Multilingual Knowledge Editing with Language-Agnostic Factual Neurons

Xue Zhang[1]*, **Yunlong Liang**[2], **Fandong Meng**[2], **Songming Zhang**[1],
Yufeng Chen[1]†, **Jinan Xu**[1], **Jie Zhou**[2]
[1]Beijing Key Lab of Traffic Data Analysis and Mining,
Beijing Jiaotong University, Beijing, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China
{zhang_xue,smzhang22,chenyf,jaxu}@bjtu.edu.cn,
{yunlonliang,fandongmeng,withtomzhou}@tencent.com

Abstract

Multilingual knowledge editing (MKE) aims to simultaneously update factual knowledge across multiple languages within large language models (LLMs). Previous research indicates that the same knowledge across different languages within LLMs exhibits a degree of shareability. However, most existing MKE methods overlook the connections of the same knowledge between different languages, resulting in knowledge conflicts and limited edit performance. To address this issue, we first investigate how LLMs process multilingual factual knowledge and discover that the same factual knowledge in different languages generally activates a shared set of neurons, which we call language-agnostic factual neurons (LAFNs). These neurons represent the same factual knowledge shared across languages and imply the semantic connections among multilingual knowledge. Inspired by this finding, we propose a new MKE method by Locating and Updating Language-Agnostic Factual Neurons (LU-LAFNs) to edit multilingual knowledge simultaneously, which avoids knowledge conflicts and thus improves edit performance. Experimental results on Bi-ZsRE and MzsRE benchmarks demonstrate that our method achieves the best edit performance, indicating the effectiveness and importance of modeling the semantic connections among multilingual knowledge.

1 Introduction

Multilingual knowledge editing (MKE) (Wang et al., 2023b) aims to simultaneously rectify factual knowledge across multiple languages within large language models (LLMs). This process poses more challenges (Wang et al., 2023a) compared to monolingual knowledge editing (KE) since the edited factual knowledge should be updated together across multiple languages.

Edit Languages	Monolingual		Multilingual	
Test Languages	en	zh	en	zh
LoRA-FT	48.46	34.33	47.06 (−1.4)	33.17 (−1.16)
M-ROME	65.84	53.66	54.82 (−**11.02**)	46.27 (−**7.40**)
M-MEMIT	68.18	62.80	58.63 (−9.55)	57.16 (−5.64)
M-PMET	68.18	60.97	61.98 (−6.19)	57.33 (−3.64)
LU-LAFNs (Ours)	**72.85**	**66.71**	**73.92** (+**1.07**)	**67.29** (+**0.58**)

Table 1: The average EM results of Reliability, Generality, Locality, and Portability on Bi-ZsRE using Llama-3.1-8B as the backbone. "Monolingual" means editing and testing on the same one language, while "Multilingual" means editing on both en and zh, and testing on each language, respectively. The values in red represent the performance decline compared to monolingual KE due to knowledge conflicts across languages. The values in green indicate that our method avoids such conflicts and further promotes the edit performance compared to monolingual KE.

Recently, some Locate-then-Edit (Yao et al., 2023) methods, such as ROME (Meng et al., 2022), MEMIT (Meng et al., 2023), and PMET (Li et al., 2024), exhibit strong edit performance in monolingual KE. These methods identify parameters corresponding to specific knowledge and directly modify them to the target parameters. When adapting them to MKE, edit performance will probably degrade due to the conflicts between different languages (as shown in the red results of Table 1). Similarly, directly fine-tuning the original model with LoRA (Hu et al., 2021) also suffers from performance degradation due to the multilingual knowledge conflicts in LoRA modules. These conflicts can be attributed to the ignoring of the potential connections between multilingual knowledge in LLMs (Chen et al., 2023). Therefore, it is important to model the connections between multilingual knowledge during the editing process to avoid such conflicts.

To address this problem, we first investigate how LLMs process the same factual knowledge in different languages. We discover that the same mul-

* This work was done during internship at Pattern Recognition Center, WeChat AI, Tencent Inc, China.
† Yufeng Chen is the corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 5775–5788
January 19–24, 2025.

tilingual factual knowledge generally activates a shared set of neurons in feed-forward networks (FFNs), which we call Language-Agnostic Factual Neurons (LAFNs). These neurons represent the same factual knowledge shared across multiple languages and imply the semantic connections among multilingual knowledge. Based on this finding, we propose a new MKE method by **L**ocating and **U**pdating **L**anguage-**A**gnostic **F**actual **N**eurons (**LU-LAFNs**) to edit multilingual knowledge simultaneously. Specifically, we generate a set of paraphrases for multilingual knowledge to precisely locate LAFNs. Then we optimize the update values for modifying these located neurons to achieve simultaneous modification of the same multilingual knowledge. Additionally, to avoid the degradation of the edited model's general abilities due to directly modifying model parameters (Gu et al., 2024), we store the update values of the edited LAFNs in the cache. When the edited subject appears in the user query, the relative update values will be retrieved and used for model inference.

To evaluate the effectiveness of our method, we conduct experiments on two multilingual KE benchmarks, Bi-ZsRE (Wang et al., 2023a) and MzsRE (Wang et al., 2023b). Experimental results demonstrate that our method outperforms existing MKE methods in terms of Reliability, Generality, and Locality. Further analysis indicates that our method avoids conflicts by modeling the semantic connections between multilingual knowledge and thus improves the edit performance.

In summary, the major contributions of this paper are as follows[1]:

- We propose a new method by locating and updating language-agnostic factual neurons to achieve MKE. Our method avoids conflicts by modeling the semantic connections between multilingual knowledge.
- Experimental results on Bi-ZsRE and MzsRE benchmarks demonstrate that our method achieves the best edit performance, which proves the effectiveness of our method.
- We further analyze the key factors that influence multilingual edit performance, including LLMs' inherent language capabilities, the updated layers, and the number of LAFNs.

[1]The code is publicly available at https://github.com/XZhang00/LU-LAFNs.

2 Related Work

Multilingual Knowledge Editing. MKE aims to update multilingual knowledge simultaneously by using parallel multilingual data. ReMaKE (Wang et al., 2023b) retrieves the multilingual aligned knowledge from a multilingual knowledge base as context to achieve MKE. Additionally, some methods, such as LiME (Xu et al., 2023) and MPN (Si et al., 2024), explore cross-lingual knowledge editing, which only utilizes monolingual knowledge to edit the model and then test the edit performance on other languages. In this work, we mainly focus on MKE, which is more practical and performs better in updating multilingual outdated knowledge.

Multilingual Knowledge Analysis. Analyzing the multilingual capabilities of language models is always a research hotspot (Pires et al., 2019; Bhattacharya and Bojar, 2023; Kojima et al., 2024; Zhao et al., 2024), especially exploring the relationship between model architecture and multilingual capabilities. Tang et al. (2024) indicate that LLMs' proficiency in processing a particular language is predominantly due to a subset of neurons within FFNs. Similar to our work, Chen et al. (2023) discover the language-independent knowledge neurons of mBERT and mGPT, which store knowledge in a form that transcends language, but ignores how to control neurons to achieve desired outputs. Differently, we first investigate the language-agnostic factual neurons related to specific fact knowledge in LLMs and then modify them to achieve MKE.

3 Methodology

In this section, we first give the definition of MKE (§3.1). Then we investigate how LLMs process multilingual factual knowledge by identifying and analyzing the associated neurons (§3.2). Subsequently, we introduce our method LU-LAFNs for MKE (§3.3).

3.1 Task Definition

MKE aims to simultaneously update multilingual knowledge with new information while preserving previous accurate knowledge within the model. Formally, we denote the original model as $\mathcal{F}_\theta$ and the multilingual group of an edit descriptor (x^e, y^e) as $G = \{(x^e_\ell, y^e_\ell) | \ell \in L\}$, where x^e_ℓ is the question for the knowledge to be edited in language ℓ and usually contains a subject and a relation, and y^e_ℓ is the new answer of x^e_ℓ. On this basis, MKE will lead to a model $\mathcal{F}'_\theta$ to correctly answer the edited ques-

tion x_ℓ^e in each language ℓ and meanwhile maintain the original prediction on other unedited questions:

$$\forall \ell \in L, \mathcal{F}'_\theta(x_\ell) = \begin{cases} y_\ell^e, & x_\ell \in I(x_\ell^e), \\ \mathcal{F}_\theta(x_\ell), & x_\ell \notin I(x_\ell^e), \end{cases} \quad (1)$$

where $I(x_\ell^e)$ denotes a broad set of inputs with the same semantics as x_ℓ^e (Wang et al., 2023a).

3.2 Language-Agnostic Factual Neurons

To investigate how LLMs process the same factual knowledge represented in different languages, we identify and analyze language-agnostic factual neurons (LAFNs) within FFNs based on two multilingual LLMs. Specifically, we first separately identify the factual neurons associated with monolingual knowledge in each language. Then, we take the intersection of these neurons across different languages to obtain the LAFNs.

Identifying LAFNs. The forward process of the FFN layer in current LLMs can generally be described as the following two formulas:

$$h^i = \text{act\_fn}(\tilde{h}^i W_1^i) \cdot W_2^i, \quad (2)$$

$$h^i = (\text{act\_fn}(\tilde{h}^i W_1^i) \otimes \tilde{h}^i W_2^i) \cdot W_3^i, \quad (3)$$

where i denotes the i-th FFN layer, $\tilde{h}^i$/h^i are the output hidden states of the i-th attention/FFN layer, and act_fn$(\cdot)$ is the activation function. Eq.(2) represents the FFN structure of older LLMs, *e.g.*, BLOOM-series (Muennighoff et al., 2023). Eq.3 shows the FFN structure of the latest LLMs, *e.g.*, Qwen2 (Yang et al., 2024) and Llama3 (Dubey et al., 2024), where W_1^i, W_2^i, W_3^i correspond to the `gate_proj`, `up_proj`, `down_proj` matrix, respectively. In this process, knowledge neurons refer to the output activations by the activation function after the first matrix of FFNs, *i.e.*, act_fn$(\tilde{h}^i W_1^i)$. Then we define that the j-th neuron of the i-th FFN layer is activated when act_fn$(\tilde{h}^i W_1^i)_j > 0$ following the previous work (Tang et al., 2024).

For the factual neurons of language ℓ, we use the factual corpus C_ℓ in language ℓ to track the activation of neurons in each FFN layer during the forward propagation. Subsequently, we identify and select the neurons that are activated most frequently to form the neuron set. For instance, the set of factual neurons in the i-th FFN layer D_ℓ^i can be obtained using C_ℓ as follows:

$$N^i = \{n_j^i | n_j^i = \sum_{c \in C_\ell} \mathbb{1}(\text{act\_fn}(\tilde{h}_c^i W_1^i)_j > 0)\}, \quad (4)$$

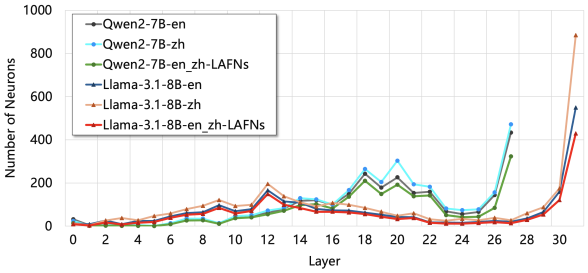

Figure 1: The identified neuron numbers in each layer of Qwen2-7B and Llama-3.1-8B. "xxx-en" and "xxx-zh" represent the English and Chinese factual neurons respectively. "xxx-LAFNs" refers to the language-agnostic factual neurons shared by English and Chinese.

$$D_\ell^i = \{j \mid \frac{n_j^i}{\max(N^i)} > \beta\}, \quad (5)$$

where $\tilde{h}_c^i$ contains $\tilde{h}^i$ at each token position in sentence c, $\mathbb{1}(\text{act\_fn}(\tilde{h}_c^i W_1^i)_j > 0)$ equals to 1 when act_fn$(\tilde{h}_c^i W_1^i)_j > 0$ otherwise 0, n_j^i is the total activation counts of the j-th neuron of the i-th FFN layer, N^i is the set of activation counts of all neurons in i-th FFN layer when processing C_ℓ, and β is the threshold to control the amount of D_ℓ^i. After obtaining the sets of factual neurons for each language in L, we calculate the intersection of all these sets in the i-th FFN layer to extract the shared knowledge neurons among all languages:

$$D^i = D_{\ell_1}^i \cap D_{\ell_2}^i \cap \cdots \cap D_{\ell_L}^i, \quad (6)$$

where we call D^i as the LAFNs in the i-th layer that imply the semantic connections of $\{C_\ell, \ell \in L\}$.

Experiments. We conduct analysis on PARAREL (Elazar et al., 2021), which contains factual knowledge with 34 relations in English. Here, we identify the LAFNs between English (en) and Chinese (zh). Firstly, we randomly choose 3000 sentences in each relation from PARAREL to build the factual corpus C_{en} (around 100k), and then utilize the `Google Translate API` to translate C_{en} to C_{zh}. We select two public multilingual LLMs: Llama-3.1-8B (Dubey et al., 2024) and Qwen2-7B (Yang et al., 2024). The layer numbers of the two LLMs are 32 and 28. The threshold β in Eq.(5) for two LLMs is set to 0.9 and 0.8. According to Eq.(5) and Eq.(6), we count the LAFNs in each layer for the two LLMs.

Results. We plot the identified neuron numbers in each layer of the two LLMs in Figure 1, including the factual neurons of each language and LAFNs, *i.e.*, D_{en}^i, D_{zh}^i and D^i. It shows that the changes

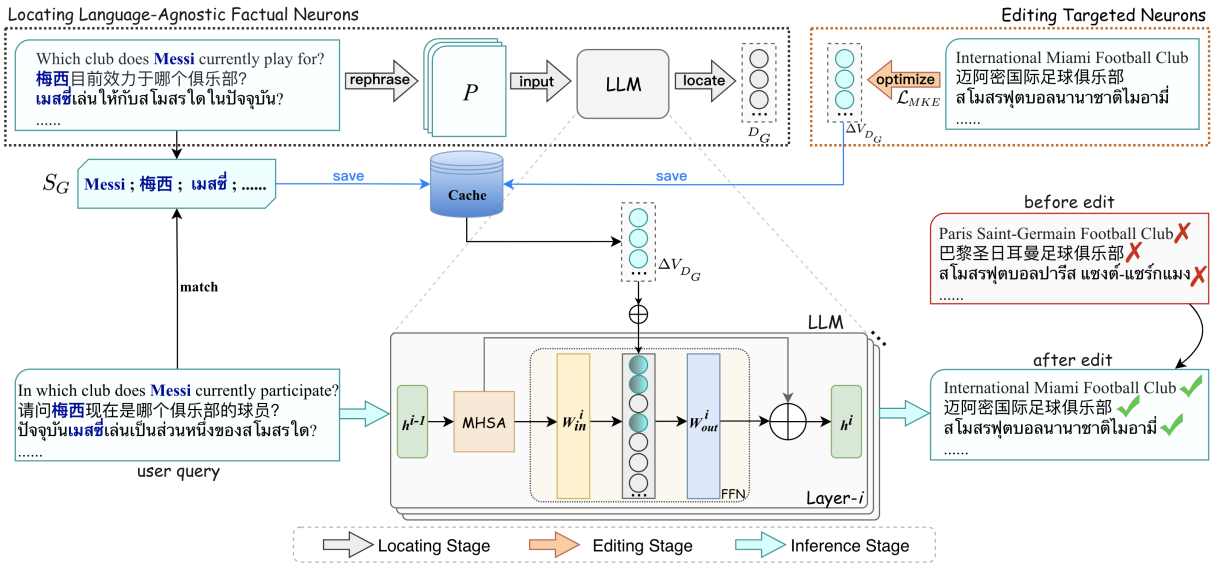

Figure 2: The architecture of our LU-LAFNs. Given the multilingual knowledge to be edited (including the aligned multilingual subject set S_G), we first locate the corresponding language-agnostic factual neurons D_G. Then the update values ΔV_{D_G} is optimized for modifying D_G, and $\{S_G : \Delta V_{D_G}\}$ is stored in cache. When the subject of the user query is matched in the cache, the relative ΔV_{D_G} is used for model inference.

of the neuron numbers for the two models exhibit similar trends, with a greater presence of language-agnostic knowledge neurons in the middle layers and the last layer (refer to the green and red lines in Figure 1). The difference is that in the middle layers, Llama-3.1-8B peaks in quantity at the 12th layer, while Qwen2-7B reaches its peak at the 18th layer. These results prove the existence of LAFNs, which represent the connections between the same factual knowledge in different languages and are mainly located in certain layers. Based on this finding, we design our method LU-LAFNs.

3.3 LU-LAFNs

Figure 2 shows the architecture of our method. We first locate the LAFNs for each group of multilingual edit descriptors, and then we optimize the update values to modify these neurons and store them in the cache. During the inference stage, when the subject of the user query is matched in the cache, the relative update values are utilized for model inference.

Locating Stage. Given the multilingual group G of an edit descriptor (x^e, y^e) ($G = \{(x_\ell^e, y_\ell^e) | \ell \in L\}$), we first locate the factual neurons D_ℓ^i in i-th layer for (x_ℓ^e, y_ℓ^e) in language ℓ according to Eq.(4) and Eq.(5). Specifically, to more precisely locate the neurons that are semantically related to x_ℓ^e, we use an LLM to generate several paraphrases for x_ℓ^e to build its paraphrase set as the factual corpus C_ℓ in Eq.(4). After obtaining D_ℓ^i in each language ℓ, we follow Eq.(6) to obtain the LAFNs set D_G^i of G in i-th layer.

Editing Stage. Given one multilingual edit description group G and its LAFNs set D_G of located layers, we aim to modify the values of D_G to edit knowledge in G simultaneously. Following the settings of MEMIT (Meng et al., 2023) and PMET (Li et al., 2024), we modify the values V_{D_G} of D_G at the last token position of the subject in the question x_ℓ^e. As for subjects, we obtain the corresponding aligned multilingual subject set S_G from G (refer to S_G in Figure 2). Then we will optimize the update values ΔV_{D_G} for adding to V_{D_G} to achieve MKE. That is, the model should generate the corresponding new answer y_ℓ^e by adding the ΔV_{D_G}:

$$\mathcal{F}_{(\theta, V_{D_G}+\Delta V_{D_G})}(x_\ell^e) = y_\ell^e. \tag{7}$$

To this end, we calculate the $\mathcal{L}_{target}$ to optimize ΔV_{D_G}:

$$\mathcal{L}_{target} = \frac{1}{|L|M} \sum_{\ell \in L} \sum_{m=1}^{M} -\log P_{\mathcal{F}'_\theta}(y_\ell^e \mid p_\ell^m + x_\ell^e), \tag{8}$$

where $\ell \in L$, $\mathcal{F}'_\theta = \mathcal{F}_{(\theta, V_{D_G}+\Delta V_{D_G})}$, and p_ℓ^m represents a randomly generated prefix to improve generalization (Meng et al., 2023) on $I(x_\ell^e)$, and M is the total number of prefixes.

Additionally, to ensure that the knowledge under the other relations of S_G is not affected, we also use

$\mathcal{L}_{kl}$ to optimize ΔV_{D_G} similar to MEMIT (Meng et al., 2023) and PMET (Li et al., 2024):

$$\mathcal{L}_{kl} = \frac{1}{|L|} \sum_{\ell \in L} \mathrm{KL}\big[P_{\mathcal{F}_\theta}(y \mid q_\ell) \,||\, P_{\mathcal{F}'_\theta}(y \mid q_\ell)\big], \tag{9}$$

where q_ℓ has the format of "$\{s_\ell\}$ is a" in language ℓ, s_ℓ is the subject in x^e_ℓ and $s_\ell \in S_G$, and $\mathrm{KL}[\cdot \,||\, \cdot]$ is the Kullback-Leibler divergence (Kullback and Leibler, 1951).

In the end, the overall optimized objective $\mathcal{L}_{\mathrm{MKE}}$ consists of the above two loss functions:

$$\mathcal{L}_{\mathrm{MKE}} = \lambda_1 \mathcal{L}_{target} + \lambda_2 \mathcal{L}_{kl}, \tag{10}$$

where λ_1 and λ_2 are hyperparameters to control the weight of two loss functions.

After obtaining ΔV_{D_G}, we store $\{S_G : \Delta V_{D_G}\}$ in the cache to avoid directly modifying the model parameters. When the subject s_ℓ of the current query x_ℓ is matched[2] in S_G, we retrieve the corresponding ΔV_{D_G} for model inference as follows:

$$\mathcal{F}'_\theta(x_\ell) = \begin{cases} \mathcal{F}_{(\theta, V_{D_G} + \Delta V_{D_G})}(x_\ell), & s_\ell \in S_G. \\ \mathcal{F}_\theta(x_\ell), & s_\ell \notin S_G. \end{cases} \tag{11}$$

4 Experiments

4.1 Experimental Settings

Datasets and Metrics. We conduct our experiments on Bi-ZsRE (Wang et al., 2023a) and MzsRE (Wang et al., 2023b). Bi-ZsRE covers English (en) and Chinese (zh) languages, and each language contains 10000/3000/1037 samples for the train/development/test set. MzsRE includes 12 languages[3], and each language consists of 10000/743 examples for the train/test set. Following Wang et al. (2023a), we calculate the F1/EM value of Reliability, Generality, Locality, and Portability as our evaluation metrics. The detailed introduction of metrics is listed in Appendix A.1.

Backbones. In our experiments, we select three public multilingual models as backbones to conduct MKE[4], including Llama-3.1-8B (Dubey et al., 2024), Qwen2-7B (Yang et al., 2024), and bloomz-7b1-mt (Muennighoff et al., 2023). Among them, each FFN layer of Llama-3.1-8B and Qwen2-7B follows Eq.(3), and bloomz-7b1-mt follows Eq.(2). The detailed supported languages of the three LLMs are introduced in Appendix A.2.

Implementation Details. When locating LAFNs in §3.3, we utilize the Qwen2-72B-instruct (Yang et al., 2024) model to generate 30 paraphrases for each x^e_ℓ. The detailed instruction is listed in Appendix A.3. The length of each randomly generated prefix p^m_ℓ in Eq.(8) is set to 5, and the total amount M of prefixes for each language is set to 4. Additionally, λ_1 in Eq.(10) is set to 1, and λ_2 is set to 0.0625 following MEMIT (Meng et al., 2023). For layers to be modified, we set (11, 12, 13, 31) for Llama-3.1-8B, (19, 20, 21, 27) for Qwen2-7B, and (9, 10, 11, 29) for bloomz-7b1-mt respectively. And the threshold β in Eq.(5) is set to 0.1, 0, and 0.2 for Llama-3.1-8B, Qwen2-7B, and bloomz-7b1-mt respectively.

4.2 Contrast Methods

Fine-tuning Method. We directly use LoRA (Hu et al., 2021) to conduct parameter-efficient tuning for the original LLM, namely LoRA-FT.

MKE Method. ReMaKE (Wang et al., 2023b) retrieves similar knowledge from a multilingual knowledge base as the context to instruct the model. Here, for the multiple languages to be edited, we retrieve[5] top-one question (with the answer) for each language and concatenate them as the context.

Adaptations of KE methods. We mainly adapt some Locate-then-Edit methods to MKE. For example, ROME (Meng et al., 2022) modifies the output matrix of one FFN layer located following causal tracing analysis. MEMIT (Meng et al., 2023) updates the output matrices of multiple FFN layers simultaneously. PMET (Li et al., 2024) conducts more precise editing based on MEMIT. We extend ROME, MEMIT, and PMET to M-ROME, M-MEMIT, and M-PMET to edit multilingual knowledge simultaneously. Specifically, since the knowledge to be edited of different languages corresponds to different answers, we train the new value for updating FFNs separately for each language. And we estimate the previously memorized keys of FFNs for each language.

[2] Here, we use the exact-match method.

[3] English (en), Chinese (zh), Czech (cz), German (de), Dutch (nl), Spanish (es), French (fr), Portuguese (pt), Russian (ru), Thai (th), Turkish (tr), and Vietnamese (vi).

[4] In the initial stage, we conduct cross-lingual experiments on Llama2-7B (Touvron et al., 2023). We list and discuss these results in Appendix B.

[5] We use XLM-RoBERTa-base shared by https://github.com/weixuan-wang123/ReMaKE.

Methods	Test Language: en				Test Language: zh				
	Reliability	Generality	Locality	Portability	Reliability	Generality	Locality	Portability	*avg*
Llama-3.1-8B (Edit Languages: en & zh)									
LoRA-FT	97.31 / 95.47	77.39 / 62.78	52.52 / 25.75	32.41 / **4.24**	86.59 / 74.06	75.56 / 52.07	27.78 / 2.41	30.32 / **4.15**	50.05
ReMaKE	43.40 / 16.30	44.84 / 17.74	**100.0 / 100.0**	34.45 / 0.96	56.66 / 20.25	57.00 / 20.44	**100.0 / 100.0**	**37.26** / 0.87	46.89
M-ROME	85.89 / 75.80	79.09 / 65.67	88.49 / 73.67	34.18 / 4.15	82.52 / 68.18	79.00 / 61.62	78.44 / 51.40	31.79 / 3.86	60.23
M-MEMIT	88.79 / 80.42	77.30 / 61.91	96.27 / 89.10	35.69 / 3.09	85.92 / 75.22	80.22 / 64.80	95.40 / 84.67	33.59 / 3.95	66.02
M-PMET	91.34 / 83.70	81.70 / 69.14	96.73 / 90.94	**35.87** / 4.15	85.39 / 73.87	81.48 / 67.02	95.28 / 84.76	33.29 / 3.66	67.40
LU-LAFNs (Ours)	**99.34 / 98.94**	**96.03 / 93.44**	**100.0 / 100.0**	29.84 / 3.28	**90.71 / 83.99**	**88.90 / 81.20**	**100.0 / 100.0**	29.01 / 3.95	**74.91**
Qwen2-7B (Edit Languages: en & zh)									
LoRA-FT	88.82 / 80.81	69.93 / 54.10	49.60 / 20.54	31.72 / **5.79**	96.17 / 92.29	83.01 / 67.89	52.31 / 21.60	34.94 / **7.71**	53.58
ReMaKE	43.64 / 16.30	44.75 / 17.55	**100.0 / 100.0**	**34.47** / 1.25	62.04 / 28.25	63.22 / 29.80	**100.0 / 100.0**	**39.07** / 3.47	48.99
M-ROME	81.89 / 70.97	74.08 / 59.59	92.74 / 83.41	33.11 / 1.83	88.15 / 77.05	81.52 / 65.86	93.41 / 81.68	33.65 / 4.15	63.94
M-MEMIT	97.17 / 94.99	89.00 / 82.26	94.17 / 86.60	34.18 / 2.89	98.56 / 96.24	93.44 / 87.17	95.62 / 87.95	33.94 / 5.21	73.71
M-PMET	88.13 / 79.27	77.91 / 64.32	93.59 / 85.25	34.10 / 2.22	90.24 / 80.42	82.53 / 67.31	95.25 / 87.17	33.38 / 3.95	66.57
LU-LAFNs (Ours)	**99.45 / 99.23**	**95.61 / 92.29**	**100.0 / 100.0**	30.27 / 2.03	**99.80 / 99.71**	**96.50 / 93.06**	**100.0 / 100.0**	30.78 / 5.11	**77.74**
bloomz-7b1-mt (Edit Languages: en & zh)									
LoRA-FT	83.76 / 75.31	64.11 / 48.60	29.63 / 7.62	23.14 / **3.66**	94.49 / 89.39	78.52 / 64.03	18.21 / 3.38	22.55 / 4.15	44.41
ReMaKE	28.78 / 2.03	28.30 / 1.16	**100.0 / 100.0**	22.29 / 0.00	61.08 / 37.99	60.77 / 38.38	**100.0 / 100.0**	**32.49 / 7.04**	45.02
M-ROME	70.27 / 52.75	63.50 / 43.30	77.29 / 59.88	26.67 / 0.58	84.58 / 71.36	78.17 / 62.49	67.99 / 48.41	26.55 / 5.01	52.43
M-MEMIT	99.09 / 98.07	90.07 / 84.57	98.44 / 96.62	**28.39** / 2.03	98.34 / 97.01	91.16 / 86.89	97.91 / 95.08	27.89 / 6.27	74.86
M-PMET	96.41 / 93.44	85.96 / 76.86	98.35 / 96.62	27.90 / 1.25	96.74 / 93.92	88.38 / 81.49	97.99 / 95.08	27.99 / 5.69	72.75
LU-LAFNs (Ours)	**99.83 / 99.71**	**96.40 / 94.41**	**100.0 / 100.0**	26.43 / 2.70	**99.65 / 99.42**	**97.78 / 96.43**	**100.0 / 100.0**	26.94 / 6.17	**77.87**

Table 2: The **F1/EM (%)** results on Bi-ZsRE using Llama-3.1-8B, Qwen2-7B, and bloomz-7b1-mt as backbones. Results in **bold** represent the best results. "***avg***" denotes the average value of all metrics in both two languages.

4.3 Experimental Results

Results on Bi-ZsRE. Table 2 shows the F1/EM results on Bi-ZsRE using Llama-3.1-8B, Qwen2-7B, and bloomz-7b1-mt as backbones. From the "***avg***" column, the average results of all metrics show that our method outperforms other baselines significantly in all three backbones. Particularly, our method exceeds other methods by almost >5 points in Reliability and Generality under F1&EM. The superiority in Reliability indicates that updating LAFNs can edit the multilingual knowledge (needs to be edited) more effectively, and in Generality means excellent generalization on the equivalent questions that have the same semantics as the edited questions. As for Locality, both our method and ReMaKE achieve the "100.00" value since the two methods do not modify the parameters of the original model during the editing process, not influencing previously learned knowledge. However, ReMaKE performs poorly in Reliability and Generality because the retrieved examples can not instruct the model to generate correct answers. LoRA-FT has a good performance in Reliability among baselines (*e.g.*, when Llama-3.1-8B testing on en, and when Qwen2-7B and bloomz-7b1-mt testing on zh), but it scores the lowest Locality since it dramatically modifies the original model parameters. Additionally, the adaptions of Locate-then-Edit methods to MKE (M-ROME, M-MEMIT, and M-PMET) perform moderately among all methods. Specifically, M-ROME is less effective than M-MEMIT and M-PMET because it only updates a single layer. M-PMET performs the second best on Llama-3.1-8B, and M-MEMIT performs the second best on Qwen2-7B and bloomz-7b1-mt, while both are inferior to our method. It demonstrates that the simple adaptations of these methods to MKE are less effective due to overlooking the connections of multilingual knowledge.

Portability, as a more difficult metric, measures whether the edited model can reason based on the edited knowledge via a portability question (Yao et al., 2023; Sun et al., 2024), where the relations and objects are out of the scope of the edited knowledge. The corresponding results show that all methods underperform on this metric without significant difference, especially when all EM results are less than 10, even than 5. We speculate that this reasoning ability is difficult to be well-measured without a reasoning process. We believe there is substantial room for measuring and improving portability in the future. Moreover, we observe that Llama-3.1-8B exhibits notably superior edit performance on English compared to Chinese since Llama-3.1-8B is not fine-tuned using Chinese instruction data. We guess that the inherent language capabilities of LLMs have a crucial impact on their edit performance.

Results on MzsRE. As for more languages, the average EM results of four metrics on MzsRE are reported in Figure 3. (Detailed EM results of each

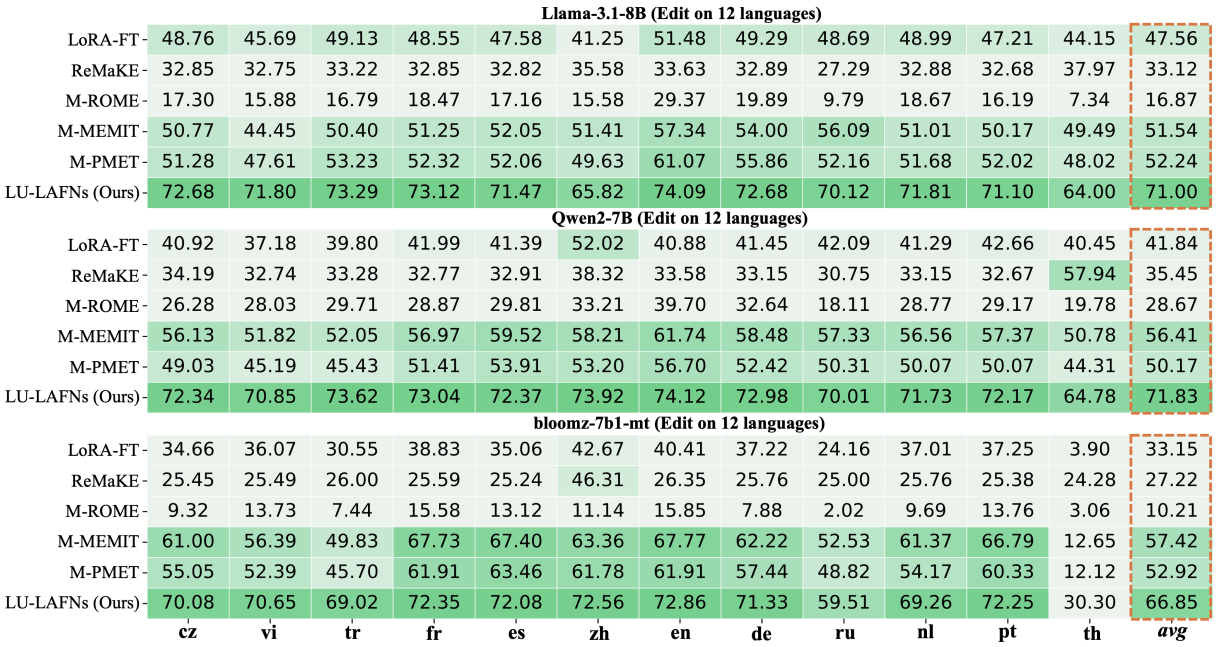

Figure 3: The average **EM (%)** results of four metrics (Reliability, Generality, Locality, and Portability) on the MzsRE dataset using Llama-3.1-8B, Qwen2-7B, and bloomz-7b1-mt as backbones. Values below 40.0 are shown in the same light color, while higher values have deeper colors indicating better performance. The orange box highlights the mean results across 12 languages. The detailed results for each metric are listed in Appendix C.

metric for the three LLMs are listed in Table 10, 11, and 12 of Appendix C.) Figure 3 shows that our method (with the deeper color) performs better than other baselines to a large extent on each language under three backbones. Among baselines, "M-ROME" performs worst since this method only updates one single layer, struggling to support simultaneous editing of more language knowledge. Other methods also underperform our method and exhibit a similar trend with the performance on Bi-ZsRE. For the edit performance of each language, most methods perform better on English than other languages under all backbones since these LLMs are primarily proficient in English (due to the existence of large-scale high-quality English data). Additionally, we also observe that the edit performance in the same language family is similar since these languages have a shared vocabulary, such as the Indo-European Family (Germanic languages: en, de, and nl, Slavic languages: cz and ru, Romance languages: es, fr, and pt). Moreover, Llama-3.1-8B has a worse performance on vi, zh, and th. Qwen2-7B also performs poorly on vi and th than other languages, while bloomz-7b1-mt performs badly on tr, ru, and th. The different edit performance of different LLMs on various languages is probably due to the language distribution of the training dataset and the linguistic characteristics of different languages. These results further demonstrate that the inherent language capabilities of LLMs determine the edit performance in different languages.

5 Analysis

In §5.1, we initially demonstrate the knowledge conflicts of other baselines. Then we explore the key factors affecting edit performance in §5.2. Subsequently, we compare different locating strategies to prove that using paraphrases during the locating stage can improve the edit performance (§5.3).

5.1 Conflicts of Editing Multilingual Knowledge

We conduct monolingual editing and multilingual editing experiments on Bi-ZsRE, and the results are reported in Table 1 and 3. Referring to the red values, we can find that most methods (*e.g.*, LoRA-FT, M-ROME, M-MEMIT, and M-PMET) on the three LLMs lead to conflicts when conducting MKE, resulting in the degradation of edit performance compared to monolingual KE. Among them, M-ROME has a dramatic decline due to the limited edit region. By contrast, our method conducts MKE by locating and updating LAFNs, which does not cause conflicts and further improves the edit performance than monolingual KE. Additionally, although the

Edit Languages	Monolingual		Multilingual	
Test Languages	en	zh	en	zh
Qwen2-7B				
LoRA-FT	40.60	48.60	40.31 (−0.29)	47.37 (−1.23)
ReMaKE	33.87	39.03	33.78 (−0.10)	40.38 (+**1.35**)
M-ROME	67.60	67.24	53.95 (−**13.65**)	57.19 (−**10.05**)
M-MEMIT	69.12	71.29	66.69 (−2.43)	69.14 (−2.15)
M-PMET	61.31	61.65	57.77 (−3.54)	59.71 (−1.93)
LU-LAFNs (Ours)	**72.35**	**74.04**	**73.39** (+**1.04**)	**74.47** (+0.43)
bloomz-7b1-mt				
LoRA-FT	33.83	40.65	33.80 (−0.03)	40.24 (−0.41)
ReMaKE	28.40	48.19	25.80 (−2.60)	45.85 (−2.34)
M-ROME	57.47	59.19	39.13 (−**18.35**)	46.82 (−**12.37**)
M-MEMIT	71.75	72.93	70.32 (−1.43)	71.31 (−1.61)
M-PMET	69.67	70.47	67.04 (−2.63)	69.05 (−1.43)
LU-LAFNs (Ours)	**72.81**	**75.34**	**74.21** (+**1.40**)	**75.51** (+**0.17**)

Table 3: The average EM results of four metrics on Bi-ZsRE using Qwen2-7B and bloomz-7b1-mt. "Monolingual" means editing and testing on the same one language. "Multilingual" means editing on both en and zh, and testing on each language, respectively. The values in red represent the performance decline compared with monolingual KE due to knowledge conflicts. The others in green represent no such conflicts.

A Single Layer	*avg*	Multiple Layers	*avg*
0	66.60	11-12	74.37
5	72.86	12-13	74.28
10	72.73	12-31	74.31
12	**73.35**	11-12-13	74.45
27	71.35	**11-12-13-31**	**74.91**
31	33.43	12-13-14-15-31	74.54

Table 4: The average F1/EM results of different layer settings on Bi-ZsRE using Llama-3.1-8B as the backbone (when $\beta = 0.1$).

ReMaKE method does not cause conflicts when Qwen2-7B testing on zh, its edit performance is much lower than our method. In summary, our method avoids the conflicts in MKE by locating and updating LAFNs, which represent the connections between multilingual knowledge.

5.2 Key Factors to Edit Performance

In this section, we explore the key factors affecting edit performance based on Llama-3.1-8B.

Updated Layers of LAFNs. Figure 1 in §3.2 has shown that the LAFNs are mostly located in some middle FFN layers and the last FFN layer. Thus, we further evaluate our method when updating LAFNs in different layers according to the distribution of LAFNs, including updating a single layer and multiple layers (the threshold β in Eq.(5) is set to 0.1 for this evaluation). The corresponding results reported in Table 4 show that in the single-layer setting, the edit performance achieves the best in the 12-th layer (which has the most LAFNs in the middle layers) and worst in the last layer. Although the last layer also has numerous LAFNs, we conjecture that these neurons are directly related to the final outputs, and thus a single update vector is difficult to fulfill answers in all languages. Moreover, we find that simultaneously editing multiple layers around the 12-th layer can further improve edit performance, with the best performance observed in (11, 12, 13, 31) layers.

β	Num (Proportion)	*avg*	β	Num (Proportion)	*avg*
0	14046 (98.0%)	74.85	0.5	1933 (13.5%)	73.26
0.1	**9738 (67.9%)**	**74.91**	0.6	1195 (8.3%)	72.08
0.2	6729 (46.9%)	74.79	0.7	720 (5.0%)	70.25
0.3	4613 (32.2%)	74.74	0.8	418 (2.9%)	66.30
0.4	3045 (21.2%)	74.44	0.9	223 (1.6%)	50.56

Table 5: The average F1/EM results of different β on Bi-ZsRE using Llama-3.1-8B as the backbone when editing (11, 12, 13, 31) layers. The "Num (Proportion)" represents the average number and proportion of LAFNs on each updated layer.

Methods	Llama-3.1-8B	Qwen2-7B	bloomz-7b1-mt
LU-LAFNs (Ours)	**74.91**	**77.74**	**77.87**
No-PGs	74.75 (↓ 0.16)	77.69 (↓ 0.05)	77.70 (↓ 0.17)
All	74.85 (↓ 0.06)	77.71 (↓ 0.03)	77.13 (↓ **0.74**)
Random	74.69 (↓ **0.22**)	77.61 (↓ **0.13**)	77.66 (↓ 0.21)

Table 6: The average F1/EM results of different locating strategies on Bi-zsRE using Llama-3.1-8B, Qwen2-7B, and bloomz-7b1-mt as backbones.

Number of LAFNs in Updated Layers. We also explore the influence of the threshold β in Eq.(5), which controls the number of LAFNs in each layer, when editing the (11, 12, 13, 31) layers. The results in Table 5 show that when $0 \leq \beta \leq 0.4$, the edit performance does not change obviously, and the best performance is achieved when $\beta = 0.1$, that is, 67.9% of neurons are located and modified in each layer. Moreover, when $\beta = 0.7$ (only updates 5.0% LAFNs for each layer), the performance (70.25) still exceeds the baselines in Table 2 (the best is 67.40 by M-PMET), proving the effectiveness of updating LAFNs. In summary, **both the updated layers and the number of LAFNs affect the edit performance, with the layers having a greater impact**. The discussions of Qwen2-7B and bloomz-7b1-mt are listed in Appendix D, which draw similar conclusions with Llama-3.1-8B.

5.3 Different Locating Strategies

To verify the effectiveness of using paraphrases during the locating stage, we compare three different

locating strategies with the original LU-LAFNs: (1) **No-PGs**: not using paraphrases to assist in locating LAFNs, *i.e.*, only using a single sentence in each language; (2) **All**: modifying all neurons of the same layers as LU-LAFNs without locating the set of LAFNs; (3) **Random**: randomly selecting the same number of neurons in the same layers to modify. The results in Table 6 show that the performance of all these three settings declines compared to the proposed method. These results demonstrate that using paraphrases during the locating stage can improve the edit performance since it can locate the LAFNs that are more semantically relevant to the multilingual knowledge to be edited.

6 Conclusion

In this work, we first identify language-agnostic factual neurons (LAFNs) in LLMs that represent the factual knowledge shared across different languages and imply semantic connections between multilingual knowledge. Then, we propose a new method LU-LAFNs to conduct MKE by locating and updating LAFNs. The experimental results demonstrate our method avoids knowledge conflicts and achieves the best MKE performance.

Limitations

In our approach, it is necessary to provide the aligned multilingual knowledge to be edited and their corresponding multilingual subjects, which is directly available in both Bi-ZsRE and MzsRE datasets. However, for other datasets that do not contain this information, we first need to preprocess the data to support our method. For example, if there is no corresponding multilingual data available, using translation API can translate the existing knowledge to be edited to other languages. If the corresponding subjects are not annotated, existing LLMs can be utilized to identify the aligned multilingual subjects in the sentences of each language. These preprocessing steps can be easily implemented by calling existing tools. Moreover, the current method for determining whether a subject exists in the cache adopts the exact-match approach, which is too strict. We will optimize it to a fuzzy matching method in future work to enhance the performance in practical application scenarios.

Furthermore, our method performs poorly in the Portability metric, which measures whether the edited model can reason based on the edited knowledge. Recently, Khandelwal et al. propose the cross-lingual multi-hop knowledge editing benchmark CROLIN-MQUAKE based on MQUAKE (Zhong et al., 2023) to test the multi-hop reasoning ability of the edited model. Next, we will test our method on this benchmark and further improve our method in reasoning scenarios.

Acknowledgments

The research work described in this paper has been supported by the National Nature Science Foundation of China (No. 62476023, 61976016, 62376019, 61976015), and the authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Sunit Bhattacharya and Ondřej Bojar. 2023. Unveiling multilinguality in transformer models: Exploring language specificity in feed-forward networks. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 120–126, Singapore. Association for Computational Linguistics.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. *Preprint*, arXiv:2308.13198.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Preprint*, arXiv:2102.01017.

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. *Preprint*, arXiv:2401.04700.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Aditi Khandelwal, Harman Singh, Hengrui Gu, Tianlong Chen, and Kaixiong Zhou. 2024. Cross-lingual multi-hop knowledge editing – benchmarks, analysis and a simple contrastive learning based approach. *Preprint*, arXiv:2407.10275.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language

models: Finding and controlling language-specific neurons. *Preprint*, arXiv:2404.02431.

S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024. Pmet: Precise model editing in a transformer. *Preprint*, arXiv:2308.08742.

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

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*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

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

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

Zengkui Sun, Yijin Liu, Jiaan Wang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2024. Outdated issue aware decoding for factual knowledge editing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9282–9293, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *Preprint*, arXiv:2402.16438.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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

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

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *Preprint*, arXiv:2402.18815.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2023.

Mquake: Assessing knowledge editing in language models via multi-hop questions. *Preprint*, arXiv:2305.14795.

A Detailed Experimental Settings

A.1 Evaluation Metrics

The details of the four metrics are as follows (Wang et al., 2023a). **Reliability** measures the average accuracy on the edit case. When receiving x_e as input, the edited model $\mathcal{F}_\theta^{'}$ should output y_e. **Generality** evaluates the average accuracy on the equivalent cases as the edit case. For instance, when receiving a rephrased sentence of x_e, the edited model $\mathcal{F}_\theta^{'}$ is also expected to output y_e. **Locality** assesses the accuracy of the edited model on the irrelevant samples. When the input x is irrelevant with x_e, $\mathcal{F}_\theta^{'}(x)$ should be the same as $\mathcal{F}_\theta(x)$ ideally. **Portability** measures the robust generalization of the edited model via a portability question that needs reasoning based on the edited knowledge. When receiving the portability question as input, the edited model $\mathcal{F}_\theta^{'}$ is expected to output the golden answer to demonstrate the model indeed learns the knowledge.

A.2 Supported Languages of LLMs

Llama-3.1-8B is fine-tuned with high-quality multilingual instruction data including English, French, German, Portuguese, Spanish, and Thai, and also has a certain degree of generalization ability to the other 6 languages in MzsRE. Qwen2-7B supports all 12 languages[6] in MzsRE. The bloomz-7b1-mt model is finetuned on the cross-lingual task mixture (xP3mt[7]) across 46 languages and 16 NLP tasks and has the capability of cross-lingual generalization to unseen tasks and languages.

A.3 The Instruction for generating paraphrases

We call the Qwen2-72B-instruct API (from the ALIYUN platform) to generate the paraphrase set P_ℓ for more precisely locating neurons. We directly use the default generation configs. The English version of the instruction for inputting Qwen2-72B-instruct is "You are an expert at sentence rewriting. Below I will give you a subject and a question containing the subject. Please give me 30 questions including this subject in English. They must have the same semantics as the given question. Subject: {}. Question containing this Subject: {}".

[6] https://qwenlm.github.io/zh/blog/qwen2/

[7] https://huggingface.co/datasets/bigscience/xP3mt

Edit on	en	en	zh	zh
Test on	en	zh	en	zh
ROME	72.96	35.11	41.46	47.61
MEMIT	76.26	36.56	42.64	48.41
PEMT	77.18	36.00	42.69	48.12
LU-LAFNs (Ours)	**79.96**	**37.34**	**43.74**	**55.57**

Table 7: The average F1 results of four metrics on Bi-ZsRE using Llama2-7B as the backbone under the cross-lingual edit setting.

B Cross-Lingual Experiments

In the initial stage, we conduct cross-lingual experiments on Llama2-7B (Touvron et al., 2023), i.e., we only utilize monolingual knowledge to edit the model and then test the edit performance on other languages. The results in Table 7 show that our method has better generalization on unseen languages than ROME/MEMIT/PEMT. However, there is still a large gap between the editing performance on unedited languages and that on edited languages. Therefore, we mainly focus on multilingual knowledge editing in this paper, which performs better in updating multilingual knowledge simultaneously.

C Detailed Results on MzsRE

The detailed EM results of four metrics on MzsRE for Llama-3.1-8B, Qwen2-7B, and bloomz-7b1-mt are listed in Table 10, 11, and 12, respectively.

D Different Settings of Qwen2-7B and bloomz-7b1-mt

We report the edit performance of Qwen2-7B and bloomz-7b1-mt under different layers (Table 8) and different values of β (Table 9). The changes in edit performance under different settings are similar to Llama-3.1-8B. Qwen2-7B achieves the best result when editing (19, 20, 21, 27) layers and $\beta = 0$, and bloomz-7b1-mt performs best when editing (9, 10, 11, 29) layers and $\beta = 0.2$. Additionally, when $\beta = 0.5$ on editing Qwen2-7B (only updates 5.1% LAFNs for each layer), the result (75.44) exceeds all baselines in Table 2 (the best result is 73.71 of M-MEMIT). And when $\beta = 0.5$ on editing bloomz-7b1-mt (only updates 1.9% LAFNs for each layer), the result (77.41) exceeds all baselines in Table 2 (the best result is 74.86 of M-MEMIT).

A Single Layer	*avg*	Multiple Layers	*avg*
Qwen2-7B			
0	61.16	19-20	77.49
5	34.35	20-21	77.29
10	76.02	19-20-21	77.51
18	76.33	18-19-20-21	77.36
20	**76.56**	**19-20-21-27**	**77.53**
27	33.43	18-19-20-21-27	77.51
bloomz-7b1-mt			
0	48.18	10-11	77.80
2	75.31	10-15	77.79
4	71.63	9-10-11	77.83
10	**77.54**	14-15-16	77.83
15	77.49	**9-10-11-29**	**77.84**
29	31.25	14-15-16-29	77.65

Table 8: The results of different layer settings on Bi-ZsRE using Qwen2-7B and bloomz-7b1-mt as backbones (when $\beta = 0.1$).

β	Num (Proportion)	*avg*	β	Num (Proportion)	*avg*
Qwen2-7B, layers=19-20-21-27					
0	**18190 (96.0%)**	**77.74**	0.5	962 (5.1%)	75.44
0.1	8907 (47.0%)	77.53	0.6	603 (3.2%)	71.84
0.2	4421 (23.3%)	77.25	0.7	375 (2.0%)	54.88
0.3	2539 (13.4%)	76.69	0.8	226 (1.2%)	38.04
0.4	1543 (8.1%)	76.29	0.9	125 (0.7%)	34.35
bloomz-7b1-mt, layers=9-10-11-29					
0	15220 (92.9%)	77.36	0.5	316 (1.9%)	77.41
0.1	8169 (49.9%)	77.84	0.6	114 (0.7%)	45.50
0.2	**4201 (25.6%)**	**77.87**	0.7	43 (0.3%)	33.03
0.3	2009 (12.3%)	77.68	0.8	17 (0.1%)	31.58
0.4	844 (5.2%)	77.21	0.9	0 (0.0%)	0.00

Table 9: The results of different β on Bi-ZsRE using Qwen2-7B and bloomz-7b1-mt as backbones under the best layer setting. The "Num (Proportion)" represents the average number and proportion of LAFNs on each updated layer.

Furthermore, the appropriate layer setting is more crucial to edit performance than the threshold β.

Methods	cz	vi	tr	fr	es	zh	en	de	ru	nl	pt	th	*avg*
Reliability													
LoRA-FT	98.65	96.10	98.79	97.17	96.77	81.29	99.19	98.38	98.25	99.19	95.96	96.23	**96.33**
ReMaKE	15.36	15.23	16.17	15.36	15.36	19.95	16.85	15.50	3.10	15.50	15.09	27.76	15.94
M-ROME	31.36	26.24	29.61	31.90	30.28	25.71	48.18	35.13	15.75	32.97	27.05	13.73	28.99
M-MEMIT	69.85	54.24	66.49	67.43	66.49	63.53	74.43	71.06	84.12	69.58	67.70	82.91	69.82
M-PMET	68.10	60.16	70.52	66.62	65.95	61.64	80.48	73.22	75.24	67.43	68.10	74.83	69.36
LU-LAFNs (Ours)	97.98	96.90	97.17	97.58	95.29	81.83	98.12	96.77	94.08	95.83	94.35	88.29	94.52
Generality													
LoRA-FT	88.83	81.83	89.10	89.77	88.69	76.99	94.89	91.52	88.16	91.66	87.21	78.20	87.24
ReMaKE	15.50	15.36	16.17	15.36	15.36	21.56	16.85	15.50	6.33	15.63	14.96	23.72	16.03
M-ROME	29.07	25.17	28.26	30.15	28.26	24.36	48.32	35.53	15.07	31.22	25.98	9.42	27.57
M-MEMIT	59.35	46.03	58.68	57.74	58.14	56.53	63.93	61.10	69.04	58.55	57.60	44.41	57.59
M-PMET	61.91	54.24	66.22	62.31	62.45	57.20	75.24	68.64	64.87	62.31	62.72	45.76	61.99
LU-LAFNs (Ours)	91.12	88.69	92.87	93.41	88.69	77.93	95.96	91.92	84.39	89.77	88.16	66.35	**87.44**
Locality													
LoRA-FT	3.77	2.42	4.31	3.63	2.42	3.36	5.92	3.63	4.17	2.56	2.83	1.08	3.34
ReMaKE	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.73	100.00	100.00	100.00	99.98
M-ROME	7.54	10.50	7.00	10.63	8.61	10.50	17.09	7.40	6.59	9.29	9.83	6.06	9.25
M-MEMIT	71.74	74.70	72.54	76.72	80.75	82.23	85.87	80.48	68.51	73.49	71.87	70.52	75.79
M-PMET	72.95	72.68	72.01	77.25	76.99	76.58	83.98	78.06	66.49	74.02	73.76	70.93	74.64
LU-LAFNs (Ours)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.73	100.00	100.00	100.00	**99.98**
Portability													
LoRA-FT	3.77	2.42	4.31	3.63	2.42	3.36	5.92	3.63	4.17	2.56	2.83	1.08	**3.34**
ReMaKE	0.54	0.40	0.54	0.67	0.54	0.81	0.81	0.54	0.00	0.40	0.67	0.40	0.53
M-ROME	1.21	1.62	2.29	1.21	1.48	1.75	3.90	1.48	1.75	1.21	1.88	0.13	1.66
M-MEMIT	2.15	2.83	3.90	3.10	2.83	3.36	5.11	3.36	2.69	2.42	3.50	0.13	2.95
M-PMET	2.15	3.36	4.17	3.10	2.83	3.10	4.58	3.50	2.02	2.96	3.50	0.54	2.98
LU-LAFNs (Ours)	1.62	1.62	3.10	1.48	1.88	3.50	2.29	2.02	2.29	1.62	1.88	1.35	2.05

Table 10: The **EM** results on the MzsRE dataset using **Llama-3.1-8B** as the backbone.

Methods	cz	vi	tr	fr	es	zh	en	de	ru	nl	pt	th	*avg*
Reliability													
LoRA-FT	85.60	78.73	83.04	83.98	85.60	94.08	80.62	83.18	90.04	85.60	88.69	91.66	85.90
ReMaKE	18.03	15.21	16.29	15.21	15.61	23.55	16.82	16.02	9.42	16.15	15.07	60.57	19.83
M-ROME	33.51	32.97	38.49	32.84	31.90	34.32	47.78	38.76	17.23	35.53	30.69	26.51	33.38
M-MEMIT	86.00	74.83	73.62	83.31	84.66	77.25	89.37	88.29	87.21	84.66	82.23	88.16	83.30
M-PMET	68.91	59.08	59.89	70.52	72.14	68.24	77.12	72.81	72.41	68.24	67.43	70.79	68.97
LU-LAFNs (Ours)	97.84	96.76	98.11	98.92	96.49	98.92	99.46	98.52	94.60	97.44	97.44	91.36	**97.16**
Generality													
LoRA-FT	72.41	63.53	70.12	76.31	73.62	84.79	74.02	75.50	73.62	74.56	77.25	65.14	73.41
ReMaKE	18.03	15.48	16.55	15.61	15.88	26.51	16.82	16.15	13.59	16.29	15.34	68.37	21.22
M-ROME	32.44	29.61	34.32	30.28	28.40	30.82	42.93	36.20	14.54	32.03	30.01	15.07	29.72
M-MEMIT	71.47	61.91	67.03	71.47	75.37	69.18	77.66	73.49	74.29	69.58	70.52	51.41	69.45
M-PMET	59.08	47.91	52.36	62.05	64.20	58.14	67.83	61.91	60.57	58.28	55.72	40.65	57.39
LU-LAFNs (Ours)	89.88	85.02	93.66	91.50	90.82	91.77	94.74	91.23	83.54	88.39	88.93	66.80	**88.02**
Locality													
LoRA-FT	3.10	4.71	3.90	4.98	3.77	21.53	3.90	4.44	1.88	3.10	2.29	3.63	5.10
ReMaKE	99.87	99.87	99.87	99.87	99.87	100.00	100.00	99.87	99.73	99.87	99.87	99.73	99.87
M-ROME	37.95	48.45	44.55	51.55	58.14	65.55	66.22	54.51	39.84	46.57	55.32	37.42	50.51
M-MEMIT	65.28	68.78	63.93	70.79	75.91	82.37	76.58	69.72	65.14	69.99	74.29	63.26	70.50
M-PMET	66.49	72.27	67.16	71.60	76.99	83.31	79.14	73.22	66.22	71.74	74.97	65.41	72.38
LU-LAFNs (Ours)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	**100.00**
Portability													
LoRA-FT	2.56	1.75	2.15	2.69	2.56	7.67	4.98	2.69	2.83	1.88	2.42	1.35	**2.96**
ReMaKE	0.81	0.40	0.40	0.40	0.27	3.23	0.67	0.54	0.27	0.27	0.40	3.10	0.90
M-ROME	1.21	1.08	1.48	0.81	0.81	2.15	1.88	1.08	0.81	0.94	0.67	0.13	1.09
M-MEMIT	1.75	1.75	3.63	2.29	2.15	4.04	3.36	2.42	2.69	2.02	2.42	0.27	2.40
M-PMET	1.62	1.48	2.29	1.48	2.29	3.10	2.69	1.75	2.02	2.02	2.15	0.40	1.94
LU-LAFNs (Ours)	1.62	1.62	2.70	1.75	2.16	4.99	2.29	2.16	1.89	1.08	2.29	0.94	2.12

Table 11: The **EM** results on the MzsRE dataset using **Qwen2-7B** as the backbone.

Methods	cz	vi	tr	fr	es	zh	en	de	ru	nl	pt	th	*avg*
Reliability													
LoRA-FT	78.06	74.43	67.97	77.12	71.47	89.64	79.14	79.68	59.62	78.47	73.89	8.88	69.86
ReMaKE	0.83	0.97	2.07	1.10	0.41	39.45	2.48	1.66	0.00	1.79	0.83	0.00	4.30
M-ROME	12.65	21.27	10.23	23.42	21.53	17.23	25.84	14.54	1.62	15.07	20.46	4.44	15.69
M-MEMIT	92.19	75.64	66.49	95.15	94.35	83.71	95.83	93.54	81.43	92.46	94.35	14.67	81.65
M-PMET	78.87	67.29	58.55	83.31	86.81	81.97	85.46	84.79	74.29	75.37	82.50	13.06	72.69
LU-LAFNs (Ours)	95.79	97.69	93.07	98.23	95.65	96.74	98.64	97.69	73.78	96.60	97.28	13.99	**87.93**
Generality													
LoRA-FT	53.57	60.30	49.53	65.14	61.91	73.49	71.60	65.81	32.57	63.66	63.39	4.85	55.49
ReMaKE	0.97	0.97	1.93	1.24	0.55	39.31	2.90	1.38	0.00	1.24	0.69	0.00	4.27
M-ROME	13.32	19.65	10.36	22.75	19.65	15.21	24.50	14.00	1.48	13.73	19.52	3.90	14.84
M-MEMIT	75.50	61.37	55.45	83.71	84.66	72.14	82.50	76.18	68.91	73.62	82.37	9.96	68.86
M-PMET	63.39	55.32	48.18	71.33	75.64	68.64	69.18	65.95	60.30	60.70	68.51	8.75	59.66
LU-LAFNs (Ours)	82.74	83.56	80.57	89.67	90.62	89.40	91.30	85.87	63.04	79.21	89.40	10.05	**77.95**
Locality													
LoRA-FT	6.06	7.54	3.77	11.17	5.38	3.90	8.34	2.29	3.23	5.11	9.29	1.75	5.65
ReMaKE	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	97.10	99.76
M-ROME	11.17	13.86	9.15	16.02	11.17	10.77	12.79	2.96	4.98	9.69	14.80	3.77	10.09
M-MEMIT	75.37	87.48	75.91	91.12	89.50	91.92	90.98	78.20	58.55	78.33	88.96	25.98	77.69
M-PMET	77.12	86.27	75.24	92.33	90.31	91.39	91.79	78.20	59.22	79.81	89.10	26.65	78.12
LU-LAFNs (Ours)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	97.15	**99.76**
Portability													
LoRA-FT	0.94	2.02	0.94	1.88	1.48	3.63	2.56	1.08	1.21	0.81	2.42	0.13	1.59
ReMaKE	0.00	0.00	0.00	0.00	0.00	6.48	0.00	0.00	0.00	0.00	0.00	0.00	0.54
M-ROME	0.13	0.13	0.00	0.13	0.13	1.35	0.27	0.00	0.00	0.27	0.27	0.13	0.23
M-MEMIT	0.94	1.08	1.48	0.94	1.08	5.65	1.75	0.94	1.21	1.08	1.48	0.00	1.47
M-PMET	0.81	0.67	0.81	0.67	1.08	5.11	1.21	0.81	1.48	0.81	1.21	0.00	1.22
LU-LAFNs (Ours)	1.77	1.36	2.45	1.49	2.04	4.08	1.49	1.77	1.22	1.22	2.31	0.00	**1.77**

Table 12: The **EM** results on the MzsRE dataset using **bloomz-7b1-mt** as the backbone.