Extrapolating Large Language Models to Non-English by Aligning Languages

Anonymous ACL submission

Abstract

Existing large language models (LLM) show disparate capability across different languages. Their performances on non-English tasks are often much worse than on English tasks. In this paper, we explore to extrapolate LLM's English ability to non-English by building semantic alignment across languages. We start from targeting individual languages by performing bilingual multi-task instruction-tuning, i.e. tuning LLM with bilingual translation task and bilingual instruction-following task. Then we formulate underlying scaling laws to quantify the impact of scaling up translation data and providing insights for devising multilingual instruction-tuning strategies, e.g., optimizing multilingual data allocation. Experiment results show that our alignment-enhanced LLMs significantly outperforms the English-dominated instruction-tuned counterpart on both translation task and other zero-shot non-English tasks, e.g., question answering, knowledge infilling and summarization. Our optimized data allocation also assists LLM in achieving better multilingual performance compared to uniform allocation. Further analysis on representation space and response content reveals additional evidence of the established language alignment.

1 Introduction

The language ability of LLMs is often imbalanced across languages (Zhu et al., 2023; Huang et al., 2023; Qi et al., 2023), because both the pre-training corpus (Blevins and Zettlemoyer, 2022) and the instruction-tuning data (Wang et al., 2023) are English-dominated. As a result, LLMs usually perform poorly on non-English languages, especially on languages that are dissimilar to English (Bang et al., 2023; Huang et al., 2023).

Previously, there have been some attempts to enhance LLMs' non-English abilities by continued pre-training with large scale monolingual corpus (Cui et al., 2023; Yang et al., 2023). However, further learning a language may require large scale data and computing.

In this paper, our objective is to enhance the proficiency of off-the-shelf LLMs on non-English languages in a more efficient manner. Specifically, we explore to extrapolate LLM's English ability to non-English languages. For this goal, we present a multi-task training recipe, which combines translation task and instruction-following task during instruction-tuning. Intuitively, the translation tasks stimulate the semantic alignment between languages and combining it with multilingual parallel instruction-following task encourages LLMs to execute non-English instructions based on its understanding of English.

At first, we target individual languages by performing bilingual instruction-tuning (as depicted on the left side of Figure 1) and formulate underlying scaling laws to investigate the impact of scaling up translation data. Guided by these scaling laws, we perform multilingual instruction-tuning with mixed resources (illustrated on the right side of Figure 1). Since we observe that more translation data usually contributes to improved alignment, combining all available resources for instruction-tuning becomes the most straightforward approach to obtain a powerful multilingual LLM. If we consider a practical scenario where instruction-tuning have to be performed under a fixed data budget, we also devise a multilingual data allocation method by formulating the problem as constrained non-linear programming based on the established scaling laws.

In the experiments, we use both LLaMA-7B (Touvron et al., 2023) and Pythia-6.9B (Biderman et al., 2023) as the pre-trained LLM and evaluate them on six challenging target languages. Experiment results on several multilingual benchmarks (FLORES-101, XQUAD, MLQA, mLAMA, XLSUM) show that our alignment-enhanced LLM outperforms its English-dominated instruction-tuned counterpart by a large margin. On

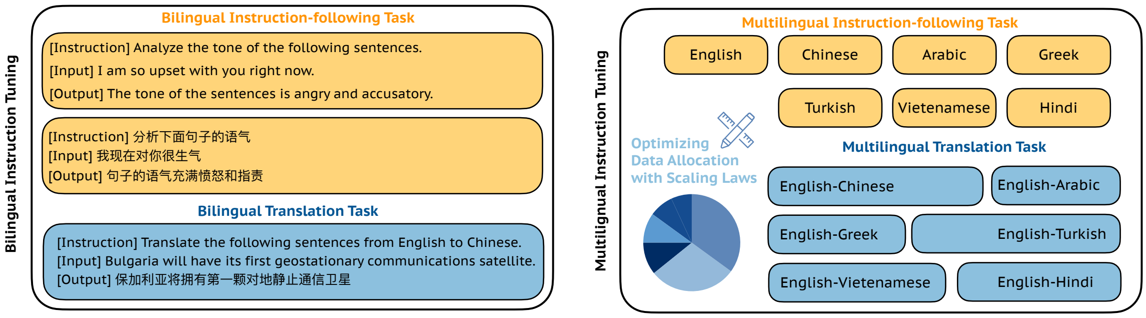

Figure 1: Illustration of our devised training recipes: bilingual instruction-tuning and multilingual instruction-tuning. We perform bilingual instruction-tuning by tuning pre-trained LLM with both bilingual instruction-following task and bilingual translation task. Guided by the scaling law in bilingual instruction-tuning, we perform multilingual instruction-tuning with mixed resources.

the translation task, our LLM has become a proficient translator, achieving a COMET score of 85. On other non-English zero-shot tasks, such as question answering, knowledge infilling and summarization, our model also achieves significant performance enhancements. In the resource-constrained setting, our optimized data allocation yields higher multilingual performance than the uniform allocation, showing a practical usage of the established scaling laws. Further analysis on response content and representation space reveals that our model has a tendency to generate non-English response based on its English memory and multilingual semantic space appears to align within the middle layers, demonstrating the effectiveness of our method.

The main contribution of this paper can be summarized as:

- We present a multi-task training recipe to elicit pre-trained LLM's non-English capability.
- We formulate the scaling law in bilingual instruction-tuning, providing insight for multilingual instruction-tuning, e.g., we devise a novel data allocation algorithm based on the established scaling law.
- Extensive experiment results on multilingual benchmarks show that our training recipe can greatly improves LLM's non-English capabilities.

2 Related Work

Instruction-tuning LLM to unlock its potential Although pre-trained LLMs memorizes vast amounts of knowledge, they often struggle to follow human instructions accurately (Ouyang et al., 2022). Therefore, Wei et al. (2022) propose instruction-tuning to teach LLM to follow human instruction and align its behavior closely with human expectations. Subsequently, numerous endeavors have been dedicated to this fine-tuning approach to unlock the potential of LLMs, such as step-by-step reasoning (Kim et al., 2023), story generation (Du and Chilton, 2023), tabular prediction (Slack and Singh, 2023). In this paper, we focus on elicit LLM's non-English ability through instruction-tuning.

Improving LLM's non-English performance Extensive empirical evidence has shown that there is a large gap between LLM's English and non-English performance (Huang et al., 2023; Qin et al., 2023). To improve LLM's non-English performance, a straightforward idea is to continued pre-train LLM with non-English corpus (Cui et al., 2023; Nguyen et al., 2023). However, this approach requires large scale monolingual corpus and computing. In contrast, we focus on the instruction-tuning stage and explore a more efficient manner, which shares the same spirit with some work conducted during the same period (Chen et al., 2023; Li et al., 2023). Compared to concurrent studies, we go beyond showing the value of non-English instruction data; we also present the advantages of incorporating the translation task to enhance LLM's non-English performance.

3 Eliciting LLM's non-English Ability

This section introduces our training methodology. We start by introducing bilingual instruction-tuning (§3.1), a technique aimed at empowering LLM for specific non-English languages. Following this, we formulate the scaling law in bilingual instruction-tuning to quantify the impact of scaling up translation data (§3.2). Lastly, we draw insights from these scaling laws to perform multilingual instruction-tuning (§3.3).

3.1 Bilingual Instruction-tuning

When we target a specific non-English language, our multi-task training framework consists of bilingual translation task and bilingual instruction-following task.

Translation task Intuitively, translation data is an invaluable resource for learning semantic alignment, which is, however, often overlooked in concurrent multilingual instruction-tuning research. In our training framework, we incorporate machine translation as an auxiliary task to teach LLM to semantically align English and non-English languages. Specifically, we position English and non-English text on the source and target sides of the translation data, respectively. This implementation can not only enhance LLM's proficiency in non-English generation, but also inherently encourage LLM to generate non-English content based on its understanding of English.

Instruction-following task It has been found that training LLM with diverse instructions can greatly improves LLM's performance on understanding (even unseen) instructions and aligning LLM's behavior with human expectations. Consequently, we also incorporate this approach into our framework. Given that commonly-used instruction-following datasets are almost in English, we translate the English dataset into the target language using a machine translation engine. During training, we utilize both the English and non-English version to establish a bilingual instruction-following task, which simultaneously elicit LLM's English and non-English capabilities. Combining it with translation task further encourage the extrapolation of LLM's English capabilities towards non-English languages.

Training Details In the end, we combine multi-task data into a training set $\mathcal{D}$ for instruction-tuning. To unify the data format, we also pair each translation data with a translation instruction. The final training objective can be written as:

$$\arg\min_{\theta} \frac{1}{|\mathcal{D}|} \sum_{\{T,X,Y\}\in\mathcal{D}} -\log p_{\theta}(Y|T,X)$$

T denotes a task instruction that describes the task requirement. X represents the optional input sequence and Y is the desired output for the given task. θ denotes learnable parameters of the LLM.

3.2 Scaling Law Formulation

The volume of translation data is an important variable in our instruction-tuning. Hence, prior to extending our approach to a multilingual setting, we strive to understand the effect of varying the size of translation data on language alignment.

Specifically, we employ bilingual translation performance as a measure of semantic alignment. To quantify the relationship between translation performance $\mathcal{S}$ and translation data scale $\mathcal{X}$, we formulate the scaling law based on following insights: (1) The upper bound of $\mathcal{S}$ is 100, representing the maximum score achievable by most translation quality metrics. (2) The translation performance generally enhances as the scale of translation data expands. (3) Languages that are less similar to English require a larger amount of translation data to achieve alignment compared to languages more similar to English. After exploring various possible formulations, we present the best-performing formulation as follows:

$$\mathcal{S} = g(\mathcal{X}) = 100 - \alpha \cdot (\gamma \cdot \mathcal{X})^{-\beta}$$

where $\gamma \in (0, 1)$ represents the language similarity between the target language and English, which can be pre-calculated with a parallel corpora through a method [1] introduced by Pan et al. (2021). Using a set of observed data points $\{(\mathcal{S}, \mathcal{X})\}$, the value of α and β can be determined through non-linear least squares [2].

In the subsequent subsection, we will illustrate how we utilize the established scaling laws as guidance to devise multilingual instruction-tuning strategies.

3.3 Multilingual Instruction-tuning

Although bilingual instruction-tuning is effective, serving customized LLMs for each language can be costly. Now we take a step further and investigate multilingual instruction-tuning. In this scenario, our multi-task training framework encompasses multilingual translation task and multilingual instruction-following task, also making multilingual data allocation a crucial aspect to consider.

[1] Specifically, we use pre-trained LLM to encode parallel data and obtain sentence representations by averaging the outputs from the final layer. Then we use English representations to retrieve non-English sentences. Consequently, a retrieval score is assigned based on the ranking of the target sentence in the retrieval results. By averaging all retrieval scores, we can calculate the language similarity.

[2] https://en.wikipedia.org/wiki/Non-linear_least_squares

Task	Dataset	Arabic	Greek	Hindi	Turkish	Vietnamese	Chinese
Translation	WikiMatrix	999.8k	620.8k	231.5k	477.7k	1073.8k	786.5k
	NewsCommentary	97.4k	-	2.8k	-	-	126.0k
Instruction-following	Alpaca	52.0k	52.0k	52.0k	52.0k	52.0k	52.0k

Table 1: Statistics of our applied training dataset.

Resource-rich setting Given that the scaling law function is monotonically increasing, it suggests that semantic alignment can continually improve with the use of more translation data. Consequently, if no data budget is specified, combining all available resources for instruction-tuning becomes a direct and effective approach to maximize multilingual performance.

Resource-constrained setting A potential concern of using all available resources for tuning is the huge computational cost it incurs. Therefore, we also explore a practical scenario where we assume that there is a fixed data budget for the multilingual translation data being used. To achieve the optimal data combination in this scenario, we propose to formulate data allocation as a constrained non-linear programming problem based on our established scaling law. The objective of this programming problem is to maximize the average multilingual translation performance:

$$\max \frac{1}{n}\sum_{i=1}^{n} g(\mathcal{X}_i), \text{ s.t.} \sum_{i=1}^{n} \mathcal{X}_i = C, \quad \text{where} \quad 0 \le \mathcal{X}_i \le \mathcal{X}_i^{max}, i = 1, 2, 3 \cdots n. \tag{1}$$

In this equation, n denotes the number of considered languages and the data budget constraint limits the total amount of translation data to a predefined budget C. $\mathcal{X}_i^{max}$ denotes the maximum number of available translation data for language i. This constrained nonlinear optimization can be solved with sequential least squares programming [3].

4 Experiment Setting

Pre-trained LLM We take LLaMA-7B (Touvron et al., 2023) and Pythia-6.9B (Biderman et al., 2023) as the pre-trained LLM and consider six target languages that LLM usually struggle to deal with: Arabic (Ar), Greek (El), Hindi (Hi), Turkish (Tr), Vietnamese (Vi) and Chinese (Zh).

Instruction tuning details For translation data, we use publicly available parallel corpora, WikiMatrix[4] (Schwenk et al., 2021) and NewsCommentary[5] (Tiedemann, 2012), which are more accessible and scalable compared to high-cost expert-annotated translation data (Jiao et al., 2023). For multilingual general task instruction data, we incorporate Alpaca dataset (Taori et al., 2023), which consists of 52k English questions and corresponding response, and we obtain its foreign version with an advanced machine translation engine [6]. The statistics of the datasets are presented in Table 1. We use *stanford_alpaca*[7] as the code base.

Evaluation Dataset We use five multilingual benchmarks to assess LLM's non-English performance, spanning several downstream tasks. Flores-101 (Goyal et al., 2022) evaluates translation performance. MLQA (Lewis et al., 2020) and XQUAD (Artetxe et al., 2020), both question answering tasks, require the model to reason over the provided context and respond to the posed question. mLAMA (Kassner et al., 2021) assesses the multilingual knowledge contained in the model. XLSum (Hasan et al., 2021) evaluates the model's summarization capabilities.

Evaluation Metrics For translation tasks, we use COMET (Rei et al., 2020), calculated by *wmt22-comet-da* model. For question answering and knowledge infilling task, we report exact-matching accuracy. For summarization task, we report ROUGE score (Lin, 2004).

5 Main Results

In this section, we present our main experiment results, show the effectiveness of our training recipes and introduce our findings.

[3] https://en.wikipedia.org/wiki/Sequential_quadratic_programming

[4] https://opus.nlpl.eu/News-Commentary.php

[5] https://github.com/facebookresearch/LASER/tree/main/tasks/WikiMatrix

[6] We employ Alibaba Translate for the translation process, which has strong translation capabilities (https://www.alibabacloud.com/product/machine-translation).

[7] https://github.com/tatsu-lab/stanford_alpaca

	Flores-101 En-X (COMET)						
[LLaMA-7B]	Hi	Tr	El	Zh	Vi	Ar	Avg.
+ English Instruction Task	39.5	43.1	43.7	53.3	42.7	46.9	44.9
+ Bilingual Instruction Task	43.3 (+3.8)	59.5 (+16.4)	64.5 (+20.8)	69.7 (+16.4)	68.7 (+26.0)	66.0 (+19.1)	62.0 (+17.1)
↪ + Bilingual Translation Task	**78.4** (+38.9)	**87.1** (+44.0)	**87.2** (+43.5)	**87.2** (+33.9)	**87.8** (+45.1)	**86.6** (+39.7)	**85.7** (+40.9)
[Pythia-6.9B]	Hi	Tr	El	Zh	Vi	Ar	Avg.
+ English Instruction Task	39.8	53.3	55.3	61.5	57.5	51.3	53.1
+ Bilingual Instruction Task	39.5 (-0.3)	57.5 (+4.2)	67.9 (+12.6)	67.6 (+6.1)	67.7 (+10.2)	59.4 (+8.1)	59.9 (+6.8)
↪ + Bilingual Translation Task	**76.0** (+36.2)	**85.8** (+32.5)	**87.8** (+32.5)	**85.9** (+24.4)	**87.3** (+29.8)	**85.8** (+34.5)	**84.8** (+31.7)
	MLQA (Accuracy)						
[LLaMA-7B]	Hi	Tr	El	Zh	Vi	Ar	Avg.
+ English Instruction Task	13.7	-	-	26.7	34.1	16.1	22.7
+ Bilingual Instruction Task	35.1 (+21.4)	-	-	48.0 (+21.3)	50.1 (+16.0)	33.1 (+17.0)	41.6 (+18.9)
↪ + Bilingual Translation Task	**42.3** (+28.6)	-	-	**51.8** (+25.1)	**50.8** (+16.7)	**37.0** (+20.9)	**45.5** (+22.8)
[Pythia-6.9B]	Hi	Tr	El	Zh	Vi	Ar	Avg.
+ English Instruction Task	6.8	-	-	18.0	25.5	9.4	14.9
+ Bilingual Instruction Task	30.6 (+23.8)	-	-	39.1 (+21.1)	37.8 (+12.3)	27.0 (+17.6)	33.6 (+18.7)
↪ + Bilingual Translation Task	**33.8** (+27.0)	-	-	**42.7** (+24.7)	**45.1** (+19.6)	**31.9** (+22.5)	**38.4** (+23.5)
	XQUAD (Accuracy)						
[LLaMA-7B]	Hi	Tr	El	Zh	Vi	Ar	Avg.
+ English Instruction Task	15.5	36.7	31.7	31.8	36.7	14.9	27.9
+ Bilingual Instruction Task	37.8 (+22.3)	**54.5** (+17.8)	**48.0** (+16.3)	51.7 (+19.9)	**54.5** (+17.8)	**39.0** (+24.1)	**47.6** (+19.7)
↪ + Bilingual Translation Task	**44.0** (+28.5)	50.9 (+14.2)	44.1 (+12.4)	**54.9** (+23.1)	50.9 (+14.2)	38.8 (+23.9)	47.3 (+19.4)
[Pythia-6.9B]	Hi	Tr	El	Zh	Vi	Ar	Avg.
+ English Instruction Task	10.1	29.4	16.9	22.0	27.4	8.8	19.1
+ Bilingual Instruction Task	29.3 (+19.2)	32.4 (+3.0)	39.2 (+22.3)	40.2 (+18.2)	41.5 (+14.1)	30.3 (+21.5)	35.5 (+16.4)
↪ + Bilingual Translation Task	**33.3** (+23.2)	**44.7** (+15.3)	**43.7** (+26.8)	**44.3** (+22.3)	**47.6** (+20.2)	**34.1** (+25.3)	**41.3** (+22.2)
	mLAMA (Accuracy)						
[LLaMA-7B]	Hi	Tr	El	Zh	Vi	Ar	Avg.
+ English Instruction Task	0.9	6.1	0.6	4.5	2.2	1.2	2.6
+ Bilingual Instruction Task	3.7 (+2.8)	11.2 (+5.1)	8.1 (+7.5)	16.9 (+12.4)	17.5 (+15.3)	18.0 (+16.8)	12.6 (+10.0)
↪ + Bilingual Translation Task	**6.7** (+5.8)	**18.8** (+12.7)	**12.4** (+11.8)	**22.4** (+17.9)	**29.2** (+27.0)	**18.9** (+17.7)	**18.1** (+15.5)
[Pythia-6.9B]	Hi	Tr	El	Zh	Vi	Ar	Avg.
+ English Instruction Task	0.3	4.5	1.3	0.5	4.7	0.3	1.9
+ Bilingual Instruction Task	**1.7** (+1.4)	6.0 (+1.5)	2.9 (+1.6)	13.2 (+12.7)	14.3 (+9.6)	2.3 (+2.0)	6.7 (+4.8)
↪ + Bilingual Translation Task	1.5 (+1.2)	**7.3** (+2.8)	**3.6** (+2.3)	**14.2** (+13.7)	**15.6** (+10.9)	**3.1** (+2.8)	**7.6** (+5.7)
	XLSum (ROUGE)						
[LLaMA-7B]	Hi	Tr	El	Zh	Vi	Ar	Avg.
+ English Instruction-task	13.9	29.7	-	9.0	32.3	15.2	20.0
+ Bilingual Instruction Task	27.0 (+13.1)	33.7 (+4.0)	-	25.5 (+16.5)	**34.1** (+1.8)	**41.5** (+26.3)	32.4 (+12.3)
↪ + Bilingual Translation Task	**30.6** (+16.7)	**37.4** (+7.7)	-	**28.3** (+19.3)	32.1 (-0.2)	40.2 (+25.0)	**33.7** (+13.7)
[Pythia-6.9B]	Hi	Tr	El	Zh	Vi	Ar	Avg.
+ English Instruction Task	21.8	38.3	-	13.1	36.7	17.3	25.4
+ Bilingual Instruction Task	**45.5** (+23.7)	**46.5** (+8.2)	-	**37.4** (+24.3)	**47.8** (+11.1)	**48.4** (+31.1)	**45.1** (+19.7)
↪ + Bilingual Translation Task	44.7 (+22.9)	46.0 (+7.7)	-	28.6 (+15.5)	46.0 (+9.3)	47.3 (+30.0)	42.5 (+17.1)

Table 2: Effects of bilingual instruction-tuning, i.e. tuning LLM with both bilingual instruction-following task and bilingual translation task. Bold text denotes the highest score across different training strategies. The number in the bracket denotes the performance improvement over the baseline approach.

5.1 Results on Bilingual Instruction-tuning

Bilingual instruction-tuning yields great improvement on non-English performance Table 2 presents the comparison results between our bilingual instruction-tuning method and the baseline approach, which tunes LLM with English-dominated instruction-following task (original Alpaca dataset). It is obvious that the baseline approach fails to fully harness the LLM's capabilities in non-English languages. Bilingual instruction-tuning significantly enhances LLM's performance on non-English tasks, yielding an average accuracy improvement of 4.8% to 23.5% on question answering and knowledge infilling tasks, and yielding an average ROUGE improvement of 12.3 to 19.7, where both the bilingual instruction-following and translation tasks contributing to this improvement. Notably, the added translation task not only augments the model's performance in translation, it also leads to performance improvements in other zero-shot tasks, demonstrating the value of this auxiliary task.

Scaling up translation data usually lifts non-English performance Now we show the impact of scaling up translation data and provide insight for subsequent multilingual instruction-tuning. Fig-

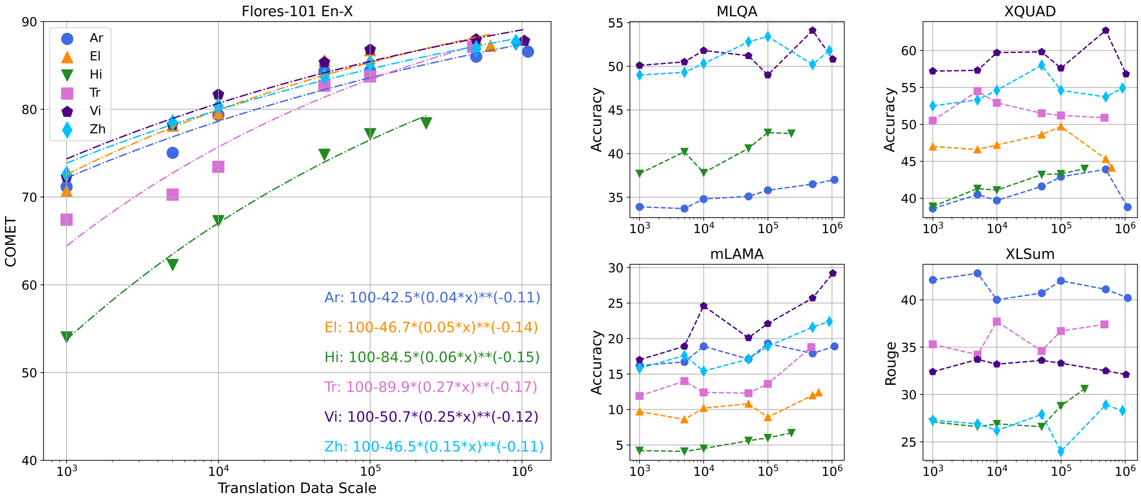

Figure 2: The relationship between translation data scale and downstream task performance. On the left subfigure, our designed formulation (the dashdotted line) well fits with the trend and the scaling laws are listed on the figure.

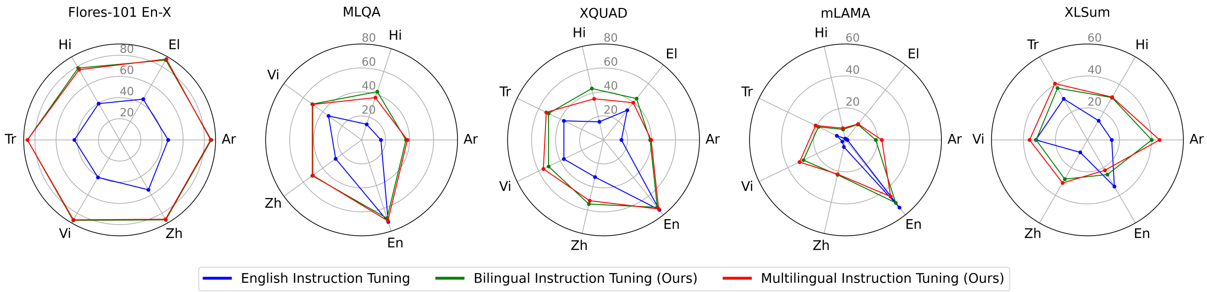

Figure 3: Multilingual performance of LLMs that are instruction-tuned with different strategies.

ure 2 illustrates our empirical results on LLaMA-7B. Incorporating more translation data usually results in improved performance on both translation task and other zero-shot tasks. After fitting our designed formulation to these observed points, we can see that the scaling law (the dashdotted line in the left subfigure) well represents the trend and describe the quantified relationship between translation performance and translation data scale. Besides, we can also interpret from the scaling curve that the rate of improvement in semantic alignment appears to diminish as the volume of translation data increases. Therefore it would be an interesting problem to investigate how to achieve the largest marginal effect in multilingual data allocation.

5.2 Results on Multilingual Instruction-tuning

Multilingual instruction tuning can simultaneously enhance LLM's capabilities across several non-English languages Building on our previous analysis of scaling laws, if there's no specific data budget, combining all available resources for instruction-tuning stands out as an intuitive strategy to maximize multilingual performance. Figure 3 displays experiment results on LLaMA-7B. Our multilingual LLM achieves performance on par with LLMs fine-tuned with bilingual data for individual languages, which also outperforms the baseline system in non-English tasks by a large margin. In terms of English tasks, our training method does not lead to severe catastrophic forgetting. However, we also notice that our approach has not yet completely closed the performance gap between English and non-English tasks, which continues to be an open challenge.

In resource-constrained setting, we can leverage the formulated scaling laws to achieve the optimal data allocation In this setting, we assume a fixed data budget for the multilingual translation being used, for example, a 1.2M data budget. Table 3 presents the comparison results between the uniform allocation and our optimized allocation. Given that our optimization objective (Equation 1) aims to maximize multilingual translation

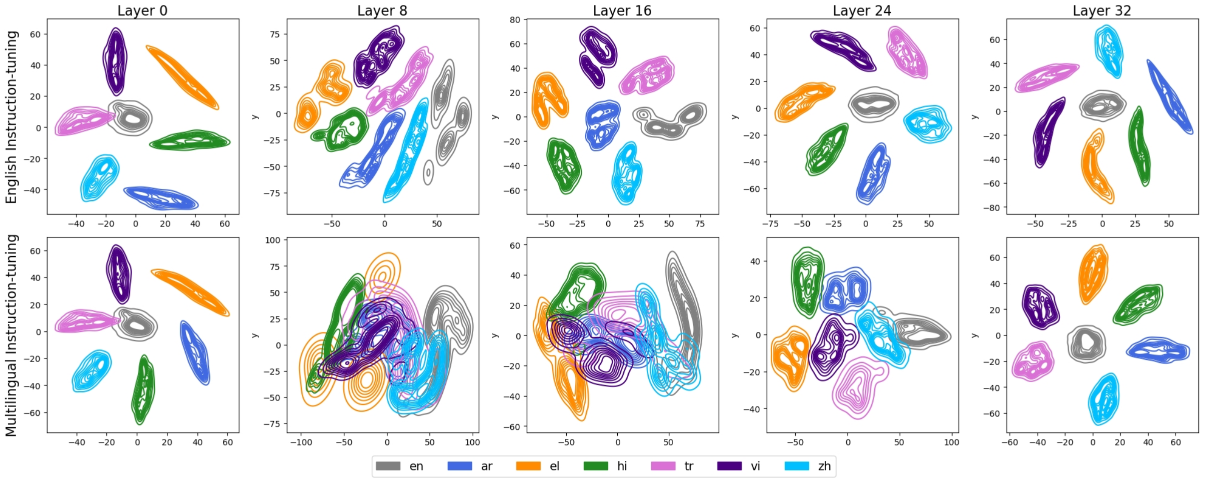

Figure 4: Visualization analysis on the representation space of LLMs that are instruction-tuned with different strategies. For English instruction-tuned model, representations of different languages always stay apart from bottom layers to top layers. In contrast, we observe representation overlap in our multilingual model, especially in middle layers.

Translation Data Allocation						Multilingual Tasks		
Ar	El	Hi	Tr	Vi	Zh	**Flores-COMET**	**Flores-BLEURT**	**Flroes-BLEU**
200,000	200,000	200,000	200,000	200,000	200,000	84.22	69.73	33.81
183,539	189,556	234,233	242,263	175,985	174,422	84.70* (+0.48)	70.42* (+0.69)	34.40* (+0.59)
Ar	El	Hi	Tr	Vi	Zh	**MLQA**	**XQUAD**	**mLAMA**
200,000	200,000	200,000	200,000	200,000	200,000	43.2	46.9	18.1
183,539	189,556	234,233	242,263	175,985	174,422	44.6* (+1.4)	49.2* (+2.3)	15.9 (-2.2)

Table 3: Comparison results between our optimized allocation and uniform allocation under a 1.2M data budget. We report averaged multilingual performance for downstream tasks. The number in the bracket denotes the performance gap between the two data allocation strategies. The annotation "*" indicates that the improvement is significant ($p<0.05$).

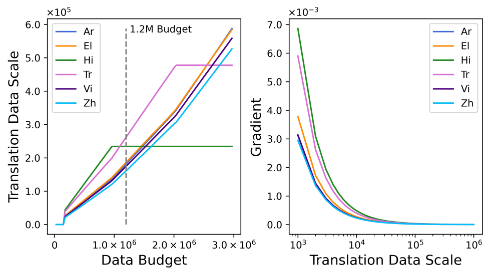

Figure 5: The left figure illustrates the changes in data allocation as the total data budget varies, while the right figure shows the gradient changes of different language scaling law functions. In the left figure, when the line becomes straight, it indicates that the data for that language has been completely used up.

performance evaluated by COMET, our method does significantly outperform the uniform allocation on the COMET metric. The improvement on other translation metrics, BLEURT and BLEU, is also significant. On zero-shot tasks, the results are mixed, our optimized allocation yields a higher average performance in question answering, but a lower average performance in knowledge infilling. We can also explain the insight behind our allocation algorithm here: as different languages demonstrate unique scaling laws, it implies that the performance gain achieved at the same scale can differ, as characterized by varying gradients (Figure 5). Consequently, when the total data volume is limited, our method allocate resources preferentially to languages with steeper gradients, such as Hindi and Turkish. In contrast, uniform allocation overlooks these differences.

6 Analysis

Visualization results show that multilingual semantic space becomes aligned in the middle layers of our model To gain a deeper understanding of the effects of our tuning method, we conduct a comparison between the representation space of our alignment-enhanced multilin-

English instruction	Chinese instruction
[Instruction] Choose three adjectives to describe a street [Output] Busy, vibrant, and lively.	[Instruction] 选择三个形容词来描述一条街道 [Output] 繁华、充满活力和美丽。(Bustling, vibrant and beautiful.)
[Instruction] What values does a democratic society strive to uphold? [Output] A democratic society strives to uphold values such as freedom, equality, justice, and human rights. It is committed to upholding the right of all citizens to participate in decision-making and to be treated with respect and dignity. It also seeks to promote freedom of speech and thought, as well as the rule of law.	[Instruction] 民主社会努力维护哪些价值观？ [Output] 民主社会努力维护的价值观包括个人自由、平等、言论和宗教自由、人身安全、公正法律和人类尊严。民主社会还促进尊重多样性和包容性，以及言论自由和集体决策的权利。(The values that democratic societies strive to uphold include individual liberty, equality, freedom of speech and religion, security of the person, just laws, and human dignity. Democratic societies also promote respect for diversity and inclusion, as well as the rights to freedom of expression and collective decision-making.)

Table 4: Two representative cases where our mulitlingual model makes similar response when given instructions in different languages. The gray text in the bracket denotes the English meaning of the Chinese response.

Stage	FLORES-101 COMET	MLQA Accuracy	XQUAD Accuracy	mLAMA Accuracy	XLSum ROUGE
Continued Pretraining	78.7	47.9	50.7	**36.8**	27.6
Bilingual Instruction-tuning	**87.2**	**51.8**	**54.9**	22.4	**28.3**

Table 5: Effects of using parallel data at different stages. Bold text denotes the highest score along the column.

gual model and the unaligned counterpart. Specifically, we use them to encode multilingual parallel data from FLORES-101 dataset and visualize dimension-reduced representations across various layer, from bottom to top, in Figure 4. For the baseline model, the representations of different languages always stay apart across layers. In contrast, our model demonstrates an overlap of representations, particularly noticeable within the middle layers. This overlap serves as additional evidence that our multilingual instruction-tuning establish better language alignment.

The alignment-enhanced LLM shows the tendency to respond multilingual instructions according to its English memory During experiments, we discover that our multilingual LLM shows the tendency to respond multilingual instructions according to its English memory. Table 4 shows two representative cases where our multilingual model produces similar response when given instructions in different languages.

The value of translation data is beyond exposing more non-English tokens to LLM For ablation study, instead of using parallel data during instruction-tuning, we use the Chinese part of the English-Chinese translation data as monolingual corpus for continued pre-training and then only use bilingual instruction-following task for instruction-tuning (denoted as "continued pretraining" in Table 5). Experimental results show that bilingual instruction-tuning exhibits better performance on all tasks except knowledge infilling, indicating that the benefits of parallel data for the model are not solely derived from exposing it to more non-English data, but also from aligning languages.

7 Conclusion

This paper aims at extrapolating pre-trained large language models to non-English by strengthening semantic alignment across languages. Specifically, we explore two multi-task training recipe: bilingual instruction-tuning and multilingual instruction-tuning, which both incorporates translation task as an important auxiliary task. Moreover, we formulate the scaling law of bilingual instruction-tuning and provide guidance for performing multilingual instruction-tuning, e.g., optimizing multilingual data allocation. Experiment results on several multilingual benchmarks show that our devised training strategies effectively enhance pre-trained LLM's non-English proficiency even these target languages share little alphabet with English. Overall, our approach and findings illuminate the potential for developing more potent LLMs for non-English languages.

Limitation

A limitation of our work is that we do not extend vocabulary for target non-English languages. The effect is dual. Our approach does not require a large-scale non-English corpus to learn embedding of extended tokens. But on the other hand, since LLaMA usually tokenizes non-English tokens to bytes, our model is slower in encoding and decoding non-English sequence than those models equipped with extended vocabulary. We leave the exploration on vocabulary manipulation as our future work.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O‘Ă‘ŹBrien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Barry Haddow, and Kenneth Heafield. 2023. Monolingual or multilingual instruction tuning: Which makes a better alpaca. *arXiv preprint arXiv:2309.08958*.

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

Yulun Du and Lydia Chilton. 2023. Storywars: A dataset and instruction tuning baselines for collaborative story understanding and generation. *arXiv preprint arXiv:2305.08152*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics (TACL)*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne 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. *arXiv preprint arXiv:2305.07004*.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Parrot: Translating during chat using large language models. *arXiv preprint arXiv:2304.02426*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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*. Association for Computational Linguistics (ACL).

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. Seallms – large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

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

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. *arXiv preprint arXiv:2310.10378*.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

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

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Dylan Slack and Sameer Singh. 2023. Tablet: Learning from instructions for tabular data. *arXiv preprint arXiv:2304.13188*.

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

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

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

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

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 (ICLR)*.

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

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. 2023. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Details of Our Instruction-tuning

For each experiment, we instruction-tune LLaMA's full parameters for 3 epoch on 8×A100. The learning rate is set as 2e-5 and batch size is set as 128. For training acceleration, we adopt FSDP training strategy (Zhao et al., 2023).

B Our Used Prompts for Downstream Tasks

We report our used prompts (English version) in Table 6. For monolingual non-English tasks, i.e. MLQA, XQUAD, MLAMA, XLSUM, we apply language-specific prompt (a foreign version of the English prompt in Table 6) when evaluating LLM's performance on the target language. For machine translation tasks, FLORES-101, we only use English instruction for multilingual translation in our experiments.

C Used Scientific Artifacts

Below lists scientific artifacts that are used in our work. For the sake of ethic, our use of these artifacts is consistent with their intended use.

- *Stanford Alpaca (Apache-2.0 license)*, a project that aims to build and share an instruction-following LLaMA model.
- *Transformers (Apache-2.0 license)*, a framework that provides thousands of pretrained models to perform tasks on different modalities such as text, vision, and audio.

Task	Dataset	Prompt
Question Answering	MLQA, XQUAD	Answer the question according to the paragraph in a few words. Context: <context> Question: <question> Answer:
Knowledge Infilling	mLAMA	Please write an answer that can be filled in [MASK].
Summarization	XLSum	Summarize this article. Article: <article> Summary:
Machine Translation	Flores-101	Translate the following sentences from <SRC> to <TGT>.

Table 6: Our used prompts for downstream tasks. "<context>", "<question>", "<article>" are placeholders for input information. "<SRC>" and "<TGT>" represent the placeholder for source and target language name in English.