Edit Once, Update Everywhere: A Simple Framework for Cross-Lingual Knowledge Synchronization in LLMs

Yuchen Wu[1], **Liang Ding**[2*], **Li Shen**[3], **Dacheng Tao**[4]
[1]Shanghai Jiao Tong University, China 200240
[2]The University of Sydney, Australia 2006
[3]Shenzhen Campus of Sun Yat-sen University, China 518107
[4]Nanyang Technological University, Singapore 639798

Abstract

Knowledge editing allows for efficient adaptation of large language models (LLMs) to new information or corrections without requiring full retraining. However, prior methods typically focus on either single-language editing or basic multilingual editing, failing to achieve true cross-linguistic knowledge synchronization. To address this, we present a simple and practical state-of-the-art (SOTA) recipe *Cross-Lingual Knowledge Democracy Edit* (X-KDE), designed to propagate knowledge from a dominant language to other languages effectively. Our X-KDE comprises two stages: (i) Cross-lingual Edition Instruction Tuning (XE-IT), which fine-tunes the model on a curated parallel dataset to modify in-scope knowledge while preserving unrelated information, and (ii) Target-language Preference Optimization (TL-PO), which applies advanced optimization techniques to ensure consistency across languages, fostering the transfer of updates. Additionally, we contribute a high-quality, cross-lingual dataset, specifically designed to enhance knowledge transfer across languages. Extensive experiments on the Bi-ZsRE and MzsRE benchmarks show that X-KDE significantly enhances cross-lingual performance, achieving an average improvement of +8.19%, while maintaining high accuracy in monolingual settings. Our code is available at: https://github.com/YukinoshitaKaren/X_KDE.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Dubey et al., 2024; Yang et al., 2024a; Guo et al., 2025) have shown strong capabilities in natural language understanding, generation, and reasoning (Wei et al., 2022; Zhong et al., 2023a; Peng et al., 2023; Zhao et al., 2023). However, as world knowledge evolves, LLMs need methods to update outdated information efficiently. Knowledge editing (Yao et al., 2023) allows modifications to specific knowledge while preserving unrelated information, making it more cost-effective than retraining from scratch.

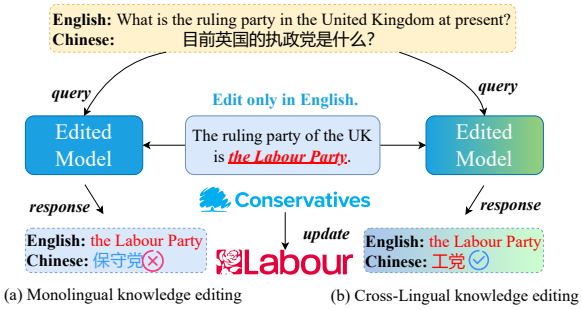

Figure 1: Examples of (a) **monolingual** and (b) **cross-lingual** knowledge editing. In the former, the editing and verification languages are the same, while in the latter, knowledge is transferred from the source language (e.g., English) to the target language (e.g., Chinese).

Despite significant progress, most existing approaches focus on monolingual editing (De Cao et al., 2021; Dai et al., 2021; Mitchell et al., 2021). As LLMs are increasingly required to handle multilingual queries (Zhang et al., 2024; Wang et al., 2023b), monolingual solutions often fail. For example, when editing the response "the Conservative Party" to "the Labor Party" in English (Figure 1(a)), this update does not propagate to the Chinese version. Thus, expanding knowledge editing to cross-lingual settings is crucial to ensure that changes made in the source language are correctly applied to target languages.

Currently, several studies on multilingual knowledge editing have emerged (Xu et al., 2022; Wang et al., 2023d; Wei et al., 2024; Xie et al., 2024; Liang et al., 2024). Some of these methods extend the edited language from single to multiple, while others prescribe source-language answers as the ground truth for multilingual queries. Both strategies fall short of achieving true cross-lingual knowledge democratization. For example, although IKE

*Correspond to Liang Ding liangding.liam@gmail.com

Findings of the Association for Computational Linguistics: ACL 2025, pages 23282–23302

was regarded as the state-of-the-art method in previous studies (Wang et al., 2023a; Xie et al., 2024), its performance on the Bi-ZsRE benchmark demonstrates significant limitations, achieving merely a 73.33 average score when editing in English. Unlike previous methods that attempt to forcefully correct LLM behaviors in both the source and target languages, we propose guiding LLMs to internalize knowledge from source language editing and apply it to target language queries. Our *Cross-Lingual Knowledge Democracy Edit* (X-KDE) with Dual-Stage Refinement, where we use parallel language datasets to transfer knowledge from the source to the target language.

The X-KDE framework involves two phases: (i) Cross-lingual edition Instruction Tuning (XE-IT), where the source language editing descriptor is paired with target language queries to create a parallel dataset, guiding the model to answer in the target language while preserving unchanged knowledge. (ii) Target-language Preference Optimization (TL-PO), where we adopt the ORPO strategy (Hong et al., 2024), further constrains cross-lingual knowledge, promoting the diffusion of updates from source to target languages, and achieving true knowledge democratization. Taking the Bi-ZsRE benchmark as an example, X-KDE outperforms others, achieving average scores of 91.04 and 88.49 when editing in English and Chinese, respectively. Our **contributions** are three-fold:

- To tackle the scarcity of high-quality resources in cross-lingual knowledge editing, we introduce new datasets that fill gaps in existing resources, enhancing the reliability of knowledge transfer across languages.
- We propose X-KDE, a simple yet highly effective method for cross-lingual knowledge editing. This approach, based on a two-stage process, ensures robust knowledge generalization across languages.
- Through extensive experiments, we establish X-KDE as a new state-of-the-art (SOTA) solution for cross-lingual knowledge editing, demonstrating significant improvements in performance while preserving original knowledge and enhancing the portability of updates.

2 Preliminary

2.1 Knowledge Editing

Knowledge editing selectively modifies in-scope knowledge while preserving out-of-scope behavior. Given a base LLM p_θ and an *edit descriptor* $\langle x_e, y_e \rangle$, where x_e is the modification description and y_e is the corresponding answer, the edited model should adhere to four key properties:

Reliability evaluates accuracy on edit descriptors:

$$\mathbb{E}_{(x_e,y_e)\sim\mathcal{X}_e} \mathbf{1}[\arg\max_y p^*_\theta(y|x_e) = y_e] \quad (1)$$

Generality assesses the precision of semantically rephrased examples:

$$\mathbb{E}_{(x_e^{par},y_e)\sim\mathcal{X}_e^{par}} \mathbf{1}[\arg\max_y p^*_\theta(y|x_e^{par}) = y_e] \quad (2)$$

Locality ensures that out-of-scope inputs remain unchanged:

$$\mathbb{E}_{(x_e,y_e)\sim\mathcal{O}_e} \mathbf{1}[p^*_\theta(y|x_e) = p_\theta(y|x_e)] \quad (3)$$

Portability measures the ability to transfer updated knowledge to related queries:

$$\mathbb{E}_{(x_e,y_e)\sim\mathcal{I}_e} \mathbf{1}[\arg\max_y p^*_\theta(y|x_e) = y_e] \quad (4)$$

2.2 Cross-Lingual Knowledge Editing

Cross-lingual knowledge editing extends monolingual knowledge editing by requiring a multilingual LLM $p_{m\theta}$ to propagate knowledge from a source language to a target language. Given an edit descriptor in the source language $\langle x_e^s, y_e^s \rangle$, the goal is to maximize:

$$\mathbb{E}_{\substack{(x_e^s,y_e^s)\sim\mathcal{X}_e^s \\ x_e^t=I^t(x_e^s),y_e^t=I^t(y_e^s)}} \mathbf{1}[\arg\max_y p^*_{m\theta}(y|x_e^t) = y_e^t] \quad (5)$$

$$\mathbb{E}_{\substack{(x_e^s,y_e^s)\sim\mathcal{O}_e^s \\ x_e^t=I^t(x_e^s),y_e^t=I^t(y_e^s)}} \mathbf{1}[p^*_{m\theta}(y|x_t) = p_{m\theta}(y|x_t)] \quad (6)$$

Here, x_e^t, y_e^t are the edit descriptors in the target language t, and $I^t(\cdot)$ translates the source language input into the target language. Cross-lingual knowledge editing demands cross-lingual comprehension, ensuring that updates in the source language lead to consistent responses in the target language.

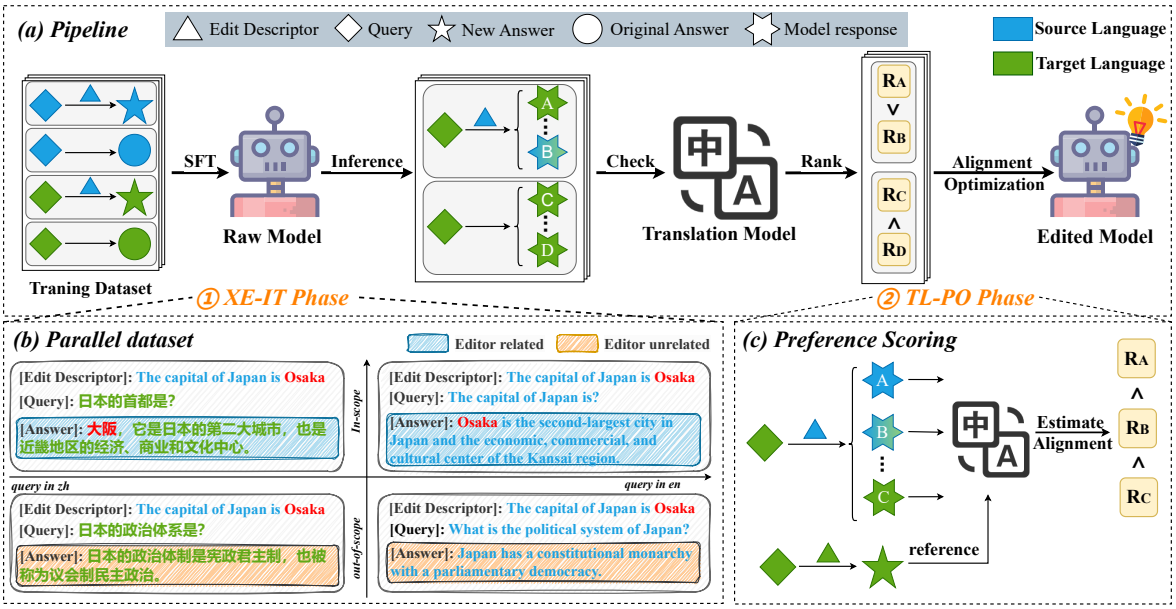

Figure 2: **Illustration of *Cross-Lingual Knowledge Democracy Edit* (X-KDE) framework**. In the XE-IT phase, we fine-tune the LLM on a carefully curated parallel dataset, enabling it to incorporate newly edited information from the source language when queried in the target language. In the TL-PO phase, multiple responses are generated, ranked based on similarity to the target language answer, and alignment optimization is applied to refine the output.

3 Methodology

To achieve the democratization of knowledge, we propose the *Cross-Lingual Knowledge Democracy Edit* (X-KDE) framework, as shown in Figure 2. This framework enables LLMs to adapt to evolving knowledge demands and facilitates the transfer of knowledge to target languages by editing only the source language. X-KDE consists of two stages: the Cross-lingual Edition Instruction Tuning (XE-IT) phase and the Target-language Preference Optimization (TL-PO) phase.

3.1 XE-IT Stage: Learning to Edit

The goal is to enable the model to leverage knowledge edits in the source language while preserving the unchanged information. To meet the requirements for cross-lingual editing, we carefully constructed a high-quality dataset and employed XE-IT to fine-tune the model.

3.1.1 Dataset Construction

Data Sources. Our goal is to enable the model to use edit descriptors effectively while maintaining unrelated information. We use several widely used knowledge editing datasets, including ZsRE (Levy et al., 2017), HalluEditBench (Huang et al., 2024a), RIPPLEEDITS (Cohen et al., 2024), WikiBio (Hartvigsen et al., 2024), MQUAKE (Zhong et al., 2023b), and COUNTERFACT (Meng et al., 2022a), to build our training data. These datasets provide edit descriptors paired with QA pairs. Additionally, following LTE (Jiang et al., 2024), we incorporate Evol-Instruct (Xu et al., 2023) to better preserve the language capabilities of the LLM. To mitigate data leakage, we randomly sample and translate subsets for training.

Sample Generation. Existing datasets often feature straightforward QA pairs, which limit the model's comprehension ability. To address this, we use Deepseek (Liu et al., 2024) to generate complex in-scope and out-of-scope query-answer pairs. This method enhances training data quality and model comprehension.

Quality Control. To ensure relevance, we use Deepseek to assess the quality of in-scope query-answer pairs. Samples are scored based on Syntactic Structure, Lexical Richness, and Edit Consistency, with low-quality samples filtered out and replaced by higher-scoring ones.

Translation Process. We use large language models, i.e., Deepseek[1] to translate the generated data from English to Chinese.

[1] https://api-docs.deepseek.com/

Parallel Data Construction. Our dataset follows a parallel structure (Figure 2(b)), with the in-scope section guiding LLMs when to use updated knowledge and the out-of-scope section minimizing the impact on unrelated knowledge. The dataset includes both monolingual and cross-lingual sections, where the source language contains edit descriptors and the target language contains queries and answers. Further details about our dataset are provided in Appendix A.3.

3.1.2 Fine-Tuning

Thanks to the flexible parallel structure, we can adaptively select the source and target languages to satisfy specific needs. We create a large-scale cross-lingual dataset and compute loss based only on the answer tokens. The model generates answers in the target language given source-language edit descriptors and target-language queries.

3.2 TL-PO Stage: Preference Optimization

3.2.1 Preference Scoring

After the XE-IT phase, the model has initially acquired the ability for cross-lingual knowledge editing. However, when faced with queries in the target language, the model may still make mistakes, such as generating responses in the source language, producing surface-level transliterations, or failing to follow target language patterns. To address this, we use a multilingual translation model to compute "alignment" scores, favoring responses aligned with the target language. More details about the "alignment" score computation can be found in Appendix B.1.

3.2.2 Alignment Optimization

When an edit is performed in the source language and the query is made in the target language, we aim for the model to generate answers in the target language with higher likelihood than in the source language: $p^*_\theta(y^t_e|x^s_e) > p^*_\theta(y^s_e|x^s_e)$. To achieve this, we employ ORPO, a state-of-the-art preference optimization method. We collect flawed outputs (Y_l) and preferred outputs (Y_w), then optimize the objective function:

$$\mathcal{L}_{ORPO} = \mathbb{E}_{(x,y_w,y_l)} \left[\mathcal{L}_{XE-IT} + \lambda \cdot \mathcal{L}_{OR} \right] \quad (7)$$

where $\mathcal{L}_{XE-IT}$ is the XE-IT loss and $\mathcal{L}_{OR}$ maximizes the likelihood ratio between the preferred response and the less preferred one.

$$\mathcal{L}_{XE-IT} = -\frac{1}{m} \sum_{t=1}^{m} \log P_\theta(y_t|x, y_{<t}) \quad (8)$$

$$\mathcal{L}_{OR} = -\log \sigma \left(\log \frac{\textbf{odds}_\theta(y_w|x)}{\textbf{odds}_\theta(y_l|x)} \right) \quad (9)$$

$$\textbf{odds}_\theta(y|x) = \frac{P_\theta(y|x)}{1 - P_\theta(y|x)} \quad (10)$$

4 Experiments

4.1 Experimental Setup

Baselines. We chose the following methods as baselines: (1) **FT-L** (Meng et al., 2022a) fine-tunes a specific layer of the feed-forward network to maximize the likelihood of target tokens; (2) **FT-M** (Zhang et al., 2024) fine-tunes the same feed-forward network layer as FT-L. Additionally, it masks the original text and applies cross-entropy loss on the target answer; (3) **ROME** (Meng et al., 2022a) employs causal mediation analysis to identify the target area for editing, and then updates the parameters of the feed-forward network layers; (4) **MEMIT** (Meng et al., 2022c), built upon the ROME framework, enables the simultaneous update of thousands of knowledge; (5) **IKE** (Zheng et al., 2023) utilizes the in-context-learning ability of the model and provides a few-shot demonstration to guide the model's responses based on the updated facts. (6) **LTE** (Jiang et al., 2024) enhances the model's instruction-following ability through supervised fine-tuning (SFT), and employs a retrieval-based mechanism to provide updated knowledge for demonstrations.

Dataset. We evaluate X-KDE and Baselines using two widely used benchmarks: Bi-ZsRE dataset (Wang et al., 2023a) and MzSRE dataset (Wang et al., 2023d). The former contains 1,037 test samples in English and an equal number in Chinese, while the latter covers twelve languages: English, Czech, German, Dutch, Spanish, French, Portuguese, Russian, Thai, Turkish, Vietnamese and Chinese, with 743 samples in each.

Backbones. We select two public models as backbones, including LLaMA2-Chat-7B (Touvron et al., 2023) and Qwen2.5-instruct-7B (Yang et al., 2024a). These models are widely used in chatbot applications, the former excels in English, while the latter demonstrates strong multilingual abilities.

Method	Test in English				Test in Chinese				
	Reliability	Generality	Locality	Portability	Reliability	Generality	Locality	Portability	*Avg.*
Edit in English									
FT-L	53.51	50.18	94.01	53.31	51.81	51.71	85.56	55.14	61.90
FT-M	99.97	95.38	97.92	57.69	56.89	56.52	94.61	52.16	76.39
ROME	96.09	84.69	98.04	58.87	49.94	50.31	97.70	51.81	73.43
MEMIT	95.21	89.14	98.56	57.77	52.05	52.01	98.76	52.19	74.46
IKE	99.59	99.61	56.95	71.27	67.83	67.88	64.54	58.97	73.33
LTE	99.91	99.81	88.97	77.40	76.86	76.82	86.99	67.49	84.28
X-KDE(Ours)	99.93	99.87	90.15	76.41	94.81	94.65	95.05	77.43	91.04
Edit in Chinese									
FT-L	40.80	40.66	94.80	55.24	54.72	53.68	66.51	48.75	56.89
FT-M	51.86	51.24	98.18	55.30	100.0	99.71	79.28	61.98	74.69
ROME	44.14	43.80	97.92	52.66	72.24	70.12	96.48	48.15	65.69
MEMIT	45.37	44.95	99.07	54.65	75.19	73.45	96.02	51.44	67.52
IKE	65.87	65.74	69.41	63.06	99.86	99.73	64.86	72.39	80.79
LTE	64.63	62.56	85.23	62.6	99.79	99.31	87.17	69.69	78.87
X-KDE(Ours)	93.49	92.22	90.56	65.55	100.00	99.11	92.85	74.14	88.49

Table 1: **Cross-lingual editing performance of different methods** on Llama2-chat-7B backbones. Results in green indicates the best results. "***Avg.***" represents the overall mean of all metrics evaluated across the two languages.

Metrics	Methods	en-en	en-cz	en-de	en-du	en-es	en-fr	en-pt	en-ru	en-th	en-tr	en-vi	en-zh	en-avg
Reliability	FT-L	52.92	41.81	39.79	39.02	39.49	39.72	39.26	39.79	36.44	36.86	46.21	51.81	41.93
	FT-M	99.96	66.93	70.16	67.17	63.69	64.98	64.22	48.96	36.46	57.54	66.80	56.89	63.65
	ROME	96.36	56.54	60.82	58.89	57.41	56.43	54.91	41.69	35.44	45.76	56.94	49.94	55.93
	MEMIT	95.44	62.37	64.82	64.12	59.46	61.90	58.69	44.54	36.40	49.15	61.34	52.05	59.19
	IKE	99.65	83.22	80.61	79.36	76.69	78.48	75.37	67.62	54.38	76.90	81.22	67.83	76.78
	LTE	100.00	84.29	81.71	80.60	77.67	79.11	77.39	72.02	62.04	78.87	81.92	76.93	79.38
	X-KDE	99.93	92.78	87.43	88.89	85.71	87.49	89.87	89.32	89.66	91.23	87.55	93.07	90.24
Generality	FT-L	49.60	40.75	38.87	38.36	39.68	39.12	39.56	38.97	36.89	37.18	45.89	51.71	41.38
	FT-M	95.53	65.45	68.15	65.09	62.39	62.28	61.63	47.69	36.88	56.87	65.97	56.52	62.04
	ROME	85.13	54.99	58.91	56.99	56.58	54.47	53.94	40.68	35.36	45.06	56.38	50.31	54.07
	MEMIT	89.59	60.71	63.80	61.98	58.10	59.40	57.63	43.31	36.77	48.68	60.51	52.01	57.71
	IKE	99.54	82.67	80.78	79.18	76.37	78.22	75.49	67.51	54.26	76.97	80.99	67.88	76.65
	LTE	99.87	84.26	81.63	81.07	77.51	78.99	77.38	71.46	61.90	78.26	81.37	76.24	79.16
	X-KDE	99.68	92.87	87.25	88.87	85.16	87.57	89.93	89.10	89.21	91.25	87.62	93.11	90.14

Table 2: **Results on MzsRE dataset for editing performed in English** using Llama2-7b-chat. Here, "en-zh" means that English serves as the source language and Chinese as the target language, with similar interpretations for the other pairs. "en-avg" denotes the average performance across cross-lingual scenarios.

For brevity, the results on Qwen2.5-instruct-7B are provided in D.2.

4.2 Results of Single Fact Editing

Table 1 and Table 2 demonstrate the main results of single fact editing, which focus on single editing cases. From these results, we can find several significant observations:

X-KDE outperforms other methods in the cross-lingual setting by a significant margin. As shown in Table 1, when edited in English, it is evident that our method brings average performance improvements of 6.76%, compared to LTE. In particular, in the cross-lingual setting, our method achieves further performance gains in portability, which demonstrates that our method not only captures surface-level changes in wording but enables the LLM to effectively internalize the knowledge edited in the source language and apply it to the target language. In summary, LTE sets a new state-of-the-art in cross-lingual knowledge editing task.

X-KDE brings consistent and effective improvements in more complex multilingual environments. Our method is effective not only in a bilingual Chinese-English setting but can also be generalized to additional languages. We conducted more extensive experiments on the Mzsre dataset, and the results are presented in Table 2. More detailed results in Appendix D.1. It is evident that, compared to LTE, our method exhibits 10.86% in reliability and 10.98% average gains in generality when edit in English. These results further demon-

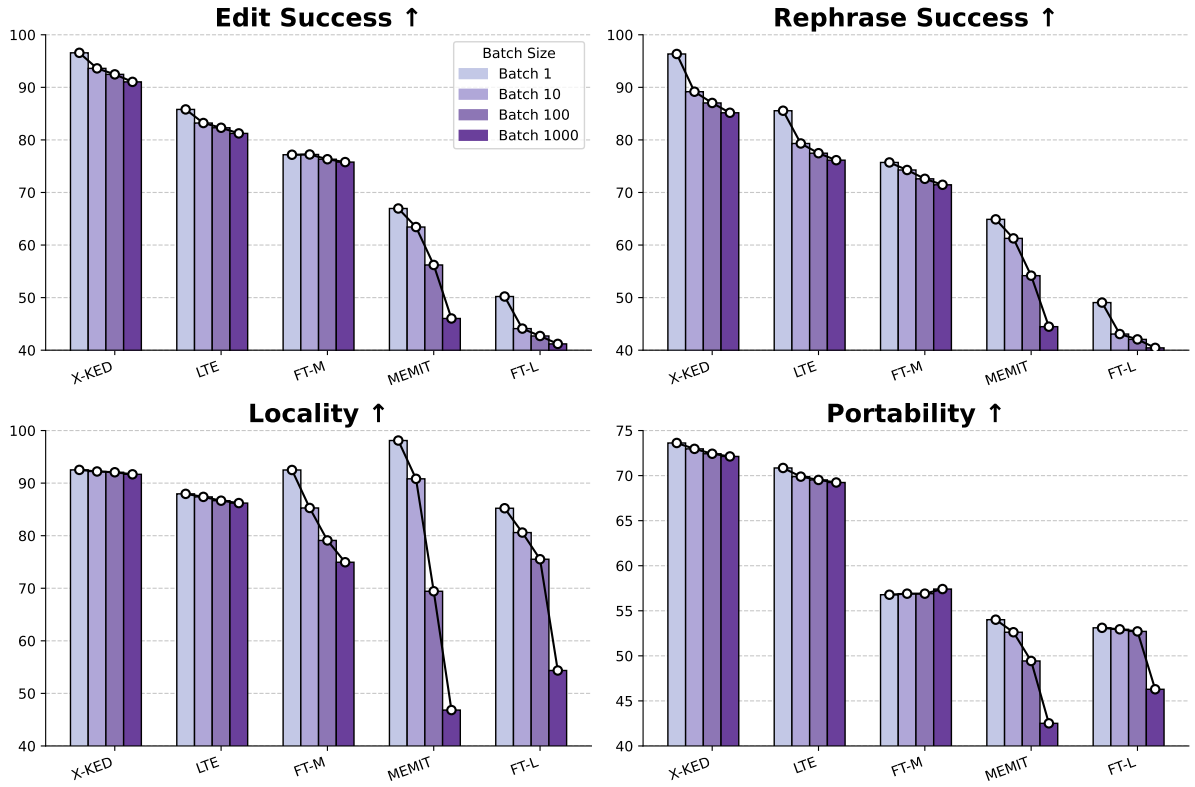

Figure 3: **Mean batch-editing performance across four benchmarks** at batch sizes 1, 10, 100, and 1000.

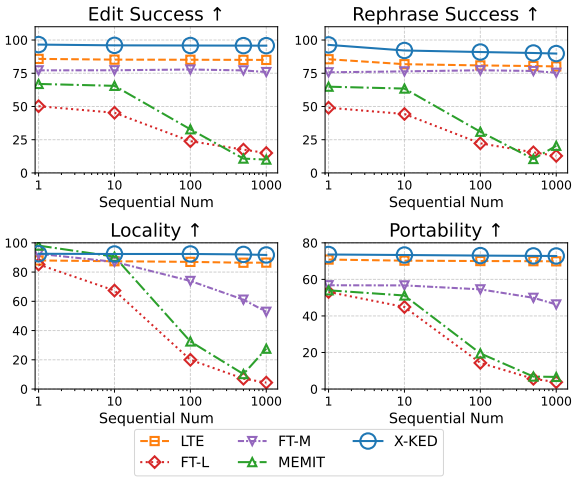

Figure 4: **Mean sequential-editing results across four knowledge editing benchmarks**, shown for data stream sizes of 1, 10, 100, 500, and 1000 (log-scale).

strate that, our approach significantly enhances the model's cross-lingual abilities, enabling it to effectively apply knowledge from a pivot language to others and marking a significant step toward the democratization of knowledge.

4.3 Results of Mass Fact Editing

In the previous section, we introduced the results of single fact editing. However, real-world scenarios are often more complex, requiring simultaneous or sequential edits to multiple pieces of knowledge. Therefore, in this section, we conduct comprehensive experiments using X-KDE alongside several methods that support mass editing (FT-L, FT-M, MEMIT, and LTE) on LLaMA2-Chat-7B in both batch-edit and sequential-edit settings, and then present the corresponding results.

X-KDE can process thousands of edits simultaneously. In line with the single-edit procedure, we evaluate both English and Chinese edits separately. For simplicity in presentation, we take the average of these two results, as shown in Figure 3. As the batch size increases, we observe a gradual decline across all performance metrics for all method. The drop is particularly severe for MEMIT and FT-L, especially in the locality metric, which is nearly cut in half. In contrast, X-KDE achieves the best performance, maintaining the highest accuracy while exhibiting the slowest degradation rate. These results indicate that our approach remains stable in cross-lingual settings, even after thousands of edits.

X-KDE can sequential acquire new knowledge without forgetting previous information. In the sequential-editing setting, the model integrates new knowledge on top of its previous edits, which leads to a gradual decline in performance over time. As

	Test Language: en					Test Language: zh				
	MMLU	**CommonSenseQA**	**PIQA**	**Xsum**	**NQ**	**CMMLU**	**CommonSenseQA_zh**	**CEval**	**NQ_zh**	***avg***
LLaMA2-Chat-7B	44.78	**64.21**	**66.43**	**21.24**	19.39	20.64	4.67	30.08	0	30.16
X-KDE	**47.67**	59.46	61.64	20.96	**29.36**	**32.63**	**45.13**	**30.75**	**14.07**	**37.96**

Table 3: **General tasks performance of X-KDE, and LLaMA2-Chat-7B**. Results in **bold** represent the best performance in each category.

illustrated in Figure 4, the performance of methods that modify model parameters typically degrades as the number of edits increases. For instance, MEMIT and FT-L remain stable only when the number of edits $n \leq 10$; beyond that, their performance deteriorates sharply. In contrast, knowledge storage-retrieval paradigms represented by X-KDE and LTE circumvent direct parameter modifications through external memory architectures. Moreover, X-KDE demonstrates superior cross-lingual transfer capabilities compared to LTE, achieving better performance across diverse data streams.

4.4 Results of General Tasks

A series of studies have demonstrated that knowledge editing can influence model performance across various scenarios (Yang et al., 2024b; Li et al., 2023b; Gu et al., 2024). To investigate whether our method impact the model's capabilities in unrelated domains, we conducted tests across a range of fields. Given that the cross-lingual knowledge editing task typically involves two languages, we use English and Chinese as representative examples. Multiple benchmarks are selected in these two languages, covering tasks such as commonsense reasoning, natural language understanding, open-domain QA, and general intelligence. For example, the benchmarks selected for English include MMLU (Hendrycks et al., 2020), CommonSenseQA (Talmor et al., 2018), PIQA (Bisk et al., 2020), XSum (Narayan et al., 2018), and Natural Questions (Kwiatkowski et al., 2019). For Chinese, the chosen benchmarks are CMMLU (Li et al., 2023a), CommonSenseQA_zh (Contributors, 2023), CEval (Huang et al., 2024b), and Natural Questions_zh (Contributors, 2023). We conducted all experiments using the OpenCompass tool (Contributors, 2023). The results are presented in Table 3. Overall, our method not only preserves the model's performance in English but also significantly enhances its capabilities in Chinese. Although certain tasks, such as CommonsenseQA, XSum, and PIQA, show a decrease in performance when tested in English, the overall results demonstrate consistent English capabilities. This highlights the robustness of our method, which achieves cross-lingual knowledge editing while preserving the model's original performance and significantly improving its proficiency in Chinese.

5 Analysis

5.1 Are both stages of X-KDE indispensable?

We examine the significance of the two stages in the X-KDE method through our ablation experiments, as shown in Table 4. Focusing solely on the improvement in performance metrics, Stage 1 undoubtedly plays a decisive role in our method, achieving significant gains (up to +25.64% average score) compared to the untrained baseline model. This stage enables cross-lingual knowledge editing via in-context learning, providing a substantial boost in model performance. While Stage 2 appears to offer a smaller improvement (a 2.25% average gain), a closer analysis highlights its practical importance. Stage 2 is particularly effective in adjusting the model's output style to align with the target language, addressing issues such as incorrect language mixing (e.g., code-switching) or failure to generate responses in the expected linguistic format. For example, after Stage 1 updates the knowledge, the model may still produce a year like "2006" in the source language format. Stage 2 ensures the correct linguistic form, such as "2006年" in Chinese. The optimization of target-language preferences using ORPO not only improves factual accuracy but also ensures stylistic appropriateness in the target language. By refining the model's preferences, ORPO helps it better adapt to the cultural and grammatical norms of the target language, addressing challenges like code-switching and maintaining consistency across multilingual contexts.

5.2 Does every composition of the training data matter?

In this section, we focus on the composition of the training data. As shown in Table 5, the absence of any specific segment of the training data leads to a

Methods	Stages		Score	
	Stage-1	Stage-2	en-avg	zh-avg
Origin	✗	✗	62.41	64.96
X-KDE	✓	✗	88.05	86.99
	✗	✓	77.14	66.38
	✓	✓	91.04	88.49

Table 4: **Ablation results on Bi-ZsRE benchmark** with Llama2-7b-chat as the base model. The *en-avg* and *zh-avg* columns denote average scores when editing in English or Chinese, respectively.

noticeable decline in editing performance, whether in monolingual or cross-lingual settings. Excluding either monolingual or cross-lingual training data causes a sharp drop in performance in the corresponding areas. When the in-scope data is omitted, the model tends to retain its original knowledge, resulting in reduced reliability, generality, and portability. On the other hand, removing the out-of-scope data causes the model to overly depend on the updated knowledge, thus diminishing locality. Similarly, removing the edit descriptors from the training data prevents the model from effectively utilizing the updated knowledge, leading to a drop in all metrics. Interestingly, training data with longer text samples seems to enhance the model's comprehension and improve portability. In summary, each component of the training data plays a unique and indispensable role, and omitting any part negatively impacts the model's performance across all key metrics.

Methods	en-en				en-zh			
	R.	G.	L.	P.	R.	G.	L.	P.
X-KDE	99.93	99.87	90.15	76.41	94.81	94.65	95.05	77.43
-w/o mono. data	78.93	76.60	77.33	68.21	94.6	94.52	94.73	76.02
-w/o coss-lin. data	99.91	99.81	88.97	77.40	76.86	76.82	86.99	67.49
-w/o in-scope	81.02	82.58	93.93	69.15	75.17	74.85	93.56	69.62
-w/o out-of-scope	99.99	99.45	70.71	76.64	92.91	92.73	76.21	73.48
-w/o edit descriptor	87.53	81.99	67.69	66.53	84.26	84.12	79.43	74.13
-w/o long-text	100	99.82	93.54	73.63	92.35	92.75	93.16	72.46

Table 5: **Ablation results in the monolingual and the cross-lingual setting**, examining "reliability" (R.), "generality" (G.), "locality" (L.), and "portability" (P.).

5.3 Why choose ORPO as the preferred optimization method?

We evaluate several popular preference optimization methods using a held-out dataset from our training data, specifically direct policy optimization (DPO)(Rafailov et al., 2023), contrastive preference optimization (CPO)(Xu et al., 2024), Kahneman-Tversky Optimization (KTO)(Ethayarajh et al., 2024), and odds ratio preference optimization (ORPO)(Hong et al., 2024) on the Bi-ZsRE benchmark. As shown in Table 6, ORPO outperforms the other methods, achieving the best overall performance. While CPO and KTO also yield similar improvements, ORPO excels in preserving irrelevant target-language samples, demonstrating superior locality. In contrast, DPO results in a performance decline, which we attribute to the absence of negative log-likelihood (NLL) constraints, potentially weakening the model's instruction-following capabilities. Based on these results, we adopt ORPO as the default optimization method for the second phase of our approach, as it provides the most significant improvement.

Method	Eff.	Gen.	Loc.a	Por.	Avg.
SFT	90.22	90.2	89.22	64.03	83.41
+DPO	88.47	88.3	89.18	61.11	83.26
+CPO	92.41	92.67	90.97	67.09	85.78
+KTO	92.23	92.01	89.22	67.23	85.17
+ORPO	**92.85**	**93.06**	**92.49**	**67.23**	**86.41**

Table 6: **Effects of different preference optimization methods** with single edit setting on en-zh. Shades of cell color represent differences between preference optimization methods and simply SFT, where blue denotes better performance while red indicates worse.

6 Related Work

Knowledge Editing The task of knowledge editing was introduced by (Sinitsin et al., 2020) to update specific knowledge while preserving unrelated information. Current methods fall into two paradigms: preserving or modifying the model's parameters. (1) Preserving LLMs' parameters involves auxiliary models or extra parameters. SERAC (Mitchell et al., 2022) uses a counterfactual model to update knowledge without altering model parameters. TPatcher (Huang et al., 2023) and CaliNET (Dong et al., 2022) add trainable parameters to edit knowledge. IKE (Zheng et al., 2023) and ICE (Cohen et al., 2024) leverage in-context learning to correct knowledge. (2) Modifying the model's parameters directly updates specific parameters to change knowledge. KE (De Cao et al., 2021) and MEND (Mitchell et al., 2021) predict weight updates for new data using a hyper-network. KN (Dai et al., 2021), ROME (Meng et al., 2022b), and MEMIT (Meng et al., 2022c) use knowledge attribution or causal mediation analysis to target

specific parameters for updating.

Cross-Lingual Knowledge Editing Cross-lingual knowledge editing extends monolingual editing by propagating edits across languages. (Wang et al., 2023a) introduced cross-lingual knowledge editing and created the Bi-ZsRE dataset to assess the applicability of monolingual methods in multilingual contexts. LiME (Xu et al., 2022) proposes language anisotropic editing to enhance cross-lingual editing, and MPN (Si et al., 2024) introduces multilingual patch neurons to update knowledge. However, these methods treat source language answers as ground truth for target language queries, falling short of achieving true cross-lingual transfer.

LLM Alignment LLM alignment (Gabriel, 2020) ensures that LLMs' behaviors align with human values. Techniques such as supervised fine-tuning (SFT) (Wei et al., 2021; Wang et al., 2023e; Mishra et al., 2021) train models to follow task descriptions in natural language. Despite SFT, models may still generate harmful content (Carlini et al., 2021; Gehman et al., 2020). To address this, reinforcement learning with human feedback (RLHF) (Stiennon et al., 2020; Ouyang et al., 2022) refines models further. Recent methods like SimPO (Meng et al., 2024) and ORPO (Hong et al., 2024) contribute to improving alignment in practical deployments.

7 Conclusion

In this paper, we present the Cross-Lingual Knowledge Democracy Edit (X-KDE) framework, which facilitates knowledge editing across languages in large language models (LLMs). By integrating Cross-lingual Edition Instruction Tuning (XE-IT) and Target-language Preference Optimization (TL-PO), X-KDE efficiently transfers knowledge from a source language to a target language while maintaining strong performance in monolingual settings. Additionally, we introduce high-quality datasets specifically designed for cross-lingual knowledge editing, filling gaps in existing resources. Our experimental results demonstrate that X-KDE outperforms current methods, offering a scalable solution for cross-lingual knowledge editing. Future research will explore applying X-KDE to other domains and optimizing its efficiency.

Limitations

While our work presents promising results, there are a few limitations to consider. First, due to computational constraints, we validate X-KDE on models with up to 7B parameters. Evaluating larger models, such as those exceeding 70B parameters, could provide more robust insights. Second, while our method has been effective in multilingual settings, its application to additional domains, such as finance or law, remains unexplored. Future research will focus on scaling X-KDE and extending its applicability to other fields.

Ethics and Reproducibility Statements

Ethics We take ethical considerations seriously and strictly adhere to the ACL Ethics Policy. All datasets used in this work are publicly available and widely adopted by the research community. Our methods focus on enhancing the multilingual capabilities of large language models without introducing harmful biases or unethical content. We ensure that all experiments are conducted in compliance with ethical guidelines, prioritizing fairness and transparency in model deployment.

Reproducibility In this paper, we discuss the detailed experimental setup, including training hyperparameters, baseline implementations, and statistical descriptions. More importantly, ***we have provided our code and data in the Supplementary Material*** to help reproduce the experimental results of this paper. Due to space limitations during uploading, the full dataset will be released upon acceptance.

Acknowledgements

We are grateful to the anonymous reviewers and the area chair for their insightful comments and suggestions. Dr Tao's research is partially supported by NTU RSR and Start Up Grants.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the*

AAAI conference on artificial intelligence, volume 34, pages 7432–7439.

Nicholas Carlini, Florian Tramer, et al. 2021. Extracting training data from large language models. In *USENIX*, pages 2633–2650.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. *arXiv preprint arXiv:2210.03329*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing can hurt general abilities of large language models. *arXiv preprint arXiv:2401.04700*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2024. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems*, 36.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. 2403.

Baixiang Huang, Canyu Chen, Xiongxiao Xu, Ali Payani, and Kai Shu. 2024a. Can knowledge editing really correct hallucinations? *arXiv preprint arXiv:2410.16251*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

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

Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, et al. 2024. Learning to edit: Aligning llms with knowledge editing. *arXiv preprint arXiv:2402.11905*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

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

Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, Jie Zhou, et al. 2024. Multilingual knowledge editing with language-agnostic factual neurons. *arXiv preprint arXiv:2406.16416*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022b. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022c. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.

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

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Nianwen Si, Hao Zhang, and Weiqiang Zhang. 2024. Mpn: Leveraging multilingual patch neuron for cross-lingual model editing. *arXiv preprint arXiv:2401.03190*.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitriy Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. *arXiv preprint arXiv:2004.00345*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

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

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2023a. Cross-lingual knowledge editing in large language models. *arXiv preprint arXiv:2309.08952*.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023b. Towards unifying multi-lingual and cross-lingual summarization. *arXiv preprint arXiv:2305.09220*.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, et al. 2023c. Easyedit: An easy-to-use knowledge editing framework for large language models. *arXiv preprint arXiv:2308.07269*.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2023d. Retrieval-augmented multilingual knowledge editing. *arXiv preprint arXiv:2312.13040*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, et al. 2023e. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36:74764–74786.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024. Mlake: Multilingual knowledge editing benchmark for large language models. *arXiv preprint arXiv:2404.04990*.

Jiakuan Xie, Pengfei Cao, Yuheng Chen, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Memla: Enhancing multilingual knowledge editing with neuron-masked low-rank adaptation. *arXiv preprint arXiv:2406.11566*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. 2022. Language anisotropic cross-lingual model editing. *arXiv preprint arXiv:2205.12677*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. 2024b. The butterfly effect of model editing: Few edits can trigger large language models collapse. *arXiv preprint arXiv:2402.09656*.

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

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

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*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023a. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

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

A Details of Dataset Construction

A.1 Prompt Details of Sample Generation

Here, we present the detailed prompts for sample generation. Specifically, we guide LLMs in producing the queries and answers via the following prompts:

Query Generation Prompt

[Edit description] will modify the knowledge. You must think according to the narrative of [Edit description].

[Edit description]: Who is Chris Klemmer affiliated with? University of Washington
[Prompt]: Please generate a question related to Chris Klemmer. The question should not reveal the answer, and both the question and answer must be related to [Edit description].
[Generated Question]: Which university is Chris Klemmer associated with?

[Edit description]: What profession does Dagmar Lurz pursue? Film director
[Prompt]: Please generate a question related to Dagmar Lurz. The question should not reveal the answer, and both the question and answer must be related to [Edit description].
[Generated Question]: In which creative field has Dagmar Lurz displayed her extraordinary talent?

[Edit description]: When was Jana Fesslová born? April 20, 1977
[Prompt]: Please generate a question related to Jana Fesslová. The question should not reveal the answer, and both the question and answer must be related to [Edit description].
[Generated Question]: What major historical or cultural event in the 1970s aligns with Jana Fesslová's birth date?

[Edit description]: What type of voice does Martin Crosby have? Contralto
[Prompt]: Please generate a question related to Martin Crosby. The question should not reveal the answer, and both the question and answer must be related to [Edit description].
[Generated Question]: What is Martin Crosby's vocal type, and how does this type bring unique timbre and expressiveness to musical performances?

[Edit description]: Who directed "The Mentor"? Jean-Luc Godard
[Prompt]: Please generate a question related to "The Mentor". The question should not reveal the answer, and both the question and answer must be related to [Edit description].
[Generated Question]: Who directed "The Mentor," and how did they use their unique directorial style to shape the film's artistic quality and influence?

[Edit description]: `<instruct>`
[Prompt]: Please generate a question related to `<subject>`. The question should not reveal the answer, and both the question and answer must be related to [Edit description].
[Generated Question]: `<answer>`

Answer Generation Prompt

Imagine that the description of [Edit description] is correct. Please answer [question] based on [Edit description].

[Edit description]: Who is Chris Klemmer affiliated with? University of Washington
[Question]: Who is Chris Klemmer affiliated with?
[Prompt]: The answer should focus on Chris Klemmer, with University of Washington as a secondary focus, avoiding unnecessary details.
[Answer]: Chris Klemmer is affiliated with the University of Washington. During his academic career at the University of Washington, he made significant achievements and contributions:
1.Academic Research: Klemmer conducted extensive research at the University of Washington, particularly in his field of expertise. His research findings not only advanced academic knowledge but also had a considerable impact on practical applications.
2.Teaching Contributions: As a professor at the University of Washington, Klemmer excelled in teaching and mentored many outstanding students. His teaching methods and course designs were well-received and inspired students to innovate.
3.Academic Leadership: Klemmer held various academic leadership positions at the University of Washington, such as department chair or research center director. In these roles, he promoted academic projects, fostered interdisciplinary collaboration, and enhanced the institution's academic reputation.
4.Publications: Klemmer published numerous high-level academic papers and books, which received widespread attention and citation in the academic community, further solidifying his standing in academia.

[Edit description]: What profession does Dagmar Lurz pursue? Film director
[Question]: In which field has Dagmar Lurz shown exceptional talent?
[Prompt]: The answer should focus on Dagmar Lurz, with film directing as secondary, avoiding unrelated content.
[Answer]: Dagmar Lurz has demonstrated her exceptional talent in the field of film directing. As a film director, she is responsible for guiding actors' performances and overseeing the film's overall artistic style and narrative pacing. Her work involves interpreting scripts, arranging scenes, guiding cinematography, and making editing decisions, all requiring rich creativity and sharp artistic insight. Through her directorial work, Dagmar Lurz has left a unique mark on film art, offering viewers deep visual and emotional experiences.

[Edit description]: `<instruct>`
[Question]: `<question>`
[Prompt]: Please generate a question related to `<subject>`. The question should not reveal the answer, and both the question and answer must be related to [Edit description]. The answer should focus on `<subject>`, with `<new answer>` as secondary, avoiding unrelated content.
[Generated Answer]: `<new answer>`

Out-of-scope Generation Prompt

In the following statements, "changed answer" represents the modified factual knowledge. When the answer is changed, other properties of the subject should remain unchanged. For example, if we edit basketball player Grant Hill to a football player, this would not affect his nationality. Therefore, for irrelevant attributes such as country, the output should remain consistent with the pre-edit version. You should recall an irrelevant attribute and generate a question and answer based on that irrelevant attribute and the "subject".

Question: Who is the father of Juan María Bordaberry?
Subject: Juan María Bordaberry
Changed answer: Gabrielle Bordaberry
Irrelevant attribute recalled: place of death
New question: Where did Juan María Bordaberry die?
New answer: Montevideo

Question: Who published the game Street Rod 2?
Subject: Street Rod 2
Changed answer: Sierra Entertainment
Irrelevant attribute recalled: release format
New question: What is the release format of Street Rod 2?
New answer: Floppy disk

Question: What is the status of the Cross River Gorilla?
Subject: Cross River Gorilla
Changed answer: Endangered
Irrelevant attribute recalled: classification level
New question: What is the classification level of the Cross River Gorilla?
New answer: Subspecies"""

[Question]: `<question>`
[Subject]: `<subject>`
[Changed Answer]: `<new answer>`

A.2 Prompt Details of Quality Control

Moreover, we employ LLMs in Quality Controls. Judging and scoring via the following prompts:

Prompt for judging

Please act as a fair judge and determine whether the [answer] answers the [question] based on the [Edit description]. Provide an explanation and strictly follow the format:
- If the answer is based on the edit description, output "[T]"
- If it is not, output "[F]".

[Edit description]: `<instruct>`
[Question]: `<question>`
[Answer]: `<answer>`

Prompt for scoring

Please act as a fair judge and rate the sentence based on the following criteria:
1. Sentence complexity: Evaluate the complexity of the sentence, such as inversion, imperative sentences, sentences with word inflections, or sentences starting with multiple adverbs, nouns, and subjects. The more complex, the higher the score.
2. Vocabulary richness: Evaluate the diversity of vocabulary used. The more diverse, the higher the score.
3. Faithfulness: Evaluate whether the [answer] faithfully adheres to [Edit description], meaning it accurately answers [question]. If the answer highly matches the description, the score is higher; if the question leaks the answer, deduct points.
Provide separate scores for each criterion (1-10), and calculate the average score, then output the sentence with the highest score. The output format should be:
[Sentence complexity: score; Vocabulary richness: score; Faithfulness: score; Overall score: score]

[Edit description]: `<instruct>`
[Question]: `<question>`
[Answer]: `<answer>`

A.3 Training Data Statistics

Table 8 lists the statistics of our constructed high-quality dataset, which includes 388k samples and covers two languages: English (En) and Chinese (Zh). To prevent data leakage, we only sample instances from the training sets.

B Experimental Setup

B.1 Alignment Score Calculation

We employ NLLB-600M-distilled[2] as the multilingual translation model. And since the translation model is trained on a large-scale parallel dataset in both source and target languages (X and Y) by maximizing the conditional generation probability $P(Y|X;\theta)$. Higher alignment between Y and X results in a higher conditional generation probability $P(Y|X)$. To compute the "alignment" score, we input the model's response and force-decode the ground-truth Y in target language.

B.2 Implementation Details

All experiments were executed on 8 NVIDIA A100 GPUs (80G). We employ EasyEdit (Wang et al., 2023c) to implement all the baselines with the default settings. We employ llamafactory (Zheng et al., 2024) to implement Cross-lingual Edition Instruction Tuning (XE-IT) phase of our method. When training on the English editing only subset, the duration is approximately 10 to 15 hours. Hyper-parameters of our X-KDE are in Table 7.

Hyperparameter	XE-IT	TL-PO
Learning rate	5e-6	1e-6
Max length	2560	1024
Optimizer	AdamW	AdamW
Scheduler	cosine	cosine
Weight decay	0.1	0.05
warmup steps	100	100

Table 7: **Hyper-parameters** for training our X-KDE.

[2] https://huggingface.co/facebook/nllb-200-distilled-600M

Data Source	Lang.	# in-scope		# out-scope		# Total	Avg Token
		w/ edit	w/o edit	w/ edit	w/o edit		
ZsRE	En	20,000	20,000	20,000	20,000	80,000	48
	Zh	20,000	20,000	20,000	20,000	80,000	84
HalluEditBench	En	2,000	2,000	2,000	2,000	8,000	38
	Zh	2,000	2,000	2,000	2,000	8,000	60
RIPPLEEDITS	En	2,250	2,250	2,250	2,250	9,000	53
	Zh	2,250	2,250	2,250	2,250	9,000	88
WikiBio	En	250	250	250	250	1,000	162
	Zh	250	250	250	250	1,000	294
MQUAKE	En	4,000	4,000	4,000	4,000	16,000	266
	Zh	4,000	4,000	4,000	4,000	16,000	334
COUNTERFACT	En	7,500	7,500	7,500	7,500	30,000	530
	Zh	7,500	7,500	7,500	7,500	30,000	888
Total	En	36,000	36,000	36,000	36,000	144,000	170
	Zh	36,000	36,000	36,000	36,000	144,000	266

Table 8: **Statistics of our training data (Lang.:language)**. "Avg Token" denotes the average length(token-level) of samples, and "edit" indicates the edit descriptor.

C Used Scientific Artifacts

We list scientific artifacts used in our work blow. And we use of these existing artifacts is consistent with their intended use.

- *DeepSpeed (Apache-2.0 license)* [3], a deep learning optimization library to improve the efficiency of training large language models.
- *Transformers (Apache-2.0 license)* [4], a framework that provides state-of-the-art pretrained models for NLP tasks
- *trl (Apache-2.0 license)* [5], a library designed to integrate reinforcement learning (RL) with transformer models.
- *vLLM (Apache-2.0 license)* [6], an optimized framework for inference with large language models.

D Supplemental Experiment Results

D.1 Detailed Results of Llama2-7b-chat

More detailed results on MzsRE of Llama2-7b-chat are listed in Table 9 and Table 10.

D.2 Detailed Results of Qwen2.5-7B-Instruct

To verify the universality of our method, we conducted experiments on Qwen2.5-7B-Instruct. More detailed results are listed in Table 11, Table 12 and Table 13.

[3] https://github.com/deepspeedai/DeepSpeed
[4] https://github.com/huggingface/transformers
[5] https://github.com/huggingface/trl
[6] https://github.com/vllm-project/vllm

Metrics	Methods	en-en	en-cz	en-de	en-du	en-es	en-fr	en-pt	en-ru	en-th	en-tr	en-vi	en-zh	en-avg
Reliability	FT-L	52.92	41.81	39.79	39.02	39.49	39.72	39.26	39.79	36.44	36.86	46.21	51.81	41.93
	FT-M	99.96	66.93	70.16	67.17	63.69	64.98	64.22	48.96	36.46	57.54	66.80	56.89	63.65
	ROME	96.36	56.54	60.82	58.89	57.41	56.43	54.91	41.69	35.44	45.76	56.94	49.94	55.93
	MEMIT	95.44	62.37	64.82	64.12	59.46	61.90	58.69	44.54	36.40	49.15	61.34	52.05	59.19
	IKE	99.65	83.22	80.61	79.36	76.69	78.48	75.37	67.62	54.38	76.90	81.22	67.83	76.78
	LTE	100.00	84.29	81.71	80.60	77.67	79.11	77.39	72.02	62.04	78.87	81.92	76.93	79.38
	X-KDE	99.93	92.78	87.43	88.89	85.71	87.49	89.87	89.32	89.66	91.23	87.55	93.07	90.24
Generality	FT-L	49.60	40.75	38.87	38.36	39.68	39.12	39.56	38.97	36.89	37.18	45.89	51.71	41.38
	FT-M	95.53	65.45	68.15	65.09	62.39	62.28	61.63	47.69	36.88	56.87	65.97	56.52	62.04
	ROME	85.13	54.99	58.91	56.99	56.58	54.47	53.94	40.68	35.36	45.06	56.38	50.31	54.07
	MEMIT	89.59	60.71	63.80	61.98	58.10	59.40	57.63	43.31	36.77	48.68	60.51	52.01	57.71
	IKE	99.54	82.67	80.78	79.18	76.37	78.22	75.49	67.51	54.26	76.97	80.99	67.88	76.65
	LTE	99.87	84.26	81.63	81.07	77.51	78.99	77.38	71.46	61.90	78.26	81.37	76.24	79.16
	X-KDE	99.68	92.87	87.25	88.87	85.16	87.57	89.93	89.10	89.21	91.25	87.62	93.11	90.14
Locality	FT-L	93.96	90.78	81.06	88.98	83.32	89.30	90.98	89.53	90.18	88.95	93.02	85.56	88.80
	FT-M	97.71	96.94	96.24	96.57	96.36	97.56	97.49	97.31	96.93	97.25	98.04	94.61	96.92
	ROME	97.81	96.12	97.57	96.80	97.36	96.98	97.14	96.70	96.28	96.83	97.60	97.70	97.07
	MEMIT	98.55	98.24	98.55	98.08	98.35	98.30	98.45	98.33	98.88	98.97	98.89	98.76	98.53
	IKE	58.13	61.35	65.57	61.52	63.93	60.42	59.42	58.90	68.84	63.97	68.40	64.54	62.91
	LTE	89.28	77.01	77.90	77.68	81.54	81.51	81.23	78.39	79.86	76.34	82.93	86.63	80.86
	X-KDE	93.12	78.76	79.88	77.19	81.29	78.97	80.00	82.78	82.08	72.62	82.11	91.91	81.73
Portability	FT-L	52.85	46.85	43.51	43.21	44.47	44.91	43.72	47.05	39.92	41.14	54.05	55.13	46.40
	FT-M	57.17	48.66	46.38	46.34	47.09	47.54	46.36	48.41	38.55	42.50	55.53	52.16	48.06
	ROME	58.46	49.79	48.58	47.06	48.29	48.83	47.30	49.21	38.11	42.38	56.62	51.81	48.87
	MEMIT	57.02	50.41	47.96	47.26	47.26	48.47	47.21	49.25	38.56	43.44	57.16	52.19	48.85
	IKE	70.97	56.44	58.87	56.91	58.05	58.41	56.33	57.96	39.46	48.69	62.86	58.97	56.99
	LTE	77.29	61.85	64.91	63.82	61.53	62.83	62.39	61.43	44.51	51.04	65.37	67.47	62.04
	X-KDE	76.13	60.53	58.74	56.94	55.19	58.85	58.81	62.89	56.18	48.69	61.78	74.04	60.73

Table 9: **Results on MzsRE dataset for editing performed in English** using Llama2-7b-chat. Here, "en-zh" means that English serves as the source language and Chinese as the target language, with similar interpretations for the other pairs. "en-avg" denotes the average performance across cross-lingual scenarios.

Metrics	Methods	zh-en	zh-cz	zh-de	zh-du	zh-es	zh-fr	zh-pt	zh-ru	zh-th	zh-tr	zh-vi	zh-zh	zh-avg
Reliability	FT-L	40.81	38.16	36.21	35.60	36.28	36.45	35.55	38.88	33.98	34.50	43.32	54.79	38.71
	FT-M	51.87	48.51	46.71	45.70	45.69	45.98	45.65	46.97	39.44	44.74	54.06	100.00	51.28
	ROME	44.15	40.06	38.04	37.62	38.44	37.99	37.69	39.25	32.94	36.49	44.82	73.48	41.75
	MEMIT	51.87	41.45	39.61	39.29	39.19	39.23	38.78	40.85	33.77	38.49	46.72	76.12	43.78
	IKE	65.88	68.68	67.63	66.75	65.06	65.63	63.82	63.52	52.39	61.32	70.90	99.85	67.62
	LTE	65.44	64.74	62.05	62.91	61.09	60.85	61.20	63.09	55.71	58.15	67.02	99.76	65.17
	X-KDE	94.64	84.40	83.05	81.08	80.33	81.22	83.38	82.56	83.09	78.69	81.47	99.99	84.49
Generality	FT-L	40.67	37.70	36.35	35.18	36.60	35.49	35.67	38.19	33.86	34.97	43.14	53.90	38.48
	FT-M	51.24	48.24	46.49	45.30	45.71	45.63	45.81	46.32	39.62	45.06	54.42	99.68	51.13
	ROME	43.80	39.72	38.01	37.83	38.26	36.74	38.62	38.46	32.76	36.46	45.02	71.13	41.40
	MEMIT	51.24	41.16	40.14	38.22	39.10	38.80	39.06	40.15	34.18	38.22	46.34	74.21	43.40
	IKE	65.75	67.90	67.48	66.39	65.01	65.35	63.50	63.44	52.72	61.03	70.27	99.28	67.34
	LTE	64.94	64.53	62.72	62.31	61.15	60.50	61.11	62.94	55.39	58.29	66.88	98.69	64.95
	X-KDE	94.51	84.27	82.43	81.46	80.12	81.08	82.69	82.19	82.81	78.33	81.07	98.89	84.15
Locality	FT-L	94.81	89.42	83.41	88.81	83.09	89.92	90.09	86.63	79.94	85.75	89.69	66.38	85.66
	FT-M	98.19	96.70	92.82	95.79	93.90	97.68	97.17	96.46	93.05	95.49	97.18	79.74	94.51
	ROME	97.93	96.17	97.41	96.23	96.81	96.66	96.99	95.89	94.36	96.15	97.23	96.42	96.52
	MEMIT	98.19	98.05	98.52	98.68	98.76	98.52	98.50	97.51	96.55	98.10	98.36	96.06	97.98
	IKE	69.41	63.74	63.22	64.02	62.24	61.69	62.10	61.15	67.15	64.73	69.31	67.91	64.72
	LTE	89.26	76.59	80.09	78.01	81.86	81.44	81.22	80.31	80.24	76.63	82.92	86.67	81.27
	X-KDE	94.07	80.53	81.79	78.13	83.49	81.18	81.72	84.22	84.14	75.66	83.15	92.20	83.36
Portability	FT-L	55.25	47.54	45.76	44.73	46.21	46.61	44.96	48.20	37.99	41.16	54.13	48.07	46.72
	FT-M	55.30	48.13	45.78	45.25	46.09	47.04	45.22	48.75	40.32	42.12	55.57	61.79	48.45
	ROME	52.66	47.19	44.60	44.89	44.56	45.64	44.11	47.96	36.40	40.23	53.87	48.34	45.87
	MEMIT	55.30	47.78	45.99	45.60	46.00	46.91	45.32	48.10	37.82	41.51	54.88	50.83	47.17
	IKE	63.06	54.77	58.40	56.25	55.59	56.72	54.58	57.65	40.80	47.15	61.58	66.44	56.08
	LTE	68.30	59.86	60.92	59.62	58.27	59.71	58.59	60.03	45.83	48.53	63.11	69.40	59.35
	X-KDE	67.95	60.34	59.10	58.10	55.89	58.43	59.37	61.73	56.27	49.69	62.78	73.43	60.26

Table 10: **Results on MzsRE dataset for editing performed in Chinese** using Llama2-7b-chat. Here, "zh-en" means that Chinese serves as the source language and English as the target, with similar interpretations for the other pairs. "zh-avg" denotes the average performance across cross-lingual scenarios.

Method	Test in English				Test in Chinese				
	Reliability	Generality	Locality	Portability	Reliability	Generality	Locality	Portability	*Avg.*
Edit in English									
FT-L	62.89	65.93	75.06	38.83	44.25	44.36	68.35	40.21	54.98
FT-M	100.0	99.35	92.44	52.59	64.03	63.61	88.48	53.22	76.72
ROME	99.48	92.83	98.63	56.69	58.91	58.44	98.47	54.93	77.30
MEMIT	96.77	89.30	98.78	55.60	59.97	59.35	98.49	54.31	76.57
IKE	97.32	98.56	50.26	67.30	69.89	69.45	57.07	57.94	70.97
LTE	99.78	99.28	87.64	74.23	73.95	74.40	84.34	61.85	81.93
X-KDE(Ours)	99.72	99.56	88.79	73.96	90.42	90.20	91.53	62.59	87.10
Edit in Chinese									
FT-L	32.59	33.17	73.39	37.89	48.85	53.49	54.88	28.17	45.30
FT-M	53.31	52.80	92.53	51.52	100.0	99.85	79.38	52.84	72.78
ROME	45.66	45.31	98.31	52.74	99.36	94.77	97.97	57.06	73.90
MEMIT	45.68	44.25	98.94	52.26	98.07	94.20	96.69	57.70	73.47
IKE	79.59	78.77	49.57	65.20	96.37	96.47	66.05	61.59	74.20
LTE	79.40	78.50	86.64	70.24	98.95	98.60	84.54	64.53	82.68
X-KDE(Ours)	94.78	94.77	95.14	67.50	99.79	98.29	90.57	61.30	87.77

Table 11: **Cross-lingual editing performance of different methods** on Qwen2.5-7B-Instruct backbones. Results in green indicates the best results. "***Avg.***" represents the overall mean of all metrics evaluated across the two languages.

Metrics	Methods	en-en	en-cz	en-de	en-du	en-es	en-fr	en-pt	en-ru	en-th	en-tr	en-vi	en-zh	en-avg
Reliability	FT-L	63.77	50.88	50.30	47.23	47.40	51.08	48.18	42.59	44.37	47.86	48.39	44.91	48.91
	FT-M	100.0	71.56	74.67	70.47	66.45	68.94	68.82	57.35	51.57	67.36	64.79	64.59	68.88
	ROME	99.44	55.82	63.27	61.38	57.41	59.44	60.23	49.78	48.74	53.73	53.06	59.23	60.13
	MEMIT	96.92	54.87	61.36	58.79	54.67	58.15	58.02	49.53	47.71	52.49	51.88	60.27	58.72
	IKE	97.89	82.71	82.84	78.05	76.34	79.26	78.69	69.95	66.43	77.28	75.93	70.53	77.99
	LTE	99.70	84.28	84.73	80.76	78.02	82.07	80.40	76.52	69.53	81.06	77.89	73.35	80.69
	X-KDE	98.56	89.60	86.69	86.94	84.84	84.87	85.30	89.05	87.61	90.00	79.47	89.74	87.72
Generality	FT-L	66.81	50.62	50.42	46.74	47.63	51.18	48.35	42.85	44.81	47.66	47.50	44.91	49.12
	FT-M	99.26	70.46	73.67	69.37	65.68	67.06	66.14	56.27	51.83	65.35	62.55	64.09	67.64
	ROME	93.67	54.71	61.05	58.82	55.48	57.48	58.01	48.59	48.32	51.49	51.57	59.13	58.19
	MEMIT	90.32	54.69	59.12	56.52	53.93	55.96	55.53	48.25	47.77	51.16	50.64	59.72	56.97
	IKE	98.52	82.75	82.71	77.83	75.86	78.92	78.30	69.57	66.83	77.26	75.26	70.68	77.87
	LTE	99.30	84.48	84.57	80.39	77.93	81.49	80.34	76.56	69.45	81.07	77.56	73.63	80.56
	X-KDE	98.00	89.91	86.55	87.07	84.88	84.53	85.49	89.03	87.49	89.82	79.42	89.95	87.68
Locality	FT-L	74.68	75.09	62.23	69.32	62.46	70.98	73.56	74.63	79.92	66.58	75.33	67.88	71.06
	FT-M	92.46	90.64	83.27	87.98	83.72	90.23	90.39	90.89	93.08	86.20	91.12	88.30	89.02
	ROME	98.71	97.43	97.32	98.05	98.35	97.37	98.41	98.49	98.46	97.53	98.53	98.35	98.08
	MEMIT	98.76	98.49	98.21	98.49	98.77	98.63	98.48	98.74	98.75	98.59	98.47	98.51	98.57
	IKE	50.52	55.71	57.39	53.66	57.51	58.40	56.95	61.17	65.64	59.46	60.03	57.45	57.82
	LTE	88.28	80.06	81.15	78.18	82.89	84.62	82.80	86.32	85.12	78.61	80.00	84.74	82.73
	X-KDE	95.22	77.82	87.16	77.71	79.86	83.98	82.35	87.42	86.40	75.54	80.53	92.59	83.88
Portability	FT-L	38.08	38.77	36.92	36.49	37.42	39.98	40.22	45.27	47.89	39.07	35.92	40.23	39.69
	FT-M	52.17	49.42	48.36	46.27	46.92	50.18	49.86	53.84	54.22	46.93	45.66	52.83	49.72
	ROME	56.40	49.68	49.33	48.00	49.17	53.04	51.94	54.93	54.53	47.56	46.14	54.95	51.31
	MEMIT	54.82	49.69	49.62	47.80	49.48	51.74	51.67	55.01	54.57	48.15	46.94	54.32	51.15
	IKE	67.00	55.75	59.25	55.11	56.31	60.00	58.76	59.04	55.32	52.44	52.16	58.27	57.45
	LTE	74.25	61.72	65.09	61.27	62.63	65.92	64.21	65.30	60.78	59.96	57.94	61.36	63.37
	X-KDE	70.36	53.71	54.93	52.23	54.76	56.20	56.57	60.27	58.51	52.41	49.23	61.18	56.70

Table 12: **Results on MzsRE dataset for editing performed in English** using Qwen2.5-7B-Instruct. Here, "en-zh" means that English serves as the source language and Chinese as the target language, with similar interpretations for the other pairs. "en-avg" denotes the average performance across cross-lingual scenarios.

Metrics	Methods	zh-en	zh-cz	zh-de	zh-du	zh-es	zh-fr	zh-pt	zh-ru	zh-th	zh-tr	zh-vi	zh-zh	zh-avg
Reliability	FT-L	33.24	34.37	32.47	31.74	30.58	33.23	32.99	35.73	40.39	29.41	28.90	48.62	34.31
	FT-M	52.97	51.26	51.88	49.22	48.13	49.82	51.21	52.25	51.33	50.00	47.02	100.00	54.59
	ROME	46.29	42.41	43.29	42.32	41.15	42.64	43.13	45.94	47.14	44.22	40.61	99.51	48.22
	MEMIT	46.36	42.81	44.02	42.29	40.87	42.69	43.75	46.06	46.78	43.18	41.53	98.59	48.25
	IKE	80.22	71.25	76.32	69.69	69.28	71.10	70.64	66.51	65.03	68.96	66.66	96.32	72.67
	LTE	79.54	71.83	75.40	71.18	67.28	70.02	68.88	72.39	66.11	70.15	68.79	99.03	73.38
	X-KDE	94.97	79.86	84.24	79.12	77.97	81.24	79.16	75.77	67.75	77.76	76.48	99.85	81.18
Generality	FT-L	32.90	34.32	32.80	31.56	31.00	33.88	33.40	35.69	40.79	29.89	29.57	53.08	34.91
	FT-M	52.66	51.13	52.15	49.12	47.66	48.95	50.46	51.47	51.51	50.22	46.62	99.96	54.33
	ROME	46.11	42.31	42.60	41.57	40.65	41.88	42.77	44.60	46.34	42.97	40.80	95.23	47.32
	MEMIT	45.03	42.58	42.79	41.50	40.83	41.81	42.51	45.20	46.48	42.65	41.33	95.06	47.31
	IKE	78.68	72.05	76.01	69.73	69.72	71.07	71.09	66.08	65.18	69.40	66.86	96.50	72.70
	LTE	78.20	71.15	75.18	70.72	66.41	69.68	69.22	72.32	66.04	69.93	68.50	98.34	72.97
	X-KDE	94.93	79.44	84.67	79.08	77.85	81.37	79.20	75.49	68.11	77.53	76.29	98.51	81.04
Locality	FT-L	73.25	67.27	58.50	63.17	58.05	65.53	68.75	67.52	69.27	56.25	64.19	54.75	63.87
	FT-M	92.53	89.49	83.74	87.61	85.73	89.50	90.45	88.96	90.36	83.69	90.04	79.06	87.60
	ROME	98.26	97.35	97.07	96.83	97.80	97.37	97.90	97.94	97.92	96.74	97.54	97.90	97.55
	MEMIT	98.94	98.36	98.42	97.94	98.49	98.69	98.81	98.36	97.69	97.96	97.76	96.67	98.18
	IKE	50.43	57.50	57.28	53.59	55.05	56.53	56.76	61.92	68.25	61.89	61.34	65.49	58.84
	LTE	87.33	79.75	81.18	76.95	83.05	83.28	80.42	85.36	84.31	78.27	80.59	84.56	82.09
	X-KDE	93.87	81.90	84.70	81.11	86.56	87.89	87.14	88.08	90.18	79.84	84.63	90.42	86.36
Portability	FT-L	37.03	36.69	34.82	34.70	34.97	37.81	37.40	41.47	44.90	33.05	30.24	27.71	35.90
	FT-M	50.78	47.32	45.81	44.66	45.38	47.77	47.23	51.59	53.59	44.78	42.89	51.91	47.81
	ROME	51.70	47.48	47.36	45.42	46.54	50.09	49.16	54.37	54.39	46.86	44.10	56.78	49.52
	MEMIT	51.54	47.74	47.22	45.78	46.19	49.21	48.87	54.48	54.24	46.25	44.32	57.65	49.46
	IKE	64.32	56.14	58.98	55.44	56.10	59.75	58.63	61.12	57.28	53.05	52.57	61.75	57.93
	LTE	70.11	61.02	64.45	60.44	62.24	65.28	63.82	65.28	60.67	59.51	57.18	64.36	62.86
	X-KDE	67.44	56.98	61.07	58.20	60.59	61.93	61.62	63.60	58.95	55.87	55.28	60.92	60.20

Table 13: **Results on MzsRE dataset for editing performed in Chinese** using Qwen2.5-7B-Instruct. Here, "zh-en" means that Chinese serves as the source language and English as the target, with similar interpretations for the other pairs. "zh-avg" denotes the average performance across cross-lingual scenarios.