Parameter-Efficient Multilingual Summarization: An Empirical Study

Anonymous ACL submission

Abstract

Although the emergence of pre-trained Large Language Models has significantly accelerated recent progress in NLP, their ever-increasing size poses significant challenges for conventional fine-tuning, especially in memory-intensive tasks. We investigate the potential of Parameter-Efficient Fine-Tuning, focusing on Low-Rank Adaptation (LoRA), in the domain of multilingual summarization, a task that is both challenging (due to typically long inputs), and relatively unexplored. We conduct an extensive study across different data availability scenarios, including high- and low-data settings, and cross-lingual transfer, leveraging models of different sizes. Our findings reveal that LoRA is competitive with full fine-tuning when trained with high quantities of data, and excels in low-data scenarios and cross-lingual transfer. We also study different strategies for few-shot cross-lingual transfer, finding that continued LoRA tuning outperforms full fine-tuning and the dynamic composition of language-specific LoRA modules.

1 Introduction

The emergence of pre-trained Large Language Models (LLMs), such as PaLM 2 (Anil et al., 2023), LLaMA 2 (Touvron et al., 2023), and the GPT family from OpenAI, has significantly accelerated recent progress in NLP. However, the ever-increasing size of LLMs poses significant challenges for traditional fine-tuning, particularly when faced with many downstream tasks or tasks with a large memory footprint, e.g., due to processing long inputs.

Parameter-Efficient Fine-Tuning (PEFT) methods have recently shown promise in adapting a pre-trained model to different tasks by selectively fine-tuning a small subset of additional parameters. Widely-adopted PEFT techniques include adapters (Houlsby et al., 2019; Pfeiffer et al., 2021), Low-Rank Adaptation (LoRA; Hu et al. 2022), prefix-tuning (Li and Liang, 2021), and prompt-tuning (Lester et al., 2021). Among these, LoRA has become one of the most popular approaches, achieving state-of-the-art performance without introducing latency at inference time. The majority of PEFT studies have focused on natural language understanding, e.g., classification tasks as exemplified in the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks, and monolingual generation, e.g., table-to-text generation or summarization (Li and Liang, 2021).

In this paper, we empirically investigate the potential of LoRA in the domain of multilingual summarization, a task that is both challenging, and relatively unexplored. Multilingual summarization often involves processing lengthy inputs (Hasan et al., 2021), providing a natural testbed for the effective use of PEFT methods. In addition to being able to understand long documents, models are expected to fluently generate sentences in many languages, requiring significant linguistic versatility. Multilingual tasks face additional challenges pertaining to the availability of resources (e.g., for training). It is unrealistic to expect that large-scale and high-quality data will be available or created for every language (Parida and Motlicek, 2019). In situations where multilingual data is scarce, PEFT methods which selectively update a small number of parameters seem more suitable, while fine-tuning can lead to overfitting or catastrophic forgetting (Kirkpatrick et al., 2017; Mitchell et al., 2022).

This motivates us to explore the following research questions: (1) Can LoRA be effectively applied to complex multilingual summarization tasks? and (2) Under which conditions does LoRA exhibit the most potential? To answer these questions, we investigate different data availability scenarios: *high-data* regime (high quantities of training data are available for all languages), *low-data* regime (training data is limited but available for all languages), and *cross-lingual transfer* (zero or only a few examples are available for some languages). In

the latter case, a model trained on a high-resource language (e.g., English) is localized to additional languages for which data is scarce or unavailable (Artetxe et al., 2020; Karthikeyan et al., 2020). In addition to mimicking real-world conditions, the cross-lingual transfer setting allows us to experiment with the composition of language-specific LoRA modules, including the recently proposed few-shot LoraHub (Huang et al., 2023). Our experiments are conducted on two multilingual summarization datasets, XLSum (Hasan et al., 2021) and XWikis (Perez-Beltrachini and Lapata, 2021), using different sizes of the PaLM 2 model, an LLM trained on multilingual text spanning more than 100 languages (Anil et al., 2023).

To summarize, our contributions are as follows: (1) we conduct a comprehensive study of the effectiveness of LoRA for multilingual summarization under different data regimes; (2) we showcase the benefits of LoRA in low-data and cross-lingual transfer settings; and (3) we investigate how to best leverage LoRA for cross-lingual transfer subject to the availability of target language examples.

2 Related Work

Parameter Efficient Fine-Tuning methods focus on enhancing computational efficiency while maintaining competitive performance compared to full fine-tuning. LoRA is one of the most popular PEFT approaches (Hu et al., 2022; Chen et al., 2022). It reduces the number of trainable parameters by learning pairs of rank-decomposition matrices while freezing the model's original weights. This vastly reduces storage requirements for large language models adapted to specific tasks and enables efficient task-switching during deployment, without introducing inference latency. More recent work explores how to adaptively adjust the rank of the matrices (Zhang et al., 2023b; Valipour et al., 2023), proposes generalizations of LoRA and related PEFT approaches under a common framework (He et al., 2022; Chavan et al., 2023), and combines LoRA with quantization (Dettmers et al., 2023). However, most of these studies focus on classification and monolingual generation tasks. In contrast, we investigate the potential of LoRA in the domain of multilingual summarization, a task that is both challenging, and relatively unexplored.

Cross-lingual Transfer requires a model to learn a task from labeled data in one language (typically English), and then perform the equivalent task in another language where no or very little labeled data is available (Artetxe et al., 2020; Karthikeyan et al., 2020; Lauscher et al., 2020; Whitehouse et al., 2022, 2023a). Previous studies focusing on PEFT methods for cross-lingual transfer have explored adapter-based approaches (Pfeiffer et al., 2020; Ansell et al., 2021) and composable sparse fine-tuning (Ansell et al., 2022), among others. Vu et al. (2022) evaluate prompt-tuning (Lester et al., 2021) in a zero-shot setting for cross-lingual summarization, focusing on the Wikilingua dataset (Ladhak et al., 2020). Their study does not cover LoRA, nor does it explore scenarios with more available data (e.g., few-shot settings).

Model Composition and Weight Merging aims to combine individually trained models to enable generalization to unseen tasks. Previous work includes weight composition guided by task similarity (Lv et al., 2023) or arithmetic operations such as addition or subtraction (Zhang et al., 2023a), multi-task prompt pre-training (Sun et al., 2023), and combining models in parameter space by minimizing prediction differences between a merged model and individual models (Jin et al., 2023). For our multilingual summarization task, we also explore the composition of language-specific LoRA matrices through weight averaging, as well as dynamic weight composition when few-shot samples are available (Huang et al., 2023).

3 LoRA for Multilingual Summarization

We now present the fundamentals of LoRA (Hu et al., 2022) and then discuss how individual LoRA modules can be combined (Huang et al., 2023) for cross-lingual transfer. We also introduce our assumptions regarding the availability of training data in the domain of multilingual summarization.

3.1 LoRA and LoraHub

LoRA Let $W_0 \in \mathbb{R}^{d\times k}$ denote the weight matrix of a pre-trained LLM (where d is the input dimension and k is the output dimension). The key idea of LoRA is to represent the fine-tuned W with a low-rank decomposition $W_0 + \Delta W = W_0 + BA$ where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d, k)$, making BA a low-rank matrix compared to W_0. During training, W is frozen, while B and A contain trainable parameters which are effectively a portion $(2r/d)$ of the parameters compared to full fine-tuning. Although LoRA can be in principle applied to any subset of weight matrices, Hu et al.

(2022) only adapt the weight matrices in the self-attention module of the Transformer architecture. We also follow this recipe in experiments and update all four attention matrices (i.e., *query*, *key*, *value*, and *out*).

LoraHub is a gradient-free, few-shot learning approach, recently proposed in Huang et al. (2023). It focuses on composing individually trained LoRA modules for cross-task generalization. Available LoRA modules m_i are synthesized into module $\hat{m} = \sum_{i=1}^{N} w_i m_i$ where w_i is a scalar weight that can assume positive and negative values. The optimal weighted sum is learned through black-box gradient-free optimization (Sun et al., 2022), based on performance metrics on a few examples representative of a new target task.

3.2 Data Regimes

We investigate the effectiveness of LoRA for multilingual summarization under the following data assumptions:

High Data This scenario assumes that *sufficient* training data is available in all languages of interest. Such data could be obtained through automatic pipelines or crowdsourcing.

Low Data In this scenario, we assume that a *limited* number of examples are available in the target languages of interest, typically in the order of dozens or a few hundred. This scenario is common when working with low-resource languages or when data cannot be easily obtained through crowdsourcing but requires input from expert annotators.

Cross-Lingual Transfer Within this context, we consider scenarios where training examples are primarily available in one or a few high-resource languages. We explore three settings corresponding to the following assumptions: (a) only English training data is available; (b) training data is available in some languages besides English, which creates a more complex multilingual setting; and (c) a small number of labeled examples are available in the target language, allowing us to study few-shot cross-lingual generalization.

4 Experimental Setup

This section introduces the datasets and models used in our study. We further elaborate on the details of our experimental setup, and the metrics used to assess the generated summaries.

Dataset	XLSUM	XWIKIS
Source	BBC News	Wikipedia
Languages	44	5
Train/Val/Test Data	1.12M / 114K / 114K	1.43M / 40K / 35K
Input/Output Words	470.2 / 22.1	1042.7 / 63.7

Table 1: Summary statistics for XLSum and XWikis multilingual summarization datasets. Train/Val/Test shows the number of examples in each split. Input/Output shows the average number of *words* in the *English* input document and output summary. XWikis has long documents and multi-sentence summaries.

4.1 Datasets

We perform experiments on two multilingual abstractive summarization datasets which differ with respect to the number of languages they cover, the number of data samples available, and the summarization task itself (short vs long summaries). Dataset statistics are presented in Table 1.

XLSum (Hasan et al., 2021) contains over one million article-summary pairs in 45 languages. The dataset was automatically collated from BBC news, under the assumption that the introductory sentence in the article is effectively a summary of its content. The number of training examples varies significantly among languages, with English having more than 300K instances, and Scottish-Gaelic just above 1K (seeTable 8 in Appendix A).

XWikis (Perez-Beltrachini and Lapata, 2021) consists of document-summary pairs with long documents and multi-sentence summaries. It was synthesized from Wikipedia articles, under the assumption that the body of the article and its lead paragraph together form a document-summary pair. XWikis covers five languages, i.e., Czech, German, English, French, and Chinese. It also includes cross-lingual document-summary instances, created by combining lead paragraphs and article bodies from Wikipedia titles that are language-aligned. In our experiments, we focus on cases where the article and the summary are in the *same* language.

4.2 Modelling Details

Our experiments focus on PaLM 2 (Anil et al., 2023), a decoder-only LLM which, compared to PaLM (Chowdhery et al., 2023), exhibits superior multilingual and reasoning capabilities, as well as better compute efficiency. Specifically, we employ two sizes (XXS and S) of the instruction-tuned FLAN-PaLM 2 model (Wei et al., 2022a). All

experiments were conducted on cloud TPUs, with a learning rate in the range of $\{1e^{-3}, 2e^{-4}, 2e^{-5}\}$. The input/output length was truncated at 2,048/128 tokens for XLSum and 2,048/256 for XWikis.

4.3 Automatic Evaluation

We evaluate the quality of the generated summaries along three dimensions, namely relevance, faithfulness, and conciseness. In terms of relevance, we employ the widely used ROUGE score (Lin, 2004), which measures the degree of n-gram overlap between generated summaries and reference text. Following Aharoni et al. (2023), we compute ROUGE over SentencePiece tokens (Kudo and Richardson, 2018) to avoid inconsistencies in tokenization among languages.

We measure the extent to which generated summaries are faithful to their input using textual entailment (Falke et al., 2019; Kumar and Talukdar, 2020; Honovich et al., 2022; Whitehouse et al., 2023b; Huot et al., 2023). Specifically, for our entailment classifier, we fine-tuned mT5-XXL (Xue et al., 2021) on two NLI datasets, namely ANLI (Nie et al., 2020) and XNLI (Conneau et al., 2018). Following previous work (Aharoni et al., 2023; Huot et al., 2023), for each sentence in the summary, we compute its entailment probability given the input and report the average across sentences.

We also assess if a summary concisely represents the information in the source article using a recently proposed metric trained on the SEAHORSE benchmark (Clark et al., 2023), which is a large-scale collection of human ratings on various dimensions of system summary quality across multiple languages, datasets, and models. We use a publicly available[1] mT5-XXL model (Xue et al., 2021) fine-tuned on SEAHORSE binary conciseness judgments.

5 Results and Analysis

5.1 High-data Regime

In the high-data regime, we use the complete training set, including all languages in XLSum and XWikis. In Table 2, we compare conventional fine-tuning on all layers (Full FT), and a more constrained setting that exclusively updates attention layers (FT-Att) against LoRA variants where attention layers are tuned with different ranks ($r = \{4, 16, 64, 512\}$). We report results with PaLM 2-XXS and select the best checkpoints based on ROUGE-L throughout.

[1] https://huggingface.co/google/seahorse-large-q6

	Params	XLSUM R-L	XLSUM NLI	XLSUM SH	XWIKIS R-L	XWIKIS NLI	XWIKIS SH
Full FT	100%	**31.11**	42.93	31.64	**34.08**	41.04	25.19
FT-Att	20%	30.88	50.32	**36.12**	32.22	37.06	24.20
LoRA-512	13.3%	29.81	42.58	30.16	33.38	40.48	24.78
LoRA-64	1.7%	29.79	45.51	31.80	34.04	45.34	27.02
LoRA-16	0.4%	29.77	48.48	33.25	33.80	46.10	27.42
LoRA-4	0.1%	29.03	**51.16**	34.42	32.92	**47.43**	**27.72**

Table 2: Results on XLSum and XWikis datasets with PaLM 2-XXS trained in the high-data regime: full fine-tuning on all layers (Full FT), on attention layers (FT-Att), and LoRA-* (with different ranks). Params denotes the proportion of trainable parameters. Best ROUGE-L (R-L), NLI, and SEAHORSE (SH) conciseness scores (area under the ROC curve) are in bold.

Perhaps unsurprisingly, conventional fine-tuning on all layers achieves the best ROUGE-L scores for XLSum and XWikis. Updating attention layers only results in competitive performance on XLSum, however, it delivers a drop of 1.86 ROUGE-L points on XWikis. All LoRA variants, even those with high ranks, update fewer parameters than constrained fine-tuning. Despite remarkable efficiency in parameter updates, LoRA with rank 4 lags behind full fine-tuning (by 2.08 ROUGE-L points on XLSum and 1.16 on XWikis). In general, we observe that expanding the parameter update space through higher ranks enhances summary relevance. For XWikis, LoRA with rank 64 is very close to full fine-tuning. However, for XLSum where language diversity and data imbalances are more pronounced, all LoRA variants lag behind full fine-tuning by more than 1 ROUGE-L point. In line with Chen et al. (2022), we observe that LoRA becomes more sensitive to learning rate with higher ranks, requiring more careful hyper-parameter tuning.

With regard to NLI, we note that LoRA achieves superior scores compared to full fine-tuning with lower rank settings exhibiting better summary faithfulness. We observe similar trends with the conciseness metric. Additional results and language-specific performance are included in Appendix B.

Takeaways When training data is available, full fine-tuning yields the most relevant and informative summaries. LoRA is a competitive alternative, particularly when considering summary faithfulness. Its performance can be further enhanced with higher ranks, although more careful hyper-parameter tuning is generally required.

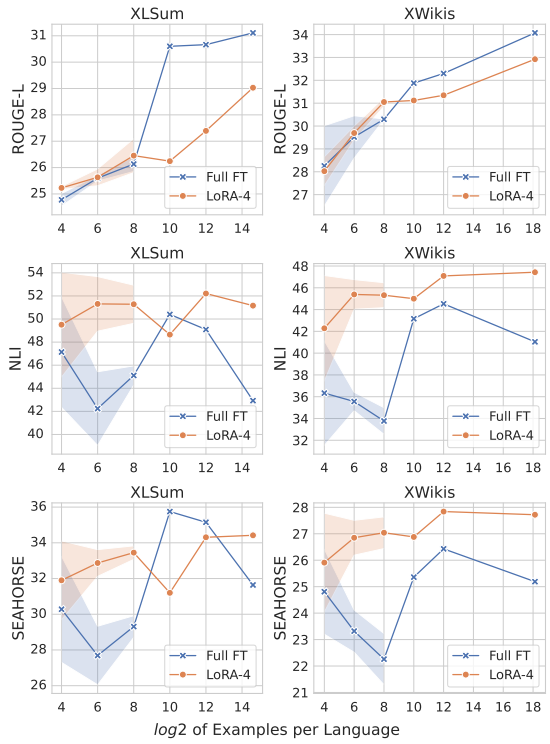


Figure 1: Results on XLSum and XWikis datasets with PaLM 2-XXS trained in the low → high-data regime: Full FT vs. LoRA-4. Results for up to 256 examples per language are averaged over three seeds, with the standard deviation shown in the shaded areas.

5.2 Low-data Regime

We then compare full fine-tuning against LoRA in the low-data regime where limited training data is available. From this section onward, we focus on LoRA with rank 4 and full fine-tuning on all layers.

We randomly sample 16, 64, and 256 training examples per language for both XLSum and XWikis. To ensure our results are robust, we conduct experiments with three different seeds, each with a unique set of samples. To examine how performance evolves as we increase our training samples, we further present experiments with 1,024 and 4,096 examples per language for both datasets.[2] We set the number of validation samples to match that of the training data. As before, we select the best checkpoint based on ROUGE-L and subsequently evaluate on the entire test set.

Figure 1 tracks the performance of PaLM 2-XXS with full fine-tuning and LoRA, when the number of training examples per language varies from 16 to the entire dataset. The x-axis shows the number of examples per language on *log* scale; the high-data setting is approximated by $\sim 2^{14.6}$ examples in XLSum and $\sim 2^{18.1}$ in XWikis. For training data with 256 or fewer samples, we show standard deviation with shaded areas. We observe that LoRA achieves overall better faithfulness (NLI) and conciseness (SH) than full fine-tuning. For ROUGE-L, LoRA demonstrates advantages in low-data scenarios, while full fine-tuning delivers a performance boost when increasing the number of examples from 256 to 1,024. In addition, full fine-tuning is sensitive to checkpoint selection in the low-data regime, due to its susceptibility to overfitting. As a result, it requires more frequent validation for optimal checkpoint selection. In comparison, the training process for LoRA is more stable.

[2]When the number of training samples is set to 4,096, three languages in XLSum already lack sufficient data, so we refrain from selecting more examples per language.

	Test Languages	XLSUM R-L	XLSUM NLI	XLSUM SH	XWIKIS R-L	XWIKIS NLI	XWIKIS SH
Full FT	Non-English	5.20	4.49	6.88	17.51	35.95	22.43
LoRA-4	Non-English	**21.13**	**39.07**	**23.08**	**23.86**	**45.54**	**25.96**
Full FT	English	**32.58**	57.09	38.01	**36.59**	**53.59**	**30.81**
LoRA-4	English	32.21	**63.13**	**43.44**	34.07	49.94	29.01

Table 3: Zero-shot cross-lingual transfer using full fine-tuning (Full FT) and LoRA (rank 4); PaLM 2-XXS models are trained and validated on English and tested on all other languages (Non-English) and English only. Best ROUGE-L (R-L), NLI, and SEAHORSE (SH) conciseness scores (area under the ROC curve) are in bold.

Takeaways In low-data scenarios, LoRA is a better alternative to full fine-tuning. It delivers consistently competitive or even superior results with the added advantage of efficient and stable training.

5.3 Cross-lingual Transfer

We now focus on cross-lingual transfer in multilingual summarization and explore two common scenarios, namely zero- and few-shot learning. For LoRA, we focus on rank 4 in all experiments.

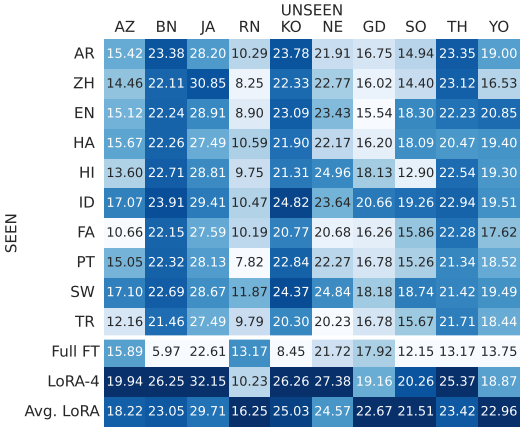
Zero-shot Transfer from English

We first consider a typical scenario where only English training data is available, i.e., training and validation are carried out using English examples, whilst the model is tested on new languages.

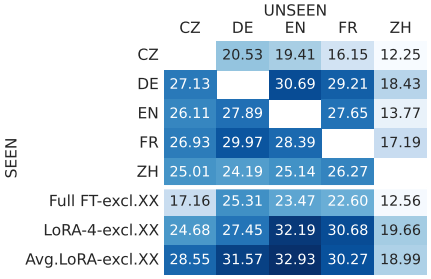
Table 3 shows the performance of PaLM 2-XXS with full fine-tuning and LoRA. We separate results on English as they are not zero-shot (second block in Table 3) and broadly align with our findings in Section 5.1 (high-data regime). Full fine-tuning generally outperforms LoRA except for NLI and SH on English XLSum. In the cross-lingual trans-

Hausa
Target: Gwamnatin Najeriya ta ce 'yan kasar sun ga irin amfani da rufe iyakokin kasar ya yi a fannin tattalin arzikinta.
Full FT: President Muhammadu Buhari has appointed his deputy, the BBC presenter and former minister, Shugaba Muhammadu Buhari, as the new chairman of the Presidential Council.
LoRA-4: Gwamnatin Nijeriya ta yi tsokacin da shawarar da zai rufe iyakokin kasar.
Indonesian
Target: Perempuan Vietnam yang dituding terlibat dalam pembunuhan Kim Jong-nam, saudara tiri dari pemimpin Korea Utara Kim Jong-un, telah dibebaskan.
Full FT: Kim Jong-nam, the wife of North Korean leader Kim Jong-un, has died in a fight with Malaysia Airlines flight MH17. Here are the key points of the ruling:
LoRA-4: Seorang wanita Vietnam yang didakwa sebagai bagian dari pembunuhan Kim Jong-nam, saudara tiri dari pemimpin Korea Utara, telah dibebaskan.

Table 4: XLSum Output Examples: zero-shot transfer from English using Full FT and LoRA with PaLM 2-XXS. Full FT fails to generate summaries in target languages and the content are off-topic.



(a) ROUGE-L scores for 10 test languages on XLSum.



(b) ROUGE-L scores for five languages on XWikis.

Figure 2: Zero-shot cross-lingual transfer on XLSum (top) and XWikis (bottom); PaLM 2-XXS models are trained on one (SEEN) language and tested on another (UNSEEN). We also show results with full fine-tuning on *all* seen languages (Full FT), LoRA, and (average) weighted combination of language-specific LoRA modules (Avg.LoRA); excl.XX in XWikis denotes *leave-one-out* training, excluding the test language.

fer scenario (first block in Table 3), fine-tuning performs exceptionally poorly across metrics and languages on XLSum. The gap is smaller for XWikis as only four non-English languages are covered and all but Chinese are Indo-European. Further examination of the model output shows that the generated text is mostly in English rather than the target language. The model appears to comprehend the new language (i.e., input documents), however, it struggles to generate output accordingly.

Table 4 illustrates XLSum examples of model output for Hausa and Indonesian. In both cases, Full FT summaries are in English, and off-topic (the Hausa article discusses the Nigerian government's decision to close its borders, while the Indonesian one reports on the murder of Kim Jong-un). In Appendix B, we provide per-language results which highlight that for zero-shot transfer from English, full fine-tuning consistently lags behind LoRA in *every* language, even in cases where languages are well-represented in the pre-training phase of PaLM 2 or are considered linguistically close to English.[3] This catastrophic forgetting behavior echoes the findings in Vu et al. (2022).

Zero-shot Transfer from Multiple Languages

We extend our study of zero-shot cross-lingual transfer to scenarios where training data is available in multiple languages rather than just English.

For XLSum, we create a training data pool of 10 languages from eight distinct linguistic families, each with substantial training data. These languages include Arabic (AR), (Simplified) Chinese (ZH), English (EN), Hausa (HA), Hindi (HI), Indonesian (ID), Persian (FA), Portuguese (PT), Swahili (SW), and Turkish (TR). Additionally, we select 10 test languages: Azerbaijani (AZ), Bengali (BN), Japanese (JA), Kirundi (RN), Korean (KO), Nepali (NE), Scottish Gaelic (GD), Somali (SO), Thai (TH), Yoruba (YO). Test languages were selected so that they are maximally diverse, each representing a unique language family.[4] For XWikis, we adopt a *leave-one-out* approach, since it only covers five languages. We rotate through the available languages, using four for training and one for testing.

In addition to full fine-tuning and LoRA, we report experiments with language-specific LoRA modules, each trained on examples from one language. An advantage of such specialized modules is their scalability and adaptability. When additional languages become available, there is no need

[3] See Table 21 in Anil et al. (2023) regarding the distribution of languages used in the pre-training of PaLM 2.

[4] See Appendix A for details of language families in XLSum.

to re-train the entire model; it is sufficient to add a new language-specific module. During inference, we can also flexibly experiment with various LoRA modules or weight composition methods. As mentioned in Section 2, weight composition is an active research area that has demonstrated effectiveness across a spectrum of applications. We adopt a simple approach that computes the weighted average of all available modules.

The heatmaps in Figure 2 represent the ROUGE-L for models trained in one language and tested on another. Rows represent source models from SEEN languages, and columns UNSEEN test languages. The color scale is column-wise normalized to provide a comparative view of the performance of the best and worst models for each UNSEEN test language. In the bottom three rows, we also illustrate the performance of models trained on multiple seen languages and tested on unseen ones. We experiment with full fine-tuning, LoRA (rank 4), and weighted average LoRA.

Based on the results of Figure 2, we observe that: (1) full fine-tuning consistently lags behind LoRA in zero-shot cross-lingual transfer, even with a diverse collection of languages besides English; (2) the weighted average of language-specific LoRA modules (Avg.LoRA) and LoRA (trained on all available languages together) benefit different unseen languages. Particularly for XLSum, lower-resource languages (i.e., RN, GD, SO, and YO), exhibit superior performance with language-specific LoRA training; and (3) languages with similarities demonstrate better transferability, as exemplified by transferring ZH to JA and SW to RN on XLSum, and the Indo-European languages on XWikis.

Few-shot Cross-lingual Transfer

Finally, we consider situations where some examples are available in the target languages and explore effective strategies for utilizing them. We follow on from the previous section and assume that models have been already trained on (seen) languages with sufficient data. One approach is to *continue* training these models using target language examples. So, if the starting checkpoint was obtained from full fine-tuning on seen languages, we continue with full fine-tuning on the new languages. We also adopt the same strategy for LoRA.

Another widely-used technique is *in-context learning*, where input and output examples are concatenated to form in-context demonstrations. Despite promising results in many LLM applications (Brown et al., 2020; Wei et al., 2022b), in-context learning becomes less practical in the domain of multilingual summarization where models are expected to process long articles, which is memory-intensive, especially as the number of examples grows. Instead, we experiment with the recently proposed few-shot LoraHub learning approach (Section 3.1). The original formulation of LoraHub (Huang et al., 2023) does not assume any prior knowledge of the available LoRA modules which are randomly sampled and initialized with zero weights (i.e., starting from a general-purpose pre-trained LLM). We initialize LoraHub with the weighted average of pre-existing language-specific LoRA modules. The composition of modules fine-tuned on the same task, albeit in different languages, offers a stronger baseline compared to a pre-trained LLM.

We consider two few-shot settings, with 16 or 64 target language examples, simulating practical scenarios where human annotators or experts create a few examples for low-resource languages. We compare few-shot *continued learning* and *LoraHub learning*, using the same examples. To ensure robustness, all experiments are run on three different sets of examples, and we report the average. For continued learning, we split the examples into training and validation using 14/2 and 60/4 splits. For LoraHub, we use the Nevergrad toolkit[5] for black-box optimization. We empirically compared ROUGE-L and loss as performance metrics guiding the optimization, and found that ROUGE-L led

	Method	XLSUM R-L	XLSUM NLI	XLSUM SH	XWIKIS R-L	XWIKIS NLI	XWIKIS SH
ZERO-SHOT	Full FT	14.48	28.87	13.71	20.22	30.17	16.26
	LoRA-4	22.59	37.39	24.21	**28.46**	48.31	26.40
	Avg. LoRA	**22.74**	**49.14**	**32.44**	26.93	**49.29**	**26.86**
16-SHOT	Full FT (cl)	22.31	30.15	18.79	26.90	34.17	21.82
	LoRA-4 (cl)	**24.71**	**41.12**	**26.47**	**30.05**	45.90	**28.20**
	LoraHub	23.37	38.95	26.07	27.59	**47.45**	25.84
64-SHOT	Full FT (cl)	24.30	30.65	19.57	28.73	39.42	24.16
	LoRA-4 (cl)	**25.94**	**42.07**	27.66	**31.08**	45.12	**28.05**
	LoraHub	24.21	41.34	**28.02**	27.66	**48.09**	26.56

Table 5: Cross-lingual transfer on 10 XLSum languages and five XWikis languages (using *leave-one-out* training) for PaLM 2-XXS model. 16- and 64-shot experiments show average results from three different seed runs. For *continued learning* (cl), we use a 14/2 and 60/4 training/validation split. Best ROUGE-L (R-L), NLI, and SEAHORSE (SH) conciseness scores (area under the ROC curve) are in bold. Results for individual languages are in Tables 11 and 14, Appendix B.

[5] https://facebookresearch.github.io/nevergrad

	Params	XLSUM R-L	XLSUM NLI	XLSUM SH	XWIKIS R-L	XWIKIS NLI	XWIKIS SH
Full FT	100%	**36.99**	58.72	41.92	**39.65**	46.03	28.01
LoRA-4	0.04%	36.29	**61.64**	**43.99**	39.25	**47.56**	**28.30**

Table 6: Results on XLSum and XWikis datasets with PaLM 2-S trained in the high-data regime: Full FT and LoRA (rank 4). Params denotes the proportion of trainable parameters. Best ROUGE-L (R-L), NLI, and SEAHORSE (SH) conciseness scores (area under the ROC curve) are in bold.

	Test Languages	XLSUM R-L	XLSUM NLI	XLSUM SH	XWIKIS R-L	XWIKIS NLI	XWIKIS SH
Full FT	Non-English	33.22	60.72	41.96	35.70	46.27	27.51
LoRA-4	Non-English	**33.31**	**64.18**	**43.98**	**36.00**	**47.23**	**28.69**
Full FT	English	**40.38**	71.21	45.82	**42.03**	**51.76**	28.95
LoRA-4	English	39.61	**78.05**	**47.02**	41.53	50.09	**29.07**

Table 7: Zero-shot transfer on XLSum and XWikis using Full FT and LoRA (rank 4). PaLM 2-S models are trained and validated on English and tested on all other languages (Non-English) and English only. Best ROUGE-L (R-L), NLI, and SEAHORSE (SH) conciseness scores (area under the ROC curve) are in bold.

to more stable results.

Table 5 presents our results on XLSum and XWikis with 16- and 64-shots, averaged across test languages. Zero-shot results are also included for comparison. Our analysis supports the following observations: (1) with only a few target language examples (e.g., 16), full fine-tuning sees a remarkable improvement, resulting in an average boost of 7.8 ROUGE-L points on XLSum and 6.7 on XWikis, corroborating the findings of Lauscher et al. (2020); (2) LoraHub slightly enhances ROUGE-L performance compared to (zero-shot) weighted-average on XLSum with only 16 examples; (3) LoRA continued learning consistently outperforms full fine-tuning and LoRAHub in terms of ROUGE-L and SH; however, LoraHub is superior in terms of NLI for XWikis.

Takeaways In cross-lingual transfer situations, LoRA is consistently superior performance compared to full fine-tuning. LoRA continued learning shows particular promise when only a small number of examples are available in the target language.

6 Scaling Up

We extend our analysis to the larger PaLM 2-S model, focusing on the high-data regime and zero-shot cross-lingual transfer using English data. Our results are summarized in Table 6 and Table 7.

Interestingly, LoRA and full fine-tuning achieve similar performance, with LoRA taking the lead in cross-lingual transfer (see first block in Table 7). We hypothesize that when using the larger PaLM 2-S model, the increased capacity makes up for the small percentage of trainable parameters in LoRA (only 0.04% of the parameters), allowing it to benefit more from high-data regime training. At the same time, the larger model is more robust and does not exhibit catastrophic forgetting during full fine-tuning. As a result, we see that full fine-tuning performs on par with LoRA in the zero-shot cross-lingual setting (see Table 7).

Takeaways For larger models such as PaLM 2-S, LoRA is on par with full fine-tuning but a better choice when considering computational efficiency.

7 Conclusions

In this paper, we explored the effectiveness of LoRA on multilingual summarization across a diverse range of scenarios primarily determined by the availability of training data. We summarize our key findings by comparing the computationally efficient LoRA against full fine-tuning.

LoRA achieves **superior performance** to full fine-tuning in zero-shot and few-shot cross-lingual transfer scenarios, and low-data settings (e.g., training data with fewer than 1K samples). This is most pronounced with smaller models (e.g., PaLM 2-XXS). In the specific case of few-shot learning, LoRA continued learning outperforms LoraHub. It also achieves overall superior summary faithfulness and conciseness across various scenarios.

For larger models like PaLM 2-S, LoRA exhibits **on-par performance** to full fine-tuning. This suggests that model capacity matters. Notably, for smaller models like PaLM 2-XXS, LoRA displays **worse performance** in the full fine-tuning (high-data) regime, when said performance is measured via ROUGE-L, but is consistently superior in terms of faithfulness and conciseness.

Taken together, our results underscore the utility of PEFT methods for complex multilingual tasks and cross-lingual transfer. Avenues for future work include few-shot transfer and effective ways to combine LoRA modules, e.g., by learning which ones to activate for different tasks/languages (Ponti et al., 2023). It would also be interesting to reproduce our results across varied LLMs and broader multilingual generation tasks, beyond summarization.

Limitations

In this work, we have focused exclusively on decoder-only models. Future work could explore a wider range of LLMs, including encoder-decoder models. We anticipate the observations gained from decoder-only models to largely align with those from encoder-decoder models, thus generalizing our findings. In our cross-lingual transfer studies, we only considered LoRA models with a rank of 4, due to computational considerations. Expanding to additional LoRA settings would allow us to perform a more thorough comparison. Finally, our experiments have exclusively focused on multilingual summarization tasks. Extending our study to a wider range of multilingual text generation tasks with long input and output would provide a more comprehensive perspective on the capabilities and limitations of LoRA.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual Summarization with Factual Consistency Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. PaLM 2 Technical Report. *arXiv preprint arXiv:2305.10403*.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable Sparse Fine-Tuning for Cross-Lingual Transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

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.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. 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.

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

Arnav Chavan, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen. 2023. One-for-All: Generalized LoRA for Parameter-Efficient Fine-tuning. *arXiv preprint arXiv:2306.07967*.

Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang. 2022. Revisiting Parameter-Efficient Tuning: Are We Really There Yet? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2612–2626, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A Multilingual, Multifaceted Dataset for Summarization Evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. 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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. 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 *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2023. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*.

Fantine Huot, Joshua Maynez, Chris Alberti, Reinald Kim Amplayo, Priyanka Agrawal, Constanza Fierro, Shashi Narayan, and Mirella Lapata. 2023. μplan: Summarizing using a content plan as cross-lingual bridge. *arXiv preprint arXiv:2305.14205*.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. 2023. Dataless knowledge fusion by merging weights of language models. In *The Eleventh International Conference on Learning Representations*.

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

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 of the United States of America*, 114(13):3521—3526.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. 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.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xingtai Lv, Ning Ding, Yujia Qin, Zhiyuan Liu, and Maosong Sun. 2023. Parameter-efficient weight ensembling facilitates task-level knowledge transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 270–282, Toronto, Canada. Association for Computational Linguistics.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

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*, pages 4885–4901, Online. Association for Computational Linguistics.

Shantipriya Parida and Petr Motlicek. 2019. Abstract text summarization: A low resource challenge. 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 5994–5998, Hong Kong, China. Association for Computational Linguistics.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. 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, Alessandro Sordoni, Yoshua Bengio, and Siva Reddy. 2023. Combining parameter-efficient modules for task-level generalisation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 687–702, Dubrovnik, Croatia. Association for Computational Linguistics.

Tianxiang Sun, Zhengfu He, Qin Zhu, Xipeng Qiu, and Xuanjing Huang. 2023. Multitask pre-training of modular prompt for Chinese few-shot learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11156–11172, Toronto, Canada. Association for Computational Linguistics.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20841–20855. PMLR.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. 2023. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3274–3287, Dubrovnik, Croatia. Association for Computational Linguistics.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Chenxi Whitehouse, Monojit Choudhury, and Alham Aji. 2023a. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore. Association for Computational Linguistics.

Chenxi Whitehouse, Fenia Christopoulou, and Ignacio Iacobacci. 2022. EntityCS: Improving zero-shot cross-lingual transfer with entity-centric code switching. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6698–6714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chenxi Whitehouse, Clara Vania, Alham Fikri Aji, Christos Christodoulopoulos, and Andrea Pierleoni. 2023b. WebIE: Faithful and robust information extraction on the web. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7734–7755, Toronto, Canada. 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*, Online. Association for Computational Linguistics.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. 2023a. Composing parameter-efficient modules with arithmetic operations. *arXiv preprint arXiv:2306.14870*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023b. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*.

A Datasets

Table 8 and Table 9 show the language families and the number of training examples per language in the XLSum and XWikis datasets.

B Additional Results

Table 10 shows ROUGE-1, ROUGE-2, and ROUGE-L scores for LoRA and full fine-tuning with PaLM 2-XXS on the two datasets. We additionally report activating LoRA tuning on Feed Forward layers with different ranks.

Table 11, Table 12, and Table 13 show ROUGE-L, NLI, and SEAHORSE few-shot learning results for individual languages on XLSum. Table 14, Table 15, and Table 16 show ROUGE-L, NLI, and SEAHORSE few-shot learning results for individual languages on XWikis.

Table 17, Table 18, and Table 19 show ROUGE-L, NLI, and SEAHORSE results for PaLM 2-XXS on XWikis for individual languages; in the high-data regime and in a zero-shot cross-lingual transfer setting from English. Table 20, Table 21, and Table 22 show ROUGE-L, NLI, and SEAHORSE results for PaLM 2-XXS on XLSum for individual languages; in the high-data regime and in a zero-shot cross-lingual transfer setting from English.

Language	ISO	Language Family	# Train
English	EN	Indo-European	306,522
Hindi	HI	Indo-European	70,778
Urdu	UR	Indo-European	67,665
Russian	RU	Indo-European	62,243
Portuguese	PT	Romance	57,402
Persian	FA	Indo-Iranian	47,251
Ukrainian	UK	Slavic	43,201
Indonesian	ID	Austronesian	38,242
Spanish	ES	Romance	38,110
Arabic	AR	Semitic	37,519
Chinese-Traditional	ZH	Sino-Tibetan	37,373
Chinese-Simplified	ZH	Sino-Tibetan	37,362
Vietnamese	VI	Austroasiatic	32,111
Turkish	TR	Turkic	27,176
Tamil	TA	Dravidian	16,222
Pashto	PS	Indo-Iranian	14,353
Marathi	MR	Indo-Aryan	10,903
Telugu	TE	Dravidian	10,421
Welsh	CY	Celtic	9,732
Pidgin	PI	Unknown	9,208
Gujarati	GU	Indo-European	9,119
French	FR	Romance	8,697
Punjabi	PA	Indo-Iranian	8,215
Bengali	BN	Indo-European	8,102
Swahili	SW	Bantu	7,898
Serbian-Latin	SR	Indo-European	7,276
Serbian-Cyrillic	SR	Indo-European	7,275
Japanese	JA	Japonic	7,113
Thai	TH	Kra-Dai Languages	6,616
Azerbaijani	AZ	Turkic	6,478
Hausa	HA	Afro-Asiatic	6,418
Yoruba	YO	Niger-Congo	6,350
Oromo	OM	Afro-Asiatic	6,063
Somali	SO	Afro-Asiatic	5,962
Nepali	NE	Indo-Aryan	5,808
Amharic	AM	Semitic	5,761
Kirundi	RN	Bantu	5,746
Tigrinya	TI	Semitic	5,451
Uzbek	UZ	Turkic	4,728
Burmese	MY	Sino-Tibetan	4,569
Korean	KO	Koreanic	4,407
Igbo	IG	Niger-Congo	4,183
Sinhala	SI	Indo-European	3,249
Kyrgyz	KY	Turkic	2,266
Scottish-Gaelic	GD	Celtic	1,313

Table 8: Language family and the Number of training examples per language in XLSum.

Language	ISO	Language Family	# Train
English	EN	Indo-European	624,178
German	DE	Indo-European	390,203
French	FR	Indo-European	323,915
Czech	CS	Indo-European	61,224
Chinese	ZH	Sino-Tibetan	31,281

Table 9: Language family and the Number of training examples per language in XWikis.

PaLM 2-XXS	Trainable Layers		Params	XLSUM R-L	XLSUM R-1	XLSUM R-2	XWIKIS R-L	XWIKIS R-1	XWIKIS R-2
Full FT	All Layers		100%	31.11	41.66	21.78	34.08	42.68	24.37
	Attention Layers		20%	30.88	41.17	21.43	32.22	41.48	22.54
LoRA	Attention Layers	*rank 512*	13.3%	29.81	40.33	20.25	33.38	41.52	23.63
		rank 64	1.7%	29.79	39.98	20.18	34.04	41.58	24.28
		rank 16	0.4%	29.77	39.75	20.09	33.80	41.34	24.14
		rank 4	0.1%	29.03	38.83	19.28	32.92	39.97	23.27
	Attention + FFN Layers	*rank 64*	5.4%	29.45	39.64	19.79	33.59	41.37	23.79
		rank 16	1.4%	29.79	39.99	20.17	33.55	41.11	23.95
		rank 4	0.3%	29.67	39.76	20.02	33.70	40.82	24.05

Table 10: Results on XLSum and XWikis datasets with PaLM 2-XXS trained in the high-data regime: full fine-tuning on all layers, full fine-tuning on attention layers, and LoRA (with different ranks). Params denotes the proportion of trainable parameters.

	PaLM 2-XXS	AVG	AZ	BN	JA	RN	KO	NE	GD	SO	TH	YO
ZERO-SHOT	Full FT	14.48	15.89	5.97	22.61	13.17	8.45	21.72	17.92	12.15	13.17	13.75
	LoRA-4	22.59	**19.94**	**26.25**	**32.15**	10.23	**26.26**	**27.38**	19.16	20.26	**25.37**	18.87
	Avg. LoRA	**22.74**	18.22	23.05	29.71	**16.25**	25.03	24.57	**22.67**	**21.51**	23.42	**22.96**
16-SHOT	Full FT + *continued learning*	22.31	16.64	22.95	28.28	17.02	24.31	26.94	19.56	19.66	23.58	24.18
	LoRA-4 + *continued learning*	**24.71**	**20.74**	**26.19**	**32.26**	**17.82**	**27.13**	**27.82**	23.00	22.26	24.30	**25.55**
	LoraHub	23.37	18.58	24.81	27.69	16.65	25.82	25.40	**24.83**	**23.13**	**24.71**	22.05
64-SHOT	Full FT + *continued learning*	24.30	17.86	22.80	32.49	**19.28**	27.09	28.89	21.72	22.17	23.90	26.83
	LoRA-4 + *continued learning*	**25.94**	**20.91**	**26.08**	**33.10**	19.09	**28.43**	**29.38**	**25.78**	**23.06**	**25.48**	**28.06**
	LoraHub	24.21	20.10	25.16	29.03	17.61	27.46	26.82	24.94	23.04	24.37	23.53

Table 11: Cross-lingual transfer results (ROUGE-L) on 10 XLSum languages XLSum for PaLM 2-XXS model. 16- and 64-shot experiments show average results from three different seed runs. For *continued learning*, we use a 14/2 and 60/4 split for training/validation.

	PaLM 2-XXS	AVG	AZ	BN	JA	RN	KO	NE	GD	SO	TH	YO
ZERO-SHOT	Full FT	28.87	19.17	28.28	35.27	24.00	30.76	46.44	**38.22**	15.23	33.02	18.26
	LoRA	37.39	37.92	52.61	62.54	9.62	54.57	47.35	16.86	21.92	53.23	17.26
	Avg. LoRA	**49.14**	**45.54**	**60.55**	**66.37**	**39.63**	**66.44**	**55.86**	33.43	**34.26**	**60.86**	**28.47**
16-SHOT	Full FT + *continued learning*	30.15	15.83	46.29	37.55	17.55	35.61	42.22	20.07	26.69	39.74	20.00
	LoRA + *continued learning*	**41.12**	37.33	**56.45**	**52.31**	30.62	**58.08**	**49.10**	24.32	**33.37**	45.02	**24.58**
	LoraHub	38.95	**37.76**	47.95	48.17	**35.52**	49.29	40.90	**32.14**	26.29	**52.08**	19.40
64-SHOT	Full FT + *continued learning*	30.65	20.06	39.10	47.13	12.54	41.70	47.93	18.01	24.03	46.00	9.96
	LoRA + *continued learning*	**42.07**	**37.16**	**53.76**	**54.85**	18.20	52.19	**52.90**	**31.29**	**37.10**	52.61	**30.60**
	LoraHub	41.34	36.56	44.52	54.47	**47.70**	**53.45**	45.16	29.60	23.98	**54.35**	23.65

Table 12: Cross-lingual transfer results (NLI) on 10 XLSum languages XLSum for PaLM 2-XXS model. 16- and 64-shot experiments show average results from three different seed runs. For *continued learning*, we use a 14/2 and 60/4 split for training/validation.

PaLM 2-XXS		AVG	AZ	BN	JA	RN	KO	NE	GD	SO	TH	YO
ZERO-SHOT	Full FT	13.71	12.38	12.97	24.94	9.43	8.40	30.71	13.57	5.18	12.57	6.99
	LoRA-4	24.21	25.40	36.15	48.70	4.15	36.86	32.84	7.10	9.02	36.71	5.17
	Avg. LoRA	**32.44**	**29.79**	**41.96**	**53.99**	**20.28**	**46.02**	**39.28**	**19.00**	**17.15**	**41.79**	**15.11**
16-SHOT	Full FT + *continued learning*	18.79	11.44	29.33	26.83	9.69	24.72	26.20	8.31	11.22	29.66	10.51
	LoRA-4 + *continued learning*	**26.47**	26.22	**37.43**	37.68	14.53	**39.02**	**32.29**	13.57	**15.95**	33.12	**14.91**
	LoraHub	26.07	**26.35**	33.69	**38.75**	**17.56**	34.90	29.20	**17.87**	13.45	**38.99**	9.92
64-SHOT	Full FT + *continued learning*	19.57	13.88	26.64	29.87	7.77	27.07	28.51	8.78	11.80	32.63	8.77
	LoRA-4 + *continued learning*	27.66	**27.02**	**37.05**	40.43	9.67	34.20	**35.36**	16.84	**18.37**	38.63	**19.03**
	LoraHub	**28.02**	26.81	31.88	**46.24**	**23.37**	**37.88**	33.13	**17.36**	11.72	**39.58**	12.17

Table 13: Cross-lingual transfer results (SEAHORSE) on 10 XLSum languages XLSum for PaLM 2-XXS model. 16- and 64-shot experiments show average results from three different seed runs. For *continued learning*, we use a 14/2 and 60/4 split for training/validation.

PaLM 2-XXS		AVG	CS	DE	EN	FR	ZH
ZERO-SHOT	Full FT	20.22	17.16	25.31	23.47	22.60	12.56
	LoRA-4	**28.46**	**28.55**	**31.57**	**32.93**	**30.27**	18.99
	Avg. LoRA	26.93	24.68	27.45	32.19	30.68	**19.66**
16-SHOT	Full FT + *continued learning*	26.90	22.53	29.23	30.50	26.16	26.11
	LoRA-4 + *continued learning*	**30.05**	**27.68**	**33.76**	31.98	**30.12**	**26.70**
	LoraHub	27.59	26.09	29.81	**32.70**	29.10	20.25
64-SHOT	Full FT + *continued learning*	28.73	26.45	30.17	32.24	28.86	25.95
	LoRA-4 + *continued learning*	**31.08**	**28.97**	**34.09**	**33.11**	**30.99**	**28.24**
	LoraHub	27.66	26.05	29.82	33.00	29.20	20.25

Table 14: Cross-lingual transfer results (ROUGE-L) on XWikis using *leave-one-out* training with PaLM 2-XXS model. 16-shot and 64-shot experiments show the average results obtained across three different seed runs. For *continued learning*, we use a 14/2 and 60/4 split for training/validation.

PaLM 2-XXS		AVG	CS	DE	EN	FR	ZH
ZERO-SHOT	Full FT	30.17	33.30	28.60	34.73	24.07	30.14
	LoRA-4	48.31	**52.80**	**44.27**	48.66	40.05	55.78
	Avg. LoRA	**49.29**	52.67	42.86	**50.34**	**43.76**	**56.80**
16-SHOT	Full FT + *continued learning*	34.17	26.93	31.51	43.51	26.73	42.17
	LoRA-4 + *continued learning*	45.90	48.17	37.67	**49.65**	39.30	**54.73**
	LoraHub	**47.45**	**52.71**	**41.32**	48.09	**43.15**	51.98
64-SHOT	Full FT + *continued learning*	39.47	35.29	32.36	54.19	35.73	39.81
	LoRA-4 + *continued learning*	45.12	48.46	36.48	**50.75**	38.31	51.62
	LoraHub	**48.09**	**52.74**	**41.31**	50.63	**42.72**	**53.05**

Table 15: Cross-lingual transfer results (NLI) on XWikis using *leave-one-out* training with PaLM 2-XXS model. 16-shot and 64-shot experiments show the average results obtained across three different seed runs. For *continued learning*, we use a 14/2 and 60/4 split for training/validation.

PaLM 2-XXS		AVG	CS	DE	EN	FR	ZH
ZERO-SHOT	Full FT	16.26	13.13	19.74	18.81	16.90	12.73
	LoRA-4	26.40	29.84	**28.59**	27.22	27.23	19.11
	Avg. LoRA	**26.86**	**30.13**	28.48	**28.07**	**28.18**	**19.44**
16-SHOT	Full FT + *continued learning*	21.82	18.10	25.53	25.00	21.47	18.98
	LoRA-4 + *continued learning*	**28.20**	27.80	**29.59**	**29.43**	**29.18**	**25.02**
	LoraHub	25.84	**28.95**	27.48	28.20	27.63	16.96
64-SHOT	Full FT + *continued learning*	24.16	22.92	23.31	29.15	26.80	18.64
	LoRA-4 + *continued learning*	**28.05**	28.02	**28.80**	**30.08**	**29.16**	**24.17**
	LoraHub	26.56	**29.34**	27.33	29.93	27.60	18.58

Table 16: Cross-lingual transfer results (SEAHORSE) on XWikis using *leave-one-out* training with PaLM 2-XXS model. 16-shot and 64-shot experiments show the average results obtained across three different seed runs. For *continued learning*, we use a 14/2 and 60/4 split for training/validation.

PaLM 2-XXS	High-Data		EN Zero-Shot	
	FT	LoRA	FT	LoRA
Average	34.08	32.92	17.51	23.86
English	35.13	34.16	—	—
German	36.97	36.08	20.64	27.89
French	34.53	33.65	16.77	27.65
Czech	31.92	30.82	19.21	26.11
Chinese	31.82	29.91	13.43	13.77

Table 17: Per language ROUGE-L results on XWikis using full fine-tuning (FT) and LoRA (rank 4), for PaLM 2-XXS in the high-data regime and in a zero-shot cross-lingual transfer setting from English.

PaLM 2-XXS	High-Data		EN Zero-Shot	
	FT	LoRA	FT	LoRA
Average	41.04	47.43	35.95	45.54
English	48.34	51.35	—	—
German	35.27	39.42	37.09	42.02
French	35.58	39.31	37.32	41.73
Czech	42.46	50.78	33.75	51.75
Chinese	43.53	56.30	35.66	46.65

Table 18: Per language NLI results on XWikis using full fine-tuning (FT) and LoRA (rank 4), for PaLM 2-XXS in the high-data regime and in a zero-shot cross-lingual transfer setting from English.

PaLM 2-XXS	High-Data		EN Zero-Shot	
	FT	LoRA	FT	LoRA
Average	25.19	24.20	22.43	25.96
English	27.80	29.37	—	—
German	26.42	28.88	25.39	28.47
French	25.95	29.15	24.87	28.46
Czech	25.36	28.40	21.86	28.97
Chinese	20.42	22.79	17.60	17.93

Table 19: Per language SEAHORSE results on XWikis using full fine-tuning (FT) and LoRA (rank 4), for PaLM 2-XXS in the high-data regime and in a zero-shot cross-lingual transfer setting from English.

PaLM 2-XXS	High-Data		EN Zero-Shot	
	FT	LoRA	FT	LoRA
Average	31.11	29.03	5.20	21.13
English	32.33	31.25	–	–
Hindi	35.41	32.80	4.43	27.24
Urdu	34.93	31.70	1.31	22.67
Russian	27.40	25.08	5.67	22.60
Portuguese	30.82	28.50	8.85	26.44
Persian	34.64	31.91	3.67	27.94
Ukrainian	27.64	24.92	5.82	19.32
Indonesian	33.42	31.28	8.49	27.87
Spanish	26.21	24.80	8.21	22.85
Arabic	29.47	27.52	4.19	23.26
Chinese-Traditional	36.74	33.44	3.03	25.86
Chinese-Simplified	36.95	33.96	2.11	28.54
Vietnamese	32.11	30.00	5.68	24.69
Turkish	31.09	28.08	7.11	24.41
Tamil	32.71	29.67	3.44	21.22
Pashto	36.41	33.81	3.13	14.76
Marathi	28.14	26.25	4.73	17.97
Telugu	29.62	27.31	4.04	19.30
Welsh	30.72	27.72	6.78	23.75
Pidgin	31.50	30.37	16.84	22.76
Gujarati	35.79	33.43	3.88	26.00
French	29.67	29.74	10.32	26.55
Punjabi	42.01	40.61	2.08	34.14
Bengali	29.52	28.26	1.85	22.29
Swahili	31.11	29.81	6.45	25.14
Serbian-Latin	22.92	21.94	5.56	18.87
Serbian-Cyrillic	24.55	23.60	5.49	15.28
Japanese	38.34	36.08	2.01	29.04
Thai	25.93	26.53	4.53	22.26
Azerbaijani	24.01	23.36	5.34	14.31
Hausa	33.03	28.85	7.55	15.82
Yoruba	33.66	29.87	8.40	20.30
Oromo	23.89	19.35	5.42	7.15
Somali	26.31	24.56	6.33	18.14
Nepali	32.28	30.81	1.79	23.07
Amharic	36.45	34.61	1.76	11.22
Kirundi	25.55	19.28	7.16	8.80
Tigrinya	39.85	36.34	1.52	16.90
Uzbek	24.21	23.09	3.51	12.39
Burmese	35.79	33.33	1.63	23.66
Korean	32.92	31.03	6.78	23.05
Igbo	30.59	28.57	7.74	16.63
Sinhala	35.75	35.26	2.93	27.88
Kyrgyz	22.61	22.18	4.31	12.24
Scottish-Gaelic	25.10	25.55	6.72	15.21

Table 20: Per language ROUGE-L results on XLSum using full fine-tuning (FT) and LoRA (rank 4), for PaLM 2-XXS in the high-data regime and in a zero-shot cross-lingual transfer setting from English.

PaLM 2-XXS	High-Data		EN Zero-Shot	
	FT	LoRA	FT	LoRA
Average	42.93	51.16	4.49	39.07
English	62.34	62.60	–	–
Hindi	55.00	56.15	2.02	49.50
Urdu	53.06	52.70	0.58	34.95
Russian	50.39	51.32	4.63	51.37
Portuguese	43.72	43.55	15.98	47.45
Persian	61.11	62.91	1.71	61.88
Ukrainian	47.34	50.00	3.35	34.03
Indonesian	57.03	60.26	8.02	61.63
Spanish	43.06	47.82	22.81	57.13
Arabic	46.92	48.78	2.09	44.86
Chinese-Traditional	57.04	60.85	0.80	58.49
Chinese-Simplified	57.13	59.94	1.84	57.74
Vietnamese	55.83	59.04	3.40	54.95
Turkish	46.84	51.02	5.44	44.23
Tamil	61.37	63.26	1.59	42.65
Pashto	45.49	46.87	1.97	24.39
Marathi	45.66	55.50	3.12	42.45
Telugu	43.73	50.49	0.88	33.70
Welsh	41.63	46.40	2.33	35.87
Pidgin	41.96	49.96	32.69	59.13
Gujarati	44.91	52.15	0.46	32.48
French	43.49	54.14	25.62	53.52
Punjabi	33.92	45.35	0.49	26.78
Bengali	51.17	63.23	2.93	47.99
Swahili	38.19	49.10	4.65	40.19
Serbian-Latin	35.07	44.41	4.91	44.72
Serbian-Cyrillic	29.72	42.63	1.12	45.97
Japanese	59.18	63.13	3.44	58.70
Thai	42.09	53.77	1.42	52.30
Azerbaijani	31.69	48.98	4.04	28.95
Hausa	31.37	38.11	2.31	19.91
Yoruba	34.71	42.99	6.02	18.23
Oromo	37.75	57.54	3.70	17.95
Somali	31.43	42.26	4.69	22.16
Nepali	46.71	59.99	0.54	40.40
Amharic	35.94	52.42	0.18	28.72
Kirundi	22.97	28.42	3.45	18.49
Tigrinya	41.04	44.01	0.53	28.01
Uzbek	29.78	46.71	2.14	17.78
Burmese	36.21	54.03	1.00	32.02
Korean	47.31	62.30	2.72	48.97
Igbo	32.43	38.73	3.76	23.81
Sinhala	31.49	57.47	0.17	34.10
Kyrgyz	27.39	47.95	0.70	26.43
Scottish-Gaelic	19.42	32.90	1.46	13.94

Table 21: Per language NLI results on XLSum using full fine-tuning (FT) and LoRA (rank 4), for PaLM 2-XXS in the high-data regime and in a zero-shot cross-lingual transfer setting from English.

PaLM 2-XXS	High-Data		EN Zero-Shot	
	FT	LoRA	FT	LoRA
Average	42.93	51.16	4.49	39.07
English	42.00	42.72	–	–
Hindi	39.78	40.39	5.51	34.22
Urdu	40.96	38.77	4.28	20.58
Russian	43.60	44.32	6.79	41.20
Portuguese	33.19	34.56	12.95	37.87
Persian	44.39	45.43	5.68	42.24
Ukrainian	40.90	43.22	6.29	26.24
Indonesian	41.22	44.43	7.04	40.13
Spanish	33.30	37.06	19.19	45.30
Arabic	36.35	38.57	5.09	32.96
Chinese-Traditional	41.70	43.34	5.09	35.71
Chinese-Simplified	41.68	42.84	5.10	39.12
Vietnamese	35.78	36.42	6.03	28.99
Turkish	46.85	48.85	7.19	44.01
Tamil	38.00	37.06	4.43	21.96
Pashto	29.78	25.35	4.10	5.60
Marathi	33.08	35.48	7.82	25.43
Telugu	25.97	26.26	5.49	10.30
Welsh	29.00	28.98	5.93	17.44
Pidgin	28.82	32.71	22.28	37.42
Gujarati	26.13	27.99	5.56	13.49
French	36.77	48.56	20.38	43.65
Punjabi	22.58	23.84	4.80	10.53
Bengali	38.57	43.09	7.72	32.00
Swahili	31.57	39.15	5.72	28.45
Serbian-Latin	29.87	38.37	9.78	36.45
Serbian-Cyrillic	24.71	32.70	4.92	22.66
Japanese	38.68	43.81	7.77	45.05
Thai	31.35	38.99	5.68	32.21
Azerbaijani	27.07	35.27	5.29	15.41
Hausa	25.95	28.11	4.71	8.14
Yoruba	26.45	26.82	5.50	7.52
Oromo	24.63	22.66	6.16	6.26
Somali	23.87	22.36	5.01	7.59
Nepali	33.83	40.11	5.11	23.67
Amharic	25.44	26.81	4.79	4.67
Kirundi	18.69	14.30	5.89	6.64
Tigrinya	26.61	17.19	4.71	5.19
Uzbek	25.59	30.67	5.17	6.05
Burmese	28.06	35.06	4.92	10.54
Korean	32.52	40.15	5.95	27.97
Igbo	18.25	19.65	5.92	7.27
Sinhala	24.01	34.99	4.62	13.99
Kyrgyz	20.81	29.62	5.22	8.29
Scottish-Gaelic	15.65	22.03	4.95	5.05

Table 22: Per language SEAHORSE results on XLSum using full fine-tuning (FT) and LoRA (rank 4), for PaLM 2-XXS in the high-data regime and in a zero-shot cross-lingual transfer setting from English.