Cross-Lingual Knowledge Editing in Large Language Models

Anonymous ACL submission

Abstract

Knowledge editing aims to change language models' performance on several special cases (*i.e.*, editing scope) by infusing the corresponding expected knowledge into them. With the recent advancements in large language models (LLMs), knowledge editing has been shown as a promising technique to adapt LLMs to new knowledge without retraining from scratch. However, most of the previous studies neglect the multi-lingual nature of some main-stream LLMs (*e.g.*, LLaMA, ChatGPT and GPT-4), and typically focus on monolingual scenarios, where LLMs are edited and evaluated in the same language. As a result, it is still unknown the effect of source language editing on a different target language. In this paper, we aim to figure out this cross-lingual effect in knowledge editing. Specifically, we first collect a large-scale cross-lingual synthetic dataset by translating ZsRE from English to Chinese. Then, we conduct English editing on various knowledge editing methods covering different paradigms, and evaluate their performance in Chinese, and vice versa. To give deeper analyses of the cross-lingual effect, the evaluation includes four aspects, *i.e.*, reliability, generality, locality and portability. Furthermore, we analyze the inconsistent behaviors of the edited models and discuss their specific challenges.[1]

1 Introduction

The goal of knowledge editing is to adjust language models' behaviors within an expected scope (*i.e.*, editing scope) and retain out-of-scope model performance ideally (Yao et al., 2023). Along with the dynamic changes in the world, knowledge editing could help models forget outdated knowledge and adapt to the new counterpart without retraining from scratch. As the example shown in Figure 1 (a), the number of Honkai-series games increases to four after the release of *Honkai: Star Rail* (on April 26, 2023). However, if we ask a model that has been trained before the date, the model might only know three Honkai-series games. In such a situation, knowledge editing could help the model efficiently update this new knowledge, and give the right answer after editing.

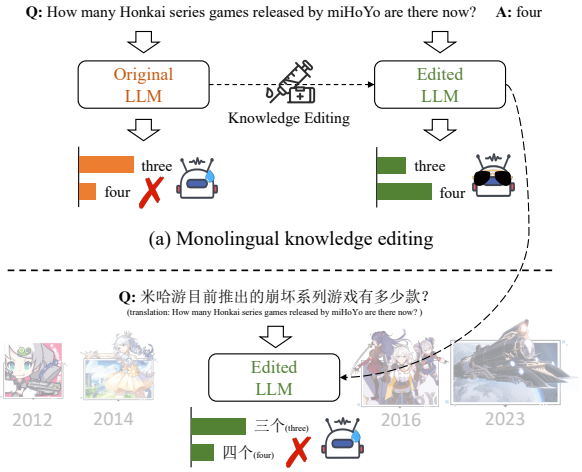


Figure 1: Illustration of (a) monolingual knowledge editing, where the model is edited and verified in the same language; and (b) cross-lingual knowledge editing, where the model is edited and verified in different languages.

Despite many efforts devoted to this research field (De Cao et al., 2021; Mitchell et al., 2022b; Dong et al., 2022; Dai et al., 2022; Meng et al., 2022; Mitchell et al., 2022a; Huang et al., 2023b; Meng et al., 2023; Zheng et al., 2023), current knowledge editing studies typically focus on monolingual scenarios, where language models are edited and evaluated within the same language, c.f., Figure 1 (a). Meanwhile, the rapid advancements in large language models (LLMs) have led to the widespread adoption of multi-lingual settings, allowing language modeling ability can be shared across different languages (Zhao et al., 2023). For example, LLMs such as LLaMA (Tou-

[1]The data and codes will be released upon publication.

vron et al., 2023a), ChatGPT (OpenAI, 2022), and GPT-4 (OpenAI, 2023) are designed to operate under multilingual setting. Under this background, the performance of a source-language edited model on other languages is still unknown. As shown in Figure 1 (b), a research question (RQ) arises, *when we utilize source-language samples to edit a multi-lingual LLM, can the model reflect consistent behaviors when faced with a different target language?*

To answer the RQ, in this paper, we explore knowledge editing in cross-lingual scenarios, and study the effects of source-language editing on a different target language. Specifically, we automatically translate the knowledge editing data from English to Chinese via cutting-edge LLMs (*i.e.*, ChatGPT and GPT-4). After carefully comparing existing datasets, we finally choose ZsRE (Levy et al., 2017) which is originally a question answering (QA) dataset and is further widely used in knowledge editing (De Cao et al., 2021; Meng et al., 2022; Mitchell et al., 2022a). More recently, Yao et al. (2023) collect a number of QA pairs that need deep reasoning based on ZsRE, and the data could be used to evaluate the portability of knowledge editing methods beyond simple paraphrasing. Therefore, we also translate these QA pairs to give a deeper understanding of cross-lingual knowledge editing performance. The translated data together with the original ones is denoted as Bi-ZsRE. Then, we conduct English/Chinese editing on several open-sourced multi-lingual LLMs (LLaMA, LLaMA2, Baichuan and Baichuan2), and evaluate their behaviors in Chinese/English in terms of reliability, generality, locality and portability. Our experiments involve seven knowledge editing methods covering three main-stream paradigms pointed by Yao et al. (2023), *i.e.*, memory-based, meta-learning and locate-then-edit methods. The experimental results reveal that (1) the language modeling gaps across different languages influence the efficiency of knowledge editing; (2) it is still hard for existing knowledge editing methods to transfer the edited knowledge from one language to another in a multi-lingual LLM; (3) when editing LLMs in a language, the model performance on the irrelevant examples in other languages could also be influenced, resulting in low locality. This presents a significant challenge for multi-lingual LLMs in maintaining consistent behaviors across different languages.

Our main contributions are concluded as follows:

- To our knowledge, we are the first to explore the cross-lingual effect of knowledge editing in LLMs. We achieved this by translating the ZsRE dataset and studying the cross-lingual effect from English (Chinese) to Chinese (English).
- We conduct experiments on various knowledge editing methods and multi-lingual LLMs. Our results indicate that it remains challenging for multi-lingual LLMs to generalize the edited knowledge to other languages.
- In-depth analysis of the inconsistent behaviors exhibited by the edited models and their specific challenges provide us with a deeper understanding of the cross-lingual effect in knowledge editing.

2 Related Work

Knowledge Editing Methods. The goal of knowledge editing is to alter the behavior of LLMs within an expected scope (*i.e.*, editing scope) without negatively impacting performance out of the scope. According to a comprehensive survey on knowledge editing (Yao et al., 2023), there are three main-stream knowledge editing paradigms: (1) *Memory-based methods* keep the original model parameters unchanged while employing another model to influence the model's behaviors. SERAC (Mitchell et al., 2022b) utilizes a scope classifier to evaluate whether new input is close to the stored editing examples, and further influences the model behaviors based on the retrieved editing examples. T-Patcher (Huang et al., 2023b) and CaliNET (Dong et al., 2022) add extra trainable parameters into the FFN layers of LLMs to edit model performance. IKE (Zheng et al., 2023) uses context-edit facts to guide the model in generating edited facts. (2) *Meta-learning methods* employ a hyper network to learn the weight updates of LLMs to edit the models. KE (De Cao et al., 2021) makes use of LSTM networks to predict the weight update for each new input. MEND (Mitchell et al., 2022a) transforms the gradient of fine-tuned language models by employing a low-rank decomposition of gradients. (3) *Locate-then-edit methods* first identify parameters corresponding to specific knowledge and then update these parameters. Among them, KN (Dai et al., 2022) specifies a key-value pair in the FFN matrix that embodies the knowledge and then proceeds to update the corresponding parameters. ROME (Meng et al., 2022) leverages

causal mediation analysis to locate the edit area, and update the whole parameters in the FFN matrix. MEMIT (Meng et al., 2023) directly updates LLMs with many memories, and thus, facilitating thousands of edits to be executed simultaneously.

More recently, Xu et al. (2023) introduce the cross-lingual model editing task and design a language anisotropic editing method. However, the proposed method is applied to mBERT (Devlin et al., 2019) (a classical pre-trained multi-lingual NLU model), making the cross-lingual effect still unknown in generative LLMs.

Knowledge Editing Datasets. ZsRE (Levy et al., 2017) is a question answering dataset whose queries require models to answer the questions based on the information within the queries. COUNTERFACT (Meng et al., 2022) evaluates whether the edited model can provide counterfactual answers when asked about the corresponding factual knowledge. MQUAKE (Zhong et al., 2023) aims to assess whether edited models correctly answer questions where the answer needs reasoning based on the edited facts. Cheng et al. (2023) propose MMEdit, a multi-modal knowledge editing benchmark dataset. PersonalityEdit (Mao et al., 2023) is proposed to edit personality traits for LLMs. Li et al. (2023) propose ConflictEdit and RoundEdit to investigate the potential pitfalls associated with knowledge editing for LLMs. Eva-KELLM (Wu et al., 2023) evaluates the edited model from reasoning with the altered knowledge and cross-lingual transfer. Though Eva-KELLM provides a subset for cross-lingual knowledge editing, the data has not yet been made public.[2] Besides, this work does not conduct experiments on any knowledge editing methods, leaving the cross-lingual effect still not known in the knowledge editing research field.

3 Bi-ZsRE

In this section, we first discuss the details of data collection, including data sources, translation process as well as quality control (§ 3.1). Then, we give the data statistics of Bi-ZsRE (§ 3.2), and finally provide the task overview of cross-lingual knowledge editing (§ 3.3).

3.1 Data Collection

Data Sources. ZsRE (Levy et al., 2017) is a Question Answering (QA) dataset whose queries require models to answer the questions based on the information within the queries. Following previous data settings (Yao et al., 2023; Wang et al., 2023), it contains 163,196 training samples and 19,086 validation samples. Each sample involves a question and a corresponding answer for editing LLMs. To evaluate the generality of edited models, a rephrased question is also provided. Besides, each sample also associates with an unrelated QA pair (selected from the NQ dataset (Kwiatkowski et al., 2019)) to evaluate the locality. Recently, Yao et al. (2023) provide a test set with 1,037 samples for a more comprehensive evaluation of knowledge editing, where each test sample additionally contains a QA pair to assess LLMs' portability to reason based on the edited fact. To control the cost of translation, we randomly selected 10,000 training samples and 3,000 validation samples, which together with all test samples are further translated.

Translation Process. We use `gpt-3.5-turbo` and `gpt-4` to translate the above knowledge editing data from English to Chinese. In particular, considering the trade-off between quality and cost, training samples and validation samples are translated by `gpt-3.5-turbo`, while test samples are translated by `gpt-4`. The translation is conducted based on the OpenAI's official APIs[3] with zero temperature. The used translation prompt is shown as follows:

> Please translate the following JSON data from English to Chinese and keep the format unchanged:
> `[JSON data]`

where each sample is organized in JSON format and further translated at the sample level.

Quality Control. To further ensure the translation quality of the test samples, we also employ three translators to correct the translations of `gpt-4`. All translators are native Chinese and are fluent in English. As a result, there are about 6.0% of samples are corrected while the remaining are unchanged. All corrected samples are further checked by a data expert who has rich experience in translation annotations. Finally, all translated data and original data are denoted as Bi-ZsRE.

3.2 Data Statistics

Table 1 lists the data statistics of Bi-ZsRE, covering two languages, English (En) and Chinese

[2] October 15, 2023

[3] https://platform.openai.com/docs/api-reference/chat/object

Splitting	Lang.	# Example	Question	Rephrased Question	Answer	Locality Question	Locality Answer	Portability Question	Portability Answer
Training	En	10,000	11.28	11.25	2.85	15.25	5.61	-	-
	Zh	10,000	10.86	10.95	4.36	14.71	6.77	-	-
Validation	En	3,000	11.19	11.20	2.79	15.39	5.50	-	-
	Zh	3,000	10.94	11.01	4.37	14.66	6.55	-	-
Test	En	1,037	11.43	11.49	3.11	15.31	5.62	18.02	4.54
	Zh	1,037	11.48	11.60	4.69	11.56	6.77	16.48	5.88

Table 1: Statistics of Bi-ZsRE (Lang.: language; En: English; Zh: Chinese). "*# Example*" indicates the number of samples in each subset. All decimals denote the average length (token-level) of different aspects in each subset.

(Zh), across three subsets. For English samples, the average question lengths are 11.28, 11.19, and 11.43 tokens in the training, validation, and test subsets, respectively, while the counterparts in Chinese are 10.86, 10.94, and 11.48. Besides, the average length of portability questions is longer than that of original questions, rephrased questions or locality questions, thus portability questions may involve more intricate reasoning based on the edited knowledge. To correctly answer the portability questions, the edited model should absorb the knowledge rather than simply memorize the word replacement.

3.3 Task Overview

Knowledge Editing. Given a language model p_θ and an edit descriptor $\langle x_e, y_e \rangle$, the goal of knowledge editing is to create an edited model p'_θ satisfy the following requirements:

$$p'_\theta(x) = \begin{cases} y_e & x \in \mathcal{X}_e \\ p_\theta(x) & x \notin \mathcal{X}_e \end{cases} \quad (1)$$

where $\mathcal{X}_e$ denotes a broad set of inputs with the same semantics as x_e. The edited model should also satisfy the following four properties: (1) *Reliability* measures the average accuracy on the edit case. When receiving x_e as input, the edited model p'_θ should output y_e. (2) *Generality* evaluates the average accuracy on the equivalent cases as the edit case. For instance, when receiving a rephrased text of x_e, the edited model p'_θ is also expected to output y_e. (3) *Locality* assesses the accuracy of the edited model on the irrelevant samples. When the input x is out of the edit scope $\mathcal{X}_e$, $p'_\theta(x)$ should be the same as $p_\theta(x)$ ideally. (4) *Portability* measures the robust generalization of the edited model via a portability question that needs reasoning based on the edited knowledge. When receiving the portability question as input, the edited model p'_θ is expected to output the golden answer to demonstrate the model indeed learns the knowledge rather than memorizing superficial changes in wording.

Cross-Lingual Knowledge Editing. Given a multi-lingual language model $p_{m\theta}$ and an edit descriptor in a source language $\langle x_e^s, y_e^s \rangle$, the goal of cross-lingual knowledge editing is to create an edited model $p'_{m\theta}$ satisfy the following requirements:

$$p'_{m\theta}(x^s) = \begin{cases} y_e^s & x^s \in \mathcal{X}_e^s \\ p_{m\theta}(x^s) & x^s \notin \mathcal{X}_e^s \end{cases} \quad (2)$$

$$p'_{m\theta}(x^t) = \begin{cases} I^t(y_e^s) & x^t \in I^t(\mathcal{X}_e^s) \\ p_{m\theta}(x^t) & x^t \notin I^t(\mathcal{X}_e^s) \end{cases} \quad (3)$$

where x^s and x^t denote the input text in the source language s and a different target language t, respectively. $\mathcal{X}_e^s$ indicates the edit scope in the source language. $I^t(\cdot)$ transforms the input text from its source language into the target language t with the same meaning, *i.e.*, translation. Therefore, in addition to learning edited knowledge in the source language, the model $p'_{m\theta}$ should also reflect consistent behaviors when querying in a different language. The cross-lingual knowledge editing also needs to satisfy the four properties, *i.e.*, reliability, generality, locality and portability. Different from the monolingual scenario, all test samples (except reliability samples) in the cross-lingual scenario are in both the source and the target languages, respectively. For example, an English edited model will be evaluated by a Chinese generality sample to indicate its cross-lingual generality.

4 Experiments

4.1 Experimental Setup

Metrics. To evaluate the edited model in terms of reliability, generality, locality and portability, different questions, which pair with the golden answers, are input to the edited model. Thus, we follow previous QA studies (Rajpurkar et al., 2016;

Editing Language	Method	Reliability	English Evaluation			Chinese Evaluation		
			Generality	Locality	Portability	Generality	Locality	Portability
				Chinese-LLaMA-Plus-7B				
English	FT	20.46 / 00.77	18.36 / 00.19	87.49 / 70.11	06.30 / 00.00	22.08 / 00.10	79.52 / 47.44	24.63 / 00.00
	SERAC	73.84 / 56.03	50.86 / 27.10	100.0 / 100.0	06.52 / 00.00	19.26 / 00.29	99.97 / 99.90	15.63 / 00.00
	IKE	99.90 / 99.90	99.24 / 98.36	62.79 / 36.26	50.86 / 17.84	92.95 / 69.72	36.16 / 06.75	33.84 / 04.24
	MEND	37.57 / 02.22	33.24 / 01.35	88.96 / 74.25	06.56 / 00.10	17.08 / 00.00	91.75 / 75.89	16.91 / 00.00
	KN	04.63 / 00.00	04.54 / 00.00	42.25 / 29.12	03.53 / 00.00	06.66 / 00.00	36.75 / 19.67	08.46 / 00.00
	ROME	98.98 / 97.20	94.58 / 87.85	92.49 / 81.49	08.48 / 00.00	26.65 / 06.75	89.08 / 67.60	17.07 / 00.00
	MEMIT	96.19 / 92.48	90.66 / 81.97	98.31 / 94.70	08.27 / 00.00	28.26 / 06.75	97.31 / 91.51	17.96 / 00.00
Chinese	FT	09.54 / 00.19	12.38 / 00.00	87.81 / 73.29	06.77 / 00.10	35.91 / 00.96	57.78 / 15.24	21.09 / 00.19
	SERAC	27.05 / 12.83	14.67 / 00.00	100.0 / 100.0	06.56 / 00.00	67.41 / 37.32	94.42 / 85.82	20.65 / 00.00
	IKE	99.90 / 99.71	85.39 / 77.24	64.14 / 37.32	40.07 / 05.01	97.31 / 95.37	52.46 / 17.36	38.39 / 07.52
	MEND	15.47 / 00.48	14.39 / 00.00	89.19 / 73.87	06.72 / 00.10	44.32 / 00.68	78.17 / 46.00	22.94 / 00.19
	KN	03.09 / 00.00	04.74 / 00.00	29.82 / 16.39	02.87 / 00.00	05.08 / 00.00	20.08 / 08.78	05.56 / 00.00
	ROME	36.63 / 20.15	24.24 / 08.87	89.21 / 74.73	06.52 / 00.00	81.83 / 39.92	86.44 / 63.74	21.33 / 00.10
	MEMIT	35.54 / 19.19	22.88 / 08.97	98.13 / 94.12	06.88 / 00.00	81.11 / 39.34	95.84 / 86.98	23.29 / 00.00
				Chinese-LLaMA-2-7B				
English	FT	36.62 / 05.98	35.01 / 07.52	81.90 / 55.06	07.33 / 00.00	20.24 / 00.10	72.95 / 32.11	17.91 / 00.00
	SERAC	98.78 / 97.01	89.62 / 82.64	100.0 / 100.0	08.75 / 00.00	21.92 / 02.60	97.67 / 93.44	17.30 / 00.00
	IKE	100.0 / 100.0	99.69 / 99.32	56.35 / 30.76	45.72 / 11.76	92.28 / 77.72	41.59 / 04.63	37.04 / 04.82
	MEND	56.57 / 00.00	49.33 / 00.00	95.46 / 86.79	07.62 / 00.00	20.66 / 00.00	95.25 / 86.21	17.34 / 00.00
	KN	10.94 / 00.00	10.96 / 00.00	49.28 / 06.85	05.75 / 00.00	12.30 / 00.00	43.65 / 09.35	14.39 / 00.00
	ROME	77.65 / 67.98	72.27 / 55.06	93.67 / 81.58	07.48 / 00.10	23.27 / 03.28	95.55 / 84.96	17.88 / 00.00
	MEMIT	83.01 / 74.64	77.63 / 61.43	98.45 / 95.37	08.08 / 00.10	23.91 / 03.95	98.13 / 93.54	17.22 / 00.00
Chinese	FT	13.03 / 01.16	16.30 / 01.06	76.68 / 48.02	07.07 / 00.00	36.01 / 00.77	59.70 / 16.59	19.25 / 00.00
	SERAC	26.76 / 20.44	19.87 / 02.31	100.0 / 100.0	08.14 / 00.00	71.76 / 49.37	77.85 / 56.89	23.67 / 02.03
	IKE	99.95 / 99.90	94.40 / 91.22	51.42 / 23.43	40.75 / 05.40	99.40 / 98.94	52.23 / 14.66	45.05 / 13.69
	MEND	20.65 / 00.00	20.40 / 00.00	96.45 / 89.87	07.06 / 00.00	47.04 / 00.00	90.13 / 70.11	22.62 / 00.00
	KN	08.40 / 00.00	10.55 / 00.00	45.10 / 04.44	05.88 / 00.00	12.19 / 00.00	37.47 / 03.95	14.02 / 00.00
	ROME	24.88 / 08.29	20.17 / 02.51	93.75 / 82.45	07.06 / 00.00	60.44 / 12.83	94.75 / 83.70	24.75 / 02.12
	MEMIT	25.84 / 09.60	20.41 / 02.12	98.67 / 95.76	07.29 / 00.00	64.16 / 13.31	96.75 / 89.49	26.10 / 02.31

Table 2: Experimental results on the Chinese-LLaMA-Plus-7B and Chinese-LLaMA-2-7B backbones in terms of F1 / EM. Grey denotes the score is less than 10.0, while green indicates the score is more than 80.0. "Editing language" denotes the model is edited by the samples of which language.

	C-Eval	MMLU
GPT-4	68.7	86.4
GPT-3.5-turbo	54.4	70.0
Baichuan2-7B	54.0	54.2
Baichuan-7B	42.8	42.3
Chinese-LLaMA-2-7B	34.4	36.8
Chinese-LLaMA-Plus-7B	25.5	31.8

Table 3: Chinese and English capability of the LLMs used in our experiments.

Yang et al., 2018) and adapt exact match (EM) and F1 as two evaluation metrics: (1) EM measures the percentage of predictions that match the golden answers exactly. (2) F1 measures the average overlap between the prediction and the golden answer. We treat the prediction and ground truth as bags of tokens, and compute their F1.

Baselines. Following Yao et al. (2023); Wang et al. (2023), we adopted 7 methods as baselines: (1) Directly fine-tuning (**FT**) the language models with L_∞ constraint; (2) **SERAC** (Mitchell et al., 2022b) utilizes a scope classifier to evaluate whether new input is close to the stored editing examples, and further influences the model behaviors based on the retrieved editing examples; (3) **IKE** (Zheng et al., 2023) uses context-edit facts to guide the model in generating edited facts; (4) **MEND** (Mitchell et al., 2022a) transforms the gradient of fine-tuned language models by employing a low-rank decomposition of gradients; (5) **KN** (Dai et al., 2022) specifies a key-value pair in the FFN matrix that embodies the knowledge and then proceeds to update the corresponding parameters; (6) **ROME** (Meng et al., 2022) leverages causal mediation analysis to locate the edit area, and updates the whole parameters in the FFN matrix; (7) **MEMIT** (Meng et al., 2023) directly updates LLMs with many memories, and thus, facilitating thousands of edits to be executed simultaneously.

Backbones. Considering the English and Chinese abilities, we adopt the following four LLMs in the

Editing Language	Method	Reliability	English Evaluation			Chinese Evaluation		
			Generality	Locality	Portability	Generality	Locality	Portability
				Baichuan-7B				
English	FT	33.33 / 13.11	27.09 / 07.43	91.71 / 83.12	09.21 / 00.19	20.79 / 00.19	87.36 / 64.71	30.77 / 00.10
	IKE	100.0 / 100.0	99.72 / 99.61	66.87 / 48.02	69.79 / 50.72	99.25 / 98.65	47.96 / 15.72	44.58 / 14.66
	KN	10.77 / 00.00	10.32 / 00.00	71.28 / 55.74	08.96 / 00.19	19.69 / 00.00	93.32 / 80.71	31.74 / 00.00
	ROME	68.97 / 52.36	60.45 / 42.53	98.31 / 96.43	09.65 / 00.29	24.45 / 01.45	98.71 / 95.85	31.61 / 00.29
	MEMIT	71.20 / 54.97	66.47 / 49.66	98.60 / 96.72	09.43 / 00.10	26.19 / 02.51	98.82 / 95.56	30.53 / 00.29
Chinese	FT	13.08 / 01.45	13.39 / 00.58	95.18 / 90.26	09.28 / 00.29	28.71 / 04.34	53.83 / 16.88	27.76 / 00.29
	IKE	100.0 / 100.0	98.20 / 97.01	70.28 / 51.40	69.82 / 51.11	100.0 / 100.0	46.92 / 14.95	48.91 / 19.38
	KN	10.22 / 00.00	10.49 / 00.00	73.43 / 58.24	09.04 / 00.29	19.52 / 00.00	84.62 / 59.98	31.64 / 00.00
	ROME	24.04 / 06.36	16.05 / 01.93	98.06 / 95.66	09.40 / 00.29	68.74 / 12.63	97.96 / 93.15	27.98 / 00.68
	MEMIT	23.95 / 06.27	19.11 / 05.59	98.47 / 96.53	09.05 / 00.19	72.29 / 14.75	96.87 / 90.55	24.49 / 00.48
				Baichuan2-7B				
English	FT	33.43 / 00.48	32.25 / 00.00	90.47 / 78.50	27.28 / 01.74	24.76 / 03.47	80.79 / 61.43	22.64 / 00.29
	IKE	77.76 / 70.40	77.71 / 70.30	71.18 / 51.01	58.11 / 37.51	97.00 / 95.85	71.97 / 47.44	65.61 / 42.43
	KN	10.06 / 00.00	09.66 / 00.00	96.22 / 93.83	31.62 / 00.29	19.77 / 00.00	95.51 / 90.55	25.21 / 00.58
	ROME	88.33 / 81.97	73.22 / 59.88	95.96 / 90.26	31.10 / 02.31	29.82 / 07.52	96.36 / 89.20	26.00 / 01.25
	MEMIT	89.34 / 83.32	82.54 / 73.19	98.61 / 96.53	30.57 / 01.74	32.11 / 11.09	98.31 / 95.08	24.59 / 00.58
Chinese	FT	11.45 / 00.19	13.29 / 00.00	92.46 / 83.03	30.11 / 00.39	34.04 / 07.04	60.57 / 29.80	26.78 / 02.31
	IKE	97.08 / 96.05	77.36 / 69.72	71.48 / 51.88	58.58 / 37.99	97.00 / 95.95	73.64 / 50.63	66.64 / 43.68
	KN	08.92 / 00.10	08.95 / 00.10	84.11 / 81.39	27.39 / 00.29	18.30 / 00.10	84.58 / 78.01	22.78 / 00.58
	ROME	35.55 / 19.38	17.12 / 01.06	95.32 / 90.07	31.85 / 00.48	90.38 / 83.70	95.00 / 85.54	28.53 / 00.96
	MEMIT	35.29 / 19.00	17.37 / 01.35	98.74 / 96.24	31.45 / 00.87	93.45 / 88.52	97.13 / 92.09	28.16 / 00.77

Table 4: Experimental results on the Baichuan-7B and Baichuan2-7B backbones in terms of F1 / EM.

experiments: (1) Chinese-LLaMA-Plus-7B[4] is created based on LLaMA-7B (Touvron et al., 2023a) with Chinese vocabulary extension and continual pre-training. (2) In the similar way, Chinese-LLaMA-2-7B[5] is created based on LLaMA-2-7B (Touvron et al., 2023b). (3) Baichuan-7B[6] and (4) BaiChuan2-7B[7] are two LLMs that support both English and Chinese. Table 3 lists the above LLMs' performance on C-Eval (Huang et al., 2023a) and MMLU (Hendrycks et al., 2021) to show their Chinese and English abilities, respectively. Baichuan2-7B performs the best among the four backbones in both two evaluation benchmark datasets.

Implementation Details. All experiments are conducted on a single NVIDIA A800 GPU (80G). The implementation of all baselines is employed by EasyEdit (Wang et al., 2023) with the default settings. The hyper-parameters of each method can be found in the corresponding GitHub repository.[8]

[4] https://github.com/ymcui/Chinese-LLaMA-Alpaca

[5] https://github.com/ymcui/Chinese-LLaMA-Alpaca-2

[6] https://github.com/baichuan-inc/Baichuan-7B/

[7] https://github.com/baichuan-inc/Baichuan2

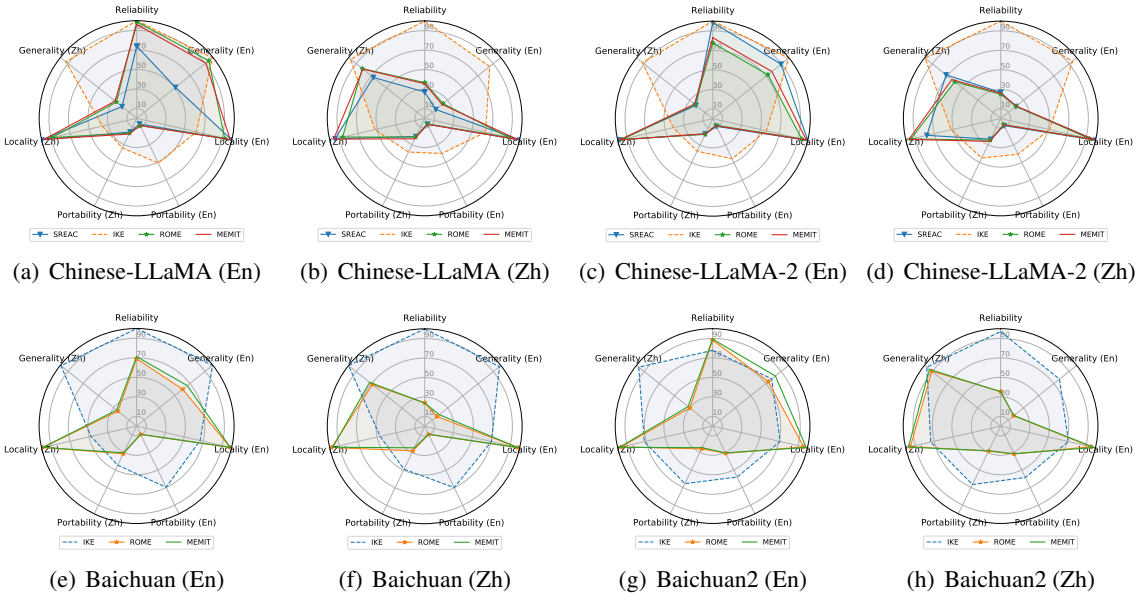
[8] https://github.com/zjunlp/EasyEdit/tree/main/hparams

4.2 Results & Analyses

Table 2 shows the experimental results on Chinese-LLaMA-Plus-7B and Chinese-LLaMA-2-7B.

Monolingual Analysis. Compared with the other three properties, portability is more challenging for knowledge editing methods to achieve. As we can see, IKE achieves the best performance in terms of portability among all baselines. However, the best EM score in portability is still less than 60.0, showing it is non-trivial to absorb the edited knowledge for most editing methods. As for reliability which directly evaluates the model performance on the edited knowledge, we find that FT and KN obtain limited performance in this property, thus failing to edit knowledge in LLMs. For example, KN (En) and KN (Zh) only achieve 4.63 and 3.09 F1 reliability with the Chinese-LLaMA-Plus-7B backbone, while the counterparts of FT (En) and FT (Zh) are 20.46 and 9.54. Given the above analyses, we next compare the cross-lingual knowledge editing performance on SERAC, IKE, MEND, ROME and MEMIT methods.

Inconsistent Behaviors in Reliability. When using different languages to edit LLMs, there might be performance gaps in terms of reliability. For example, SERAC (En) achieves 73.84 F1 while SERAC (Zh) only achieves 27.05 F1 on Chinese-LLaMA-Plus-7B. A similar situation could also be found in MEND, ROME and MEMIT. When



(a) Chinese-LLaMA (En) (b) Chinese-LLaMA (Zh) (c) Chinese-LLaMA-2 (En) (d) Chinese-LLaMA-2 (Zh)

(e) Baichuan (En) (f) Baichuan (Zh) (g) Baichuan2 (En) (h) Baichuan2 (Zh)

Figure 2: Radar chart of knowledge editing performance when editing different LLMs with different languages. The language identifiers (*i.e.*, En and Zh) after the name of LLMs indicate the model is edited by the samples of which language, while those after the name of properties indicate the language of the corresponding testing samples.

using English samples to edit LLMs via these four methods, the achieved reliability is significantly higher than using the Chinese samples. This is because the language modeling ability of different languages might be different in a single integrated multi-lingual LLM. Many LLMs show their strong English ability perhaps due to the high-quality English data dominating the pre-training corpora (Touvron et al., 2023a,b). The language modeling gaps of different languages might influence the efficiency of knowledge editing in different languages. Moreover, we find that when editing LLMs via IKE using different languages, the achieved reliability scores are similar. For example, IKE (En) and IKE (Zh) achieve the same F1 score (99.90) in terms of reliability on Chinese-LLaMA-Plus-7B. This indicates that not all knowledge editing methods are sensitive to the choice of editing languages. Given the strong in-context learning ability of LLMs, IKE can efficiently edit them with demonstration samples in different languages.

Inconsistent Behaviors in Generality. It is intuitive that when using one language to edit LLMs, the generality in this language is significantly higher than in others. For example, SERAC (En) achieves 50.86 F1 in English generality but only performs 19.26 F1 in Chinese generality (with the Chinese-LLaMA-Plus-7B backbone). In contrast, SERAC (Zh) achieves better Chinese generality than English (67.41 F1 vs. 14.67 F1). Almost all knowledge editing methods have this phenomenon. This finding also indicates that the cross-lingual performance of knowledge editing is still limited. It is hard for existing knowledge editing methods to transfer the edited knowledge from one language to others in multi-lingual LLMs, and reflect consistent behaviors when querying with different languages.

Cross-Lingual Influence on Locality. When editing LLMs in a source language, the locality in other languages could also be influenced. When editing Chinese-LLaMA-Plus-7B, MEND (En) achieves 88.96 F1 and 91.75 F1 in English and Chinese locality scores, while the counterparts in MEND (Zh) are 89.19 and 78.17. Ideally, when editing LLMs in a source language, its performance on irrelevant target-language samples should remain unchanged. Previous work typically only studies locality in the same language and neglects the cross-lingual locality. We also find that though IKE works well in terms of reliability and generality, its locality is generally less than that of SERAC, ROME or MEMIT. The low-level locality makes its usefulness need to be carefully verified in real applications.

Limited Portability in both Languages. As shown in Figure 2, when editing multi-lingual LLMs in English or Chinese, their portability performance in both languages is extremely limited compared with other properties. This finding indi-

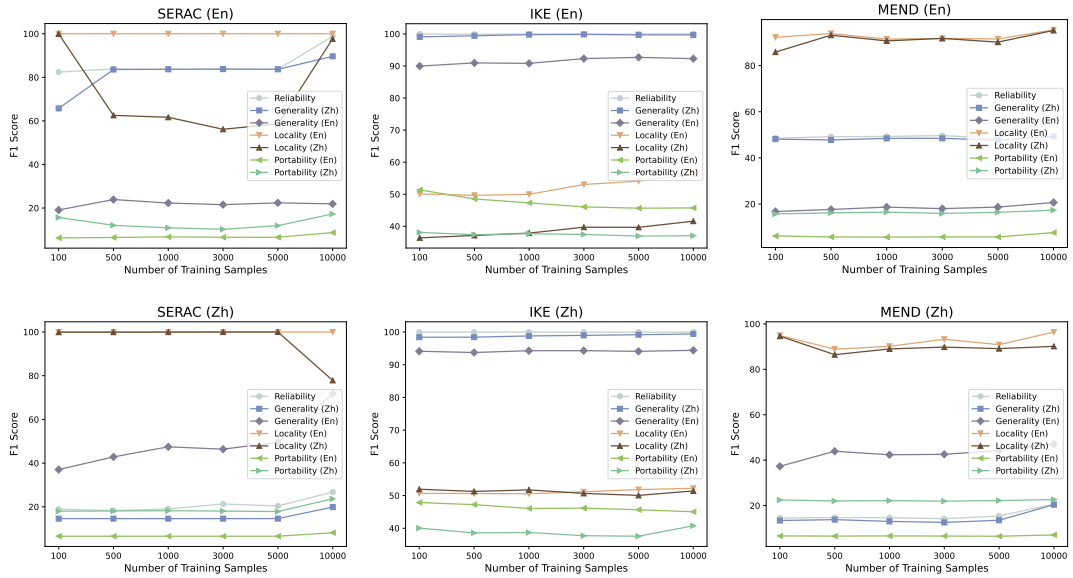


Figure 3: F1 scores of the edited Chinese-LLaMA-2-7B using different numbers of training samples.

cates that most existing knowledge editing methods only memorize the superficial changes in wording rather than absorbing the edited knowledge. This phenomenon shows that sharing knowledge across different languages is tricky. As a result, the currently edited LLMs reflect inconsistent behaviors on the edited knowledge in different languages.

Knowledge Editing Performance on Baichuan. Table 4 shows the knowledge editing performance on Baichuan-7B and Baichuan2-7B. The results show similar situations to those of Chinese-LLaMA-Plus-7B and Chinese-LLaMA-2-7B, verifying the generality of the previous phenomena and our analyses.

4.3 The Influence of Training Scale

Due to the high cost of translating all ZsRE training samples (163K), we randomly select and translate 10K samples via gpt-3.5-turbo (Section 3.1). We further conduct experiments to investigate the model performance when the training set is limited. Specifically, we randomly choose 100, 500, 1K, 3K and 5K training samples to conduct experiments. Among all baselines, SERAC, IKE and MEND are three knowledge editing methods that need additional training. Thus, we use different numbers of training samples to edit LLMs via these methods, and evaluate their performance in terms of reliability, generality, locality and portability. Figure 3 shows the results using Chinese-LLaMA-2-7B as an example backbone. As we can see, the reliability, generality and portability of SERAC typically increase with the number of training samples, especially SERAC (Zh). The English locality of SERAC is stable while the Chinese locality is sensitive to the training data. As for IKE and MEND, their performances are relatively stable in terms of all metrics. We also find that the portability of IKE and MEND may decrease or remain the same as the number of training samples increases. Therefore, simply adding more training samples in these two methods cannot increase the ability of the edited models to absorb and reason the edited knowledge. Future work could explore more effective methods or design more reasonable training paradigms to let LLMs go beyond memorizing the superficial changes in wording.

5 Conclusion

In this paper, we first explore the cross-lingual effect of knowledge editing in large language models. To achieve that, we automatically construct Bi-ZsRE dataset by translating the previous ZsRE dataset from English to Chinese. Based on Bi-ZsRE, we conduct experiments on various knowledge editing methods, and study the cross-lingual effect from English to Chinese and vice versa. Our results indicate that it is still hard for existing knowledge editing methods to transfer the edited knowledge from one language to another in a multilingual LLM. We also analyze the inconsistent behaviors of the edited models and discuss their specific challenges to provide a deeper understanding of the cross-lingual effect in knowledge editing.

Ethical Considerations

In this section, we consider the potential ethical issues of our work. In this paper, we propose the Bi-ZsRE dataset which is collected based on the publicly-available datasets, *i.e.*, ZsRE (Levy et al., 2017) and portability QA pairs provided by Yao et al. (2023). Therefore, Bi-ZsRE might involve the same biases and toxic behaviors exhibited by these datasets. Besides, we obtain our Bi-ZsRE dataset by translating these datasets, and their corresponding license is the MIT License which is granted to copy, distribute and modify the contents.

During manually correcting the results of machine translation, the salary for each annotator is determined by the average time of annotation and local labor compensation standard.

Limitation

While we show the cross-lingual effect in knowledge editing, there are some limitations worth considering in future work: (1) Bi-ZsRE only involves English and Chinese, and future work could extend our Bi-ZsRE to more languages and give more comprehensive analyses w.r.t different language families. (2) The backbones used in our experiments are several LLMs with 7B parameters. Future work can extend our analyses to other LLMs with more parameters (*e.g.*, 13B and 70B).

References

Siyuan Cheng, Bo Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. 2023. Can we edit multimodal large language models? *arXiv preprint arXiv:2310.08475*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5937–5947, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023a. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023b. Transformer-patcher: One mistake worth one neuron. *arXiv preprint arXiv:2301.09785*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. 2023. Unveiling the pitfalls of knowledge editing for large language models. *arXiv preprint arXiv:2310.02129*.

Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Meng Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2023. Editing personality for llms.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, et al. 2023. Easyedit: An easy-to-use knowledge editing framework for large language models. *arXiv preprint arXiv:2308.07269*.

Suhang Wu, Minlong Peng, Yue Chen, Jinsong Su, and Mingming Sun. 2023. Eva-kellm: A new benchmark for evaluating knowledge editing of llms. *arXiv preprint arXiv:2308.09954*.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. 2023. Language anisotropic cross-lingual model editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5554–5569, Toronto, Canada. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Yunzhi Yao, Peng Wang, Bo Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. *ArXiv*, abs/2305.13172.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.