Empowering Cross-lingual Abilities of Instruction-tuned Large Language Models by Translation-following Demonstrations

Leonardo Ranaldi (†), **Giulia Pucci** (⋆), **André Freitas**(†,∗)
(†) Idiap Research Institute, Martigny, Switzerland
(⋆) Department of Computing Science, University of Aberdeen, UK
(∗)Department of Computer Science, University of Manchester, UK
[first_name].[last_name]@idiap.ch

Abstract

The language ability of Large Language Models (LLMs) is often unbalanced towards English because of the imbalance in the distribution of the pre-training data. This disparity is demanded in further fine-tuning and affecting the cross-lingual abilities of LLMs. In this paper, we propose to empower Instruction-tuned LLMs (It-LLMs) in languages other than English by building semantic alignment between them. Hence, we propose *CrossAlpaca*, an It-LLM with cross-lingual Instruction-following and Translation-following demonstrations to improve semantic alignment between languages. We validate our approach on the multilingual Question Answering (QA) benchmarks XQUAD and MLQA and adapted versions of MMLU and BBH. Our models, tested over six different languages, outperform the It-LLMs tuned on monolingual data. The final results show that instruction tuning on non-English data is not enough and that semantic alignment can be further improved by Translation-following demonstrations.

1 Introduction

Large Language Models (LLMs) achieve comprehensive language abilities through pre-training on large corpora (Brown et al., 2020; Touvron et al., 2023; Ranaldi and Pucci, 2023b). Hence, the acquired language abilities follow the corpora features, mostly available in English (Lin et al., 2021; Pucci and Ranaldi, 2024). This phenomenon produces an imbalance in both pre-training (Blevins and Zettlemoyer, 2022) and fine-tuning (Le et al., 2021). Consequently, LLMs underperform in languages different from English (Huang et al., 2023).

Efforts to increase multilingual abilities propose continuing pre-training with large-scale monolingual data (Imani et al., 2023; Cui et al., 2023; Yang et al., 2023). However, learning a language from monolingual data requires considerable data and computational resources.

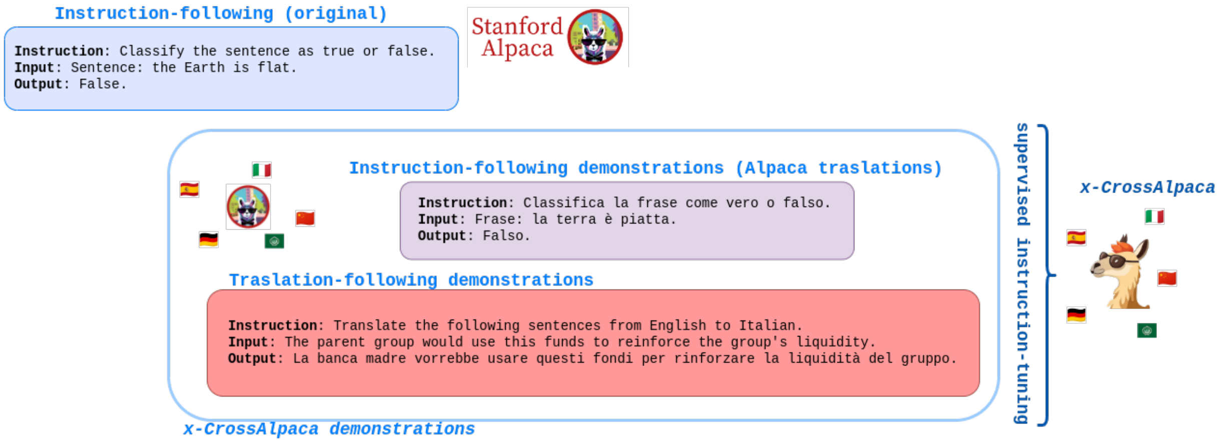
In this paper, we propose *CrossAlpaca*, an Instruction-tuned LLM (It-LLM) empowered with a semantic alignment between English and other languages. We analyze how to elicit the non-English abilities of It-LLMs by focusing on the crucial phase: instruction-tuning over Instruction-following demonstrations. Hence, we study the effect of cross-lingual alignment by proposing Translation-following demonstrations to improve the instruction-tuning phase.

In our experiments, we use *LLaMA-7B* (Touvron et al., 2023) as the LLM backbone and consider six target languages selected from among the available data (shown in Table 4); where data are absent, we perform a translation task. As Instruction-following demonstrations, we use the Stanford Alpaca dataset (Taori et al., 2023) and translated versions in the corresponding languages, while for the Translation-following, we use a publicly available translation resource news_commentary (Tiedemann, 2012), the most accessible and extendable to multiple languages (i.e., *CrossAlpaca demonstrations* on Figure 1).

Behind an instruction-tuning phase, we evaluate the performance of our six language-specific *CrossAlpacas* using four different benchmarks: two native-multilingual, i.e., XQUAD (Artetxe et al., 2019) and MLQA (Lewis et al., 2020), and two native-monolinguals, MMLU (Hendrycks et al., 2021) and BBH (Suzgun et al., 2022). The results show that *CrossAlpacas* instructed with language-specific instruction and translation data far outperform Alpacas instructed only with non-English demonstrations. However, even though *CrossAlpaca* reduces the gap, the Translation-following models of the original Alpaca still perform best. This result shows that LLaMA's learning abilities on English data are superior to those on non-English data. The semantic alignment task also improves the cross-lingual abilities of It-LLMs.

Our findings can be summarized as follows:

Findings of the Association for Computational Linguistics: ACL 2024, pages 7961–7973
August 11-16, 2024 ©2024 Association for Computational Linguistics


Instruction-following (original)
Stanford Alpaca
Instruction-following demonstrations (Alpaca traslations)
Traslation-following demonstrations
x-CrossAlpaca demonstrations
supervised instruction-tuning
x-CrossAlpaca

Instruction: Classify the sentence as true or false.
Input: Sentence: the Earth is flat.
Output: False.

Instruction: Classifica la frase come vero o falso.
Input: Frase: la terra è piatta.
Output: Falso.

Instruction: Translate the following sentences from English to Italian.
Input: The parent group would use this funds to reinforce the group's liquidity.
Output: La banca madre vorrebbe usare questi fondi per rinforzare la liquidità del gruppo.

Figure 1: Our *x-CrossAlpacas* are fine-tuned on Instruction-following and Translation-following demonstrations. This example shows the *it-CrossAlpaca*, fine-tuned on it-Alpaca and Translation-following demonstrations.

- The learning abilities of LLMs on non-English Instruction-tuning tasks are limited;
- The multi-lingual abilities of Instruction-tuned LLMs could be empowered through cross-lingual alignment;
- Thus, we propose *CrossAlpaca*, an Instruction-tuning approach for non-English models that are based on Instruction-following and Translation-following demonstration for target language;
- We show that It-LLMs are able to semantically align through cross-lingual Translation-following demonstrations via an extensive evaluation on four multi-lingual QA benchmarks.

2 Related Work

2.1 Pre-trained Language Model on Cross-Lingual Corpora

The next token prediction based on the prefix sequence, as well-known as language modeling, is the everlasting task of modern NLP (Tenney et al., 2019). The extensive knowledge of today's Large Language Models (LLMs) depends on the billions of neurons trained on large-scale corpora with derivatives of the language modeling task. 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) PILE (Gao et al., 2020), C4 (Dodge et al., 2021), RefinedWeb (Penedo et al., 2023); therefore, LLMs usually have much better knowledge of English than other languages.

Aulamo and Tiedemann (2019); Abadji et al. (2022), in order to solve this problem, propose forward corpora translated into several languages. However, these corpora are not as huge as their competitors, and the lack of massively parallel data in the pre-training corpora also prevents LLMs from aligning the different languages well (Li et al., 2023).

2.2 The Instruction-tuning Paradigm

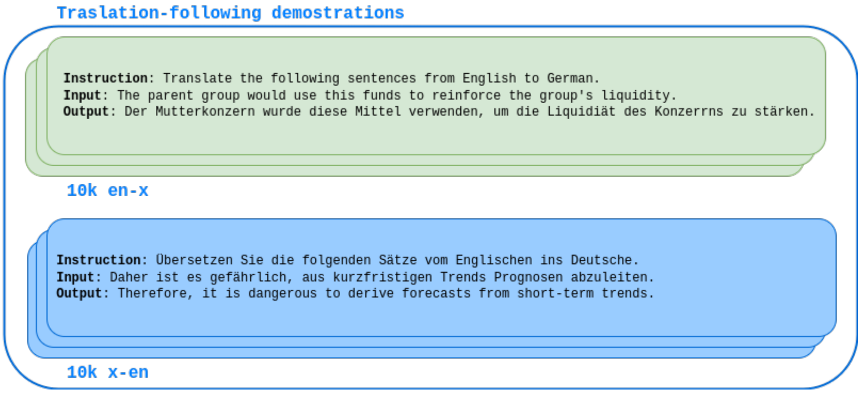
Ouyang et al. (2022); Wei et al. (2022) fine-tuned LLMs using the Instruction-tuning method based on Instruction-tuning data, which are instruction-response corpora, to make LLMs more scalable and improve zero-shot performance. In this method, the LLM backbone is fed with data from the instruction (I, 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 goal of this method is to minimize the function $f(Y)$:

$$f(Y) = \arg\min_{\theta} \log p_{\theta}(Y \mid I, X) \qquad (1)$$

where θ are model learnable parameters.

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) or a teacher-student alignment approaches (Ranaldi and Freitas, 2024), conceived



Instruction: Translate the following sentences from English to German.
Input: The parent group would use this funds to reinforce the group's liquidity.
Output: Der Mutterkonzern wurde diese Mittel verwenden, um die Liquidiät des Konzerrns zu stärken.

Instruction: Übersetzen Sie die folgenden Sätze vom Englischen ins Deutsche.
Input: Daher ist es gefährlich, aus kurzfristigen Trends Prognosen abzuleiten.
Output: Therefore, it is dangerous to derive forecasts from short-term trends.

Figure 2: Examples of Translation-following demonstrations. In this particular example, there are two demonstrations with two different directions from English to German (en-x) and vice versa.

in English, which limits the prospect of LLMs to follow non-English instructions and therefore solve related tasks.

2.3 Instruction-tuning is at hand

Large Language Models have considerable success with many techniques in vogue, such as Instruction-tuning. However, their prohibitive size only allows part of the scientific community to experiment with these models.

The latest advances that make these models and techniques accessible involve efficient parameter tuning. Parameter-Efficient Tuning (PEFT) is an efficient technique to adjust a small part of the model parameters and freeze the others. The main goal is to significantly reduce computational and storage costs while maintaining the performance of the original models. The standard techniques for PEFT are LoRA (Hu et al., 2021a), Prefix Tuning (Li and Liang, 2021), P-Tuning (Liu et al., 2022). The basic idea is to keep the pre-trained model weights and incorporate low-rank matrices in each architecture layer. This approach significantly reduces the number of parameters that require training for subsequent tasks, thereby increasing overall efficiency. This building block has been and continues to be very important because it skills the scientific community to fair research and enables the development of multiple open-source works.

2.4 Cross-lingual Instruction-tuning Challenge

Recent work has shown the remarkable abilities of LLMs in learning instruction in different languages. Santilli and Rodolà (2023); Chen et al. (2023) proposed monolingual fine-tuning LLaMA adaptations on language-specific translated instructions. The use of optimization techniques introduces in Section 2.3, to propose engaging custom adapters for various tasks. Zhang et al. (2023); Ranaldi and Pucci (2023a) investigated the cross-lingual abilities of It-LLMs by promoting the effects of augmenting instruction demonstrations. In particular, Zhang et al. (2023) use fine-grained custom translations that are not easily accessible, while Ranaldi and Pucci (2023a) propose a multilingual approach by considerably increasing instruction and translation tasks. However, these studies have opened an interesting avenue in analyzing the cross-lingual abilities of It-LLMs.

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

Table 1: Details and names of mono-lingual Instruction-tuned Large Language Models that use a language-specific version of Alpaca as instruction-tuning data.

In this paper, we take the next step by proposing our *CrossAlpaca*. Our methods show of the cross-lingual learning abilities of It-LLMs on a large scale. Using four well-known benchmarks, we show that the weaknesses of It-LLMs trained on non-English data can be strengthened with cross-lingual alignment approaches. Hence, our analyses aim to understand the role of Instruction-following and Translation-following demonstrations in closing the gap in learning and adapting LLMs' abilities to non-English languages.

3 Data & Methods

Pre-training from scratch a Large Language Model (LLM) in multiple languages is cost-prohibitive for data collection and parameter learning. This is the reason why the trend is to do further fine-tuning to strengthen the models' abilities in a specific language (Tanti et al., 2021; Moslem et al., 2023). In this paper, we aim to exploit the abilities of pre-trained LLMs for non-English languages by further improving the alignment between English and the target language. Hence, in Section 3.1, we introduce the challenge of fine-tuning an LLM on a mono-lingual custom scenario. Furthermore, in Section 3.2, we propose our approach to fine-tune cross-lingual LLMs.

3.1 Mono-lingual Instruction-tuning

The limited accessibility and transparency of paid API services of state-of-the-art LLMs have pushed research toward developing open-source models. Using the instruction-tuning paradigm (Section 2.2) and Stanford Alpaca (Taori et al., 2023), a corpus consisting of 52k of English instruction-output pairs generated by `text-davinci-003`, several Instruction-tuned versions of LLaMA were released.

Following this approach, multiple mono-lingual Instruction-tuned versions of LLaMA were proposed by translating the Stanford Alpaca Instruction-following data into the specific language. Table 4 (Appendix B) shows a set of alpaca versions available as open source. Following a systematic analysis of the translated versions of Alpaca 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 languages that share the most already available data. Table 1 shows the custom versions used in this work, which for simplicity will be renamed *x*-Alpaca, where *x* indicates the specific language.

3.2 Cross-lingual Instruction-tuning

Mono-lingual Instruction-tuning approaches (Section 3.1) reward LLMs' multi-lingual abilities. However, the use of translated Alpaca alone for the specific language is not sufficient to elicit the non-English ability of LLMs. For this reason, we propose *CrossAlpaca* (Figure 1), which enriches the cross-lingual Instruction-tuning with the Translation-following demonstrations. We aim to highlight LLMs' English and non-English abilities by proposing a semantic alignment task to enrich versions of Alpaca.

Instruction-following demonstrations The original version of Alpaca is in English. However, as described in Section 3.1, different open-source translations are produced with a translation engine. In our experiments, we propose the Instruction-tuning phase with the original English version (en-Alpaca) and the translated language-specific versions. Moreover, we propose the *CrossAlpaca* versions built with the different Alpacas translated into the specific language and the cross-lingual translations (introduced later). In this way, we investigate the abilities of the LLM backbone in understanding multilingual instructions and cross-lingual alignment.

Translation-following demonstrations Learning general instruction data is a sensible strategy for building models to solve multi-tasks directed by instructions (Wang et al., 2023; Zeng et al., 2023). However, translation data could contribute to learning semantic alignment. Zhang et al. (2023) showed that the translation performance of LLMs could be improved by using expertly annotated translation data, and Zhu et al. (2023) showed that it could be beneficial for instruction-tuning.

Inspired by Zhu et al. (2023), 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 20k demonstrations. Hence, following the Alpaca style (Instruction, Input, and Output) (see Figure 1), we randomly selected 10k English to non-English translations and 10k random non-English to English translations (all translations come from news_commentary).We have constructed these correspondences for the languages mentioned above in Section 3.1. Figure 2 shows a case of Translation-following for English-German direction (en-de) and vice versa (de-en).

4 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 *CrossAlpaca*. Our approach is based on instruction-tuning on

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

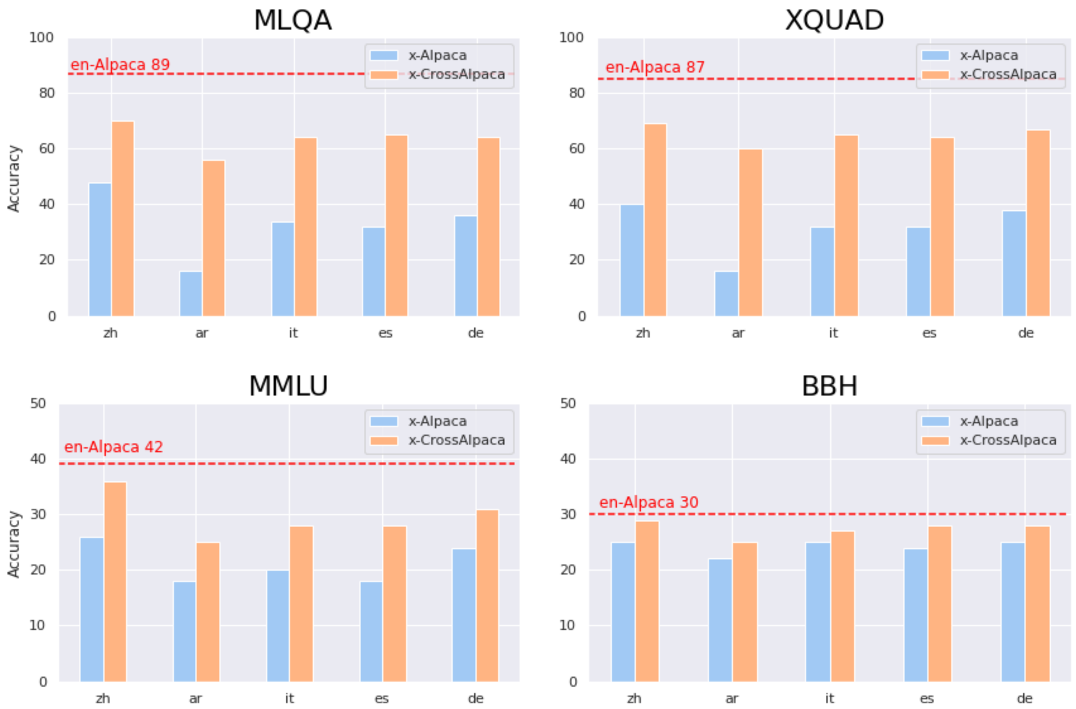


Figure 3: Evaluation results on proposed benchmarks. The dotted line represents the performance of the original Alpaca (Taori et al., 2023) on English tasks.

language-specific data augmented with a cross-lingual semantic alignment. Hence, we set several baseline models (Section 4.1), which we augmented with our *CrossAlpaca* approach (Section 4.2). Finally, we performed a series of systematic evaluations (Section 4.3.1) to observe the impact of the proposed intervention.

4.1 Baseline Instruction-tuned 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-Alpaca* for *x* specific languages: Chinese (zh), Italian (it), Arabic (ar), Spanish (es), German (de) and the original English version (en).

4.2 Cross-lingual Instruction-tuned LLMs

The *CrossAlpacas* are instruction-tuned on Instruction-following and Translation-following demonstrations (*CrossAlpacas demonstrations*). The first ones stem from the resources introduced in Section 4.1. The second comes from *news_commentary* (Tiedemann, 2012).

Our approach generates a series of instruction-tuned versions of the data shown in Figure 1. We have named the versions *x-CrossAlpaca* where *x* denotes Chinese (zh), Arabic (ar), Italian (it), Spanish (es), and German (de).

4.3 Experimental Setup

In order to assess the performance of the *CrossAlpaca*, we defined several benchmarks (Section 4.3.1) on which we applied systematic tuning (Section 4.3.2) and evaluation (Section 4.3.3) pipelines.

4.3.1 Benchmarks

To evaluate the performance of the It-LLMs and the impact of the translation-based 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. XQUAD and MLQA focus on understanding questions and answers through translation into different languages. MMLU and BBH, being multi-task benchmarks, include subtasks related to Boolean expressions and QA on basic-level subjects (e.g., chemistry, physics). However, we decided to introduce them to observe whether our approach degrades performance in these tasks. The first two

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

datasets selected are appropriately constructed for multi-language testing, while the second two are available only in English. So we do a preliminary translation step as outlined below.

MultiLingual Question Answering (MLQA) (Lewis et al., 2020) evaluatates cross-lingual question answering performance. The benchmark comprises over 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 are not represented, such as Italian. 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. Here again, the benchmark is native English, and we have translated it into five languages[3].

4.3.2 Models Setup

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

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

4.3.3 Evaluation

As an evaluation metric, we use accuracy. Hence, we estimate accuracy by measuring exact match values in the zero-shot setting. For each model, the parts of benchmarks related to the specific language are used (e.g., for zh-Alpaca and zh-CrossAlpaca, data from MLQA, XQUAD, MMLU, and BBH in Chinese are used).

5 Results

Improving non-English abilities in Instruction-tuned Large Language Models (It-LLMs) remains challenging. However, the *x-CrossAlpacas* revealed improved results in cross-lingual Question Answering (QA) benchmarks and maintained logical-mathematical skills. From the results of Figure 3 (further detailed in Table 3), it is possible to observe the weaknesses emerging from the fine-tuning of the translated versions of Alpaca (Section 5.1), the improvement obtained from the alignment phase is encouraging (Section 5.2) but it is not enough to outperform the English one. Therefore, we investigated the impact of demonstrations on downstream performance (Section 5.3).

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.

5.1 Translating the Alpaca is not the right way

The instruction-tuning on LLaMA, predominantly pre-trained in English, affects x-Alpaca. Figure 3 and in terms of numbers, Table 3 show that in both MLQA and XQUAD, there is a gap of 55 and 53 average points between en-Alpaca and the x-Alpacas. This phenomenon is mitigated for MMLU and BBH, where we observed an average gap of 18 and 14 points. Instructing an LLM on Alpaca-style demonstrations translated into different languages is not always a good strategy. However, some x-Alpacas, such as zh-Alpaca and de-Alpaca, have performed better. We hypothesize that this phenomenon is related to the scale of the pre-training data in the respective languages and, thus, the abilities of the LLaMA. In future developments, we plan to extend the study on other LLMs beyond LLaMA to observe whether the phenomenon is similar, milder, or more significant.

[3] We performed translations using the Google translator API from English to Chinese (zh), Italian (it), Arabic (ar), Spanish (es), German (de). Resources will be made available along with the publication

QA Task	en-Alpaca	avg-Alpaca	avg-CrossAlpaca	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 2: Averages of the results on proposed benchmarks. The column δ indicates the difference between avg-Alpacas and our avg-CrossAlpacas.

5.2 *CrossAlpaca*: A cross-lingual solution

Semantic alignment through Translation-following demonstrations during fine-tuning could have a valuable impact on the cross-lingual abilities of It-LLMs. The *x-CrossAlpacas* outperformed the x-Alpacas by 30 average points on MLQA, 34 average points on XQUAD, 8 on MMLU, and 4 on BBH (see Table 2 and more detailed in Table 3). They also brought their performances closer to the sota obtained from the original Alpaca by 25 points on MLQA and 22 average points on XQUAD. In MMLU and BBH, the gap became very close, with averages of 10 and 2 points (see Table 2 or the line 'en-Alpaca vs avg-CrossAlpaca' in Table 3).

Enriching Translation-following demonstrations has outstanding influences on the cross-lingual abilities of the It-LLMs. However, even in this case, Chinese and German models (zh- and de-CrossAlpaca) outperformed Arabic by many points and, in some specific cases, Spanish and Italian as well. This phenomenon, we hypothesize, is related to the diversity in corpus representation within the pre-training data, as shown in (Yang et al., 2023). Therefore, cross-lingual approaches do not have an incisive impact as in languages less present in the pre-training phases of the language model.

5.3 Ablation Study

Our *CrossAlpacas*, distinguished by the construction of the demonstrations pairs (Section 4.2), achieves significant performance improvements and contributes to closing the gap between the original Alpaca (en-Alpaca) and a series of x-Alpacas in different languages. In order to show the impact of enrichment with cross-lingual demonstrations, we propose two different analyses. In the first analysis, we incrementally decrease the training data, in particular on the Translation-following demonstrations side (Section 5.4). In the second, still 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 5.4.1).

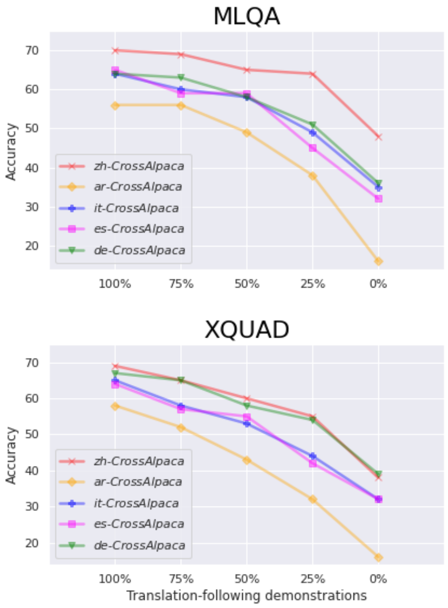


Figure 4: Evaluation of proposed benchmarks of the demonstrations used for instruction-tuning our CrossAlpacas.

5.4 Translation-following demonstrations empower non-English abilities.

Figure 4 (for a more extensive analysis please refer to Figure 6 in the Appendix) shows the non-English performance of the x-CrossAlpaca on the benchmarks under different scales of Translation-following demonstrations[4]. The x-CrossAlpacas achieve higher accuracy when more demonstrations are used, revealing the benefits of cross-lingual alignment to improve non-English performance. However, Translation-following demonstrations are built in two directions: x-en and en-x (foreign-English and English-foreign). Although equally distributed, we are still determining whether there is an asymmetry between the two types of contributions.

Finally, this result is less pronounced for the multi-task benchmarks (see Figure 6 MMLU and BBH). This could be related to the fact that in the benchmarks, there are logical-mathematical subtasks which are less language dependent (comparatively simpler syntax, smaller vocabularies), and hence, our cross-lingual approach does not influence the results for MLQA and XQUAD as

[4] The Traslation-following demonstrations are equally selected in a random way from the 10k en-x and x-en.

much. Furthermore, it is possible to observe a trend in language groups, particularly Chinese-German, Italian-Spanish, and Arabic. This trend, which recurs frequently in the results of our experiments, will be further investigated in future work.

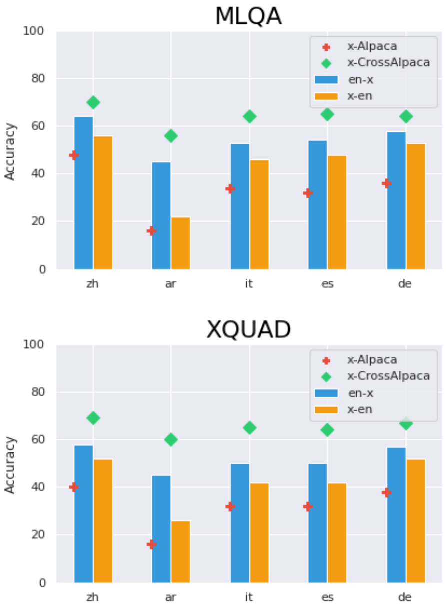


Figure 5: Evaluation of proposed benchmarks using one-direction Translation-following demonstrations. For en-x for English-foreigner and x-en for foreign English.

5.4.1 Demonstration direction matters on non-English abilities

Figure 5 shows the evaluation conducted in varying directions of the translation-following demonstrations. In particular, demonstrations with English-foreign direction (en-x) better impact downstream models. Conversely, foreign-English (x-en) demonstrations perform better than baselines but underperform demonstrations in the opposite direction. However, as shown in Figure 5 (and complete Figure 7), the x-CrossAlpacas continue to outperform. Nevertheless, the trend of translation-following demonstrations with one direction is interesting. Again, as was the case in the previous ablation study, the multi-task benchmarks (see MMLU and BBH in Figure 7) do not seem to obtain significant influences, which reinforces the hypothesis that the models are greatly affected by cross-lingual capabilities in tasks where there is a strong presence of natural language.

6 Future Works

The cross-lingual abilities of Intruction-tuned Large Language Models (It-LLMs) seem to be supported by LLMs, such as in the case of Alpaca, the LLaMA backbone. However, it seems that low-impact demonstrations at the data level can enrich these abilities. We obtained valuable results from our experiments by proposing strategic demonstrations, i.e., Translation-following demonstrations. These results were proposed by performing fine-tuning on LLaMA-7B as done by Taori et al. (2023).

In future work, we would like to continue to investigate by increasing the number of parameters in LLaMA and including additional backbone models. In addition, it might be interesting to evaluate the impact on low-resource languages as done for in-context settings in (Ranaldi et al., 2024). Hence, we would like to get to the underside of performances obtained in some experiments (see Section 5.4) by extending previous epistemic approaches (Ranaldi et al., 2023a,c,b) to It-LLMs. In parallel, plans include analyzing the translation capabilities of general It-LLMs and those enhanced 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 change, proposing different probing experiments on (original) English data enhanced with translations.

7 Conclusion

In this paper, we proposed *CrossAlpaca*, an approach to empowering the instruction-tuning of LLMs on non-English data. Specifically, we coupled Instruction-following (Alpaca-style) demonstrations with Translation-following demonstrations. Our method seeks to instruct the LLM to semantic alignment between English and non-English overperforms models instructed on non-English texts. In particular, thanks to our *CrossAlpaca demonstrations*, the instructed models achieved significant performance improvements on four Question Answering benchmarks XQUAD, MLQA, MMLU, and BBH. In addition, we observe that semantic alignment strengthens with increasing Translation-following data; this demonstrates the de-facto abilities of It-LLMs to learn from instructions. Our approach and results contribute to improved research on the potential for producing more powerful LLMs for non-English languages.

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.

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.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

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.

Kohaku-Blueleaf. 2023. Guanaco-lora: LoRA for trainin Multilingual Instruction-following LM based on LLaMA. https://huggingface.co/plncmm/guanaco-lora-7b .

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.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Giulia Pucci and Leonardo Ranaldi. 2024. Does the language matter? curriculum learning over neo-Latin languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5212–5220, Torino, Italia. ELRA and ICCL.

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 and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Leonardo Ranaldi, Aria Nourbakhsh, Elena Sofia Ruzzetti, Arianna Patrizi, Dario Onorati, Michele Mastromattei, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2023b. The dark side of the language: Pre-trained transformers in the DarkNet. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 949–960, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Leonardo Ranaldi and Giulia Pucci. 2023a. Does the English matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183, Singapore. Association for Computational Linguistics.

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

Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2024. Empowering multi-step reasoning across languages via tree-of-thoughts.

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

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, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages.

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)*.

A Appendix

	Model	**MLQA**	**XQUAD**	**MMLU**	**BBH**
	en-Alpaca	*0.89*	*0.87*	*0.42*	*0.30*
Alpaca	zh-Alpaca	**0.48**	0.38	**0.26**	**0.25**
	ar-Alpaca	0.17	0.16	0.18	0.20
	it-Alpaca	0.35	0.32	0.21	0.25
	es-Alpaca	0.32	0.33	0.19	0.24
	de-Alpaca	0.36	**0.39**	0.24	0.25
	***avg*-Alpaca**	0.34	0.31	0.24	0.24
CrossAlpaca	zh-CrossAlpaca	**0.70**	**0.69**	**0.36**	0.28
	ar-CrossAlpaca	0.56	0.60	0.25	0.25
	it-CrossAlpaca	0.64	0.65	0.28	0.27
	es-CrossAlpaca	0.65	0.64	0.28	0.28
	de-CrossAlpaca	0.64	0.67	0.32	**0.29**
	***avg*-CrossAlpaca**	0.64	0.65	0.32	0.28
	en-Alpaca vs *avg*-Alpaca	0.34***(-0.55)***	0.31***(-0.56)***	0.24***(-0.18)***	0.24***(-0.06)***
	en-Alpaca vs *avg*-CrossAlpaca	0.64***(-0.25)***	0.65***(-0.22)***	0.32***(-0.10)***	0.28***(-0.20)***
	***avg*-CrossAlpaca vs *avg*-Alpaca**	***(+0.30)***	***(+0.34)***	***(+0.08)***	***(+0.04)***

Table 3: Evaluation results on proposed benchmarks. The ***avg-*** lines are the averages of x-Alpacas and x-CorssAlpacas. The last three lines indicate the comparisons between en-Alpaca. The last line indicates the comparisons between avg-Alpacas and avg-CrossAlpacas.

B Appendix

Language	Alpaca	MLQA	XQUAD	MMLU	BBH	News_commentary
Arabic	*x*	*x*	*x*	-	-	*x*
Chinese	*x*	*x*	*x*	-	-	*x*
English	*x*	*x*	*x*	*x*	*x*	*x*
German	*x*	*x*	*x*	-	-	*x*
Greek	-	*x*	*x*	-	-	*x*
Hindi	*x*	*x*	*x*	-	-	*x*
Italian	*x*	-	-	*x*	*x*	*x*
Russian	*x*	*x*	*x*	-	-	*x*
Spanish	*x*	*x*	*x*	-	-	*x*
Turkish	*x*	-	*x*	-	-	-
Vietnamese	-	*x*	*x*	-	-	-

Table 4: List of available state-of-the-art resources.

C Appendix

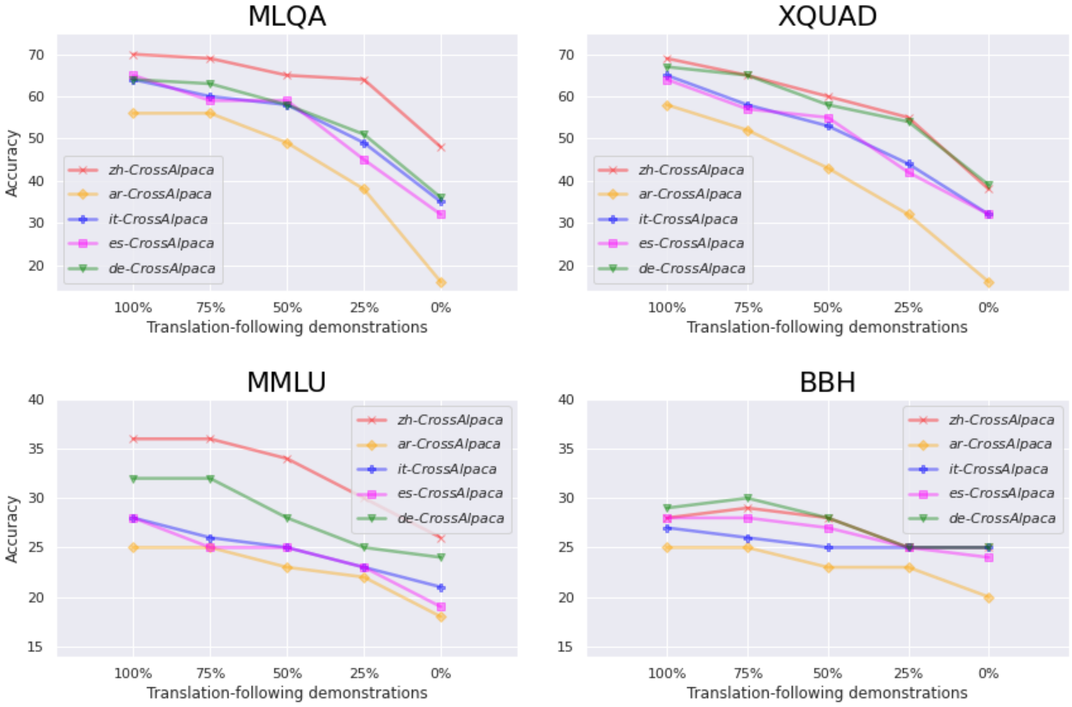


Figure 6: Evaluation of all proposed benchmarks of the demonstrations used for instruction-tuning our CrossAlpacas.

D Appendix

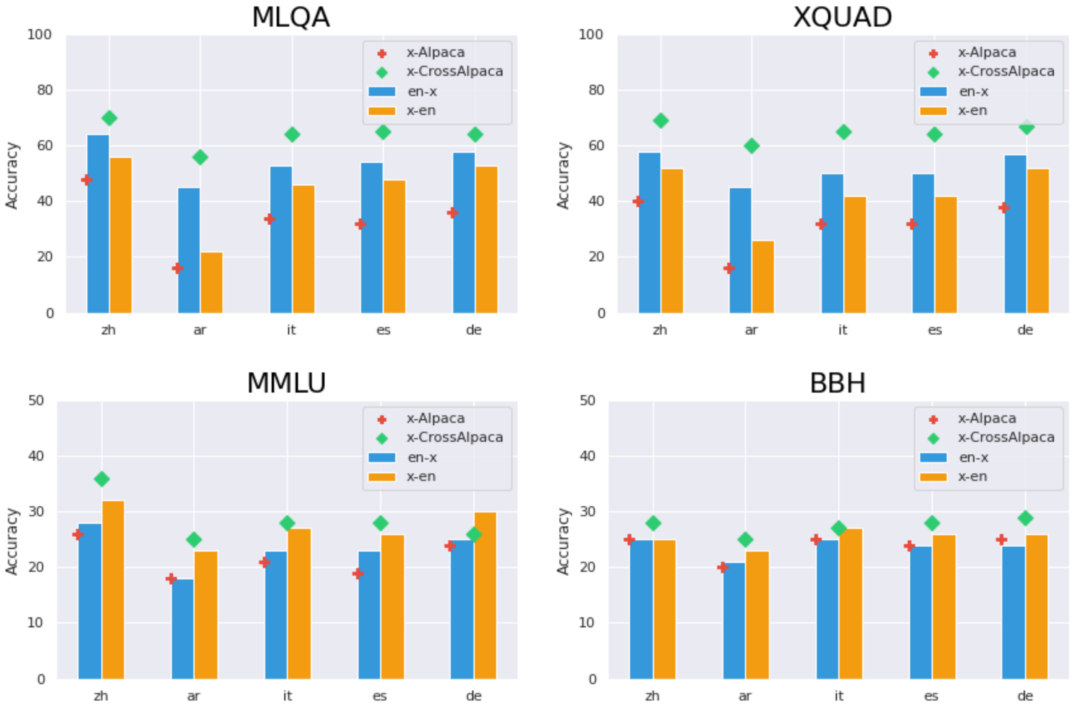


Figure 7: Evaluation of all proposed benchmarks using one-direction Translation-following demonstrations. For en-x for English-foreigner and x-en for foreign English. With the red cross, we indicate the results of the x-Alpaca standards, and with the green diamond, the results of our x-CrossAlpaca.