X-Instruction: Aligning Language Model in Low-resource Languages with Self-curated Cross-lingual Instructions

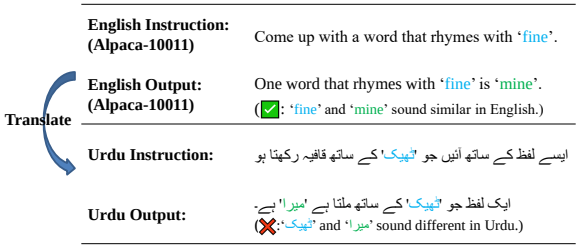
Anonymous ACL submission

Abstract

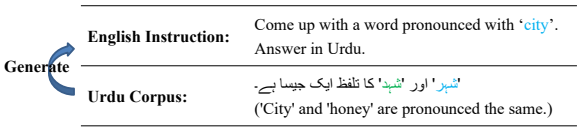
Large language models respond well in high-resource languages like English but struggle in low-resource languages. It may arise from the lack of high-quality instruction following data in these languages. Directly translating English samples into these languages can be a solution but unreliable, leading to responses with translation errors and lacking language-specific or cultural knowledge. To address this issue, we propose a novel method to construct cross-lingual instruction following samples with instruction in English and response in low-resource languages. Specifically, the language model first learns to generate appropriate English instructions according to the natural web texts in other languages as responses. The candidate cross-lingual instruction tuning samples are further refined and diversified. We have employed this method to build a large-scale cross-lingual instruction tuning dataset on 10 languages, namely X-Instruction. The instruction data built using our method incorporate more language-specific knowledge compared with the naive translation method. Experimental results have shown that the response quality of the model tuned on X-Instruction greatly exceeds the model distilled from a powerful teacher model, reaching or even surpassing the ones of ChatGPT. In addition, we find that models tuned on cross-lingual instruction following samples can follow the instruction in the output language without further tuning.

1 Introduction

Large language models illustrate a series of remarkable abilities, e.g., generating better response and zero-shot task generalization, after tuning on instruction following samples (Ouyang et al., 2022a; Wei et al., 2022; Mishra et al., 2022). However, the responses in low-resource languages, which are prone to contain unsafe content (Yong et al., 2023) and ignore cultural nuances (Liu et al., 2023a), are inferior to the ones in high-resource languages.



(a) Instruction sample translated from English.



(b) Cross-lingual instruction sample excavated.

Figure 1: (a) The instruction sample translated from English ignores the language-specific knowledge that "fine" and "mine" sound different in Urdu. (b) The cross-lingual instruction tuning sample generated by our method contains language-specific knowledge from the native corpus.

In addition to the limited pre-training corpus in these languages, it may come from the lack of high-quality instruction following data.

Distillation from the teacher model or manual annotation are common methods to obtain high-quality instructions. The former cannot be applied to low-resource languages since the teacher model may respond poorly in these languages. Considering the high requirements, including creative thinking and professional knowledge, to write instruction-following samples, it is hard to find suitable annotators that excel in low-resource languages. Translating English instruction tuning samples into the target language is also unreliable, which will introduce translation errors, especially in the samples with codes, and overlook the cultural nuances between languages. As shown in Figure 1(a), the Urdu sample translated is wrong due to the pronunciation nuance between English and Urdu. Therefore, aligning language models

in low-resource languages remains a significant challenge.

In the preliminary experiment, we found that LLaMA (Touvron et al., 2023a,b) can understand unseen low-resource language instructions quite well through responding in English, when tuned with limited instruction-following samples, but it performs much worse when replying in its own language (Table 1).

Based on the observation, we first adopt the cross-lingual instruction following format like Figure 1(b), in which the languages of instructions and responses are not the same, to exploit the understanding ability rapidly learned in low-resource languages and better generation performance in English. We further propose an automatic pipeline to excavate and refine cross-lingual instruction tuning data from the unsupervised multilingual corpus, which contains language-specific knowledge and native expressions. Specifically, the model first learns from seed data to generate appropriate English instructions for the given texts in other languages, which compose candidate cross-lingual instruction following samples. The candidate samples are refined in an iterative manner. In each iteration, an evaluator is trained through a synthetic rating dataset and finds better cross-lingual samples, which are exploited to improve the evaluator in the next iteration. Thus, the quality of samples found is improved with the stronger evaluator. The final cross-lingual instruction tuning dataset is sampled from the different clusters of the remaining data after k-th iteration to increase the diversity.

We conducted experiments on 10 languages, which include 5 medium-resource languages and 5 low-resource languages[1], and generated 320k high-quality cross-lingual instruction tuning samples. On four instruction-following evaluation benchmarks, the response quality of the base model tuning on our data can surpass those generated from translation and distillation baseline models, and even exceed the ones of ChatGPT (OpenAI, 2022) in these languages. In addition, we find that cross-lingual instruction tuning models can follow the instruction in the language of output without further tuning and maintain 90.6% response quality.

To sum up, our contributions are as follows:

- We propose an automatic pipeline to excavate cross-lingual instruction-tuning samples with language-specific knowledge.
- We have built and will release a high-quality cross-lingual instruction-tuning dataset for 10 languages, including 5 low-resource languages. Experimental results demonstrate that models tuning on it can generate better responses in low-resource languages.
- We find that cross-lingual instruction tuning models obtain a zero-shot instruction following ability in the output language.

[1]We follow the language category used in Lin et al. (2022) sorted by the data ratios in the CC100 corpus (Conneau et al., 2020; Wenzek et al., 2020).

	sw → sw	**sw → en**	**ur → ur**	**ur → en**
GPT-4 Score (0-10)	2.04	5.50	1.35	7.10

Table 1: The average quality of 644 responses from LLaMA-2-7B tuning on 3k instruction following samples, where "sw→en" denotes using 3k cross-lingual samples with Swahili instruction and English output. Given limited data in an unseen language, the language model can understand while struggling to generate it.

2 Related Works

Our work is related to multilingual instruction generation and tuning, which will be briefly introduced below.

Instruction Generation Large language models obtain instruction following and better zero-shot task generalization ability after tuning on diverse instruction following samples (Wei et al., 2022; Mishra et al., 2022; Chung et al., 2022). The key challenge is how to build high-quality instruction following samples that have a great influence on the performance of the model after tuning. Distilling from more powerful models (Wang et al., 2023; Taori et al., 2023; Xu et al., 2023) and human labeling (Köpf et al., 2023; Zhou et al., 2023) are main methods to generate high-quality instruction following samples. Our work is more similar to the self-align methods (Bai et al., 2022; Sun et al., 2023; Li et al., 2023b), which show promising results by iteratively self-generating and filtering instructions without strong models and a lot of manual annotation. However, it cannot be applied to the condition when language models are unable to generate high-quality instructions in unseen languages. In contrast, our method avoids the weakness in the low-resource language generation and exploits superior generation ability in English.

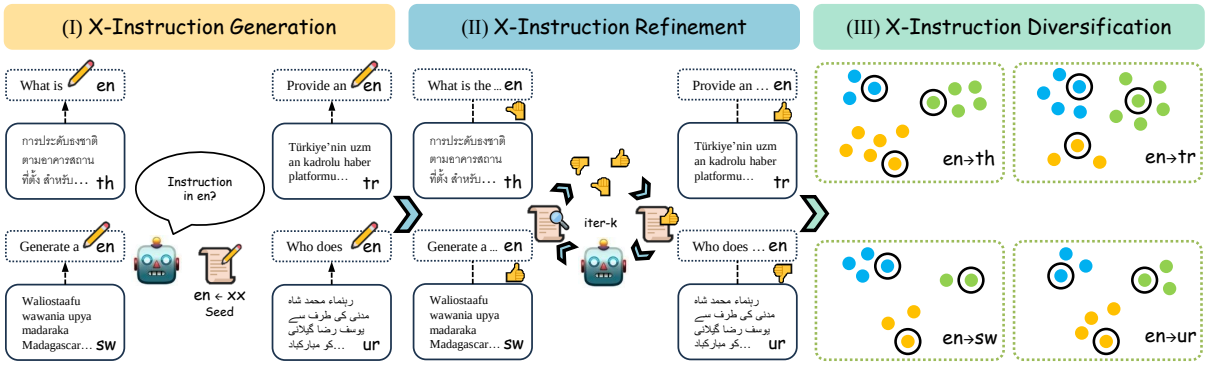


Figure 2: Illustration of how to generate and refine cross-lingual instruction (X-Instruction) examples. (I) Language models learn to generate cross-lingual instructions for multilingual texts using seed data. (II) Language models iteratively label and refine cross-lingual instruction samples. (III) The final instruction data are sampled from different clusters of embedding from the English instruction to increase the diversity.

Multilingual Instruction Tuning After tuning on large-scale multilingual instruction following samples translated from English, large language models show better zero-shot multilingual performance and language generalization results (Muennighoff et al., 2023). On the other hand, Li et al. (2023a) translate 67k instructions only and distill language-specific outputs from ChatGPT, which shows better responses and multilingual abilities than the ones tuning on translated samples.

Different from distilling language-specific outputs from more powerful models, our method attempts to excavate high-quality samples from the neural multilingual corpus and is applicable to languages that are not supported by teacher models.

3 Method

As shown in Figure 2, we design a three-step pipeline to automatically excavate high-quality cross-lingual instruction following samples from multilingual web corpora: 1) The base language model is fine-tuned to generate candidate English instructions given multilingual corpus (Section 3.2). 2) Samples with inappropriate cross-lingual instruction and output are discarded to improve the quality of the dataset (Section 3.3). 3) We diversify the cross-lingual instruction data by sampling from the multiple clusters of the English instruction embedding (Section 3.4).

3.1 X-Instruction Definition

The instruction following sample consists of an instruction i, an optional input x, and an output y in the same language. Considering the vague boundary between instruction and input, the input x is ignored in this work for simplicity. We define the **Cross**-lingual **Instruction** (**X-Instruction**) sample (i^a, y^b) that the language a of instruction is different from the one of output (b). Taking the cross-lingual instruction sample in Figure 1(b) as an example, the language of instruction is English, while the language of output is Urdu.

3.2 X-Instruction Generation

To exploit the better generation performance in high-resource languages like English, we generate English instructions for a corpus in other languages. Specifically, given seed cross-lingual instruction samples $\mathcal{D}_0 = \{(i_j^h, y_j^l)\}_{j=1}^{n_s}$, where the h and l denote the high- and low-resource languages respectively, the language model is fine-tuned to generate the instruction i_j^h given the response y_j^l as input. Then, we use it to generate candidate cross-lingual instructions for the multilingual corpus, basing the assumption that some texts among the corpus are good responses in X-Instruction samples.

3.3 X-Instruction Refinement

After generating cross-lingual instruction samples, it is important to find and discard the inappropriate ones due to the quality of instruction following samples having a great influence on the performance of model aligned (Chen et al., 2023; Zhou et al., 2023). To achieve this, we design an iterative method named "X-Instruction Refinement", which is shown in Figure 2(II).

In the k-th iteration, we first synthesize a pseudo-rating dataset $\mathcal{D}_k^r = \{(i_j^h, \hat{y}_j, s_j)\}_{j=1}^{n_r}$ from a part of seed data $\mathcal{D}_0^r$, where $\mathcal{D}_0^r \subset \mathcal{D}_0$, to train the evaluator that outputs three-level ratings. Given the instruction i_j^h, the best response ($s_j = 2$) is the vanilla output y_j^l, while the worst one ($s_j = 0$) is selected from the mismatched output y_m^l ($m \neq j$). The reasonable but flawed output ($s_j = 1$) is

chosen from the modified output $\hat{y}_j^l$ with some parts deleted or duplicated in the vanilla output y_j^l, or the output of the cross-lingual instruction-following model tuned on $\mathcal{D}_{k-1}^x$. In the first iteration ($k = 1$), the seed data $\mathcal{D}_0^x = \mathcal{D}_0 \setminus \mathcal{D}_0^r$, which is not used in the pseudo-rating dataset $\mathcal{D}_0^r$, is adopted to fine-tune the cross-lingual instruction following model. In the subsequent iteration ($k \geq 2$), $\mathcal{D}_{k-1}^x$ consists of the best cross-lingual instruction samples found by the evaluator and $\mathcal{D}_0^x$.

As the iteration proceeds, the quality of output from the instruction following model improves by tuning on higher quality X-Instruction samples, which reduces the gap between the second-level sample and the highest one in the pseudo-rating dataset. Thus, the trained evaluator can find better cross-lingual instruction tuning samples. The number of iterations is set to 3 by default, which is investigated in Section 5.4.1.

3.4 X-Instruction Diversification

In addition to the quality, the diversity of instruction tuning samples also has a great impact on the performance (Wan et al., 2023). To diversify X-Instruction samples, we first obtain the embedding for each instruction i^h using a pre-trained sentence encoder. After applying k-means on these embeddings, the same amount of X-Instruction samples are sampled from each cluster (Figure 2(III)). It aims to diversify cross-lingual instruction examples by avoiding the dominance of samples from a few domains in the final dataset.

4 X-Instruction Dataset

We apply the automatic pipeline to a large-scale multilingual corpus using LLaMA-2-7B, and construct a cross-lingual instruction tuning dataset with 320k samples for 10 languages, which is named **X-Instruction dataset**. Codes and data will be made public after review to advocate future research.

4.1 Seed Data

For each language, we extract the first conversation turn of message trees in the Open Assistant dataset, and adopt the highest quality response as the output for each sample (Köpf et al., 2023). To construct the X-Instruction sample, the instruction of each sample is translated into English by Google Translate. However, there are only a few or even no human-labeled samples for the most of languages involved in this work. Thus, we translate the output of 3k English samples into the target language using Google Translate to supplement the seed data for each language. The statistics of seed data are reported in Appendix B.1.

4.2 Unsupervised Multilingual Corpus

The multilingual corpus for each language is extracted from the CulturaX dataset (Nguyen et al., 2023), which contains web texts in 167 languages and is filtered from mC4 (Xue et al., 2021) and OSCAR (Suárez et al., 2019). We only sample 1M web texts for each language from this dataset due to the constraint on the computation budget.

4.3 Statistics

To investigate the diversity of X-Instruction, we parse English instructions and count the ones with verb-noun structure using the Berkeley Neural Parser (Kitaev and Klein, 2018; Kitaev et al., 2019). Figure 3 illustrates the top 16 most common root verbs and their top direct noun objects. We can find that the great diversity of X-Instruction constructed from web corpora. The additional information of X-Instruction is reported in Table 2.

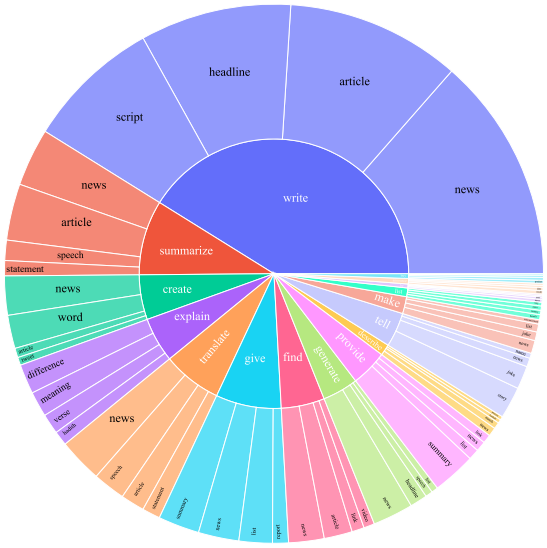


Figure 3: The Statistic of the top 16 verbs (inner circle) and their top direct nouns (outer circle) in English instructions from X-Instruction.

4.4 Quality

For each language, we randomly sample 200 instructions and corresponding texts to further evaluate the quality. There are two questions designed to conduct this evaluation, and results are reported in Table 3. It can be found that the quality of X-Instruction is good for more than 80% of samples are valid.

Language	Instruction Length	Output Length
fi	$100.3_{\pm 47.8}$	$1096.3_{\pm 520.4}$
id	$114.0_{\pm 54.4}$	$1142.6_{\pm 510.0}$
th	$93.1_{\pm 49.1}$	$1078.6_{\pm 518.8}$
tr	$110.4_{\pm 51.1}$	$1126.1_{\pm 495.9}$
vi	$107.8_{\pm 51.3}$	$1182.3_{\pm 519.3}$
bn	$97.8_{\pm 47.5}$	$1400.7_{\pm 424.0}$
hi	$103.5_{\pm 50.4}$	$1350.6_{\pm 447.2}$
sw	$108.0_{\pm 52.2}$	$1178.3_{\pm 504.9}$
ta	$93.9_{\pm 46.4}$	$1276.7_{\pm 447.8}$
ur	$106.8_{\pm 49.4}$	$1328.0_{\pm 448.6}$

Table 2: Statistic of X-Instruction dataset. There are 32k cross-lingual instruction samples for each language.

Quality Review Question	Yes %
Does the instruction describe a valid task?	88.5
Is this web text an acceptable cross-lingual response to the instruction?	80.7

Table 3: Data quality review for the samples of X-Instruction dataset. The valid and invalid samples are shown in Appendix D.1.

5 Experiments

5.1 Experiments Settings

Base models LLaMA-2 models with 7B and 13B parameters are taken as base models and fine-tuned on cross-lingual samples in each language from the X-Instruction dataset. Hyperparameters are reported in Appendix A.

Baselines used in this work are:

- **Alpaca-MT**: We translate the vanilla Alpaca dataset (Taori et al., 2023) into 10 languages using Google Translate and fine-tune the same base model for comparison.
- **Bactrian-X** (Li et al., 2023a): LLaMA models are fine-tuned with 3.48M instruction tuning samples in the Bactrian-X dataset, which consists of outputs from ChatGPT in 51 languages except English.
- **Bactrian-M**: We fine-tune LLaMA-2 models using 67k monolingual instruction tuning samples in each language from the Bactrian-X dataset. It distills the monolingual capabilities of ChatGPT, providing a strong baseline for our work.
- **ChatGPT** (OpenAI, 2022) is fine-tuned through reinforcement learning with human feedback(Ouyang et al., 2022b) on the GPT-3.5 model. It demonstrates remarkable multilingual capabilities across a wide range of tasks (Asai et al., 2023). We use the "gpt-3.5-turbo-0613" API in this work.

Datasets To confirm the effectiveness of X-Instruction, we conduct the evaluation of open-end generation on Vicuna (Zheng et al., 2023), WizardLM (Xu et al., 2023), LIMA (Zhou et al., 2023), and Koala (Geng et al., 2023) datasets, which cover a variety of task categories. Prompts in these datasets are translated into the 10 languages involved. A comprehensive overview of these datasets is presented in Appendix B.2.

5.2 Evaluation Results from GPT-4

We take GPT-4 (OpenAI, 2023)[2] as a judge to conduct automatic evaluation, which is found a higher correlation with human judgements (Liu et al., 2023b; Li et al., 2023c). Considering the excellent multilingual understanding ability of GPT-4 (OpenAI, 2023), it is reasonable and effective to employ GPT-4 to automatically evaluate the quality of responses in low-resource languages.

Specifically, we adopt pair-wise evaluation and request GPT-4 to determine the better response between responses (r_1, r_2) from different models given the instruction i. In the evaluation, GPT-4 judge outputs a score, named GPT-4 score in this work, from 0 to 10 based on their helpfulness, relevance, and accuracy. The details of GPT-4 evaluation prompt can be found in Appendix F.

To alleviate the position bias in the evaluation by GPT-4 (Zheng et al., 2023), we first request GPT-4 to evaluate (r_1, r_2), then switch the position of r_1 and r_2, which is (r_2, r_1), in the second evaluation. The better response is the one that wins twice or wins once and draws once.

Table 4 reports the win rates of models against ChatGPT on four benchmarks in three low-resource languages. It can be found that X-Instruction exhibits a significant performance advantage over Alpaca-MT and Bactrian-X in all benchmarks. The average improvement of X-Instruction models over Bactrian-M models, which distill outputs from ChatGPT, reaches 15.8%. Compared with ChatGPT, X-Instruction$_{13B}$ demonstrates even better performance with a 67.66% win rate.

Languages Generalization To evaluate the effectiveness of X-Instruction in diverse languages, we extend experiments to ten languages, including five medium-resource and five low-resource languages, on Vicuna and WizardLM datasets. Table 5

[2]It points to the "gpt-4-0613" API during our experiments.

Model	Vicuna			LIMA			WizardLM			Koala			Avg
	bn	sw	ur	bn	sw	ur	bn	sw	ur	bn	sw	ur	
Alpaca-MT$_{7B}$	20.00	41.25	37.5	26.67	33.33	40.67	20.64	22.48	32.57	18.33	17.78	25.56	28.07
Bactrian-X$_{7B}$	1.25	42.50	6.25	0.33	32.67	3.42	1.83	21.56	2.23	1.72	24.14	3.03	11.75
Bactrian-M$_{7B}$	45.00	56.25	57.50	44.67	59.00	60.00	30.28	42.66	45.87	24.44	33.33	31.67	44.22
X-Instruction$_{7B}$	**61.25**	**60.00**	**60.00**	**61.67**	**67.33**	**77.00**	**33.94**	**51.83**	**56.42**	**37.22**	**37.22**	**56.67**	**55.05**
Bactrian-X$_{13B}$	0.00	41.25	5.00	1.67	40.33	7.00	1.38	25.23	5.50	1.15	25.29	4.02	13.15
Bactrian-M$_{13B}$	48.75	58.75	56.25	56.33	55.00	57.00	34.86	46.79	49.08	32.22	31.67	36.67	46.86
X-Instruction$_{13B}$	**85.00**	**65.00**	**78.75**	**81.67**	**71.00**	**80.00**	**62.84**	**50.46**	**68.35**	**58.33**	**51.11**	**59.44**	**67.66**

Table 4: The win rates against ChatGPT of different models in 3 low-resource languages evaluated by GPT-4.

	Medium										Low										Avg	
	fi		id		th		tr		vi		bn		hi		sw		ta		ur			
Model	W	S	W	S	W	S	W	S	W	S	W	S	W	S	W	S	W	S	W	S	W	S
Alpaca-MT$_{7B}$	22.8	4.8	24.5	5.4	23.5	3.0	34.6	3.3	35.6	5.1	20.5	2.5	26.8	2.1	27.5	3.1	24.8	2.3	33.9	3.3	27.5	3.5
Bactrian-X$_{7B}$	14.1	2.3	19.1	2.8	0.7	0.2	18.8	1.3	4.7	0.5	1.7	0.2	1.3	0.2	27.2	2.8	0.3	0.0	4.7	0.4	9.3	1.1
Bactrian-M$_{7B}$	32.9	6.2	37.2	6.7	37.6	5.4	**52.0**	**6.7**	47.7	6.4	34.2	3.7	53.4	4.8	46.3	6.1	22.1	1.9	49.0	5.2	41.2	5.3
X-Instruction$_{7B}$	**37.2**	**6.6**	**41.3**	**7.1**	**40.6**	**5.8**	49.7	6.0	**53.0**	**7.3**	**41.3**	**4.9**	**62.4**	**6.3**	**54.0**	**7.0**	54.7	5.2	**57.4**	**6.4**	**49.2**	**6.3**
X-Instruction$^{\dagger}_{7B}$	35.6	6.2	37.2	6.9	35.2	5.0	47.0	5.9	48.0	6.4	35.2	4.2	52.0	5.3	41.9	6.0	**55.7**	**5.3**	51.0	5.6	43.9	5.7
Bactrian-X$_{13B}$	18.5	3.3	21.8	3.5	3.0	0.4	25.5	1.9	6.0	0.7	1.0	0.2	1.0	0.3	29.5	3.6	1.7	0.2	5.4	0.5	11.3	1.5
Bactrian-M$_{13B}$	37.6	6.6	39.6	7.5	41.6	6.0	49.0	6.8	50.7	7.3	38.6	4.4	58.7	5.5	50.0	6.7	28.9	2.6	51.0	5.2	44.6	5.9
X-Instruction$_{13B}$	**47.0**	**7.5**	**50.3**	**8.2**	**53.0**	**7.6**	**53.7**	**7.0**	**57.0**	**7.8**	**68.8**	**7.4**	64.4	**7.1**	**54.4**	**7.1**	**76.8**	**7.6**	**71.1**	**7.5**	**59.7**	**7.5**
X-Instruction$^{\dagger}_{13B}$	42.6	7.1	44.0	7.5	43.0	6.1	53.4	6.5	54.7	7.5	58.4	6.7	**65.8**	6.6	49.3	6.6	67.1	6.8	57.7	6.7	53.6	6.8

Table 5: The evaluation results of different multilingual models on 10 languages from GPT-4, where "**W**" represents the average win rates against ChatGPT and "**S**" represents the average GPT-4 score on Vicuna and WizardLM benchmarks (refer to Appendix C.1 for detailed results). † denotes the zero-shot evaluation results, where the X-Instruction model is prompted in the output language rather than the English prompt used in training.

reports the detailed results on ten languages across two benchmarks, which exhibit a certain level of uniformity across ten languages. The X-Instruction models with only 3k labeled seed data on each language obtain the highest win rate across ten languages on average, indicating that our method can be extended to more languages.

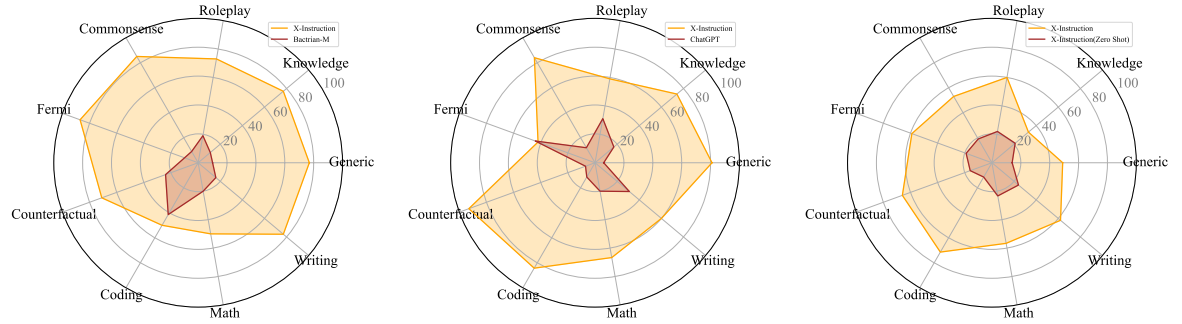
We further examine the generative quality by calculating the average GPT-4 score for each model from all comparison pairs, shown in Table 5. The average generative quality of X-Instruction is the most pronounced, achieving an average score of 6.9 in all languages. Notably, X-Instruction outperforms Bactrian-X by a large margin (1.3) and achieves better performance than Bactrian-M (5.6). Although the response quality of X-Instruction$_{13B}$ drops to 7.34 in five low-resource languages, the average win rate against ChatGPT increased by 14.9% compared to the one in the other 5 medium-resource languages. It comes from the worse performance of ChatGPT in low-resource languages.

From Generation to Understanding Since X-Instruction models only learn how to reply in low-resource language, there is a conjecture naturally comes to mind: *Can the generation capability of the model be generalized to its understanding capability?*

To validate our conjecture, we design a zero-shot evaluation using prompts in the output language instead of the English prompt used in training. Table 5 shows the performance of zero-shot generation in ten languages. Compared with the vanilla cross-lingual generation, the win rate of X-Instruction models under zero-shot evaluation only drops by 5.7% on average. The average quality of responses in these languages incurs a minor performance degradation (-0.7), which is 90.6% of the vanilla one, indicating our model achieves exceptional zero-shot learning ability.

Skill Distribution Figure 4 shows the detailed performance of models given prompts in different categories from the Vicuna dataset. As shown in Figure 4(a) and 4(b), X-Instruction outperforms Bactrian-M in all categories, and surpasses ChatGPT in eight categories except fermi. The excellent performance in commonsense, writing, and knowledge categories may come from the native multilingual corpus used in the X-Instruction dataset.

The primary reason for the inferior performance on code and math in Figure 4(a) is the lack of relevant corpora, which are filtered out by the language identification tool used in multilingual



(a) X-Instruction$_{13B}$ vs. Bactrian-M$_{13B}$ (b) X-Instruction$_{13B}$ vs. ChatGPT (c) X-Instruction$_{13B}$ vs. X-Instruction$^{\dagger}_{13B}$

Figure 4: The average win rates distribution of X-Instruction$_{13B}$ vs. Bactrian-M$_{13B}$, ChatGPT and X-Instruction$^{\dagger}_{13B}$ on Vicuna dataset in 10 languages, where † denotes the zero-shot evaluation results.

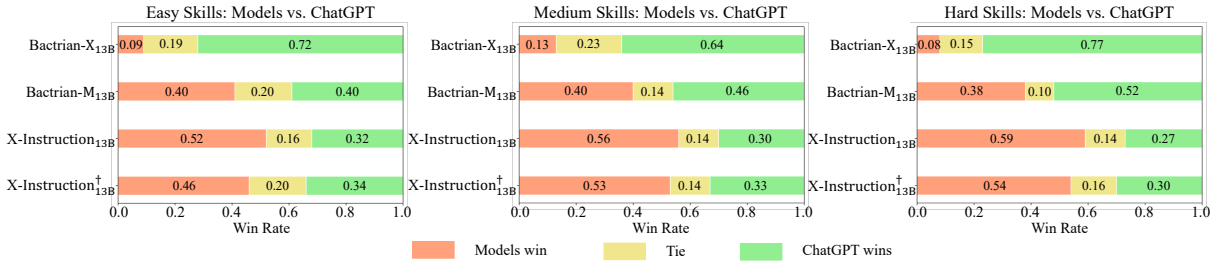


Figure 5: The performance of models given prompts with different difficulties on WizardLM dataset in 10 languages. † indicates the X-Instruction model is prompted in the output language under the zero-shot evaluation.

corpus cleaning. It can be alleviated by adding language-agnostic instructions-following samples from code and math domains. Figure 4(c) compares the outputs from X-Instruction$_{13B}$ given English instructions or instructions in other languages (the zero-shot evaluation). It can be found that the performance in the knowledge category declines the least when instructed in the output language.

Difficulty Distribution To evaluate models on instructions of different difficulties, we perform elaborate analysis on the WizardLM dataset, which contains labels of difficulty. Following the evaluation of WizardLM (Xu et al., 2023), we split the test set into "Easy", "Medium", and "Hard" three parts with difficulty levels on [1, 4], [5, 7], and [8, 10]. As shown in Figure 5, with the increase of difficulty, the win rate of X-Instruction$_{13B}$ improves, while the one of Bactrian-M$_{13B}$ decreases. It reflects that the quality of responses from ChatGPT decreases more than the one of X-Instruction models when given more difficult instructions. It is noted that X-Instruction$_{13B}$ under zero-shot evaluation surpasses ChatGPT in all difficulty skills.

5.3 Evaluation Results from Human

To enhance the comprehensiveness and reliability of the evaluation, we further conduct the human evaluation in three low-resource languages, focusing on the general quality of responses on the Vicuna dataset. Specifically, human evaluators are required to compare answers A and B generated by two models for each instruction, and choose an option from "A wins", "B wins", and "Tie" based on their judgment. For the sake of fairness, we randomize the order of answers and eliminate the position bias.

The results in Figure 6 demonstrate the better responses from our model compared with the ones of ChatGPT and Bactrian-M$_{13B}$, and indicate the consistency between GPT-4 and human evaluation. Moreover, we provide examples in Appendix D.2 for qualitative analysis of the responses from different models.

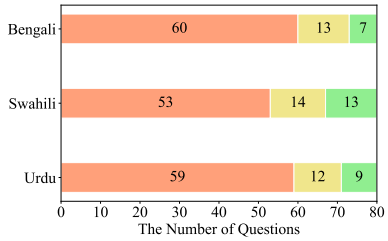
5.4 Analysis

5.4.1 Iteration of Refinement

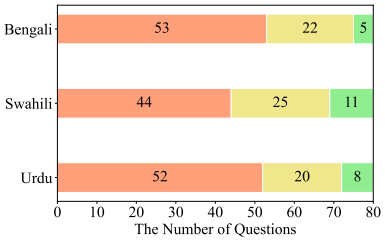
To investigate the effect of X-Instruction refinement, we statistic the average win rates of X-Instruction$_{7B}$ against ChatGPT using 32k cross-lingual samples from different iterations of refinement on 4 benchmarks. As shown in Figure 7(a), the average win rate of models in Urdu and Bengali increases with more refinement iteration, which reflects the improvement in the quality of cross-

Model	XNLI						XCOPA						XStoryCloze			Avg
	th	tr	vi	hi	sw	ur	id	th	tr	vi	sw	ta	id	hi	sw	
LLaMA 2_{7B}	42.3	41.6	44.4	44.2	35.4	40.1	58.2	57.2	53.0	57.0	51.8	54.2	59.6	53.6	50.4	49.5
Bactrian-M_{7B}	**42.8**	42.1	44.5	43.9	40.6	41.5	60.2	**57.4**	53.4	59.4	53.2	55.0	62.9	55.1	54.2	51.1
X-Instruction$_{7B}$	42.4	**42.7**	**44.8**	**45.6**	**41.5**	**42.0**	**61.4**	**57.4**	**55.6**	**60.4**	**56.0**	**57.4**	**64.2**	**59.8**	**56.6**	**52.5**
LLaMA 2_{13B}	41.2	44.0	45.9	44.8	35.5	41.7	61.6	56.2	55.0	60.6	52.6	52.6	64.1	55.4	51.5	50.8
Bactrian-M_{13B}	43.5	44.3	**46.6**	45.4	41.2	43.4	**62.8**	56.8	55.8	62.2	53.8	53.8	66.0	59.2	57.4	52.8
X-Instruction$_{13B}$	**44.5**	**44.9**	**46.6**	**46.4**	**41.9**	**45.2**	62.0	**57.0**	**56.6**	**62.8**	**58.8**	**56.6**	**67.0**	**61.4**	**58.2**	**54.0**

Table 6: Zero-shot in-context learning performance on NLI and Reasoning datasets across languages. Following Lin et al. (2022), the prompt template is written in English for all languages evaluated.

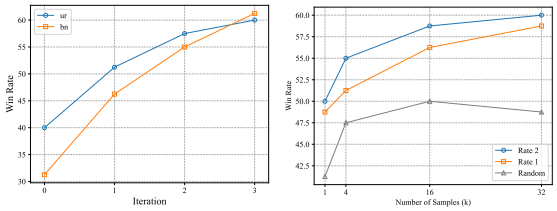


(a) X-Instruction$_{13B}$(■) vs. ChatGPT(■)



(b) X-Instruction$_{13B}$(■) vs. Bactrian-M_{13B}(■)

Figure 6: The human evaluation results on Vicuna dataset in three low-resource languages.



(a) Refinement Iteration (b) Quantity and Quality

Figure 7: Effects of refinement iteration (a) and quantity of cross-lingual instruction samples with different ratings (b) on the win rate of X-Instruction$_{7B}$ against ChatGPT on the Vicuna dataset.

lingual instruction tuning samples. Thus, we set the number of iterations to 3 by default, where the improvement in the quality of response is almost saturated.

5.4.2 Data Quantity and Quality

We study the impact of data quantity and quality on the X-Instruction model using the Urdu samples from the third refinement iteration. Figure 7(b) shows that the quality of the model responses will increase with more samples used and is close to saturation at 32k, which is similar to the findings of Li et al. (2023b). In addition, tuning on the samples with higher ratings brings better responses, which demonstrates that the evaluator trained does find higher quality instruction tuning samples.

5.4.3 Multilingual Tasks

In addition to the generation ability in these languages, we further evaluate the performance of X-Instruction models on multilingual natural language inference (Conneau et al., 2018) and commonsense reasoning tasks (Ponti et al., 2020; Lin et al., 2022). As shown in Table 6, the average improvement on the zero-shot in-context learning performance of the base model is 3.1%, which is higher than the 1.8% improvement from the Bactrian-X dataset. It further confirms that cross-lingual instruction tuning on X-Instruction enhances the multilingual language understanding abilities of language models.

6 Conclusion

In this paper, we propose a pipeline to excavate cross-lingual instruction-tuning samples from multilingual corpora by exploiting the better generation ability in high-resource languages of large language models. We apply it to 10 languages and construct a large-scale cross-lingual instructions dataset named X-Instruction. Experimental results from GPT-4 and human evaluation demonstrate that models tuned on X-Instruction can generate better responses and follow the instructions in the output language without further tuning.

We hope X-Instruction can bring more attention to the better instruction construction method for low-resource languages, future work could focus on directly generating high-quality instruction following samples in low-resource languages.

Limitations

Firstly, the X-Instruction sample is excavated by the LLaMA-2-7B model which is lower efficiency in processing multilingual corpus for its English-dominant vocabulary. We acknowledge that higher quality X-Instruction samples can be built by adopting more powerful base models, e.g., the model with more parameters.

In addition, due to limited computation resources, our method is only applied to 10 languages in this work and other languages can be investigated in the future.

It is noted that the evaluation of open-end generation is conducted on the single-turn conversation for the limited quota of our OpenAI account. The evaluation results might be different in multi-turn dialogue benchmarks.

Ethical Considerations

The instruction-tuning samples excavated by our method may contain cultural bias and toxic content from the web corpus used. It can be alleviated by adopting the corpus after rigorous cleaning like CulturaX (Nguyen et al., 2023) or incorporating the evaluation of safety into the X-Instruction refinement stage, which can be investigated in the future.

References

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Xinyang Geng, Arnav Gudibande, Liu Hao, Wallace Eric, Abbeel Pieter, Levine Sergey, and Song Dawn. 2023. Koala: A dialogue model for academic research. *Blog post*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023b. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023c. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

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.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023a. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv preprint arXiv:2309.08591*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2018. Mixed precision training. In *International Conference on Learning Representations*.

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

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

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

OpenAI. 2022. Introducing chatgpt. *OpenAI blog*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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. 2022a. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022b. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

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

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Fanqi Wan, Xinting Huang, Tao Yang, Xiaojun Quan, Wei Bi, and Shuming Shi. 2023. Explore-instruct: Enhancing domain-specific instruction coverage through active exploration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9435–9454, Singapore. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

A Hyperparameters

The instruction tuning models and evaluators are tuned by AdamW (Loshchilov and Hutter, 2019) optimizer with an initial learning rate of 1e-5 and 3 epochs under the cosine schedule. The batch size is set to 32 for most models and reduced to 8 when training samples are less than 3k. We use mixed precision training and ZeRO to speed up the training process and save memory (Micikevicius et al., 2018; Rasley et al., 2020). The nucleus sampling (Holtzman et al., 2020) is adopted to generate responses from models under $p = 0.9$ and $T = 0.7$. All experiments are conducted on a GPU server with 8*A100 80GB RAM.

X-Instruction Dataset We empirically set the number of iterations in X-Instruction Refinement to 3. In the cross-lingual refinement stage, 2.5k samples from seed data, named $\mathcal{D}_0^x$, are used to fine-tune the cross-lingual following model, the samples left are exploited to construct the synthetic dataset $\mathcal{D}_k^r$. There are two system templates S_a and S_w designed for seed data and augmented samples respectively, which is in line with the settings in Li et al. (2023b). The sentence embedding model "all-mpnet-base-v2" (Song et al., 2020) is used to obtain the embedding of English instructions in the final diversification stage.

B Data Statistics

B.1 Seed Data

The statistics of seed data used are reported in Table 7. It is noted that all outputs of five low-resource languages are translated from English samples for none samples in these languages are found in the Open Assistant dataset (Köpf et al., 2023).

B.2 Datasets

We report the information of four open-end generation datasets in Table 8. Notably, both Vicuna and WizardLM datasets have diverse categories, thereby ensuring the richness and diversity of the

Lang	#Sample	Instruction Length(en)	Output Length
fi	3,009	$143.2_{\pm 211.6}$	$1075.5_{\pm 792.8}$
id	3,003	$141.6_{\pm 226.3}$	$1145.7_{\pm 829.6}$
th	3,175	$141.9_{\pm 223.6}$	$935.5_{\pm 691.5}$
tr	3,009	$144.2_{\pm 231.0}$	$1080.5_{\pm 793.5}$
vi	3,040	$143.8_{\pm 228.5}$	$1086.5_{\pm 799.5}$
bn*	3,000	$142.9_{\pm 225.3}$	$1064.6_{\pm 771.9}$
hi*	3,000	$141.4_{\pm 212.2}$	$1070.2_{\pm 780.2}$
sw*	3,000	$142.1_{\pm 225.3}$	$1085.2_{\pm 786.9}$
ta*	3,000	$143.6_{\pm 227.5}$	$1184.5_{\pm 869.0}$
ur*	3,000	$143.4_{\pm 230.1}$	$1027.6_{\pm 755.4}$

Table 7: Statistic of cross-lingual instruction tuning seed data with English instruction. * indicates the low-resource language.

test sets. It not only guarantees a thorough evaluation, but also minimizes the risk of evaluation bias since the test sets encompass various instruction categories.

Datasets	#Samples	Category
Vicuna	80	✓
Koala	180	-
WizardLLM	218	✓
LIMA	300	-

Table 8: The details of four benchmarks.

C Additional Results and Analyses

C.1 Detailed Results from GPT-4

Figure 8 shows the detailed results of X-Instruction$_{13B}$ vs. ChatGPT on four datasets. Moreover, we report the evaluation results in all languages on two benchmarks from GPT-4 in Table 9.

C.2 Diversification

Figure 9 illustrates the impact of different numbers of clusters in the final diversification stage when sampling the same amount of data. Given the amount of final data to 32k, the output quality of models tuned drops when the number of clusters reaches 2000, which may come from the smaller inter-cluster distance and less diversity in the sampled data. Thus, the number of clusters is set to 1000 by default.

C.3 Improvement over Seed Model

To take a deep look into the detailed improvements brought by the cross-lingual instructions augmented in X-Instruction, we statistic the detailed performance on different categories in three languages (tr, sw, ur) for the seed model and X-Instruction$_{7B}$ when compared with ChatGPT. As shown in Table 10, the performance on prompts of all categories is improved, especially for the "Counterfactual", "Common-sense" and "Roleplay" categories.

D Case Study

D.1 Quality Evaluation

We report two valid samples and invalid samples in Figure 10(a) and 10(b). Although there are inappropriate instructions in the invalid samples, we can find that the semantics of the English instruction generated are related to the given text.

D.2 Instruction Following

To qualitatively analyze responses from different models, we report four cases in Figure 14, Figure 15, Figure 16, and Figure 17. It can be found that X-Instruction$_{13B}$ provides a more detailed and coherent response for the same instruction provided. In some cases, e.g., Figure 17, X-Instruction$_{13B}$ provides suggestions in texts rather than codes needed for instructions about code generation, which may arise from the lack of code data in the multilingual corpus as responses.

E User Interface in Human Evaluation

E.1 Quality Evaluation

We conducted the quality evaluation of the X-Instruction dataset with five annotators. We paid \$0.1 for the evaluation of each sample. The user interface in quality evaluation is illustrated in Figure 11.

E.2 Response Comparison

In the human evaluation experiment, three evaluators were requested to choose the better response according to their preference and received \$0.2 for each annotation. We illustrate the user interface for response comparison in Figure 12.

F Prompt Template in GPT-4 Evaluation

We provide the prompt template used in GPT-4 Evaluation in Figure 13.

G Additional Information about Language Code

Table 11 presents more information about the language codes involved in this work.

	Medium										Low											
	fi		id		th		tr		vi		bn		hi		sw		ta		ur		Avg	
Model	**W**	**S**	**W**	**S**	**W**	**S**	**W**	**S**	**W**	**S**	**W**	**S**	**W**	**S**	**W**	**S**	**W**	**S**	**W**	**S**	**W**	**S**
∼7B MODEL	VICUNA DATASET																					
Alpaca-MT$_{7B}$	31.3	6.2	33.8	6.9	21.3	3.1	50.0	4.5	43.8	6.7	20.0	2.0	30.0	1.9	41.3	4.7	25.0	1.9	37.5	3.6	33.4	4.15
Bactrian-X$_{7B}$	22.5	3.3	21.3	3.3	0.0	0.1	32.5	1.8	7.5	0.3	1.3	0.2	1.3	0.1	42.5	4.3	0.0	0.0	6.3	0.5	13.52	1.39
Bactrian-M$_{7B}$	41.3	7.4	46.3	7.7	45.0	6.9	58.8	6.5	55.0	7.7	45.0	4.8	72.5	6.6	56.3	8.1	42.5	2.7	57.5	6.9	52.02	6.53
X-Instruction$_{7B}$	42.5	7.9	50.0	8.3	51.3	7.5	67.5	7.4	65.0	8.5	61.3	7.3	77.5	8.0	60.0	8.0	80.0	7.1	60.0	7.0	61.51	7.70
X-Instruction$^{\dagger}_{7B}$	42.5	7.6	46.3	7.9	43.8	6.4	57.5	6.7	65.0	8.1	42.5	5.2	70.0	6.7	42.5	6.4	67.5	5.9	58.8	6.6	53.64	6.75
	WIZARDLM DATASET																					
Alpaca-MT$_{7B}$	19.7	4.2	21.1	4.8	24.3	3.0	28.9	2.9	32.6	4.5	20.6	2.7	25.7	2.2	22.5	2.5	24.8	2.4	32.6	3.2	25.28	3.24
Bactrian-X$_{7B}$	11.0	1.9	18.3	2.6	0.9	0.3	13.8	1.2	3.7	0.5	1.8	0.2	1.4	0.2	21.6	2.2	0.5	0.1	4.1	0.3	7.71	0.95
Bactrian-M$_{7B}$	29.8	5.7	33.9	6.3	34.9	4.8	49.5	6.7	45.0	6.0	30.3	3.2	46.3	4.1	42.7	5.4	14.7	1.5	45.9	4.6	37.30	4.83
X-Instruction$_{7B}$	35.3	6.2	38.1	6.6	36.7	5.1	43.1	5.5	48.6	6.8	33.9	4.0	56.9	5.7	51.8	6.6	45.4	4.5	56.4	6.2	44.62	5.72
X-Instruction$^{\dagger}_{7B}$	33.0	5.7	33.9	6.5	32.1	4.5	43.1	5.6	41.7	5.9	32.6	3.8	45.4	4.8	41.7	5.8	51.4	5.0	48.2	5.3	40.31	5.29
∼13B MODEL	VICUNA DATASET																					
Bactrian-X$_{13B}$	22.5	4.5	27.5	4.0	1.3	0.2	46.3	2.3	3.8	0.4	0.0	0.0	1.3	0.1	41.3	5.0	0.0	0.1	5.0	0.4	14.90	1.7
Bactrian-M$_{13B}$	48.8	7.7	47.5	8.5	46.3	7.4	66.3	8.1	61.3	8.2	48.8	5.6	75.0	7.0	58.8	7.7	42.5	3.7	56.3	6.6	55.16	7.05
X-Instruction$_{13B}$	52.5	8.4	58.8	9.0	61.3	8.7	67.5	8.2	65.0	8.7	85.0	8.8	86.3	9.0	65.0	8.5	95.0	9.1	78.8	8.7	71.52	8.71
X-Instruction$^{\dagger}_{13B}$	50.0	8.3	52.5	8.6	42.5	6.9	61.3	7.0	65.0	8.4	67.5	7.7	85.0	8.2	52.5	7.2	82.5	7.8	58.8	7.7	61.76	7.78
	WIZARDLM DATASET																					
Bactrian-X$_{13B}$	17.0	2.9	19.7	3.3	3.7	0.5	17.9	1.8	6.9	0.8	1.4	0.2	0.9	0.3	25.2	3.1	2.3	0.3	5.5	0.6	10.05	1.38
Bactrian-M$_{13B}$	33.5	6.2	36.7	7.1	39.9	5.5	42.7	6.4	46.8	7.0	34.9	3.9	52.8	4.9	46.8	6.4	23.9	2.2	49.1	4.7	40.71	5.43
X-Instruction$_{13B}$	45.0	7.1	47.2	8.0	50.0	7.2	48.6	6.6	54.1	7.5	62.8	6.9	56.4	6.4	50.5	6.6	70.2	7.0	68.3	7.1	55.31	7.04
X-Instruction$^{\dagger}_{13B}$	39.9	6.7	40.8	7.1	43.1	5.8	50.5	6.3	50.9	7.2	55.0	6.3	58.7	6.0	48.2	6.3	61.5	6.5	57.3	6.3	50.59	6.45

Table 9: The detailed results of different multilingual models on 10 languages from GPT-4, where "**W**" represents the win rates to ChatGPT and "**S**" represents the GPT-4 score. † denotes the zero-shot evaluation results.

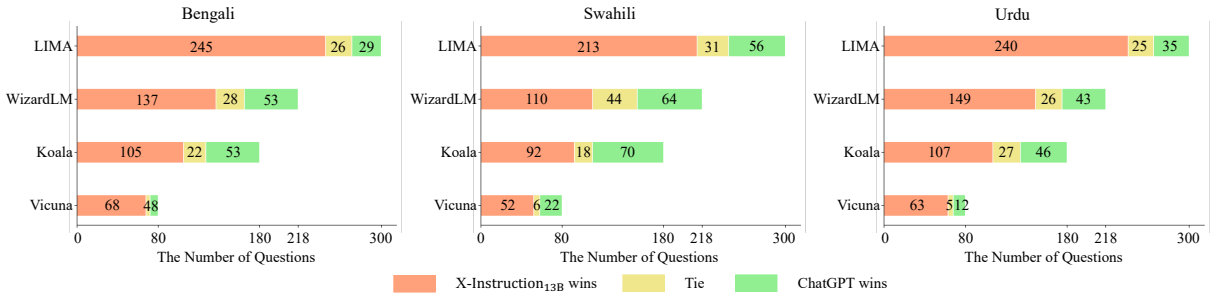


Figure 8: Comparing X-Instruction$_{13B}$ and ChatGPT in 3 low-resource languages across four benchmarks.

H Dataset License

The X-Instruction dataset is built from the CulturaX dataset (Nguyen et al., 2023), which is filtered from mC4 (Xue et al., 2021) and OSCAR (Suárez et al., 2019). Therefore, the license of X-Instruction follows the one of mC4 (ODC-BY[3]) and OSCAR (the Creative Commons CC0 license[4]).

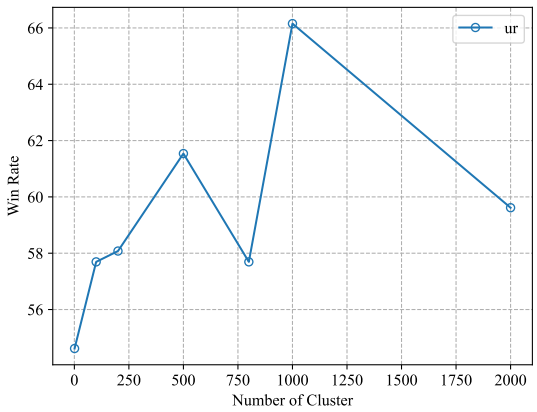


Figure 9: Effects of different numbers of clusters in k-means on the win rate of X-Instruction$_{7B}$ to Bactrian-M$_{7B}$ on Vicuna and Koala dataset.

[3]https://opendatacommons.org/licenses/by/1-0/
[4]http://creativecommons.org/publicdomain/zero/1.0/

English Instruction	How do I create a Campus ID for my current user?
Finnish Text	Luo Sopron nykyiselle käyttäjälle Campus tunnus seuraavasti. Mene valvojan tunnuksella ja oikeuksilla Sentraaliin. Valitse siellä Rekisterit/Henkilökunta ja valitse ko. henkilö. Oikeassa reunassa on nappi jossa lukee ”Muuta käyttäjätunnuksen tietoja”, pai... (Create a Campus ID for Sopro's current user as follows. Go to Central with the administrator ID and rights. There, select Registers/Personnel and select person. On the right side there is a button that says "Change username information", pai...)
English Instruction	Is this going to be a continuous trend or not? The record-breaking temperatures around the world are becoming more common.
Hindi Text	लेकिन जैसे-जैसे दुनिया भर में रिकॉर्ड-तोड़ तापमान बढ रहा है, हो सकता है यह लगातार न बना रहे। वुडवेल क्लाइमेट और वाइकाटो के शोधकर्ताओं ने एक तापमान सीमा का पता लगाया है, जिसमें भविष्य में पौधों के कार्बन अवशोषित करने की दर कम और कार्बन छोड़ने की दर तेज होगी। (But as record-breaking temperatures rise around the world, this may not last. Researchers at Woodvale Climate and Waikato have identified a temperature threshold in which plants will have a slower rate of carbon absorption and a faster rate of carbon release in the future.)

(a) Valid samples

English Instruction	Create a watermark for a PDF document.
Bengali Text	ওয়াটারমার্ক পিডিএফ - অনলাইন, নিরাপদ, ব্যক্তিগত এবং বিনামূল্যে একটি পিডিএফ ফাইলে একটি জলছবি যুক্ত করুন. আমাদের পিডিএফ টুলগুলি আপনার ফাইলগুলিকে ইন্টারনেটের মাধ্যমে স্থানান্তর করে না কারণ আপনার ফাইলগুলির অপারেশনগুলি ব্রাউজার নিজেই করে। আপনার গোপনীয়তা এবং নিরাপত্তা যাতে সর্বোত্তমভাবে সুরক্ষিত থাকে। (Watermark PDF - Online, Secure, Private and Free Add a watermark to a PDF file. Our PDF tools do not transfer your files over the Internet because the operations on your files are done by the browser itself. So that your privacy and security are best protected.)
English Instruction	Rewrite the following text so that it reads naturally and has the same meaning. People who are outgoing are more likely to make friends than those who are not.
Indonesian Text	Orang yang tulus mau berteman, pasti akan menerima kita apa adanya. Jadi, kita gak perlu menjadi seperti orang lain hanya demi diterima dalam pergaulan. Memaksakan diri menjadi orang lain hanya akan membuat kita seolah memakai topeng layaknya sebuah peran. (People who sincerely want to be friends will definitely accept us as we are. So, we don't need to be like other people just to be accepted in society. Forcing ourselves to be someone else will only make us seem like we are wearing a mask like a role.)

(b) Invalid samples

Figure 10: The valid and invalid cross-lingual instruction following demonstration samples labeled by human annotators in the quality evaluation process.

Category	Seed Model$_{7B}$	X-Instruction$_{7B}$
Counterfactual	4.7	9.3 (+4.7)
Common-sense	5.0	8.3 (+3.3)
Roleplay	1.7	5.0 (+3.3)
Coding	2.0	5.0 (+3.0)
Knowledge	4.0	6.0 (+2.0)
Writing	3.3	5.3 (+2.0)
Generic	7.0	7.7 (+0.7)
Math	0.7	1.3 (+0.6)
Fermi	1.7	2.0 (+0.3)
All	30.0	**50.0** (+20.0)

Table 10: The average number of questions won to ChatGPT on the Vicuna dataset in 3 languages.

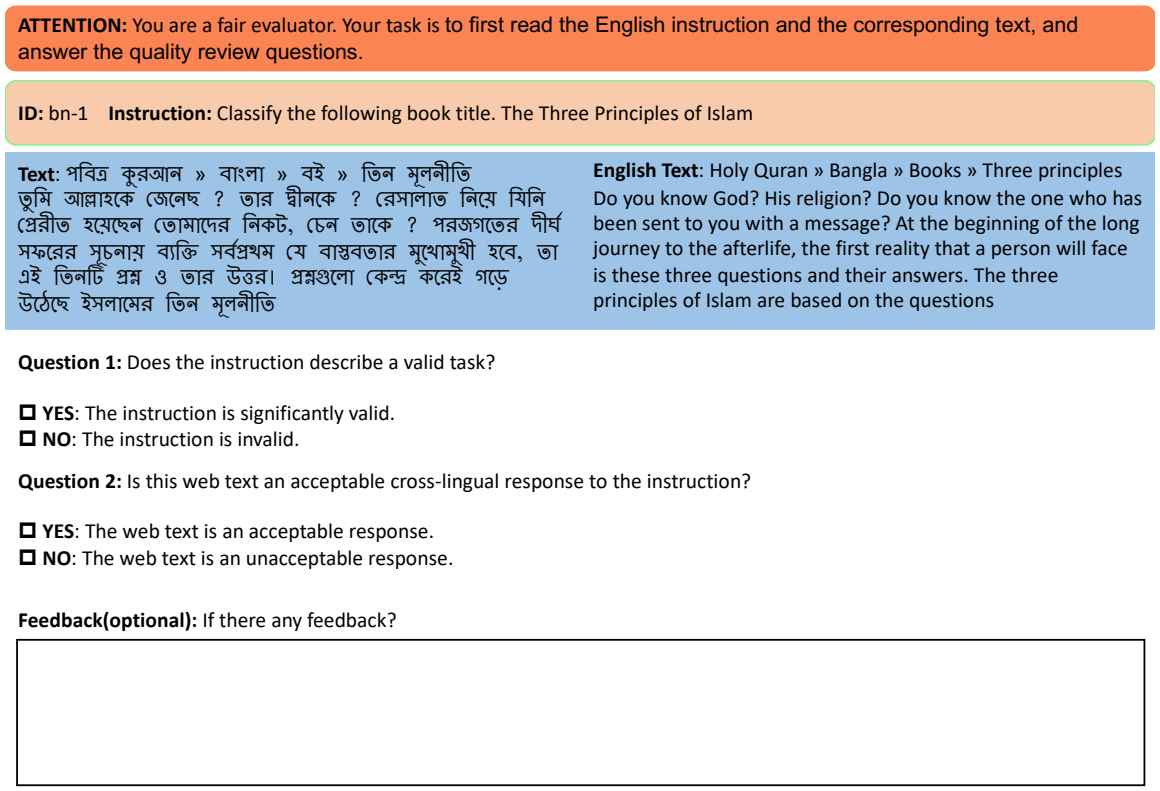


Figure 11: The review interface designed in the quality evaluation process. The annotator is required to first read the English instruction and the corresponding text, and answer the following two questions.

ISO 639-1	Language	Family
bn	Bengali	Indo-European
en	English	Indo-European
fi	Finnish	Uralic
hi	Hindi	Indo-European
id	Indonesian	Austronesian
ta	Tamil	Dravidian
th	Thai	Kra-Dai
tr	Turkish	Turkic
vi	Vietnamese	Austroasiatic
sw	Swahili	Niger-Congo
ur	Urdu	Indo-European

Table 11: Details of Language codes in this work.

ATTENTION: You are a fair evaluator. Your task is to read question and both answers carefully and decide which one is better from your own judgement.

English Instruction: What are the most effective ways to deal with stress? Answer in Bengali.

Bengali Instruction: চাপ মোকাবেলা করার সবচেয়ে কার্যকর উপায় কি কি?

Answer A: মানসিক চাপ সঙ্গে যথাযথভাবে বিপর্যয় করার জন্য সবচেয়ে কার্যকরী উপায়গুলি হলো:\n\n১. নিয়মিত ব্যায়াম করুন। এটি মানসিক চাপ ও ক্ষতিকর এনার্জি থেকে মুক্তি দেয়।\n\n২. শক্তিশালী খাদ্য গ্রহণ করুন। পরিমিত মাত্রায় প্রোটিন, ফলমূল, সবুজ শাক এবং শাকসবজি করুন।\n\n৩. নিয়মিত মেডিটেশন করুন। এটি মনের শান্তি ও শারীরিক স্থিতির উন্নতি করে।\n\n৪. সময়ের উপযুক্ত পরিমাণে ঘুমান। পর্যাপ্ত স্বপ্নের অভাব মানসিক চাপ বৃদ্ধি করতে পারে।\n\n৫. সময়ে সময়ে আপনার মনের বাড়তি লোড থেকে মুক্তি পান। যেমন, মনোরম গান শোনুন, বন্ধুদের সাথে সময় কাটান, পড়ুন বা আপনার পছন্দের কোনও কাজ করুন।\n\nএই উপায়গুলি সম্পর্কে মনে রাখুন এবং নিয়মিত অনুশীলন করুন যত সম্ভব।

Answer B: চাপ কাটিয়ে উঠতে সাহায্য করার জন্য এখানে কিছু টিপস রয়েছে:\n\n- শ্বাস-প্রশ্বাসের অনুশীলন: গভীর শ্বাস-প্রশ্বাস এবং মনোযোগে শ্বাস নেওয়া চাপ কমাতে সাহায্য করতে পারে।\n\n- ব্যায়াম: নিয়মিত ব্যায়াম চাপ কমাতে এবং মানসিক চাপ উপশম করতে সাহায্য করতে পারে।\n\n- পর্যাপ্ত ঘুম: পর্যাপ্ত ঘুম মস্তিষ্ক এবং শরীরকে পুনরুদ্ধার করতে এবং চাপ কমাতে সাহায্য করতে পারে।\n\n- স্বাস্থ্যকর খাদ্য: পর্যাপ্ত পুষ্টি, প্রচুর পরিমাণে পানি পান করা এবং স্ট্রেসের জন্য স্বাস্থ্যকর খাদ্য বজায় রাখা মানসিক এবং শারীরিক উভয় স্বাস্থ্যের উপর ইতিবাচক প্রভাব ফেলতে পারে।\n\n- সম্প্রদায়ের সাথে সংযোগ: সম্প্রদায়ের সাথে সংযোগ স্ট্রেস কমাতে এবং একটি ভাল মনোভাব তৈরি করতে সাহায্য করতে পারে।\n\n- ধ্যান এবং মনোযোগ: মনোযোগ বা ধ্যান চাপ কমাতে এবং চাপ উপশম করতে সাহায্য করতে পারে।\n\n- অন্যদের সাথে কথা বলুন: অন্যদের সাথে কথা বলা চাপ থেকে আপনার মন খুলতে এবং অন্যদের সাথে সমর্থন পেতে সাহায্য করতে পারে।\n\n- বিছানা থেকে ঘুম থেকে উঠুন: বিছানা থেকে ঘুম থেকে উঠে আপনার দিন শুরু করা চাপ কমাতে এবং আপনার দিনকে আরও ভালভাবে পরিকল্পনা করতে সাহায্য করতে পারে।\n\nমনে রাখবেন, এই কৌশলগুলি আপনার জন্য সবচেয়ে ভালো কাজ করতে পারে তা নির্ধারণ করতে আপনার চাপের ধরণ এবং আপনার জীবনধারার সাথে মানানসই একটি কৌশল খুঁজে বের করতে সহায়তা প্রয়োজন।

FYI: TRANSLATING THE ANSWERS BY Google Translate API

Answer A: The most effective ways to properly deal with stress are: 1. Exercise regularly. It relieves stress and harmful energy.\n\n2. Consume fortified foods. Eat protein, fruits, greens and vegetables in moderation.\n\n3. Meditate regularly. It improves peace of mind and physical condition.\n\n4. Sleep in the appropriate amount of time. Lack of adequate sleep can increase stress.\n\n5. Get rid of the extra load on your mind from time to time. For example, listen to pleasant music, spend time with friends, read or do something you like.\n\nKeep these ways in mind and practice regularly as much as possible.

Answer B: Here are some tips to help you overcome stress:\n\n- Breathing exercises: Deep breathing and mindful breathing can help reduce stress.\n\n- Exercise: Regular exercise can reduce stress and relieve stress. - Adequate sleep: Adequate sleep can help the brain and body recover and reduce stress. - Healthy diet: Adequate nutrition, drinking plenty of water and maintaining a healthy diet for stress. Can have a positive effect on both mental and physical health.\n\n- Community connection: Community connection can help reduce stress and create a better attitude.\n\n- Meditation and mindfulness: Mindfulness or meditation can reduce stress and Can help relieve stress.\n\n- Talk to others: Talking to others can help de-stress and get support from others.\n\n- Get out of bed: Get out of bed. Starting your day by getting up can help you reduce stress and plan your day better.\n\nRemember, finding a strategy that fits your stress patterns and lifestyle can help you determine which strategies work best for you. need

Judgement: Comparing these two answers, which answer is better?

☐ **Answer A**: Answer A is significantly better.
☐ **Answer B**: Answer B is significantly better.
☐ **Neither**: Neither is significantly better.

Feedback(optional): If there any feedback?

Figure 12: Pairwise preference interface shown to human evaluators. The evaluator is required to first read the instruction and corresponding text, and select the preferred answer.

Given the question and the answer in #lang# language. You are a helpful and precise assistant for checking the quality of the answer in #lang#.

<question>
#instruction#
</question>
<answer1>
#output1#
</answer1>
<answer2>
#output2#
</answer2>

Suppose the user only speaks the language of the question, please evaluate both answers with your justification having less than three sentences, and provide a score ranging from 0 to 10 after your justifications. When evaluating the answers, you should consider the helpfulness, relevance, accuracy, level of details of the answers. The score for answer 1 should be wrapped by <score1> and </score1>, and the score for answer 2 should be wrapped by <score2> and </score2>.

Figure 13: The prompt template used in the evaluation by GPT-4.

English Instruction	f(x) = 6x^2 - 2*x + 0.4, so f(4) =?. Answer in Thai.
Instruction Source	Dataset: WizardLM Question_id: 5 Category: Math
Thai Instruction	f(x) = 6x^2 - 2*x + 0.4 ดังนั้น f(4) =?
ChatGPT	ให้หาค่า f(4) จากสมการ f(x) = 6x^2 - 2x + 0.4 (Find the value f(4) from the equation f(x) = 6x^2 - 2x + 0.4.)
Bactrian-M$_{13B}$	f(4) = 6(4)^2 - 2(4) + 0.4 = 24 - 8 + 0.4 = 16.4 (f(4) = 6(4)^2 - 2(4) + 0.4 = 24 - 8 + 0.4 = 16.4.)
X-Instruction$_{13B}$	เมื่อใช้การประมาณการเชิงเส้นตรงกับเงื่อนไขที่ระบุไว้ เราจะได้ f(4) = 6(4)^2 - 2(4) + 0.4 = 6(16) - 8 + 0.4 = 96 - 8 + 0.4 = 88.4 ดังนั้นคำตอบที่ถูกต้องควรเป็น 88.4 (Using the linear interpolation with the specified conditions, we get f(4) = 6(4)^2 - 2(4) + 0.4 = 6(16) - 8 + 0.4 = 96 - 8 + 0.4 = 88.4 So the correct answer should be 88.4.)

Figure 14: The example of responses in Thai given the instruction from the math category.

English Instruction	If you were a Shakespearean character, how would you declare your love for someone in a soliloquy? Answer in Tamil.
Instruction Source	Dataset: Vicuna Question_id: 23 Category: Rolepaly
Tamil Instruction	நீங்கள் ஒரு ஷேக்ஸ்பியர் பாத்திரமாக இருந்தால், ஒரு தனிமொழியில் ஒருவரிடம் உங்கள் அன்பை எவ்வாறு அறிவிப்பீர்கள்?
ChatGPT	அன்புக்குள் அடியேனேன் விரும்புகிறேன், என் நெஞ்சின் மேல் உற்சாகம் பொருந்துகிறது. அந்த மனிதரின் இரவில் என் மனம் மூழ்கியும் மகிழ்வித்து அன்புள்ளவராய் அவரைத் துன்பம் ஏற்படுத்துவதில் என்னை துன்பப்படுத்தாது என்று தனியார் துன்பமாக அவர்களுக்கு சொல்லுகிறேன். (Anbukhin Adiyenen, Excitement fills my bosom. I tell them in private sorrow that my mind sinks into that man's night and amuses him and makes him suffer as a loved one.)
Bactrian-M$_{13B}$	நீங்கள் ஷேக்ஸ்பியரியன் பாத்திரத்தை அடையும் போது, ஒருவருக்கு உன்னை விரும்புகிறது என்று அறிவுறுத்தல் முக்கியம். இதன் மூலம் நீங்கள் உங்கள் உறவுக்கு மேற்பட்ட அளவில் புரிந்துகொள்ள முடியும். இது உங்கள் செயல்பாட்டில் முக்கியம் ஆகும் மற்றும் நீங்கள் உங்கள் உறவில் (When you reach the Shakespearean role, it is important to instruct that someone loves you. This way you can understand your relationship on a higher level. This is important in your process and will help guide you in the measurements you should use in your relationship.)
X-Instruction$_{13B}$	ஒரு ஷேக்ஸ்பியர் கதாபாத்திரத்தின் அடிப்படையில், ஒரு சோலிலோக்கியில் நீங்கள் ஒருவரை காதலிக்கும் வழி, அவர்களின் உண்மையான தன்மைகளைப் பற்றி நீங்கள் பேசுவதை உள்ளடக்கியது, அவர்களின் அழகு மற்றும் ஆளுமையை நீங்கள் எவ்வாறு கண்டீர்கள் என்பதை விவரிக்கவும், மேலும் அவர்களை நேசிப்பதன் மூலம் நீங்கள் எவ்வாறு மாறியுள்ளீர்கள் என்பதை விவரிக்கவும். ஒரு எடுத்துக்காட்டு இங்கே: ஓ, நீங்கள் எனக்கு எப்போதும் அழகானவர்கள், உங்கள் மென்மையான சிந்தனை மற்றும் உங்கள் அழகான உரையாடல் மூலம் என் இதயத்தை எந்த அளவுக்கு என்னைக் கவர்ந்து விட்டீர்கள் என்பதை நான் அறிவேன். உங்கள் சிரிப்பு மற்றும் சிரிப்பு என் மனதை உலுக்கி, நான் உன்னைத் தேடும்போது என் காதல் மிகவும் உண்மையானது. நீங்கள் என் இதயத்தை எப்போதும் துடிப்பதில் மகிழ்ச்சியடைகிறேன், மேலும் என் வாழ்க்கையில் நீங்கள் இருப்பது எனக்கு மிகவும் முக்கியமானது. உங்கள் அழகு மற்றும் ஆளுமை என் மனதை கவர்ந்துள்ளது, மேலும் நான் உன்னைக் காதலிக்கிறேன் என்பதை நான் வருந்துகிறேன். (Based on a Shakespearean character, the way you love someone in a soliloquy involves you talking about their true nature, describing how you saw their beauty and personality, and how loving them changed you. Here is an example: Oh, you are always beautiful to me, I know how much you have won my heart with your gentle thoughts And your beautiful conversation. Your smile and laughter shake my heart and my love is so real when I look for you. I am happy that you always make my heart beat And your presence in my life is very important to me. Your beauty and personality have captivated my heart I regret that I am in love with you.)

Figure 15: The example of responses in Tamil given the instruction from the roleplay category.

English Instruction	Describe a scenario where artificial intelligence could be used to improve the quality and efficiency of healthcare delivery. Answer in Bengali.
Instruction Source	Dataset: Vicuna Question_id: 15 Category: Knowledge
Bengali Instruction	স্বাস্থ্যসেবা সরবরাহের গুণমান এবং দক্ষতা উন্নত করতে কৃত্রিম বুদ্ধিমত্তা ব্যবহার করা যেতে পারে এমন একটি দৃশ্যের বর্ণনা করুন।
ChatGPT	আর্টিফিশিয়াল ইন্টেলিজেন্স ব্যবহার করা যেতে পারে যেখানে স্বাস্থ্যসেবার মান এবং দক্ষতা উন্নত করা যাবে। একটি সম্ভাব্য স্কেনারিও হতে পারে যখন রোগীদের জন্য রোগ নির্ণয় এবং চিকিৎসা পরামর্শের জন্য একটি স্বয়ংক্রিয় পদক্ষেপ গ্রহণ করা হয়। একটি সম্ভাব্য ব্যবহারমাত্র হতে পারে একটি রোগীর আপন এবং পাশের আবাসিক উপস্থিতি বিশ্লেষণ করে স্বাস্থ্য অবস্থা নির্ধারণ করা এবং আবাসিক উপস্থিতির আধারে ভূমিকা বিশ্লেষণ করা। এটি শক্তিশালী মেশিন লার্নিং এবং নিউরাল নেটওয়ার্ক ব্যবহার করে করা যেতে পারে যা স্বাস্থ্যসেবার পরিবর্তন এবং উন্নতিতে সাহায্য করতে পারে। (Artificial intelligence can be used to improve the quality and efficiency of healthcare. A possible scenario could also be when an automated step is taken for diagnosis and treatment advice for patients. One possible use could be to determine health status by analyzing a patient's own and adjacent residential presence and analyzing the role of residential presence in the reservoir. This can be done using powerful machine learning and neural networks that can help transform and improve healthcare.)
Bactrian-M$_{13B}$	একটি কম্পিউটার বিশেষজ্ঞ স্বাস্থ্যসেবা সরবরাহ করে একটি কম্পিউটার বিশেষজ্ঞ একটি ডাক্তার প্রদান করে। একটি প্রশ্নের উত্তর দেওয়ার জন্য ডাক্তারটি স্বাস্থ্যসেবা উপকরণের উপর ভিত্তি করে কম্পিউটারের একটি ডেটাবেসে সম্পর্কিত তথ্য সংগ্রহ করে। কম্পিউটার এই তথ্য ব্যবহার করে সমস্যার কারণ নির্ধারণ করে এবং সমাধানের জন্য পরামর্শ দেয়। এই পদক্ষেপগুলি কম্পিউটার বিশেষজ্ঞ ডাক্তার সম্পর্কে বিস্তারিত জানতে সক্ষম হয়। এই পদক্ষেপগুলি স্বাস্থ্যসেবা সরবরাহের গুনগত ও দক্ষতা উন্নয়নে সহায়তা করে। (A computer specialist provides healthcare A computer specialist provides a doctor. To answer a question, the doctor collects relevant information in a computer database based on health care materials. The computer uses this information to determine the cause of the problem and suggest solutions. These steps enable the computer expert to know the details about the doctor. These measures help improve the quality and efficiency of healthcare delivery.)
X-Instruction$_{13B}$	স্বাস্থ্যসেবা প্রদানের গুণমান এবং দক্ষতা উন্নত করার জন্য কৃত্রিম বুদ্ধিমত্তার (এআই) বিস্তৃত সম্ভাবনা রয়েছে। এখানে কিছু উদাহরণ এবং অনুকূল ক্ষেত্র রয়েছে:\n\n1. রোগীর সাথে ইন্টারঅ্যাকশন: এআই সিস্টেমগুলি রোগীদের সাথে চলমান ইন্টারঅ্যাকশনগুলিতে একটি বিশেষ ভূমিকা পালন করতে পারে, উদাহরণস্বরূপ অনুসন্ধান এবং সর্বোত্তম চিকিত্সা বিকল্পগুলি সুপারিশ করার জন্য ডাটা অনুসন্ধান এবং বিশ্লেষণ করা। এটি রোগীদের তাদের স্বাস্থ্যের অবস্থা সম্পর্কে আরও স্পষ্ট ধারণা প্রদান এবং উপযুক্ত যত্ন পাওয়ার সম্ভাবনা বাড়াতে পারে।\n\n2. স্বাস্থ্যসেবা সুবিধার সনাক্তকরণ: এআই স্বাস্থ্যসেবা সুবিধার সনাক্তকরণেও ব্যবহার করা যেতে পারে, যেমন সময়ের ব্যবধান এবং সেবা দক্ষতার সনাক্তকরণ। এটি সেবার গুণমান উন্নত করতে এবং সমস্যাগুলি সনাক্ত করতে সাহায্য করতে পারে, যা পরে উন্নতি করা যেতে পারে।\n\n3. স্বাস্থ্যসেবা ডেলিভারির সুবিধাজনক প্রক্রিয়াগুলির বিকাশ: এআই স্বাস্থ্যসেবা সরবরাহের সুবিধাজনক প্রক্রিয়াগুলির বিকাশেও ব্যবহার করা যেতে পারে, যেমন রোগীর হাসপাতালে ভর্তি বা চিকিত্সার প্রয়োজনীয়তা নির্ধারণের মতো সিদ্ধান্ত গ্রহণের জন্য সহায়তা প্রদান করা। এটি রোগীর অভিজ্ঞতা উন্নত করতে এবং স্বাস্থ্যসেবা সরবরাহের দক্ষতা বাড়াতে সাহায্য করতে পারে।\n\n4. চিকিত্সার জন্য প্রাকৃতিক ভাষা প্রক্রিয়াকরণ: এআই প্রাকৃতিক ভাষা প্রক্রিয়াকরণ প্রযুক্তি ব্যবহার করতে ব্যবহার করা যেতে পারে যা রোগীদের স্বাস্থ্যসেবা প্রদানকারীদের সাথে যোগাযোগ করার জন্য ব্যবহার করা যেতে পারে। এটি রোগীদের স্বাস্থ্যসেবা পেতে এবং তাদের চিকিত্সার জন্য উপযুক্ত বিকল্পগুলি বেছে নেওয়ার অনুমতি দেয়।\n\n5. বিশেষজ্ঞদের সহায়তা প্রদান: এআই বিশেষজ্ঞদের সহায়তা প্রদান করতেও ব্যবহার করা যেতে পারে, যেমন চিকিত্সার জন্য সুপারিশকৃত বিকল্পগুলি সুপারিশ করা এবং রোগীর সম্পর্কে তথ্য বিশ্লেষণ করা। এটি বিশেষজ্ঞদের তাদের ক্লায়েন্টদের স্বাস্থ্যের অবস্থা সম্পর্কে আরও ভাল ধারণা প্রদান করতে এবং তাদের চিকিত্সা পরিকল্পনা তৈরি করতে সাহায্য করতে (Artificial intelligence (AI) has vast potential to improve the quality and efficiency of healthcare delivery. Here are some examples and favorable areas:\n\n1. Interaction with the patient: AI systems can play a special role in ongoing interactions with patients, for example searching and analyzing data to recommend the best treatment options. This can provide patients with a clearer understanding of their health status and increase the likelihood of receiving appropriate care.\n\n2. Identification of healthcare facilities: AI can also be used for identification of healthcare facilities, such as identification of time intervals and service efficiency. This can help improve service quality and identify problems, which can then be improved.\n\n3. Development of convenient processes in healthcare delivery: AI can also be used in the development of convenient processes in healthcare delivery, such as providing support for decision-making such as determining a patient's hospitalization or treatment needs. This can help improve the patient experience and improve the efficiency of healthcare delivery.\n\n4. Natural Language Processing for Medicine: AI can be used to develop natural language processing technology that can be used to help patients communicate with healthcare providers. It allows patients to access healthcare and choose appropriate options for their treatment.\n\n5. Providing support to experts: AI can also be used to provide support to experts, such as recommending recommended options for treatment and analyzing information about the patient. This is to help specialists provide a better understanding of their clients' health conditions and formulate their treatment plans)

Figure 16: The example of responses in Bengali given the instruction from the knowledge category.

English Instruction	I want to write a modern operating system from scratch for AMD64 systems with your assistance. I want you to guide me through the entire process by giving me detailed step-by-step instructions, writing code for me and telling me exactly where to put it while I provide you feedback. Where do I start? Answer in Swahili.
Instruction Source	Dataset: WizardLM Question_id: 36 Category: Code Generation
Swahili Instruction	Ninataka kuandika mfumo wa uendeshaji wa kisasa kutoka mwanzo kwa mifumo ya AMD64 kwa usaidizi wako. Nataka uniongoze katika mchakato mzima kwa kunipa maelekezo ya kina ya hatua kwa hatua, kuniandikia msimbo na kuniambia ni wapi pa kuiweka huku nikitoa maoni. Nianzie wapi?
X-Instruction$_{13B}$	Kuandika mfumo wa uendeshaji kutoka mwanzo ni kazi ngumu na yenye changamoto sana, na inategemea uzoefu wako wa kutengeneza programu na uelewa wa kiufundi. Ikiwa huna uzoefu wowote wa kutengeneza programu au kiufundi, ninapendekeza kuanza na kozi maalum au kitabu cha mafunzo ili kupata ufahamu wa msingi wa jinsi mifumo ya uendeshaji inavyofanya kazi. Ikiwa unaamua kuendelea na mradi huu, hapa kuna hatua za kuanzia: 1. Chagua lugha ya programu: Mifumo ya uendeshaji inaweza kuandikwa katika lugha nyingi tofauti za programu, ikiwa ni pamoja na C, C++, Rust, na zaidi. Chagua lugha ambayo inafaa zaidi kwako na ambayo ina jumuiya inayoendelea na rasilimali nyingi za mtandaoni. 2. Jifunze misingi ya AMD64: Kabla ya kuanza kuandika mfumo wa uendeshaji, ni muhimu kuelewa jinsi mfumo wa AMD64 unavyofanya kazi. Hii itajumuisha kuelewa muundo wa hifadhidata, mfumo wa uendeshaji wa kutofautisha, na michakato ya kuingiza na kutolea. … ***(some content is skipped.)*** 9. Sanidi na uchapishe mfumo wako wa uendeshaji: Mara tu unapokamilisha mfumo wako wa uendeshaji, unaweza kuifanye iwe rahisi kutumia na kusanidi kwa kutumia zana kama vile Make au CMake. Unaweza pia kuchapisha mfumo wako wa uendeshaji kama mradi wa GitHub ili wengine waweze kuuliza masuala na kuchangia. Kumbuka kwamba kuandika mfumo wa uendeshaji kutoka mwanzo ni mchakato mgumu na yenye changamoto sana, na inaweza kuchukua muda mrefu na juhudi kukamilisha. Ninapendekeza kuanza na kozi maalum au kitabu cha mafunzo ili kupata ufahamu wa msingi wa jinsi mifumo ya uendeshaji inavyofanya kazi, na kisha uanze na hatua iliyo hapo juu. (Writing an operating system from scratch is a very difficult and challenging task, and it depends on your programming experience and technical understanding. If you don't have any programming or technical experience, I recommend starting with a specialized course or textbook to get a basic understanding of how operating systems work. If you decide to continue with this project, here are the steps to start: 1. Choose a programming language: Operating systems can be written in many different programming languages, including C, C++, Rust, and more. Choose the language that works best for you and that has a thriving community and many online resources. 2. Learn the basics of AMD64: Before starting to write an operating system, it is important to understand how the AMD64 system works. This will include understanding database structure, variable operating system, and input and output processes. … ***(some content is skipped.)*** 9. Build and publish your operating system: Once you've finalized your operating system, you can make it easy to use and configure using tools like Make or CMake. You can also publish your operating system as a GitHub project so that others can ask for issues and contribution. Remember that writing an operating system from scratch is a very difficult and challenging process, and it can take a long time and effort to complete. I recommend starting with a special course or textbook to get a basic understanding of how operating systems work, and then start with the above step.)

Figure 17: The example of responses in Swahili given the instruction from the code generation category.