UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

English Intermediate-Task Training Improves Zero-Shot Cross-Lingual Transfer Too

Phang, J.; Calixto, I.; Htut, P.M.; Pruksachatkun, Y.; Liu, H.; Vania, C.; Kann, K.; Bowman, S.R.

Publication date
2020

Document Version
Final published version

Published in
The 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing

License
CC BY

Link to publication

Citation for published version (APA):
Phang, J., Calixto, I., Htut, P. M., Pruksachatkun, Y., Liu, H., Vania, C., Kann, K., & Bowman, S. R. (2020). English Intermediate-Task Training Improves Zero-Shot Cross-Lingual Transfer Too. In K-F. Wong, K. Knight, & H. Wu (Eds.), *The 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: AACL-IJCNLP 2020 : proceedings of the conference : December 4-7, 2020* (pp. 557–575). The Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.aacl-main.56/

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:10 Mar 2023

English Intermediate-Task Training Improves Zero-Shot Cross-Lingual Transfer Too

Jason Phang[1,*] **Iacer Calixto**[1,2,*] **Phu Mon Htut**[1] **Yada Pruksachatkun**[1]
Haokun Liu[1] **Clara Vania**[1] **Katharina Kann**[3] **Samuel R. Bowman**[1]

[1]New York University [2]ILLC, University of Amsterdam [3]University of Colorado Boulder
{jasonphang,iacer.calixto,bowman}@nyu.edu

Abstract

Intermediate-task training—fine-tuning a pretrained model on an *intermediate* task before fine-tuning again on the target task—often improves model performance substantially on language understanding tasks in monolingual English settings. We investigate whether English intermediate-task training is still helpful on *non*-English target tasks. Using nine intermediate language-understanding tasks, we evaluate intermediate-task transfer in a zero-shot cross-lingual setting on the XTREME benchmark. We see large improvements from intermediate training on the BUCC and Tatoeba sentence retrieval tasks and moderate improvements on question-answering target tasks. MNLI, SQuAD and HellaSwag achieve the best overall results as intermediate tasks, while multi-task intermediate offers small additional improvements. Using our best intermediate-task models for each target task, we obtain a 5.4 point improvement over XLM-R Large on the XTREME benchmark, setting the state of the art[1] as of June 2020. We also investigate continuing multilingual MLM during intermediate-task training and using machine-translated intermediate-task data, but neither consistently outperforms simply performing English intermediate-task training.

1 Introduction

Zero-shot cross-lingual transfer involves training a model on task data in one set of languages (or language pairs, in the case of translation) and evaluating the model on the same task in unseen languages (or pairs). In the context of natural language understanding tasks, this is generally done using a pretrained multilingual language-encoding model such as mBERT (Devlin et al., 2019a), XLM (Conneau and Lample, 2019) or XLM-R (Conneau et al., 2020) that has been pretrained with a masked language modeling (MLM) objective on large corpora of multilingual data, fine-tune it on task data in one language, and evaluate the tuned model on the same task in other languages.

Intermediate-task training (STILTs; Phang et al., 2018) consists of fine-tuning a pretrained model on a data-rich *intermediate* task, before fine-tuning a second time on the target task. Despite its simplicity, this two-phase training setup has been shown to be helpful across a range of Transformer models and target tasks (Wang et al., 2019a; Pruksachatkun et al., 2020), at least within English settings.

In this work, we propose to use intermediate training on English tasks to improve zero-shot cross-lingual transfer performance. Starting with a pretrained multilingual language encoder, we perform intermediate-task training on one or more English tasks, then fine-tune on the target task in English, and finally evaluate zero-shot on the same task in other languages.

Intermediate-task training on English data introduces a potential issue: We train the pretrained multilingual model extensively on only English data before evaluating it on non-English target task data, potentially causing the model to lose the knowledge of the other languages that was acquired during pretraining (Kirkpatrick et al., 2017; Yogatama et al., 2019). To mitigate this issue, we experiment with mixing in multilingual MLM training updates during the intermediate-task training. In the same vein, we also conduct a case study where we machine-translate intermediate task data from English into three other languages (German, Russian and Swahili) to investigate whether intermediate training on these languages improves target task performance in the same languages.

Concretely, we use the pretrained XLM-R (Con-

*Equal contribution.

[1]The state of art on XTREME at the time of final publication in September 2020 is held by Fang et al. (2020), who introduce an orthogonal method.

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 557–575
December 4 - 7, 2020. ©2020 Association for Computational Linguistics

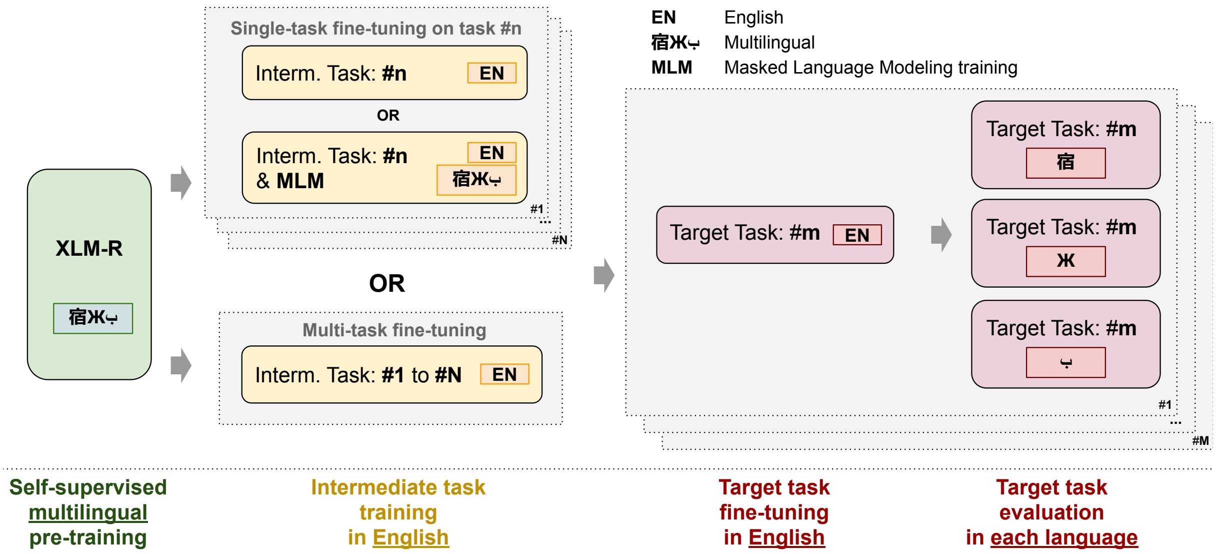

Figure 1: We investigate the benefit of injecting an additional phase of intermediate-task training on English language task data. We also consider variants using multi-task intermediate-task training, as well as continuing multilingual MLM during intermediate-task training. Best viewed in color.

neau et al., 2020) encoder and perform experiments on 9 target tasks from the recently introduced XTREME benchmark (Hu et al., 2020), which aims to evaluate zero-shot cross-lingual transfer performance across diverse target tasks across up to 40 languages each. We investigate how training on 9 different intermediate tasks, including question answering, sentence tagging, sentence completion, paraphrase detection, and natural language inference impacts zero-shot cross-lingual transfer performance. We find the following:

- Intermediate-task training on SQuAD, MNLI, and HellaSwag yields large target-task improvements of 8.2, 7.5, and 7.0 points on the development set, respectively. Multi-task intermediate-task training on all 9 tasks performs best, improving by 8.7 points.

- Applying intermediate-task training to BUCC and Tatoeba, the two sentence retrieval target tasks that have no training data of their own, yields dramatic improvements with almost every intermediate training configuration. TyDiQA shows consistent improvements with many intermediate tasks, whereas XNLI does not see benefits from intermediate training.

- Evaluating our best performing models for each target task on the XTREME benchmark yields an average improvement of **5.4 points**, setting the state of the art as of writing.

- Training on English intermediate tasks outperforms the more complex alternatives of (i) continuing multilingual MLM during intermediate-task training, and (ii) using machine-translated intermediate-task data.

2 Approach

We follow a three-phase approach to training, illustrated in Figure 1: (i) we use a publicly available model pretrained on raw multilingual text using MLM; (ii) we perform intermediate-task training on one or more English intermediate tasks; and (iii) we fine-tune the model on English target-task training data, before evaluating it on target-task test data in each target language.

In phase (ii), our intermediate tasks have English input data. In Section 2.4, we investigate an alternative where we machine-translate intermediate-task data to other languages, which we use for training. We experiment with both single- and multi-task training for intermediate-task training. We use target tasks from the recent XTREME benchmark for zero-shot cross-lingual transfer.

2.1 Intermediate Tasks

We study the effect of intermediate-task training (STILTs; Phang et al., 2018) with nine different English intermediate tasks, described in Table 1.

We choose the tasks below based to cover a variety of task formats (classification, question answering, and multiple choice) and based on evidence

Name	|Train|	|Dev|	|Test|	Task	Genre/Source
Intermediate tasks					
ANLI$^+$	1,104,934	22,857	–	natural language inference	Misc.
MNLI	392,702	20,000	–	natural language inference	Misc.
QQP	363,846	40,430	–	paraphrase detection	Quora questions
SQuAD v2.0	130,319	11,873	–	span extraction	Wikipedia
SQuAD v1.1	87,599	10,570	–	span extraction	Wikipedia
HellaSwag	39,905	10,042	–	sentence completion	Video captions & Wikihow
CCG	38,015	5,484	–	tagging	Wall Street Journal
Cosmos QA	25,588	3,000	–	question answering	Blogs
CommonsenseQA	9,741	1,221	–	question answering	Crowdsourced responses
Target tasks (XTREME Benchmark)					
XNLI	392,702	2,490	5,010	natural language inference	Misc.
PAWS-X	49,401	2,000	2,000	paraphrase detection	Wiki/Quora
POS	21,253	3,974	47–20,436	tagging	Misc.
NER	20,000	10,000	1,000–10,000	named entity recognition	Wikipedia
XQuAD	87,599	34,726	1,190	question answering	Wikipedia
MLQA	87,599	34,726	4,517–11,590	question answering	Wikipedia
TyDiQA-GoldP	3,696	634	323–2,719	question answering	Wikipedia
BUCC	–	–	1,896–14,330	sentence retrieval	Wiki / news
Tatoeba	–	–	1,000	sentence retrieval	Misc.

Table 1: Overview of the intermediate tasks (top) and target tasks (bottom) in our experiments. For target tasks, *Train* and *Dev* correspond to the English training and development sets, while *Test* shows the range of sizes for the target-language test sets for each task. XQuAD, TyDiQA and Tateoba do not have separate held-out development sets.

of positive transfer from literature. Pruksachatkun et al. (2020) shows that MNLI (of which ANLI$^+$is a superset), CommonsenseQA, Cosmos QA and HellaSwag yield positive transfer to a range of downstream English-language tasks in intermediate training. CCG involves token-wise prediction and is similar to the POS and NER target tasks. Both versions of SQuAD are widely-used question-answering tasks, while QQP is semantically similar to sentence retrieval target tasks (BUCC and Tatoeba) as well as PAWS-X, another paraphrase-detection task.

ANLI + MNLI + SNLI (ANLI$^+$) The Adversarial Natural Language Inference dataset (Nie et al., 2020) is collected using model-in-the-loop crowdsourcing as an extension of the Stanford Natural Language Inference (SNLI; Bowman et al., 2015) and Multi-Genre Natural Language Inference (MNLI; Williams et al., 2018) corpora. We follow Nie et al. (2020) and use the concatenated ANLI, MNLI and SNLI training sets, which we refer to as ANLI$^+$. For all three natural language inference tasks, examples consist of premise and hypothesis sentence pairs, and the task is to classify the relationship between the premise and hypothesis as entailment, contradiction, or neutral.

CCG CCGbank (Hockenmaier and Steedman, 2007) is a conversion of the Penn Treebank into Combinatory Categorial Grammar (CCG) derivations. The CCG supertagging task that we use consists of assigning lexical categories to individual word tokens, which together roughly determine a full parse.[2]

CommonsenseQA CommonsenseQA (Talmor et al., 2019) is a multiple-choice QA dataset generated by crowdworkers based on clusters of concepts from ConceptNet (Speer et al., 2017).

Cosmos QA Cosmos QA is multiple-choice commonsense-based *reading comprehension* dataset (Huang et al., 2019b) generated by crowdworkers, with a focus on the causes and effects of events.

HellaSwag HellaSwag (Zellers et al., 2019) is a commonsense reasoning dataset framed as a four-way multiple choice task, where examples consist of an incomplete paragraph and four choices of spans, only one of which is a plausible continuation of the scenario. It is built using adversarial filtering (Zellers et al., 2018; Le Bras et al., 2020) with BERT.

[2]If a word is tokenized into sub-word tokens, we use the representation of the first token for the tag prediction for that word as in Devlin et al. (2019a).

MNLI In additional to the full ANLI$^+$, we also consider the MNLI task as a standalone intermediate task because of its already large and diverse training set.

QQP Quora Question Pairs[3] is a paraphrase detection dataset. Examples in the dataset consist of two questions, labeled for whether they are semantically equivalent.

SQuAD Stanford Question Answering Dataset (Rajpurkar et al., 2016, 2018) is a question-answering dataset consisting of passages extracted from Wikipedia articles and crowd-sourced questions and answers. In SQuAD version 1.1, each example consists of a context passage and a question, and the answer is a text span from the context. SQuAD version 2.0 includes additional questions with no answers, written adversarially by crowd-workers. We use both versions in our experiments.

2.2 Target Tasks

We use the 9 target tasks from the XTREME benchmark, which span 40 different languages (hereafter referred to as the *target languages*): Cross-lingual Question Answering (**XQuAD**; Artetxe et al., 2020b); Multilingual Question Answering (**MLQA**; Lewis et al., 2020); Typologically Diverse Question Answering (**TyDiQA-GoldP**; Clark et al., 2020); Cross-lingual Natural Language Inference (**XNLI**; Conneau et al., 2018); Cross-lingual Paraphrase Adversaries from Word Scrambling (**PAWS-X**; Yang et al., 2019); Universal Dependencies v2.5 (Nivre et al., 2018) **POS** tagging; Wikiann **NER** (Pan et al., 2017); **BUCC** (Zweigenbaum et al., 2017, 2018), which requires identifying parallel sentences from corpora of different languages; and **Tatoeba** (Artetxe and Schwenk, 2019), which involves aligning pairs of sentences with the same meaning.

Among the 9 tasks, BUCC and Tatoeba are sentence retrieval tasks that do not include training sets, and are scored based on the similarity of learned representations (see Appendix A). XQuAD, TyDiQA and Tatoeba do not include development sets separate from the test sets.[4] For all XTREME tasks, we follow the training and evaluation protocol described in the benchmark paper (Hu et al., 2020) and their sample implementation.[5] Intermediate- and target-task statistics are shown in Table 1.

2.3 Multilingual Masked Language Modeling

Our setup requires that we train the pretrained multilingual model extensively on English data before using it on a non-English target task, which can lead to the catastrophic forgetting of other languages acquired during pretraining. We investigate whether continuing to train on the multilingual MLM pretraining objective while fine-tuning on an English intermediate task can prevent catastrophic forgetting of the target languages and improve downstream transfer performance.

We construct a multilingual corpus across the 40 languages covered by the XTREME benchmark using Wikipedia dumps from April 14, 2020 for each language and the MLM data creation scripts from the `jiant` 1.3 library (Phang et al., 2020). In total, we use 2 million sentences sampled across all 40 languages using the sampling ratio from Conneau and Lample (2019) with $\alpha = 0.3$.

2.4 Translated Intermediate-Task Training

Large-scale labeled datasets are rarely available in languages other than English for most language-understanding benchmark tasks. Given the availability of increasingly performant machine translation models, we investigate if using machine-translated intermediate-task data can improve same-language transfer performance, compared to using English intermediate task data.

We translate training and validation data of three intermediate tasks: QQP, HellaSwag, and MNLI. We choose these tasks based on the size of the training sets and because their example-level (rather than word-level) labels can be easily mapped onto translated data. To translate QQP and HellaSwag, we use pretrained machine translation models from OPUS-MT (Tiedemann and Thottingal, 2020). These models are trained with Marian-NMT (Junczys-Dowmunt et al., 2018) on OPUS data (Tiedemann, 2012), which integrates several resources depending on the available corpora for the language pair. For MNLI, we use the publicly available machine-translated training data of XNLI provided by the XNLI authors.[6] We use German, Russian, and Swahili translations of

[3] `http://data.quora.com/First-Quora-DatasetRelease-Question-Pairs`

[4] UDPOS also does not include development sets for Kazakh, Thai, Tagalog or Yoruba.

[5] `https://github.com/google-research/xtreme`

[6] According to Conneau et al. (2018), these data are translated using a Facebook internal machine translation system.

all three datasets instead of English data for the intermediate-task training.

3 Experiments and Results

3.1 Models

We use the pretrained XLM-R Large model (Conneau et al., 2020) as a starting point for all our experiments, as it currently achieves state-of-the-art performance on many zero-shot cross-lingual transfer tasks.[7] Details on intermediate- and target-task training can be found in Appendix A.

XLM-R For our baseline, we directly fine-tune the pretrained XLM-R model on each target task's English training data (if available) and evaluate zero-shot on non-English data, closely following the sample implementation for the XTREME benchmark.

XLM-R + Intermediate Task In our main approach, as described in Figure 1, we include an additional intermediate-task training phase before training and evaluating on the target tasks as described above.

We also experiment with multi-task training on all available intermediate tasks. We follow Raffel et al. (2020) and sample batches of examples for each task with probability $r_m = \frac{min(e_m,K)}{\sum(min(e_m,K))}$, where e_m is the number of examples in task m and the constant $K = 2^{17}$ limits the oversampling of data-rich tasks.

XLM-R + Intermediate Task + MLM To incorporate multilingual MLM into the intermediate-task training, we treat multilingual MLM as an additional task for intermediate training, using the same multi-task sampling strategy as above.

XLM-R + Translated Intermediate Task We translate intermediate-task training and validation data for three tasks and fine-tune XLM-R on translated intermediate-task data before we train and evaluate on the target tasks.

3.2 Software

Experiments were carried out using the jiant (Phang et al., 2020) library (2.0 alpha), based on PyTorch (Paszke et al., 2019) and Transformers (Wolf et al., 2019).

[7]XLM-R Large (Conneau et al., 2020) is a 550m-parameter variant of the RoBERTa masked language model (Liu et al., 2019b) trained on a cleaned version of CommonCrawl on 100 languages. Notably, Yoruba is used in the POS and NER XTREME tasks but not is not in the set of 100 languages.

3.3 Results

We train three versions of each intermediate-task model with different random seeds. For each run, we compute the average target-task performance across languages, and report the median performance across the three random seeds.

Intermediate-Task Training As shown in Table 2, no single intermediate task yields positive transfer across all target tasks. The target tasks TyDiQA, BUCC and Tatoeba see consistent gains from most or all intermediate tasks. In particular, BUCC and Tatoeba, the two sentence retrieval tasks with no training data, benefit universally from intermediate-task training. PAWS-X, NER, XQuAD and MLQA also exhibit gains with the additional intermediate-task training on some intermediate tasks. On the other hand, we find generally no or negative transfer to XNLI and POS.

Among the intermediate tasks, we find that MNLI performs best; with meaningful improvements across the PAWS-X, TyDiQA, BUCC and Tatoeba tasks. ANLI$^+$, SQuAD v1.1, SQuAD v2.0 and HellaSwag also show strong positive transfer performance: SQuAD v1.1 shows strong positive transfer across all three QA tasks, SQuAD v2.0 shows the most positive transfer to TyDiQA, while HellaSwag shows the most positive transfer to NER and BUCC tasks. ANLI$^+$does not show any improvement over MNLI (of which it is a superset), even on XNLI for which it offers additional directly relevant training data. This mirrors negative findings from Nie et al. (2020) on NLI evaluations and Bowman et al. (2020) on transfer within English. QQP significantly improves sentence retrieval-task performance, but has broadly negative transfer to the other target tasks.[8] CCG also has relatively poor transfer performance, consistent with Pruksachatkun et al. (2020).

Among our intermediate tasks, both SQuAD v1.1 and MNLI also serve as training sets for target tasks (for XNLI and XQuAD/MLQA respectively). While both tasks show overall positive transfer, SQuAD v1.1 actually markedly improves the performance in XQuAD and MLQA, while MNLI slightly hurts XNLI performance. We hypothesize that the somewhat surprising improvements to XQuAD and MLQA performance from SQuAD v1.1 arise due to the baseline XQuAD and MLQA

[8]For QQP, on 2 of the 3 random seeds the NER model performed extremely poorly, leading to the large negative transfer of -45.4.

		Target tasks									
		XNLI	**PAWS-X**	**POS**	**NER**	**XQuAD**	**MLQA**	**TyDiQA**	**BUCC**	**Tatoeba**	**Avg.**
	Metric	*acc.*	*acc.*	*F1*	*F1*	*F1 / EM*	*F1 / EM*	*F1 / EM*	*F1*	*acc.*	–
	# langs.	15	7	33	40	11	7	9	5	37	–
	XLM-R	**80.1**	86.5	75.7	62.8	76.1 / 60.0	70.1 / 51.5	65.6 / 48.2	71.5	31.0	67.2
Without MLM	**ANLI+**	- 0.8	- 0.0	- 1.4	- 3.5	- 1.1 / - 0.5	- 0.6 / - 0.8	- 0.6 / - 3.0	+19.9	+48.2	+ 6.6
	MNLI	- 1.2	+ 1.4	- 0.7	+ 0.5	- 0.3 / - 0.1	+ 0.2 / + 0.2	- 1.0 / - 1.6	+20.0	+48.8	+ 7.5
	QQP	- 4.4	- 4.8	- 6.5	-45.4	- 3.8 / - 3.8	- 3.9 / - 4.4	-11.1 / -10.2	+17.1	+49.5	- 1.5
	SQuADv1.1	- 1.9	+ 1.2	- 0.8	- 0.4	**+ 1.8** / **+ 2.5**	**+ 2.2** / **+ 2.6**	+ 9.7 / +10.8	+18.9	+41.3	+ 8.1
	SQuADv2	- 1.6	+ 1.9	- 1.1	+ 0.8	- 0.5 / + 0.7	- 0.4 / + 0.1	**+10.4** / **+11.3**	+19.3	+43.4	+ 8.2
	HellaSwag	- 7.1	+ 1.8	- 0.7	+ 1.6	- 0.0 / + 0.5	- 0.1 / + 0.2	- 0.0 / - 1.0	**+20.3**	+47.6	+ 7.0
	CCG	- 2.6	- 3.4	- 2.0	- 1.5	- 1.5 / - 1.3	- 1.6 / - 1.5	- 2.8 / - 6.2	+11.7	+41.9	+ 4.1
	CosmosQA	- 2.1	- 0.3	- 1.4	- 1.5	- 0.9 / - 1.3	- 1.5 / - 2.0	+ 0.5 / - 0.6	+19.2	+43.9	+ 6.1
	CSQA	- 2.9	- 2.8	- 1.7	- 1.6	- 1.0 / - 1.8	- 1.0 / - 0.6	+ 3.5 / + 2.9	+18.1	+48.6	+ 6.5
	Multi-task	- 0.9	+ 1.7	- 1.0	**+ 1.8**	+ 0.3 / + 0.9	+ 0.2 / + 0.5	+ 5.8 / + 6.0	+19.6	**+49.9**	**+ 8.7**
With MLM	**ANLI+**	- 1.1	+ 1.4	**+ 0.0**	+ 0.4	- 1.9 / - 1.7	- 0.7 / - 0.6	+ 0.9 / + 0.5	+18.6	+46.2	+ 7.1
	MNLI	- 0.7	+ 1.6	- 1.6	+ 1.0	- 0.7 / + 0.1	+ 0.4 / + 0.8	- 1.8 / - 3.2	+17.1	+44.3	+ 6.6
	QQP	- 1.3	- 1.1	- 2.4	- 0.9	- 0.3 / - 0.2	+ 0.0 / + 0.2	- 1.6 / - 4.2	+14.4	+39.8	+ 5.0
	SQuADv1.1	- 2.6	+ 0.3	- 2.0	- 0.9	+ 0.2 / + 1.6	+ 0.1 / + 1.1	+ 8.5 / + 9.5	+16.0	+40.3	+ 6.8
	SQuADv2	- 1.7	**+ 2.1**	- 1.4	+ 1.0	- 0.8 / + 0.1	- 0.8 / - 0.5	+ 8.3 / + 8.9	+15.6	+31.3	+ 6.1
	HellaSwag	- 3.3	+ 2.0	- 0.7	+ 0.8	- 0.8 / - 0.0	+ 0.1 / + 0.6	+ 0.3 / + 1.0	+ 6.3	+22.3	+ 3.1
	CCG	- 1.0	- 1.3	- 1.2	- 1.9	- 1.9 / - 2.2	- 2.1 / - 2.6	- 5.5 / - 6.2	+ 8.8	+36.1	+ 3.3
	CosmosQA	- 1.0	- 1.0	- 1.6	- 3.8	- 3.1 / - 3.3	- 3.7 / - 4.2	- 0.6 / - 3.2	+15.5	+42.7	+ 4.7
	CSQA	- 0.5	+ 0.3	- 1.0	- 0.7	- 0.9 / - 1.0	- 0.7 / - 0.6	+ 2.1 / + 0.4	+11.6	+17.2	+ 2.9
		XTREME Benchmark Scores†									
	XLM-R (Hu et al., 2020)	79.2	86.4	72.6	**65.4**	76.6 / 60.8	71.6 / 53.2	65.1 / 45.0	66.0	57.3	68.1
	XLM-R (Ours)	79.5	86.2	74.0	62.6	76.1 / 60.0	70.2 / 51.2	65.6 / 48.2	64.5	31.0	64.8
	Our Best Models‡	**80.0**	**87.9**	**74.4**	64.0	**78.7 / 63.3**	**72.4 / 53.7**	**76.0 / 59.5**	**71.9**	**81.2**	**73.5**
	Human (Hu et al., 2020)	92.8	97.5	97.0	-	91.2 / 82.3	91.2 / 82.3	90.1 / -	-	-	-

Table 2: Intermediate-task training results. We compute the average target task performance across all languages, and report the median over 3 separate runs with different random seeds. Multi-task experiments use all intermediate tasks. We underline the best results per target task with and without intermediate MLM co-training, and bold-face the best overall scores for each target task. †: XQuAD, TyDiQA and Tatoeba do not have held-out test data and are scored using development sets in the benchmark. ‡: Results obtained with our best-performing intermediate task configuration for each target task, selected based on the development set. The results for individual languages can be found in Appendix B.

models being under-trained. For all target-task fine-tuning, we follow the sample implementation for target task training in the XTREME benchmark, which trains on SQuAD for only 2 epochs. This may explain why an additional phase of SQuAD training can improve performance. Conversely, the MNLI-to-XNLI model might be over-trained, given the MNLI training set is approximately 4 times as large as the SQuAD v1.1 training set.

Multi-Task Training Multi-task training on all intermediate tasks attains the best overall average performance on the XTREME tasks, and has the most positive transfer to NER and Tatoeba tasks. However, the overall margin of improvement over the best single intermediate-task model is relatively small (only 0.3, over MNLI), while requiring significantly more training resources. Many single intermediate-task models also outperform the multi-task model in individual target tasks. Wang et al. (2019b) also found more mixed results from a having an initial phase of multi-task training, albeit only among English language tasks across a different set of tasks. On the other hand, multi-task training precludes the need to do intermediate-task model selection, and is a useful method for incorporating multiple, diverse intermediate tasks.

MLM Incorporating MLM during intermediate-task training shows no clear trend. It reduces negative transfer, as seen in the cases of CommonsenseQA and QQP, but it also tends to somewhat reduce positive transfer. The reductions in positive transfer are particularly significant for the BUCC and Tatoeba tasks, although the impact on TyDiQA is more mixed. On balance, we do not see that incorporating MLM improves transfer performance.

XTREME Benchmark Results At the bottom of Table 2, we show results obtained by XLM-R on the XTREME benchmark as reported by Hu et al. (2020), results obtained with our re-implementation of XLM-R (i.e. our baseline), and results obtained with our best models, which use intermediate-task configuration selected according

TL	Model	XNLI	PAWS-X	POS	NER	XQuAD	MLQA	TyDiQA	BUCC	Tatoeba
English	XLM-R	**89.3**	93.4	95.9	81.6	**86.3** / 74.2	81.6 / 68.6	70.4 / 56.6	–	–
	MNLI$_{en}$	- 1.2	**+ 1.6**	+ 0.3	+ 2.6	- 2.1 / - 1.6	+ 1.1 / + 1.4	+ 1.1 / + 1.1	–	–
	QQP$_{en}$	- 3.2	- 0.4	- 2.2	- 5.8	- 4.0 / - 3.6	- 2.6 / - 2.6	- 6.2 / - 5.0	–	–
	HellaSwag$_{en}$	- 0.8	+ 1.5	**+ 0.6**	**+ 2.7**	- 0.2 / **+ 1.4**	**+ 1.8 / + 2.3**	**+ 1.7 / + 2.5**	–	–
German	XLM-R	**83.8**	88.1	88.6	78.6	77.7 / 61.2	**69.1** / 52.0	–	77.7	63.9
	MNLI$_{en}$	- 0.8	**+ 0.9**	- 0.1	- 0.8	- 0.3 / - 1.0	- 1.0 / - 0.2	–	+16.5	+32.7
	MNLI$_{de}$	- 0.4	+ 0.5	- 0.3	- 0.9	+ 0.2 / - 0.3	- 2.4 / - 2.0	–	**+17.0**	+33.7
	QQP$_{en}$	- 2.2	- 4.2	- 3.2	- 7.3	- 4.5 / - 4.7	- 6.7 / - 6.4	–	+16.5	+32.6
	QQP$_{de}$	- 2.6	- 9.1	- 3.2	-22.9	- 6.6 / - 5.9	- 7.7 / - 6.6	–	+16.0	+33.5
	HellaSwag$_{en}$	- 0.3	+ 0.3	**+ 0.1**	**+ 0.5**	**+ 1.0 / + 0.2**	- 0.3 / **+ 0.4**	–	+16.9	**+33.8**
	HellaSwag$_{de}$	- 0.2	+ 0.2	- 0.4	- 0.4	+ 0.2 / - 0.2	- 3.5 / - 2.5	–	+16.3	+33.5
Russian	XLM-R	79.2	–	**89.5**	69.3	77.7 / 59.8	–	65.4 / 43.6	79.2	42.1
	MNLI$_{en}$	**+ 0.3**	–	- 0.0	+ 0.8	+ 0.1 / + 1.5	–	- 1.5 / - 4.6	+14.3	+47.1
	MNLI$_{ru}$	- 0.6	–	- 0.3	+ 1.9	- 0.4 / + 1.3	–	**+11.2 / +16.1**	+13.1	+48.3
	QQP$_{en}$	- 0.7	–	- 2.9	-18.6	- 3.5 / - 2.4	–	- 8.1 / - 5.4	+14.1	+49.5
	QQP$_{ru}$	- 3.0	–	-10.6	-59.1	- 5.2 / - 3.9	–	-14.4 / -12.1	+13.3	+46.7
	HellaSwag$_{en}$	- 0.9	–	- 0.0	+ 1.4	**+ 0.8 / + 2.9**	–	- 4.0 / -10.6	**+14.7**	**+49.9**
	HellaSwag$_{ru}$	- 0.3	–	- 0.4	**+ 2.8**	+ 0.2 / + 0.2	–	+ 8.5 / +13.2	-71.6	-23.5
Swahili	XLM-R	**72.4**	–	–	69.8	–	–	67.2 / 48.7	–	7.9
	MNLI$_{en}$	- 3.0	–	–	**+ 0.6**	–	–	- 0.3 / - 0.2	–	+24.9
	MNLI$_{sw}$	- 1.1	–	–	- 2.4	–	–	+13.8 / +23.4	–	**+47.9**
	QQP$_{en}$	- 2.8	–	–	- 4.6	–	–	-12.7 / -12.2	–	+27.2
	QQP$_{sw}$	- 7.1	–	–	-32.1	–	–	- 7.0 / - 0.4	–	+41.8
	HellaSwag$_{en}$	- 0.4	–	–	+ 0.1	–	–	- 0.9 / - 0.4	–	+27.2
	HellaSwag$_{sw}$	- 9.8	–	–	+ 0.4	–	–	**+15.6 / +26.3**	–	- 0.5

Table 3: Experiments with translated intermediate-task training and validation data evaluated on all XTREME target tasks. In each target language (TL) block, models are evaluated on a single target language. We show results for models trained on original intermediate-task training data (`en`) and compare it to models trained on translated data {`de`, `ru`, `sw`}. '–' indicates that target task data is not available for that target language.

to development set performance on each target task. Based on the results in Table 2, which reflect the median over 3 runs, we pick the best intermediate-task configuration for each target task, and then choose the best model out of the 3 runs. Scores on the XTREME benchmark are computed based on the respective test sets where available, and based on development sets for target tasks without separate held-out test sets. We are generally able to replicate the best reported XLM-R baseline results, except for Tatoeba, where our implementation significantly underperforms the reported scores in Hu et al. (2020), and TyDiQA, where our implementation outperforms the reported scores. We also highlight that there is a large margin of difference between development and test set scores for BUCC–this is likely because BUCC is evaluated based on sentence retrieval over the given set of input sentences, and the test sets for BUCC are generally much larger than the development sets.

Our best models show gains in 8 out of the 9 XTREME tasks relative to both baseline implementations, attaining an average score of 73.5 across target tasks, a 5.4 point improvement over the previous best reported average score of 68.1. We set the state of the art on the XTREME benchmark as of June 2020, though Fang et al. (2020) achieve higher results and hold the state of the art using an orthogonal approach at the time of our final publication in September 2020.

Translated Intermediate-Task Training Data In Table 3, we show results for experiments using machine-translated intermediate-training data, and evaluated on the available target-task languages. Surprisingly, even when evaluating in-language, using target-language intermediate-task data does not consistently outperform using English intermediate-task data in any of the intermediate tasks on average.

In general, cross-lingual transfer to XNLI is negative regardless of the intermediate-task or the target language. In contrast, we observe mostly positive transfer on BUCC, and Tatoeba, with a few notable exceptions where models fail catastrophically. TyDiQA exhibits positive transfer where the intermediate- and target-task languages aligned: intermediate training on Russian or German helps TyDiQA performance in that respective language, whereas intermediate training on English hurts non-English performance somewhat. For the remaining tasks, there appears to be little correlation between performance and the alignment of intermediate- and target-task languages. English language QQP already has mostly negative transfer to all target tasks except for BUCC and Tatoeba (see Table 2), and also shows a similar trend when translated into any of the three target languages.

We note that the quality of translations may affect the transfer performance. While validation performance on the translated intermediate tasks (Table 15) for MNLI and QQP is only slightly worse than the original English versions, the performance for the Russian and Swahili HellaSwag is much worse and close to chance. Despite this, intermediate-task training on Russian and Swahili HellaSwag improve performance on PAN-X and TyDiQA, while we see generally poor transfer performance from QQP. The interaction between translated intermediate-task data and transfer performance continues to be a complex open question. Artetxe et al. (2020a) found that translating or back-translating training data for a task can improve zero-shot cross-lingual performance for tasks such as XNLI depending on how the multilingual datasets are created. In contrast, we train on translated intermediate-task data and then fine-tune on a target task with English training data (excluding BUCC2018 and Tatoeba). The authors of the XTREME benchmark have also recently released translated versions of all the XTREME task training data, which we hope will prompt further investigation into this matter.

4 Related work

Sequential transfer learning using pretrained Transformer-based encoders (Phang et al., 2018) has been shown to be effective for many text classification tasks. This setup generally involves fine-tuning on a single task (Pruksachatkun et al., 2020; Vu et al., 2020) or multiple tasks (Liu et al., 2019a; Wang et al., 2019b; Raffel et al., 2020), sometimes referred to as the intermediate task(s), before fine-tuning on the target task. We build upon this line of work, focusing on intermediate-task training for improving cross-lingual transfer.

Early work on cross-lingual transfer mostly relies on the availability of parallel data, where one can perform translation (Mayhew et al., 2017) or project annotations from one language into another (Hwa et al., 2005; Agić et al., 2016). For dependency parsing, McDonald et al. (2011) use delexicalized parsers trained on source languages and labeled training data for parsing target-language data. Agić (2017) proposes a parser selection method to select the single best parser for a target language.

For large-scale cross-lingual transfer outside NLU, Johnson et al. (2017) train a single multilingual neural machine translation system with up to 7 languages and perform zero-shot translation without explicit bridging between the source and target languages. Aharoni et al. (2019) expand this approach to cover over 100 languages in a single model. Recent works on extending pretrained Transformer-based encoders to multilingual settings show that these models are effective for cross-lingual tasks and competitive with strong monolingual models on the XNLI benchmark (Devlin et al., 2019b; Conneau and Lample, 2019; Conneau et al., 2020; Huang et al., 2019a). More recently, Artetxe et al. (2020a) showed that cross-lingual transfer performance can be sensitive to translation artifacts arising from a multilingual datasets' creation procedure.

Finally, Pfeiffer et al. (2020) propose adapter modules that learn language and task representations for cross-lingual transfer, which allow adaptation to languages not seen during pretraining.

5 Conclusion

We evaluate the impact of intermediate-task training on zero-shot cross-lingual transfer. We investigate 9 intermediate tasks and how intermediate-task training impacts the zero-shot cross-lingual transfer to the 9 target tasks in the XTREME benchmark.

Overall, intermediate-task training significantly improves the performance on BUCC and Tatoeba, the two sentence retrieval target tasks in the XTREME benchmark, across almost every intermediate-task configuration. Our best models obtain 5.9 and 23.9 point gains on BUCC and Tatoeba, respectively, compared to the best available XLM-R baseline scores (Hu et al., 2020). We also observed gains in question-answering tasks, particularly using SQuAD v1.1 and v2.0 as intermediate tasks, with absolute gains of 2.1 F1 for XQuAD, 0.8 F1 for MLQA, and 10.4 for F1 TyDiQA, again over the best available baseline scores. We improve over XLM-R by 5.4 points on average on the XTREME benchmark. Additionally, we found multi-task training on all 9 intermediate tasks to slightly outperform individual intermediate training. On the other hand, we found that neither incorporating multilingual MLM into the intermediate-task training phase nor translating intermediate-task data consistently led to improved transfer performance.

While we have explored the extent to which English intermediate-task training can improve cross-lingual transfer, a clear next avenue of investigation for future work is how the choice of intermediate- and target-task languages influences transfer across different tasks.

Acknowledgments

This project has benefited from support to SB by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program), by Samsung Research (under the project *Improving Deep Learning using Latent Structure*), by Intuit, Inc., by NVIDIA Corporation (with the donation of a Titan V GPU), by Google (with the donation of Google Cloud credits). IC has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 838188. This project has benefited from direct support by the NYU IT High Performance Computing Center. This material is based upon work supported by the National Science Foundation under Grant No. 1922658. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Samuel R. Bowman, Gabor Angeli, Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Samuel R Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. Collecting entailment data for pretraining: New protocols and negative results. *arXiv preprint arXiv:2004.11997*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 7059–7069. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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

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

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2020. FILTER: An Enhanced Fusion Method for Cross-lingual Language Understanding. *arXiv e-prints*, page arXiv:2009.05166.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn treebank. *Computational Linguistics*, 33(3):355–396.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *Proceedings of the 37th International Conference on Machine Learning*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019a. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019b. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11(3):311–325.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. *Proceedings of the 37th International Conference on Machine Learning*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, and et al. 2018. Universal Dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP)*.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. Unpublished manuscript available on arXiv.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Phu Mon Htut, , Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Edouard Grave, Najoung Kim, Thibault Févry, Berlin Chen, Nikita Nangia, Anhad Mohananey, Katharina Kann, Shikha Bordia, Nicolas Patry, David Benton, Ellie Pavlick, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

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

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019a. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019b. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. Unpublished manuscript available on arXiv.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-x: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

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

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

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third BUCC shared task:spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshopon Building and Using Comparable Corpora*, pages 39–42.

A Implementation Details

A.1 Intermediate Tasks

For intermediate-task training, we use a learning rate of 1e-5 without MLM, and 5e-6 with MLM. Hyperparameters in the Table 4 were chosen based on intermediate task validation performance in an preliminary search. We use a warmup of 10% of the total number of steps, and perform early stopping based on the first 500 development set examples of each task with a patience of 30. For CCG, where tags are assigned for each word, we use the representation of first sub-word token of each word for prediction.

Task	Batch size	# Epochs
ANLI$^+$	24	2
MNLI	24	2
CCG	24	15
CommonsenseQA	4	10
Cosmos QA	4	15
HellaSwag	24	7
QQP	24	3
SQuAD	8	3
MLM	8	-
Multi-task	Mixed	3

Table 4: Intermediate-task training configuration.

A.2 XTREME Benchmark Target Tasks

We follow the sample implementation for the XTREME benchmark unless otherwise stated. We use a learning rate of 3e-6, and use the same optimization procedure as for intermediate tasks. Hyperparameters in the Table 5 follow the sample implementation. For POS and NER, we use the same strategy as for CCG for matching tags to tokens. For BUCC and Tatoeba, we extract the representations for each token from the 13th self-attention layer, and use the mean-pooled representation as the embedding for that example, as in the sample implementation. Similarly, we follow the sample implementation and set an optimal threshold for each language sub-task for BUCC as a similarity score cut-off for extracting parallel sentences based on the development set and applied to the test set.

We randomly initialize the corresponding output heads for each task, regardless of the similarity between intermediate and target tasks (e.g. even if both the intermediate and target tasks train on SQuAD, we randomly initialize the output head in between phases).

Task	Batch size	# Epochs
XNLI (MNLI)	4	2
PAWS-X	32	5
XQuAD (SQuAD)	16	2
MLQA (SQuAD)	16	2
TyDiQA	16	2
POS	32	10
NER	32	10
BUCC	-	-
Tatoeba	-	-

Table 5: Target-task training configuration.

B Per-Language Results

		ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg
	XLM-R	**79.8**	82.7	**83.8**	81.3	89.3	84.4	83.7	77.3	79.2	72.4	77.1	78.9	72.6	80.0	79.6	80.1
Without MLM	ANLI^{+}	77.5	82.5	82.3	80.8	87.6	83.5	83.6	76.5	79.1	70.4	77.3	78.0	73.5	79.2	79.3	79.4
	MNLI	78.4	82.8	83.0	81.3	88.2	84.0	83.6	77.2	79.5	69.4	77.6	77.9	73.2	79.8	79.1	79.7
	QQP	77.1	81.0	81.6	81.6	86.1	83.6	82.0	75.4	78.5	69.6	76.9	77.1	72.7	79.2	78.6	78.7
	SQuAD v2.0	77.9	81.3	81.7	79.9	85.6	83.5	81.8	75.5	78.5	70.6	77.2	77.2	73.7	78.9	79.6	78.9
	SQuAD v1.1	77.1	82.1	81.8	79.9	87.1	82.8	82.7	75.5	78.6	71.3	76.3	77.3	71.2	79.2	78.6	78.8
	HellaSwag	78.6	82.6	83.5	80.6	88.5	83.7	83.1	77.4	78.2	72.0	77.4	78.7	73.5	80.0	79.4	79.8
	CCG	77.3	81.9	81.7	79.8	88.1	82.9	83.2	75.4	78.8	69.9	76.5	76.9	71.4	79.7	78.6	78.8
	Cosmos QA	77.1	81.1	81.7	80.1	87.4	83.2	81.7	74.3	77.7	72.0	75.2	76.7	71.1	78.3	78.4	78.4
	CSQA	77.3	80.8	81.9	80.0	87.5	83.5	82.5	76.3	78.4	70.6	76.3	77.5	72.5	79.6	78.5	78.9
	Multi-task	76.9	82.2	82.9	81.0	88.5	84.4	82.5	75.8	79.1	71.1	77.1	79.1	72.0	79.6	79.2	79.4
With MLM	ANLI^{+}	78.5	82.8	**83.8**	81.5	89.2	84.1	82.5	76.5	79.2	72.7	77.4	78.6	72.7	80.7	80.1	80.0
	MNLI	78.0	82.9	83.1	81.1	88.8	84.3	83.4	76.7	**80.3**	72.2	**78.4**	79.3	73.4	80.5	80.2	80.2
	QQP	78.0	81.7	83.3	80.8	88.6	**84.5**	82.9	75.9	78.3	72.2	77.7	78.6	72.7	79.9	78.9	79.6
	SQuAD v2.0	77.5	82.8	83.3	80.4	88.8	83.6	82.7	76.0	79.6	71.6	77.0	78.7	72.9	79.9	78.9	79.6
	SQuAD v1.1	77.9	81.7	82.2	79.7	87.0	82.8	82.1	74.4	78.4	71.2	76.6	78.1	71.3	79.0	78.6	78.7
	HellaSwag	79.3	**83.5**	83.7	**81.8**	**89.6**	**84.5**	**84.1**	**78.2**	79.9	72.9	78.1	**80.1**	**74.5**	**81.3**	**80.7**	**80.8**
	CCG	77.9	82.5	82.4	80.8	87.1	83.8	82.6	76.6	78.9	72.0	76.7	78.2	72.2	80.2	78.4	79.4
	Cosmos QA	78.1	82.7	82.7	80.4	87.6	83.9	82.9	76.2	79.5	**73.7**	77.8	79.0	72.7	80.4	79.6	79.8
	CSQA	79.0	83.4	83.7	81.2	89.0	83.8	83.3	76.9	79.9	72.3	78.0	79.1	73.3	80.4	80.6	80.2

Table 6: Full XNLI Results

		de	en	es	fr	ja	ko	zh	Avg
	XLM-R	88.1	93.4	89.2	89.3	81.8	81.8	82.0	86.5
Without MLM	ANLI^{+}	88.0	94.1	89.6	90.7	82.0	82.2	81.9	87.0
	MNLI	89.0	95.0	90.7	90.9	82.9	83.8	84.2	88.1
	QQP	83.9	93.0	87.7	88.7	79.2	78.6	79.7	84.4
	SQuADv2.0	88.9	95.2	**91.7**	91.3	84.7	84.5	**85.4**	88.8
	SQuADv1.1	89.4	94.2	91.1	91.1	83.8	83.5	83.9	88.1
	HellaSwag	88.4	95.0	90.2	91.1	84.8	84.6	84.5	88.4
	CCG	83.5	92.3	86.5	88.1	78.0	77.0	78.6	83.5
	Cosmos QA	88.4	93.8	90.4	90.3	84.3	84.3	85.0	88.1
	CSQA	85.9	93.7	88.6	89.8	81.7	80.4	81.5	86.0
	Multi-task	89.0	95.0	90.2	91.1	83.8	83.5	85.5	88.3
With MLM	ANLI^{+}	88.1	94.5	90.1	90.4	84.0	84.2	84.2	87.9
	MNLI	90.1	**95.5**	91.3	91.3	84.4	84.1	84.5	88.7
	QQP	88.6	94.3	89.8	90.6	81.7	82.8	82.3	87.1
	SQuADv2.0	88.9	95.0	**91.7**	**92.0**	**85.2**	83.9	84.7	88.8
	SQuADv1.1	89.0	93.8	90.3	88.9	82.7	82.2	82.2	87.0
	HellaSwag	**90.3**	95.0	91.0	90.5	84.9	**85.9**	84.8	**88.9**
	CCG	87.5	93.3	88.3	88.4	81.5	81.2	81.3	85.9
	Cosmos QA	88.1	94.0	89.4	90.0	82.5	82.4	82.3	87.0
	CSQA	88.7	94.1	89.1	89.8	82.5	82.9	82.2	87.0

Table 7: Full PAWS-X Results

		af	ar	bg	de	el	en	es	et	eu	fa	fi	fr	he	hi	hu	id	it
	XLM-R	87.7	56.3	87.9	88.6	85.6	95.9	89.8	87.6	72.8	70.0	84.9	**65.5**	68.1	73.2	81.3	**81.7**	88.8
Without MLM	ANLI$^+$	87.9	57.6	88.3	88.8	85.6	95.7	89.4	87.3	73.4	72.0	84.9	65.4	70.9	70.1	**82.9**	81.0	88.3
	MNLI	87.9	56.6	87.8	88.5	84.6	96.2	88.9	86.9	70.4	69.5	84.1	51.8	70.1	72.4	81.2	81.1	88.6
	QQP	83.9	52.6	86.0	85.3	81.7	93.7	87.7	82.1	70.1	66.7	79.3	62.5	61.1	62.5	78.3	79.2	86.8
	SQuADv2.0	87.5	58.0	88.0	87.9	83.6	96.2	88.7	86.6	69.9	69.1	83.9	51.8	**71.3**	69.7	82.6	81.0	89.0
	SQuADv1.1	87.7	58.1	**88.6**	88.4	85.8	95.7	89.4	87.2	73.4	70.1	84.3	65.1	70.9	72.2	81.8	81.3	88.5
	HellaSwag	88.3	57.3	88.5	88.7	85.6	**96.5**	89.2	87.6	72.6	69.5	84.7	52.5	69.6	74.8	81.6	81.1	**89.6**
	CCG	88.2	56.2	86.5	**89.4**	85.9	95.8	87.8	87.9	73.7	69.1	**85.6**	53.5	68.8	75.1	81.8	80.8	86.8
	Cosmos QA	88.4	56.4	86.2	88.0	84.4	95.9	88.9	87.1	73.5	71.2	84.5	65.3	67.5	75.6	81.1	81.0	88.8
	CSQA	87.1	55.7	87.6	87.8	85.8	95.4	88.6	87.3	**76.4**	69.3	84.7	64.6	65.3	67.6	81.2	80.9	86.6
	Multi-task	87.7	58.5	89.7	88.8	85.2	96.3	89.4	87.1	67.7	71.6	84.7	52.7	71.0	68.2	81.5	80.7	89.8
With MLM	ANLI$^+$	87.9	**58.4**	88.3	88.9	**86.3**	95.8	**90.3**	87.8	**76.4**	**72.5**	85.1	53.3	69.0	72.5	82.4	80.7	88.6
	MNLI	**89.1**	57.2	87.6	88.6	85.1	96.2	88.8	88.0	73.4	69.5	85.1	52.7	68.0	76.9	80.6	80.4	88.7
	QQP	87.7	56.3	87.6	88.6	84.2	95.9	89.6	**88.1**	76.3	71.2	84.5	59.7	67.5	**78.0**	81.8	81.2	88.8
	SQuADv2.0	88.5	57.8	87.8	88.5	85.8	96.2	89.0	86.1	74.7	71.0	84.6	49.1	68.2	73.2	81.4	80.8	85.8
	SQuADv1.1	88.0	55.1	**88.6**	88.9	85.3	95.7	89.7	85.7	73.5	70.2	83.5	64.5	66.7	74.4	79.7	81.5	86.8
	HellaSwag	88.3	58.0	87.8	88.3	85.7	96.4	87.2	86.8	74.0	70.2	84.3	51.5	70.9	74.8	79.9	81.0	88.4
	CCG	88.1	54.5	86.7	89.2	**86.3**	95.9	87.5	87.6	77.2	71.4	84.0	64.4	66.3	76.7	81.1	81.4	89.0
	Cosmos QA	87.5	57.8	87.7	88.6	85.5	95.8	89.5	**88.1**	71.7	70.1	84.9	64.4	68.9	76.6	81.0	80.0	88.3
	CSQA	87.6	55.9	87.4	88.7	85.1	95.6	88.5	87.2	**76.4**	70.4	84.2	65.1	68.2	68.3	81.6	81.2	88.4

		ja	kk	ko	mr	nl	pt	ru	ta	te	th	tl	tr	ur	vi	yo	zh	Avg
	XLM-R	31.9	-	50.4	80.0	90.1	**90.2**	89.5	67.1	**90.0**	-	-	76.0	65.6	56.4	-	40.9	75.7
Without MLM	ANLI$^+$	19.4	-	50.7	79.6	90.1	89.7	**90.0**	69.2	86.6	-	-	75.0	66.2	55.3	-	27.2	74.8
	MNLI	38.1	-	50.7	79.1	90.4	89.7	89.4	69.4	86.7	-	-	74.8	67.6	54.4	-	**48.6**	75.4
	QQP	6.2	-	45.9	73.5	88.4	88.2	86.6	65.1	81.7	-	-	71.5	59.1	54.5	-	12.0	70.1
	SQuADv2.0	**39.4**	-	50.8	80.5	90.3	90.1	89.1	68.5	86.1	-	-	74.1	60.6	54.1	-	45.3	75.0
	SQuADv1.1	30.9	-	49.7	78.7	**90.5**	89.7	89.3	66.8	84.9	-	-	74.4	65.4	56.2	-	37.7	75.3
	HellaSwag	31.1	-	50.5	83.7	90.1	89.8	89.5	**69.7**	86.2	-	-	74.2	67.4	54.5	-	35.1	75.2
	CCG	17.8	-	50.3	81.0	90.1	88.0	88.9	66.8	88.4	-	-	75.9	70.7	55.5	-	23.1	74.1
	Cosmos QA	16.4	-	50.3	77.7	89.9	89.7	89.4	67.9	88.1	-	-	**76.5**	69.2	56.3	-	23.2	74.4
	CSQA	32.4	-	49.3	82.8	89.4	88.5	88.5	66.9	86.3	-	-	74.5	63.5	56.0	-	29.6	74.5
	Multi-task	36.4	-	50.7	79.6	90.0	89.8	88.9	68.4	86.2	-	-	74.4	62.2	55.5	-	44.3	75.1
With MLM	ANLI$^+$	39.0	-	51.2	80.7	90.2	90.0	89.8	68.7	87.6	-	-	76.4	66.2	56.7	-	45.7	**76.1**
	MNLI	30.1	-	51.0	80.1	90.0	88.8	89.1	68.8	85.5	-	-	75.1	69.6	55.4	-	38.4	75.1
	QQP	27.6	-	50.8	81.0	90.1	89.5	89.4	67.2	88.0	-	-	76.2	**70.3**	56.5	-	34.0	75.4
	SQuADv2.0	35.3	-	51.0	80.2	89.9	88.1	89.3	67.1	84.3	-	-	75.5	68.8	56.9	-	39.0	75.0
	SQuADv1.1	16.3	-	49.7	79.4	90.2	90.0	89.2	68.0	83.3	-	-	75.8	64.6	**57.3**	-	19.0	73.8
	HellaSwag	35.4	-	50.9	78.4	90.0	87.9	89.3	68.7	86.4	-	-	75.4	69.3	54.8	-	43.6	75.3
	CCG	25.7	-	50.7	**86.1**	89.8	88.8	88.4	68.0	86.6	-	-	76.2	68.2	55.5	-	23.9	75.0
	Cosmos QA	16.5	-	51.0	80.9	89.7	88.9	89.0	67.4	87.9	-	-	76.3	70.1	56.0	-	19.6	74.5
	CSQA	30.8	-	**51.8**	80.5	**90.5**	89.6	89.0	66.8	86.5	-	-	74.8	61.9	56.3	-	31.3	74.8

Table 8: Full POS Results. `kk`, `th`, `tl` and `yo` do not have development set data.

		af	ar	bg	bn	de	el	en	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv	ka
	XLM-R	77.7	47.1	81.9	74.9	78.6	76.3	81.6	74.7	77.2	61.2	58.2	78.3	78.3	50.2	68.7	80.6	53.7	80.8	15.6	56.2	61.4
Without MLM	ANLI$^+$	75.4	52.7	78.1	72.7	76.4	76.3	80.9	71.6	72.8	52.2	60.7	75.8	77.4	49.1	69.6	79.6	52.7	78.9	13.1	54.3	62.1
	MNLI	76.9	48.3	80.5	72.8	77.7	77.9	84.2	76.9	78.5	62.1	58.3	78.7	81.1	55.1	69.0	81.1	55.7	80.8	16.4	54.2	68.1
	QQP	73.8	40.9	75.5	66.0	71.3	71.6	75.8	65.5	69.3	55.5	49.9	73.1	72.8	42.6	59.8	74.3	49.2	75.9	5.7	54.4	51.1
	SQuADv2.0	76.0	48.0	81.1	71.8	78.4	78.2	84.3	74.7	78.4	53.9	56.9	78.9	**82.5**	56.0	68.9	79.8	56.4	80.8	18.1	61.8	67.3
	SQuADv1.1	**79.1**	52.6	80.1	75.5	77.8	78.1	80.8	75.3	76.7	54.3	61.9	78.7	78.4	52.8	65.6	80.3	54.6	80.8	18.7	52.1	62.4
	HellaSwag	77.0	54.9	82.7	**76.6**	79.1	**78.9**	84.3	**77.8**	78.0	58.8	65.0	77.5	80.3	57.0	**71.2**	81.8	54.3	**81.4**	19.6	56.9	70.6
	CCG	77.4	51.5	78.7	72.5	78.4	76.2	80.8	73.0	78.0	56.9	62.1	78.2	77.3	48.6	67.3	79.7	54.9	79.9	15.9	60.3	58.9
	Cosmos QA	76.6	49.3	79.2	76.0	77.8	76.1	81.2	73.2	76.6	59.8	55.8	77.8	77.0	46.8	67.8	79.4	53.2	80.0	14.1	55.5	57.8
	CSQA	77.6	46.1	78.9	75.4	78.4	76.2	81.3	77.3	75.2	59.8	61.9	78.0	78.2	48.9	67.6	79.6	55.6	80.1	11.6	53.8	57.7
	Multi-task	78.5	49.2	82.0	73.3	78.9	80.1	84.5	76.6	78.5	59.4	49.4	79.1	81.2	56.4	70.6	81.0	57.0	80.7	20.7	64.7	68.6
With MLM	ANLI$^+$	76.4	51.5	80.7	73.3	79.2	77.8	84.3	75.4	78.0	57.7	49.7	77.6	80.1	54.8	68.9	80.8	54.8	80.5	14.4	54.9	64.5
	MNLI	78.0	52.3	81.7	73.0	79.6	78.1	84.4	77.2	79.4	59.6	60.6	79.2	81.4	55.1	68.6	81.0	51.3	81.0	14.0	62.0	64.3
	QQP	77.1	46.7	79.0	72.9	79.4	76.3	81.9	74.2	78.7	61.8	**66.0**	78.3	78.0	50.4	69.1	81.6	53.2	80.1	15.1	**62.6**	60.7
	SQuADv2.0	78.0	46.5	**82.8**	71.7	79.0	77.3	84.2	74.8	79.0	61.6	63.3	79.5	80.0	**57.6**	67.5	**81.9**	**62.0**	80.7	**20.0**	62.3	68.2
	SQuADv1.1	77.7	**58.0**	81.4	75.2	78.0	77.4	82.1	69.6	76.1	54.1	58.4	77.5	78.7	54.8	67.5	78.8	49.9	79.5	14.5	55.9	**68.3**
	HellaSwag	78.7	47.0	81.8	73.8	**79.7**	78.2	**84.8**	73.6	79.2	55.8	55.6	78.2	79.4	55.0	69.8	81.3	54.1	81.3	18.5	58.1	67.5
	CCG	74.5	46.4	76.7	74.5	76.9	75.7	80.5	72.6	77.7	58.9	59.6	77.7	77.0	48.1	66.3	80.1	53.4	78.7	13.8	57.1	58.2
	Cosmos QA	78.2	39.1	80.0	73.8	79.0	77.2	81.4	70.3	78.8	65.4	48.9	78.7	77.7	48.3	68.0	80.8	55.1	81.2	13.2	58.9	59.0
	CSQA	77.4	48.8	78.9	73.9	78.8	76.3	81.9	75.2	**79.5**	**66.7**	58.6	**79.6**	78.5	47.7	68.2	81.0	55.3	81.3	12.2	60.4	58.9

		kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh	Avg	
	XLM-R	48.7	54.5	58.8	61.8	54.1	53.7	83.2	80.7	69.3	69.8	58.2	50.8	2.2	73.2	81.1	67.0	74.9	33.2	23.6	62.8	-
Without MLM	ANLI$^+$	50.2	52.6	61.2	63.0	66.8	46.5	81.8	78.7	67.0	66.9	55.0	52.1	2.5	71.2	78.0	67.3	73.9	43.3	18.9	62.0	-
	MNLI	51.7	58.8	64.8	61.3	69.8	54.9	83.0	80.8	70.2	70.3	59.3	**55.4**	1.0	74.8	80.5	56.9	78.1	38.9	25.2	64.2	-
	QQP	50.4	40.1	51.2	51.4	61.4	32.5	78.2	73.0	50.8	65.1	47.3	41.4	1.6	67.4	72.3	57.2	67.9	43.9	8.6	55.9	-
	SQuADv2.0	49.9	58.1	61.6	62.5	72.1	50.0	83.1	82.3	70.8	65.4	62.6	53.6	0.6	74.8	80.0	63.2	**78.9**	41.2	22.5	64.1	-
	SQuADv1.1	51.8	57.1	61.7	59.8	50.4	52.2	83.3	80.8	69.8	69.2	58.3	49.5	0.8	71.6	79.1	58.6	76.3	**47.5**	26.2	63.0	-
	HellaSwag	50.5	58.4	56.6	**66.6**	72.8	**59.4**	83.2	**82.5**	70.8	69.9	**63.7**	53.0	1.1	75.1	78.0	70.0	75.0	42.1	**29.7**	**65.5**	-
	CCG	52.4	52.7	57.7	59.6	52.3	50.0	82.5	79.0	67.1	67.0	55.3	49.1	**2.6**	70.0	81.0	65.3	74.2	37.6	23.3	62.1	-
	Cosmos QA	48.4	52.4	60.3	62.1	56.9	50.2	82.8	79.5	67.4	67.8	57.2	51.4	1.3	74.6	80.7	60.8	74.9	34.8	19.5	61.8	-
	CSQA	49.7	52.0	59.1	62.9	62.4	46.1	82.5	80.3	65.4	69.0	57.1	51.2	1.8	73.1	80.2	**73.3**	73.5	35.3	19.3	62.3	-
	Multi-task	53.2	57.8	60.8	61.0	69.3	54.2	83.8	80.8	69.4	70.6	58.9	53.7	2.2	75.2	77.2	57.7	75.6	46.1	30.4	64.7	-
With MLM	ANLI$^+$	52.9	56.8	60.0	61.1	**75.4**	49.5	**83.4**	80.9	68.3	71.0	57.2	49.8	0.9	74.5	79.0	59.8	76.3	31.7	22.5	63.2	-
	MNLI	**54.7**	57.5	63.5	63.3	66.3	49.6	**83.4**	81.1	70.3	72.2	57.0	53.5	1.1	74.1	80.9	61.1	75.1	43.4	22.8	64.3	-
	QQP	49.9	54.5	63.3	64.6	54.7	49.0	82.9	78.9	68.7	70.9	58.0	50.7	1.1	74.0	**82.3**	70.2	77.1	40.3	24.9	63.5	-
	SQuADv2.0	52.1	**60.8**	**65.1**	63.2	54.7	54.8	**83.4**	80.9	**71.6**	**72.6**	63.0	54.1	0.4	**75.3**	80.4	59.8	77.6	33.6	28.0	64.7	-
	SQuADv1.1	51.6	57.7	62.7	60.2	62.2	52.9	81.8	77.7	71.4	68.5	59.7	49.9	1.5	72.9	78.1	54.2	71.5	34.3	22.4	62.6	-
	HellaSwag	53.6	58.9	62.5	63.2	72.4	54.7	82.8	80.9	71.3	70.6	59.5	52.0	2.4	73.6	80.1	58.4	78.3	36.8	24.9	64.2	-
	CCG	54.6	53.5	60.6	62.8	69.1	41.6	80.7	78.1	65.4	68.1	55.1	51.6	1.3	68.7	79.8	61.9	68.8	37.9	19.8	61.6	-
	Cosmos QA	49.7	52.5	55.7	60.2	52.1	48.1	82.9	78.9	67.1	66.6	55.3	47.7	0.9	74.7	80.8	59.5	74.0	34.9	19.3	61.3	-
	CSQA	52.2	54.4	60.4	61.1	52.9	47.8	**83.4**	80.7	68.5	69.0	57.9	50.1	1.4	73.6	81.5	63.2	74.0	43.6	19.3	62.9	-

Table 9: Full NER Results

		ar	de	el	en	es	hi	ru	th	tr	vi	zh	Avg
	XLM-R	72.5 / 53.4	77.7 / 61.2	77.6 / 59.2	86.3 / 74.2	80.0 / 61.0	73.7 / 57.5	77.7 / 59.8	72.8 / 62.3	72.6 / 54.8	77.6 / 58.0	68.7 / 58.2	76.1 / 60.0
Without MLM	ANLI$^+$	72.9 / 55.0	77.2 / 60.7	75.8 / 58.3	84.9 / 73.1	78.4 / 59.5	73.1 / 56.9	76.8 / 59.9	73.0 / 63.3	72.1 / 55.0	78.0 / 57.6	68.3 / 59.0	75.5 / 59.8
	MNLI	70.7 / 53.2	77.4 / 60.2	76.8 / 59.1	84.2 / 72.6	80.3 / 62.5	72.2 / 55.9	77.8 / 61.3	72.9 / 63.5	71.9 / 56.3	78.1 / 59.7	68.0 / 60.0	75.5 / 60.4
	QQP	68.4 / 50.4	73.2 / 56.5	73.3 / 55.9	82.3 / 70.6	75.4 / 57.3	68.5 / 52.5	74.2 / 57.5	68.6 / 60.2	68.3 / 51.4	72.9 / 53.4	66.3 / 58.0	72.0 / 56.7
	SQuADv2.0	73.8 / 56.0	79.5 / 62.0	78.6 / 60.6	86.7 / 75.5	81.5 / **63.6**	72.7 / 56.2	79.2 / 61.8	71.0 / 56.8	75.0 / 59.1	78.6 / 58.9	68.8 / 57.6	76.9 / 60.7
	SQuADv1.1	**75.9 / 59.9**	**80.3 / 63.6**	**80.3 / 62.1**	**88.3 / 77.4**	**81.8** / 63.2	**76.1 / 59.2**	**80.0 / 64.1**	**75.6 / 65.5**	**75.8 / 59.2**	**80.5 / 61.2**	70.8 / **61.3**	**78.7 / 63.3**
	HellaSwag	73.9 / 56.9	78.7 / 61.3	77.9 / 58.8	86.1 / 75.6	79.6 / 60.1	74.3 / 57.5	78.5 / 62.8	73.6 / 64.5	73.5 / 56.6	78.8 / 59.1	69.2 / 59.4	76.7 / 61.1
	CCG	71.5 / 54.2	76.3 / 58.5	75.9 / 58.2	84.2 / 72.3	79.0 / 60.1	72.3 / 54.9	76.7 / 60.0	71.2 / 60.9	71.7 / 55.3	76.4 / 56.9	67.9 / 58.2	74.8 / 59.0
	Cosmos QA	73.2 / 53.8	78.1 / 62.2	77.3 / 58.3	86.7 / 75.4	79.9 / 61.9	74.2 / 57.7	77.9 / 59.4	72.3 / 61.5	73.3 / 55.6	78.2 / 58.0	68.3 / 58.5	76.3 / 60.2
	CSQA	72.6 / 53.4	79.5 / 62.4	78.3 / 59.4	87.1 / 76.1	81.0 / 62.9	74.9 / 58.5	77.6 / 60.3	69.7 / 58.9	73.4 / 56.5	78.2 / 58.1	67.5 / 57.3	76.3 / 60.3
	Multi-task	73.2 / 56.4	79.1 / 61.8	78.3 / 60.0	85.5 / 74.2	81.1 / 62.9	74.0 / 56.5	77.7 / 61.7	71.6 / 61.8	73.7 / 57.6	78.8 / 59.1	68.1 / 57.0	76.5 / 60.8
With MLM	ANLI$^+$	72.1 / 52.4	77.3 / 59.8	76.1 / 57.6	85.8 / 74.1	78.7 / 58.8	72.9 / 55.3	76.9 / 59.4	73.0 / 63.4	72.3 / 55.3	78.5 / 57.8	70.9 / 61.0	75.9 / 59.5
	MNLI	72.5 / 54.8	78.4 / 60.7	77.8 / 60.4	86.4 / 75.5	80.4 / 61.3	73.6 / 56.6	78.2 / 61.7	73.9 / 64.5	72.5 / 57.5	79.0 / 60.3	69.0 / 59.7	76.5 / 61.2
	QQP	72.8 / 55.3	78.8 / 61.6	76.9 / 58.8	85.9 / 74.4	79.8 / 61.2	73.9 / 56.3	78.1 / 61.3	72.0 / 61.0	73.4 / 57.7	78.2 / 59.0	67.6 / 57.2	76.1 / 60.4
	SQuADv2.0	72.3 / 55.0	79.0 / 63.3	76.9 / 58.6	85.3 / 73.9	80.3 / 61.9	73.1 / 56.9	77.8 / 61.7	72.5 / 61.1	72.8 / 55.8	77.8 / 58.2	68.4 / 58.6	76.0 / 60.4
	SQuADv1.1	73.3 / 56.1	79.0 / 62.9	78.8 / 60.5	86.6 / 75.5	80.7 / 62.4	74.6 / 57.2	79.2 / 62.8	71.2 / 58.9	73.8 / 56.3	79.4 / 60.6	69.3 / 59.6	76.9 / 61.2
	HellaSwag	73.3 / 56.2	77.4 / 59.7	78.0 / 58.7	85.1 / 73.6	79.8 / 61.2	74.7 / 57.6	77.9 / 61.0	72.7 / 61.8	73.2 / 57.6	77.8 / 58.8	67.7 / 58.3	76.1 / 60.4
	CCG	71.8 / 53.2	77.4 / 60.5	75.7 / 56.9	84.8 / 72.9	79.3 / 60.1	73.1 / 55.8	75.8 / 57.1	70.3 / 58.3	71.7 / 55.6	77.2 / 57.0	66.9 / 57.4	74.9 / 58.6
	Cosmos QA	72.5 / 53.9	77.2 / 61.2	76.9 / 59.1	85.1 / 72.9	79.2 / 60.6	73.4 / 57.5	76.4 / 57.7	72.0 / 61.7	72.1 / 55.1	77.4 / 57.6	68.6 / 59.0	75.5 / 59.6
	CSQA	73.0 / 54.0	77.6 / 60.7	77.4 / 58.7	86.2 / 74.5	80.3 / 61.1	73.1 / 57.3	77.8 / 59.9	71.4 / 59.6	72.1 / 55.0	77.9 / 58.7	**71.2** / 60.7	76.2 / 60.0

Table 10: Full XQuAD Results

		ar	de	en	es	hi	vi	zh	Avg
	XLM-R	62.7 / 42.4	69.1 / 52.0	81.6 / 68.6	72.2 / 53.0	68.0 / 50.7	69.5 / 47.6	67.9 / 46.2	70.1 / 51.5
Without MLM	ANLI^{+}	64.1 / 43.9	66.8 / 49.8	82.5 / 69.4	71.9 / 52.6	69.2 / 50.5	70.5 / 49.7	66.9 / 44.8	70.3 / 51.5
	MNLI	64.2 / 43.5	68.1 / 51.8	82.7 / 70.0	73.7 / 54.8	70.3 / 52.7	68.9 / 49.5	67.1 / 46.0	70.7 / 52.6
	QQP	60.5 / 39.7	62.4 / 45.5	79.0 / 66.0	70.7 / 51.6	62.9 / 45.4	67.0 / 47.6	63.5 / 41.1	66.6 / 48.1
	SQuADv2.0	66.1 / 45.3	68.2 / 50.2	83.5 / **71.1**	73.6 / 55.4	68.5 / 51.5	71.7 / **52.4**	68.2 / 46.4	71.4 / 53.2
	SQuADv1.1	**67.4** / **46.4**	**69.6** / 52.9	**84.1** / 70.8	**75.3** / **56.8**	**72.5** / **54.8**	70.9 / 51.7	**69.4** / 47.0	**72.8** / **54.4**
	HellaSwag	64.2 / 43.1	68.8 / 52.3	83.5 / 70.9	73.0 / 53.6	69.2 / 51.7	69.8 / 48.7	68.5 / 46.2	71.0 / 52.4
	CCG	62.7 / 41.6	67.5 / 50.4	82.9 / 70.0	72.9 / 54.6	66.1 / 50.1	68.9 / 48.9	66.4 / 45.6	69.6 / 51.6
	Cosmos QA	63.8 / 43.9	68.2 / 50.4	82.2 / 69.0	72.9 / 54.2	69.4 / 51.7	70.8 / 50.1	66.6 / 44.4	70.6 / 52.0
	CSQA	64.0 / 43.9	68.8 / 52.0	83.4 / 70.6	75.2 / 55.0	69.1 / 51.5	**72.6** / 52.1	69.2 / 46.6	71.8 / 53.1
	Multi-task	65.1 / 44.1	70.2 / 54.9	82.9 / 69.4	75.2 / 56.4	70.1 / 52.3	72.0 / 51.7	68.6 / 46.2	72.0 / 53.6
With MLM	ANLI^{+}	62.7 / 41.8	68.5 / 51.4	82.1 / 69.0	73.6 / 54.2	66.7 / 48.7	69.5 / 49.3	66.2 / 44.2	69.9 / 51.2
	MNLI	62.9 / 41.0	69.2 / **53.5**	82.6 / 69.4	74.3 / 54.4	68.0 / 50.7	70.5 / 50.5	68.0 / 45.8	70.8 / 52.2
	QQP	64.6 / 44.9	68.1 / 51.2	83.2 / 70.4	74.0 / 55.6	70.4 / 53.1	69.1 / 49.3	68.3 / 45.6	71.1 / 52.9
	SQuADv2.0	64.7 / 43.9	66.6 / 51.0	82.1 / 69.6	73.1 / 55.2	70.2 / 53.1	69.0 / 51.1	68.6 / **47.2**	70.6 / 53.0
	SQuADv1.1	64.4 / 43.3	68.0 / 50.0	83.1 / 70.0	75.2 / 56.2	68.5 / 51.9	71.2 / 51.9	66.8 / 44.6	71.0 / 52.6
	HellaSwag	64.7 / 44.3	68.4 / 52.3	83.3 / 70.4	73.9 / 55.0	69.5 / 52.1	69.9 / 47.9	67.7 / 44.8	71.1 / 52.4
	CCG	60.4 / 41.4	66.5 / 50.8	81.8 / 68.6	72.8 / 54.2	66.2 / 48.7	67.7 / 46.2	64.5 / 44.6	68.6 / 50.7
	Cosmos QA	63.4 / 43.1	69.0 / 51.0	81.9 / 68.9	72.3 / 53.6	66.3 / 48.9	69.1 / 47.6	66.0 / 45.2	69.7 / 51.2
	CSQA	64.3 / 43.7	69.5 / 51.8	82.6 / 69.4	73.4 / 54.4	68.0 / 50.7	70.9 / 48.7	67.7 / 45.8	70.9 / 52.1

Table 11: Full MLQA Results

		ar	bn	en	fi	id	ko	ru	sw	te	Avg
	XLM-R	64.5 / 46.9	59.5 / 41.6	70.4 / 56.6	64.9 / 49.2	75.1 / 59.8	54.7 / 39.5	65.4 / 43.6	67.2 / 48.7	68.8 / 48.3	65.6 / 48.2
Without MLM	ANLI^{+}	67.3 / 47.8	54.9 / 37.2	71.0 / 57.3	64.7 / 47.8	74.9 / 57.5	54.5 / 41.3	62.4 / 33.0	67.2 / 47.3	68.2 / 46.9	65.0 / 46.2
	MNLI	67.8 / 49.7	60.6 / 40.7	71.6 / 57.7	66.5 / 48.6	76.6 / 61.9	55.3 / 42.4	63.9 / 39.0	66.9 / 48.5	71.0 / 51.4	66.7 / 48.9
	QQP	63.2 / 44.4	43.8 / 26.5	64.4 / 52.7	56.3 / 39.9	71.6 / 57.0	47.5 / 32.6	57.4 / 38.2	54.5 / 36.5	45.5 / 26.2	56.0 / 39.3
	SQuADv2.0	76.5 / 59.8	77.7 / 63.7	76.1 / 63.2	78.3 / 64.3	83.1 / 69.9	68.1 / 56.5	73.0 / 51.5	79.1 / 67.1	79.2 / 61.1	76.8 / 61.9
	SQuADv1.1	76.1 / 60.0	75.6 / 61.9	77.6 / 66.6	76.0 / 61.3	82.5 / 68.3	63.7 / 51.4	71.1 / 44.7	76.5 / 63.5	79.0 / 61.6	75.3 / 59.9
	HellaSwag	69.9 / 49.4	60.6 / 42.5	72.2 / 59.1	63.0 / 44.1	76.7 / 60.4	54.7 / 39.1	61.4 / 33.0	66.3 / 48.3	70.6 / 47.8	66.1 / 47.1
	CCG	63.6 / 41.8	54.1 / 37.2	68.5 / 55.9	59.6 / 41.7	73.2 / 57.5	50.8 / 37.7	60.2 / 33.4	66.8 / 49.7	66.2 / 43.8	62.6 / 44.3
	Cosmos QA	71.7 / 51.9	65.9 / 48.7	73.3 / 61.6	66.7 / 50.9	78.5 / 63.4	52.6 / 36.6	66.2 / 44.1	68.0 / 51.3	74.5 / 54.7	68.6 / 51.5
	CSQA	70.9 / 52.1	67.8 / 49.6	74.6 / 60.9	69.6 / 52.6	77.0 / 60.2	60.8 / 46.4	63.6 / 36.0	70.8 / 53.5	73.3 / 54.7	69.8 / 51.8
	Multi-task	73.3 / 52.3	66.7 / 48.7	75.6 / 63.6	74.7 / 59.6	81.7 / 67.3	60.2 / 46.4	71.0 / 43.0	76.0 / 64.3	77.2 / 58.4	72.9 / 56.0
With MLM	ANLI^{+}	67.1 / 48.9	59.5 / 42.5	72.2 / 58.9	67.2 / 51.4	76.8 / 60.7	54.9 / 42.0	62.4 / 35.3	70.3 / 52.1	70.4 / 53.1	66.8 / 49.4
	MNLI	67.3 / 49.7	60.0 / 41.6	71.2 / 59.3	66.8 / 50.4	78.1 / 62.1	56.4 / 42.0	62.2 / 33.9	68.5 / 50.7	70.0 / 48.4	66.7 / 48.7
	QQP	67.8 / 49.0	55.7 / 37.2	69.8 / 56.1	64.1 / 47.1	74.2 / 58.6	49.0 / 34.4	60.0 / 34.5	64.5 / 45.7	70.1 / 45.6	63.9 / 45.3
	SQuADv2.0	76.9 / 60.5	70.1 / 54.9	76.6 / 64.5	74.4 / 59.6	83.4 / 69.7	61.6 / 48.6	71.3 / 45.2	74.0 / 61.5	76.7 / 59.3	73.9 / 58.2
	SQuADv1.1	77.0 / 59.3	68.5 / 51.3	75.4 / 64.3	77.2 / 63.4	83.3 / 71.0	63.7 / 51.8	71.7 / 47.9	73.1 / 56.5	76.4 / 59.0	74.0 / 58.3
	HellaSwag	68.8 / 50.4	62.6 / 47.8	70.9 / 56.8	64.0 / 48.6	77.4 / 61.8	54.6 / 40.9	61.2 / 31.7	68.2 / 49.5	71.4 / 50.5	66.6 / 48.7
	CCG	68.1 / 49.1	57.5 / 39.8	69.0 / 55.9	65.9 / 48.6	76.5 / 61.9	55.0 / 39.9	61.6 / 31.9	67.5 / 49.3	56.3 / 30.3	64.2 / 45.2
	Cosmos QA	66.6 / 46.6	56.8 / 37.2	71.5 / 58.0	64.2 / 45.0	75.0 / 57.0	56.3 / 41.3	63.6 / 39.0	69.0 / 51.1	63.6 / 46.3	65.2 / 46.8
	CSQA	68.8 / 50.4	60.2 / 43.4	71.3 / 59.1	67.6 / 50.5	76.9 / 59.8	54.0 / 41.3	63.5 / 38.1	69.5 / 52.9	72.8 / 54.1	67.2 / 49.9

Table 12: Full TyDiQA Results

		de	fr	ru	zh	Avg
	XLM-R	77.7	62.7	79.2	66.5	71.5
Without MLM	ANLI$^+$	94.6	89.8	93.5	88.6	91.6
	MNLI	94.2	90.2	93.5	**89.9**	92.0
	QQP	94.2	91.0	93.3	88.5	91.8
	SQuADv2.0	94.0	89.8	93.0	**89.9**	91.7
	SQuADv1.1	94.2	90.5	93.1	87.0	91.2
	HellaSwag	94.6	**91.9**	**93.9**	88.9	**92.3**
	CCG	88.3	82.9	86.6	78.0	83.9
	Cosmos QA	94.1	90.2	93.2	88.6	91.5
	CSQA	**95.1**	90.6	93.5	89.1	92.1
	Multi-task	94.3	90.4	93.4	87.0	91.3
With MLM	ANLI$^+$	93.4	88.0	92.9	86.5	90.2
	MNLI	92.7	89.0	93.2	86.1	90.3
	QQP	90.8	86.9	90.6	83.6	88.0
	SQuADv2.0	92.8	87.0	91.4	85.8	89.2
	SQuADv1.1	92.9	89.5	92.7	85.3	90.1
	HellaSwag	92.6	87.5	91.4	86.6	89.5
	CCG	87.6	78.5	87.6	75.7	82.4
	Cosmos QA	91.8	86.9	91.7	88.4	89.7
	CSQA	86.1	80.8	87.9	81.6	84.1

Table 13: Full BUCC Results

		af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv
	XLM-R	30.5	20.4	39.0	13.3	63.9	18.9	48.0	25.8	19.9	42.0	41.5	48.1	28.0	38.3	42.5	47.0	42.3	41.8	10.2
Without MLM	ANLI$^+$	78.8	74.0	88.0	72.3	97.4	82.4	91.2	70.9	53.3	91.5	88.6	89.8	82.1	92.8	86.2	**92.1**	82.6	88.7	31.7
	MNLI	79.6	70.7	84.8	71.2	96.6	82.5	93.1	74.3	59.2	90.0	89.0	89.6	81.8	91.7	86.0	91.7	86.3	89.5	30.7
	QQP	**80.4**	74.9	87.3	74.3	96.5	84.1	**93.8**	74.7	60.2	91.0	90.3	89.9	**86.0**	93.3	88.4	**92.1**	86.3	89.9	35.6
	SQuADv2.0	73.7	67.7	84.2	63.2	96.0	74.3	89.2	70.5	54.0	87.9	85.5	87.1	77.1	88.0	83.5	89.5	80.2	86.4	32.2
	SQuADv1.1	76.9	68.9	85.7	65.7	96.4	76.3	89.5	76.9	58.4	88.0	88.5	88.5	77.3	89.9	84.0	90.4	83.0	88.7	30.2
	HellaSwag	78.9	**75.4**	**89.9**	**75.4**	**97.7**	**84.8**	93.1	**79.8**	64.8	**91.8**	**92.0**	**92.2**	84.9	**93.4**	**89.5**	**92.1**	**86.7**	**91.6**	**37.1**
	CCG	71.9	59.1	82.1	62.5	95.5	74.4	87.0	67.3	49.0	84.7	82.6	84.4	77.2	85.4	80.7	87.2	79.1	78.7	24.9
	Cosmos QA	78.6	70.6	86.6	71.0	96.4	80.5	91.8	77.6	60.7	89.8	91.3	89.4	83.0	91.5	87.7	91.4	83.7	88.2	37.1
	CSQA	79.5	74.5	87.7	74.0	96.9	83.6	92.9	79.1	65.8	90.0	**92.0**	90.7	83.1	92.2	88.4	91.8	85.4	88.9	33.7
	Multi-task	81.2	71.9	88.0	73.6	97.1	82.9	92.6	73.1	58.6	90.4	89.6	89.6	84.1	92.6	87.2	92.6	83.9	91.0	34.1
With MLM	ANLI$^+$	78.6	65.2	86.6	67.8	97.0	78.2	90.2	79.1	59.3	89.3	89.1	90.4	78.7	89.3	86.5	91.0	84.6	87.0	26.3
	MNLI	77.3	65.2	83.8	64.9	97.2	76.1	92.1	77.7	57.3	88.1	88.8	87.5	81.0	89.0	87.1	90.5	82.6	85.6	27.3
	QQP	74.4	61.3	83.7	64.6	96.2	75.7	88.1	76.7	59.4	86.3	87.0	86.9	76.6	85.9	84.2	89.8	79.8	84.0	28.8
	SQuADv2.0	70.8	57.6	80.9	52.7	96.6	63.4	84.5	71.5	47.4	85.4	86.9	85.1	71.9	85.2	83.9	90.4	78.1	83.2	16.1
	SQuADv1.1	79.2	67.7	86.5	71.4	96.7	80.4	91.6	83.1	**66.3**	90.8	91.1	89.8	77.5	92.3	87.4	91.8	84.6	87.4	26.3
	HellaSwag	57.1	45.2	69.4	40.4	89.7	57.8	73.4	64.0	42.2	77.1	76.4	76.5	62.6	75.1	76.2	82.5	69.7	77.5	22.0
	CCG	71.9	52.3	80.4	51.0	95.0	72.6	86.0	73.5	51.0	83.3	84.1	81.8	71.3	79.1	81.6	87.2	78.7	76.2	12.7
	Cosmos QA	69.7	63.7	84.0	58.8	95.1	74.2	84.6	76.5	58.6	85.7	85.2	84.5	76.2	87.1	84.7	88.5	81.4	85.5	24.9
	CSQA	54.3	45.3	63.6	33.5	87.0	50.5	70.0	58.8	35.7	74.1	71.0	70.7	58.2	70.2	72.5	80.4	64.2	75.5	16.6
		ka	kk	ko	ml	mr	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	zh	Avg	
	XLM-R	11.8	17.4	35.5	19.4	15.2	52.6	47.2	42.1	7.9	9.1	19.7	27.4	10.3	37.8	22.5	38.3	41.2	31.0	-
Without MLM	ANLI$^+$	76.9	67.3	84.6	90.8	80.5	**93.6**	91.0	90.5	30.8	76.5	85.5	**91.2**	59.9	87.9	79.7	94.6	**93.0**	80.8	-
	MNLI	77.9	67.7	84.3	89.8	80.4	92.5	91.3	89.2	32.8	70.0	78.2	86.7	60.9	88.8	74.5	92.5	91.2	80.2	-
	QQP	78.7	69.4	86.4	**92.9**	**82.9**	93.3	**92.5**	91.6	35.1	**81.4**	**90.6**	90.0	64.6	**91.4**	81.7	95.0	92.3	82.7	-
	SQuADv2.0	67.0	63.0	80.8	82.8	71.6	89.7	90.4	86.9	27.7	60.9	74.4	80.7	54.2	85.9	70.6	92.5	89.3	76.1	-
	SQuADv1.1	70.9	63.7	83.3	87.3	74.7	91.7	90.2	89.1	31.5	60.6	77.8	82.3	59.3	88.3	68.3	92.8	90.8	77.9	-
	HellaSwag	**80.8**	**72.0**	**86.5**	92.1	81.1	93.2	91.9	**92.0**	35.1	79.2	87.2	89.6	64.5	90.6	**82.4**	**95.1**	92.6	**83.3**	-
	CCG	65.1	56.9	76.8	82.5	70.3	88.9	88.8	84.5	24.9	60.3	65.4	72.8	53.3	82.6	64.7	89.7	84.8	72.9	-
	Cosmos QA	75.7	69.9	83.6	90.1	78.7	92.0	91.3	89.7	34.1	72.3	84.6	89.1	59.7	89.6	79.8	93.3	90.9	80.9	-
	CSQA	**80.8**	70.3	85.5	91.7	82.7	93.3	91.4	90.4	**35.9**	73.3	84.6	89.4	**65.4**	90.2	77.1	94.8	92.9	82.2	-
	Multi-task	78.7	68.2	85.0	91.4	80.4	92.1	92.0	90.2	34.4	68.7	83.8	89.1	62.3	88.9	77.6	95.0	92.8	81.2	-
With MLM	ANLI$^+$	70.6	64.7	83.6	88.9	75.6	92.0	91.0	88.1	29.0	70.0	76.9	84.7	51.6	88.0	71.7	93.6	91.6	78.5	-
	MNLI	67.7	63.3	81.8	84.3	75.0	90.8	90.5	87.8	29.7	62.2	73.5	85.2	53.4	87.6	71.2	93.3	88.5	77.4	-
	QQP	66.0	64.2	80.2	82.0	70.6	89.4	89.8	86.7	30.5	60.9	76.1	83.6	52.3	84.9	72.7	90.5	88.0	76.0	-
	SQuADv2.0	53.8	54.8	77.5	72.5	61.5	90.0	87.0	87.2	20.3	41.7	51.7	80.5	38.0	81.8	63.3	90.6	89.1	70.4	-
	SQuADv1.1	73.2	66.8	83.9	89.8	78.9	93.0	90.4	89.7	33.8	76.2	85.0	90.0	54.5	90.0	78.6	93.6	90.9	80.6	-
	HellaSwag	38.5	43.1	70.5	63.2	39.7	79.1	78.4	80.0	19.2	30.9	55.6	66.6	33.1	71.5	49.8	80.4	77.7	61.4	-
	CCG	58.3	51.3	74.6	76.3	58.4	89.0	86.9	82.9	23.3	46.9	60.3	72.6	40.9	82.5	55.8	87.9	80.3	69.4	-
	Cosmos QA	63.3	56.0	80.7	79.0	63.1	89.4	87.2	86.1	26.2	55.7	71.8	80.5	44.6	83.0	63.7	91.0	85.1	73.8	-
	CSQA	33.4	36.2	65.9	47.0	30.9	76.6	74.7	75.5	19.0	28.3	49.6	64.1	26.0	64.1	53.0	78.4	75.1	56.9	-

Table 14: Full Tatoeba Results

	MNLI	QQP	HellaSwag
`en`	87.1	88.0	71.6
Translated to `de`	82.2	84.6	55.1
Translated to `ru`	70.1	83.8	27.4
Translated to `sw`	70.8	79.3	25.1

Table 15: Intermediate task performance on trained and evaluated on translated data. We report the median result for English (original) task data.