Exploring Very Low-Resource Translation with LLMs: The University of Edinburgh's Submission to AmericasNLP 2024 Translation Task

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

Abstract

This paper describes the University of Edinburgh's submission to the AmericasNLP 2024 shared task on the translation of Spanish into 11 indigenous American languages. We explore the ability of multilingual Large Language Models (LLMs) to model low-resource languages by continued pre-training with LoRA, and conduct instruction fine-tuning using a variety of datasets, demonstrating that this improves LLM performance. Furthermore, we demonstrate the efficacy of checkpoint averaging alongside decoding techniques like beam search and sampling, resulting in further improvements. We participate in all 11 translation directions. Our models are released here: https://tinyurl.com/edi-amnlp24

1 Introduction

We participated in AmericasNLP 2024's shared task on machine translation (MT). It requires participants to translate from Spanish to 11 indigenous American languages: Aymara (aym), Bribri (bzd), Ashaninka (cni), Chatino (ctp), Guarani (gn), Huichol (hch), Nahuatl (nhe), Otomi (ote), Quechua (quy), Shipibo-Konibo (shp) and Tarahumara (tar). We adopted multilingual large language models (LLMs) and our workflow consists of data curation, continued pre-training, instruction tuning, and several decoding techniques. We submitted to all 11 translation directions.

We study and report the feasibility of using LLMs for very low-resource machine translation tasks. LLMs have recently been the focus of recent research interest, and in machine translation, they have demonstrated competitive or better performance against traditional neural MT systems in high-resource languages (Hendy et al., 2023; Robinson et al., 2023; Iyer et al., 2023; Alves et al., 2024). Nonetheless, research has shown that these models struggle in low-resource settings if used off-the-shelf (Robinson et al., 2023), and there has been limited exploration of adapting LLMs to extremely low-resource MT. Existing approaches rely on massively multilingual dictionaries (Lu et al., 2023) or a series of complex grammatical and linguistic tools (Zhang et al., 2024). Despite their effectiveness, a pitfall of these approaches is that it can be hard to scale them up to build multilingual, low-resource LLMs. Moreover, it is unclear how the (scarce) monolingual and parallel data available for these languages can be effectively utilised, and how recent developments in MT of high-resource languages (Xu et al., 2024; Alves et al., 2024) scale to very low-resource settings.

This work attempts to take a step towards answering these questions. We build multilingual LLMs for these indigenous American languages by fine-tuning Llama-2 7B (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023) and MaLA-500 (Lin et al., 2024). We explore continued pre-training with LoRA on various monolingual and parallel data sources. We then conduct instruction tuning using a variety of tasks and language pairs, and show this contributes to performance improvements in MT. We end this work by demonstrating how familiar techniques such as checkpoint averaging, beam search, and sampling help boost LLM performance for low-resourced translation as well.

2 Data

2.1 Monolingual data

We summarize statistics of the monolingual data used in our experiments in Table 1. We curate this data from various sources:

MADLAD-400 (Kudugunta et al., 2024): This is a manually audited general domain dataset sourced from Common Crawl, spanning 419 languages. Given this corpus has many dialects among

*denotes equal contribution

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 209–220
June 21, 2024

the American languages of interest, we create a dictionary[1] mapping each language to the ISO 639-3 codes of all its dialects, and download all of them. We remark on various strategies we tried for handling dialects in Section 3.1. We sample 150000 sentences from the English and Spanish splits to maintain comparable data quantities.

Glot500 (ImaniGooghari et al., 2023): This dataset belongs to multiple domains, covers 500 languages and spans multiple licences. We downloaded the publicly available version of this dataset from Hugging Face, for the languages of interest to us, and concatenated the train, dev, and test splits for these languages. We handled dialects similar to the MADLAD-400 corpus.

Wikipedia: We download Wikipedia dumps for the languages of interest and parse them with WikiExtractor (Attardi, 2015) for downstream use.

Helsinki'23 datasets (De Gibert et al., 2023): We reuse the monolingual data extracted by the winning team from the AmericasNLP 2023 Shared Task, University of Helsinki. We separate out the Bibles, UDHR, Wikipedia, and Miscellaneous (Misc) domains.

OCR data: In the pursuit of additional data, we utilized alternative external resources. We manually extracted[2] various text resources (summarised in Table 9 and classified them into groups and languages. The extracted files were converted to PDF format. Each page of the file was transformed into PNG format and upscaled to a resolution of 600 DPI. Our approach employed **ocrmac**[3](based on the Apple Vision Framework) for OCR. The methodology focused solely on bounding box text spans, without the application of sentence or paragraph restoration. We summarize statistics of the OCR data in Tables 3, 8, 9.

2.2 Instruction Tuning data

Inspired from Alves et al. (2024), we try to make our instruction tuning dataset as diverse as possible, and observe that multi-task instruction tuning yields performance gains on the singular task of Machine Translation as well. We summarize the statistics of our instruction tuning dataset in Table 4, and detail our sources as follows:

Aya (Singh et al., 2024): We use the Cohere Aya Dataset for the English, Portuguese and Spanish languages which consist of about 3.8K, 3.8K and 9K instructions respectively. The Aya Dataset consists of freshly created human annotations to existing prompts, as well as re-annotations by humans of machine-generated prompt completions. Given that this dataset relies strongly on human annotation, we include it in our instruction tuning dataset - even though the languages provided are not the indigenous American languages we are interested in. We could not find any data for these American languages in the Aya project.

MT Data: We use the official datasets provided by the organizers (`official`), the NLLB and the FLORES 200 corpora (Costa-jussà et al., 2022), the Helsinki'23 OPUS parallel corpora (De Gibert et al., 2023) as well as our own extraction of the OPUS dataset (Tiedemann, 2009) – from which we were able to extract more languages and pairs than the original Helsinki collection. For the NLLB corpus, which is sorted in decreasing order of scores indicating translation quality, we sample sentences from the top to ensure the highest quality sentences are chosen for instruction tuning. Finally, as far as possible, we try to ensure uniform sampling across all these languages and corpora to prevent imbalance.

Cross-lingual QA: We also generate synthetic cross-lingual instruction data using a powerful open-source LLM, Mixtral-8x7B-Instruct (Jiang et al., 2024), for data augmentation. Our generation process is illustrated in Figure 1. Given a translation pair (X, Y), where X is from a high-resource language and Y is from a low-resource language, we follow the prompt of Köksal et al. (2024) and ask Mixtral to generate a question Q based on X. As X and Y are semantically equivalent, Y is now used as the answer to the question Q. Finally, we add an instruction at the end of the prompt to generate in the target language. This is, thus, similar to a cross-lingual QA task - where the question is in a high-resource language, but the answer is in the indigenous American language and the LLM is instructed to generate its response in the latter. In this way, we use (Q, Y) as synthetic cross-lingual instruction data.

During training, we convert all our instruction-

[1] https://tinyurl.com/uedin-dialectsdict

[2] We are not speakers of any indigenous languages in this shared task.

[3] https://github.com/straussmaximilian/ocrmac v0.1.6 with parameters: recognition_level="accurate", language_preference=["es-ES", "en-US", "ru-RU", "fr-FR", "de-DE"]

Language	Total	MADLAD 400	GLOT 500	Wikipedia	Helsinki'23 (Bibles)	Helsinki'23 (Misc)	Helsinki'23 (UDHR)	Helsinki'23 (Wikipedia)	*OCR (multilingual)*[†]
Aymara (ay)	779835	58572	355229	19272	61182	0	120	16081	269379
Bribri (bzd)	41123	0	0	0	7659	0	0	0	33464
Asháninka (cni)	74964	0	0	0	0	0	0	0	74964
Chatino (ctp)	113415	0	0	0	23764	0	0	0	89651
Guarani (gn)	531478	98351	97470	39546	7849	0	102	39593	248567
Huichol (hch)	68411	0	0	0	7936	373	0	0	60102
Nahuatl (nhe)	547187	84647	23615	0	70988	0	91	8641	359205
Otomi (oto)	284988	131139	7991	0	7943	443	156	0	137316
Quechuan (quy)	986947	113640	168189	62777	61131	0	277	58073	522860
Shipibo-Konibo (shp)	32326	4897	0	0	16025	0	122	0	11282
Tarahumara (tar)	63438	0	0	0	7894	0	0	0	55544
Total	3384364	491246	652494	121595	272371	816	868	122388	1862334

Table 1: Monolingual dataset used for continued pre-training, in terms of number of sentences, for the indigenous American languages. [†]OCR data is inherently multilingual, with significant amounts of English and/or Spanish, so the data per language is likely overestimated.

Corpus	English	Spanish
MADLAD 400	150000	150000
Wikipedia	100000	100000
Helsinki'23 (Bibles)	148060	487006
Helsinki'23 (UDHR)	0	120
Total	398060	737126

Table 2: Monolingual dataset used for continued pre-training, in terms of number of sentences, for high-resourced languages (English, Spanish) we use as replay data to prevent catastrophic forgetting.

tuning datasets to the Alpaca format.

3 Approach

To adapt LLMs for the task of translating indigenous American languages, we follow the 2-stage training paradigm proposed in related work (Xu et al., 2024; Alves et al., 2024) and explore its effectiveness for low-resource languages.

3.1 Stage 1: Continued Pre-training with LoRA

In order to "teach" our LLMs the indigenous American languages, we first fine-tuned LLMs with monolingual data for each of these languages. Given these low-resource languages are out-of-distribution from the original pre-training data, we also included replay data from two high-resource languages (English and Spanish) to prevent catastrophic forgetting (Ibrahim et al., 2024). For each American language, given that there were often several (distinctive) dialects, we found that the easiest setting, i.e., to concatenate all of them together, performed very similarly to more careful dialect separation techniques. Inspired by Nguyen et al. (2023), who filtered data from various domains into quality buckets, we segregated our data based on dialects - we assigned the test/dev set dialects to "higher-quality" buckets, and the rest to lower quality. We then tried out a variety of approaches in our preliminary experiments that involved pre-training on various buckets at various stages, but none of these settings performed significantly better[4] than our earlier baseline that concatenated all dialects. Our conclusion here was that these LLMs are only just beginning to learn to model these very low-resourced languages, and cannot separate between dialects at this stage.

For efficiency reasons, we opted for low-rank (LoRA) adaptation (Hu et al., 2021), rather than full-fine tuning. We attached rank 8 LoRA adapters to query and value matrices, following Hu et al. (2021), and also fine-tuned input and output (LM head) embeddings – which we empirically observed to yield significant gains in validation performance. We used average cross-entropy loss σ on the official development set as our validation metric, which we computed as the weighted average of average perplexity on high-resource languages (English and Spanish) and that of the indigenous American languages:

$$\sigma = 0.9 \cdot \sigma_{\text{avg}}^{\{En,Es\}} + 0.1 \cdot \sigma_{\text{avg}}^{\{American\}}$$

where $\sigma_{\text{avg}}^{\{En,Es\}}$ and $\sigma_{\text{avg}}^{\{American\}}$ are the average perplexities on English and Spanish, as well as the indigenous American languages respectively.

We explored adaptation of four LLMs: Llama-2 7B (Touvron et al., 2023), MaLA-500 (Lin et al.,

[4]from a validation loss perspective

Source	Files	Characters
Grammar/Education Book	156 (52.2%)	39,971,932 (46.6%)
Scientific Paper	58 (19.4%)	9,880,833 (11.5%)
Dictionary	55 (18.4%)	28,579,012 (33.3%)
Book	16 (5.4%)	3,360,407 (3.9%)
Other	14 (4.7%)	4,009,128 (4.7%)
Total	299	85,801,312

Table 3: Summary statistics of the OCR data grouped by **source**. We exclude whitespaces while counting **characters**. Percentages of the total are displayed in parentheses.

Task(s)	Dataset	Languages	Instruction Count
Human-annotated Prompt Completions	Aya Dataset	{es, pt, en}	16795
Cross-lingual QA	Synthetic	{es} → All	82538
Machine Translation	Official	{es} → All	76511
	NLLB	{en} → {aym, gn}	13276
	FLORES 200	{es, en, pt} → {aym, gn, quy}	18081
	Helsinki'23	{es} → {gn, hch, nhe, quy, shp}	27976
	OPUS	{es, en, pt} → {aym, cni, gn, nhe, quy}	112681

Table 4: Datasets used for instruction tuning. Languages are denoted by their ISO 639 codes.

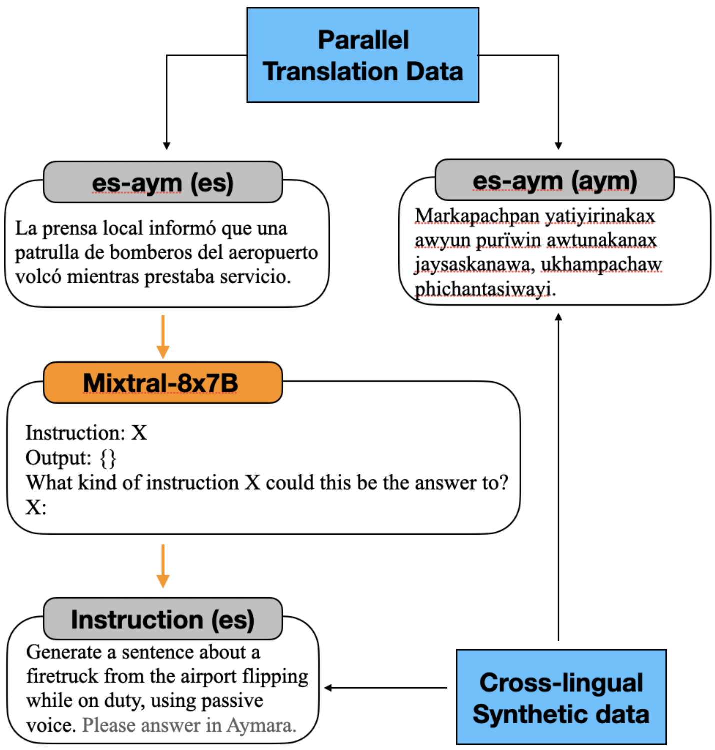

Figure 1: Illustration of our designed process of generating cross-lingual synthetic instruction data.

2024), Mistral 7B (Jiang et al., 2023) and Mistral 7B v0.2[5] for this task. We chose Llama-2 and Mistral since they are the most widely used general-purpose models while MaLA-500, which is the Llama-2 model scaled to 500 languages using LoRA adapters, could potentially enable better cross-lingual transfer.

To examine in greater detail the role of parallel data for continued pre-training under low-resource settings, we trained primarily 2 sets of models, dubbed `v1` and `v2`. `v1` used a concatenation of all available monolingual data[6], while the `v2` models integrated not only monolingual data from `v1`, but also the parallel corpora. Inspired from related work, we explored 3 techniques of leveraging this parallel data: i) `v2.0:` considering the target side of es-X bitext as additional monolingual data, and using the same for pre-training, ii) `v2.1:` following Alves et al. (2024), concatenating[7] the source and target sentences of a certain percentage of sentences (25%, in our experiments[8]), while the rest is used for its target-side data, and iii) `v2.2:` 'interleaving' concatenated Es-X and X-Es parallel text, closely following Guo et al. (2024), and fine-tuning with the same after pre-training on exclusively monolingual data (i.e. `v1` models in our case). For our best-performing model, Mistral 7B, we found `v2.2` baselines overfit and lead to divergence of validation loss, as a result we discard these models.

Given that validation loss cannot be compared fairly across models with different tokenizers, and may not correlate well with downstream MT performance (Iyer et al., 2023), a key challenge we faced was our inability to reliably estimate downstream MT performance after stage 1 pre-training. We, thus, resorted to instruction-tuning all our top-performing models and directly evaluated downstream MT quality– similar to related works (Xu et al., 2024; Alves et al., 2024).

3.2 Stage 2: Instruction Tuning

For instruction-tuning, we continue fine-tuning the stage 1 LoRA adapters on our curated multi-task dataset (Table 4). We fine-tune both input and output embeddings, along with the LoRA adapters, since we observe that this leads to marginal improvements in MT quality. We show these results in Table 5, along with ablations showing how each dataset contributes to improving our overall average performance.

4 Experiments

4.1 Experimental Settings

Stage 1: We used temperature sampling with $\tau = 80$ to ensure uniform data distribution across the relatively higher-resourced (English, Spanish, Quechua, Aymara, Guarani) and the other lower-resourced languages in this setup – since our objective in this work was to build a multilingual LLM that generalizes well to all the languages in this task. However, given the temperature is quite high, and low-resource languages might thus be oversampled excessively, we used a 'clipping factor' of 10 to ensure oversampling does not exceed 10x the original data size.

We conducted continued pre-training of our models using Hugging Face PEFT (Mangrulkar et al., 2022) with the DeepSpeed ZeRO3 configuration (Rajbhandari et al., 2020) on 2 A100-80GB GPUs. We used LoRA adapters on the query and value matrices of rank 4, alpha 8, and dropout 0.1. We used a batch size of 3 per GPU and 16 gradient accumulation steps. We used a learning rate of 2e-5 and a cosine scheduler. We did not use warm-up since we also provided replay data, and empirically found this to be a better choice for validation performance. We saved and evaluated every 100 steps, with a patience value of 5 for early stopping and average evaluation loss as the validation metric. We pre-trained our models for 1 epoch only, due to the enormous training costs.

Stage 2: For instruction tuning, we used the LLaMa-Factory (Zheng et al., 2024) library – which is an easy-to-use package for instruction tuning, built on top of Hugging Face libraries. We continued to tune the LoRA adapters from Stage 1 for 4 epochs using tf32 floating point precision. We used a learning rate of 1e-4, with a cosine scheduler and warm-up ratio of 3%. We used a batch size of 8 per GPU and 16 gradient accumulation steps.

Decoding: We used LLaMa-Factory for decoding on the test set. We used the following default parameters for sampling: a sampling temperature

[5] `https://models.mistralcdn.com/mistral-7b-v0-2/mistral-7B-v0.2.tar`

[6] except the OCR data, which we were only able to obtain for `v2` pre-training

[7] During concatenation, we prepend the language code `L` before each sentence `X`, like so `[L]: X`. Source and target sentences are then joined with the newline character `\n`.

[8] We observed higher percentages (like 75%) decreased validation perplexity more significantly.

No.	Base Model	Tuned Part	Data	Avg. chrF++
1	Llama-2-7B	LoRA	Parallel	7.09
2	Llama-2-7B	LoRA	Parallel+Aya	8.11
3	Mistral-7B	LoRA	Parallel	9.54
4	Mistral-7B	LoRA	Parallel+Aya	9.85
5	Llama-2-7B-Stage1	LoRA	Parallel+Aya	15.17
6	Llama-2-7B-Stage1	LoRA+Emb	Parallel+Aya	15.20
7	Mistral-7B-Stage1	LoRA+Emb	Parallel+Aya	16.24
8	Mistral-7B-Stage1-v1	LoRA+Emb	Parallel+Aya+Syn	16.81
9	MaLA-7B-Stage1	LoRA+Emb	Parallel+Aya+Syn	17.41
10	Mistral-7B-Stage1-v2	LoRA+Emb	Parallel+Aya+Syn	17.32

Table 5: chrF++ scores on the AmericasNLP24 development set using greedy decoding.

of 0.95, top-p sampling with p=0.7 and top-k sampling with k=50. We used beam search with a beam size of 10, repetition and length penalty of 1.0. We used a batch size of 16 and set the maximum number of new tokens for generation to 512.

4.2 Instruction Tuning Experiments

We report our empirical experiment results in Table 5 and introduce our main findings below.

Continued pre-training is crucial. As evident from the instruction-tuning experiments performed on two raw LLMs, i.e. Llama-2-7B & Mistral-7B, and their corresponding stage 1 variants (Llama-2-7B-Stage1 & Mistral-7B-Stage1), we can see that the pre-trained stage 1 models outperform raw instruction-tuned models by a large margin – indicating that LLMs benefit significantly from in-domain monolingual data, even if it is scarce compared to usual high-resourced setups.

However, these gains can potentially suffer from limited returns over time. For the Stage 1 `v2.0` models, which have been trained on 2.5M sentences (78M tokens) more, and obtained a gain in stage 1 validation loss of almost -1.0 point, the corresponding gains in downstream performance (chrF++) was not as significant. Further research is required to verify and analyse the findings from these preliminary experiments.

The general purpose Aya instruction dataset boosts MT performance. This was a surprising finding that showed that even though: a) the language of the generation is not an American indigenous language, and b) the task is not Machine Translation, general-purpose instruction data do not focus on the translation task - we still found significant gains in MT performance. This is likely because this data helps the LLM to reason and follow instructions better.

Adding cross-lingual synthetic instruction data also helps Another interesting exploration in our work is the usage of cross-lingual synthetic instruction data (Section 2.2). While we observe that the quality of the synthetic is not perfect and contains some degree of noise, it does improve the system's translation quality on average. Preliminary experiments also suggested that substituting this with higher quality (but less quantity) data end up performing worse, suggesting that LLMs likely do not know how to generate in these low-resource languages and more data, even if synthetic, can help.

Fine-tuned Mistral usually outperforms Llama-2 Mistral 7B, which has been shown to consistently outperform Llama 13B (Jiang et al., 2023), seems to be more effective in low-resource settings as well. It consistently beats the latter by significant margins. Hence, we choose Mistral as our primary LLM and decide to improve on the same for our final models.

4.3 Checkpoint Averaging

Inspired by (Gao et al., 2022), we use a straightforward low computational approach to boost the performance of our instruction-tuned LLMs. We selected the last 4 model checkpoints from the same run and averaged the model (LoRA) parameters to obtain a better model. Checkpoint averaging is relatively cheaper and does not require storing and querying multiple models at test time. Additionally, we explore all 10 combinations of the last 4 model checkpoints, combining them in triplets and pairs. However, the most significant improvement was observed when averaging the last 4 models checkpoints.

We perform decoding using default parameters of LLaMa-Factory— a sampling temperature of 0.95, top-p and top-k sampling with p=0.7 and

#	Checkpoint	Avg. chrF++ score per model		
		Mistral-7B-v1	MaLA-500	Mistral-7B-v2
(a)	Final checkpoint (step=8151)	19.05	19.18	19.34
(b)	Checkpoint 8000	19.42	19.20	19.16
(c)	Checkpoint 7500	19.18	19.34	18.82
(d)	Checkpoint 7000	19.27	19.08	19.14
(e)	AVG(a,b,c,d)	**20.29**	**19.94**	**20.07**

Table 6: Checkpoint averaging with different models on AmericasNLP development set using default generation parameters of LLaMa-Factory.

k=50 respectively, beam size 1, length and repetition penalty of 1.0 and maximum number of new tokens for generation 512. In Table 6, it's evident that the model with averaged checkpoints consistently outperforms the others. We believe the reason behind its superior performance is that checkpoint averaging acts as a form of regularization.

During the training process, it is possible for a few layers of the model to start over-fitting after certain steps, leading to a degradation in performance if training continues. However, by averaging later checkpoints with the initial ones from earlier in the training process, the effects of over-fitting can be mitigated. This combination helps to regularize the model, preventing it from over-fitting to the training data while still leveraging the useful information learned during the later stages of training.

For future work, we will explore two approaches: a) combining last k checkpoints instead of last 4 during model averaging. b) Weighted averaging of checkpoints, where checkpoints with better performance on the development set receive higher weights. Our hypothesis is that these methods could improve model performance over the current unweighted averaging of the last 4 checkpoints.

4.4 Final Test Set Results

The final systems we submit to the shared task are, therefore (all model IDs are from Table 6 and are open-sourced at `https://tinyurl.com/edi-amnlp24`):

- System 1: Checkpoint `e` i.e. average of checkpoints `a`, `b`, `c` and `d`, for `Mistral-7B-v1`
- System 2: Checkpoint `e` i.e. average of checkpoints `a`, `b`, `c` and `d`, for `MALA-7B-stage2`
- System 3: Average of checkpoints `a`, `c` and `d` for `Mistral-7b-stage2-v2`

For final inference, we use a beam size of 10 expecting a performance boost. Other decoding parameters remained the same. We show our final results on the AmericasNLP 2024 test sets in Table 7. We observe that while our models do not outperform the best systems, the gap is relatively lower for lower resourced languages like Huichol, Nahuatl and Otomi. While this does align with our stated goal of building a general purpose LLM for the languages in this task, as part of future research, we shall explore how we can model better across the other pairs too and increase our competitiveness.

Ethical Considerations

None of the authors of this paper speak any indigenous American languages in this shared task. We rely on the language-labelled datasets suggested by the task organizers and from other reputable sources. We actively sought data manual inspection using Google Translate.

Acknowledgments

This work has received funding from UK Research and Innovation under the UK government's Horizon Europe funding guarantee [grant numbers 10039436 and 10052546].

The computations described in this research were performed using the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk/). Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham.

Language	Metrics	Best system 1	Best system 2	UEdin Submission 1	UEdin Submission 2	UEdin Submission 3
aym	BLEU	3.49	3.23	1.14	1.06	1.13
	chrF++	30.97	29.39	21.77	21.37	21.89
bzd	BLEU	4.84	4.56	2.21	1.89	1.75
	chrF++	23.47	23.41	16.54	16.32	15.56
cni	BLEU	2.41	3.49	0.41	0.37	0.43
	chrF++	23.20	22.98	14.82	13.68	14.50
ctp	BLEU	13.44	4.65	3.35	4.30	3.38
	chrF++	37.38	23.64	17.66	20.70	17.57
gn	BLEU	12.04	11.28	3.38	1.78	3.21
	chrF++	38.93	37.64	29.20	24.61	29.13
hch	BLEU	10.08	9.62	9.87	7.03	9.60
	chrF++	27.64	26.46	24.41	22.03	24.37
nah	BLEU	2.30	1.09	0.48	0.37	0.44
	chrF++	22.87	21.71	18.12	17.21	18.98
oto	BLEU	1.42	1.55	0.43	0.21	0.44
	chrF++	12.98	12.63	8.91	7.81	9.19
quy	BLEU	4.85	4.83	1.32	0.94	1.31
	chrF++	38.21	38.19	25.23	22.77	25.04
shp	BLEU	4.45	4.14	1.34	1.56	1.55
	chrF++	29.37	27.04	22.04	22.43	22.86
tar	BLEU	0.92	1.01	0.11	0.11	0.15
	chrF++	17.03	15.42	9.65	9.49	9.48

Table 7: AmericasNLP 2024 test set results. We show the performances of the top 2 best systems from each language, as well as each of the 3 systems we submit. Languages are denoted by their ISO 639 codes.

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*.

Giusepppe Attardi. 2015. Wikiextractor. `https://github.com/attardi/wikiextractor`.

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.

Yingbo Gao, Christian Herold, Zijian Yang, and Hermann Ney. 2022. Revisiting checkpoint averaging for neural machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 188–196, Online only. Association for Computational Linguistics.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. *arXiv preprint arXiv:2403.11430*.

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. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*.

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. *arXiv preprint arXiv:2309.11668*.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. 2024. Longform: Effective instruction tuning with reverse instructions. *Preprint*, arXiv:2304.08460.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

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*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20:*

International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–16. IEEE.

Nathaniel R Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high-(but not low-) resource languages. *arXiv preprint arXiv:2309.07423*.

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*.

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.

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

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*.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *arXiv preprint arXiv:2402.18025*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Appendix

Combinations of languages	**Source type**	**Files**	**Characters**
Aymara	Mono	8	682,766
English/Asháninka	Mixed	2	1,605,073
English/Aymara	Mixed	9	2,945,037
English/Chatino	Mixed	8	2,708,631
English/Guaraní	Mixed	12	2,773,253
English/Hñähñu	Mixed	5	2,181,855
English/Nahuatl	Mixed	24	8,950,757
English/Quechua	Mixed	7	1,429,763
English/Spanish/Aymara	Mixed	1	246,850
English/Spanish/Quechua	Mixed	3	953,120
English/Spanish/Rarámuri	Mixed	2	1,250,289
English/Wixarika	Mixed	1	544,090
French/Aymara	Mixed	1	52,022
French/Bribri	Mixed	1	1,099,198
French/Hñähñu	Mixed	1	111,296
French/Quechua	Mixed	1	194,163
French/Rarámuri	Mixed	1	23,418
German/Guaraní	Mixed	1	178,220
German/Quechua	Mixed	2	1,361,053
Nahuatl	Mono	2	224,394
Quechua	Mono	10	492,504
Russian/Guaraní	Mixed	1	51,939
Russian/Nahuatl	Mixed	1	75,205
Russian/Quechua	Mixed	2	193,794
Spanish/Asháninka	Mixed	9	2,133,942
Spanish/Asháninka/Quechua	Mixed	1	65,046
Spanish/Aymara	Mixed	45	9,546,160
Spanish/Aymara/Nahuatl/Quechua	Mixed	1	208,828
Spanish/Bribri	Mixed	4	801,911
Spanish/Chatino	Mixed	3	1,162,349
Spanish/Guaraní	Mixed	20	8,101,890
Spanish/Hñähñu	Mixed	10	3,059,227
Spanish/Nahuatl	Mixed	17	6,171,836
Spanish/Quechua	Mixed	67	19,830,467
Spanish/Rarámuri	Mixed	6	1,311,759
Spanish/Shipibo-Konibo	Mixed	4	461,930
Spanish/Wixarika	Mixed	5	1,478,789
Wixarika	Mono	1	1,138,488

Table 8: Summary statistics of the OCR data, grouped by **Combinations of languages**. **Characters** counted without whitespaces.

Source	**Low-resource languages**	**Source type**	**Files**	**Characters**
Book	Nahuatl	Mono	1	195,009
	Quechua	Mixed	8	1,727,827
		Mono	6	299,083
	Wixarika	Mono	1	1,138,488
Dictionary	Asháninka	Mixed	3	783,665
	Aymara	Mixed	15	4,792,382
	Chatino	Mixed	2	1,012,744
	Guaraní	Mixed	8	5,509,379
	Nahuatl	Mixed	5	3,424,235
	Quechua	Mixed	19	12,354,240
	Rarámuri	Mixed	3	702,367
Grammar/Education Book	Asháninka	Mixed	5	2,279,964
	Aymara	Mixed	25	6,212,691
		Mono	8	682,766
	Bribri	Mixed	3	714,131
	Chatino	Mixed	1	149,605
	Guaraní	Mixed	16	4,585,622
	Hñähñu	Mixed	13	4,441,870
	Nahuatl	Mixed	24	9,877,127
		Mono	1	29,385
	Quechua	Mixed	47	9,072,258
		Mono	5	247,344
	Rarámuri	Mixed	3	1,146,458
	Shipibo-Konibo	Mixed	3	314,443
	Wixarika	Mixed	2	218,268
Other	Aymara	Mixed	4	1,136,545
	Hñähñu	Mixed	1	95,944
	Nahuatl	Mixed	5	1,461,840
	Rarámuri	Mixed	1	54,278
	Wixarika	Mixed	3	1,260,521
Scientific Paper	Asháninka	Mixed	3	675,386
	Asháninka/Quechua	Mixed	1	65,046
	Aymara	Mixed	12	648,451
	Aymara/Nahuatl/Quechua	Mixed	1	208,828
	Bribri	Mixed	2	1,186,978
	Chatino	Mixed	8	2,708,631
	Guaraní	Mixed	10	1,010,301
	Hñähñu	Mixed	2	814,564
	Nahuatl	Mixed	8	434,596
	Quechua	Mixed	7	754,112
	Rarámuri	Mixed	2	682,363
	Shipibo-Konibo	Mixed	1	147,487
	Wixarika	Mixed	1	544,090

Table 9: Summary statistics of the OCR data, grouped by **Source** and **Low-resource languages**. **Characters** counted without whitespaces.