CROSSFIT: A Few-shot Learning Challenge for Cross-task Generalization in NLP

Qinyuan Ye Bill Yuchen Lin Xiang Ren
University of Southern California
{qinyuany, yuchen.lin, xiangren}@usc.edu

Abstract

Humans can learn a new language task more efficiently than machines, conceivably by leveraging their prior experience and knowledge in learning other tasks. In this paper, we explore whether such *cross-task generalization* ability can be acquired, and further applied to build better *few-shot learners* across diverse NLP tasks. We introduce CROSSFIT, a task setup for studying cross-task few-shot learning ability, which standardizes seen/unseen task splits, data access during different learning stages, and the evaluation protocols. In addition, we present NLP Few-shot Gym, a repository of 160 few-shot NLP tasks, covering diverse task categories and applications, and converted to a unified text-to-text format.

Our empirical analysis reveals that the few-shot learning ability on unseen tasks can be improved via an upstream learning stage using a set of seen tasks. Additionally, the advantage lasts into medium-resource scenarios when thousands of training examples are available. We also observe that selection of upstream learning tasks can significantly influence few-shot performance on unseen tasks, asking further analysis on task similarity and transferability.[1]

1 Introduction

With recent progress in pre-trained language representations, models can learn to perform a new natural language processing (NLP) task competently with only a handful of examples (*i.e.*, few-shot learning). Moving towards this direction, researchers have developed approaches to further improve learning efficiency by re-formulating the target task into cloze questions (Schick and Schütze, 2020a,b), generating prompts and using demonstrations (Gao et al., 2020), and densifying the supervision signals (Tam et al., 2021).

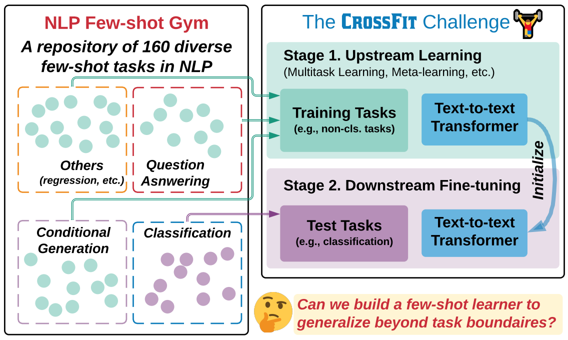

Figure 1: We present the CROSSFIT Challenge to study cross-task few-shot learning ability of a system, where the tasks are selected from a diverse distribution. To support this problem setting, we introduce NLP Few-shot Gym, a repository of 160 diverse few-shot tasks in NLP, formulated in a unified text-to-text format.

Recent advances in pre-training and fine-tuning have primarily focused on improving *instance-level generalization*, *i.e.*, within the scope of one task (dataset), how to make predictions about unseen instances given only few demonstrations. On the other hand, few-shot learning ability can potentially be improved with *task-level generalization*[2], *i.e.*, how to learn a new task efficiently given previous experiences on learning tasks. This idea of "learning to learn" has been widely explored in computer vision and robotics community (Yu et al., 2020; Triantafillou et al., 2020). For language tasks, the same intuition holds: human learners develop high-level skills by learning language tasks and apply these skills when encountering new tasks. For example, a good text classification learner may become a good reading comprehension learner, since

[1] Our code and data are publicly available at https://github.com/INK-USC/CrossFit/

[2] We use the term "task-level generalization" and "cross-task generalization" interchangeably. The former is mainly used for comparison with "instance-level generalization"

both tasks require language understanding; to learn to answer open-ended questions, experiences in learning summarization may help, since both tasks need writing coherent and informative sentences.

In fact, several attempts have already been made towards this direction in NLP. However, the tasks of interests are usually drawn from a narrow distribution. For example, both Han et al. (2018) and Bansal et al. (2020a) focus on generalization within the scope of classification tasks. We anticipate more human-like learning ability that allows generalization across different task formats (classification, span extraction, multiple choice, generation, etc.), goals (question answering, summarization, fact checking, etc.) and domains (academic, biomedical, social media, etc.).

Towards acquiring and evaluating such generalization, we propose the CROSSFIT Challenge, a task setup to investigate a system's cross-task few-shot learning ability, with standardized training pipeline, data access and evaluation protocol. In short, a system for the CROSSFIT challenge may go through an upstream learning stage on a set of seen tasks, and is then evaluated on a set of unseen tasks in few-shot scenario, as illustrated in Fig. 1. To further analyze the capability and limitation of existing approaches, we present NLP Few-shot Gym, a repository of 160 open-access NLP tasks covering a wide range of formats and goals, and formulated into a unified text-to-text few-shot setting. We then instantiate the CROSSFIT challenge with eight different seen/unseen task partitions created with NLP Few-shot Gym. With these resources, we investigate the following research questions:

- **Q1.** Do upstream learning methods, such as multitask learning and meta-learning, improve few-shot learning ability on unseen tasks?
- **Q2.** How does the selection of seen tasks influence unseen tasks performance?
- **Q3.** Does improved few-shot learning ability last when more data is available?

To answer the first two questions, we empirically analyze the performance of multi-task learning and MAML (Finn et al., 2017), a meta-learning algorithm, in the CROSSFIT setup and with the eight different task partitions. For Q1, we found that the few-shot performance is improved on a wide range of tasks after upstream learning, with significant boost on CommonsenseQA, Ropes, MNLI. These encouraging observations showcase the potential power of acquiring and leveraging cross-task generalization for few-shot learning. For Q2, we observe that performance of individual unseen tasks varies with different selection of seen tasks. In addition, we observe that non-classification tasks and classification tasks are equivalently helpful for a set of held-out unseen classification tasks. There observations call for more thorough investigation of the relationship between task similarity and transferability. For Q3, we take the three successful cases in Q1 and further examine the performance when "more shots" become available. We find that the improvements brought by upstream learning last in medium-resource scenarios (e.g., 2048 examples). For CommonsenseQA, this lasts when the full dataset is available. These findings suggest the wide use cases of CROSSFIT systems, as the improvement lasts beyond the few-shot setting.

2 Related Work

Few-shot Fine-tuning. Few-shot learning is the problem to teach models a new task with an extremely small number of annotated examples. Large-scale pre-trained language models (e.g., BERT (Devlin et al., 2019), T5 (Raffel et al., 2020)) have demonstrated great ability to learn new tasks efficiently via *fine-tuning*. Zhang et al. (2021) empirically examined fine-tuning BERT models in few-shot scenarios and provided practical suggestions to improve performance and reduce instability. Schick and Schütze (2020a,b) proposed *pattern-exploiting training* (PET), which formulates text classification and NLI tasks into cloze questions (or "prompts"). These prompts share the same format of masked language modeling, the pre-training tasks of many pre-trained LMs, and thus leads to improved few-shot performance. Extending from PET, Gao et al. (2020) proposed LM-BFF which learns to generate prompts *automatically* and incorporates demonstrations into the input; Tam et al. (2021) proposed ADAPET which densifies the supervision signal with a label conditioning objective.

While successful, in these approaches the downstream tasks are learned in isolation. Our work aims to boost few-shot learning ability on unseen tasks via acquiring cross-task generalization ability from diverse seen tasks.

Meta-learning in NLP. Recent works have explored meta-learning methods for relation classification (Han et al., 2018; Gao et al., 2019), general text classification tasks (Dou et al., 2019;

Bansal et al., 2020a,b), low-resource machine translation (Gu et al., 2018), cross-lingual NLI/QA (Nooralahzadeh et al., 2020), and syllable structure learning (McCoy et al., 2020). In general, these works formulate sub-tasks and apply meta-learning algorithms; however the sub-tasks are either *synthetic* (e.g., a new set of five relations for classification is a new sub-task) or drawn from a rather *narrow* distribution (e.g., QA in one language is a sub-task). In our work, we explore a more realistic setting of learning from a much more *diverse* set of NLP tasks: classification, question answering in different formats, conditional generation (e.g., summarization), etc.

Unifying NLP Task Formats. Recent works explored unifying the formats of different tasks, in order to better enable transfer learning. DecaNLP (McCann et al., 2018) is a benchmark including 10 different and complex NLP tasks, and all tasks are processed into a unified question answering format. UFO-Entail (Yin et al., 2020) formulates multiple-choice QA and co-reference resolution as textual entailment tasks and examines the performance in few-shot settings. T5 (Raffel et al., 2020) studies unifying all tasks in text-to-text format, including discriminative tasks that were typically solved with classification heads attached to the pre-trained model. UnifiedQA (Khashabi et al., 2020) further examines the feasibility of training a general, cross-format QA model. Our work also extends the idea of unifying different tasks into a general text-to-text format, and we significantly enlarge the task repository to 160 to broaden the coverage, in hope of building a general-purpose few-shot learner.

3 The CrossFit Challenge

In this section, we present the CrossFit Challenge, a task setup for acquiring and evaluating cross-task few-shot learning ability. Ideally, a strong CrossFit system can capture cross-task generalization ability from a set of seen tasks and adapts to new unseen tasks efficiently.

In the following, we first introduce the notations and definitions in §3.1, then present the formulation of our CrossFit challenge (§3.2) with its two learning stages (§3.3), and finally present the evaluation protocol in §3.4.

3.1 Preliminaries

Task. We define a task T as a tuple of $(\mathcal{D}_{train}, \mathcal{D}_{dev}, \mathcal{D}_{test}, E)$. Each set $\mathcal{D}$ consists of a set of annotated examples $\{(x_i, y_i)\}$. As we reformulate each task into text-to-text format, x_i and y_i are both sequences of tokens in a shared vocabulary. E denotes a function to *evaluate* the performance of a system on a task based on certain metrics of interest. We use $E(M, \mathcal{D}_{test})$ to represent the performance of a model M based on its predictions and ground-truth labels in $\mathcal{D}_{test}$.

Few-shot Task. For few-shot tasks, the size of $\mathcal{D}_{train}$ and $\mathcal{D}_{dev}$ are required to be small. For classification and regression tasks, we follow (Gao et al., 2020) and include $K = 16$ training examples *per class* in $\mathcal{D}_{train}$. For other types of tasks, we include $K = 32$ examples in $\mathcal{D}_{train}$. In conformity with real-world situations where labeled data are scarce, we assume a development set $\mathcal{D}_{dev}$ which shares the same size with $\mathcal{D}_{train}$, following (Gao et al., 2020). We defer the details of gathering different few-shot tasks from existing open-source datasets in §4.

3.2 Problem Formulation

To acquire and evaluate cross-task generalization ability, we build three non-overlapping sets of *few-shot tasks*, $\mathcal{T}_{train}$, $\mathcal{T}_{dev}$, $\mathcal{T}_{test}$. A CrossFit approach is expected to first learn from the **training tasks** $\mathcal{T}_{train}$, and (optionally) tune the hyper-parameters with **developing tasks** $\mathcal{T}_{dev}$. Finally, we evaluate the few-shot learning ability on all **test tasks** in $\mathcal{T}_{test}$. Specifically, for each test task $T = (\mathcal{D}^T_{train}, \mathcal{D}^T_{dev}, \mathcal{D}^T_{test}, E^T) \in \mathcal{T}_{test}$, we apply a few-shot fine-tuning method to obtain a model M, and assess its performance on $\mathcal{D}^T_{test}$ by executing $E^T(M, \mathcal{D}^T_{test})$.

In our experiments, we manually design several different partitions of $\mathcal{T}_{train}$, $\mathcal{T}_{dev}$, $\mathcal{T}_{test}$ (e.g., random partition, withholding a specific subcategory of tasks, etc.), in hope to examine the capability and limitation of a CrossFit approach in different settings and answer our research questions. More details are deferred in §4.4 and Table 1.

3.3 The Two Learning Stages

A CrossFit system may learn from $\mathcal{T}_{train}$ in the upstream learning stage; it is then evaluated for task-specific few-shot learning with $\mathcal{T}_{test}$:

- **Upstream learning stage.** At first, the algorithm only has access to the $\mathcal{D}_{train}$ and $\mathcal{D}_{dev}$ for each training task in $\mathcal{T}_{train}$, while the performance on $\mathcal{D}_{test}$ is not available at this stage.

- **Few-shot learning stage.** Then, the $\mathcal{T}_{dev}$ and $\mathcal{T}_{test}$ are available for the model to be fine-tuned on. A few-shot learning method (e.g., direct fine-tuning) is applied for the model to learn from $\mathcal{D}_{train}$. The few-shot learning performance is reported on $\mathcal{D}_{test}$. [3]

3.4 Evaluation Protocol

Evaluating the few-shot learning ability over a list of diverse NLP tasks can be tricky, because different tasks use different evaluation metrics. For example, classification tasks typically use *F1* score or *accuracy*, while conditional generation tasks use *exact match* or *BLEU/Rouge*. To develop a unified evaluation protocol for analyzing the performance on 160 different datasets, as shown in §4, we narrow down to a collection of 7 evaluation metrics: classification F1, accuracy, question answering F1, exact match (EM), Matthew correlation, and Pearson correlation. These metrics cover all tasks we considered in the NLP Few-shot Gym benchmark.

To aggregate over multiple tasks in evaluation, we define *Average Relative Gain* (**ARG**), a metric that computes the average relative performance changes between with/without the *upstream learning* stage for each task in evaluation. Suppose we have $\mathcal{T}_{test} = \{T_A, T_B\}$. If an upstream learning algorithm helps improve the few-shot learning performance from 50% F1 score to 70% *F1 score* on task T_A (i.e., a 40% relative improvement), and from 40% accuracy to 30% *accuracy* on task T_B (i.e., -25% relative improvement), the final ARG on $\mathcal{T}_{test}$ would be computed as $\frac{40\%+(-25\%)}{2} = 7.5\%$.

The ARG metric reflects the *overall* performance gain on all tasks in $\mathcal{T}_{dev}$ or $\mathcal{T}_{test}$, no matter what specific metrics each task uses. We use ARG for a high-level comparison, and we still report the improvement on each task for in-depth analysis.

4 NLP Few-shot Gym

In support of CROSSFIT learning, we introduce the NLP Few-shot Gym, a repository of 160 few-shot learning tasks in NLP, covering a wide range of NLP applications and language skills in multiple distinct task formats. In this section, we introduce the dataset selection criteria as well as the ontology we create to facilitate analysis (§4.1), and the details about unifying task formats (§4.2) and data sampling (§4.3).

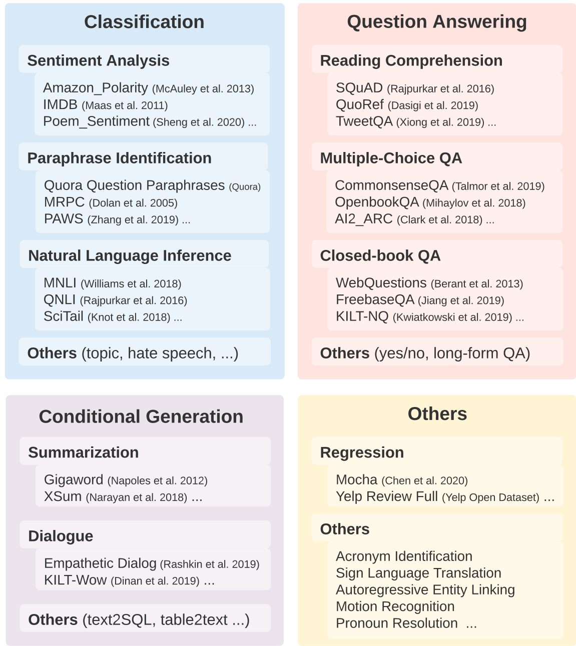

Figure 2: Task Ontology for NLP Few-shot Gym

4.1 Dataset Selection

We choose to use *Huggingface Datasets* [4] as the pool of our candidate tasks and datasets. *Huggingface Datasets* is an extensible and open-source library and provides access to numerous open-access NLP tasks with a unified API. We further select datasets based on the following criteria:

1. We focus on English monolingual datasets.
2. We exclude tasks leveraging external knowledge sources or information retrieval technique.
3. We exclude sequence labeling tasks (e.g., dependency parsing, NER), which is highly dependent on tokenization, and is hard to evaluate when converted into sequence-to-sequence format.
4. We exclude datasets that aim for special domains, e.g., COVID-19 related dataset;
5. We exclude datasets dealing with extremely long documents (e.g., a scientific paper) as input, as most pre-trained models cannot process such long input sequences.

After filtering tasks that conflict with any criteria, we finalize with 160 datasets, the details of which

[3] The performance on the $\mathcal{D}_{dev}$ of a task in $\mathcal{T}_{dev}$ or $\mathcal{T}_{test}$ will be used for tuning task-specific model-level hyper-parameters. The overall performance on $\mathcal{T}_{dev}$ is used for tuning the hyper-parameters for upstream learning.

[4] https://huggingface.co/datasets. As of February 25, 2021, there are 626 datasets on Huggingface Datasets

are listed in Appendix. We manually classify the 160 datasets and form a **task ontology** with categories and sub-categories as, shown in Fig. 2. This ontology enables us to analyze the cross-task generalization performance grouped by their categories.

4.2 A Unified Sequence-to-Sequence Format

We follow Raffel et al. (2020) to convert all of our tasks into one unified *text-to-text format* similar to the T5 model's fine-tuning. For example, the task of natural language inference (originally a sentence-pair classification format) becomes: `premise: <premise> hypothesis: <hypothesis>`, and the target sequence is either the word `entailment`, `contradiction` or `neutral`. As for machine reading comprehension tasks, the input format is `question: <question> context: <context>` and the target sequence is the correct answer span. We also reference the format for QA tasks from (Khashabi et al., 2020).

4.3 Few-shot Sampling

We mainly follow the practice in (Gao et al., 2020) by randomly sampling $\mathcal{D}_{train}$ and $\mathcal{D}_{dev}$ splits from each dataset's original train set with 5 different random seeds. This helps us reduce variance during evaluation, and also enlarges the number of few-shot tasks used for learning. Consequently, the "effective size" of the NLP Few-shot Gym is $160 \times 5 = 800$, while we use the number 160 in the following to avoid possible confusion.

We use the original development set for each dataset as $\mathcal{D}_{test}$, or held-out 20% of the dataset when the official development split is not available. The held-out test examples are sampled *once* before sampling $\mathcal{D}_{train}$ and $\mathcal{D}_{dev}$.

4.4 Task Partitions

To comprehensively evaluate a CrossFit system in different scenarios we design 8 different partitions of $(\mathcal{T}_{train}, \mathcal{T}_{dev}, \mathcal{T}_{test})$. We list the details in 1. Our Partition 1 randomly split all 160 few-shot tasks into the three sets, where $|\mathcal{T}_{train}| = 120$ and $|\mathcal{T}_{dev}| = |\mathcal{T}_{test}| = 20$. The design of Partition 1 mimics the real-world language learning environment where the goal is to build a general purpose few-shot learner, and a set of diverse tasks are seen to the learner.

Our Partition 2.1-2.3 withhold 10 classification tasks for development and 10 more for testing. The

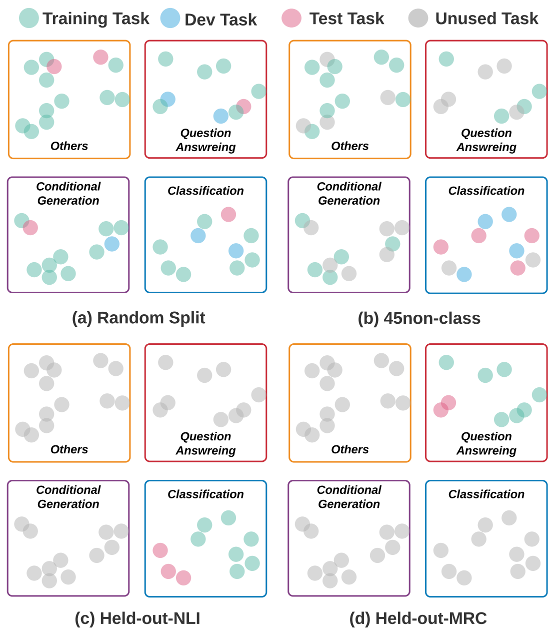

Figure 3: **Illustration for different task partitions.** We evaluate a CrossFit approach on different task partitions to examine its generalization ability in different scenarios. Full details in Table 1.

$|\mathcal{T}_{train}|$ is controlled to have either 100% classification tasks, 100% non-classification tasks, or half-and-half. These three partitions help us to understand the influence brought by different task distribution in $\mathcal{T}_{train}$. These experiments will also help us to examine the capability of a CrossFit system's task-level generalization across drastically different task formats.

The remaining four partitions still focus on cross-task generalization, but in a finer granularity: seen and unseen tasks are still in the same category, but not the same sub-category. For example, Partition 3.1 has 57 non-NLI classification tasks as $\mathcal{T}_{train}$, and 8 NLI tasks as $\mathcal{T}_{test}$. These partitions help us to understand whether cross-task generalization in this finer granularity is easier for model to acquire.

5 Methods to CrossFit

We use BART-Base (Lewis et al., 2020) as the text-to-text transformer for our initial analysis in the CrossFit setup.[5] We compare the following three methods.

Direct Fine-tuning. This serves as the basic baseline method for the CrossFit challenge, which does not make use of the training or development tasks $(\mathcal{T}_{train}, \mathcal{T}_{dev})$ at all. For each task

[5]We plan to extend to T5 (non-multitask) models in our future version, which share similar techniques as BART.

No.	Shorthand	$\mathcal{T}_{train}$	$\mathcal{T}_{dev}$	$\mathcal{T}_{test}$	ARG(Multi,$\mathcal{T}_{test}$)	ARG(Meta,$\mathcal{T}_{test}$)	Details
1	Random	120	20	20	35.06%	28.50%	Fig. 4(a)
2.1	45cls	45 cls.	10 cls.	10 cls.	11.68%	9.37%	Fig. 6
2.2	23cls+22non-cls	23 cls. + 22 non-cls.	10 cls.	10 cls.	11.82%	9.69%	
2.3	45non-cls	45 non-cls.	10 cls.	10 cls.	11.91%	9.33%	
3.1	Held-out-NLI	57 non-NLI cls.	/	8 NLI	16.94%	12.30%	Fig. 4(b)
3.2	Held-out-Para	61 non-Paraphrase cls.	/	4 Para. Iden.	18.21%	17.90%	Fig. 4(c)
4.1	Held-out-MRC	42 non-MRC QA	/	9 MRC	32.81%	27.28%	Fig. 4(d)
4.2	Held-out-MCQA	29 non-MC QA	/	22 MC QA	12.20%	4.69%	Fig. 4(e)

Table 1: Details about ($\mathcal{T}_{train}$,$\mathcal{T}_{dev}$,$\mathcal{T}_{test}$) splits used in the study, and their results. "cls." stands for "classification", "Para. Iden." stands for "paraphrase identification", "MRC" for "machine reading comprehension" and "MCQA" for "multiple-choice QA".

$T \in \mathcal{T}_{test}$, we directly fine-tune the BART-Base model with its $\mathcal{D}_{train}$, tune the hyper-parameters on the $\mathcal{D}_{dev}$, and assess its performance on T with the test dataset $\mathcal{D}_{test}$. Note that this method does nothing in the first *upstream learning* stage (§3.3), and thus an effective method to the CROSSFIT challenge should have better performance on testing tasks than it. Therefore, we choose to use the performance of direct fine-tuning as the base for computing **ARG** (§3.4) scores of other CROSSFIT approaches.

Multi-task Learning. A straightforward yet effective method is to combine the data[6] in the training tasks to learn a multi-task model, before fine-tuning it on each test task. Specifically, we gather source-target examples for all tasks in $\mathcal{T}_{train}$ and fine-tune the BART-Base model with these examples. Then we use the resulting checkpoint as initialization and perform the same procedure in "direct fine-tuning" for each test task T in $\mathcal{T}_{test}$. The performance gain over the *direct fine-tuning* is thus used for computing its overall ARG score.

Meta-Learning. We use MAML (Finn et al., 2017), a representative meta-learning approach, which trains the model to adapt fast to new tasks. In MAML training, we iterate through tasks in $\mathcal{T}_{train}$ to update the model. For each train task $(\mathcal{D}_{train}, \mathcal{D}_{dev})$, we first sample a support batch $\mathcal{B}_{support}$ from $\mathcal{D}_{train}$ and a query batch $\mathcal{B}_{query}$ from $\mathcal{D}_{dev}$. We use f_θ to denote the text-to-text model with parameters θ. Using $\mathcal{B}_{support}$, we first compute the updated parameters θ' with gradient descent (*i.e.*, the inner loop). Due to the size of pre-trained text-to-text models, we use one gradient update in the inner loop.

$$\theta' = \theta - \alpha \nabla_\theta \mathcal{L}(f_\theta, \mathcal{B}_{support}). \quad (1)$$

Then we apply the updated text-to-text model $f_{\theta'}$ to $\mathcal{B}_{query}$, and do one step of meta-optimization (*i.e.*, the outer loop),

$$\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}(f_{\theta'}, \mathcal{B}_{query}). \quad (2)$$

After the meta-learning stage, we use "direct fine-tuning" for each task in $\mathcal{T}_{test}$, similar to the practice in multi-task learning.

6 Empirical Analysis

We list the ARG results in Table 1 and we plot the performance of each test task in each partition in Fig. 4 and Fig. 6. We aim to interpret the results and answer the research questions we raised.

Q1. Do upstream learning methods help address the CROSSFIT challenge? From Table 1, we observe that, on average, both upstream learning methods (i.e., *multi-task learning* and *meta-learning*) are helpful — both ARG scores are positive, meaning that they are better than *direct fine-tuning* (ARG=0%). In addition, we have the following observations:

(1) There are a few cases with negative performance gain, such as `Glue-COLA` (measuring linguistic acceptability) and `Domain Crawl` (separating domain names into tokens) in the setting with *Random* train/test split. For `Glue-COLA`, similar observations are reported by (Pruksachatkun et al., 2020) in an intermediate-task transfer learning setting, where the authors conjecture *catastrophic forgetting* of the masked language modeling (MLM) tasks may be the cause. The BART model that we use in our study uses denoising pre-training

[6]Both $\mathcal{D}_{train}$ and $\mathcal{D}_{dev}$ are used, as $\mathcal{D}_{dev}$ is used for gradient updates in meta-learning algorithm. We do so to make sure that the data access for the two methods is fair.

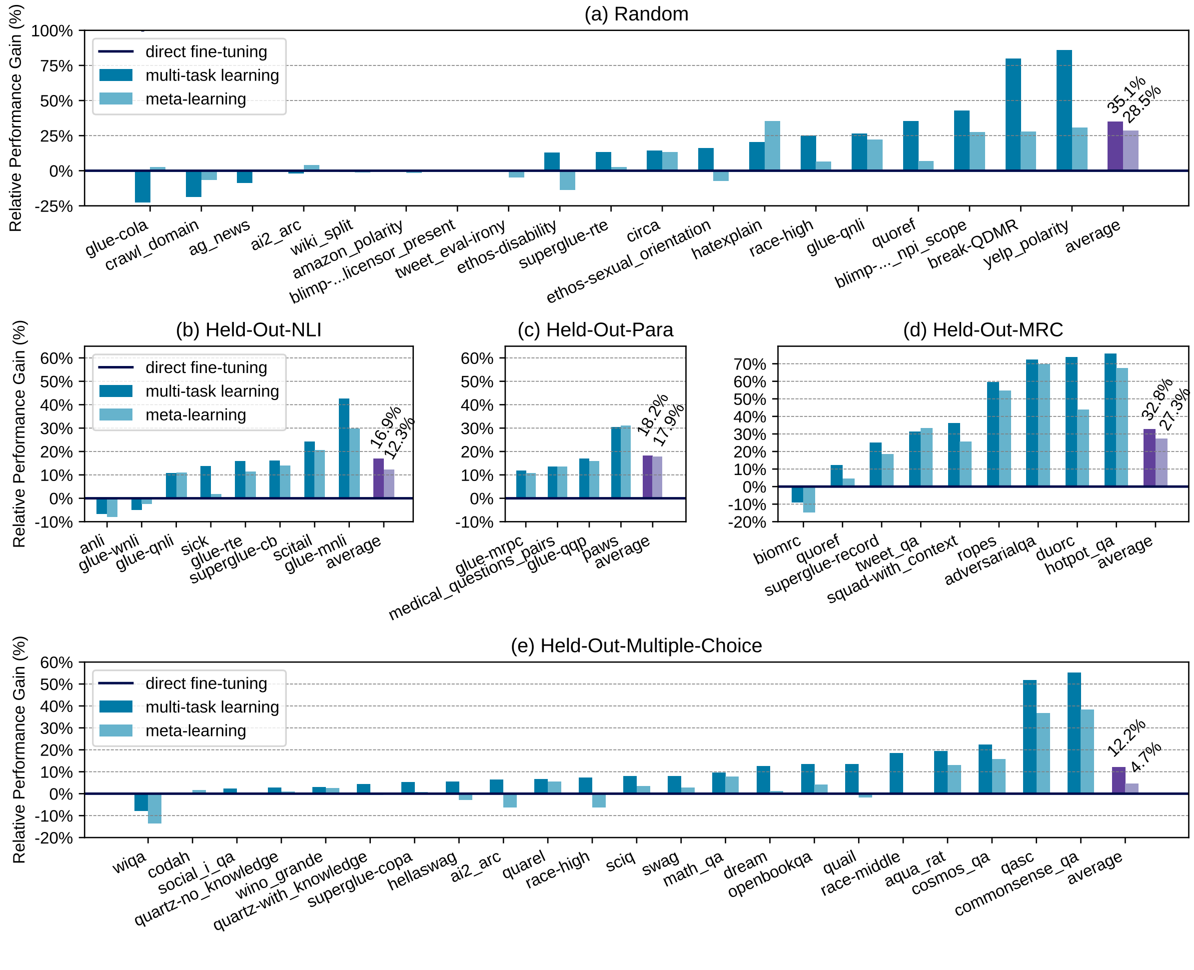

Figure 4: Experimental results for the CrossFit challenge with different task partitions. The details of each partition is shown in Table 1. Relative performance gain is computed based on the results of *direct fine-tuning*.

objective, a variant of MLM. Intuitively, `Domain Crawl` is also one of the most similar tasks to denoising in all test tasks, which further supports this conjecture. We thus conclude that for test tasks that resemble pre-training objectives (e.g., MLM), upstream learning could hurt performance due to the *catastrophic forgetting* phenomena.

(2) The performance gain obtained with the two upstream learning methods are correlated with each other — i.e., tasks that benefit from multi-task learning is likely to also benefit from meta-learning. For the *Random partition*, the *Spearman Correlation* between the improvement brought by multi-task learning and meta-learning is 0.66, with p value equals to 0.0015. This suggests that the two methods, while being significantly different, are capturing similar inductive bias from $\mathcal{T}_{train}$.

(3) Surprisingly, the multi-task learning method generally outperforms the MAML method, even though MAML is designed for fast adaptation to unseen tasks, a similar objective to our CrossFit Challenge. We conjecture there are two possible reasons: a) we suspect MAML is not used to its full extend (e.g., we use only one inner loop update), due to computation constraints; b) alternatively, MAML may struggle to learn from $\mathcal{T}_{train}$ that contains highly-diverse tasks (Yu et al., 2020). We leave further analysis as future work, and we believe it is promising to improve the performance by applying memory-efficient approaches or customized upstream learning algorithms.

Q2. How does the distribution in $\mathcal{T}_{train}$ influence the performance on unseen tasks? To study this, we first look at the tasks that appear in the $\mathcal{T}_{test}$ of more than one partitions. For example, `AI2_ARC` and `Race-High` are in the $\mathcal{T}_{test}$ of both *Random* partition and *Held-out-MCQA* partition. We present the results in Table 2. The performance of these tasks vary when different $\mathcal{T}_{train}$ sets are used. Notably, we observe significant performance drop with *Held-out-MCQA* par-

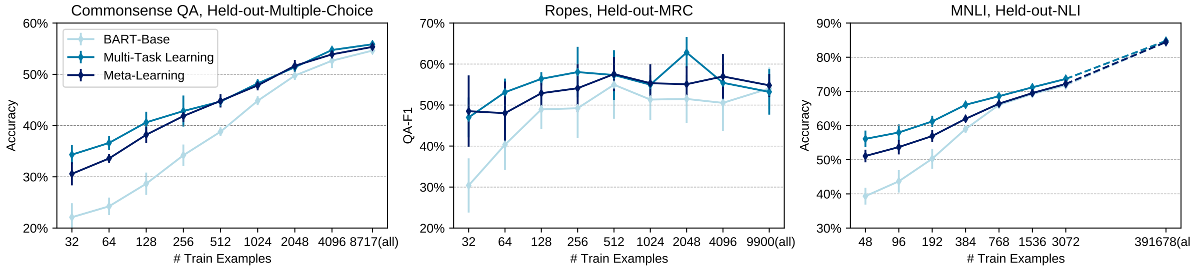

Figure 5: Performance comparisons in medium and high-resource scenarios. Benefits brought by upstream learning lasts in medium-resource scenarios.

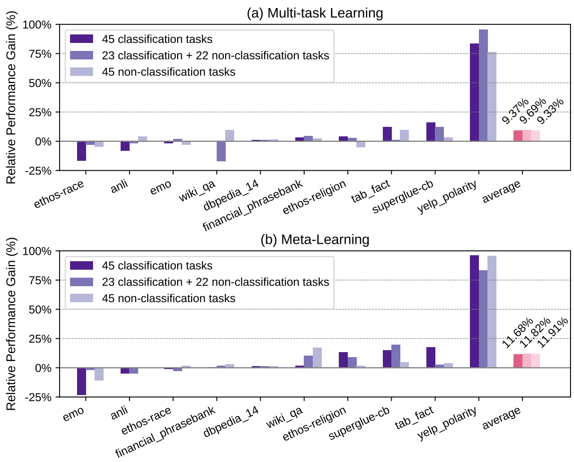

Figure 6: Performance comparison for the controlled experiment on Partition 2.1-2.3. $\mathcal{T}_{test}$ is a fixed set of 10 classification tasks, while $\mathcal{T}_{train}$ varies.

Test Task	Partition	Δ_{multi}	Δ_{meta}
Glue-QNLI	Random	15.89%	11.55%
	Held-Out-NLI	10.88%	10.94%
AI2_ARC	Random	1.30%	4.22%
	Held-Out-MCQA	6.49%	−6.22%
Race-High	Random	26.71%	6.59%
	Held-Out-MCQA	7.27%	−6.28%
QuoRef	Random	25.47%	3.99%
	Held-Out-MRC	12.25%	4.64%

Table 2: Performance comparison of test task performance when different $\mathcal{T}_{train}$ sets are used in upstream learning. See text in Q2 for in-depth analysis.

tition and meta-learning. We suspect this is due to the smaller size of $\mathcal{T}_{train}$ in *Held-out-MCQA* partition, as 22 QA tasks may not be sufficient for the meta-learning method to capture task-level generalization ability, especially when the train and test tasks have different formats (non-MCQA vs. MCQA). Apart from that, we have not found consistent patterns of what type of $\mathcal{T}_{train}$ lead to better performance for a specific test task.

We also conduct a set of controlled experiments with *Partition 2.1-2.3*, where $\mathcal{T}_{test}$ is a fixed set of classification tasks, and $\mathcal{T}_{train}$ varies. The performance analysis is plotted in Fig. 6. Ideally, we would expect upstream learning with all classification tasks (*Partition 2.1*) to achieve the best performance, while upstream learning with all non-classification tasks (*Partition 2.3*) to be the worst. However, the three partitions achieved comparable improvement in terms of the ARG score. Meanwhile, we observe several counter-intuitive cases: `ANLI` benefits most from *Partition 2.3* (all non-classification tasks) and least from *Partition 2.1* with multi-task learning, and similarly for WikiQA with meta-learning.[7]

Firstly, it is encouraging that non-classification tasks and classification tasks are equivalently helpful in the controlled experiment, demonstrating that acquiring cross-task generalization is feasible and promising. Yet, the two counter-intuitive cases suggest that we still lack clear understanding of these upstream learning methods, and our conventional perception about task affinity may not align with how models learn during upstream learning: selecting $\mathcal{T}_{train}$ tasks that have similar task format as the test task may not be an optimal solution. We believe that selecting appropriate $\mathcal{T}_{train}$ to learn for a target set of tasks is an interesting open problem. In addition, a more thorough investigation for the inner mechanism of upstream learning should be obtained by extending our study.

Q3. Does improved few-shot learning ability last when more data is available? We observe significant improvement for CommonsenseQA in *Held-out-Multiple-Choice* setting (Δ_{multi}=55.19% / Δ_{meta}=38.30%), ROPES in *Held-out-MRC* setting (Δ_{multi}=59.59% / Δ_{meta}=54.58%), and MNLI in *Held-out-NLI* setting (Δ_{multi}=42.61%

[7] We formulate WikiQA as a classification task to determine whether an answer is correct.

/ Δ_{meta}=29.87%). We further take these initialization and conduct experiments in medium and high-resource scenarios. That is, we randomly sample $\{32, 64, \ldots, 4096\}$ examples from these three datasets, and use them as $\mathcal{D}_{train}$. We then sample a $\mathcal{D}_{dev}$ which has the same size as $\mathcal{D}_{train}$, or has the size of 1024 if $|\mathcal{D}_{train}| > 1024$. We also try using the full dataset.[8] The performance of these settings is shown in Fig. 5. From the results we see that the benefits brought by upstream learning methods extend into medium resource cases with up to 2048 training examples. For Commonsense QA, checkpoints from upstream learning outperform direct fine-tuning significantly, even when the full dataset is used (Multi: $p = 0.01$ / Meta: $p = 0.07$). This generalization ability is particularly useful when users continue to collect more data to improve downstream performance.

7 Conclusion and Future Work

In this paper, we study the problem of building better few-shot learners via acquiring cross-task generalization ability from diverse NLP tasks. Towards our goal, we introduce the CROSSFIT Challenge, an task setup that standardizes the training pipeline, data access and evaluation protocol. We also present NLP Few-shot Gym, a repository of 160 diverse few-shot NLP tasks, to support CROSSFIT learning in different scenarios. We empirically demonstrated that cross-task generalization can be acquired via multi-task learning and meta-learning; confirmed that the selection of seen tasks would influence the few-shot performance on unseen tasks; and observed that the performance gain in few-shot scenarios last in medium-resource scenarios.

Our work focuses on cross-task generalization, which is non-conflicting to few-shot fine-tuning methods that focus on instance-level generalization; combining these two and check whether they're complementary to each other would be an interesting future direction. We also hope the CROSSFIT Challenge and NLP Few-shot Gym can serve as the testbed for many interesting "meta-problems", such as (1) learning to generate prompt for diverse task formats and further improve learning efficiency; (2) learning to select appropriate source tasks to learn from during upstream learning; (3) learning to accumulate knowledge and avoid catastrophic forgetting in an continual learning setup.

[8] We do five random samples of 1024 examples as $\mathcal{D}_{dev}$ and use the remaining examples in the original train set as $\mathcal{D}_{train}$. We use the original dev set for testing.

Acknowledgment

We thank authors and crowd-workers of all datasets used in our study. We thank huggingface datasets team for making datasets more accessible. We thank members of INK Lab at USC for their valuable feedback on this project.

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.

Trapit Bansal, Rishikesh Jha, and Andrew McCallum. 2020a. Learning to few-shot learn across diverse natural language classification tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5108–5123, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020b. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534, Online. Association for Computational Linguistics.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4. Venice.

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.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

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 Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

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.

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.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020a. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. 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. 2020b. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

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*, 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, 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 *International Conference on Learning Representations*.

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*, abs/1704.05179.

Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proc. 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 Languages 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, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Tianyu Gao, A. Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *ArXiv*, abs/2012.15723.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. 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 6250–6255, Hong Kong, China. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.

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.

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.

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.

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*.

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.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

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.

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.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

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 *AAAI*.

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.

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*, 7:453–466.

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.

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.

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.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 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.

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.

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-2014)*, pages 216–223, Reykjavik, Iceland. European Languages Resources Association (ELRA).

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.

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 arXiv:2012.10289*.

Julian McAuley and J. Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. *Proceedings of the 7th ACM conference on Recommender systems*.

Bryan McCann, N. Keskar, Caiming Xiong, and R. Socher. 2018. The natural language decathlon: Multitask learning as question answering. *ArXiv*, abs/1806.08730.

R. Thomas McCoy, Erin Grant, Paul Smolensky, Thomas L. Griffiths, and Tal Linzen. 2020. Universal linguistic inductive biases via meta-learning. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

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 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3458–3465, New York, NY, USA. Association for Computing Machinery.

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.

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

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.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

A. Othman and M. Jemni. 2012. English-asl gloss parallel corpus 2012: Aslg-pc12.

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.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

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

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

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.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. 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.

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. 2020a. Exploiting cloze questions for few-shot text classification and natural language inference. *Computing Research Repository*, arXiv:2001.07676.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *Computing Research Repository*, arXiv:2009.07118.

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.

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.

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.

Derek Tam, R. R. Menon, M. Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. *ArXiv*, abs/2103.11955.

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, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. 2020. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*.

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.

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.

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.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239, Online. 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.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. 2020. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 1094–1100. PMLR.

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*, abs/1810.12885.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting few-sample {bert} fine-tuning. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

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.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, 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.

A Selected Tasks in NLP Few-shot Gym

Table 3: Tasks in NLP Few-shot Gym.

Task Name	Ontology	Reference
acronym_identification	other	Pouran Ben Veyseh et al. 2020
ade_corpus_v2-classification	cls/other	Gurulingappa et al. 2012
ade_corpus_v2-dosage	other/slot filling	Gurulingappa et al. 2012
ade_corpus_v2-effect	other/slot filling	Gurulingappa et al. 2012
adversarialqa	qa/machine reading comprehension	Bartolo et al. 2020
aeslc	cg/summarization	Zhang and Tetreault 2019
ag_news	cls/topic	Gulli (link)
ai2_arc	qa/multiple-choice qa	Clark et al. 2018
amazon_polarity	cls/sentiment analysis	McAuley and Leskovec 2013
anli	cls/nli	Nie et al. 2020
app_reviews	other/regression	Missing
aqua_rat	qa/multiple-choice qa	Ling et al. 2017
art (abductive nli)	other	Bhagavatula et al. 2020
aslg_pc12	other	Othman and Jemni 2012
biomrc	qa/machine reading comprehension	Pappas et al. 2020
blimp-anaphor_gender_agreement	other/linguistic phenomenon	Warstadt et al. 2020
blimp-anaphor_number_agreement	other/linguistic phenomenon	Warstadt et al. 2020
blimp-determiner_noun_agreement_with_adj_irregular_1	other/linguistic phenomenon	Warstadt et al. 2020
blimp-ellipsis_n_bar_1	other/linguistic phenomenon	Warstadt et al. 2020
blimp-ellipsis_n_bar_2	other/linguistic phenomenon	Warstadt et al. 2020
blimp-existential_there_quantifiers_1	other/linguistic phenomenon	Warstadt et al. 2020
blimp-irregular_past_participle_adjectives	other/linguistic phenomenon	Warstadt et al. 2020
blimp-sentential_negation_npi_licensor_present	other/linguistic phenomenon	Warstadt et al. 2020
blimp-sentential_negation_npi_scope	other/linguistic phenomenon	Warstadt et al. 2020
blimp-wh_questions_object_gap	other/linguistic phenomenon	Warstadt et al. 2020
boolq	qa/binary	Clark et al. 2019
break-QDMR	other	Wolfson et al. 2020
break-QDMR-high-level	other	Wolfson et al. 2020
circa	cls/other	Louis et al. 2020
climate_fever	cls/fact checking	Diggelmann et al. 2020
codah	qa/multiple-choice qa	Chen et al. 2019
common_gen	other	Lin et al. 2020b
commonsense_qa	qa/multiple-choice qa	Talmor et al. 2019
cos_e	other/generate explanation	Rajani et al. 2019
cosmos_qa	qa/multiple-choice qa	Huang et al. 2019
crawl_domain	other	Zhang et al. 2020
crows_pairs	other	Nangia et al. 2020
dbpedia_14	cls/topic	Lehmann et al. 2015
definite_pronoun_resolution	other	Rahman and Ng 2012
discovery	cls/other	Sileo et al. 2019
dream	qa/multiple-choice qa	Sun et al. 2019
duorc	qa/machine reading comprehension	Saha et al. 2018
e2e_nlg_cleaned	other	Dušek et al. 2020, 2019
eli5-askh	qa/long-form qa	Fan et al. 2019
eli5-asks	qa/long-form qa	Fan et al. 2019
eli5-eli5	qa/long-form qa	Fan et al. 2019
emo	cls/emotion	Chatterjee et al. 2019
emotion	cls/emotion	Saravia et al. 2018
empathetic_dialogues	cg/dialogue	Rashkin et al. 2019
ethos-directed_vs_generalized	cls/hate speech detection	Mollas et al. 2020
ethos-disability	cls/hate speech detection	Mollas et al. 2020
ethos-gender	cls/hate speech detection	Mollas et al. 2020
ethos-national_origin	cls/hate speech detection	Mollas et al. 2020
ethos-race	cls/hate speech detection	Mollas et al. 2020
ethos-religion	cls/hate speech detection	Mollas et al. 2020
ethos-sexual_orientation	cls/hate speech detection	Mollas et al. 2020
financial_phrasebank	cls/sentiment analysis	Malo et al. 2014
freebase_qa	qa/closed-book qa	Jiang et al. 2019
gigaword	cg/summarization	Napoles et al. 2012
glue-cola	cls/other	Warstadt et al. 2019
glue-mnli	cls/nli	Williams et al. 2018
glue-mrpc	cls/paraphrase	Dolan and Brockett 2005
glue-qnli	cls/nli	Rajpurkar et al. 2016
glue-qqp	cls/paraphrase	(link)
glue-rte	cls/nli	Dagan et al. 2005; Bar-Haim et al. 2006 Giampiccolo et al. 2007; Bentivogli et al. 2009
glue-sst2	cls/sentiment analysis	Socher et al. 2013
glue-wnli	cls/nli	Levesque et al. 2012
google_wellformed_query	cls/other	Faruqui and Das 2018
hate_speech18	cls/hate speech detection	de Gibert et al. 2018
hate_speech_offensive	cls/hate speech detection	Davidson et al. 2017
hatexplain	cls/hate speech detection	Mathew et al. 2020
health_fact	cls/fact checking	Kotonya and Toni 2020
hellaswag	qa/multiple-choice qa	Zellers et al. 2019
hotpot_qa	qa/machine reading comprehension	Yang et al. 2018
imdb	cls/sentiment analysis	Maas et al. 2011
jeopardy	qa/closed-book qa	(link)
kilt_ay2	other/entity linking	Hoffart et al. 2011
kilt_fever	cls/fact checking	Thorne et al. 2018

Continued on next page

Task Name	Ontology	Reference
kilt_hotpotqa	qa/closed-book qa	Yang et al. 2018
kilt_nq	qa/closed-book qa	Kwiatkowski et al. 2019
kilt_trex	qa/closed-book qa	Elsahar et al. 2018
kilt_wow	cg/dialogue	Dinan et al. 2019
kilt_zsre	qa/closed-book qa	Levy et al. 2017
lama-conceptnet	qa/closed-book qa	Petroni et al. 2019, 2020
lama-google_re	qa/closed-book qa	Petroni et al. 2019, 2020
lama-squad	qa/closed-book qa	Petroni et al. 2019, 2020
lama-trex	qa/closed-book qa	Petroni et al. 2019, 2020
liar	cls/fact checking	Wang 2017
limit	other	Manotas et al. 2020
math_qa	qa/multiple-choice qa	Amini et al. 2019
mc_taco	qa/binary	Zhou et al. 2019
medical_questions_pairs	cls/paraphrase	McCreery et al. 2020
mocha	other/regression	Chen et al. 2020a
multi_news	cg/summarization	Fabbri et al. 2019
numer_sense	qa/closed-book qa	Lin et al. 2020a
onestop_english	cls/other	Vajjala and Lučić 2018
openbookqa	qa/multiple-choice qa	Mihaylov et al. 2018
paws	cls/paraphrase	Zhang et al. 2019
piqa	other	Bisk et al. 2020
poem_sentiment	cls/sentiment analysis	Sheng and Uthus 2020
proto_qa	other	Boratko et al. 2020
qa_srl	other	He et al. 2015
qasc	qa/multiple-choice qa	Khot et al. 2020
quail	qa/multiple-choice qa	Rogers et al. 2020
quarel	qa/multiple-choice qa	Tafjord et al. 2019a
quartz-no_knowledge	qa/multiple-choice qa	Tafjord et al. 2019b
quartz-with_knowledge	qa/multiple-choice qa	Tafjord et al. 2019b
quoref	qa/machine reading comprehension	Dasigi et al. 2019
race-high	qa/multiple-choice qa	Lai et al. 2017
race-middle	qa/multiple-choice qa	Lai et al. 2017
reddit_tifu-title	cg/summarization	Kim et al. 2019
reddit_tifu-tldr	cg/summarization	Kim et al. 2019
ropes	qa/machine reading comprehension	Lin et al. 2019
rotten_tomatoes	cls/sentiment analysis	Pang and Lee 2005
samsum	cg/summarization	Gliwa et al. 2019
scicite	cls/other	Cohan et al. 2019
sciq	qa/multiple-choice qa	Welbl et al. 2017
scitail	cls/nli	Khot et al. 2018
search_qa	qa/closed-book qa	Dunn et al. 2017
sick	cls/nli	Marelli et al. 2014
sms_spam	cls/other	Almeida et al. 2011
social_i_qa	qa/multiple-choice qa	Sap et al. 2019
spider	cg/other	Yu et al. 2018
squad-no_context	qa/closed-book qa	Rajpurkar et al. 2016
squad-with_context	qa/machine reading comprehension	Rajpurkar et al. 2016
superglue-cb	cls/nli	de Marneffe et al. 2019
superglue-copa	qa/multiple-choice qa	Gordon et al. 2012
superglue-multirc	qa/multiple-choice qa	Khashabi et al. 2018
superglue-record	qa/machine reading comprehension	Zhang et al. 2018
superglue-rte	cls/nli	Dagan et al. 2005; Bar-Haim et al. 2006 Giampiccolo et al. 2007; Bentivogli et al. 2009
superglue-wic	cls/other	Pilehvar and Camacho-Collados 2019
superglue-wsc	cls/other	Levesque et al. 2012
swag	qa/multiple-choice qa	Zellers et al. 2018
tab_fact	cls/fact checking	Chen et al. 2020b
trec	cls/other	Li and Roth 2002; Hovy et al. 2001
trec-finegrained	cls/other	Li and Roth 2002; Hovy et al. 2001
tweet_eval-emoji	cls/emotion	Barbieri et al. 2020
tweet_eval-emotion	cls/emotion	Barbieri et al. 2020
tweet_eval-hate	cls/emotion	Barbieri et al. 2020
tweet_eval-irony	cls/emotion	Barbieri et al. 2020
tweet_eval-offensive	cls/emotion	Barbieri et al. 2020
tweet_eval-sentiment	cls/emotion	Barbieri et al. 2020
tweet_eval-stance_abortion	cls/emotion	Barbieri et al. 2020
tweet_eval-stance_atheism	cls/emotion	Barbieri et al. 2020
tweet_eval-stance_climate	cls/emotion	Barbieri et al. 2020
tweet_eval-stance_feminist	cls/emotion	Barbieri et al. 2020
tweet_eval-stance_hillary	cls/emotion	Barbieri et al. 2020
tweet_qa	qa/machine reading comprehension	Xiong et al. 2019
web_questions	qa/closed-book qa	Berant et al. 2013
wiki_auto	cls/other	Jiang et al. 2020
wiki_bio	cg/other	Lebret et al. 2016
wiki_qa	cls/other	Yang et al. 2015
wiki_split	cg/other	Botha et al. 2018
wikisql	cg/other	Zhong et al. 2017
wino_grande	qa/multiple-choice qa	Sakaguchi et al. 2020
wiqa	qa/multiple-choice qa	Tandon et al. 2019
xsum	cg/summarization	Narayan et al. 2018
yahoo_answers_topics	cls/topic	(link)
yelp_polarity	cls/sentiment analysis	Zhang et al. 2015; (link)
yelp_review_full	other/regression	Zhang et al. 2015; (link)

B Details about Task Partition

B.1 Partition 1. Random

```
{
    "train": [’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’, ’
        definite_pronoun_resolution’, ’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’, ’jeopardy’, ’biomrc’, ’superglue-multirc’, ’ethos-race’, ’eli5-askh’,
        ’glue-qqp’, ’paws’, ’ethos-directed_vs_generalized’, ’glue-sst2’, ’mocha’, ’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’, ’yelp_review_full’, ’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’, ’wiki_auto’, ’mc_taco’, ’blimp-
        wh_questions_object_gap’, ’hotpot_qa’, ’emo’, ’kilt_nq’, ’kilt_trex’, ’quartz-with_knowledge’, ’
        dbpedia_14’, ’yahoo_answers_topics’, ’app_reviews’, ’superglue-copa’, ’blimp-
        anaphor_gender_agreement’, ’hate_speech18’, ’gigaword’, ’multi_news’, ’aeslc’, ’quail’],
    "dev": [’cos_e’, ’kilt_fever’, ’eli5-asks’, ’trec’, ’eli5-eli5’, ’art’, ’empathetic_dialogues’, ’
        tweet_qa’, ’wikisql’, ’lama-trex’, ’tweet_eval-stance_hillary’, ’discovery’, ’tweet_eval-emotion’,
        ’liar’, ’wiki_bio’, ’dream’, ’ade_corpus_v2-classification’, ’health_fact’, ’samsum’, ’
        financial_phrasebank’],
    "test": [’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’, ’break-QDMR’, ’
        crawl_domain’, ’freebase_qa’, ’glue-qnli’, ’hatexplain’, ’ag_news’, ’circa’],
}
```

B.2 Partition 2.1. 45cls

```
{
    "train": ["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", "wiki_auto", "hate_speech_offensive", "superglue-wic", "google_wellformed_query",
        "tweet_eval-irony", "ethos-gender", "onestop_english", "trec", "rotten_tomatoes", "kilt_fever"],
    "dev": ["tweet_eval-stance_feminist", "ethos-national_origin", "tweet_eval-hate", "ag_news", "
        amazon_polarity", "hate_speech18", "poem_sentiment", "climate_fever", "medical_questions_pairs", "
        tweet_eval-stance_atheism"],
    "test": ["superglue-cb", "dbpedia_14", "wiki_qa", "emo", "yelp_polarity", "ethos-religion", "
        financial_phrasebank", "tab_fact", "anli", "ethos-race"],
}
```

B.3 Partition 2.2. 23cls+22non-cls

```
{
    "train": ["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"],
    "dev": ["tweet_eval-stance_feminist", "ethos-national_origin", "tweet_eval-hate", "ag_news", "
        amazon_polarity", "hate_speech18", "poem_sentiment", "climate_fever", "medical_questions_pairs", "
        tweet_eval-stance_atheism"],
    "test": ["superglue-cb", "dbpedia_14", "wiki_qa", "emo", "yelp_polarity", "ethos-religion", "
        financial_phrasebank", "tab_fact", "anli", "ethos-race"]
}
```

B.4 Partition 2.3. 45non-cls

```
{
```

```
    "train": ["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"],
    "dev": ["tweet_eval-stance_feminist", "ethos-national_origin", "tweet_eval-hate", "ag_news", "
         amazon_polarity", "hate_speech18", "poem_sentiment", "climate_fever", "medical_questions_pairs", "
         tweet_eval-stance_atheism"],
    "test": ["superglue-cb", "dbpedia_14", "wiki_qa", "emo", "yelp_polarity", "ethos-religion", "
         financial_phrasebank", "tab_fact", "anli", "ethos-race"]
}
```

B.5 Partition 3.1. Held-out-NLI

```
{
    "train": [
        "ade_corpus_v2-classification",
        "ag_news",
        "amazon_polarity",
        "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-mrpc",
        "glue-qqp",
        "glue-sst2",
        "google_wellformed_query",
        "hate_speech18",
        "hate_speech_offensive",
        "hatexplain",
        "health_fact",
        "imdb",
        "kilt_fever",
        "liar",
        "medical_questions_pairs",
        "onestop_english",
        "paws",
        "poem_sentiment",
        "rotten_tomatoes",
        "scicite",
        "sick",
        "sms_spam",
        "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_auto",
        "wiki_qa",
        "yahoo_answers_topics",
        "yelp_polarity"
    ],
    "dev": [],
    "test": ["anli", "glue-mnli", "glue-qnli", "glue-rte", "glue-wnli", "scitail", "sick", "superglue-cb"]
}
```

B.6 Partition 3.2. Held-out-Para

```
{
    "train": ["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_auto",
        "wiki_qa",
        "yahoo_answers_topics",
        "yelp_polarity"],
    "dev": [],
    "test": [
        "glue-mrpc",
        "glue-qqp",
        "medical_questions_pairs",
        "paws"
    ]
}
```

B.7 Partition 4.1. Held-out-MRC

```
{
    "train": [
        "ai2_arc",
        "aqua_rat",
        "boolq",
        "codah",
        "commonsense_qa",
        "cosmos_qa",
        "dream",
        "eli5-askh",
        "eli5-asks",
```

```
        "eli5-eli5",
        "freebase_qa",
        "hellaswag",
        "jeopardy",
        "kilt_hotpotqa",
        "kilt_nq",
        "kilt_trex",
        "kilt_zsre",
        "lama-conceptnet",
        "lama-google_re",
        "lama-squad",
        "lama-trex",
        "math_qa",
        "mc_taco",
        "numer_sense",
        "openbookqa",
        "qasc",
        "quail",
        "quarel",
        "quartz-no_knowledge",
        "quartz-with_knowledge",
        "race-high",
        "race-middle",
        "sciq",
        "search_qa",
        "social_i_qa",
        "squad-no_context",
        "superglue-copa",
        "superglue-multirc",
        "swag",
        "web_questions",
        "wino_grande",
        "wiqa"
    ],
    "dev": [],
    "test": [
        "adversarialqa",
        "biomrc",
        "duorc",
        "hotpot_qa",
        "quoref",
        "ropes",
        "squad-with_context",
        "superglue-record",
        "tweet_qa"
    ],
}
```

B.8 Partition 4.2. Held-out-MCQA

```
{
    "train": [
        "adversarialqa",
        "biomrc",
        "boolq",
        "duorc",
        "eli5-askh",
        "eli5-asks",
        "eli5-eli5",
        "freebase_qa",
        "hotpot_qa",
        "jeopardy",
        "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",
        "squad-with_context",
        "superglue-multirc",
        "superglue-record",
        "tweet_qa",
        "web_questions"
    ],
    "dev": [],
    "test": [
        "ai2_arc",
```

```
        "aqua_rat",
        "codah",
        "commonsense_qa",
        "cosmos_qa",
        "dream",
        "hellaswag",
        "math_qa",
        "openbookqa",
        "qasc",
        "quail",
        "quarel",
        "quartz-no_knowledge",
        "quartz-with_knowledge",
        "race-high",
        "race-middle",
        "sciq",
        "social_i_qa",
        "superglue-copa",
        "swag",
        "wino_grande",
        "wiqa"
    ]
}
```