →QA	Non-QA →QA	NP →P
MAML	$\mathbf{8.78}_{\pm 0.69}$	$\mathbf{20.16}_{\pm 0.84}$	$10.57_{\pm 1.03}$	$6.34_{\pm 0.48}$	$0.32_{\pm 0.04}$	$7.54_{\pm 0.73}$	$6.71_{\pm 0.39}$	$-16.59_{\pm 1.36}$	$3.26_{\pm 0.24}$	$11.14_{\pm 0.93}$
FoMAML	$1.24_{\pm 0.18}$	$18.80_{\pm 1.13}$	$\mathbf{17.84}_{\pm 1.21}$	$7.32_{\pm 0.42}$	$6.42_{\pm 0.51}$	$9.81_{\pm 0.64}$	$3.88_{\pm 0.31}$	$16.63_{\pm 1.58}$	$9.83_{\pm 0.76}$	$-0.68_{\pm 0.07}$
Reptile	$8.42_{\pm 0.46}$	$-5.17_{\pm 0.71}$	$-4.18_{\pm 0.37}$	$2.42_{\pm 0.21}$	$-1.54_{\pm 0.18}$	$-3.38_{\pm 0.49}$	$0.78_{\pm 0.07}$	$0.77_{\pm 0.09}$	$-0.09_{\pm 0.01}$	$\mathbf{20.44}_{\pm 1.34}$
Multi-task learning	$7.14_{\pm 0.62}$	$-5.64_{\pm 0.92}$	$5.73_{\pm 0.43}$	$4.97_{\pm 0.39}$	$\mathbf{8.51}_{\pm 1.16}$	$\mathbf{13.47}_{\pm 0.97}$	$\mathbf{19.67}_{\pm 1.72}$	$\mathbf{25.65}_{\pm 1.93}$	$\mathbf{17.23}_{\pm 1.08}$	$-5.19_{\pm 0.86}$
Fine-tuning	$-12.61_{\pm 1.57}$	$16.02_{\pm 1.44}$	$16.02_{\pm 1.44}$	$\mathbf{16.02}_{\pm 1.44}$	$-35.70_{\pm 2.73}$	$-35.70_{\pm 2.73}$	$-35.70_{\pm 2.73}$	$-47.37_{\pm 2.97}$	$-47.37_{\pm 2.97}$	$1.56_{\pm 0.12}$

Table 2: **Average relative gain (ARG %) of different methods with respect to prompt tuning (PT) in various settings. Bold** indicates the best ARG score. 'Cls', 'QA', 'P' and 'NP' respectively stand for 'classification', 'question answering', 'paraphrase' and 'non-paraphrase classification'.

them typically requires a Seq2Seq formulation. These tasks typically lack shared task semantics. For example, the structure of QA is context + question + answer, requiring reasoning ability. In contrast, the structure of summarization is long document + short summary, requiring summarizing ability. Although it has been shown that QA can help summarization in content selection (Arumae and Liu, 2019), it is more difficult for MPT to capture transferable knowledge as success of meta learning eventually depends on how much the tasks share (Finn, 2022).

To provide an in-depth analysis of the difference between classification and non-classification tasks, we consider from the perspective of task similarity. Following (Lin et al., 2022), the correlation between input subspaces (the norm of projected subspace onto the other subspace) for two tasks could serve as the similarity score between them. We randomly pick 5 (cls,cls) task pairs as similar tasks. For dissimilar tasks, we randomly pick 5 (QA, summarization) task pairs. The average similarity score for similar task pairs is 0.768 while the average similarity score for dissimilar task pairs is only 0.306 (see Appendix A.6 for detailed results), which verifies that classification tasks share more structure than non-classification tasks.

Given the performance gap between MPT and MTL in some settings, we believe that exploring more advanced MPT methods could be a promising research direction.

Q2. What happens with more labelled data for source/target tasks (beyond few-shot settings)?

As mentioned in §5.1, we mainly explore how MPT improves cross-task generalization when both the source and target tasks are few-shot, which corresponds to the way humans learn (Lake et al., 2017). We used 16 samples per class for classification tasks, and 32 samples per dataset for non-classification tasks. To validate whether more labelled data for source/target tasks can influence the performance of MPT, we conduct controlled experiments with $\{32, 64, 128\}$ samples per class for source/target tasks in the Cls→Cls setting.

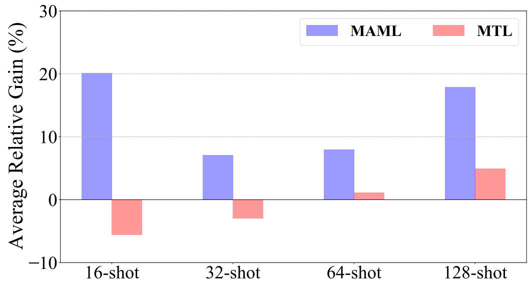

Figure 3: ARG (%) of MPT (MAML) and multi-task learning *w.r.t.* prompt tuning (ARG = 0) for **varying data size of source tasks** in the Cls→Cls setting.

• **Source** We report the results of MAML and MTL with more labelled data for the source tasks in Fig. 3. We can observe that: (*i*) MPT outperforms PT (ARG = 0) and MTL in all cases, showing its robustness to data sizes. (*ii*) Increasing the number of samples in source tasks *does not* necessarily lead to better cross-task generalization for MPT. The best ARG is achieved for 16-shot, which justifies using few-shot source tasks. (*iii*) The performance of MTL improves with more data for source tasks, showing a different learning pattern from MPT.

• **Target** Table 3 shows the results for increasing the number of examples in target tasks. We can see that: (*i*) The performance gain of MPT is evident even at 128-shot (8.36%), demonstrating that it *does* help cross-task generalization beyond few-shot. (*ii*) MPT outperforms MTL by a large margin in all settings. (*iii*) MTL is unstable in terms of ARG scores; while it outperforms PT in 64-shot (1.96%), it falls behind PT in all other settings, indicating that MPT is a better choice when adapting to classification tasks.

Q3. Does MPT help with more diverse source tasks?

Method	Shot			
	16	32	64	128
MPT (MAML)	**20.16**	**9.10**	**5.64**	**8.36**
Multi-task learning	−5.64	−14.17	1.96	−0.20

Table 3: ARG (%) of different methods when more labelled data is used in target tasks.

Method	Source task number		
	12	24	45
MPT (MAML)	8.44	12.89	**20.16**

Table 4: ARG (%) of MPT (MAML) when using different number of source tasks in the Cls→Cls setting.

MPT aims to learn to initialize the prompt embeddings from source tasks, which may cover different types. We hypothesize that the diversity of source tasks might influence its performance. To verify this, we analyze the influence of different source task selections on the same target tasks in two settings: varying the type and number of tasks.

- **Type of tasks.** The results of learning from different types of source tasks are reported in Table 2. The performance of MPT on non-classification target tasks improves when using more diverse source tasks, *e.g.,* from Non-Cls/Cls→Non-Cls to Both→Non-Cls. However, for adapting to classification task, the best ARG is achieved when all source tasks are classification, *i.e.,* the Cls→Cls setting. Hence, we can conclude that increasing the type diversity of source tasks *does not* necessarily improve cross-task generalization, which is consistent with the finding in (Ye et al., 2021).
- **Number of tasks.** To investigate the impact of the number of source tasks, we conduct controlled experiments on $\{12, 24\}$ source tasks sampled from the original 45 source tasks in the Cls→Cls setting (see Appendix A.3 for a full list). From Table 4, we can observe that the performance of MPT keeps improving as the number of source tasks increases, showing better cross-task generalization.

It is worthwhile to note that while our work provides some insights on the choice of source tasks, more systematic studies on how to select the most suitable source tasks given a set of target tasks are needed. We hope that future analysis can provide a more comprehensive understanding of the relationship between source and target tasks.

Q4. Is the performance gain of MPT consistent across different backbone language models?

Method	MAML	FoMAML	Reptile	MTL	Fine-tuning
T5-Large	11.14	−0.68	**20.44**	−5.19	1.56
T5-Base	**9.24**	4.15	7.96	1.64	7.41
T5-XLarge	**14.35**	2.46	10.74	5.72	−9.61
BART-Large	7.63	1.16	**8.94**	−2.37	2.74
GPT2-Large	3.19	−2.68	**4.62**	−1.43	3.75

Table 5: Average relative gain (ARG %) of all methods with different backbone models in the NP→P setting. 'MTL' stands for 'multi-task learning'.

Our experiments and analysis so far use T5-Large as the backbone model. To verify whether the performance gain of MPT is consistent across different backbone models, we extend the experiments to T5-Base, T5-XLarge, BART-Large and GPT2-Large in the NP→P setting. From the results shown in Table 5, we can see that MPT still outperforms PT and MTL by a large margin when using other PLMs as the backbone model, showing its robustness to model size and type. In addition, the consistent gain of MPT with T5-XLarge could also verify the effectiveness of MPT for huge PLMs which have been shown to perform better in prompt tuning (Lester et al., 2021).

6.1 Further Analysis

Prompt tuning (PT) vs. Fine-tuning (FT). While PT shows strong few-shot learning ability, FT remains the dominant paradigm. As shown in Table 2, FT outperforms PT when adapting to classification tasks even in few-shot settings, which might be because PT has only a few tunable parameters. Though MPT is based on PT, its performance gain over FT in all cases suggests that it can learn to initialize the prompt embeddings from source tasks, enabling effective knowledge transfer.

7 Conclusion

In this paper, we have introduced meta prompt tuning (MPT), which learns to initialize the prompt embeddings for adapting to a target task. We have identified key research questions and systematically studied where and how meta learning can improve cross-task generalization in prompt tuning. We have empirically analyzed a representative set of meta learning methods in a variety of adaptation settings on a large, diverse collection of few-shot tasks. Extensive experimental results and analysis verify the effectiveness of MPT. Given the findings, in the future, we would like to explore more advanced meta learning algorithms which can consistently outperform multi-task learning.

References

Tiago A. Almeida, José María G. Hidalgo, and Akebo Yamakami. 2011. Contributions to the study of sms spam filtering: New collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering*, DocEng '11, page 259–262, New York, NY, USA. Association for Computing Machinery.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. 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 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Antreas Antoniou, Harrison Edwards, and Amos Storkey. 2019. How to train your MAML. In *International Conference on Learning Representations*.

Kristjan Arumae and Fei Liu. 2019. Guiding extractive summarization with question-answering rewards. 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 2566–2577, Minneapolis, Minnesota. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. 2020. ProtoQA: A question answering dataset for prototypical common-sense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1122–1136, Online. Association for Computational Linguistics.

Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 732–737, Brussels, Belgium. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Susan Carey and E. Bartlett. 1978. Acquiring a single new word. *Proceedings of the Stanford Child Language Conference*, 15:17–29.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020a. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xilun Chen, Asish Ghoshal, Yashar Mehdad, Luke Zettlemoyer, and Sonal Gupta. 2020b. Low-resource domain adaptation for compositional task-oriented semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5090–5100, Online. Association for Computational Linguistics.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730, Dublin, Ireland. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv preprint*, abs/1803.05457.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 5925–5932, Hong Kong, China. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

T. Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *ArXiv preprint*, abs/2012.00614.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1192–1197, Hong Kong, China. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V. U. Güney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *ArXiv preprint*, abs/1704.05179.

Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Computer Speech & Language*, 59:123–156.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Manaal Faruqui and Dipanjan Das. 2018. Identifying well-formed natural language questions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 798–803, Brussels, Belgium. Association for Computational Linguistics.

Chelsea Finn. 2022. Deep multi-task and meta learning.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Tianyu Gao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2020. Neural snowball for few-shot relation learning.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. 2018. Conditional neural processes. In *International Conference on Machine Learning*, pages 1704–1713. PMLR.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Edward Grefenstette, Brandon Amos, Denis Yarats, Phu Mon Htut, Artem Molchanov, Franziska Meier, Douwe Kiela, Kyunghyun Cho, and Soumith Chintala. 2019. Generalized inner loop meta-learning. *arXiv preprint arXiv:1910.01727*.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892. Text Mining and Natural Language Processing in Pharmacogenomics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Yukun Huang, Kun Qian, and Zhou Yu. 2022. Learning a better initialization for soft prompts via meta-learning. *arXiv preprint arXiv:2205.12471*.

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: A new factoid QA data set matching trivia-style question-answer pairs with Freebase. 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 318–323, Minneapolis, Minnesota. Association for Computational Linguistics.

Akhil Kedia, Sai Chetan Chinthakindi, and Wonho Ryu. 2021. Beyond reptile: Meta-learned dot-product maximization between gradients for improved single-task regularization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 407–420, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8082–8090.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5189–5197. AAAI Press.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. 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 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Hung-yi Lee, Shang-Wen Li, and Ngoc Thang Vu. 2022. Meta learning for natural language processing: A survey. *arXiv preprint arXiv:2205.01500*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, D. Kontokostas, Pablo N. Mendes, Sebastian Hellmann, M. Morsey, Patrick van Kleef, S. Auer, and C. Bizer. 2015. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6:167–195.

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.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Jian-Yun Nie, Ji-Rong Wen, and Xin Zhao. 2022. Learning to transfer prompts for text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3506–3518, Seattle, United States. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. 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.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020a. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020b. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. 2022. TRGP: Trust region gradient projection for continual learning. In *International Conference on Learning Representations*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.*, 65(4):782–796.

Irene Manotas, Ngoc Phuoc An Vo, and Vadim Sheinin. 2020. LiMiT: The literal motion in text dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 991–1000, Online. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *ArXiv preprint*, abs/2012.10289.

Julian J. McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*, pages 165–172. ACM.

Clara H. McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2020. Effective transfer learning for identifying similar questions: Matching user questions to COVID-19 faqs. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3458–3465. ACM.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. A simple neural attentive meta-learner. In *International Conference on Learning Representations*.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. *ArXiv preprint*, abs/2006.08328.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100, Montréal, Canada. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

A. Othman and M. Jemni. 2012. English-asl gloss parallel corpus 2012: Aslg-pc12. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon LREC*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online. Association for Computational Linguistics.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Quan Hung Tran, and Thien Huu Nguyen. 2020. What does this acronym mean? introducing a new dataset for acronym identification and disambiguation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3285–

3301, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chengwei Qin and Shafiq Joty. 2022a. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2776–2789, Dublin, Ireland. Association for Computational Linguistics.

Chengwei Qin and Shafiq Joty. 2022b. LFPT5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5. In *International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

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

Jurgen Schmidhuber. 1987. Evolutionary principles in self-referential learning. on learning now to learn: The meta-meta-meta...-hook. Diploma thesis, Technische Universitat Munchen, Germany, 14 May.

Emily Sheng and David Uthus. 2020. Investigating societal biases in a poetry composition system. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 93–106, Barcelona, Spain (Online). Association for Computational Linguistics.

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

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. 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 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages

1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019a. Quarel: A dataset and models for answering questions about qualitative relationships. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7063–7071.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019b. QuaRTz: An open-domain dataset of qualitative relationship questions. 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 5941–5946, Hong Kong, China. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for “what if...” reasoning over procedural text. 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 6076–6085, Hong Kong, China. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Eleni Triantafillou, Richard S. Zemel, and Raquel Urtasun. 2017. Few-shot learning through an information retrieval lens.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou’, and Daniel Cer. 2022. 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.

William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan

Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. TWEETQA: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5020–5031, Florence, Italy. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hao Zhang, Jae Ro, and Richard Sproat. 2020. Semi-supervised URL segmentation with recurrent neural networks pre-trained on knowledge graph entities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4667–4675, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Rui Zhang and Joel Tetreault. 2019. This email could save your life: Introducing the task of email subject line generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 446–456, Florence, Italy. Association for Computational Linguistics.

Sheng Zhang, X. Liu, J. Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *ArXiv preprint*, abs/1810.12885.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language usin. *ArXiv preprint*, abs/1709.00103.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. 2022. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*.

A Appendix

A.1 Task List

We report the full list of tasks used in ten different settings in Table 8. All tasks are taken from CROSSFIT (Ye et al., 2021).

A.2 Detailed Results on Every Target Task

We mainly report average relative gain (ARG) in our experiments (§6). In this section, we show detailed results on each target task in Fig. 4 $\sim$ Fig. 13.

A.3 Details of Sampled Tasks

We sample $\{12, 24\}$ tasks from the original 45 source tasks in the Cls→Cls setting to investigate the influence of the number of source tasks. The details of sampled tasks are shown in Table 9.

A.4 Details of Hyperparameters

For meta-training, we set the inner and outer learning rates to $3\mathrm{e}{-5}$ and $5\mathrm{e}{-1}$, respectively. We use 5000 for total training steps. We set the inner batch size to 2, 4 and 4, and inner update steps to 1, 1 and 10 for MAML, FoMAML and Reptile, respectively. For multi-task learning, we set the learning rate, batch size and number of epochs to $5\mathrm{e}{-1}$, 4 and 20, respectively. For MAML, we select the inner learning rate from $\{2\mathrm{e}{-5}, 3\mathrm{e}{-5}, 5\mathrm{e}{-5}\}$, the outer learning rate from $\{2\mathrm{e}{-1}, 3\mathrm{e}{-1}, 5\mathrm{e}{-1}\}$, and total training steps from $\{2500, 5000, 10000\}$. We adopt the same three hyperparameters for FoMAML and Reptile. The search range for the inner update steps of Reptile is $\{2, 4, 6, 8, 10\}$. For multi-task learning, we select the learning rate from $\{2\mathrm{e}{-1}, 3\mathrm{e}{-1}, 5\mathrm{e}{-1}\}$, the batch size from $\{2, 4, 6, 8\}$, and the number of epochs from $\{5, 10, 20\}$.

For downstream learning, we mainly follow the settings in Ye et al. (2021). For prompt tuning, we select the learning rate from $\{5\mathrm{e}{-1}, 4\mathrm{e}{-1}, 3\mathrm{e}{-1}, 2\mathrm{e}{-1}\}$ based on the validation performance. For fine-tuning, the search range for the learning rate is $\{5\mathrm{e}{-4}, 3\mathrm{e}{-4}, 2\mathrm{e}{-4}, 1\mathrm{e}{-4}\}$. We set the batch size, total training steps and evaluation interval to 8, 3000 and 50, respectively.

A.5 Hyperparameter Sensitivity Analysis

As mentioned in Appendix A.4, for MAML, we select the inner learning rate from $\{2\mathrm{e}{-5}, 3\mathrm{e}{-5}, 5\mathrm{e}{-5}\}$, the outer learning rate from $\{2\mathrm{e}{-1}, 3\mathrm{e}{-1}, 5\mathrm{e}{-1}\}$, and total training steps from $\{2500, 5000, 10000\}$ in the R→R setting. The best validation performance (10.14% ARG) is achieved with $\{3\mathrm{e}{-5}, 5\mathrm{e}{-1}, 5000\}$, while the worst validation ARG is -16.21% when using $\{5\mathrm{e}{-5}, 2\mathrm{e}{-1}, 2500\}$. We can see that MPT is quite sensitive to hyperparameters. It performs even worse than PT with inappropriate hyperparameters.

A.6 Task Similarity Analysis

As discussed in §6, we use the correlation between input subspaces for two tasks as the similarity score between them. Detailed results of randomly picked similar and dissimilar task pairs are shown in Table 6.

A.7 Case Study

To take a closer look at the influence of different source task types on a particular target task, we further conduct a case study where we ensure that the task under consideration appears in the target task partitions.[2] Results are shown in Table 7; for example, the first block indicates that `Amazon_Polarity` appears as a target task in both R→R and Cls→Cls settings. We can observe that there is no consistent conclusion on how we should choose the source tasks for a specific target task, which is consistent with our view in Q3.

A.8 Limitations

Although comprehensive, our study of MPT in this work has couple of limitations:

- As mentioned in §5.3, because of infeasiblity to search for optimal hyperparameters for each of the meta learning methods in each of the ten settings, we choose to use the R→R setting as our main representative setting. This could be one of the reasons for MPT underperforming MTL in some non-classification tasks (noted in §6-Q1).
- We mainly focus on how upstream meta learning can improve the performance on target tasks. However, meta learning also enables faster convergence. We leave how it could help reduce the convergence time of PT as future work.

[2] As before, we ensure it does not appear in the source.

	Task Pair Index					Average
	1	2	3	4	5	
Similar	0.772	0.695	0.754	0.819	0.802	0.768
Dissimilar	0.326	0.311	0.283	0.315	0.297	0.306

Table 6: Similarity scores of randomly picked similar and dissimilar task pairs.

Target Task	**Partition**	Δ_{MPT}	Δ_{MTL}
Amazon_Polarity	R→R	3.10	2.25
	Cls→Cls	7.40	10.45
AI2_ARC	R→R	12.54	5.55
	Both→Non-Cls	8.17	6.69
Samsum	R→R	1.97	6.77
	Both→Non-Cls	2.50	5.71
Superglue-Copa	Both→Non-Cls	1.20	10.00
	QA→QA	−3.20	4.80

Table 7: Relative gain in % for MPT and MTL when the same target task appears in different patitions.

Aside from that, meta prompt tuning (MPT) as a method has a limitation that it is Memory-intensive. Optimization-based meta learning methods, especially MAML, are memory-intensive, which limits the tuning of the inner batch size and inner update steps (Appendix A.4). One potential solution is to build more memory-efficient meta learning libraries.

Partition: **Random Source** glue-mrpc, math_qa, quarel, e2e_nlg_cleaned, tweet_eval-stance_atheism, lama-squad, tab_fact, aqua_rat, tweet_eval-emoji, glue-wnli, codah, tweet_eval-offensive, wiki_qa, blimp-ellipsis_n_bar_1, openbookqa, sms_spam, acronym_identification, blimp-determiner_noun_agreement_with_adj_irregular_1, ethos-national_origin, spider, hellaswag, superglue-wsc, numer_sense, ade_corpus_v2-dosage, blimp-ellipsis_n_bar_2, kilt_ay2, squad-no_context, google_wellformed_query, xsum, wiqa, tweet_eval-stance_abortion, reddit_tifu-tldr, ade_corpus_v2-effect, qa_srl, ethos-religion, commonsense_qa, biomrc, superglue-multirc, ethos-race, eli5-askh, glue-qqp, paws, ethos-directed_vs_generalized, glue-sst2, tweet_eval-hate, glue-rte, blimp-anaphor_number_agreement, lama-conceptnet, hate_speech_offensive, superglue-wic, boolq, kilt_hotpotqa, quartz-no_knowledge, aslg_pc12, sick, tweet_eval-stance_climate, tweet_eval-sentiment, crows_pairs, glue-mnli, medical_questions_pairs, break-QDMR-high-level, qasc, imdb, ethos-gender, trec-finegrained, adversarialqa, onestop_english, web_questions, duorc, swag, proto_qa, scitail, tweet_eval-stance_feminist, limit, common_gen, scicite, blimp-irregular_past_participle_adjectives, social_i_qa, anli, kilt_zsre, cosmos_qa, superglue-record, squad-with_context, emotion, blimp-existential_there_quantifiers_1, race-middle, kilt_wow, sciq, wino_grande, rotten_tomatoes, superglue-cb, poem_sentiment, ropes, reddit_tifu-title, piqa, climate_fever, lama-google_re, search_qa, mc_taco, blimp-wh_questions_object_gap, hotpot_qa, emo, kilt_nq, kilt_trex, quartz-with_knowledge, dbpedia_14, yahoo_answers_topics, superglue-copa, blimp-anaphor_gender_agreement, hate_speech18, gigaword, multi_news, aeslc, quail
Partition: **Random Target** quoref, wiki_split, ethos-disability, yelp_polarity, superglue-rte, glue-cola, ethos-sexual_orientation, blimp-sentential_negation_npi_scope, ai2_arc, amazon_polarity, race-high, blimp-sentential_negation_npi_licensor_present, tweet_eval-irony, crawl_domain, freebase_qa, glue-qnli, hatexplain, ag_news, circa, samsum
Partition: **Classification Source** superglue-rte, tweet_eval-sentiment, discovery, glue-rte, superglue-wsc, scicite, glue-mrpc, tweet_eval-stance_hillary, tweet_eval-offensive, emotion, hatexplain, glue-cola, sick, paws, ethos-sexual_orientation, glue-qqp, tweet_eval-emotion, sms_spam, health_fact, glue-mnli, imdb, ethos-disability, glue-wnli, scitail, trec-finegrained, yahoo_answers_topics, liar, glue-sst2, tweet_eval-stance_abortion, circa, tweet_eval-stance_climate, glue-qnli, tweet_eval-emoji, ethos-directed_vs_generalized, ade_corpus_v2-classification, ag_news, hate_speech_offensive, superglue-wic, google_wellformed_query, tweet_eval-irony, ethos-gender, onestop_english, trec, rotten_tomatoes, kilt_fever
Partition: **Non-Classification Source** ade_corpus_v2-dosage, art, biomrc, blimp-anaphor_number_agreement, blimp-ellipsis_n_bar_2, blimp-sentential_negation_npi_licensor_present, blimp-sentential_negation_npi_scope, break-QDMR-high-level, commonsense_qa, crows_pairs, dream, duorc, eli5-asks, eli5-eli5, freebase_qa, gigaword, hellaswag, hotpot_qa, kilt_ay2, kilt_hotpotqa, kilt_trex, kilt_zsre, lama-conceptnet, lama-google_re, lama-squad, math_qa, numer_sense, openbookqa, piqa, proto_qa, qa_srl, quarel, quartz-no_knowledge, race-high, reddit_tifu-title, reddit_tifu-tldr, ropes, sciq, social_i_qa, spider, superglue-multirc, wiki_bio, wikisql, xsum, yelp_review_full
Partition: **Both (Classification + Non-Classification) Source** ade_corpus_v2-dosage, biomrc, blimp-ellipsis_n_bar_2, blimp-sentential_negation_npi_scope, commonsense_qa, crows_pairs, duorc, hellaswag, kilt_zsre, lama-google_re, lama-squad, math_qa, numer_sense, openbookqa, piqa, proto_qa, quartz-no_knowledge, race-high, reddit_tifu-tldr, ropes, sciq, wiki_bio, discovery, emotion, ethos-disability, ethos-sexual_orientation, glue-cola, glue-mnli, glue-mrpc, glue-qqp, glue-rte, glue-wnli, hatexplain, health_fact, imdb, paws, scicite, sick, sms_spam, superglue-rte, superglue-wsc, tweet_eval-emotion, tweet_eval-offensive, tweet_eval-sentiment, tweet_eval-stance_hillary
Partition: **Classification Target** superglue-cb,dbpedia_14,wiki_qa,emo,yelp_polarity,ethos-religion,amazon_polarity,tab_fact,anli,ethos-race
Partition: **Non-Classification Target** multi_news, superglue-copa, quail, blimp-anaphor_gender_agreement, common_gen, acronym_identification, quoref, wiki_split, ai2_arc, break-QDMR, crawl_domain, samsum
Partition: **QA Source** biomrc, boolq, freebase_qa, hotpot_qa, kilt_hotpotqa, kilt_nq, kilt_trex, kilt_zsre, lama-conceptnet, lama-google_re, lama-squad, lama-trex, mc_taco, numer_sense, quoref, ropes, search_qa, squad-no_context, superglue-multirc, superglue-record, tweet_qa, web_questions
Partition: **Non-QA Source** hate_speech_offensive, google_wellformed_query, circa, glue-sst2, scitail, emo, ag_news, art, paws, kilt_ay2, glue-qnli, ade_corpus_v2-classification, hatexplain, emotion, glue-qqp, kilt_fever, dbpedia_14, glue-mnli, discovery, gigaword, amazon_polarity, tab_fact, tweet_eval-emoji, tweet_eval-offensive, tweet_eval-sentiment, imdb, liar, anli, wikisql, xsum, yahoo_answers_topics, yelp_polarity, yelp_review_full
Partition: **QA Target** ai2_arc, codah, cosmos_qa, dream, hellaswag, qasc, quail, quarel, quartz-no_knowledge, quartz-with_knowledge, sciq, superglue-copa, swag, wino_grande, wiqa
Partition: **Non-Paraphrase Classification Source** ade_corpus_v2-classification, ag_news, amazon_polarity, anli, circa, climate_fever, dbpedia_14, discovery, emo, emotion, ethos-directed_vs_generalized, ethos-disability, ethos-gender, ethos-national_origin, ethos-race, ethos-religion, ethos-sexual_orientation, financial_phrasebank, glue-cola, glue-mnli, glue-qnli, glue-rte, glue-sst2, glue-wnli, google_wellformed_query, hate_speech18, hate_speech_offensive, hatexplain, health_fact, imdb, kilt_fever, liar, onestop_english, poem_sentiment, rotten_tomatoes, scicite, scitail, sick, sms_spam, superglue-cb, superglue-rte, superglue-wic, superglue-wsc, tab_fact, trec, trec-finegrained, tweet_eval-emoji, tweet_eval-emotion, tweet_eval-hate, tweet_eval-irony, tweet_eval-offensive, tweet_eval-sentiment, tweet_eval-stance_abortion, tweet_eval-stance_atheism, tweet_eval-stance_climate, tweet_eval-stance_feminist, tweet_eval-stance_hillary, wiki_qa, yahoo_answers_topics, yelp_polarity
Partition: **Paraphrase Target** glue-mrpc, glue-qqp, medical_questions_pairs, paws

Table 8: Full datasets for all settings described in Section 5.1. We provide references for all datasets in Table 10.

12 source tasks superglue-rte, tweet_eval-sentiment, discovery, glue-rte, hatexplain, glue-cola, health_fact, glue-mnli, imdb, ethos-disability, glue-wnli, scitail
24 source tasks superglue-rte, tweet_eval-sentiment, discovery, glue-rte, superglue-wsc, scicite, hatexplain, glue-cola, tweet_eval-emotion, sms_spam, health_fact, glue-mnli, imdb, ethos-disability, glue-wnli, scitail, glue-sst2, tweet_eval-stance_abortion, glue-qnli, ethos-directed_vs_generalized, ag_news, hate_speech_offensive, ethos-gender, kilt_fever

Table 9: Details of sampled $\{12, 24\}$ tasks for investigating the impact of the number of source tasks.

Task Name	Reference
eli5-eli5	Fan et al. 2019
ethos-race	Mollas et al. 2020
tweet_qa	Xiong et al. 2019
tweet_eval-stance_hillary	Barbieri et al. 2020
piqa	Bisk et al. 2020
acronym_identification	Pouran Ben Veyseh et al. 2020
wiki_split	Botha et al. 2018
scitail	Khot et al. 2018
emotion	Saravia et al. 2018
medical_questions_pairs	McCreery et al. 2020
blimp-anaphor_gender_agreement	Warstadt et al. 2020
sciq	Welbl et al. 2017
paws	Zhang et al. 2019
yelp_review_full	Zhang et al. 2015; (link)
freebase_qa	Jiang et al. 2019
anli	Nie et al. 2020
quartz-with_knowledge	Tafjord et al. 2019b
hatexplain	Mathew et al. 2020
yahoo_answers_topics	(link)
search_qa	Dunn et al. 2017
tweet_eval-stance_feminist	Barbieri et al. 2020
codah	Chen et al. 2019
lama-squad	Petroni et al. 2019, 2020
superglue-record	Zhang et al. 2018
spider	Yu et al. 2018
mc_taco	Zhou et al. 2019
glue-mrpc	Dolan and Brockett 2005
kilt_fever	Thorne et al. 2018
eli5-asks qa	Fan et al. 2019
imdb	Maas et al. 2011
tweet_eval-stance_abortion	Barbieri et al. 2020
aqua_rat	Ling et al. 2017
duorc	Saha et al. 2018
lama-trex	Petroni et al. 2019, 2020
tweet_eval-stance_atheism	Barbieri et al. 2020
ropes	Lin et al. 2019
squad-no_context	Rajpurkar et al. 2016
superglue-rte	Dagan et al. 2005
qasc	Khot et al. 2020
hate_speech_offensive	Davidson et al. 2017
trec-finegrained	Li and Roth 2002; Hovy et al. 2001
glue-wnli	Levesque et al. 2012
yelp_polarity	Zhang et al. 2015; (link)
kilt_hotpotqa	Yang et al. 2018
glue-sst2	Socher et al. 2013
xsum	Narayan et al. 2018
tweet_eval-offensive	Barbieri et al. 2020
aeslc	Zhang and Tetreault 2019
emo	Chatterjee et al. 2019
hellaswag	Zellers et al. 2019
social_i_qa	Sap et al. 2019
kilt_wow	Dinan et al. 2019
scicite	Cohan et al. 2019
superglue-wsc	Levesque et al. 2012
hate_speech18	de Gibert et al. 2018
adversarialqa	Bartolo et al. 2020
break-QDMR	Wolfson et al. 2020
dream	Sun et al. 2019
circa	Louis et al. 2020
wiki_qa	Yang et al. 2015
ethos-directed_vs_generalized	Mollas et al. 2020
wiqa	Tandon et al. 2019
poem_sentiment	Sheng and Uthus 2020
kilt_ay2	Hoffart et al. 2011
cosmos_qa	Huang et al. 2019
reddit_tifu-title	Kim et al. 2019
superglue-cb	de Marneffe et al. 2019
kilt_nq	Kwiatkowski et al. 2019
quarel	Tafjord et al. 2019a
race-high	Lai et al. 2017
wino_grande	Sakaguchi et al. 2020
break-QDMR-high-level	Wolfson et al. 2020
tweet_eval-irony	Barbieri et al. 2020
liar	Wang 2017
openbookqa	Mihaylov et al. 2018
superglue-multirc	Khashabi et al. 2018
race-middle	Lai et al. 2017
quoref	Dasigi et al. 2019
cos_e	Rajani et al. 2019
reddit_tifu-tldr	Kim et al. 2019
ai2_arc	Clark et al. 2018
quail	Rogers et al. 2020
crawl_domain	Zhang et al. 2020
glue-cola	Warstadt et al. 2019

Task Name	Reference
art	Bhagavatula et al. 2020
rotten_tomatoes	Pang and Lee 2005
tweet_eval-emoji	Barbieri et al. 2020
numer_sense	Lin et al. 2020a
blimp-existential_there_quantifiers_1	Warstadt et al. 2020
eli5-askh qa	Fan et al. 2019
ethos-national_origin	Mollas et al. 2020
boolq	Clark et al. 2019
qa_srl	He et al. 2015
sms_spam	Almeida et al. 2011
samsum	Gliwa et al. 2019
ade_corpus_v2-classification	Gurulingappa et al. 2012
superglue-wic	Pilehvar and Camacho-Collados 2019
ade_corpus_v2-dosage	Gurulingappa et al. 2012
tweet_eval-stance_climate	Barbieri et al. 2020
e2e_nlg_cleaned	Dušek et al. 2020, 2019
aslg_pc12	Othman and Jemni 2012
ag_news	Gulli (link)
math_qa	Amini et al. 2019
commonsense_qa	Talmor et al. 2019
web_questions	Berant et al. 2013
biomrc	Pappas et al. 2020
swag	Zellers et al. 2018
blimp-determiner_noun_agreement_with_adj_irregular_1	Warstadt et al. 2020
glue-mnli	Williams et al. 2018
squad-with_context	Rajpurkar et al. 2016
blimp-ellipsis_n_bar_2	Warstadt et al. 2020
financial_phrasebank	Malo et al. 2014
sick	Marelli et al. 2014
ethos-religion	Mollas et al. 2020
hotpot_qa	Yang et al. 2018
tweet_eval-emotion	Barbieri et al. 2020
dbpedia_14	Lehmann et al. 2015
ethos-gender	Mollas et al. 2020
tweet_eval-hate	Barbieri et al. 2020
ethos-sexual_orientation	Mollas et al. 2020
health_fact	Kotonya and Toni 2020
common_gen	Lin et al. 2020b
crows_pairs	Nangia et al. 2020
ade_corpus_v2-effect	Gurulingappa et al. 2012
blimp-sentential_negation_npi_scope	Warstadt et al. 2020
lama-conceptnet	Petroni et al. 2019, 2020
glue-qnli	Rajpurkar et al. 2016
quartz-no_knowledge	Tafjord et al. 2019b
google_wellformed_query	Faruqui and Das 2018
kilt_trex	Elsahar et al. 2018
blimp-ellipsis_n_bar_1	Warstadt et al. 2020
trec	Li and Roth 2002; Hovy et al. 2001
superglue-copa	Gordon et al. 2012
ethos-disability	Mollas et al. 2020
lama-google_re	Petroni et al. 2019, 2020
discovery	Sileo et al. 2019
blimp-anaphor_number_agreement	Warstadt et al. 2020
climate_fever	Diggelmann et al. 2020
blimp-irregular_past_participle_adjectives	Warstadt et al. 2020
tab_fact	Chen et al. 2020a
gigaword	Napoles et al. 2012
glue-rte	Dagan et al. 2005
tweet_eval-sentiment	Barbieri et al. 2020
limit	Manotas et al. 2020
wikisql	Zhong et al. 2017
glue-qqp	(link)
onestop_english	Vajjala and Lučić 2018
amazon_polarity	McAuley and Leskovec 2013
blimp-wh_questions_object_gap	Warstadt et al. 2020
multi_news	Fabbri et al. 2019
proto_qa	Boratko et al. 2020
wiki_bio	Lebret et al. 2016
kilt_zsre	Levy et al. 2017
blimp-sentential_negation_npi_licensor_present	Warstadt et al. 2020

Table 10: References for all datasets.

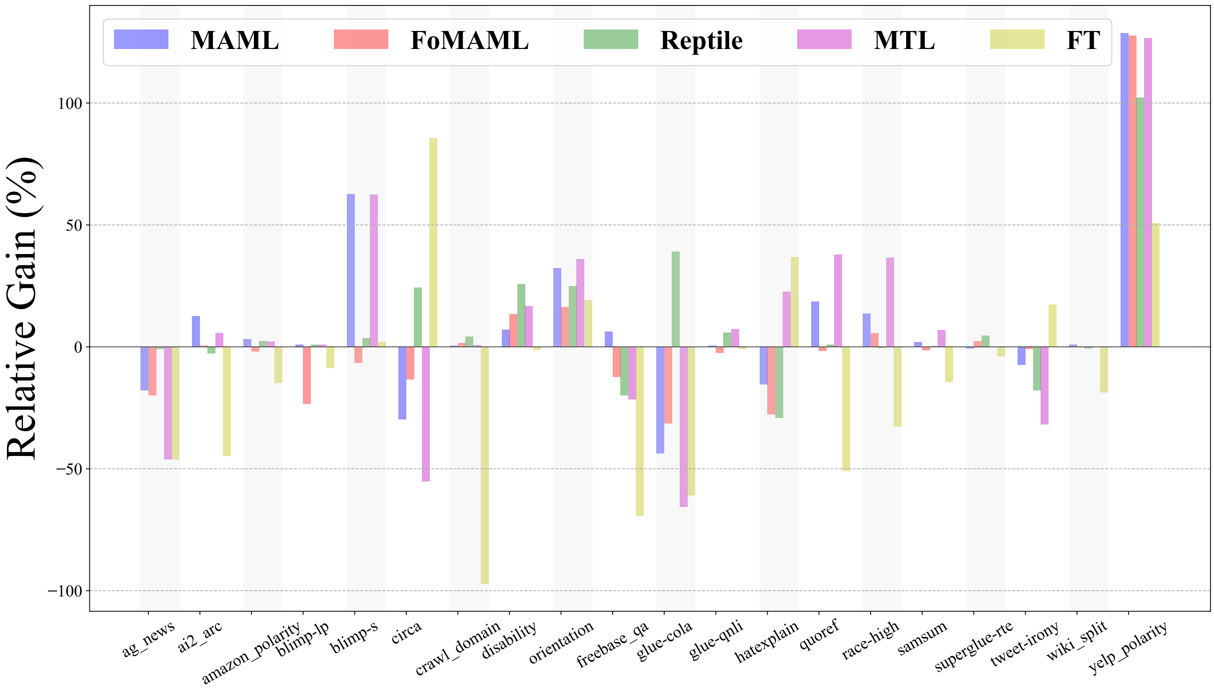

Figure 4: Random to Random

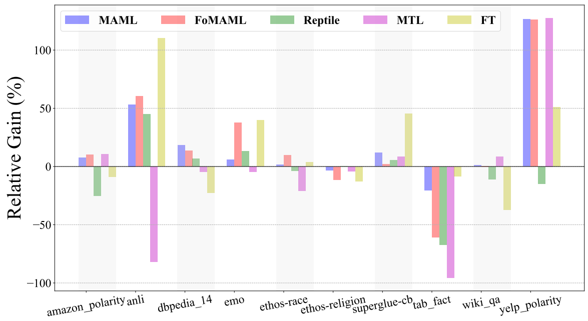

Figure 5: Classification to Classification

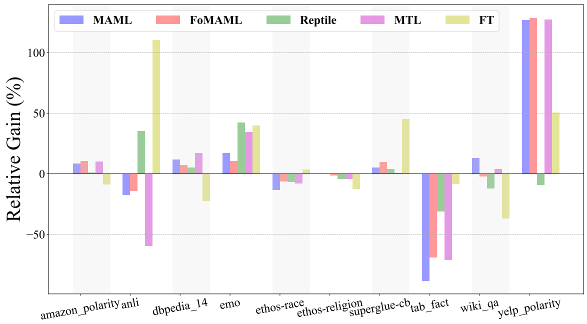

Figure 6: Non-Classification to Classification

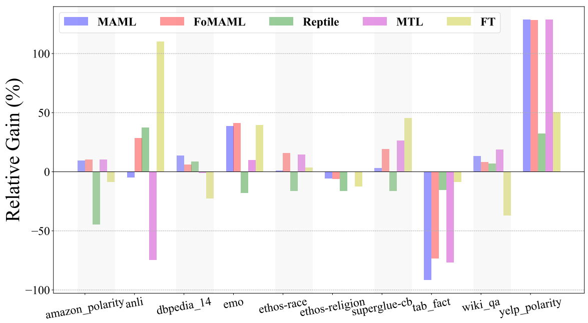

Figure 7: Both (Classification + Non-Classification) to Classification

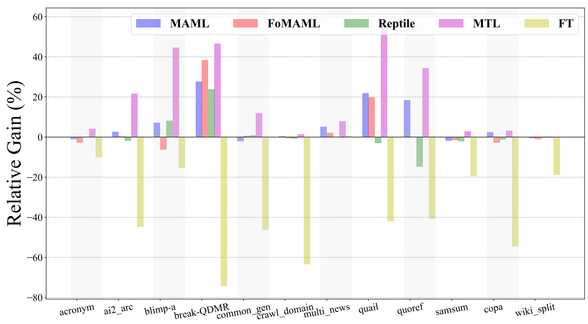

Figure 8: Non-Classification to Non-Classification

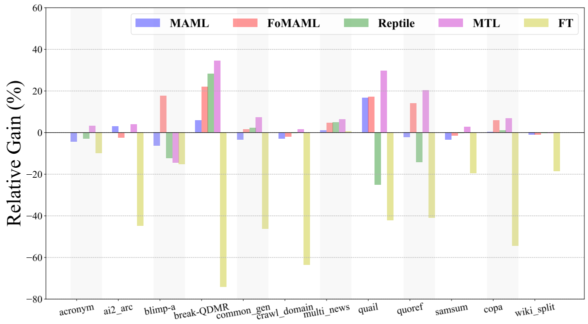

Figure 9: Classification to Non-Classification

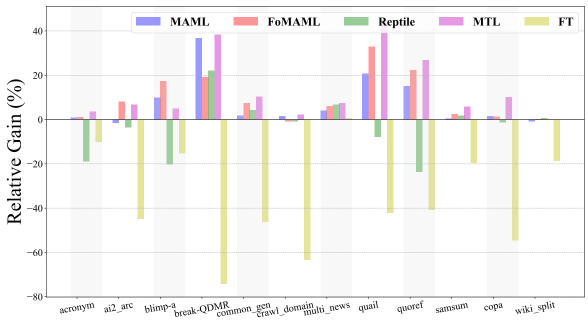

Figure 10: Both (Classification + Non-Classification) to Non-Classification

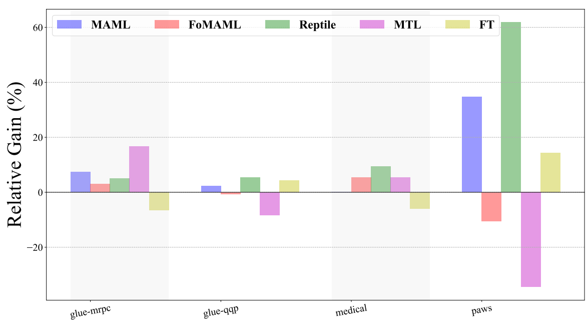

Figure 11: Non-Paraphrase Classification to Paraphrase

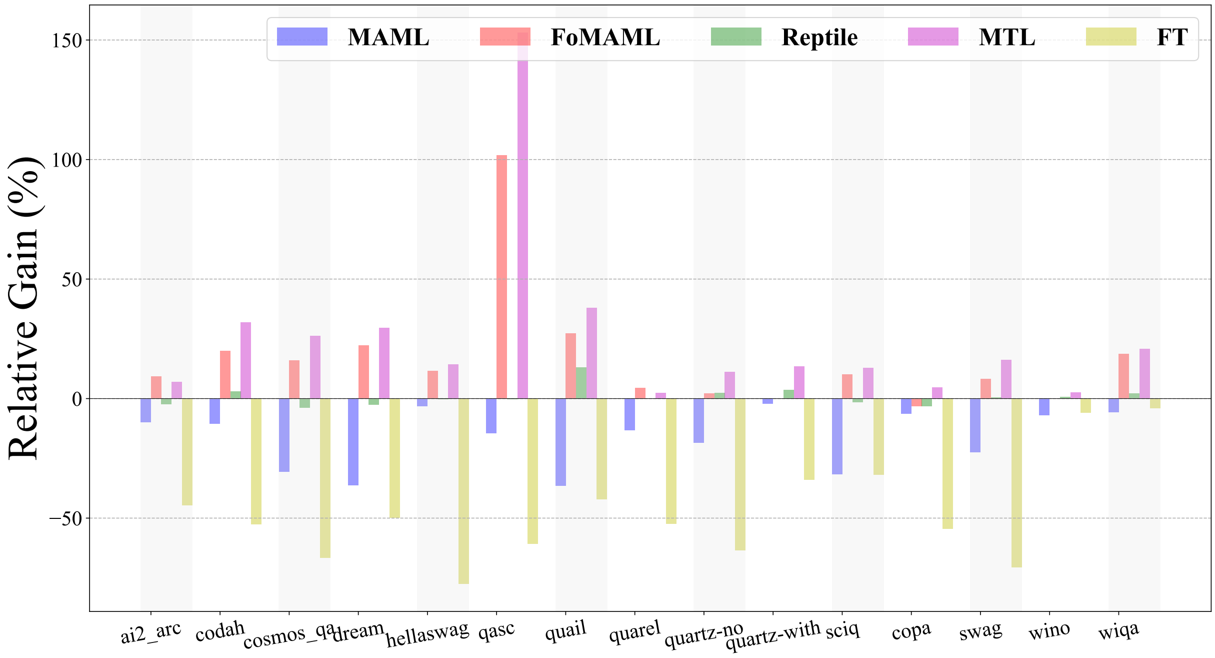

Figure 12: QA to QA

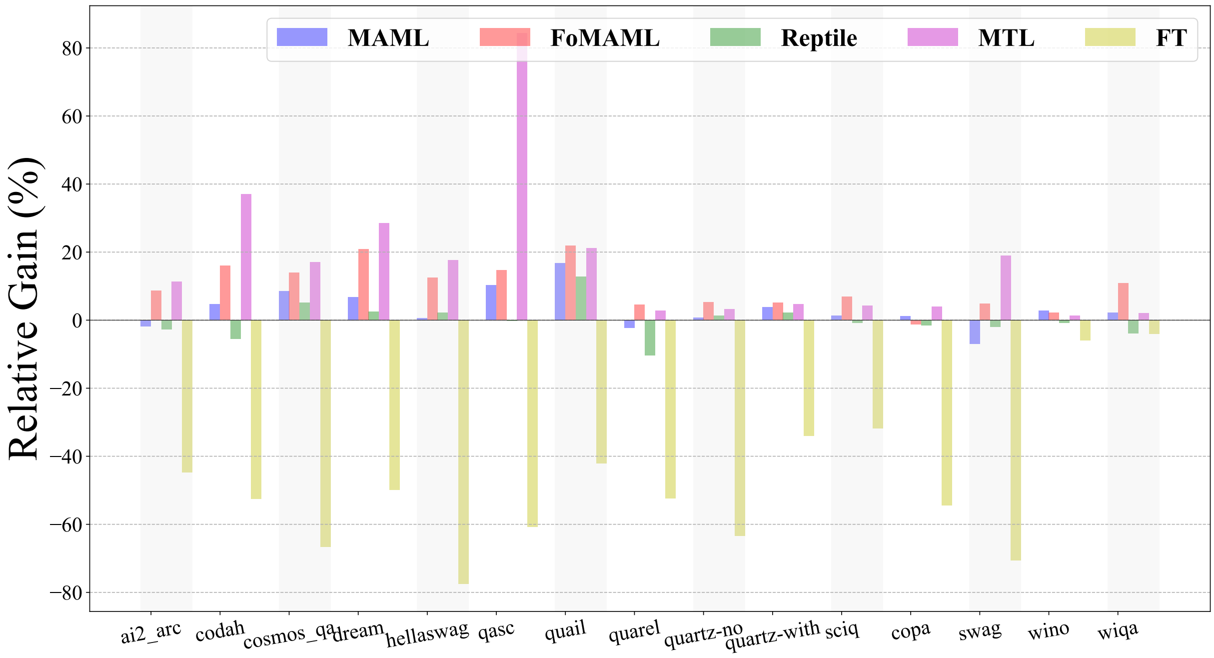

Figure 13: Non-QA to QA