Multilingual Pretraining and Instruction Tuning Improve Cross-Lingual Knowledge Alignment, But Only Shallowly

Changjiang Gao[1] **Hongda Hu**[1] **Peng Hu**[1] **Jiajun Chen**[1] **Jixing Li**[2] **Shujian Huang**[1]*
[1]National Key Laboratory for Novel Software Technology, Nanjing University
[2]Department of Linguistics and Translation, City University of Hong Kong
{gaocj, huhd, hup}@smail.nju.edu.cn chenjj@nju.edu.cn
jixingli@cityu.edu.hk huangsj@nju.edu.cn

Abstract

Despite their strong ability to retrieve knowledge in English, current large language models show imbalance abilities in different languages. Two approaches are proposed to address this, i.e., multilingual pretraining and multilingual instruction tuning. However, whether and how do such methods contribute to the cross-lingual knowledge alignment inside the models is unknown. In this paper, we propose CLiKA, a systematic framework to assess the cross-lingual knowledge alignment of LLMs in the Performance, Consistency and Conductivity levels, and explored the effect of multilingual pretraining and instruction tuning on the degree of alignment. Results show that: while both multilingual pretraining and instruction tuning are beneficial for cross-lingual knowledge alignment, the training strategy needs to be carefully designed. Namely, continued pretraining improves the alignment of the target language at the cost of other languages, while mixed pretraining affect other languages less. Also, the overall cross-lingual knowledge alignment, especially in the conductivity level, is unsatisfactory for all tested LLMs, and neither multilingual pretraining nor instruction tuning can substantially improve the cross-lingual knowledge conductivity. [1]

1 Introduction

The language imbalance of modern NLP systems has long been discussed (Hupkes et al., 2023). Many studies have shown that the performance of current LLMs on English tasks is much higher than non-English tasks (Wang et al., 2023; Zhang et al., 2023c). One possible explanation is that the knowledge required for completing tasks are learnt mainly from English text. So it could be better retrieved with English than with other languages.

Recent studies suggest that cross-lingual consistency may be a possible way to narrow the gap between languages (Qi et al., 2023). Ideally, if the knowledge of a fact could be aligned to a "true" representation regardless of the language it is described with, it may be retrieved in any required language, helping the model to generalize across languages. In this paper, we refer to this internal mechanism as *cross-lingual knowledge alignment*.

To improve LLMs' performance in non-English languages, two approaches are proposed. The first is *multilingual pretraining*, which add non-English data in the pretraining corpus. The second is *multilingual instruction tuning*, i.e., using tasks in different languages or translation-related tasks, to finetune a foundation model (Zhang et al., 2023a; Zhu et al., 2023). Although these methods do improve LLMs' non-English performance, whether they can bring real cross-lingual knowledge alignment is not well investigated.

Therefore, the aim of this study is to evaluate the effect of multilingual pretraining and instruction tuning on the cross-lingual knowledge alignment mechanism. However, the evaluation is challenging, because the improvement of performance may come from the improvement of language ability in a specific language or the improvement of knowledge alignment. It is hard to discriminate the effects of the two by performance as the single clue. Furthermore, even if LLMs show higher consistency between two languages, there is still possibility that the knowledge in the two languages are learned correctly but separately.

To meet this challenge, we propose to assess cross-lingual knowledge alignment systematically, by using 3 deepening levels of measurement:

- **Performance** (PF): achieving similar performance for tasks in different languages;
- **Consistency** (CT): generating the same output for the same input in different languages;
- **Conductivity** (CD): retrieving knowledge learned in one language with another.

*Corresponding author
[1]Our code and data are available on GitHub.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6101–6117

Most previous evaluations of the multilingualism only focus on the PF level (Kassner et al., 2021; Yin et al., 2022) and the CT level (Qi et al., 2023), but the CD level is closer to the nature of knowledge alignment.

The evaluation for the CD level is non-trivial, because the successful retrieval of a factual knowledge learned in another language depends not only on the alignment of the knowledge, but also on the basic language ability in the current language. For example, even if the alignment is correct, retrieving this knowledge in non-English languages such as Japanese is harder than doing it in English, because the model's basic ability is not as good.

In this regard, we propose a systematic framework, CLiKA (standing for Cross-Lingual Knowledge Alignment), to reveal the effects of different multilingualism. CLiKA considers all three levels of the alignment, with specific metrics for each level. CLiKA includes three comparative settings: *Factual*, *Basic*, and *Fictional*, to further discriminate the effects of language abilities and knowledge alignment in knowledge retrieval.

We apply CLiKA to popular LLMs, including BLOOM (Workshop et al., 2023), LLaMA (Touvron et al., 2023a,b), ChatGPT and their multilingual variants (Cui et al., 2023; Zhang et al., 2023a). Our results indicate that:

- The general cross-lingual knowledge alignment of current multilingual LLMs is unsatisfactory. They show imbalanced basic abilities and knowledge PF in English and non-English, and their high knowledge CT comes with low CD, suggesting low cross-lingual knowledge conduction.
- Mixed multilingual pretraining improves the basic ability, knowledge PF and CT in multiple languages, while continued pretraining can only improve the knowledge PF in the target language at the cost of other languages. However, both of them cannot improve the knowledge CD of LLMs.
- Multilingual instruction tuning improves the basic ability in the target language, and can lower the knowledge PF drop brought by instruction tuning. However, it can hardly improve the knowledge CT and CD.

2 Related Work

Multilingualism of language models. Due to the "incident bilingualism" (Briakou et al., 2023) and cross-lingual data sharing (Choenni et al., 2023) in the training corpus, pretrained models, including those English-centered ones, will have multilingual ability and cross-lingual alignment of representations to some extent. On that basis, multilingualism can be further strengthened by adding monolingual data in different languages in the pretraining corpus, resulting in multilingual PLMs such as mBERT (Devlin et al., 2019), mBART (Liu et al., 2020), mT5 (Xue et al., 2021). However, because the multilingual data is not parallel in these models, their language balance and cross-lingual knowledge alignment is much limited (Pires et al., 2019). To address this issue, some work uses supervised parallel data in the pretraining stage to enhance the model's multilingualism, e.g. XLM (Conneau and Lample, 2019) and BLOOM (Workshop et al., 2023). Such method is also used nowadays to train LLMs with better multilingualism, bringing models such as PaLM 2 (Anil et al., 2023) on multiple languages; and ChatGLM (Du et al., 2022; Zeng et al., 2022) and Baichuan 2 (Yang et al., 2023a) mainly focusing on English and Chinese. Also, another popular practise is to fine-tune an English-centered foundation model with translation and instruction data in English and other languages (Cahyawijaya et al., 2023), resulting in models with better translation ability and instruction-following ability in those languages, such as BigTranslate(Yang et al., 2023b), BayLing (Zhang et al., 2023a), x-LLaMA/m-LLaMA (Zhu et al., 2023), and mFTI (Li et al., 2023). However, despite the performance gain on multilingual benchmarks, the effect of these training methods on cross-lingual knowledge alignment is still to be examined.

Multilingual benchmarks and evaluations. Evaluation work is rapidly updating in the field of cross-lingual knowledge alignment. In the PLM era, many cross-lingual NLP benchmark datasets were proposed to test the models' performance on certain aspects in different languages, such as XCOPA (Ponti et al., 2020) and X-CSQA (Lin et al., 2021) for commonsense reasoning, and X-FACTR (Jiang et al., 2020) and multilingual versions of LAMA (Kassner et al., 2021; Yin et al., 2022; Qi et al., 2023) for factual knowledge. Some work also tested LLMs' multilingual performance on different NLP tasks (Lai et al., 2023; Zhang

Knowledge	Dataset	Example
Basic	xCSQA	**Question:** The dental office handled a lot of patients who experienced traumatic mouth injury, where were these patients coming from? A. town B. michigan C. hospital D. schools E. office building **Answer:** C. hospital
	xCOPA	**Premise:** The item was packaged in bubble wrap. **Question:** What was the cause of this? A. It was fragile. B. It was small. **Answer:** A. It was fragile.
Factual	xGeo	**Question:** What administrative division of Egypt is Alexandria in? A. Red Sea Governorate B. Alexandria Governorate C. Cairo Governorate D. Emirate of Dubai **Answer:** B. Alexandria Governorate
	xPeo	**Question:** In what year was Houari Boumediene born? A. 1820 B. 1828 C. 1838 D. 1932 **Answer:** D. 1932
Fictional	Translation	**Question:** Could you convert the upcoming English text to German? Tempest Hollow **Answer:** Sturmhain
	QA	**Question:** Which continent is Tempest Hollow located in? **Answer:** Vividora

Table 1: Example of questions used in each of the testing datasets.

et al., 2023b; Ahuja et al., 2023).

Knowledge misalignment of language models. Previous work have pointed out the imbalance of multilingual pretrained language models (PLMs) (Pires et al., 2019; Qi et al., 2023). However, since the "incident multilingualism" in pertraining increased a lot for LLMs, this conclusion needs to be reevaluated. Zhang et al. (2023c) found that ChatGPT does not perform consistently on tasks in different languages, while exhibiting a translation-like thinking mode. Wang et al. (2023) concluded that multilingually-trained models have not attained "balanced multilingual" capabilities, especially on commonsense or factual knowledge. However, they did not differentiate between the two sources of language misalignment. Qi et al. (2023) further evaluated the cross-lingual consistency of PLMs and the factors affecting it using a rank-based metric. However, an evaluation with all three levels of cross-lingual knowledge alignment considered is yet to be done.

3 Methods

Because the cross-lingual knowledge retrieval is affected by both the basic language ability and the three levels of cross-lingual knowledge alignment, our CLiKA framework adopts special testing datasets and metrics to evaluate them separately.

3.1 Constructing testing data

We constructed three testing datasets: *Basic*, *Factual*, and *Fictional*. The datasets are all in the multiple choice format for easier evaluation. Also, the data is parallel in the same 10 chosen languages: `en`, `de`, `fr`, `it`, `ru`, `pl`, `ar`, `he`, `zh`, `ja` (Details are listed in Table 13 in Appendix A). These languages are chosen because they are widely used, and they enable comparison between and within language families.[2] Table 1 shows the example questions.

***Basic* knowledge.** We consider commonsense as *Basic* knowledge to measure the basic language ability of models in the selected languages. There are two reasons for this. Firstly, commonsense is indispensable for LLMs to generate meaningful answers, so if a model lacks commonsense in some languages, it will be very likely to show poor overall ability in these languages. Secondly, because commonsense is so elementary, that they are unlikely to be explicitly stated in any text in any language (Lenat, 1995), it is difficult to be learned through short-cuts such as remembering training samples. The two parts of this dataset are:

- xCOPA (500 samples per language). COPA (Roemmele et al., 2011) is an English dataset focusing on commonsense causality, where each question is a 1-out-of-2 choice. Although there is already a cross-lingual version of COPA, i.e. XCOPA (Ponti et al., 2020), it does not cover the languages considered in this study. Instead, we use DeepL and Google Translate to translate the COPA `test` set into the other 9 languages.
- xCSQA (1000 samples per language). CSQA (Talmor et al., 2019) is a challenging English dataset focusing on the semantic relation of

[2]The datasets will be publicly available along with the publication of this paper.

common concepts, where each question is a 1-out-of-5 choice. There is also a cross-lingual version of this dataset, i.e. X-CSQA (Lin et al., 2021). However, we still use the updated translation of the `val` set for higher quality.

***Factual* knowledge.** This represents the real-life knowledge retrieval scenario, and is deliberately balanced among the tested languages, i.e., the knowledge originates evenly from the 10 languages, and is presented parallelly in all of them. Currently, there is no off-the-shelf dataset that meets the requirements. The dataset contains two parts originating from Wikidata:

- xGeo (200 samples per language), about cities and the administrative division they belong. For each of the 10 languages, we choose 20 cities in the major countries speaking this language (see Table 13 in Appendix A), and collect their names, and their administrative divisions' names in the 10 languages with WikiData. Then, we construct a 1-out-of-4 choice for each city-division with 3 randomly picked wrong options. After that, we use templates in the 10 languages (see Appendix C) to write the questions. There are thus 200 samples presented in each language in total, 20 for each original language.

- xPeo (180 samples per language), about famous people and their years of birth/death (YOBs/YODs). For each language, we choose 10 famous historical figures from the major countries speaking this language. [3] Then, we collect their YOBs, DOBs, and names in the 10 languages from WikiData. We again construct a 1-out-of-4 choice for each person-year with 3 randomly picked wrong options, and then use templates in the 10 languages (see Appendix C for all templates) to write the questions. Specially, the xPeo dataset does not contain historical figures originating from Hebrew, because they are either with multiple nationalities, or are contemporary.

This *Factual* dataset can be used to evaluate the PF and CT level of cross-lingual knowledge alignment, but it cannot accurately measure the CD level, because even though we have identified the language origins of the factual knowledge, which language is the knowledge first learned in a model, i.e. the source of conductivity, is unknown. To help measure CD, we then construct the *Fictional* knowledge dataset to test the knowledge conductivity from English to other languages.

***Fictional* knowledge.** This dataset consists of artificial entity-relation knowledge, which do not exist in the training of LLMs, making it possible to observe the learning and transferring of knowledge in different languages. While the entity names and their translations in all the 10 languages are provided for training, the tested relations between entities are only presented in English. Therefore, to answer the non-English relations, the models need to conduct knowledge from English to non-English. The dataset is built by the following steps:

1. Names of 400 fictional places and 20 fictional continents are generated in English and translated into the other 9 languages by ChatGPT as the entities. Then, 10 translation templates (see Appendix C) are used to construct the translation training data from English to the other 9 languages (4200 samples per language).

2. Each place is randomly assigned to a continent to build relations between the entities ($20 \times 20 = 400$ relations). Then, all the relations are filled in 10 English QA templates (see Appendix C) and used as the first part of the training data (4000 samples, English only); Meanwhile, half of the relations are filled in the QA templates in the other 9 languages and respectively used as the second part of the training data (2000 samples per language). Note that in each conductivity experiment, only one non-English presents in the training data.

3. The other half of the relations excluded in the non-English training data are filled in the first template ("Which continent is {PLACE} located in?") and used as the testing data (200 samples per language).

The tested models will be tuned with LoRA (Hu et al., 2021) on the two instruction sets, and the performance on the test set is collected.

3.2 CLiKA measurements

The basic ability is measured with the *Basic* knowledge, while the PF and CT alignments are mea-

[3] Cases of multiple nationalities and unclear YOBs/YODs are ruled out.

sured with the *Factual* knowledge, and the CD aligment is measured on the *Fictional* knowledge. We design three measurements to score these aspects.

PF: Re-scaled accuracy (RA). The raw accuracy is affected by the question difficulty, and there exists a random baseline for multiple choice questions, making it hard to compare the performance across languages and aspects of model ability. Thus, to focus on the difference across languages and models, we re-scale the accuracy. Suppose the raw accuracy of a model in one language is A, we use the accuracy of ChatGPT in English (noted A_g) as a reference for difficulty, and the expected accuracy of random choice $A_r = 1/n_{\text{choice}}$ as the baseline. The re-scaled accuracy (RA) is calculated as:

$$\text{RA} = \frac{\max\{A - A_r,\ 0\}}{\max\{A_g - A_r, 0\}}$$

Note that RA can exceed 1 as long as the raw accuracy is higher than that of ChatGPT in English. Also, since ChatGPT performs better than random on almost all tested tasks, the denominator is larger than 0. More balanced RAs in English and non-English means better PF alignment.

CT: Correct prediction overlap with English (en-CO). Similar to Jiang et al. (2020), we use the ratio of consistent and correct predictions between English and another language to measure their CT. Suppose the model gives n_X correct answers in language X, and among them, $n_{\text{en}X}$ are consistent with its answers in English, the en-CO is:

$$\text{CO}(\text{en}, X) = \frac{n_{\text{en}X}}{n_X}$$

The en-CO ranges from 0 to 1, higher value meaning better CT with English.

CD: Cross-retrieval ratio (XRR). Suppose n_{en} is the number of correct answers in English and $n_{\text{en}X}$ is correct in both English and language X on *Fictional* knowledge, and A_r is the random accuracy baseline (0.05 for the *Fictional* dataset). The cross-retrieval ratio (XRR) is then calculated as:

$$\text{XRR}(X) = \max\left\{\frac{n_{\text{en}X}}{n_{\text{en}}} - A_r, 0\right\}$$

XRR is non-negative and can exceed 1, higher value meaning better CD from English to another language.

Mixed	Cont'	Model
N	N	LLaMA
N	Y	Chinese-LLaMA
Y	N	Baichuan2-base, LLaMA2
Y	Y	Chinese-LLaMA2

Table 2: List of foundation models used in the Chinese case study. "Mixed" stands for mixed pretraining in Chinese, and "Cont'" stands for continued pretraining in Chinese.

PT	FT	Model
N	N	Alpaca
N	Y	BayLing
Y	N	LLaMA2-Chat
Y	Y	Vicuna v1.5

Table 3: List of instruction-tuned LLMs used in the Chinese case study. "PT" stands for whether the model has Chinese pretraining, and "FT" stands for whether the model has Chinese instruction tuning.

4 Experiment Settings

4.1 Models

Our CLiKA analysis of cross-lingual knowledge alignment is two-fold. First, we assess the basic language ability and cross-lingual knowledge alignment of popular multilingual LLMs among all tested languages; then, we examine the effect of multilingual pretraining and instruction tuning on their basic language ability and cross-lingual knowledge alignment, taking Chinese as a representative for high-resource, non-English language.

The **popular multilingual LLMs** we selected are ChatGPT (called with the OpenAI API `gpt-3.5-turbo` in Octorber, 2023), LLaMA 2-Chat 70B and 13B (Touvron et al., 2023b), Vicuna v1.5 13B (Chiang et al., 2023) and BLOOMZ-7B1-MT (Workshop et al., 2023).

The models for **the Chinese case study** include foundation models (Table 2) and LLMs (Table 3) that allows comparison of having or not having continued/mixed pretraining and instruction tuning in Chinese (See their information in Appendix E).

4.2 Assisted inference

Because some of the models tested are not instruction tuned, directly given the question and options, they may not provide a valid choice. To help them inference, we: (1) use an in-context demonstration for every question to inform them the correct answering format (Figure 2); (2) force the models to generate only the options (e.g. from A to E).

4.3 Tuning strategy

Instruction tuning is needed in the CD evaluation, where instruction templates are required. For

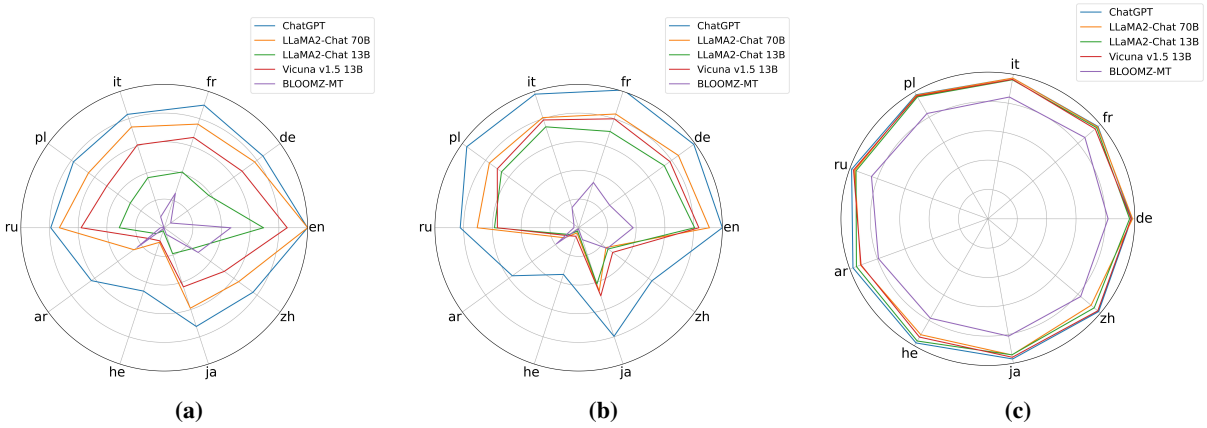

Figure 1: Results of the general cross-lingual knowledge alignment evaluation. The outer circle of the radar graphs is 1.0 and the center is 0.0, and each circle represents a 0.2 span. a) The RA scores on the *Basic* knowledge (the mean of xCSQA and xCOPA scores; b) The RA scores on the *Factual* knowledge (the mean of xGeo and xPeo scores); c) The en-CO scores on the *Factual* knowledge (the mean of xGeo and xPeo scores).

Model	de	fr	it	pl	ru	ar	he	ja	zh
LLaMA2-Chat 13B	.0800	.1200	.0900	.1050	.0050	.0003	.0000	.0000	.0000
Vicuna v1.5 13B	.1309	.0200	.0050	.0000	.0800	.0800	.0000	.0800	.0000
BLOOMZ-7B1-MT	.0200	.1350	.0350	.0656	.0050	.0000	.0000	.0100	.0000

Table 4: The XRR scores of representative LLMs on the *Fictional* knowledge. Scores below 0.01 (less than 1% above-random-accuracy) are colored red.

instruction-tuned LLMs, we use their own instruction templates; and for foundation models, we use the Alpaca template. Also, we use LoRA (Hu et al., 2021) on the attention blocks to lower the training cost.[4] The hyper-parameters and computational resource used in the experiments are listed in Appendix D.

5 Main Results

5.1 General cross-lingual knowledge alignment of multilingual LLMs

In this part, we assess the basic ability and the cross-lingual knowledge alignment of representative multilingual LLMs among the 10 tested langauges. The findings are:

Basic abilities: imbalanced. Figure 1a shows the models' RA scores on the *Basic* knowledge, which reflects the imbalance of basic abilities across different languages, which could be affected by language similarity and resources. For instance, `en`, `de`, `fr`, `it`, `pl` and `ru` belong to the Indo-European family, and they also witness better cross-lingual knowledge alignment with English; On the other hand, `ar` , `he`, `ja` and `zh` belong to other families, and are non-Latin. Among them, `ar` and `he` are also lower-resourced, so it is not surprising that the models show the worst performance on `ar` and `he`. Compared with ChatGPT, the LLaMA models and BLOOMZ show larger imbalance.

Factual knowledge alignment: imbalanced PF, but high CT. Figure 1b and 1c show the RA and en-CO on the *Factual* knowledge, corresponding to the PF and CT levels of cross-lingual knowledge alignment. One can see the RA scores are also imbalanced, especially in `zh`, where the factual knowledge performance is too low to match the basic ability. However, the en-CO scores are quite high in all languages, suggesting that the right answers given in non-English languages are very likely the same as English answers.

Factual knowledge alignment: low CD. There are two possible causes of the high CT: A. The knowledge is conducted from English to non-English; B. The non-English training data is a translated subset of the English training data. Here, our XRR results (Table 4) supports the latter. It shows the XRR scores are low across all non-English languages, especially in non-Latin languages. This result suggests that, the high English-CT revealed by the models are more likely an outcome of overlapping training data, instead of knowledge conductivity.

[4]We use the recommended setting in the LLaMA-Factory repository (see Appendix D).

Model	mixed	cont'	en	zh (/en)	others (/en)
LLaMA	**N**	**N**	50.36	15.98 (0.32)	17.95 (0.36)
Chinese-LLaMA	N	Y	29.38	21.12 (0.72)	5.88 (0.20)
LLaMA2	**Y**	**N**	73.29	43.57 (0.59)	35.78 (0.49)
Chinese-LLaMA2	Y	Y	60.93	32.32 (0.53)	22.59 (0.37)
LLaMA	**N**	**N**	50.36	15.98 (0.32)	17.95 (0.36)
LLaMA2	Y	N	73.29	43.57 (0.59)	35.78 (0.49)
Baichuan2-base	Y	N	87.58	59.30 (0.68)	38.61 (0.44)
Chinese-LLaMA	**N**	**Y**	29.38	21.12 (0.72)	5.88 (0.20)
Chinese-LLaMA2	Y	Y	60.93	32.32 (0.53)	22.59 (0.37)

Table 5: The comparison of the selected models' RA scores on the *Basic* knowledge (the mean of xCSQA and xCOPA scores), where "mixed" and "cont'" means having Chinese mixed or continued pretraining, "/en" means the ratio to the English scores, and "others" refers to the mean scores in the other 8 languages. The first lines in each division (LLaMA, LLaMA2, LLaMA and Chinese-LLaMA) are the baseline values (in black). The green values are higher than baseline, and the red ones are lower than baseline. (Same notations in below.)

Model	mixed	cont'	en	zh (/en)	others (/en)
LLaMA	**N**	**N**	79.28	7.49 (0.09)	42.19 (0.53)
Chinese-LLaMA	N	Y	44.23	13.01 (0.29)	15.44 (0.35)
LLaMA2	**Y**	**N**	91.58	40.39 (0.44)	56.00 (0.61)
Chinese-LLaMA2	Y	Y	79.84	45.92 (0.58)	43.70 (0.55)
LLaMA	**N**	**N**	79.28	7.49 (0.09)	42.19 (0.53)
LLaMA2	Y	N	91.58	40.39 (0.44)	56.00 (0.61)
Baichuan2-base	Y	N	86.34	65.81 (0.76)	50.51 (0.59)
Chinese-LLaMA	**N**	**Y**	44.23	13.01 (0.29)	15.44 (0.35)
Chinese-LLaMA2	Y	Y	79.84	45.92 (0.58)	43.70 (0.55)

Table 6: The comparison of the selected models' RA scores on the *Factual* knowledge (the mean of xGeo and xPeo scores).

Model	mixed	cont'	zh	others
LLaMA	**N**	**N**	.8327	.8975
Chinese-LLaMA	N	Y	.8597	.7764
LLaMA2	**Y**	**N**	.9648	.9498
Chinese-LLaMA2	Y	Y	.9536	.9276
LLaMA	**N**	**N**	.8327	.8975
LLaMA2	Y	N	.9648	.9498
Baichuan2-base	Y	N	.9410	.9453
Chinese-LLaMA	**N**	**Y**	.8597	.7764
Chinese-LLaMA2	Y	Y	.9536	.9276

Table 7: The comparison of the selected models' en-CO scores on the *Factual* knowledge (the mean of xGeo and xPeo scores).

Model	mixed	cont'	zh	others
LLaMA	**N**	**N**	.0000	.0443
Chinese-LLaMA	N	Y	.0204	.0436
LLaMA2	**Y**	**N**	.0153	.0570
Chinese-LLaMA2	Y	Y	.0050	.0337
LLaMA	**N**	**N**	.0000	.0443
LLaMA2	Y	N	.0153	.0570
Baichuan2-base	Y	N	.0000	.0421
Chinese-LLaMA	**N**	**Y**	.0204	.0436
Chinese-LLaMA2	Y	Y	.0050	.0337

Table 8: The comparison of the selected models' XRR scores on the *Fictional* knowledge.

5.2 Chinese case study on the effect of multilingual pretraining and finetuning

In this part, we show the effect of multilingual pretraining and instruction tuning by comparing the basic ability and the cross-lingual knowledge alignment of several selected models in Chinese.

5.2.1 Effect of multilingual pretraining

Mixed pretraining improves basic abilities, while continued pretraining does not. Table 5 shows the RA scores on the *Basic* knowledge. Comparing models with and without continued and mixed Chinese pretraining, one can see that mixed pretraining improves the models' basic language abilities in all languages, while continued pretraining has negative effect on them (even in Chinese). This suggest that continued pretraining in a certain language may not be as useful as adding the language in the mixed pretraining process, in order to enhance the model's overall basic language abilities.

Mixed pretraining improves PF and CT alignment. Table 6 and 7 show the RA and en-CO scores on the *Factual* knowledge. Similar to the *Basic* results, mixed pretraining improves the performance in all languages, as well as enhancing the English consistency of non-English languages. However, continued pretraining in Chinese only improves the Chinese performance, at the cost of lowering performance in other languages. Also, continued pretraining contributes little to the English consistency of non-English languages, including Chinese. This result suggests that, for mixed

Model	pre	tune	en	zh	others
Alpaca	**N**	**N**	+25.47	+6.70	+9.19
BayLing	N	Y	+25.24	+22.37	+8.01
LLaMA2 Chat	**Y**	**N**	-4.04	-18.97	-10.17
Vicuna v1.5	Y	Y	+12.51	+8.24	+10.17

Table 9: The change in RA scores on the *Basic* knowledge (the mean of xCSQA and xCOPA scores) after instruction tuning, compared with their foundation models (LLaMA and LLaMA2). "pre" and "tune" means whether the model has Chinese pretraining or instruction tuning, and "others" refers to the mean scores in the other 8 languages. (Same notations in below.)

Model	pre	tune	en	zh	others
Alpaca	**N**	**N**	-2.00	-2.98	+3.13
BayLing	N	Y	-13.67	+2.07	-9.70
LLaMA2-Chat	**Y**	**N**	-10.90	-15.13	-6.63
Vicuna v1.5	Y	Y	-7.94	-11.11	-2.30

Table 10: The change in RA scores on the *Factual* knowledge (the mean of xGeo and xPeo scores) after instruction tuning.

pretraining, the performance gain is spread in all languages, and the improvement of cross-lingual knowledge alignment is down to the consistency level; However, for continued pretraining, despite the surfacial performance gain in the trained language, it is risky of harming performance in other languages, and does not improve the cross-lingual knowledge alignment in deeper levels.

Multilingual pretraining hardly improves CD alignment. Table 8 shows the XRR scores on the *Fictional* knowledge. One can see that neither continued nor mixed pretraining can bring stable and significant increase to the XRR scores, meaning the knowledge conductivity from English to Chinese is still near zero after Chinese pretraining. This result suggest that current multilingual pretraining methods cannot improve the cross-lingual knowledge alignment in the CD level, which again supported the hypothesis that the high consistency between non-English and English found in current LLMs is an outcome of overlapping training data, not knowldege transfer from English.

5.2.2 Effect of multilingual finetuning

Multilingual finetuning improves basic abilities. Table 9 shows the RA scores of instruction-tuned LLMs on the *Basic* knowledge. Compared with English-only instruction tuning, adding Chinese data in the tuning process significantly improves the RA scores in Chinese, while not hurting the English performance. This results suggests that multilingual instruction tuning is suitable for fostering basic language abilities in non-English languages.

Model	pre	tune	zh	others
Alpaca	**N**	**N**	+.0800	+.0338
BayLing	N	Y	+.0456	-.0376
LLaMA2-Chat	**Y**	**N**	-.0203	+.0081
Vicuna v1.5	Y	Y	+.0116	+.0054

Table 11: The change in en-CO scores on the *Factual* knowledge (the mean of xGeo and xPeo scores) after instruction tuning.

Model	pre	tune	zh	others
Alpaca	**N**	**N**	+.0000	-.0034
BayLing	N	Y	+.0003	-.0078
LLaMA2-Chat	**Y**	**N**	-.0153	-.0070
Vicuna v1.5	Y	Y	-.0153	-.0076

Table 12: The change in XRR scores on the *Fictional* knowledge after instruction tuning.

Multilingual finetuning lowers performance drop in factual knowledge. Table 10 shows the RA scores of instruction-tuned LLMs on the *Factual* knowledge. Surprisingly, both English-only and multilingual instruction tuning causes drop in the RA scores, which indicates a performance drop in factual knowledge after the tuning. Since this phenomenon is not observed on the *Basic* knowledge, this cannot be explained by the "chat bot" preference, but may suggest a shared disadvantage of current instruction tuning strategies. However, compared with English-only tuning, multilingual tuning causes less damage to the factual knowledge performance, contributing to the PF level of cross-lingual knowledge alignment.

Multilingual finetuning can hardly improve CT or CD alignment. Table 11 and 12 shows the en-CO scores on the *Factual* knowledge and the XRR scores on the *Fictional* knowledge. One can see that the changes in the two scores brought by English-only and multilingual instructiong tuning are both minor, and multilingual instruction tuning shows no significant advantage over English-only tuning. This result suggests that instruction tuning cannot improve cross-lingual knowledge alignment deeper than the PF level.

6 Supplement Experiments

The main results of this paper has shown that the PF and CD levels of current open source LLMs are unsatisfactory, and the CT and CD of them cannot be substantially enhanced by multilingual pretraining or finetuning. However, the result of low CD from English to Chinese can also be due to the linguistic (e.g. lexical, morphological) difference between Chinese and English, or the deficiency of our

LoRA finetuning. Thus, adding CD experiments on another Indo-European language and using other finetuning strategies will make our findings more grounded.

6.1 German case study

To avoid the effect of low resource on CD, we choose German as a high-resource, Indo-European (Germanic) language to test the models' conductivity to it from English. We adopt a trending German LLM, LeoLM[5] (13B, base version), which is based on LLaMA2 and has gone through continued pretraining in German. Also, we compare LLaMA2 and Vicuna v1.5 for having or not having German instruction tuning. (See Appendix F.)

German continued pretraining harms basic ability and knowledge alignment. Table 23 shows the result of LLaMA2 and LeoLM-base on the *Basic* knowledge. Similar to the result of Chinese-LLaMA2, the German continued pretraining of LeoLM leads to the overall decline of basic ability in English, German and other languages. Also, from Table 24, 15 and 16, we can see the PF, CT and CD levels of cross-lingual knowledge alignment drop for all the tested languages after the continued pretraining. This is consistent with our findings in the Chinese case study.

German finetuning can improve basic ability and knowledge alignment. Table 17 shows the result of LLaMA2-chat and Vicuna v1.5 on the *Basic* knowledge, where we can see the multilingual (including German) instruction tuning of Vicuna v1.5 improves the basic ability in German and other languages. Then, from Table 18, 21 and 22, one can see that it lowers the performance drop in factual knowledge, slightly improves CT, and raises CD. Interestingly, the increase in CD is above chance level (+0.0709 XRR), which is not observed in the Chinese case. This suggest that improving the knowledge conductivity from English to a similar language may be easier than to a less similar one. However, the increased XRR score is only 0.13, which is still not satisfactory for a language like German.

6.2 Alternative finetuning strategies

Apart from LoRA tuning on attention blocks only, we add two other experiments using LoRA on all blocks on LLaMA-13B, and fully finetuning on LLaMA 7B. Besides, another experiment that adds extra translation data of the entities from non-English to English is performed to see whether the "reversal curse" (Berglund et al., 2023) of the unidirectional translation data causes the low XRR results. (See Appendix G.)

LoRA-all and fully finetuning cannot improve overall CD. Table 19 shows the comparison of LoRA-attention, LoRA-all and fully finetuning on the LLaMA 13B and 7B models. Although the XRR scores are improved in certain languages such as French, Italian and Polish, the overall improvement is minor, and the scores even drops in some other languages. This result shows the low CD results still hold with larger scale finetuning.

Adding reversed translation data cannot improve overall CD. Table 20 shows the comparison of XRR scores between using unidirectional or bidirectional translation data in the CD experiment on LLaMA2-13B. The results show that adding translation data from non-English to English does not significantly improve the XRR scores, thus the low CD results still hold.

7 Conclusion

In this paper, we evaluated the cross-lingual knowledge alignment of representative multilingual LLMs, and systematically assessed the effect of multilingual pretraining and instruction tuning on it, using the proposed CLiKA framework.

The first part of our results shows that the cross-lingual knowledge alignment of current multilingual LLMs is unsatisfactory, and even though they show high cross-lingual consistency, it is more likely to come from overlapping training data, instead of knowledge conduction between languages.

The second part of our results demonstrates the effect on basic language ability and knowledge alignment of adding multilingualism in pretraining and instruction tuning, which shows that mixed multilingual pretraining and multilingual instruction tuning is beneficial. However, our results also point out that neither of the two techniques can improve the knowledge conductivity of LLMs, meaning the cross-lingual alignment in current models are still shallow and requires novel strategies for improvement.

[5] https://huggingface.co/LeoLM/leo-hessianai-13b

Limitations

One key limitation of this paper is that the evaluation is restricted to several selected models, which may lead to over-simplification of models in a wider range. Also, the assessment of the effect of multilingual pretraining and instruction tuning only takes English and Chinese into account (and German for supplement), which only covers a narrow set of linguistic features (Littell et al., 2017) and cannot represent the whole picture of multilingual research. These two limitations are partly due to our computational resources and the lack of suitable models for comparison. With adequate resources, our CLiKA framework can be applied to more models and languages to further examine our findings.

Another limitation is that the *Fictional* knowledge requires 2-hop inference to conduct (one for translating the city name, the other for translating the continent name), which is consistent with questions in the xGeo dataset, but it may add too much difficulty of CD alignment, leading to underestimation of the models' knowledge conductivity. To address this issue, we did a small-scale experiment on LLaMA2-Chat using fictional city names and their founding years. The result shows that the XRR of de rises to 0.26, but that of zh is still very low (around 0.01), which suggests that the low conductivity issue is still existing in questions with lower difficulty.

Ethics Statement

The authors declare no competing interests. The datasets used in the evaluation come from publicly available sources and do not contain sensitive contents such as personal information. The adaptation and use of data (CSQA, COPA, Wikidata) are under their licenses. The data generated by ChatGPT other models are non-toxic and used for research only, which is consistent with their intended use.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by National Science Foundation of China (No. 62376116, 62176120), the Liaoning Provincial Research Foundation for Basic Research (No. 2022-KF-26-02), research project of Nanjing University-China Mobile Joint Institute.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on "a is b" fail to learn "b is a". In The Twelfth International Conference on Learning Representations.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. Instructalign: High-and-low resource language alignment via continual crosslingual instruction tuning.

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

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. How do languages influence each other? studying cross-lingual data sharing during llm fine-tuning.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. Advances in neural information processing systems, 32.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in NLP. Nature Machine Intelligence, 5(10):1161–1174.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5943–5959, Online. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3250–3258, Online. Association for Computational Linguistics.

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

Douglas B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. Commun. ACM, 38(11):33–38.

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.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning. 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 1274–1287. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8:726–742.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2362–2376. Association for Computational Linguistics.

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

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In 2011 AAAI Spring Symposium Series.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F. Chen. 2023. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng

Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023a. Baichuan 2: Open large-scale language models.

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

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. In The Eleventh International Conference on Learning Representations.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu,

Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023a. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023b. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023c. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms.

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.

A Language choice

Table 13 shows the languages tested in CLiKA.

ISO	Countries	Langauge Family
en de	US, UK Germany, Austria	Germanic
fr it	France, Canada Italy	Romance
pl ru	Poland Russia, Belarus	Slavic
ar he	Egypt, Algeria Israel	Afro-Asiatic
ja	Japan	Japonic
zh	China (Mainland)	Chinese-Tibetan

Table 13: Correspondence between Languages, Countries, and Language Families

Instruction: The following are multiple choice questions. Please choose the most reasonable one from the following options.
Example:
Question: 2+3=?
A. 1
B. 2
C. 3
D. 5
E. 6
Answer: D. 5
Question: 请问胡阿里·布迈丁出生于哪一年？
(In what year was Houari Boumediene born?)
A. 1820
B. 1828
C. 1838
D. 1932
Answer:

Figure 2: Example prompt for testing models on all our datasets.

B Example Fictional knowledge

Figure 5 shows some example continents and places in the Fictional dataset.

Language	Question template
en	What administrative division of [COUNTRY] is [CITY] in?
de	In welchem Verwaltungsbezirk von [COUNTRY] liegt [CITY]?
fr	Dans quelle division administrative de [COUNTRY] se trouve [CITY] ?
it	In quale divisione amministrativa del [COUNTRY] si trova [CITY]?
pl	W jakim podziale administracyjnym [COUNTRY] znajduje się [CITY]?
ru	В каком административном делении [COUNTRY] находится [CITY]?
ar	في [COUNTRY] لإداري تقسيم أي في [CITY] مدينة تقع؟
he	של מנהלית חלוקה באיזו [COUNTRY] נמצאת [CITY]?
ja	[CITY] は [COUNTRY] のどの行政区にありますか?
zh	[CITY] 位于 [COUNTRY] 的哪个行政区划?

Figure 3: Question templates for the xGeo part of the *Factual* dataset.

C Question templates

Figure 2 shows the example prompt for testing models in all the datasets. Figure 3 and 4 show the templates we use to construct the *Factual* dataset. Table 14 shows the templates we use to construct the *Fictional* dataset.

D Experiment details

This section provides the details in our experiments for replication.

Infrastructure We use PyTorch and HuggingFace Transformers to load and run the LLMs. For the conductivity experiments, we use the LoRA components in the LLaMA-Factory repository to finetune the models.

Hyper-parameters For model inference, we set temperature as 0 for ChatGPT and use forced decoding on all other models. For finetuning (in the conductivity experiments), we set training batch size to 32, gradient accumulation steps to 4, training epochs to 3, LoRA rank to 128, LoRA alpha to 16, LoRA dropout to 0.1, and learning rate to 2e-4, which keeps the highest performance of the majority of the models on the *Basic* and *Factual* knowledge after the LoRA finetuning.

LoRA target The reported CD results in the main body are derived from experiments using LoRA on the attention blocks only. However, we also conducted some experiments using fully finetuning and LoRA on all blocks in the Supplement Experiments.

Computational resources All the model inference can be done on 8 Nvidia Tesla v100 32 GB GPUs. Each of the finetuning experiments of 13B models on the *Fictional* knowledge can be done within 2 hours on 4 of those GPUs.

Type	Templates
Translation	Could you convert the upcoming English text to {lang}? {ENTITY} I'd appreciate it if you could transform the following English sentence into {lang}. {ENTITY} Please change the following English expression into {lang}. {ENTITY} Kindly rewrite the next English phrase in {lang}. {ENTITY} Can you transmute the subsequent English words into {lang}? {ENTITY} I need the ensuing English to be translated into {lang}, please. {ENTITY} Would you mind translating the forthcoming English into {lang}? {ENTITY} Can you render the English text that follows into {lang}? {ENTITY} Please transform the subsequent English language into {lang}. {ENTITY} I require a translation of the upcoming English sentence into {lang}, please. {ENTITY}
QA	Which continent is {PLACE} located in? What is the continent of {PLACE}? Where is {PLACE}? Which continent? In which continent can you find {PLACE}? Tell me the continent where {PLACE} is located. What continent does {PLACE} belong to? Where can {PLACE} be found continent-wise? What's the continental location of {PLACE}? Which part of the world is {PLACE} in, continent-wise? Could you specify the continent for {PLACE}?

Table 14: Translation and QA templates used to construct the *Fictional* knowledge dataset.

Language	Question template
en	In what year was [PERSON] born?
de	In welchem Jahr wurde [PERSON] geboren?
fr	En quelle année est né [PERSON] ?
it	In che anno è nato l'[PERSON]?
pl	W którym roku urodziło się [PERSON]?
ru	В каком году родился [PERSON]?
ar	ولد عام أي في [PERSON]؟
he	נולד שנה באיזו [PERSON]?
ja	[PERSON] は何年に誕生しましたか?
zh	请问 [PERSON] 出生于哪一年?

Language	Question template
en	In what year did [PERSON] die?
de	In welchem Jahr ist [PERSON] gestorben?
fr	En quelle année [PERSON] est-il décédé ?
it	In che anno è morto [PERSON]?
pl	W którym roku zmarło [PERSON]?
ru	В каком году умер [PERSON]?
ar	توفي عام أي في [PERSON]؟
he	מת שנה באיזו [PERSON]?
ja	[PERSON] が亡くなったのは何年ですか?
zh	请问 [PERSON] 逝世于哪一年?

Figure 4: Question templates for the xPeo part of the *Factual* dataset.

Model	mixed	cont'	de	others
LLaMA2	**Y**	**N**	.9720	.9498
LeoLM-base	Y	Y	.9102	.8685

Table 15: The en-CO scores of LLaMA2 and LeoLM-base on the *Factual* knowledge (the mean of xGeo and xPeo scores).

Model	mixed	cont'	de	others
LLaMA2	**Y**	**N**	.0600	.0515
LeoLM-base	Y	Y	.0580	.0399

Table 16: The XRR scores of LLaMA2 and LeoLM-base on the *Fictional* knowledge.

Model	de-tune	en	de	others
LLaMA2-Chat	**N**	-4.04	-7.79	-11.57
Vicuna v1.5	Y	+12.51	+20.99	+8.57

Table 17: The difference in RA scores after the instruction tuning of LLaMA2-Chat and Vicuna v1.5 on the *Basic* knowledge.

Model	de-tune	en	de	others
LLaMA2-Chat	**N**	-10.90	-8.85	-7.42
Vicuna v1.5	Y	-7.94	-3.98	+4.23

Table 18: The difference in RA scores after the instruction tuning of LLaMA2-Chat and Vicuna v1.5 on the *Factual* knowledge.

E Introduction of the tested models

This section introduces the models used in this research. The model parameter sizes are all 13B unless specified.

Foundation models. We use the following foundation models:

- LLaMA 1&2 (Touvron et al., 2023a,b). They are pretrained on mainly English data and a small portion of non-English data. For LLaMA 1, the multilingual pretraining data is basically Wikipedia (4.5% of the total 1.4T tokens) in 20 languages, including en, fr, it, fr, po and ru. For LLaMA 2, the multilin-

Language	Example Continent Names	Example Place Names
en	Mythosia, Veridica, Chronostead	Phoenixfire Ridge, Lunar Enclave, Titan's Summit
de	Mythosien, Veridika, Chronostätte	Phönixfeuergrat, Lunarenklave, Titanengipfel
fr	Mythosie, Véridique, Chronoséjour	Crête de Phénixfeu, Enclave Lunaire, Sommet du Titan
it	Mitosa, Veridica, Cronostallo	Cresta di Phoenixfire, Enclave Lunare, Vetta del Titano
pl	Mytozja, Veridyka, Chronostad	Grzbiet Feniksowego Ognia, Enklawa Księżycowa, Szczyt Tytana
ru	Мифосия, Веридика, Хроностэд	Гребень Фениксового Огня, Лунная Анклава, Вершина Титана
ar	ميثوسيا | فيريديكا | كرونوستيد	فينيكسفاير قمة | لونار جيب | التيتان قمة
he	מיתוסיה | וריריקה | כרונוסטד	פניקספייר הר | ירח מושבת | הטיטן שיא
ja	ミトシア, ヴェリディカ, クロノステッド	フェニックスファイア山脈, ルナーエンクレイブ, タイタンズサミット
zh	神话之地, 真实之域, 时光之地	凤凰火脊, 月亮飞地, 泰坦顶峰

Figure 5: Examples of the continents and places in the *Fictional* data.

Model	Strategy	de	fr	it	pl	ru	ar	he	ja	zh
LLaMA-13B	LoRA-Attn	.0965	.0950	.1023	.0606	.0000	.0000	.0000	.0000	.0000
LLaMA-13B	LoRA-All	.0850	.1850	.1300	.0750	.0050	.0050	.0000	.0300	.0000
LLaMA-7B	LoRA-Attn	.1250	.1450	.1500	.1250	.0150	.0050	.0250	.0000	.0000
LLaMA-7B	Fully	.0600	.1950	.1700	.1850	.0000	.0000	.0000	.0000	.0000

Table 19: The XRR scores measured by different tuning strategies on the *Fictional* knowledge, where "LoRA-Attn" means using LoRA only on the attention blocks, "LoRA-All" means using LoRA on all blocks and "Fully" stands for fully finetuning.

Translation	de	fr	it	pl	ru	ar	he	ja	zh
en-x	.0600	.1268	.1100	.1350	.0000	.0000	.0000	.0245	.0153
en-x, x-en	.0806	.0800	.0950	.1350	.0000	.0000	.0000	.0050	.0000

Table 20: The XRR scores measured with LLaMA2-13B on the *Fictional* knowledge using different translation training, where "en-x" means only translation pairs from English to other languages are provided, and "en-x, x-en" means a equal size of reversed translation data being added using the same templates.

Model	de-tune	de	others
LLaMA2-Chat	**N**	-.0075	+.0065
Vicuna v1.5	Y	+.0022	+.0066

Table 21: The difference in en-CO scores after the instruction tuning of LLaMA2 and Vicuna v1.5 on the *Factual* knowledge.

Model	de-tune	de	others
LLaMA2-Chat	**N**	+.0200	-.0114
Vicuna v1.5	Y	+.0709	-.0183

Table 22: The difference in XRR scores after the instruction tuning of LLaMA2 and Vicuna v1.5 on the *Fictional* knowledge.

gual data are extended both in quantity and language coverage (`zh` and `ja` are added). We use the 70B and 13B versions of LLaMA 2.

- Chinese-LLaMA 1&2 (Cui et al., 2023). They are built on LLaMA 1 and 2 respectively, with vocabularies and tokenizers adapted for Chinese, and continued pretraining on 120GB Chinese data. [6]

- Baichuan 2-Base (Yang et al., 2023a). It is pretrained on 2.6T tokens of multilingual, especially English-Chinese bilingual data.

- BLOOM (Workshop et al., 2023). It is a foundation model pretrained on 46 languages and 13 programming languages, including `en`, `fr`, `zh` and `ar`. We use the 7.1B version of it.

Instruction-tuned LLMs. We use the following instruction tuned models:

- Stanford Alpaca (Taori et al., 2023). It is LLaMA tuned with 52K English instruction data, and shows improved performance on several LLM benchmarks such as MMLU.

- Vicuna (Chiang et al., 2023). We use the v1.5 version of it, which is LLaMA 2 tuned with 70K user conversations with ChatGPT, collected from the ShareGPT website. Because the data is shared by users worldwide, it contains multilingual instructions in various languages.

- BayLing (Zhang et al., 2023a). It is LLaMA tuned on interactive translation data and instruction data in `en`, `de` and `zh`. It is reported to show high translation and instruction-following performance on these languages.

[6] We use the "plus" version of Chinese-LLaMA and the "pro" version of Chinese-Alpaca since they are recommended on the GitHub page.

Model	mixed	cont'	en	de (/en)	others (/en)
LLaMA2	Y	N	73.29	46.19 (0.63)	35.45 (0.48)
LeoLM-base	Y	Y	61.10	41.44 (0.68)	22.13 (0.36)

Table 23: The RA scores of LLaMA2 and LeoLM-base on the *Basic* knowledge (the mean of xCSQA and xCOPA scores), where "mixed" and "cont'" means having German mixed or continued pretraining, "/en" means the ratio to the English scores, and "others" refers to the mean scores in the other 8 languages.

Model	mixed	cont'	en	de (/en)	others (/en)
LLaMA2	Y	N	91.58	82.52 (0.90)	50.73 (0.55)
LeoLM-base	Y	Y	48.56	41.91 (0.86)	18.68 (0.38)

Table 24: The RA scores of LLaMA2 and LeoLM-base on the *Factual* knowledge (the mean of xGeo and xPeo scores).

- Chinese-Alpaca 1&2 (Cui et al., 2023). They are the Chinese-LLaMAs added Alpaca-style (Taori et al., 2023) instruction tuning. The tuning data is bilingual in English and Chinese. They show much improved performance in Chinese compared with the LLaMA models.

- LLaMA 2-Chat (Touvron et al., 2023b). It is LLaMA 2 with instruction tuning and RLHF. Although not clearly stated, the data used in the finetuning process is inferred to be mainly in English. It shows chatting ability in multiple languages.

- Baichuan 2-Chat (Yang et al., 2023a). It is Baichuan 2-Base undergone instruction tuning and reinforcement learning. The training is also multilingual, especially bilingual in English and Chinese.

- BLOOMZ-MT (Workshop et al., 2023). BLOOMZ-MT is BLOOM tuned on the xP3 instruction dataset (Muennighoff et al., 2023) in 46 languages and translation data in 9 languages. The two datasets covers `en`, `fr`, `zh`, `ar`, `de` and `ru`.

F Results of the German case study

For multilingual pretraining, Table 23 shows the basic ability of LLaMA2 and LeoLM-base; Table 24 and 15 show their PF and CT alignment on the *Factual* knowledge; and Table 16 shows their CD alignment on the *Fictional* knowledge.

For multilingual finetuning, Tabel 17 shows the basic ability of LLaMA2-Chat and Vicuna v1.5; Table 18 and 21 show their PF and CT alignment on the *Factual* knowledge; and Table 22 shows their CD alignment on the *Fictional* knowledge.

G Results of the alternative finetuning strategies

Table 19 shows the CD results of the LLaMA models using different tuning techniques; and Table 20 shows the comparison of adding or not adding reversed translation data in the tuning process on the LLaMA2 model.