Does the English Matter? Elicit Cross-lingual Abilities of Large Language Models

Giulia Pucci and Leonardo Ranaldi
Human-Centric ART Group, Department of Enterprise Engineering,
University of Rome Tor Vergata.
[first_name].[last_name]@uniroma2.it,

Abstract

Large Language Models reveal diverse abilities across different languages due to the disproportionate amount of English data they are trained on. Their performances on English tasks are often more robust than in other languages.

In this paper, we propose a method to empower the cross-lingual abilities of instruction-tuned LLMs (It-LLMs) by building semantic alignment between languages. To achieve this, we introduce translation-following demonstrations to elicit better semantic alignment across languages. Our evaluations on multilingual question-answering benchmarks reveal that our models, tested in five distinct languages, outperform the performance of It-LLMs trained on monolingual datasets. The findings highlight the impact of translation-following demonstrations on non-English data, eliciting instruction-tuning and empowering semantic alignment.

1 Introduction

Large Language Models (LLMs) achieve comprehensive language abilities through pre-training on large corpora (Brown et al., 2020). Hence, the acquired language abilities follow the corpora features, primarily available in English (Lin et al., 2021; Zhang et al., 2023; Zhu et al., 2023). This phenomenon produces an imbalance in pre-training (Blevins and Zettlemoyer, 2022) and fine-tuning (Le et al., 2021). Thus, performance is usually lower for non-English languages, especially for low-resource ones (Huang et al., 2023; Bang et al., 2023). The most common approaches to mitigate this problem propose continuing pre-training with large-scale monolingual data (Imani et al., 2023; Cui et al., 2023; Yang et al., 2023), which requires considerable data and computational resources.

In this paper, we propose an approach to empower the It-LLM that elicits semantic alignment between English and other languages. We focus on exploiting the latent multilingual abilities of It-LLMs by empowering the pivotal phase of instruction-tuning using instruction-following demonstrations. To this end, we explore the potential of cross-lingual alignment by integrating translation-following demonstrations to refine the instruction-tuning process.

In our experiments, we use *Llama-7b* (Touvron et al., 2023) as the foundational LLM and target five languages. In instances where data is lacking, we undertake translation tasks. We use the Stanford Alpaca dataset (Taori et al., 2023) and its translated versions in the corresponding languages, while for the translation-following, we use a publicly available translation resource (Tiedemann, 2012), the most accessible and extendable to multiple languages (i.e., *translation-following demonstrations* on Figure 1).

Following the instruction-tuning phase, we assessed the efficacy of our five distinct *Alpaca* tailored for specific languages. Our evaluation leveraged four benchmarks: two inherently multilingual, i.e., XQUAD (Artetxe et al., 2019) and MLQA (Lewis et al., 2020), and two intrinsically monolingual, MMLU (Hendrycks et al., 2021) and BBH (Suzgun et al., 2022). The empirical results indicate that when trained using language-specific instructions combined with translation data, the instruction-tuned models significantly surpass the performance of models trained exclusively with non-English demonstrations. While our models bridge the gap among performances, the translation-following models exhibit optimal alignments. This highlights the pronounced proficiency of Llama when trained on English-centered datasets compared to non-English ones. Furthermore, the semantic alignment effort significantly strengthens the cross-lingual abilities of It-LLMs.

Our findings can be summarized as follows:

- The learning abilities of LLMs on non-English instruction-tuning tasks are limited;
- The multi-lingual abilities of instruction-tuned

Proceedings of the The 3rd Workshop on Multi-lingual Representation Learning (MRL), pages 173–183

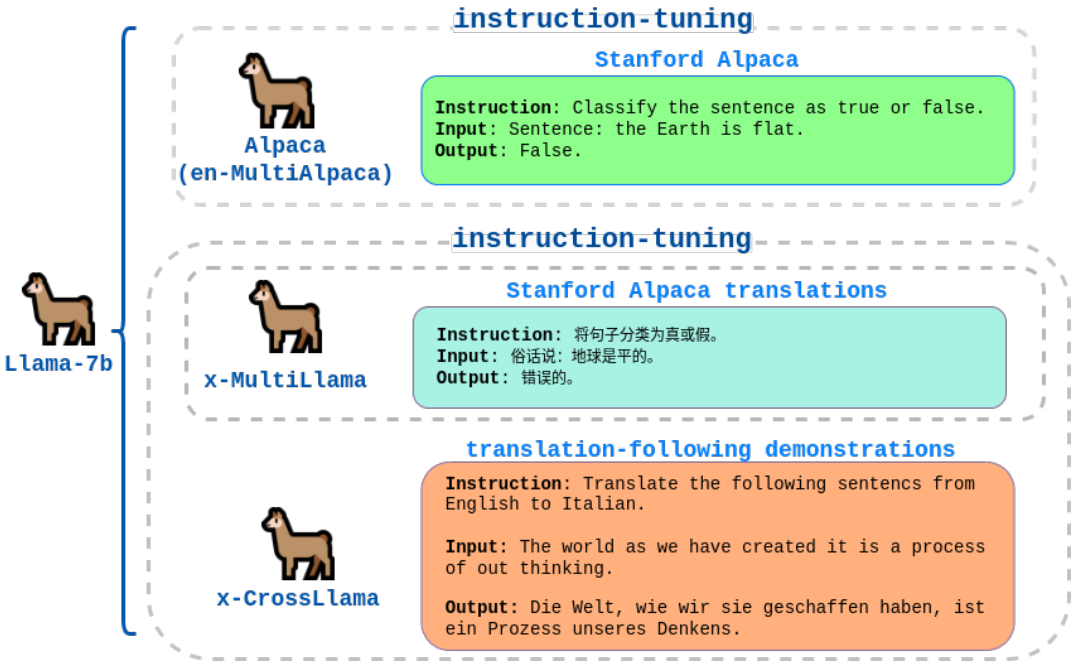

Figure 1: Our *x-CrossLlama* are instruction-tuned on instruction-following and translation-following demonstrations.

LLMs could be empowered through cross-lingual alignment;

- Thus, we propose to elicit the instruction-tuning approach for non-English models based on instruction-following and translation-following demonstrations for the target language. Hence, we show that It-LLMs can semantically align through cross-lingual translation-following demonstrations via an extensive evaluation.

2 Methods

Pre-training from scratch a Large Language Model (LLM) to fill the imbalance language problem is cost-prohibitive for data collection and parameter learning. This is why the trend is to do further fine-tuning to empower the models' abilities in a specific language (Tanti et al., 2021; Moslem et al., 2023). Hence, we aim to elicit the abilities of pre-trained LLMs for non-English languages by further improving the alignment between English and the target language. In the following Sections, we investigate the difficulties of fine-tuning a monolingual scenario (Section 2.1). Based on this, we propose our approach to empower the cross-lingual abilities of It-LLMs (Section 2.2).

2.1 Alpaca Instruction-tuning

The restricted availability and clarity of premium API services for cutting-edge LLMs have driven researchers to focus on creating open-source alternatives. Using the instruction-tuning paradigm, presented in Section 5.2, and resources as Stanford Alpaca (Taori et al., 2023) that is a corpus consisting of 52k of English instruction-output pairs generated by `text-davinci-003`, several instruction-tuned versions of instructed-Llama were released.

Following this approach, multiple monolingual versions of instructed-Llama were proposed by translating the Stanford Alpaca data into the specific language. Table 1 shows a set of versions available as open source. Following an analysis of the translated versions of instructed-Llama in official repositories[1], the languages of the benchmark datasets, and the translation pairs present in news_commentary, which will be introduced later, we selected the speeches that share the most already available data. Table 1 shows the custom versions used in this work, which for simplicity will be renamed *x*-MultiLlama, where *x* indicates the specific language.

2.2 Cross-lingual Instruction-tuning

Although monolingual techniques (presented in Section 2.1) play a key role in enhancing the multilingual strengths of LLMs, simply focusing on translated versions of Alpacas for specific languages does not allow the non-English capabilities

[1]official versions on https://github.com/tloen/alpaca-lora and https://huggingface.co/models

Model	Language	Name
Alpaca (Taori et al., 2023)	English	en-Llama
Alpaca-Chinese (Chen et al., 2023)	Chinese	zh-Llama
Camoscio (Santilli and Rodolà, 2023)	Italian	it-Llama
German (Thissen, 2023)	German	de-Llama
Arabic (Yasbok)	Arabic	ar-Llama

Table 1: The monolingual Instruction-tuned Large Language Models that use a language-specific version of *MultiLlama* as instruction-tuning data.

of LLMs to be exploited. To overcome this overlaps, we present CrossLlama, shown in Figure 1). This method empowers cross-lingual instruction-tuning by integrating translation-following demonstrations. We aim to elicit LLMs' English and non-English abilities by stimulating a semantic alignment challenge.

Instruction-following Although the version of the Alpaca dataset is in English, there are many derivatives. However, derived versions of the Alpaca dataset, as described in 2.1, have been produced with translation systems. Our work starts with the instruction-tuned Llama on Alpaca (native English) and its versions adapted for distinct languages (which we called *x-MultiLlama*). We also propose the *CrossLlama* variations, built from Alpaca translations specific to each language and augmented with translations (explained further). With this methodology, we intend to elicit the LLM backbone's capability to interpret multilingual instructions and ensure cross-lingual consistency.

Translation-following Challenge Using general instruction information is a logical approach when creating models to tackle multiple tasks guided by instructions(Wang et al., 2023; Zeng et al., 2023). Nevertheless, data from translations might aid in grasping semantic alignment.

We use publicly available sentence-level translation datasets, such as *news_commentary* (Tiedemann, 2012), to construct the translation task instruction demonstrations. We also propose extending this to additional languages, which we release as an open-source dataset. In particular, for each specific language, we constructed specific sets of demonstrations. Hence, following the Alpaca style (Instruction, Input, and Output) (see Table 1), we selected the same number of English to non-English translations non-English to English translations.

3 Experiments

In order to observe the English and non-English abilities of Large Language Models (LLMs) and the impact of the instruction-tuning approach in cross-lingual scenarios, we propose *CrossLlama*. Our approach is based on instruction-tuning on language-specific data augmented with a cross-lingual semantic alignment. Hence, we set several baseline models explained in Section 3.1, which we augmented with our approach introduced in Section 3.2. Finally, we performed a series of systematic evaluations (Section 3.3.1) to observe the impact of the proposed method.

3.1 Baseline LLMs

The common denominator among the It-LLMs shown in Table 1 is the LLM backbone Llama-7b (Touvron et al., 2023). Starting from instruction-following data from the original Alpaca (Taori et al., 2023) and its open-source non-English versions[2], we reproduced *x-MultiLlama* for *x* specific languages: Chinese (zh), Italian (it), Arabic (ar), German (de) and the original English version (en).

3.2 Cross-lingual LLMs

Our method produces *x-CrossLlama* that are instruction-tuned on standard instruction-following empowered with translation-following demonstrations.

Our approach generates a series of instruction-tuned versions of the data shown in Figure 1. We have named the versions *x-CrossLlama*.

3.3 Experimental Setup

To assess the performance of the *x-CrossLlama*, we defined several benchmarks (Section 3.3.1) on which we applied systematic instruction-tuning pipelines in Section 3.3.2.

3.3.1 Benchmarks

To evaluate the performance of the It-LLMs and the impact of the semantic alignment approach, we used two cross-lingual (XQUAD (Artetxe et al., 2019), MLQA (Lewis et al., 2020)) and two multi-task (MMLU (Hendrycks et al., 2021) and BBH (Suzgun et al., 2022)) benchmarks. While XQUAD and MLQA are very focused and require the model to reason about the given context and answer the given question, MMLU, and BBH are much more

[2]open-source code is available on https://github.com/tloen/alpaca-lora

Instruction
Translate the following sentences from **English** to **German**.
Input
The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.
Output
Die Welt, wie wir sie geschaffen haben, ist ein Prozess unseres Denkens. Es kann nicht geändert werden, ohne unser Denken zu ändern.

Instruction
Translate the following sentences from **German** to **English**.
Input
Die Welt, wie wir sie geschaffen haben, ist ein Prozess unseres Denkens. Es kann nicht geändert werden, ohne unser Denken zu ändern.
Output
The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.

Table 2: Examples of translation-following demonstrations. In particular, in this example, there are two demonstrations with the same directions from English to German (en-x).

open but require the models' ability to solve logical mathematical tasks less related to the language.

However, we decided to introduce them to observe whether our approach degrades performance in these tasks. The first two datasets selected are appropriately constructed for multi-language testing, while the second two are available only in English. Hence, we do a preliminary translation step as outlined below. Thus, descriptions of the benchmarks follow in the next paragraphs:

MultiLingual Question Answering (MLQA) (Lewis et al., 2020) evaluates cross-lingual question answering performance using 5K extractive QA instances in the SQuAD (Rajpurkar et al., 2016) format in several languages. MLQA is highly parallel, with QA instances aligned across four languages on average. Although comprising different languages, some languages, such as Italian, are not represented. To conduct the experiments uniformly, we have translated the examples as also done in the forthcoming MMLU and BBH.

Cross-lingual Question Answering Dataset (XQUAD) (Artetxe et al., 2019) consists of a subset of 240 paragraphs and 1190 question-answer pairs from the development set of SQuAD v1.1 (Rajpurkar et al., 2016) with their manual translations into several languages. Consequently, the dataset is entirely parallel across 11 languages.

Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) measures knowledge of the world and problem-solving problems in multiple subjects with 57 subjects across STEM, humanities, social sciences, and other areas. The benchmark is native in English; however, we translated it into five additional languages[3].

BIG-Bench Hard (BBH) (Suzgun et al., 2022) is a subset of challenging tasks related to navigation, logical deduction, and fallacy detection. Again, the benchmark is native English, and we have translated it into five languages[??].

3.3.2 Models Setup & Evaluation

We used the alpaca_LoRA (Hu et al., 2021a) code[2], adopting the same hyperparameters to align the results with the state-of-the-art models.

We performed the fine-tuning with a single epoch and a batch-size of 128 examples, running our experiments on a workstation equipped with one Nvidia RTX A6000 with 48 GB of VRAM.

As an evaluation metric, we use accuracy. Hence, we estimate accuracy by measuring exact match values in the zero-shot setting. The parts of benchmarks related to the specific language are used for each model.

[3]We performed translations using the Google translator API from English to Chinese (zh), Italian (it), Arabic (ar), and German (de).

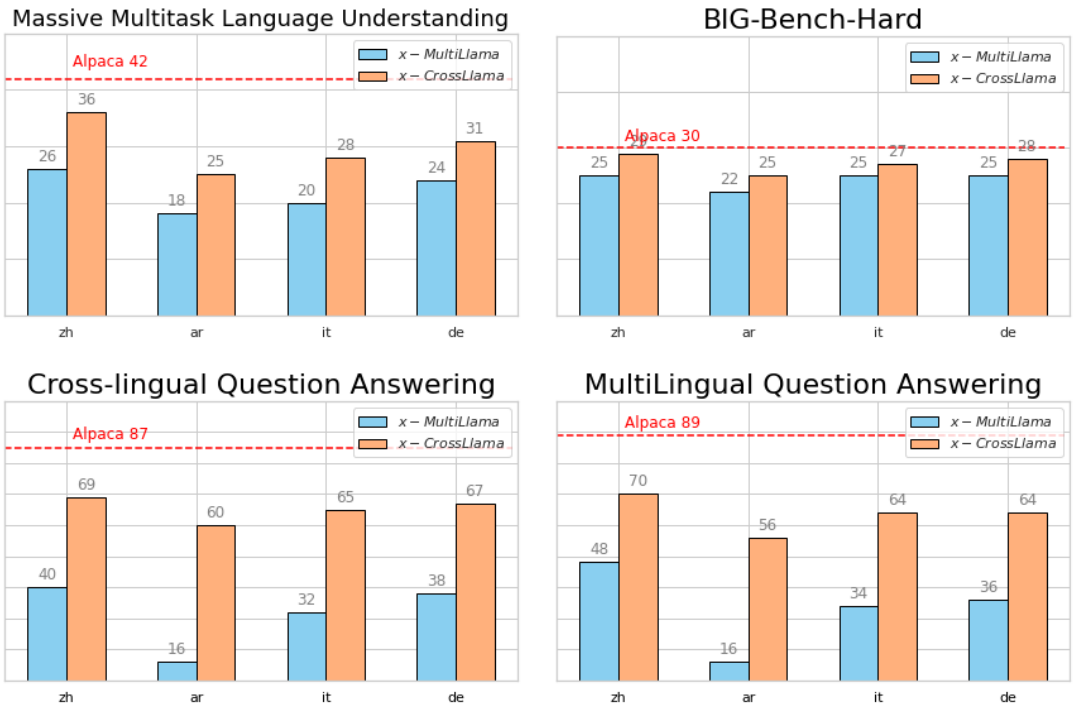

Figure 2: Accuracies (%) on proposed benchmarks. The dotted line represents the performance of the original version of Llama instructed on English data (Taori et al., 2023), which we call Alpaca.

4 Results & Discussion

Eliciting non-English abilities in instruction-tuned Large Language Models (It-LLMs) remains challenging. However, our *x-CrossLlama* revealed improved results in cross-lingual Question Answering (QA) benchmarks. Moreover, at the same time, the instructed models maintained logical-mathematical skills. From the results of Figure 2, it is possible to observe the weaknesses emerging from the fine-tuning of the translated versions of Alpaca (Section 4.1), the improvement obtained from the alignment phase is encouraging (Section 4.2) but it is not enough to outperform the English one.

The fine-grained analysis highlighted the importance of cross-lingual alignment data and the critical issues with non-English data. This opens the way for new hypotheses regarding the imbalance of pre-training languages and learning abilities via instruction-tuning.

4.1 Alpacas problems on Translations

The Instruction-tuning task on LLMs, in our case, Llama-7b, is primarily pre-trained in English, and has implications for the derivated models. As shown in Figure 2, both MLQA and XQUAD benchmarks reveal a notable disparity, with an average point gap of 55 and 53, respectively, between the original tuned Llama-7b (called Alpaca) and the various x-MultiLlama. This discrepancy is attenuated in the case of MMLU and BBH, where the average gaps are 18 and 14 points. Hence, relying exclusively on translations of Alpaca-style demonstrations for instruction in various languages only sometimes yields optimal effects. However, models, for example, zh-MultiLlama and de-MultiLlama, have exhibited better performances. This variation may be attributed to the volume of pre-training data available for the respective languages and, consequently, the inherent abilities of Llama. In future work, we aim to expand our analysis to include LLMs beyond Llama to see if similar, less pronounced, or more accentuated trends emerge.

QA Task	en-Llama	avg-Llama	avg-CrossLlama	δ
MLQA	*0.89*	0.34	0.64	+0.30
XQUAD	*0.97*	0.31	0.65	+0.30
MMLU	*0.42*	0.24	0.32	+0.08
BBH	*0.30*	0.24	0.28	+0.04

Table 3: Averages accuracies on proposed benchmarks.

Figure 3: Accuracies (%) of proposed benchmarks using one-direction Translation-following demonstrations. For en-x for English-foreigner and x-en for foreign English.

4.2 A Cross-lingual solution

Using the translation-following demonstrations close to instruction-following ones during instruction-tuning significantly empowers the cross-lingual performances of It-LLMs. In fact, *x-CrossLlama* consistently surpassed the x-MultiLlama, obtaining an improvement of 30 average points on MLQA, 34 on XQUAD, 8 on MMLU, and 4 on BBH, as detailed in Table 3. This approach brought their performance metrics closer to the benchmark set by the original version of Llama (Alpaca), bridging the gap in different situations. For MMLU and BBH, the performance difference was even more marginal, with average gaps of 10 and 2 points, respectively, as indicated in Table 3 and the 'en-Llama vs avg-CrossLlama'.

The inclusion of translation-following demonstrations has undeniably elevated the cross-lingual abilities of It-LLMs. Moreover, specific models, specifically the Chinese and German, surpassed the Arabic version by a significant margin. This disparity might be attributed to the varied representation of corpora within the pre-training datasets, as highlighted in (Yang et al., 2023). Consequently, cross-lingual strategies might not yield as pronounced benefits for underrepresented languages during the initial pre-training stages of the language model.

In conclusion, our strategy shifted to be high-performance and sustainable. As regards the performances, as merely discussed following the systematic analysis, we found empirical evidence to support this statement. While sustainability, our method uses a limited number of demonstrations, around 20k, which, combined with those of Alpaca, around 52k, remain a meager number, allowing the downstream models to obtain performances comparable to those of more robust models.

4.3 Ablation Study

Our *CrossLlama*, distinguished by the construction of the demonstrations pairs presented in Section 3.2, achieves significant performance improvements and contributes to closing the gap between the original version of tuned Llama and a series of x-MultiLlama in different languages. We propose an additional analysis. Working on the translation-following part (defined by half en-x and half x-en demonstrations), we analyze the impact of the demonstrations by splitting the experiments into en-x and x-en (Section 4.3.1).

4.3.1 Demonstration Direction matters

The evaluations in Figure 3 shed light on the impact of varying the directionality of translation-following demonstrations. In particular, demonstra-

tions that transition from English to a non-English language (en-x) appear to have a more pronounced positive effect on subsequent models. On the other hand, demonstrations transitioning from a foreign language to English (x-en) exhibit superior performance compared to baseline models, yet they lag behind when juxtaposed with demonstrations in the reverse direction.

However, as further illustrated in Figure 3, the x-CrossLlama consistently maintains its edge in performance. The observed trend, where translation-following demonstrations in one specific direction seem more influential, is intriguing. Mirroring our prior ablation analysis observations, multi-task benchmarks do not exhibit substantial variances. This observation lends further credence to the hypothesis that cross-lingual capabilities predominantly influence models in tasks heavily imbued with natural language elements.

5 Related Work

In the NLP field, multilingual and cross-linguistic methods have solid foundations and a long-standing tradition, with in-depth studies on feature adaptation (Section 5.1). However, the new Large Language Models (LLMs) no longer require such interventions. After extensive pre-training on massive corpora, cross-linguistic skills are inherently present in LLMs (Section 5.2 and Section 5.3). Nevertheless, although these abilities appear embedded, most LLMs must be elicited to show them exhaustively. Our study introduces a method to empower these cross-linguistic abilities through a cross-linguistic semantic alignment approach (Section 5.4).

5.1 Multilingual Pre-training

The next token prediction based on the prefix sequence, also well-known as language modeling, is the everlasting task of modern NLP (Tenney et al., 2019). The profound linguistic knowledge embedded within today's Large Language Models (LLMs) depends on the billions of neurons trained on large-scale corpora with derivatives of the language modeling task (Zanzotto et al., 2020; Ranaldi et al., 2022). Consequently, the pre-training corpora are predominantly in English, e.g., BooksCorpus (Zhu et al., 2015), MEGATRON-LM (Shoeybi et al., 2019), Gutenberg Dataset (Lahiri, 2014) therefore, LLMs usually have much better knowledge of English than other languages.

Researchers like Aulamo and Tiedemann (2019); Abadji et al. (2022) have proposed forward corpora translated into multiple languages to address this linguistic imbalance. However, these translated datasets, while valuable, are not as voluminous as their English-focused counterparts. The absence of extensive parallel data in these pre-training corpora further hinders the ability of LLMs to align and understand diverse languages effectively (Li et al., 2023).

5.2 Instruction-tuning Paradigm

Ouyang et al. (2022); Wei et al. (2022) fine-tuned LLMs using the instruction-tuning method based on instruction-tuning data, which are instruction-response corpora, to make LLMs more scalable and improve zero-shot performance. In this method, the LLM backbone is fed with data from the instruction (I, X, Y), where I is an instruction describing the task's requirements, X is the input, which can be optional, and Y is the output for the given task. The method aims to minimize the function $f(Y)$ based on the $\log$ likelihood with model parameters θ.

Earlier studies show that the instruction-tuning method of LLMs with both human (Wang et al., 2023) and synthetic-generated instructions (Taori et al., 2023; Xu et al., 2023) empowers the ability of LLMs to solve considerable tasks in zero-shot scenarios.

However, we state that the generally used instruction-tuning datasets, alpaca (Taori et al., 2023), Self-Instruct (Wang et al., 2023), Self-Chat (Xu et al., 2023), conceived in English, which limits the prospect of LLMs to follow non-English instructions and therefore solve related tasks.

5.3 Instruction-tuning is at hand

While Large Language Models (LLMs) have achieved remarkable outcomes using prevalent techniques like instruction-tuning, their vastness limits the breadth of the scientific community that can actively experiment with them.

Recent innovations aimed at democratizing access to these models and techniques focus on optimizing parameter tuning. One such method, Parameter-Efficient Tuning (PEFT), strategically adjusts a subset of the model's parameters while keeping the rest static. The overarching objective is to substantially curtail computational and storage overheads without compromising the performance exhibited by the original models (Ranaldi et al., 2023b). Established methodologies under the

PEFT umbrella include LoRA (Hu et al., 2021b), Prefix Tuning (Li and Liang, 2021), and P-Tuning (Liu et al., 2022). The fundamental principle behind these techniques is to retain the weights of the pre-trained model and integrate low-rank matrices at each architectural layer. This strategy considerably diminishes the parameter count that necessitates training for subsequent tasks, thereby enhancing efficiency. Such foundational advancements play a pivotal role in leveling the playing field for the scientific community, eliciting equitable research opportunities, and catalyzing the proliferation of open-source contributions.

5.4 Multilingual Instruction-tuning

Recent studies have highlighted the impressive capabilities of LLMs in assimilating instructions across diverse languages. Researchers such as Santilli and Rodolà (2023); Chen et al. (2023) have ventured into monolingual fine-tuning of Llama, focusing on instructions translated specifically to each language. Adopting optimization techniques elaborated further in Section 5.3, to design bespoke adapters tailored for various tasks has gained momentum. In exploring the cross-lingual potential of It-LLMs, Zhang et al. (2023) emphasized the benefits of enhancing instruction demonstrations.

In this paper, we propose *CrossLlama*, with a series of It-LLMs models with the Llama-7b backbone as the common denominator. The factor of our method is based on the inclusion of translation-following demonstrations that elicit semantic alignment between languages. We present empirical evidence underscoring the expansive cross-lingual learning prowess of It-LLMs. Through evaluations of four benchmarks, we demonstrate that the inherent limitations of It-LLMs can be effectively mitigated using cross-lingual alignment strategies when trained on non-English data. Consequently, our investigation seeks to elucidate the significance of instruction-following and translation-following demonstrations in bridging the linguistic divide, thereby enhancing the adaptability of LLMs to languages beyond English.

6 Future Works

The multilingual abilities of instruction-tuned Large Language Models (It-LLMs) are supported by LLMs, as seen with the Llama backbone in Alpaca's instance. Interestingly, small data-level stimuli improve downstream skills. Our experiments yielded significant insights when introducing strategic demonstrations, specifically translation-following demonstrations. We achieved these outcomes by fine-tuning Llama-7b, following the approach used in Taori et al. (2023).

In subsequent research, we aim to delve deeper by extending the number of parameters in Llama and integrating more backbone models. We are also intrigued by the potential effects on languages with limited resources. Furthermore, we aspire to fully understand the results from specific experiments by applying epistemic approaches (Ranaldi et al., 2023a,c) to It-LLMs.

In parallel, plans include analyzing the translation abilities of general It-LLMs and those empowered with translation tasks, including some specialized translation tasks among our evaluation benchmarks. Finally, we would like to investigate the learning abilities of the original Alpaca as the translation data changes, proposing different probing experiments on (original) English data enhanced with translations. Finally, we would like to investigate explainability techniques to understand better the underlying mechanisms, as done in (Ranaldi and Pucci, 2023), that enable these models to solve multiple tasks in complex scenarios using a small number of instances.

7 Conclusion

In this paper, we proposed *CrossLlama*, a novel methodology designed to empower the instruction-tuning of LLMs for non-English datasets. Our approach uniquely integrates instruction-following demonstrations, reminiscent of the Alpaca style, with translation-following demonstrations. The primary objective of this method is to elicit the LLM towards achieving semantic alignment between English and non-English languages, thereby outperforming models that are instructed using non-English texts. Leveraging the proposed demonstrations led to marked performance enhancements across four Question Answering benchmarks: XQUAD, MLQA, MMLU, and BBH. Furthermore, the depth of semantic alignment amplifies with the direction of the translation data, underscoring the inherent abilities of It-LLMs to assimilate from instruction-following demonstrations. Our innovative approach and the ensuing findings pave the way for advanced research, eliciting the development of more adept LLMs tailored for non-English linguistic contexts.

Limitations

Although the performance achieved by our method is consistently superior to that of several Instruction-tuned on custom corpora, our work has limitations:

- The proposed method was only analyzed on the Large Language Model Llama-7b; consequently, we can only report the results. We intend to extend our work using larger and different models in future developments.
- Although the proposed method performed well, it is only sometimes applicable as it requires an additional data set, the translation-following set.
- Finally, a significant limitation is that it is impossible to conduct correlations between the composition percentages of the training data and the downstream results, as the corpora used for pre-training are not always accessible, and the technical reports do not essay precise estimations.

Ethical Statement

This work used open-source corpora that do not deal with hate speech or inequality topics. The evaluation phase was also done on solid benchmarks commonly used for evaluation in Large Language Models. Finally, the concept of 'disparity' in the multilingual abilities of the Large Language Models in this work is understood as unbalancing the pre-training data used in the training phase.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. In *Annual Meeting of the Association for Computational Linguistics*.

Mikko Aulamo and Jörg Tiedemann. 2019. The OPUS resource repository: An open package for creating parallel corpora and machine translation services. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 389–394, Turku, Finland. Linköping University Electronic Press.

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.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Wei-Lin Chen, Cheng-Kuang Wu, and Hsin-Hsi Chen. 2023. Traditional-chinese alpaca: Models and datasets. `https://github.com/ntunlplab/traditional-chinese-alpaca`.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021a. Lora: Low-rank adaptation of large language models.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021b. Lora: Low-rank adaptation of large language models.

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.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André F. T. Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages.

Shibamouli Lahiri. 2014. Complexity of Word Collocation Networks: A Preliminary Structural Analysis. In *Proceedings of the Student Research Workshop at the*

14th Conference of the European Chapter of the Association for Computational Linguistics, pages 96–105, Gothenburg, Sweden. Association for Computational Linguistics.

Hang Le, Juan Miguel Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 817–824. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2023. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2021. Pre-training multilingual neural machine translation by leveraging alignment information.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models.

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 Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Federico Ranaldi, Elena Sofia Ruzzetti, Felicia Logozzo, Michele Mastromattei, Leonardo Ranaldi, and Fabio Massimo Zanzotto. 2023a. Exploring linguistic properties of monolingual berts with typological classification among languages.

Leonardo Ranaldi, Aria Nourbakhsh, Arianna Patrizi, Elena Sofia Ruzzetti, Dario Onorati, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2022. The dark side of the language: Pre-trained transformers in the darknet.

Leonardo Ranaldi and Giulia Pucci. 2023. Knowing knowledge: Epistemological study of knowledge in transformers. *Applied Sciences*, 13(2).

Leonardo Ranaldi, Elena Sofia Ruzzetti, Davide Venditti, Dario Onorati, and Fabio Massimo Zanzotto. 2023b. A trip towards fairness: Bias and de-biasing in large language models.

Leonardo Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2023c. Precog: Exploring the relation between memorization and performance in pre-trained language models.

Andrea Santilli and Emanuele Rodolà. 2023. Camoscio: an italian instruction-tuned llama.

Mohammad Shoeybi, Mostofa Ali Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training multi-billion parameter language models using model parallelism. *ArXiv*, abs/1909.08053.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Marc Tanti, Lonneke van der Plas, Claudia Borg, and Albert Gatt. 2021. On the language-specificity of multilingual bert and the impact of fine-tuning.

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`.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Martin Thissen. 2023. Fine-tune alpaca for any language. `https://github.com/thisserand/alpaca-lora-finetune-language`.

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

Evaluation (LREC'12), pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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.

Jason Wei, Maarten Bosma, Vincent Y. 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.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

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

Yasbok. Alpaca Instruction Fine-Tuning for Arabic. https://huggingface.co/Yasbok .

Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online. Association for Computational Linguistics.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison.

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

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.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.