Kardeş-NLU: Transfer to Low-Resource Languages with the Help of a High-Resource Cousin – A Benchmark and Evaluation for Turkic Languages

Lütfi Kerem Şenel[1,2,*], Benedikt Ebing[3,*], Konul Baghirova[1]
Hinrich Schütze[1,2] and Goran Glavaš[3]
[1]Center for Information and Language Processing (CIS), LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Germany
[3]University of Würzburg
lksenel@cis.lmu.de, {benedikt.ebing, goran.glavas}@uni-wuerzburg.de

Abstract

Cross-lingual transfer (XLT) driven by massively multilingual language models (mmLMs) has been shown largely ineffective for low-resource (LR) target languages with little (or no) representation in mmLM's pretraining, especially if they are linguistically distant from the high-resource (HR) source language. Much of the recent focus in XLT research has been dedicated to *LR language families*, i.e., families without any HR languages (e.g., families of African languages or indigenous languages of the Americas). In this work, in contrast, we investigate a configuration that is arguably of practical relevance for more of the world's languages: XLT to LR languages that do have a close HR relative. To explore the extent to which a HR language can facilitate transfer to its LR relatives, we (1) introduce Kardeş-NLU,[1] an evaluation benchmark with language understanding datasets in five LR Turkic languages: Azerbaijani, Kazakh, Kyrgyz, Uzbek, and Uyghur; and (2) investigate (a) intermediate training and (b) fine-tuning strategies that leverage Turkish in XLT to these target languages. Our experimental results show that both—integrating Turkish in intermediate training and in downstream fine-tuning—yield substantial improvements in XLT to LR Turkic languages. Finally, we benchmark cutting-edge instruction-tuned large language models on Kardeş-NLU, showing that their performance is highly task- and language-dependent.

1 Introduction

Transformer-based massively multilingual language models (mmLMs), such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020a), and mT5 (Xue et al., 2021), have substantially advanced multilingual NLP. These models have enabled rapid development of language technologies for a wide range of low-resource (LR) languages by means of cross-lingual transfer (XLT) from high-resource (HR) languages, using zero-shot (Wu and Dredze, 2019; Karthikeyan et al., 2020) or few-shot transfer techniques (Lauscher et al., 2020; Schmidt et al., 2022). mmLMs are, however, biased towards HR languages and XLT with mmLMs yields especially poor transfer performance for LR target languages that are (i) underrepresented in mmLMs' pretraining corpora and (ii) linguistically distant from the source language (Lauscher et al., 2020). Besides these reasons, such poor XLT is also a consequence of the *curse of multilinguality* (Conneau et al., 2020a; Pfeiffer et al., 2022), i.e., a reduced representational quality of supported languages, stemming from mmLMs' parameters being shared by many linguistically diverse languages.

In recent years, a large body of work focused on improving XLT abilities of mmLMs, ranging from models that aim to better align representation subspaces of source and target language with cross-lingual supervision (Cao et al., 2020; Hu et al., 2021; Conneau et al., 2020b; Minixhofer et al., 2022; Wang et al., 2022) to those that improve the mmLMs' representational capacity for individual, mostly LR languages (Pfeiffer et al., 2020; Parović et al., 2022; Ansell et al., 2021; Pfeiffer et al., 2022). At the same time, an incredible amount of effort has also been dedicated to the creation of new multilingual evaluation benchmarks that either encompass sets of linguistically diverse languages (Clark et al., 2020; Ponti et al., 2020; Ruder et al., 2021) or focus on LR languages (Adelani et al., 2021; Muhammad et al., 2022; Ebrahimi et al., 2022; Armstrong et al., 2022; Winata et al., 2023; Khanuja et al., 2023, *inter alia*). The vast majority of existing work, however, assumes (i) zero-shot downstream transfer from (ii) English as the source. That is primarily because, on the one hand, for most tasks, training data is only available in English. On the

[*] These authors contributed equally.
[1] https://github.com/lksenel/Kardes-NLU

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1672–1688
March 17-22, 2024 ©2024 Association for Computational Linguistics

other hand, many of the recent benchmarks cover *LR language families*, i.e., families without *any* HR languages (e.g., some African language families or indigenous languages of the Americas): this prevents the creation of high-quality silver-standard training data in a (closely) related HR language (e.g., via machine translation (MT)), as no such language exists.

Contributions. **1)** In this work, we contribute to the body of evaluation resources for LR XLT with Kardeş-NLU,[2] an evaluation benchmark covering three natural language understanding (NLU) tasks—natural language inference (NLI), semantic text similarity (STS), and commonsense reasoning, in particular choice of plausible alternatives (COPA)—for five Turkic languages—Azerbaijani (az), Kazakh (kk), Kyrgyz (ky), Uyghur (ug), and Uzbek (uz). We focus on Turkic languages because, unlike most concurrent work, we aim to explore a highly underinvestigated XLT research question: to what extent can LR languages that *do have* a linguistically and genealogically (close) HR relatives profit from those relatives (Snæbjarnarson et al., 2023). **2)** We extend a number of established (i) intermediate training and (ii) fine-tuning approaches (covering both zero-shot and few-shot XLT) for improving LR XLT by incorporating Turkish as the HR sibling of the Kardeş-NLU languages; and show that the mixture of incorporating Turkish in intermediate training and in task-specific fine-tuning results in substantial performance gains. **3)** Given the praised generalization abilities of large instruction-based language models (LLMs) (Chung et al., 2022; Ahuja et al., 2023; Asai et al., 2023), we additionally evaluate (zero-shot) two multilingual LLMs on Kardeş-NLU—the open mT0 (Muennighoff et al., 2023) and commercial ChatGPT—showing that their performance is highly task- and language-dependent and in some cases substantially trails that of XLT with traditionally fine-tuned "small" mmLMs.

2 Kardeş-NLU Benchmark

Language and Task Selection. We selected languages for Kardeş-NLU based on two criteria: (i) linguistic and genealogical diversity within the Turkic language family and (ii) availability of native speakers of those languages who are also fluent in English.[3] Our final selection contains five languages from the Common Turkic branch, covering three different sub-branches: Western Oghuz languages (Azerbaijani; Turkish, as the HR language in our experiments, also belongs to this branch), Kipchak languages (Kazakh and Kyrgyz) and Karluk languages (Uzbek and Uyghur). Moreover, Kardeş-NLU covers languages with two different scripts: Latin (Azerbaijani and Uzbek) and Cyrillic (Kazakh, Kyrgyz, and Uyghur).[4]

We select three tasks that are (i) among the most prominent NLU tasks, included in popular NLU benchmarks (Wang et al., 2018, 2019), and (ii) already have existing evaluation datasets in a number of languages (commonly translations of an original English dataset): NLI (Conneau et al., 2018; Aggarwal et al., 2022; Ebrahimi et al., 2022), STS (Cer et al., 2017), and COPA (Gordon et al., 2012; Ponti et al., 2020).

Dataset Translation. We adopt a widely used two-step translation approach to obtain translations in which a native speaker of the target language, fluent in English, post-edits the output of MT.[5] This way, we translated English instances from the following datasets: XNLI (Conneau et al., 2018) (2000 instances from the test portion and 1000 instances from the validation portion), STS-Benchmark (Cer et al., 2017) (800 test instances and 200 validation instances), and XCOPA (Ponti et al., 2020) (500 test instances and 100 validation instances). We initially manually compared, on a small subsample of instances from all three datasets, translation (i) with Google Translate (GT) vs. the open Turkic Interlingua MT models (Mirzakhalov et al., 2021) and (ii) from English vs. from Turkish (with Turkish instances that were, in turn, machine translated from English) and have found that GT from English produces the best output. Due to MT in the first step, we instructed the annotators to pay special attention to the idiomaticity of the source English sentences during post-editing. This particularly refers to finding suitable translations for culture-specific concepts that do not have a direct translation (e.g.,

[2] *kardeş* is a Turkish gender-neutral word for *sibling*. Referring to a brother (*erkek kardeş*) or sister (*kız kardeş*), requires an additional gender denotation: *kız* (*girl*) or *erkek* (*boy*).

[3] For example, we wanted to include Chuvash, the only living language of the Oghur branch of Turkic languages, but we could not find annotators native in that language.

[4] While Uyghur is more commonly written in the Arabic script (e.g., in CC-100 or Wikipedia), our Uyghur annotator was unfamiliar with it and was only able to produce Uyghur translations in the Cyrillic script.

[5] We hired one annotator per target language.

"passing for white" has no direct translation in our target languages since *racial passing* is not a native concept in respective cultures). Table 1 displays several instances from Kardeş-NLU.

Annotation Costs. Given the high post-editing costs, Kardeş-NLU contains only subsets of the original English development and test portions of STS-B and XNLI. All of our annotators were university students who were paid the equivalent of 14$ per hour for their effort. On average, post-editing took 92 hours per language, bringing the total cost of creating Kardeş-NLU to 6,440$.

3 Kardeş Transfer: Leveraging Turkish

We next attempt to improve XLT to LR Kardeş-NLU languages by explicitly incorporating Turkish as the close HR relative into the process. We try to (1) increase mmLMs' capacity for the target languages as well as their alignment with Turkish via intermediate LM training and (2) leverage Turkish as an additional source language in downstream zero-shot and few-shot transfer.

3.1 Intermediate Language Modeling

Adapting pretrained mmLMs to target distributions—different languages, domains, or datasets—through further LM-ing can bring significant performance gains (Howard and Ruder, 2018; Gururangan et al., 2020; Muller et al., 2021; Wang et al., 2022; Hung et al., 2022). Building upon these findings, we investigate the benefit of additional LM-ing in transfer to LR Kardeş-NLU languages. Specifically, we explore the potential benefits of incorporating Turkish into the mmLM adaptation process and the extent to which this inclusion can improve the downstream performance for LR Turkic languages. We experiment with three different intermediate training strategies detailed below: in all cases, we (1) use the standard masked language modeling (MLM) as the training objective and (2) update all of the mmLM's pretrained weights.

Target Language LM-ing (TLLM). In this case, we perform additional MLM-ing only on the limited-size corpora of the target language. Turkish, as the HR relative, is not leveraged in TLLM.

Bilingual Alternating LM-ing (BALM). Here we alternately update the mmLM by MLM-ing on one batch of target language data, followed by one batch of Turkish data. BALM is similar to the bilingual training procedure of Parović et al. (2022): they, however, opt for parameter-efficient training with adapters, whereas we update all of the mmLM's parameters.

Bilingual Joint LM-ing (BJLM). Like BALM, in BJLM we perform bilingual MLM-ing on both the LR target language and the related HR language (Turkish). However, while in BALM monolingual batches are alternated, in BJLM batches are bilingual, i.e., they consist of instances of both languages. Importantly, both languages have the same number of instances in each batch (i.e., B/2 with B as the batch size). Although such balancing leads to frequent repetition of instances from the LR language corpus, these repeating instances are, in different batches, "regularized" with different source-language instances, which prevents overfitting to small-sized corpora of LR languages. Schmidt et al. (2022) demonstrate the effectiveness of BJLM in task-specific few-shot fine-tuning; here, we test it in intermediate MLM-ing.

Parameter-Efficient LM-ing. Besides full fine-tuning, we also carried out intermediate training (for TLLM and BALM) in a parameter-efficient manner with adapters (Houlsby et al., 2019) in the vein of prior work on XLT (Pfeiffer et al., 2020; Parović et al., 2022). Adapter-based variants yielded consistently weaker performance compared to tuning all mmLM's parameters. For brevity, we report these results in the Appendix (§C).

3.2 Downstream Cross-Lingual Transfer

We investigate two common setups for downstream cross-lingual transfer: (1) zero-shot XLT, in which we assume that we do not have any labeled task instances in the target language, and (2) few-shot transfer, in which a small number of labeled instances in the target language exists. We follow the fair XLT evaluation procedure of Schmidt et al. (2022), which does not allow for model selection based on target-language validation data. Relying on target-language validation violates the assumption of true zero-shot XLT. Moreover, Schmidt et al. (2022, 2023a) show that any labeled target-language instances are better leveraged for training. We thus use the validation portions of Kardeş-NLU

Language	Task	Instance	Label
Azerbaijani	NLI	*Premise*: Bütün hallarda müştərinin iddialarına xələl gətirməmək üçün mühüm addımlar atılmalıdır. (*In all cases, significant steps would have to be taken to avoid prejudicing the client's claims.*) *Hypothesis*: Bu addımlara müştərilərin həqiqi şəxsiyyətinin müstəntiqlərdən gizlədilməsi daxildir (*These steps include hiding the real identity of clients from investigators.*)	Neutral
Kazakh	STS	*Sent. 1*: Бір адам қазанға күріш слаып жатыр. (*A man pours rice into a pot.*) *Sent. 2*: Ер адам табаққа күріш салып жатыр. (*A man is putting rice in a bowling pot.*)	4.2
Kyrgyz	COPA	*Premise*: Кыз кодду жаттап калды. (*The girl memorized the code.*) *Choice 1* (Cause): Ал өзүнө өзү окуду. (*She recited it to herself.*) *Choice 2* (Cause): Ал муну жазууну унутуп калды. (*She forgot to write it down.*)	Choice 1
Uzbek	STS	*Sent. 1*: Okapi daraxtdan yemoqda. (*An okapi is eating from a tree.*) *Sent. 2*: Sichqon suv purkagichdan ichadi. (*A moose drinks from a sprinkler.*)	0.3
Uyghur	COPA	*Premise*: Дәрәх йопурмақлирини төкти. (*The tree shed its leaves.*) *Choice 1* (Effect): Йопурмақ рәңгигә боялди. (*The leaves turned colors.*) *Choice 2* (Effect): Йопурмақлар йәргә йиғилип қалди. (*The leaves accumulated on the ground.*)	Choice 2

Table 1: Examples from Kardeş-NLU one for each language and at least one for each task.

only for training in few-shot XLT.

Zero-Shot Transfer. We explore three zero-shot XLT setups: (i) monolingual training on English data, (ii) monolingual training on Turkish data, machine translated from the original English training data, and (iii) bilingual training on both English and machine-translated Turkish data, with joint bilingual batches.

Few-Shot Transfer. In few-shot fine-tuning, we additionally train on a small number of instances in the target language. We evaluate two different few-shot fine-tuning strategies: (1) in *sequential* transfer (Lauscher et al., 2020; Zhao et al., 2021), large(r)-scale fine-tuning on data from the source language(s)—in our case, English, Turkish, or bilingually English and Turkish—is followed by efficient target-language fine-tuning on the few shots; (2) in *joint* fine-tuning, we follow Schmidt et al. (2022) and, after initial source-only training, interleave source- and target-language instances at the batch level—the final batch loss is then the macro-average of the language-specific losses. Note that this results in joint trilingual fine-tuning when the source datasets are both English and Turkish.

	az	kk	ky	ug	uz
script	Latin	Cyrillic	Cyrillic	Arabic	Latin
monolingual corpus sizes (in bytes)					
CC-100	1.3G	889M	173M	46M	155M
Wiki	315M	354M	126M	36M	136M
Avg no. tokens in test instances (XLM-R tokenizer)					
NLI	44	46	47	79	52
COPA	22	24	24	34	26
STS	34	36	36	56	40

Table 2: Dataset statistics for Wikipedias and CC-100 portions of Kardeş-NLU languages along with average no. tokens in the test instances of Kardeş-NLU (as per XLM-R tokenizer)

4 Experimental Setup

Data. We carry out intermediate training for five Kardeş-NLU languages, monolingually (i.e., TLLM) or bilingually with Turkish (BALM and BAJM, see §3.1) using Wikipedias of the respective languages. Table 2 summarizes the base statistics of Wikipedias of Kardeş-NLU languages,[6] together with the size of their corresponding monolingual corpora in CC-100.[7] The sizes of the Turkish Wikipedia and Turkish CC-100 portions are 631MB and 5.4GB, respectively. Table 2 additionally shows the average number of tokens in test instances after XLM-R tokenization. Uyghur yields substantially more tokens than the other four languages. This is because most of Uyghur's pretraining corpus in XLM-R's is in the Arabic script, whereas Uyghur instances in Kardeş-NLU are written in Cyrillic.

In downstream XLT, we use the existing training data in English and respective automatic translations to Turkish. For NLI, we train on MNLI (Williams et al., 2018) and (automatically translated) Turkish training data from XNLI (Conneau et al., 2018). For STS, we train on the English training portions of STS-B (Cer et al., 2017) and its existing (automatic) translation to Turkish.[8] Due to

[6] The Wikipedia dumps were obtained from https://dumps.wikimedia.org/ on 10.12.2022. The text is extracted using the standard wikiextractor script.

[7] We report CC-100 portions, as XLM-R—the mmLM that we use in our experiments—was pretrained on it.

[8] https://huggingface.co/datasets/emrecan/stsb-mt-turkish

the small size of the English training data for COPA (400 instances) (Gordon et al., 2012), reported to hinder convergence of mmLM-based models (Sap et al., 2019; Ponti et al., 2020), we follow this prior work and first fine-tune on (English) SocialIQa (SIQA)—a closely related causal commonsense reasoning dataset (Sap et al., 2019) before fine-tuning on (English and/or Turkish) COPA data[9].

Intermediate Training Details. In all our main experiments, we use XLM-R (Base size) (Conneau et al., 2020a) as our mmLM. For the bilingual intermediate training procedure (e.g., BALM and BJLM), we train for a full epoch on Turkish Wikipedia: this results in multiple passes over the target language Wikipedias, given that those are substantially smaller. Thus, in the interest of fair evaluation, we train TLLM for multiple epochs: 2 for Azerbaijani and Kazakh, 5 for Kyrgyz and Uzbek, and 18 for Uyghur. We set the batch size to 32 and limit the sequence length to 128 tokens. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a fixed learning rate of $5e{-}5$.

Downstream Training Details. We adopt standard fine-tuning and add a task-specific classifier on top of the mmLM. Unless explicitly said otherwise, we perform full fine-tuning updating all parameters of the encoder together with the classifier's parameters. For NLI and STS, we encode the pair of sentences with the mmLM and feed the transformed representation of the [CLS] token to the classifier. For the multiple-choice tasks—COPA and SIQA (which we use as a "pre-fine-tuning" task to stabilize COPA training)—we face a varying number of answer choices per dataset (i.e., there are 3 possible answers in SIQA and 2 in COPA). We follow prior work Sap et al. 2019; Ponti et al. 2020 and encode the premise together with each answer choice. We feed the resulting output [CLS] token into a feed-forward regressor that produces a single score for each answer choice. Afterwards, the individual scores of all choices are concatenated and fed to the softmax classifier.

We train the models for 10 epochs with mixed precision using AdamW (Loshchilov and Hutter, 2019) with a weight decay of 0.05 and the initial learning rate set to $2e{-}5$. We use a linear scheduler with 10% linear warm-up and decay. We deviate from this configuration (i) in the *joint* few-shot fine-tuning, where we train for 50 epochs without a scheduler, following recommendations of (Schmidt et al., 2022), and (ii) for all NLI experiments, where we train for 5 epochs due to the size of the MNLI training data (ca. 400K instances). The sequence length is limited to 128 tokens for all tasks, matching the input size of the intermediate MLM-ing. We fine-tune with a batch size of 32, except in the trilingual *joint* few-shot fine-tuning (English-Turkish-target language), where we sample 10 instances per language (i.e., batch size 30). For each experiment, we execute three runs with different random seeds and report the average performance (accuracy for NLI and COPA and Pearson correlation for STS). In zero-shot XLT, we report the performance of the last checkpoint obtained at the end of the training. In few-shot XLT, we start training from the last snapshot of the source training (English, Turkish, or English and Turkish) and select the last snapshot of the second—*sequential* or *joint*—training step.

5 Results and Discussion

Zero-Shot Transfer. Table 3 displays the zero-shot XLT performance for all five Kardeş-NLU languages on NLI, COPA and STS. Generally, we reach the best performance when Turkish is integrated into both intermediate training (rows BALM and BAJM) *and* as the source language in fine-tuning (columns TR and EN,TR). On average, across all five languages, BJLM combined with source fine-tuning on concatenated English and Turkish instances (EN,TR) yields a 6.6% and 2.1% boost over zero-shot XLT from English only with the vanilla XLM-R (Base) on NLI and COPA, respectively. On these two tasks, this observation holds for all individual languages except Kazakh. The gains over the vanilla zero-shot XLT for STS, however, are much smaller, with only BALM combined with English and Turkish fine-tuning surpassing the default zero-shot XLT performance of XLM-R (Base, EN) and that by a narrower margin (+0.6). We speculate that this is because (i) fine-grained sentence similarity is more sensitive to slight semantic misalignment and (ii) while our bilingual intermediate training improves the semantic links between Turkish and the target language, it is not of an adequate scale to establish alignments of such semantic precision.

Including Turkish as a fine-tuning source language (TR and EN,TR) brings consistent gains

[9] We translate the COPA training set to Turkish with GT.

over transfer from English only, regardless of the intermediate training strategy. The best results are almost always obtained when we fine-tune on both English and Turkish (EN,TR): we hypothesize that such fine-tuning establishes task-specific representational associations between the two languages and allows the transfer to benefit from both (i) XLM-R's unmatched representational quality for English and (ii) proximity of Turkish to the target languages. The effect is then further amplified when intermediate training (BALM and BJLM) increases the XLM-R's capacity for Turkish and the target language and strengthens the alignments between them. This is confirmed by the fact that intermediate training on the target language alone (TLLM) brings downstream gains (compared to Base) for NLI but not for the other two tasks.

Looking at individual languages, we observe the least (and smallest) gains for Azerbaijani and Kazakh, the two most-resourced Kardeş-NLU languages, and the most (and largest) gains for the three less-resourced languages: Uyghur, Uzbek, and Kyrgyz (e.g., compared to Base transfer from EN on NLI, BJLM with transfer from EN,TR leads to gains of 5.0% for Kyrgyz, 5.1% for Uzbek, and 17.2% for Uyghur). We see the largest gains (by a wide margin) for Uyghur, despite the script mismatch between the intermediate training (Arabic script) and evaluation (Uyghur in Cyrillic script). The intermediate bilingual training for Uyghur, which improves representations of Arabic-script tokens, would thus likely yield even larger gains if the Uyghur test instances were in the Arabic script.

Few-Shot Transfer. Table 4 summarizes the few-shot XLT results. We observe mixed results compared to the strongest zero-shot approaches: while there is a small improvement on STS (+1.0%), we see virtually no gains for COPA (+0.1%) and NLI (-0.3%). Consistent with zero-shot XLT findings, few-shot XLT yields best results when we start the few-shot target language training from models trained on both English and Turkish (EN,TR). Additionally, we observe that few-shot XLT with models that were intermediately trained on Turkish and the target languages (BALM, BAJM) yields stronger performance than with those MLM-ed on the target language alone (TLLM). Nonetheless, there is no bilingual intermediate training strategy that is consistently best: BJLM yields better scores on COPA, whereas BALM reaches better STS performance; on NLI, both strategies perform comparably. Concerning the number of target language shots, we observe that we typically need at least 50 shots to match or surpass the zero-shot XLT performance. Comparing few-shot transfer procedures, we observe task-dependent variability. On NLI, sequential fine-tuning substantially outperforms the joint approach. Conversely, on COPA and STS, joint few-shot transfer shows better performance, with a more pronounced gap on STS.

Kardeş-NLU: A Difficult Few-Shot XLT Benchmark. Not only does the comparison of zero-shot and few-shot results in Table 4 render Kardeş-NLU as a difficult few-shot XLT benchmark but also does Kardeş-NLU involve two tasks—STS and COPA—that are underrepresented in the current body of work on (few-shot) XLT (Lauscher et al., 2020; Zhao et al., 2021; Schmidt et al., 2022). This makes Kardeş-NLU a valuable evaluation resource for XLT research.

Instruction-Based LLMs on Kardeş-NLU. Given the recent popularity of instruction-tuned LLMs as competent "generalizers" (Ouyang et al., 2022; Ahuja et al., 2023), we additionally evaluate (zero-shot) two state-of-the-art multilingual LLMs on Kardeş-NLU:[10] mT0 (Muennighoff et al., 2023), as the open model tuned on instructions derived from NLP tasks, and ChatGPT, as the commercial model tuned from human instructions and feedback. To this end, we slightly modify the instructions and prompts proposed by Ahuja et al. (2023): we provide further details in the Appendix §A.

Figure 1 compares the best zero-shot XLT performance (based on XLM-R) for each language from Table 3 against zero-shot inference with mT0 and ChatGPT. The NLI results, in which both LLMs dramatically underperform our language-adapted zero-shot XLT (-23.9% and -15.1% for ChatGPT and mT0, respectively), diametrically oppose those on COPA, where both LLMs (and especially mT0) excel and surpass our best zero-shot XLT (the gap is full 10% in favor of mT0, albeit only 1.1% for ChatGPT). We believe that this is because mT0 was instruction-tuned, multilingually, on a large number of different multi-choice QA datasets (including, e.g., SIQA). ChatGPT, in contrast, being fine-tuned based on open-ended instruction-reply

[10] Regression (i.e., score prediction) tasks are inherently difficult to cast as text generation tasks; we thus omit STS from this evaluation.

		Azerbaijani			Kazakh			Kyrgyz			Uyghur			Uzbek			Average		
		EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR
NLI	Base	76.5	**80.1**	79.6	73.8	76.3	**77.3**	70.4	73.9	74.1	42.2	44.4	42.9	70.7	72.0	71.8	66.7	69.4	69.1
	TLLM	77.3	79.0	79.2	75.3	76.3	76.8	72.4	74.1	74.4	56.7	57.1	56.9	73.1	74.3	74.8	71.0	72.2	72.4
	BALM	77.3	78.8	79.3	74.4	75.3	77.0	71.6	73.4	74.0	57.4	58.7	58.0	73.1	74.5	75.0	70.8	72.1	72.7
	BJLM	76.4	78.4	79.3	74.9	75.1	76.8	71.9	74.3	**75.5**	57.2	59.2	**59.4**	73.4	74.6	**75.7**	70.7	72.3	**73.3**
COPA	Base	60.1	61.1	60.9	60.7	**60.8**	59.9	59.7	60.0	59.4	51.8	52.7	52.7	57.3	59.5	60.1	57.9	58.8	58.6
	TLLM	62.1	62.1	61.5	55.7	55.8	56.1	57.5	59.7	58.9	49.9	50.3	49.3	62.9	**63.2**	62.5	57.6	58.2	57.7
	BALM	57.2	58.3	59.4	59.1	59.5	59.7	56.1	59.9	59.1	51.1	**53.9**	52.5	60.5	61.7	61.9	56.8	58.6	58.5
	BJLM	61.8	**63.3**	**63.3**	58.4	58.6	57.7	56.8	61.5	**62.0**	50.9	52.2	**53.9**	61.7	60.5	62.9	57.9	59.2	**60.0**
STS	Base	80.3	78.9	**80.4**	**85.8**	84.1	84.8	78.2	77.9	**78.7**	69.2	64.8	64.2	78.3	77.2	77.1	78.4	76.6	77.1
	TLLM	75.8	75.5	78.1	80.6	80.1	81.9	71.3	71.8	74.2	70.6	69.3	71.3	70.6	67.0	76.9	73.8	72.7	76.5
	BALM	72.7	78.7	79.7	81.4	83.2	83.9	71.1	77.3	78.3	72.8	72.3	**73.5**	72.5	77.6	**79.3**	74.1	77.8	**79.0**
	BJLM	69.3	77.0	78.3	78.6	83.2	84.6	69.9	75.1	77.3	65.7	66.9	69.0	71.1	76.8	77.3	70.9	75.8	77.3

Table 3: Zero-Shot XLT results on Kardeş-NLU for three intermediate LM-ing strategies (TLLM, BALM, and BJLM) and source fine-tuning datasets (English only, Turkish only, and English and Turkish combined). The best results for each language-task pair are shown in **bold**. The evaluation metrics are accuracy (%) for NLI and COPA, and Pearson correlation for STS.

		Zero-Shot			Few-Shot																	
					Sequential									Joint								
		EN	TR	EN,TR	EN			TR			EN,TR			EN			TR			EN,TR		
Shots		-	-	-	10	50	100	10	50	100	10	50	100	10	50	100	10	50	100	10	50	100
NLI	Base	66.7	69.4	69.1	63.5	67.9	68.1	65.7	69.0	69.3	66.0	69.5	70.1	65.0	66.2	66.4	67.0	67.4	67.5	66.7	68.0	69.0
	TLLM	71.0	72.2	72.4	68.1	70.7	71.7	69.3	71.9	72.3	70.6	72.6	72.5	69.3	70.3	70.7	70.1	71.3	70.7	70.4	71.2	71.9
	BALM	70.8	72.1	72.7	67.9	70.9	71.2	69.0	71.8	72.0	70.0	72.6	73.0	69.1	70.0	70.4	70.5	71.5	71.3	70.5	71.0	71.6
	BJLM	70.7	72.3	**73.3**	67.5	71.0	71.3	69.2	71.7	71.5	69.9	72.7	73.0	69.4	70.3	69.9	70.7	71.3	71.2	70.6	71.5	71.8
COPA	Base	57.9	58.8	58.6	56.4	57.9	58.8	56.8	57.6	58.2	57.0	57.8	58.3	57.6	57.9	59.0	58.7	58.5	58.5	59.0	59.0	59.5
	TLLM	57.6	58.2	57.7	56.8	57.4	58.4	57.1	57.9	59.5	56.7	58.0	58.9	57.2	57.5	58.3	58.1	58.7	58.6	58.6	59.0	59.8
	BALM	56.8	58.6	58.5	56.6	57.2	58.1	56.8	58.0	58.5	57.6	58.0	58.4	56.8	57.8	57.2	59.0	58.7	58.2	59.1	59.4	58.3
	BJLM	57.9	59.2	**60.0**	57.2	58.6	59.3	58.0	59.3	59.7	58.0	59.8	59.8	58.1	58.8	58.8	58.9	59.9	59.3	60.1	59.9	59.8
STS	Base	78.4	76.6	77.1	73.5	75.5	75.4	74.5	76.5	75.7	75.4	77.1	77.1	76.3	77.6	77.6	77.0	78.9	78.9	77.1	79.0	79.3
	TLLM	73.8	72.7	76.5	73.6	75.3	75.6	74.9	76.1	76.2	76.4	77.3	77.6	75.1	76.8	76.9	75.2	77.0	77.6	77.2	78.5	78.8
	BALM	74.1	77.8	**79.0**	74.5	76.0	76.3	76.2	77.6	77.8	77.3	78.6	78.4	77.1	77.2	76.9	78.3	79.4	79.6	79.4	80.0	80.0
	BJLM	70.9	75.8	77.3	72.8	74.9	75.4	75.2	76.9	76.8	76.1	77.7	78.1	74.0	76.2	76.6	76.8	78.3	78.5	77.9	79.3	79.4

Table 4: Results of *sequential* and *joint* few-shot XLT on Kardeş-NLU: performance with 10, 50, and 100 target-language shots. The best zero-shot result per task is shown in **bold**, the best few-shot result is underlined. The evaluation metrics are accuracy (%) for NLI and COPA, and Pearson correlation for STS.

pairs, has a weaker inductive bias for both COPA and NLI. The two LLMs yield the best performance on both tasks for Azerbaijani, the most resourced language in Kardeş-NLU—the performance drops for the remaining languages are drastic, especially for ChatGPT. This is in line with findings from concurrent work (Ahuja et al., 2023; Asai et al., 2023) and shows that even the largest instruction-tuned LLMs are bound by the language distribution of their (pre)training data, indicating that there is still a long way to go to enable truly multilingual NLP.

6 Related Work

Multilingual Evaluation Benchmarks. Reliable evaluation of the multilingual abilities of mmLMs requires that they are tested against a large set of diverse languages (Joshi et al., 2020). On the one hand, multilingual benchmarks that encompass many tasks, such as XGLUE (Liang et al., 2020) and XTREME (Hu et al., 2020; Ruder et al., 2021), comprise diverse but predominantly highly or moderately resourced languages: their coverage of LR languages is small and varies across tasks. On the other hand, many recent efforts introduce dedicated benchmarks for specific families of LR languages (Armstrong et al., 2022; Adelani et al., 2022; Ebrahimi et al., 2022; Winata et al., 2023, *inter alia*). While these target truly underrepresented languages, they typically focus on a single task only, e.g., NLI or NER. With Kardeş-NLU we, (i) cover multiple languages from an underrepresented language family while (ii) including various tasks (NLI, COPA, and STS) that require different

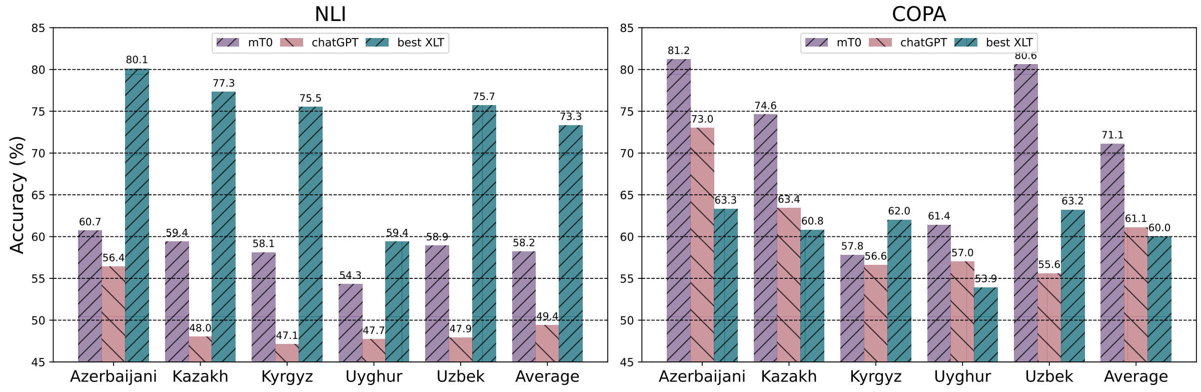

Figure 1: Performance of mT0-XXL, chatGPT, and our best performing zero-shot XLT strategy on NLI and COPA.

degrees of precision in language understanding.

Cross-Lingual Transfer with mmLMs. mmLMs still play an important role in multilingual NLU and XLT, exhibiting good performance in zero-shot XLT (Wu and Dredze, 2019; Hu et al., 2020) to HR languages. They, however, perform much worse in XLT to LR languages distant from English (as the common source). The body of work on improving XLT is threefold. The first line of work seeks to improve XLT via post-hoc alignment of representational subspaces of individual languages, guided by parallel data (Cao et al., 2020; Conneau et al., 2020b; Hu et al., 2021; Wang et al., 2022; Minixhofer et al., 2022, *inter alia*) and driven by cross-lingual supervision. These efforts, however, offer little gain for LR languages, whose representational subspaces are of low semantic quality, to begin with. The second line of work seeks to improve the representational quality for LR languages through additional language modeling training (Pfeiffer et al., 2020; Ansell et al., 2021; Parović et al., 2022; Pfeiffer et al., 2022), resulting in moderate downstream performance gains. Finally, the third line of work (Lauscher et al., 2020; Zhao et al., 2021; Xu and Murray, 2022; Schmidt et al., 2022, 2023a,b) focuses on the actual downstream transfer, rather than the task-agnostic adaptation of mmLMs, investigating how to best utilize the limited number of annotated task-specific target-language instances (Lauscher et al., 2020; Schmidt et al., 2022, 2023a) or tailor source-language instances to resemble target language ones (Xu and Murray, 2022).

In this work, we adopt the latter two ideas and seek to improve XLT to Turkic LR languages via both intermediate LM-ing and few-shot XLT: unlike most existing work, however, we seek to leverage a close HR language (Turkish) to facilitate the transfer. The work of Snæbjarnarson et al. (2023) is conceptually most similar; they, however, target a single LR language (Faroese) from a HR family (Germanic branch of the Indo-European family) with many HR relatives (Scandinavian languages).

The three mentioned lines of work typically propose methods to improve XLT starting from a single, given source language (usually EN). Complementary to these lines of work, the work of Lin et al. (2019) and Glavaš and Vulić (2021) instead focus on identifying the best source languages to transfer from for a given target language. Their work considers linguistic and dataset related factors beyond the sole language family. Their findings are complementary to our work, suggesting that even for LR languages that do not have a closely related HR language within their family, it might still be possible to infer such a closely related HR language from another language family.

7 Conclusion

In this work, we contribute to the body of evaluation resources for low-resource (LR) cross-lingual transfer (XLT) by introducing Kardeş-NLU, an evaluation benchmark covering three NLU tasks (NLI, STS, and COPA)—for five Turkic languages: Azerbaijani, Kazakh, Kyrgyz, Uyghur, and Uzbek. Kardeş-NLU allows investigation of an understudied XLT approach: leveraging a high-resource (HR) language to improve transfer to linguistically and genealogically related LR languages. We extend existing intermediate training and fine-tuning approaches for improving LR XLT to integrate Turkish as the HR "sibling" of the Kardeş-NLU languages. Through comprehensive experimentation and analysis, we demonstrated that adding Turkish in task-specific fine-tuning can provide significant XLT gains for Kardeş-NLU languages that are further amplified by incorporating Turkish in bilingual intermediate training strategies. What is more, we also find that Kardeş-NLU is a difficult benchmark for few-shot XLT, observing that established few-shot transfer methods are not effective. Finally, we evaluated two cutting-edge instruction-tuned large language models—mT0 and chatGPT—on Kardeş-NLU, showing that their (zero-shot) performance is inferior on lower-resourced Kardeş-NLU languages (Uyghur, Uzbek, Kyrgyz) and greatly varies across tasks. This proves that there is still a long way to (truly) multilingual NLP. In our subsequent efforts, we will not only seek to extend Kardeş-NLU with additional LR Turkic languages, but also explore how to leverage HR siblings in LR XLT for other language families.

8 Limitations

We strove for both a representative NLU benchmark for Turkic languages and a comprehensive study of XLT to LR target languages with the help of a closely related HR language. Nonetheless, our work is limited in several aspects. Out of 23 live Turkic languages, Kardeş-NLU covers only five. Two main factors determined the set of initially included languages: a limited annotation budget and the ability to find native speakers. The latter is why we ended up with languages that are among the largest Turkic languages in terms of number of native speakers (Kyrgyz, as the smallest, has ca. 5M native speakers). Further, there is a mismatch between the more common Arabic script used for Uyghur and the Cyrillic script we use for it in Kardeş-NLU because our Uyghur annotator was unfamiliar with the Arabic script.

The Kardeş-NLU benchmark is obtained through automatic translations from the existing English test sets to the target languages. This is followed by manual annotation and curation through native speakers to ensure high quality. In order to have suitable translations for culture specific concepts, we instructed our annotators to pay special attention to the idiomaticity of the English sentences during the editing. Despite our best efforts, the resulting datasets might not perfectly reflect the cultural and social elements of the target low-resource languages since their content is tied to original English datasets.

Next, we employed Wikipedias as corpora for our intermediate pretraining. Albeit curated, Wikipedia content is subject to biased, missing or simply incorrect information that can lead to undesired behavior in the resulting models.

Concerning the methodology, we limited our study exclusively to mainstream approaches: (i) intermediate LM-ing for improving the representational quality of mmLMs for a language of interest and (ii) established protocols for downstream zero-shot and few-shot XLT. We acknowledge the existence of more sophisticated (and more recent) XLT methods based, e.g., on gradient manipulation (Wang and Tsvetkov, 2021; Xu and Murray, 2022) or dedicated representational alignment of lexical units (i.e., embedding spaces) (Minixhofer et al., 2022). We hope the research community will use Kardeş-NLU to evaluate and profile existing and future state-of-the-art XLT approaches.

Finally, for the prompt-based evaluation of LLMs, we experiment only with a single instruction (i.e., prompt) adapted from Ahuja et al. (2023). It is reasonable to expect that some prompt engineering effort yields better results.

Acknowledgements

This work was partially supported by DFG (grant SCHU 2246/14-1).

References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. IndicXNLI: Evaluating multilingual inference for Indian languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruth-Ann Armstrong, John Hewitt, and Christopher Manning. 2022. JamPatoisNLI: A jamaican patois natural language inference dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5307–5320, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. 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, 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.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.

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

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2021. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.

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 generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Chia-Chien Hung, Anne Lauscher, Simone Ponzetto, and Goran Glavaš. 2022. DS-TOD: Efficient domain specialization for task-oriented dialog. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 891–904, Dublin, Ireland. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

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. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

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)*, pages 7654–7673, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2023a. Free lunch: Robust cross-lingual transfer via model checkpoint averaging. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5712–5730, Toronto, Canada. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2023b. One for all & all for one: Bypassing hyperparameter tuning with model averaging for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12186–12193.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Zirui Wang and Yulia Tsvetkov. 2021. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Haoran Xu and Kenton Murray. 2022. Por qué não utiliser alla språk? mixed training with gradient optimization in few-shot cross-lingual transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2043–2059, Seattle, United States. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.

A LLMs: mT0 and ChatGPT

For mT0, we only use the instance-based prompts, without the task instruction, following Ahuja et al. (2023) (and accept exact matches as correct answers only):

NLI. {PREMISE} *Question:* {HYPOTHESIS} *True, False, or Neither?*

COPA. {PREMISE} {% if question == "cause" %} *This happened because...* {% else %} *As a consequence...* {% endif %} *Help me pick the more plausible option:* -{CHOICE1}-{CHOICE2}

For ChatGPT, we slightly modify the prompts from Ahuja et al. (2023) due to the fact that they perform in-context few-shot learning, whereas we carry out zero-shot prediction:

NLI. *You are an NLP assistant whose purpose is to solve Natural Language Inference (NLI) problems. NLI is the task of determining the inference relation between two (short, ordered) texts. For the given two sentences, you need to predict one of the following: 1. Entailment, 2. Contradiction, or 3. Neither (Neutral). Sentence 1:* {PREMISE}. *Sentence 2:* {HYPOTHESIS}. *Answer:*

COPA. *You are an AI assistant whose purpose is to perform open-domain commonsense causal reasoning. You will be provided a premise and two alternatives, where the task is to select the alternative that more plausibly has a causal relation with the premise. Answer as concisely as possible.* PREMISE {% if question == "cause" %} *This happened because...* {% else %} *As a consequence...* {% endif %}: *Alternative 1:* CHOICE1 *Alternative 2:* CHOICE2

For NLI, the model's output is compared directly against the target label (*True*, *False*, or *Neither*). For COPA, it is compared against the correct alternative ({CHOICE1} or {CHOICE2}). Since the models are free to generate any text, they can theoretically perform below the random baseline (33% for NLI and 50% for COPA).

Table 5 displays per language and average results for zero-shot evaluations on NLI and COPA for the XLM-R base versions that we experiment with, mT0 of various sizes, and ChatGPT. We also experiment with the templates that are translated to the target language using Google Translate. However, those versions overall performed worse than the English versions, most likely because of the low translation quality. We can see that mT0's performance on COPA improves drastically when it is scaled to XL and XXL versions. It should be noted that mT0's instruction tuning dataset includes the Social IQA dataset, which is similar to the COPA dataset. This might explain the larger model's strong performance on this dataset outperforms zero-shot XLM-R variants.

B Computational Resources

All the experiments were run on a single V100 with 32GB VRAM. We roughly estimate that total GPU time accumulates to 2800 hours across all experiments.

C Adapter Fine-Tuning Experiments

In preliminary experiments, we investigated the adapter-based equivalents to TLLM and BALM (on STS and NLI) (Pfeiffer et al., 2020; Parović et al., 2022). We report per-language and averaged scores in Table 6. Full fine-tuning of the mmLM outperformed the adapter-based tuning, especially on lower-resourced languages.

Target Language LM-ing Adapters (TLLM-AD). We first train monolingual language adapters on target languages via MLM-ing. We then stack a task adapter on top and fine-tune it on the corresponding downstream data—English, Turkish or English and Turkish jointly—while keeping the language adapter frozen.

Bilingual Alternating LM-ing Adapters (BALM-AD). Here, we stick to Parović et al. 2022 and update the language adapter´s parameters alternately by one batch on the target language data followed by one batch on Turkish data. Afterwards, we fine-tune task adapters on either English, Turkish or English and Turkish jointly, while keeping the language adapter frozen.

Adapter Training Details. We trained monolingual language adapters for 25000 steps and bilingual ones for 50000. We set the learning rate to $1e{-}4$ and the batch size to 64. For task adapters, we applied the same hyperparameters used for our full fine-tuning experiments explained in section 4 but lowered the learning rate to $1e{-}4$, as suggested by Pfeiffer et al. 2020.

		Azerbaijani			Kazakh			Kyrgyz			Uyghur			Uzbek			Average		
		EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR
NLI	Base	76.5	**80.1**	79.6	73.8	76.3	**77.3**	70.4	73.9	74.1	42.2	44.4	42.9	70.7	72.0	71.8	66.7	69.4	69.1
	TLM	77.3	79.0	79.2	75.3	76.3	76.8	72.4	74.1	74.4	56.7	57.1	56.9	73.1	74.3	74.8	71.0	72.2	72.4
	BALM	77.3	78.8	79.3	74.4	75.3	77.0	71.6	73.4	74.0	57.4	58.7	58.0	73.1	74.5	75.0	70.8	72.1	72.7
	BJLM	76.4	78.4	79.3	74.9	75.1	76.8	71.9	74.3	**75.5**	57.2	59.2	**59.4**	73.4	74.6	**75.7**	70.7	72.3	**73.3**
	mT0$_{small}$	35.3			34.9			36.8			36.6			35.3			35.8		
	mT0$_{base}$	40.5			40.3			39.8			38.3			40.4			39.8		
	mT0$_{large}$	40.8			42.5			42.0			41.9			41.2			41.7		
	mT0$_{XL}$	56.9			55.7			53.0			49.4			55.6			54.1		
	mT0$_{XXL}$	60.7			59.4			58.1			54.3			58.9			58.2		
	chatGPT	56.4			48.0			47.1			47.7			47.9			49.4		
COPA	Base	60.1	61.1	60.9	60.7	60.8	59.9	59.7	60.0	59.4	51.8	52.7	52.7	57.3	59.5	60.1	57.9	58.8	58.6
	TLM	62.1	62.1	61.5	55.7	55.8	56.1	57.5	59.7	58.9	49.9	50.3	49.3	62.9	63.2	62.5	57.6	58.2	57.7
	BALM	57.2	58.3	59.4	59.1	59.5	59.7	56.1	59.9	59.1	51.1	53.9	52.5	60.5	61.7	61.9	56.8	58.6	57.9
	BJLM	61.8	63.3	63.3	58.4	58.6	57.7	56.8	61.5	**62.0**	50.9	52.2	53.9	61.7	60.5	62.9	57.9	59.2	60.0
	mT0$_{small}$	34.2			7.6			3.4			5.6			43.6			18.8		
	mT0$_{base}$	32.0			3.6			5.8			4.2			39.8			17.1		
	mT0$_{large}$	38.0			38.2			30.4			24.2			38.4			33.8		
	mT0$_{XL}$	60.4			62.8			50.4			47.6			63.2			56.9		
	mT0$_{XXL}$	**81.2**			**74.6**			57.8			**61.4**			**80.6**			**71.1**		
	chatGPT	73.0			63.4			56.6			57.0			55.6			61.1		

Table 5: Zero-Shot results for the target languages and the average results across the five languages for XLM-R base, mT0 and chatGPT models. The best results for each language-task pair are shown in **bold**.

		Azerbaijani			Kazakh			Kyrgyz			Uyghur			Uzbek			Average		
		EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR	EN	TR	EN,TR
NLI	TLLM	77.3	79.0	79.2	75.3	76.3	76.8	72.4	74.1	74.4	56.7	57.1	56.9	73.1	74.3	74.8	71.0	72.2	72.4
	BALM	77.3	78.8	79.3	74.4	75.3	77.0	71.6	73.4	74.0	57.4	58.7	**58.0**	73.1	74.5	**75.0**	70.8	72.1	**72.7**
	TLLM-AD	77.1	78.2	**80.3**	74.0	74.8	76.8	70.1	72.7	74.5	48.3	47.0	48.3	71.1	71.1	73.4	68.1	68.8	70.6
	BALM-AD	77.9	78.0	80.1	73.3	75.2	**77.6**	70.7	73.2	**74.7**	47.8	46.4	46.8	70.5	71.8	73.1	68.1	69.0	70.5
STS	TLLM	75.8	75.5	78.1	80.6	80.1	81.9	71.3	71.8	74.2	70.6	69.3	71.3	70.6	67.0	76.9	73.8	72.7	76.5
	BALM	72.7	78.7	79.7	81.4	83.2	83.9	71.1	77.3	**78.3**	72.8	72.3	**73.5**	72.5	77.6	**79.3**	74.1	77.8	**79.0**
	TLLM-AD	76.1	77.5	79.5	82.0	81.4	**84.3**	74.0	75.4	77.8	69.7	68.4	70.5	75.2	75.5	77.4	75.4	75.6	77.9
	BALM-AD	76.2	77.5	**79.9**	82.3	81.6	84.1	73.2	75.5	77.3	68.2	67.3	70.0	75.1	75.0	77.3	75.1	75.4	77.7

Table 6: Zero-Shot XLT results on Kardeş-NLU (NLI and STS) for two adapter strategies (TLLM-AD and BALM-AD) and source fine-tuning datasets (English only, Turkish only, and English and Turkish combined). The best results for each language-task pair are shown in **bold**.

D Few-Shot Results

		Zero-Shot			Few-Shot																	
					Sequential									Joint								
		EN	TR	EN,TR	EN			TR			EN,TR			EN			TR			EN,TR		
Shots		-	-	-	10	50	100	10	50	100	10	50	100	10	50	100	10	50	100	10	50	100
Azerbaijani	Base	76.5	80.1	79.6	73.3	76.6	76.3	74.9	78.5	77.9	75.2	78.8	79.0	75.0	74.7	74.1	77.7	76.9	76.8	76.7	77.1	77.3
	TLM	77.3	79.0	79.2	75.7	77.7	77.8	75.7	78.7	79.3	76.9	79.1	78.9	76.4	77.0	76.7	77.8	77.7	77.2	78.0	78.3	78.2
	BALM	77.3	79.0	79.2	75.4	77.2	77.3	76.5	78.1	78.1	76.7	78.9	79.2	74.8	76.0	76.3	78.0	78.4	78.1	77.6	77.5	78.0
	BJLM	77.3	78.8	79.3	72.3	77.5	77.3	75.8	78.7	78.3	77.3	79.1	79.2	76.6	76.9	75.7	77.8	78.2	77.3	78.3	78.4	77.7
Kazakh	Base	73.8	76.3	77.3	69.7	73.6	73.5	72.0	75.0	75.3	73.3	75.5	76.0	71.1	71.5	71.4	74.3	73.0	72.7	74.6	74.4	74.3
	TLM	75.3	76.3	76.8	72.4	75.5	76.3	75.1	75.9	75.7	74.8	76.8	76.1	73.8	75.2	74.8	75.2	75.6	74.6	76.0	75.8	76.4
	BALM	74.4	75.3	77.0	72.8	75.3	74.7	72.9	75.8	75.7	75.1	76.4	76.9	73.8	73.8	74.5	74.6	74.8	74.2	74.9	74.7	75.8
	BJLM	74.9	75.1	76.8	73.2	74.8	75.0	73.0	74.5	74.6	74.5	76.8	76.4	73.3	74.1	73.6	74.1	75.0	74.3	75.2	75.2	74.7
Kyrgyz	Base	70.4	73.9	74.1	66.6	70.6	70.5	69.4	72.3	72.7	70.3	73.1	73.6	68.9	69.7	69.2	70.7	69.4	69.5	70.8	70.5	71.7
	TLM	72.4	74.1	74.4	71.0	73.6	73.1	72.2	73.6	74.0	72.9	75.4	75.4	71.4	71.6	71.9	72.4	73.4	72.6	72.8	73.0	73.2
	BALM	71.6	73.4	74.0	69.2	73.2	72.6	71.2	73.4	73.0	73.0	74.5	74.7	71.0	71.4	71.8	71.7	72.3	71.9	73.0	73.2	73.0
	BJLM	71.9	74.3	75.5	71.7	73.1	73.3	72.9	74.0	73.5	73.7	75.8	75.7	72.0	72.8	72.0	73.4	72.8	73.6	72.6	73.6	73.8
Uyghur	Base	42.2	44.4	42.9	41.5	49.2	50.1	45.0	47.9	50.5	43.5	48.6	49.6	43.2	47.8	49.9	43.8	48.4	49.8	42.2	47.9	48.3
	TLM	56.7	57.1	56.9	50.1	53.7	58.0	52.1	57.3	58.8	55.3	56.8	57.9	52.6	54.6	56.6	52.9	56.5	56.2	52.4	55.7	58.1
	BALM	57.4	58.7	58.0	51.4	57.0	58.3	53.0	58.0	59.5	51.9	58.3	59.4	53.7	56.3	55.8	54.9	57.9	58.9	54.0	56.4	57.4
	BJLM	57.2	59.2	59.4	51.1	56.4	57.8	52.8	57.3	57.3	51.6	57.0	58.8	52.8	54.4	55.9	54.5	56.4	57.1	54.0	56.1	57.9
Uzbek	Base	70.7	72.0	71.8	66.5	69.5	69.8	67.1	71.6	70.2	67.6	71.3	72.3	66.5	67.5	67.4	68.6	69.0	68.6	67.9	68.6	69.0
	TLM	73.1	74.3	74.8	71.3	73.3	73.4	71.3	74.1	73.9	73.1	74.9	74.4	72.4	73.1	73.3	72.4	73.2	72.9	72.7	73.2	73.5
	BALM	73.1	74.5	75.0	70.9	71.6	73.4	71.4	73.9	73.8	73.3	74.7	75.1	72.1	72.4	73.5	73.4	73.9	73.2	73.1	73.2	73.7
	BJLM	73.4	74.6	75.7	69.3	73.1	73.3	71.4	74.0	74.0	72.2	74.8	75.0	72.4	73.4	72.3	73.4	74.1	73.7	73.1	74.0	75.1

Table 7: Per-language results of *sequential* and *joint* transfer on Kardeş-NLI.

		Zero-Shot			Few-Shot																	
					Squential									Joint								
		EN	TR	EN,TR	EN			TR			EN,TR			EN			TR			EN,TR		
Shots		-	-	-	10	50	100	10	50	100	10	50	100	10	50	100	10	50	100	10	50	100
Azerbaijani	Base	60.1	61.1	60.9	62.3	62.5	63.8	61.5	61.3	62.5	61.9	62.3	62.5	60.3	62.2	61.9	62.3	62.8	62.7	61.7	62.8	62.9
	TLM	62.1	62.1	61.5	60.1	60.7	60.6	60.3	60.3	62.1	59.9	60.8	61.1	60.8	61.2	62.1	62.3	60.8	60.6	61.6	61.7	62.6
	BALM	57.2	58.3	59.4	58.5	58.3	59.2	58.8	58.0	59.2	60.1	58.7	59.8	59.5	59.8	57.7	58.9	59.3	59.1	62.7	60.6	59.3
	BJLM	61.8	63.3	63.3	61.1	62.4	62.1	62.5	61.9	62.9	61.0	62.1	61.7	62.0	62.8	61.9	62.1	63.7	61.9	61.9	62.3	62.4
Kazakh	Base	60.7	60.8	59.9	55.6	59.3	60.1	57.6	60.7	60.3	56.7	60.4	60.3	58.7	59.2	60.8	60.2	60.7	60.9	60.7	60.8	61.9
	TLM	55.7	55.8	56.1	54.4	56.1	57.2	54.8	55.5	57.9	54.9	56.5	57.9	55.4	56.4	56.5	56.3	57.6	58.4	56.6	58.3	59.5
	BALM	59.1	59.5	59.7	58.6	59.4	60.3	55.9	59.5	59.5	57.1	58.7	59.9	57.5	57.9	60.3	60.0	59.3	59.8	59.9	60.7	59.3
	BJLM	58.4	58.6	57.7	56.0	57.9	60.1	58.3	58.9	60.5	58.3	59.5	60.5	57.5	59.8	58.9	58.5	59.5	59.2	59.6	59.8	59.7
Kyrgyz	Base	59.7	60.0	59.4	56.6	59.0	59.7	58.0	58.5	59.0	59.3	59.3	59.7	60.1	60.1	61.1	61.1	60.5	60.2	61.3	61.1	61.1
	TLM	57.5	59.7	58.9	58.5	58.9	61.2	59.7	60.9	61.9	58.7	60.0	60.2	58.7	58.2	59.7	60.1	60.6	59.5	61.3	61.5	61.7
	BALM	56.1	59.9	59.1	57.6	58.1	58.3	58.1	61.7	60.7	57.6	59.8	60.3	56.1	58.1	57.7	60.7	61.7	60.1	58.5	60.9	58.9
	BJLM	56.8	61.5	62.0	57.3	59.5	60.8	60.5	63.1	61.3	60.1	62.4	62.1	59.5	59.3	60.1	61.3	61.9	62.3	62.2	62.9	60.9
Uyghur	Base	51.8	52.7	52.7	51.7	50.7	52.5	51.3	50.3	51.9	50.7	51.3	51.7	51.3	50.9	52.4	51.1	50.5	50.1	51.5	50.6	51.7
	TLM	49.9	50.3	49.3	50.9	48.1	50.5	48.6	49.1	52.7	48.7	49.7	51.1	49.2	49.9	50.2	49.9	49.9	50.4	49.5	49.8	52.3
	BALM	51.1	53.9	52.5	51.1	49.4	50.7	53.3	51.2	51.7	52.9	51.2	50.7	50.8	50.9	49.6	54.2	52.5	51.5	52.5	52.5	51.7
	BJLM	50.9	52.2	53.9	50.7	49.9	51.5	49.7	50.6	51.6	49.5	50.7	52.4	50.6	50.1	50.5	51.0	51.9	51.4	52.9	51.9	51.7
Uzbek	Base	57.3	59.5	60.1	55.9	57.9	57.6	55.7	57.1	57.1	56.6	55.9	57.1	57.3	57.2	58.7	58.9	58.0	58.6	59.5	59.6	59.7
	TLM	62.9	63.2	62.5	59.9	63.1	62.7	62.1	63.5	63.1	61.1	62.8	64.1	62.1	61.7	63.1	61.9	64.7	64.1	63.9	63.7	62.8
	BALM	60.5	61.7	61.9	56.9	60.7	62.3	58.2	59.8	61.3	60.3	61.4	61.2	60.3	62.3	60.6	61.3	60.9	60.3	61.7	62.3	62.1
	BJLM	61.7	60.5	62.9	60.7	63.3	62.1	59.3	61.9	62.4	61.2	64.2	62.3	60.9	61.9	62.7	61.5	62.3	61.7	63.9	62.7	64.4

Table 8: Per-language results of *sequential* and *joint* few-shot transfer on Kardeş-COPA.

		Zero-Shot			Few-Shot																	
					Squential									Joint								
		EN	TR	EN,TR	EN			TR			EN,TR			EN			TR			EN,TR		
Shots		-	-	-	10	50	100	10	50	100	10	50	100	10	50	100	10	50	100	10	50	100
Azerbaijani	Base	80.3	78.9	80.4	74.5	76.7	76.9	75.7	77.2	77.0	77.6	78.8	78.2	79.3	78.8	79.2	79.7	80.2	80.0	80.4	80.8	80.8
	TLM	75.8	75.5	78.1	75.0	76.2	76.3	75.1	76.6	77.2	77.5	78.0	78.9	77.5	77.4	78.0	76.2	77.4	77.9	78.8	79.2	79.7
	BALM	72.7	78.7	79.7	75.6	76.3	76.3	76.0	77.2	78.1	77.6	78.7	79.4	75.8	76.4	77.1	79.4	79.6	80.1	80.1	80.6	80.5
	BJLM	69.3	77.0	78.3	73.9	74.8	75.6	76.6	77.5	77.9	77.3	78.2	78.5	75.3	75.9	76.4	78.1	79.1	79.5	79.6	80.2	80.5
Kazakh	Base	85.8	84.1	84.8	81.6	82.1	82.4	81.2	82.3	82.3	82.5	83.1	83.8	84.5	84.4	84.9	84.5	85.1	85.4	85.0	85.6	85.6
	TLM	80.6	80.1	81.9	81.1	82.0	82.2	81.2	81.2	81.9	82.5	84.0	83.8	81.8	83.2	83.5	80.9	82.6	83.3	82.6	84.0	84.3
	BALM	81.4	83.2	83.9	81.5	82.7	82.6	82.0	83.2	84.3	82.5	84.6	84.4	82.6	83.7	84.2	83.9	84.7	85.0	84.7	85.6	85.9
	BJLM	78.6	83.2	84.6	79.6	81.5	82.0	80.9	83.1	83.3	82.4	83.7	84.5	80.5	82.3	82.6	83.9	84.5	84.9	85.1	85.6	85.8
Kyrgyz	Base	78.2	77.9	78.7	71.3	72.1	73.3	73.7	74.7	73.4	74.0	75.1	75.9	76.4	76.0	75.8	78.7	79.5	79.4	78.8	79.8	79.5
	TLM	71.3	71.8	74.2	71.2	70.8	71.6	72.5	73.6	73.4	73.4	73.2	73.6	72.7	73.8	73.8	74.1	75.7	76.8	76.0	77.2	77.1
	BALM	71.1	77.3	78.3	69.4	71.3	72.3	74.5	76.5	75.5	75.7	77.0	75.4	72.3	72.8	73.6	77.7	78.6	78.4	78.1	78.7	79.3
	BJLM	69.9	75.1	77.3	68.8	70.6	72.4	73.6	75.0	74.1	74.8	75.8	76.1	71.7	73.3	74.3	76.4	77.2	76.9	77.4	77.9	78.0
Uyghur	Base	69.2	64.8	64.2	65.7	71.2	69.2	67.4	71.8	69.7	66.1	71.1	70.9	64.7	71.1	71.3	64.2	70.9	70.9	63.7	70.0	71.5
	TLM	70.6	69.3	71.3	68.4	71.8	72.4	71.5	72.6	72.0	71.9	73.0	73.8	69.3	72.5	72.6	69.6	72.1	72.7	70.8	73.2	73.6
	BALM	72.8	72.3	73.5	71.5	74.1	74.3	72.8	74.2	74.2	73.2	74.5	74.8	71.3	74.7	74.6	71.7	74.9	75.0	72.9	75.3	75.6
	BJLM	65.7	66.9	69.0	69.0	72.7	71.7	70.5	72.1	71.4	70.4	73.2	73.1	68.5	73.3	73.2	68.3	72.4	72.4	69.8	73.7	73.7
Uzbek	Base	78.3	77.2	77.1	74.2	75.4	75.2	74.6	76.2	75.7	76.6	77.6	76.7	76.7	77.5	77.1	77.9	78.7	78.5	77.8	78.8	78.9
	TLM	70.6	67.0	76.9	72.5	75.6	75.5	74.2	75.6	76.1	77.0	78.2	78.0	74.1	77.0	76.7	75.4	77.2	77.2	77.8	79.0	79.2
	BALM	72.5	77.6	79.3	74.4	75.7	76.1	75.9	76.9	76.9	77.4	78.1	78.1	75.4	77.2	77.6	78.6	79.3	79.3	79.9	80.3	80.5
	BJLM	71.1	76.8	77.3	72.6	74.7	75.2	74.5	76.8	77.3	75.7	77.8	78.1	74.0	76.1	76.4	77.1	78.5	78.7	77.8	79.0	79.1

Table 9: Per-language results of *sequential* and *joint* few-shot transfer on Kardeş-STS.