Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback

Viet Dac Lai[1], **Chien Van Nguyen**[1], **Nghia Trung Ngo**[1], **Thuat Nguyen**[1]
Franck Dernoncourt[2], **Ryan A. Rossi**[2], **Thien Huu Nguyen**[1]
[1]Dept. of Computer Science, University of Oregon, OR, USA
[2]Adobe Research, USA
{vietl@cs,chienn,nghian@cs,thien@cs}@uoregon.edu
{franck.dernoncourt,ryrossi}@adobe.com

Abstract

A key technology for large language models (LLMs) involves instruction tuning that helps align the models' responses with human expectations to realize impressive learning abilities. Two major approaches for instruction tuning characterize supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF), which are applied to produce the best commercial LLMs. To improve the accessibility of LLMs, various instruction-tuned open-source LLMs have also been introduced recently. However, existing open-source LLMs have only been instruction-tuned for English and a few popular languages, thus hindering their accessibility to many other languages in the world. In addition, SFT has been used as the only approach to instruction-tune open-source LLMs for multiple languages. This has left a significant gap for fine-tuned LLMs based on RLHF in diverse languages and raised important questions on how RLHF can boost the performance of multilingual instruction tuning. To overcome this issue, we present Okapi, the first system with instruction-tuned LLMs based on RLHF for multiple languages. Okapi introduces instruction and response-ranked data in 26 diverse languages to facilitate the experiments and development of future multilingual LLM research. We also present benchmark datasets to enable the evaluation of generative LLMs in multiple languages. Our experiments demonstrate the advantages of RLHF for multilingual instruction over SFT for different base models and datasets. Our framework with created resources, fine-tuned LLMs, interaction scripts are released at https://github.com/nlp-uoregon/Okapi. A demo video to show our framework can also be found at: https://youtu.be/QFV2fkPwvi0.

1 Introduction

Pre-trained on massive data, large language models (LLMs) with hundreds of billions of parameters such as GPT-3 (Rae et al., 2021) can unlock new emergent abilities that cannot be achieved with smaller models (Wei et al., 2022; Choi et al., 2023; Jiao et al., 2023). However, as LLMs are trained with the autoregressive learning objective, they might exhibit unintended behaviours from human expectations (Tamkin et al., 2021; Weidinger et al., 2021; Kenton et al., 2021). To overcome this issue, instruction fine-tuning has been proposed as a prominent approach to improve capabilities in following human instructions for LLMs and align them with human intentions in conversations (Christiano et al., 2017; Stiennon et al., 2020; Sanh et al., 2021; Ouyang et al., 2022). As such, two major techniques for instruction tuning feature supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) that are leveraged by the best commercial LLMs such as ChatGPT and GPT-4 to deliver outstanding dialog performance.

Another issue with LLMs pertains to the massive scales and closed-source nature of the commercial LLMs that greatly restrict accessibility and the extent of interactions with the technology. To this end, there have been growing efforts from the open-source community to create more accessible LLMs with affordable scales while securing competitive performance as the proprietary LLMs, e.g., LLaMA (Touvron et al., 2023), StableLM (StabilityAI, 2023), Falcon (Almazrouei et al., 2023), and MTP (MosaicML, 2023). Instruction tuning has also been applied to these open-source LLMs to improve their abilities to engage with human, and different instruction datasets have been collected to facilitate the process, e.g., Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), LaMini-LM (Wu et al., 2023), and Dolly (Conover et al., 2023).

However, the instruction-following abilities of existing open-source LLMs have been developed mainly for English and some popular languages (i.e., using instruction data for those languages), failing to support many other languages of the world to serve a broader population (Taori et al.,

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 318–327
December 6-10, 2023 ©2023 Association for Computational Linguistics

2023; Wu et al., 2023). To overcome this challenge, a few contemporary frameworks have explored instruction tuning of LLMs for multiple languages, i.e., Phoenix (Chen et al., 2023) and Bactrian-X (Li et al., 2023). However, their multilingual instruction tuning efforts are limited to only supervised fine-tuning, which is unable to examine reinforcement learning with human feedback (RLHF) to further boost the performance for multilingual LLMs.

To fill in this gap, our work aims to develop Okapi, an open-source framework with RLHF-based instruction-tuned LLMs for multiple languages to provide resources and shed light on their performance for multilingual LLM learning. Okapi will emphasize on less studied languages and open-source LLMs to better democratize the benefits of instruction-tuned LLMs. In particular, an example in the instruction datasets involves an instruction, an input text, and a desired response output/demonstration. In SFT, the pre-trained LLMs are fine-tuned over the instruction triples (*instruction, input, output*) via supervised learning to promote their alignment with human expectations. In RLHF, generated outputs from the SFT-tuned LLMs are first ranked to provide training signals for the reward functions. Afterward, the SFT-tuned models will be further optimized via reinforcement learning utilizing rewards from the trained reward models. As such, RLHF has been successfully employed to create effective commercial LLMs (e.g., InstructGPT, ChatGPT), owning to its ability to learn beyond positive examples associated with only desired demonstrations. By leveraging the reward models, RLHF can observe lower ranking scores for less accurate demonstrations to obtain richer training signals for LLMs. To our knowledge, Okapi is the first work to perform instruction tuning with RLHF for open-source LLMs over multiple languages.

To develop Okapi, we need to overcome the scarcity of instruction datasets in multiple languages to train and evaluate RLHF models. Motivated by the 52K instructions from Alpaca (Taori et al., 2023), we leverage Self-Instruct (Wang et al., 2023) to generate 106K additional instructions in English, introducing a larger dataset to facilitate RLHF evaluation. Afterward, we utilize ChatGPT to translate the instructions into a diverse set of 26 languages, including high-, medium-, and low-resource languages (e.g., Telugu, Ukrainian, Nepali, and Kannada) to offer comprehensive resources and insights for multilingual instruction-tuning. In addition, we introduce a translation-based prompt for ChatGPT to produce rankings for multiple responses of the same instructions from the LLMs, which will be used to train the reward models for RLHF experiments. Finally, we obtain the multilingual evaluation datasets for our fine-tuned LLMs by translating three benchmark datasets for LLMs in the widely-used HuggingFace Open LLM Leaderboard (HuggingFace, 2023; Gao et al., 2021) into 26 languages, i.e., ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and MMLU (Hendrycks et al., 2021).

Using BLOOM (Scao et al., 2022) and LLaMa (Touvron et al., 2023) as the base LLMs, our experiments illustrate that RLHF generally performs better than SFT for multilingual instruction tuning. We also highlight the greater challenges of low-resource languages for multilingual instruction-tuning of LLMs that can be focused in future research. Finally, we release our framework with the created resources and fine-tuned RLHF models. We also provide scripts to interact with our models at `https://github.com/nlp-uoregon/Okapi`.

2 Data Preparation

A key requirement for our development of instruction-tuned LLMs with RLHF involves instruction, ranking, and evaluation datasets in multiple languages. To this end, we perform a comprehensive data collection process to prepare necessary data for our multilingual framework Okapi in 26 languages, divided into four major steps: English instruction generation, instruction translation, ranking data production, and evaluation data creation.

English Instruction Generation: An instruction example to tune LLMs often has three components: an instruction to specify the task, an input text, and an associated output text (i.e., demonstration or label) (Ouyang et al., 2022). As such, current public instruction datasets for LLMs mainly cover English or some popular languages. Also, we note that a few recent instruction datasets such as xP3 (Muennighoff et al., 2022) and Flan (Chung et al., 2022; Longpre et al., 2023) include multilingual data; however, their instructions are still written in English. Additionally, these datasets tend to be converted from NLP datasets with template instructions, which cannot reflect the flexibility of human-written prompts (Wang et al., 2023). Consequently, our goal is to develop instruction datasets with instructions, inputs, and output texts in multiple languages, including low-resource ones, to better realize general prompts from human.

To achieve this goal, our strategy is to first obtain English instructions and then translate them into other languages. The benefits of our approach involve consistent instruction content across languages to facilitate performance comparison while taking advantages of translation systems to enable examination for more languages. As such, to conveniently scale our data, we follow the instruction generation method in Alpaca, which in turn employs the Self-Instruct procedure in (Wang et al., 2023), to produce our English dataset.

Starting with a pool of 175 human-written seed instructions in English, at each time, Alpaca samples several instructions from the seeds to form an in-context example to prompt the text-davinci-003 model of OpenAI for new instruction generation. Overall, Alpaca releases 52K instructions for tuning LLMs. In this work, we apply the same Self-Instruct procedure as Alpaca to generate 106K additional English instructions, resulting in a larger combined dataset of 158K instructions for our RLHF-based models in Okapi. Notably, we condition our generation process on the 52K instructions from Alpaca so a new instruction is only saved if it is different enough from Alpaca's and previous instructions per the ROUGE score criteria in Alpaca (Taori et al., 2023).

Instruction Translation: Given the 158K English instructions, we aim to translate them into multiple other languages to obtain data for our multilingual models in Okapi. Table 1 presents 26 selected languages in our framework. Using the data ratios r of the languages in CommonCrawl[1] to classify languages as in previous work (Bang et al., 2023; Lai et al., 2023), our study encompasses a diverse set of languages, including 8 high-resource languages ($r > 1.0$), 11 medium-resource languages ($r > 0.1$), and 7 low-resource languages ($r < 0.1$). Notably, several of our languages, such as Marathi, Gujarati, and Kannada, have received limited attention in NLP and instruction-tuning.

Language	Code	Pop.	CC Size		B	L
		(M)	(%)	Cat.		
English	en	1,452	45.8786	H	✓	✓
Russian	ru	258	5.9692	H	✓	✓
German	de	134	5.8811	H	✓	✓
Chinese	zh	1,118	4.8747	H	✓	
French	fr	274	4.7254	H	✓	✓
Spanish	es	548	4.4690	H	✓	✓
Italian	it	68	2.5712	H	✓	✓
Dutch	nl	30	2.0585	H	✓	✓
Vietnamese	vi	85	1.0299	H	✓	
Indonesian	id	199	0.7991	M	✓	
Arabic	ar	274	0.6658	M	✓	
Hungarian	hu	17	0.6093	M	✓	✓
Romanian	ro	29	0.5637	M	✓	✓
Danish	da	6	0.4301	M	✓	✓
Slovak	sk	7	0.3777	M	✓	✓
Ukrainian	uk	33	0.3304	M	✓	✓
Catalan	ca	10	0.2314	M	✓	✓
Serbian	sr	12	0.2205	M	✓	✓
Croatian	hr	14	0.1979	M	✓	✓
Hindi	hi	602	0.1588	M	✓	
Bengali	bn	272	0.0930	L	✓	
Tamil	ta	86	0.0446	L	✓	
Nepali	ne	25	0.0304	L	✓	
Malayalam	ml	36	0.0222	L	✓	
Marathi	mr	99	0.0213	L	✓	
Telugu	te	95	0.0183	L	✓	
Kannada	kn	64	0.0122	L	✓	

Table 1: List of 26 non-English languages in Okapi along with their codes, numbers of first and second speakers (the "*Pop.*" column), data ratios in CommonCrawl, and categories. The languages are grouped into categories based on their data ratios in CommomCrawl: High- (H, $> 1\%$), Medium- (M, $> 0.1\%$), and Low-Resource (L, $> 0.01\%$). Columns "*B*" and "*L*" indicate if a language is supported by the LLMs BLOOM and LLaMa (respectively) or not.

We utilize ChatGPT to translate the 158K English instructions into 26 target languages for Okapi. Compared to traditional machine translation systems, an advantage of ChatGPT is the ability to use prompts to specify different expectations for the translated texts to facilitate diverse types of instructions. For example, we can instruct ChatGPT to preserve code in the instruction examples about programming as we expect code to be the same in the instructions across natural languages. It is important to note that we directly translate the instruction, input text, and associated output in each English instruction of our data. This is in contrast to the other multilingual instruction-tuning approaches (Li et al., 2023) that only translate instructions and input texts into a target language (using Google Translate), and then prompt ChatGPT to generate response outputs in the target language based on the translated instructions and inputs. The intuition for our approach concerns various potential issues of ChatGPT, e.g., hallucination, bias, mathematical reasoning, and toxic content (Bang et al., 2023; Borji, 2023), that can be exaggerated if ChatGPT is used to produce responses in non-English languages for different tasks (Lai et al., 2023). By generating the instructions and responses in English, we aim to capitalize on the greater performance of LLMs for different

[1] http://commoncrawl.org

Translation Prompt: Translate the values in the following JSON object into <target language> language. You must keep the keys in the JSON object in English. If a value contains programming code, only translate the comments while preserving the code. Your translations must convey all the content in the original text and cannot involve explanations or other unnecessary information. Please ensure that the translated text is natural for native speakers with correct grammar and proper word choices. Your translation must also use exact terminology to provide accurate information even for the experts in the related fields. Your output must only contain a JSON object with translated text and cannot include explanations or other information.

Figure 1: Translation prompt for ChatGPT for multiple languages in Okapi. We organize our instruction examples into JSON objects with fields for translation prompts, instructions, inputs, and outputs send to ChatGPT. <target language> is replaced with the selected languages in our dataset.

NLP tasks in English to avoid the exaggeration issues and achieve higher quality instructions.

Ranking Data Production: To perform RLHF for a LLM, we need to obtain ranked response outputs from the model for the same instruction and input to train a reward model. Concretely, given a LLM M and a dataset $S = \{inst_k, input_k\}_{k=1}^{N}$ with N pairs of instructions $inst_k$ and input texts $input_k$ for a target language, we first prompt M to generate T output responses $output_k = \{output_k^1, \ldots, output_k^T\}$ for each pair of instruction and input text $(inst_k, input_k)$ $(T > 1)$. Afterward, the responses in $output_k$ are ranked according to their fitness and quality for the instruction $inst_k$ and input text $input_k$. This ranking data $\{inst_k, input_k, output_k\}$ can then be leveraged to train our reward models in Okapi.

We also employ ChatGPT to rank the response outputs for multilingual LLMs. Similar to the motivation for our translation-based approach to obtain instruction data in multiple languages, our ranking strategy first asks ChatGPT to translate the instructions and responses $\{inst_k, input_k, output_k\}$ of a target language into English; the ranking of the responses is then done over the translated English data to exploit the greater quality of ChatGPT for English (using the translation and ranking prompts in Figure 2). For each example $\{inst_k, input_k, output_k\}$, the translation and ranking prompts are wrapped in a two-turn dialog with ChatGPT to allow the ranking process to condition on the resulting translations. It also ensures the same output format for the ranking prompts for convenient parsing. Overall, we obtain ranked response outputs for 42K instructions from the 106K generated instructions for each language in Okapi.

•**Turn 1: Translation Prompt** You will be given an instruction, an input for the instruction, and four possible responses for the instruction. The input can be empty, shown as <empty>. You need to translate the provided instruction, input, and responses into English.
Instruction: . . .
Input: . . .
Response 1: . . .
Response 2: . . .
Response 3: . . .
Response 4: . . .

• **Turn 2: Ranking Prompt** Given the translated instruction, input, and responses, you will need to rank the responses according to three factors: correctness with respect to the instruction and input, coherence, and naturalness.
You will need to provide an overall rank for each response when all the three factors are considered. The overall rank for a response must be an integer between 1 and 4 where 1 is for the best response and 4 is the worst response. You cannot assign the same rank for two different responses.
The format of your output must be: for each response: "<Response r>: overall rank: <1/2/3/4>". The responses must be in original order. Do not include explanation in your output.

An Example Output from ChatGPT:
Response 1: 3
Response 2: 1
Response 3: 4
Response 4: 2

Figure 2: Prompts to translate and rank responses.

Evaluation Data Creation: We employ three datasets in the HuggingFace Open LLM Leaderboard (HuggingFace, 2023) i.e., ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and MMLU (Hendrycks et al., 2021), to evaluate the model performance for our Okapi framework. All the datasets are organized as multiple-choice question-answering tasks although they focus on different types of knowledge and reasoning aspects. ARC involves 1170 grade-school science questions; HellaSwag provides 9162 commonsense inference questions that are easy for humans, but difficult for many state-of-the-art models; and MMLU assesses accuracy for 13062 questions over various branches of knowledge (STEM, humanities, social sciences, and more). Nevertheless, although the LLM community has widely adopted the HuggingFace leaderboard for performance examination, the datasets are only provided for English, thus unable to evaluate LLMs for the languages in our work. To this end, we translate the examples of the three datasets into 26 selected languages using ChatGPT and the translation prompt in Figure 1. The translated datasets are then reserved to evaluate the LLMs in our Okapi framework.

3 Instruction-tuning with RLHF

We follow three steps to develop a fine-tuned LLM with RLHF for each target language in our Okapi framework: supervised fine-tuning, reward model training, and reinforcement learning.

Supervised Fine-tuning (SFT): Starting with a multilingual LLM as the base, e.g., BLOOM (Scao et al., 2022), we fine-tune the model with our instruction dataset for the target language using supervised learning with the autoregressive objective. Here, we fine-tune the entire base LLM for all of its parameters with SFT to accurately understand the model performance for multilingual settings.

Reward Model Training: The goal of this step is to train a reward model for the target language that will compute reward signals for reinforcement learning to further optimize the SFT-tuned model from the previous step. For each pair of a prompt and potential response, our reward model returns a scalar value to quantify the appropriateness of the response with respect to the instruction and input text in the prompt. We exploit the instructions with multiple ranked responses in the data collection step for this training step. An example to train our reward model for a language involves an instruction and an input text (to form a prompt x) along with two sampled responses y_c and y_r for x from our datasets. Based on the ranking information, we can assume one of the responses (i.e., y_c) is more preferable than the other (i.e., y_r). In the next step, the binary ranking loss (Ouyang et al., 2022) is employed to train our reward model, aiming to assign a higher score $r(x, y_c)$ for the preferred response y_c than the score $r(x, y_r)$ for y_r: $L_{reward}(\theta) = -\mathbb{E}_{(x,y_c,y_r)}\left[\log \sigma(r_\theta(x, y_c) - r_\theta(x, y_r))\right]$.

Reinforcement Learning (RL): With the reward model established for the target language, the SFT model undergoes additional fine-tuning through RL to align it with human preferences. For this purpose, we employ the Proximal Policy Optimization (PPO) algorithm (Ouyang et al., 2022) that maximizes the mean reward of the model via the objective: $L_{RL}(\phi) = -\mathbb{E}_{x\sim D_{RL}, y\sim\pi_\phi(y|x)}\left[r_\theta(x, y) - \beta KL(x, y)\right]$. Here, D_{RL} corresponds to the prompt distribution, and $\pi_\phi(y|x)$ denotes the policy or language model that requires optimization. $\pi_\phi(y|x)$ is initialized with the SFT-tuned model $\pi_\phi(y|x)$. Also, $KL(x, y) = D_{KL}(\pi_\phi(y|x)||\pi_0(y|x))$ is the Kullback–Leibler divergence to penalize large deviation of π_ϕ from the initial SFT policy π_0.

4 Experiments

Our Okapi framework utilizes two multilingual LLMs: BLOOM (Scao et al., 2022) and LLaMA (Touvron et al., 2023) as the base models for the fine-tuning processes. We focus on their 7B-parameter versions to facilitate the computing resources and achieve fairer comparison. For each base model and target language, we carry out both SFT-based and RLHF-based instruction-tuning:

- SFT: The base model is fine-tuned over our entire set of 158K translated instructions for the target language in the supervised manner.
- RLHF: The base model is first fine-tuned with supervised training over 52K translated instructions from Alpaca. Afterward, a reward model is trained using the 42K instructions with ranked responses obtained in the data collection. Note that the ranked responses are sampled from the SFT-tuned base model over the 52K Alpaca instructions from previous step. Finally, given the reward model, the SFT-tuned model is further optimized via reinforcement learning over the 64K remaining translated instructions from our generation set.

Following the HuggingFace Open LLM Leaderboard, the Eleuther AI Language Model Evaluation Harness framework (Gao et al., 2021) is used to compute the model performance over the translated datasets ARC, HellaSwag, and MMLU for each language in our framework. As a reference, we also report the performance of the base models BLOOM and LLaMA in the experiments. Finally, for BLOOM, we further compare with BLOOMZ (Muennighoff et al., 2022), which is the fine-tuned version of BLOOM over the cross-lingual task mixture dataset xP3 with millions of multilingual instructions to achieve instruction-following ability.

Evaluation: Tables 2 and 3 present the performance of the models on ARC, HellaSwag, and MMLU when BLOOM and LLaMa are used as the base models (respectively). In the tables, for each language group (i.e., high-, medium-, and low-resource), we report the average performance over the languages and the performance for two example languages in the group. We also include the average performance over all languages in Okapi. As some of our languages in Okapi (especially the low-resource ones) are not supported by LLaMA, Table 3 will omit those languages (see Table 1). Finally, Appendix A provides performance of the models over all languages and datasets in Okapi.

The first observation from the tables is that

Data	Language	BLOOM	BLOOMZ	SFT	RLHF
ARC	Chinese	37.3	37.0	37.9	**40.0**
	French	36.7	37.6	37.6	**41.2**
	Average High	31.5	30.7	32.3	**34.0**
	Indonesian	36.0	35.9	37.4	**38.8**
	Arabic	31.4	31.2	32.1	**33.2**
	Average Medium	27.7	26.7	28.0	**29.8**
	Bengali	26.2	25.5	26.8	**28.9**
	Kannada	**24.7**	24.6	24.5	24.6
	Average Low	25.1	24.9	24.7	**25.6**
	Average All	28.2	27.4	28.4	**30.0**
HellaSwag	Chinese	51.2	42.6	51.8	**53.8**
	French	56.6	45.7	55.9	**58.7**
	Average High	43.8	39.6	44.5	**46.6**
	Indonesian	49.5	42.0	50.0	**52.2**
	Arabic	43.3	39.5	44.3	**47.0**
	Average Medium	35.7	33.5	36.9	**38.9**
	Bengali	32.8	31.5	33.9	**35.4**
	Kannada	30.3	30.9	30.7	**32.1**
	Average Low	30.3	30.9	31.2	**32.3**
	Average All	36.8	34.7	37.7	**39.5**
MMLU	Chinese	**29.1**	27.2	27.7	28.2
	French	27.4	27.7	27.7	**28.4**
	Average High	**27.5**	26.4	26.9	**27.5**
	Indonesian	26.9	26.3	26.8	**27.5**
	Arabic	27.5	24.4	27.4	**27.7**
	Average Medium	**27.1**	25.8	26.7	**27.1**
	Bengali	**28.2**	25.9	27.1	26.8
	Kannada	26.7	26.0	26.6	**26.8**
	Average Low	**26.7**	25.9	26.1	26.1
	Average All	**27.1**	26.0	26.6	26.9

Table 2: Performance of the models using BLOOM 7B.

Data	Language	LLaMA	SFT	RLHF
ARC	German	35.1	37.5	**39.7**
	French	37.3	38.4	**38.8**
	Average High	35.1	36.5	**38.7**
	Danish	32.7	35.1	**36.8**
	Ukrainian	32.9	35.7	**36.4**
	Average Medium	32.0	34.3	**36.2**
	Average All	33.3	35.2	**37.3**
HellaSwag	German	49.9	49.0	**52.6**
	French	55.7	55.6	**56.9**
	Average High	51.4	51.2	**53.7**
	Danish	46.7	47.7	**51.7**
	Ukrainian	44.1	46.9	**47.7**
	Average Medium	42.7	44.0	**46.5**
	Average All	46.4	47.1	**49.6**
MMLU	German	29.9	30.4	**31.7**
	French	30.5	**31.0**	30.7
	Average High	30.1	30.4	**30.9**
	Danish	30.0	30.9	**31.8**
	Ukrainian	29.4	30.8	**31.6**
	Average Medium	29.5	29.9	**30.7**
	Average All	29.8	30.1	**30.8**

Table 3: Performance of the models using LLaMa 7B.

RLHF is generally better than SFT for multilingual fine-tuning of LLMs over different datasets, base models, and language groups. It is also evident that the RLHF-tuned models can significantly improve the performance of the original base models (i.e., BLOOM and LLaMa) for almost all the language groups and datasets. In all, it highlights the quality of the generated instruction data and the effectiveness of RLHF in Okapi.

Comparing the performance across language groups, the models tend to achieve the highest performance for the high-resource languages, followed by the medium-resource and low-resource languages. The performance improvement of RLHF for low-resource languages is also the least (based on BLOOM). Interestingly, our fine-tuned BLOOM models with 158K generated instructions can significantly outperform BLOOMZ over almost all the languages for the ARC, HellaSwag, and MMLU datasets using either SFT or RLHF. As BLOOMZ has fine-tuned BLOOM over more than 78M multilingual instructions converted from NLP datasets (Muennighoff et al., 2022), it demonstrates the higher quality of our generated instructions for multilingual instruction tuning of LLMs.

5 Related Work

The most advanced methods for NLP involve fine-tuning the pre-trained language models (PLMs) on training data of the downstream tasks (Min et al., 2023). Instruction tuning can be considered as a special type of fine-tuning techniques for PLMs where generative PLMs (e.g., GPT) are further trained with instruction data to accomplish the instruction following abilities. SFT is the most popular instruction tuning approach that is leveraged by most of the existing LLMs, including ChatGPT, Apaca (Taori et al., 2023), and Vicuna (Chiang et al., 2023). RLHF can also be used to further enhance LLMs (Wei et al., 2021; Ouyang et al., 2022) although it has been less explored by current open-source LLMs due to the challenges in obtaining ranking data for the reward models. For multilingual learning, instruction tuning is only applied in the form of SFT for non-English languages using multilingual LLMs, e.g., BLOOM and LLaMA, in a few contemporary work (Chen et al., 2023; Li et al., 2023; Muennighoff et al., 2022).

6 Conclusion

We present the first framework, called Okapi, on instruction tuning for LLMs in multiple language using RLHF. We introduce instruction, ranked response, and evaluation data in 26 diverse languages to enable the training of RLHF methods. Our results reveal the benefits of RLHF for multilingual fine-tuning of LLMs and the challenging problems of low-resource languages in this area.

Acknowledgement

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112, the NSF grant CNS-1747798 to the IUCRC Center for Big Learning, and the NSF grant # 2239570.

Ethical Statement

Our framework utilizes the multilingual LLMs BLOOM-7B and LLaMa-7B to develop instruction-tuned models with reinforcement learning from human feedback. To obtain necessary resources to train and evaluation our models, we also apply Self-Instruct (Taori et al., 2023) with GPT-3 to generate English instruction data, and ChatGPT to translate and rank our response data in different languages. As such, the models in our framework might inherit potential issues in the underlying models of BLOOM, LLaMa, GPT-3, and ChatGPT, such as hallucination, biases, and toxic content. Regrettably, the data required to train such LLMs, even in the case of purportedly open-source models such as LLaMa and BLOOM, remains unreleased to enable essential investigation into these matters for our models. Future research can explore open-source datasets, such as CulturaX (Nguyen et al., 2023) and RedPajama (Computer, 2023), to develop truly open LLMs, enabling deeper attribution of the problems and better understanding of the models' operations. To maximally minimize the impacts of these issues in the current work, our framework will fully release the generated instruction, ranking, and evaluation data to enable comprehensive exploration and research for the techniques. We will also restrict the release of our models to research purpose, respecting the policy of the underlying models such as LLaMa and ChatGPT, to facilitate future research for LLMs while limiting the potential ethical issues for the society. Consequently, we do not believe our framework poses any greater societal risks than existing published research in this area for LLMs (Wang et al., 2023). Finally, we confirm that our work fully complies with the ACL Ethnics Policy and there is no other ethical issues associated with our work, to the best of our knowledge.

References

Ebtesam Almazrouei, Hamza Alobeidli, and Abdulaziz Alshamsi et al. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*, abs/2302.04023.

Ali Borji. 2023. A categorical archive of chatgpt failures. *ArXiv*, abs/2302.03494.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Phoenix: Democratizing chatgpt across languages. *ArXiv*, abs/2304.10453.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jonathan Choi, Kristin Hickman, Amy Monahan, and Daniel Schwarcz. 2023. Chatgpt goes to law school. *Available at SSRN*.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Hyung Won Chung, Le Hou, S. Longpre, and Barret Zoph et al. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Mike Conover, Matt Hayes, and Ankit Mathur et al. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. `https://www.databricks.com`.

Leo Gao, Jonathan Tow, Stella Biderman, and et al. 2021. A framework for few-shot language model evaluation.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

HuggingFace. 2023. Open llm leaderboard. `https://github.com/tatsu-lab/stanford_alpaca`.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *ArXiv*, 2301.08745.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *ArXiv*, abs/2103.14659.

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

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*, abs/2305.15011.

S. Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. *ArXiv*, abs/2301.13688.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Survey*.

MosaicML. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. https://www.mosaicml.com/blog/mpt-7b.

Niklas Muennighoff, Thomas Wang, and Lintang Sutawika et al. 2022. Crosslingual generalization through multitask finetuning. *ArXiv*, abs/2211.01786.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Jack Rae, Sebastian Borgeaud, and et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446.

Victor Sanh, Albert Webson, and Colin Raffel et al. 2021. Multitask prompted training enables zero-shot task generalization. *ArXiv*, abs/2110.08207.

Teven Scao, Angela Fan, and et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

StabilityAI. 2023. Stablelm: Stability ai language models. https://github.com/stability-AI/stableLM.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan J. Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *ArXiv*, abs/2009.01325.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *ArXiv*, abs/2102.02503.

Rohan Taori, Ishaan Gulrajani, and Tianyi Zhang et al. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, and Gautier Izacard et al. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

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. 2021. Finetuned language models are zero-shot learners. In *Proceedings of the International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Laura Weidinger, John F. J. Mellor, and Maribeth Rauh et al. 2021. Ethical and social risks of harm from language models. *ArXiv*, abs/2112.04359.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *ArXiv*, abs/2304.14402.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Model Performance

Tables 4, 5, and 6 present the performance of the models on the ARC, HellaSwag, and MMLU datasets (respectively) across all languages when BLOOM is used as the base model. Similarly, Tables 7, 8, and 9 report the performance with the base model LLaMA over the three datasets. In the tables, in addition to the average scores over all languages for the models, we also include the average scores for each group of languages (i.e., rows "*Ave Group*" for high-, medium-, and low-resource languages) to facilitate the comparisons.

	Language	BLOOM	BLOOMZ	SFT	RLHF
High-Resource	Russian	27.5	25.5	29.2	30.3
	German	26.3	25.4	24.9	25.5
	Chinese	37.3	37.0	37.9	40.0
	French	36.7	37.6	37.6	41.2
	Spanish	38.1	37.2	39.7	41.5
	Italian	29.0	27.5	29.3	31.3
	Dutch	23.1	21.5	24.8	26.1
	Vietnamese	33.7	33.5	35.0	36.2
	Ave Group	31.5	30.7	32.3	**34.0**
Medium-Resource	Indonesian	36.0	35.9	37.4	38.8
	Arabic	31.4	31.2	32.1	33.2
	Hungarian	25.9	22.8	25.2	27.5
	Romanian	26.9	23.4	27.5	30.3
	Danish	24.6	24.6	23.6	25.2
	Slovak	24.9	22.5	26.2	27.3
	Ukrainian	22.8	23.1	23.6	25.2
	Catalan	34.7	35.8	35.1	38.9
	Serbian	25.1	23.6	25.6	27.8
	Croatian	23.7	22.8	22.7	24.1
	Hindi	29.2	28.2	28.5	29.6
	Ave Group	27.7	26.7	28.0	**29.8**
Low-Resource	Bengali	26.2	25.5	26.8	28.9
	Tamil	24.2	25.6	23.7	25.1
	Nepali	22.3	22.7	23.4	25.7
	Malayalam	26.4	25.1	24.6	24.7
	Marathi	27.3	24.8	25.8	26.0
	Telugu	24.3	25.8	23.9	24.5
	Kannada	24.7	24.6	24.5	24.6
	Ave Group	25.1	24.9	24.7	**25.6**
	Average	28.2	27.4	28.4	**30.0**

Table 4: Performance of the models on the translated ARC dataset over different languages in Okapi. BLOOM 7B is used as the base LLM.

	Language	BLOOM	BLOOMZ	SFT	RLHF
High-Resource	Russian	32.5	33.1	32.9	34.2
	German	32.4	33.1	34.7	35.9
	Chinese	51.2	42.6	51.8	53.8
	French	56.6	45.7	55.9	58.7
	Spanish	56.7	48.7	56.1	59.0
	Italian	40.8	40.3	43.1	44.6
	Dutch	31.7	32.3	32.6	34.9
	Vietnamese	48.3	40.6	49.0	51.3
	Ave Group	43.8	39.6	44.5	**46.6**
Medium-Resource	Indonesian	49.5	42.0	50.0	52.2
	Arabic	43.3	39.5	44.3	47.0
	Hungarian	30.1	29.8	30.8	32.7
	Romanian	31.8	32.3	33.1	35.2
	Danish	31.2	31.5	33.8	35.7
	Slovak	29.8	29.6	31.4	32.9
	Ukrainian	30.0	30.4	32.2	33.6
	Catalan	51.2	40.3	50.9	53.8
	Serbian	29.9	30.1	30.7	33.7
	Croatian	30.0	29.4	30.5	31.6
	Hindi	36.4	34.0	37.7	39.7
	Ave Group	35.7	33.5	36.9	**38.9**
Low-Resource	Bengali	32.8	31.5	33.9	35.4
	Tamil	29.4	29.5	30.0	30.4
	Nepali	30.9	31.9	32.5	34.1
	Malayalam	28.8	29.8	29.7	30.2
	Marathi	31.0	31.9	31.7	32.5
	Telugu	29.2	30.7	30.0	31.7
	Kannada	30.3	30.9	30.7	32.1
	Ave Group	30.3	30.9	31.2	**32.3**
	Average	36.8	34.7	37.7	**39.5**

Table 5: Performance of the models on the translated HellaSwag dataset over different languages in Okapi. BLOOM 7B is used as the base LLM.

	Language	**BLOOM**	**BLOOMZ**	**SFT**	**RLHF**
High-Resource	Russian	26.2	25.4	26.5	26.8
	German	28.1	25.6	27.0	28.6
	Chinese	29.1	27.2	27.7	28.2
	French	27.4	27.7	27.7	28.4
	Spanish	28.9	27.1	27.8	28.1
	Italian	25.7	25.8	25.1	26.0
	Dutch	26.4	26.0	26.1	26.0
	Vietnamese	28.1	26.3	27.0	27.5
	Ave Group	**27.5**	26.4	26.9	**27.5**
Medium-Resource	Indonesian	26.9	26.3	26.8	27.5
	Arabic	27.5	24.4	27.4	27.7
	Hungarian	26.9	26.1	25.4	26.3
	Romanian	27.4	25.9	27.6	27.4
	Danish	27.1	25.2	27.2	26.9
	Slovak	26.1	26.3	26.4	26.1
	Ukrainian	26.6	25.8	25.9	26.4
	Catalan	28.8	26.0	26.7	27.6
	Serbian	27.2	25.7	27.5	27.6
	Croatian	26.0	26.1	26.4	27.7
	Hindi	27.5	25.9	26.8	26.5
	Ave Group	**27.1**	25.8	26.7	**27.1**
Low-Resource	Bengali	28.2	25.9	27.1	26.8
	Tamil	26.6	26.7	26.1	26.0
	Nepali	26.6	25.6	25.5	25.2
	Malayalam	26.4	25.2	25.8	25.8
	Marathi	26.3	26.0	26.1	26.1
	Telugu	26.2	25.7	25.4	25.9
	Kannada	26.7	26.0	26.6	26.8
	Ave Group	**26.7**	25.9	26.1	26.1
	Average	**27.1**	26.0	26.6	26.9

Table 6: Performance of the models on the translated MMLU dataset over different languages in Okapi. BLOOM 7B is used as the base LLM.

	Language	**LLaMA**	**SFT**	**RLHF**
High-Resource	Russian	32.1	32.8	37.7
	German	35.1	37.5	39.7
	French	37.3	38.4	38.8
	Spanish	36.8	38.7	39.3
	Italian	35.8	36.3	39.4
	Dutch	33.6	35.2	37.5
	Ave Group	35.1	36.5	**38.7**
Medium-Resource	Hungarian	29.8	31.4	33.2
	Romanian	32.4	33.8	37.5
	Danish	32.7	35.1	36.8
	Slovak	29.0	34.3	37.2
	Ukrainian	32.9	35.7	36.4
	Catalan	35.1	36.8	36.9
	Serbian	30.8	33.5	35.8
	Croatian	33.0	33.8	35.9
	Ave Group	32.0	34.3	**36.2**
	Average	33.3	35.2	**37.3**

Table 7: Performance of the models on the translated ARC dataset over different languages in Okapi. LLaMA 7B is used as the base LLM.

	Language	**LLaMA**	**SFT**	**RLHF**
High-Resource	Russian	45.7	46.0	49.1
	German	49.9	49.0	52.6
	French	55.7	55.6	56.9
	Spanish	56.4	55.7	56.6
	Italian	52.0	52.5	55.9
	Dutch	48.7	48.1	51.3
	Ave Group	51.4	51.2	**53.7**
Medium-Resource	Hungarian	37.9	38.7	41.0
	Romanian	44.9	45.1	48.7
	Danish	46.7	47.7	51.7
	Slovak	35.9	39.5	43.6
	Ukrainian	44.1	46.9	47.7
	Catalan	49.6	49.2	49.0
	Serbian	41.1	42.6	45.0
	Croatian	41.1	42.4	45.2
	Ave Group	42.7	44.0	**46.5**
	Average	46.4	47.1	**49.6**

Table 8: Performance of the models on the translated HellaSwag dataset over different languages in Okapi. LLaMA 7B is used as the base LLM.

	Language	**LLaMA**	**SFT**	**RLHF**
High-Resource	Russian	30.2	30.0	30.6
	German	29.9	30.4	31.7
	French	30.5	31.0	30.7
	Spanish	30.3	30.4	30.9
	Italian	29.9	30.6	30.4
	Dutch	29.8	30.0	31.1
	Ave Group	30.1	30.4	**30.9**
Medium-Resource	Hungarian	29.0	29.2	30.1
	Romanian	29.7	29.8	30.9
	Danish	30.0	30.9	31.8
	Slovak	29.4	29.6	30.2
	Ukrainian	29.4	30.8	31.6
	Catalan	30.2	30.3	30.5
	Serbian	29.2	29.7	30.4
	Croatian	29.3	29.2	30.0
	Ave Group	29.5	29.9	**30.7**
	Average	29.8	30.1	**30.8**

Table 9: Performance of the models on the translated MMLU dataset over different languages in Okapi. LLaMA 7B is used as the base LLM.