Improving Multilingual Instruction Finetuning via Linguistically Natural and Diverse Datasets

Sathish Reddy Indurthi†**, Wenxuan Zhou, Shamil Chollampatt, Ravi Agrawal**
Kaiqiang Song, Lingxiao Zhao, Chenguang Zhu
Zoom Video Communications

Abstract

Advancements in Large Language Models (LLMs) have significantly enhanced instruction-following capabilities. However, most Instruction Fine-Tuning (IFT) datasets are predominantly in English, limiting model performance in other languages. Traditional methods for creating multilingual IFT datasets—such as translating existing English IFT datasets or converting existing NLP datasets into IFT datasets by templating—struggle to capture linguistic nuances and ensure prompt (instruction) diversity. To address this issue, we propose a novel method for collecting multilingual IFT datasets that preserves linguistic naturalness and ensures prompt diversity. This approach leverages English-focused LLMs, monolingual corpora, and a scoring function to create high-quality, diversified IFT datasets in multiple languages. Experiments demonstrate that LLMs finetuned using these IFT datasets show notable improvements in both generative and discriminative tasks, indicating enhanced language comprehension by LLMs in non-English contexts. Specifically, on the multilingual summarization task, LLMs using our IFT dataset achieved 17.57% and 15.23% improvements over LLMs fine-tuned with translation-based and template-based datasets, respectively.

1 Introduction

Recent advancements in natural language processing (NLP) have showcased remarkable progress, particularly in its instruction-following capabilities. Notably, Large Language Models (LLMs) like GPT-4, Gemini-1.5, Claude-3, Llama-3, and Mistral (OpenAI, 2024; Team et al., 2024; AI@Meta, 2024; Jiang et al., 2023) have demonstrated significant prowess in this area (Brown et al., 2020; Le Scao et al., 2023; Chowdhery et al., 2023). After the pretraining stage, LLMs are fine-tuned on

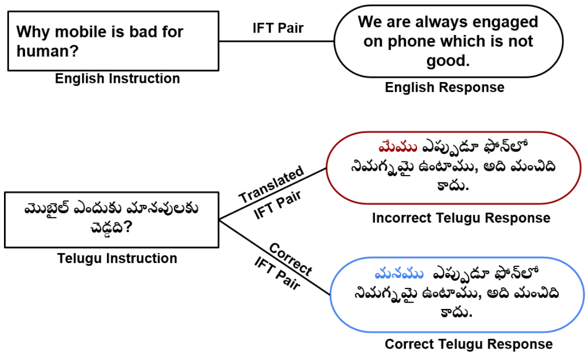

Figure 1: The incorrectly translated Telugu instruction-response pair is from the Aya collection (Üstün et al., 2024), which was translated from an English instruction-response pair in the Dolly v2 dataset (Conover et al., 2023). The correct Telugu instruction-response pair was provided by a native Telugu speaker.

Instruction Fine-Tuning (IFT) datasets followed by an optional Alignment Tuning (AT) based on the availability of the training datasets. IFT datasets consist of instruction prompt-response pairs and have proven instrumental in enhancing the efficacy and overall instruction-following abilities of LLMs (Anil et al., 2023; Sanh et al., 2022; Wei et al., 2023; Iyer et al., 2023; Chung et al., 2022; Wang et al., 2022a; Zhang et al., 2024a). However, a notable disparity persists between the abundance of instruction prompts available in English compared to other languages. While over 7k[1] languages are spoken worldwide, a staggering 73% of prevalent IFT datasets primarily cater to English alone (Longpre et al., 2023).

While LLMs often demonstrate proficiency in understanding and generating text across multiple languages, the language imbalance in training datasets has led to suboptimal performance in non-English contexts (Ahuja et al., 2023; Lai et al., 2023a; Zhang et al., 2023c). To enhance LLMs' ability to follow non-English instructions,

†Correspondence to sathishreddy.indurthi@zoom.us

[1]https://www.ethnologue.com/

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 2306–2323
November 12-16, 2024

Telugu Instruction

మునుపటి ప్రశ్నను బట్టి, జవాబును కలిగి ఉన్న సందర్భం వ్రాయండి. ఇది 1 - 20 వాక్యాలు కావచ్చు. సారాంశం

Given the previous question, write a context containing the answer. It can be 1 - 20 sentences. Summary

Figure 2: Lack of diversity in templated datasets: The template created by human annotators has been repeated several thousand times in the templated adversarial QA dataset from the Aya collection (Üstün et al., 2024)

various studies have explored fine-tuning LLMs on multilingual Instruction Fine-Tuning (IFT) datasets (Muennighoff et al., 2023; Wei et al., 2023; Lai et al., 2023b; Zhang et al., 2024b; Shaham et al., 2024; Chen et al., 2024a; Üstün et al., 2024). However, creating such multilingual IFT datasets is challenging. Previous efforts have primarily focused on two approaches: translating existing English IFT datasets or templating existing Natural Language Processing (NLP) datasets in non-English languages through native speakers to form IFT-style datasets. Each approach has its drawbacks, highlighting the need for more effective methods.

Translating English IFT datasets poses significant challenges, primarily because it fails to capture the nuances and intricacies unique to each language (Liu et al., 2024; Zhang et al., 2023b). Additionally, the translation process often introduces errors, leading to suboptimal performance when fine-tuning LLMs on these translated datasets, as the models absorb these errors during training (Xu et al., 2023; Zhou et al., 2023; Kong et al., 2023). For example, in Figure 1, the first translated response (red) is incorrect, even though it differs from the correct response (blue) by just one word. The red and blue words are used in different contexts in Telugu and do not have direct translations in English. Despite being generated by a state-of-the-art translation model, the first translation (red) fails to capture the correct meaning. Thus, relying entirely on translated data poses significant challenges in accurately reflecting the nuances of non-English languages.

Comparatively, the templating approach avoids the introduction of translation errors. However, achieving high diversity through templated approaches is challenging and often tedious due to the manual effort required (Muennighoff et al., 2023; Sanh et al., 2022). For instance, as shown in Figure 2, one of the templated datasets contains the same instruction repeated several thousand times, resulting in a lack of diversity in the IFT dataset.

To address the issues of translation and templated approaches, we introduce an efficient method to collect high-quality multilingual IFT datasets. The proposed method preserves the nuances of languages, avoids errors, and creates a diverse set of IFT examples for multiple languages. This is achieved by leveraging an English-focused LLM and the availability of monolingual corpora in each non-English language. We also employ a scoring function to control the quality of generated IFT examples. By relying on English-focused LLMs, we can tap into their extensive capabilities and transfer these abilities across diverse linguistic contexts. Utilizing monolingual corpora allows us to capture the unique linguistic and cultural nuances of each language, enhancing performance and accuracy in multilingual applications. Additionally, the robust scoring function ensures that the knowledge and capabilities derived from English-centric LLMs are appropriately adapted and optimized for non-English languages.

Extensive experiments on both generative and discriminative tasks demonstrate the effectiveness of the multilingual IFT datasets resulting from our proposed method. Compared to models fine-tuned on IFT datasets created using translation and templated approaches, the model fine-tuned on IFT datasets from our method achieves an average improvement of 11.1% in generative tasks and 6.9% in discriminative tasks. Furthermore, these improvements are obtained with an IFT dataset less than half the size of those created using templated and translation methods, highlighting the superior quality and diversity of the IFT dataset generated by our approach.

2 Method

A fundamental component in the development of Multilingual Large Language Models (MLLMs) lies in the acquisition of training datasets, crucially needed throughout distinct phases: Pretraining (PT), Instruction fine-tuning (IFT), and Alignment Tuning (AT).

While obtaining the necessary monolingual datasets for pretraining is relatively straightforward, acquiring datasets for instruction fine-tuning and

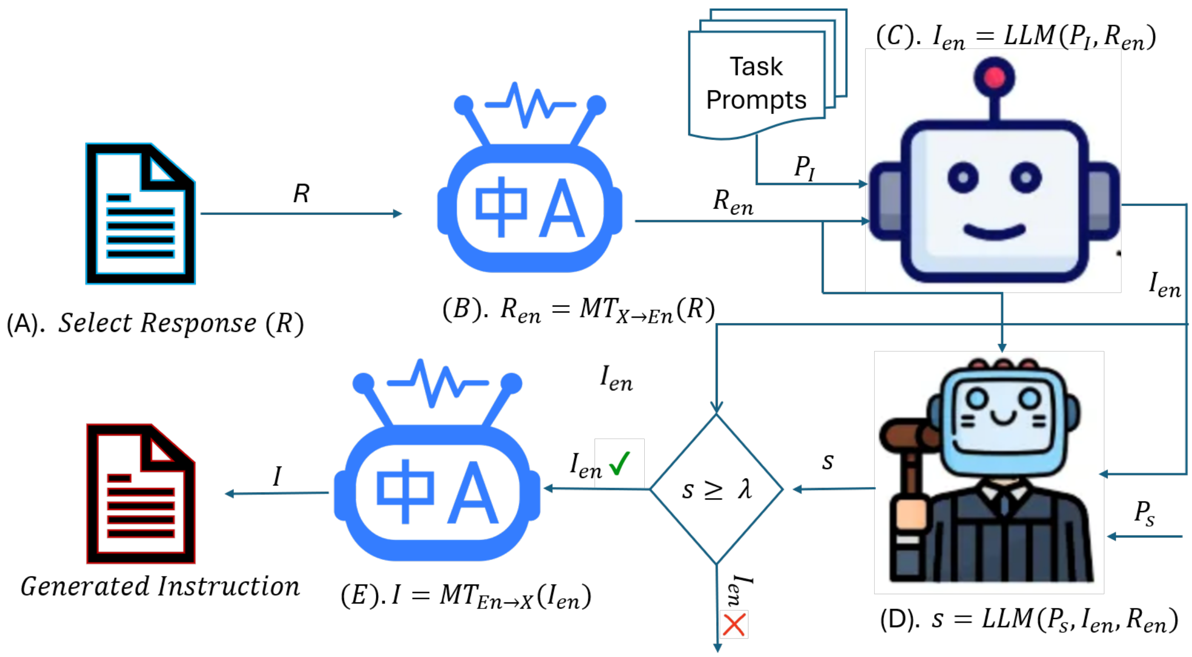

Figure 3: Overview of the proposed method: (A) Select Response, (B) Translating Response to English, (C) Generating English instructions using the English Response and task-specific prompt, (D) Scoring the generated English instruction against the translated response, and (E) Translating the English instruction back to the language of the original response.

alignment tuning presents significant challenges due to the costs and human effort involved. To address these challenges while maintaining linguistic characteristics and diversity, we propose a framework for creating IFT datasets for multiple languages. The framework consists of five stages, illustrated in Figure 3 and described below:

(A). Select Responses: We utilize a monolingual corpus as the primary source of response, supplemented by answers from existing NLP datasets for each non-English language (x). We extract several thousand text fragments from these non-English corpora, deduplicate, and apply various heuristics to filter out potentially low-quality fragments. These heuristics include criteria such as the prevalence of capitalized letters and specialized symbols. These text fragments are natural and most likely error-free output since they are from the monolingual corpus or human-curated answers from existing NLP datasets. Each fragment, denoted as $\mathcal{R}_x$, which varies in length to resemble responses in real-world scenarios, is then used to generate pseudo instructions through the following steps. By doing this, we ensure the output quality of the multilingual IFT data.

(B). Translating Responses: Given the availability of competent English LLMs in both open-source and closed environments, we have chosen to generate pseudo-instructions in English. This strategic decision allows us to leverage the strength of these models, ensuring the generation of high-quality and diverse instructions that cater to a wide range of NLP tasks, we translate the selected response $(\mathcal{R}_x)$ into English.

$$\mathcal{R}_{en} = \mathcal{MT}_{x \to en}(\mathcal{R}_x)$$

(C). Generating Instruction: We generate English instructions by utilizing English-focused LLM, a translated response $(\mathcal{R}_{en})$, and a randomly selected prompt $(\mathcal{P}_I)$ from a pool of predefined task prompts. Our approach involves designing a range of prompts specifically tailored to support various NLP tasks, including question-answering, summarization, and sentiment analysis. Additionally, the prompt allows for open-ended instruction generation, providing LLMs with the opportunity to produce the most plausible instructions for a given response. Focusing on generating instructions in English enables us to tap into the extensive resources and capabilities available for this language, thereby enhancing the adaptability and effectiveness of our approach across diverse linguistic contexts. This emphasis on English instruction generation also ensures seamless integration with existing English-centric NLP systems, further augmenting the versatility and applicability of our methodology in real-world scenarios. Formally, the English instruction ($\mathcal{I}_{en}$) is generated by:

$$\mathcal{I}_{en} = \mathcal{LLM}(\mathcal{P}_I, \mathcal{R}_{en}) \quad (1)$$

(D). Scoring The instructions generated through (1) do not always yield high-quality examples due to misalignment in the prompt-response pair or LLM's failure to generate appropriate instruction. Thus we rely on a scoring function to filter and identify high-quality examples while maintaining diversity in the generated dataset.

We use LLM as a judge, employing the prompt $\mathcal{P}_s$ to assess the quality of $(\mathcal{I}_{en}, \mathcal{R}_{en})$ pair. This results in a score, denoted as s:

$$s = \mathcal{LLM}(\mathcal{P}_s, \mathcal{I}_{en}, \mathcal{R}_{en}) \quad (2)$$

Pairs with a score greater than or equal to a predefined threshold (λ) are used for fine-tuning, while those below this threshold are removed from fine-tuning phase.

(E). Translating Instruction: Following the scoring phase, we proceed to translate $\mathcal{I}_{en}$ into the same language as $\mathcal{R}_x$:

$$\mathcal{I}_x = \mathcal{MT}_{en \to x}(\mathcal{I}_{en})$$

Subsequently, we form a training pair $(\mathcal{I}_x, \mathcal{R}_x)$. Here, $\mathcal{I}_x$ serves as a pseudo instruction, while $\mathcal{R}_x$ represents natural text in the same non-English language. During the LLM fine-tuning stage, despite potential unnaturalness and errors in $\mathcal{I}_x$ arising from the instruction generation and translation process, the model is trained to generate $\mathcal{R}_x$, which is typically a natural and error-free output sourced from the monolingual corpus or existing human-curated NLP datasets. Leveraging such pairs enhances the model's ability to handle instruction errors and improves its overall language comprehension.

The sample task prompts ($\mathcal{P}_I$) and scoring prompt ($\mathcal{P}s$) used in Equation 1 and Equation 2 are provided in Table 8 and Table 9 in the Appendix.

3 Experimental Settings

3.1 Dataset creation

We utilize the CC-100 monolingual dataset (Conneau et al., 2020). We also utilize answers from the templated examples in the aya dataset (Üstün et al., 2024). In both cases, the texts are written in the native language not derived from other languages (Wenzek et al., 2020). We selected the text based on the criteria described in Section 2. We choose four languages: Telugu (tel), Hindi (hin), Japanese (jpn), and Spanish (spa) to create IFT datasets through our approach. According to Aya and Okapi (Üstün et al., 2024; Lai et al., 2023b), Telugu and Nepali are low-resource, Indonesian and Hindi are mid-resource, and Japanese and Spanish are high-resource languages. We collected approximately one million text fragments for each language.

Language	TM	TR	GR
Telugu	1,312,185	2,596,857	523,739
Hindi	1,171,530	2,540,447	570,467
Japanese	2,392,691	3,029,014	531,163
Spanish	1,220,649	2,560,149	557,563

Table 1: Total number of instruction-response pairs used for fine-tuning the LLMs by Templated (TM), Translation (TR), and Generated (GR) approaches.

.

In creating multilingual datasets using the proposed approach, we utilize open source *meta-llama/Meta-Llama-3-70B-Instruct* (AI@Meta, 2024) as our LLM to generate instructions and also to score instruction-response pairs. However, this LLM can be replaced with more powerful open-source or closed-source LLMs to improve the quality of generated instructions further.

We utilize NLLB-200 (Costa-jussà et al., 2022)[2], which has support for 200 languages with state-of-the-art translation quality. The same model is used for translating the response ($\mathcal{R}$) to English as well as for translating ($\mathcal{I}_{en}$) into the language of ($\mathcal{R}$). After the translation, we use the COMET score (Rei et al., 2020) to remove low-quality translated responses ($\mathcal{R}_{en}$) and generated instructions ($\mathcal{I}_x$). Specifically, we use *Unbabel/wmt23-cometkiwi-da-xl* model (Rei et al., 2023), which is a reference-free model with 3.5 billion parameters. We retain examples with COMET scores greater than or equal to 0.7.

3.2 Training details

We use *Meta-Llama-3-8B* (AI@Meta, 2024)[3] as the base model to fine-tune on our multilingual IFT dataset. We also fine-tune non-English focused models such as *Rakuten-ai-7B-Instrcut* (Rakuten

[2]https://huggingface.co/facebook/nllb-200-3.3B

[3]https://huggingface.co/meta-llama/Meta-Llama-3-8B

Language	Templated		Translated		Generated	
	Instruction	Response	Instruction	Response	Instruction	Response
Telugu (tel)	344(±312)	221(±297)	223(±295)	204(±179)	381(±917)	308(±482)
Hindi (hin)	401(±450)	290(±315)	228(±397)	203(±181)	475(±897)	358(±582)
Japanese (jpn)	67(±79)	95(±115)	94(±172)	86(±78)	162(±473)	98(±116)
Spanish (spa)	306(±280)	138(±215)	238(±435)	215(±196)	425(±723)	289(±475)

Table 2: Average lengths (#characters) of instruction-response pairs in templated, translated, and generated approaches.

Group et al., 2024), *Aya-23* (Aryabumi et al., 2024). During training, we only optimize the loss on the output tokens, not the input tokens, thus deviating from the standard language modeling loss. We apply the same hyperparameters as existing instruction fine-tuning (IFT) methods (Zhou et al., 2023; Touvron et al., 2023): a learning rate of $1e^{-5}$ that linearly decays to $9e^{-6}$ by the end of the training, weight decay of 0.1, batch size of 128 examples, and dropout of 0.1. For generation, we use nucleus sampling (Holtzman et al., 2020) with a temperature of $T = 0.7$ and $p = 0.9$. We use 8 NVIDIA H100 GPUs for fine-tuning the model.

RougeLsum				
	tel	**hin**	**jpn**	**spa**
Templated Approaches				
Aya-TM	18.0	33.8	7.9	24.2
Llama-3-8B-TM	19.6	36.4	17.8	26.8
Translated Approaches				
Bactrian-X	12.1	23.5	5.2	15.7
Aya-TR	16.9	32.8	6.7	22.1
Llama-3-8B-TR	18.4	35.9	18.4	25.9
Ours				
Llama-3-8B-GR	**24.3**	**39.5**	**22.6**	**29.5**

Table 3: Performance of models on XLSUM.

4 Results

In Table 1, we present the statistics of datasets created using various approaches. The statistics of datasets created using the template-based and translation-based approaches are from *aya_collection* (Üstün et al., 2024). Please see the Appendix for more details. Using our approach, we generated approximately 500K instruction-response pairs from the initial pool of 1M text fragments for each language.

We evaluate the performance of models fine-tuned on datasets collected using our approach against models fine-tuned on datasets obtained through translation and template-based methods. Specifically, we compare the *Aya-TM* and Llama-3-8B-TM models, which are trained on template-based datasets as described in Üstün et al. (2024). Additionally, we assess the *Aya-TR* and Llama-3-8B-TR models, which are trained on translation-based datasets detailed in Üstün et al. (2024). Both types of datasets include the Aya-human annotated dataset[4]. Furthermore, we compare these with the *Bactrian-X* model (Li et al., 2023), fine-tuned on a dataset comprising translated English instructions and their corresponding multilingual responses generated using ChatGPT. Our final model, Llama-3-8B-GR, is trained using the created instruction-response dataset along with the Aya human-annotated dataset. In all the approaches, the percentage of training examples collected through the human annotation process corresponds to less than 0.1%.

4.1 Generative Tasks

We evaluated the models on two generative tasks: summarization using *XLSUM* (Hasan et al., 2021) and machine translation using *FLORES-200* (Costa-jussà et al., 2022). These tasks were selected because they include responses written in native languages, not derived from other languages. We present the performance of our model, Llama-3-8B-GR-H, and its variants, comparing them to baseline models across the four languages used for creating multilingual IFT datasets. For the summarization task, we employed the RougeLsum metric (Lin, 2004), and for the translation task, we utilized sp-BLEU (Goyal et al., 2021) and chrF++ (Popović, 2017)[5].

Tables 3 and 4 present the results for the summarization and machine translation tasks using the XLSUM and FLORES-200 datasets, respec-

[4]https://huggingface.co/datasets/CohereForAI/aya_dataset

[5]https://github.com/mjpost/sacrebleu

	tel	hin	jpn	spa
spBleu				
Templated datasets				
Aya-TM	21.9	22.7	18.2	27.1
LLama-3-8B-TM	24.6	25.3	21.6	30.7
Translated datasets				
Bactrian-X	17.3	19.2	11.78	22.4
Aya-TR	21.0	22.8	14.7	28.4
Llama-3-8B-TR	23.5	24.9	20.2	31.2
Ours				
Llama-3-8B-GR	**27.2**	**28.4**	**24.8**	**33.9**
chrF++				
Templated datasets				
Aya-TM	44.7	44.1	29.7	50.3
Llama-3-8B-TM	47.1	46.9	34.7	58.4
Translated datasets				
Bactrian-X	35.8	36.9	22.1	42.8
Aya-TR	45.5	44.9	29.9	51.9
Llama-3-8B-TR	47.7	46.4	35.0	57.7
Ours				
Llama-3-8B-GR	**49.8**	**50.2**	**38.3**	**63.2**

Table 4: Performance of models on FLORES-200 devtest set (en→xxx).

	tel	hin	spa
Translated datasets			
Bactrian-X	24.5	26.2	27.2
Okapi	26.9	27.9	30.3
Aya-TR	32.1	38.7	39.7
Llama-3-8B-TR	34.1	41.4	42.9
Ours			
Llama-3-8B-GR	**36.3**	**44.7**	**45.6**

Table 5: Performance of models on multilingual MMLU

Figure 4: Instruction diversity in the generated IFT dataset. The inner circle displays common root verbs, while the outer circle shows the corresponding noun objects, based on approximately 15 percent of instructions generated across 4 languages. The figure only represents 13.1% of verb-noun pairs since not all instructions have the parsed verb-noun structure.

tively. From the results presented in both tables, models trained with translated datasets do not exhibit any improvement over those trained with template datasets. In contrast, the Llama-3-8B-GR model, fine-tuned on datasets created using our method, demonstrates significant performance enhancements across both tasks compared to all other dataset types. Our dataset, free from translation errors and rich in diversity, enables the model to better capture the authentic form of language, leading to superior performance.

4.2 Discriminative Tasks

We also evaluate the models on a discriminative task to assess whether introducing high-quality, diversified, and native-written responses enhances the model's language comprehension and overall performance. Specifically, we use the multilingual MMLU task (Lai et al., 2023b), a machine-translated version of the English MMLU task (Hendrycks et al., 2021), to compare the performance of models trained extensively on translated datasets versus those trained on native datasets created using our approach. This task was unseen during the models' fine-tuning stage, so we employ a few-shot evaluation to compare performance. The *Llama-3-8B* and *Aya* models use a 5-shot evaluation, while the *Bactrian-X* and *Okapi* models use a 25-shot evaluation. The task comprises 13,000 questions covering 57 topics, ranging from STEM and humanities to social sciences.

Table 5 shows the multilingual results in three languages. The model trained with our dataset (*Llama-3-8B-GR*), outperforms the models trained with datasets collected using other approaches. Our model outperforms Okapi, Aya, and our baseline by 48.74%, 13.8%, and 6.9%, respectively. These results indicate that the diversity and quality of the datasets lead to better performance.

Despite our dataset being 2.7 and 4.9 times smaller than the templated and translated datasets, respectively, the model fine-tuned on our dataset

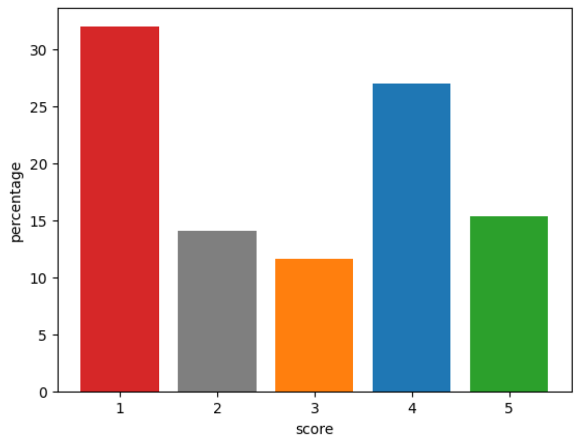

Figure 5: Scores assigned by LLM judge on Instruction-Response pairs. The scores are averaged across all languages.

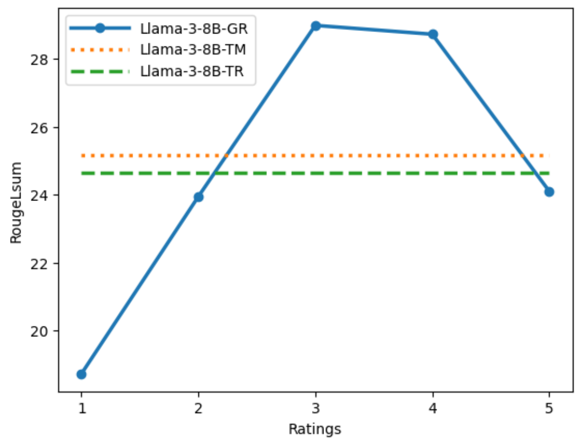

Figure 6: Importance of scoring function in creating high-quality IFT dataset. The x-axis represents the scoring threshold used to filter the IFT dataset. The Y-axis represents the average RougeLsum score of Telugu, Hindi, Japanese, and Spanish languages from the XLSUM summarization task.

achieved significant improvements in both generative and discriminative tasks. This underscores the importance of high-quality, diversified datasets in developing efficient multilingual LLMs.

4.3 Analysis

4.3.1 Instruction diversity

To understand the diversity of the generated instructions, we plot the verb-noun structure of instructions in Figure 4. The figure visualizes the distribution of the most frequent root verbs and their corresponding most common direct noun objects from 15% of the generated instructions across four languages. These noun-verb pairs represent 8.1% of the entire set, which exhibits diverse intents and patterns in our generated instructions. We also provide a few generated samples in the Appendix.

We also report the average length of instructions and responses from all data creation approaches. As shown in Table 2, the average number of characters in the instructions generated using our approach varies significantly compared to the other two approaches. This variation arises from using different types of task prompts when generating an instruction for a given response.

4.3.2 Effect of Scoring Function:

The frequency of average scores obtained using the LLM judge is shown in Figure 5. To evaluate the impact of the scoring function on the creation of high-quality multilingual IFT datasets, we fine-tuned the *Llama-3-8B-GR* model on datasets filtered using different scoring thresholds, $\lambda = \{1, 2, 3, 4, 5\}$. For each specific threshold λ_i, all examples below that score were excluded from the training set. We then compared the performance of the *Llama-3-8B-GR* model trained on these filtered datasets against models (*Llama-3-8B-TM* and *Llama-3-8B-TR*) trained on template-based and translation-based datasets. As illustrated in Figure 6, the performance of *Llama-3-8B-GR* improves as the scoring threshold increases up to $\lambda = 3$, achieving superior performance compared to the *Llama-3-8B-TM* and *Llama-3-8B-TR* models. Beyond $\lambda = 3$, performance declines due to the reduced size of the training dataset. These results underscore the critical role of the scoring function in creating high-quality multilingual IFT datasets.

Model	XLSUM (Rouge-2)	MMLU (Acc.)
RakutenAI-7B (Rakuten Group et al., 2024)	14.1	61.3
RakutenAI-7B-GR (w/ our IFT dataset)	**18.5**	**63.2**

Table 6: Performance of Japanese-focused LLMs on XLSUM and MMLU Japanese tasks.

4.3.3 Effect on non-English focused models.

To evaluate the diversity and quality of our IFT datasets, we conducted further fine-tuning on two robust non-English-focused LLMs using our IFT datasets. First, we assessed the impact on the Japanese-focused model (Rakuten Group et al., 2024). This model was initially pre-trained on Japanese texts and fine-tuned on Japanese

Model	XLSUM (RougeL)	MMLU (Acc.)
Aya-23-8B (Aryabumi et al., 2024)	29.7	45.3
Aya-23-8B-GR (w/ our IFT dataset)	**31.4**	**46.8**

Table 7: Performance of Aya-23-8B LLM on XLSUM and MMLU Hindi and Spanish tasks. The *Aya-23-8B-GR* model is obtained by further finetuning of *Aya-23-8B* model on our Hindi and Spanish IFT datasets.

instruction-response pairs. Second, we evaluated the performance of a state-of-the-art multilingual LLM named Aya-23 (Aryabumi et al., 2024). This model is based on Cohere's Command model[6] and was instruction-tuned on 23 languages using the template-based dataset from Üstün et al. (2024).

As shown in Table 6 and Table 7, fine-tuning further on our IFT dataset significantly enhances the performance of these non-English-focused LLMs.

5 Related Work

Multilingual LLMs. LLMs (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023; OpenAI, 2024) have achieved remarkable results on various NLP tasks (Hendrycks et al., 2021; Srivastava et al., 2022). With over 7,000 languages spoken worldwide and approximately 2,500 classified as low-resource by Joshi et al. (2020), which are spoken by more than 1 billion people, there is a growing need to expand the language coverage of LLMs. To develop LLMs with multilingual capabilities, one straightforward approach is to pretrain them on a diverse set of languages. For example, BLOOM (Le Scao et al., 2023) is pretrained on 46 languages and 13 programming languages, while Llama-2 (Touvron et al., 2023) is pretrained primarily on English with additional data from 27 other languages. Despite these efforts, the language coverage of these models remains limited and predominantly focused on English. Another approach is to continually train LLMs with additional languages (Cui et al., 2023; Basile et al., 2023; ImaniGooghari et al., 2023). In particular, Chinese-Llama (Cui et al., 2023) continually trains Llama on Chinese corpora and integrates additional Chinese tokens into the original vocabulary to further improve the Chinese ability.

Instruction Tuning. Instruction tuning has been a key paradigm for LLMs to improve their general performance and ability to follow instructions (Wei et al., 2022; Wang et al., 2022b; Ding et al., 2023). However, these models are predominantly tuned using English, resulting in significant discrepancies in performance across languages (Huang et al., 2023; Etxaniz et al., 2023). Multilingual instruction tuning has effectively narrowed this performance gap (Kew et al., 2023; Chen et al., 2024b). Typically, the data for multilingual instruction tuning is derived through translation from English data (Li et al., 2023; Zhang et al., 2023a; Üstün et al., 2024), but this approach often misses cultural nuances and can introduce unnatural responses. Some efforts (Üstün et al., 2024) utilize templates to automatically create large amounts of multilingual data, but this method is constrained by limited diversity in the instructions. We propose to generate instructions directly from original multilingual responses, which preserves the naturalness of responses and enhances the diversity of instructions.

[6]https://cohere.com/command

6 Conclusion

In conclusion, our research addresses the notable disparity in Instruction Fine-Tuning (IFT) datasets, predominantly centered on English, by proposing a novel method for collecting multilingual IFT datasets. By leveraging English-focused LLMs and monolingual corpora, our approach maintains the naturalness of specific languages and ensures diversity in the datasets. The quality control through a scoring function further enhances the effectiveness of the generated datasets.

Our extensive experiments on generative tasks demonstrate that the models trained with our multilingual IFT datasets significantly outperform those trained on traditional translated and templated datasets. Moreover, our models show substantial improvements in discriminative tasks, indicating a better comprehension of language.

These results underscore the importance of diverse and high-quality multilingual datasets in enhancing the performance of large language models across various languages. Our method provides a viable solution to the challenges faced in creating effective multilingual IFT datasets, paving the way for more inclusive and capable language models. Future research can build upon this approach to further refine and expand the capabilities of LLMs in a broader range of linguistic contexts.

Limitations

Since the instructions were generated by LLMs, there may be inherent biases originating from the underlying models used in this study. Nevertheless, the models used are open-source, extensively utilized by the community, and trained with the goals of reducing bias and enhancing safety and usefulness.

This study aims to systematically assess the effectiveness of generated instructions for given responses in various languages. Due to limitations in computing resources, we were unable to extend the proposed data creation framework beyond four languages. However, we endeavored to cover low, medium, and high-resource languages and evaluated our approach on several NLP tasks.

In our evaluation of LLMs using different IFT-style datasets, we selected two generative tasks and one discriminative task to demonstrate the impact of our dataset. The study was limited to three tasks due to computational and time constraints. However, these tasks are popular and widely used in evaluating multilingual LLMs.

In future work, we plan to extend our evaluation to LLMs optimized for additional languages and explore multiple benchmarks within each language to better understand the native aspects of LLM performance.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai. *Preprint*, arXiv:2303.12528.

AI@Meta. 2024. Llama 3 model card.

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 Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan 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, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *arXiv preprint arXiv:2312.09993*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024a. Monolingual or multilingual instruction tuning: Which makes a better alpaca. *Preprint*, arXiv:2309.08958.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul

Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *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. *Preprint*, arXiv:1911.02116.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

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

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english? *arXiv preprint arXiv:2308.01223*.

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

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *Preprint*, arXiv:2106.13822.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *Preprint*, arXiv:1904.09751.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

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

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *Preprint*, arXiv:2212.12017.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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

Kezhi Kong, Jiuhai Chen, John Kirchenbauer, Renkun Ni, C. Bayan Bruss, and Tom Goldstein. 2023.

GOAT: A global transformer on large-scale graphs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17375–17390. PMLR.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023a. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *Preprint*, arXiv:2304.05613.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023b. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *Preprint*, arXiv:2307.16039.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176Bxx3-parameter open-access multilingual language model.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *Preprint*, arXiv:2403.10258.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2023. The data provenance initiative: A large scale audit of dataset licensing & attribution in AI. *Preprint*, arXiv:2310.16787.

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. *Preprint*, arXiv:2211.01786.

OpenAI. 2024. Gpt-4 technical report. *Computation and Language*, arXiv:2303.08774. Version 6.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Rakuten Group, Aaron Levine, Connie Huang, Chenguang Wang, Eduardo Batista, Ewa Szymanska, Hongyi Ding, Hou Wei Chou, Jean-François Pessiot, Johanes Effendi, Justin Chiu, Kai Torben Ohlhus, Karan Chopra, Keiji Shinzato, Koji Murakami, Lee Xiong, Lei Chen, Maki Kubota, Maksim Tkachenko, Miroku Lee, Naoki Takahashi, Prathyusha Jwalapuram, Ryutaro Tatsushima, Saurabh Jain, Sunil Kumar Yadav, Ting Cai, Wei-Te Chen, Yandi Xia, Yuki Nakayama, and Yutaka Higashiyama. 2024. Rakutenai-7b: Extending large language models for japanese. *Preprint*, arXiv:2403.15484.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. *Preprint*, arXiv:2309.11925.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. *Preprint*, arXiv:2110.08207.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *Preprint*, arXiv:2401.01854.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Gemini Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomas Kocisky, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, Juliette Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Yingjie Miao, Lukas Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontanon, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, Anja Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matthew Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara Sainath, Shane Gu, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, Sébastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Josh Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, Sébastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Kiran Vodrahalli, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravi Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Lučić, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjösund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Zeyncep Cankara, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Lora Aroyo, Zhufeng Pan, Zachary Nado, Jakub Sygnowski, Stephanie Winkler, Dian Yu, Mohammad Saleh, Loren Maggiore, Yamini Bansal, Xavier Garcia, Mehran Kazemi, Piyush Patil, Ishita Dasgupta, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Qingze Wang, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Raphaël Lopez Kaufman, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Chris Welty, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabe Barth-Maron, Craig Swanson, Dominika Rogozińska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Renshen Wang, Dave Lacey, Anastasija Ilić, Yao Zhao, Adam Iwanicki, Alejandro Lince, Alexander Chen, Christina Lyu, Carl Lebsack, Jordan Griffith, Meenu Gaba, Paramjit Sandhu, Phil Chen, Anna Koop, Ravi Rajwar, Soheil Hassas Yeganeh, Solomon Chang, Rui Zhu, Soroush Radpour, Elnaz Davoodi, Ving Ian Lei, Yang Xu, Daniel Toyama, Constant Segal, Martin Wicke, Hanzhao Lin, Anna Bulanova, Adrià Puigdomènech Badia, Nemanja Rakićević, Pablo Sprechmann, Angelos Filos, Shaobo Hou, Víctor Campos, Nora Kassner, Devendra Sachan, Meire Fortunato, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, Tina Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anaïs White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Ying Xu, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, Çağlar Ünlü, David Reid, Zora Tung, Daniel Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Ricardo Aguilar, Mai Giménez, Jiawei Xia, Olivier Dousse, Willi Gierke, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Dan Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nick Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, DongHyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alex Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clement Farabet, Pedro Valenzuela, Quan Yuan, Ananth Agarwal, Mia Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, Andrey Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiri Simsa, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas FitzGerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel Kaed, Jing Li, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Alanna Walton, Alicia Parrish, Mark Epstein, Sara McCarthy, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022a. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *Preprint*, arXiv:2204.07705.

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. 2022b. 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 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*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *Preprint*, arXiv:2307.06018.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4003–4012.

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. *Preprint*, arXiv:2304.12244.

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

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024a. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023b. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Preprint*, arXiv:2306.05179.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023c. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. *Preprint*, arXiv:2305.16339.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024b. Plug: Leveraging pivot language in cross-lingual instruction tuning. *Preprint*, arXiv:2311.08711.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

A Prompt details

The task prompts are inspired by several NLP tasks. A few of the sample prompts are shown in Table 8. The scoring prompt used to evaluate our generated instruction and given response pair is given in Table 9.

B Examples

A few examples of multilingual IFT datasets created using our approach are shown in Figure 7, Figure 8.

C Templated and Translated Datasets

The templated and translated datasets in aya are constructed using *Aya_dataset*, *Aya_collection* datasets by following the ratios described in Table 10. The *Aya_dataset* is created by using native speakers from each language and contains approximately 6k examples per language. The *Aya_collection* is created by templating existing NLP datasets of each language as well as translating 19 datasets covering 93 languages. In total *Aya_collection* includes 513 million instances making it the largest open-source multilingual IFT dataset. For our experiments, we collected templated and translated datasets from *Aya_collection* for *Telugu, Hindi, Chinese, and Spanish* languages.

Response: హైదరాబాద్ లో అర్ధరాత్రి భారీవర్షం పడింది. పలు ప్రాంతాల్లో వర్షం దంచికొట్టింది. జూబ్లీహిల్స్, అమీర్పేట్, పంజాగుట్ట, మణికొండ, మాదాపూర్, రాజేంద్రనగర్, సికింద్రాబాద్, రామాంతపూర్, దిల్ షుఖ్ నగర్, ఎల్బీనగర్ తదితర ప్రాంతాల్లో వర్షం పడింది. గత రెండు రోజులుగా ఎడతెరిపిలేని వానతో నగరవాసులు ఇబ్బందులు పడుతున్నారు. ఇవాళ ఉదయం నుంచి 7.33 సెంటీమీటర్ల వర్షపాతం నమోదయ్యింది. భారీ వర్షంతో రోడ్లన్ని జలమయమయ్యాయి. ప్రజలకు రాకపోకలకు ఇబ్బందికరంగా మారిందని ప్రజలు వాపోతున్నారు.

Heavy rain fell in Hyderabad at midnight. Rain lashed many areas. Jubilee Hills, Ameerpet, Panjagutta, Manikonda, Madapur, Rajendranagar, Secunderabad, Ramantapur, Dil Shukh Nagar, LB Nagar and other places received rain. The residents of the city are suffering due to incessant rain for the last two days. 7.33 cm rainfall has been recorded since this morning. All the roads were flooded due to heavy rain. People complain that it has become difficult for people to travel.

Instruction: దయచేసి క్రింది వార్తా కథనాన్ని సంగ్రహించండి.
హైదరాబాద్ నగరంలో అర్ధరాత్రి నుంచి భారీ వర్షం కురుస్తోంది, పలు ప్రాంతాల్లో భారీ వర్షాలు కురుస్తున్నాయి. ఎడతెరిపి లేకుండా కురుస్తున్న వర్షం కారణంగా నగరంలోని పలు ప్రాంతాల్లో నీటి ఎద్దడి నెలకొంది. జూబ్లీ హిల్స్, ఒక నాగరిక ప్రాంతం, అత్యంత ప్రభావితమైన ప్రాంతాలలో ఒకటిగా ఉంది, నివాసితులు వరదలకు గురైన వీధుల్లో నావిగేట్ చేయడానికి కష్టపడుతున్నారు. అమీర్‌పేట్ మరియు పంజాగుట్ట సమీపంలోని ప్రాంతాలలో కూడా భారీ వర్షాలు కురిశాయి, డ్రెయిన్లు పొంగిపొర్లుతున్నాయని మరియు రోడ్లు ముంపునకు గురయ్యాయని నివేదించబడింది.\n\nనగరం యొక్క తూర్పు భాగం కూడా తీవ్రంగా ప్రభావితమైంది, మణికొండ మరియు మాదాపూర్ వంటి ప్రాంతాల్లో భారీ వర్షాలు కురుస్తున్నాయి. వర్షం కారణంగా రోజువారీ జీవనానికి అంతరాయం ఏర్పడింది, చాలా మంది ప్రయాణికులు తమ గమ్యస్థానాలకు చేరుకోవడానికి కష్టపడుతున్నారు. సాధారణంగా రద్దీగా ఉండే రాజేంద్రనగర్ మరియు సికింద్రాబాద్ రోడ్లు నిర్మానుష్యంగా మారాయి, కొన్ని వాహనాలు మాత్రమే భారీ వర్షాన్ని తట్టుకుని వెళ్తున్నాయి.\n\nరామనాట్‌పూర్ మరియు దిల్‌షుక్ నగర్ వంటి ప్రాంతాల్లో భారీ వర్షపాతం నమోదవడంతో నగరంలోని దక్షిణ భాగం కూడా అతలాకుతలమైంది. వర్షం కారణంగా కొన్ని ప్రాంతాల్లో విద్యుత్ సరఫరా నిలిచిపోవడంతో ప్రజలు కరెంటు లేకుండా పోయారు. ఎల్బీనగర్, జనసాంద్రత అధికంగా ఉండే ప్రాంతం, నీటి ఎద్దడి మరియు వరదల నివేదికలతో కూడా ప్రభావితమైంది.\n\nభారీ వర్షపాతం నగరాన్ని స్తంభింపజేసింది, చాలా మంది నివాసితులు ఇళ్లలో ఉండటానికి ఇష్టపడతారు. ముంపునకు గురైన రోడ్లను క్లియర్ చేసి విద్యుత్ సరఫరాను పునరుద్ధరించేందుకు అధికారులు కసరత్తు చేస్తున్నారు, అయితే వర్షం తీవ్రతతో ఆ పని సవాలుగా మారింది. నగరం యొక్క డ్రైనేజీ వ్యవస్థ పూర్తిగా నిండిపోయింది, ఇది విస్తృతమైన నీటి ఎద్దడి మరియు వరదలకు దారితీసింది. వర్షం కురుస్తూనే, 7.33 సెంటీమీటర్ల వర్షపాతానికి చేరుకోవడంతో, నివాసితులు కష్టతరమైన రోజు కోసం తమను తాము సిద్ధం చేసుకుంటున్నారు.

Please summarize the following news article.
The city of Hyderabad has been experiencing heavy rainfall since midnight, with several areas receiving significant amounts of rain. The downpour has been relentless, causing waterlogging in many parts of the city. Jubilee Hills, a posh locality, has been one of the worst affected areas, with residents struggling to navigate the flooded streets. The nearby areas of Amirpet and Panjagutta have also received heavy rainfall, with reports of overflowing drains and inundated roads.\n\nThe eastern part of the city has also been severely affected, with areas like Maniconda and Madapur receiving heavy rainfall. The rain has caused disruptions to daily life, with many commuters struggling to reach their destinations. The usually busy roads of Rajendranagar and Sikandrabad have been deserted, with only a few vehicles braving the heavy rain.\n\nThe southern part of the city has also been hit hard, with areas like Ramanatpur and Dilshuk Nagar experiencing heavy rainfall. The rain has caused power outages in some areas, leaving residents without electricity. Elbinagar, a densely populated area, has also been affected, with reports of waterlogging and flooding.\n\nThe heavy rainfall has brought the city to a standstill, with many residents preferring to stay indoors. The authorities have been working to clear the flooded roads and restore power supply, but the task has been challenging due to the intensity of the rain. The city's drainage system has been overwhelmed, leading to widespread waterlogging and flooding. As the rain continues to pour down, reaching 7.33 centimeters of rainfall, residents are bracing themselves for a difficult day ahead. (**Telugu Example**)

Figure 7: Telugu example based on summarization task

Response: భారత్ లో కరోనా సెకండ్ వేవ్ కలకలం సృష్టిస్తోంది. తగ్గినట్టే తగ్గిన కరోనా ఒక్కసారిగా విజృంభిస్తోంది. దీంతో దేశ వ్యాప్తంగా ఆందోళన మొదలైంది. సెకండ్ వేవ్ లో కరోనా వైరస్ వేగంగా విస్తరిస్తోంది. న్యూఢిల్లీలో గత 24 గంటల్లో 24 వేలకు పైగా కేసులు నమోదయ్యాయి. దీంతో ట్రేడర్స్ ఫెడరేషన్ ఆఫ్ ఇండియా ఆందోళన వ్యక్తం చేసింది. ప్రభుత్వం వెంటనే కఠిన చర్యలు తీసుకోవాలని కోరుతూ లేఖ రాసింది. ఢిల్లీ లెఫ్టినెంట్ గవర్నర్, సీఎం కేజ్రీవాల్ కు వారు లేఖ రాశారు. అందులో ఏమని పలు సూచనలు చేశారు. అందులో ఏమని పేర్కొన్నారంటే. "కరోనా సెకండ్ వేవ్ వేగంగా విస్తరిస్తోంది. ఈ వేవ్ ను అడ్డుకోవాలంటే కనీసం 15 రోజులు లాక్ డౌన్ ను విధించాలి. ఢిల్లీకి చేరుకునే అన్ని మార్గాలను మూసేయాలి. రైల్వే స్టేషన్లు, విమానాశ్రయాల్లో కోవిడ్ నిబంధనలను కఠినం చేయాలి." అని ఆల్ ఇండియా ట్రేడర్స్ ఫెడరేషన్ లేఖలో కోరింది.

The second wave of Corona is creating chaos in India. Corona, which has decreased, is suddenly booming. This started a panic across the country. Corona virus is spreading rapidly in the second wave. More than 24 thousand cases have been reported in New Delhi in the last 24 hours. Traders Federation of India has expressed concern over this. The government has written a letter asking to take strict action immediately. They wrote a letter to Lt Governor of Delhi and CM Kejriwal. Many suggestions have been made as to what is in it. What is mentioned in it. "The second wave of Corona is spreading rapidly. At least 15 days of lockdown should be imposed to stop this wave. All roads leading to Delhi should be closed. Covid rules should be tightened at railway stations and airports." All India Traders Federation demanded in the letter.

Instruction: భారతదేశంలో, ప్రత్యేకించి న్యూఢిల్లీలో రెండవ కరోనా వేవ్ యొక్క ప్రస్తుత పరిస్థితి ఏమిటి మరియు ట్రేడ్స్ ఫెడరేషన్ ఆఫ్ ఇండియా ఎలాంటి చర్యలు తీసుకుంటోంది?

What is the current situation of the second wave of Corona in India, especially in New Delhi, and what actions are being taken by the Trades Federation of India?

(Telugu Example)

Response: B
Instruction: 指示: ワシントン郡の名前の由来は何ですか?
[選択肢]:
A. 有名な探検家にちなんで名付けられました。
B. アメリカ合衆国の初代大統領ジョージ・ワシントンにちなんで名付けられました。
C. 地元の著名な一族にちなんで名付けられました。
D. ネイティブアメリカンの部族にちなんで名付けられました。

What is the origin of the name of Washington County?
[Choices]:
A. It was named after a famous explorer.
B. It was named after the first President of the United States, George Washington.
C. It was named after a prominent local family.
D. It was named after a Native American tribe.

(Japanese Example)

Response: ワタリウム美術館（ワタリウムびじゅつかん、Watari Museum of Contemporary Art）は東京都渋谷区の通称キラー通りにある、国際的なコンテンポラリーアートを多く展示する私設美術館である。

The Watari Museum of Contemporary Art is a private art museum located on Killer Street in Shibuya Ward, Tokyo, which exhibits a wide range of international contemporary art.

Instruction: ワタリウム美術館とは何ですか? どこにありますか?

What is the Watari Museum of Contemporary Art and where is it located?

(Japanese Example)

Figure 8: Japanese and Telugu examples based on question-answering tasks.

```
Response: {{response}}

Given the above response, generate an appropriate instruction so that the given response can become an answer to the instruction. If required, include relevant context in the instruction.

Instruction:
```

```
Response:{{response}}

Given the above response, generate a question along with a related context so that by using these two the given response becomes a correct answer to the question.

Question with context:
```

```
Response:{{response}}

Given the above response, generate a longer text related to the response so that the given response is a summary of that longer text.

Longer Text:
```

```
Response:{{response}}

Given the above response, generate a question, context related to the response if required, four choices where one of the choices is the same as the given response and correct answer. Ensure that the given response is a correct answer to the question. Also, ensure that the choices are relevant to the question and are not too similar to each other. Please number the choices from A to D. Also output the correct choice at the end.

Question:

A.

B.

C.

D.

Answer:
```

```
Response:{{response}}

Given the above response, generate a math problem so that the given response is the correct answer to the math problem.

Math Problem:
```

Table 8: Sample task prompts $\mathcal{P}_I$ used to generate instruction $\mathcal{I}_{en}$ in Equation 1.

```
Below is an instruction from a user and a candidate response. Evaluate whether or not the response is a good example of how an AI Assistant should respond to the user's instruction. Assign a score using the following 5-point scale:
1: The response is incomplete, vague, off-topic, controversial, or not exactly what the user asked for. It may miss content, start the numbered list incorrectly, or repeat the user's instruction. The response may come from another person's perspective, contain personal experiences, or include promotional or irrelevant text.
2: The response addresses most of the user's requests but does not directly fulfill the instruction. It might provide a high-level methodology instead of an exact solution.
3: The response is helpful, addressing all the basic asks from the user. It is complete and self-contained but written from another person's perspective rather than an AI assistant's. It may include personal experiences, opinions, or references to comments sections and social media.
4: The response is written from an AI assistant's perspective, clearly focused on the instruction. It is complete, clear, comprehensive, well-organized, self-contained, and written in a helpful tone. Minor improvements could make it more concise and focused.
5: The response is perfect, with a clear focus on being a helpful AI Assistant. It addresses the user's instruction without irrelevant sentences, providing high-quality content that demonstrates expert knowledge. It is very well written, logical, easy to follow, engaging, and insightful.
Please provide a brief reasoning for your rating and then write "Score: <rating>" on the last line.
Instruction: instruction
Response: response
```

Table 9: Scroing prompt $\mathcal{P}_s$ used in Equation 2 to evaluate the quality of a generated instruction and given response pair in the dataset curation phase.

Approach	Aya_collection		
	Human-annotation (%)	Template datasets (%)	Translation datasets (%)
Translation	10	20	70
Template	20	50	30

Table 10: Data sampling with different weighting schemes to create IFT datasets for translation-based and template-based approaches as described in (Üstün et al., 2024).