

# Parameter-Efficient Multilingual Summarisation: An Empirical Study

Chenxi Whitehouse<sup>1,2,\*</sup> Fantine Huot<sup>2</sup> Jasmijn Bastings<sup>2</sup>

Mostafa Dehghani<sup>2</sup> Chu-Cheng Lin<sup>3</sup> Mirella Lapata<sup>2</sup>

<sup>1</sup>City, University of London <sup>2</sup>Google DeepMind <sup>3</sup>Google

chenxi.whitehouse@city.ac.uk

{fantinehuot, bastings, dehghani, kitsing, lapata}@google.com

## Abstract

With the increasing prevalence of Large Language Models, traditional full fine-tuning approaches face growing challenges, especially in memory-intensive tasks. This paper investigates the potential of Parameter-Efficient Fine-Tuning, focusing on Low-Rank Adaptation (LoRA), for complex and under-explored multilingual summarisation tasks. We conduct an extensive study across different multilingual summarisation scenarios based on data availability, including full-data, low-data, and cross-lingual transfer, leveraging models of different sizes. Our findings reveal that LoRA lags behind full fine-tuning when trained with full data, however, it excels in low-data scenarios and cross-lingual transfer. Interestingly, as models scale up, the performance gap between LoRA and full fine-tuning diminishes. Additionally, we investigate effective strategies for few-shot cross-lingual transfer, finding that continued LoRA tuning achieves the best performance compared to both full fine-tuning and dynamic composition of language-specific LoRA modules.

## 1 Introduction

The emergence of powerful 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 advances in NLP. However, the continuous expansion of LLM sizes presents a significant challenge to conventional full fine-tuning approaches when applied to downstream tasks, especially those involving extensive memory, such as handling long input text.

Parameter-Efficient Fine-Tuning (PEFT) has gained increasing importance in addressing this issue. Existing PEFT methods typically freeze most of the LLM parameters while tuning only a small number of (additional) parameters, substantially

reducing the computational cost. Widely-adopted PEFT approaches include adapters (Houlsby et al., 2019; Pfeiffer et al., 2021), prefix-tuning (Li and Liang, 2021), prompt-tuning (Lester et al., 2021), and Low-Rank Adaption (LoRA) (Hu et al., 2022). Among these, LoRA achieves state-of-the-art performance without introducing latency at inference time. However, most PEFT studies have focused on classification or monolingual generation tasks, such as the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks, and the E2E NLG Challenge (Puzikov and Gurevych, 2018).

This paper explores the potential of PEFT, particularly LoRA, in a challenging yet under-explored domain: multilingual summarisation. Multilingual summarisation often involves processing lengthy inputs, such as news articles (Hasan et al., 2021), making the effective use of PEFT methods highly advantageous. Nonetheless, this task is complex: models are expected to not only comprehend but also fluently generate sentences in many languages, requiring significant linguistic versatility. Moreover, multilingual tasks frequently face imbalances in language availability and data resources, making it particularly challenging to collect data for very low-resource languages (Parida and Motlicek, 2019). In situations where multilingual training data is limited, full fine-tuning can lead to overfitting or catastrophic forgetting (Kirkpatrick et al., 2017; Mitchell et al., 2022). In such cases, PEFT methods may be more suitable as they only update marginal parameters of LLMs.

This motivates us to study the following research questions: Can LoRA be effectively applied to complex multilingual summarisation tasks? In particular, under what resource availability conditions does LoRA exhibit the most potential? To answer these questions, we explore various scenarios for multilingual summarisation based on data availability: full-data regime, low-data regime, and cross-lingual transfer. The latter has proven effective

\*Work conducted as Research Intern at Google DeepMind.

when data is scarce or unavailable in some languages (Artetxe et al., 2020; K et al., 2020).

We select two multilingual summarisation datasets: XLSum (Hasan et al., 2021) and XWikis (Perez-Beltrachini and Lapata, 2021), and conduct extensive experiments using models of different sizes, including PaLM 2-XXS and PaLM 2-S. A focus is given to the more constrained setup with the smaller XXS model. Specifically, in the cross-lingual transfer scenario, we further experiment with the composition of language-specific LoRA modules, including the recently proposed few-shot LoraHub learning (Huang et al., 2023).

To summarise, our contributions are as follows: (1) We conduct a comprehensive study of the effectiveness of LoRA in multilingual summarisation compared to full fine-tuning in various scenarios. (2) We highlight the benefits of LoRA in low-data and cross-lingual transfer scenarios. (3) We investigate the most promising strategies for achieving zero-shot and few-shot cross-lingual transfer performance, depending on the availability of examples in target languages.

## 2 Related Work

**Parameter Efficient Fine-Tuning** methods focus on enhancing computational efficiency while maintaining competitive performance compared to full fine-tuning. LoRA stands out as one of the most widely adopted PEFT approaches (Hu et al., 2022; Chen et al., 2022), which freezes the pre-trained weights of LLMs and adds trainable low-rank matrices to approximate the parameter update process during full fine-tuning. These low-rank matrices can be merged back into the frozen parameters, without introducing latency during inference. Subsequent work explores the adaptability in LoRA ranks (Zhang et al., 2023b; Valipour et al., 2023), formulating general PEFT approaches (He et al., 2022; Chavan et al., 2023), and proposing further optimisation through methods such as combining LoRA with quantisation (Dettmers et al., 2023).

However, the majority of these studies focus on classification and monolingual generation tasks. In contrast, this paper investigates the application of LoRA to multilingual summarisation, a complex and under-explored domain.

**Cross-lingual Transfer** addresses the data scarcity challenge through training models on high-resource data and applying them to low-resource languages (Artetxe et al., 2020; K et al., 2020;

Lauscher et al., 2020; Whitehouse et al., 2022, 2023a). Various studies in PEFT for cross-lingual transfer have explored adapter-based approaches (Pfeiffer et al., 2020; Ansell et al., 2021), composable sparse fine-tuning (Ansell et al., 2022), among others. Vu et al. (2022) investigate zero-shot cross-lingual transfer with PEFT on a multilingual dataset Wikilingua (Ladhak et al., 2020), however, the study does not cover LoRA.

Our work primarily focuses on LoRA, one of the most effective PEFT methods effective PEFT method that excels in both efficiency and performance (Hu et al., 2022; Chen et al., 2022), and explores its potential for both zero-shot and few-shot cross-lingual transferability in multilingual summarisation.

**Model Composition and Weight Merging** encompasses a broad research area that investigates optimal approaches for combining individually trained modules, which offers benefits such as in model ensembling or multi-task learning, including unseen tasks. Prior work includes composition guided by task similarity (Lv et al., 2023), employing specific arithmetic operations (Zhang et al., 2023a), multi-task modular prompt pre-training (Sun et al., 2023), dataless composition (Jin et al., 2023), to name a few.

This paper explores the composition of language-specific LoRA matrices, including weight averaging and dynamic composition with few-shot LoraHub learning, the latter is achieved through black-box optimisation of the performance on a few examples from the unseen tasks.

## 3 LoRA vs Full Fine-tuning for Multilingual Summarisation

This section first presents the fundamentals of LoRA and few-shot LoraHub learning and then introduces the different scenarios we study for multilingual summarisation tasks.

### 3.1 LoRA and LoraHub

**LoRA** freezes pre-trained weight matrices  $W \in \mathbb{R}^{d \times k}$  in LLMs and adds pairs of trainable rank-decomposition matrices  $B \in \mathbb{R}^{d \times r}$  and  $A \in \mathbb{R}^{r \times k}$ , with  $r$  representing the rank. The adapted weight matrices can then be approximated as  $W + BA$ , which allows effectively updating only a portion of  $2r/d$  of the parameters compared to full fine-tuning, without introducing latency at inference.

Hu et al. (2022) find that LoRA tuning only attention layers achieves competitive results. In our experiments, we also focus on updating attention layers to further boost efficiency.<sup>1</sup>

**LoraHub** is a gradient-free, few-shot learning approach, recently proposed by Huang et al. (2023), focusing on the composability of LoRA modules for cross-task generalisation. LoraHub utilises a collection of available LoRA modules, each fine-tuned for specific tasks, along with a few labelled examples from the target task. It aims to learn the optimal weighted sum of these LoRA modules through black-box gradient-free optimisation (Sun et al., 2022), based on the performance metrics of the target task, such as loss. Results on the Big-Bench Hard benchmark (Suzgun et al., 2022) indicate that LoraHub learning achieves competitive results compared to few-shot In-Context Learning, without the additional complexity of concatenating in-context examples to the input.

### 3.2 Data Regimes

We investigate the effectiveness of LoRA for multilingual summarisation in the following scenarios:

**Full-Data Regime:** This scenario assumes the availability of sufficient training data in the languages of interest. The data could be sourced through automatic pipelines or crowdsourcing in high-resource languages.

**Low-Data Regime:** In this scenario, we examine situations where there are only a limited number of examples, typically in the order of dozens or a few hundred, available in the target languages. These scenarios often involve low-resource languages or data collected via human annotation, which may not be suitable for crowdsourcing.

**Cross-Lingual Transfer:** Within this context, we consider scenarios where training examples are primarily available in high-resource languages, such as English. We further subdivide this category into three cases: (a) Only English training data is available; (b) Training data is available in multiple languages beyond English, which creates a more complex multilingual setting; (c) Additionally, a small number of labelled examples in the target languages are available, allowing us to study few-shot cross-lingual transfer performance.

<sup>1</sup>We note that all four attention matrices, *query*, *key*, *value*, and *out*, are being updated in our LoRA tuning.

Dataset	XLSum	XWikis
Source	BBC News	Wikipedia
Languages	45	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 of the two multilingual summarisation tasks. XWikis contains multi-sentence summaries. Train/Val/Test shows the number of examples in each split. Input/Output shows average number of *words* in the *English* input document and output summary.

## 4 Experimental Setup

This section introduces the datasets and models we study, details of the experiment setup, and the evaluation metrics for the generated summaries.

### 4.1 Datasets

We experiment on the following two multilingual abstractive summarisation datasets:

**XLSum** comprises over one million professionally annotated article-summary pairs sourced from BBC News, covering 45 diverse languages. However, the number of training examples varies significantly among these languages, resulting in notable imbalances.<sup>2</sup> Summaries in XLSum are extracted from the paragraph at the beginning of each news article, typically presented in bold text consisting of one or two sentences.

**XWikis** consists of multi-sentence document-summary pairs. It is constructed using Wikipedia articles, with the assumption that the body of the article and its lead paragraph together form a document-summary pair. XWikis has five languages: 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 specifically focus on cases where the input article and the summary are in the same language.

A summary of the datasets is included in Table 1.

### 4.2 Modelling Details

For our experiments, we focus on PaLM 2 (Anil et al., 2023), a decoder-only LLM that is improved upon PaLM (Chowdhery et al., 2022), achieving superior multilingual and reasoning capabilities, as well as better compute-efficiency. Specifically, we choose the PaLM 2-XXS and S models.

<sup>2</sup>See Appendix A for the language distribution of XLSum.

All the experiments are conducted on cloud TPUs. We search the best learning rate from  $\{1e^{-3}, 2e^{-4}, 2e^{-5}\}$ . The input/output length is truncated at 2048/128 for XLSum and 2048/256 for XWikis.

### 4.3 Evaluation

We evaluate the quality of the generated summaries from two key aspects: summary relevance and summary faithfulness.

In terms of relevance, we employ the widely used ROUGE score (Lin, 2004), which measures the degree of local n-gram overlap between the generated summary and the reference text. Given the variations in word tokenizers for non-English languages, we follow Aharoni et al. (2023) and employ a SentencePiece tokenizer (Kudo and Richardson, 2018) trained on the mC4 dataset (Xue et al., 2021), which is capable of tokenizing all languages in the two datasets we study.

Existing n-gram-based automatic evaluation metrics, including ROUGE, are insufficient for assessing the faithfulness or truthfulness of the generated text. To address this, we use entailment-based metrics, typically utilising the entailment score from Natural Language Inference (NLI) models, which has shown to be an effective approach to determining whether the input sequence supports the content of the output text (Falke et al., 2019; Kumar and Talukdar, 2020; Honovich et al., 2022; Whitehouse et al., 2023b; Huot et al., 2023).

Specifically, we use mT5-XXL (Xue et al., 2021) fine-tuned on two NLI datasets, ANLI (Nie et al., 2020) and XNLI (Conneau et al., 2018), as our entailment classifier. Consistent with prior research (Aharoni et al., 2023; Huot et al., 2023), we segment the summary into individual sentences for a more detailed assessment. The average entailment scores across all sentences are reported.

## 5 Results

This section presents the ROUGE-L and NLI scores of different training regimes on the two datasets.

### 5.1 Full-Data Regime

In the full-data regime, we use the complete training set, including all languages in XLSum and XWikis. For full fine-tuning, we compare tuning on all layers and a more constrained setting that updates solely the attention layers. For LoRA, we experiment with tuning the attention layers with various ranks. We note that for all experiments in

PaLM 2 -XXS	Trainable Layers	Trainable Params.	XLSUM		XWikis	
			R-L	NLI	R-L	NLI
Full FT	All	100%	<b>31.11</b>	42.93	<b>34.08</b>	41.04
	Attention	20%	30.88	50.32	32.22	37.06
LoRA	Attention	$r=512$	29.81	42.58	33.38	40.48
		$r=64$	29.79	45.51	34.04	45.34
		$r=16$	29.77	48.48	33.80	46.10
		$r=4$	29.03	<b>51.16</b>	32.92	<b>47.43</b>

Table 2: Results of Full FT and LoRA with PaLM 2-XXS trained under the full-data regime on XLSum and XWikis. *Trainable Params.* represents the proportion of trainable parameters. The best R-L and NLI scores are highlighted in bold.

this paper, we select the best checkpoints based on the ROUGE-L score.

Table 2 presents the ROUGE-L and NLI entailment scores with PaLM 2-XXS on the two datasets.<sup>3</sup> We can see that the conventional fine-tuning scheme: full fine-tuning on all layers, achieves the best ROUGE-L scores. Updating only attention layers results in competitive performance in XLSum, however, exhibits a decline of 1.86 ROUGE-L in XWikis. All LoRA settings, even those with high ranks like 512, update fewer parameters than the constrained full fine-tuning, with as little as a mere 0.1% parameter update in the case of very low rank settings (rank=4) on attention layers. Despite the efficiency in parameter updates in LoRA settings, noticeable differences in ROUGE-L scores are observed compared to the best-performing approach (2.03 in XLSum and 1.16 in XWikis).

Generally, expanding the parameter update space by increasing the rank enhances summary relevance. For XWikis, the optimal LoRA setting with a rank of 64 approaches very close performance to full fine-tuning. However, for XLSum where language diversity and imbalances are more pronounced, all LoRA experiments fall behind full fine-tuning by more than 1 ROUGE-L score. It is worth noting that with higher ranks, LoRA becomes more sensitive to learning rate, requiring more careful hyper-parameter tuning, consistent with findings in Chen et al. (2022).

Different behaviours can be observed with NLI, where LoRA achieves superior NLI scores than full fine-tuning. Lower rank settings also exhibit better summary faithfulness.

<sup>3</sup>ROUGE-1 and ROUGE-2 scores, results of LoRA on additional Feed-Forward Network (FFN) layers, as well as Language-specific performance are included in Appendix B.



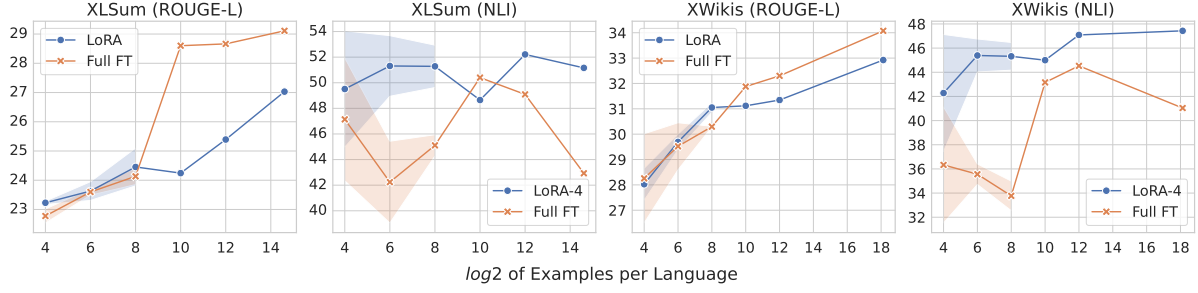


Figure 1: Comparison of ROUGE-L and NLI scores of LoRA (rank=4) and Full FT, when training on low  $\rightarrow$  full data regime with the PaLM 2-XXS model. Results for up to 256 examples per language are averaged over three seeds, with the standard deviation shown in the shaded areas.

**Key Takeaway:** For optimal summary relevancy with the PaLM 2-XXS model in a full-data regime, opt for full fine-tuning when resources allow. LoRA is a competitive choice, especially in summary faithfulness. The performance could be further enhanced with higher ranks though more careful hyper-parameter tuning is generally required.

## 5.2 Low-Data Regime

We proceed to study the performance of full fine-tuning and LoRA in the low-data regime, where the availability of data is limited. From this section onward, we primarily focus on LoRA with rank 4.

We start by randomly sampling 16, 64, and 256 training examples per language for both XLSum and XWikis. To improve the robustness of the results, we conduct experiments with three different seeds, each with a unique set of samples.

Furthermore, it is desirable to see how performance evolves as we increase the number of training data. Therefore, we further include experiments with 1024 and 4096 examples per language for both datasets.<sup>4</sup> We approximate the average examples per language in a full-data regime by averaging the total training data across languages (note that the actual dataset is imbalanced). The number of validation samples is set to be the same as the training data. As before, we select the best checkpoint based on ROUGE-L and subsequently evaluate the entire test set.

Figure 1 shows the ROUGE-L and NLI scores of PaLM 2-XXS on the two datasets with full fine-tuning and LoRA, ranging from low-data of 16 examples per language to the full-data regime. The x-axis shows  $\log_2$  of the number of examples

per language, where the full data is approximated as  $\sim 2^{14.6}$  examples per language for XLSum and  $\sim 2^{18.1}$  for XWikis. For low data up to 256 examples per language, we show the standard deviation with the shaded areas. We observe that LoRA achieves overall better NLI scores than full fine-tuning. For ROUGE-L, LoRA demonstrates advantages in low-data scenarios, while full fine-tuning exhibits a noticeable performance boost when increasing the number of examples from 256 to 1024.

Furthermore, full fine-tuning is sensitive to checkpoint selection in the low-data regime, due to its susceptibility to overfitting. Consequently, it necessitates more frequent validation for optimal checkpoint selection. In comparison, the training process for LoRA is more stable.

**Key Takeaway:** In low-data scenarios, LoRA emerges as a preferable option compared to full fine-tuning. The advantages include efficient and stable training, along with consistently competitive or even superior results.

## 5.3 Cross-lingual Transfer

This section explores zero-shot and few-shot cross-lingual transfer in multilingual summarisation. For LoRA, we focus on rank 4 in all experiments.

### 5.3.1 Zero-shot Transfer from English

For zero-shot transfer, we first consider only English training data is available, a typical scenario for most monolingual datasets. Training and validation are conducted using English examples, and evaluation spans all languages in the test set.

Table 3 presents the ROUGE-L and NLI scores for full fine-tuning and LoRA on the two datasets. The English test scores reveal that full fine-tuning generally outperforms LoRA except for NLI score on English XLSum, aligning with the findings

<sup>4</sup>With the setting of 4096, three languages in XLSum already lack sufficient data, therefore we refrain from selecting more examples per language.

PaLM 2 -XXS	XLSum				XWikis			
	Non-EN		EN		Non-EN		EN	
	R-L	NLI	R-L	NLI	R-L	NLI	R-L	NLI
Full FT	5.20	4.49	<b>32.58</b>	57.09	17.51	35.95	<b>36.59</b>	<b>53.59</b>
LoRA	<b>21.13</b>	<b>39.07</b>	32.21	<b>63.13</b>	<b>23.86</b>	<b>45.54</b>	34.07	49.94

Table 3: ROUGE-L and NLI scores for zero-shot cross-lingual transfer using Full FT and LoRA (rank=4) on the PaLM 2-XXS model. Models are trained and validated on English and tested on all languages. Non-EN shows the average score of all non-English languages.

discussed in §5.1 for the full-data regime. However, when comparing the cross-lingual transfer performance, full fine-tuning performs exceptionally poorly, with a notably low average ROUGE-L score of 5.2 and NLI score of 4.49 across the 44 non-English languages in XLSum. The gap in XWikis is narrower as only four non-languages are covered and all but Chinese are Indo-European languages. Further examination of the model output shows that the generated text remains in English rather than in target languages. The model appears to comprehend input examples in the target languages, however, struggles to generate output accordingly.

In Appendix B, we provide per-language results. These results 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.<sup>5</sup> This catastrophic forgetting behaviour echoes the findings in Vu et al. (2022).

### 5.3.2 Zero-shot: Multiple Languages

We extend the study of zero-shot cross-lingual transfer to scenarios with more training languages.

#### Available Languages.

For XLSum, we curate a set of 10 languages from eight distinct linguistic families, each with substantial training data, to form the pool of available languages. These languages include Arabic: AR, Chinese (Simplified): 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, each repre-

<sup>5</sup>Refer to Table 21 in Anil et al. (2023) regarding the distribution of languages used in the pre-training of PaLM 2.

		UNSEEN									
		AZ	BN	JA	RN	KO	NE	GD	SO	TH	YO
SEEN	AR	15.42	23.38	28.20	10.29	23.78	21.91	16.75	14.94	23.35	19.00
	ZH	14.46	22.11	30.85	8.25	22.33	22.77	16.02	14.40	23.12	16.53
	EN	15.12	22.24	28.91	8.90	23.09	23.43	15.54	18.30	22.23	20.85
	HA	15.67	22.26	27.49	10.59	21.90	22.17	16.20	18.09	20.47	19.40
	HI	13.60	22.71	28.81	9.75	21.31	24.96	18.13	12.90	22.54	19.30
	ID	17.07	23.91	29.41	10.47	24.82	23.64	20.66	19.26	22.94	19.51
	FA	10.66	22.15	27.59	10.19	20.77	20.68	16.26	15.86	22.28	17.62
	PT	15.05	22.32	28.13	7.82	22.84	22.27	16.78	15.26	21.34	18.52
	SW	17.10	22.69	28.67	11.87	24.37	24.84	18.18	18.74	21.42	19.49
	TR	12.16	21.46	27.49	9.79	20.30	20.23	16.78	15.67	21.71	18.44
	Ave.LoRA	18.19	23.16	29.77	16.08	24.93	24.64	22.28	21.31	23.47	23.11
	LoRA	19.94	26.25	32.15	10.23	26.26	27.38	19.16	20.26	25.37	18.87
	Full FT	15.89	5.97	22.61	13.17	8.45	21.72	17.92	12.15	13.17	13.75

(a) ROUGE-L Scores for the 10 Test Languages in XLSum.

		UNSEEN				
		CZ	DE	EN	FR	ZH
SEEN	CZ		20.53	19.41	16.15	12.25
	DE	27.13		30.69	29.21	18.43
	EN	26.11	27.89		27.65	13.77
	FR	26.93	29.97	28.39		17.19
	ZH	25.01	24.19	25.14	26.27	
	Ave.LoRA-excl.XX	26.07	29.33	32.87	28.96	20.00
	LoRA-excl.XX	28.55	31.57	32.93	30.27	18.99
	Full FT-excl.XX	17.16	25.31	23.47	22.60	12.56

(b) ROUGE-L Scores for the Five Languages in XWikis.

Figure 2: ROUGE-L scores for zero-shot cross-lingual transfer in XLSum (top) and XWikis (bottom) with PaLM 2-XXS using language-specific LoRA, weighted-average LoRA (Ave.LoRA), and LoRA and Full FT models trained on all selected languages. excl.XX in XWikis denotes *leave-one-out* training, excluding the test language.

sending a unique language family to offer a diverse range of linguistic characteristics.<sup>6</sup>

For XWikis, as the dataset only covers five languages, we adopt a *leave-one-out* cross-validation approach. This entails rotating through the available languages, using four for training and reserving one for testing.

#### Training Language-specific LoRA Modules.

In addition to full fine-tuning and LoRA tuning models on all selected available languages together, we experiment with training language-specific LoRA modules, where we only include examples from one language.

An advantage of training language-specific LoRA lies in its scalability and adaptability. When additional languages become available, there is no need to re-train the entire model; rather, the addi-

<sup>6</sup>See Appendix A for details of language families in XLSum.

tion of a new module specific to the new language suffices. During the inference phase, we can also flexibly experiment with various LoRA modules or weight composition methods.

As outlined in §2, weight composition is an active research area that has demonstrated effectiveness across a spectrum of applications. One straightforward option involves computing the weighted average of all available modules.

We illustrate the ROUGE-L scores across the selected test languages in both datasets for different source models in the heatmaps in Figure 2, with rows representing the source models from SEEN languages. The colour scale is column-wise normalised to provide a comparative view of performance concerning the best and worst source models for each UNSEEN test language.

Several **key observations** can be drawn: (1) As indicated by the last three rows in each heatmap, even with a diverse collection of available languages beyond English, full fine-tuning consistently lags behind LoRA in zero-shot cross-lingual transfer. (2) Ave. LoRA (weighted average of language-specific LoRA modules) and LoRA (trained on all available languages together) benefit different unseen languages. Particularly in XLSum, lower-resource languages, RN, GD, SO, and YO, exhibit superior performance with language-specific LoRA training. (3) Languages with similarities demonstrate better transferability, exemplified by cases like ZH to JA and SW to RN in XLSum, as well as the European languages in XWikis.

### 5.3.3 Few-shot Cross-lingual Transfer

In the last scenario, we consider situations where some examples in the target languages are available and explore effective strategies for utilising them. We start with models trained on languages with sufficient training data, as discussed in §5.3.2.

One approach is to *continue* training these models using target language examples, following the same fine-tuning methodology, i.e. if the starting checkpoint was obtained from full fine-tuning on the available languages, we continue with full fine-tuning, and the same applies to LoRA. Another widely-used technique is few-shot *In-Context Learning* (ICL), where input and output examples are concatenated to form in-context demonstrations. While ICL has shown promising results in many LLMs applications (Brown et al., 2020; Wei et al., 2022), it becomes less practical for multilingual summarisation that handles long article inputs, es-

PaLM 2-XXS		XLSUM	XWIKIS
<b>ZERO-SHOT</b>	Full FT	14.48	<b>28.46</b>
	LoRA	22.59	18.87
	Ave. LoRA	<b>22.69</b>	27.45
<b>16-SHOT</b>	Full FT + <i>continued learning</i>	22.31	26.90
	LoRA + <i>continued learning</i>	<b>24.89</b>	<b>30.05</b>
	Ave. LoRA + <i>LoraHub</i>	23.30	25.67
<b>64-SHOT</b>	Full FT + <i>continued learning</i>	24.30	28.64
	LoRA + <i>continued learning</i>	<b>25.94</b>	<b>31.08</b>
	Ave. LoRA + <i>LoraHub</i>	23.34	25.95

Table 4: Average ROUGE-L scores for cross-lingual transfer on the 10 test languages in XLSum and *leave-one-out* cross-validation in XWikis using the PaLM 2-XXS model. The 16- and 64-shot experiments show the average results from three different seed runs. For *continued learning*, we use a 14-2 and 60-4 split for train-validation examples. We include the per-language results in Table 10 and Table 11 in Appendix B.

pecially as the number of examples increases.

Our study instead experiments with the recently proposed few-shot LoraHub learning approach (§3.1). We note that the original LoraHub setup from Huang et al. (2023) significantly differs from ours. Their experiments do not assume any prior knowledge of the available LoRA modules: the candidate LoRA modules are randomly sampled without pre-filtering and initialised with zero weights (i.e. starting from a general-purpose pre-trained LLM). In contrast, we initialise LoraHub learning with the weighted average of available language-specific LoRA modules, as all of these modules are fine-tuned on the same task in different languages, offering a stronger baseline compared to pre-trained LLM.

We consider two budget constraints, with 16 or 64 examples in the target languages, simulating practical scenarios where human annotators or experts craft 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 the average scores are reported. For the *continued learning* approach, we also employ random splits for training and validation, using 14-2 and 60-4 splits. For LoraHub learning, we use the Nevergrad toolkit<sup>7</sup> for black-box optimisation, and conduct various experiments comparing utilising ROUGE-L score, loss, and their combination as the performance metric

<sup>7</sup><https://facebookresearch.github.io/nevergrad>

guiding the optimisation, finding that more stable results are achieved based on loss.

Table 4 presents the 16-shot and 64-shot ROUGE-L scores in XLSum and XWikis, averaged across all test languages. Zero-shot results are also included for comparison.

Our analysis reveals several **key observations**: (1) With only a few, e.g. 16 examples in the target languages, cross-lingual transfer via full fine-tuning sees a remarkable improvement, resulting in an average of 7.8 ROUGE-L score increase in XLSum and 6.7 in XWikis, echoing the observation in Lauscher et al. (2020). (2) LoraHub learning demonstrates the potential to enhance performance compared to a weighted-average baseline (+0.6 ROUGE-L score for XLSum) with a small number of examples. However, increasing the number of examples to 64 does not lead to further improvements. (3) LoRA continued learning consistently outperforms other approaches across the few-shot learning scenarios studied.

**Key Take-away:** In diverse cross-lingual transfer situations, LoRA consistently delivers superior performance compared to full fine-tuning. LoRA continued learning demonstrates the best potential 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 full-data regime and zero-shot cross-lingual transfer using English data. The ROUGE-L scores for both experiments are included in Table 5 and Table 6.

We observe an intriguing trend where both LoRA and full fine-tuning achieve remarkably similar performance with the larger model. LoRA marginally outperforms full fine-tuning in cross-lingual transfer but falls slightly behind in other aspects. However, these distinctions remain minimal.

We hypothesise that when using the larger parameter model PaLM 2-S, the lower percentage of trainable parameters in LoRA (only 0.04% of the parameters in the Attention and FFN layers) is compensated by the increased model capacity, allowing LoRA to benefit from full-data regime training. At the same time, the model maintains higher robustness in addressing catastrophic forgetting during full fine-tuning. Consequently, we observe competitive zero-shot cross-lingual results in full fine-tuning compared to LoRA.

PaLM 2-S	Trainable Parameters	XLSUM	XWIKIS
Full FT	100%	36.99	39.65
LoRA	0.04%	36.29	39.25

Table 5: ROUGE-L scores for XLSum and XWikis in the full-data regime, with Full FT and LoRA (rank=4) on PaLM 2-S.

PaLM 2-S	XLSUM		XWIKIS	
	Non-EN (ave.)	EN	Non-EN (ave.)	EN
Full FT	33.22	40.38	35.70	42.03
LoRA	33.31	39.61	36.00	41.53

Table 6: ROUGE-L scores for zero-shot cross-lingual transfer using Full FT and LoRA (rank=4) on PaLM 2-S. Models are trained and validated on English and tested on all languages.

**Key Take-away:** For larger models such as PaLM 2-S, LoRA achieves very close performance compared to full fine-tuning. LoRA is a better choice when considering computational efficiency.

## 7 Conclusions

This paper explores the effectiveness of LoRA on multilingual summarisation across a diverse range of scenarios. Given the inherent computational efficiency in LoRA, as only marginal parameters are updated, we summarise our key findings based on the performance compared to full fine-tuning.

**Superior Performance:** LoRA particularly in zero-shot and few-shot cross-lingual transfer scenarios, as well as in low-data regime training with fewer than e.g. 1K data points. This is most pronounced in smaller models, for example in the scale of PaLM 2-XXS. Specifically, for few-shot learning, LoRA continued learning shows the best performance compared to LoraHub learning. LoRA also achieves overall superior summary faithfulness across various scenarios. **On-par Performance:** For larger models like PaLM 2-S, LoRA exhibits highly competitive results in comparison to full fine-tuning. **Worse Performance:** Notably, for smaller models like PaLM 2-XXS, LoRA shows limited capacity in summary relevance under the full-data regime for multilingual summarisation.

These observations are encouraging for future research on PEFT methods for complex multilingual tasks and cross-lingual transfer. This may extend to more diverse LLMs and broader multilingual generation tasks, beyond summarisation.



## Limitations

We identify the following limitations in this work: (1) Focus on decoder-only models: Our research primarily experiments with decoder-only models. Future work could explore more LLMs including encoder-decoder models for complementary insights. We anticipate that the observation gained from decoder-only models could largely align with those from encoder-decoder models, thus expanding the scope of applicability. (2) Fixed rank setting in cross-lingual transfer studies: In our cross-lingual transfer studies, we only considered a rank of 4. Expanding to additional LoRA settings would enable a more thorough comparison. (3) Focus on summarisation: Our experiments are restricted to multilingual summarisation tasks. Extending the study to cover a broader 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

- Roei 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, et al. 2022. [Palm: Scaling language modeling with pathways](#). *arXiv preprint arXiv:2204.02311*.
- 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, Roei 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*.
- Karthikeyan K, 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.
- Yevgeniy Puzikov and Iryna Gurevych. 2018. [E2E NLG challenge: Neural models vs. templates](#). In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 463–471, Tilburg University, The Netherlands. 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.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. [Challenging big-bench tasks and whether chain-of-thought can solve them](#). *arXiv preprint arXiv:2210.09261*.
- 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, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*.
- Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023a. [Llm-powered data augmentation for enhanced crosslingual performance](#). *arXiv preprint arXiv:2305.14288*.
- 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

We include the language distribution of the two datasets in [Table 7](#) and [Table 8](#).

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 7: Language distribution of XWikis. # Train shows the number of training examples per language.

## B Additional Results

[Table 9](#) shows the ROUGE-1, ROUGE-2, and ROUGE-L scores for LoRA and full fine-tuning with PaLM 2-XXS on the two datasets.

[Table 10](#) [Table 11](#) shows the per-language result of few-shot learning results in XWikis.

[Table 12](#) and [Table 13](#) show the ROUGE-L score per language for full-data and zero-shot cross-lingual transfer (from English) in the two datasets with PaLM 2-XXS.

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 distribution of XLSum. # Train shows the number of training examples per language.



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

Table 9: ROUGE-L, ROUGE-1, and ROUGE-2 scores of Full FT and LoRA with PaLM 2-XXS trained under the full-data regime on XLSum and XWikis. Attention + FFN shows LoRA tuning on both attention and FFN layers.

PaLM 2-XXS		AVE	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	22.59	<b>19.94</b>	<b>26.25</b>	<b>32.15</b>	10.23	<b>26.26</b>	<b>27.38</b>	19.16	20.26	<b>25.37</b>	18.87
	Ave. LoRA	<b>22.69</b>	18.19	23.16	29.77	<b>16.08</b>	24.93	24.64	<b>22.28</b>	<b>21.31</b>	23.47	<b>23.11</b>
16-SHOT	Full FT + <i>continue learning</i>	22.31	16.64	22.95	28.28	17.02	24.31	26.94	19.56	19.66	23.58	24.18
	LoRA + <i>continue learning</i>	<b>24.89</b>	<b>20.74</b>	<b>26.19</b>	<b>32.26</b>	<b>17.82</b>	<b>27.13</b>	<b>27.82</b>	<b>23.00</b>	<b>22.26</b>	<b>24.30</b>	<b>27.44</b>
	Ave. LoRA + <i>LoraHub</i>	23.30	19.59	24.36	30.37	16.63	25.91	26.90	22.25	20.70	23.80	22.53
64-SHOT	Full FT + <i>continue learning</i>	24.30	17.86	22.80	32.49	<b>19.28</b>	27.09	28.89	21.72	22.17	23.90	26.83
	LoRA + <i>continue learning</i>	<b>25.94</b>	<b>20.91</b>	<b>26.08</b>	<b>33.10</b>	19.09	<b>28.43</b>	<b>29.38</b>	<b>25.78</b>	<b>23.06</b>	<b>25.48</b>	<b>28.06</b>
	Ave. LoRA + <i>LoraHub</i>	23.34	19.90	25.34	30.40	16.38	26.15	26.48	21.71	20.67	23.93	22.44

Table 10: ROUGE-L scores for cross-lingual transfer on the 10 test languages in XLSum using the PaLM 2-XXS model. The 16-shot and 64-shot experiments show the average results obtained across three different seed runs. For *continue learning*, we use a 14-2 and 60-4 split for train/validation examples.

PaLM 2-XXS		AVE	CZ	DE	EN	FR	ZH
ZERO-SHOT	Full FT	20.22	17.16	25.31	23.47	22.60	12.56
	LoRA	<b>28.46</b>	<b>28.55</b>	<b>31.57</b>	<b>32.93</b>	<b>30.27</b>	18.99
	Ave. LoRA	27.45	26.07	29.33	32.87	28.96	<b>20.00</b>
16-SHOT	Full FT + <i>continue learning</i>	26.90	22.53	29.23	30.50	26.16	26.11
	LoRA + <i>continue learning</i>	<b>30.05</b>	<b>27.68</b>	<b>33.76</b>	<b>31.98</b>	<b>30.12</b>	<b>26.70</b>
	Ave. LoRA + <i>LoraHub</i>	25.67	26.36	29.64	29.47	29.17	13.70
64-SHOT	Full FT + <i>continue learning</i>	28.64	26.45	30.17	31.79	28.86	25.95
	LoRA + <i>continue learning</i>	<b>31.08</b>	<b>28.97</b>	<b>34.09</b>	<b>33.11</b>	<b>30.99</b>	<b>28.24</b>
	Ave. LoRA + <i>LoraHub</i>	25.95	25.98	29.63	30.71	29.37	14.08

Table 11: ROUGE-L scores for cross-lingual transfer on XWikis using *leave-one-out* cross-validation with the PaLM 2-XXS model. The 16-shot and 64-shot experiments show the average results obtained across three different seed runs. For *continue learning*, we use a 14-2 and 60-4 split for train/validation examples.

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

Table 12: Per language ROUGE-L scores of XWikis with PaLM 2-XXS. We compare full fine-tuning and LoRA (rank=4), each on full-data regime and zero-shot cross-lingual transfer from English.

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

Table 13: Per language ROUGE-L scores of XLSum with PaLM 2-XXS. We compare full fine-tuning and LoRA (rank=4), each on full-data regime and zero-shot cross-lingual transfer from English.