Exploring the Benefits of Training Expert Language Models over Instruction Tuning

Joel Jang [1] **Seungone Kim** [1] **Seonghyeon Ye** [1] **Doyoung Kim** [1] **Lajanugen Logeswaran** [2] **Moontae Lee** [2 3] **Kyungjae Lee** [2] **Minjoon Seo** [1]

Abstract

Recently, Language Models (LMs) instruction-tuned on multiple tasks, also known as multitask-prompted fine-tuning (MT), have shown the capability to generalize to unseen tasks. Previous work has shown that scaling the number of training tasks is the key component in making stronger MT LMs. In this work, we report an unexpected finding that an *expert* LM fine-tuned on just a single task can outperform an MT LM trained with 300+ different tasks on 11 different unseen datasets and on 13 datasets of the BIG-bench benchmark by a mean accuracy of 3.20% and 1.29%, respectively. This finding casts doubt on the previously held belief that simply scaling the number of tasks makes stronger MT LMs. Leveraging this finding, we further show that this distributed approach of training a separate expert LM per training task instead of a single MT LM for zero-shot inference possesses many benefits including (1) avoiding negative task transfer that often occurs during instruction tuning, (2) being able to continually learn new tasks without having to re-train on previous tasks to avoid catastrophic forgetting, and (3) showing *compositional* capabilities when merging individual experts together. The code is available at https://github.com/joeljang/ELM.

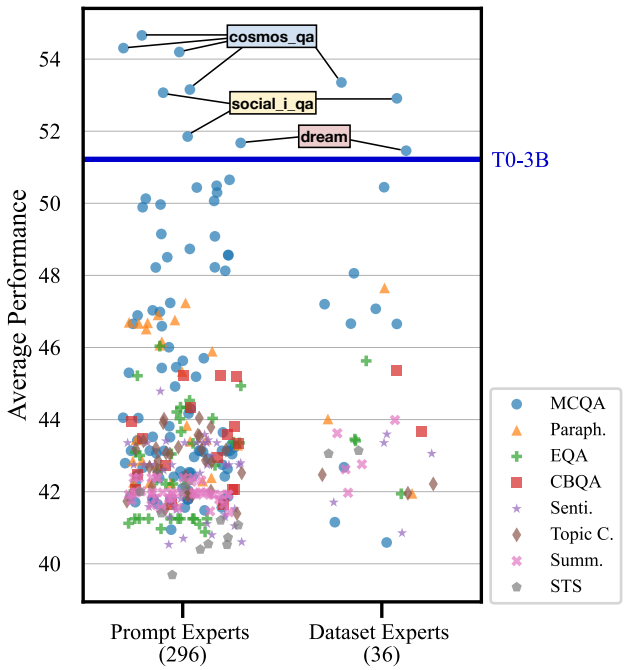

Figure 1. Mean accuracy performance of Expert LMs (each trained on a single task) on 11 unseen datasets compared to an instruction-tuned LM, T0-3B. Results show some Expert LMs surpassing T0-3B, challenging the commonly held belief that simply scaling the total number of training tasks is the key component to enhancing the capability of MT LMs.

1. Introduction

Recent works show pretrained Language Models (LMs) that have been fine-tuned on multiple tasks with instructions (prompted instances), also known as multitask-prompted fine-tuned LMs and referred to as MT LMs in this work, can generalize to unseen tasks without task-specific fine-tuning (Wei et al., 2021; Sanh et al., 2021; Chung et al., 2022; Ye et al., 2022b; Ouyang et al., 2022; Wang et al., 2022a; Muennighoff et al., 2022). This paper raises some questions regarding the current paradigm of training MT LMs and is mainly divided into two parts. In Part 1, we report an unexpected finding regarding *expert* LMs (trained only on a single task) compared to MT LMs. In Part 2, we leverage the finding to highlight some of the benefits of *expert* LMs over MT LMs.

Part 1 (Section 5) Previously, the key component to enhancing the unseen task generalization performance of MT

Work done while JJ and SY were interns at LG AI Research. [1]KAIST [2]LG AI Research [3]University of Illinois Chicago. Correspondence to: Joel Jang <joeljang@kaist.ac.kr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

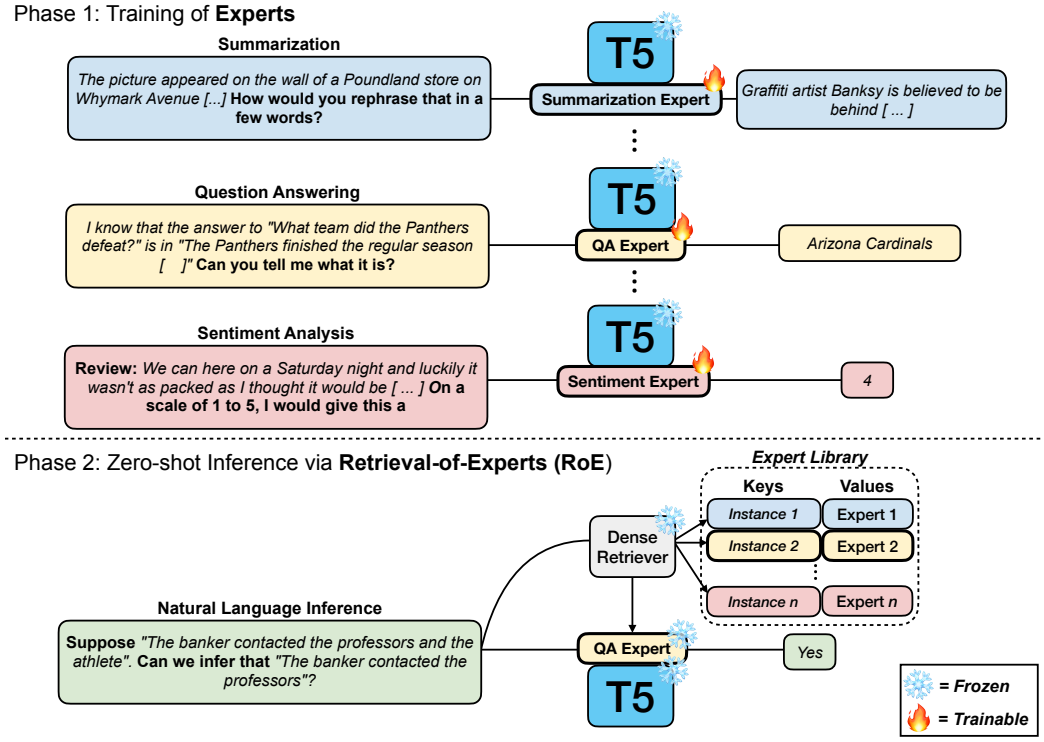

Figure 2. Independent training and Retrieval-of-Experts (RoE) for zero-shot task generalization. During training, only the additional adapters (experts) are trained while the backbone LM is frozen. After training separate experts per training task, we construct an Expert Library that stores samples of the training task as *keys*, and the specific expert id as *values*. During zero-shot inference, the most relevant expert is retrieved for an unseen task.

LMs was thought to be scaling the total number of tasks used in training (Wei et al., 2021; Chung et al., 2022; Wang et al., 2022a). However, in this work, we show that training a single expert LM on *one*[1] out of the 300+ tasks used to train an MT LM (T0-3B (Sanh et al., 2021)) can outperform the MT LM by a non-trivial margin on 24 unseen tasks on mean accuracy.

Specifically, following the same experimental setup (training and evaluation) as T0-3B (Sanh et al., 2021), one of the most widely used MT LM, we first train *expert* LMs for each given training task (296) by freezing the underlying LM and updating adapters (Houlsby et al., 2019). We report a finding that shows 7 out of the 296 experts surpass T0-3B on the capability to generalize to unseen tasks on mean accuracy (shown in Figure 1). Using the top performing expert for all of the unseen task evaluation tasks surpasses T0-3B by a mean accuracy of 3.20% and 1.29% on 11 unseen datasets and 13 datasets of the BIG-Bench benchmark, respectively. We also show that applying a simple mechanism to retrieve relevant experts for each individual unseen task results in comparable performance to T0-3B. Considering the significant room for improvement when retrieving the best-performing expert for each unseen task (+11.94% compared to T0-3B), these results imply that choosing the right expert rather than naïvely utilizing a single MT LM for all of the unseen tasks can be a more efficient and effective approach.

[1]Training task: cosmos_qa, Prompt Name: no_prompt_text from Bach et al. (2022).

Part 2 (Section 6) Leveraging the finding of expert LMs showing improved unseen task generalization capability, we highlight three other advantages of training multiple expert LMs for each task and retrieving the relevant expert during inference (shown in Figure 2) compared to training MT LMs.

#1. MT LMs do not show the optimal performance for *seen* tasks because of negative task transfer, where learning multiple tasks at once hinders the learning of some specific tasks (Aghajanyan et al., 2021; Asai et al., 2022a; Zhang et al., 2022). Expert LMs, on the other hand, are not subject to negative task transfer (Levine et al., 2022) since each task is learned independently; We show our approach of selecting relevant experts during inference results in a +10.4% mean accuracy improvement on validation datasets of the 36 training tasks compared to T0-3B.

#2. MT LMs are susceptible to catastrophic forgetting (McCloskey & Cohen, 1989) of previous tasks when learning new tasks and require re-training on previous tasks to mitigate forgetting (Chakrabarty et al., 2022). Results show our *distributed* (training individual tasks in an independent manner) approach results in absolutely no degradation of seen tasks, even when adding the 8 new experts to the Expert Library, without re-training on previous tasks when learning 8 new generative tasks.

#3. We show that MT LMs show poor ability in performing *composition* of previously learned tasks given via concatenation of the corresponding instructions as a single *compositional* instruction. On the other hand, we show that *merging* the two experts trained on the individual tasks with mT5-3B (Xue et al., 2021) as the underlying pre-trained LM results in an expert that can outperform its MT LM counterpart, mT0-3B (Muennighoff et al., 2022), by a mean ROUGE-L score of +2.71 on 5 novel compositional tasks (summarization & translation). Details of the merging mechanism are provided in Section 3.3.

2. Related Work

2.1. Multitask Prompted Fine-tuning of Language Models

Several studies have demonstrated that multitask fine-tuning moderately sized LMs with instructions, also referred to as *instruction tuning*, enables zero-shot task generalization. Specifically, Sanh et al. (2021); Wang et al. (2022a) have shown that scaling the number of training tasks, the number of prompts per task, and the size of the LM helps boost zero-shot task generalization performance. In addition to scaling these aspects, Chung et al. (2022) include Chain-of-Thought (Wei et al., 2022) tasks during instruction tuning, reaching state-of-the-art performance on zero-shot and few-shot settings with PaLM 540B (Chowdhery et al., 2022) as the underlying LM. Lin et al. (2022) improve MT LMs by adapting MT LMs on subsets of the training data retrieved given a few unlabeled examples of the unseen task. Ouyang et al. (2022) adapt MT LMs to align with human preferences through reinforcement learning. Muennighoff et al. (2022) include multilingual tasks to show cross-lingual generalization capability. Ye et al. (2022b) flip the instruction and label space to enhance generalization capability to novel unseen labels. Asai et al. (2022b) utilize instruction tuning to construct a general-purpose retrieval system. Similarly, Su et al. (2022) utilize instruction tuning to construct a general-purpose embedding model that can be used to perform different unseen tasks requiring text embeddings.

While previous literature has mostly asserted that the primary key component of MT LMs is scaling the total number of training tasks, in this paper, we propose an alternative perspective and instead show experimental results and findings that the *feature* of the tasks may be a more critical factor (analysis provided in Section 5); Similar findings are shown in the setting of few-shot adaptation (Chan et al., 2022) as well.

2.2. Retrieving task-specific embeddings

Retrieving task-specific parameters has the advantage of rapid target task adaptation, especially for low-resource scenarios (Vu et al., 2022; Asai et al., 2022a; Ye et al., 2022a; Qin & Eisner, 2021; Wang et al., 2022b; Bari et al., 2022). Vu et al. (2022) show that retrieving an optimal source soft prompt leads to better initialization for adapting to the target task. Asai et al. (2022a) also focus on retrieval of soft prompts for initialization for the target task but utilize the idea of attention weights to effectively interpolate between multiple training soft prompts. Similarly, Ye et al. (2022a) extend this idea of retrieving soft prompts, but utilize an MT LM as the underlying LM and do not fine-tune the LM to the target task, performing the target task in a zero-shot manner. Our work is motivated by Ye et al. (2022a), but proposes to replace the instruction tuning stage altogether, using vanilla pretrained LMs as the underlying LM instead of MT LMs. We accomplish this by training experts whereas previous work trained soft prompts on top of MT LMs.

2.3. Distributed Training of Language Models

Recent work has shown the possibilities and benefits of distributed training of LMs. Li et al. (2022) have shown that it is possible to merge individual LMs pretrained on different subsets (domains) of the training corpora to construct a single LM that shows lower overall perplexity compared to an LM trained on all of the corpora at once. Another line of work that explores merging individually fine-tuned LMs is Wortsman et al. (2022b), where they merge LMs fine-tuned on the same task with different configurations to boost performance. Similarly, Wortsman et al. (2022a) merge LMs fine-tuned on the same task, but with subsets of the training data for efficiency. Don-Yehiya et al. (2022) explore merging LMs fine-tuned on different tasks to make a multitask fine-tuned LM in a distributed manner, which has many benefits including federated learning (McMahan et al., 2017).

Other interesting extensions of distributed LM training include performing task arithmetic with task vectors (Ilharco et al., 2022), training and performing inference of several billion parameter LMs on distributed compute (Borzunov et al., 2022), introducing language-specific modules for growing the total capacity of multilingual LMs (Pfeiffer et al., 2022), finding theoretical guarantees of why merging works (Frankle et al., 2020; Ainsworth et al., 2022) and proposing novel methods of merging model weights (Matena & Raffel, 2021).

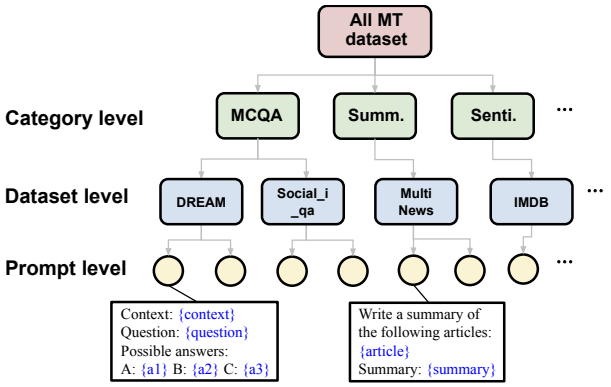

Figure 3. The hierarchy of all the training datasets used to train MT LMs. In this work, we explore training Dataset level Experts (DE) and Prompt level Experts (PE).

In our work, we also show the benefits of distributed LM training by showing that the capability of expert LMs can be further amplified through *merging* individual experts.

3. Expert Language Models

In this section, we describe the framework of our proposed method. We train each expert by training adapters for each training task (Section 3.1). During inference, we retrieve the relevant experts from the Expert Library (Section 3.2). We additionally explore the effect of merging experts to observe the benefits of distributed training (Section 3.3).

3.1. Training Experts

For training the experts, we mainly explore parameter-efficient fine-tuning via adapters while freezing the underlying LM to train individual experts. We train experts for each task with the corresponding *prompts* and denote the resulting experts as Prompt Experts (PE).[2] We also explore training experts for each *dataset*, which consists of multiple training prompts, referred to as Dataset Experts (DE). For training DE, instead of utilizing a parameter-efficient fine-tuning approach (adapters), we instead simply train the entire LM to observe the merging capability of expert LMs.[3] Figure 3 shows the hierarchy of the training datasets and the level at which PE and DE are trained on.

Adapters We apply a parameter-efficient method of representing experts by training additional adapters while freezing the original parameters (Houlsby et al., 2019). Specifically, given a standard Transformer LM with l layers, input sequence X containing T tokens, the output for a single layer $\mathbf{h}^l_{1:T}$ is calculated by

$$\mathbf{h}^l_t = \text{FFN}_d(\mathbf{u}^l_t) + \mathbf{u}^l_t, \tag{1}$$

$$\mathbf{u}^l_{1:T} = \textsc{Self-Att}(\mathbf{h}^{l-1}_{1:T}) + \mathbf{h}^{l-1}_{1:T}, \tag{2}$$

where $\mathbf{h}^l_t$ is the hidden state of t-token after the l-th layer, $\textsc{Self-Att}(\cdot)$ is the self-attention module, and $\text{FFN}_d(\cdot)$ is the feed-forward network with hidden dimensions d. When fine-tuning the LM with an adapter expert, each layer before the self-attention layer (Equation 1) changes into the following format:

$$\mathbf{h}^l_t = \text{FFN}_e(\mathbf{u}^l_t) + \text{FFN}_d(\mathbf{u}^l_t) + \mathbf{u}^l_t, \tag{3}$$

where e represents the hidden dimension of the adapter feed-forward network. When using adapters to represent experts, parameters of FFN_e are the only trainable parameters and the rest of the parameters in the LM are frozen.

3.2. Retrieval-of-Experts (RoE)

After independent (distributed) training of individual experts, we retrieve one of the experts to use during inference (Ye et al., 2022a). To this end, we construct an *Expert Library* and use dense retrieval to retrieve a relevant expert from the library to use during inference.

Expert Library We first construct the *Expert Library*. This library contains keys that are each embedding representations of a single instance from the training tasks and values that are unique ids of the corresponding trained experts. For each unique expert, S training instances are randomly sampled and stored in the library. This results in $[S \times \#\ of\ experts]$ entries in the Expert Library. To get the embedding representation of the training instances, we employ a simple Sentence Transformer (Reimers & Gurevych, 2019) as the dense retriever.[4] For the text format of the training instance that is given to the embedding model as input, we simply concatenate the answer choice (e.g. Yes|No, A|B|C|D) to the Prompted Input. The answer choice for generative tasks is given as 'None'. We report ablation results of varying the text format given as the input to the embedding model in Appendix B.

Retrieval Following the approach of Lin et al. (2022); Ye et al. (2022a), given a target task during inference, we first randomly select Q instances from the target task[5]. Next, we use the same text format (concatenation of Prompted

[2]Each prompt (instruction) is referred to as *tasks*, following Chung et al. (2022).

[3]Experimental results show that merging adapter experts does not lead to improved positive task transfer on mean accuracy (shown in Section 5).

[4]We explore other text embedding models for the retriever such as Sentence-T5, SimCSE, INSTRUCTOR, etc., in Appendix B. Sentence Transformer shows the best performance among the embedding models.

[5]We assume a scenario where we can perform batch-inference.

Input and Answer Choice) and the same embedding model used to construct the Expert Library to obtain embedding representations of each of the Q target queries. We then use MIPS (maximum inner product search) on our Expert Library to identify the most similar training instance (key) for each query instance, resulting in a total of Q corresponding experts (value). We select the most frequently retrieved expert as the expert for solving the given target task.

3.3. Merging of Experts

Previous work has shown the possibility of distributed multitask fine-tuning by *merging* individually fine-tuned LMs (Don-Yehiya et al., 2022). Along with selecting the most retrieved expert, we observe how merging fully fine-tuned LMs (DE) affects the generalization performance on the unseen tasks.

A fully fine-tuned LM can be represented in the form of a vector $\tau_d = \theta_d - \theta_{pre}$ where θ_{pre} represent the full parameters of the vanilla pretrained LM and θ_d represents the full parameters of the LM fine-tuned on the training dataset d (Ilharco et al., 2022). The formula for merging of *N* experts can be denoted as follows:

$$\theta_{new} = \theta_{pre} + (\sum_{i}^{N} \lambda_i \tau_i) \tag{4}$$

where $\lambda_i = \frac{1}{N}$ as default if not stated otherwise. Note that when $\lambda_i = \frac{1}{N}$, it results in merging experts uniformly. In some cases, however, performance was optimal when $\sum_i \lambda_i > 1$ and each λ_i (representing the importance to place on τ_i) and was set to a different value determined using a held-out validation dataset following Ilharco et al. (2022). A concrete example is provided in Appendix C.

4. Experimental Setup

Training Setup Following the setting of Sanh et al. (2021), we use a total of 36 training datasets of T0 for training our experts.[6] For each dataset, we use all of the prompts used to train T0 from the Promptsource Library (Bach et al., 2022) which results in a total of 296 prompts to train the corresponding experts ($\sim$8 prompts per training dataset). This results in 36 Dataset Experts (DE) represented via fully fine-tuned LMs, and 296 Prompt Experts (PE) via adapter training. For each individual fine-tuning, we randomly sample $K = 50,000$ training instances for each classification task and $K = 10,000$ for each generative task.[7] We use the LM-adapted T5 model (Lester et al., 2021) checkpoint as our base model, and train for 5 epochs with a constant learning rate of 1e-4 for both adapter fine-tuning and full LM fine-tuning. For the construction of the Expert Library, much smaller $S = 100$ training instances are randomly sampled for each expert following Ye et al. (2022a).

Evaluation Setup We evaluate the baseline MT LMs (T0-3B, T0-11B) and our proposed method (T5-3B + DE/PE) on the same evaluation setting as the original T0 paper (Sanh et al., 2021): 11 unseen datasets that can be categorized into 4 task categories and on 13 datasets from BIG-Bench benchmark (Srivastava et al., 2022), which are diverse and challenging tasks that are not encountered during training.[8] We further evaluate the models on 8 new generative tasks[9] that were not included in the original T0 paper evaluation setting. We use a *rank classification* evaluation by selecting the label option with higher log-likelihood following Brown et al. (2020); Sanh et al. (2021) for the classification tasks. For the generative tasks, we use the ROUGE-L score as the default metric if not stated otherwise. The details of each training and evaluation dataset are provided in Appendix A.

During inference, we set $Q = 32$ for applying our Retrieval-of-Expert (RoE) mechanism. We do not separately perform ablations of S and Q, simply following the optimal setting of Ye et al. (2022a).

5. Expert LMs Can Generalize to Unseen Tasks

In this section, we show experimental results of expert LMs and show their potential for becoming a new paradigm over instruction tuning. Since this is a fairly novel approach of endowing LMs the capability to generalize to unseen tasks, we focus on providing *proof-of-concept* of some core research questions instead of making head-to-head comparisons with all of the baselines. We leave other extensive comparisons and exhaustive ablations for future work.

Main Results Table 1 shows the evaluation results on the 11 unseen datasets, Table 2 shows the results on the 13 unseen BIG-Bench tasks, and Table 3 shows the results on the 8 unseen generative tasks. Results from the three tables show that (1) a single PE significantly outperforms T0-3B, (2) the RoE (ORC.) outperforms other baselines

[6]The original T0 (Sanh et al., 2021) paper includes 38 training datasets. However, we could not load 4 datasets from the Huggingface Dataset library: adversarial_qa/dbidaf, adversarial_qa/dbert, adversarial_qa/droberta, and duorc/SelfRC. Instead, we utilize the adversarial_qa/adversarialQA dataset and also additionally train on commonsense_qa dataset which is a variant of the cos_e dataset, resulting in a total of 36 training datasets.

[7]We train with less number of instances for the generative tasks because the training generative tasks required longer max token length, and thus longer training time.

[8]We exclude NOVEL CONCEPTS task from the original T0 evaluation setting because it is a multi-label classification task. Multiple prompts are evaluated for each evaluation dataset.

[9]The dataset details of the 8 new generative tasks are provided in Appendix A.

Method	NLI					Sentence Completion			Coreference Resolut.		WSD	Total Avg.
	RTE	CB	AN. R1	AN. R2	AN. R3	COPA	Hellasw.	StoryC.	Winogr.	WSC	WiC	
T0-11B	80.83	70.12	43.56	38.68	41.26	90.02	33.58	92.40	59.94	61.45	56.58	60.76
GPT-3(175B)	63.50	46.40	34.60	35.40	34.50	91.00	78.90	83.20	70.20	65.40	45.92	59.00
T0-3B	60.61	48.81	35.10	33.27	33.52	75.13	27.18	84.91	50.91	**65.00**	51.27	51.43
T5(3B) + COS PE	49.53	**49.52**	**36.21**	**36.11**	**36.38**	**89.63**	**43.77**	**97.06**	56.65	57.02	49.01	**54.63**
T5(3B) + PE W/ ROE	**64.01**	43.57	35.49	34.64	31.22	79.25	34.60	86.33	**61.60**	62.21	**52.97**	53.48
T5(3B) + PE W/ ROE (ORC.)	70.32	70.12	40.02	40.11	42.07	92.88	55.00	97.47	64.40	65.77	58.90	63.37

Table 1. Evaluation performance on 11 different unseen datasets categorized into 4 task categories. PE represents Prompt Experts. PE W/ ROE (ORC.) represents retrieving the best-performing (oracle) expert for each evaluation task. COS PE represents the PE trained on COSMOS-QA with the prompt NO-PROMPT-TEXT which showed the highest mean accuracy on the 11 unseen tasks. PE W/ ROE represents Retrieval-of-Expert (RoE) for each individual unseen task. Note that PE adds 100M additional parameters while freezing the 3B paramters of T5 during training. The best comparable performances are **bolded** and second best underlined.

Dataset (metric)	T0 3B	COS PE 3B	T0 11B	GPT-3 175B	PALM 540B
Known Un.	47.83	**58.70**	65.22	60.87	56.52
Logic Grid	**32.10**	30.70	33.67	31.20	32.10
Strategy.	**53.23**	42.36	54.67	52.30	64.00
Hindu Kn.	34.86	**51.43**	42.86	32.57	56.00
Movie D.	**53.22**	46.72	57.33	51.40	49.10
Code D.	53.33	**66.67**	51.67	31.67	25.00
Concept	67.25	**72.92**	71.72	26.78	59.26
Language	14.94	**25.95**	18.33	15.90	20.10
Vitamin	**58.18**	46.55	57.33	12.30	14.10
Syllogism	**52.27**	50.00	48.33	50.50	49.90
Misconcept.	**52.05**	47.03	52.97	47.95	47.47
Logical	**45.33**	42.40	54.67	23.42	24.22
Winowhy	44.29	**44.33**	55.00	51.50	45.30
BIG-bench AVG	46.84	**48.13**	51.06	37.57	41.77

Table 2. Evaluation performance on 13 BIG-bench tasks. The best comparable performances are **bolded**.

by a non-trivial margin, and (3) our simple ROE approach outperforms T0-3B on the classification tasks, but not on generative tasks. Details of each finding are provided in the following paragraphs.

#1. In Table 1, surprisingly, T5(3B) + COS PE, which is a Prompt Expert (PE) that is only trained on a single prompt ('no_prompt_text' prompt of COSMOS-QA dataset), outperforms its MT LM counterpart (T0-3B) on 8 out of 11 evaluation datasets and +3.20% on mean accuracy. Prior work shows that scaling the total number of training tasks during instruction tuning leads to better generalization; in our case, training an expert on a single task outperforms an LM trained on 300+ tasks (T0-3B). This finding is bolstered in Table 2 where the same COS PE that shows the highest mean accuracy for the 11 unseen tasks outperforms T0-3B by +1.29% on the mean accuracy performance on 13 datasets of BIG-Bench Benchmark and in Table 3 where T5(3B) + SAM PE, which is a PE trained on ('Given the above dialogue write a summary' prompt of SAMSUM dataset), outperforms T0-3B by +6.83 mean score on the 8 generative tasks.

#2. In Table 1, we can see that T5(3B) + PE W/ ROE (ORC.), which is the upper-bound performance of choosing the best-performing expert based on the accuracy for each unseen task, outperforms T0-3B, much larger GPT-3(175B) and T0-11B by +11.94%, +4.37% and +2.61%, respectively, on the mean accuracy. T5(3B) + PE W/ ROE (ORC.) also outperforms T0-3B by +13.69 mean score on the 8 unseen generative tasks shown in Table 3. This means that ROE has a potential for strong unseen task generalization when the proper expert is chosen.

#3. T5(3B) + PE W/ ROE, which is a simple method of retrieving an expert for each unseen task leveraging an off-the-shelf retriever (Sentence Transformer (Reimers & Gurevych, 2019)), outperforms T0-3B on 8 out of 11 evaluation datasets and by +2.05% on mean accuracy. However, T5(3B) + PE W/ ROE underperforms T0-3B by -5.37 mean score on the 8 unseen generative tasks (Table 3). Considering that T5(3B) + PE W/ ROE still shows a significant performance gap compared to retrieving the best-performing expert (T5(3B) + PE W/ ROE (ORC.)), there is much room for improvement on the retriever side. One way to close the gap is to train a *supervised* retrieval model, which we leave for future work.

Merging of Experts Table 4 shows the merging capability of expert LMs. The first three rows show the merging results of PE which are represented in the form of adapters. While COS&SOC PE (MER.), which is an expert constructed by performing uniform merging with COS PE and SOC PE[10] shows positive task transfers for some evaluation datasets (Copa & Story Cloze), not all of the results are the best or second best (RTE, Hellaswag, & Winogrande). This means that there was a negative task transfer when merging the adapter experts.

Thus, in order to further explore the merging capability of expert LMs, we train DE via full LM fine-tuning, known to be effective in previous literature (Ilharco et al., 2022),

[10] SOC PE is a PE that was trained on SOCIAL_I_QA with prompt 'no_prompt_text' that showed the second highest mean accuracy on the 11 unseen tasks other than PE trained with COSMOS-QA.

Method	wiki auto (BLEU)	HGen (ROUGE)	haiku (ROUGE)	covid qa (BS)	eli5 (BS)	emdg (BS)	esnli (BS)	twitter (BS)	Total Avg.
T0-3B	21.76	33.29	19.93	**50.00**	**59.86**	47.76	42.80	28.40	37.98
T5(3B) + SAM PE	**30.69**	25.49	25.25	49.93	47.94	**51.36**	**58.28**	**69.55**	**44.81**
T5(3B) + PE W/ ROE	3.88	**35.55**	**26.53**	33.52	33.66	49.90	28.61	49.22	32.61
T5(3B) + PE W/ ROE (ORC.)	31.56	35.55	30.16	52.49	63.20	58.36	60.02	82.08	51.67

Table 3. Evaluation performance on 8 unseen generative tasks. SAM PE represents the PE trained on SAMSUM with the prompt GIVEN THE ABOVE DIALOGUE WRITE A SUMMARY which showed the highest mean score on the 8 unseen generative tasks. The best comparable performances are **bolded** and second best underlined.

Method	NLI					Sentence Completion			Coreference Resolut.		WSD	Total Avg.
	RTE	CB	AN. R1	AN. R2	AN. R3	COPA	Hellasw.	StoryC.	Winogr.	WSC	WiC	
T5(3B) + COS PE	49.53	**49.52**	**36.21**	**36.11**	**36.38**	89.63	**43.77**	97.06	**56.65**	**57.02**	49.01	**54.63**
T5(3B) + SOC PE	**61.26**	38.81	33.16	33.63	33.46	90.50	37.21	97.09	55.28	50.00	**50.11**	52.77
T5(3B) + COS&SOC PE (MER.)	49.10	39.40	33.80	34.28	34.18	**91.63**	36.29	**97.25**	55.06	51.25	49.62	51.99
T5(3B) + COS DE	59.71	**57.62**	33.45	33.93	34.54	90.00	**36.58**	96.29	53.37	42.88	49.91	53.48
T5(3B) + SOC DE	**65.52**	48.69	**35.20**	**35.39**	**37.11**	83.25	30.38	87.18	54.27	**54.62**	**51.39**	53.00
T5(3B) + COS&SOC DE (MER.)	60.43	54.17	35.01	34.53	35.52	**91.25**	35.59	**96.73**	**54.33**	42.88	50.05	**53.68**

Table 4. Evaluation performance on 11 different unseen datasets categorized into 4 task categories. PE represents Prompt Experts. COS PE represents the PE trained on COSMOS-QA dataset and NO PROMPT TEXT prompt and SOC PE represents the PE trained on SOCIAL-I-QA dataset and SHOW CHOICES AND GENERATE ANSWER prompt. COS&SOC PE (MER.) represents expert constructed by performing *uniform* merging with the COS PE and SOC PE. COS DE represents the DE trained on the COSMOS-QA dataset with all of the prompts and SOC DE represents the DE trained on SOCIAL-I-QA on all of the propmts. COS&SOC DE (MER.) represents expert constructed by performing merging with the COS DE and SOC DE. The best comparable performances are **bolded** and second best underlined.

and merge them as shown in the last three rows in Table 4. COS DE (COSMOS-QA) and SOC DE (SOCIAL-I-QA) are the two highest performing DE based on the mean accuracy performance on the 11 unseen tasks. While COS&SOC DE (MER.) shows only a +0.20% enhancement compared to COS DE on mean accuracy, it still shows either the best or second best performance compared to the individual COS and SOC DE. This implies that merging the two experts results in a composition of abilities. This opens up new possibilities of leveraging the merging of experts to unlock new capabilities which are further explored in Section 6 with the composition of instructions.

Overall, Table 4 shows that merging with adapters does not always result in positive task transfer while merging with full parameters seems to. Thus, future work should explore developing more parameter-efficient methods of merging expert LMs since always training and utilizing the entire LM weights is computationally demanding.

Analysis of Experts Figure 1 shows the mean accuracies of all the PE and DE results on the 11 unseen datasets. We highlight three main analyses from the figure and from the tables.

First, among the 8 training task categories, Multiple-Choice Question Answering (MCQA) training tasks generally show the strongest generalization capability. We hypothesize this to be the case because all of the 11 evaluation datasets are classification tasks and require some form of question answering via instructions. This extends the findings of Khashabi et al. (2020) that Multiple-Choice Question Answering (MCQA) generalizes well to not only different format QA tasks, but also different types of tasks such as natural language inference, story completion, coreference resolution, and word sense disambiguation as well.

Second, among the 36 training datasets, 3 datasets consistently ensure high performance for both PE and DE: COSMOS-QA (Huang et al., 2019), SOCIAL-I-QA (Sap et al., 2019), and DREAM (Sun et al., 2019). All three datasets are commonsense reasoning datasets, which have been considered to be crucial for generalization to unseen tasks (Lourie et al., 2021). We provide the full ranking of the PE and DE for the 11 unseen tasks shown in Figure 1 in Appendix E.

Lastly, T5(3B) + SAM PE which is a PE trained on SAMSUM (Gliwa et al., 2019), a dataset with abstractive dialogue summaries, shows the best mean score on the 8 unseen generative tasks in Table 3, outperforming T0-3B by +6.83 mean score. However, the same PE shows one of the lowest ranks for the 11 unseen (classification) tasks (shown in Appendix E) underperforming T0-3B by -9.15% mean accuracy. This shows that there is *no free lunch*: a PE that shows high mean performance for unseen generative tasks do not show high mean performance for unseen classification tasks. This also implies that it is more-so important to retrieve the correct expert dynamically depending on the given context (target task).

Method	MCQA (12) (ACC)	Senti. (5) (ACC)	Topic C. (3) (ACC)	Paraph. (3) (ACC)	STS (2) (ROUGE-L)	Summ. (5) (ROUGE-L)	EQA (4) (ROUGE-L)	CBQA (2) (ROUGE-L)	Total Avg.
T0-3B	46.97	66.40	59.99	**76.63**	41.90	33.10	28.79	24.67	47.30
T0-11B	51.32	64.03	60.95	73.64	45.42	33.10	**41.20**	30.37	50.00
T5(3B)+ PE w/ RoE	**58.95**	**70.18**	**96.52**	72.97	**47.57**	**33.14**	30.36	**51.89**	**57.70**
T5(3B)+ PE w/ RoE (Orc.)	56.28	84.52	96.91	79.34	47.94	35.40	40.34	43.24	60.50

Table 5. Evaluation performance on 300 sample instances from each validation dataset of the 36 training tasks categorized into 8 task categories. The number in the () represents the # of datasets in the task category. The best comparable performances are **bolded** and second best underlined.

Method	Seen Avg.	Unseen Avg.	Gen Avg.
Before Continual Learning			*Unseen*
T0-3B	47.30	51.43	**37.98**
T5(3B) + PE w/ RoE	**57.70**	**53.48**	32.61
After Continual Learning			*Seen*
CT0-3B	47.54	50.84	54.52 (↑)
T5(3B) + PE$^+$ w/ RoE	**57.70**	**53.33**	**55.60** (↑)

Table 6. **Seen Avg.** represents the mean accuracy of the 36 seen tasks in Table 5. **Unseen Avg.** represents the mean accuracy of the 11 unseen tasks in Table 1. **Gen Avg.** represents the mean score of the 8 (unseen) generative tasks in Table 3. (BS) represents BertScore. PE$^+$ represents augmenting the Expert Library with 8 PE trained on the 8 generative tasks. We use the LM checkpoint from Chakrabarty et al. (2022) for CT0-3B, T0-3B continually fine-tuned on the 8 generative tasks is a sequential manner while rehearsing previous tasks. The best comparable performances are **bolded**.

6. Benefits of Expert LMs over MT LMs

In this section, we highlight the 3 main benefits of expert LMs and RoE over MT LMs.

Seen Task Performance First, we show that expert LMs are less susceptible to negative task transfer by comparing the performance of T5(3B) + PE w/ RoE on the validation datasets of the 36 training datasets with two MT LMs, T0-3B and T0-11B. As shown in Table 5, our distributed approach outperforms T0-3B and T0-11B by +10.40% and +7.70% on mean accuracy, respectively.

This is because since evaluation is done with *seen* instructions, our simple retrieval mechanism is highly likely to select the best-performing expert from the Expert Library, showing comparable performance to T5(3B) + PE w/ RoE (Orc.). In fact, T5(3B) + PE w/ RoE retrieves the PE from the same training dataset on 280 out of 296 seen tasks, and the PE trained with both the same prompt and dataset (oracle) on 185 out of 296 seen tasks.

Continual Learning of New Tasks In some scenarios when we want to additionally fine-tune LMs on additional datasets *after* model deployment, making finetuned LMs continual learners is important (Chakrabarty et al., 2022). This is because performing instruction tuning on the entire set of original and additional tasks in each update would lead to heavy computation. Previous work mitigates this issue through a rehearsal-based method, continually training the instruction-tuned LM on *samples* of the original and additional tasks (Chakrabarty et al., 2022). However, this approach (1) assumes that we have access to the original datasets and (2) still leads to additional computational overhead, especially when scaling the total number of seen tasks during instruction tuning.

We show that we can accomplish the same feat through distributed training of experts without any access to original, seen datasets by training separate experts for each additional task and simply adding them to the Expert Library. Specifically, we show the comparison between continually training an MT LM (T0-3B) which is referred to as CT0-3B through a rehearsal-based approach, and our distributed approach on 8 new generative tasks in Table 6. The 8 generative tasks for continual learning were chosen following the previous work (Chakrabarty et al., 2022).

The table shows that our distributed approach results in absolutely no degradation of performance for the seen task, a minor (-0.15%) degradation for unseen tasks, and superior mean performance (+1.08) for the 8 target tasks compared to the MT LM counterpart, outperforming CT0-3B on 7 out of the 8 target tasks. This shows that without any access to original datasets or heavy computational cost, our distributed approach is mostly able to retain its original ability (seen & unseen) as well as outperform CT0-3B on the target tasks. We leave scaling the number of new target tasks and how our distributed approach performs against its instruction-tuned counterpart for future work.

Compositional Instructions Prior work has shown the need for performing *compositional* instructions (Logeswaran et al., 2021; Corona et al., 2021; Khot et al., 2022). For example, we can give the following instruction to the LM: "*Write a summary of the following English text and translate the sentence into Korean.*" where "*Write a summary of the following English text.*" and "*Translate the sentence into Korean.*" are two separate instructions seen during training. To test this compositional capability, especially in a multi-lingual setting, we utilize the mT0-3B (Muennighoff et al., 2022) as our MT LM and evaluate the composition of

Method	xsum en→ko	xsum en→es	xsum en→zh	xsum en→fr	xsum en→ja	Total Avg.
MT0-3B	1.84	16.14	**6.74**	20.37	3.44	9.71
MT5-3B + MER. EX.	**8.23**	**16.97**	2.40	**20.55**	**13.98**	**12.43**

Table 7. Comparison of compositional abilities of both summarization and translation task for MT LM (MT0-3B) and our distributed approach (MT5-3B + MER. EX.) which involves merging the corresponding experts. ROUGE-L is used as the evaluation metric. ko, es, zh, fr, ja stand for Korean, Spanish, Chinese, French, and Japanese, respectively. The best comparable performances are **bolded**

Task	EXAMPLE
xsum en→es	**MT0**: El asesinato de un niño de tres años de edad en Francia fue atribuido a su hermano mayor. **M.E.**: La policía francesa arrestó a cuatro miembros de la familia del niño por su presunta implicación en el crimen, que ocurrió el 17 de septiembre en la casa familiar en Mulhouse, al oeste del país, y de más de 100.000 habitantes. **G.T.**: La polic ıa que investiga el fallecimiento de un ni no de 9 a nos en Francia, supuestamente golpeado hasta la muerte, arrest o este martes a cuatro miembros de su familia, seg un declaraciones de los fiscales a la agencia de noticias AFP.
xsum en→fr	**MT0**: Le président de la République démocratique du Malawi a été condamné à cinq ans de prison pour complicité dans l'assassinat de Paul Mphwiyo. **M.E.**: Le 8 novembre 2013, l'ancien ministre de la Justice du Malawi, M. Ralph Kasambara, a été arrêté après avoir commis le meurtre de Paul Paul MPHWIYO, le directeur du budget du ministère des Finances. **G.T.**: La Haute Cour de Lilongwe a condamn e mardi l'ancien ministre de la Justice, Raphael Kasambara, 'a 13 ans d'emprisonnement et de travaux forc es pour complicit e de meurtre
xsum en→ja	**MT0**: 副裁 Meng Ship 和副 Meng Teng 被加拿大警方逮捕,被指控侵犯公民利 。 **M.E.**: カナダの最高裁判所(CFO)は 12月に逮捕された創設者の息子であり 、副社長はカナダ政府とカナダ移民局(CBSA)と警察を告訴した 。 **G.T.**: 中の通信機器最大手 華技術(ファウェイ)の最高財務責任者(CFO)の孟晩舟副長は 昨年12月にカナダ局がアメリカの要請で自分を逮捕したことについて 、カナダを提訴した 。
xsum en→zh	**MT0**: The Sierra Leonean nurse who was isolated for seven hours at the airport terminal has said that the isolation experience is "terrifying" and may make other medical workers reluctant to go to West Africa. **M.E.**: 一名感染埃博拉病毒的生Craig Spencer目前正在大都院接受隔治,但只得到了一粮食棒的 。 **G.T.**: 一位曾在塞拉利治埃博拉病人的美士返回美后被隔察,批了瓦克机待的方式
xsum **en→ko**	**MT0**: Korean peninsula has had its warmest winter since 1973, according to the Meteorological Administration. **M.E.**: 지난해 1월은 국내에서 가장 따뜻한 겨울이었다. **G.T.**: 올겨울, 추위가 실종됐다. 따뜻한 날씨가 이어지면서 눈 구경도 어려워졌다.

Table 8. Example outputs from the 5 Compositional Tasks given the input "*Write a summary of the following English text and translate the sentence into [Language]: [English Summary]*.". **M.E.** stands for Merged Experts. **G.T.** stands for Ground Truth. es, fr, ja, zh, and ko stand for Spanish, French, Japanese, Chinese, and Korean, respectively. The actual input for the examples are provided in Appendix C.

performing 5 novel compositional tasks of summarization and translation. To explore the benefits of merging experts for performing compositional instructions, we perform 6 full fine-tuning with mT5-3B (Xue et al., 2021) as the underlying vanilla pretrained multilingual LLM: We use XSUM to train one English Summarization expert and use five translation pairs in TATOEBA (en→es, en→fr, en→ja, en→zh, en→ko) to train the corresponding five translation experts. During inference, we merge the summarization expert with each of the five translation experts[11]. Note that both XSUM and TATOEBA are part of the training tasks used during instruction tuning of mT0-3B.

Evaluation results on the five compositional tasks are shown in Table 7. Our distributed approach, MT5-3B + MER. EX., outperforms its MT LM counterpart, MT0-3B on 4 out of the 5 tasks and by a mean ROUGLE-L score of +2.71; This is due to a significant performance gap for the tasks involving low-resource languages (Korean and Japanese) because the low-resource languages are protected from negative transfer when doing distributed training. Cherry-picked output examples of the MT LM and the merged experts are provided in Table 8.

7. Conclusion

In this work, we provide an interesting finding that *expert* LMs trained on single tasks show strong generalization capability to unseen tasks, even surpassing MT LMs trained on multiple tasks (300+) by a non-trivial margin. We leverage this capability and show three main benefits of training and retrieving experts for inference over MT LMs, demonstrating that our proposed distributed approach is more robust against negative task transfer, more adapt at learning new tasks, and can perform compositional instructions. To this end, we urge the research community to further explore distributed and collaborative training of experts which may have other future benefits including efficiency, privacy, and personalization not explicitly explored in this paper. We provide limitations and discussion of this work in Appendix D.

ACKNOWLEDGMENTS

We thank Colin Raffel, Sungdong Kim, Sejune Joo, Miyoung Ko, Eunbi Choi, Hyunji Lee, Dongkeun Yoon, Yoonjoo Lee, and Yujin Kim for the useful discussion and feedback on the paper. This work was partly supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00113, Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework, 80%; No.2021-0-02068, Artificial Intelligence Innovation Hub, 20%).

References

Aghajanyan, A., Gupta, A., Shrivastava, A., Chen, X., Zettlemoyer, L., and Gupta, S. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5799–5811, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.468. URL https://

[11] We provide the specific configurations used for merging such as the λ_i values for each task vector τ_i and the training and validation stats in Appendix C

aclanthology.org/2021.emnlp-main.468.

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Asai, A., Salehi, M., Peters, M. E., and Hajishirzi, H. Attempt: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6655–6672, 2022a.

Asai, A., Schick, T., Lewis, P., Chen, X., Izacard, G., Riedel, S., Hajishirzi, H., and Yih, W.-t. Task-aware retrieval with instructions. *arXiv preprint arXiv:2211.09260*, 2022b.

Bach, S., Sanh, V., Yong, Z. X., Webson, A., Raffel, C., Nayak, N. V., Sharma, A., Kim, T., Bari, M. S., Fevry, T., Alyafeai, Z., Dey, M., Santilli, A., Sun, Z., Ben-david, S., Xu, C., Chhablani, G., Wang, H., Fries, J., Al-shaibani, M., Sharma, S., Thakker, U., Almubarak, K., Tang, X., Radev, D., Jiang, M. T.-j., and Rush, A. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 93–104, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-demo.9. URL https://aclanthology.org/2022.acl-demo.9.

Bari, M. S., Zhang, A., Zheng, S., Shi, X., Zhu, Y., Joty, S., and Li, M. Spt: Semi-parametric prompt tuning for multitask prompted learning. *arXiv preprint arXiv:2212.10929*, 2022.

Bartolo, M., Roberts, A., Welbl, J., Riedel, S., and Stenetorp, P. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020a. doi: 10.1162/tacl_a_00338. URL https://aclanthology.org/2020.tacl-1.43.

Bartolo, M., Roberts, A., Welbl, J., Riedel, S., and Stenetorp, P. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8: 662–678, 2020b. doi: 10.1162/tacl_a_00338. URL https://doi.org/10.1162/tacl_a_00338.

Borzunov, A., Baranchuk, D., Dettmers, T., Ryabinin, M., Belkada, Y., Chumachenko, A., Samygin, P., and Raffel, C. Petals: Collaborative inference and fine-tuning of large models. *arXiv preprint arXiv:2209.01188*, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., and Blunsom, P. e-snli: Natural language inference with natural language explanations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 9539–9549. Curran Associates, Inc., 2018.

Chakrabarty, T., Scialom, T., and Muresan, S. Fine-tuned language models can be continual learners. In *Challenges & Perspectives in Creating Large Language Models*, 2022. URL https://openreview.net/forum?id=rbMH3zBIbc.

Chan, J. S., Pieler, M., Jao, J., Scheurer, J., and Perez, E. Few-shot adaptation works with unpredictable data. *arXiv preprint arXiv:2208.01009*, 2022.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Chuang, Y.-S., Dangovski, R., Luo, H., Zhang, Y., Chang, S., Soljačić, M., Li, S.-W., Yih, S., Kim, Y., and Glass, J. Diffcse: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4207–4218, 2022.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Corona, R., Fried, D., Devin, C., Klein, D., and Darrell, T. Modular networks for compositional instruction following. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1033–1040, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.81. URL https://aclanthology.org/2021.naacl-main.81.

Dagan, I., Glickman, O., and Magnini, B. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pp. 177–190. Springer, 2005.

De Marneffe, M.-C., Simons, M., and Tonhauser, J. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pp. 107–124, 2019.

Don-Yehiya, S., Venezian, E., Raffel, C., Slonim, N., Katz, Y., and Choshen, L. Cold fusion: Collaborative descent for distributed multitask finetuning. *arXiv preprint arXiv:2212.01378*, 2022.

Fabbri, A., Li, I., She, T., Li, S., and Radev, D. 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*, pp. 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1102. URL https://aclanthology.org/P19-1102.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346. URL https://aclanthology.org/P19-1346.

Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269. PMLR, 2020.

Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021.

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409.

Graff, D., Kong, J., Chen, K., and Maeda, K. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1): 34, 2003.

Hasan, T., Bhattacharjee, A., Islam, M. S., Samin, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., and Shahriyar, R. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages, 2021.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for NLP. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Huang, L., Le Bras, R., Bhagavatula, C., and Choi, Y. 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)*, pp. 2391–2401, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1243. URL https://aclanthology.org/D19-1243.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Jiang, C., Maddela, M., Lan, W., Zhong, Y., and Xu, W. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7943–7960, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.709. URL https://aclanthology.org/2020.acl-main.709.

Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., and Hajishirzi, H. Unifiedqa: Crossing format boundaries with a single qa system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1896–1907, 2020.

Khot, T., Clark, P., Guerquin, M., Jansen, P., and Sabharwal, A. Qasc: A dataset for question answering via sentence composition. *arXiv:1910.11473v2*, 2020.

Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., and Sabharwal, A. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*, 2022.

Lebret, R., Grangier, D., and Auli, M. 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*, pp. 1203–1213, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1128. URL https://aclanthology.org/D16-1128.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6:167–195, 2015.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Levesque, H., Davis, E., and Morgenstern, L. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

Levine, Y., Dalmedigos, I., Ram, O., Zeldes, Y., Jannai, D., Muhlgay, D., Osin, Y., Lieber, O., Lenz, B., Shalev-Shwartz, S., et al. Standing on the shoulders of giant frozen language models. *arXiv preprint arXiv:2204.10019*, 2022.

Li, M., Gururangan, S., Dettmers, T., Lewis, M., Althoff, T., Smith, N. A., and Zettlemoyer, L. Branch-train-merge: Embarrassingly parallel training of expert language models. *arXiv preprint arXiv:2208.03306*, 2022.

Li, X. and Roth, D. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://www.aclweb.org/anthology/C02-1150.

Lin, B. Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y., and Ren, X. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1823–1840, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.165. URL https://aclanthology.org/2020.findings-emnlp.165.

Lin, B. Y., Tan, K., Miller, C., Tian, B., and Ren, X. Unsupervised cross-task generalization via retrieval augmentation. *arXiv preprint arXiv:2204.07937*, 2022.

Lin, K., Tafjord, O., Clark, P., and Gardner, M. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 58–62, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5808. URL https://aclanthology.org/D19-5808.

Logeswaran, L., Carvalho, W. T., and Lee, H. Learning compositional tasks from language instructions. In *Deep RL Workshop NeurIPS 2021*, 2021. URL https://openreview.net/forum?id=CoMFsP9Vs-k.

Lourie, N., Le Bras, R., Bhagavatula, C., and Choi, Y. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13480–13488, 2021.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015.

Matena, M. and Raffel, C. Merging models with fisher-weighted averaging. *arXiv preprint arXiv:2111.09832*, 2021.

McAuley, J. J. and Leskovec, J. Hidden factors and hidden topics: understanding rating dimensions with review text. In Yang, Q., King, I., Li, Q., Pu, P., and Karypis, G. (eds.), *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*, pp. 165–172. ACM, 2013. doi: 10.1145/2507157.2507163. URL https://doi.org/10.1145/2507157.2507163.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation*, 24:109–165, 1989.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Möller, T., Reina, A., Jayakumar, R., and Pietsch, M. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.nlpcovid19-acl.18.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1098. URL https://aclanthology.org/N16-1098.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

Narayan, S., Cohen, S. B., and Lapata, M. 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*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206.

Ni, J., Qu, C., Lu, J., Dai, Z., Ábrego, G. H., Ma, J., Zhao, V. Y., Luan, Y., Hall, K. B., Chang, M.-W., et al. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*, 2021.

Ni, J., Abrego, G. H., Constant, N., Ma, J., Hall, K., Cer, D., and Yang, Y. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1864–1874, 2022.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Pang, B. and Lee, L. 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)*, pp. 115–124, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219855. URL https://www.aclweb.org/anthology/P05-1015.

Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., and Riedel, S. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200.

Pfeiffer, J., Goyal, N., Lin, X. V., Li, X., Cross, J., Riedel, S., and Artetxe, M. Lifting the curse of multilinguality by pre-training modular transformers. *arXiv preprint arXiv:2205.06266*, 2022.

Pilehvar, M. T. and Camacho-Collados, J. 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)*, pp. 1267–1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1128. URL https://aclanthology.org/N19-1128.

Qin, G. and Eisner, J. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5203–5212, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.410. URL https://aclanthology.org/2021.naacl-main.410.

Rajani, N. F., McCann, B., Xiong, C., and Socher, R. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4932–4942, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1487. URL https://aclanthology.org/P19-1487.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Roemmele, M., Bejan, C. A., and Gordon, A. S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95, 2011.

Rogers, A., Kovaleva, O., Downey, M., and Rumshisky, A. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*

2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pp. 8722–8731. AAAI Press, 2020a. URL https://aaai.org/ojs/index.php/AAAI/article/view/6398.

Rogers, A., Kovaleva, O., Downey, M., and Rumshisky, A. 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, Apr. 2020b. doi: 10.1609/aaai.v34i05.6398. URL https://ojs.aaai.org/index.php/AAAI/article/view/6398.

Saha, A., Aralikatte, R., Khapra, M. M., and Sankaranarayanan, K. 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)*, pp. 1683–1693, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1156. URL https://aclanthology.org/P18-1156.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

Sap, M., Rashkin, H., Chen, D., Le Bras, R., and Choi, Y. 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)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454.

See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://www.aclweb.org/anthology/P17-1099.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Su, H., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.-t., Smith, N. A., Zettlemoyer, L., Yu, T., et al. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*, 2022.

Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., and Cardie, C. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, 2019. doi: 10.1162/tacl_a_00264. URL https://aclanthology.org/Q19-1014.

Tafjord, O., Clark, P., Gardner, M., Yih, W.-t., and Sabharwal, A. Quarel: A dataset and models for answering questions about qualitative relationships. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7063–7071, 2019.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. 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)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Tandon, N., Dalvi, B., Sakaguchi, K., Clark, P., and Bosselut, A. 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)*, pp. 6076–6085, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1629. URL https://aclanthology.org/D19-1629.

Vu, T., Lester, B., Constant, N., Al-Rfou', R., and Cer, D. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5039–5059, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.346. URL https://aclanthology.org/2022.acl-long.346.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings*

of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *URL https://arxiv. org/abs/2204.07705*, 2022a.

Wang, Z., Zhang, Z., Lee, C.-Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., and Pfister, T. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022b.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Welbl, J., Liu, N. F., and Gardner, M. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413.

Welbl, J., Stenetorp, P., and Riedel, S. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018. doi: 10.1162/tacl_a_00021. URL https://aclanthology.org/Q18-1021.

Wortsman, M., Gururangan, S., Li, S., Farhadi, A., Schmidt, L., Rabbat, M., and Morcos, A. S. lo-fi: distributed fine-tuning without communication. *arXiv preprint arXiv:2210.11948*, 2022a.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998. PMLR, 2022b.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Yamada, K., Hitomi, Y., Tamori, H., Sasano, R., Okazaki, N., Inui, K., and Takeda, K. Transformer-based lexically constrained headline generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4085–4090, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.335. URL https://aclanthology.org/2021.emnlp-main.335.

Yang, Y., Yih, W.-t., and Meek, C. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2013–2018, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1237. URL https://aclanthology.org/D15-1237.

Ye, S., Jang, J., Kim, D., Jo, Y., and Seo, M. Retrieval of soft prompt enhances zero-shot task generalization. *arXiv preprint arXiv:2210.03029*, 2022a.

Ye, S., Kim, D., Jang, J., Shin, J., and Seo, M. Guess the instruction! making language models stronger zero-shot learners. *arXiv preprint arXiv:2210.02969*, 2022b.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Zhang, W., Deng, L., Zhang, L., and Wu, D. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 2022.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015a. URL https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Zhang, X., Zhao, J. J., and LeCun, Y. Character-level convolutional networks for text classification. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and

Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 649–657, 2015b. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Zhang, Y., Baldridge, J., and He, L. 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)*, pp. 1298–1308, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1131. URL https://www.aclweb.org/anthology/N19-1131.

Embedding Models	Hellasw.	StoryC.	AN. R1	AN. R2	AN. R3	COPA	CB	RTE	WSC	WiC	Winogr.	Total Avg.
RANDOM	31.25	47.38	32.94	33.38	32.12	61.00	38.57	54.01	46.35	49.03	54.27	43.66
ALL-MINILM-L6-V2	34.60	86.33	**35.49**	**34.64**	31.22	79.25	43.57	**64.01**	62.21	52.97	**61.60**	**53.48**
ALL-MINILM-L12-V2	32.33	67.13	33.84	33.38	33.69	63.00	47.38	58.48	49.52	51.17	56.80	47.88
ALL-MPNET-BASE-V2	31.53	59.33	33.71	33.02	31.73	61.38	46.43	53.97	44.62	52.33	54.93	47.73
NLI-MPNET-BASE-V2	22.60	50.87	34.02	33.69	34.53	58.75	38.57	48.59	52.21	49.77	51.07	43.15
SUP-SIMCSE-ROBERTA-LARGE	26.93	59.67	34.58	33.29	**34.73**	84.75	41.90	52.06	50.67	**56.03**	51.67	47.84
UNSUP-SIMCSE-ROBERTA-LARGE	24.27	71.93	33.98	32.22	33.78	69.75	43.33	50.72	55.38	50.33	50.93	46.97
HKUNLP/INSTRUCTOR-LARGE	19.80	57.33	33.16	33.78	32.93	54.50	39.64	47.80	55.96	49.20	51.20	43.21
HKUNLP/INSTRUCTOR-XL	19.60	44.53	32.62	32.82	32.31	57.88	44.52	47.83	60.77	48.77	51.80	43.04
GTR-T5-LARGE	29.60	70.47	33.04	31.64	32.31	58.38	**50.95**	54.69	57.79	51.50	50.80	47.38
GTR-T5-XL	**37.20**	84.80	33.24	33.27	33.58	83.00	43.69	58.59	45.00	50.73	51.07	50.38
SENTENCE-T5-LARGE	33.33	78.53	33.11	33.76	33.31	**87.25**	46.19	58.34	**63.08**	52.13	54.27	52.12
SENTENCE-T5-XL	25.67	**87.13**	35.27	33.38	32.98	68.63	46.19	59.10	61.63	52.33	51.67	50.36
VOIDISM/DIFFCSE-BERT-BASE-UNCASED-STS	21.93	46.53	33.07	32.91	32.47	58.75	45.60	49.71	60.77	49.70	50.33	43.80
T0-SMALL (YE ET AL., 2022A)	39.55	97.09	33.89	33.96	34.38	88.00	41.55	62.53	53.95	52.45	70.20	55.23

Table 9. Comparison of different embedding models, measured on 11 different unseen datasets using Prompt Experts(PE). For instance, ALL-MINILM-L6-V2 refers to T5(3B) + PE W/ ROE in Table 1. All the task format are fixed to *'Answer Choices: {answer choice}, Instance: {instance}'*. The best comparable performances are **bolded** and second best underlined. Note that evaluation is performed with 300 samples from each evaluation dataset for efficiency.

A. Details of Training and Evaluation Datasets

Details of Training Dataset Following Sanh et al. (2021), we use 36 training datasets from the 8 task categories for training our experts. We provide the official names given in Huggingface Datasets: **Sentiment Classification (Senti.)** imdb (Maas et al., 2011), amazon_polarity (McAuley & Leskovec, 2013), rotten_tomatoes (Pang & Lee, 2005), yelp_review_full (Zhang et al., 2015b), and app_reviews. **Paraphrase Identification (Para.)** glue/qqp (Wang et al., 2018), glue/mrpc (Wang et al., 2018), and paws/labeled_final (Zhang et al., 2019). **Topic Classification (Topic C.** ag_news (Zhang et al., 2015a), dbpedia_14 (Lehmann et al., 2015), and trec (Li & Roth, 2002). **Summarization (Summ.)** gigaword (Graff et al., 2003), multi_news (Fabbri et al., 2019), samsum (Gliwa et al., 2019), xsum (Narayan et al., 2018), and cnn_dailymail/3.0.0 (See et al., 2017). **Structure-To-Text (STS)** common_gen (Lin et al., 2020) and wiki_bio (Lebret et al., 2016). **Multiple-Choice Question Answering (MCQA)** commonsense_qa (Talmor et al., 2019), dream (Sun et al., 2019), quail (Rogers et al., 2020a), qasc (Khot et al., 2020), quarel (Tafjord et al., 2019), cos_e/v1.11 (Rajani et al., 2019), quail (Rogers et al., 2020b), social_i_qa (Sap et al., 2019), wiqa (Tandon et al., 2019), cosmos_qa (Huang et al., 2019), sciq (Welbl et al., 2017), and wiki_hop/original (Welbl et al., 2018) **Extractive Question Answering (EQA)** adversarial_qa/adversarial_qa (Bartolo et al., 2020b), quoref (Bartolo et al., 2020a), ropes (Lin et al., 2019), and duorc/Paraphrase IdentificationRC (Saha et al., 2018) **Closed Book Question Answering (CBQA)** kilt_tasks/hotpotqa (Petroni et al., 2021) and wiki_qa (Yang et al., 2015).

Details of Evaluation Dataset Following Sanh et al. (2021), we include 11 evaluation datasets as follows: RTE (Dagan et al., 2005), CB (De Marneffe et al., 2019), ANLI (Nie et al., 2020) for natural language inference task, COPA (Roemmele et al., 2011), Hellaswag (Zellers et al., 2019), Storycloze (Mostafazadeh et al., 2016) for sentence completion task, Winogrande (Sakaguchi et al., 2021), WSC (Levesque et al., 2012) for coreference resolution task, and WiC (Pilehvar & Camacho-Collados, 2019) for word sense disambiguation task.

For BIG-bench tasks, we evaluate on 13 tasks, following Sanh et al. (2021): Known Unknown, Logic Grid, StrategyQA, Hindu Knowledge, Movie Dialog, Code Description, Conceptual, Language ID, Vitamin C, Syllogisms, Misconceptions, Logical Deduction, and Winowhy.

For the generative evaluation tasks, we follow Chakrabarty et al. (2022) and utilize 8 tasks: Text Simplification (Wiki-Auto) (Jiang et al., 2020), Headline Generation with constraint (HGen) (Yamada et al., 2021), Haiku Generation (Haiku), Covid QA (Möller et al., 2020), Inquisitive Question Generation (ELI5) (Fan et al., 2019), Empathetic Dialogue Generation (EmDg) (Rashkin et al., 2019), Explanation Generation (eSNLI) (Camburu et al., 2018), and Twitter Stylometry (Twitter)

B. Varying the Embedding Model and Text Format for Retrieval of Experts

Performance of Different Embedding Models While Ye et al. (2022a) used T0 (Sanh et al., 2021) as the base embedding model to retrieve prompt embeddings, we explore 13 different sentence embedding models to waive the need of using instruction tuned models for retrieval of expert LMs.

Text Format	Hellasw.	StoryC.	AN. R1	AN. R2	AN. R3	COPA	CB	RTE	WSC	WiC	Winogr.	Total Avg.
'Instance: {instance}'	24.67	78.07	33.53	32.67	32.91	64.13	40.36	54.55	50.48	52.47	52.73	46.96
'Answer Choices: {label list}'	24.93	51.47	33.80	34.29	33.20	58.38	42.38	50.83	51.54	53.47	51.13	44.13
'Answer Choices: {answer choice}'	31.60	50.53	32.09	32.16	**35.98**	84.75	44.05	50.83	51.54	53.47	**63.40**	48.22
'Answer Choices: {label list}, Instance: {instance}'	32.27	56.40	**35.76**	**34.73**	31.11	67.13	46.31	59.17	61.15	52.30	52.67	48.09
'Answer Choices: {answer choice}, Instance: {instance}'	34.60	**86.33**	35.49	34.64	31.22	79.25	43.57	**64.01**	62.21	52.97	61.60	**53.48**
'{instance}'	24.27	82.40	33.53	33.47	33.89	58.25	43.81	51.66	51.92	52.60	51.13	46.99
'{label list}'	24.53	50.53	33.67	32.76	32.58	58.38	42.02	50.83	51.54	53.47	51.13	43.77
'{answer choice}'	24.00	49.87	32.09	32.16	**35.98**	**86.00**	44.05	50.83	51.54	53.47	**63.40**	47.58
'{label list}</s>{instance}'	25.53	65.60	**35.76**	33.91	31.07	62.38	**46.90**	60.14	**62.69**	**53.70**	50.73	48.04
'{answer choice}</s>{instance}'	**35.93**	60.53	35.29	32.51	33.00	68.75	43.93	59.03	62.60	52.40	60.73	49.52

Table 10. Comparison of different text formats, measured on 11 different unseen datasets using Prompt Experts(PE). For instance, *'Answer Choices: {answer choice}, Instance: {instance}'* refers to T5(3B) + PE W/ ROE in Table 1. All the embedding model are fixed to ALL-MINILM-L6-V2. The best comparable performances are **bolded** and second best underlined. Note that evaluation is performed with 300 samples from each evaluation dataset for efficiency.

More specifically, we list of embedding models we use are as follows: (a) 4 different variants of SENTENCE TRANSFORMER model (Reimers & Gurevych, 2019): all-MiniLM-L6-v2, all-MiniLM-L12-v2, all-mpnet-base-v2, nli-mpnet-base-v2, (b) 2 different variants of SIMCSE model (Gao et al., 2021): sup-simcse-roberta-large, unsup-simcse-roberta-large, (c) 2 different variants of INSTRUCTOR model (Su et al., 2022): hkunlp/instructor-large, hkunlp/instructor-xl, (d) 2 different variants of GTR model (Ni et al., 2021): gtr-t5-large, gtr-tr-xl, (e) 2 different variants of SENTENCET5 model (Ni et al., 2022): sentence-t5-large, sentence-t5-xl, and (f) DIFFCSE model (Chuang et al., 2022): voidism/diffcse-bert-base-uncased-sts which are all available on HuggingFace. Note that we try different embedding models in an unsupervised manner, i.e., not requiring any supervision to train the embedding model, but using it off-the-shelf. The results are shown in Table 9.

Performance of Different Text Formats We also try different variants of text format given to the embedding model. Using Promptsource (Bach et al., 2022), we compare including the instance, label list, answer choice in 2 different formats. Specifically, the full list of text formats are as follows: (a) *'Instance: {instance}'*, (b) *'Answer Choices: {label list}'*, (c) *'Answer Choices: {answer choice}'*, (d) *'Answer Choices: {label list}, Instance: {instance}'*, (e) *'Answer Choices: {answer choice}, Instance: {instance}'*, (f) *'{instance}'*, (g) *'{label list}'*, (h) *'{answer choice}'*, (i) *'{label list}</s>{instance}'*, (j) *'{answer choice}</s>{instance}'*. Label list and answer choice differ in that while label list uses the actual label options (e.g., ['swim','fly','walk','run']), answer choice organizes them with a '—' deliminator in the middle (e.g. A|B|C|D). The results are shown in Table 10.

Results While we tried different variants, the oldest, yet most chosen model ALL-MINILM-L6-V2 outperforms other options. We conjecture that this is because most of the model variants we tested were trained as sentence embedding models, not for embedding prompted instances. Prompted instances are some how structural and formatted compared to natural language sentences used for training sentence embedding models. In terms of text format, using both the prompted instance and the answer choice showed the best results. These results show that for the dense retriever to map instances, it should rely on both components, which are orthogonally important. Also, using the actual label option harms performance compared to using the answer choice, which indicates that the output format itself is important to retrieve well-matched expert LMs.

C. Details of Performing Compositional Instructions

Our compositional instruction setting consists of a total of 400 instances for each task (300 instances for the validation set, and 100 instances for the test set.) per language that was obtained using google translate to change the input of the **XL-Sum** (Hasan et al., 2021) dataset. We thus use the ground truth label in the specified language and the input is the machine-translated version. The reason for this is that we measure the λ_i values (the importance to place on each task vector τ_i) by performing evaluation on the validation datasets. Empirically, setting 1.0 for each λ_i value resulted in the best performance. Thus, as mentioned in the method section, the total $\sum \lambda_i$ results in 2.0, greater than 1.0.

We also vary the decoding strategies to check the performance of merging two experts finetuned from MT5-3B compared with naive MT0-3B on **XL-Sum** dataset. The detailed optimal setting we found is as follows:

- LAMBDA1: 1.0
- LAMBDA2: 1.0

- NO_REPEAT_NGRAM_SIZE:2
- TEMPERATURE:1.0
- EARLY_STOPPING:True
- DO_SAMPLE:True
- TOP_P:0.95

Here are the actual inputs for the LM generated & ground truth output examples shown in Table 8. The *compositional* instruction portion is shown in **bold**.

ENGLISH → SPANISH: "**Write a summary of the following English text and translate the sentence into Spanish:** The French police arrested four members of the child's family for their alleged involvement on Tuesday. Police sources told local media that the child refused to do his homework and that he was beaten with the stick of a broom. The 20 -year -old sister, his older brother and his girlfriend were present at the time of the incident and were arrested. The three called emergency services, which could not save the child. The alleged crime occurred on September 17 at the family's home in the town of Mulhouse, in the east of the country, and of just over 100,000 inhabitants. Although the child's mother was not at home because she was on a trip for work reasons, she was also arrested. The authorities say it will be questioned to confirm whether it encouraged the punishment. The four family members remain in police custody and must appear before the Mulhouse Prosecutor's Office for a judicial investigation. Prosecutor Edwige Roux-Morizot will investigate the case. Moretones after the death of the child, victim of cardiac arrest, several neighbors celebrated a vigil in their honor and met with the child's parents to offer them comfort. However, the results of the autopsy motivated the police to carry out an investigation into what happened. The child's body presented several bruises, especially at his feet, according to AFP. Despite the confirmation of cardiac arrest, pathologists said the cause of death was probably the blows he had suffered. A police source said the child was beaten with blunt objects. Although the main suspect of the murder is the older brother, the French authorities hope that the investigation will shed light on what happened. France is one of the 13 countries of the European Union where corporal punishment is legal. A legal practice The National Assembly of France is considering approveing a law to prohibit corporal punishments for children. There are two new law proposals that would grant children a violence -free education, venting parents to use "forms of humiliation such as physical or verbal..."

ENGLISH → FRENCH: "**Write a summary of the following English text and translate the sentence into French:** The former Minister of Justice of Malawi, Ralph Kasambara, was arrested on November 8, 2013. Mr. Kasambara was found guilty of conspiracy in the assassination in September 2013 of the former budget director at the Ministry of Finance, Paul Paul MPHWIYO. The murder of Mr. Mphwiyo had led to the discovery of the scandal of "cashgate", the systematic looting of public resources, during the administration of President Joyce Banda. Nearly 250 million had been fraudulently paid to businessmen for services who have never been rendered. A few days before the tragedy, a subordinate official would have been found with gold bars belonging to the cash, the equivalent of more than $ 300 million, in the trunk of his car. Money was also confiscated at the home of certain officials and in chests from their vehicles. Immediately after his conviction last month, Kasambara had suggested that he would not appeal the court verdict."

ENGLISH → JAPANESE: "**Write a summary of the following English text and translate the sentence into Japanese:** Vice Chairman Meng Ship, the highest financial manager (CFO), was the daughter of the founder arrested in Vancouver, Canada last December, and Vice President Meng Teng was sanctioned at Vancouver Airport last December. He was arrested for violating and associated scams and was charged at the end of January this year. The United States authorities are seeking to hand over the vice chairman, but they deny the charges. Defendant Meng filed an administrative lawsuit for the Canadian government, the immigration bureau, and the police for "significantly infringing" their citizenship. China has accused the defendant's arrest and delivery procedure as a "political project." ¡Related article¿ Introduction is "illegal" and "Dandridy" British Columbia Senior Court on the 1st, and Meng is the Canadian government and the Royal Canadian equestrian police (RCMP), and the Canadian Immigration Bureau (CBSA). He is complaining of civil rights infringement. Before the arrest of RCMP, CBSA complained that he had detained himself on unfair claims, investigated and interrogated his belongings. The vice chairman was bail and was at Vancouver's home, and the authorities arrested Vice Chairman Meng on the spot. He complained that it infringed on the rights based on the Canadian Characters of Human Rights. In addition, Vice -Chairman's detention was "illegal" and "arbitrary", and authorities pointed out that "the reason for detention, the right to call lawyers, or the right to be paid to be silent." What is the reaction of each country? The relationship between China, Canada and the United States has deteriorated over the arrest of Vice Chairman Meng. In January, the U.S. Department was charged with 23

cases of Huawei and Vice Chairman Meng. In addition to bank fraud, communication fraud, judicial obstruction, a major US telecommunications equipment T -mobile has been charged with trying to steal technology. China accused these movements as ”abuse of the handover agreement” between the United States and Canada, and stated that they...”

ENGLSIH → CHINESE: “**Write a summary of the following English text and translate the sentence into Chinese:** Dr. Craig Spencer, who is infected with Ebola virus, is currently being hospitalized at the New York Metropolitan Hospital. Caisyex said that the isolation experience is very scary and may also make other medical workers reluctant to go to West Africa to help curb the Ebola epidemic. Following New York and New Jersey, Illinois has also adopted a strict isolation policy. New measures means that those who have come into contact with any Ebola patient in West Africa will be forced to isolate for 21 days. U.S. President Obama Obama said in a weekly radio speech on September (October 25) that Americans should believe in the facts rather than being dominated. He also reiterated that he can infect the virus only with direct body fluids with Ebola patients. Higkos, who was ”criminals”, who was an isolated person, said that she had witnessed ”confusion, panic, and the most terrifying isolation” when she returned from Sierra Leone on Friday (24th). Hekox wrote a newspaper in the United States: ”I don’t know how many medical workers who fought with Ebola virus in the West African epidemic area will have the same encounter.” She said, ”Will they feel like criminals like criminals like criminals? She also said that she was isolated for seven hours at the airport terminal, but she only got a grain rod to fill her hunger. She denied that she had had a fever and said that she was just blushing at the time because she was not satisfied with the treatment at the airport. Even though Hiccoks was negative in Ebola virus testing, she was still being isolated for three weeks and was monitored by medical officials. Frontline medical staff was deeply influenced by the Ebola outbreak. After being diagnosed with Ebola patients, a doctor of New York, who had worked in Guinea last week, was diagnosed with Ebola patients, New York State and New Jersey have strengthened their isolation measures. Spencer is currently receiving isolation treatment in a hospital in New York. Mali has also recently appeared in Ebola, and President Ibrahim... ”

ENGLISH → KOREAN: “**Write a summary of the following English text and translate the sentence into Korean:** According to the Korea Meteorological Administration, January this year was the warmest winter since 1973, when the weather observation began in the Korean peninsula. The average temperature in January last month was 2.8 degrees. This is 3.8 degrees higher than the average of minus 1.0 degrees in January, 1981 2010. The previous average temperature record was 1.6 degrees in 1979. Except for the first day of the new year, the average temperature in the country was higher than normal. Due to the high temperature, the snowfall was the lowest. The Korea Meteorological Administration cited the introduction of warm southwestern air flow into the Siberian region, and the fact that the ’pole whirl’, which traps cold air in the Arctic, was strong as an abnormal temperature. It also analyzed that the warm south wind flow was introduced to the Korean peninsula due to the high sea level temperature of the Western Pacific. Nationwide weather data in January, the average temperature in the coldest January of the year has continued to rise in recent years. According to the weather data released by the Korea Meteorological Administration in January 1973-2020, the average temperature in January in Korea is steadily rising. Choi Jung -hee, the Korea Meteorological Agency, said that the warming of winter is ”global warming impact,” and ”most of the monthly weather data tends to be similar.” Detection of the ecosystem change is detected throughout the ecosystem. The first spawning season of ’Bukbangsan Guri’, a climate change indicator, has been faster. Mudeungsan National Park Eastern Office said on the 24th of last month that the first spawning of the North Bangsan Gogi, a species designated by the Ministry of Environment, was observed. It was first observed. It is 27 days earlier than February 19 last year. This is the first time that spawning has been observed in January since 2010, when the survey began. Researchers at the Park Industrial Complex believed that the spawning day was advanced due to the exceptionally warm... ”

D. Limitations and Discussions

While we highlight some of the major drawbacks of instruction tuning and propose an alternative approach of instead training and retrieving experts in this paper, we do not perform experimental results over MT LMS that have more than $>$11B parameters. For example, MT LMs with $>$11B parameters may be less susceptible to negative task transfer because of increased model capacity. Also, during the inference of unseen tasks, our retrieval mechanism assumes batch inference (i.e. having access to 32 samples of the target tasks without labels). Finally, when showing the compositional instruction experiments, we assume the two optimal experts could be retrieved from the compositional instruction (concatenation of the two seen instructions) given as the input along with the evaluation instance. This might not necessarily be the case with more complex, compositional instructions, which might require a separate *decomposition* stage. We instead focus on showing the possibility merging experts can bring and leave developing novel methods of retrieving the optimal experts during inference for future work.

E. Full List of PE and DE ranked on the 11 unseen datasets

Table 11 shows the full list of DE and Table 12 shows the full list of PE, both lists sorted in descending order with regards to the mean accuracy on 11 unseen tasks.

Dataset	AVG	Categories
cosmos_qa	53.35229377	MCQA
social_i_qa	52.9111819	MCQA
dream	51.45885188	MCQA
quail	50.4459655	MCQA
qasc	48.05781887	MCQA
paws/labeled_final	47.65196514	Paraph.
commonsense_qa	47.20113697	MCQA
sciq	47.07330356	MCQA
cos_e/v1.11	46.66113821	MCQA
quartz	46.65265672	MCQA
adversarial_qa/adversarialQA	45.62737167	EQA
wiki_qa	45.36088559	CBQA
glue/qqp	44.0165991	Paraph.
cnn_dailymail/3.0.0	43.98887691	Summ.
hotpot_qa/fullwiki	43.66845602	CBQA
xsum	43.62089761	Summ.
amazon_polarity	43.5926426	Senti.
ropes	43.45845826	EQA
quoref	43.41009006	EQA
rotten_tomatoes	43.35511468	Senti.
common_gen	43.1382362	STS
app_reviews	43.05588093	Senti.
wiki_bio	43.05367126	STS
samsum	42.7618847	Summ.
wiki_hop/original	42.67778976	MCQA
gigaword	42.61971626	Summ.
trec	42.46916224	Topic C.
dbpedia_14	42.21388133	Topic C.
multi_news	41.97036069	Summ.
ag_news	41.95621965	Topic C.
glue/mrpc	41.95418826	Paraph.
duorc/ParaphraseRC	41.94062218	EQA
imdb	41.70437975	Senti.
wiqa	41.1534245	MCQA
yelp_review_full	40.85474309	Senti.
quarel	40.59043188	MCQA

Table 11. The full list of Dataset Experts (DE) ranked in the mean accuracy on the 11 unseen tasks. The evaluations are performed on 300 sample instances of each unseen task for efficiency.

Dataset	Prompt	AVG	Task Category
cosmos_qa	no_prompt_text	54.65821845	MCQA
cosmos_qa	context_question_description_answer_text	54.3060466	MCQA
cosmos_qa	context_description_question_answer_text	54.19701579	MCQA
cosmos_qa	description_context_question_answer_text	53.15591518	MCQA
social_i_qa	Show choices and generate answer	53.06841536	MCQA
dream	baseline	51.85164999	MCQA
dream	read_the_following_conversation_and_answer_the_question	51.67431073	MCQA
cos_e/v1.11	description_question_option_text	50.65180447	MCQA
cosmos_qa	context_question_description_answer_id	50.48691808	MCQA
social_i_qa	Show choices and generate index	50.43707145	MCQA
cos_e/v1.11	description_question_option_id	50.29845396	MCQA
sciq	Multiple Choice (Closed Book)	50.12860827	MCQA
commonsense_qa	most_suitable_answer	50.06566011	MCQA
commonsense_qa	question_answering	49.96578376	MCQA
cosmos_qa	context_description_question_answer_id	49.89036173	MCQA
qasc	qa_with_separated_facts_1	49.14814303	MCQA
cos_e/v1.11	question_option_description_text	49.08282529	MCQA
sciq	Multiple Choice	48.73448898	MCQA
cosmos_qa	no_prompt_id	48.56936806	MCQA
cos_e/v1.11	question_option_description_id	48.55509469	MCQA
sciq	Multiple Choice Question First	48.50439309	MCQA
cosmos_qa	description_context_question_answer_id	48.22390771	MCQA
qasc	qa_with_separated_facts_2	48.2197083	MCQA
qasc	qa_with_combined_facts_1	48.12678008	MCQA
cos_e/v1.11	question_description_option_text	47.23675042	MCQA
paws/labeled_final	task_description-no-label	47.23675042	Paraph.
cos_e/v1.11	question_description_option_id	47.03021282	MCQA
social_i_qa	Check if a random answer is valid or not	46.98766238	MCQA
paws/labeled_final	Rewrite	46.90427355	Paraph.
quartz	paragraph_question_plain_concat	46.88892082	MCQA
paws/labeled_final	Concatenation	46.76229133	Paraph.
paws/labeled_final	context-question	46.69767805	Paraph.
paws/labeled_final	PAWS-ANLI GPT3-no-label	46.68362131	Paraph.
paws/labeled_final	Rewrite-no-label	46.66735722	Paraph.
quartz	given_the_fact_answer_the_q	46.65622609	MCQA
commonsense_qa	question_to_answer_index	46.59109421	MCQA
paws/labeled_final	Concatenation-no-label	46.51096254	Paraph.
paws/labeled_final	Meaning	46.15932052	Paraph.
paws/labeled_final	context-question-no-label	46.06366702	Paraph.
ropes	prompt_beginning	46.03684758	EQA
quartz	use_info_from_question_paragraph	46.00687505	MCQA
paws/labeled_final	Meaning-no-label	45.89445599	Paraph.
quartz	answer_question_below	45.70112461	MCQA
qasc	qa_with_separated_facts_4	45.63098518	MCQA
quartz	read_passage_below_choose	45.45333529	MCQA
dream	generate-last-utterance	45.43172606	MCQA
paws/labeled_final	PAWS-ANLI GPT3	45.33228586	Paraph.
quartz	use_info_from_paragraph_question	45.29788178	MCQA
ropes	plain_bottom_hint	45.21541083	EQA

Dataset	Prompt	AVG	Task Category
wiki_qa	Decide_good_answer	45.21394529	CBQA
wiki_qa	automatic_system	45.21245106	CBQA
quartz	answer_question_based_on	45.18935019	MCQA
wiki_qa	exercise	45.18589809	CBQA
ropes	prompt_mix	44.93591101	EQA
quartz	having_read_above_passage	44.91798422	MCQA
rotten_tomatoes	Reviewer Opinion bad good choices	44.78559829	Senti.
ropes	plain_no_background	44.53005571	EQA
wiki_qa	Generate Question from Topic	44.34881059	CBQA
ropes	new_situation_background_answer	44.34412958	EQA
adversarial_qa/adversarialQA	based_on	44.32984883	EQA
cos_e/v1.11	explain_why_human	44.30354265	MCQA
ropes	background_situation_middle	44.21042071	EQA
wiqa	effect_with_string_answer	44.16834366	MCQA
commonsense_qa	answer_given_question_without_options	44.051375	MCQA
trec	pick_the_best_descriptor	44.04277181	Topic C.
social_i_qa	Generate the question from the answer	44.04212031	MCQA
adversarial_qa/adversarialQA	answer_the_following_q	44.03344043	EQA
ropes	plain_background_situation	44.02607152	EQA
ag_news	classify_with_choices	44.01140825	Topic C.
wiki_qa	Topic Prediction - Question and Answer Pair	43.95941542	CBQA
trec	fine_grained_DESC_context_first	43.91506114	Topic C.
glue/qqp	quora	43.83700658	Paraph.
qasc	is_correct_1	43.81501204	MCQA
hotpot_qa/fullwiki	classify_question_type	43.80908285	CBQA
trec	which_category_best_describes	43.78976077	Topic C.
ropes	prompt_bottom_no_hint	43.66584294	EQA
cos_e/v1.11	aligned_with_common_sense	43.6452535	MCQA
app_reviews	convert_to_rating	43.58299315	Senti.
wiki_qa	Is This True?	43.57268207	CBQA
dbpedia_14	given_list_what_category_does_the_paragraph_belong_to	43.57149988	Topic C.
trec	fine_grained_HUM_context_first	43.53804635	Topic C.
cos_e/v1.11	i_think	43.52231837	MCQA
quarel	heres_a_story	43.5107948	MCQA
wiki_qa	Jeopardy style	43.46830805	CBQA
glue/qqp	answer	43.44450543	Paraph.
glue/qqp	duplicate or not	43.43977509	Paraph.
app_reviews	convert_to_star_rating	43.43198943	Senti.
quail	description_context_question_answer_text	43.42121948	MCQA
trec	trec1	43.41024144	Topic C.
app_reviews	generate_review	43.40556677	Senti.
glue/qqp	same thing	43.39970221	Paraph.
ropes	prompt_bottom_hint_beginning	43.37137146	EQA
yelp_review_full	so_i_would	43.35330514	Senti.
yelp_review_full	based_on_that	43.35330514	Senti.
yelp_review_full	format_star	43.35330514	Senti.
yelp_review_full	this_place	43.35330514	Senti.
yelp_review_full	format_score	43.35330514	Senti.
yelp_review_full	on_a_scale	43.35330514	Senti.
yelp_review_full	format_rating	43.35330514	Senti.
ropes	given_background_situation	43.35288364	EQA
adversarial_qa/adversarialQA	tell_what_it_is	43.34066211	EQA
wiki_qa	Direct Answer to Question	43.33163471	CBQA

Dataset	Prompt	AVG	Task Category
cos_e/v1.11	rationale	43.30650279	MCQA
glue/qqp	meaning	43.28847032	Paraph.
ag_news	which_section_choices	43.23247086	Topic C.
wiqa	effect_with_label_answer	43.22751795	MCQA
trec	fine_grained_NUM_context_first	43.20388525	Topic C.
ag_news	which_section	43.18354617	Topic C.
dbpedia_14	pick_one_category_for_the_following_text	43.17307426	Topic C.
dbpedia_14	given_a_list_of_category_what_does_the_title_belong_to	43.15357419	Topic C.
qasc	is_correct_2	43.13462473	MCQA
quail	context_question_answer_description_text	43.13447545	MCQA
quail	context_question_description_answer_text	43.13447545	MCQA
quail	context_question_description_text	43.13447545	MCQA
quail	context_description_question_text	43.13447545	MCQA
quail	no_prompt_text	43.13447545	MCQA
social_i_qa	Generate answer	43.13447545	MCQA
quail	context_description_question_answer_text	43.1284649	MCQA
quail	context_description_question_answer_id	43.10641223	MCQA
ropes	read_background_situation	43.09034626	EQA
ag_news	classify_with_choices_question_first	43.07124399	Topic C.
quail	description_context_question_text	43.06430021	MCQA
adversarial_qa/adversarialQA	question_context_answer	43.04578872	EQA
trec	fine_grained_open_context_first	43.04356783	Topic C.
dream	generate-first-utterance	43.04159372	MCQA
ropes	background_new_situation_answer	43.00629035	EQA
rotten_tomatoes	Reviewer Enjoyment Yes No	42.97922952	Senti.
quarel	do_not_use	42.96312743	MCQA
wiki_qa	Topic Prediction - Question Only	42.95471099	CBQA
quail	description_context_question_answer_id	42.91664826	MCQA
glue/qqp	duplicate	42.87048524	Paraph.
trec	fine_grained_ENTY	42.86820792	Topic C.
trec	fine_grained_LOC_context_first	42.8602283	Topic C.
glue/mrpc	generate_sentence	42.84869263	Paraph.
trec	fine_grained_NUM	42.82283103	Topic C.
imdb	Reviewer Expressed Sentiment	42.79593794	Senti.
sciq	Direct Question	42.79083732	MCQA
cos_e/v1.11	generate_explanation_given_text	42.74883356	MCQA
amazon_polarity	Is_this_review	42.74325079	Senti.
amazon_polarity	User_recommend_this_product	42.74325079	Senti.
amazon_polarity	Is_this_product_review_positive	42.74325079	Senti.
amazon_polarity	Is_this_review_negative	42.74325079	Senti.
amazon_polarity	convey_negative_or_positive_sentiment	42.74325079	Senti.
amazon_polarity	negative_or_positive_tone	42.74325079	Senti.
amazon_polarity	user_satisfied	42.74325079	Senti.
amazon_polarity	would_you_buy	42.74325079	Senti.
glue/mrpc	generate_paraphrase	42.74325079	Paraph.
amazon_polarity	flattering_or_not	42.7424637	Senti.
wiki_qa	found_on_google	42.73480328	CBQA
quoref	Guess Title For Context	42.73108831	EQA
trec	trec2	42.67551711	Topic C.
wiqa	what_is_the_final_step_of_the_following_process	42.66352026	MCQA
quarel	choose_between	42.63029283	MCQA
commonsense_qa	answer_to_question	42.62117703	MCQA
quoref	Guess Answer	42.61963732	EQA
imdb	Reviewer Enjoyment Yes No	42.59507536	Senti.

Dataset	Prompt	AVG	Task Category
qasc	qa_with_separated_facts_5	42.56217194	MCQA
cosmos_qa	context_question_description_text	42.53956247	MCQA
ag_news	classify_question_first	42.53946803	Topic C.
social_i_qa	I was wondering	42.52275144	MCQA
ag_news	recommend	42.5213931	Topic C.
imdb	Reviewer Opinion bad good choices	42.51140462	Senti.
wiki_qa	Topic Prediction - Answer Only	42.46767372	CBQA
qasc	qa_with_separated_facts_3	42.45034098	MCQA
trec	fine_grained_HUM	42.43031616	Topic C.
quail	context_question_answer_description_id	42.42340357	MCQA
quail	context_question_description_answer_id	42.42340357	MCQA
quarel	logic_test	42.42340357	MCQA
quail	no_prompt_id	42.41294351	MCQA
paws/labeled_final	paraphrase-task	42.38669957	Paraph.
xsum	DOC_write_summary_of_above	42.38486858	Summ.
xsum	article_DOC_summary	42.38486858	Summ.
xsum	DOC_how_would_you_rephrase_few_words	42.38486858	Summ.
xsum	college_roommate_asked_DOC_so_I_recap	42.38486858	Summ.
xsum	DOC_boils_down_to_simple_idea_that	42.38486858	Summ.
xsum	summarize_DOC	42.38486858	Summ.
xsum	summarize_this_DOC_summary	42.38486858	Summ.
cosmos_qa	context_description_question_text	42.3758318	MCQA
quoref	What Is The Answer	42.32137551	EQA
samsum	Generate a summary for this dialogue	42.31155911	Summ.
glue/mrpc	want to know	42.29343352	Paraph.
samsum	Given the above dialogue write a summary	42.27633128	Summ.
sciq	Direct Question (Closed Book)	42.26387475	MCQA
glue/mrpc	equivalent	42.26079671	Paraph.
glue/mrpc	paraphrase	42.24289148	Paraph.
glue/mrpc	replace	42.2412404	Paraph.
quoref	Context Contains Answer	42.23229654	EQA
quoref	Given Context Answer Question	42.2152412	EQA
quoref	Read And Extract ’	42.21343959	EQA
common_gen	sentence to concepts	42.14160703	STS
trec	fine_grained_open	42.13572199	Topic C.
quarel	testing_students	42.09377162	MCQA
hotpot_qa/fullwiki	generate_answer_affirmative	42.05311313	CBQA
hotpot_qa/fullwiki	generate_explanations_affirmative	42.05311313	CBQA
hotpot_qa/fullwiki	generate_answer_interrogative	42.05311313	CBQA
cosmos_qa	only_question_answer	42.03758485	MCQA
quoref	Found Context Online	42.02555959	EQA
trec	fine_grained_ABBR	42.01818176	Topic C.
samsum	To sum up this dialog	42.01224255	Summ.
common_gen	topics from the sentence	42.00149943	STS

Dataset	Prompt	AVG	Task Category
trec	fine_grained_DESC	41.97978705	Topic C.
gigaword	generate_summary_for_this	41.97889741	Summ.
gigaword	reverse_writing	41.97889741	Summ.
gigaword	make_a_title	41.97889741	Summ.
gigaword	first_sentence_title	41.97889741	Summ.
gigaword	TLDR	41.97889741	Summ.
gigaword	write_its_sentence	41.97889741	Summ.
gigaword	write_a_title_for_this_sentence	41.97889741	Summ.
gigaword	in_a_nutshell	41.97889741	Summ.
samsum	Write a dialogue that match this summary	41.97889741	Summ.
gigaword	write_an_article	41.93445375	Summ.
trec	fine_grained_ABBR_context_first	41.91844542	Topic C.
cnn_dailymail/3.0.0	write_an_outline	41.91841535	Summ.
cnn_dailymail/3.0.0	news_summary	41.91841535	Summ.
cnn_dailymail/3.0.0	2_or_3_sentences	41.91841535	Summ.
cnn_dailymail/3.0.0	tldr_summary	41.91841535	Summ.
cnn_dailymail/3.0.0	news_card_view	41.91841535	Summ.
cnn_dailymail/3.0.0	generate_story	41.91841535	Summ.
cnn_dailymail/3.0.0	sum_in_brief	41.91841535	Summ.
cnn_dailymail/3.0.0	news_stock	41.91841535	Summ.
quoref	Answer Friend Question	41.91841535	EQA
cnn_dailymail/3.0.0	spice_up_story	41.91295723	Summ.
trec	what_category_best_describe	41.89413219	Topic C.
wiqa	which_of_the_following_is_the_supposed_perturbation	41.8674171	MCQA
cosmos_qa	context_answer_to_question	41.86422765	MCQA
xsum	DOC_given_above_write_one_sentence	41.81263384	Summ.
xsum	read_below_DOC_write_abstract	41.81263384	Summ.
xsum	DOC_tldr	41.81263384	Summ.
rotten_tomatoes	Writer Expressed Sentiment	41.80526158	Senti.
imdb	Movie Expressed Sentiment 2	41.80336688	Senti.
wiki_hop/original	choose_best_object_interrogative_1	41.78715174	MCQA
wiki_hop/original	explain_relation	41.78715174	MCQA
wiki_hop/original	generate_object	41.78715174	MCQA
wiki_hop/original	generate_subject	41.78715174	MCQA
wiki_hop/original	choose_best_object_affirmative_1	41.78715174	MCQA
wiki_hop/original	choose_best_object_affirmative_3	41.78715174	MCQA
wiki_hop/original	generate_subject_and_object	41.78715174	MCQA
wiki_hop/original	choose_best_object_affirmative_2	41.78715174	MCQA
wiki_hop/original	choose_best_object_interrogative_2	41.78715174	MCQA
app_reviews	categorize_rating_using_review	41.7833793	Senti.
samsum	Summarize this dialogue:	41.78235121	Summ.
samsum	Sum up the following dialogue	41.75107511	Summ.
trec	fine_grained_LOC	41.73465262	Topic C.
rotten_tomatoes	Reviewer Expressed Sentiment	41.72418821	Senti.
glue/mrpc	same thing	41.72027244	Paraph.
wiqa	what_is_the_missing_first_step	41.70884339	MCQA
wiqa	what_might_be_the_first_step_of_the_process	41.6543053	MCQA
samsum	Summarize:	41.6530481	Summ.

Dataset	Prompt	AVG	Task Category
hotpot_qa/fullwiki	generate_title_affirmative	41.64987718	CBQA
hotpot_qa/fullwiki	generate_question	41.640237	CBQA
multi_news	summary scenario	41.62718689	Summ.
imdb	Writer Expressed Sentiment	41.60178406	Senti.
rotten_tomatoes	Reviewer Enjoyment	41.58082141	Senti.
dream	answer-to-dialogue	41.56118159	MCQA
cosmos_qa	description_context_question_text	41.5598663	MCQA
multi_news	what are the key points	41.55348468	Summ.
multi_news	distill	41.55348468	Summ.
ag_news	classify	41.52154068	Topic C.
rotten_tomatoes	Text Expressed Sentiment	41.51669372	Senti.
multi_news	expand (reverse task)	41.49696057	Summ.
rotten_tomatoes	Sentiment with choices ’	41.49571862	Senti.
wiqa	what_might_be_the_last_step_of_the_process	41.4814863	MCQA
multi_news	summarize	41.45677025	Summ.
multi_news	synthesize	41.428813	Summ.
common_gen	choice in concept centric sentence generation	41.40527643	STS
dbpedia_14	given_a_choice_of_categories ‘	41.39273021	Topic C.
rotten_tomatoes	Movie Expressed Sentiment	41.35952481	Senti.
rotten_tomatoes	Reviewer Sentiment Feeling	41.29297692	Senti.
imdb	Movie Expressed Sentiment	41.29017	Senti.
duorc/ParaphraseRC	build_story_around_qa	41.25012619	EQA
duorc/ParaphraseRC	decide_worth_it	41.25012619	EQA
duorc/ParaphraseRC	question_answering	41.25012619	EQA
duorc/ParaphraseRC	movie_director	41.25012619	EQA
duorc/ParaphraseRC	generate_question	41.25012619	EQA
duorc/ParaphraseRC	extract_answer	41.25012619	EQA
duorc/ParaphraseRC	title_generation	41.25012619	EQA
duorc/ParaphraseRC	answer_question	41.25012619	EQA
duorc/ParaphraseRC	generate_question_by_answer	41.25012619	EQA
common_gen	Put together	41.20526211	STS
quoref	Find Answer	41.12144463	EQA
rotten_tomatoes	Movie Expressed Sentiment 2	41.10981068	Senti.
quoref	Answer Question Given Context	41.09694099	EQA
wiki_bio	who	41.07422576	STS
imdb	Reviewer Sentiment Feeling	41.04883277	Senti.
adversarial_qa/adversarialQA	generate_question	40.97089459	EQA
wiqa	does_the_supposed_perturbation_have_an_effect	40.94586331	MCQA
quoref	Answer Test	40.88342121	EQA
imdb	Negation template for positive and negative	40.80008389	Senti.
common_gen	Given concepts - type 2	40.72623213	STS
imdb	Reviewer Enjoyment	40.70140793	Senti.
imdb	Sentiment with choices ’	40.60427787	Senti.
common_gen	topic to sentence	40.54846736	STS
imdb	Text Expressed Sentiment	40.53260931	Senti.
common_gen	Given concepts type 1	40.52827679	STS
common_gen	random task template prompt	40.3974667	STS
common_gen	Example prompt	39.6913846	STS

Table 12. The full list of Prompt Experts (PE) ranked in the mean accuracy on the 11 unseen tasks. The evaluations are performed on 300 sample instances of each unseen task for efficiency.