MLaKE: Multilingual Knowledge Editing Benchmark for Large Language Models

Zihao Wei[1,2 *] **Jingcheng Deng**[1,2 *] **Liang Pang**[1†]
Hanxing Ding[1,2] **Huawei Shen**[1,2] **Xueqi Cheng**[1,2]
[1] Institute of Computing Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
{weizihao22z, dengjingcheng23s, pangliang}@ict.ac.cn
{dinghanxing18s, shenhuawei, cxq}@ict.ac.cn

Abstract

The extensive utilization of large language models (LLMs) underscores the crucial necessity for precise and contemporary knowledge embedded within their intrinsic parameters. Existing research on knowledge editing primarily concentrates on monolingual scenarios, neglecting the complexities presented by multilingual contexts and multi-hop reasoning. To address these challenges, our study introduces MLaKE (Multilingual Knowledge Editing), a novel benchmark comprising 4072 multi-hop and 5360 single-hop questions designed to evaluate the adaptability of knowledge editing methods across five languages: English, Chinese, Japanese, French, and German. MLaKE aggregates fact chains from Wikipedia across languages and utilizes LLMs to generate questions and answer. We assessed the effectiveness of current multilingual knowledge editing methods using the MLaKE dataset. Our results show that due to considerable inconsistencies in both multilingual performance and encoding efficiency, these methods struggle to generalize effectively across languages. The accuracy of these methods when editing English is notably higher than for other languages. The experimental results further demonstrate that models encode knowledge and generation capabilities for different languages using distinct parameters, leading to poor cross-lingual transfer performance in current methods. Transfer performance is notably better within the same language family compared to across different families. These findings emphasize the urgent need to improve multilingual knowledge editing methods.[1]

1 Introduction

With the widespread deployment of large language models, ensuring that the knowledge stored in their intrinsic parameters is correct and up-to-date has become a very important topic (Sinitsin et al., 2020; Chen and Shu, 2023, 2024). knowledge editing serves as a promising solution to this challenge, necessitating timely updates to the knowledge embedded within LLMs (Zhu et al., 2020; De Cao et al., 2021; Ding et al., 2024; Meng et al., 2022; Mitchell et al., 2022a; Xu et al., 2024; Chen et al., 2024).

Despite considerable efforts devoted to this research field, current studies on knowledge editing typically concentrate on monolingual scenarioi (Li et al., 2023; Zhang et al., 2024; Huang et al., 2023a), where language models are edited and evaluated within the same language (Meng et al., 2022, 2023; Mitchell et al., 2022a). However, the rapid advancements in large language models (LLMs) have facilitated the widespread adoption of multilingual settings (Zhao et al., 2023; Wang et al., 2023a). Given this context, the performance of a source-language edited model on other languages remains largely unexplored.

To address this challenge, our study introduces MLaKE (Multilingual Language Knowledge Editing), a novel benchmark for multilingual multi-hop knowledge editing. MLaKE comprises 5360 single-hop questions and 4072 multi-hop questions designed to test the adaptability of knowledge editing methods across various languages. To ensure the quality and currency of knowledge, we begin by collecting fact chains across languages from Wikipedia. Subsequently, we leverage powerful LLM (e.g., ChatGPT) to generate questions in both free-form and multiple-choice formats using fact chains as input. Consequently, The MLaKE dataset comprises single-hop and multi-hop questions in English, Chinese, Japanese, French, and German. This diverse dataset serves as a robust foundation for evaluating the effectiveness of knowledge editing techniques in diverse linguistic environments.

We assessed the effectiveness of various knowl-

[*] Equal Contributions
[†] Corresponding Author
[1] Our benchmark and source code are available at https://github.com/Hi-archers/MLaKE.

Proceedings of the 31st International Conference on Computational Linguistics, pages 4457–4473
January 19–24, 2025. ©2025 Association for Computational Linguistics

edge editing methods on MLaKE, with a focus on their performance in multilingual contexts. The results show that current mainstream knowledge editing methods demonstrate weak generalization in multilingual editing and poor cross-lingual transfer performance. Specifically, these methods not only struggle to edit knowledge in one language while transferring across others, but also exhibits weak performance in the foundational task of editing knowledge within a single language. Moreover, the aforementioned challenges become even more pronounced in multi-hop reasoning scenarios. Our findings underscore the significant impact of language differences on the performance of knowledge editing. To better understand this challenge, we conduct a series of experiments to analyze the effects of linguistic and structural differences on knowledge editing performance.

The main contributions of our work are as:

- We collect the multilingual knowledge editing dataset, MLaKE, which comprises 5360 single-hop questions and 4072 multi-hop questions designed to test the adaptability of existing methods across various languages.
- We demonstrate that existing knowledge editing methods, when applied to LLMs, suffer from significant shortcomings in multilingual generalization and cross-lingual transfer performance.
- Our analysis shows that weak multilingual generalization is primarily due to the models' insufficient multilingual performance and encoding inefficiency. The limited cross-lingual transferability is largely caused by LLMs using distinct parameter sets to encode knowledge for different languages.

2 MLaKE: Multi-Lingual Knowledge Editing Benchmark

In our study, we construct the MLaKE (**M**ulti**L**ingu**a**l **K**nowledge **E**diting) benchmark, an comprehensive and challenging dataset that encompasses five languages (English, Chinese, Japanese, French, German) and intricate logical structures (single-hop and multi-hop). In this section, we first present the data construction process of MLaKE, followed by a detailed description of the dataset. Lastly, we elaborate on the evaluation settings and metrics utilized.

2.1 Data Construction of MLaKE

Figure 1 illustrates the construction process of MLaKE, which is primarily composed of three sequential steps: selection and alignment of fact chains, generation of raw data, and construction of question and answer.

2.1.1 Select chains of facts

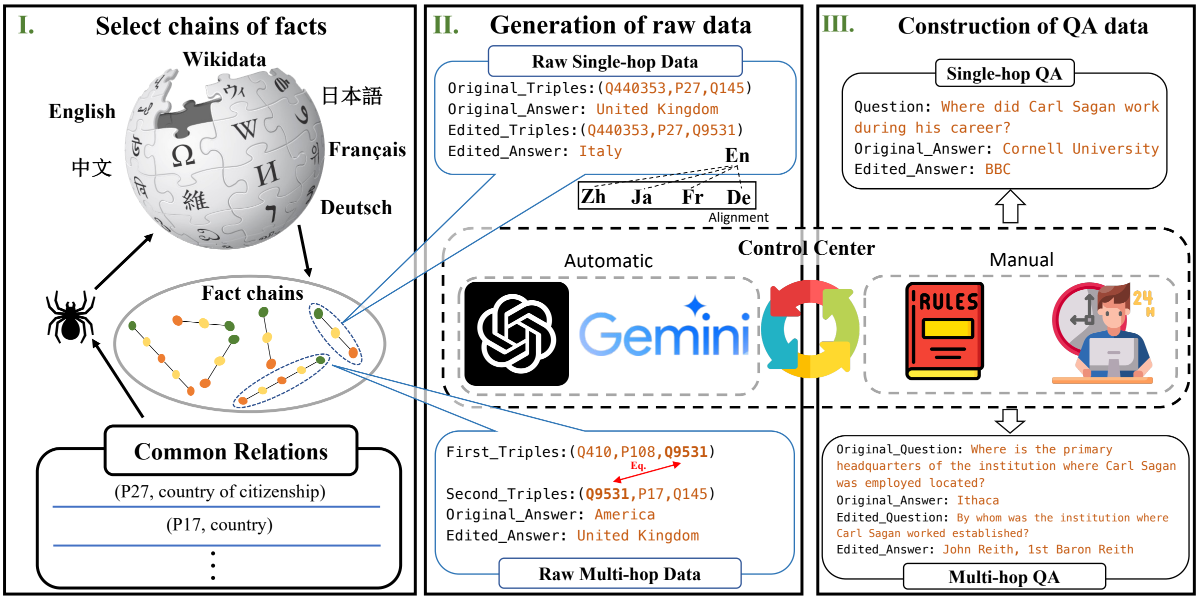
We define fact chains as tuples, where a single-hop fact chain is represented as (s_1, r_1, o_1). In this representation, s_1, r_1 and o_1 denote the subject, relationship, and object of the single-hop, respectively. Similarly, a multi-hop fact chain is expressed as $(s_1, r_1, o_1, r_2, o_2)$, where r_2 represents the multi-hop relationship, and o_2 signifies the multi-hop object. Notably, the single-hop object is equivalent to the multi-hop subject.

Inspired by the work of Zhong et al. (2023), We gather fact chains by crawling data from Wikipedia[2]. Initially, we manually create a relational dataset consisting of 43 common relations, the same as the approach taken in previous work (Petroni et al., 2019; Meng et al., 2022). Subsequently, we collect single-hop fact chains from Wikidata across five languages, leveraging the relational dataset. During this process, we employ rules to ensure that the single-hop fact chain satisfies specific predefined conditions. For instance, to facilitate batch editing, we enforce restrictions that prevent repetitive modifications of the relationship associated with an entity in the single-hop fact chains. For additional filtering rules applied to fact chains, please refer to the Appendix C.1. To assess the generalizability of the knowledge editing method across languages, we perform alignment on the collected single-hop fact chains. Specifically, we retain only single-hop fact chains that were simultaneously available in all five languages. Given that Wikipedia is written by local communities worldwide, this approach allow us to gather authentic localized expressions. In addition, we continue to collect knowledge from Wikidata for the objects in the single-hop fact chain to form a multi-hop fact chain. Figure 1 provides a simple example showcasing this process.

2.1.2 Generation of raw data

Once the fact chain is generated, additional data is required to compose the raw data. This additional data primarily encompasses edited answers and

[2] We collect data from Wikipedia via the Wikidata API: https://www.wikidata.org/w/api.php



```
Original_Triples:(Q440353,P27,Q145)
Original_Answer: United Kingdom
Edited_Triples:(Q440353,P27,Q9531)
Edited_Answer: Italy
```

```
First_Triples:(Q410,P108,Q9531)
Second_Triples:(Q9531,P17,Q145)
Original_Answer: America
Edited_Answer: United Kingdom
```

```
Question: Where did Carl Sagan work
during his career?
Original_Answer: Cornell University
Edited_Answer: BBC
```

```
Original_Question: Where is the primary
headquarters of the institution where Carl Sagan
was employed located?
Original_Answer: Ithaca
Edited_Question: By whom was the institution where
Carl Sagan worked established?
Edited_Answer: John Reith, 1st Baron Reith
```

Figure 1: Construction of MLaKE. Firstly, we gather a set of common relations and utilize them to extract single-hop and multi-hop fact chains from Wikidata, encompassing five languages. Then, we combine ChatGPT and manual collaboration to generate edited objects for them, and align the single-hop fact chains. Finally, we utilize the organized raw data to create QA data.

answer aliases, as depicted in the Raw Data of Figure 1.

To generate edited answers, we develop instructions to leverage powerful language models like ChatGPT. These instructions ensure the similarity between the edited answer and the original answer while avoiding conflicts with common knowledge within the LLM. For instance, it would be illogical to edit the single-hop knowledge chain ('Carl Sagan', 'employer', 'Cornell University') as ('Carl Sagan', 'employer', 'Glass Cup'). We represent the edited object as o^*. The editing process for a single-hop (or multi-hop) fact chain can be expressed as $(s_1, r_1, o_1 \rightarrow o_1^*)$ or $(s_1, r_1, o_1 \rightarrow o_1^*, r_2, o_2^*)$, where o_2^* is the object corresponding to r_2 and o_1^*.

We initially retrieve the answer alias using the Wikidata API. However, we observe that for certain languages, such as French and German, answers often have diverse variations that are not present in Wikidata. To ensure that these variations do not impact the evaluation (refer to Section 2.3), we design specific instructions for ChatGPT to expand the answer and incorporate appropriate qualifiers into the answer alias.

2.1.3 Construction of question and answer

Considering the complexities of inflection in French and German sentence structures, we avoid the template-driven methods often used in prior studies to convert triples into questions. Instead, we utilize ChatGPT to generate fluent and coherent multilingual questions, along with their corresponding answers, based on the collected triplet data in several languages. To minimize potential errors in this transformation process, five experts in the relevant languages were invited to review the model-generated texts, following the criteria detailed in Appendix C.1.

2.2 Description of the MLaKE

Dataset statistics Table 5 summarizes the statistics of the MLaKE dataset. The MLaKE dataset consists of more than 13K samples. We align all single-hop problems and employ them as a means to investigate the generalizability of existing knowledge editing methods across different languages following editing in a single language. Multi-hop problems are not aligned across languages, and we use them to assess the generalizability and transferability of existing editing methods.

Dataset analysis Figure 8 briefly analyzes the characteristics of the MLaKE dataset. Figure 8(a) depicts all first relations and their corresponding top 3 second relations, demonstrating the diversity of relations in MLaKE. The majority of questions pertain to nationality, names of individuals and locations, and typically adhere to the following structure: "From which nation does Gwendoline Christie hold citizenship?" (single-hop question) or "In which country is the institution where Carl Sagan was employed located?" (multi-hop question). In Figure 8(b), we further examine the relation PIDs that account for more than 1%, mainly including "country of citizenship" (p27), "country" (P17), "continent" (P30), etc. For the corresponding table of relationship PID and relationship label, please refer to the Appendix C. Figure 8(c) depicts the distribution of entities that have an occurrence rate exceeding 0.5%. The prominent entities in this distribution include 'United Kingdom' (Q145), 'Canada' (Q36), and 'United States of America' (Q30). Figure 8(d) illustrates the distribution of question lengths. The majority of questions fall within the 10-20 word range, which allows for precise expression of the subject and relationship without the inclusion of extraneous information. To accommodate various answer preferences of Large Language Models (LLMs), we strive to generate multiple aliases for each answer in MLaKE. Figure 8(e) demonstrates that the majority of answers possess 2-13 aliases, while there are even several answers with more than 20 aliases. The analysis data presented in Figure 8 only includes English samples, both single-hop and multi-hop.

2.3 Evaluation Metrics

Diverging from other benchmarks, MLaKE primarily focuses on assessing the generalizability and transferability of knowledge editing methods across multilingual scenarios. For different models, we use corresponding question to guide them to generate answers. We evaluate the accuracy of the question-answering task by determining whether the model-generated sentences contained the correct answer or its aliases.

3 Experiments

This section first explains the experimental settings, then analyzes the generalization ability of multilingual knowledge editing and the transfer ability of cross-language knowledge editing, and finally explores the potential reasons that affect the performance of knowledge editing.

3.1 Experimental Setup

Language Models We use the following two language models in our experiments: 1) **Vicuna-7B-v1.5** is fine-tuned from Llama-2 using supervised instruction fine-tuning with training data sourced from ShareGPT.com. 2) **Qwen1.5-7B-Chat** is a transformer-based aligned chat model pre-trained on extensive data, developed by Alibaba Cloud.

Knowledge Editing Methods Building on previous research (Yao et al., 2023; Wang et al., 2023c), we incorporate four strong knowledge editing methods as baselines: 1) **MEND** (Mitchell et al., 2022a) uses small auxiliary networks and gradient decomposition for efficient and localized post-hoc editing. 2) **ROME** (Meng et al., 2022) employs causal mediation analysis to pinpoint the area for edits, and then adjusts crucial feedforward weights using rank-one model editing. 3) **MEMIT** (Meng et al., 2023) extends ROME to edit a large set of facts and facilitate thousands of edits to be executed simultaneously. 4) **StableKE** (Wei et al., 2024), which leverages knowledge augmentation rather than focusing solely on knowledge localization, exhibits stability across various knowledge editing settings.

Evaluation Dimensions

- **Multilingual generalization** refers to the accuracy of knowledge editing across different languages, especially when edited knowledge is tested using single language.
- **Cross-lingual transferability** refers to the ability of the editing model to apply the knowledge edited in one language to another language, i.e., the accuracy of editing in other languages after performing editing in one language.

3.2 Generalization of Multilingual Knowledge Editing

To evaluate the multilingual generalization performance of commonly used knowledge editing methods, we select four widely used knowledge editing methods and evaluated their robustness across five different languages using the Vicuna and Qwen models. Each method are both edited and tested within the same language context. Table 1 presents the experimental results, and the main conclusion are as follows.

Conclusion 1: All evaluated knowledge editing methods exhibit limited multilingual generalization. As can be seen in Table 1, all methods demonstrate significantly higher accuracy in English knowledge editing compared to other languages, regardless of the base model used. This may be due to the fact that the quality and scale of English are the highest among the training corpora used by existing models.

Methods	Single-hop QA (%) - Vicuna					Multi-hop QA (%) - Vicuna				
	EN	DE	FR	JA	ZH	EN	DE	FR	JA	ZH
MEND	0.19	0.09	0.09	0.00	0.00	0.00	0.14	0.86	0.04	0.25
ROME	14.93	1.59	4.38	0.09	0.28	2.44	0.00	0.57	0.11	0.25
MEMIT	60.73	23.79	44.22	4.94	3.73	20.70	7.11	14.12	3.19	3.30
StableKE	88.43	37.31	37.31	32.09	28.73	26.66	13.68	5.99	11.11	11.66

Methods	Single-hop QA (%) - Qwen					Multi-hop QA (%) - Qwen				
	EN	DE	FR	JA	ZH	EN	DE	FR	JA	ZH
ROME	44.59	21.92	20.24	0.28	33.68	15.72	4.51	5.85	0.11	16.48
MEMIT	75.47	59.51	51.12	20.24	43.38	41.31	19.56	18.12	9.46	27.50
StableKE	92.82	45.34	44.96	37.31	67.44	55.37	22.02	10.56	17.05	31.94

Table 1: Single-hop and Multi-hop QA performance comparison between Vicuna and Qwen models using various knowledge editing methods across different languages.

Conclusion 2: The performance of the knowledge editing method is related to the performance of the base model. Except for StableKE, the other three methods perform poorly in editing Chinese and Japanese knowledge using the Vicuna model. In contrast, these methods achieve significantly higher accuracy with the Qwen model. The Figure 2 shows that the better the base model performs in a particular language, the more effective the knowledge editing method becomes.

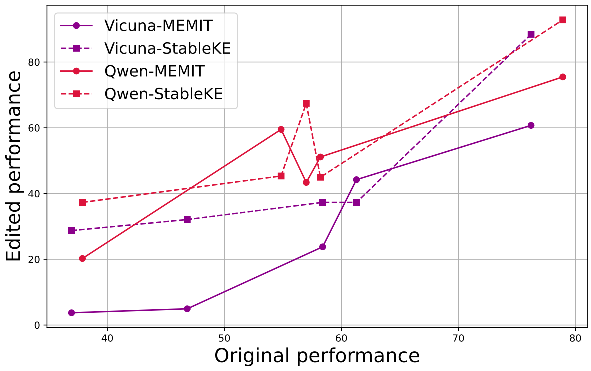


Figure 2: The relationship between the original performance and the edited performance of the model. With the same knowledge editing method, the better the model's original performance, the more effective the knowledge editing becomes.

Additionally, the table shows that the accuracy of multi-hop knowledge editing is substantially lower than that of single-hop editing. This finding aligns with previous studies focused on English, and our study extends these conclusions to a multilingual scenario.

3.3 Transferability of Cross-Language Knowledge Editing

To evaluate the transferability performance of existing knowledge editing methods in cross-language scenarios, we conduct knowledge editing in one language and assessed the editing accuracy in different languages. Our analysis focus on two representative methods: MEMIT and StableKE. Results from additional methods are presented in the appendix, and the main conclusion are as follows.

Conclusion 3: All knowledge editing methods have limited cross-language transferability, and the accuracy of the model in answering questions drops significantly when using different languages for reasoning. As illustrated in Figures 3, the editing performance along the diagonal is notably superior, indicating that the cross-language transferability ability of MEMIT and StableKE is limited. Notably, when performing knowledge editing in English on the Vicuna model, the accuracy in other languages was significantly lower than in English. A similar pattern was observed in the Qwen model during both Chinese and English editing.

Conclusion 4: The greater the similarity between the editing language and the reasoning language, the stronger the cross-language transferability. According to linguistic classification, English, German, and French all belong to the Indo-European language family (Joseph, 2005). Although Chinese and Japanese do not belong to the same language family, Japanese has been significantly influenced by Chinese (Handel, 2008; Jinlian, 2004). The figure illustrates a trend: in the top-left corner, cross-linguistic transfer among the three Indo-European languages is notably stronger,

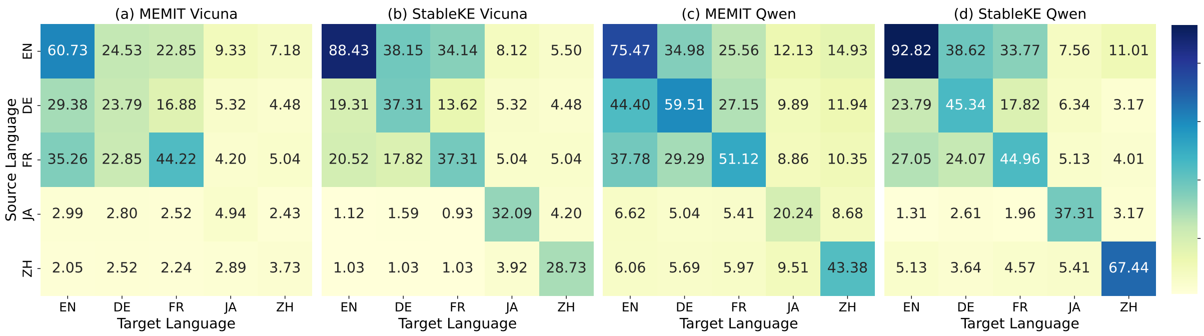


Figure 3: Performance of MEMIT and StableKE on different source and edit languages on the vicuna1.5 and Qwen1.5 model.

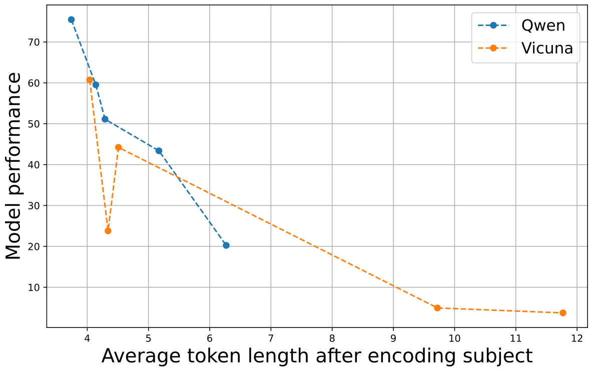


Figure 4: Relationship between the average number of tokens needed to encode each subject and the success rate of editing.

Method	Accuracy	
	Use Subject	No Subject
ROME	14.93	6.16
MEMIT	60.73	4.91

Table 2: Accuracy of MEMIT and ROME with and without using subject.

whereas the transfer from these languages to Chinese and Japanese is significantly weaker than within the Indo-European family. While there is some linguistic connection between Chinese and Japanese, it is weaker compared to the Indo-European languages, resulting in stronger transfer between Chinese and Japanese than their transfer to other languages, but still weaker than the transfer observed within the Indo-European language family.

4 Causes of Multilingual Editing Challenges

We identify two causes contributing to the suboptimal performance of multilingual knowledge editing: significant disparities in the multilingual capabilities of models and the structural independence of multilingual representations. The inconsistency in multilingual performance refers to the model's varying capability and encoding efficiency across different languages, which undermines the generalization of knowledge editing in multilingual contexts. Structural independence in multilingual models suggests that knowledge and abilities in different languages are encoded by distinct parameters. This separation limits the effectiveness of knowledge editing in achieving cross-linguistic transfer.

4.1 Potential Factors Limiting Multilingual Editing Generalization

Besides the model's multilingual performance, another crucial factor influencing the generalization ability of multilingual knowledge editing methods is its text encoding performance. To examine the relationship between text encoding performance and the generalization of multilingual knowledge editing, we analyze the ROME and MEMIT methods. Both methods follow the locate-then-edit approach, with their core process focused on identifying the subject's final token to enhance editing accuracy. However, as shown in Table 2, when the subject's final token is replaced with the sentence's last token, disrupting this process, the knowledge editing accuracy of both MEMIT and ROME declines significantly. Additionally, the Vicuna and Qwen models exhibit substantial differences in encoding efficiency across languages. For instance, as demonstrated in Table 3, the Vicuna model requires an average of 11.77 tokens to represent a Chinese subject, while only 4.04 tokens are needed for an English subject. As a result, the information density in the final token of a Chinese subject in Vicuna is lower than that of an English subject, negatively impacting target localization accuracy and reducing editing precision. As illustrated in Figure 4, as a model's encoding efficiency decreases (i.e., more tokens are required to encode a target),

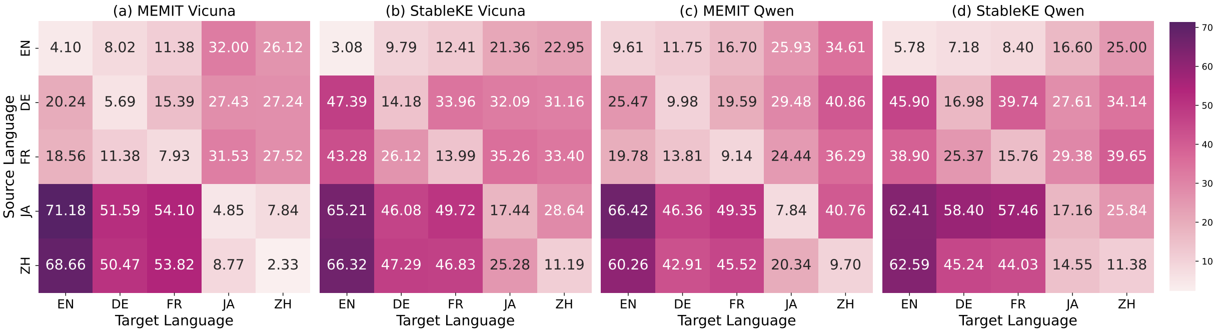


Figure 5: Proportion of responses that remain identical to the original, unedited outputs after applying MEMIT and StableKE with different source and edit languages on the Vicuna1.5 and Qwen1.5 model.

the success rate of knowledge editing correspondingly diminishes. In addition to the more apparent factor of the model's multilingual performance influencing generalization, we further demonstrate through experiments that the model's multilingual text encoding efficiency is another potential factor limiting multilingual editing generalization.

4.2 Examining Factors Affecting Cross-lingual Knowledge Transfer

Conclusion 5: Knowledge editing has a greater impact on the knowledge of languages that are more closely related to each other. To thoroughly investigate the factors influencing cross-lingual transferability in multilingual knowledge editing, we focus on evaluating how edits to knowledge affect content generated in languages that were not directly edited. we further analyze how MEMIT and StableKE impacted the models' ability to generate multilingual text. As shown in Figures 5, our results further illustrate the findings presented in Section 3.3. Knowledge editing exerts a stronger influence on the knowledge of languages closely related to the target language, while exerting a weaker influence on the knowledge of more distantly related languages. For example, when knowledge editing is performed in Chinese or Japanese, the accuracy of the original responses in three Indo-European languages decreases only slightly, indicating minimal disruption. These results suggest that existing knowledge editing methods have a smaller effect on knowledge associated with languages that are less closely related. For instance, when knowledge editing is conducted in Chinese or Japanese, the accuracy of responses in three Indo-European languages declines only marginally, indicating minimal interference. Similarly, when Indo-European languages are used for knowledge editing, the accuracy of original responses in Chinese and Japanese shows a comparable marginal decline, reflecting limited impact. These findings suggest that current knowledge editing methods have a reduced impact on knowledge related to languages that are less closely connected.

Model	EN	DE	FR	JA	ZH
Vicuna	4.04	4.34	4.51	9.72	11.77
Qwen	3.74	4.14	4.29	6.27	5.17

Table 3: Average number of tokens needed to encode a subject word by Vicuna and Qwen.

Conclusion 6: Knowledge editing has a greater impact on the generation ability of languages that are more closely related to each other. To comprehensively evaluate the impact of multilingual knowledge editing on non-edited languages, we assess its impact on the generative ability of non-edited languages. Without considering the accuracy of the text, we used ChatGPT-4 to evaluate the fluency of the content. As presented in Figure 6, knowledge editing minimally affects the generative performance of distantly related languages. Notably, when MEMIT is applied to edit Vicuna's knowledge in Chinese, the model's ability to generate coherent Chinese text deteriorated significantly, resulting in repetitive, incoherent, and meaningless output. In contrast, the impact on English text generation was less pronounced. Specifically, the fluency score for Chinese dropped sharply from 3.84 to 2.29, while the English fluency score declined only slightly from 4.03 to 3.81 following knowledge editing. This trend is observed consistently across multiple language families. We hypothesize that the model encodes knowledge and generative abilities for different languages in distinct parameters, with greater divergence in encoding for more distantly related languages.

To test this hypothesis, we analyze the differences in the multilayer perceptrons (MLPs) that are significantly affected at the same layer dur-

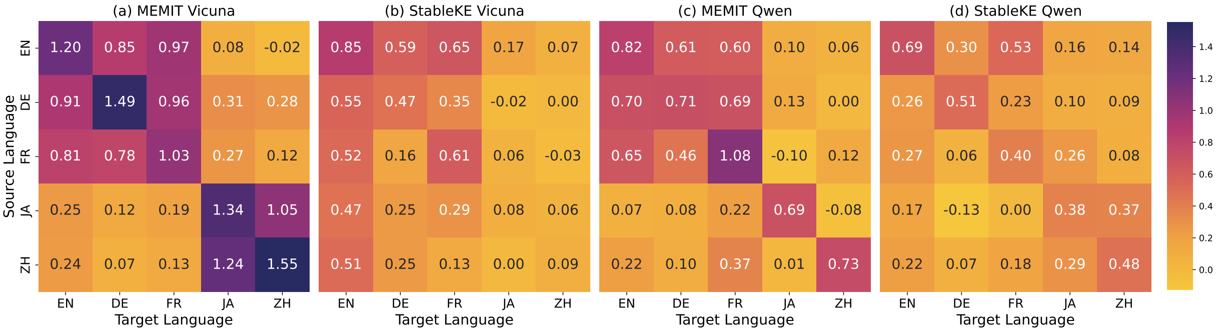


Figure 6: Fluency Performance of MEMIT and StableKE on different source and edit languages on the Vicuna1.5 and Qwen1.5 model.

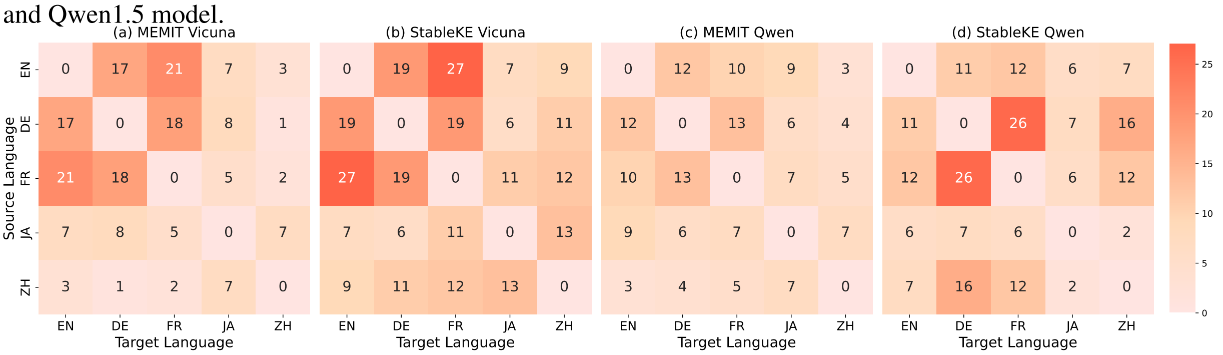


Figure 7: Overlap of the Top 100 and 200 Most Impacted Neurons Across Different Languages When Editing Vicuna and Qwen Models Using MEMIT and StableKE.

ing knowledge editing across different languages. Specifically, we identify the neurons that exhibited the largest changes when editing the Vicuna and Qwen models using the MEMIT and StaleKE methods. We then compare the extent of overlap in these neurons across various languages. As illustrated in Figure 7, the MEMIT method, which focuses on layers 4-8 for editing, operates within a narrower range but exerts a more pronounced effect on the corresponding parameters. We identified the top 100 neurons most impacted by MEMIT and analyzed their cross-linguistic overlap. Additionally, We analyze the overlap between the top 200 neurons primarily influenced by the StaleKE approach, which modifies all parameters of the model. Although StaleKE affects a broader range of parameters, its influence on individual neurons is comparatively smaller. The results indicate that the cross-linguistic overlap of the most impacted neurons is generally low, with all overlaps remaining below 20%. However, closely related languages demonstrate a higher degree of neuron overlap. For example, languages within the Indo-European family exhibit greater overlap compared to those from different language families. Given that knowledge and capabilities in different languages are encoded in distinct parameters, current knowledge-editing methods face limitations in their ability to generalize edits across languages when applied to a single language.

5 Conclusion

In this paper, we explore the effectiveness of current knowledge editing methods in multilingual settings. To this end, we create the MLaKE dataset by extracting knowledge tuples from Wikipedia and generating single-hop and multi-hop questions using ChatGPT. Leveraging MLaKE, we conduct experiments employing various methods and multilingual LLMs to investigate the generalization and transferability of knowledge editing from English and other languages. Our research findings reveal that: (1) Current knowledge-editing methods have limited generalization performance in multilingual settings. Beyond the inconsistent model performance across different languages, experiments suggest that insufficient multilingual encoding may contribute to their weak generalization. (2) The cross-lingual transferability of knowledge-editing methods is similarly constrained. Even within the same language family, transferability remains limited and is especially weak across distinct language families. A key finding from the experiments is that LLMs encode knowledge in different languages using distinct parameter sets, which reduces cross-lingual transfer performance.

Limitations

In this paper, we primarily explore knowledge-editing methods, focusing on parameter adjustment. These methods specifically target knowledge encoded in model parameters, and through a comprehensive analysis of their impact on the parameters, we highlight existing limitations in multilingual generalization and cross-lingual knowledge transfer. Our findings provide insights for advancing the development of more robust multilingual knowledge-editing methods. Given time and resource constraints, we do not extensively explore or compare knowledge-editing methods unrelated to parameter changes, like in-context editing.

Acknowledgements

This work was supported by the Strategic Priority Research Program of the CAS under Grants No.XDB0680302, the National Natural Science Foundation of China (NSFC) under Grants No. 62276248, and the Youth Innovation Promotion Association CAS under Grants No. 2023111.

Ethics Statement

This study thoroughly investigates multilingual knowledge editing in large language models and introduces the Multilingual Knowledge Editing Dataset-MLaKE. Using this dataset, we extensively demonstrate and analyze the performance of existing knowledge editing methods across different languages. The MLaKE dataset reveals notable disparities in current multilingual knowledge editing methods, specifically that the success rate of knowledge editing in English is significantly higher than in other languages. Consequently, MLaKE facilitates the advancement of more equitable and generalizable multilingual knowledge editing methods.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, Xifeng Yan, William Yang Wang, Philip Torr, Dawn Song, and Kai Shu. 2024. Can editing llms inject harm? *CoRR*, abs/2407.20224.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *CoRR*, abs/2311.05656.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *CoRR*, abs/2307.12976.

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.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *CoRR*, abs/2402.10612.

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.

Zev Handel. 2008. What is sino-tibetan? snapshot of a field and a language family in flux. *Language and Linguistics Compass*, 2(3):422–441.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *CoRR*, abs/2301.04213.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023a. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023b. Transformer-patcher: One mistake worth one neuron. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Lin Jin-lian. 2004. The chinese characters and vocabulary in japanese. *Shandong Foreign Languages Teaching Journal*.

Brian D Joseph. 2005. The indo-european family—the linguistic evidence. *History of the Greek Language from the beginnings up to later antiquity*, pages 128–134.

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), Vancouver, Canada, August 3-4, 2017*, pages 333–342. Association for Computational Linguistics.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. CMMLU: measuring massive multitask language understanding in chinese. *CoRR*, abs/2306.09212.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *NeurIPS*.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry V. Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/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. *CoRR*, abs/2303.04048.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. 2023b. Cross-lingual knowledge editing in large language models. *CoRR*, abs/2309.08952.

Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bozhong Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Guozhou Zheng, and Huajun Chen. 2023c. Easyedit: An easy-to-use knowledge editing framework for large language models. *CoRR*, abs/2308.07269.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2023d. Retrieval-augmented multilingual knowledge editing. *CoRR*, abs/2312.13040.

Yifan Wei, Xiaoyan Yu, Huanhuan Ma, Fangyu Lei, Yixuan Weng, Ran Song, and Kang Liu. 2023. Assessing knowledge editing in language models via relation perspective. *CoRR*, abs/2311.09053.

Zihao Wei, Liang Pang, Hanxing Ding, Jingcheng Deng, Huawei Shen, and Xueqi Cheng. 2024. Stable knowledge editing in large language models. *arXiv preprint arXiv:2402.13048*.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024. Unsupervised information refinement training of large language models for retrieval-augmented generation. *CoRR*, abs/2402.18150.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

Yue Zhang, Ming Zhang, Haipeng Yuan, Shichun Liu, Yongyao Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Llmeval: A preliminary study on how to evaluate large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19615–19622. AAAI Press.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

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

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix X. Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *CoRR*, abs/2012.00363.

A Evaluating QA Performance of Five Widely Used LLMs on MLaKE

To evaluate the effectiveness of LLMs in various languages and their content generation capabilities, we assessed five instruction-tuned LLMs: LLaMa-2-7B-chat, Vicuna-7B-v1.5, Qwen1.5-7B-Chat, Gemma-7B-IT, and Mistral-7B-Instruct-v0.2 s (Touvron et al., 2023; Team et al., 2024; Bai et al., 2023; Jiang et al., 2023). Using the unedited MLaKE data, we assessed two capabilities of these five models. The results for single-hop and multi-hop QA are presented in Table 4. Our key findings include: 1. The same model can display notable differences in QA performance when tested in different languages. English consistently performs better in both single-hop and multi-hop QA, indicating that LLMs have a deeper grasp of English. 2. Qwen1.5-7B-Chat demonstrates a more consistent performance across various languages than other LLMs, achieving especially strong results in Chinese and Japanese. We attribute this advantage to its pre-training corpus, which includes a diverse range of languages (Bai et al., 2023). 3. Conversely, Vicuna-7B-v1.5 shows markedly improved generative capabilities in Chinese and Japanese. This enhancement is mainly attributed to the inclusion of Chinese and Japanese data in the ShareGPT instruction-tuning dataset, which strengthens its ability to generate responses in these languages. 4. Gemma-7B-IT shows significant differences in accuracy in free-form QA compared to other models. Notably, it falls behind LLaMa-2-7B-Chat by 27.8% in single-hop Free form QA and 15.11% in multi-hop Free form QA. This gap is mainly attributed to the RLHF optimization in the Gemma model, which often leads to its refusal to generate responses.

B Implementation Details

All experiments are conducted on a single NVIDIA A800 GPU (80GB). We re-implement MEND, ROME and MEMIT using EasyEdit[3] (Wang et al., 2023c) with default settings. Additionally, we reproduce the results of StableKE by utilizing their official repository[4]. In the case of MEND, MEMIT, and StableKE, a batch size of 100 was employed for actual editing. As for ROME, which does not support batch editing, a batch size of 1 was used,

[3] https://github.com/zjunlp/EasyEdit

[4] https://github.com/Hi-archers/StableKE

Models	Single-hop QA (%)					Multi-hop QA (%)				
	EN	DE	FR	JA	ZH	EN	DE	FR	JA	ZH
LLaMa-2-7B-Chat	78.17	59.05	63.62	0.0	0.0	49.94	25.99	29.92	0.0	0.0
Vicuna-7B-v1.5	76.21	58.40	61.29	46.83	36.94	47.60	28.50	26.98	26.50	21.32
Qwen1.5-7B-Chat	78.92	54.85	58.21	37.87	57.00	60.81	30.46	30.56	30.51	53.55
Gemma-7B-IT	50.37	42.07	42.35	43.00	34.33	34.83	12.94	22.51	24.26	22.11
Mistral-7B-Ins.-v0.2	83.21	60.73	65.95	26.77	28.17	62.12	40.39	37.21	18.85	29.21

Table 4: Single-hop and Multi-hop Free-form QA performance of five LLMs on MLaKE.

and 100 samples were iteratively edited before testing.

C Data Details

In this section, we present the rules employed for filtering the data, followed by the instructions utilized for constructing the data.

C.1 Filter rules

We hired five professionals, each specializing in one of the five languages used in this paper, to conduct a comprehensive data cleaning exercise on the dataset according to the following guidelines:

1. All single-hop data has corresponding five language representations, which is the alignment operation mentioned in the main text. It serves as the foundational basis for conducting experiments on the generalization of multilingual knowledge editing.
2. Among all data, the relationship corresponding to an entity cannot be modified repeatedly. To facilitate evaluation and avoid knowledge conflicts, we ensure that the fact chain of an entity is not modified multiple times. This is particularly relevant considering that several knowledge editing methods support batch editing capabilities.
3. The dataset is free of toxic information, including content related to politics, violence, or pornography.
4. Fact chains, particularly multi-hop ones, are free from circular dependencies.
5. Identify and eliminate any semantic or syntactic errors in the questions and answers generated by the model.

C.2 Instructions

Tables 6 and 7 respectively display the prompts that guided ChatGPT in generating single-hop and multi-hop questions, as well as their corresponding answers.

D Related Work

We introduce recent datasets and knowledge editing methods in this section.

D.1 Knowledge Editing Datasets

Current research in knowledge editing datasets predominantly focuses on monolingual contexts. For instance, RIPPLEEDITS (Cohen et al., 2023), with its 5,000 instances of factual edits in English, serves as a pivotal benchmark designed to examine the ripple effects in knowledge editing processes. Similarly, MQuAKE delves into English multi-hop queries (Zhong et al., 2023), evaluating how edits influence intricate chains of knowledge. KEbench offers a thorough assessment of the stability of various knowledge editing methods using a tree-structured dataset in English. In contrast, the works of Bi-ZsRE (Wang et al., 2023b) and MzsRE (Wang et al., 2023d) extend to a multilingual knowledge editing dataset by translating the English Zero-Shot Relation Extraction (ZsRE) (Levy et al., 2017) dataset into various languages. Nonetheless, such endeavors in translation might introduce discrepancies in entity alignment, thereby possibly diminishing the quality of datasets (Wang et al., 2023b).

D.2 Knowledge Editing Methods

Current knowledge editing methods can be classified into four paradigms based on knowledge storage and learning techniques: locate-then-edit, memory-based models, meta-learning, and knowledge augmentation. **Locate-Then-Edit** involves identifying and updating a subset of model parameters associated with the edited knowledge. For instance, Dai et al. (2022) manipulates 'knowledge neurons' (KN) within pretrained transformers to

Statistics	**Language**				
	EN	ZH	JA	FR	DE
# single-hop questions	1072	1072	1072	1072	1072
# multi-hop questions (original)	916	760	849	782	765
# multi-hop questions (edited)	1024	798	909	701	731
# single-hop entities	602	596	596	596	594
# multi-hop entities (original)	1004	858	968	988	994
# multi-hop entities (edited)	1303	1053	1253	1267	1285
# relations	43	43	43	43	43

Table 5: Data statistics of MLaKE. Multi-hop questions are not aligned cross five languages, so we mark them with `original` and `edited` respectively. EN denotes English, ZH denotes Chinese, JA denotes Japanese, FR denotes French, and DE denotes German.

update facts. Similarly, Meng et al. (2022) introduces a method called Rank-One Model Editing (ROME) that modifies key feedforward weights to edit factual associations in LLMs. However, both ROME and KN are limited to modifying one piece of knowledge at a time. To address this limitation, Meng et al. (2023) extended the capabilities of ROME and developed MEMIT, enabling batch modification of knowledge simultaneously (Hase et al., 2023; Wei et al., 2023). **Memory-based Model** facilitates editing by introducing a small auxiliary model or additional parameters within the MLP layer, without altering the parameters of the original model. SERAC is a method that modifies knowledge by optimizing a counterfactual model (Mitchell et al., 2022b). On the other hand, T-Patcher achieves knowledge editing by integrating a few trainable neuron patches into the MLP layer (Huang et al., 2023b). In addition, CALINET leverages the characteristics of MLP layers to directly calibrate factual knowledge within LLMs (Dong et al., 2022). **Meta-learning** utilizes a hypernetwork specifically designed to handle the necessary alterations for manipulating knowledge within the MLP layers of models. KnowledgeEditor (De Cao et al., 2021) make use of hypernetworks to facilitate efficient edits in language models. MEND (Mitchell et al., 2022a) introduces auxiliary networks and enables scalable edits by decomposing gradients. **Knowledge augmentation** mainly includes StableKE (Wei et al., 2024) method enhances the stability and effectiveness of knowledge editing in large language models through two automated knowledge augmentation strategies: Semantic Paraphrase Enhancement and Contextual Description Enrichment.

E Multilingual Knowledge Editing Experiment

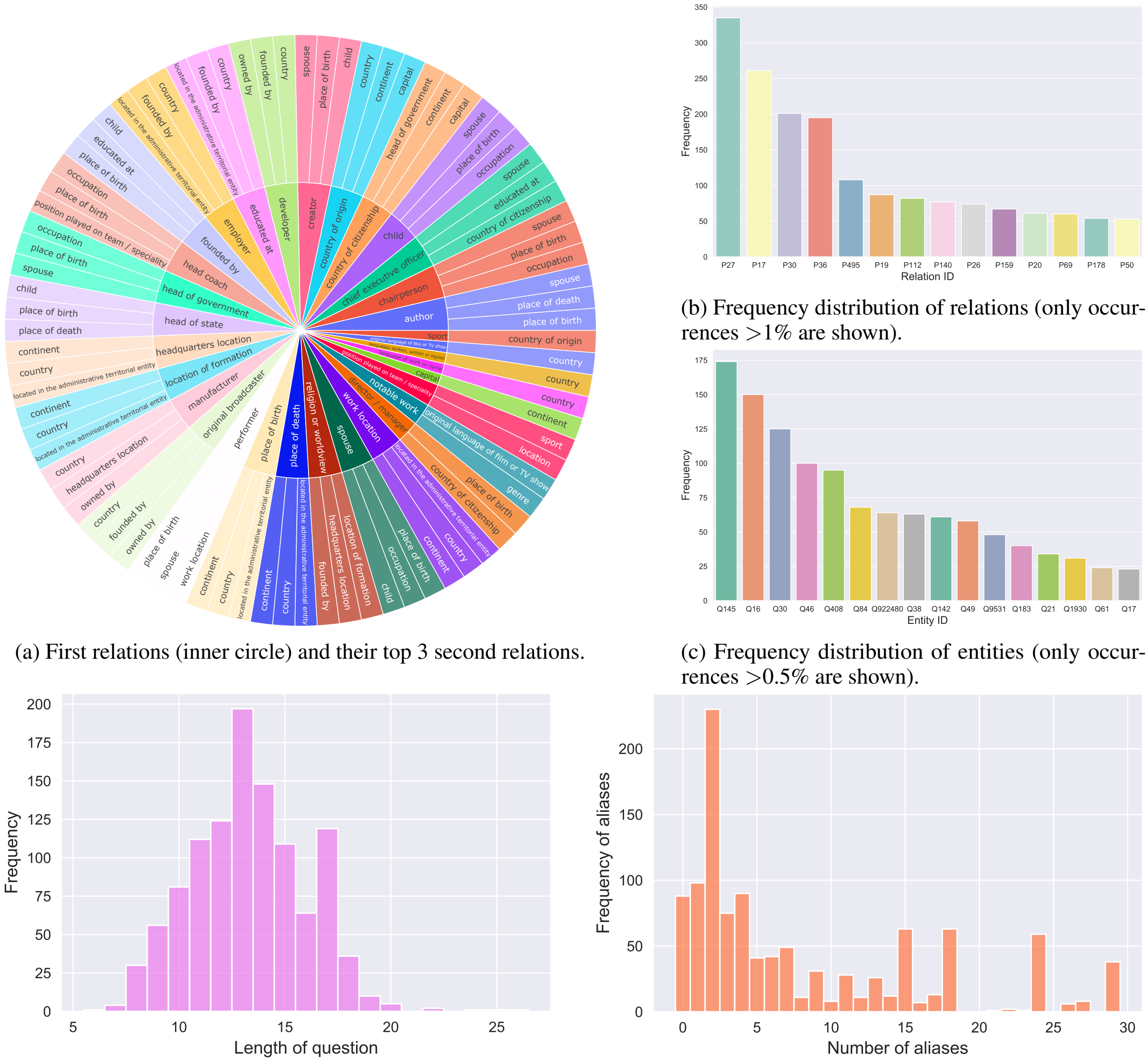
We present Figure 3 in table form, as shown in Table 8 and Table 9. The accuracy of English responses after editing in Chinese and Japanese exhibits a notable decrease compared to editing in French and German. This highlights the substantial influence of language disparities on the performance of cross-language knowledge editing.

English instructions for single-hop data	
Question Generation Instruction	Given the Wikidata knowledge triplet structure [subject, relation, object] where subject is s_1, relation is r_1, use this information to guide ChatGPT in creating a question that aims to identify the o_1 of the triplet. Your challenge is to create a detailed question that prompts LLM to identify the r_1 of s_1, without giving away the o_1 or making any reference to it.
Answer Generation Instruction	Using the provided Wikidata knowledge triple [s_1, r_1, o_1], craft a concise answer to the question (question). Your response should clearly link the s_1 with the o_1 as the answer, without delving into additional details or context. The aim is to directly address the query with the information given in the triple.
Chinese instructions for single-hop data	
Question Generation Instruction	给定Wikidata知识三元组[subject, relation, object]，subject是s_1，relation是r_1,。根据这些信息，设计一个中文问题，旨在根据s_1和r_1询问三元组中的object，即o_1。问题应详细且具体，让便让LLM准确的回答出s_1的r_1，同时避免直接提到或暗示o_1。
Answer Generation Instruction	根据给定的Wikidata知识三元组[s_1, r_1, o_1], 就问题(question)形成一个简洁的回答。您的回答应明确地将s_1与o_1联系起来作为答案，避免深入其他细节或背景。请力求简明扼要，利用特定的三元组关系，确保回答简洁且直接相关。请直接生成结果，不要说无关的内容。
Japanese instructions for single-hop data	
Question Generation Instruction	Wikidataの知トリプレット造[subject, relation, object]が与えられた合、ここでsubjectはs_1、relationはr_1です。この情を使用して、トリプレットのo_1を特定することを目的としたをChatGPTが作成するよういてください。あなたの挑は、o_1を明かしたり、それに言及したりすることなく、s_1のr_1を特定するようLLMに促すなを作成することです。
Answer Generation Instruction	ウィキデタの知トリプル[s_1, r_1, o_1]を考すると、、「(question)」にする正な答を作成します。あなたの答えは、「r_1」によって提供されるコンテキストを通じて、「s_1」を「r_1」に直接付ける必要があります。トリプルで示される特定のを活用したな明を目指し、答がでに直接していることをします。
French instructions for single-hop data	
Question Generation Instruction	Étant donné la structure de triplet de connaissances Wikidata [subject, relation, object] où subject est s_1, relation est r_1, utilisez cette information pour créer une question avec ChatGPT pour identifier l'élément o_1 du triplet. Votre défi est de créer une question détaillée qui incite LLM à identifier le r_1 de o_1, sans révéler le o_1 ou faire référence à celui-ci.
Answer Generation Instruction	Étant donné le triplet de connaissances Wikidata [s_1, r_1, o_1], formulez une réponse concise à la question (question). Votre réponse devrait clairement lier le s_1 avec le o_1 comme réponse, sans entrer dans des détails ou contextes supplémentaires. Visez une explication simple qui tire parti de la relation spécifique dénotée par le triplet, en vous assurant que la réponse est succincte et directement pertinente à la question. Veuillez générer les résultats directement et ne dites pas de contenu hors sujet.
German instructions for single-hop data	
Question Generation Instruction	Angesichts der Struktur eines Wissens-Tripels in Wikidata (subject, relation, object), wobei 'subject' s_1 ist, 'relation' r_1 ist, nutzen Sie diese Informationen, um ChatGPT bei der Erstellung einer Frage zu leiten, die darauf abzielt, das o_1 des Tripels zu identifizieren. Ihre Herausforderung besteht darin, eine detaillierte Frage zu formulieren, die LLM dazu anregt, das r_1 von s_1 zu identifizieren, ohne das o_1 preiszugeben oder darauf Bezug zu nehmen.
Answer Generation Instruction	Verwendung des bereitgestellten Wikidata-Wissenstripels (s_1, r_1, o_1), verfassen Sie eine prägnante Antwort auf die Frage (question). Ihre Antwort sollte s_1 eindeutig mit o_1 als Antwort verknüpfen, ohne auf zusätzliche Details oder den Kontext einzugehen. Ziel ist es, mit den im Tripel enthaltenen Informationen direkt auf die Anfrage einzugehen.

Table 6: Instructions required to generate single-hop data in five languages.

English instructions for multi-hop data	
Question Generation Instruction	Given the Wikidata triples: (s_1, r_1, o_1) and (o_1, r_2, x2), craft a multi-hop question in natural English about s_1 that explicitly involves the relationships r_1 and r_2. The question must ensure there is no implicit or explicit reference to or information leakage about s_1, leading to an inquiry about x2 without revealing s_1.
Option Generation Instruction	Given the information: 'question' and the known facts: (s_1, r_1, o_1) and (o_1, r_2, x2), please generate a correct option A and provide three incorrect but plausible options B, C, and D. Ensure that all options are presented in a sentence, not just single words or phrases. The incorrect options should be related enough to the correct answer to pose a challenge, but there's no need to mention the intermediary connecting entity (s_1) or any other detailed information.
Chinese instructions for multi-hop data	
Question Generation Instruction	请根据以下Wikidata知识三元组：（s_1, r_1, o_1）和（o_1, r_2, x2），用流畅的中文提出一个关于s_1的多跳问题，用于通过s_1查询得到x2，该问题必须明确涉及关系r_1和r_2，从而引出关于x2的询问。(问题要确保没有对s_1的直接或间接引用或信息泄露)。
Option Generation Instruction	给定信息：'question'，已知事实包括：(s_1, r_1, o_1)和(o_1, r_2, x2)。请根据这些信息，生成一个正确的选项A，并提供三个错误但听起来合理的选项B、C和D。请确保这些选项是以完整的句子形式提出的，而不仅仅是单个词或短语，错误选项应该与正确答案有一定的关联度，使问题具有一定的挑战性，但无需提及中间的连接实体(s_1)或其他详细信息。
Japanese instructions for multi-hop data	
Question Generation Instruction	次のWikidataの知トリプルに基づいて、s_1についての1つのを作成してください：（s_1, r_1, o_1）および（o_1, r_2, x2）。このは、r_1とr_2のを通じてs_1からx2への理的なを明にし、s_1にする直接または接的な参照を避けなければなりません。
Option Generation Instruction	情「question」と既知の事「s_1, r_1, o_1」および「o_1, r_2, x2」に基づき、正しい肢Aを生成し、不正解ながらも妥当な肢B、C、Dを3つ提供してください。すべての肢は、やフレズだけでなく、文章で提示する必要があります。不正解の肢は、正解と十分していて挑となる必要がありますが、中の接エンティティ「s_1」や他のな情に言及する必要はありません。
French instructions for multi-hop data	
Question Generation Instruction	Veuillez poser une question multi-sauts en français fluide basée sur les triplets de connaissances Wikidata suivants : (s_1, r_1, o_1) et (o_1, r_2, x2). La question doit concerner s_1 et servir à interroger pour obtenir x2, en mentionnant explicitement les relations r_1 et r_2 pour introduire une question sur x2. (La question doit s'assurer qu'il n'y ait aucune référence directe ou indirecte à s_1 ou de fuite d'informations à son sujet).
Option Generation Instruction	Informations fournies: 'question' Les faits connus incluent: (s_1, r_1, o_1) et (o_1, r_2, x2). Sur la base de ces informations, veuillez générer une option correcte A et trois options incorrectes mais raisonnables B, C et D. Assurez-vous que les options sont présentées sous forme de phrases complètes et pas seulement de mots ou d'expressions simples. Les mauvaises options doivent avoir un certain degré de pertinence par rapport à la bonne réponse pour rendre la question difficile, mais sans mentionner l'entité de connexion entre les deux (s_1) ou d'autres détails.
German instructions for multi-hop data	
Question Generation Instruction	Ich möchte eine deutsche Multi-Hop-Frage formulieren, basierend auf den gegebenen Wikidata-Wissens-Tripeln: (s_1, r_1, o_1) und (o_1, r_2, x2). Die Frage soll flüssig in deutscher Sprache gestellt werden und es ermöglichen, durch Abfrage von x0 die Information über x2 zu erhalten. Dabei muss die Frage die Beziehungen r1 und r2 explizit einbeziehen, ohne jedoch direkte oder indirekte Hinweise auf die Brückenentität x1 zu geben.
Option Generation Instruction	Gegebene Informationen: 'question' Zu den bekannten Fakten gehören: (s_1, r_1, o_1) und (o_1, r_2, x2). Generieren Sie auf der Grundlage dieser Informationen bitte eine korrekte Option A und drei falsche, aber vernünftig klingende Optionen B, C und D. Stellen Sie sicher, dass die Optionen als vollständige Sätze und nicht nur als einzelne Wörter oder Phrasen dargestellt werden. Die falschen Optionen sollten einen gewissen Grad an Relevanz für die richtige Antwort haben, um die Frage herausfordernd zu gestalten, ohne jedoch die verbindende Entität dazwischen zu erwähnen (s_1) oder andere Details.

Table 7: Instructions required to generate multi-hop data in five languages.



(a) First relations (inner circle) and their top 3 second relations.

(b) Frequency distribution of relations (only occurrences >1% are shown).

(c) Frequency distribution of entities (only occurrences >0.5% are shown).

(d) Distribution of the length of questions.

(e) Distribution of the number of edited answer aliases.

Figure 8: Analysis of MLaKE Dataset. **(a)** We illustrate the connections between the first relations (inner circle) and their corresponding second relations (outer circle). **(b)** We depict the distribution of relations occurring more than 1%. **(c)** We visualize the distribution of entities occurring more than 0.5%. **(d)** We present the distribution of question lengths. **(e)** We display the distribution of the number of edited answer aliases.

Source Language	**Single-hop Free-form QA** (%)					**Multi-hop Free-form QA** (%)				
	EN	DE	FR	JA	ZH	EN	DE	FR	JA	ZH
EN	**60.73**	24.53	22.85	9.33	7.18	**20.70**	15.18	14.12	10.01	9.13
DE	**29.38**	23.79	16.88	5.32	4.48	**17.97**	7.11	10.13	10.45	7.86
FR	35.26	22.85	**44.22**	4.20	5.04	**21.29**	9.03	14.12	9.90	8.49
JA	2.99	2.80	2.52	**4.94**	2.43	**9.67**	6.84	8.42	3.19	3.17
ZH	2.05	2.52	2.24	2.89	**3.73**	**10.94**	6.16	7.28	3.74	3.30

Table 8: The capability of MEMIT to edit knowledge in the source language and to generate accurate responses in a different target language.

Source Language	Single-hop Free-form QA (%)					Multi-hop Free-form QA (%)				
	EN	DE	FR	JA	ZH	EN	DE	FR	JA	ZH
EN	**88.43**	38.15	34.14	8.12	5.50	**26.66**	17.10	15.41	6.21	8.69
DE	19.31	**37.31**	13.62	5.32	4.48	**17.77**	13.68	12.98	6.97	8.14
FR	20.52	17.82	**37.31**	5.04	5.04	15.14	9.03	5.99	7.60	**15.51**
JA	1.12	1.59	0.93	**32.09**	4.20	9.67	6.84	7.28	6.34	**11.11**
ZH	1.03	1.03	1.03	3.92	**28.73**	8.59	5.61	6.13	**11.66**	7.59

Table 9: The capability of StableKE to edit knowledge in the source language and to generate accurate responses in a different target language.

	English Single Hop Free Form QA
$\mathcal{C}$	Question : Who is the mastermind behind the character Hannibal Lecter? original answer : Thomas Harris original answer aliases: William Thomas Harris III edit answer : Spede Pasanen edit answer aliases: None
	German Single Hop Free Form QA
$\mathcal{C}$	Question : Wer ist der Schöpfer des Charakters, der als Hannibal Lecter bekannt ist? original answer : Thomas Harris original answer aliases: None edit answer : Spede Pasanen edit answer aliases: None
	French Single Hop Free Form QA
$\mathcal{C}$	Question : Qui est l'esprit derrière le personnage de Hannibal Lecter ? original answer : Thomas Harris original answer aliases: None edit answer : Spede Pasanen edit answer aliases: None
	Japanese Single Hop Free Form QA
$\mathcal{C}$	Question : ハンニバル・レクターの物語を生み出した作家は誰ですか？ original answer : トマス・ハリス original answer aliases: トーマス・ハリス edit answer : スペデ・パサネン edit answer aliases: None
	Chinese Single Hop Free Form QA
$\mathcal{C}$	Question : 汉尼拔·莱克特的作品是由哪位作者创作的? original answer : 托马斯·哈里斯 original answer aliases: None edit answer : 斯佩德·帕萨宁 edit answer aliases: None

Table 10: Qualitative examples of the generated multi-hop questions on.