

# Cross-Lingual Knowledge Editing in Large Language Models

Jiaan Wang<sup>♣\*</sup>, Yunlong Liang<sup>◇</sup>, Zengkui Sun<sup>◇</sup>, Yuxuan Cao<sup>♣</sup>, Jiarong Xu<sup>♡</sup>

<sup>♣</sup>Soochow University

<sup>♣</sup>Zhejiang University

<sup>◇</sup>Beijing Jiaotong University

<sup>♡</sup>Fudan University

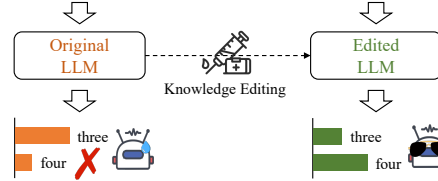
## Abstract

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.<sup>1</sup>

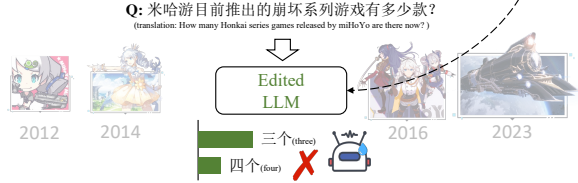
## 1 Introduction

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 Along with the dynamic changes in the world, knowledge editing could help models forget outdated knowledge and adapt to the new counterpart without retraining from scratch.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

Q: How many Honkai series games released by miHoYo are there now? A: four



(a) Monolingual knowledge editing



(b) Cross-Lingual knowledge 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.

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.

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.,

\*Email: jawang.nlp@gmail.com

<sup>1</sup>Data and codes are available at <https://github.com/krystalan/Bi-ZsRE>

2023; Wang et al., 2023a). For example, LLMs such as LLaMA (Touvron 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 and BaiChuan), 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 locality in the other languages could also be influenced. 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 in knowledge editing. We achieved this by automatically 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 mainstream 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.

**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. 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.<sup>2</sup> Besides, this work does not conduct experiments with 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., 2023b), 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 a 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<sup>3</sup> 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. Finally, 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 (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 11.01. 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 which should be carefully considered during editing knowledge.

#### 3.3 Task Overview

**Knowledge Editing.** Given a language model  $p_\theta$  and an edit descriptor  $\langle x_e, y_e \rangle$ , the goal of knowledge editing is to create an edited model  $p'_\theta$  satisfy

<sup>2</sup>September 11, 2023

<sup>3</sup><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	11.01	11.10	4.51	14.53	6.55	16.11	5.67

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.

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  $p'_\theta$  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  $p'_\theta$  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  $p'_\theta$  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; 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. (2023b), we adopted 7 methods as baselines: (1) Directly fine-tuning (FT) the language models with  $L_\infty$  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.,



Method	Reliability	Generality		Locality		Portability	
		En	Zh	En	Zh	En	Zh
FT (En)	20.46 / 00.77	18.36 / 00.19	26.65 / 00.10	87.49 / 70.11	74.39 / 40.02	06.30 / 00.00	23.61 / 00.00
FT (Zh)	20.03 / 01.06	16.69 / 00.19	16.43 / 01.06	86.48 / 71.07	61.78 / 21.99	06.87 / 00.10	17.15 / 00.00
SERAC (En)	83.61 / 77.15	83.64 / 77.24	21.67 / 11.96	100.0 / 100.0	92.87 / 85.44	06.49 / 00.00	08.69 / 00.00
SERAC (Zh)	29.72 / 15.91	14.85 / 00.00	67.93 / 40.89	100.0 / 100.0	100.0 / 100.0	06.57 / 00.00	19.72 / 00.10
IKE (En)	99.90 / 99.90	99.24 / 98.36	93.74 / 72.52	62.79 / 36.26	44.57 / 12.34	50.86 / 17.84	35.34 / 04.53
IKE (Zh)	99.97 / 99.90	85.51 / 77.82	97.37 / 95.66	63.64 / 36.35	51.00 / 16.10	39.64 / 04.44	38.70 / 07.43
MEND (En)	49.27 / 00.77	48.26 / 00.58	17.41 / 00.00	90.50 / 77.24	89.75 / 70.40	05.86 / 00.10	14.74 / 00.10
MEND (Zh)	17.20 / 00.58	15.65 / 00.29	42.84 / 00.87	89.74 / 74.93	87.85 / 65.09	06.69 / 00.10	22.03 / 00.29
KN (En)	04.63 / 00.00	04.54 / 00.00	06.03 / 00.00	42.25 / 29.12	35.81 / 20.15	03.53 / 00.00	08.15 / 00.00
KN (Zh)	03.41 / 00.00	04.48 / 00.00	05.03 / 00.00	30.85 / 18.61	20.97 / 09.35	03.06 / 00.00	05.49 / 00.00
ROME (En)	99.17 / 97.88	94.44 / 88.81	29.01 / 09.64	99.01 / 96.43	97.23 / 91.13	08.72 / 00.00	16.16 / 00.00
ROME (Zh)	98.99 / 97.30	55.94 / 38.48	35.02 / 21.89	90.09 / 76.95	85.98 / 63.07	07.81 / 00.00	11.63 / 00.00
MEMIT (En)	95.74 / 91.22	89.09 / 78.69	30.48 / 07.91	98.41 / 95.18	97.64 / 92.38	08.23 / 00.00	17.91 / 00.00
MEMIT (Zh)	94.28 / 89.59	51.85 / 35.39	36.88 / 20.64	98.08 / 94.12	96.27 / 87.66	07.72 / 00.19	14.69 / 00.00

Table 2: Experimental results on the Chinese-LLaMA-Plus-7B backbone in terms of F1 / EM. Grey denotes the score is less than 10.0, while green indicates the corresponding score is more than 80.0.

	C-Eval	MMLU
GPT-4	68.7	86.4
GPT-3.5-turbo	54.4	70.0
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.

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 update 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 three LLMs in the experiments: (1) Chinese-LLaMA-Plus-7B<sup>4</sup> is created based on LLaMA-7B (Touvron et al., 2023a) with vocabulary extension and continual pre-training on the Chinese corpora. (2) In the same way, Chinese-LLaMA-2-7B<sup>5</sup> is created based on LLaMA-2-7B (Touvron et al., 2023b). (3) Baichuan-7B<sup>6</sup> is another LLM that sup-

ports 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 indicate their abilities in Chinese and English, respectively. Baichuan-7B performs the best among the three 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., 2023b) with the default settings. The hyper-parameters of each method can be found in the corresponding GitHub repository.<sup>7</sup>

## 4.2 Results & Analyses

Table 2, Table 4 and Table 5 show the experimental results on Chinese-LLaMA-Plus-7B, Chinese-LLaMA-2-7B and Baichuan-7B, respectively. The language identifier in the method name indicates its source language. For example, SERAC (En) means the method is edited in English, while SERAC (Zh) is in Chinese.

**Monolingual Analysis.** Compared with the other three properties, portability is more challenging for knowledge editing methods to achieve. As we can see, only *IKE* achieves decent performance in terms of portability while other methods fail. 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. Given the above analyses, we next compare the cross-lingual

<sup>4</sup><https://github.com/ymcui/Chinese-LLaMA-Alpaca>

<sup>5</sup><https://github.com/ymcui/Chinese-LLaMA-Alpaca-2>

<sup>6</sup><https://github.com/baichuan-inc/>

Baichuan-7B/

<sup>7</sup><https://github.com/zjunlp/EasyEdit/tree/main/hparams>

Method	Reliability	Generality		Locality		Portability	
		En	Zh	En	Zh	En	Zh
FT (En)	36.62 / 05.98	35.01 / 07.52	20.45 / 00.39	81.90 / 55.06	76.64 / 38.57	07.33 / 00.00	16.94 / 00.00
FT (Zh)	13.91 / 01.93	17.13 / 01.35	36.15 / 01.06	76.79 / 47.44	68.44 / 26.13	07.09 / 00.00	18.50 / 00.00
SERAC (En)	99.25 / 98.07	99.26 / 98.07	32.28 / 20.06	100.0 / 100.0	93.64 / 88.72	07.12 / 00.19	11.28 / 00.19
SERAC (Zh)	29.97 / 23.43	20.36 / 02.70	72.19 / 51.98	100.0 / 100.0	100.0 / 100.0	08.10 / 00.00	22.09 / 01.83
IKE (En)	100.0 / 100.0	99.69 / 99.32	91.90 / 77.15	56.35 / 30.76	54.40 / 12.15	45.72 / 11.76	37.50 / 05.11
IKE (Zh)	99.95 / 99.90	94.24 / 90.84	99.25 / 98.75	49.90 / 21.50	51.59 / 13.21	40.91 / 05.69	44.99 / 12.63
MEND (En)	61.25 / 00.00	60.78 / 00.00	22.84 / 00.00	93.55 / 81.68	81.68 / 50.92	07.74 / 00.00	14.70 / 00.00
MEND (Zh)	22.26 / 00.00	21.64 / 00.00	45.95 / 00.00	96.68 / 90.45	96.26 / 88.24	07.04 / 00.00	21.72 / 00.00
KN (En)	10.94 / 00.00	10.96 / 00.00	11.71 / 00.00	49.28 / 06.85	43.65 / 09.74	05.75 / 00.00	13.54 / 00.00
KN (Zh)	08.36 / 00.00	10.24 / 00.00	11.38 / 00.00	45.06 / 03.76	36.65 / 03.95	05.84 / 00.00	13.03 / 00.00
ROME (En)	78.19 / 68.76	72.91 / 57.47	24.43 / 05.40	94.17 / 83.70	96.02 / 87.08	07.66 / 00.19	16.52 / 00.00
ROME (Zh)	27.47 / 10.80	22.10 / 03.18	60.90 / 15.53	94.09 / 82.55	94.71 / 82.16	06.63 / 00.10	24.21 / 01.93
MEMIT (En)	83.67 / 76.76	77.55 / 62.20	25.20 / 06.17	98.41 / 95.37	97.87 / 93.35	08.20 / 00.19	16.79 / 00.10
MEMIT (Zh)	28.56 / 11.76	22.89 / 04.05	63.98 / 16.39	98.53 / 95.56	97.82 / 92.48	07.13 / 00.10	24.22 / 01.64

Table 4: Experimental results on the Chinese-LLaMA-2-7B backbone in terms of F1 / EM.

Method	Reliability	Generality		Locality		Portability	
		En	Zh	En	Zh	En	Zh
FT (En)	33.33 / 13.11	27.09 / 07.43	20.79 / 00.29	91.71 / 83.12	85.08 / 63.74	09.21 / 00.19	29.40 / 00.00
FT (Zh)	13.76 / 02.31	14.06 / 00.77	28.45 / 04.73	95.38 / 90.07	60.78 / 20.25	09.31 / 00.29	26.88 / 00.29
KN (En)	10.77 / 00.00	10.32 / 00.00	18.97 / 00.00	71.28 / 55.74	91.99 / 79.27	08.96 / 00.19	29.98 / 00.00
KN (Zh)	10.10 / 00.00	10.53 / 00.00	18.71 / 00.00	73.16 / 58.15	85.95 / 65.38	09.08 / 00.29	30.02 / 00.10
ROME (En)	68.70 / 52.36	59.08 / 40.50	24.76 / 01.64	98.28 / 96.05	98.92 / 95.27	09.90 / 00.19	30.24 / 00.29
ROME (Zh)	25.75 / 08.39	18.69 / 03.38	69.95 / 14.75	97.86 / 95.47	97.70 / 92.57	09.54 / 00.29	26.21 / 01.06
MEMIT (En)	70.78 / 54.29	64.46 / 46.48	27.49 / 03.57	99.22 / 97.69	98.93 / 96.14	09.72 / 00.10	28.83 / 00.29
MEMIT (Zh)	25.60 / 08.78	22.77 / 07.62	71.44 / 16.30	98.47 / 96.53	97.92 / 93.54	09.13 / 00.39	23.83 / 00.29

Table 5: Experimental results on the Baichuan-7B backbone in terms of F1 / EM.

knowledge editing performance on *SERAC*, *IKE*, *MEND*, *ROME* and *MEMIT*.

**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 83.61 F1 while *SERAC* (Zh) only achieves 29.72 F1 on Chinese-LLaMA-Plus-7B. 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. Besides, when comparing three backbones, we find that *ROME* (En) and *ROME* (Zh) achieve similar reliability (99.17 F1 vs. 98.99 F1) on Chinese-LLaMA-Plus-7B, but significantly different on Chinese-LLaMA-2-7B (78.19 F1 vs. 27.47 F1) and Baichuan-7B (70.78 F1 vs. 25.60 F1). *MEMIT* (En) and *MEMIT* (Zh) also show this situation.

**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 83.64 F1 in English generality but only performs 28.46 F1 in Chinese generality. In contrast, *SERAC* (Zh) achieves better Chinese generality than English (67.93 F1 vs. 15.00 F1). 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. The degree of influence appears to be similar in different languages. For example, *MEND* (En) achieves 90.50 F1 and 89.75 F1 in English and Chinese locality, while the counterparts in *MEND* (Zh) are 89.74 and 87.85. We also find that though *IKE* works well in terms of reliability, 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. The lower the locality, the higher the potential risk.

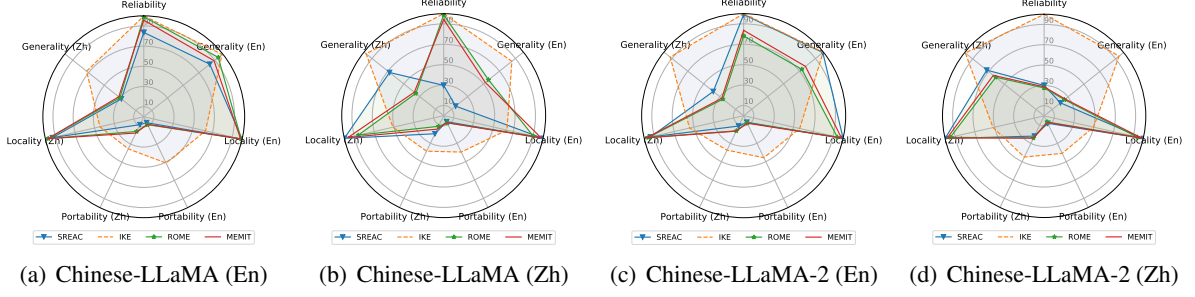


Figure 2: Radar chart of knowledge editing performance when editing different LLMs with different languages.

Method	Cross-Lingual Generality		
	Before	After	$\Delta$
FT (En)	20.45 / 00.39	21.34 / 00.39	+0.89 / 00.00
FT (Zh)	17.13 / 01.35	18.75 / 01.54	+1.62 / +0.19
SERAC (En)	32.28 / 20.06	38.10 / 21.22	+5.82 / +1.16
SERAC (Zh)	20.36 / 02.70	20.55 / 02.70	+0.19 / 00.00
IKE (En)	91.90 / 77.15	92.12 / 77.24	+0.22 / +0.09
IKE (Zh)	94.24 / 90.84	94.27 / 90.84	+0.03 / 00.00
MEND (En)	22.84 / 00.00	25.94 / 00.00	+3.10 / 00.00
MEND (Zh)	21.64 / 00.00	24.08 / 00.00	+2.44 / 00.00
KN (En)	11.71 / 00.00	11.81 / 00.00	+0.10 / 00.00
KN (Zh)	10.24 / 00.00	11.14 / 00.00	+0.90 / 00.00
ROME (En)	24.43 / 05.40	26.05 / 05.50	+1.62 / +0.10
ROME (Zh)	22.10 / 03.18	25.60 / 04.73	+3.50 / +1.55
MEMIT (En)	25.20 / 06.17	27.42 / 06.46	+2.22 / +0.29
MEMIT (Zh)	22.89 / 04.05	26.51 / 06.08	+3.62 / +2.03

Table 6: The cross-lingual generality performance before and after relaxing the evaluation settings (backbone: Chinese-LLaMA-2-7B).

**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 indicates 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, LLMs might reflect inconsistent behaviors on the edited knowledge in different languages.

### 4.3 The Influence of Language Mismatch

When querying an edited multi-lingual LLM with target-language questions, it may output source-language answers, named language mismatch. To figure out the influence of language mismatch, we attempt to relax the settings for assessing generality: there are two golden answers for a single generality question, one is in English while the other is in Chinese, and both two share the same se-

mantics. When calculating EM and F1 scores using questions in the target language, both two golden answers are used for comparisons, and the highest score will be recorded. In this manner, when querying the model with target-language questions, the source-language answers could also be accepted. As shown in Table 6, the cross-lingual generality slightly increases when we relax the evaluation settings, indicating the influence of language mismatch indeed exists despite being slight.

## 5 Conclusion

In this paper, we first explore the cross-lingual effect of knowledge editing. 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 multi-lingual LLMs, and study the cross-lingual effect from English to Chinese and vice versa. Our results indicate that (1) the language modeling gaps of different languages might influence the efficiency of knowledge editing in different languages; (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 locality in the other languages could also be influenced. 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.

## References

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.
- 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.
- 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*.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.
- Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, et al. 2023b. 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*.
- 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*.