EXT5: TOWARDS EXTREME MULTI-TASK SCALING FOR TRANSFER LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the recent success of multi-task learning and transfer learning for natural language processing (NLP), few works have systematically studied the effect of scaling up the number of tasks during pre-training. Towards this goal, this paper introduces EXMIX (**Ex**treme **Mix**ture): a massive collection of 107 supervised NLP tasks across diverse domains and task-families. Using EXMIX, we study the effect of multi-task pre-training at the largest scale to date, and analyze co-training transfer amongst common families of tasks. Through this analysis, we show that manually curating an ideal set of tasks for multi-task pre-training is not straightforward, and that multi-task scaling can vastly improve models on its own. Finally, we propose EXT5: a model pre-trained using a multi-task objective of self-supervised span denoising and supervised EXMIX. Via extensive experiments, we show that EXT5 outperforms strong T5 baselines on SuperGLUE, GEM, Rainbow, Closed-Book QA tasks, and several tasks outside of EXMIX. EXT5 also significantly improves sample efficiency while pre-training.

1 INTRODUCTION

Transfer learning (Schmidhuber, 1987; Pratt et al., 1991; Caruana et al., 1995) has been the cornerstone of recent progress in natural language processing (Ruder et al., 2019; Devlin et al., 2019; Raffel et al., 2020). While self-supervised pre-training has been shown to be highly effective at exploiting large amounts of unlabeled data without relying on human annotation, there is still much to explore regarding transfer learning in a multi-task co-training setup.

Prior seminal works like T5 (Raffel et al., 2020) and MT-DNN (Liu et al., 2019a) have demonstrated a degree of promise in the paradigm of multi-task co-training (Caruana, 1997). However, the challenge of catastrophic forgetting remains. Tasks often have to be carefully selected in order to demonstrate positive affinity with regards to downstream transfer. In many cases, it is not unreasonable to expect negative transfer (Rosenstein et al., 2005; Vu et al., 2020). This makes the process of empirically curating a set of tasks to include in a transfer learning setup both computationally prohibitive and specific to downstream tasks.

While standard pre-training typically employs a variant of the self-supervised language modeling objective (Raffel et al., 2020), certain types of skills such as commonsense knowledge are only acquired at a slow rate even using massive amounts of unlabeled data (Zhang et al., 2021). As ever larger models are trained, the development of much more sample-efficient pre-training settings becomes thus more important, and could be addressed via multi-task learning.

For the first time, we explore and propose *Extreme Multi-task Scaling* — a new paradigm for multi-task pre-training. Compared to the largest prior work (Aghajanyan et al., 2021), our study doubles the number of tasks and focuses on multi-task pre-training rather than fine-tuning, which enables a direct comparison to standard pre-training. Our proposal is based on the insight that despite negative transfer being common during fine-tuning, a massive and diverse collection of pre-training tasks is generally preferable to an expensive search for the best combination of pre-training tasks.

To this end, we introduce EXMIX: a massive collection of 107 supervised NLP tasks to be included in a multi-task pre-training setup. We process all tasks in an encoder-decoder friendly format to readily support the sharing of all parameters across all tasks. We postulate that an *ensembling* effect across as many tasks, distributions and domains as possible results in a consistently net-positive

outcome. This echoes early multi-task learning results (Caruana, 1997; Baxter, 2000) indicating that a bias that is learned on sufficiently many tasks is likely to generalize to unseen tasks drawn from the same environment. Moreover, our experiments verify that our ExMix mixture outperforms a best-effort mixture of manually curated tasks.

Finally, we propose ExT5: a T5 model (Raffel et al., 2020) pre-trained on a mixture of supervised ExMix and self-supervised C4 span denoising. ExT5 outperforms state-of-the-art T5 models on well-established benchmarks such as SuperGLUE (Wang et al., 2019a), GEM (Gehrmann et al., 2021), and Rainbow (Lourie et al., 2021); as well as Closed-Book QA (Roberts et al., 2020) tasks. Notably, our experimental findings also suggest that including ExMix may reduce the number of pre-training steps required to achieve strong performance, bringing about substantial sample efficiency benefits.

To summarize, this paper contributes the following:

- We propose ExMix (§2): a collection of 107 supervised NLP tasks for *Extreme Multi-task Scaling*, formatted for encoder-decoder training. ExMix has approximately twice as many tasks as the largest prior study to date (Aghajanyan et al., 2021), totaling 18M labeled examples across diverse task families.
- Given this large collection of tasks, we conduct rigorous empirical studies evaluating transfer between common task families (§2.1). Our experiments show that curating a pre-training mixture based on fine-tuning transfer is not straightforward (§2.2). Hence, efficiently searching for the best subset of tasks to include in a multi-task pre-training setup is challenging and prohibitive.
- Using ExMix, we pre-train a model alongside the C4 span-denoising objective introduced by Raffel et al. (2020), resulting in a new pre-trained model which we call ExT5 (§3). ExT5 outperforms state-of-the-art T5 on well-established benchmarks such as SuperGLUE, GEM, Rainbow, Closed Book Question Answering, and several other tasks that are outside of ExMix (§3.2), while also being more sample-efficient (§2.6).

2 The ExMix Task Collection

To explore the Extreme Task Scaling paradigm, we introduce ExMix (**Ex**treme **Mix**ture), a collection of 107 diverse English NLP tasks totaling 18M examples. Following Raffel et al. (2020), we format all tasks as text-to-text examples to readily allow for multi-task training. This unified format also enables simple implementations without the need for task-specific heads/losses, loss scaling, or explicit gradient accumulation for heterogenous batches as in prior works (Liu et al., 2019a; Aghajanyan et al., 2021). When selecting examples from ExMix, examples from each dataset are sampled proportionate to the individual dataset's size, with each dataset's sampling rate capped at 3×10^5 maximum effective examples to ensure a balance between large and small datasets. We refer readers to Appendix A for a comprehensive breakdown of ExMix. Additionally, we discuss future multilingual variants of ExMix and ExT5 in §5.

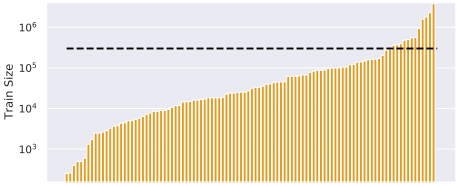


Figure 1: ExMix task sizes in log scale. The dashed line is the 3×10^5 sampling rate cap.

Although tasks in ExMix come from various datasets and all have their own nuances, ExMix tasks broadly represents the following *task families*: classification (single segment), natural language inference (two segments), reading comprehension, closed-book question answering (CBQA), commonsense reasoning, semantic parsing, dialogue, and summarization. While other task groups exist as well, the aforementioned task families cover most of ExMix. This task collection also enables multi-task learning at a scale twice that of previous work (Aghajanyan et al., 2021).

2.1 Transfer Relations Between ExMix Tasks

As discussed in §1, our goal is to pre-train a model on ExMix to improve downstream performance. One natural question to ask is *which tasks have a negative impact on downstream performance?*

Specifically, is there a subset of ExMix that leads to better representations when used for multi-task pre-training? Obviously, testing all possible $2^{|\text{ExMix}|}$ combinations is impractical since the pre-training process is expensive to run. Instead, we experiment with the less expensive *co-training* procedure (i.e., multi-task fine-tuning), using representative subsets of similar tasks. Later, in §2.2, we explore whether these results can be used to inform the task selection for multi-task pre-training.

To study transfer amongst task-families in ExMix, we construct subsets (Table 1) of 3 tasks each that are partitioned along their task family. Using these subsets, we evaluate transfer among task families in a multi-task learning (co-training) setup. While other types of tasks are available in ExMix, we did not include them because they were not diverse enough to be representative of a task family, and would scale the number of models needing to be trained at a quadratic rate.

Task Family	Datasets
Summarization	CNN/DailyMail XSum Wiki Lingua
Dialogue	Schema-guided dialogue Wizard-of-Wikipedia Dialoglue-TOP
NLI	ANLI MNLI αNLI
Classification	IMDb reviews GoEmotions Civil Comments
Semantic Parsing	ATIS to FunQL GEO to FunQL COGS
Commonsense	PhysicaliQA SocialiQA WinoGrande
Closed-Book QA	Natural Questions Trivia QA Hotpot QA
Reading Comprehension	SQuAD BoolQ TweetQA

Table 1: Representative datasets used for task-family transfer learning experiments (§2.1).

Experimental Setting We fine-tune a model on each pair of task families (i.e., 6 datasets at a time). To ensure a fair balance of tasks, we sample tasks proportional within their family, but uniformly between task families. For example, while evaluating how classification tasks and NLI tasks transfer amongst each other, the sampling ratio of MNLI:ANLI will be proportional (approximately 2.4:1), but the overall ratio of NLI examples to classification examples will be 1:1. For reference, we also train a model on each individual task family using proportional example mixing (Sanh et al., 2019).

In total, this results in $F + \binom{F}{2}$ models trained, where F is the number of task families. Our experiments use $F = 8$ as shown in Table 1, resulting in 34 models trained in total. Each model is fine-tuned on top of the released $\text{T5.1.1}_{\text{BASE}}$ checkpoint for 200k steps using a batch size of 128 and a constant learning rate of 10^{-3}.

Observations We summarize our results in Table 2. We observe that although there exist particular task-family pairs that show positive transfer (e.g., co-training with NLI helps most other tasks), negative transfer is more common when training across task families compared to intra-family training. 21 out of the 56 inter-family relationships perform worse than intra-family models with the same data budget, which grows to 38 out of 56 for a fixed compute-budget. While the abundance of negative transfer among diverse task families is an expected result, interesting trends manifest in the individual relationships. For example, summarization tasks generally seem to hurt performance on most other task families; and CBQA tasks are highly sensitive to multi-task fine-tuning.

We also report correlations for intra-family datasets in Figure 2 using the same models as in Table 2. In most cases, we see positive correlations between datasets in the same family. In a few cases, however, we observe an opposite trend. For example, fine-tuned models that performed better on the GEM schema-guided dialog dataset achieved lower scores on KILT Wizard-of-Wikipedia.

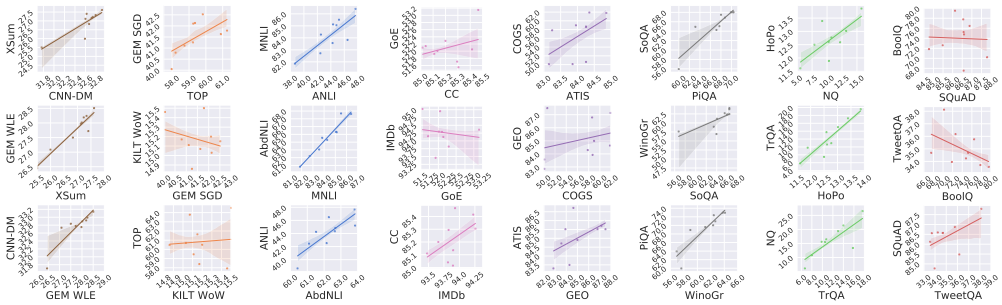
This initial exploratory analysis highlights both the potential of ExMix as a tool to systematically study task relationships, as well as the potential challenges in leveraging multi-task learning naively on top of pre-trained representations.

2.2 Can fine-tuning task relationships help curate a pre-training mixture?

Our observations in §2.1 showed that multi-task co-training on top of existing pre-trained checkpoints is not straightforward, and often results in negative transfer. However, the uncovered task relationships might help efficiently search for an ideal subset of ExMix for multi-task pre-training. To this end, we select a set of the most promising task families to be included in a multi-task pre-training setup, ranking task families by the average relative improvement they provide to other target families (the last column in Table 2). Specifically, we include NLI, commonsense, classification,

	SUM	DLG	NLI	CLS	SEM	CMNS	CBQA	RC	Δ_{AVG}
SUM	27.89 29.36	37.81	60.45	77.10	78.25	61.92	7.84	65.37	-6.9%
DLG	29.05	38.56 39.76	63.62	77.10	75.55	64.05	13.39	64.75	+0.1%
NLI	28.61	40.60	64.91 67.23	77.29	77.72	67.60	15.24	66.40	+4.3%
CLS	29.52	40.16	66.69	77.14 77.47	76.05	65.29	12.93	65.20	+1.4%
SEM	29.30	38.86	62.46	76.83	72.09 72.79	57.84	12.44	63.52	-2.5%
CMNS	29.28	39.27	65.08	77.05	76.29	68.24 68.35	16.48	66.01	+4.7%
CBQA	29.75	39.29	64.96	77.66	75.21	66.84	14.68 19.98	66.37	+1.2%
RC	29.45	38.12	63.70	77.14	76.98	66.62	10.26	62.94 65.60	-2.4%
$\text{AVG}_{\setminus\text{diag}}$	29.28	39.16	63.77	77.17	76.43	64.31	12.65	65.37	

Table 2: Co-training transfer among task families. The entry at (row i, column j) indicates average performance on family j using a model co-trained on families i and j. For intra-family models (diagonal cells) we report results upto 100k steps (consistent data-budget) and 200k steps (consistent compute-budget). Averages are calculated excluding the intra-family models (i.e. the diagonal cells). The last column denotes the average gain that a source family provides to other task families.



(a) SUM (b) DLG (c) NLI (d) CLS (e) SEM (f) CMNS (g) CBQA (h) RC

Figure 2: Within-family correlations for each dataset in a task family, using models from Table 2. Performance on datasets from some task families are highly correlated (e.g., NLI) whereas other task families have more erratic results across their datasets (e.g., Dialogue)

and closed-book QA tasks from ExMix to form a mixture of 48 tasks to include in a multi-task pre-training setup. We then fine-tune the resulting model on SuperGLUE, comparing it to T5.1.1 and ExT5 in Table 3.

While this model narrowly outperformed T5.1.1, it did not yield better results than including all of ExMix in the multi-task pre-training mixture, as we report in §3.2. Moreover, it did not outperform a random selection of 55 tasks on average, as we report in our task scaling experiments (§2.5).

We conclude that the **negative transfer during multi-task fine-tuning does not necessarily inhibit pre-training**. While we cannot directly conclude that an ideal subset of ExMix does not exist to be mixed with self-supervised pre-training for specific downstream tasks, our experiments show that randomly including more diverse pre-training tasks generally improves downstream performance. It must also be noted that the end-goal is to find a mixture that leads to a *general* pre-trained model that can be used for a large variety of downstream tasks, and that a setup to find a pre-training mixture tailored for SuperGLUE would be different.

Mixture	# Tasks	SuperGLUE
Vanilla	0	76.1
Best-effort	48	76.4
Random-55 (§2.5)	55	77.0
ExMix (§3.2)	107	**79.9**

Table 3: A best-effort mixture from fine-tuning transfer results does not beat increasing the number of tasks.

2.3 MULTI-TASK PRE-TRAINING VS PRE-FINETUNING

Instead of pre-training on EXMIX, another way to leverage multi-task learning is as an intermediate step between pre-training and fine-tuning. This is referred to as *pre-finetuning* by Aghajanyan et al. (2021). We conduct controlled experiments to compare pre-training with pre-finetuning. We begin with a standard T5 base checkpoint and pre-finetune it with EXMIX. After this phase, we fine-tune on SuperGLUE.

Table 4 compares pre-finetuning and our proposed multi-task pre-training. We also report the total compute (in total number of tokens processed) by the model in both schemes. The results show that multi-task pre-training is significantly superior to pre-finetuning. A potential hypothesis is that multi-task pre-training narrows the gap between pre-training and finetuning data distributions, as the pre-training stage more closely resembles fine-tuning. Conversely, segregating pre-training and pre-finetuning into two different stages may induce catastrophic forgetting of the pre-training task. Hence, in EXT5, we opt for multi-task pre-training over pre-finetuning.

Method	Compute	SuperGLUE
Vanilla	1.2M	76.1
Pre-finetuning (200k)	1.4M	78.1
Multi-task Pre-training	1.2M	**79.9**

Table 4: Comparison of Pre-finetuning and Multi-task Pre-training on EXMIX.

2.4 HOW MUCH LABELED DATA SHOULD BE MIXED?

In this section, we explore how the ratio of C4 to EXMIX examples during massive multi-task pre-training affects performance. As mentioned later in §3, this is controlled by a hyperparameter R, where a pre-training batch will have approximately R times as many C4 examples compared to EXMIX. From our results in Figure 3, we find that despite EXMIX improving downstream performance when mixed with self-supervised C4 pre-training at many rates, a model trained with $R = 0$ suffers greatly in comparison. This result is significant, as it shows that while EXMIX improves the pre-training process, self-supervised training over a large unstructured corpus is still crucial.

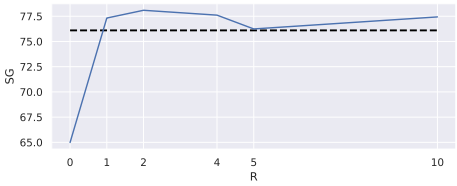


Figure 3: How the ratio of C4 span denoising examples to EXMIX affects SuperGLUE results on $\text{EXT5}_{\text{BASE}}$. The dashed line is performance without using EXMIX ($R \to \infty$)

2.5 DOES ADDING MORE TASKS HELP? TASK SCALING EXPERIMENTS

In this section, we explore how model performance changes as the number of tasks included in a massive multi-task pre-training setup is scaled up. We choose random sets of 30, 55, and 80 tasks (each a superset of the last), pre-train a BASE-sized model for 524k steps, and fine-tune them on SuperGLUE. We train our models with batch sizes of 128 and 512 and $R = 2$ (the ratio of C4 to EXMIX examples) as this configuration worked best for our BASE-sized models (§2.4). We repeat this over three random seeds, and report average scores in Figure 4.

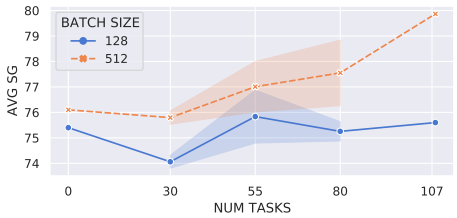


Figure 4: Scaling the number of tasks during multi-task pre-training generally helps,

Overall, with large batches, we can see that increasing the number of tasks being mixed generally helps downstream performance. This reinforces our intuition that task scaling indeed helps. With small batches, there is less of an upward trend, signifying that large batches are essential for a large number of tasks. This is intuitive, given that multi-task learning may cause gradients to be noisy (Yu et al., 2020). Another explanation as to why this happens is that large-batch training can offer benefits even for single-task models (Smith et al., 2018) — a trend formalized by McCandlish et al. (2018).

2.6 IMPROVING SAMPLE EFFICIENCY WITH EXMIX

We hypothesize that extreme multi-task scaling also improves the sample efficiency of pre-training. To test this, we exclude SuperGLUE from EXMIX, pre-train a large model for 200k steps, and fine-tune it on SuperGLUE at several intervals during early pre-training stages. We find that EXMIX pre-training is significantly more sample-efficient than vanilla self-supervised pre-training. Note that at only 20k pre-training steps, our ExT5 model already achieves 75.8 SuperGLUE score, which outperforms a **fully pre-trained** BERT large model by about $+4\%$ (Wang et al., 2019a).

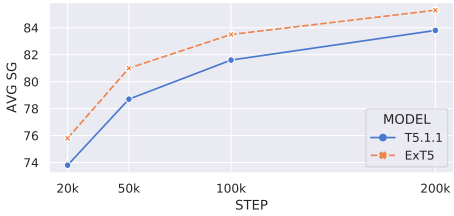


Figure 5: SuperGLUE score of $\text{ExT5}_{\text{LARGE}}$ vs T5_{LARGE} as a function of number of pre-training steps.

3 THE EXT5 MODEL

To overcome the challenges of multi-task co-training at scale, i.e. negative transfer and catastrophic forgetting explored in §2.1, the rest of this paper revisits the multi-task pre-training paradigm introduced by Raffel et al. (2020) via extreme multi-task scaling. This section introduces EXT5: a pre-trained sequence-to-sequence Transformer encoder-decoder model (Vaswani et al., 2017) based on the popular T5 framework.

3.1 TRAINING EXT5

Pre-training We pre-train on a mixture of C4 and EXMIX (§2), and combine them with a hyper-parameter R that is the ratio at which C4 examples are sampled with respect to EXMIX examples. The C4 objective we use is the same as that used by Raffel et al. (2020), and every task optimizes the standard sequence-to-sequence cross-entropy loss. We pre-train EXT5 on the same number of steps as T5, and EXT5 sees an identical number of tokens to the released T5 models. Concretely, we pre-train our models for 1M total steps with a batch size of 2048 and sequence length 512, resulting in a total of approximately 1T tokens seen by the model during pre-training (both unsupervised and supervised inclusive). We use the T5.1.1 architecture (Shazeer, 2020) for all of our experiments — which uses GEGLU-activated layers instead of ReLU in classic Transformer models (Vaswani et al., 2017). For optimization, we use Adafactor with an inverse square root learning rate schedule that kicks in after a a constant phase of 0.01 for 10k steps.

Fine-tuning We follow the same fine-tuning procedure for T5 and EXT5 for fair comparison, although we found that EXT5 generally benefitted from a smaller learning rate while fine-tuning (10^{-4} worked well for EXT5 vs 10^{-3} for T5 variants). Fine-grained details can be found in Appendix B.

3.2 EXPERIMENTAL SETUP

Our experiments consider both within-mixture and out-of-mixture tasks (i.e., whether a task is included in EXMIX). Within-mixture tasks measure the amount the task benefits from multi-task pre-training and extreme task scaling. Similar to the co-trained models in Raffel et al. (2020), we continue to fine-tune on the target task from a pre-trained ExT5 checkpoint. For out-of-mixture tasks, we consider possibly new unseen tasks or collections that were not included in the EXMIX mixture to test the effect of generalizing to unseen tasks. For the sake of brevity, the fine-grained details of these experimental setups can be found in the Appendix.

3.3 EXPERIMENTAL RESULTS

WITHIN-MIXTURE RESULTS

We report results on **SuperGLUE** (Table 5), **GEM** (Table 6), **Rainbow** (Table 7), **MsMarco** (Table 8) and **CBQA** datasets (Table 9). On the whole, we observe that ExT5 consistently outperforms strong T5 baselines across a range of model sizes. On SuperGLUE, we achieve $+5\%$, 2.3% and

$+0.7\%$ gain on BASE, LARGE and XL respectively. On GEM, ExT5 outperforms T5 on 6 out of 9 collections while remaining on-par on the other 3 collections. Notably, the gain on datasets such as WebNLG are approximately $+11\%$ ROUGE for the large model and generally range from $+1\%$ to $+6\%$ on different collections. On Rainbow, ExT5 outperforms our own run of T5 by +0.7% on average and $+4.6\%$ improvement over the best multi-task (sequential) setup in Lourie et al. (2021). Finally, on question answering and ranking, ExT5 substantially outperforms T5 at two different sizes.

Model	BoolQ	CB	Copa	MultiRC	ReC	RTE	WiC	WSC	AVG
$T5.1.1_{BASE}$	82.3	91.7/92.9	60.0	76.9/39.6	80.9/80.2	84.5	69.3	81.7	76.1
$ExT5_{BASE}$	**82.6**	**98.7/98.2**	**73.0**	**79.5/45.4**	80.8/80.0	**87.0**	**71.3**	**83.7**	**79.9**
$T5.1.1_{LARGE}$	88.3	94.3/96.4	87.0	85.4/55.1	89.2/88.5	90.6	73.5	88.5	85.3
$ExT5_{LARGE}$	**88.4**	**98.7/98.2**	**89.0**	**85.5/58.0**	88.6/87.9	**93.1**	73.4	**96.2**	**87.3**
$T5.1.1_{XL}$	89.6	92.0/96.4	96.0	88.2/64.1	92.4/91.7	91.7	74.3	95.2	88.7
$ExT5_{XL}$	89.4	**100/100**	**97.0**	87.5/62.7	91.4/90.9	**94.2**	**74.6**	93.3	**89.4**

Table 5: Comparisons of T5 and ExT5 on SuperGLUE validation sets.

Model	Metric	WebNLG	DART	SGD	E2E	CG	ToTTo	WiA-A	WiA-T	WLE
$T5.1.1_{BASE}$	METEOR	0.323	0.364	0.325	0.383	0.201	0.366	0.302	0.368	0.189
	ROUGE-2	39.46	45.62	36.25	47.40	17.32	49.8	38.58	**51.54**	19.19
	BLEU	29.06	34.75	33.44	43.17	8.34	39.59	29.53	42.71	14.72
$ExT5_{BASE}$	METEOR	**0.349**	**0.367**	**0.330**	0.382	**0.206**	**0.368**	**0.306**	0.367	**0.192**
	ROUGE-2	**45.07**	**46.87**	**37.46**	47.32	**18.13**	**50.17**	**39.10**	51.35	**19.41**
	BLEU	**32.36**	**35.15**	**34.34**	42.71	**9.39**	**40.01**	**30.04**	**43.39**	**14.96**
$T5.1.1_{LARGE}$	METEOR	0.344	0.363	0.324	0.382	0.202	0.368	0.301	0.362	0.196
	ROUGE-2	43.31	45.22	36.17	46.60	17.01	49.90	38.37	50.52	20.47
	BLEU	31.67	34.31	33.15	42.57	8.38	39.79	29.30	42.12	15.55
$ExT5_{LARGE}$	METEOR	**0.365**	**0.376**	**0.330**	0.381	**0.214**	**0.369**	0.300	0.358	**0.204**
	ROUGE-2	**48.17**	**48.14**	**37.77**	**46.70**	**19.04**	**50.33**	37.98	50.38	**21.16**
	BLEU	**35.03**	**36.62**	**34.74**	42.25	**9.68**	**40.14**	29.23	41.39	**16.64**

Table 6: Comparisons of T5 and ExT5 on GEM (English) test sets.

Model	αNLI	CosmosQA	HellaSwag	PIQA	SocialIQA	Winogrande	AVG
$T5_{LARGE}$ (multitask)†	78.40	81.10	81.30	80.70	74.80	72.10	78.07
$T5_{LARGE}$ (sequential)†	79.50	83.20	83.00	82.20	75.50	78.70	80.35
$T5.1.1_{LARGE}$	**82.51**	85.59	88.57	**85.53**	78.51	79.79	83.42
$ExT5_{LARGE}$	82.25	**85.86**	**88.99**	85.04	**79.73**	**82.53**	**84.07**
% Gain	-0.3%	+0.3%	+0.5%	-0.6%	+1.6%	+3.4%	+0.8%

Table 7: Results on the Rainbow Commonsense Reasoning benchmark validation sets. Results with † are from Lourie et al. (2021).

Model	MRR@10
$T5_{LARGE}$ (Nogueira et al., 2020)	0.393
$ExT5_{LARGE}$	**0.402**
% Gain	+2.3%
$T5_{XL}$ (Nogueira et al., 2020)	0.398
$ExT5_{XL}$	**0.403**
% Gain	+1.3%

Table 8: Results on MSMarco.

Model	NQ	WQ	TQA
$T5.1.1_{LARGE}$	27.3	29.5	28.5
$ExT5_{LARGE}$	**28.6**	**30.5**	**30.7**
% Gain	+4.8%	+3.4%	+7.7%
$T5.1.1_{XL}$	29.5	32.4	36.0
$ExT5_{XL}$	**30.6**	**35.2**	**37.0**
% Gain	+3.7%	+0.9%	+2.8%

Table 9: Results on CBQA dev sets.

OUT-OF-MIXTURE RESULTS

We are also interested in evaluating EXT5 on tasks outside of EXMIX, and hypothesize that the extreme multi-task pre-training of EXT5 will lead to better performance on new unseen settings. Concretely, we fine-tune and evaluate on **Machine Translation**: translating sentences from English to other languages (Bojar et al., 2014; 2015; 2016); **Reasoning**: answering scientific questions on ARC (Clark et al., 2018); and **Named Entity Recognition**: extracting all entities from sentences on the CoNLL-2003 NER dataset (Tjong Kim Sang & De Meulder, 2003).

Model	Machine Translation			QA	NER	
	EnDe	EnFr	EnRo	ARC	Dev	Test
Raffel et al. (2020)	26.98	39.82	27.65	-	-	-
$T5.1.1_{BASE}$	28.30	41.01	28.11	26.45	92.55	85.75
$ExT5_{BASE}$	**28.32**	**41.89**	**28.38**	**36.35**	**92.68**	**86.53**
% Gain	±0%	+2.1%	+1.0%	+37.4%	+0.13%	+0.91%
$T5.1.1_{LARGE}$	28.68	41.34	29.01	55.80	92.80	86.56
$ExT5_{LARGE}$	**28.98**	**42.71**	**29.49**	**63.99**	**93.63**	**87.34**
% Gain	+1.0%	+3.3%	+1.7%	+14.7%	+0.90%	+0.90%

Table 10: Experimental results on tasks that are not in EXMIX. For ARC, we report test scores on the challenge set with retrieval. For NER, we report accuracy on a sentence level (see Appendix B.2).

Table 10 summarizes the results on the out-of-mixture tasks. Across all tasks, we see that EXT5 outperforms upon T5 baselines. The largest improvement is on the ARC scientific reasoning task, perhaps due to the large amount of QA tasks in EXMIX. Though, the trend is consistent also with the NER and MT tasks that do not have any similar dataset in EXMIX. This suggests that the representations learned by EXT5 are more general adaptable to a new objective, even when the output is in a new language.

This improved generalization of EXT5 is very encouraging from a practical stand point, since pre-training again with EXMIX $\cup \{t\}$ for any new target task t would be very expensive. Instead, we see that the extreme multi-task pre-training of EXT5 already provides improved results. Therefore, it might only be worth repeating pre-training when the collection of training datasets grows by a significant amount (see §2.5).

4 RELATED WORK

Improving NLP models with Multi-task Learning Phang et al. (2019) proposed an intermediate fine-tuning stage on labeled data, using four tasks for which ample training data is available. Similarly, Liu et al. (2019a) proposed MT-DNN, which uses a setup at a scale of around 30 tasks and up to 440M parameters. Most recently, Aghajanyan et al. (2021) use around 50 tasks and models of sizes upto 440M parameters. Gururangan et al. (2020) take an alternative approach, which is to continue pre-training a model but use domain-specific data as an intermediate step. Lourie et al. (2021) proposed a universal commonsense model co-trained on various commonsense-based question answering or language inference datasets. Recently, Wei et al. (2021) also illustrated how a multi-task learning stage can greatly improve the zero-shot prompting performance of large language models at the scale of ~137B parameters. Efforts have also been made to tailor pre-training objectives to specific tasks, e.g., question answering (Ram et al., 2021; Jia et al., 2021), dialogue (Li et al., 2020), and span selection tasks (Joshi et al., 2020).

Relationships amongst different tasks Bingel & Søgaard (2017) conducted a study similar to ours in §2.1 but for more traditional NLP tasks like chunking, CCG tagging, POS tagging, etc. More recently, Vu et al. (2020) conducted an in-depth study of relationships between various classification/regression, question-answering, and sequence-labeling tasks, and proposed a task-embedding framework to predict such relationships. Khashabi et al. (2020) also conducted similar experiments but specific to question-answering datasets/formats, resulting in a strong QA model known as UnifiedQA that is also based on the T5 framework. Outside of NLP, Zhang & Yeung (2010) introduced a convex optimization objective for *learning* task relationships, and Li et al. (2018) explore and exploit task relationships on a variety of diverse datasets.

Choosing which tasks to transfer from Our experiments in §2.2 attempted to empirically select a set of tasks to transfer from. Along these lines, Ruder & Plank (2017) use a Bayesian Optimization method with similarity measures to automatically select relevant data from different domains. Similarly, Guo et al. (2019) use multi-armed bandits to select tasks and a Gaussian Process to control the mixing rates for the selected tasks. Another strand of recent work selects appropriate transfer languages based on manually defined features (Lin et al., 2019; Sun et al., 2021). Aside from the NLP domain, Fifty et al. (2021) proposed a method to select which tasks to transfer to based on task gradients. All of the aforementioned works select data tailored to a downstream task of interest. If a general pre-trained model was attempted to be trained in a similar fashion, computational bottlenecks similar to those motivating §2.1 and §2.2 would arise.

Pre-trained Transformers Transformer models (Vaswani et al., 2017) such as T5 (Raffel et al., 2020), BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) rely on large unlabeled corpus for self-supervised learning. Given the wild success of the pre-train-finetune paradigm, the search for suitable pre-training tasks has also become an active area of research (Lewis et al., 2019; Lan et al., 2019; Chang et al., 2020; Zhang et al., 2019; Lourie et al., 2021). While there has been evidence that supplementary pre-training tasks can help improve performance, this work is the first massive-scale multi-task pre-trained model.

Scaling Laws Scaling laws for Transformers have attracted much attention recently, especially pertaining to model size (Kaplan et al., 2020; Zhai et al., 2021; Tay et al., 2021a). In Kaplan et al. (2020), the authors further investigate scaling with respect to dataset size (on the same pre-training corpus). To this end, this work can be interpreted as an attempt of scaling up with respect to the number of high quality, diverse labeled tasks that can be used for pre-training.

5 Epilogue

Limitations Despite our best efforts to evaluate on as many representative tasks as possible while also maintaining a balance among task partitions for a given set of transfer learning experiments, any study that explicitly abstracts datasets into "task families" is highly dependent on nuances pertaining to the nature, domain, and expressiveness of the task family's representative datasets. For this paper, the subsets were constructed so as to include a diverse set of datasets to evaluate on, and we tried to partition task-families to be as mutually exclusive as possible. However, it must be acknowledged that no dataset is perfectly isolated, and any set of them only a proxy for a larger "task family". On a separate note, lexical metrics like BLEU/ROUGE are useful but do not paint the full picture of how well a model truly performs on text-generation tasks.

Future Work We believe that a multilingual version of ExT5 would be a natural extension of this work. Such a model will require extra care with regard to balancing not only task families, but also task languages. A multilingual version of ExMix could provide a more robust foundation for the analysis of task families in existing works that analyze how multilingual NLP models transfer amongst different languages (Kudugunta et al., 2019; Hu et al., 2020; Wang et al., 2021). For example, it would be interesting to understand whether our results in §2.1 hold across different languages (and language families), and to explore cross-lingual cross-task generalization. We also hypothesize that modeling innovations that introduce inductive biases designed to exploit multi-task learning setups (Ha et al., 2016; Tay et al., 2021b) can push the boundary of the strong performance displayed by ExT5. Other solutions like gradient manipulation (Yu et al., 2020; Wang et al., 2021) might also further improve extreme multi-task scaling, albeit at the cost of more complex implementations.

Conclusion This paper explores how supervised multi-task learning at a massive scale can be used to improve existing self-supervised pre-training strategies for NLP models, and does so by introducing ExMix (§2) and ExT5 (§3). Our experiments showed that while negative transfer is common when fine-tuning on diverse tasks (§2.1), scaling up the number of tasks to include in a multi-task pre-training setup enables strong downstream performance (§3.2) with better sample-efficiency (§2.6). We hope that this paper motivates future research on how existing labeled datasets can be used to further improve NLP models within the pre-train/fine-tune paradigm.

ETHICS STATEMENT

Large language models have been shown to capture certain biases about the data they have been pre-trained on (Bender et al., 2020). While a comprehensive analysis of such biases is outside of the scope of this work, it is a compelling direction to investigate to what extent the inclusion of supervised data during pre-training can help mitigate such biases. An alternative consideration is the addition of diverse values-targeted data (Solaiman & Dennison, 2021) during pre-training in order to instill beneficial biases in a model.

Another challenge when training large models is their energy consumption and environmental impact (Strubell et al., 2019). To ablate different task combinations, we performed experiments using the more computationally efficient fine-tuning setup. We have shown that EXMIX leads to more sample-efficient pre-training compared to standard self-supervision, which we hope will save compute in future experiments.

REPRODUCABILITY STATEMENT

All of the code used to train EXT5 and its variants is already open-sourced as a part of the Mesh Tensorflow[1] (Shazeer et al., 2018) and T5[2] (Raffel et al., 2020) Libraries.

REFERENCES

Amazon reviews dataset. https://s3.amazonaws.com/amazon-reviews-pds/readme.html.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. Muppet: Massive multi-task representations with pre-finetuning, 2021.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2018.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 24–33, 2018.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://aclanthology.org/2020.findings-emnlp.148.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 54–63, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL https://www.aclweb.org/anthology/S19-2007.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.

Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: can language models be too big? In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, volume 1, pp. 271–278. Association for Computing Machinery, 2020. ISBN 9781450375856. doi: 10.1145/3442188.3445922.

[1] https://github.com/tensorflow/mesh

[2] https://github.com/google-research/text-to-text-transfer-transformer

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1160.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Byg1v1HKDB.

Joachim Bingel and Anders Søgaard. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 164–169, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2026.

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

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3302. URL https://aclanthology.org/W14-3302.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 1–46, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3001. URL https://aclanthology.org/W15-3001.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2301. URL https://aclanthology.org/W16-2301.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pp. 491–500, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3317593. URL https://doi.org/10.1145/3308560.3317593.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

R. Caruana, D. L. Silver, J. Baxter, T. M. Mitchell, L. Y. Pratt, and S. Thrun. Learning to learn: knowledge consolidation and transfer in inductive systems, 1995.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders, 2020.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. The 2020 bilingual, bi-directional webnlg+ shared task overview and evaluation results (webnlg+ 2020). In *Proceedings of the 3rd WebNLG Workshop on Natural Language Generation from the Semantic Web (WebNLG+ 2020)*, pp. 55–76, Dublin, Ireland (Virtual), 2020. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001.

Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. Pre-training tasks for embedding-based large-scale retrieval. *arXiv preprint arXiv:2002.03932*, 2020.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2174–2184, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1241. URL https://aclanthology.org/D18-1241.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300.

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

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc (eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pp. 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-33428-6.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. 2019. To appear in proceedings of Sinn und Bedeutung 23. Data can be found at https://github.com/mcdm/CommitmentBank/.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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. 2368–2378, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246.

Ondřej Dušek, David M Howcroft, and Verena Rieser. Semantic Noise Matters for Neural Natural Language Generation. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG 2019)*, pp. 421–426, Tokyo, Japan, 2019. URL https://www.aclweb.org/anthology/W19-8652/.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 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, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1102. URL https://aclanthology.org/P19-1102.

Claire Cardie Faisal Ladhak, Esin Durmus and Kathleen McKeown. Wikilingua: A new benchmark dataset for multilingual abstractive summarization. In *Findings of EMNLP, 2020*, 2020.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering, 2019.

Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *arXiv preprint arXiv:2109.04617*, 2021.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for nlg micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 179–188. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1017. URL http://www.aclweb.org/anthology/P17-1017.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pp. 96–120, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.10. URL https://aclanthology.org/2021.gem-1.10.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 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, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision, 2009. URL http://help.sentiment140.com/home.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. AutoSeM: Automatic task selection and mixing in multi-task learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3520–3531, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1355. URL https://aclanthology.org/N19-1355.

Jiaqi Guo, Qian Liu, Jian-Guang Lou, Zhenwen Li, Xueqing Liu, Tao Xie, and Ting Liu. Benchmarking meaning representations in neural semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1520–1540, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.118. URL https://aclanthology.org/2020.emnlp-main.118.

Zhaochen Guo and Denilson Barbosa. Robust entity linking via random walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pp. 499–508, 2014. doi: 10.1145/2661829.2661887. URL http://doi.acm.org/10.1145/2661829.2661887.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. Semantic parsing for task oriented dialog using hierarchical representations, 2018.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740.

David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks, 2016.

Jonathan Herzig, Peter Shaw, Ming-Wei Chang, Kelvin Guu, Panupong Pasupat, and Yuan Zhang. Unlocking compositional generalization in pre-trained models using intermediate representations, 2021.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*, 2001. URL https://aclanthology.org/H01-1069.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. In *Proceedings of ICML 2020*, 2020.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 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.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pp. 10–19, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4902. URL https://aclanthology.org/W17-4902.

Robin Jia, Mike Lewis, and Luke Zettlemoyer. Question answering infused pre-training of general-purpose contextualized representations, 2021.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 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://www.aclweb.org/anthology/2020.acl-main.709.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 252–262, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1023. URL https://aclanthology.org/N18-1023.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1896–1907, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.171. URL https://aclanthology.org/2020.findings-emnlp.171.

Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.731. URL https://aclanthology.org/2020.emnlp-main.731.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. Investigating multilingual NMT representations at scale. 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. 1565–1575, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1167. URL https://aclanthology.org/D19-1167.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 08 2019. ISSN 2307-387X. doi: 10.1162/tacl_a_00276. URL https://doi.org/10.1162/tacl_a_00276.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. 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. 1311–1316, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1131. URL https://aclanthology.org/D19-1131.

Adam D. Lelkes, Vinh Q. Tran, and Cong Yu. *Quiz-Style Question Generation for News Stories*, pp. 2501–2511. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383127. URL https://doi.org/10.1145/3442381.3449892.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, pp. 47, 2011.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Junlong Li, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. Task-specific objectives of pre-trained language models for dialogue adaptation, 2020.

Ya Li, Xinmei Tian, Tongliang Liu, and Dacheng Tao. On better exploring and exploiting task relationships in multitask learning: Joint model and feature learning. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1975–1985, May 2018. ISSN 2162-2388. doi: 10.1109/tnnls.2017.2690683. URL http://dx.doi.org/10.1109/TNNLS.2017.2690683.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 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. URL https://www.aclweb.org/anthology/2020.findings-emnlp.165.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of ACL 2019*, 2019. URL http://arxiv.org/abs/1905.12688.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496, Florence, Italy, 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1441. URL https://aclanthology.org/P19-1441.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents, 2019b.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *AAAI*, 2021.

Yiwei Lyu, Paul Pu Liang, Hai Pham, Eduard Hovy, Barnabás Póczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. StylePTB: A compositional benchmark for fine-grained controllable text style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2116–2138, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.171. URL https://aclanthology.org/2021.naacl-main.171.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 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, 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training, 2018.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. Dialoglue: A natural language understanding benchmark for task-oriented dialogue, 2020.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 31–41, 2016.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pp. 1–17, 2018.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 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, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 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, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 708–718, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.63. URL https://aclanthology.org/2020.findings-emnlp.63.

Richard Yuanzhe Pang, Adam Lelkes, Vinh Tran, and Cong Yu. AgreeSum: Agreement-oriented multi-document summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3377–3391, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.299. URL https://aclanthology.org/2021.findings-acl.299.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*, 2020.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 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, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200.

Jason Phang, Thibault Févry, and Samuel R. Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks, 2019.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 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, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1128. URL https://aclanthology.org/N19-1128.

Lorien Y Pratt, Jack Mostow, Candace A Kamm, Ace A Kamm, et al. Direct transfer of learned information among neural networks. In *Aaai*, volume 91, pp. 584–589, 1991.

Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Nazneen Fatema Rajani, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Murori Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, and Richard Socher. Dart: Open-domain structured data record to text generation. *arXiv preprint arXiv:2007.02871*, 2020.

Alec Radford, Jong Wook Kim, and Jeff Wu. Gpt-2 output dataset. https://github.com/openai/gpt-2-output-dataset, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. Few-shot question answering by pretraining span selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3066–3079, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.239. URL https://aclanthology.org/2021.acl-long.239.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset, 2020.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5418–5426, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.437. URL https://aclanthology.org/2020.emnlp-main.437.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.

Michael T. Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G. Dietterich. To transfer or not to transfer. In *In NIPS'05 Workshop, Inductive Transfer: 10 Years Later*, 2005.

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pp. 502–518, 2017.

Sebastian Ruder and Barbara Plank. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 372–382, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1038. URL https://aclanthology.org/D17-1038.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 15–18, 2019.

Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1044. URL https://aclanthology.org/D15-1044.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *ArXiv*, abs/1907.10641, 2020.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL https://doi.org/10.1145/3474381.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. A Hierarchical Multi-task Approach for Learning Embeddings from Semantic Tasks. In *Proceedings of AAAI 2019*, 2019.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 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.

J. H. Schmidhuber. Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook. 1987.

Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.52. URL https://aclanthology.org/2021.naacl-main.52.

Abigail See, Peter J. Liu, and Christopher D. Manning. 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, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://aclanthology.org/P17-1099.

Noam Shazeer. Glu variants improve transformer, 2020.

Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee, Mingsheng Hong, Cliff Young, Ryan Sepassi, and Blake A. Hechtman. Mesh-tensorflow: Deep learning for supercomputers. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 10435–10444, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3a37abdeefe1dab1b30f7c5c7e581b93-Abstract.html.

Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. Don't decay the learning rate, increase the batch size. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=B1Yy1BxCZ.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

Irene Solaiman and Christy Dennison. Process for adapting language models to society (palms) with values-targeted datasets. *arXiv preprint arXiv:2106.10328*, 2021.

Christian Stab, Tristan Miller, Pranav Rai, Benjamin Schiller, and Iryna Gurevych. Ukp sentential argument mining corpus, 2018. URL https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2345.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. Ranking Transfer Languages with Pragmatically-Motivated Features for Multilingual Sentiment Analysis. In *Proceedings of EACL 2021*, 2021.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*, 2021a.

Yi Tay, Zhe Zhao, Dara Bahri, Donald Metzler, and Da-Cheng Juan. Hypergrid transformers: Towards a single model for multiple tasks. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=hiq1rHO8pNT.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, 2018.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pp. 142–147, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119176.1119195. URL https://doi.org/10.3115/1119176.1119195.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 39–50, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7882–7926, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.635. URL https://aclanthology.org/2020.emnlp-main.635.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3261–3275, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019b. URL https://openreview.net/forum?id=rJ4km2R5t7.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=F1vEjWK-lH_.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl_a_00290. URL https://aclanthology.org/Q19-1040.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (eds.), *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pp. 1391–1399. ACM, 2017. doi: 10.1145/3038912.3052591. URL https://doi.org/10.1145/3038912.3052591.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. TWEETQA: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5020–5031, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1496. URL https://aclanthology.org/P19-1496.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016. URL https://cocoxu.github.io/publications/tacl2016-smt-simplification.pdf.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4913–4922, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.435. URL https://aclanthology.org/2021.findings-acl.435.

Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proc. VLDB Endow.*, 4(12):1450–1453, August 2011. ISSN 2150-8097. doi: 10.14778/3402755.3402793. URL https://doi.org/10.14778/3402755.3402793.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning, 2020.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 75–86, 2019.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 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.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

Rui Zhang and Joel Tetreault. This email could save your life: Introducing the task of email subject line generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 446–456, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1043. URL https://aclanthology.org/P19-1043.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint 1810.12885*, 2018.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (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, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1112–1125, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.90. URL https://aclanthology.org/2021.acl-long.90.

Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pp. 733–742, Arlington, Virginia, USA, 2010. AUAI Press. ISBN 9780974903965.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019.

A DATASETS

Dataset(s)	Description	No. Train Datasets	$\lvert\mathcal{D}\rvert$	Citation
GLUE	General Language Understanding	7	949,101	Wang et al. (2019b)
SuperGLUE	General Language Understanding	8	185,673	Wang et al. (2019a)
KILT	Knowledge-Intensive Language Tasks	9	3,129,859	Petroni et al. (2021)
Rainbow	Commonsense Reasoning	6	324,742	Lourie et al. (2021)
GEM (en)	Natural Language Generation	8	1,067,955	Gehrmann et al. (2021)
DialoGLUE	Dialogue Understanding	6	76,122	Mehri et al. (2020)
TweetEval	Twitter Classification Benchmark	8	120,104	Barbieri et al. (2020)
CNN/Dailymail	News Summarization	1	287,113	See et al. (2017)
XSum	News Summarization	1	203,577	Narayan et al. (2018)
Multi-News	News Summarization	1	44,972	Fabbri et al. (2019)
AESLC	Email Summarization	1	14,436	Zhang & Tetreault (2019)
Gigaword	Summarization	1	3,803,957	Rush et al. (2015)
SamSum	Dialogue Summarization	1	14,372	Gliwa et al. (2019)
ANLI	Adverserial NLI	1	162,865	Nie et al. (2020)
ESNLI	Explainable NLI	1	549,367	DeYoung et al. (2020)
AgreeSum Entailment	Article-Summary NLI	1	7,750	Pang et al. (2021)
DocNLI	Document NLI	1	942,314	Yin et al. (2021)
Vitamin C	Fact-checking NLI	1	370,653	Schuster et al. (2021)
Web Questions	QA (open)	1	3778	Berant et al. (2013)
SQuAD	QA (context)	1	87,599	Rajpurkar et al. (2016)
QuAC	QA (context)	1	83,568	Choi et al. (2018)
DROP	QA (Discrete Reasoning)	1	77,409	Dua et al. (2019)
RACE	School QA (MCQ)	4	113,013	Lai et al. (2017)
Eraser MultiRC	Explainable QA (MCQ)	1	24,029	DeYoung et al. (2020)
TweetQA	QA (context)	1	10692	Xiong et al. (2019)
NewsQuizQA	Question-Answer Generation	1	16,160	Lelkes et al. (2021)
Amazon Reviews	Review Classification	1	100,000	ama
GoEmotions	Emotion Classification	1	43.410	Demszky et al. (2020)
IMDb Reviews	Sentiment Classification	1	25,000	Maas et al. (2011)
Sentiment140	Sentiment Classification	1	1,600,000	Go et al. (2009)
Yelp Reviews	Sentiment Classification	1	560,000	Zhang et al. (2015)
AGNews	News Classification	1	120,000	Zhang et al. (2015)
TreqQC	Question Classification	1	5000	Hovy et al. (2001)
Civil Comments	Toxicity Classification	1	1,804,874	Borkan et al. (2019)
Wiki Toxicity	Toxicity Classification	1	159,571	Wulczyn et al. (2017)
Yahoo! Answers	Topic Classification	1	140,000	Zhang et al. (2015)
UKP Arg. Mining	Argument Classification	1	18,341	Stab et al. (2018)
Parsing to FunQL	Semantic Parsing	3	5,565	Guo et al. (2020)
Parsing to interm. repr.	Semantic Parsing	4	117,318	Herzig et al. (2021)
COGS	Semantic Parsing (Comp. Gen.)	1	24,155	Kim & Linzen (2020)
GPT Deepfake detection	Generated-text classification	8	500,000	Radford et al. (2019)
StylePTB	Style Transfer	4	53,546	Lyu et al. (2021)
Shakespearizing English	Style Transfer	2	36,790	Jhamtani et al. (2017)
MS-MARCO	Pointwise Ranking	1	100,000	Bajaj et al. (2018)
Total	EXMIX	107	18,085,040	-

Table 11: All of the training datasets used to construct ExMix.

Table 11 summarizes the 107 datasets included in EXMIX. Some of the lines in the table represent existing benchmarks that group several tasks together. From each collection, we use the datasets that include English training data:

- **GLUE**: CoLA (Warstadt et al., 2019), SST-2 (Socher et al., 2013), MRPC (Dolan & Brockett, 2005), QQP, STS-B (Cer et al., 2017), MNLI (Williams et al., 2018), QNLI (Converted from Rajpurkar et al. (2016), RTE (Dagan et al., 2006), WNLI (Sakaguchi et al., 2020).
- **SuperGLUE**: BoolQ (Clark et al., 2019), CB (De Marneffe et al., 2019), COPA (Roemmele et al., 2011), MultiRC (Khashabi et al., 2018), ReCoRD (Zhang et al., 2018), RTE (Dagan et al., 2006), WiC (Pilehvar & Camacho-Collados, 2019), WSC (Levesque et al., 2011).
- **KILT**: FEVER (Thorne et al., 2018), AIDA (Yosef et al., 2011), WNED (Guo & Barbosa, 2014), T-REx (Guo & Barbosa, 2014), NQ (Kwiatkowski et al., 2019), HoPo (Yang et al., 2018), TQA (Joshi et al., 2017), ELI5 (Fan et al., 2019), WoW (Dinan et al., 2019).
- **Rainbow**: αNLI (Bhagavatula et al., 2020), CosmosQA (Huang et al., 2019), HellaSWAG (Zellers et al., 2019), PIQA (Bisk et al., 2020), SocialIQA (Sap et al., 2019), WinoGrande (Sakaguchi et al., 2021).
- **GEM (en)**: Wiki-Lingua (Faisal Ladhak & McKeown, 2020), WenNLG (Gardent et al., 2017; Castro Ferreira et al., 2020), CommonGEN (Lin et al., 2020), E2E (Dušek et al., 2019), DART

(Radev et al., 2020), ToTTo (Parikh et al., 2020), Wiki-Auto (Jiang et al., 2020), TurkCorpus(Xu et al., 2016)

- **DialoGLUE**: Banking77 (Casanueva et al., 2020), HWU64 (Liu et al., 2019b), CLINC150 (Larson et al., 2019), SGD (Rastogi et al., 2020), TOP (Gupta et al., 2018).
- **TweetEval**: Emotion Recognition (Mohammad et al., 2018), Emoji Prediction (Barbieri et al., 2018), Irony Detection (Van Hee et al., 2018), Hate Speech Detection (Basile et al., 2019), Offensive Language Identification (Zampieri et al., 2019), Sentiment Analysis (Rosenthal et al., 2017), Stance Detection (Mohammad et al., 2016).

B EXPERIMENTAL DETAILS

This section describes the experimental details

B.1 IMPLEMENTATION DETAILS

Our models were trained using Mesh Tensorflow (Shazeer et al., 2018) using the T5 library (Raffel et al., 2020).

B.2 DATASET EXPERIMENTAL SETUP

This section reports the dataset and experimental setup on each individual target task/dataset.

SuperGLUE We finetune on the entire SuperGLUE as a mixture with proportionate sampling in similar fashion to (Raffel et al., 2020). We finetune for a total of 200k steps with a batch size of 128. When selecting checkpoints on SuperGLUE, we follow the same convention as Raffel et al. (2020) in selecting the best checkpoint for each task for a fair comparison to models that are fine-tuned on the individual tasks instead of co-training on all of them.

GEM We report test set results on all datasets except CommonGen and ToTTo, on which we report validation scores. We sweep over learning rates of 10^{-3}, 5×10^{-4} and 10^{-4}. All results are computed using GEM-metrics[3]. For each dataset, we select the best model checkpoint using average of BLEU, ROUGE-1, ROUGE-2 and ROUGE-L scores on the validation set. We use the greedy decoding strategy to be consistent with the original GEM paper (Gehrmann et al., 2021).

CBQA We report validation set results, and sweep over learning rates of 10^{-3} and 10^{-4}.

Rainbow We multi-task co-train on all datasets, and sweep over learning rates of 10^{-3} and 10^{-4}.

WMT Machine Translation We finetune our models on three collections of WMT, namely EnDe, EnFr and EnRo. We use a constant learning rate of 10^{-3} and dropout of 0.1. We train with a batch size of 4096 for a maximum of $400k$ steps and report peak validation BLEU score. We use a beam size of 4 and a length penalty of 0.6.

ARC We report scores on the Challenge set, and train with a batch size of 32 and sweep over learning rates of 10^{-3} and 10^{-4}.

CoNLL-03 NER We convert NER to seq2seq by writing the target as the ordered sequence of tags and entities (for example *"When Alice visited New York"* $\rightarrow$ *"[PER] Alice [LOC] New York"*). Accuracy is measured on a sentence level, considering a prediction to be correct only if it exactly matches the reference sequence.

[3] https://github.com/GEM-benchmark/GEM-metrics