Cross-Task Generalization Abilities of Large Language Models

Qinyuan Ye
University of Southern California
qinyuany@usc.edu

Abstract

Humans can learn a new language task efficiently with only few examples, by leveraging their knowledge and experience obtained when learning prior tasks. Enabling similar *cross-task generalization* abilities in NLP systems is fundamental for approaching the goal of general intelligence and expanding the reach of language technology in the future. In this thesis proposal, I will present my work on (1) benchmarking cross-task generalization abilities with diverse NLP tasks; (2) developing model architectures for improving cross-task generalization abilities; (3) analyzing and predicting the generalization landscape of current state-of-the-art large language models. Additionally, I will outline future research directions, along with preliminary thoughts on addressing them.

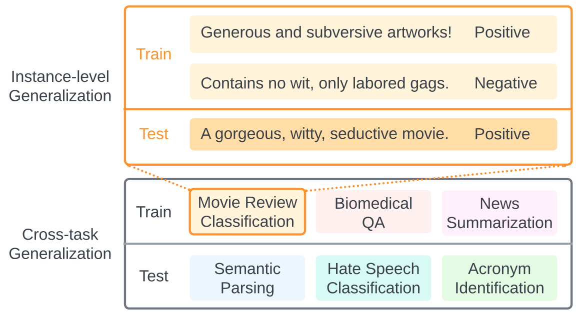

Figure 1: **Instance-level Generalization vs. Cross-task Generalization.** This thesis proposal advocates for the crucial role of cross-task generalization in NLP systems and presents my research efforts in this area.

1 Introduction

In recent years, large language models (LLMs) have greatly revolutionized natural language processing research, demonstrating remarkable capabilities in various natural language processing benchmarks (Devlin et al. 2019; Radford et al. 2019; Raffel et al. 2020; Brown et al. 2020, *inter alia*). As their capabilities have expanded, there has been a corresponding increase in their adoption. LLM-powered tools are now playing an essential role in daily activities, from translation and search engines, to personalized chatbots and tutors. Looking ahead, we can expect LLMs to be applied to a wider spectrum of downstream applications with increasing complexity and intricacy.

However, building these applications still requires extensive *task-specific* efforts. This involves data collection, model architecture modifications and training procedure design. Even with the most powerful LLMs, manual selection of in-context examples or prompt engineering is often required to fully unlock their performance.

From a *practical* perspective, these task-specific approaches lack scalability. Every new application in the future will demand repeating these tedious and costly processes. From a *research* perspective, achieving human-level performance on individual tasks through extensive data collection and engineering efforts falls short of the ideal general intelligence. A truly intelligent system should be able to "reuse previously acquired knowledge about a language and adapt to a new task quickly" (Yogatama et al., 2019; Linzen, 2020). Evaluating these systems based on their "skill-acquisition efficiency" (Chollet, 2019) becomes crucial in this context.

Existing work has approached the problem of learning efficiency by developing better few-shot learning algorithms, *e.g.*, re-formulating tasks into formats that resembles the pre-training objective (Schick and Schütze, 2020a,b). Such progress primarily focus on improving *instance-level generalization*, *i.e.*, how to better generalize from a few labeled instances to make predictions about new instances, *within the scope of one individual task*. From a broader perspective, human-like learning efficiency also benefits from *task-level generaliza-*

†Presented at NAACL 2024 Student Research Workshop (Thesis Proposal Track).

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 255–262
June 18, 2024 ©2024 Association for Computational Linguistics

tion, or *cross-task generalization* (Fig. 1). Humans accumulate their learning experience on previous seen tasks, so that when confronted with a novel task, we are able to grasp the essence of it quickly and learn it efficiently.

My research goal is to enable human-like adaptability and learning efficiency in NLP systems. I argue that achieving cross-task generalization is an essential building block for this goal. In the following, I will first revisit the background and prior works (§2). Next, I will introduce my contributions in three areas: (1) benchmarking cross-task generalization with diverse NLP tasks (§3.1); (2) developing new model architectures that not only improve cross-task generalization but also enhance interpretability (§3.2.1) and inference speed (§3.2.2). (3) analyzing the generalization landscape of LLMs and predicting their performance across different model families, model scales and tasks (§3.3). Finally, I will discuss future directions for my research, including (1) pushing the limits of in-context learning with various types of contexts, and (2) developing autonomous learning agents that can acquire their own learning materials (§4).

2 Background

Few-shot Fine-tuning. Pre-trained language models (e.g., BERT, Devlin et al. 2019) have demonstrated great few-shot learning ability via fine-tuning (Zhang et al., 2021). Schick and Schütze (2020a,b) proposed *pattern-exploiting training* (PET), which formulates text classification and NLI tasks into cloze questions that resemble the masked language modeling objective. PET can be further improved by incorporating demonstrations into the input (Gao et al., 2021); and by densifying the supervision signal with label conditioning (Tam et al., 2021). While successful, these approaches focus on instance-level generalization (Fig. 1), and different downstream tasks are learned in isolation. Our research work aims to boost few-shot learning ability on unseen tasks via acquiring cross-task generalization ability from seen tasks.

Few-shot In-Context Learning. In-context learning (ICL) is an alternative approach for few-shot learning by simply concatenating the few-shot examples and using them as a prompt before the inference example. Popularized by more recent language models like GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022), ICL allows models to learn from a few examples without any gradient updates and achieve competitive performance. While this approach works well for very large models, smaller models requires *meta-training* to gain similar capabilities (Chen et al., 2022; Min et al., 2022). Our research on cross-task generalization aligns more with the latter approach. However, the former approach remains relevant, as the next-token prediction objective during pre-training can be seen as a superset of language tasks, and ICL can be viewed as generalizing to unseen tasks at inference time.

Meta-learning in NLP. The goal of rapid task adaptation and cross-task generalization is closely related to the research field of *meta-learning*, or *learning to learn* (Schmidhuber, 1987). While widely explored in computer vision and robotics community (Yu et al., 2020; Triantafillou et al., 2020), meta-learning is relatively underexplored in NLP. Existing NLP research has primarily focused on applying meta-learning algorithms to a *narrow* distribution of tasks, *e.g.*, relation classification (Han et al., 2018; Gao et al., 2019), text classification (Dou et al., 2019; Bansal et al., 2020a,b), low-resource machine translation (Gu et al., 2018). Our work explores a more realistic scenario: learning from NLP tasks covering *diverse* formats, goals and domains. To emphasize our focus on task-level meta-learning, as opposed to cross-domain or cross-lingual meta-learning, we primarily adopt the term "cross-task generalization" in this work.

Unifying NLP Task Formats. Researchers have explored unifying the formats of different tasks, in order to better enable knowledge transfer, *e.g.*, DecaNLP (McCann et al., 2018), UFO-Entail (Yin et al., 2020) and EFL (Wang et al., 2021). Following T5 (Raffel et al., 2020), we adopt a unified text-to-text format that subsumes all text-based tasks of interest. Related to our work, UnifiedQA (Khashabi et al., 2020) examines the feasibility of training a general cross-format QA model with multi-task learning. Our work extends from these ideas, and we significantly scale the number of tasks to 160 to broaden the coverage, in hopes to build a general-purpose data-efficient learner.

3 Research Work

3.1 Benchmarking Cross-Task Generalization

To investigate and enable cross-task generalization abilities in large language models (LLMs), a suitable benchmark is essential as a starting point. In the following, we describe our efforts in building the CrossFit benchmark (Ye et al., 2021).

Problem Setting. We define a task T as a tuple of $(\mathcal{D}_{train}, \mathcal{D}_{dev}, \mathcal{D}_{test})$. Each set $\mathcal{D}$ is a set of annotated examples $\{(x_i, y_i)\}$ in text-to-text format. To benchmark cross-task generalization, we first gather a large repository of few-shot tasks $\mathcal{T}$, and partition them into three non-overlapping sets $\mathcal{T}_{train}$, $\mathcal{T}_{dev}$, $\mathcal{T}_{test}$. A method for this proposed setting is expected to first learn from $\mathcal{T}_{train}$ and perform necessary hyperparameter tuning with $\mathcal{T}_{dev}$ in an *upstream* learning stage; it is then evaluated on each task in $\mathcal{T}_{test}$ in an *downstream* learning stage.

Data. We use huggingface datasets library (Lhoest et al., 2021) and collect 160 tasks to formulate our task repository $\mathcal{T}$. They cover diverse formats (classification, multiple choice, etc.), goals (question answering, fact checking, etc.) and domains (biomedical, social media, etc.). We subsample the training sets for each task to simulate the few-shot setting (16 shots per class for classification tasks, 32 shots for other tasks). In our main experiments, we randomly partition $\mathcal{T}$ into ($\mathcal{T}_{train}$, $\mathcal{T}_{dev}$, $\mathcal{T}_{test}$). In later analysis, we also create partitions according to a task taxonomy we created for the 160 tasks (Fig. 2).

Experiments. For the upstream learning stage with $\mathcal{T}_{train}$, we compare simple multi-task learning and three meta-learning algorithms: (1) Model-Agnostic Meta-Learning (MAML; Finn et al. 2017), (2) the first-order variant of MAML, and (3) Reptile (Nichol et al., 2018), another memory-efficient, first-order meta-learning algorithm. After the upstream learning stage, we fine-tune the resulting models on each task in $\mathcal{T}_{test}$. We report the performance gains achieved by models trained with upstream learning compared to those trained without, expressed as the relative percentage increase.

Main Findings. **(1)** An upstream learning stage can improve the model's few-shot learning performance on unseen tasks. By aggregating results from all upstream learning methods and task partitions investigated, we find that the performance on 51.47% test tasks are significantly improved (>5% relative improvement compared to direct fine-tuning); 35.93% tasks are relatively unaffected (between $\pm 5\%$); and 12.60% tasks suffer from worse performance ($<-5\%$). We also find that the most straight-forward multi-task learning method outperforms more sophisticated meta-learning algo-

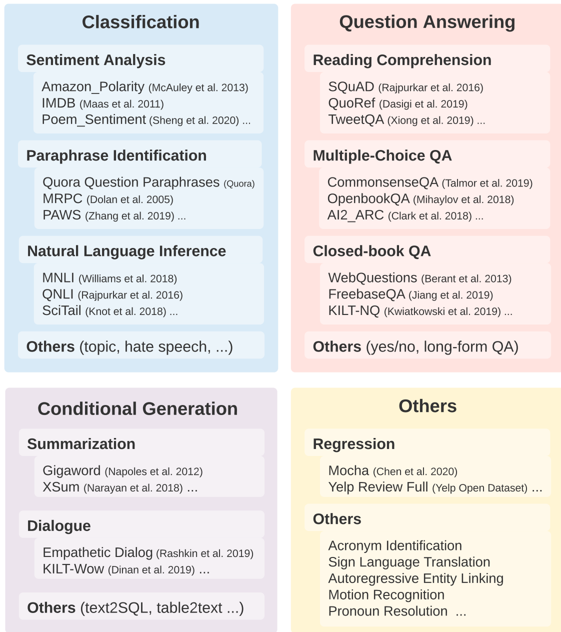

Figure 2: **Taxonomy of NLP tasks included in the CrossFit benchmark (§3.1).**

rithms. **(2)** The selection of tasks in the upstream learning stage plays an important role in performance on unseen tasks. Meanwhile, the transfer mechanism does not clearly align with our naive categorization of tasks based on task format (*e.g.*, classification, QA). For example, when controlling the composition of upstream tasks ($\mathcal{T}_{train}$) to be 100% classification, 100% non-classification, or 50%-50%, the average performance on unseen tasks are comparable. **(3)** We find that enlarging the size of D_{train} in upstream tasks does not necessitate better cross-task generalization. By enlarging D_{train} of upstream tasks by 8x, the downstream performance is improved by merely 4%.

3.2 Improved Modeling Techniques

3.2.1 Task-level Mixture-of-Experts

Our CrossFit work in §3.1 and recent work (Aghajanyan et al., 2021) suggest that training language models to multi-task on a diverse collection of NLP tasks is beneficial. The resulting model is not only better at handling seen tasks, but also better at adapting to unseen tasks in the few-shot setting. However, the potential of these multi-task models may be limited as the exact *same* set of weights is applied, and the *same* computation is executed, for very *different* tasks. Humans, on the other hand, develop modular skill sets and accumulate knowledge during learning, and can readily

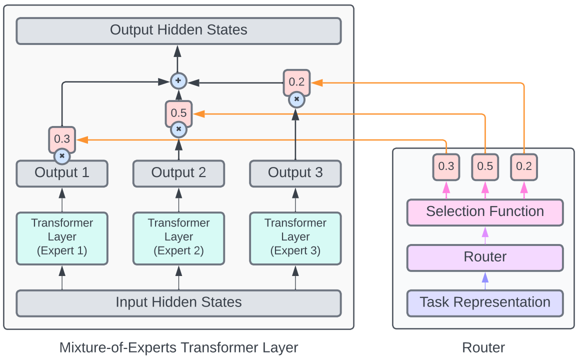

Figure 3: **Task-level Mixture-of-experts Transformer models used in §3.2.1. Right:** A router takes in a task representation and make decisions on expert selection. **Left:** the weighted sum of the outputs from each expert are considered the final output for this layer.

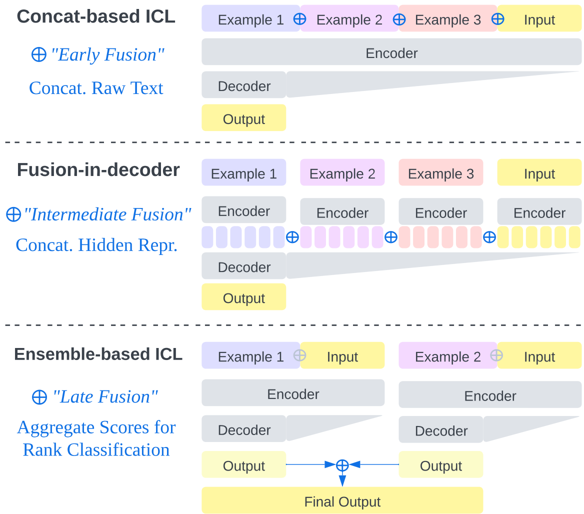

Figure 4: **Investigation on Fusion Methods for In-context Learning.** In §3.2.2, we compare different methods to incorporate examples for in-context learning. We term these as "fusion methods". ⊕ marks where and how fusion is implemented.

reuse and recompose only the necessary ones when facing a task. Although multi-task models may develop latent skills within their weights, we are interested in enabling this modular, skill-sharing process more explicitly.

A natural fit for our goal would be task-level mixture-of-expert models (Jacobs et al., 1991; Kudugunta et al., 2021), where the model computation is dependent on the task at hand. In our CrossTask-MoE work (Ye et al., 2022), we adapt and train such mixture-of-expert models in the cross-task generalization setting. Our model contains a collection of experts and a router that chooses from the experts. For a given task $T_k \in \mathcal{T}$, with k as its task index, the router first takes the task representation ($\mathbf{T}_k$) from a look-up embedding table ($\mathbf{T}$). The router network outputs a matrix $\mathbf{L} \in \mathbb{R}^{m\times n}$, where $\mathbf{L}_{i,j}$ represents the logits of using expert $E^{(i,j)}$ in layer i. $\mathbf{L}$ goes through a selection function f to normalize the routing decisions in each layer, resulting in a final decision matrix $\mathbf{D} \in \mathbb{R}^{m\times n}$. We then use the decision matrix $\mathbf{D}$ from the router to control the computation conducted by the experts. In layer i, given input hidden states $\mathbf{h}_{in}^{(i)}$, the output $\mathbf{h}_{out}^{(i)}$ would be the weighted sum of all experts in the layer, and the weights are specified in $\mathbf{D}_{i,\cdot}$, *i.e.*, $\mathbf{h}_{out}^{(i)} = \sum_{j=1}^{m} \mathbf{D}_{i,j} E^{(i,j)}(\mathbf{h}_{in}^{(i)})$.

We first conduct detailed ablations on different design choices of Task-level MoEs and converge to a final method. Our results suggest that training task-level mixture-of-experts can alleviate negative transfer and achieve better few-shot performance on unseen tasks. We find that these models help improve the average performance gain (ARG) metric by 2.6% when adapting to unseen tasks in the few-shot setting and by 5.6% in the zeroshot generalization setting. In our interpretability analysis, we find that the learned routing decisions and experts partially align with human categorization of NLP tasks – certain experts are strongly associated with extractive tasks, some with classification tasks, and some with tasks requiring world knowledge. By disabling these experts with high associations, performance will deteriorate significantly. In one extreme case, disabling 3 experts for the emotion classification task results in a dramatic drop in F1 score, from 82% to a mere 16%.

3.2.2 Fusion-in-Decoders for Efficient In-Context Learning

As previously described in §2, in-context learning (ICL) is a new way to perform few-shot learning without updating model weights, by concatenating a few demonstrations and preprending them before the test input. One limitation of in-context learning is that the concatenated demonstrations are often excessively long and induce additional computation costs. Inspired by fusion-in-decoder (FiD; Izacard and Grave 2021) models which efficiently aggregate passages and thus outperforms concatenation-based models in open-domain QA, we hypothesize that similar techniques can be applied to improve the efficiency and end-task performance of ICL.

In our FiD-ICL work (Ye et al., 2023a),

we present a comprehensive study on three methods—concatenation-based (early fusion), FiD (intermediate), and ensemble-based (late)—to aggregate few-shot examples in ICL. See Figure 4 for an illustration of these three methods. We adopt a cross-task generalization setup where a model is first trained to perform ICL on a mixture of tasks using one selected fusion method, then evaluated on held-out tasks for ICL (Sanh et al., 2022).

Results on 11 held-out tasks show that FiD-ICL matches or outperforms the other two fusion methods across three different model scales (250M, 800M, 3B). Notably, FiD-ICL, a gradient-free in-context learning method, narrows the performance gap between ICL and T-Few (Liu et al., 2022), a state-of-the-art few-shot fine-tuning method, to be less than 3%. Additionally, we show that FiD-ICL is 10x faster at inference time compared to concat-based and ensemble-based ICL, as we can pre-compute the representations of in-context examples and reuse them. FiD-ICL also enables scaling up to meta-training 3B-sized models, which would lead to out-of-memory errors with concat-based ICL when on an academic budget.

3.3 Modeling and Predicting the LLM Generalization Landscape

Because a large language model excels at one task, can we expect it to perform well on another task? Are there any patterns that govern how well state-of-the-art LLMs generalize across different tasks? To answer these questions, we use data-driven approaches to investigate the predictability of large language model capabilities across different tasks, model families, model scales and numbers of in-context examples (Ye et al., 2023b).

We investigate this question using experiment records from BIG-bench (BIG-bench authors, 2023), a collaborative benchmark that contains a diverse set of tasks contributed by the community, covering "problem from linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development, and beyond." We gather and carefully filter these records, yielding a total of 56k records which we use as the "dataset" for our analysis.

Through extensive experiments, we find that LLMs' performance on BIG-bench follows predictable patterns. In the default setting where we create train and test sets with random sampling, our best predictor, an MLP model, achieves an RMSE lower than 0.05 (*i.e.*, on average mis-predict by < 0.05 when the range is $[0, 1]$) and an R^2 greater than 95% (*i.e.*, explains more than 95% variance in the target variable). However, the predictor's performance is dependent on the assumptions of the train-test distribution. In a more challenging setting where we hold out the Cartesian product of complete model families (all model scales) and complete tasks (all numbers of shots), the predictor's performance decreases ($R^2 : 95\% \rightarrow 86\%$).

We further explore to what extent emergent abilities (Wei et al., 2022a) can be predicted, and how our performance prediction models can be used to create more efficient benchmarks for future LLMs.

4 Future Directions

Pushing the Limit of In-Context Learning. As an alternative to model fine-tuning, in-context learning has shown to be effective in adapting an LLM to perform novel tasks. Existing works on in-context learning mostly focus on conditioning on demonstrations of *one single task*. It is possible to break this convention by conditioning on diverse and heterogeneous contexts. For example, Pruksachatkun et al. (2020); Vu et al. (2020) highlight the benefits of intermediate task transfer in the fine-tuning paradigm. Revisiting this technique with in-context learning may help improve end-task performance and also enhance our understanding of in-context learning. Recent progresses on long-context LMs open up new opportunities for scaling not only the length, but also the diversity and composition of "contexts" for in-context learning, which we plan to investigate in the future.

From Data-Efficient Learners to Self-Sufficient Learners. So far in our efforts, the models are expected to perform few-shot learning when the few-shot training data are provided and fixed. A more ambitious goal will be to build intelligent systems that can acquire their own learning material and learn in the open-endedness. As the capabilities of LLMs continue to grow, they demonstrate agentic behaviors such as reasoning (Wei et al., 2022b), planning (Wang et al., 2023), tool use (Schick et al., 2023), self-refinement (Madaan et al., 2023), etc. All of these are also fundamental aspects of human learning processes. In the future, we plan to incorporate these latest advances into building an autonomous, self-sufficient learning agent capable of devising a learning plan, executing it, reflecting on its own limitations, and dynamically adjusting the plan throughout the course of learning.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic. 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.

BIG-bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

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

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.

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.

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

Tianyu Gao, Xu Han, 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.

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.

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.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th*

Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 874–880, Online. Association for Computational Linguistics.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*.

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.

Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. 2021. Beyond distillation: Task-level mixture-of-experts for efficient inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3577–3599, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

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

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

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *ArXiv*, abs/1803.02999.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Victor Sanh, Albert Webson, Colin Raffel, and Stephen Bach *et al.* 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc.

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.

Jürgen Schmidhuber. 1987. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. Ph.D. thesis, Technische Universität München.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, Online and Punta Cana, Dominican Republic. 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 *ICLR*.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Qinyuan Ye, Iz Beltagy, Matthew Peters, Xiang Ren, and Hannaneh Hajishirzi. 2023a. FiD-ICL: A fusion-in-decoder approach for efficient in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8158–8185, Toronto, Canada. Association for Computational Linguistics.

Qinyuan Ye, Harvey Fu, Xiang Ren, and Robin Jia. 2023b. How predictable are large language model capabilities? a case study on BIG-bench. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7493–7517, Singapore. Association for Computational Linguistics.

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

Qinyuan Ye, Juan Zha, and Xiang Ren. 2022. Eliciting and understanding cross-task skills with task-level mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2567–2592, Abu Dhabi, United Arab Emirates. 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.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, A. Lazaridou, Wang Ling, L. Yu, Chris Dyer, and P. Blunsom. 2019. Learning and evaluating general linguistic intelligence. *ArXiv*, abs/1901.11373.

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.

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*.