Tuning LLMs with Contrastive Alignment Instructions for Machine Translation in Unseen, Low-resource Languages

Anonymous ACL submission

Abstract

This article introduces contrastive alignment instructions (**AlignInstruct**) to address two challenges in machine translation (MT) on large language models (LLMs). One is the expansion of supported languages to previously unseen ones. The second relates to the lack of data in low-resource languages. Model fine-tuning through MT instructions (**MTInstruct**) is a straightforward approach to the first challenge. However, MTInstruct is limited by weak cross-lingual signals inherent in the second challenge. AlignInstruct emphasizes cross-lingual supervision via a cross-lingual discriminator built using statistical word alignments. Our results based on fine-tuning the BLOOMZ models (1b1, 3b, and 7b1) in up to 24 unseen languages showed that: (1) LLMs can effectively translate unseen languages using MTInstruct; (2) AlignInstruct led to consistent improvements in translation quality across 48 translation directions involving English; (3) Discriminator-based instructions outperformed their generative counterparts as cross-lingual instructions; (4) AlignInstruct improved performance in 30 zero-shot directions.

1 Introduction

Large language models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2022; Touvron et al., 2023a; Muennighoff et al., 2023; OpenAI, 2023; Anil et al., 2023; Touvron et al., 2023b) achieved good performance for a wide range of NLP tasks for prevalent languages. However, insufficient coverage for low-resource languages remains to be one significant limitation. Low-resource languages are either not present, or orders of magnitude smaller in size than dominant languages in the pre-training dataset. This limitation is in part due to the prohibitive cost incurred by curating good quality and adequately sized datasets for pre-training. Incrementally adapting existing multilingual LLMs to incorporate an

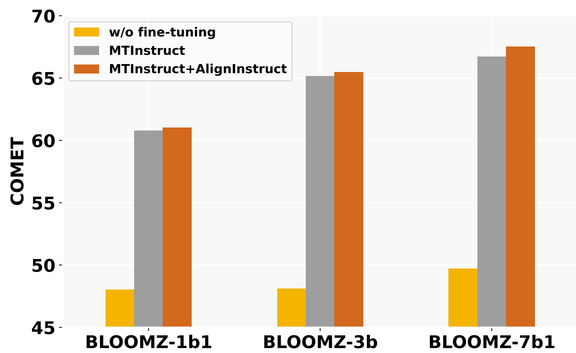

Figure 1: **Average COMET scores of BLOOMZ models across 24 unseen languages**, comparing settings of without fine-tuning, fine-tuning with MTInstruct, and fine-tuning that combines MTInstruct and AlignInstruct.

unseen, low-resource language thus becomes a cost-effective priority to address this limitation. Previous study (de la Rosa and Fernández, 2022; Müller and Laurent, 2022; Yong et al., 2023) explored extending language support using either continual pre-training (Neubig and Hu, 2018; Artetxe et al., 2020; Muller et al., 2021; Ebrahimi and Kann, 2021), or parameter efficient fine-tuning (PEFT) methods (Pfeiffer et al., 2020; Hu et al., 2022; Liu et al., 2022) on monolingual tasks. Extending language support for cross-lingual tasks remains underexplored due to the challenge of incrementally inducing cross-lingual understanding and generation abilities in LLMs (Yong et al., 2023).

This study focused on machine translation (MT) to highlight the cross-lingual LLM adaptation challenge. The challenge lies in enabling translation for low-resource languages that often lack robust cross-lingual signals. We first explored the efficacy of fine-tuning LLMs with MT instructions (MTInstruct) in unseen, low-resource languages. MTInstruct is a method previously shown to bolster the translation proficiency of LLMs for supported languages (Li et al., 2023). Subsequently, given that cross-lingual alignments are suboptimal

mal in LLMs as a result of data scarcity of low-resource languages, we proposed contrastive alignment instructions (AlignInstruct) to explicitly provide cross-lingual supervision during MT fine-tuning. AlignInstruct is a cross-lingual discriminator formulated using statistical word alignments. Our approach was inspired by prior studies (Lambert et al., 2012; Ren et al., 2019; Lin et al., 2020; Mao et al., 2022), which indicated the utility of word alignments in enhancing MT. In addition to AlignInstruct, we discussed two word-level cross-lingual instruction alternatives cast as generative tasks, for comparison with AlignInstruct.

Our experiments fine-tuned the BLOOMZ models (Muennighoff et al., 2023) of varying sizes (1b1, 3b, and 7b1) for 24 unseen, low-resource languages, and evaluated translation on OPUS-100 (Zhang et al., 2020) and Flores-200 (Costa-jussà et al., 2022). We first showed that MTInstruct effectively induced the translation capabilities of LLMs for these languages. Building on the MTInstruct baseline, the multi-task learning combining AlignInstruct and MTInstruct resulted in stronger translation performance without the need for additional training corpora. The performance improved with larger BLOOMZ models, as illustrated in Fig. 1, indicating that AlignInstruct is particularly beneficial for larger LLMs during MT fine-tuning. When compared with the generative variants of AlignInstruct, our results indicated that discriminator-style instructions better complemented MTInstruct. Furthermore, merging AlignInstruct with its generative counterparts did not further improve translation quality, underscoring the efficacy and sufficiency of AlignInstruct in leveraging word alignments for MT.

In zero-shot translation evaluations on the OPUS benchmark, AlignInstruct exhibited improvements over the MTInstruct baseline in 30 zero-shot directions not involving English, when exclusively fine-tuned with three unseen languages (German, Dutch, and Russian). However, when the fine-tuning data incorporated supported languages (Arabic, French, and Chinese), the benefits of AlignInstruct were only evident in zero-shot translations where the target language was a supported language.

To interpret the inherent modifications within the BLOOMZ models after applying MTInstruct or AlignInstruct, we conducted a visualization of the layer-wise cross-lingual alignment capabilities of the model representations. In addition, we discussed the effect of monolingual instructions in the resource-constrained scenario.

2 Methodology

This section presents MTInstruct as the baseline, and AlignInstruct. The MTInstruct baseline involved fine-tuning LLMs using MT instructions. AlignInstruct dealt with the lack of cross-lingual signals stemming from the limited parallel training data in low-resource languages. The expectation was enhanced cross-lingual supervision cast as a discriminative task without extra training corpora. Following this, we introduced two generative variants of AlignInstruct for comparison and discussed monolingual instructions for MT fine-tuning.

2.1 Baseline: MTInstruct

Instruction tuning (Wang et al., 2022; Mishra et al., 2022; Chung et al., 2022; Ouyang et al., 2022; Sanh et al., 2022; Wei et al., 2022) has been shown to generalize LLMs' ability to perform various downstream tasks, including MT (Li et al., 2023).

Given a pair of the parallel sentences, $\left((x_i)_1^N, (y_j)_1^M\right)$, where $(x_i)_1^N := x_1 x_2 \ldots x_N$, $(y_i)_1^N := y_1 y_2 \ldots y_N$. $x_i, y_j \in \mathcal{V}$ are members of the vocabulary $\mathcal{V}$ containing unique tokens that accommodate languages X and Y. Li et al. (2023) showed that the following MT instructions (MTInstruct) can improve the translation ability in an LLM with a limited number of parallel sentences:

- **Input:** "Translate from Y to X.
  Y: $y_1 y_2 \ldots y_M$.
  X: "
- **Output**: "$x_1 x_2 \ldots x_N$."

Note that Li et al. (2023) demonstrated the utility of MTInstruct solely within the context of fine-tuning for languages acquired at pre-training phase. This study called for an assessment of MTInstruct on its efficacy for adapting to previously unsupported languages, denoted as X, accompanied by the parallel data in a supported language Y.

2.2 AlignInstruct

Word alignments have been demonstrated to enhance MT performance (Lambert et al., 2012; Ren et al., 2019; Lin et al., 2020; Mao et al., 2022), both in the fields of statistical machine translation (SMT) (Brown et al., 1993) and neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015). Ren et al. (2019) and Mao et al. (2022) reported the utility of SMT-derived

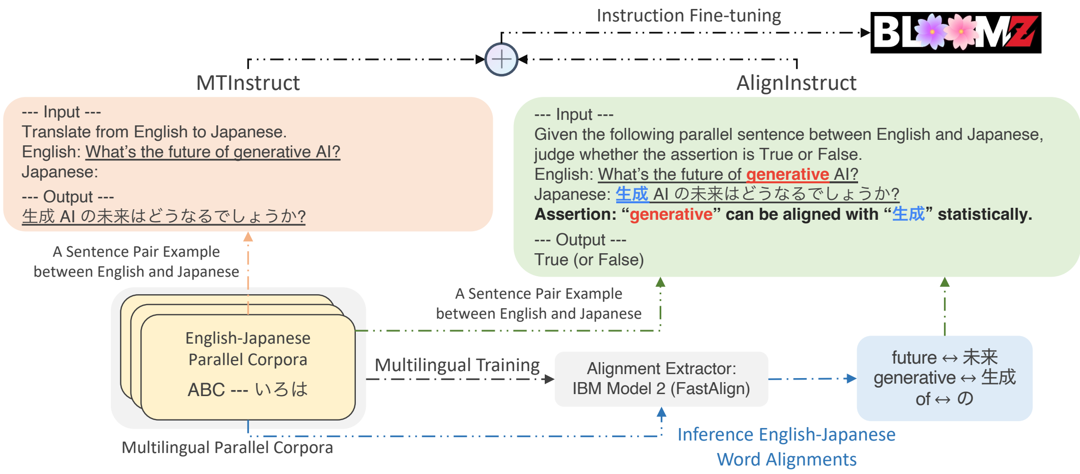

Figure 2: **Proposed instruction tuning methods combining MTInstruct (Sec. 2.1) and AlignInstruct (Sec. 2.2) for LLMs in MT tasks.** $\oplus$ denotes combining multiple instruction patters with a specific fine-tuning curriculum (Sec. 3.2). IBM Model 2 indicates word alignment model of statistical machine translation (Brown et al., 1993).

contrastive word alignments in guiding encoder-decoder NMT model training. Built upon their findings, we introduced AlignInstruct for bolstering cross-lingual alignments in LLMs. We expected AlignInstruct to enhancing translation performance particularly for languages with no pre-training data and limited fine-tuning data.

As shown in Fig. 2, we employed FastAlign (Dyer et al., 2013) to extract statistical word alignments from parallel corpora. Our approach depended on a trained FastAlign model (IBM Model 2, Brown et al., 1993) to ensure the quality of the extracted word pairs. These high-quality word alignment pairs were regarded as "gold" word pairs for constructing AlignInstruct instructions.[1] Assuming one gold word pair $(x_k x_{k+1}, y_l y_{l+1} y_{l+2})$ was provided for the sentence pair $\left((x_i)_1^N, (y_j)_1^M\right)$, the AlignInstruct instruction reads:

- **Input:** "Given the following parallel sentence between Y and X, judge whether the assertion is True or False.
  Y: $y_1 y_2 \dots y_M$.
  X: $x_1 x_2 \dots x_N$.
  Assertion: "$y_l y_{l+1} y_{l+2}$" can be aligned with "$x_k x_{k+1}$" statistically."
- **Output**: "True" (or "False")

Instructions with the "False" output were constructed by uniformly swapping out part of the word pair to crate misalignment. We anticipated that this treatment forced the model to learn to infer the output by recognizing true alignment-enriched instructions. This would require the model to encode word-level cross-lingual representation, a crucial characteristic for MT tasks.

[1]Note that these word pairs may not necessarily represent direct translations of each other; instead, they are word pairs identified based on their co-occurrence probability within the similar context. Refer to IBM model 2 in SMT.

2.3 Generative Counterparts of AlignInstruct

Previous studies (Liang et al., 2022; Yu et al., 2023) have suggested the importance of both discriminative and generative tasks in fine-tuning LLMs. We accordingly considered two generative variants of AlignInstruct. We then compared them with AlignInstruct to determine the most effective training task. As detailed in Sec. 4, our results indicated that these variants underperformed AlignInstruct when applied to unseen, low-resource languages.

2.3.1 HintInstruct

HintInstruct as a generative variant of AlignInstruct was instructions containing word alignment hints. It was inspired by Ghazvininejad et al. (2023), where dictionary hints were shown to improve few-shot in-context leaning. Instead of relying on additional dictionaries, we used the same word alignments described in Sec. 2.2, which were motivated by the common unavailability of high-quality dictionaries for unseen, low-resource languages. Let $\{(x_{k_s} x_{k_s+1} \dots x_{k_s+n_s}, y_{l_s} y_{l_s+1} \dots y_{l_s+m_s})\}_{s=1}^S$ be S word pairs extracted from the sentence pair $\left((x_i)_1^N, (y_j)_1^M\right)$. HintInstruct follows the instruction pattern:

- **Input**: "Use the following alignment hints and translate from Y to X.

Alignments between X and Y:
– $(x_{k_1}x_{k_1+1}\dots x_{k_1+n_1}, y_{l_1}y_{l_1+1}\dots y_{l_1+m_1})$,
– $(x_{k_2}x_{k_2+1}\dots x_{k_1+n_1}, y_{l_2}y_{l_2+1}\dots y_{l_2+m_2})$,
$\dots$,
– $(x_{k_S}x_{k_S+1}\dots x_{k_S+n_S}, y_{l_S}y_{l_S+1}\dots y_{l_S+m_S})$,
Y: $y_1y_2\dots y_M$.
X: "

- **Output**: "$x_1x_2\dots x_N$."

where S denotes the number of the word alignment pairs used to compose the instructions. Different from AlignInstruct, HintInstruct expects the translation targets to be generated.

2.3.2 ReviseInstruct

ReviseInstruct was inspired by Ren et al. (2019) and Liu et al. (2020) for the notion of generating parallel words or phrases, thereby encouraging a model to encode cross-lingual alignments. A ReviseInstruct instruction contained a partially corrupted translation target, as well as a directive to identify and revise these erroneous tokens. Tokens are intentionally corrupted at the granularity of individual words, aligning with the word-level granularity in AlignInstruct and HintInstruct. ReviseInstruct follows the instruction pattern:

- **Input**: "Given the following translation of X from Y, output the incorrectly translated word and correct it.
Y: $y_1y_2\dots y_M$.
X: $x_1x_2\dots x_kx_{k+1}\dots x_{k+n}\dots x_N$."
- **Output**: "The incorrectly translated word is "$x_kx_{k+1}\dots x_{k+n}$". It should be "$x_jx_{j+1}\dots x_{j+m}$"."

2.4 Monolingual Instructions

New language capabilities may be induced through continual pre-training on monolingual next-word prediction tasks (Yong et al., 2023). The coherence of the generated sentences is crucial in MT (Wang et al., 2020; Liu et al., 2020), especially when the target languages are unseen and low-resource. We examined the significance of this approach in fostering the translation quality. We reused the same parallel corpora to avoid introducing additional monolingual datasets.

Given a monolingual sentence, $(x_i)_1^N$, with length N in an unseen language X. The LLM is incrementally trained on the following task:

- **Input**: "Given the context, complete the following sentence: $x_1x_2\dots x_{l<N}$,"

- **Output**: "$x_{l+1}x_{l+2}\dots x_N$."

3 Experimental Settings

3.1 Backbone Models and Unseen Languages

Our experiments fine-tuned the BLOOMZ models (Muennighoff et al., 2023) for MT in unseen, low-resource languages. BLOOMZ is an instruction fine-tuned multilingual LLM from BLOOM (Scao et al., 2022) that supports translation across 46 languages. Two lines of experiments evaluated the effectiveness of the MTInstruct baseline and AlignInstruct:

BLOOMZ+24 Tuning BLOOMZ-7b1, BLOOMZ-3b, and BLOOMZ-1b1[2] for 24 unseen, low-resource languages. These experiments aimed to: (1) assess the effectiveness of AlignInstruct in multilingual, low-resource scenarios; (2) offer comparison across various model sizes. We used the OPUS-100 (Zhang et al., 2020)[3] datasets as training data. OPUS-100 is an English-centric parallel corpora, with around 4.5M parallel sentences in total for 24 selected languages, averaging 187k sentence pairs for each language and English. Refer to App. A for training data statistics. We used OPUS-100 and Flores-200 (Costa-jussà et al., 2022)[4] for evaluating translation between English and 24 unseen languages (48 directions in total) on in-domain and out-of-domain test sets, respectively. The identical prompt as introduced in Sec. 2.1 was employed for inference. Inferences using alternative MT prompts are discussed in App.E.

BLOOMZ+3 Tuning BLOOMZ-7b1 with three unseen languages, German, Dutch, and Russian, or a combination of these three unseen languages and another three seen (Arabic, French, and Chinese). We denote the respective setting as **de-nl-ru** and **ar-de-fr-nl-ru-zh**. These experiments assessed the efficacy of AlignInstruct in zero-shot translation scenarios, where translation directions were not presented during fine-tuning, as well as the translation performance when incorporating supported languages as either source or target languages. To simulate the low-resource fine-tuning scenario, we randomly sampled 200k parallel sentences for each language. For evaluation, we used the OPUS-100 supervised and zero-shot test sets, comprising 12 supervised directions involving English and 30 zero-shot directions without English among six languages.

[2] https://huggingface.co/bigscience/bloomz
[3] https://opus.nlpl.eu/opus-100.php
[4] https://github.com/facebookresearch/flores/blob/main/flores200/README.md

BLOOMZ model	Objective	OPUS en→xx BLEU	chrF++	COMET	OPUS xx→en BLEU	chrF++	COMET	Flores en→xx BLEU	chrF++	COMET	Flores xx→en BLEU	chrF++	COMET
BLOOMZ-7b1	w/o fine-tuning	3.61	8.82	47.94	6.70	18.49	51.49	2.00	9.35	37.04	9.95	24.47	52.18
	Individual objectives												
	MTInstruct	11.54	25.33	64.68	18.59	33.25	68.75	3.30	17.10	42.62	11.37	27.14	55.82
	AlignInstruct	4.73	9.23	49.11	5.32	12.90	53.05	1.97	8.90	40.64	3.47	11.93	39.20
	Multiple objectives with different curricula												
	MT+Align	**12.28**	**26.17**	**65.28**	**18.72**	**34.02**	**69.75**	3.26	**17.20**	**43.05**	**11.60**	**27.38**	**56.28**
	Align→MT	**11.73**	**25.48**	64.64	17.54	32.62	68.70	**3.35**	**17.21**	**42.76**	11.32	**27.21**	55.81
	MT+Align→MT	**12.10**	**26.16**	**65.14**	18.23	**33.54**	**69.56**	3.28	**17.26**	**43.13**	**11.48**	**27.34**	**56.12**
BLOOMZ-3b	w/o fine-tuning	4.63	9.93	48.38	5.90	16.38	47.88	2.00	9.09	38.88	5.86	18.56	46.47
	Individual objectives												
	MTInstruct	10.40	23.08	62.66	16.10	31.15	67.67	2.85	16.23	41.30	8.92	24.57	52.77
	AlignInstruct	1.70	4.05	44.10	0.87	3.20	42.32	0.16	3.09	31.10	0.10	1.80	29.27
	Multiple objectives with different curricula												
	MT+Align	**10.61**	**23.64**	**63.03**	**16.73**	**31.51**	**67.94**	**2.95**	**16.62**	**41.86**	**9.50**	**25.16**	**53.63**
	Align→MT	10.22	22.53	62.22	15.90	30.31	66.79	**3.02**	**16.43**	**41.67**	**9.07**	**24.70**	**53.11**
	MT+Align→MT	**10.60**	**23.35**	**62.69**	**16.58**	**31.64**	**68.29**	**2.93**	**16.57**	**41.74**	**9.41**	**25.08**	**53.44**
BLOOMZ-1b1	w/o fine-tuning	3.76	7.57	46.81	4.78	14.11	49.27	1.24	6.93	37.38	3.49	14.56	43.05
	Individual objectives												
	MTInstruct	7.42	17.85	58.05	11.99	25.59	63.50	2.11	14.40	38.90	5.33	20.65	48.42
	AlignInstruct	2.51	5.29	45.56	3.13	8.92	48.73	0.35	3.79	31.21	1.35	6.43	33.24
	Multiple objectives with different curricula												
	MT+Align	**7.80**	**18.48**	**58.58**	**12.57**	**25.92**	63.49	**2.16**	**14.54**	**39.36**	**5.46**	**20.90**	**48.81**
	Align→MT	**7.49**	**18.09**	**58.38**	11.80	24.70	62.58	2.08	14.28	**39.04**	5.24	20.53	48.37
	MT+Align→MT	**7.98**	**18.61**	**58.74**	**12.43**	**25.78**	63.30	**2.16**	**14.46**	**39.25**	**5.37**	**20.67**	**48.57**

Table 1: **Results of BLOOMZ+24 fine-tuned with MTInstruct and AlignInstruct on different curricula** as described in 3.2. Scores that surpass the MTInstruct baseline are marked in **bold**.

3.2 Training Details and Curricula

The PEFT method, LoRA (Hu et al., 2022), was chosen to satisfy the parameter efficiency requirement for low-resource languages, as full-parameter fine-tuning would likely under-specify the models.See App. B for implementation details. How AlignInstruct and MTInstruct are integrated into training remained undetermined. To that end, we investigated three training curricula:

Multi-task Fine-tuning combined multiple tasks in a single training session (Caruana, 1997). This was realized by joining MTInstruct and AlignInstruct training data, denoted as **MT+Align**.[5]

Pre-fine-tuning & Fine-tuning arranges AlignInstruct and MTInstruct into two stages; namely, curriculum learning (Bengio et al., 2009).[6] This configuration, denoted as **Align→MT**, validates whether AlignInstruct should precede MTInstruct.

Mixed Fine-tuning (Chu et al., 2017) arranged the two aforementioned curricula to start with MT+Align, followed by MTInstruct, denoted as **MT+Align→MT**.

4 Evaluation and Analysis

This section reports BLEU (Papineni et al., 2002; Post, 2018), chrF++ (Popović, 2015), and COMET (Rei et al., 2020) scores for respective experimental configurations. We further characterized of the degree to which intermediate embeddings were language-agnostic after fine-tuning.

4.1 BLOOMZ+24 Results

Tab. 1 shows the scores for the unmodified BLOOMZ models, as well as BLOOMZ+24 under MTInstruct, AlignInstruct, and the three distinct curricula. Non-trivial improvements in all metrics were evident for BLOOMZ+24 under MTInstruct. This suggests that MTInstruct can induce translation capabilities in unseen languages. Applying AlignInstruct and MTInstruct via the curricula further showed better scores than the baselines, suggesting the role of AlignInstruct as complementing MTInstruct. Align→MT was an exception, performing similarly to MTInstruct. This may indicate AlignInstruct's complementarity depends on its cadence relative to MTInstruct in a curriculum.

Superior OPUS and Flores scores under the xx→en direction were evident, compared to the reverse direction, en→xx. This suggests that our treatments induced understanding capabilities more than generative ones. This may be attributed to the fact that BLOOMZ had significant exposure to English, and that we used English-centric corpora. Finally, we noted the inferior performance of Flores than OPUS. This speaks to the challenge of instilling translation abilities in unseen languages

[5] Note that AlignInstruct and MTInstruct were derived from the same parallel corpora.

[6] An effective curriculum often starts with a simple and general task, followed by a task-specific task.

BLOOMZ model	Objective	OPUS en→xx			OPUS xx→en			Flores en→xx			Flores xx→en		
		BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
BLOOMZ-7b1	MTInstruct	11.54	25.33	64.68	18.59	33.25	68.75	3.30	17.10	42.62	11.37	27.14	55.82
	MT+Align	**12.28**	**26.17**	**65.28**	**18.72**	**34.02**	**69.75**	3.26	**17.20**	**43.05**	**11.60**	**27.38**	**56.28**
	MT+Hint	**12.12**	**25.92**	**64.82**	18.25	33.18	**69.21**	**3.34**	**17.13**	**42.95**	**11.45**	**27.37**	**56.21**
	MT+Revise	**11.96**	**25.73**	**64.99**	**18.69**	**33.74**	**69.30**	**3.34**	17.10	**43.01**	**11.44**	**27.37**	**56.08**
BLOOMZ-3b	MTInstruct	10.40	23.08	62.66	16.10	31.15	67.67	2.85	16.23	41.30	8.92	24.57	52.77
	MT+Align	**10.61**	**23.64**	**63.03**	**16.73**	**31.51**	**67.94**	**2.95**	**16.62**	**41.86**	**9.50**	**25.16**	**53.63**
	MT+Hint	**10.49**	**23.34**	62.66	**16.29**	**31.43**	**68.16**	**3.11**	**16.95**	**42.17**	**9.52**	**25.25**	**53.72**
	MT+Revise	**10.52**	23.03	62.38	**16.22**	30.98	67.27	**2.99**	**16.83**	**41.84**	**9.47**	**25.21**	**53.29**
BLOOMZ-1b1	MTInstruct	7.42	17.85	58.05	11.99	25.59	63.50	2.11	14.40	38.90	5.33	20.65	48.42
	MT+Align	**7.80**	**18.48**	**58.58**	**12.57**	**25.92**	63.49	**2.16**	**14.54**	**39.36**	**5.46**	**20.90**	**48.81**
	MT+Hint	**7.71**	**18.15**	**58.26**	11.52	24.88	62.98	**2.21**	**14.61**	**39.59**	**5.47**	**20.78**	**48.56**
	MT+Revise	7.31	**17.99**	**58.18**	**12.00**	25.33	63.11	2.07	14.32	**38.97**	**5.41**	**20.91**	**48.67**

Table 2: **Results of BLOOMZ+24 fine-tuned combining MTInstruct with AlignInstruct (or its generative variants).** Scores that surpass the MTInstruct baseline are marked in **bold**.

Objective	OPUS en→xx			OPUS xx→en			Flores en→xx			Flores xx→en		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
MTInstruct	11.54	25.33	64.68	18.59	33.25	68.75	3.30	17.10	42.62	11.37	27.14	55.82
MT+Align	**12.28**	**26.17**	**65.28**	**18.72**	**34.02**	**69.75**	3.26	**17.20**	**43.05**	**11.60**	**27.38**	**56.28**
MT+Align+Revise	**12.08**	**25.73**	64.67	**19.23**	**34.32**	**69.65**	**3.33**	**17.25**	**43.05**	**11.60**	**27.61**	**56.51**
MT+Align+Hint	**12.02**	**25.51**	64.68	**19.40**	**34.44**	**69.54**	3.25	16.87	42.85	**11.58**	**27.48**	**56.31**
MT+Hint+Revise	**12.10**	**25.69**	**64.71**	**19.58**	**34.49**	**69.46**	**3.34**	**17.24**	**43.07**	**11.70**	**27.62**	**56.48**
MT+Align+Hint+Revise	**12.00**	**25.39**	64.35	**19.68**	**34.48**	**69.58**	**3.40**	**17.17**	**43.09**	**11.67**	**27.54**	**56.44**

Table 3: **Results of BLOOMZ+24 combining MTInstruct with multiple objectives among AlignInstruct, HintInstruct, and ReviseInstruct on BLOOMZ-7b1.** Scores that surpass MTInstruct are marked in **bold**.

when dealing with the out-of-domain MT task.

4.2 Assessing AlignInstruct Variants

From the results reported in Tab. 2, we observed the objectives with AlignInstruct consistently outperformed those with HintInstruct or ReviseInstruct across metrics and model sizes. Namely, easy, discriminative instructions, rather than hard, generative ones, may be preferred for experiments under similar data constraints. The low-resource constraint likely made MTInstruct more sensitive to the difficulty of its accompanying tasks.

Further, combining more than two instruction tuning tasks simultaneously did not guarantee consistent improvements, see Tab. 3. Notably, MT+Align either outperformed or matched the performance of other objective configurations. While merging multiple instruction tuning tasks occasionally resulted in superior BLEU and chrF++ scores for OPUS xx→en, it fell short in COMET scores compared to MT+Align. This indicated that while such configurations might enhance word-level translation quality, as reflected by BLEU and chrF++ scores, due to increased exposure to cross-lingual word alignments, MT+Align better captured the context of the source sentence as reflected by COMET scores. Overall, these instruction tuning tasks did not demonstrate significant synergistic effects for fine-tuning for unseen languages.

4.3 Assessing Monolingual Instructions

We conducted experiments with two MonoInstruct settings: **MonoInstruct-full**, an objective to generate the entire sentence, and **MonoInstruct-half** for generating the latter half of the sentence given the first half, inspired by GPT (Radford et al., 2018) and MASS (Song et al., 2019), respectively. We reported the MonoInstruct results in Tab. 4. Firstly, we observed that fine-tuning MTInstruct in conjunction with either MonoInstruct-full or MonoInstruct-half harms the MT performance, which could be attributed to the inherent difficulty of monolingual instruction tasks and the limited amount of monolingual data. We found that the simpler MT+Mono-half yielded better results than MT+Mono-full as richer contexts were provided. However, MonoInstruct still did not improve the MTInstruct baseline. Secondly, further combining MonoInstrcut with AlignInstruct variants yielded improvements compared with MT+Mono-full (or half), but underperformed the MTInstruct baseline. This suggested that improving MT performance with monolingual instructions is challenging without access to additional monolingual data.

4.4 BLOOMZ+3 Zero-shot Evaluation

Tab. 5 reports the results of the two settings, de-nl-ru and ar-de-fr-nl-ru-zh. Results of

Objective	OPUS en→xx			OPUS xx→en			Flores en→xx			Flores xx→en		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
MTInstruct	11.54	25.33	64.68	18.59	33.25	68.75	3.30	17.10	42.62	11.37	27.14	55.82
MT+Mono-full	9.89	22.42	62.56	15.43	29.04	65.45	3.00	16.68	42.34	10.26	25.15	53.67
MT+Mono-half	10.23	22.45	62.59	15.51	29.65	66.18	3.18	16.91	**42.69**	10.66	26.15	54.41
MT+Mono-full+Align	10.15	22.35	62.39	15.72	29.86	66.54	3.07	16.59	42.54	10.61	25.58	54.59
MT+Mono-half+Align	10.09	22.61	63.01	16.00	30.34	67.15	3.10	16.75	**42.63**	10.79	26.27	54.87
MT+Mono-full+Align+Hint+Revise	10.33	23.04	63.06	17.16	31.61	67.40	3.23	16.70	**42.74**	10.98	26.18	54.97
MT+Mono-half+Align+Hint+Revise	10.62	23.10	63.07	17.32	31.80	67.43	3.20	16.93	**42.97**	11.09	26.77	55.41

Table 4: **Results of BLOOMZ+24 fine-tuned incorporating monolingual instructions on BLOOMZ-7b1.** Scores that surpass the MTInstruct baseline are marked in **bold**.

Fine-tuned Languages	Objective	Zero-shot Directions				Supervised Directions			
		Directions	BLEU	chrF++	COMET	Directions	BLEU	chrF++	COMET
-	w/o fine-tuning	overall	6.89	19.14	57.95	en→xx	13.38	26.65	64.28
						xx→en	21.70	42.05	72.72
		seen→seen	16.95	30.78	74.58	en→seen	20.13	32.87	76.99
		seen→unseen	2.30	13.31	49.98	en→unseen	6.63	20.43	51.56
		unseen→seen	7.78	20.07	62.74	seen→en	26.30	48.70	78.22
		unseen→unseen	2.37	14.83	46.06	unseen→en	17.10	35.40	67.23
de-nl-ru	MTInstruct	overall	8.38	22.75	59.93	en→xx	17.05	32.02	69.26
						xx→en	25.13	45.02	76.29
		seen→seen	14.52	27.25	70.48	en→seen	17.60	29.87	73.81
		seen→unseen	6.14	22.82	54.75	en→unseen	16.50	34.17	64.70
		unseen→seen	7.56	19.22	61.99	seen→en	25.73	47.07	77.52
		unseen→unseen	6.85	23.45	54.07	unseen→en	24.53	42.97	75.06
	MT+Align	overall	**8.86**	**23.30**	**60.70**	en→xx	16.63	31.73	68.79
						xx→en	**25.62**	**45.37**	**76.45**
		seen→seen	**14.77**	**27.80**	**71.07**	en→seen	15.80	28.47	72.35
		seen→unseen	**6.31**	**23.08**	**54.81**	en→unseen	**17.47**	**35.00**	**65.24**
		unseen→seen	**8.61**	**20.24**	**63.81**	seen→en	**25.90**	**47.13**	77.47
		unseen→unseen	**7.15**	**23.70**	**54.51**	unseen→en	**25.33**	**43.60**	**75.43**
ar-de-fr-nl-ru-zh	MTInstruct	overall	11.79	26.36	63.22	en→xx	21.18	35.52	70.86
						xx→en	28.35	48.00	77.30
		seen→seen	22.68	35.32	76.39	en→seen	26.20	37.77	78.22
		seen→unseen	7.10	24.50	55.18	en→unseen	16.17	33.27	63.50
		unseen→seen	12.56	24.74	68.83	seen→en	31.97	52.93	79.72
		unseen→unseen	6.78	22.62	53.69	unseen→en	24.73	43.07	74.88
	MT+Align	overall	**12.13**	**26.65**	**63.23**	en→xx	**21.33**	**35.65**	**70.99**
						xx→en	**28.60**	**48.27**	**77.49**
		seen→seen	**23.67**	**36.53**	**76.89**	en→seen	**26.30**	37.63	**78.25**
		seen→unseen	**7.27**	24.32	54.96	en→unseen	**16.37**	**33.67**	**63.73**
		unseen→seen	**12.92**	**25.29**	**69.10**	seen→en	**32.03**	**53.07**	**79.93**
		unseen→unseen	6.68	22.30	53.19	unseen→en	**25.17**	**43.47**	**75.05**

Table 5: **Results of BLOOMZ+3 without fine-tuning or fine-tuned with MTInstruct, or MT+Align.** Scores that surpass the MTInstruct baseline are marked in **bold**. xx includes seen and unseen languages.

MT+Align+Hint+Revise and pivot-based translation are reported in App. C and F. In the de-nl-ru setting, where BLOOMZ was fine-tuned with the three unseen languages, we noticed MT+Align consistently outperformed the MTInstruct baseline across all evaluated zero-shot directions. Notably, MT+Align enhanced the translation quality for unseen→seen and seen→unseen directions compared to w/o fine-tuning and MTInstruct, given that the model was solely fine-tuned on de, nl, and ru data. This suggested AlignInstruct not only benefits the languages supplied in the data but also has a positive impact on other languages through cross-lingual alignment supervision. In terms of supervised directions involving English, we noticed performance improvements associated with unseen languages, and regression in seen ones. The regression may be attributed to forgetting for the absence of seen languages in fine-tuning data. Indeed, continuous exposure to English maintained the translation quality for seen→en. As LoRA is modular, the regression can be mitigated by detaching the LoRA parameters for seen languages.

The ar-de-fr-nl-ru-zh setting yielded a consistently higher translation quality across all directions when compared with the de-nl-ru setting. This improvement was expected, as all the six languages were included. Translation quality improved for when generating seen languages under the zero-shot scenario. However, the same observation cannot be made for unseen languages. This phenomenon underscored the effectiveness of

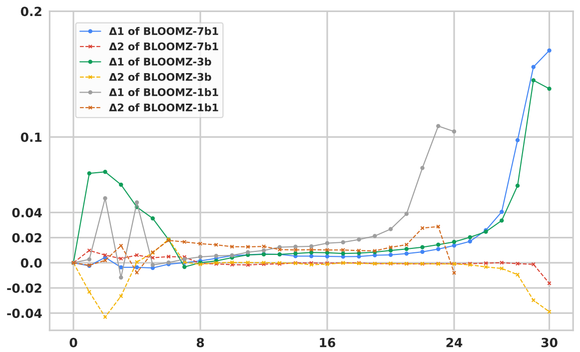

Figure 3: **Differences in cosine similarity of layer-wise embeddings for BLOOMZ+24.** $\Delta 1$ represents the changes from the unmodified BLOOMZ to the one on MTInstruct, and $\Delta 2$ from MTInstruct to MT+Align.

AlignInstruct in enhancing translation quality for BLOOMZ's supported languages, but suggested limitations for unseen languages when mixed with supported languages in zero-shot scenarios. In the supervised directions, we found all translation directions surpassed the performance of the MTInstruct baseline. This highlighted the overall effectiveness of AlignInstruct in enhancing translation quality across a range of supervised directions.

4.5 How did MTInstruct and AlignInstruct Impact BLOOMZ's Representations?

This section analyzed the layer-wise cosine similarities between the embeddings of parallel sentences to understand the changes in internal representations after fine-tuning. The parallel sentences were prepared from the English-centric validation datasets. We then mean-pool the outputs at each layer as sentence embeddings and compute the cosine similarities, as illustrated in Fig. 3. Results for BLOOMZ+3 are discussed in App. D.

We observed that, after MTInstruct fine-tuning, the cosine similarities rose in nearly all layers ($\Delta 1$, Fig. 3). This may be interpreted as enhanced cross-lingual alignment, and as indicating the acquisition of translation capabilities. Upon further combination with AlignInstruct ($\Delta 2$, Fig. 3), the degree of cross-lingual alignment rose in the early layers (layers 4 - 7) then diminished in the final layers (layers 29 & 30). This pattern aligned with the characteristics of encoder-decoder multilingual NMT models, where language-agnostic encoder representations with language-specific decoder representations improve multilingual NMT performance (Liu et al., 2021; Wu et al., 2021; Mao et al., 2023). This highlights the beneficial impact of AlignInstruct.

5 Related Work

Prompting LLMs for MT LLMs have shown good performance for multilingual MT through few-shot in-context learning (ICL) (Jiao et al., 2023). Vilar et al. (2023) showed that high-quality examples can improve MT based on PaLM (Chowdhery et al., 2022). Agrawal et al. (2023) and Zhang et al. (2023a) explored strategies to compose better examples for few-shot prompting for XGLM-7.5B (Lin et al., 2022) and GLM-130B (Zeng et al., 2023). Ghazvininejad et al. (2023), Peng et al. (2023), and Moslem et al. (2023) claimed that dictionary-based hints and domain-specific style information can improve prompting OPT (Zhang et al., 2022), GPT-3.5 (Brown et al., 2020), and BLOOM (Scao et al., 2022) for MT. He et al. (2023) used LLMs to mine useful knowledge for prompting GPT-3.5 for MT.

Fine-tuning LLMs for MT ICL-based methods do not support languages unseen during pre-training. Current approaches address this issue via fine-tuning. Zhang et al. (2023b) explored adding new languages to LLaMA (Touvron et al., 2023a) with interactive translation task for unseen high-resource languages. However, similar task datasets are usually not available for most unseen, low-resource languages. Li et al. (2023) and Xu et al. (2023a) showed multilingual fine-tuning with translation instructions can improve the translation ability in supported languages. Our study extended their finding to apply in the context of unseen, low-resource languages. In parallel research, Yang et al. (2023) undertook MT instruction fine-tuning in a massively multilingual context for unseen languages. However, their emphasis was on fine-tuning curriculum based on resource availability of languages, whereas we exclusively centered on low-resource languages and instruction tuning tasks.

6 Conclusion

In this study, we introduced AlignInstruct for enhancing the fine-tuning of LLMs for MT in unseen, low-resource languages while limiting the use of additional training corpora. Our multilingual and zero-shot findings demonstrated the strength of AlignInstruct over the MTInstruct baseline and other instruction variants. Our future work pertains to exploring using large monolingual corpora of unseen languages for MT and refining the model capability to generalize across diverse MT prompts.

Limitations

Multilingual LLMs In this study, our investigations were confined to the fine-tuning of BLOOMZ models with sizes of 1.1B, 3B, and 7.1B. We did not experiment with the 175B BLOOMZ model due to computational resource constraints. However, examining this model could provide valuable insights into the efficacy of our proposed techniques. Additionally, it would be instructive to experiment with other recent open-source multilingual LLMs, such as mGPT (Shliazhko et al., 2022) and LLaMa2 (Touvron et al., 2023b).

PEFT Methods and Adapters As discussed in the BLOOM+1 paper (Yong et al., 2023), alternative PEFT techniques, such as $(IA)^3$ (Liu et al., 2022), have the potential to enhance the adaptation performance of LLM pre-training for previously unseen languages. These approaches are worth exploring for MT fine-tuning in such languages, in addition to the LoRA methods employed in this study. Furthermore, our exploration was limited to fine-tuning multiple languages using shared additional parameters. Investigating efficient adaptation through the use of the mixture of experts (MoE) approach for MT tasks (Fan et al., 2021; Costa-jussà et al., 2022; Mohammadshahi et al., 2022; Koishekenov et al., 2023; Xu et al., 2023b) presents another intriguing avenue for LLM fine-tuning.

Instruction Fine-tuning Data Another limitation of our study is that we exclusively explored MT instruction fine-tuning using fixed templates to create MT and alignment instructions. Investigating varied templates (either manually (Yang et al., 2023) or automatically constructed (Zhou et al., 2023)) might enhance the fine-tuned MT model's ability to generalize across different MT task descriptions. Additionally, leveraging large monolingual corpora in unseen languages could potentially enhance the effectiveness of monolingual instructions for MT downstream tasks, offering further insights beyond the resource-constrained scenarios examined in this work. Furthermore, the creation and utilization of instruction tuning datasets, akin to xP3 (Muennighoff et al., 2023), for unseen, low-resource languages could potentially amplify LLMs' proficiency in following instructions in such languages. Zhu et al. (2023) has investigated multilingual instruction tuning datasets. However, the scalability of such high-quality datasets to thousands of low-resource languages still remains to be addressed.

Comparison with the State-of-the-art Multilingual NMT Models In this study, we refrained from contrasting translations in low-resource languages with best-performing multilingual NMT models like NLLB-200 (Costa-jussà et al., 2022), as our primary objective centered on enhancing the MTInstruct baseline through improved cross-lingual alignment within LLMs, rather than delving into the best combination of techniques for MT fine-tuning in LLMs. In future exploration, our methods can potentially be integrated with the MT fine-tuning paradigm proposed by the concurrent work of Xu et al. (2023a), paving the way for elevating the state-of-the-art translation quality using LLMs.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. Palm 2 technical report. *CoRR*, abs/2305.10403.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In

Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Javier de la Rosa and Andrés Fernández. 2022. Zero-shot reading comprehension and reasoning for spanish with BERTIN GPT-J-6B. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022*, volume 3202 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. CoRR, abs/2302.07856.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *CoRR*, abs/2305.04118.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585, Toronto, Canada. Association for Computational Linguistics.

Patrik Lambert, Simon Petitrenaud, Yanjun Ma, and Andy Way. 2012. What types of word alignment improve statistical machine translation? *Mach. Transl.*, 26(4):289–323.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Chen, and Jiajun Chen. 2023. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *CoRR*, abs/2305.15083.

Xiaozhuan Liang, Ningyu Zhang, Siyuan Cheng, Zhenru Zhang, Chuanqi Tan, and Huajun Chen. 2022. Contrastive demonstration tuning for pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 799–811, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1259–1273, Online. Association for Computational Linguistics.

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

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zhuoyuan Mao, Chenhui Chu, Raj Dabre, Haiyue Song, Zhen Wan, and Sadao Kurohashi. 2022. When do contrastive word alignments improve many-to-many neural machine translation? In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1766–1775, Seattle, United States. Association for Computational Linguistics.

Zhuoyuan Mao, Raj Dabre, Qianying Liu, Haiyue Song, Chenhui Chu, and Sadao Kurohashi. 2023. Exploring the impact of layer normalization for zero-shot neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1300–1316, Toronto, Canada. Association for Computational Linguistics.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

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.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Martin Müller and Florian Laurent. 2022. Cedille: A large autoregressive french language model. *CoRR*, abs/2202.03371.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

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

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *CoRR*, abs/2303.13780.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Explicit cross-lingual pre-training for unsupervised machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 770–779, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. 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 V. 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 Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *CoRR*, abs/2204.07580.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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.

Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1022–1034, Online. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2021. Language tags matter for zero-shot neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3001–3007, Online. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023a. A paradigm shift in machine translation: Boosting translation performance of large language models. *CoRR*, abs/2309.11674.

Haoran Xu, Weiting Tan, Shuyue Stella Li, Yunmo Chen, Benjamin Van Durme, Philipp Koehn, and Kenton Murray. 2023b. Condensing multilingual knowledge with lightweight language-specific modules. *CoRR*, abs/2305.13993.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *CoRR*, abs/2305.18098.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Zhang Ze Yu, Lau Jia Jaw, Wong Qin Jiang, and Zhang Hui. 2023. Fine-tuning language models with generative adversarial feedback. *CoRR*, abs/2305.06176.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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. *CoRR*, abs/2308.04948.

A Training Data Statistics

Training data statistics of BLOOMZ+24 are shown in Tab. 6. Several selected languages involved previously unseen scripts by BLOOMZ, but such fine-tuning is practical as BLOOMZ is a byte-level model with the potential to adapt to any language. Note that our proposed methods can be applied to any byte-level generative LLMs.

B Implementation Details

We employed 128 V100 GPUs for the BLOOMZ+24 and 32 V100 GPUs for the BLOOMZ+3 experiments. The batch sizes were configured at 4 sentences for BLOOMZ-7b1 and 8 sentences for both BLOOMZ-3b and BLOOMZ-1b1, per GPU device. We configured LoRA with a rank of 8, an alpha of 32, and a dropout of 0.1. Consequently, the BLOOMZ-7b1, BLOOMZ-3b, and BLOOMZ-1b1 models had 3.9M, 2.5M, and 1.2M trainable parameters, respectively, constituting approximately 0.05 - 0.10% of the parameters in the original models. We conducted training for 5 epochs, ensuring a stable convergence is achieved. To facilitate this stability, we introduced a warm-up ratio of 0.03 into our training process. Maximum input and output length were set as 384. S for HintInstruct was set as 5 at most. Additionally, we used mixed precision training (Micikevicius et al., 2018) to expedite computation using DeepSpeed (Rasley

Language	ISO 639-1	Language Family	Subgrouping	Script	Seen Script	#sent.
Afrikaans	af	Indo-European	Germanic	Latin	✓	275,512
Amharic	am	Afro-Asiatic	Semitic	Ge'ez	✗	89,027
Belarusian	be	Indo-European	Balto-Slavic	Cyrillic	✗	67,312
Welsh	cy	Indo-European	Celtic	Latin	✓	289,521
Irish	ga	Indo-European	Celtic	Latin	✓	289,524
Scottish Gaelic	gd	Indo-European	Celtic	Latin	✓	16,316
Galician	gl	Indo-European	Italic	Latin	✓	515,344
Hausa	ha	Afro-Asiatic	Chadic	Latin	✓	97,983
Georgian	ka	Kartvelian	Georgian-Zan	Georgian	✗	377,306
Kazakh	kk	Turkic	Common Turkic	Cyrillic	✗	79,927
Khmer	km	Austroasiatic	Khmeric	Khmer	✗	111,483
Kyrgyz	ky	Turkic	Common Turkic	Cyrillic	✗	27,215
Limburgish	li	Indo-European	Germanic	Latin	✓	25,535
Burmese	my	Sino-Tibetan	Burmo-Qiangic	Myanmar	✗	24,594
Norwegian Bokmål	nb	Indo-European	Germanic	Latin	✓	142,906
Norwegian Nynorsk	nn	Indo-European	Germanic	Latin	✓	486,055
Occitan	oc	Indo-European	Italic	Latin	✓	35,791
Sinhala	si	Indo-European	Indo-Aryan	Sinhala	✗	979,109
Tajik	tg	Indo-European	Iranian	Cyrillic	✗	193,882
Turkmen	tk	Turkic	Common Turkic	Latin	✓	13,110
Tatar	tt	Turkic	Common Turkic	Cyrillic	✗	100,843
Uyghur	ug	Turkic	Common Turkic	Arabic	✓	72,170
Northern Uzbek	uz	Turkic	Common Turkic	Latin	✓	173,157
Eastern Yiddish	yi	Indo-European	Germanic	Hebrew	✗	15,010
Total						4,498,632

Table 6: **Statistics of training data for BLOOMZ+24**: 24 unseen, low-resource languages for BLOOMZ. ✓ and ✗ indicate whether script is seen or unseen.

et al., 2020). We tuned the optimal learning rate for each individual experiment according to validation loss. We conducted all experiments once due to computational resource constraints and reported the average scores across all languages.

C Results of MT+Align+Hint+Revise for BLOOMZ+3

We present the results in Tab. 7. Co-referencing the results in Tab. 5, compared with MT+Align, we observed a clear advantage for the MT+Align+Hint+Revise setting in supervised directions involving English (en→seen and seen→en) in the ar-fr-de-nl-ru-zh setting. This result suggested that AlignInstruct's variants played a crucial role in preserving the BLOOMZ's capabilities for supported languages. However, in all other scenarios, AlignInstruct alone proved sufficient to enhance the performance beyond the MTInstruct baseline, but hard to achieve further improvements with additional instructions.

D Representation Change of BLOOMZ+3

The representation change observed in de-nl-ru was consistent with the findings presented in Sec. 4.5, which highlighted an initial increase in cross-lingual alignment in the early layers, followed by a decrease in the final layers. When mixing fine-tuning data with supported languages, the changes exhibited more intricate patterns. As illustrated by ar-fr-zh in ar-de-fr-nl-ru-zh in Fig. 4, sentence alignment declined after MTInstruct fine-tuning but elevated after further combining with AlignInstruct. We leave the interpretation of this nuanced behavior in future work.

E Inference using Different MT Prompts

We investigated the performance of fine-tuned models when using various MT prompts during inference, aiming to understand models' generalization capabilities with different test prompts. We examined five MT prompts for the fine-tuned models of BLOOMZ-7b1, following Zhang et al. (2023a), which are presented in Tab. 8. The results, showcased in Tab. 9, revealed that in comparison to the

Languages	Zero-shot Directions				Supervised Directions			
	Directions	BLEU	chrF++	COMET	Directions	BLEU	chrF++	COMET
de-nl-ru	overall	**8.94**	**23.53**	**60.67**	en→xx	16.70	31.83	68.98
					xx→en	**25.18**	45.00	**76.45**
	seen→seen	14.00	**27.58**	**70.59**	en→seen	15.97	28.53	**72.69**
	seen→unseen	**6.49**	**23.01**	**54.92**	en→unseen	**17.43**	**35.13**	**65.27**
	unseen→seen	**9.50**	**21.90**	**64.69**	seen→en	25.33	46.70	77.51
	unseen→unseen	6.73	22.70	53.34	unseen→en	**25.03**	**43.30**	**75.39**
ar-de-fr-nl-ru-zh	overall	**12.07**	**26.67**	63.13	en→xx	**21.62**	**36.12**	**70.94**
					xx→en	**28.92**	**48.60**	**77.50**
	seen→seen	**23.52**	**36.13**	**76.62**	en→seen	**26.87**	**38.40**	**78.40**
	seen→unseen	**7.16**	24.48	55.02	en→unseen	**16.37**	**33.83**	63.49
	unseen→seen	**12.91**	**25.23**	**68.91**	seen→en	**32.57**	**53.70**	**80.06**
	unseen→unseen	6.73	**22.65**	53.12	unseen→en	**25.27**	**43.50**	**74.93**

Table 7: **Results of BLOOMZ+3 with MT+Align+Hint+Revise.** Co-referencing Tab. 5, scores that surpass the MTInstruct baseline are marked in **bold**.

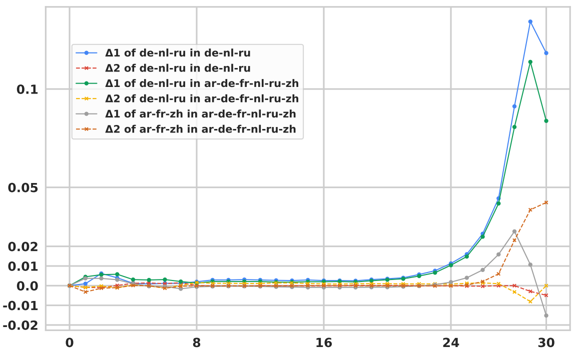

Figure 4: **Differences in cosine similarity of layer-wise embeddings for BLOOMZ+3.** $\Delta 1$ represents the changes from the unmodified BLOOMZ to the one on MTInstruct, and $\Delta 2$ from MTInstruct to MT+Align.

Prompt	Definition
PROMPT-default	Translate from Y to X. Y: $y_1 y_2 \dots y_M$. X:
PROMPT-1	Y: $y_1 y_2 \dots y_M$. X:
PROMPT-2	$y_1 y_2 \dots y_M$. X:
PROMPT-3	Translate to X. Y: $y_1 y_2 \dots y_M$. X:
PROMPT-4	Translate from Y to X. $y_1 y_2 \dots y_M$. X:
PROMPT-5	Translate to X. $y_1 y_2 \dots y_M$. X:

Table 8: **MT prompt variants investigated for fine-tuned models.** These MT prompts are following the design in Zhang et al. (2023a).

default prompt used during fine-tuning, the translation performance tended to decline when using other MT prompts. We observed that MT+Align consistently surpasses MTInstruct for xx→en translations, though the results were mixed for en→xx directions. Certain prompts, such as PROMPT-3 and PROMPT-4, exhibited a minor performance drop, while others significantly impacted translation quality. These findings underscored the need for enhancing the models' ability to generalize across diverse MT prompts, potentially by incorporating a range of MT prompt templates during the fine-tuning process, as stated in the Limitations section.

F Zero-shot Translation using English as Pivot

Pivot translation serves as a robust technique for zero-shot translation, especially given that we used English-centric data during fine-tuning. In Tab. 10, we present results that utilize English as an intermediary pivot for translations between non-English language pairs. Our findings indicated that employing the English pivot typically yielded an enhancement of approximately 1.1 - 1.2 BLEU points compared to direct translations in zero-shot directions when fine-tuning BLOOMZ. When contrasting the MTInstruct baseline with our proposed MT+Align,

Prompt	Objective	en→xx			xx→en		
		BLEU	chrF++	COMET	BLEU	chrF++	COMET
PROMPT-default	MTInstruct	11.54	25.33	64.68	18.59	33.25	68.75
	MT+Align	**12.28**	**26.17**	**65.28**	**18.72**	**34.02**	**69.75**
PROMPT-1	MTInstruct	5.29	11.31	50.74	7.87	20.08	57.10
	MT+Align	**5.30**	**11.38**	**51.29**	**8.93**	**20.77**	**58.01**
PROMPT-2	MTInstruct	2.20	6.68	45.78	7.15	19.08	57.03
	MT+Align	1.91	5.35	43.92	**7.61**	18.80	56.40
PROMPT-3	MTInstruct	10.59	22.69	62.77	15.85	29.93	66.64
	MT+Align	9.20	20.80	61.45	**16.17**	**30.58**	**67.75**
PROMPT-4	MTInstruct	8.67	20.73	61.32	15.20	28.95	65.51
	MT+Align	**8.91**	20.53	**61.55**	**16.25**	**30.67**	**67.06**
PROMPT-5	MTInstruct	6.61	14.55	55.93	10.88	22.41	60.48
	MT+Align	6.02	12.28	52.72	**11.83**	**23.85**	**61.28**

Table 9: **Results of using different MT prompts for BLOOMZ-7b1 fine-tuned models during inference.** Refer to Tab. 8 for details about definitions of different MT prompts. We report the average results for the BLOOMZ+24 setting. Results better than the MTInstruct baseline are marked in **bold**.

MTInstruct	BLEU	chrF++	COMET	**MT+Align**	BLEU	chrF++	COMET
overall	11.79	26.36	63.22	overall	**12.13**	**26.65**	**63.23**
seen→seen	22.68	35.32	76.39	seen→seen	**23.67**	**36.53**	**76.89**
seen→unseen	7.10	24.50	55.18	seen→unseen	**7.27**	24.32	54.96
unseen→seen	12.56	24.74	68.83	unseen→seen	**12.92**	**25.29**	**69.10**
unseen→unseen	6.78	22.62	53.69	unseen→unseen	6.68	22.30	53.19
MTInstruct with English pivot	BLEU	chrF++	COMET	**MT+Align with English pivot**	BLEU	chrF++	COMET
overall	12.99	28.01	65.38	overall	**13.25**	**28.30**	**65.57**
seen→seen	23.10	35.30	76.30	seen→seen	**23.48**	**35.57**	**76.43**
seen→unseen	9.00	27.67	59.54	seen→unseen	**9.28**	**28.03**	**59.73**
unseen→seen	13.18	24.98	68.77	unseen→seen	**13.36**	**25.22**	**68.94**
unseen→unseen	8.57	25.77	58.17	unseen→unseen	**8.83**	**26.07**	**58.42**

Table 10: **Results of BLOOMZ+3 using English as a pivot language for zero-shot translation evaluation.** Results of MT+Align surpassing corresponding those of MTInstruct are marked in **bold**.

we observed that combining AlignInstruct consistently boosted performance in pivot translation scenarios.

G Per Language Result Details of BLOOMZ+24 and BLOOMZ+3

We present per language detailed results of original BLOOMZ-7b1 and fine-tuned BLOOMZ-7b1 models in Tab. 11, 12, 13, 14, 15, 16, 17, 18, respectively for the BLOOMZ+24 and BLOOMZ+3 settings.

Language	OPUS en→xx			OPUS xx→en			Flores en→xx			Flores xx→en		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
af	3.8	13.2	56.38	7.6	22.0	59.14	2.6	14.9	33.60	20.1	38.0	65.61
am	0.1	0.3	33.17	0.5	8.3	43.57	0.3	0.6	30.65	1.9	12.6	46.24
be	4.2	5.1	47.26	7.3	17.5	48.57	0.4	3.3	31.58	4.2	22.3	49.27
cy	2.7	10.5	53.21	6.2	16.0	53.25	1.2	11.2	34.17	6.0	20.3	53.45
ga	1.2	10.6	42.85	4.0	16.4	46.05	1.2	11.6	33.94	5.5	19.6	46.97
gd	9.3	16.0	51.40	47.6	55.9	59.30	1.2	11.2	36.28	4.2	18.8	43.73
gl	4.5	25.6	64.93	17.2	36.7	66.07	13.4	38.5	74.77	51.0	67.8	85.77
ha	0.1	5.4	38.42	0.3	11.2	42.58	1.5	10.2	35.77	6.9	18.9	47.37
ka	0.3	1.9	31.97	0.6	9.2	44.48	0.4	1.4	28.81	2.4	17.0	47.57
kk	4.3	4.9	50.51	5.1	14.2	51.51	0.5	1.6	33.66	5.1	19.8	51.40
km	2.8	4.5	51.68	3.9	11.1	50.40	0.8	2.9	39.56	5.6	16.2	50.42
ky	10.0	10.6	54.23	10.3	24.0	55.99	0.6	1.6	30.19	3.8	17.9	48.05
li	6.6	16.2	61.39	5.9	24.8	61.65	2.0	14.9	41.01	9.8	29.8	46.92
my	1.8	2.4	45.44	3.0	5.0	48.33	0.4	0.8	29.58	1.0	3.7	44.15
nb	5.8	18.2	57.01	13.9	33.0	56.37	3.9	19.3	46.74	19.8	40.3	63.56
nn	6.3	18.6	62.33	8.9	25.3	56.28	3.7	19.7	41.75	16.9	37.5	62.37
oc	6.0	13.6	60.16	5.1	18.6	58.51	9.6	33.6	67.22	53.0	68.5	79.57
si	0.6	2.0	41.84	1.6	9.4	48.58	0.5	1.4	28.08	1.6	9.1	42.67
tg	0.4	1.4	36.26	1.1	11.8	43.54	0.4	1.5	26.63	3.3	18.0	43.79
tk	7.9	10.6	55.34	5.3	13.0	47.33	0.7	8.7	31.94	4.2	20.1	45.05
tt	0.0	1.0	28.98	0.2	13.3	42.85	0.3	1.4	27.86	4.2	20.2	48.15
ug	0.0	0.4	32.44	0.3	11.2	45.69	0.3	0.9	31.34	3.0	16.5	48.99
uz	0.7	2.1	35.94	1.0	12.8	41.86	1.5	11.5	40.65	3.1	18.7	49.43
yi	7.3	16.5	57.47	4.0	23.0	63.91	0.7	1.7	33.22	2.1	15.6	41.87
avg.	3.61	8.82	47.94	6.70	18.49	51.49	2.00	9.35	37.04	9.95	24.47	52.18

Table 11: Detailed results of BLOOMZ-7b1 without fine-tuning.

Language	OPUS en→xx			OPUS xx→en			Flores en→xx			Flores xx→en		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
af	25.0	41.4	71.05	38.5	52.3	78.94	10.1	31.0	45.42	33.9	51.1	72.66
am	3.0	12.8	59.55	3.4	19.8	59.71	0.2	5.2	42.97	1.4	16.0	49.47
be	8.9	14.9	55.16	14.0	24.9	62.37	0.7	12.3	30.90	3.7	21.0	49.99
cy	20.2	38.0	71.55	33.2	49.3	77.72	5.0	20.3	38.38	13.1	30.2	57.47
ga	15.6	37.1	63.87	29.2	49.1	75.94	3.7	21.2	39.17	12.5	30.3	57.53
gd	13.1	24.7	62.14	66.0	69.6	77.70	2.2	19.6	40.75	7.1	22.3	50.05
gl	16.9	37.6	70.62	24.7	43.6	75.62	21.9	45.2	77.26	46.6	64.5	86.86
ha	12.3	32.7	71.75	10.0	29.8	64.51	1.9	17.1	49.24	6.8	22.1	48.81
ka	4.6	18.1	67.39	10.0	24.3	60.50	0.3	6.8	27.46	1.5	14.9	46.10
kk	12.6	19.5	66.07	14.6	28.2	71.80	0.8	13.0	35.76	3.9	19.7	52.24
km	19.7	25.2	63.24	13.9	32.1	75.02	0.5	12.3	35.60	6.2	22.4	56.45
ky	16.0	20.5	66.27	21.1	33.8	73.06	0.9	12.7	36.10	3.0	17.5	50.40
li	13.5	32.8	70.97	21.3	35.7	67.20	3.3	19.9	42.21	14.6	31.4	55.94
my	6.2	14.3	58.04	5.2	15.6	63.65	0.2	12.9	40.37	1.3	12.7	48.38
nb	12.7	30.4	63.27	22.2	42.1	76.74	7.9	28.4	44.15	25.6	44.3	72.56
nn	18.3	38.0	77.18	27.1	47.7	81.80	7.3	25.7	45.35	24.3	42.9	70.06
oc	10.0	20.0	63.31	13.4	27.1	69.89	8.0	27.5	51.48	46.9	63.5	79.64
si	5.2	21.4	68.16	11.5	26.4	70.79	0.9	12.9	41.73	3.7	19.2	57.41
tg	5.5	22.0	66.08	8.0	25.9	60.54	1.1	15.8	65.14	3.1	19.6	45.06
tk	24.4	26.7	65.53	30.4	37.8	70.39	0.7	10.8	42.36	3.9	18.8	46.23
tt	1.9	17.6	60.01	3.6	19.6	54.99	0.4	13.7	50.78	1.6	14.3	42.58
ug	1.2	19.7	49.76	4.2	21.2	61.34	0.4	12.9	35.88	1.7	16.7	50.29
uz	3.1	18.2	62.12	5.7	22.0	61.12	0.5	3.6	34.67	3.9	18.8	50.32
yi	7.1	24.3	59.13	14.9	20.2	58.66	0.3	9.5	29.77	2.5	17.2	43.27
avg.	11.54	25.33	64.68	18.6	33.25	68.75	3.30	17.10	42.62	11.37	27.14	55.82

Table 12: Detailed results of BLOOMZ-7b1 fine-tuned with MTInstruct for BLOOMZ+24.

Language	OPUS en→xx			OPUS xx→en			Flores en→xx			Flores xx→en		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
af	25.0	41.9	70.72	36.9	52.2	78.68	10.6	31.9	45.84	33.5	51.1	72.84
am	3.4	13.2	60.62	4.9	22.8	62.43	0.3	5.4	44.20	1.4	16.4	51.05
be	8.3	14.5	55.23	13.9	25.1	62.72	0.8	12.5	30.93	3.6	20.6	49.14
cy	20.6	39.0	71.73	33.8	49.4	77.55	4.7	20.3	38.70	14.6	31.5	58.34
ga	17.6	39.3	65.76	32.6	52.7	77.49	3.4	21.4	39.99	13.6	31.6	58.73
gd	15.6	27.2	62.09	48.1	55.4	75.90	2.3	20.3	40.81	7.4	22.0	49.99
gl	17.1	37.2	70.85	24.4	43.3	75.90	21.7	44.9	77.09	45.6	63.5	86.60
ha	14.6	35.0	73.34	11.4	31.3	65.69	1.9	17.3	50.88	7.4	22.5	49.57
ka	4.9	18.9	67.54	10.5	25.3	61.27	0.3	6.9	27.61	2.1	16.0	47.04
kk	12.3	19.3	65.73	15.6	28.0	71.01	0.9	13.0	35.86	4.1	19.8	52.43
km	20.4	26.5	63.38	14.4	35.2	75.62	0.6	12.5	35.44	7.1	22.9	57.81
ky	15.8	19.6	64.74	23.3	35.8	74.70	0.9	13.3	36.71	2.9	17.4	50.06
li	13.2	29.4	65.18	22.3	38.2	71.93	3.1	19.7	42.58	12.5	28.7	54.60
my	7.6	15.4	58.84	6.3	18.0	66.45	0.3	13.3	40.97	1.2	14.4	50.79
nb	13.5	31.4	64.08	24.2	44.2	77.58	7.9	28.7	44.12	25.5	44.9	72.72
nn	19.0	38.0	77.61	28.5	47.7	81.68	7.0	26.7	46.14	25.8	44.1	70.55
oc	9.1	19.3	63.25	13.5	27.5	70.13	7.5	25.9	50.48	47.3	63.8	79.39
si	5.1	22.1	69.60	13.9	29.1	72.51	1.1	13.1	43.01	5.6	22.7	61.89
tg	6.6	23.7	66.31	8.8	27.2	61.52	0.9	15.6	65.51	3.4	19.9	45.45
tk	27.2	26.2	66.11	31.2	38.7	70.47	0.7	11.4	43.64	3.8	18.2	45.87
tt	2.1	18.6	60.75	5.0	21.5	56.95	0.4	13.3	50.64	1.5	13.7	42.76
ug	1.1	20.7	51.12	5.5	23.4	63.42	0.4	13.8	37.51	2.1	16.3	50.45
uz	3.5	18.6	62.09	7.4	23.3	62.01	0.2	1.9	34.50	3.7	18.2	50.09
yi	11.1	33.1	70.13	12.8	21.2	60.47	0.4	9.8	30.08	2.6	17.0	42.57
avg.	12.28	26.17	65.28	18.72	34.02	69.75	3.26	17.20	43.05	11.60	27.38	56.28

Table 13: Detailed results of BLOOMZ-7b1 fine-tuned with MT+Align for BLOOMZ+24.

Zero-shot	BLEU	chrF++	COMET	Supervised	BLEU	chrF++	COMET
ar-de	1.4	14.8	56.19	en-ar	11.1	32.4	75.66
ar-fr	21.9	46.1	74.19	en-de	12.2	29.2	59.16
ar-nl	0.6	11.2	56.59	en-fr	26.8	49.2	77.42
ar-ru	3.1	6.2	48.41	en-nl	2.0	16.0	46.52
ar-zh	18.4	14.4	73.65	en-ru	5.7	16.1	49.00
de-ar	2.0	17.8	64.91	en-zh	22.5	17.0	77.90
de-fr	12.0	33.4	63.45	avg.	13.38	26.65	64.28
de-nl	3.7	17.9	47.30				
de-ru	1.3	11.8	45.53				
de-zh	8.9	7.6	61.52				
fr-ar	11.2	33.4	74.20		BLEU	chrF++	COMET
fr-de	4.6	23.4	48.83	ar-en	26.7	48.4	78.12
fr-nl	2.8	17.2	52.14	de-en	21.1	38.5	71.99
fr-ru	3.1	10.4	45.12	fr-en	27.7	49.8	79.46
fr-zh	20.9	17.0	76.20	nl-en	12.3	31.1	61.29
nl-ar	1.3	13.2	59.46	ru-en	17.9	36.6	68.40
nl-de	5.9	22.8	46.49	zh-en	24.5	47.9	77.08
nl-fr	9.6	29.6	58.30	avg.	21.70	42.05	72.72
nl-ru	0.8	9.0	42.83				
nl-zh	3.3	3.7	53.96				
ru-ar	6.5	25.3	68.38				
ru-de	2.0	17.0	48.06				
ru-fr	15.7	38.7	67.54				
ru-nl	0.5	10.5	46.14				
ru-zh	10.7	11.3	67.18				
zh-ar	8.6	29.7	73.47				
zh-de	1.6	17.6	49.90				
zh-fr	20.7	44.1	75.79				
zh-nl	0.6	10.4	48.53				
zh-ru	2.9	8.6	44.13				
avg.	6.89	19.14	57.95				
seen→seen	16.95	30.78	74.58	en→seen	20.13	32.87	76.99
seen→unseen	2.30	13.31	49.98	en→unseen	6.63	20.43	51.56
unseen→seen	7.78	20.07	62.74	seen→en	26.30	48.70	78.22
unseen→unseen	2.37	14.83	46.06	unseen→en	17.10	35.40	67.23

Table 14: Detailed results of BLOOMZ-7b1 without fine-tuning.

Zero-shot	BLEU	chrF++	COMET	Supervised	BLEU	chrF++	COMET
ar-de	4.7	20.9	56.43	en-ar	9.1	27.2	71.47
ar-fr	20.8	42.5	71.47	en-de	19.8	36.1	66.53
ar-nl	7.2	22.9	58.29	en-fr	23.0	44.5	74.98
ar-ru	5.0	21.0	54.73	en-nl	15.5	36.1	64.76
ar-zh	14.0	12.4	67.94	en-ru	14.2	30.3	62.82
de-ar	2.4	16.2	64.53	en-zh	20.7	17.9	74.97
de-fr	11.9	31.2	64.44	avg.	17.05	32.02	69.26
de-nl	9.4	28.1	54.22				
de-ru	5.1	19.6	55.41				
de-zh	4.2	5.8	55.26				
fr-ar	10.1	29.1	70.72		BLEU	chrF++	COMET
fr-de	8.6	27.7	53.77	ar-en	26.5	46.9	76.92
fr-nl	10.3	30.1	57.55	de-en	27.0	44.0	76.97
fr-ru	7.9	26.0	56.82	fr-en	27.5	49.0	78.80
fr-zh	18.1	18.5	72.24	nl-en	21.8	41.3	73.99
nl-ar	2.0	15.1	63.73	ru-en	24.8	43.6	74.23
nl-de	9.7	28.1	52.58	zh-en	23.2	45.3	76.83
nl-fr	13.2	32.3	65.17	avg.	25.13	45.02	76.29
nl-ru	5.1	18.6	55.13				
nl-zh	3.0	5.4	54.34				
ru-ar	5.9	15.0	60.36				
ru-de	5.6	23.8	52.66				
ru-fr	17.9	38.4	68.66				
ru-nl	6.2	22.5	54.41				
ru-zh	7.5	13.6	61.40				
zh-ar	6.7	22.1	67.48				
zh-de	3.3	19.6	51.75				
zh-fr	17.4	38.9	73.00				
zh-nl	4.8	19.3	54.41				
zh-ru	3.5	17.9	49.02				
avg.	8.38	22.75	59.93				
seen→seen	14.52	27.25	70.48	en→seen	17.60	29.87	73.81
seen→unseen	6.14	22.82	54.75	en→unseen	16.50	34.17	64.70
unseen→seen	7.56	19.22	61.99	seen→en	25.73	47.07	77.52
unseen→unseen	6.85	23.45	54.07	unseen→en	24.53	42.97	75.06

Table 15: Detailed results of BLOOMZ-7b1 fine-tuned with MTInstruct for BLOOMZ+3 de-nl-ru.

Zero-shot	BLEU	chrF++	COMET	Supervised	BLEU	chrF++	COMET
ar-de	5.1	20.8	55.25	en-ar	8.4	26.0	70.45
ar-fr	20.3	42.5	71.78	en-de	21.1	36.7	67.15
ar-nl	6.4	21.6	57.48	en-fr	22.9	44.4	74.67
ar-ru	5.2	21.5	55.51	en-nl	16.1	36.8	65.26
ar-zh	16.0	14.1	69.55	en-ru	15.2	31.5	63.30
de-ar	2.4	16.3	64.01	en-zh	16.1	15.0	71.93
de-fr	13.5	34.3	66.25	avg.	16.63	31.73	68.79
de-nl	9.7	28.0	55.00				
de-ru	5.3	19.6	55.61				
de-zh	7.2	7.3	60.64				
fr-ar	10.0	28.2	69.86		BLEU	chrF++	COMET
fr-de	9.2	27.8	54.03	ar-en	27.1	47.0	76.54
fr-nl	10.8	31.0	58.50	de-en	27.8	44.4	77.57
fr-ru	8.6	26.7	57.07	fr-en	27.1	48.7	78.82
fr-zh	15.9	15.8	70.78	nl-en	22.6	42.2	74.25
nl-ar	2.2	15.4	63.47	ru-en	25.6	44.2	74.46
nl-de	10.2	28.5	53.65	zh-en	23.5	45.7	77.04
nl-fr	14.4	34.4	66.55	avg.	25.62	45.37	76.45
nl-ru	5.3	19.3	55.53				
nl-zh	5.5	6.2	58.77				
ru-ar	6.5	16.0	62.69				
ru-de	6.1	24.3	52.89				
ru-fr	18.2	39.0	69.95				
ru-nl	6.3	22.5	54.36				
ru-zh	7.6	13.3	61.94				
zh-ar	8.7	26.5	70.88				
zh-de	3.0	19.5	50.82				
zh-fr	17.7	39.7	73.56				
zh-nl	4.4	19.3	54.20				
zh-ru	4.1	19.5	50.47				
avg.	8.86	23.30	60.70				
seen→seen	14.77	27.80	71.07	en→seen	15.80	28.47	72.35
seen→unseen	6.31	23.08	54.81	en→unseen	17.47	35.00	65.24
unseen→seen	8.61	20.24	63.81	seen→en	25.90	47.13	77.47
unseen→unseen	7.15	23.70	54.51	unseen→en	25.33	43.60	75.43

Table 16: Detailed results of BLOOMZ-7b1 fine-tuned with MT+Align for BLOOMZ+3 de-nl-ru.

Zero-shot	BLEU	chrF++	COMET	Supervised	BLEU	chrF++	COMET
ar-de	6.9	24.7	58.10	en-ar	14.6	35.6	76.70
ar-fr	26.2	48.2	74.96	en-de	20.4	36.0	65.96
ar-nl	8.8	24.7	59.53	en-fr	27.9	50.0	77.65
ar-ru	6.5	22.7	55.33	en-nl	14.8	34.8	63.11
ar-zh	28.6	22.3	77.64	en-ru	13.3	29.0	61.43
de-ar	3.3	19.8	68.27	en-zh	36.1	27.7	80.31
de-fr	15.2	35.8	67.05	avg.	21.18	35.52	70.86
de-nl	8.2	26.0	53.35				
de-ru	4.4	17.9	54.79				
de-zh	12.0	9.9	65.20				
fr-ar	14.2	35.2	74.84		BLEU	chrF++	COMET
fr-de	8.9	28.4	53.81	ar-en	33.7	53.5	79.81
fr-nl	10.1	29.9	56.92	de-en	27.1	43.9	77.04
fr-ru	8.1	26.0	55.96	fr-en	29.6	51.0	79.60
fr-zh	30.2	25.6	79.43	nl-en	22.0	41.4	73.54
nl-ar	3.1	18.2	67.72	ru-en	25.1	43.9	74.05
nl-de	10.4	27.7	52.67	zh-en	32.6	54.3	79.75
nl-fr	16.9	37.3	68.46	avg.	28.35	48.00	77.30
nl-ru	4.8	17.8	54.71				
nl-zh	8.1	7.0	63.96				
ru-ar	11.9	31.5	72.45				
ru-de	6.1	23.7	52.74				
ru-fr	21.2	42.5	71.71				
ru-nl	6.8	22.6	53.91				
ru-zh	21.3	20.7	74.63				
zh-ar	13.1	34.1	74.92				
zh-de	4.1	22.3	52.13				
zh-fr	23.8	46.5	76.54				
zh-nl	4.8	19.9	54.26				
zh-ru	5.7	21.9	50.60				
avg.	11.79	26.36	63.22				
seen→seen	22.68	35.32	76.39	en→seen	26.20	37.77	78.22
seen→unseen	7.10	24.50	55.18	en→unseen	16.17	33.27	63.50
unseen→seen	12.56	24.74	68.83	seen→en	31.97	52.93	79.72
unseen→unseen	6.78	22.62	53.69	unseen→en	24.73	43.07	74.88

Table 17: Detailed results of BLOOMZ-7b1 fine-tuned with MTInstruct for BLOOMZ+3 ar-de-fr-nl-ru-zh.

Zero-shot	BLEU	chrF++	COMET	Supervised	BLEU	chrF++	COMET
ar-de	6.7	24.2	57.45	en-ar	15.1	35.8	76.76
ar-fr	27.5	49.2	75.21	en-de	20.6	35.9	65.88
ar-nl	8.7	24.8	59.14	en-fr	27.5	49.4	77.46
ar-ru	6.7	21.6	55.04	en-nl	15.0	35.6	63.70
ar-zh	30.1	24.4	78.54	en-ru	13.5	29.5	61.62
de-ar	3.5	19.7	68.39	en-zh	36.3	27.7	80.52
de-fr	15.4	35.8	67.81	avg.	21.33	35.65	70.99
de-nl	9.6	27.3	53.74				
de-ru	4.7	17.9	54.23				
de-zh	12.0	9.9	65.40				
fr-ar	14.9	36.3	74.98		BLEU	chrF++	COMET
fr-de	9.2	28.3	52.96	ar-en	33.9	53.7	79.74
fr-nl	11.3	31.1	57.62	de-en	27.1	43.6	77.13
fr-ru	8.8	26.2	56.31	fr-en	29.7	51.0	80.03
fr-zh	31.1	26.9	79.93	nl-en	22.6	42.3	73.94
nl-ar	3.3	18.5	68.02	ru-en	25.8	44.5	74.07
nl-de	9.4	26.5	52.33	zh-en	32.5	54.5	80.01
nl-fr	17.2	37.3	68.38	avg.	28.60	48.27	77.49
nl-ru	4.4	17.1	53.63				
nl-zh	8.3	7.0	64.08				
ru-ar	12.4	32.1	72.40				
ru-de	5.7	22.9	51.90				
ru-fr	21.5	42.7	72.08				
ru-nl	6.3	22.1	53.32				
ru-zh	22.7	24.6	75.36				
zh-ar	13.9	35.4	75.68				
zh-de	3.6	21.3	51.32				
zh-fr	24.5	47.0	76.98				
zh-nl	4.9	20.3	54.30				
zh-ru	5.5	21.1	50.49				
avg.	12.13	26.65	63.23				
seen→seen	23.67	36.53	76.89	en→seen	26.30	37.63	78.25
seen→unseen	7.27	24.32	54.96	en→unseen	16.37	33.67	63.73
unseen→seen	12.92	25.29	69.10	seen→en	32.03	53.07	79.93
unseen→unseen	6.68	22.30	53.19	unseen→en	25.17	43.47	75.05

Table 18: Detailed results of BLOOMZ-7b1 fine-tuned with MT+Align for BLOOMZ+3 ar-de-fr-nl-ru-zh.