Quality or Quantity? On Data Scale and Diversity in Adapting Large Language Models for Low-Resource Translation

Vivek Iyer   Bhavitvya Malik*   Pavel Stepachev*
Pinzhen Chen   Barry Haddow   Alexandra Birch
School of Informatics, University of Edinburgh
`vivek.iyer@ed.ac.uk`

Abstract

Despite the recent popularity of Large Language Models (LLMs) in Machine Translation (MT), their performance in low-resource languages (LRLs) still lags significantly behind Neural Machine Translation (NMT) models. In this work, we explore what it would take to adapt LLMs for the low-resource setting. Particularly, we re-examine the role of two factors: a) the importance and application of parallel data, and b) diversity in Supervised Fine-Tuning (SFT). Recently, parallel data has seen reduced use in adapting LLMs for MT, while data diversity has been embraced to promote transfer across languages and tasks. However, for low-resource LLM-MT, we show that the opposite is true for both considerations: a) *parallel data* is critical during both pre-training and SFT; b) diversity tends to cause *interference* instead of transfer. Our experiments with three LLMs across two low-resourced language groups—Indigenous American and North-East Indian—reveal consistent trends, underscoring the generalizability of our findings. We believe these insights will be valuable for scaling to massively multilingual LLM-MT models that can effectively serve LRLs.

1 Introduction

Large Language Models (LLMs) have been dominating recent research in Machine Translation (MT), showing good few-shot prompting (Garcia et al., 2023; Hendy et al., 2023) and stronger instruction-tuning (Alves et al., 2024; Xu et al., 2024) performances—recently even outperforming commercial Neural Machine Translation (NMT) models (Kocmi et al., 2024). However, LLM translation for low-resource languages (LRLs) still lags significantly behind NMT models (Robinson et al., 2023; Zhu et al., 2024b). While the strong performance of LLMs on high-resource languages can be

---

*denotes equal contribution

	Base LLM	#Tokens	#Langs
Zhang et al., 2024a	From scratch	283B	2
Fujii et al., 2024	Llama2	100B	2
Lu et al., 2024	Llama{2,3}	~82B[†]	101
Alves et al., 2024	Llama2	20B	11
Xu et al., 2024	Llama2	20B	6
Ours	Mistral/Llama3	0.7B	12

Table 1: Comparing data scales of previous works with ours, in terms of pre-training token counts, the base LLM (if pre-training continued from one) and how many languages this spanned. [†]Estimated from the reported sentence count assuming 100 tokens per sentence.

attributed to the skewed language distribution during pre-training and the unintentional consumption of parallel data at scale (Briakou et al., 2023), no such relief exists for LRLs. This leads to the main question motivating this paper: *What would it take to adapt LLMs for low-resource MT*?

Recent work on LRL translation with LLMs has explored using resources like multilingual lexicons (Lu et al., 2023), word alignments (Mao and Yu, 2024) or linguistic tools (Zhang et al., 2024b). While effective, reliance on such tools hinders ease of extensibility across languages. Instead, in this work, we take inspiration from research done for high-resource translation with LLMs, where a 2-stage training paradigm of Continued Pre-Training (CPT), followed by small-scale Supervised Fine-Tuning (SFT; (Xu et al., 2024; Alves et al., 2024)) has been successful. Aiming to adapt this framework for low-resource MT, we re-examine the role of two factors influencing the performance of translation LLMs: a) *how best to leverage parallel data*, and b) *the interplay between diversity and transfer* during SFT (also known as 'instruction tuning').

Recently, the role of **parallel data at scale**, long viewed as fundamental to the success of NMT models, has come into question in the era of LLM-MT systems. Motivated by the modest gains of training on 300M parallel sentences (Yang et al., 2023),

1393

and the surprising benefits of scaling *down* during SFT (Zhou et al., 2023), subsequent works have used only tens of thousands of human-written bitext for LLM-MT (Zhang et al., 2023b; Alves et al., 2024; Xu et al., 2024)—with SFT scaling laws further showing the early plateau of LLM-MT performance (Zhang et al., 2024a). Even more surprisingly, Zhu et al. (2024a) showed MT abilities emerging with just 32 SFT examples! However, these explorations concern LLMs pre-trained on several billions of tokens in the languages in question. We revisit these notions for low-resource MT and work with languages having datasets that are 2-3 orders of magnitude smaller. In Table 1 we compare the scale of the datasets used in our work and related research. We discover that for low-resource MT, *parallel data is critical* not just during CPT, but even more so during SFT—in direct contrast with research on high-resource languages (HRLs).

Next, **diversity** in tasks, prompts, and datasets during SFT has been shown to significantly improve model performance across a range of tasks (Mishra et al., 2022; Chung et al., 2024). MT instructions have been shown to not just boost translation performance in unseen languages (Muennighoff et al., 2023), but also enhance LLM capabilities across diverse multilingual generation tasks (Ranaldi and Pucci, 2023; Zhu et al., 2023). Inspired by these findings, we study if SFT diversity could benefit low-resource LLM-MT systems too. By conducting experiments across a range of tasks and language pairs with SFT datasets of varying compositions, we establish that *diversity leads to negative interference* and fine-tuning on multilingual MT is the optimal strategy. Further, we observe that training for more epochs on MT data is more effective than curating and training on a diverse dataset of the same size.

Our contributions are thus as follows:

1. In contrast to findings for high-resource LLM-MT (Xu et al., 2024), we observe that for LRLs, LLMs benefit hugely from scale of parallel data, during both CPT and SFT stages

2. Linguistic and task diversity during SFT leads to negative interference for LRL LLM-MT, with focused multilingual MT fine-tuning for more epochs being the most effective recipe.

To ensure the generalizability of our findings, we conduct 2 sets of experiments training multilingual LLMs on different sets of languages: a) 11 Indigenous American and b) 4 North East Indian languages, wherein the former follows a Latin script and the latter includes languages that do not. Our focus in this work is on the eng/spa → X directions since generation in an LRL is known to be a much harder task than in an HRL like English or Spanish, and we are interested in studying the challenges involved. We experiment with 3 base LLMs of varying sizes—Gemma 2B (Gemma Team et al., 2024), Mistral 7B (Jiang et al., 2023), and Llama 3 8B (Dubey et al., 2024), and observe that findings are mostly consistent across these models.

By applying our findings to 2-stage training, our methods achieved a +16.5 average chrF++ improvement over few-shot prompting—with the largest gains coming from the 8 least-resourced American languages in our setup, all of which have about 10K-50K parallel sentences each. We hope that the findings of this work will be useful when scaling to LLMs that can effectively translate into lower-resource languages.

2 Related Work

High-Resource Translation with LLMs There has been considerable interest in using LLMs as MT systems recently. Following initial success in prompting LLMs for high-resourced pairs (Vilar et al., 2023; Garcia et al., 2023; Hendy et al., 2023; Zhang et al., 2023a; Iyer et al., 2023), subsequent works have attempted to train LLMs on parallel data at scale (Yang et al., 2023; Lu et al., 2024), but these yielded modest gains and underperformed smaller encoder-decoder baselines such as NLLB-200 (Costa-jussà et al., 2022). Zhang et al. (2024a) showed through scaling laws for SFT that LLMs pre-trained at the order of 50B-300B tokens saturate in MT performance with 20K-30K instructions. Concurrently, Xu et al. (2024) discovered excess parallel data washed out LLM knowledge, so they proposed a 2-stage paradigm called ALMA that involved pre-training on scaled-up monolingual data, followed by SFT on much smaller high-quality bitext (~60K lines). ALMA outperformed NLLB-200. Following their success, Alves et al. (2024) adopted the ALMA framework to train Tower 7B for 10 high-resourced languages, outperforming ALMA and also matching GPT-4.

It is worth noting that Xu et al. (2024) did not include parallel data during CPT. However, inspired by research showing LLMs unintentionally consume parallel data at scale (Briakou et al., 2023),

several works have included it at the order of several billions of tokens for the top 10-20 high-resource languages (Anil et al., 2023; Wei et al., 2023; Fujii et al., 2024; Alves et al., 2024). Concurrent work has explored pre-training on synthetic, code-switched parallel data for 101 languages (Lu et al., 2024) with a total of 400M sentences. This work explores the impact of parallel data exclusively for low-resource performance and experiments with 1M–13M parallel sentences (50M–750M tokens) during pre-training, two orders of magnitude smaller than prior work.

Low-Resource Explorations in LLM-MT LLMs have been shown to perform poorly in low-resource MT (Robinson et al., 2023; Zhu et al., 2024b). In response, there have been efforts to leverage external resources in the MT pipeline, including multilingual lexicons (Lu et al., 2023), rule-based linguistic tools (Zhang et al., 2024b), word alignments (Mao and Yu, 2024) and even entire grammar books (Reid et al., 2024). However, such approaches create dependencies on resources and hinder extensibility across languages. Instead, we focus on optimal data utilisation strategies during CPT and SFT, prioritizing extensibility.

Cross-Lingual Instruction Tuning There has been a body of work exploring multilingual instruction tuning that have touched on diversity and data scale, but most of it is limited to HRLs. MT examples were shown to improve cross-lingual generation (Ranaldi and Pucci, 2023; Zhu et al., 2023), while Chen et al. (2024) showed multilingual SFT on machine-translated Alpaca datasets matches or beats monolingual tuning. Kew et al. (2023) and Shaham et al. (2024) showed a small quantity of multilingual SFT data can improve cross-lingual generation capabilities in medium and high-resource languages, while Zhu et al. (2024a) showed only 32 examples in HRLs suffice to elicit MT capabilities from LLMs. In our work, we hypothesize that these findings have the common denominator of pre-training on HRLs at scale, and show that when moving to LRLs, these trends reverse and diversity is no longer beneficial.

3 Approach

We now describe our efforts to adapt the widely-used ALMA framework, originally designed for fine-tuning LLMs for HRL translation (Xu et al., 2024; Alves et al., 2024), for MT in LRLs.

3.1 Stage 1: Continued Pre-training (CPT)

CPT on Monolingual Data The objective of this stage is to 'teach' an LLM to model LRLs, which are scarce in the pre-training corpus. We conduct CPT on monolingual data with the standard Causal Language Modelling objective. We train with low-rank adaptation (LoRA) and attach rank 8 adapters to query and value matrices (Hu et al., 2022). We also fine-tune input and output embeddings.

CPT on Parallel Data In a scenario where monolingual data is scarce, it is crucial to investigate the most effective way to use parallel data—which, for our indigenous American languages, was found to surprisingly be more abundant than the former[1]. We investigate 3 methods of mixing all available parallel and monolingual data:

1. **All Mono**: Here, we merge monolingual data with only *the target side* of all available bitext—essentially using it as extra monolingual data

2. **Mono + parallel (concat)**: Here, we merge monolingual data with *concatenated source-target pairs* from parallel data. We prepend source and target language codes before concatenation and, following Guo et al. (2024), use a newline delimiter to separate them.

3. **Mono + parallel (separate)**: To ablate the impact of concatenation, we provide the source and target sides of parallel data as *separate* sentences, and shuffle with monolingual data.

We depict our approach in Figure 1. In the first technique, the motivation is that it might be hard for the LLM to learn to model concatenated sequences, given that they were likely scarce in the original pre-training corpus, with the added challenge of pre-training on 'new' low-resource languages. On the other hand, if the model is able to adapt to concatenated sequences, it could make the LLM more adjusted to the task of translation. Finally, the third method verifies whether the results of the 'concatenated' model are due to the concatenation itself, rather than simply being exposed to additional tokens in the source language. We use this terminology for all experiments in this work.

While these methods control how parallel data is incorporated in pre-training, we are also interested

---

[1]While monolingual online data is scarce, many translations of constitutions, articles etc. from Spanish do exist
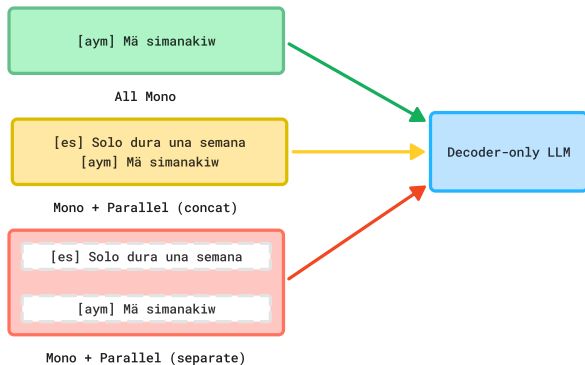
Figure 1: Strategies explored for incorporating parallel data during Continued Pre-Training. We show a Spanish (es) to Aymara (aym) example from our parallel data.

in adjusting the ratio of parallel to monolingual data in the corpus, particularly for the *'concat'* method given pre-training on 100% parallel data can be suboptimal (Alves et al., 2024). To get a desired mixing ratio for *'concat'*, we include bitext only until it comprises a given percentage of the training corpus. Once this threshold is crossed, we use the *'all mono'* method to include it as monolingual data instead. Next, we use temperature sampling (Arivazhagan et al., 2019) to control the language-wise distribution in our monolingual and parallel pre-training data, since these are quite heterogeneous and certain languages are extremely low-resourced. We sample monolingual and parallel data independently if using the 'concat' method, else we just mix them all together and shuffle at the instance level to create our final pre-training corpus.

3.2 Stage 2: Supervised Fine-Tuning (SFT)

Next, we fine-tune with LoRA on supervised instruction data and detail the tasks explored below. Note that we convert all instructions to the standard Alpaca format, and compute loss on the target tokens only (Taori et al., 2023).

Low-Resource MT Given our use case, the most intuitive task to include would be MT itself. The `instruction` for each example is chosen randomly from a set of translation prompts (Table 11), while the `input` and `output` fields are the source and target sentences respectively.

High-Resource MT Apart from MT data in the LRLs, we also experiment with adding HRL MT data since it is more abundant and of higher quality, known to be important during SFT (Xu et al., 2024). To explore the impact of transfer learning, we work with HRL data that is in some way related to the source/target language, e.g. Spanish-English

data for experiments on Spanish-X. Instructions are formatted in the same way as the LRL data.

General-Purpose Instruction Tuning Apart from MT data, we also explore adding widely used general-purpose instruction tuning datasets, such as Alpaca (Taori et al., 2023) and Aya (Singh et al., 2024), and use data from high-resource languages (comprising the source side in our tasks) to improve the model's overall instruction-following capabilities. However, for the most part, we are unable to find similar data in the LRLs we experiment on.

Synthetic Cross-Lingual QA (XQA) We do not find any instruction tuning data for most LRLs, so we follow Iyer et al. (2024) to create synthetic Question Answering (QA) data. Starting from a parallel sentence pair (X, Y), where X is from an HRL (in our case, English/Spanish) and Y is from an LRL, we prompt an LLM (Mixtral-8x7B-Instruct (Jiang et al., 2024) in this work) to generate a question Q for which X would be an answer. Since X and Y are semantically equivalent, Y is treated as the answer to question Q. We add a requirement at the end of Q to generate in the target language. Thus, we use (Q, Y) as synthetic cross-lingual instruction data. We provide the templates used for generating XQA examples in Table 12.

4 Experiments and Discussions

We fine-tune 2 separate sets of multilingual LLMs for 2 different language groups to facilitate evaluation on test sets from 2 different low-resource MT shared tasks: AmericasNLP 2024 (Ebrahimi et al., 2024) and the Indic track of WMT 2023 (Pal et al., 2023). The former involves 11 Indigenous Central & South American languages, while the latter focuses on 4 North-East (NE) Indian languages. The first group includes Aymara (aym), Bribri (bzd), Asháninka (cni), Chatino (ctp), Guarani (grn), Huichol (hch), Nahuatl (nhe), Otomi (ote), Quechua (quy), Shipibo-Konibo (shp) and Tarahumara (tar). The second consists of Khasi (kha), Meitei (mni), Mizo (lus) and Assamese (asm). Our motivation in choosing these languages was to experiment with LRLs containing both Latin (American) and non-Latin (Indic) scripts; that also had widely used, high-quality test sets. We use the former for our main experiments and replicate the most interesting baselines in the Indic languages.

Lang	#Tokens	Lang	#Tokens	Lang	#Tokens
aym	23.4M	bzd	2.6M	cni	2.2M
ctp	5.4M	grn	37.6M	hch	3.5M
nhe	32.1M	oto	23.6M	quy	45.1M
shp	3.3M	tar	2.3M	**Total**	**181.3M**
eng	9.8M	spa	27.9M	**Replay**	**37.6M**

(a) Indigenous American Languages

Lang	#Tokens	Lang	#Tokens	Lang	#Tokens
asm	1.1B	kha	39.3M	lus	165.1M
mni	16.5M	eng	7.3M	**Total**	**1.3B**

(b) North-East Indian Languages

Table 2: Monolingual data statistics, with token counts calculated using the Llama3 8B tokenizer. English and Spanish are included as replay data. Note that LRL token counts overestimate data sizes, due to poor tokenization, and cannot be directly compared with HRLs.

4.1 Data

Monolingual Data Table 2 shows the token counts for the monolingual data collected for the 2 language groups. We note that the American languages are very low-resource, with 6 of 11 having 5M tokens or less. The Indic languages have relatively more data, with Assamese being medium-resourced but still likely low-resource in the original LLM pre-training corpus. Assamese and Meitei follow the Assamese-Bengali script, while Mizo and Khasi use the Latin script. Using the Llama3 tokenizer, we observe average fertilities of 2.87 and 3.83 for the American and Indic languages respectively, almost 3x that of high-resource languages, illustrating the under-representation of non-Latin scripts in SOTA LLMs. Finally, we include some data in English and Spanish as *replay data* to prevent catastrophic forgetting (Ibrahim et al., 2024).

Parallel data We curate parallel data from various sources, for use in both CPT and SFT. Tables 3 and 4 show the sizes for the American and Indian languages respectively. Given that our primary exploration is for the American languages with limited spa-X data, we also sample eng-X and por-X from OPUS (Tiedemann, 2012). Note the heavily skewed language distribution, with the 3 HRLs constituting 80% of spa-X data and 96% of the overall data. The lesser skew for spa-X is due to the efforts of AmericasNLP to collect data for these pairs and the prevalence of Spanish in Latin American countries. A similar skew also exists for the Indic languages, with English-Khasi being the least-resourced pair. We list sources for all curated

data, along with cleaning steps, in Appendix A.

4.2 Evaluation

To evaluate MT into the 11 American languages, we use AmericasNLP'23 validation sets (Ebrahimi et al., 2024) containing spa-X translation pairs. For the Indic pairs, we use the WMT 2023 test sets from the Indic track (Pal et al., 2023) which consist of eng-X pairs. Both evaluation datasets are multi-domain, as are the curated monolingual and parallel corpora. We show test set statistics in Table 5. Given the absence of neural metrics for these languages, we evaluate using ChrF++ (Popović, 2017), since both the American and Indic languages are morphologically rich wherein chrF++ is particularly effective (Popović, 2017). We use SacreBLEU (Post, 2018) for computing this. We also report confidence intervals with bootstrap resampling (Koehn, 2004), which we implement for the multilingual setting by computing the macro-average across all languages for each resample, and then computing the mean and variance across all resamples. We report the standard deviation as the confidence interval.

4.3 Experimental Settings

For temperature sampling our data, we use $\tau = 30$ for CPT and $\tau = 80$ for SFT. We used a batch size of 8 and gradient accumulation every 16 steps. We used a learning rate of 1e-4, with a cosine scheduler and a warmup ratio of 3%. We train all models on bf16 precision for 1 epoch. We use Llama-Factory (Zheng et al., 2024) for training and evaluating all models, with Deepspeed ZeRO3 (Rasley et al., 2020) for distributed training. For inference, we used a batch size of 16 with greedy decoding, since we found higher beam sizes yielded minimal gains.

4.4 Foundational Results

We first establish the importance of fine-tuning input and output embeddings in Table 6 and then show our foundational results for the American languages in Table 7, using 5-shot prompting as a baseline. Our findings are:

1. **Fine-tuning embeddings is critical.** Across the board, we observe that fine-tuning embeddings along with LoRA modules yields huge gains (Table 6), almost doubling chrF++ scores, indicating that this step is crucial to helping LLMs adapt to these new languages. Given this, we expect that full-weight fine-tuning would perform better, but stick to

	Total	HRL	LRL	aym	bzd	cni	ctp	grn	hch	nhe	oto	quy	shp	tar
spa	1M	0.8M	0.2M	442K	8K	20K	4K	80K	21K	57K	16K	226K	62K	28K
por	1.9M	1.9M	15K	3.6K	0	6.9K	0	410K	0	8K	0	1520K	0	0
eng	5.8M	5.8M	36K	1053K	0	13.9K	0	2489K	0	22K	0	2271K	0	0
Total	8.8M	8.5M	0.3M	1.5M	7.7K	41K	4.1K	3M	21K	87K	16K	4.0M	62K	28K

Table 3: Parallel data sentence counts for American languages, from source to each target language. HRL and LRL refer to the 3 high-resource languages (Aymara, Guarani, Quechua) and the other 8 low-resource ones respectively.

Pair	eng-lus	eng-asm	eng-kha	eng-mni
#Sents	6.5M	5.0M	25K	443K

Table 4: Indic Parallel Data Sizes (Sentence counts)

Indigenous American						NE Indic	
Pair	#Lines	Pair	#Lines	Pair	#Lines	Pair	#Lines
spa-aym	996	spa-nhe	672	spa-quy	996	eng-asm	2000
spa-bzd	996	spa-oto	599	spa-shp	996	eng-kha	1000
spa-cni	883	spa-gn	995	spa-tar	995	eng-mni	1000
spa-ctp	499	spa-hch	994			eng-lus	2000

Table 5: Evaluation data statistics for the low-resourced American and Indic language experiments in this paper.

Fine-Tuned Modules	Gemma 2B	Mistral 7B	Llama3 8B
LoRA only	4.5 ± 0.1	8.8 ± 0.2	8.3 ± 0.2
LoRA + embeddings	9.6 ± 0.3	15.6 ± 0.4	15.6 ± 0.4

Table 6: Impact of fine-tuning input/output embeddings along with LoRA adapters is shown. Both models follow the "CPT *all mono*, SFT" recipe from Table 7.

Method	Gemma 2B	Mistral 7B	Llama3 8B
5-shot prompting	2.8 ± 0.1	5.1 ± 0.1	3.9 ± 0.1
SFT only	8.7 ± 0.2	16.3 ± 0.4	14.8 ± 0.4
CPT *all mono*, SFT	9.6 ± 0.3	15.6 ± 0.4	15.6 ± 0.4
CPT *mono+parallel*, SFT	**10.1 ± 0.3**	**16.7 ± 0.4**	**17.2 ± 0.4**

Table 7: chrF++ scores for spa-X LLM-MT in the American languages, for LLMs of varying sizes. Both LoRA modules and embeddings are fine-tuned. For SFT, all models use 500K spa-X MT examples. Confidence estimates are computed using bootstrap resampling. *'mono+parallel'* uses concatenated bitext.

LoRA fine-tuning for cost-efficiency reasons. We fine-tune embeddings for all future results.

2. **Choosing larger base LLMs is crucial.** For under-represented (zero-resource) languages, is it more effective to train smaller LLMs with larger vocabularies (and thus, improved fertility), or vice versa?[2] We note that Gemma 2B has the largest vocabulary (256K tokens), followed by Llama3 (128K) and Mistral (32K tokens), resulting in improved fertility (2.36 vs 2.87 for Llama3/Mistral). Regardless, the larger models, Mistral 7B and Llama3 8B, vastly outperform the Gemma 2B model, suggesting fine-tuning smaller vocabulary LLMs like Mistral might be a better option from both cost and performance standpoints.

3. **SFT alone is effective, but CPT+SFT yields best results.** While SFT yields large gains over prompting, combining CPT and SFT seems optimal for both Gemma and Llama3, although the gap is smaller than that of high-resource MT (Xu et al., 2024), ostensibly due to the difference in scale. For Mistral, SFT alone seems to suffice—we hypothesize that

this might be due to the smaller vocabulary being effectively fine-tuned on SFT data alone.

4. **Pre-training on parallel data yields major gains consistently.** Lastly, we observe that pre-training on a mixture of *concatenated* monolingual and parallel data (*'mono + parallel'*) yields statistically significant gains over converting both as monolingual data (ie. *'all mono'*; refer Section 3.1). This trend is consistent for all 3 LLMs, with larger gains for the more effective models Mistral and Llama3. We note that for 'mono+parallel', we mix parallel and monolingual data in a 1:1 ratio, since we observed it worked best empirically, and show in Table 14 how increasing the ratio of parallel data during CPT monotonically improves performance. Given these gains, we explore the importance of bitext for low-resource LLM-MT further in Section 4.5.

4.5 Analysis: Importance of Parallel Data

How important is concatenated parallel data at various scales of low-resource pre-training?

---

[2]We found it prohibitively expensive to fine-tune Gemma 7B which has a larger vocabulary *and* a larger capacity.
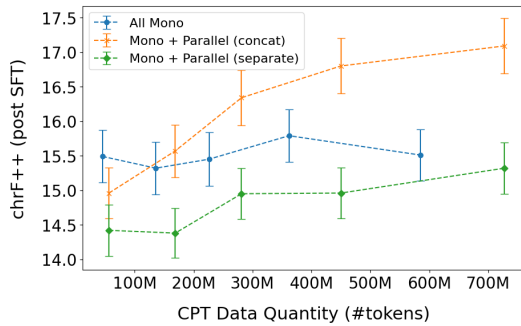
Figure 2: Comparing Llama3 8B models pre-trained on monolingual data alone versus those included parallel data too—concatenated, or as separate texts at various scales. All models were pre-trained on 1M, 3M, 5M, 8M, and 13M sentences respectively, and markers denote the corresponding token counts. The y-axis shows chrF++ post SFT on 500K spa-X MT data for 1 epoch.

Having seen improvements in pre-training on the entire corpus (consisting of 13M 'mono+parallel' sentences, or 730M tokens) in Table 7, we now study the importance of parallel data as we scale *down*—an important consideration when one moves to even lower-resource settings. We pre-train on subsets of varying sizes and mix monolingual and parallel data in 3 ways: 'All mono', 'mono+parallel (concat)' and 'mono+parallel (separate)', as defined in Section 3.1). We fine-tune all these pre-trained models on the same SFT dataset: 500K spa-X MT instructions, and plot the resulting chrF++ scores in Figure 2, including error bars from bootstrap resampling. We note that 'all mono' has different markers than the others, as the token counts on including both source and target-side data in the corpus are obviously larger than only the latter. We find that a) starting at around 5M sentences (~300M tokens), it is *consistently advantageous* to include concatenated parallel data during pre-training. b) Given 'mono+parallel (separate)' severely underperforms, we establish that it is **concatenation** that adapts the LLM for the task of MT, not the extra data alone. Our findings complement those of Alves et al. (2024), who also observe gains from pre-training on parallel data in the 1B to 20B tokens range, using 10 high-resourced European languages. In contrast, our focus here is on investigating the *minimum data threshold* at which CPT leveraging parallel data becomes beneficial.

How does scaling LRL parallel data during SFT impact performance? We now turn our focus to scaling during SFT, which has not yielded gains in high-resource LLM-MT after 20K-30K sentences

(Zhang et al., 2024a). Motivated by our previous results, we re-examine this question for low-resource LLM-MT in Figure 3, wherein we evaluate at steps of 50K until ~1M SFT instructions, the point at which spa-X MT data runs out. Our findings are:

1. **Overall MT quality improves steadily with scale.** Unlike high-resource LLM-MT, we observe that, despite fluctuations, MT quality continues to grow until 1M sentences, particularly for LRLs, which have a steeper slope.

2. **The gains of CPT on parallel data carry over to SFT scaling, particularly for HRLs.** We observe that CPT on a mixture of concatenated monolingual and parallel data yields the largest initial gains, followed by CPT on monolingual-only data. We note that the former is particularly beneficial for HRLs, which have likely consumed a lot of parallel data during pre-training, and this helps them maintain a huge lead at lower SFT scales, suggesting that *concatenated parallel data teaches the LLM the task of translation and helps it adapt to MT more naturally during SFT*. Given the prevalence of bitext in LLM pre-training corpora (Briakou et al., 2023), the gains of concatenation would help explain why few-shot prompting (Garcia et al., 2023) and tiny-scale SFT (32 examples; c.f. Zhu et al. (2024a)) can elicit MT in the highest-resource languages!

3. **LRLs need SFT scaling much more than HRLs** So, while HRLs benefit hugely from CPT on bitext, the opposite is true for LRLs— the scarce amount of parallel data observed during CPT (see Table 3) is likely not enough to outperform 'all mono' pre-training. Instead, LLMs are far more responsive to the scale of LRL MT data during SFT, showing consistent performance improvement. Meanwhile, MT quality plateaus for the HRLs[3] (Figure 3b), suggesting that the 'less is more' (Zhou et al., 2023) trend popular in high-resource LLM-MT does *not* hold for LRLs, where *scale continues to remain the most effective option.*

For generalizability, we also provide scaling graphs for Mistral 7B in Figure 7 and report trends similar to Llama3. Moreover, we also compute

---

[3]although Aymara, Guarani and Quechua *are* LRLs for the original pre-training corpus of Mistral/Llama3

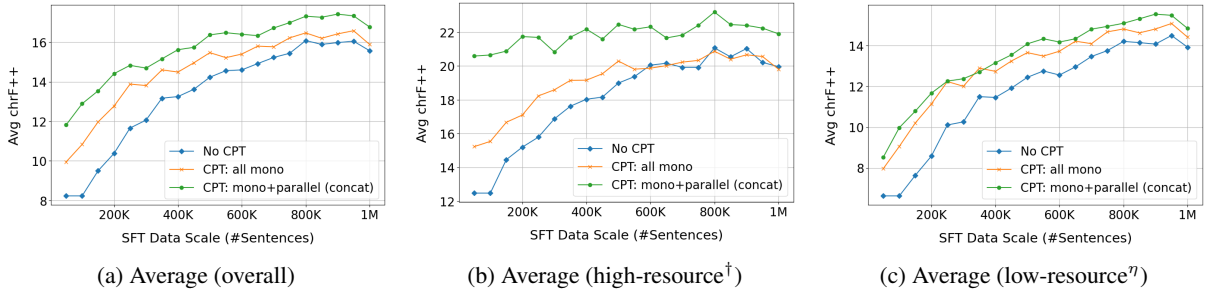(a) Average (overall)   (b) Average (high-resource†)   (c) Average (low-resource$^{\eta}$)

Figure 3: Scaling up Llama3 8B models with different CPT recipes (no CPT, CPT with monolingual data, and CPT with a mixture of monolingual and parallel data) on MT data for the American languages. †'High-resource' refers to the relatively higher-resourced languages in our low-resource setup (Aymara, Guarani and Quechua) while the other 8 are grouped as low-resource$^{\eta}$.
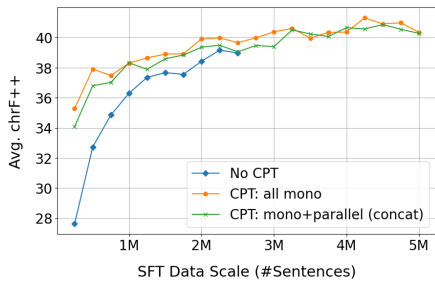

Figure 4: Scaling up Llama 3 8B models with the 3 CPT recipes for the 4 Indic languages, until 5M sentences. We were forced to stop training 'No CPT' at 2.5M sentences, constrained by budget.
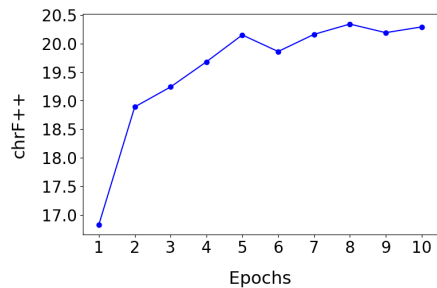

Figure 5: Epoch vs performance graph for low-resource LLM-MT. We use the entire 1M spa-X MT dataset, and plot average chrF++ for the Indigenous American languages, using Llama3 (Mono+Parallel (concat)) model.

scaling graphs for the 4 Indic languages in Figure 4 using the Llama3 model, until 5M eng-X sentences. Here, we observe that the gap between 'CPT: mono+parallel (concat)' and 'CPT (all mono)' is relatively lesser, and only significant until 500K sentences, at which point they start becoming comparable. This might be because the Indic parallel corpora used for CPT is relatively less diverse, consisting of only eng-X pairs in 4 languages, whereas our parallel data for the American languages is more heterogeneous (see Table 3)—which we observed to yield more gains in our preliminary experiments. Thus, we expect a more diverse corpus for the Indic languages to yield greater long-term gains, but we leave the verification of this to future work. In other respects, the trends are quite similar to the American languages: the gains of CPT carry over to SFT here too and interestingly, MT quality continues to improve until 5M SFT examples, once again in direct contrast to high-resourced LLM-MT research.

More epochs: Scale through repetition Finally, given the consistent gains from scale, we evaluate how effective data repetition could be during

SFT. We train for up to 10 epochs on the entire 1M spa-X MT dataset and plot our results in Figure 5. We note a strong monotonic gain of +3.3 average chrF++ until 5 epochs, with the largest coming from the 1st to the 2nd epoch (+2.0 chrF++). While the graph does plateau later (likely due to overfitting) this shows how simply training for more epochs can be an easy way to boost performance, given the data constraints for LRLs.

4.6 Analysis: Diversity in SFT is not *always* helpful

We now look at the importance of *diversity* in SFT, which plays a significant role in boosting LLM performance during general-purpose fine-tuning (Sanh et al., 2022; Longpre et al., 2023), but has also been favoured in training LLM-MT systems (Alves et al., 2024). Alves et al. (2024) observed improved performance on non-MT tasks by including non-MT instructions, but mixed results on MT quality—concluding that the "transfer/interference relations between tasks are complex". In this work, we try to tease apart these interference relations and study the effect of non-MT tasks on low-resource translation in depth. In particular, we look at 3 kinds of

1400

Method	All	HRL	LRL
Bilingual LLM-MT	14.51	24.55	10.75
Multilingual LLM-MT	18.73	23.59	16.90

Table 8: Average chrF++ of SFT on a single language pair (bilingual) vs all language pairs (multilingual). As before, "HRL" includes Aymara, Guarani and Quechua, while the other eight are grouped under LRL.

diversity: *prompt*, *linguistic*, and *task* diversity, and examine their impact on MT quality when mixed in varying proportions in the SFT dataset.

Bilingual or Multilingual fine-tuning? First, we ask whether given some multilingual low-resource data and a fixed FLOPS budget, would it be beneficial to do SFT on multiple, diverse low-resource MT pairs (multilingual LLM-MT) or only on a single pair (bilingual LLM-MT)? For this experiment, we simply concatenate all available data for the multilingual setting, while the bilingual setting consists of fine-tuning separate models for each pair, with the base LLM being constant in both cases: Llama3 (CPT with mono+parallel data). We show our results in Table 8. We observe that, on average, multilingual SFT outperforms the bilingual models by +4 chrF++ points. Looking closer, we notice that these gains mostly come from the lower-resourced languages (detailed results in Table 13). This is potentially because the bilingual setting has too little data for effective FT in these languages, while the multilingual option offers better transfer and scale. For higher-resourced pairs, the opposite is true: the performance is slightly lower, ostensibly due to negative interference. This is not unlike conventional multilingual NMT models, such as NLLB, wherein high-resource MT can often be worse than bilingual baselines.

On Linguistic, Task and Prompt Diversity Inspired by the previous results showing target-side diversity in MT pairs boosts performance, we now broaden the scope of diversity during SFT. In Table 9a, we show varying mixtures of SFT data that all have the same size (500K examples) but are composed of different tasks. For *prompt diversity*, we ablate randomly sampling from a list of potential prompt templates (listed in Table 11) versus using a constant one (the first in Table 11). We observe a statistically significant gain in MT quality, similar to general-purpose SFT (Longpre et al., 2023).

Next, we study the more interesting question of *linguistic diversity*: can data in other MT pairs

SFT Mixture	chrF++
Prompt Diversity	
500K spa-X MT (Same prompt)	16.18±0.37
500K spa-X MT (Random prompts)	**17.22±0.40**
Linguistic Diversity	
166K spa-X MT + 166K eng-X MT + 166K por-X MT	14.26±0.37
250K spa-eng MT + 250K spa-X MT	15.73±0.38
500K spa-X MT only	**17.22±0.40**
Task Diversity	
250K spa-X MT + 250K XQA	15.45±0.39
250K spa-X MT + 250K Aya (spa)	15.68±0.39
250K spa-X MT (x 2 epochs)	17.20±0.40
500K spa-X MT	**17.22±0.40**

(a) Prompt, Linguistic and Task Diversity in American pairs

SFT Mixture	chrF++
250K spa-X MT + 250K Aya (asm, mni)	26.91 ± 0.42
250K spa-X MT + 250K Alpaca (eng)	28.02 ± 0.41
500K spa-X MT only	**29.57 ± 0.43**

(b) Task Diversity in Indic pairs

Table 9: Exploring interference due to diversity in SFT for our best Llama3 8B model (CPT on both monolingual and parallel data) on the American and Indic languages. The dataset size (example count) is prepended before each task. Scores shown are average chrF++.

transfer for test languages? In one baseline, we divide our SFT dataset equally into spa-X, eng-X, and por-X MT examples, into the American languages (statistics in Table 3). In another baseline, we use Spanish-English data from ParaCrawl (Bañón et al., 2020) and combine a 250K sentence subset with 250K pairs from our usual spa-X MT data. We find that this type of linguistic diversity leads to interference, and significantly underperforms the 500K spa-X MT baseline. This suggests that while *target diversity* in related languages might help performance, source diversity or unrelated high-resource languages like English may not.

Then, we look at *task diversity*: can non-MT tasks that elicit better instruction-following and general reasoning capabilities in the source or target language, benefit LLM-MT in LRLs? We mix Aya and XQA (Section 3.2) instructions with a 250K subset of spa-X MT examples. We also include an ablation that, in place of curating non-MT data, merely trains for 2 epochs on the 250K spa-X MT subset. Here we also discover that general-purpose tasks lead to interference and that training for 2 epochs on 250K MT examples is a more effective strategy comparable to 1 epoch on 500K.

We also look at the impact of task diversity on the Indic languages in Table 9b. Here, we are able
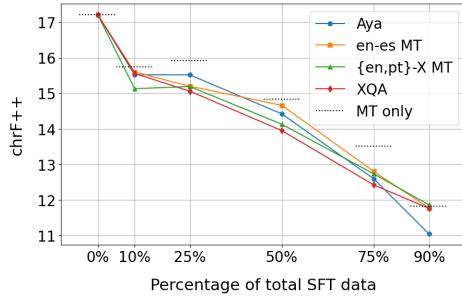
Figure 6: Mixing varying percentages of SFT tasks with spa-X MT, versus the impact on spa-X chrF++, for 500K examples. 'MT only' is a topline which *only* uses the spa-X MT data excluding all other tasks. Thus, it uses *lesser* data relatively and helps to estimate the interference of non-MT tasks on MT quality.

to find general-purpose instruction-tuning data in Assamese and Meitei, as mentioned in Section 3.2. Interestingly, including data in these *target languages*—the languages we want the model to be better at generating in—degrades performance the most! A similar amount of Alpaca data in English reduces performance far less, suggesting that reasoning in LRLs is such a hard task, that even data where the model *should* learn to generate in the languages of interest, is *worse* than generic English instruction-following data. Here, too, providing 500K spa-X examples is the most optimal strategy, underscoring the generalizability of our findings.

Does non-MT SFT always cause interference? Finally, we seek to establish conclusively if the negative results of diversity in Table 9a arise from factors like task mismatch or lack of quality; or if interference is simply *the norm* when fine-tuning on non-MT data for low-resourced LLMs. In Figure 6, we plot 5 graphs: the first 4 are graphs showing non-task-specific data (including XQA, Aya, {eng/por}-X MT and eng-spa MT) mixed in varying proportions with spa-X MT. Thus, for example, the second point of each plot represents 10% of that task mixed with 90% of spa-X MT, with the total always being 500K examples. We also include an ablation (dotted lines), which has the same quantity of spa-X MT data at each data point as other plots but without data from any other task listed.

We observe that *performance is always better using task-specific MT data*, and negative interference is indeed consistent. We conclude that transfer is challenging to achieve in low-resource settings, likely because at such scales, the LLM-MT model is still learning to generate translations, and reasoning is still a formidable challenge. Downstream

performance in such settings depends not on the diversity/composition of the SFT dataset, but only on the amount of LRL MT data provided—making optimizing for *quantity* the most effective strategy.

4.7 Discussion: LLMs vs NMT models

For the Indigenous American languages, the current SOTA systems are NordicAlps (Attieh et al., 2024) and DC-DMV (DeGenaro and Lupicki, 2024) that report average chrF++ scores of 26.73 and 23.76 respectively. Both are NMT models. NordicAlps trained a multilingual model from scratch on varying mixtures of eng-spa and spa-x data at scale, with their most major gains (+4 chrF++ points) coming from a novel redundancy-driven tokenization method. DeGenaro and Lupicki (2024) fine-tune NLLB-200 using a variety of parallel data sources, but unlike us, they generate a large amount of synthetic data. Our best-performing model, Llama3 with CPT on parallel data and 5 epochs of SFT, yields a score of 20.93 in the best setting. While this is a gain of +17.06 chrF++ from a 5-shot prompted Llama3, it is also almost 6 chrF++ points behind the SOTA. Since ours was not a shared task effort, we did not attempt explorations with tokenization or synthetic data, which are concurrent to our findings and would likely boost performance further. Overall, our results suggest that low-resourced LLM-MT systems, while promising, are still behind the curve compared to SOTA NMT models.

5 Conclusion

In this paper, we approach low-resource LLM-MT from a data-centric perspective and study performance along two axes: i) the size of parallel data and ii) diversity in SFT tasks. Through experiments and analyses, we conclude that quantity plays a dominant role in downstream performance. Specifically, parallel data drives performance significantly during both CPT and SFT, with HRLs and LRLs displaying different behaviours—making bitext a critical resource even in the modern era of LLM-MT. Moreover, we establish that diversity (on multiple fronts) consistently declines MT quality, with multilingual fine-tuning on task-specific data being the most effective option, reaffirming our previous findings on the value of scale. We hope these findings will be useful considerations when scaling to massively multilingual LLMs of the future.

Limitations

One of the limitations of this work is that due to the unavailability of robust neural metrics for LRLs, we are forced to use string-based metrics like chrF++ as our primary evaluation metric. We note that while this is not optimal, it is also not unlike most other related works on low-resource translation, and chrF++ has been shown to align reasonably well with human assessment scores for morphologically rich languages. Secondly, we only focus on two specific language groups in this work. We expect that expanding the scope to a massively multilingual setting might yield even larger improvements, owing to scale and transfer learning.

Acknowledgments

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

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*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Joseph Attieh, Zachary Hopton, Yves Scherrer, and Tanja Samardzic. 2024. System description of the nordicsalps submission to the americasnlp 2024 machine translation shared task. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 150–158.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. Four approaches to low-resource multilingual NMT: The Helsinki submission to the AmericasNLP 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 177–191, Toronto, Canada. Association for Computational Linguistics.

Dan DeGenaro and Tom Lupicki. 2024. Experiments in mamba sequence modeling and nllb-200 fine-tuning

for low resource multilingual machine translation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 188–194.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *arXiv preprint arXiv:2404.17790*.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *International Conference on Machine Learning*, pages 10867–10878. PMLR.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 639–649, Mexico City, Mexico. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

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*.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495, Singapore. Association for Computational Linguistics.

Vivek Iyer, Bhavitvya Malik, Wenhao Zhu, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Exploring very low-resource translation with LLMs: The University of Edinburgh's submission to AmericasNLP 2024 translation task. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 209–220, Mexico City, Mexico. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed? *arXiv preprint arXiv:2312.12683*.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad G, Varun Balan G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar,

Raj Dabre, and Mitesh M. Khapra. 2024. Indicllm-suite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. *arXiv preprint arXiv: 2403.06350*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024. Preliminary wmt24 ranking of general mt systems and llms. *Preprint*, arXiv:2407.19884.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *arXiv preprint arXiv:2305.06575*.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*.

Zhuoyuan Mao and Yen Yu. 2024. Tuning LLMs with contrastive alignment instructions for machine translation in unseen, low-resource languages. In *Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 1–25, Bangkok, Thailand. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao,

M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Leonardo Ranaldi and Giulia Pucci. 2023. Does the English matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183, Singapore. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

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. Multi-task prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2304–2317, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Jörg Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024a. When scaling meets LLM finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations*.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15654–15669, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, et al. 2023b. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024a. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? *arXiv preprint arXiv:2404.14122*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

A Data

A.1 Monolingual Data

We provide the detailed monolingual data statistics for the indigenous American languages used for CPT in Table 10. We collect this data from various sources. For the indigenous American languages, we use MADLAD-400 (Kudugunta et al., 2024), GLOT 500 (ImaniGooghari et al., 2023), Wikipedia, data curated by the University of Helsinki (De Gibert et al., 2023) and OCR data collected by Iyer et al. (2024). The English and Spanish data used for replay comes from Wikipedia and MADLAD-400 too. We use MADLAD-400 and GLOT 500 again for the Indic languages again, along with the verified split of Sangraha (Khan et al., 2024)—a large corpus for Indian languages.

A.2 Parallel Data

We provide the templates used for generating Machine Translation and XQA Instructions in Tables 11 and 12 respectively. For MT, we randomly use one of these prompts to create an instruction for doing SFT of our model. For XQA, the provided instructions are used to prompt Mixtral-8x-7B to generate a question, which is later used for creating the synthetic XQA datasets as described in Section 3.2.

We source the parallel data for generating these from the work of Iyer et al. (2024), which includes data sourced from AmericasNLP'23 official training datasets (Ebrahimi et al., 2024), NLLB-200 and FLORES-200 (Costa-jussà et al., 2022), OPUS (Tiedemann, 2009) and data curated by Helsinki in their 2023 submission (De Gibert et al., 2023).

A.2.1 Cleaning Parallel Data

Rule-based Filtering We clean parallel data by following standard filtering rules. First, we remove pairs with more digits or non-alphanumeric characters than alphabetic ones on either side. Then, we remove pairs where either the source or the target has less than 3 words or more than 120 words. To adjust for pairs containing partial or incomplete translations, we apply a combination of two filtering rules: a) for sentences with less than 25 words, the character length difference must be less than 65, and b) for those longer than 25 words, the character length ratio between either source and target, or vice versa, must be at most 1.55. Lastly, we filter our sentences with non-Latin characters in either sentence, as well as pairs with identical source and

target sentences to create our final training dataset.

Neural Quality Estimation We also experimented with using neural quality estimation using models like LASER-3 (Heffernan et al., 2022). LASER-3 supports 3 of the 11 American languages—the HRLs (Aymara, Guarani and Quechua)—and we used it for scoring parallel sentences for these languages. We then sorted them in order of decreasing quality, calculated as the cosine similarity of LASER representations. For the other 8 LRLs, we used random scoring to simulate standard shuffling behaviour. We found that training on this 'sorted' data actually performed *worse* than our default baseline trained on unscored corpora, consistently at various SFT scales. As none of the authors spoke these languages, we could not substantively explain why, but we hypothesize that LASER representations might not be very reliable for these very LRLs, and might introduce certain undesirable biases in the sorted data, making the default baseline more robust to different kinds of bitext. In practice, we found that the standard rule-based filtering approaches worked best and thus, we stuck to them for cleaning our data.

Finally, we note that both our monolingual and parallel corpora span a variety of domains similar to our test data. Also, due to the paucity of data, we use all available sources.

B Results

B.1 Parallel vs Monolingual Data Ratio

In Table 14, we find that mixing higher ratios of parallel data with monolingual data either performs comparably or improves performance. We do not go higher than 50% since our parallel data runs out at this stage, and to ensure higher ratios we would have to oversample the existing dataset—which would not lead to a fair comparison with the other baselines. Nevertheless, given the monotonic trend, it would be interesting to explore if mixing higher ratios of parallel data continues to improve the performance even more. However, it is likely that there is some ceiling as to how much parallel data one should mix. For instance, Alves et al. (2024) show how a baseline trained on 100% parallel data underperforms compared to mixing it with monolingual data.

B.2 Scaling up SFT: Mistral 7B

Figure 7 shows the scaling behaviour of the Mistral 7B model with different pre-training recipes (no

Language	#Sentences	#Tokens	Language	#Sentences	#Tokens	Language	#Sentences	#Tokens
Aymara (aym)	1M	23.4M	Quechua (quy)	1.9M	45.1M	Guarani (grn)	1.1M	37.6M
Nahuatl (nhe)	1.1M	32.1M	Otomi (oto)	0.6M	23.6M	English (eng)	0.4M	9.8M
Spanish (spa)	0.7M	27.9M	Shipibo-Konibo (shp)	0.1M	3.3M	Bribri (bzd)	0.1M	2.6M
Asháninka (cni)	0.2M	2.2M	Chatino (ctp)	0.3M	5.4M	Huichol (hch)	0.2M	3.5M
Tarahumara (tar)	0.2M	2.3M	**Total (Replay)**	**1.1M**	**37.6M**	**Total**	**7.8M**	**218.9M**

Table 10: Detailed monolingual data statistics for the American languages

Translation Instructions

1. Translate the following sentence from {src_lang} to {tgt_lang}.
2. Can you convert the following sentence from {src_lang} to {tgt_lang}.
3. Kindly translate this sentence from {src_lang} into {tgt_lang}.
4. Could you translate the following from {src_lang} to {tgt_lang}?
5. Proceed to translate the subsequent sentence from {src_lang} to {tgt_lang}.
6. Change the following sentence from {src_lang} to {tgt_lang}.
7. Render the sentence below from {src_lang} into {tgt_lang}.
8. Switch the following sentence from {src_lang} into {tgt_lang} language.
9. Rephrase the following sentence into {tgt_lang} from {src_lang}.
10. Transform the following text from {src_lang} to {tgt_lang}.
11. Can you restate the following sentence from {src_lang} in {tgt_lang}?
12. Please provide a translation for this sentence from {src_lang} to {tgt_lang}.
13. Adapt the following into {tgt_lang} from the original {src_lang}.
14. Translate the subsequent text from {src_lang} into the {tgt_lang} language.

Table 11: MT Instruction Templates used during Supervised Fine-Tuning (SFT)

XQA Instruction

"Consider this sentence: {input}\nWhat kind of specific instruction X could this be the unique answer to? Output ONLY the instruction, followed by a newline."

Table 12: Template used for generating XQA instructions

Method	Avg	es-aym	es-bzd	es-cni	es-ctp	es-gn	es-hch	es-nhe	es-oto	es-quy	es-shp	es-tar
Bilingual LLM-MT	14.5	**21.2**	6.1	12.3	8.4	**27.9**	15.0	**14.1**	9.5	24.6	14.5	6.1
Multilingual LLM-MT	**18.7**	20.2	**15.8**	**18.0**	**26.0**	25.8	**21.6**	18.2	**11.2**	**24.9**	**14.8**	**9.5**

Table 13: Comparison of Bilingual and Multilingual Es-X FT (MT only) Methods

CPT, CPT with monolingual data, and CPT with a mixture of monolingual and parallel data), where terminology is the same as that defined in Section 3.1. We observe that the trends for Mistral are largely similar to Llama3:

1. 'No CPT' underperforms the CPT baseline, but they become comparable at about ~500K sentences. As hypothesised in Section 4.4, due to a 4x larger vocabulary, Llama3 has 4x more parameters to fine-tune[4], meaning our

Llama3 models are likely more data hungry and retain the benefits of CPT over longer periods, but smaller models overfit sooner, leading to a shorter 'cross-over' threshold.

2. 'CPT (mono only)' consistently underperforms 'CPT (mono + parallel)', very similar to Llama3, lending further credence to our conclusion that concatenated parallel data adapts

---

[4]LoRA module parameters are negligible in comparison.

For fine-tuning input and output embeddings, Llama3 has $128K * 4096 * 2 \approx 1B$ parameters, whereas Mistral has $32K * 4096 * 2 \approx 250M$ parameters. LoRA parameters are on the order of 3M, for comparison.

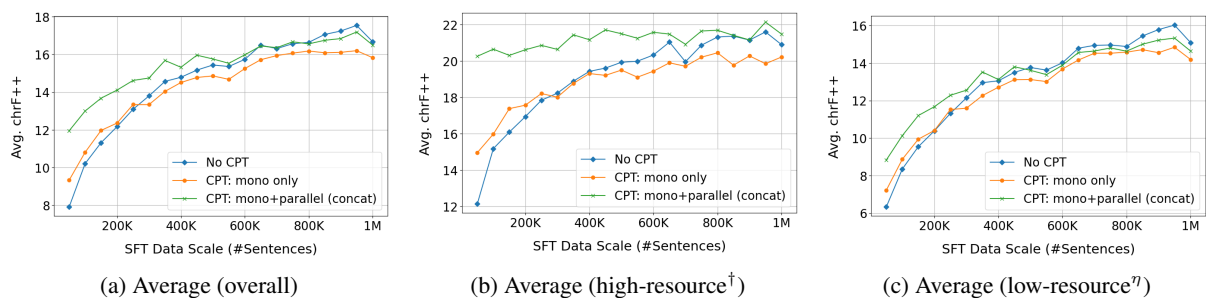| (a) Average (overall) | (b) Average (high-resource[†]) | (c) Average (low-resource[η]) |

Figure 7: Scaling up Mistral 7B models with different CPT recipes (no CPT, CPT with monolingual data, and CPT with a mixture of monolingual and parallel data) on MT data for the American languages. We observe that the trends are very similar to Llama3, with the sole exception that the gains of CPT diminish faster at around 500K sentences. This might be because Mistral, due to a 4x smaller vocabulary, gets fine-tuned effectively with less data. [†]'High-resource' refers to the relatively higher-resourced languages in our low-resource setup (Aymara, Guarani and Quechua) while the other 8 are grouped as low-resource[η].

Ratio	Gemma 2B	Mistral 7B	Llama3 8B
0%	9.6 ± 0.3	15.6 ± 0.4	15.6 ± 0.4
10%	10.1 ± 0.3	15.9 ± 0.4	16.5 ± 0.4
25%	10.1 ± 0.3	16.6 ± 0.4	16.6 ± 0.4
50%	10.1 ± 0.3	16.7 ± 0.4	17.2 ± 0.4

Table 14: Model Performance vs. Parallel Data Ratio. While the gap between ratios is not always statistically significant, it is clear that the trend is monotonic and having 50% parallel data is consistently better than 0% (ie. the fully monolingual setting).

the model to MT in a much better way

3. For HRLs, the performance plateaus quite quickly, while for LRLs SFT quality continues to grow with scale, once again following Llama3's trend

Our experiments with Mistral thus help in providing more evidence to support our claims with regard to the benefit of parallel data for low-resource LLM-MT systems.

B.3 Bilingual vs Multilingual LLM-MT: Detailed Results

We provide detailed, language-specific results for bilingual vs multilingual LLM-MT in Table 13. We observe that multilingual LLM-MT mostly outperforms bilingual baselines with the exception of 3 relatively higher-resourced languages (Aymara, Nahuatl and Guarani) where there is a bit of a gap.