LLMs Beyond English: Scaling the Multilingual Capability of LLMs with Cross-Lingual Feedback

Wen Lai[1,2], Mohsen Mesgar[3], Alexander Fraser[1,2]

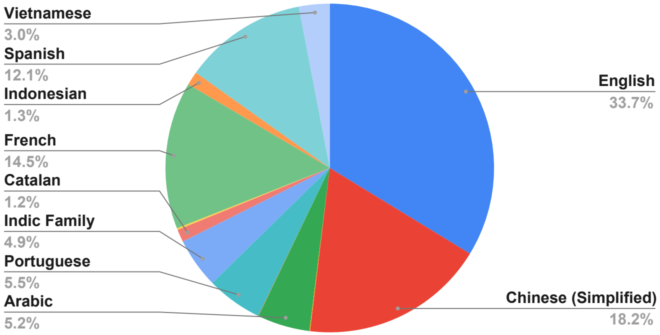
[1]School of CIT, Technical University of Munich, Germany
[2]Munich Center for Machine Learning, Germany
[3]Bosch Center for Artificial Intelligence, Renningen, Germany

12th August, 2024

Munich Center for Machine Learning

Background (I)

- Language Distribution in BLOOM



- Most of the languages in the training corpus are English, with only a few in other languages.

Background (II)

- Language Proportion in LLMs

Model	Language	Language Proportion
GPT-3	95 languages	English (92.7%); French (1.8%); German (1.5%); Others (5.9%)
chatGPT	58 languages	English (50-60%); Spanish, French, German, Chinese, Japanese, Portuguese, Russian, Italian, Korean (2-5%); Others (10-20%)
BLOOM	46 languages	English (30.03%); Simplified Chinese (16.16%); French (12.9%); Spanish (10.85%); Portuguese (4.91%); Arabic (4.6%); Others (20.55%)
LLaMA	over 30 language	English (mostly); 30+ languages (5%); Others (unknown)
Aya[1]	101 languages	High-Resource (29.1%), Mid-Resource (14.7%), Low-Resource (56.2%)

- Existing LLMs support only limited number of language.

[1] Üstün et al., 2022; Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model

Motivation

- Multilingual capability of LLMs
  - A LLM should be able to understand and generate a text in multiple natural languages.
- We consider two types of multilingual capability in LLMs:
  - **Understanding Capability:** when the instructions for LLMs are expressed in different languages, LLMs should understand these instructions and generate a correct output.
  - **Generating Capability:** LLMs should be able to generate the correct response in the target language and perform consistently well on (almost) all languages when a fixed language (e.g., English) is used as the instruction language.

	Language of inputs and outputs						
LLM	**En**	**Zh**	**Vi**	**Tr**	**Ar**	**El**	**Hi**
Understanding capability (Instruction identical to inputs)							
ChatGPT	**56.0**	20.5	26.8	18.3	24.1	17.7	0.6
LLaMA	**76.6**	27.2	36.6	27.8	11.8	22.3	14.3
BLOOM	**83.9**	83.0	79.9	27.4	79.2	22.8	82.7
Generation capability (Instructions in English)							
ChatGPT	**56.0**	37.1	36.1	34.5	32.0	29.7	17.5
LLaMA	**76.6**	66.3	42.9	38.1	24.2	40.7	30.8
BLOOM	**83.9**	81.8	79.2	27.6	77.2	49.2	80.8

Table 1 A primary evaluation for the multilingual capability (understanding and generation) of LLMs (ChatGPT, LLaMA and BLOOM) on the XQuAD dataset in terms of exact match (EM).

Related Work

- Continue training using more data
  - Yang et al., 2023; Zhu et al., 2023; Lai et al., 2023; Li et al., 2023; Luo et al., 2023; Groebeveld et al., 2024; Üstün et al., 2024
- Align non-English instructions with English instructions through cross-lingual prompting in the inference stage.
  - Huang et al., 2023; Etxaniz et al., 2023

Our Goal

- We aim to scale the two multilingual capabilities of LLMs at the same time by:
  - Constructing a multilingual instruction dataset with 100 languages by translating instructions from Alpaca (Taori et al., 2023) via ChatGPT [2] and Google Translate API [3]. **[Improving Understanding Capability]**
  - Constructing a cross-lingual human feedback data (i.e., instruction and output are different languages) covering 30 languages. **[Improving Generating Capability]**

[2] https://chat.openai.com/

[3] https://translate.google.com/

Multilingual Instruction Dataset

- Languages Coverage
  - 100 languages covered in FLORES-101 dataset (Goyal et al., 2022).
- Two Components
  - Instruction Translation
    - Google Translate API
    - NLLB model (Costa-jussà et al., 2022) for languages not supported in Google Translate API.
  - Hybrid Response Generation
    - Translation-based Approach: Languages with poor translation quality have their responses translated by Google Translate API or NLLB model.
    - Generation-based Approach: Languages with good translation quality have their responses generated by ChatGPT.

Cross-Lingual Human Feedback Dataset

- **Traditional Human Feedback**
  - Instructions, inputs and outputs are in the same language.
- **Cross-Lingual Human Feedback**
  - Instructions and outputs are in different languages.
  - Cover 30 languages, which simulate up to $30 \times 29 = 870$ generation scenarios.

Traditional Human Feedback

Given two texts, modify Text 1 to make it more concise and follow the structure of Text 2.

Text 1: I work for a small company with less than ten people. I was able to learn a lot on the job and I have obtained the skills I need to be a successful professional.
Text 2: My current job is in a very small team. I have gained a very good set of skills that will help me move forward in my career.

I worked on a small team, less than 10 people, and gained valuable skills to be successful professionally.

I'm very lucky to have found a job with a micro team of less than 10 people. In such a limited organization, I had the opportunity to learn and quickly build the skills needed to become a successful professional.

Cross-Lingual Human Feedback

Given two texts, modify Text 1 to make it more concise and follow the structure of Text 2. Please return the results in Chinese.

Text 1: I work for a small company with less than ten people. I was able to learn a lot on the job and I have obtained the skills I need to be a successful professional.
Text 2: My current job is in a very small team. I have gained a very good set of skills that will help me move forward in my career.

我在一个小团队工作，不到10人，获得了有价值的技能，可以在职业上获得成功。

我非常庆幸能找到一份只有10人不到的微型团队的工作。在这么有限的组织中，我有机会学习并且快速建立成为成功专业人士所需的技能。

Instruction and Response Quality

- **Instruction Quality**

	BLEU	COMET
[0,10)	2	0
[10,20)	7	0
[20,30)	15	0
[30,40)	18	0
[40,50)	26	3
[50,60)	16	8
(60,70]	9	19
(70,80]	5	29
(80,90]	2	32
(90,100]	0	9

Table 2 The number of languages for BLEU and COMET scores fall within each interval, obtained by back-translating from 100 languages into English.

- **Response Quality**

High			Low		
	BLEU	CP		BLEU	CP
Arabic	73.16	0.82	Armenian	47.16	0.64
Chinese	80.27	0.91	Gujarati	39.68	0.55
French	77.71	0.85	Kannada	41.72	0.57
German	75.50	0.84	Malayalam	45.24	0.62
Hindi	73.26	0.81	Marathi	41.37	0.56
Avg.	75.98	0.85	**Avg.**	43.03	0.59

Table 3 BLEU and content preservation (CP) of the response quality for 5 high-resource laguages and 5 low-resource languages.

Multilingual Instruction Tuning

- **Supervised Fine-tuning (SFT)**
  - We perform supervised fine-tuning on the LLM (e.g., LLaMA-2 and BLOOM) using our constructed multilingual instruction dataset.
- **Aligning LLMs with human feedback**
  - We further fine-tune the trained SFT model from the last step with the DPO algorithm (Rafailov et al., 2023) with our constructed cross-linguistic human feedback dataset.

Datasets and Tasks

- We evaluate xLLMs-100 on 5 typical benchmarks
  - **Understanding Task:** PAWS-X (Yang et al., 2019)
  - **Generation Task:** FLORES-101 (Goyal et al., 2022) and XL-Sum (Hasan et al., 2021).
  - **Reasoning Task:** XCOPA (Ponti et al., 2020)
  - **Expert-written Task:** Self-Instruct*, we use the Google Translate API to translate the Self-Instruct (Wang et al., 2023) benchmark from English to five high-resource languages (Arabic, Czech, German, Chinese, Hindi) and five low-resource languages (Armenian, Kyrgyz, Yoruba, Tamil, Mongolian).

Baselines

- **Off-the-shelf LLMs**
  - We evaluate LLaMA-2 (Touvron et al., 2023) and BLOOM (BigScience et al., 2022) as vanilla LLM baselines without additional finetuning.
- **Publicly available multilingual instruction-tuned models**
  - Bactrian-X is the instruction-tuned model proposed by Li et al., (2023). These models were instruction-tuned on 52 languages. They released models based on LLaMA and BLOOM. We refer to them as BX_{LLaMA} and BX_{BLOOM}, respectively.
- **Supervised Fine-Tuning (SFT)**
  - We performed instruction tuning by utilizing our constructed multilingual instruction dataset. We denote these models as $\text{SFT}_{\text{LLaMA}}$ and $\text{SFT}_{\text{BLOOM}}$.

Evaluation Metrics

- For FLORES-101, we report case-sensitive detokenized BLEU with SacreBLEU[4] (Post et al., 2018).
- For the XCOPA and PAWS-X benchmarks, we utilize the accuracy score for evaluation.
- For the XL-Sum and Self-Instruct* benchmark, we report the multilingual ROUGE-1 score implemented by Lin (2004).

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

Main Results

Understanding Capabilities												
	PAWS-X	XCOPA		Self-Instruct*		XL-Sum			FLORES(f)		FLORES(t)	
		low	high	low	high	low	mid	high	low	high	low	high
LLaMA	38.10	47.44	47.22	7.09	12.57	4.07	5.44	2.84	3.07	4.95	2.96	6.61
BX_{LLaMA}	37.28	49.53	49.00	6.31	11.88	2.17	5.52	7.89	2.69	2.38	3.15	5.31
SFT_{LLaMA}	42.32	50.19	49.86	7.32	12.72	4.70	7.34	7.55	3.13	3.93	3.16	6.92
xLLMs-100	**46.95**	**51.53**	**51.96**	**12.94**	**15.35**	**8.83**	**13.90**	**17.29**	**3.27**	**8.09**	**4.04**	**14.18**
BLOOM	36.47	44.27	49.14	7.56	8.67	9.03	14.06	16.80	2.54	2.04	2.05	2.56
BX_{BLOOM}	36.42	46.28	50.35	4.81	8.11	4.89	8.47	11.71	2.14	1.74	2.41	1.57
SFT_{BLOOM}	36.67	49.42	52.31	6.31	11.88	5.62	10.12	14.33	**3.12**	3.79	2.62	2.52
xLLMs-100	**39.83**	**52.50**	**55.59**	**7.94**	**13.35**	**12.87**	**15.23**	**18.38**	3.02	**4.71**	**3.94**	**6.54**

Generating Capabilities												
	PAWS-X	XCOPA		Self-Instruct*		XL-Sum			FLORES(f)		FLORES(t)	
		low	high	low	high	low	mid	high	low	high	low	high
LLaMA	50.22	49.33	51.52	5.38	8.81	6.26	5.80	8.08	1.35	3.90	2.11	4.95
BX_{LLaMA}	48.41	48.00	49.85	7.01	9.80	1.11	2.74	1.70	1.56	5.33	1.37	1.61
SFT_{LLaMA}	50.36	48.93	50.05	7.10	12.15	4.51	6.06	9.21	2.42	4.56	2.71	7.29
xLLMs-100	**61.94**	**49.71**	**54.68**	**9.16**	**14.71**	**9.99**	**13.57**	**16.61**	**2.89**	**9.07**	**5.64**	**16.98**
BLOOM	47.39	49.85	49.47	4.07	7.01	6.08	7.77	8.91	0.78	1.20	0.99	1.49
BX_{BLOOM}	47.26	47.72	49.98	5.88	8.21	1.98	3.59	4.58	0.47	0.82	1.95	2.33
SFT_{BLOOM}	48.50	49.13	49.28	7.78	11.51	3.89	8.87	10.89	2.59	3.12	2.05	2.56
xLLMs-100	**50.53**	**52.36**	**52.26**	**10.17**	**13.62**	**8.77**	**11.74**	**12.36**	**3.97**	**5.79**	**4.22**	**7.68**

1. Introduction
2. Method
3. Experiments
4. Results
5. **Analysis**
6. Conclusion

Different Human Feedback Datasets

- We employ the DPO algorithm (Rafailov et al., 2023) to finetune our model, xLLMs-100, on two distinct datasets: traditional monolingual human feedback dataset and our constructed cross-lingual human feedback dataset.

	Low		High	
	mono	cross	mono	cross
PAWS-X	-	-	58.43	**61.94**
XCOPA	47.26	**49.71**	52.15	**54.68**
Self-Instruct*	3.25	**9.16**	12.14	**14.71**
XL-Sum	3.38	**9.99**	12.52	**16.61**
FLORES(f)	0.85	**2.89**	4.57	**9.07**
FLORES(t)	1.55	**5.64**	8.45	**16.98**

Table 4 An ablation study of xLLMs-100 using mono-lingual and cross-lingual human feedback data on low- and high-resource languages.

Different Datasets for Multilingual Tuning

- we conduct comprehensive comparison experiments on these two types of dataset: multilingual parallel dataset and multilingual instruction tuning dataset.

	Low		High	
	para	instruct	para	instruct
PAWS-X	-	-	40.17	**50.36**
XCOPA	37.14	**48.93**	42.13	**50.05**
Self-Instruct*	2.63	**7.10**	5.48	**12.15**
XL-Sum	1.10	**4.51**	5.12	**9.21**
FLORES(f)	**5.06**	2.42	**13.27**	4.56
FLORES(t)	**12.36**	2.71	**18.27**	7.29

Table 5 Multilingual Tuning on multilingual parallel corpora and multilingual instruction dataset in high-resource and low-resource languages.

Off-Target Analysis

- Off-target (Zhang et al., 2020) refers to the generation of output in an incorrect language, which is a common issue in multilingual models.

	FLORES(f)		FLORES(t)	
	Low	High	Low	High
LLaMA	23.26	16.76	14.15	10.16
BX_{LLaMA}	14.13	8.32	12.17	8.24
SFT_{LLaMA}	10.26	6.34	8.72	6.23
xLLMs-100	**8.82**	**3.47**	**6.95**	**1.46**

Table 6 OTR scores (lower is better) of examined multilingual LLMs on the FLORES benchmark.

Language Democratization

- Language democratization, as proposed by Huang et al., (2023), is a metric used to evaluate the level of task democratization across different languages of a multilingual model. This metric is obtained by calculating the average percentage of different languages relative to the best performing language among all languages.

	LLaMA	BX_{LLaMA}	SFT_{LLaMA}	xLLMs-100
PAWS-X	60.56	58.77	60.63	**66.43**
XCOPA	93.33	98.52	**99.31**	89.63
Self-Instruct*	57.85	68.68	62.63	**73.92**
XL-Sum	47.09	8.90	50.35	**67.21**
FLORES(f)	34.33	34.00	25.84	**34.68**
FLORES(t)	49.84	**58.28**	35.53	48.28

Table 7 Language Democratization: Mitigating the gap between the average performance and the best performances of each task in different languages.

1. Introduction
2. Method
3. Experiments
4. Results
5. Analysis
6. Conclusion

Conclusion

- To enhance the multilingual capability of LLMs in two dimensions (understanding and generating), we construct two dataset: multilingual instruction dataset covering 100 languages and cross-lingual human feedback dataset covering 30 languages.
- We introduced xLLMs-100, a multilingual LLMs finetuned based on LLaMA and BLOOM, which obtains strong generating and understanding capabilities.

Thank You!

Email: lavine@cis.lmu.de
Homepage: https://lavine-lmu.github.io
Address: Bildungscampus 3, 74076 Heilbronn, Germany

Paper

Code