Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation

Tu Vu[1,2]★, **Aditya Barua**[1], **Brian Lester**[1], **Daniel Cer**[1], **Mohit Iyyer**[2], **Noah Constant**[1]
Google Research[1]
University of Massachusetts Amherst[2]
{ttvu, adityabarua, brianlester, cer, nconstant}@google.com
{tuvu, miyyer}@cs.umass.edu

Abstract

In this paper, we explore the challenging problem of performing a generative task in a target language when labeled data is only available in English, using summarization as a case study. We assume a strict setting with no access to parallel data or machine translation and find that common transfer learning approaches struggle in this setting, as a generative multilingual model fine-tuned purely on English catastrophically forgets how to generate non-English. Given the recent rise of parameter-efficient adaptation techniques, we conduct the first investigation into how one such method, prompt tuning (Lester et al., 2021), can overcome catastrophic forgetting to enable zero-shot cross-lingual generation. Our experiments show that parameter-efficient prompt tuning provides gains over standard fine-tuning when transferring between less-related languages, e.g., from English to Thai. However, a significant gap still remains between these methods and fully-supervised baselines. To improve cross-lingual transfer further, we explore several approaches, including: (1) mixing in unlabeled multilingual data, and (2) explicitly factoring prompts into recombinable language and task components. Our approaches can provide further quality gains, suggesting that robust zero-shot cross-lingual generation is within reach.

1 Introduction

Cross-lingual language understanding is an important area of ongoing research (Conneau et al., 2020; Hu et al., 2020; Ruder et al., 2021). With vastly differing amounts of data (both labeled and unlabeled) available across languages, there is significant value to developing techniques that can transfer knowledge from higher-resource languages to improve performance in lower-resource languages. *Zero-shot* cross-lingual benchmarks push on the limiting case where no labeled data is available in the target language. Remarkable progress has been made on zero-shot cross-lingual tasks by scaling up the size of pre-trained multilingual models (Conneau et al., 2020; Xue et al., 2021). However, prior work has focused nearly exclusively on *non-generative tasks* (e.g., classification, extractive question answering, and sequence labeling).

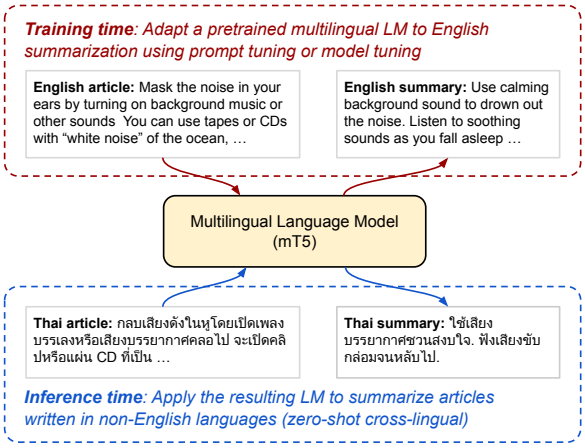

Figure 1: A demonstration of WIKILINGUA-0, a challenging zero-shot cross-lingual generation (XGEN) task, which requires a model to learn a generative task from labeled data in one language (i.e., English), and then perform the equivalent task in another language at inference time.

In this paper, we turn our attention to zero-shot cross-lingual *generation*, or "XGEN", which requires a model to learn a generative task from labeled data in one language (typically English), and then perform the equivalent generative task in another language. This problem is particularly challenging because generative models trained on one language are known to exhibit catastrophic forgetting, losing the ability to generate coherent text in other languages (Xue et al., 2021; Maurya et al., 2021; Shakeri et al., 2021). In particular, we focus on the relatively under-explored task of zero-shot cross-lingual summarization. We construct a new

★ Work done as a student researcher at Google Brain.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9279–9300
December 7-11, 2022 ©2022 Association for Computational Linguistics

zero-shot task WIKILINGUA-0 from the WIKILINGUA dataset (Ladhak et al., 2020), allowing us to test XGEN capabilities across 18 languages. We motivate a new evaluation metric for our task, SP-ROUGE, and show that it correlates well with human judgments of summary quality.

Maurya et al. (2021) show improved performance on XGEN tasks by freezing model parameters in the input and output layers during fine-tuning. Inspired by recent parameter-efficient adaptation techniques (Houlsby et al., 2019; Zaken et al., 2021; Li and Liang, 2021; Lester et al., 2021), we take this approach further: can we overcome catastrophic forgetting by freezing *all* of the pre-trained model parameters, and only tuning a much smaller set of task-specific parameters? Parameter-efficient tuning methods are particularly appealing for multilingual NLP, as they would enable reuse of a single frozen model across many combinations of task and language, reducing storage and serving costs.

To this end, we conduct the first investigation of the XGEN performance of PROMPTTUNING (Lester et al., 2021), a simple parameter-efficient adaptation technique that limits learned parameters to a set of virtual tokens prepended to the text input. We compare PROMPTTUNING with standard fine-tuning (or MODELTUNING, where all model weights are tuned) across different languages and model scales. We find that increasing model size and decreasing tunable parameter capacity are key for overcoming catastrophic forgetting. Despite its inferior performance on the training language (English), PROMPTTUNING with scale typically outperforms MODELTUNING when evaluated on non-English languages, especially on languages more distantly related to English, such as Thai. This corroborates previous findings (Li and Liang, 2021; Lester et al., 2021) that parameter-efficient methods are more robust to domain shifts between training and inference.

Motivated by our initial findings, we investigate two approaches to further improve the XGEN performance of PROMPTTUNING and MODELTUNING. Our first approach involves mixing unlabeled data in the target language into the supervised training stage. We show this dramatically alleviates catastrophic forgetting on WIKILINGUA-0. We also introduce a novel approach, "factorized prompts", which is specifically designed for PROMPTTUNING. We train prompts on a multi-task multilingual mixture, where each prompt is factorized into composable language and task modules—the first half of the prompt encodes language knowledge, while the second half captures language-agnostic task knowledge. During inference in the zero-shot cross-lingual setting, the source language module is replaced with the target language module, while the task module remains unchanged. We demonstrate that factorized prompts provide an effective means of improving XGEN performance.

To summarize, our main contributions are:

- We present the first large-scale empirical investigation of parameter-efficient PROMPTTUNING and standard MODELTUNING for zero-shot cross-lingual generation (XGEN). We show that increasing model scale and decreasing tunable parameter capacity are key for overcoming catastrophic forgetting on XGEN.
- We propose WIKILINGUA-0, a challenging XGEN benchmark and an associated SP-ROUGE evaluation metric, which we hope will facilitate future work evaluating multilingual summarization.
- We show that mixing in unsupervised multilingual data can boost XGEN performance, and are the first to combine this approach with PROMPTTUNING.
- We propose "factorized prompts", a novel approach that can also help PROMPTTUNING overcome severe catastrophic forgetting.
- To facilitate future work, we release our data, pretrained models, and code at: https://github.com/google-research/prompt-tuning/tree/main/prompt_tuning/x_gen.

2 Challenge of zero-shot cross-lingual generation

Much recent progress in multilingual NLP has been driven by zero-shot cross-lingual benchmarks that require a model to perform classification (Conneau et al., 2018; Yang et al., 2019), extractive QA (Artetxe et al., 2020; Lewis et al., 2020; Clark et al., 2020), or sequence labeling (Pan et al., 2017).[1] Here, we are interested in a more challenging task of zero-shot cross-lingual generation

[1] We refer the interested reader to Appendix A for a comprehensive comparison of MODELTUNING and PROMPTTUNING on these benchmarks. Overall, we find that MODELTUNING typically performs better than PROMPTTUNING, although PROMPTTUNING at scale matches the performance of MODELTUNING on English and can yield better results on some languages.

(XGEN) where a model is trained on a generative task in one language (typically English), and then asked to perform the equivalent task in another language during inference. We construct a novel zero-shot cross-lingual summarization task and show that state-of-the-art text-to-text models adapted using MODELTUNING and PROMPTTUNING techniques are not able to successfully perform our task. Our analysis reveals that both techniques suffer from catastrophic forgetting, causing them to often generate text in the wrong language.

2.1 Problem formulation

Defining WIKILINGUA-0 zero-shot cross-lingual summarization: We leverage the WIKILINGUA dataset (Ladhak et al., 2020; Gehrmann et al., 2021) to create a novel zero-shot cross-lingual summarization task, which we dub WIKILINGUA-0.[2] While WIKILINGUA provides labeled training data in 18 languages (including English), we are interested in a more realistic experimental setup where no training data is provided in non-English languages, as it is less practical to obtain labeled data for real low-resource languages.[3] As such, we discard all training data for non-English languages, with the exception of ablation experiments, and cast WIKILINGUA as training a model with English summarization data and feeding it non-English articles during zero-shot evaluation.[4]

Defining SP-RG for multilingual summarization evaluation: ROUGE (Lin, 2004) has been the metric of choice for evaluating summarization systems. However, it assumes that the input text uses spaces to separate words, which is not the case for many languages (e.g., Chinese, Japanese, and Thai).[5] One possible solution is to use language-specific tokenizers, as done in Conneau and Lample (2019). To avoid language-specific preprocessing, we use SentencePiece sub-word tokenization (Kudo and Richardson, 2018), which is data-driven and language independent.[6] We call our metric SP-ROUGE (SentencePiece-based ROUGE) or SP-RG for short, and report SP-RG-LSUM in our experiments.[7] In Appendix B, we demonstrate that SP-ROUGE produces a similar correlation to human judgments as BLEURT (Sellam et al., 2020) while being significantly more computationally efficient.

2.2 Experimental setup

2.2.1 Baselines

In addition to vanilla MODELTUNING and PROMPTTUNING, we consider the following baselines:

LEAD-64: This baseline simply copies the first 64 SentencePiece tokens from the input article.[8]

TRANS-TRAIN: We perform MODELTUNING or PROMPTTUNING on WIKILINGUA-0 English summarization data that is translated into the target language using GOOGLE TRANSLATE.

TRANS-TEST: We train on English summarization data and evaluate on validation data that is translated from the target language to English.

SUP & SUP-ALL: To ablate the impact of using the labeled training data provided in the original WIKILINGUA dataset for all languages, we either train on supervised data for each individual target language (SUP) or a mixture of supervised data from all languages (SUP-ALL).[9]

2.2.2 Training and implementation details

We perform MODELTUNING and PROMPTTUNING on top of pretrained mT5 checkpoints (Xue et al., 2021) of all sizes: SMALL, BASE, LARGE, XL, XXL,[10] using T5X (Roberts et al., 2022). For PROMPTTUNING, we create an LM adapted version of these checkpoints by further training them for 100K steps with the "prefix LM" objective (Raffel et al., 2020) using mC4 (Xue et al., 2021) data for all languages.[11] Except for ablations, we use 100 prompt tokens and initialize the prompt by sampling from the vocabulary embeddings. Training inputs and targets are clipped to 1024 and 512 SentencePiece

[2] Note that the original WIKILINGUA task is not suitable for direct use in our XGEN setting, as it aims to generate English summaries from non-English articles.

[3] While one might rely on machine translation (MT) to obtain labeled data in a language of interest, this is not particularly appealing due to: (i) extra computation required, (ii) varied translation quality across languages (Ruder et al., 2021), (iii) potential loss of discourse structure (Li et al., 2014), and (iv) limited understanding of black box MT systems.

[4] See Ladhak et al. (2020) for data statistics.

[5] In preliminary experiments, we found that standard ROUGE yielded extremely poor ROUGE scores in many languages, despite systems producing reasonably good summaries.

[6] Goyal et al. (2021) also use a similar approach for BLEU (Papineni et al., 2002).

[7] ROUGE-LSUM is the summary-level ROUGE-L metric used in See et al. (2017).

[8] In our preliminary experiments, $n = 64$ performed best among a range of values $\{32, 64, 128, 256\}$.

[9] This is an upper bound and is not in the XGEN setting.

[10] These are 300M, 580M, 1.2B, 3.7B, and 13B parameters.

[11] A similar approach was used in Lester et al. (2021) for PROMPTTUNING with T5.

tokens, respectively. We always train for 100,000 steps for both MODELTUNING and PROMPTTUNING. We save a checkpoint every 5,000 steps and report results on the model checkpoint corresponding to the highest performance on a target language using 250 validation examples for all languages.[12]

2.3 Results and Discussion

WIKILINGUA-0 is challenging for both MODELTUNING and PROMPTTUNING: Our zero-shot evaluation results on WIKILINGUA-0 for French (**FR**), Vietnamese (**VI**), Russian (**RU**), and Thai (**TH**) are shown in Figure 2a.[13] For comparison, we also include results on English. Overall, we find that zero-shot inference on an unseen language leads to a substantial performance drop for both model adaptation techniques, especially when feeding in articles in non-Latin script languages like Russian and Thai. Consistent with the findings in An et al. (2022) for other generative tasks, we find that PROMPTTUNING, even with scale, falls far below MODELTUNING on monolingual English summarization.[14]

PROMPTTUNING is better on larger language shifts: Interestingly, PROMPTTUNING is competitive with or out-performs MODELTUNING when evaluated on other languages. For instance, at the XXL scale, PROMPTTUNING outperforms MODELTUNING by a large margin of +7.3 SP-ROUGE (37.4 vs. 30.1) on Thai. A closer look at these results reveals an interesting pattern: as model size increases, PROMPTTUNING usually produces better results than MODELTUNING when there is a significant language shift at inference time (e.g., from English to a non-Latin script language).[15] This corroborates the view in Lester et al. (2021) that MODELTUNING may be over-parameterized and thus more prone to overfit the training task and less robust to domain shifts.

Both MODELTUNING and PROMPTTUNING suffer from catastrophic forgetting and this effect is more pronounced for MODELTUNING: When performing zero-shot evaluation on non-English languages, we discover that both MODELTUNING and PROMPTTUNING often partially summarize non-English articles into English instead of the target language. This suggests that they suffer from overfitting on the training task. To probe more deeply into this problem, we evaluate performance for each saved checkpoint, and additionally measure: (i) LID_{lang}—the average confidence score given by `cld3`[16] when detecting the language $lang$, and (ii) ASCII—the average percentage of ASCII characters present in the model's predictions, with a higher value indicating a larger amount of English in the model's output for non-Latin script languages. Figure 3 shows our evaluation results as training progresses. For PROMPTTUNING, we observe a clear "deteriorating" trend, where the longer the prompt is tuned on English, the more unwanted English is generated, and the lower summarization quality becomes for Russian and Thai. For MODELTUNING, even by the first checkpoint, the model has already heavily overfit to English, outputting >60% ASCII for Russian and Thai inputs. There is a modest recovery later in training, but quality as measured by SP-ROUGE remains low.

Bigger models are less prone to forget: In Figure 2b, we observe that moving to larger model sizes mitigates catastrophic forgetting to a large extent. This is true both for MODELTUNING (in line with the findings of Xue et al. (2021)), as well as for PROMPTTUNING. For example, at SMALL size, MODELTUNING and PROMPTTUNING only successfully generate Russian text 0.0% and 10.1% of the time respectively, whereas at XXL size, these numbers jump to 57.5% and 84.4%.

Too much capacity is harmful: Figure 2c shows an interesting "paradox of capacity" with regard to the prompt length for PROMPTTUNING. On the one hand, greater capacity (in the form of longer prompts) clearly helps to better learn the summarization task. On the other hand, the greater the capacity to learn from English training data, the more the model forgets other languages. We observe that at the beginning of training, the little amount of English introduced in generated outputs is eclipsed by the improvement in summarization quality, which results in a better SP-ROUGE score. As training continues, however, the increased capacity becomes harmful as more and more English is introduced in the model's output, which domi-

[12] For inference, we use beam search with a beam size of 4 and a length penalty of $\alpha = 0.6$. To avoid severe penalties for predictions that repeat a phrase indefinitely, we heuristically remove all but one occurrence of any prediction-final repeated substring.

[13] See Table 10 in Appendix C for full results (including variance statistics) and Table 8 in Appendix A for results across all target languages.

[14] This is somewhat surprising since across the other tasks we tried above, PROMPTTUNING at XXL can match the performance of MODELTUNING when evaluated on English.

[15] With the exception of a few languages (e.g., Chinese).

[16] https://github.com/google/cld3

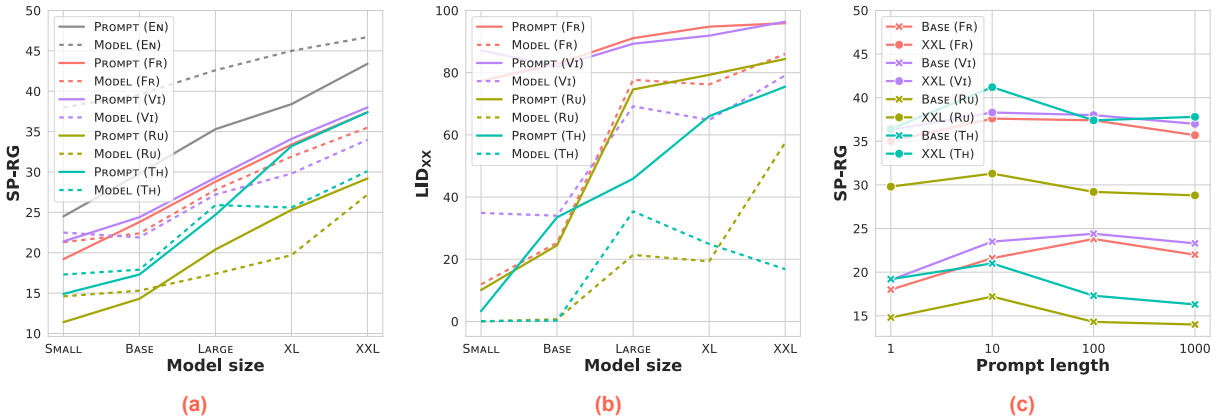

Figure 2: **(a)** Zero-shot XGEN summarization quality (SP-RG) and **(b)** target language accuracy (LID_{XX}) of PROMPTTUNING and MODELTUNING models across five model sizes and four target languages: French (FR), Vietnamese (VI), Russian (RU), and Thai (TH). English (EN) performance is provided as a point of comparison, but is no longer a zero-shot task. **(c)** The effect of prompt length on PROMPTTUNING performance at BASE and XXL model sizes.

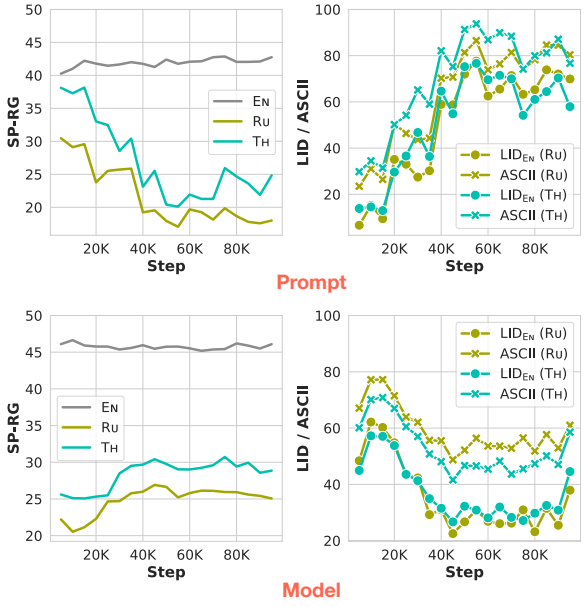

Figure 3: Learning curves showing how PROMPTTUNING (top) and MODELTUNING (bottom) progress in terms of summarization quality (left) and unwanted English output (right), at the XXL model size. Note, MODELTUNING quality is lower overall, and predictions contain high (>40%) levels of unwanted ASCII.

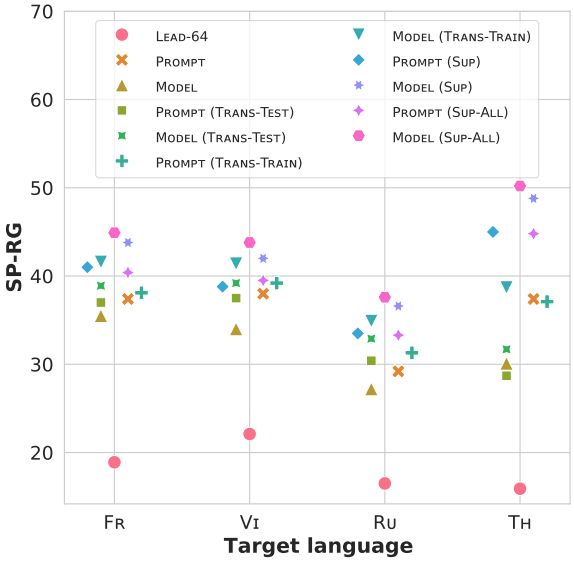

Figure 4: SP-ROUGE scores of our baselines (LEAD-64, PROMPTTUNING, MODELTUNING) at the XXL model size, in the zero-shot XGEN setting. For comparison, we also show the headroom available if a machine translation system is used (TRANS-TRAIN, TRANS-TEST), or if gold data in target languages is used (SUP, SUP-ALL).

nates the improvement in summarization quality and leads to lower SP-ROUGE. For each language and model size, we observe a critical point past which adding extra capacity becomes harmful. For instance, in Thai at the XXL size, increasing capacity from 1 to 10 prompt tokens improves summarization quality (SP-ROUGE +4.8) despite a drop in language accuracy (LID_{TH} −8.0), and increasing capacity further to 100 tokens hurts both metrics.

Significant headroom remains: The supervised baselines in Figure 4 highlight that significant headroom remains on this XGEN task. When tuning the XXL model directly on supervised training data in all languages, SP-ROUGE scores are between +5.8 (VI) and +12.8 points (TH) higher than our highest zero-shot results. We also note that for some languages, like Thai, the supervised baseline greatly exceeds any approach using machine translation. This highlights that machine translation quality is still low in some languages, so pursuing stronger zero-shot solutions is worthwhile.

3 Mitigating catastrophic forgetting

We have seen that increasing model scale and decreasing tunable parameter capacity are both effective in improving XGEN performance. Can we obtain further gains by devising methods that explicitly tackle catastrophic forgetting? Here, we investigate two approaches: mixing unlabeled training data with English supervised data, and factorizing the learned prompts into composable language and task modules. We show that both methods can provide substantially better results when there is severe catastrophic forgetting. Below, we describe each method and analyze our findings in detail.

3.1 Methods

Mixing in unlabeled training data: This approach involves multi-task learning by mixing an unsupervised training task (**UNSUP**) into the WIKILINGUA-0 data. Mixing is controlled by a mixing rate κ, resulting in a final mixture that is $\kappa\%$ UNSUP data and $(100-\kappa)\%$ WIKILINGUA-0. As a data augmentation scheme, this method can be applied in all settings. We use the span corruption pretraining objective from T5 (Raffel et al., 2020) with mC4 data. We create separate multilingual datasets for each target language (**MIX-UNSUP**) as well as a single multilingual dataset that includes all of the WIKILINGUA-0 languages (**MIX-UNSUP-ALL**). Our goal is to encourage the model not to forget about other languages during training on English summarization. In our experiments, we use $\kappa = 1$.[17] An alternative approach is to perform model or prompt tuning on an intermediate task before tuning on WIKILINGUA-0. This intermediate tuning approach has been used to boost performance on English tasks for both MODELTUNING (Phang et al., 2019; Vu et al., 2020) and PROMPTTUNING (Vu et al., 2022), and has been successfully applied to the zero-shot cross-lingual transfer setting (Phang et al., 2020; Maurya et al., 2021) for MODELTUNING. In Appendix F, we show that intermediate tuning does not give reliable gains for XGEN.

Factorized prompts: Inspired by the MAD-X (Pfeiffer et al., 2020) adapter-based framework that learns modular language and task representations to adapt a multilingual model to arbitrary tasks and languages, we propose a novel method, dubbed "factorized prompts" (**FP**) and specifically designed for PROMPTTUNING. We attempt to decompose a soft prompt into "task" and "language" components that can be recombined in novel pairings (see Figure 5) with the goal of learning soft prompts that consist of disentangled and interpretable components. Unlike MAD-X, which learns language and task adapters separately for each language and each task, we learn language and task sub-prompts jointly for all languages and tasks. While we do not actively incentivize disentanglement, our multi-task multilingual pretraining procedure encourages the general language and task-specific knowledge to be stored in separate regions of the prompt. Intuitively, we vary languages while keeping the task sub-prompt fixed to train one side of the prompt, and vary tasks while keeping the language sub-prompt fixed to learn the other side.

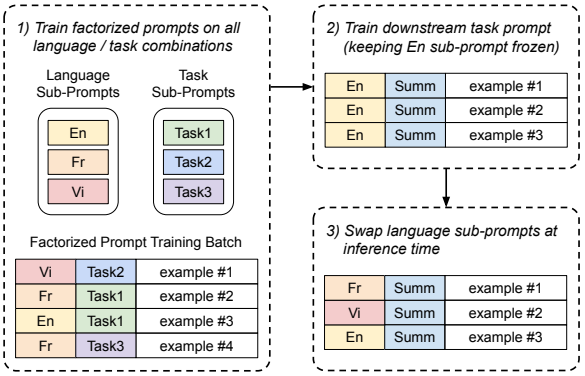

Figure 5: Our "factorized prompts" approach learns recomposable language and task sub-prompts by training on all language / task combinations from a set of unsupervised tasks covering all target languages.

We use mC4 data for all 18 WIKILINGUA-0 languages to create 7 unsupervised tasks per language. We randomly initialize language and task sub-prompts, each 50 tokens long. For each training example in our multi-task multilingual mixture, the relevant task and language sub-prompts are concatenated to form a full 100-token prompt. This training yields a set of learned language and task sub-prompts.[18] Next, we train a new task sub-prompt on WIKILINGUA-0 English summarization while using a frozen copy of the English language sub-prompt. Finally, when performing inference in another language, we replace the English sub-prompt with the target language sub-prompt, while

[17] In our preliminary experiments, $\kappa = 1$ performed best among a range of values $\{1, 5, 10, 30, 50\}$. We conjecture that a value of $\kappa > 1$ would prevent the model from focusing on the main task of summarization as more unsupervised data is added.

[18] As our mixture of tasks is large, we tuned for 200,000 steps for this training procedure.

continuing to use the learned summarization sub-prompt. To ablate the impact of the target language sub-prompt, we also report the performance using the English sub-prompt for all languages (**FP-EN**).

We use 7 unsupervised tasks per language, including: the PREFIX LM, SPAN CORRUPTION, and I.I.D. DENOISING tasks described in Raffel et al. (2020); LM, the causal left-to-right LM task with no context provided, i.e., the encoder's input is empty; MISSING PREFIX PREDICTION, predicting a missing prefix from the input; N-TOKEN PREFIX PREDICTION, copying the first n-tokens of the input; and MISSING N-TOKEN PREFIX PREDICTION, predicting the missing n-token prefix of the input. When training on WIKILINGUA-0, we initialize the task sub-prompt with the learned SPAN CORRUPTION task sub-prompt.

To confirm that language-specific prompts trained in this way encode meaningful differences between languages, we visualize a clustered heatmap of the cosine similarities between prompts trained on a classic LM task for each language in mC4. We observe meaningful clusters reflecting both linguistic and geographical similarities across languages. See Appendix D for details.

3.2 Results and Discussion

Mixing in multilingual data prevents catastrophic forgetting: In Figure 6, we observe that mixing in unsupervised multilingual data helps prevent catastrophic forgetting in all conditions, increasing the likelihood of predicting text in the target language. With MODELTUNING, this improved language accuracy reliably translates into higher end task performance (SP-ROUGE). For PROMPTTUNING, mixing provides gains for non-Latin script languages (RU and TH) where catastrophic forgetting is more severe; for Latin-script languages (FR and VI), mixing harms the overall summarization quality, despite achieving higher language accuracy.

Mixing in multilingual data in *all* WIKILINGUA languages leads to similar results, with a marginal drop in performance. Thus, if the desired target language is known ahead of time, the simpler strategy of mixing in just that language should be preferred. However, in cases where the inference language is unknown, mixing many languages is also effective.

Factorized prompts are helpful for overcoming severe catastrophic forgetting: Factorized prompts are successful at improving target language accuracy in all conditions. However, this does not always translate to higher SP-ROUGE.

Prompt Tuning	Model Tuning
वयस्क व्यक्ति का दांत निकलवाने के लिए डॅटिस्ट के पास जाएँ. वयस्क व्यक्ति का दांत खुद न निकालें.	Go to a डॅटिस्ट. Do not try to loose the दांत on your own.
Giảm độ ẩm trong nhà. Pha loãng giấm với nước. Xịt hỗn hợp lên thảm. Rắc muối nở lên mặt thảm. Làm khô thảm. Nhờ chuyên gia xử lý.	Lower the humidity. Mix giấm với nước. Apply giấm mixture lên thảm. Sprinkle muối nở lên thảm. Allow thảm to dry. Use quạt to làm khô thảm. Consider xử lý thảm bị hư hại

Table 1: Sample Hindi (top) and Vietnamese (bottom) predictions of our XXL model tuned with PROMPTTUNING and MODELTUNING. While the summaries are all understandable to a bilingual speaker, PROMPTTUNING tends to stay within the target language, whereas MODELTUNING is more prone to code switching between English (red) and the target language.

When language accuracy is already relatively high (for Latin-script languages, and for XXL models), factorized prompts are not helpful. However, in settings where vanilla PROMPTTUNING shows the most severe forgetting (e.g., at BASE size, on non-Latin script languages), factorized prompts provide large gains, similar to or exceeding our mixing approach.

4 Qualitative Analysis

To better understand qualitative differences between the solutions reached by MODELTUNING and PROMPTTUNING, two authors who were native speakers of Vietnamese and Hindi inspected 50 predictions of each method at the XXL model size.

For both languages, we observed that the MODELTUNING predictions were much more likely to include "code-switching", alternating between English and the target language, sometimes several times within a single sentence, as seen in Table 1. By comparison, the PROMPTTUNING predictions were more likely to use a consistent language throughout—typically staying entirely within the target language, but for some predictions resorting entirely to English. For both methods and both languages, we found code-switching predictions to generally be well-formed, in the sense that a bilingual speaker could extract the intended meaning, and that it served as a reasonable summary.

For Hindi, the PROMPTTUNING method showed lower mean SP-ROUGE scores than MODELTUNING (17.9 vs. 23.1), and had higher variances across runs (std: 5.1 vs. 0.7). Manual inspection showed that the lower-scoring PROMPTTUNING runs had far more predictions that were entirely English, explaining the lower SP-ROUGE scores.

For Vietnamese, PROMPTTUNING achieved higher

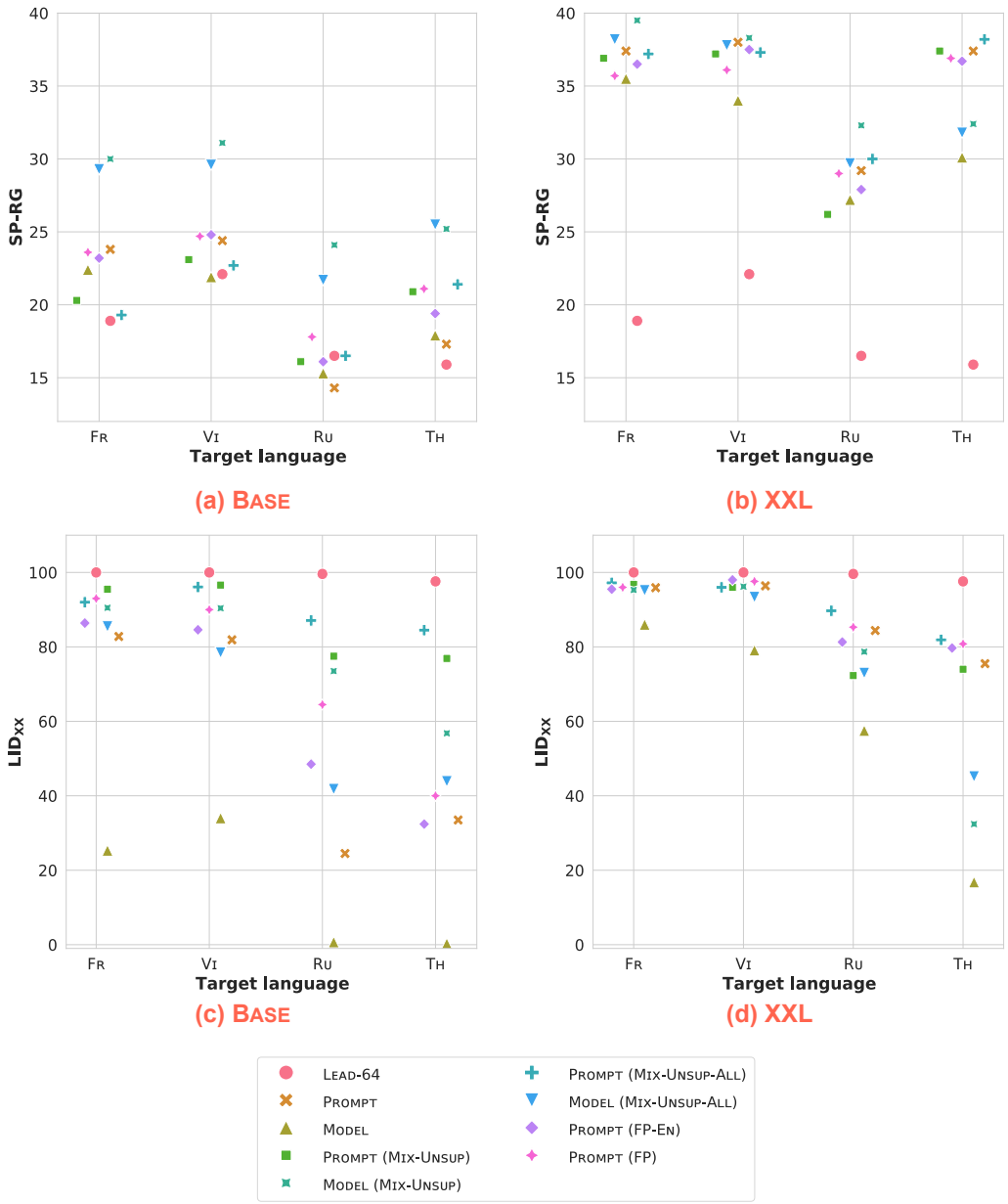

Figure 6: SP-ROUGE (top) and language accuracy (bottom) performance at BASE and XXL sizes of our proposed approaches: mixing unsupervised data (MIX), and factorized prompts (FP). See Appendix E for full results.

SP-ROUGE than MODELTUNING (38.0 vs. 34.0), with low variance in both cases (std: ≤ 0.5). On inspection, we found that most PROMPTTUNING predictions were entirely in Vietnamese, whereas MODELTUNING predictions typically contained at least some English. The PROMPTTUNING summaries tended to be shorter, but were often judged to be as good or better than the ground truth summaries. The MODELTUNING summaries tended to be a bit longer. If mentally translating any English words back to Vietnamese, the quality was judged to be similar to the prompt tuning summaries, suggesting that the lower SP-ROUGE score is primarily due to the presence of intervening English.

5 Related Work

Mixing unlabeled multilingual data in during fine-tuning can be viewed a version of rehearsal (Robins, 1995), commonly used to mitigate catastrophic forgetting. Related work has used this mixing (Xue et al., 2021; Shakeri et al., 2021) to combat "accidental translation", a symptom of English overfitting. However, these works are concerned with MODELTUNING, whereas we apply it to PROMPTTUNING. Other methods of combatting catastrophic forgetting include the slowing (or stopping) of updates for some parameters. Kirkpatrick et al. (2017) reduce the learning rate of parameters important for

earlier tasks as they train on new ones. Maurya et al. (2021) similarly stop learning for some parameters by only training input and output layers. In the context of prompt tuning, Qin and Joty (2022) address catastrophic forgetting during continual learning of new domains by combining the new training data with pseudo-labeled data of previous domains.

Previous work has also explored intermediate adaptation of pre-trained models, which has been shown to be effective for ModelTuning (Howard and Ruder, 2018; Phang et al., 2019; Vu et al., 2020, 2021) and PromptTuning (Vu et al., 2022). Phang et al. (2020) apply intermediate adaptation in the multilingual domain, but use English in the adaption instead of the target language. Maurya et al. (2021) use a cross-lingual intermediate task. Unlike our task, theirs is designed to closely match the downstream task. Several works use intermediate adaptation to create a model that is better in the zero- or few-shot settings (Wei et al., 2022; Sanh et al., 2022; Min et al., 2022), but these target generalization to new tasks, whereas we focus on generalizing to new languages within one task.

Many parameter-efficient adaption methods exist (Rebuffi et al., 2017; Houlsby et al., 2019; Karimi Mahabadi et al., 2021; Zaken et al., 2021; Hu et al., 2022) and some have shown strong performance under domain shift (Lester et al., 2021; Li and Liang, 2021). We chose PromptTuning due to its simplicity and the localization of parameters—making the implementation of factorized prompts easy. See Liu et al. (2021), He et al. (2022), and Liu et al. (2022) for detailed discussion of the differences between these methods.

Other work explores cross-lingual transfer learning with parameter-efficient methods. Zhao and Schütze (2021) find that soft prompts can effectively be used in cross-lingual settings, but their work is constrained to classification. Pfeiffer et al. (2020) use adapters rather than prompts and leverage parameter-efficient learning to create separate language and task understanding modules that can be combined at inference time.

There has been recent interest in cross-lingual generation. Maurya et al. (2021) and Chi et al. (2020) evaluate their methods using cross-lingual generation, including summarization as we do. However, Chi et al. (2020) use parallel data during pre-training to "align" representations across languages during pre-training while our approach does not.

6 Conclusion

In this work, we explored how different adaptation methods fare on the challenging "XGen" task of zero-shot cross-lingual summarization. While many methods struggled with catastrophic forgetting (outputting English rather than the target language), we observed two factors helped to mitigate this problem: (1) increasing model scale, and (2) decreasing the number of parameters tuned during adaptation. When all of a model's weights are tuned on English (ModelTuning), forgetting is quick and severe. By contrast, limiting the tunable parameters to a smaller soft prompt (PromptTuning) helps to combat forgetting, though prompt size is an important variable to control.

To further close the gap with supervised methods, we explored two adaptation techniques—one entirely novel, and one that has been used before, but not in combination with parameter-efficient methods like PromptTuning. We find that mixing in unsupervised multilingual data is always helpful. Our novel approach, "factorized prompts", is helpful at smaller model sizes, but has no benefit at larger sizes. We hope that future work will continue to explore XGen tasks including WikiLingua-0, and develop stronger zero-shot adaptation techniques to allow multilingual models to reliably generate coherent text in any target language.

7 Limitations

Our work focuses on a single XGen task, WikiLingua-0 summarization. In future work, it would be valuable to see if our findings generalize to additional domains and tasks, including those beyond summarization.

WikiLingua-0 is not a traditional summarization task. Rather than news articles, the input documents are how-to guides, and the summaries are "headings" for each step, which may be more terse than a traditional summary. We observed some minor data quality issues in WikiLingua-0, including HTML code present in some target strings, and artifacts of machine translation evident in some non-English documents. Nevertheless, we believe that WikiLingua-0 is a meaningful and challenging XGen task, with the notable advantage of covering a range of high- and low-resource languages from diverse language families and with diverse scripts.

In evaluating parameter-efficient methods, we focused on PromptTuning due to its simplicity. There are a growing number of other parameter-

efficient methods that could also be tested, including ADAPTERS (Rebuffi et al., 2017; Houlsby et al., 2019), BITFIT (Zaken et al., 2021), PREFIX-TUNING (Li and Liang, 2021), $(IA)^3$ (Liu et al., 2022), and many more; see Liu et al. (2021), He et al. (2022), and Liu et al. (2022) for detailed discussion of the differences between these methods. We expect many of the benefits of tuning fewer parameters to persist across methods, but this remains to be explored.

Acknowledgements

We thank Thibault Sellam, Sebastian Gehrmann, Kalpesh Krishna, Marzena Karpinska, and the members of the UMass NLP group for helpful discussion and feedback. We would also like to thank Grady Simon, Xavier Garcia, and Douglas Eck for their comments on this manuscript. Vu and Iyyer are partially supported by awards IIS-1955567 and IIS-2046248 from the National Science Foundation (NSF).

References

Shengnan An, Yifei Li, Zeqi Lin, Qian Liu, Bei Chen, Qiang Fu, Weizhu Chen, Nanning Zheng, and Jian-Guang Lou. 2022. Input-tuning: Adapting unfamiliar inputs to frozen pretrained models. *arXiv preprint arXiv:2203.03131*.

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 (ACL 2020)*, pages 4623–4637.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2020)*, 34(05):7570–7577.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics (TACL 2020)*, 8:454–470.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019)*, volume 32.

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 (EMNLP 2018)*, pages 2475–2485.

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

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

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning.

Proceedings of the 10th International Conference on Learning Representations (ICLR 2022).

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 (PMLR 2019)*, volume 97, pages 2790–2799.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS 2021)*, volume 34, pages 1022–1035.

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 (PNAS 2017)*, 114(13):3521–3526.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Evaluating the efficacy of summarization evaluation across languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (Findings of ACL-IJCNLP 2021)*, pages 801–812.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics (TACL 2022)*, 10:50–72.

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 (EMNLP 2018 System Demonstrations)*, pages 66–71.

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 (Findings of EMNLP 2020))*, pages 4034–4048.

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 (EMNLP 2021)*, pages 3045–3059.

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

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 283–288.

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 (ACL 2021)*, pages 4582–4597.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop of Text Summarization Branches Out (WAS 2004)*, pages 74–81.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. ZmBART: An unsupervised cross-lingual transfer framework for language generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (Findings of ACL-IJCNLP 2021)*, pages 2804–2818.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to Learn In Context. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2022)*.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318.

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 2020)*, pages 7654–7673.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL 2020)*, pages 557–575.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2019. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Chengwei Qin and Shafiq Joty. 2022. Lfpt5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5. *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*.

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

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Proceedings of the 31th Conference on Neural Information Processing Systems (NeurIPS 2017)*, volume 30.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

Anthony Robins. 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146.

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

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 1073–1083.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 7881–7892.

Siamak Shakeri, Noah Constant, Mihir Kale, and Linting Xue. 2021. Towards zero-shot multilingual synthetic question and answer generation for cross-lingual reading comprehension. In *Proceedings of the 14th International Conference on Natural Language Generation (INLG 2021)*, pages 35–45.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. 2022. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 5039–5059.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. STraTA: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pages 5715–5731.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 7882–7926.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned Language Models Are Zero-Shot Learners. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*.

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 (NAACL-HLT 2021)*, pages 483–498.

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

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pages 8547–8555.

Appendices

A Evaluation on zero-shot cross-lingual benchmarks

From Table 2 to Table 8, we show our results for MODELTUNING and PROMPTTUNING across different zero-shot cross-lingual benchmarks. Overall, we find that MODELTUNING typically performs better than PROMPTTUNING, although PROMPTTUNING at scale (i.e., XXL) matches the performance of MODELTUNING on English and can yield better results on some languages.

Size	Method	Language														
		en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh
BASE	PROMPT	$78.8_{1.2}$	$64.7_{0.4}$	$68.9_{0.5}$	$68.4_{1.0}$	$70.1_{0.8}$	$73.7_{0.5}$	$75.6_{1.1}$	$65.1_{0.4}$	$68.0_{0.6}$	$62.5_{0.2}$	$69.7_{0.9}$	$67.6_{0.3}$	$60.9_{0.7}$	$70.7_{1.5}$	$70.3_{1.3}$
	MODEL	$87.1_{0.2}$	$72.3_{0.2}$	$78.4_{0.7}$	$77.7_{0.2}$	$82.0_{1.0}$	$84.5_{0.8}$	$80.8_{0.6}$	$70.3_{1.1}$	$74.8_{0.7}$	$69.3_{1.0}$	$74.3_{1.0}$	$73.2_{1.0}$	$68.0_{0.3}$	$77.7_{0.5}$	$72.9_{1.0}$
	$\Delta_{\text{P-M}}$	-8.3	-7.6	-9.5	-9.3	-11.9	-10.8	-5.2	-5.2	-6.8	-6.8	-4.6	-5.6	-7.1	-7.0	-2.6
XXL	PROMPT	$91.5_{0.2}$	$81.5_{0.2}$	$87.1_{0.4}$	$88.5_{0.4}$	$88.9_{0.8}$	$90.1_{0.4}$	$88.4_{1.1}$	$84.5_{0.4}$	$83.3_{0.4}$	$80.7_{0.7}$	$81.6_{0.3}$	$83.7_{0.4}$	$78.9_{0.4}$	$85.1_{1.0}$	$83.7_{0.4}$
	MODEL	$92.8_{0.6}$	$85.6_{0.6}$	$89.3_{0.5}$	$89.2_{0.3}$	$89.5_{0.8}$	$90.8_{0.0}$	$88.5_{0.8}$	$84.5_{0.5}$	$82.9_{0.8}$	$83.7_{0.7}$	$78.8_{0.9}$	$83.3_{1.0}$	$81.5_{0.7}$	$87.6_{0.9}$	$84.1_{0.2}$
	$\Delta_{\text{P-M}}$	-1.3	-4.1	-2.2	-0.7	-0.6	-0.7	-0.1	0.0	0.4	-3.0	2.8	0.4	-2.6	-2.5	-0.4

Table 2: Best validation accuracy per language on XNLI.

Size	Method	Language										
		en	ar	de	el	es	hi	ru	th	tr	vi	zh
BASE	PROMPT	$83.9_{0.3}$	$63.0_{1.5}$	$70.7_{0.8}$	$63.5_{0.7}$	$75.6_{1.1}$	$61.4_{1.7}$	$61.6_{1.4}$	$58.3_{1.0}$	$60.9_{0.8}$	$68.7_{0.8}$	$45.3_{1.2}$
	MODEL	$91.9_{0.3}$	$72.9_{0.9}$	$76.9_{0.7}$	$68.4_{0.4}$	$84.9_{0.7}$	$67.5_{0.9}$	$69.8_{1.0}$	$63.4_{1.1}$	$69.3_{0.9}$	$77.2_{0.2}$	$53.3_{0.4}$
	$\Delta_{\text{P-M}}$	-8.0	-9.9	-6.2	-4.9	-9.3	-6.1	-8.2	-5.1	-8.4	-8.5	-8.0
XXL	PROMPT	$95.0_{0.1}$	$83.6_{0.3}$	$84.9_{0.9}$	$76.6_{0.6}$	$92.5_{0.5}$	$77.7_{1.1}$	$80.3_{0.6}$	$71.6_{1.5}$	$81.9_{0.5}$	$85.5_{0.2}$	$60.8_{0.7}$
	MODEL	$95.5_{0.2}$	$88.6_{0.1}$	$86.3_{0.9}$	$81.8_{0.7}$	$92.4_{0.4}$	$82.1_{0.8}$	$85.0_{0.5}$	$75.8_{0.8}$	$84.6_{0.2}$	$88.5_{0.5}$	$64.9_{0.8}$
	$\Delta_{\text{P-M}}$	-0.5	-5.0	-1.4	-5.2	0.1	-4.4	-4.7	-4.2	-2.7	-3.0	-4.1

Table 3: Best validation F1 per language on XQUAD.

Size	Method	Language						
		en	ar	de	es	hi	vi	zh
BASE	PROMPT	$75.5_{0.4}$	$45.1_{1.3}$	$55.2_{0.8}$	$63.0_{1.0}$	$47.9_{2.1}$	$53.8_{0.6}$	$55.1_{0.6}$
	MODEL	$79.4_{0.4}$	$53.8_{0.2}$	$62.8_{0.4}$	$69.7_{0.6}$	$55.7_{0.3}$	$62.8_{0.6}$	$62.7_{0.5}$
	$\Delta_{\text{P-M}}$	-3.9	-8.7	-7.6	-6.7	-7.8	-9.0	-7.6
XXL	PROMPT	$85.4_{0.4}$	$63.7_{0.8}$	$72.0_{0.8}$	$76.3_{0.4}$	$68.4_{0.5}$	$70.1_{0.6}$	$71.0_{0.4}$
	MODEL	$84.7_{0.5}$	$71.1_{0.5}$	$72.8_{0.1}$	$79.0_{0.2}$	$73.9_{0.3}$	$71.4_{0.3}$	$75.4_{0.5}$
	$\Delta_{\text{P-M}}$	0.7	-7.4	-0.8	-2.7	-5.5	-1.3	-4.4

Table 4: Best validation F1 per language on MLQA.

Size	Method	Language								
		en	ar	bn	fi	id	ko	ru	sw	te
BASE	PROMPT	$68.1_{1.5}$	$61.6_{4.3}$	$37.2_{0.5}$	$56.6_{2.0}$	$59.6_{3.4}$	$35.3_{0.7}$	$58.2_{1.0}$	$46.7_{3.1}$	$41.1_{3.3}$
	MODEL	$71.5_{0.9}$	$69.9_{1.2}$	$41.5_{1.3}$	$67.6_{0.7}$	$77.5_{0.6}$	$48.8_{0.4}$	$57.8_{1.1}$	$61.2_{1.0}$	$48.1_{1.5}$
	$\Delta_{\text{P-M}}$	-3.4	-8.3	-4.3	-11.0	-17.9	-13.5	0.4	-14.5	-7.0
XXL	PROMPT	$82.8_{0.5}$	$78.1_{0.6}$	$73.9_{1.4}$	$76.6_{0.8}$	$83.9_{0.4}$	$73.7_{2.0}$	$69.7_{0.7}$	$71.8_{2.3}$	$77.9_{1.1}$
	MODEL	$85.1_{0.4}$	$85.7_{0.1}$	$83.4_{1.5}$	$82.3_{0.5}$	$88.7_{0.3}$	$76.3_{0.5}$	$76.5_{1.1}$	$82.9_{0.6}$	$79.7_{0.1}$
	$\Delta_{\text{P-M}}$	-2.3	-7.6	-9.5	-5.7	-4.8	-2.6	-6.8	-11.1	-1.8

Table 5: Best validation F1 per language on TYDIQA.

Size	Method	Language						
		en	de	es	fr	ja	ko	zh
BASE	PROMPT	$94.3_{0.8}$	$85.3_{0.5}$	$88.1_{0.8}$	$88.9_{0.8}$	$80.8_{0.3}$	$79.7_{0.8}$	$82.3_{1.1}$
	MODEL	$94.9_{0.7}$	$89.1_{0.4}$	$90.8_{0.3}$	$90.7_{0.5}$	$84.1_{0.8}$	$83.5_{0.9}$	$84.3_{0.5}$
	$\Delta_{\text{P-M}}$	-0.6	-3.8	-2.7	-1.8	-3.3	-3.8	-2.0
XXL	PROMPT	$96.8_{0.6}$	$90.7_{0.2}$	$92.9_{0.4}$	$93.5_{0.4}$	$88.1_{0.4}$	$86.0_{0.9}$	$88.8_{1.0}$
	MODEL	$96.4_{0.3}$	$91.9_{0.2}$	$92.8_{0.3}$	$93.9_{0.5}$	$87.2_{1.2}$	$89.6_{0.3}$	$91.9_{0.4}$
	$\Delta_{\text{P-M}}$	0.4	-1.2	0.1	-0.4	0.9	-3.6	-3.1

Table 6: Best validation accuracy per language on PAWS-X.

Size	Method	Language																			
		en	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv
BASE	PROMPT	$83.3_{0.4}$	$73.3_{1.3}$	$43.6_{1.0}$	$74.8_{0.7}$	$64.5_{0.8}$	$68.9_{1.0}$	$68.2_{1.0}$	$74.3_{1.1}$	$62.5_{2.3}$	$46.2_{2.4}$	$31.9_{1.8}$	$63.2_{0.4}$	$78.3_{0.1}$	$48.9_{1.3}$	$65.8_{2.9}$	$68.4_{0.3}$	$54.5_{1.0}$	$81.9_{1.2}$	$34.9_{0.9}$	$53.7_{0.7}$
	MODEL	$87.3_{0.8}$	$72.3_{1.5}$	$52.1_{2.4}$	$63.2_{2.1}$	$72.0_{1.3}$	$64.5_{0.6}$	$59.2_{1.7}$	$66.7_{1.3}$	$68.6_{1.2}$	$49.1_{1.8}$	$30.4_{1.9}$	$71.5_{0.3}$	$79.5_{0.3}$	$44.4_{1.0}$	$65.7_{0.9}$	$66.8_{0.4}$	$51.6_{0.2}$	$82.5_{0.7}$	$35.7_{1.4}$	$46.5_{1.2}$
	$\Delta_{\text{P-M}}$	-4.0	1.0	-8.5	11.6	-7.5	4.4	9.0	7.6	-6.1	-2.9	1.5	-8.3	-1.2	4.5	0.1	1.6	2.9	-0.6	-0.8	7.2
XXL	PROMPT	$91.5_{0.4}$	$83.3_{0.8}$	$51.0_{0.8}$	$84.1_{1.0}$	$79.1_{1.2}$	$77.4_{0.3}$	$79.3_{0.8}$	$83.4_{1.1}$	$78.4_{1.7}$	$67.2_{1.8}$	$41.0_{1.9}$	$77.1_{2.6}$	$86.6_{1.2}$	$61.2_{0.3}$	$75.4_{2.1}$	$79.9_{0.9}$	$71.1_{4.1}$	$89.0_{1.2}$	$41.8_{1.7}$	$71.8_{0.5}$
	MODEL	$91.7_{0.4}$	$82.1_{1.1}$	$62.0_{2.3}$	$86.6_{0.7}$	$82.7_{0.5}$	$79.3_{0.6}$	$79.6_{0.7}$	$83.0_{0.2}$	$74.7_{1.1}$	$57.2_{1.9}$	$52.5_{0.9}$	$69.5_{1.1}$	$88.5_{0.7}$	$60.5_{0.7}$	$81.1_{0.3}$	$74.1_{0.4}$	$71.0_{3.4}$	$88.5_{0.5}$	$50.7_{0.8}$	$65.7_{0.7}$
	$\Delta_{\text{P-M}}$	-0.2	1.2	-11.0	-2.5	-3.6	-1.9	-0.3	0.4	3.7	10.0	-11.5	7.6	-1.9	0.7	-5.7	5.8	0.1	0.5	-8.9	6.1

Size	Method	Language																			
		ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh
BASE	PROMPT	$56.0_{1.4}$	$44.7_{2.0}$	$33.0_{1.0}$	$47.0_{1.7}$	$39.4_{1.5}$	$76.8_{0.9}$	$27.6_{2.0}$	$79.0_{1.2}$	$76.3_{0.7}$	$58.0_{1.6}$	$62.2_{1.3}$	$45.6_{2.5}$	$47.8_{3.2}$	$10.9_{0.1}$	$74.9_{0.7}$	$68.3_{0.8}$	$50.4_{6.9}$	$69.6_{0.7}$	$61.9_{3.1}$	$33.9_{3.0}$
	MODEL	$53.7_{1.5}$	$20.7_{1.9}$	$33.2_{0.4}$	$45.1_{0.5}$	$39.8_{0.9}$	$75.4_{0.8}$	$28.0_{1.3}$	$80.2_{2.3}$	$75.1_{1.8}$	$50.3_{1.2}$	$66.6_{0.4}$	$43.2_{0.7}$	$44.2_{1.4}$	$9.9_{0.7}$	$78.2_{1.9}$	$60.3_{1.6}$	$37.6_{3.4}$	$74.8_{1.8}$	$59.9_{1.5}$	$41.0_{1.8}$
	$\Delta_{\text{P-M}}$	2.3	24.0	-0.2	1.9	-0.4	1.4	-0.4	-1.2	1.2	7.7	-4.4	2.4	3.6	1.0	-3.3	8.0	12.8	-5.2	2.0	-7.1
XXL	PROMPT	$70.5_{2.5}$	$50.8_{2.1}$	$51.2_{1.4}$	$62.6_{1.4}$	$57.2_{3.8}$	$84.7_{0.9}$	$42.5_{1.8}$	$89.1_{0.6}$	$86.9_{0.9}$	$71.7_{1.5}$	$77.8_{0.7}$	$59.8_{1.8}$	$57.8_{0.1}$	$9.4_{1.2}$	$83.3_{1.6}$	$87.6_{0.4}$	$81.7_{3.0}$	$79.7_{2.2}$	$60.3_{0.5}$	$49.1_{1.7}$
	MODEL	$71.9_{1.1}$	$37.0_{2.1}$	$46.1_{0.6}$	$55.6_{0.1}$	$54.8_{0.6}$	$81.1_{0.7}$	$38.5_{0.8}$	$89.1_{0.3}$	$87.4_{0.7}$	$72.8_{2.3}$	$78.3_{0.5}$	$53.6_{0.6}$	$53.6_{0.9}$	$16.7_{0.7}$	$84.4_{0.6}$	$74.2_{0.4}$	$82.8_{0.7}$	$84.6_{0.3}$	$68.2_{3.1}$	$56.2_{0.6}$
	$\Delta_{\text{P-M}}$	-1.4	13.8	5.1	7.0	2.4	3.6	4.0	0.0	-0.5	-1.1	-0.5	6.2	4.2	-7.3	-1.1	13.4	-1.1	-4.9	-7.9	-7.1

Table 7: Best validation span F1 per language on WIKIANN NER.

Size	Method	Language																	
		en	es	pt	fr	de	ru	it	id	nl	ar	zh	vi	th	ja	ko	hi	cs	tr
SMALL	PROMPT	$24.5_{0.2}$	$20.2_{0.6}$	$20.7_{0.3}$	$19.2_{0.1}$	$15.4_{0.3}$	$11.4_{0.1}$	$18.3_{0.5}$	$19.0_{0.6}$	$16.9_{0.2}$	$16.0_{0.3}$	$12.8_{0.5}$	$21.4_{0.2}$	$14.9_{0.4}$	$12.1_{0.1}$	$14.8_{0.3}$	$11.2_{0.6}$	$14.0_{0.1}$	$13.7_{0.0}$
	MODEL	$38.0_{0.4}$	$22.3_{0.1}$	$23.2_{0.4}$	$21.3_{0.3}$	$17.8_{0.2}$	$14.6_{0.2}$	$20.1_{0.2}$	$21.0_{0.2}$	$19.7_{0.2}$	$17.0_{0.1}$	$14.1_{0.3}$	$22.5_{0.1}$	$17.3_{0.1}$	$14.1_{0.1}$	$17.8_{0.4}$	$9.5_{0.0}$	$17.2_{0.1}$	$16.0_{0.1}$
	$\Delta_{\text{P-M}}$	-13.5	-2.1	-2.5	-2.1	-2.4	-3.2	-1.8	-2.0	-2.8	-1.0	-1.3	-1.1	-2.4	-2.0	-3.0	1.7	-3.2	-2.3
BASE	PROMPT	$29.8_{0.4}$	$24.2_{0.7}$	$25.0_{0.5}$	$23.8_{0.1}$	$19.2_{0.5}$	$14.3_{0.6}$	$20.2_{0.1}$	$22.1_{0.7}$	$20.4_{0.7}$	$18.5_{0.7}$	$13.1_{0.8}$	$24.4_{0.6}$	$17.3_{0.6}$	$14.2_{0.3}$	$16.2_{0.4}$	$10.6_{0.7}$	$16.5_{0.3}$	$14.4_{0.1}$
	MODEL	$39.6_{0.4}$	$23.3_{0.1}$	$23.8_{0.8}$	$22.4_{0.3}$	$18.8_{0.2}$	$15.3_{0.2}$	$20.3_{0.2}$	$23.0_{0.2}$	$20.1_{0.2}$	$17.4_{0.2}$	$15.1_{0.5}$	$21.9_{0.3}$	$17.9_{0.2}$	$14.6_{0.3}$	$17.3_{0.1}$	$9.1_{0.2}$	$17.8_{0.1}$	$17.5_{0.1}$
	$\Delta_{\text{P-M}}$	-9.8	0.9	1.2	1.4	0.4	-1.0	-0.1	-0.9	0.3	1.1	-2.0	2.5	-0.6	-0.4	-1.1	1.5	-1.3	-3.1
LARGE	PROMPT	$35.3_{0.3}$	$29.4_{0.3}$	$29.0_{0.0}$	$28.8_{0.1}$	$24.8_{0.5}$	$20.4_{0.8}$	$24.3_{0.2}$	$27.2_{0.1}$	$27.0_{0.3}$	$24.1_{0.5}$	$20.8_{1.1}$	$29.3_{0.3}$	$24.7_{1.0}$	$19.4_{0.3}$	$23.4_{0.7}$	$17.3_{0.1}$	$22.7_{0.4}$	$19.5_{0.2}$
	MODEL	$42.6_{0.2}$	$29.7_{0.1}$	$30.3_{0.3}$	$27.8_{0.6}$	$23.5_{0.8}$	$17.4_{1.0}$	$25.6_{0.7}$	$26.9_{0.7}$	$25.3_{0.5}$	$23.7_{1.7}$	$19.2_{0.6}$	$27.2_{0.8}$	$25.9_{0.7}$	$22.1_{0.7}$	$23.9_{0.7}$	$12.7_{0.4}$	$22.1_{0.4}$	$20.6_{0.6}$
	$\Delta_{\text{P-M}}$	-7.3	-0.3	-1.3	1.0	1.3	3.0	-1.3	0.3	1.7	0.4	1.6	2.1	-1.2	-2.7	-0.5	4.6	0.6	-1.1
XL	PROMPT	$38.4_{0.2}$	$34.8_{0.4}$	$33.3_{0.3}$	$33.4_{0.3}$	$28.9_{0.1}$	$25.3_{0.3}$	$28.7_{0.3}$	$33.1_{0.1}$	$32.3_{0.2}$	$30.4_{0.4}$	$24.4_{2.0}$	$34.1_{0.4}$	$33.2_{0.4}$	$23.1_{2.3}$	$27.4_{1.3}$	$17.3_{2.3}$	$26.8_{0.4}$	$23.5_{0.2}$
	MODEL	$45.0_{0.3}$	$32.2_{0.3}$	$33.1_{0.3}$	$31.9_{0.5}$	$25.3_{1.0}$	$19.7_{0.6}$	$28.6_{0.2}$	$28.3_{0.5}$	$28.4_{0.7}$	$27.3_{0.8}$	$30.0_{0.8}$	$29.8_{0.7}$	$25.6_{0.5}$	$25.4_{0.7}$	$29.1_{0.4}$	$16.3_{0.5}$	$23.5_{0.6}$	$22.9_{0.3}$
	$\Delta_{\text{P-M}}$	-6.6	2.6	0.2	1.5	3.6	5.6	0.1	4.8	3.9	3.1	-5.6	4.3	7.6	-2.3	-1.7	1.0	3.3	0.6
XXL	PROMPT	$43.4_{0.4}$	$36.8_{0.4}$	$36.1_{0.4}$	$37.4_{0.2}$	$30.3_{0.4}$	$29.2_{1.0}$	$30.9_{0.5}$	$35.1_{0.6}$	$35.1_{0.5}$	$32.9_{0.3}$	$31.9_{3.2}$	$38.0_{0.0}$	$37.4_{0.7}$	$27.0_{1.6}$	$33.6_{0.9}$	$17.9_{5.1}$	$30.7_{0.1}$	$25.8_{0.9}$
	MODEL	$46.7_{0.1}$	$37.1_{0.6}$	$35.8_{0.3}$	$35.5_{0.6}$	$30.2_{0.3}$	$27.2_{0.4}$	$31.6_{0.5}$	$32.6_{0.1}$	$30.9_{0.8}$	$30.1_{0.9}$	$40.8_{0.3}$	$34.0_{0.5}$	$30.1_{0.5}$	$29.8_{0.4}$	$31.2_{0.6}$	$23.1_{0.7}$	$26.0_{0.7}$	$26.2_{0.3}$
	$\Delta_{\text{P-M}}$	-3.3	-0.3	0.3	1.9	0.1	2.0	-0.7	2.5	4.2	2.8	-8.9	4.0	7.3	-2.8	2.4	-5.2	4.7	-0.4

Table 8: Best validation SP-ROUGE per language on WIKILINGUA-0.

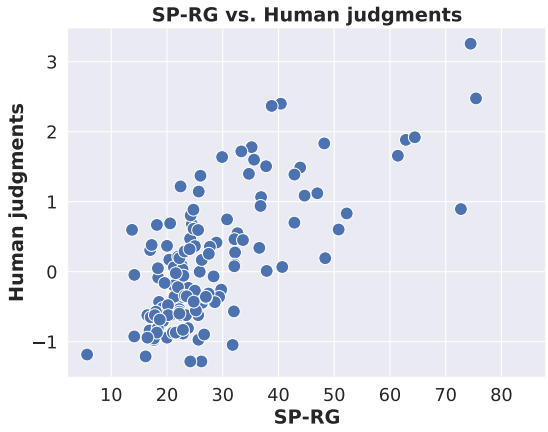

Figure 7: A scatterplot demonstrating the linear relationship between SP-ROUGE and human judgments on FOCUS for French summaries. As shown in Table 9, SP-ROUGE also correlates well with human judgments on other languages.

B Measuring the correlation between SP-RG and human judgments

To evaluate how well our proposed SP-ROUGE metric correlates with human judgments, we use the MULTISUMM EVAL dataset introduced by Koto et al. (2021), which is a manually-annotated multilingual resource for summarization evaluation with 4,320 human annotations on FOCUS (precision) and COVERAGE (recall) between machine-generated summaries and ground-truth summaries. We compare SP-ROUGE to BLEURT (Sellam et al., 2020), which is a learned evaluation metric based on BERT (Devlin et al., 2019). Table 9 shows the Pearson correlation coefficient between these metrics and human judgments across 8 MULTISUMM EVAL languages, including German (DE), English (EN), Spanish (ES), French (FR), Indonesian (ID), Russian (RU), Turkish (TR), and Mandarin Chinese (ZH). Overall, we found that the performance of SP-ROUGE and the more computationally expensive BLEURT metric were similar. Specifically, SP-ROUGE achieved an average FOCUS score of 0.68 and an average COVERAGE score of 0.65, whereas BLEURT achieved 0.68 and 0.70, respectively. Figure 7 demonstrates the linear relationship between SP-ROUGE-LSUM vs FOCUS scores on French.

C Zero-shot evaluation results on WIKILINGUA-0

Our zero-shot evaluation results on WIKILINGUA-0 for French (**FR**), Vietnamese (**VI**), Russian (**RU**), and Thai (**TH**) are shown in Table 10. See Table 8 for results across all target languages. Our results suggest that WIKILINGUA-0 is a challenging task for both MODELTUNING and PROMPTTUNING. As model size increases, PROMPTTUNING usually produces better results than MODELTUNING when there is a significant language shift at inference time. Longer prompts help to better learn the English summarization task. However, the increased capacity leads the model to forgets other languages.

D Language-Specific Prompt Clustering Analysis

To confirm that language-specific prompts trained on an LM task encode meaningful differences between languages, we train 107 prompts, one for each language in the mC4 corpus. Specifically, we train prompts for the mT5-BASE model, with a prompt length of 1, for 10K training steps, using a batch size of 32. The training task consists of classic causal language modeling, with an empty string fed as inputs to the encoder, and the document text passed as targets. Each prompt is trained exclusively on data from a single language bucket; however, we note that mC4 contains a non-trivial number of language ID errors, particularly for lower-resource languages (Kreutzer et al., 2022).

Figure 8 shows a clustered heatmap of the cosine similarities between the trained prompts. We observe a number of interpretable clusters that give us confidence that the learned prompts encode meaningful language representations. For example, the leftmost 25 languages form a visible cluster and are all nearly all languages of Europe,[19] with meaningful sub-clusters for different European regions: Northern (NO, SV, DA, NL), Central (CS, PL, SK, LT, SL), South-Western (ES, PT, FR, IT) and Eastern (KK, AZ, TR, BG, MK, BE, UK). Another prominent cluster covers languages of India, Pakistan and Nepal (ML, TE, NE, KA, KN, GU, HI, SI, BN, TA), despite the fact that these languages cover different linguistic families and are written with different scripts. While geography seems to be the primary factor influencing prompt similarity, linguistic relationships also play a role. For instance, we observe that Finnish (FI) and Hungarian (HU), both Finno-Ugric languages, form a cluster despite their geographic distance. Similarly, Igbo (IG), spoken mainly in

[19]The only exceptions are Vietnamese (VI) and Indonesian (ID), which are both written with Latin(-derived) scripts. We also note that Indonesian has a high language ID error rate within mC4.

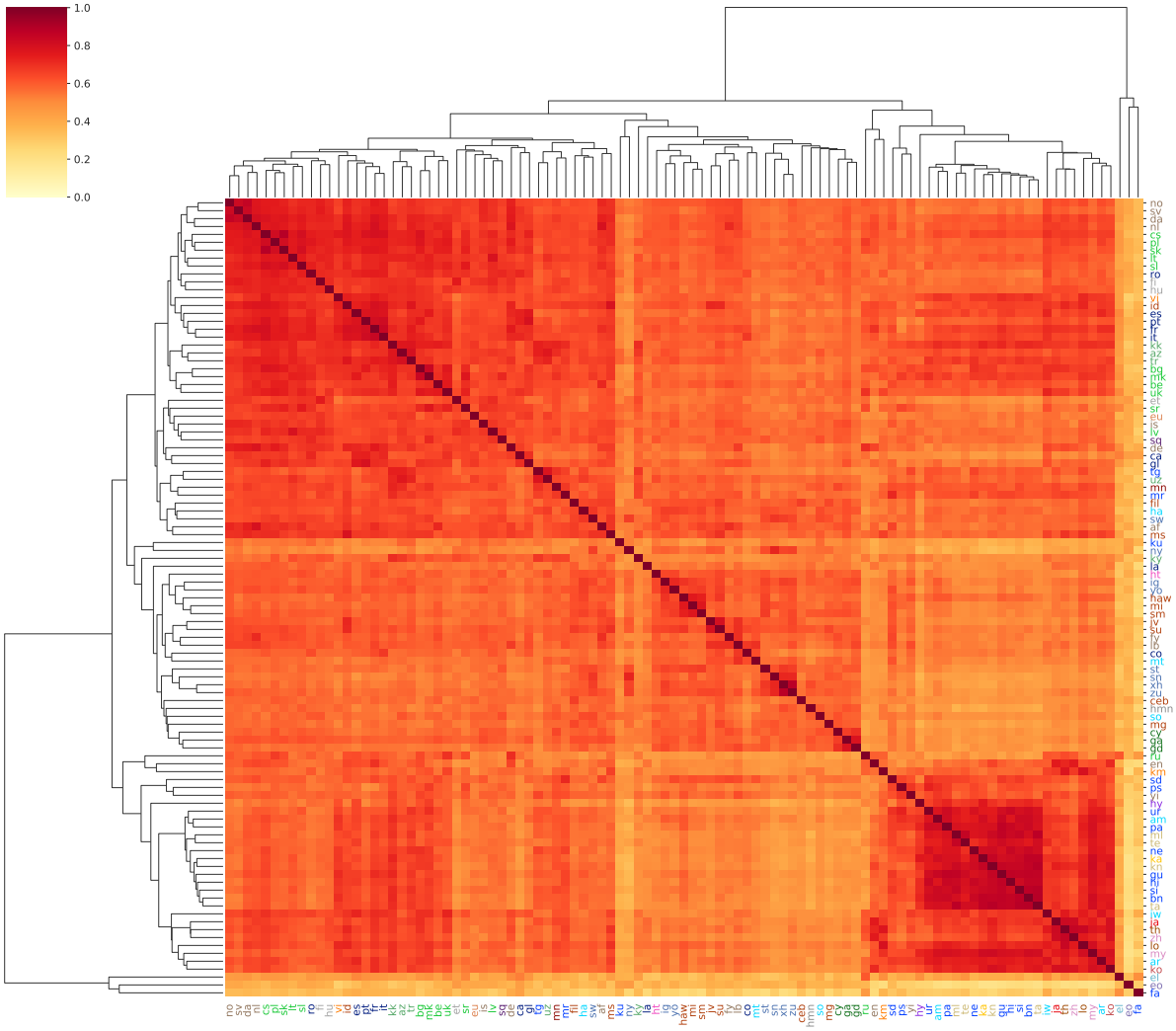

Figure 8: A clustered heatmap of cosine similarities between 107 mT5-BASE prompts trained on language-specific LM tasks. Language codes with the same color share a linguistic family.

Metric	Focus									Coverage								
	De	En	Es	Fr	Id	Ru	Tr	Zh	Avg.	De	En	Es	Fr	Id	Ru	Tr	Zh	Avg.
SP-RG	0.88	0.53	0.60	0.67	0.67	0.49	0.82	0.77	0.68	0.88	0.53	0.65	0.62	0.68	0.37	0.75	0.72	0.65
BLEURT	0.87	0.52	0.66	0.70	0.61	0.56	0.79	0.73	0.68	0.88	0.60	0.65	0.71	0.62	0.59	0.79	0.75	0.70

Table 9: SP-ROUGE correlates well with human judgments, providing a similar correlation to BLEURT while being significantly less computationally expensive.

Size	Method	En		Fr			Ru			Vi			Th		
		SP-RG	LID_{En}	SP-RG	LID_{En}	LID_{Fr}	SP-RG	LID_{En}	LID_{Ru}	SP-RG	LID_{En}	LID_{Vi}	SP-RG	LID_{En}	LID_{Th}
-	Lead-64	$20.7_{0.0}$	$99.6_{0.0}$	$18.9_{0.0}$	$0.0_{0.0}$	$100.0_{0.0}$	$16.5_{0.0}$	$0.0_{0.0}$	$99.6_{0.0}$	$22.1_{0.0}$	$0.0_{0.0}$	$100.0_{0.0}$	$15.9_{0.0}$	$0.0_{0.0}$	$97.6_{0.0}$
XXL	Prompt	$43.4_{0.4}$	$92.0_{0.5}$	$37.4_{0.2}$	$2.9_{1.5}$	$95.9_{1.5}$	$29.2_{1.0}$	$9.1_{2.4}$	$84.4_{1.8}$	$38.0_{0.0}$	$1.8_{1.1}$	$96.4_{0.8}$	$37.4_{0.7}$	$13.5_{2.0}$	$75.5_{1.5}$
XXL	Prompt, trans-test	-	-	$37.0_{0.4}$	$0.0_{0.0}$	$98.9_{0.2}$	$30.4_{0.4}$	$0.0_{0.0}$	$93.2_{0.3}$	$37.5_{0.1}$	$0.0_{0.0}$	$99.9_{0.1}$	$28.7_{0.5}$	$0.0_{0.0}$	$100.0_{0.0}$
XXL	Prompt, trans-train	-	-	$38.1_{1.5}$	$0.0_{0.0}$	$98.8_{0.2}$	$31.3_{0.2}$	$0.0_{0.0}$	$94.3_{0.8}$	$39.2_{0.1}$	$0.0_{0.0}$	$100.0_{0.0}$	$37.1_{0.3}$	$0.0_{0.0}$	$100.0_{0.0}$
XXL	Prompt, sup	$43.4_{0.4}$	$92.0_{0.5}$	$41.0_{0.1}$	$0.0_{0.0}$	$99.3_{0.1}$	$33.5_{0.3}$	$0.0_{0.0}$	$92.5_{0.5}$	$38.8_{0.3}$	$0.6_{0.4}$	$96.7_{0.9}$	$45.0_{0.1}$	$0.1_{0.1}$	$99.6_{0.3}$
XXL	Prompt, sup-all	$41.0_{0.4}$	$90.4_{0.7}$	$40.4_{0.1}$	$0.2_{0.3}$	$98.1_{0.2}$	$33.3_{0.2}$	$0.1_{0.1}$	$91.4_{1.6}$	$39.5_{0.1}$	$0.4_{0.3}$	$98.3_{0.6}$	$44.8_{0.7}$	$0.0_{0.0}$	$100.0_{0.0}$
XXL	Model	$46.7_{0.1}$	$94.4_{0.8}$	$35.5_{0.6}$	$9.1_{3.1}$	$86.0_{3.1}$	$27.2_{0.4}$	$19.7_{2.5}$	$57.5_{2.8}$	$34.0_{0.5}$	$14.8_{3.5}$	$79.1_{3.5}$	$30.1_{0.5}$	$32.7_{6.6}$	$16.8_{3.6}$
XXL	Model, trans-test	-	-	$38.9_{0.1}$	$0.0_{0.0}$	$98.9_{0.1}$	$32.9_{0.2}$	$0.0_{0.0}$	$93.1_{1.3}$	$39.2_{0.4}$	$0.0_{0.0}$	$99.5_{0.4}$	$31.7_{0.4}$	$0.0_{0.0}$	$100.0_{0.0}$
XXL	Model, trans-train	-	-	$41.6_{0.0}$	$0.4_{0.0}$	$98.5_{0.0}$	$34.9_{0.1}$	$0.0_{0.0}$	$95.4_{0.6}$	$41.4_{0.2}$	$0.0_{0.0}$	$100.0_{0.0}$	$38.7_{0.5}$	$0.0_{0.0}$	$100.0_{0.0}$
XXL	Model, sup	$46.7_{0.1}$	$94.4_{0.8}$	$43.8_{0.2}$	$0.1_{0.2}$	$99.2_{0.6}$	$36.6_{0.1}$	$0.0_{0.0}$	$95.5_{1.0}$	$42.0_{0.2}$	$0.0_{0.0}$	$99.7_{0.1}$	$48.8_{0.5}$	$0.0_{0.0}$	$99.9_{0.2}$
XXL	Model, sup-all	$47.1_{0.0}$	$93.8_{0.8}$	$44.9_{0.1}$	$0.0_{0.0}$	$98.8_{0.5}$	$37.6_{0.2}$	$0.1_{0.2}$	$93.7_{1.0}$	$43.8_{0.2}$	$0.0_{0.0}$	$99.7_{0.2}$	$50.2_{0.1}$	$0.0_{0.0}$	$100.0_{0.0}$
Small	Prompt	$24.5_{0.2}$	$82.8_{0.9}$	$19.2_{0.1}$	$3.3_{0.7}$	$77.4_{2.7}$	$11.4_{0.1}$	$29.6_{1.7}$	$10.1_{1.0}$	$21.4_{0.2}$	$2.3_{0.7}$	$87.2_{2.4}$	$14.9_{0.4}$	$45.9_{2.6}$	$3.3_{0.4}$
Base	Prompt	$29.8_{0.4}$	$85.2_{0.9}$	$23.8_{0.1}$	$5.6_{2.9}$	$82.8_{2.9}$	$14.3_{0.6}$	$39.2_{3.2}$	$24.5_{5.9}$	$24.4_{0.6}$	$6.0_{1.4}$	$81.9_{2.4}$	$17.3_{0.6}$	$34.3_{1.5}$	$33.5_{2.5}$
Large	Prompt	$35.3_{0.3}$	$89.4_{0.7}$	$28.8_{0.1}$	$3.6_{0.9}$	$91.1_{0.8}$	$20.4_{0.8}$	$13.3_{2.6}$	$74.6_{3.8}$	$29.3_{0.3}$	$3.0_{0.5}$	$89.3_{2.0}$	$24.7_{1.0}$	$29.0_{7.6}$	$45.9_{9.3}$
XL	Prompt	$38.4_{0.2}$	$90.5_{0.4}$	$33.4_{0.3}$	$2.4_{0.8}$	$94.8_{0.5}$	$25.3_{0.3}$	$9.6_{1.5}$	$79.3_{1.6}$	$34.1_{0.4}$	$3.4_{0.3}$	$91.9_{0.5}$	$33.2_{0.4}$	$19.8_{5.5}$	$66.0_{6.8}$
XXL	Prompt	$43.4_{0.4}$	$92.0_{0.5}$	$\mathbf{37.4}_{0.2}$	$2.9_{1.5}$	$95.9_{1.5}$	$\mathbf{29.2}_{1.0}$	$9.1_{2.4}$	$84.4_{1.8}$	$\mathbf{38.0}_{0.0}$	$1.8_{1.1}$	$96.4_{0.8}$	$\mathbf{37.4}_{0.7}$	$13.5_{2.0}$	$75.5_{1.5}$
Small	Model	$38.0_{0.4}$	$92.4_{0.4}$	$21.3_{0.3}$	$72.4_{2.2}$	$11.9_{1.6}$	$14.6_{0.2}$	$82.5_{1.0}$	$0.0_{0.1}$	$22.5_{0.1}$	$39.9_{4.8}$	$34.9_{2.9}$	$17.3_{0.1}$	$78.1_{4.2}$	$0.1_{0.1}$
Base	Model	$39.6_{0.4}$	$92.0_{1.0}$	$22.4_{0.3}$	$51.0_{10.2}$	$25.3_{7.4}$	$15.3_{0.2}$	$79.0_{11.7}$	$0.7_{1.1}$	$21.9_{0.3}$	$41.0_{10.5}$	$34.0_{8.3}$	$17.9_{0.2}$	$89.0_{0.8}$	$0.3_{0.2}$
Large	Model	$42.6_{0.2}$	$92.8_{0.3}$	$27.8_{0.6}$	$9.9_{4.1}$	$77.7_{5.4}$	$17.4_{1.0}$	$50.0_{3.2}$	$21.4_{3.8}$	$27.2_{0.8}$	$13.6_{6.0}$	$69.2_{7.6}$	$25.9_{0.7}$	$36.5_{4.6}$	$35.4_{2.1}$
XL	Model	$45.0_{0.3}$	$94.2_{1.6}$	$31.9_{0.5}$	$15.7_{2.6}$	$76.2_{3.9}$	$19.7_{0.6}$	$61.6_{15.8}$	$19.3_{13.1}$	$29.8_{0.7}$	$21.6_{3.5}$	$64.8_{4.5}$	$25.6_{0.5}$	$54.7_{14.5}$	$24.9_{13.7}$
XXL	Model	$46.7_{0.1}$	$94.4_{0.8}$	$\mathbf{35.5}_{0.6}$	$9.1_{3.1}$	$86.0_{3.1}$	$\mathbf{27.2}_{0.4}$	$19.7_{2.5}$	$57.5_{2.8}$	$\mathbf{34.0}_{0.5}$	$14.8_{3.5}$	$79.1_{3.5}$	$\mathbf{30.1}_{0.5}$	$32.7_{6.6}$	$16.8_{3.6}$
Base	Prompt, l=1	$19.7_{0.1}$	$75.9_{0.8}$	$18.0_{0.1}$	$0.9_{0.2}$	$89.0_{0.2}$	$14.8_{0.1}$	$2.1_{0.3}$	$83.4_{0.2}$	$19.1_{0.1}$	$0.2_{0.0}$	$92.4_{0.5}$	$19.2_{0.1}$	$3.3_{2.4}$	$80.2_{12.2}$
	Prompt, l=10	$25.1_{0.1}$	$84.4_{1.2}$	$21.6_{0.2}$	$0.3_{0.1}$	$91.7_{1.0}$	$\mathbf{17.2}_{0.5}$	$6.6_{3.1}$	$76.4_{6.4}$	$23.5_{0.1}$	$0.5_{0.2}$	$94.8_{2.1}$	$\mathbf{21.0}_{0.5}$	$11.8_{0.8}$	$53.7_{2.1}$
	Prompt, l=100	$29.8_{0.4}$	$85.2_{0.9}$	$\mathbf{23.8}_{0.1}$	$5.6_{2.9}$	$82.8_{2.9}$	$14.3_{0.6}$	$39.2_{3.2}$	$24.5_{5.9}$	$\mathbf{24.4}_{0.6}$	$6.0_{1.4}$	$81.9_{2.4}$	$17.3_{0.6}$	$34.3_{1.5}$	$33.5_{2.5}$
	Prompt, l=1000	$32.4_{0.3}$	$86.2_{1.1}$	$22.0_{0.9}$	$8.8_{2.0}$	$77.1_{4.3}$	$14.0_{0.5}$	$41.9_{4.6}$	$19.5_{3.9}$	$23.3_{0.5}$	$8.4_{0.8}$	$79.4_{1.4}$	$16.3_{1.0}$	$47.5_{3.7}$	$18.9_{4.9}$
XXL	Prompt, l=1	$37.8_{0.1}$	$88.8_{0.6}$	$35.0_{0.3}$	$0.0_{0.0}$	$99.2_{0.2}$	$29.8_{0.2}$	$0.3_{0.2}$	$93.7_{0.5}$	$36.3_{0.2}$	$0.0_{0.0}$	$98.7_{0.3}$	$36.4_{1.7}$	$0.1_{0.1}$	$99.3_{0.2}$
	Prompt, l=10	$41.2_{0.4}$	$89.8_{1.0}$	$\mathbf{37.6}_{0.3}$	$0.0_{0.0}$	$99.2_{0.5}$	$\mathbf{31.3}_{0.1}$	$1.0_{0.1}$	$92.7_{1.1}$	$\mathbf{38.3}_{0.1}$	$0.0_{0.0}$	$99.5_{0.2}$	$\mathbf{41.2}_{0.2}$	$2.0_{1.2}$	$91.3_{1.3}$
	Prompt, l=100	$43.4_{0.4}$	$92.0_{0.5}$	$37.4_{0.2}$	$2.9_{1.5}$	$95.9_{1.5}$	$29.2_{1.0}$	$9.1_{2.4}$	$84.4_{1.8}$	$38.0_{0.0}$	$1.8_{1.1}$	$96.4_{0.8}$	$37.4_{0.7}$	$13.5_{2.0}$	$75.5_{1.5}$
	Prompt, l=1000	$40.8_{2.2}$	$92.0_{2.0}$	$35.7_{1.0}$	$1.5_{0.5}$	$97.3_{0.6}$	$28.8_{0.4}$	$7.0_{2.1}$	$85.9_{2.7}$	$37.0_{1.2}$	$0.8_{0.6}$	$97.8_{1.3}$	$37.8_{1.2}$	$7.4_{0.1}$	$81.7_{3.4}$

Table 10: Summarization quality (SP-ROUGE) and language identification confidence scores (LID) across model sizes and methods (numbers in the subscript indicate the standard deviation across 3 random seeds). Our results suggest that WIKILINGUA-0 is a challenging task for both MODELTUNING and PROMPTTUNING. As model size increases, PROMPTTUNING usually produces better results than MODELTUNING when there is a significant language shift at inference time. Longer prompts help to better learn the English summarization task. However, the increased capacity leads the model to forgets other languages.

Nigeria, is clustered nearby Haitian Creole (HT), whose grammar derives from Igbo.

E Mitigating catastrophic forgetting

Table 11 shows our experiment results for different approaches described in §3.1. As can been seen, mixing in unlabeled multilingual data (MIX-UNSUP/MIX-UNSUP-ALL) helps prevent catastrophic forgetting for MODELTUNING. Intermediate tuning (IT-GIGAWORD/IT-LM) does not result in reliable gains. Finally, factorized prompts (FP-EN/ FP) lead to an improvement in target language accuracy, and an improvement in SP-RG in cases where vanilla PROMPTTUNING shows the worst performance.

F Intermediate tuning

As an adaptation step, we perform model or prompt tuning on an intermediate task before training on WIKILINGUA-0. Intermediate tuning has been used to boost performance on English tasks for both MODELTUNING (Phang et al., 2019; Vu et al., 2020) and PROMPTTUNING (Vu et al., 2022), and has been successfully applied to the zero-shot cross-lingual transfer setting (Phang et al., 2020; Maurya et al., 2021) for MODELTUNING. Maurya et al. (2021) show that intermediate tuning on an auxiliary unsuper-

Size	Method	EN		FR			RU			VI			TH		
		SP-RG	LID_{EN}	SP-RG	LID_{EN}	LID_{FR}	SP-RG	LID_{EN}	LID_{RU}	SP-RG	LID_{EN}	LID_{VI}	SP-RG	LID_{EN}	LID_{TH}
-	LEAD-64	$20.7_{0.0}$	$99.6_{0.0}$	$18.9_{0.0}$	$0.0_{0.0}$	$100.0_{0.0}$	$16.5_{0.0}$	$0.0_{0.0}$	$99.6_{0.0}$	$22.1_{0.0}$	$0.0_{0.0}$	$100.0_{0.0}$	$15.9_{0.0}$	$0.0_{0.0}$	$97.6_{0.0}$
BASE	PROMPT	$29.8_{0.4}$	$85.2_{0.9}$	$23.8_{0.1}$	$5.6_{2.9}$	$82.8_{2.9}$	$14.3_{0.6}$	$39.2_{3.2}$	$24.5_{5.9}$	$24.4_{0.6}$	$6.0_{1.4}$	$81.9_{2.4}$	$17.3_{0.6}$	$34.3_{1.5}$	$33.5_{2.5}$
BASE	PROMPT, MIX-UNSUP	$23.5_{0.1}$	$83.4_{1.4}$	$20.3_{0.8}$	$0.2_{0.3}$	$95.5_{2.3}$	$16.1_{0.3}$	$6.7_{4.0}$	$77.5_{8.2}$	$23.1_{0.3}$	$0.3_{0.2}$	$96.6_{1.0}$	$20.9_{0.8}$	$4.1_{2.1}$	$76.9_{7.0}$
BASE	PROMPT, MIX-UNSUP-ALL	$23.0_{0.4}$	$81.1_{1.6}$	$19.3_{1.0}$	$0.2_{0.2}$	$92.0_{2.2}$	$16.5_{1.0}$	$2.1_{1.1}$	$87.1_{1.5}$	$22.7_{0.8}$	$0.8_{0.8}$	$96.1_{1.5}$	$21.4_{0.7}$	$2.8_{1.1}$	$84.5_{5.7}$
BASE	PROMPT, IT-GIGAWORD	$30.8_{0.2}$	$86.0_{0.5}$	$24.0_{0.2}$	$3.1_{1.6}$	$85.5_{0.7}$	$15.1_{0.6}$	$41.7_{5.5}$	$25.8_{7.9}$	$24.8_{0.0}$	$6.5_{1.2}$	$81.4_{0.9}$	$19.3_{0.3}$	$33.5_{3.1}$	$28.4_{4.0}$
BASE	PROMPT, IT-LM	$30.3_{0.2}$	$86.2_{0.2}$	$\mathbf{24.2}_{0.1}$	$5.4_{2.0}$	$83.0_{2.3}$	$15.7_{0.5}$	$36.0_{2.1}$	$34.4_{5.2}$	$24.3_{0.2}$	$6.2_{1.8}$	$81.0_{1.4}$	$17.8_{1.4}$	$41.2_{6.6}$	$24.1_{7.7}$
BASE	PROMPT, FP-EN	$28.9_{0.2}$	$84.7_{0.3}$	$23.2_{0.4}$	$3.4_{0.9}$	$86.4_{1.7}$	$16.1_{0.6}$	$26.4_{3.3}$	$48.5_{4.2}$	$\mathbf{24.8}_{0.7}$	$4.3_{1.3}$	$84.6_{3.1}$	$19.4_{0.4}$	$28.6_{4.4}$	$32.4_{4.0}$
BASE	PROMPT, FP	$28.9_{0.2}$	$84.7_{0.3}$	$23.6_{0.4}$	$1.2_{0.7}$	$93.0_{1.3}$	$\mathbf{17.8}_{0.8}$	$15.3_{2.1}$	$64.5_{1.1}$	$24.7_{0.5}$	$2.1_{0.8}$	$90.0_{2.2}$	$\mathbf{21.1}_{0.8}$	$19.8_{5.1}$	$40.0_{13.5}$
BASE	MODEL	$39.6_{0.4}$	$92.0_{1.0}$	$22.4_{0.3}$	$51.0_{10.2}$	$25.3_{7.4}$	$15.3_{0.2}$	$79.0_{11.7}$	$0.7_{1.1}$	$21.9_{0.3}$	$41.0_{10.5}$	$34.0_{8.3}$	$17.9_{0.2}$	$89.0_{0.8}$	$0.3_{0.2}$
BASE	MODEL, MIX-UNSUP	$39.9_{0.8}$	$93.6_{1.4}$	$\mathbf{30.0}_{0.5}$	$2.6_{0.6}$	$90.5_{1.1}$	$\mathbf{24.1}_{0.8}$	$6.6_{0.9}$	$73.5_{4.2}$	$\mathbf{31.1}_{0.2}$	$3.2_{0.1}$	$90.4_{0.7}$	$25.2_{0.4}$	$16.2_{2.9}$	$56.8_{0.6}$
BASE	MODEL, MIX-UNSUP-ALL	$39.7_{0.3}$	$93.0_{1.3}$	$29.3_{0.1}$	$5.5_{1.0}$	$85.5_{0.5}$	$21.7_{0.4}$	$26.9_{4.6}$	$41.8_{2.2}$	$29.6_{0.5}$	$8.9_{1.3}$	$78.5_{1.6}$	$\mathbf{25.5}_{0.3}$	$24.6_{2.2}$	$43.9_{7.2}$
BASE	MODEL, IT-GIGAWORD	$40.5_{0.3}$	$93.0_{0.7}$	$20.8_{0.1}$	$86.0_{4.4}$	$4.0_{1.1}$	$15.5_{0.1}$	$92.5_{0.3}$	$0.0_{0.0}$	$21.2_{0.1}$	$81.1_{3.5}$	$6.3_{1.6}$	$17.3_{0.1}$	$93.4_{2.2}$	$0.0_{0.0}$
BASE	MODEL, IT-LM	$40.9_{0.2}$	$93.3_{1.1}$	$18.7_{0.8}$	$61.8_{43.7}$	$9.9_{11.6}$	$15.7_{0.1}$	$90.7_{1.8}$	$0.2_{0.2}$	$21.3_{0.2}$	$65.9_{5.1}$	$14.4_{3.6}$	$17.2_{0.1}$	$92.4_{1.9}$	$0.1_{0.2}$
XXL	PROMPT	$43.4_{0.4}$	$92.0_{0.5}$	$\mathbf{37.4}_{0.2}$	$2.9_{1.5}$	$95.9_{1.5}$	$29.2_{1.0}$	$9.1_{2.4}$	$84.4_{1.8}$	$\mathbf{38.0}_{0.0}$	$1.8_{1.1}$	$96.4_{0.8}$	$37.4_{0.7}$	$13.5_{2.0}$	$75.5_{1.5}$
XXL	PROMPT, MIX-UNSUP	$41.9_{0.2}$	$90.1_{0.8}$	$36.9_{1.1}$	$1.1_{0.6}$	$96.8_{0.9}$	$26.2_{3.0}$	$14.5_{10.1}$	$72.3_{13.1}$	$37.2_{0.8}$	$1.3_{0.9}$	$96.0_{2.1}$	$37.4_{2.0}$	$16.2_{9.9}$	$74.0_{10.8}$
XXL	PROMPT, MIX-UNSUP-ALL	$41.2_{1.6}$	$91.2_{1.1}$	$37.2_{0.9}$	$1.5_{0.6}$	$97.2_{0.4}$	$\mathbf{30.0}_{0.4}$	$3.9_{1.1}$	$89.7_{1.5}$	$37.3_{1.1}$	$1.8_{0.8}$	$96.0_{1.7}$	$\mathbf{38.2}_{2.0}$	$9.7_{6.6}$	$81.9_{6.4}$
XXL	PROMPT, IT-GIGAWORD	$43.5_{0.1}$	$92.6_{0.2}$	$36.6_{0.5}$	$3.9_{1.1}$	$94.2_{1.5}$	$24.0_{1.1}$	$37.5_{5.7}$	$54.6_{6.8}$	$37.2_{0.2}$	$5.1_{1.4}$	$93.2_{1.5}$	$32.2_{1.7}$	$33.7_{6.0}$	$52.8_{7.0}$
XXL	PROMPT, IT-LM	$42.9_{0.1}$	$92.8_{2.2}$	$36.4_{0.5}$	$6.6_{1.2}$	$91.4_{2.0}$	$26.9_{1.8}$	$17.9_{7.2}$	$73.1_{7.8}$	$37.2_{0.3}$	$2.2_{0.7}$	$94.3_{1.4}$	$\mathbf{38.2}_{0.2}$	$6.5_{0.4}$	$83.1_{1.8}$
XXL	PROMPT, FP-EN	$40.8_{2.6}$	$90.0_{3.0}$	$36.5_{1.2}$	$2.5_{1.6}$	$95.5_{1.9}$	$27.9_{1.2}$	$9.4_{7.8}$	$81.3_{9.3}$	$37.5_{1.3}$	$0.4_{0.3}$	$98.0_{0.8}$	$36.7_{0.6}$	$10.8_{6.1}$	$79.7_{9.4}$
XXL	PROMPT, FP	$40.8_{2.6}$	$90.0_{3.0}$	$35.7_{1.6}$	$2.2_{1.5}$	$96.0_{1.4}$	$29.0_{0.5}$	$5.3_{5.1}$	$85.3_{5.1}$	$36.1_{2.6}$	$0.6_{0.5}$	$97.6_{1.3}$	$36.9_{1.2}$	$9.0_{5.4}$	$80.8_{9.3}$
XXL	MODEL	$46.7_{0.1}$	$94.4_{0.8}$	$35.5_{0.6}$	$9.1_{3.1}$	$86.0_{3.1}$	$27.2_{0.4}$	$19.7_{2.5}$	$57.5_{2.8}$	$34.0_{0.5}$	$14.8_{3.5}$	$79.1_{3.5}$	$30.1_{0.5}$	$32.7_{6.6}$	$16.8_{3.6}$
XXL	MODEL, MIX-UNSUP	$46.7_{0.1}$	$95.5_{1.3}$	$\mathbf{39.5}_{0.1}$	$2.2_{0.4}$	$95.3_{0.9}$	$\mathbf{32.3}_{0.3}$	$6.2_{1.1}$	$78.7_{2.7}$	$\mathbf{38.3}_{0.2}$	$1.5_{0.7}$	$96.2_{0.9}$	$\mathbf{32.4}_{0.7}$	$17.0_{2.1}$	$32.4_{8.4}$
XXL	MODEL, MIX-UNSUP-ALL	$46.3_{0.1}$	$94.5_{0.3}$	$38.2_{0.1}$	$2.4_{0.5}$	$95.2_{0.7}$	$29.7_{0.2}$	$13.0_{0.3}$	$73.0_{0.7}$	$37.8_{0.2}$	$2.5_{0.9}$	$93.4_{0.5}$	$31.8_{0.6}$	$17.4_{1.6}$	$45.2_{4.0}$
XXL	MODEL, IT-GIGAWORD	$46.3_{0.2}$	$95.6_{0.4}$	$24.8_{0.2}$	$81.2_{1.9}$	$9.9_{1.4}$	$20.8_{0.1}$	$78.8_{1.9}$	$3.8_{0.2}$	$31.3_{0.3}$	$32.5_{4.1}$	$54.4_{4.1}$	$22.8_{0.2}$	$87.2_{0.9}$	$2.4_{0.6}$
XXL	MODEL, IT-LM	$46.3_{0.0}$	$95.2_{0.9}$	$25.7_{0.1}$	$72.7_{5.2}$	$16.6_{4.5}$	$22.4_{0.2}$	$59.5_{1.6}$	$15.7_{1.3}$	$19.5_{4.1}$	$82.0_{14.2}$	$10.8_{12.5}$	$25.1_{0.1}$	$66.5_{1.1}$	$6.4_{1.4}$

Table 11: Summarization quality (SP-ROUGE) and language identification confidence scores (LID) across two model sizes (BASE and XXL) and methods (numbers in the subscript indicate the standard deviation across 3 random seeds). Mixing in unlabeled multilingual data (MIX-UNSUP/MIX-UNSUP-ALL) helps prevent catastrophic forgetting for MODELTUNING. Intermediate tuning (IT-GIGAWORD/IT-LM) does not result in reliable gains. Factorized prompts (FP-EN/ FP) lead to an improvement in target language accuracy, and an improvement in SP-ROUGE in cases where vanilla PROMPTTUNING shows the worst performance.

vised task from the target language is helpful in conjunction with freezing some model components for MODELTUNING. Previous work has used an auxiliary task designed to be close to the main task, while we simply use mC4 data. For each target language we create a causal, left-to-right LM task by providing no context, i.e., the encoder's input is empty (**IT-LM**). To further explore the effect of continued training on English data, we include an additional experiment where the GIGAWORD (Graff et al., 2003) summarization dataset is used as the intermediate task (**IT-GIGAWORD**).[20]

Intermediate tuning does not give reliable gains: As can be seen in Table 11, intermediate tuning on English summarization (IT-GIGAWORD) improves English performance, but generally hurts XGEN capabilities. For MODELTUNING, it exacerbates catastrophic forgetting and harms overall performance across all model sizes. For PROMPTTUNING, English intermediate tuning provides small gains at BASE size, but is harmful at XXL size. Intermediate tuning on an LM task in the target language (IT-LM) has a neutral or negative effect in most cases, running somewhat counter to the findings of Maurya et al. (2021).[21] Compared to directly mixing in unlabeled multilingual data, intermediate tuning has little benefit on language accuracy. This smaller effect is to be expected, given that the final stage of English-only training is still ample opportunity to overfit on English and catastrophically forget other languages.

[20] We found that additional tuning was helpful for intermediate tuning on large datasets. As such, we performed 200,000 steps during tuning on an intermediate task and selected the best prompt checkpoint based on validation performance on that task.

[21] Note, however that their unsupervised task was designed to be well-aligned with their downstream tasks of choice.