Retrieval-Augmented Multilingual Knowledge Editing

Anonymous ACL submission

Abstract

Knowledge represented in Large Language Models (LLMs) is quite often incorrect and can also become obsolete over time. Updating knowledge via fine-tuning is computationally resource-hungry and not reliable, and so knowledge editing (KE) has developed as an effective and economical alternative to inject new knowledge or to fix factual errors in LLMs. Although there has been considerable interest in this area, current KE research exclusively focuses on monolingual settings, typically in English. However, what happens if the new knowledge is supplied in one language, but we would like to query an LLM in a different language? To address the problem of multilingual knowledge editing, we propose **Retrieval-Augmented Multilingual Knowledge Editor (ReMaKE)** to update knowledge in LLMs. ReMaKE can be used to perform model-agnostic knowledge editing in a multilingual setting. ReMaKE concatenates the new knowledge retrieved from a multilingual knowledge base with users' prompts before querying an LLM. Our experimental results show that ReMaKE outperforms baseline knowledge editing methods by a significant margin and is scalable to real-word application scenarios. Our multilingual knowledge editing dataset (MzsRE) in 12 languages, the code, and additional project information will be made available.

1 Introduction

Large Language Models (LLMs) are being used as sources of factual knowledge for search engines and other downstream tasks. Despite their considerable progress, knowledge generated by LLMs can be incorrect or become obsolete in a changing world. Pre-training from scratch or fine-tuning LLMs to adapt them to new knowledge is computationally expensive and not guaranteed to work. Knowledge editing (KE) methods (Zhu et al., 2020; Cao et al., 2021; Mitchell et al., 2022b; Zheng et al., 2023) have been proposed as effective and economic alternatives to fine-tuning when specific factual knowledge needs to be added or updated. KE involves either updating the parameters of a model (Dai et al., 2022a; Mitchell et al., 2022a; Meng et al., 2022, 2023; Dai et al., 2022b) or adding extra components to an LLM (Mitchell et al., 2022b; Zheng et al., 2023; Dong et al., 2022; Hartvigsen et al., 2022). For example, KE can be used to correct the answer to this question *"Who is the foreign secretary of the UK?"* from *"James Cleverly"* (true until mid November 2023) to *"David Cameron"*, who has recently been appointed to the post.

Despite significant interest in this problem, current research on KE predominantly concentrates on a monolingual setting, where both the injected knowledge and the subsequent queries to the LLM are in English (Mitchell et al., 2022a; Meng et al., 2022, 2023; Mitchell et al., 2022b; Zheng et al., 2023). Companies serving a multilingual customer base need to consider the multilingual KE case, where KE is done in one language and this propagates to queries and answers in all other languages. While Wang et al. (2023a) explored the cross-lingual applicability of knowledge editing to the English-Chinese cross-lingual scenario, their primary focus was to highlight the challenges rather than develop a functional KE approach in a multilingual setting.

Drawing inspiration from in-context learning (ICL), in-context knowledge editing (IKE) uses prompts to edit factual knowledge. It is noted that IKE is so far the only method demonstrating positive results in the cross-lingual KE task setting (Wang et al., 2023a). However, IKE requires explicit provision of new knowledge every time an LLM is used, confining its practicality and scalability in real-world applications. In addition, IKE suffers when irrelevant facts are included in the prompt (Wang et al., 2023c) especially in scenarios where a substantial number of facts are being edited.

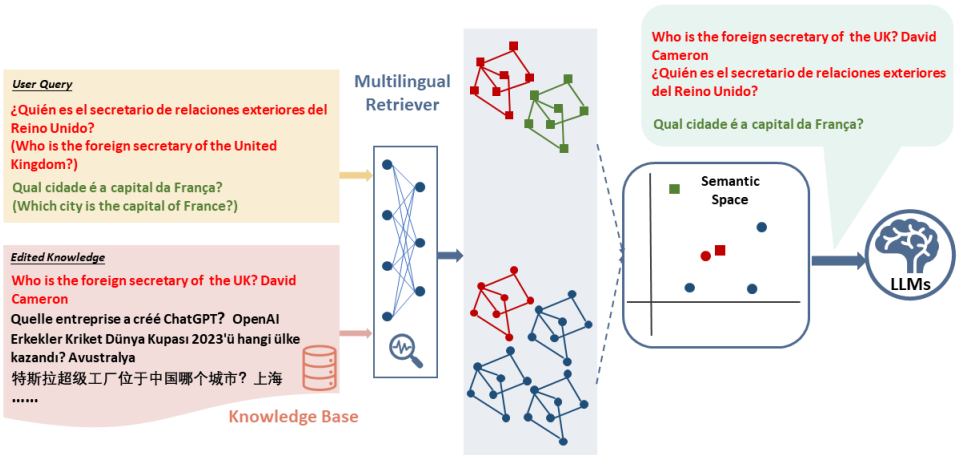


Figure 1: ReMaKE attaches in-context knowledge to an LLM prompt when it is retrieved (red example where the edited knowledge is in English and a user query is in Spanish) from a customer-defined multilingual knowledge base. When no edited knowledge is retrieved (green example) the prompt is passed to the LLM unchanged.

In this paper, we propose **Retrieval-Augmented Multilingual Knowledge Editor (ReMaKE)** that integrates multilingual retrieval from a knowledge base with in-context learning. ReMaKE concatenates the retrieved knowledge from an external database with a user query to create the prompt. The proposed multilingual retriever grounds the ReMaKE to the retrieved accurate and up-to-date information highly relevant to user queries, therefore effectively mitigating the contextual interference due to irrelevant context. In this way, the generated prompts are able to guide the LLMs in producing accurate responses associated with the injected knowledge. ReMaKE leverages a knowledge base's ability to scale to further enhance IKE's knowledge editing performance in real-world application scenarios where large volumes of edits are in scope. Figure 1 shows the architecture of the proposed retrieval-augmented multilingual knowledge editor. Our main contributions are listed below:

- **Multilingual knowledge editing**: ReMaKE extends the scope of knowledge editing practices across language boundaries. Given that the multilingual knowledge base and multilingual retriever operate independently to a specific LLM, ReMaKE is a **plug-and-play** tool applicable to any LLM. It is **scalable**, capable of editing a large number of knowledge. Experiments show ReMaKE outperforms baseline methods by a significant margin in the average accuracy score (up to +40.53%).
- **Multilingual editing dataset**: We build a machine-translated multilingual knowledge editing dataset (**MzsRE**) in 12 languages: English, Czech, German, Dutch, Spanish, French, Portuguese, Russian, Thai, Turkish, Vietnamese, and Chinese using the zsRE test-set (Levy et al., 2017). The dataset will be made available to the community.

2 Related Work

Knowledge editing: Monolingual knowledge editing methods can be categorized into four main paradigms (Yao et al., 2023; Zhang et al., 2023): **1. Hypernetwork editors** (Cao et al., 2021; Mitchell et al., 2022a; Hernandez et al., 2023) re-frame knowledge editing as a learning-to-update problem with the help of gradient shift, which is predicted by extrinsic editors. While the research scope extends beyond a single editing, the success rate of edits diminishes remarkably when more edits are executed simultaneously. **2. Locate-and-edit editors** (Dai et al., 2022a; Meng et al., 2022, 2023; Dai et al., 2022b) are used to locate the parameters related to factual knowledge and subsequently modify them. It is worth noting that this method requires an error-prone analytic step to identify parameters. It is model-specific and not efficient, as the locations are unique for each LLM. **3. Plug-in editors** (Cao et al., 2021; Mitchell et al., 2022a; Hernandez et al., 2023) can be used to add extra components to generate predictions about new knowledge without impacting on the parameters of the LLMs. Although this method is less influenced by unrelated inputs, it often cannot achieve precise editing. **4. Prompt-based editors** like IKE (Zheng et al., 2023) use ICL to attach knowledge

to the context of a prompt. Compared with other KE methods, IKE achieves a considerably stronger editing performance, together with far fewer side effects. However, IKE needs explicit provision of new knowledge every time, limiting its practicality and scalability in real-world applications. All above-mentioned editors are limited to a monolingual setting. For cross-lingual editing, Xu et al. (2023) and Chen et al. (2023) require training and manipulating model parameters. They rely on the masking mechanism, limiting the scope of application to a one-word output scenario. Most KE approaches are designed to process single editing and, therefore, have limited practicality. Our proposed editor, ReMaKE, addresses these problems and scales multilingual KE to cover a considerable amount of knowledge.

Retrieval-augmented in-context learning: Off-the-shelf search engines are often used to enhance retrieval-augmented ICL (Gao et al., 2021; Shi et al., 2023; Liu et al., 2023), attaching semantically similar examples to the context to improve the performance of LLMs in few-shot learning. In cross-lingual scenarios, a search engine first uses a low-resource language input sample as a query to find the most semantically similar high-resource language sample in the corpus. The retrieved high-resource language sample is concatenated with the input sample to form a prompt for an LLM. For instance, Nie et al. (2023) retrieves semantically similar cross-lingual sentences as prompts to improve the performance of sentiment classification for low-resource languages. While ICL can be used to support cross-lingual tasks, editing knowledge across language boundaries has not been explored. KE in a cross-lingual setting remains a challenge (Wang et al., 2023a).

3 Retrieval-Augmented Multilingual Knowledge Editing

ReMaKE enables knowledge in an LLM to be edited in one language and subsequently queried in multiple languages. ReMaKE consists of two stages: multilingual knowledge retrieval and multilingual in-context editing.

3.1 Multilingual Knowledge Retrieval

We propose a simple multilingual retrieval model to search for the most relevant knowledge stored in the knowledge base for a query. As shown in Figure 1, the proposed retrieval model initially maps a query and knowledge base entries to a shared multilingual semantic space. We train a classifier on top of these semantic representations to determine if the knowledge is semantically related to a query. The classifier is based on a sentence transformer (i.e., XLM-R).

More specifically, we finetune the multilingual retrieval model f_θ with a binary classification head on the multilingual parallel dataset constructed by translating our English training dataset using Google Translate. We use the separator token `</s>` to concatenate the sentence x and its corresponding translation $I(x)$ to format the input, predicting whether they are semantically related (related: $f_\theta(x, I(x)) \geq 0.5$ vs. unrelated: $f_\theta(x, I(x)) < 0.5$). Negative examples are constructed by pairing unrelated sentences between languages.

Once trained, the multilingual retriever f_θ is used to map a query x_{l_1} in language l_1 to knowledge k_{l_2} in language l_2. From a knowledge base $K_{l_2} = \{k^0_{l_2}, .., k^i_{l_2}, ..., k^K_{l_2}\}$, we use the retriever f_θ to score each knowledge item for the query and either select the most related knowledge, or an empty $R(x_{l_1})$:

$$k_{l_2} = R(x_{l_1}) = \begin{cases} k^{i^*}_{l_2} & f_\theta(x_{l_1}, k^{i^*}_{l_2}) \geq 0.5 \\ \textbf{None} & f_\theta(x_{l_1}, k^{i^*}_{l_2}) < 0.5 \end{cases} \quad (1)$$

where $i^* = argmax_i f_\theta(x_{l_1}, k^i_{l_2})$ is the index that maximizes the relevance $f_\theta(x_{l_1}, k^i_{l_2})$.

The multilingual retrieval model gives a solid performance on our test set with an average retrieval accuracy 99.31%, as shown in Figure 6 (Appendix A.2), indicating its solid performance in supporting ReMaKE. ReMaKE can be extended to accommodate a more efficient and performant Information Retrieval model for real-world deployment. We leave this extension as one of our future endeavors.

3.2 Multilingual In-context Editing

As shown in Figure 2, ReMaKE performs zero-shot and few-shot editing. In **zero-shot editing** (ReMaKE-zero), the retrieved result ("*new knowledge*" in Figure 2) is concatenated with a user's query ("*test input*" in Figure 2) to form a prompt ("*zero-shot prompt*" in Figure 2) for an LLM to predict the output $P(y_{l_1}|x_{l_1}, k_{l_2})$.

In **few-shot editing** (ReMaKE-few-bi), bilingual examples $S = \{(s^1_{l_1}, s^1_{l_2}), ..., (s^q_{l_1}, s^q_{l_2})\}$ are prepended to the "new knowledge" and the "test input", where $s^j_{l_1}$ and $s^j_{l_2}$ are the same statement in

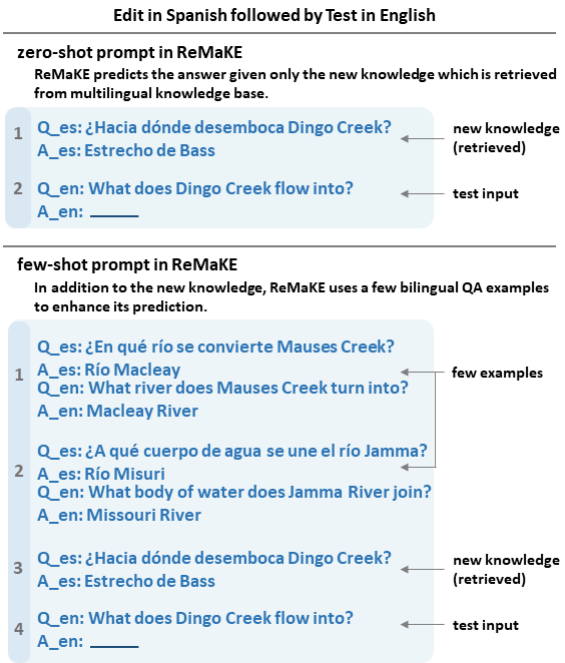


Figure 2: **Zero-shot and few-shot editing with ReMaKE.** The panels above show two methods of multilingual KE, in which a knowledge edited in Spanish is subsequently evaluated using an English question. "Q_en, A_en" and "Q_es, A_es" are QA pairs in English and Spanish.

languages l_1 and l_2, corresponding to the "*Q_es: ... A_es: ...*" and "*Q_en: ... A_en: ...*" ("*few examples*" in Figure 2). In few-shot editing, we concatenate "*few examples*", "*new knowledge*", "*test input*" as a prompt ("*few-shot prompt*" in Figure 2). The goal of predicting an edited knowledge is $P(y_{l_1}|x_{l_1}, k_{l_2}, S)$. For the few-shot setting, we follow Zheng et al. (2023) in selecting examples using an unsupervised method from the corpus based on their cosine similarity to the inputs (with the all-MiniLM-L6-v2 model[1]). The selected examples are included in the context to perform in-context learning. ReMaKE leveraging a search-based strategy can significantly outperform that based on a random selection (see our experiment in Appendix A.8).

4 Metrics, Data and Model

4.1 Metrics

Following Wang et al. (2023a), we evaluate multilingual knowledge editing with the following four metrics: (1) *Reliability* evaluates the average accuracy of an LLM on all edited instances. In other words, we update the LLM with new knowledge and test the accuracy of querying the LLM for this new knowledge. (2) *Generality* measures the average accuracy of an LLM for the paraphrased inputs for all edited instances. It indicates ReMaKE's effectiveness under the prompting frame bias (Wang et al., 2023c) induced by paraphrasing. (3) *Locality* assesses the average accuracy of an LLM in response to queries on irrelevant semantics after knowledge editing. It tests the knowledge editors' ability to update only the desired knowledge without affecting other knowledge in the model. (4) *Portability* estimates the average accuracy of an LLM for questions requiring reasoning after knowledge editing. Questions are constructed to test an LLM's ability to provide answers requiring it to reason. Portability can indicate whether KE can effectively adapt knowledge to support reasoning.

4.2 Data Construction

Zero-Shot Relation Extraction (zsRE) (Levy et al., 2017) is a monolingual question-answering test set containing 1,038 samples widely used in the knowledge editing task. There is a question-answer pair for each knowledge where the answer is an alternative counterfactual prediction (Cao et al., 2021). The counterfactual answer is expected to be generated by the post-edited LLMs. Additionally, a paraphrased question, an unrelated question, and a portability question are provided to evaluate the generality, the locality, and the portability of the editing. An example of editing knowledge in Spanish and testing in English for four metrics is shown in Table 1. We translate the zsRE from English (EN) to ten languages: Czech (CS), German (DE), Dutch (NL), Spanish (ES), French (FR), Portuguese (PT), Russian (RU), Thai (TH), Turkish (TR), and Vietnamese (VI) with Google Translate and use the Chinese (ZH) zsRE test set (which was also machine translated from zsRE) from (Wang et al., 2023a) to construct the multilingual zsRE test set (MzsRE). As there are multiple counterfactual answers for the same question in zsRE, we de-duplicate MzsRE to 743 items to ensure each question corresponds to a single unique answer, avoiding conflicting entries in the knowledge base. Table 5 (Appendix A.1) lists the statistics of MzsRE.

4.3 Base LLMs

Two representative multilingual LLMs are selected as backbones for us to test various KE methods in

[1]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Edited Knowledge	¿Qué ciudad fue el lugar de nacimiento de Henning Löhlein? Munich		
Metrics	**Question**	**Answer**	**Ground Truth**
Reliability	Which city was the birthplace of Henning Löhlein?	Munich	Bonn
Generality	In which city is Henning Löhlein born?	Munich	Bonn
Locality	Who is the lead singer of collective soul?	Ed Roland	Ed Roland
Portability	In which German state was Henning Löhlein born?	Bavaria	North Rhine

Table 1: An example of editing knowledge in Spanish and testing in English for four metrics. "Answer" represents the counterfactual post-edited knowledge which is needed to predict, and "Ground Truth" is the factual knowledge.

Metrics	Edit on EN	Test on												AVG
		EN	**CS**	**DE**	**NL**	**ES**	**FR**	**PT**	**RU**	**TH**	**TR**	**VI**	**ZH**	
Reliability	LLaMA	1.08	0.13	0.54	0.27	0.13	0.27	0.27	0.40	0.27	0.54	0.13	1.21	0.44
	SERAC	91.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.60
	ROME	68.91	10.77	16.02	15.48	12.25	10.09	12.11	0.13	0.13	1.21	4.31	1.21	12.72
	IKE	100.0	50.34	51.49	44.26	36.45	43.39	38.09	3.86	3.18	39.44	40.02	6.36	38.07
	ReMaKE-zero	96.37	61.10	64.87	54.91	52.62	53.43	54.51	27.73	5.92	45.22	48.32	25.44	49.20
	ReMaKE-few-mono	100.0	56.26	57.87	49.93	43.47	48.32	45.49	19.78	5.65	43.47	41.72	17.63	44.13
	ReMaKE-few-bi	**100.0**	**75.10**	**81.70**	**72.68**	**68.10**	**73.35**	**71.20**	**62.58**	**32.44**	**70.79**	**68.37**	**54.78**	**69.26**
Generality	LLaMA	0.94	0.13	0.94	0.40	0.13	0.13	0.13	0.27	0.13	0.13	0.13	1.48	0.41
	SERAC	26.78	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.23
	ROME	56.53	10.90	14.40	11.96	11.71	8.34	9.56	0.13	0.00	1.48	4.17	0.81	10.83
	IKE	**98.65**	49.76	51.49	43.88	35.39	42.91	37.61	3.38	3.18	39.15	39.34	5.98	37.56
	ReMaKE-zero	86.81	57.60	62.85	53.16	50.34	50.74	51.01	24.50	6.06	42.66	46.03	23.01	46.23
	ReMaKE-few-mono	98.25	55.59	57.34	48.59	43.61	47.64	44.68	18.57	5.52	42.4	41.18	17.23	43.38
	ReMaKE-few-bi	98.25	**73.76**	**80.62**	**71.60**	**67.97**	**71.60**	**70.66**	**62.45**	**32.97**	**70.12**	**67.83**	**53.57**	**68.45**
Locality	**SERAC**	99.46	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.87	99.94
	ROME	92.87	84.25	87.48	87.08	88.83	88.56	86.54	84.39	97.31	87.21	95.02	91.92	89.29
	IKE	38.48	0.39	5.69	1.54	1.74	0.48	0.48	0.19	1.35	0.96	0.96	0.96	4.44
	ReMaKE-zero	**99.46**	**98.65**	**99.73**	**99.87**	**98.52**	**99.06**	**99.19**	**97.58**	**95.29**	**97.17**	**97.71**	**94.48**	**98.06**
	ReMaKE-few-mono	99.46	98.38	99.6	99.73	98.52	99.06	99.19	97.58	95.29	97.04	97.71	94.48	98.00
	ReMaKE-few-bi	99.46	98.25	99.60	99.73	98.25	98.92	99.19	97.44	95.29	97.04	97.71	93.94	97.90
Portability	LLaMA	8.48	2.29	3.50	2.83	3.90	2.29	3.10	0.54	0.27	0.94	1.88	1.08	2.59
	SERAC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	ROME	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IKE	17.26	1.54	4.63	3.28	1.93	2.51	2.89	0.10	0.10	0.87	1.74	0.10	3.08
	ReMaKE-zero	**34.59**	**12.11**	**18.30**	**13.73**	**11.71**	**12.25**	**12.92**	**3.50**	**0.27**	**5.38**	**9.83**	3.63	**11.52**
	ReMaKE-few-mono	31.49	6.46	11.57	9.69	10.23	8.48	10.23	2.02	0.13	4.04	5.79	2.42	8.55
	ReMaKE-few-bi	31.49	7.67	11.31	9.02	8.61	8.08	9.83	5.79	0.67	3.50	5.25	**5.92**	8.93

Table 2: Exact Match (EM) results and average score (AVG) on the LLaMA backbone obtained from testing in 12 languages after performing KE on knowledge in English. "ReMaKE-few-bi" means the proposed knowledge editor leveraging few-shot learning based on 16 bilingual examples concatenated in the context. "ReMaKE-few-mono" and "IKE" use 16 monolingual (English) examples in the context. "LLaMA" are the results of pre-editing.

the experiments: LLaMA2-7b and BLOOMZ-7b1-mt, where LLaMA2-7b[2] (Touvron et al., 2023) is a foundation model and BLOOMZ-7b1-mt[3] (Muennighoff et al., 2023) is an instruction-finetuned model. We translate a random sample of 10,000 instances from the zsRE training dataset into the other 11 languages and finetune an XLM-RoBERTa-base[4] (Conneau et al., 2020) on this multilingual dataset to develop our multilingual retriever.

4.4 Implementation Details

All experiments are conducted on a single NVIDIA A-100 GPU (80G). The implementation is based on the EasyEdit (Wang et al., 2023b) framework.

4.5 Baseline

We choose three top-performed KE methods based on the LLMs in Wang et al. (2023a) as our baselines for the experiments. **IKE** (Zheng et al., 2023) uses in-context learning for knowledge editing, where the prompt consists of one explicit piece of knowledge in the editing language, one query in the test language, and a certain number of examples in the editing language (16 in this case, following the setting in Wang et al. (2023a)). We also test a memory-based KE method **SERAC** (Mitchell et al., 2022b)). with a memory size K ($K = 10$ is the default parameter in Mitchell et al. (2022b)). The classifier and counterfactual model in SERAC are pre-trained on the monolingual dataset in the editing language. The parameters of the LLM for both methods mentioned above are frozen. To compare the effect of parameter-updating KE, We evaluate the **ROME** (Meng et al., 2022) method, which locates the knowledge in the editing language first and then edits it. After updating the parameters, we evaluate the performance with a query in the testing language. For all baselines, we use their original proposed default parameters and LLaMA2-7b as the backbone in the experiments. All KE methods are tested on the multilingual knowledge editing.

[2]https://huggingface.co/meta-LLaMA/LLaMA-2-7b-hf
[3]https://huggingface.co/bigscience/bloomz-7b1-mt
[4]https://huggingface.co/xlm-roberta-base

Metrics	Test on EN	Edit on											AVG
		CS	DE	NL	ES	FR	PT	RU	TH	TR	VI	ZH	
Reliability	LLaMA	1.08	1.08	1.08	1.08	1.08	1.08	1.08	1.08	1.08	1.08	1.08	1.08
	ROME	16.96	37.55	35.13	32.84	32.57	31.49	1.35	0.00	3.90	4.85	0.94	17.96
	IKE	57.67	55.45	50.05	40.21	46.38	43.20	52.36	2.03	40.31	41.85	20.54	40.91
	ReMaKE-zero	69.18	65.68	60.97	62.31	66.22	59.76	59.49	9.96	50.47	51.14	44.68	54.53
	ReMaKE-few-mono	62.72	61.37	55.05	45.76	56.8	48.72	60.16	2.83	49.93	50.2	41.86	48.67
	ReMaKE-few-bi	**87.89**	**90.17**	**87.21**	**86.41**	**86.41**	**86.68**	**82.91**	**49.26**	**82.10**	**84.66**	**72.14**	**81.44**
Generality	LLaMA	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	ROME	16.02	35.53	32.30	30.15	30.15	26.78	1.62	0.00	3.90	3.50	0.81	16.43
	IKE	56.41	54.39	49.08	39.25	45.03	42.91	49.47	2.03	39.15	40.89	20.64	39.93
	ReMaKE-zero	63.26	62.05	54.37	53.84	61.10	54.64	53.84	9.29	47.51	46.70	40.38	49.73
	ReMaKE-few-mono	61.1	60.43	53.3	45.09	56.66	48.32	57.6	2.56	48.86	49.66	42.13	47.79
	ReMaKE-few-bi	**87.75**	**88.96**	**86.00**	**84.66**	**84.93**	**85.60**	**82.10**	**48.99**	**80.35**	**84.25**	**69.99**	**80.33**
Locality	**ROME**	84.66	87.89	86.94	88.83	87.89	85.33	82.37	97.04	90.04	93.54	92.73	88.84
	IKE	1.25	1.16	1.16	1.06	1.16	1.25	0.87	0.10	1.16	1.06	0.96	1.02
	ReMaKE-zero	98.92	**99.06**	**99.46**	**98.52**	**98.92**	**98.92**	**98.12**	**97.58**	97.31	**98.79**	**99.33**	**98.63**
	ReMaKE-few-mono	**99.06**	98.52	98.92	98.52	98.65	98.92	98.12	97.04	**97.44**	98.79	99.06	98.46
	ReMaKE-few-bi	98.79	98.38	98.79	98.52	98.79	98.92	97.98	97.17	97.31	98.79	99.19	98.42
Portability	LLaMA	8.48	8.48	8.48	8.48	8.48	8.48	8.48	8.48	8.48	8.48	8.48	8.48
	ROME	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	IKE	5.69	7.43	5.88	5.50	2.89	5.11	7.62	0.10	2.12	4.34	1.06	4.34
	ReMaKE-zero	**25.71**	**27.99**	**26.65**	**25.44**	**24.63**	**26.11**	**20.86**	**11.57**	**22.48**	**24.09**	**19.65**	**23.20**
	ReMaKE-few-mono	19.38	23.42	19.65	18.98	20.59	20.59	16.55	3.36	13.73	14.27	16.29	16.98
	ReMaKE-few-bi	17.50	21.53	19.92	18.71	19.11	19.25	13.19	13.06	17.77	19.65	16.15	17.80

Table 3: EM (Exact Match) results and average score (AVG) on the LLaMA backbone obtained from testing in English after performing KE on knowledge in other 11 languages.

5 Experimental Results

We define a standard notion ("LANGUAGE 1 (edit) → LANGUAGE 2 (test)") to refer to the multilingual KE experiments. For example, "ES (edit) → EN (test)" refers to an experiment in which Spanish is the editing language and English is the language we tested the knowledge, as shown in Figure 2. The knowledge base of the experiments consists of all the above-mentioned multilingual knowledge instances in the MzsRE test set.

5.1 English-based Multilingual KE

In this subsection, we conduct experiments in which English is used as either the editing or testing language, investigating the effects corresponding to all other languages. The evaluation results of LLMs on the LLaMA backbone in 12 languages after editing in English (aka "EN (edit) → ALL (test)") are shown in Table 2 (based on Exact Match (EM)) and Table 9 (based on F1). Experimental results on the LLaMA backbone obtained from "ALL (edit) → EN (test)" are shown in Table 3 and Table 10. We examine ReMaKE with LLaMA backbone in various scenarios, including zero-shot ("ReMaKE-zero"), monolingual few-shot setting ("ReMaKE-few-mono"), and bilingual few-shot settings ("ReMaKE-few-bi"). We compare results of these settings with those from three baseline methods and the pre-editing setup ("LLaMA").

As shown in Table 2, current KE approaches perform reasonably well in the monolingual case (see Reliability for SERAC, IKE, and ROME for "EN (edit) → EN (test)") but do not work in the multilingual setting. "LLaMA" fails with poor pre-editing results (less than 2%) as the knowledge editing test examples are counterfactual. SERAC scores all zeros in the multilingual case except for the Locality metric in response to irrelevant queries. ROME performs similarly poorly in the multilingual setting. The top performer, IKE, shows 100% accuracy for monolingual KE, which is considerably better than other baselines. ReMaKE reveals a significant improvement over IKE in multilingual language conditions. ReMaKE, although fundamentally similar to IKE, provides bilingual few-shot examples and an additional means to filter out irrelevant queries (by returning null knowledge), leading to significant improvements in all four metrics. Furthermore, the scalability and the precision of the editing are also improved by an accurate multilingual retriever.

For Reliability (average accuracy), ReMaKE averagely outperforms baseline IKE by up to +31.19%. Take "EN (edit) → ES (test)" as an example, SERAC has the worst reliability score (0.00%) as the counterfactual model (used to generate predictions about new knowledge) in SERAC is monolingual, and IKE and ROME have reliability scores of 36.45% and 12.25%, respectively. ReMaKE-zero achieves a reliability score of 52.62% instead. When scaled up to a few-shot setting, ReMaKE-few-mono drops to 43.47% due to the negative influence of the monolingual context, but adding bilingual examples to the context makes ReMaKE-few-bi the most capable KE with a reliability score of 68.10%.

With regard to the results of "ALL (edit) → EN (test)", ReMaKE-few-bi achieves the highest scores, outperforming baselines by significant mar-

gins (up to +40.53% and +40.40%) in the average reliability and generality scores[5]. It records a reliability score 86.41% in "ES (edit) → EN (test)".

The discrepancies in ReMaKE scores across languages reflect the disparate capabilities of multilingual LLMs. After editing knowledge expressed in English, ReMaKE-few-bi attains the highest cross-lingual reliability score of 81.70% ("EN (edit) → DE (test)") when testing the LLM in German. The lowest reliability score 32.44% is obtained from ("EN (edit) → TH (test)"), indicating the effect of KE on LLMs is sensitive to language settings. A similar phenomenon can be observed for the same KE method (ReMaKE-few-bi) on a different backbone LLM (i.e., BLOOMZ in Appendix A.3). The sensitivity may be caused by unbalanced distributions of training data across languages. It can be observed that a high-resourced language and a powerful LLM are preferable choices. Even though ReMaKE appears sensitive to language settings and backbone LLMs, it consistently enhances the multilingual KE performances of baselines.

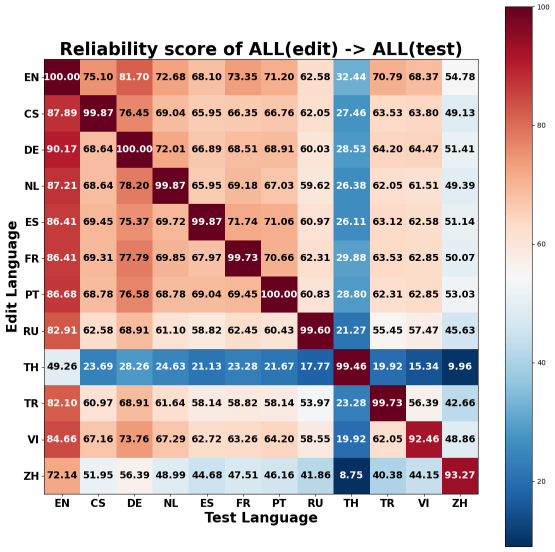


Figure 3: Reliability score of multilingual knowledge editing between all twelve languages.

After knowledge editing, the locality of an LLM can be significantly influenced, as shown in Tables 2-3. The locality is calculated by comparing pre-edit and post-edit predictions to show the degree an LLM is affected by irrelevant input after editing. IKE performs poorly in locality, with most of the EM scores below 1%. It can be observed that ReMaKE can achieve consistently high locality scores across language settings and backbone LLMs due to its robustness against contextual interference. The English classifier implemented in SERAC treats all non-English multilingual queries as irrelevant information, resulting in abnormally high scores in the locality metric while producing a minimal score (0%) in the reliability assessment (Table 2).

It is expected that all KE methods record notably low portability, given their limited capability in impacting LLMs' reasoning capability. Understanding the mechanism responsible for the reasoning capability of an LLM remains a challenge. We notice that ReMaKE-zero outperforms all other KE methods and its few-shot counterpart (ReMaKE-few), possibly due to a reduced level of context interference imposed on an LLM. We leave this hypothesis to a future study.

5.2 Multilingual KE between All Languages

The results of the above-mentioned metrics based on EM are illustrated as heat maps in Figures 3 and 7 (Appendix A.5) for multilingual KE between all twelve languages ("ALL (edit) → ALL (test)"). The discrepancies in the reliability and generality scores between a certain language group (i.e., ZH, TH, and TR) and the rest of the language groups are significant. It appears natural segregation exists between these special languages and the rest of Indo-European languages, probably due to their linguistic characteristics and the language distribution in the training dataset.

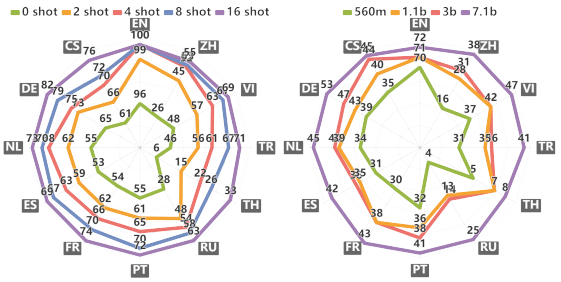
Moreover, the portability scores captured in Figure 7(c) (Appendix A.5) are below 10%, which are much lower than those shown in English-based multilingual KE (Tables 2-3). It is more difficult for multilingual KE methods to impact the reasoning capability of an LLM when English is not in the loop. Whether we should employ English as a pivot for multilingual KE to enhance the portability score, particularly when the editing and testing languages are non-English, requires a future study.

6 Analysis and Discussion

6.1 Ablation Study on Few-shot Editing

We have shown that ReMaKE-few-bi outperforms ReMaKE-zero significantly on the reliability and the generality scores. We conduct ablation studies in this subsection to delve deeper into the impact of two key factors on the performance of multilingual

[5]A counterfactual model is required for each language for SERAC, leading to significant computation overhead. It is not included in this experiment to this end.



(a) Quantity of in-context example.

(b) Backbone model scaling.

Figure 4: Effects of the quantity of bilingual examples (i.e., in a series of 2, 4, 8, 16) and the size of backbone model on multilingual few-shot knowledge editing.

KE in few-shot editing: the quantity of bilingual examples and the size of backbone model.

The results of ReMaKE in response to the quantitative change in bilingual examples are illustrated in Figure 4(a) for "EN (edit) → ALL (test)". The increase in the number of in-context examples is associated with an observable rise in the reliability score. The generality, locality, and portability scores are shown in Figure 8 in Appendix A.6.

We also analyze the effect of the backbone (i.e., BLOOMZ series) with different model sizes on 16-shot editing performance. Figures 4(b) and 12 demonstrate a convincing win of BLOOMZ-7b1 over its siblings with fewer numbers of parameters in all four metrics. Even though it is hard to differentiate the performance of BLOOMZ-1b1 and BLOOMZ-3b in some specific languages, they outperform BLOOMZ-560m in all four metrics. A scale-law pattern is observed between the backbone model size and multilingual KE performance.

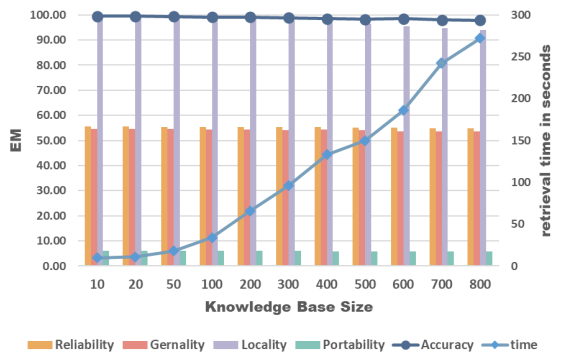


Figure 5: Effects of the size of the knowledge base on a variety of benchmark metrics when performing ReMaKE-few-bi editing on "EN (edit) → ZH (test)" on the LLaMA backbone. The retrieval time consumed is assessed for the entire test set.

6.2 Ablation Study on Size of Knowledge Base

We conduct an ablation study to investigate the effects of the size of the knowledge base on a range of benchmark metrics, including the above-mentioned four metrics, retrieval accuracy and retrieval time consumed. A few-shot KE scenario ("EN (edit) → ZH (test)") is performed on the LLaMA backbone. It can be observed in Figure 5 that all benchmark metrics remain stable except for a minor decrease (-0.81% Reliability) with the increase in the size of the knowledge base. This decrease may be attributed to a minor degradation of the retriever's accuracy (-1.61%) as the knowledge scale increases.

6.3 Computing Cost

Editor	IKE	SERAC	ROME	ReMaKE 0shot	ReMaKE 4shot	ReMaKE 8shot	ReMaKE 16shot
time	0.94s	0.46s	5.92s	**0.70s**	0.85s	1.07s	1.35s

Table 4: Time cost for each knowledge editing method conducting 1 edit on the LLaMA backbone using 1X A100 GPU.

We gather the time consumed in KE on "EN (edit) → ZH (test)" in Table 4 to show the computation cost associated with various editors. It is noted that the proposed ReMaKE-0shot achieves the second-best computation efficiency measured in time. SERAC, the top performer in this test, is not suitable for multilingual KE, making ReMaKE the state-of-the-art editor in terms of computational cost efficiency. ROME requires a substantial time to train and update the parameters through gradient calculation and back-propagation, resulting in the least efficient editor in this category.

7 Conclusion

In this paper, we propose ReMaKE, a retrieval-augmented multilingual knowledge editor, to update multilingual knowledge in LLMs by leveraging prompts composed of retrieved new knowledge and user inputs. To achieve multilingual knowledge editing, we automatically construct the MzsRE dataset to cover 12 languages. ReMaKE is a model and language-agnostic knowledge editor not restricted to a specific LLM and language setting. Our experimental results show that ReMaKE achieves SOTA multilingual knowledge editing performance. We also share the characteristics of multilingual knowledge editing with the community to foster research along this line.

Limitations

As we extend the initial zsRE test set to implement a multilingual knowledge base of the proposed ReMaKE, the volume of the knowledge base is limited to 743 entries. Although ReMaKE can be easily scaled up to cope with real-world applications, the implication of implementing a large-capacity knowledge base on the proposed key metrics warrants a future study. A predefined question-and-answering template is used to define multilingual knowledge contained in the knowledge base. Future work will focus on developing a formal template to accommodate a more comprehensive scope of tasks. We found KE methods, in general, have low portability scores, indicating their limitations to impact knowledge that requires an LLM's reasoning capability. How to perform multilingual knowledge editing to secondary associated knowledge needs further investigation.

References

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. 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. *arXiv preprint arXiv:2308.13198*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022a. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022b. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

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

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2022. Aging with GRACE: lifelong model editing with discrete key-value adaptors. *CoRR*, abs/2211.11031.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. 2023. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*.

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.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics.

Yanchen Liu, Timo Schick, and Hinrich Schtze. 2023. Semantic-oriented unlabeled priming for large-scale language models. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing, SustaiNLP 2023, Toronto, Canada (Hybrid), July 13, 2023*, pages 32–38. Association for Computational Linguistics.

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.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8320–8340. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: retrieval-augmented black-box language models. *CoRR*, abs/2301.12652.

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, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. 2023a. 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. 2023b. Easyedit: An easy-to-use knowledge editing framework for large language models. *CoRR*, abs/2308.07269.

Weixuan Wang, Barry Haddow, Alexandra Birch, and Wei Peng. 2023c. Assessing the reliability of large language model knowledge. *CoRR*, abs/2310.09820.

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, Toronto, Canada, July 9-14, 2023*, pages 5554–5569. Association for Computational Linguistics.

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. *CoRR*, abs/2305.13172.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? A review of recent advances. *CoRR*, abs/2310.07343.

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

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 Appendix

A.1 Statistics of MzsRE

We list the statistics of MzsRE in 12 languages in Table 5.

A.2 Retriever Accuracy

We further investigate the accuracy of the multilingual retriever of ReMaKE using sampled sentence pairs in the MzsRE dataset. The results are captured in Figure 6. The retriever achieves an accuracy of over 94% for all languages. The sub-optimal retrieval accuracy for some languages (i.e., Chinese, Russian) may contribute to the sub-optimal performance of multilingual KE results in these languages.

Lang	Reliability Question	Generality Question	Answer	Locality Question	Locality Answer	Portability Question	Portability Answer
EN	7.89	2.01	7.69	11.11	3.68	12.74	2.87
CS	6.62	1.90	6.58	7.29	3.38	10.76	2.68
DE	7.21	1.86	7.23	8.39	3.56	12.12	2.69
NL	7.55	1.91	7.54	8.83	3.80	12.60	2.75
ES	7.94	2.28	7.87	9.69	4.21	13.19	3.13
FR	9.12	2.17	9.04	9.71	4.11	14.24	3.11
PT	7.98	2.23	7.88	9.27	4.04	12.57	3.04
RU	6.21	2.02	6.18	7.10	3.51	10.10	2.59
TH	31.72	11.06	31.76	32.06	17.82	52.29	14.99
TR	5.58	1.90	5.55	6.65	3.22	8.95	2.62
VI	8.66	2.71	8.63	11.02	4.94	14.98	3.78
ZH	19.46	6.05	19.61	16.90	9.05	27.16	7.05

Table 5: Statistics of sentence length (in word count) of MzsRE. Lang: language, EN: English, CS: Czech, DE: German, NL: Dutch, ES: Spanish, FR: French, PT: Portuguese, RU: Russian, TH: Thai, TR: Turkish, VI: Vietnamese, ZH: Chinese.

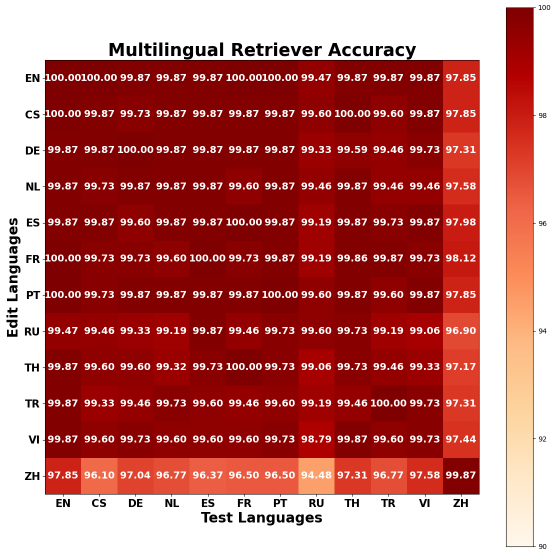


Figure 6: The retrieval accuracy among 12 languages evaluated on the MzsRE dataset reliability subset.

Furthermore, we evaluate the retriever for the different test subsets (reliability, generality, locality, and portability) as shown in Table 6. The results demonstrate that the retrieval accuracy of portability subset is lower than other subsets, which means that the retriever lacks reasoning ability.

A.3 Exact Match Scores of ReMaKE with BLOOMZ Backbone

In order to compare the results of ReMaKE with different base LLMs, we implement ReMaKE on the BLOOMZ backbone, and the exact match scores are show in Table 7 and Table 8.

A.4 Results of English-based multilingual KE with F1

We demonstrate the F1 scores of editing in English and testing in other languages and vise versa on the LLaMA and BLOOMZ backbones in Table 9 and Table 10.

A.5 Supplemental Results Multilingual KE

We supplement the experimental results (Generality, Locality, Portability) of multilingual knowledge editing on ReMaKE-16shot-bi on the LLaMA backbone based on "ALL (edit) → ALL (test)" editing in Figure 7.

A.6 Effects of the Quantity of In-context Examples

We supplement the experimental results (Generality, Locality, Portability) of few-shot learning on ReMaKE-few-bi on the LLaMA backbone in Figure 8. We reach a similar conclusion with the finding obtained in subsection 5.1, in which ReMaKE-zero takes the lead instead in the portability score as few-shot examples tend to introduce contextual interference to the KE process.

We supplement the comparison results of ReMaKE-few-bi on the BLOOMZ backbone under the 0-shot, 2-shot, 4-shot, 8-shot, 16-shot settings with editing in English and testing in other languages in Figure 9. Also we conduct the same setting with editing in other languages and testing in English in Figure 10 and Figure 11. From the results, it proves that few-shot learning could greatly improve the performance compared to zero-shot for the reliability and generality.

A.7 Supplemental Results of Model Size

We supplement the experimental results (Generality, Locality, Portability) of different model sizes on ReMaKE-16shot-bi on the BLOOMZ backbone in Figure 12 with editing in English and testing in other languages.

Subsets	EN	CS	DE	NL	ES	FR	PT	RU	TH	TR	VI	ZH
Reliability	100	100	99.86	99.87	99.87	100	100	99.46	99.87	99.87	99.87	97.85
Generality	99.87	99.19	99.33	99.33	99.73	99.46	99.73	98.38	99.33	99.06	99.06	96.1
Locality	100	100	100	100	100	100	100	100	100	100	100	100
Portability	91.79	88.16	89.23	89.5	89.64	89.77	89.23	81.43	85.6	89.23	89.23	84.25

Table 6: Retrieval accuracy for different test subsets. We evaluate retriever for the reliability, generality, locality, portability test part in MzsRE for editing in English and testing in other languages.

Metrics	Edit on EN	Test on											
		EN	CS	DE	NL	ES	FR	PT	RU	TH	TR	VI	ZH
Reliability	BLOOMZ	1.88	0.13	0.40	0.13	0.40	1.21	0.81	0.00	0.00	0.13	0.40	1.75
	ReMaKE-zero	69.04	29.21	34.59	28.26	25.03	27.59	25.98	0.13	4.44	21.53	28.80	18.98
	ReMaKE-few-bi	71.20	44.55	52.76	44.68	41.59	42.53	40.65	24.50	7.54	40.65	46.30	37.28
Generality	BLOOMZ	1.35	0.13	0.27	0.13	0.27	0.81	0.67	0.00	0.00	0.13	0.27	1.88
	ReMaKE-zero	63.26	28.67	33.24	27.59	24.63	26.65	25.30	0.13	4.71	21.27	27.05	17.50
	ReMaKE-few-bi	65.81	43.47	51.14	43.07	39.70	41.32	39.84	23.28	7.13	38.63	44.01	35.94
Locality	**ReMaKE**-zero	99.19	98.25	99.60	99.73	97.85	98.92	99.19	97.44	95.29	97.04	97.44	94.62
	ReMaKE-few-bi	99.19	98.25	99.60	99.73	97.85	99.06	98.92	97.44	95.29	97.04	97.44	94.62
Portability	BLOOMZ	6.59	0.13	1.35	0.13	2.29	2.15	2.15	0.00	0.00	0.00	2.29	4.58
	ReMaKE-zero	12.65	0.40	2.29	0.94	4.44	4.98	4.71	0.00	0.13	0.40	3.77	**7.67**
	ReMaKE-few-bi	7.81	0.54	1.62	1.08	4.71	4.04	4.04	0.54	0.13	0.54	2.83	6.86

Table 7: Exact Match (EM) results on the BLOOMZ backbone obtained from testing in 12 languages after performing KE on knowledge in English. "ReMaKE-few-bi" means the proposed knowledge editing method leveraging few-shot learning based on 16 bilingual examples concatenated in the context. "BLOOMZ" are the results of pre-editing.

Metrics	Test on EN	Edit on										
		CS	DE	NL	ES	FR	PT	RU	TH	TR	VI	ZH
Reliability	BLOOMZ	1.88	1.88	1.88	1.88	1.88	1.88	1.88	1.88	1.88	1.88	1.88
	ReMaKE-zero	34.05	39.43	30.55	33.65	34.05	32.71	23.82	5.38	27.86	36.47	23.55
	ReMaKE-few-bi	48.32	55.99	49.53	52.89	48.86	53.43	36.74	14.00	46.16	54.91	45.76
Generality	BLOOMZ	1.35	1.35	1.35	1.35	1.35	1.35	1.35	1.35	1.35	1.35	1.35
	ReMaKE-zero	32.44	36.34	28.80	32.57	32.71	31.36	21.27	5.11	25.84	33.38	21.27
	ReMaKE-few-bi	46.70	54.91	48.99	51.68	48.18	51.95	34.86	13.73	44.82	52.49	42.93
Locality	**ReMaKE**-zero	98.52	98.38	98.52	98.38	98.38	98.52	97.71	97.17	96.64	98.52	98.92
	ReMaKE-few-bi	98.52	98.38	98.52	98.38	98.52	98.52	97.71	97.31	96.77	98.52	98.92
Portability	BLOOMZ	6.59	6.59	6.59	6.59	6.59	6.59	6.59	6.59	6.59	6.59	6.59
	ReMaKE-zero	9.02	9.69	9.69	9.96	10.90	11.57	7.67	6.19	7.54	9.29	8.88
	ReMaKE-few-bi	5.79	7.27	6.33	6.19	7.81	6.73	5.38	4.98	6.33	7.40	5.38

Table 8: Exact Match(EM) results on the BLOOMZ backbone obtained from testing in English after performing KE on knowledge in other languages. "ReMaKE-few-bi" means the proposed knowledge editing method leveraging few-shot learning based on 16 bilingual examples concatenated in the context. "BLOOMZ" are the results of pre-editing.

Metrics	Edit on EN	Test on											
		EN	CS	DE	NL	ES	FR	PT	RU	TH	TR	VI	ZH
Reliability	**SERAC**	96.25	19.38	18.08	18.32	16.61	17.78	17.33	19.91	4.01	15.59	12.68	10.01
	IKE	100.0	74.62	74.16	70.91	64.55	70.24	65.16	55.83	43.89	65.18	73.60	42.25
	ROME	83.80	37.00	43.56	42.47	36.99	39.15	38.40	18.10	2.72	23.25	22.59	11.18
	ReMaKE-zero-BLOOMZ	89.69	49.41	57.23	50.17	52.18	56.39	52.50	19.91	23.01	43.75	58.17	55.64
	ReMaKE-zero-LLaMA	98.05	79.91	82.43	75.81	71.99	75.13	74.13	67.37	48.91	69.23	76.28	68.53
	ReMaKE-few-BLOOMZ	91.43	68.58	74.47	67.88	69.11	71.44	70.71	49.47	46.32	66.92	72.60	70.08
	ReMaKE-few-LLaMA	**100.0**	**87.61**	**90.45**	**85.99**	**82.86**	**86.99**	**85.25**	**84.28**	**69.77**	**84.38**	**86.42**	**80.23**
Generality	**SERAC**	54.25	19.14	18.28	18.52	16.69	17.27	17.30	19.61	3.91	15.54	12.66	10.33
	IKE	99.10	73.85	73.94	70.42	63.81	69.62	64.62	55.11	44.21	64.63	73.32	42.11
	ROME	68.91	36.35	41.83	40.73	36.98	37.67	35.97	17.82	2.97	23.52	22.56	10.66
	ReMaKE-zero-BLOOMZ	85.02	48.83	56.05	49.42	51.28	55.28	51.53	19.77	23.30	43.13	56.00	53.82
	ReMaKE-zero-LLaMA	92.48	78.03	80.38	74.37	70.55	72.71	71.84	65.04	49.23	67.22	75.14	66.36
	ReMaKE-few-BLOOMZ	87.16	67.72	73.21	66.46	67.72	70.41	68.96	48.62	46.06	66.06	71.10	68.63
	ReMaKE-few-LLaMA	**99.07**	**87.02**	**89.77**	**85.24**	**82.67**	**85.98**	**84.91**	**83.74**	**69.75**	**83.83**	**86.00**	**79.44**
Locality	**SERAC**	99.80	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.98
	IKE	67.50	32.71	38.60	33.96	34.41	32.94	33.26	34.88	53.54	34.04	41.32	45.65
	ROME	97.83	95.46	95.99	95.94	97.20	96.32	96.20	95.72	97.53	95.57	97.91	97.80
	ReMaKE-zero-BLOOMZ	99.50	98.46	99.63	99.82	98.55	99.34	99.40	97.91	97.38	97.43	98.14	96.53
	ReMaKE-zero-LLaMA	99.76	99.15	99.84	99.94	99.03	99.51	99.50	98.49	97.79	98.26	98.67	97.17
	ReMaKE-few-BLOOMZ	99.39	98.48	99.63	99.78	98.51	99.46	99.24	97.68	97.34	97.37	98.05	96.43
	ReMaKE-few-LLaMA	99.76	98.97	99.71	99.80	98.71	99.47	99.47	98.14	96.73	98.01	98.49	96.60
Portability	**SERAC**	10.06	2.52	4.65	4.82	4.44	4.78	6.11	4.31	0.74	1.02	0.47	0.67
	IKE	51.96	35.51	38.48	36.57	34.74	37.87	37.23	39.55	30.60	28.44	44.83	23.83
	ROME	9.28	3.10	5.61	4.73	4.46	5.02	5.73	4.32	0.75	1.13	0.61	0.73
	ReMaKE-zero-BLOOMZ	44.06	12.83	20.21	14.22	30.23	32.77	28.65	6.14	17.19	13.65	32.07	43.19
	ReMaKE-zero-LLaMA	**64.07**	**45.38**	**49.08**	**45.42**	**44.25**	**45.90**	**45.17**	44.39	32.14	**34.18**	**51.41**	47.75
	ReMaKE-few-BLOOMZ	37.63	12.73	19.55	14.70	29.92	30.45	28.74	8.44	20.47	14.50	30.21	42.47
	ReMaKE-few-LLaMA	62.27	42.89	44.03	41.84	41.70	43.41	42.71	**44.64**	**33.45**	33.37	47.00	**50.23**

Table 9: F1 results obtained from testing in 12 languages after performing KE on knowledge in English. "ReMaKE-few" means the proposed knowledge editing method leveraging few-shot learning based on 16 bilingual examples concatenated in the context.

Metrics	Test on EN	Edit on										
		CS	DE	NL	ES	FR	PT	RU	TH	TR	VI	ZH
Reliability	**IKE**	76.58	75.06	72.94	64.87	67.62	66.74	72.03	4.27	64.38	60.42	42.94
	ROME	50.18	66.81	66.78	61.67	64.70	65.76	26.88	4.36	37.05	27.93	13.00
	ReMaKE-zero-BLOOMZ	60.77	65.55	58.36	60.26	60.77	59.46	46.07	23.23	54.48	61.93	49.76
	ReMaKE-zero-LLaMA	83.71	81.11	78.45	78.63	81.67	77.57	76.91	33.64	70.14	71.41	67.56
	ReMaKE-few-BLOOMZ	73.49	78.54	74.50	76.67	75.91	77.24	61.49	35.62	71.65	77.49	71.07
	ReMaKE-few-LLaMA	**93.76**	**94.59**	**92.83**	**92.38**	**92.53**	**92.78**	**90.52**	**67.22**	**89.85**	**91.14**	**84.15**
Generality	**IKE**	75.19	74.38	71.77	63.82	66.57	65.60	70.00	4.38	63.24	59.53	43.34
	ROME	48.80	65.89	65.62	60.11	62.19	61.57	26.93	5.03	36.75	27.49	12.74
	ReMaKE-zero-BLOOMZ	59.11	63.78	56.71	58.95	59.88	58.18	44.22	22.51	52.40	60.16	47.51
	ReMaKE-zero-LLaMA	79.30	78.57	74.19	73.05	78.05	73.60	72.87	32.44	67.90	67.69	63.53
	ReMaKE-few-BLOOMZ	72.21	78.17	73.86	75.71	74.72	75.78	60.04	33.86	69.98	75.84	68.80
	ReMaKE-few-LLaMA	**93.44**	**94.00**	**92.22**	**91.23**	**91.61**	**91.70**	**89.62**	**66.81**	**88.49**	**90.70**	**82.52**
Locality	**IKE**	36.39	36.35	36.18	35.72	35.37	36.70	37.69	3.46	35.49	33.59	36.13
	ROME	95.47	96.35	96.06	97.20	96.16	95.60	95.19	97.65	96.30	97.66	97.71
	ReMaKE-zero-BLOOMZ	98.78	98.93	98.91	98.93	98.69	99.01	98.16	98.00	97.57	98.87	99.28
	ReMaKE-zero-LLaMA	99.30	99.51	99.69	99.36	99.48	99.46	98.82	98.54	98.60	99.42	99.59
	ReMaKE-few-BLOOMZ	98.80	98.98	98.96	99.00	98.77	98.98	98.09	98.05	97.78	98.82	99.15
	ReMaKE-few-LLaMA	99.09	99.25	99.34	99.35	99.18	99.31	98.78	98.26	98.44	99.42	99.46
Portability	**IKE**	41.42	43.34	41.74	41.90	38.64	41.45	42.30	2.26	36.81	36.67	32.50
	ROME	3.23	5.80	4.72	4.46	5.03	5.72	4.26	0.77	1.46	0.57	0.85
	ReMaKE-zero-BLOOMZ	38.42	39.80	39.52	40.49	41.01	41.40	34.56	31.81	37.30	39.35	38.04
	ReMaKE-zero LLaMA	**57.57**	**59.04**	**57.45**	**57.01**	**56.89**	**57.10**	**52.87**	41.94	**54.61**	**55.28**	49.87
	ReMaKE-few-BLOOMZ	34.30	35.97	34.83	35.21	36.70	35.65	31.39	31.00	34.71	34.93	34.34
	ReMaKE-few-LLaMA	52.39	55.25	54.32	53.68	54.03	53.75	48.41	**44.70**	51.70	53.67	**49.96**

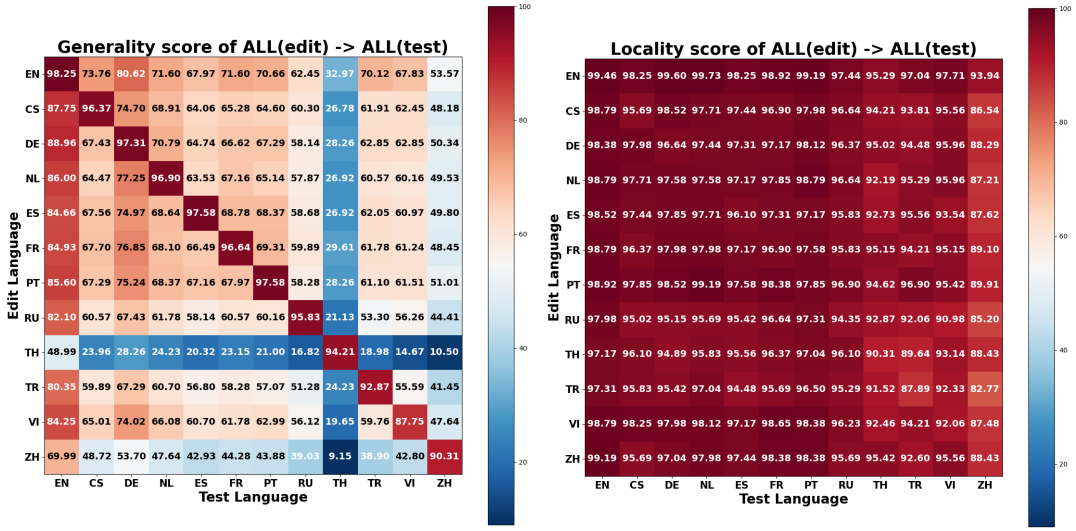
Table 10: F1 results obtained from testing testing in English after performing KE on knowledge in other languages.

A.8 In-context Examples Selection Method

Liu et al. (2022) has demonstrated that the search-based examples selection approach consistently outperforms the random selection baseline. All the above few-shot experimental results are conducted with the unsupervised prompt searching method. We compare the results of random selection and search-based strategy for examples in Table 11. It follows the conclusion of Liu et al. (2022) that search-based selection could increase the performance (reliability score), such as from 41.45% to 67.97% on "EN(edit) → ES (test)".

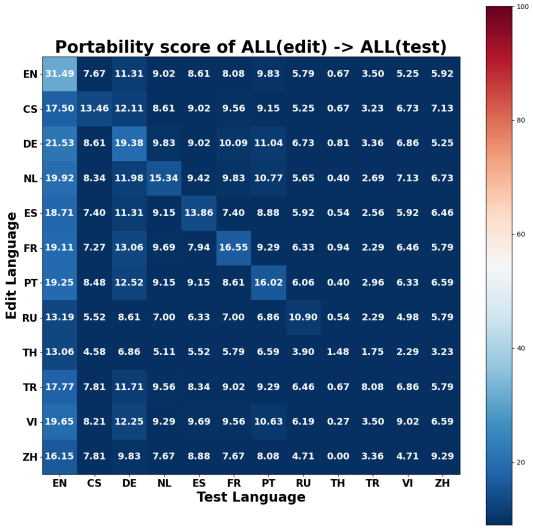
Edit on EN	Test on											
	EN	CS	DE	NL	ES	FR	PT	RU	TH	TR	VI	ZH
ReMaKE-random-BLOOMZ	41.32	35.67	37.15	30.28	32.17	36.47	36.34	3.63	4.98	31.22	35.4	26.78
ReMaKE-search-BLOOMZ	70.93	44.41	52.62	44.55	41.45	42.40	40.51	23.82	7.40	39.97	45.9	37.01
ReMaKE-random-LLaMA	54.64	61.91	79.04	59.89	55.85	61.91	61.24	43.88	8.48	52.22	55.59	34.05
ReMaKE-search-LLaMA	**99.33**	**75.10**	**81.16**	**72.54**	**67.97**	**73.08**	**71.06**	**61.78**	**32.17**	**69.99**	**67.97**	**53.70**

Table 11: The reliability scores base on EM comparison of ReMaKE-16shot-bi between selected examples with an unsupervised method (ReMaKE-search) and random examples (ReMaKE-random) in "EN (edit) → ALL (test)" editing.



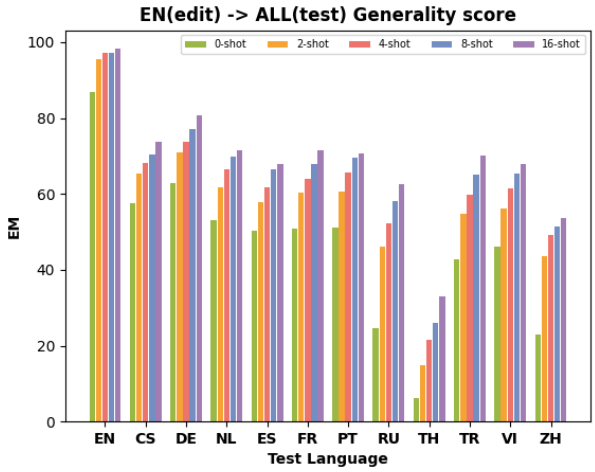
(a) Generality score of multilingual knowledge editing between all twelve languages.

(b) Locality score of multilingual knowledge editing between all twelve languages.

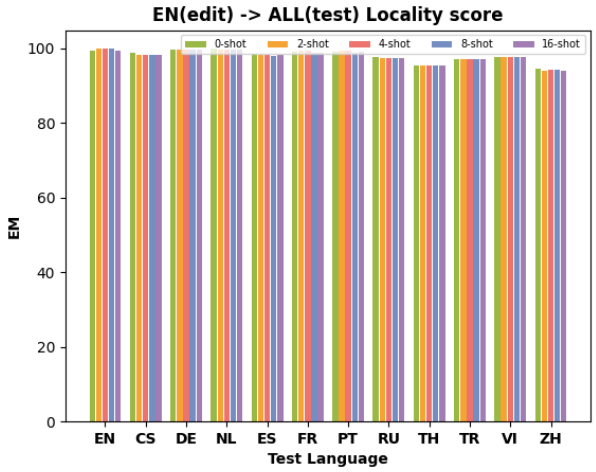


(c) Portability score of multilingual knowledge editing between all twelve languages.

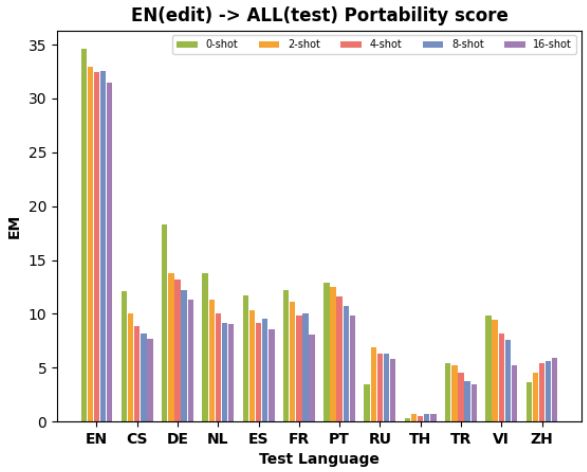
Figure 7: Metrics based on "ALL (edit) → ALL (test)" editing, where "ALL" represents 12 languages.



(a) Generality score of ReMaKE-few-bi with different quantities of examples.

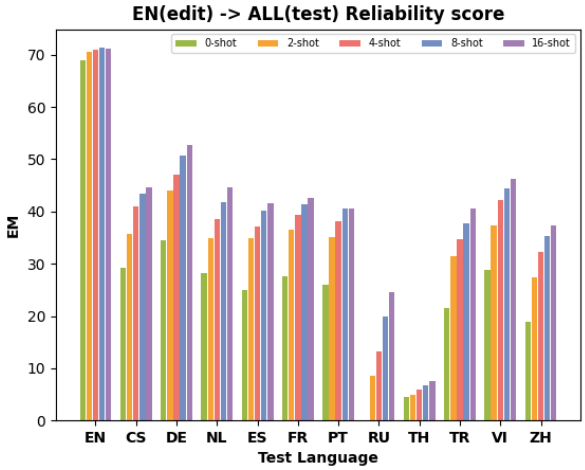


(b) Locality score of ReMaKE-few-bi with different quantities of examples.

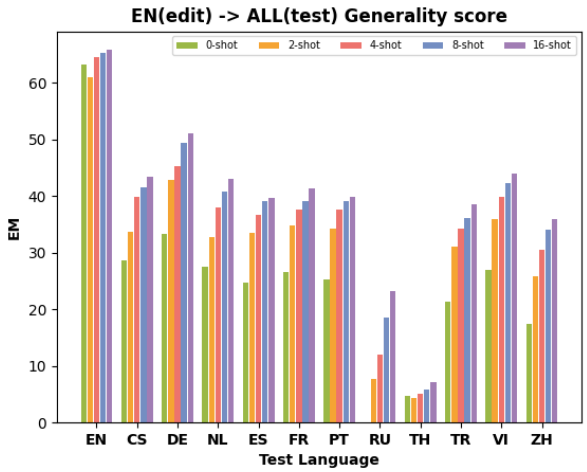


(c) Portability score of ReMaKE-few-bi with different quantities of examples.

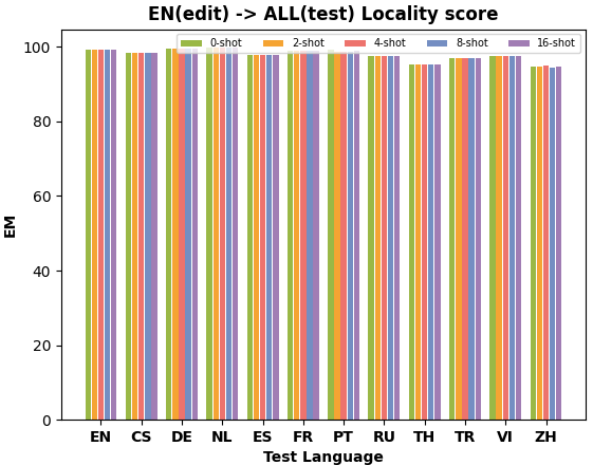
Figure 8: Effects of the quantity of bilingual in-context examples on the LLaMA backbone with editing in English and testing in 12 languages.



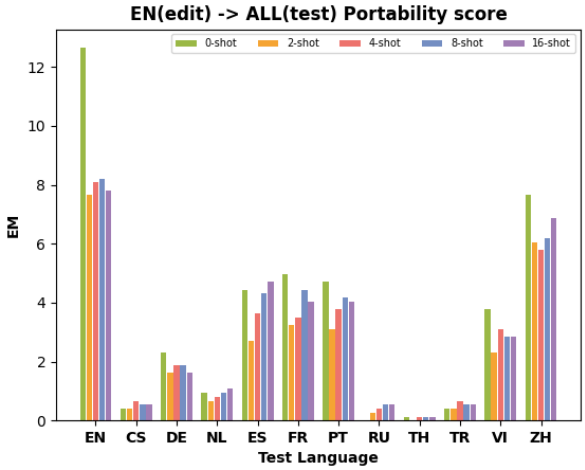
(a) Reliability score of ReMaKE-few-bi with different quantities of examples.



(b) Generality score of ReMaKE-few-bi with different quantities of examples.

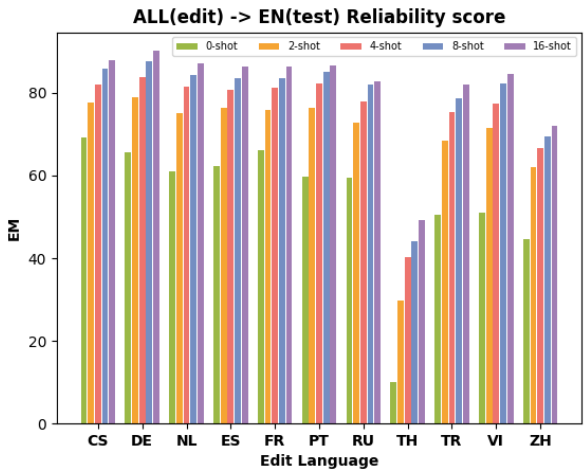


(c) Locality score of ReMaKE-few-bi with different quantities of examples.

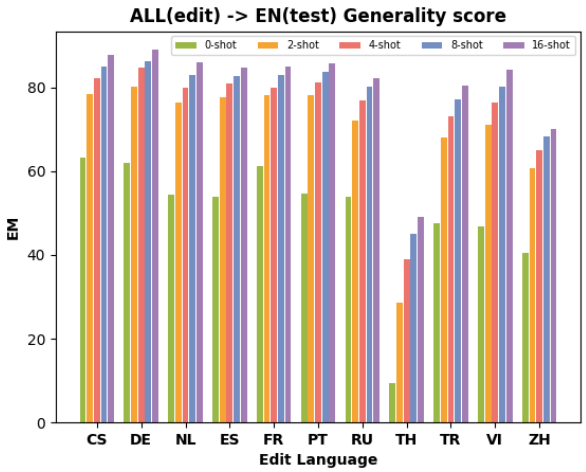


(d) Portability score of ReMaKE-few-bi with different quantities of examples.

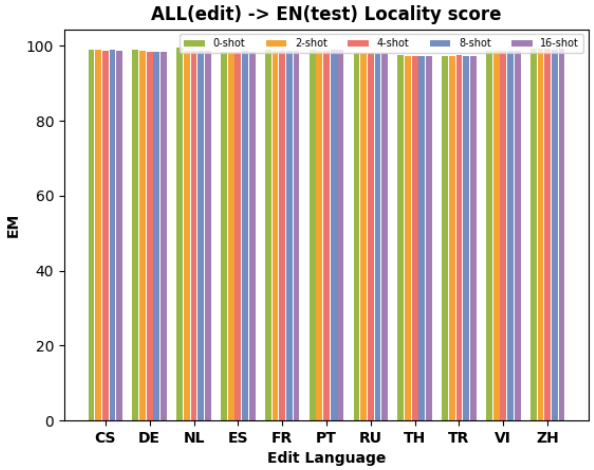
Figure 9: Effects of the quantity of bilingual demonstrations on the BLOOMZ backbone with editing in English and testing in 12 languages.



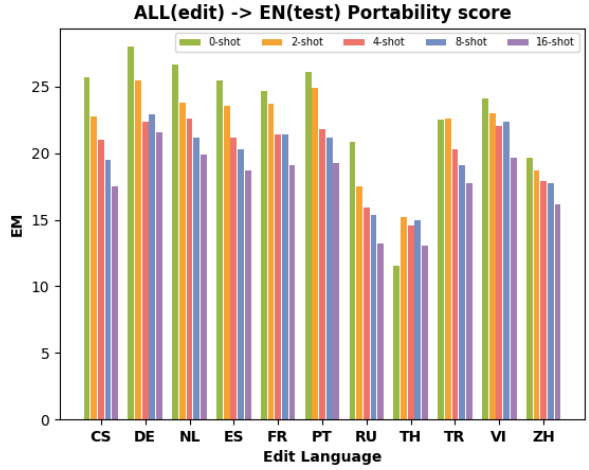
(a) Reliability score of ReMaKE-few-bi with different quantities of examples.



(b) Generality score of ReMaKE-few-bi with different quantities of examples.

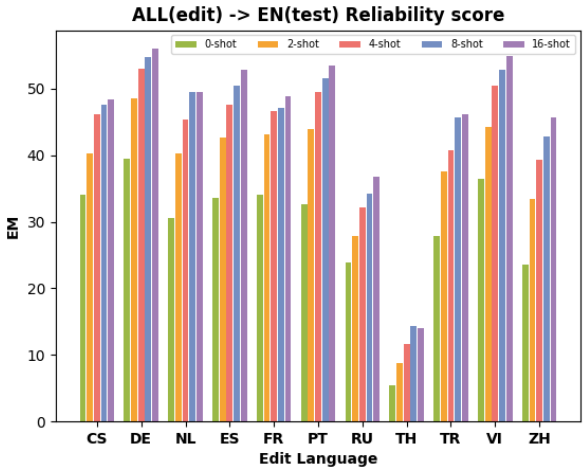


(c) Locality score of ReMaKE-few-bi with different quantities of examples.

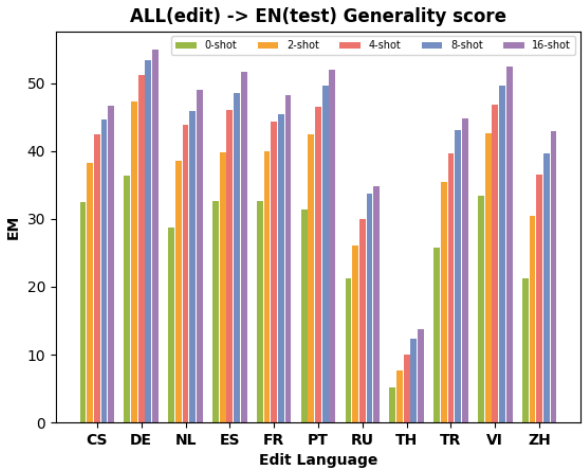


(d) Portability score of ReMaKE-few-bi with different quantities of examples.

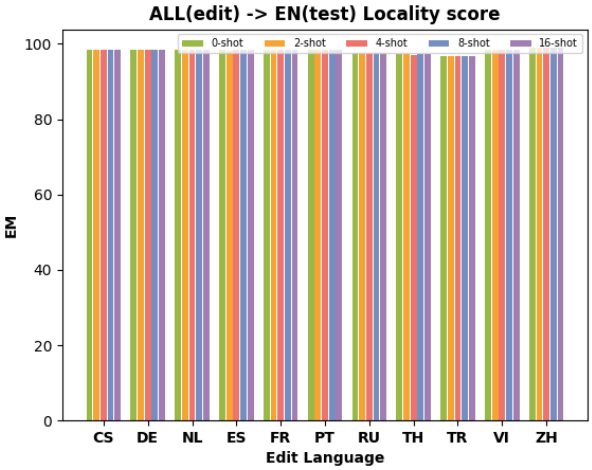
Figure 10: Effects of the quantity of bilingual demonstrations on the LLaMA backbone with editing in other languages and testing in English.



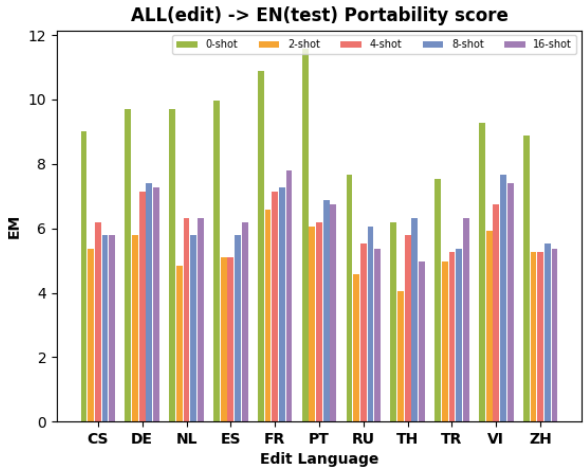
(a) Reliability score of ReMaKE-few-bi with different quantities of examples.



(b) Generality score of ReMaKE-few-bi with different quantities of examples.

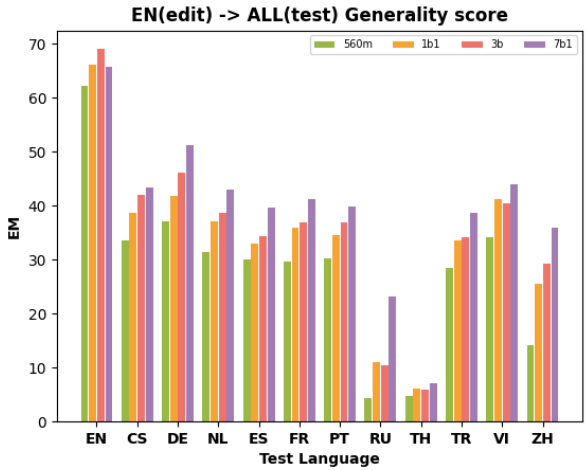


(c) Locality score of ReMaKE-few-bi with different quantities of examples.

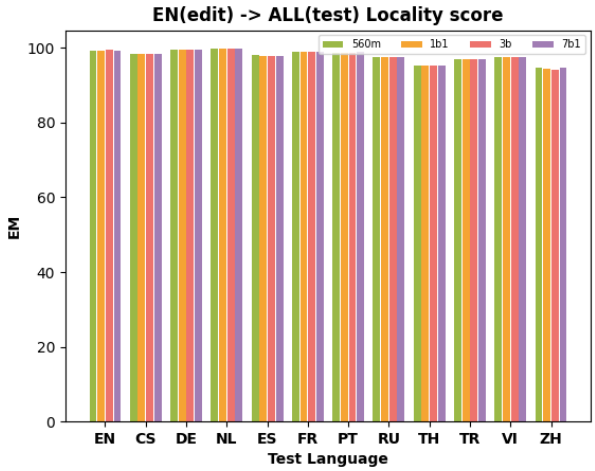


(d) Portability score of ReMaKE-few-bi with different quantities of examples.

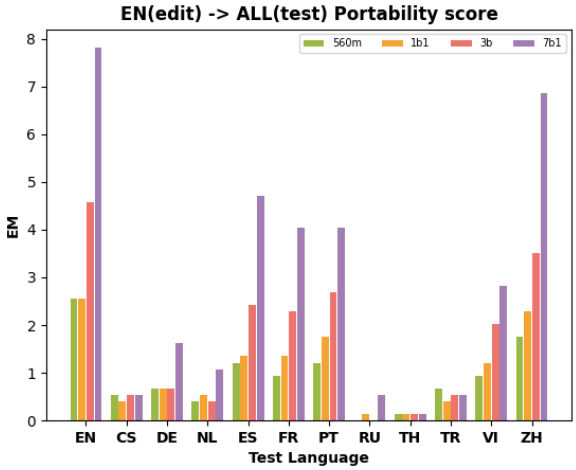
Figure 11: Effects of the quantity of bilingual demonstrations on the BLOOMZ backbone with editing in other languages and testing in English.



(a) Generality score of ReMaKE-few-bi with different model sizes.



(b) Locality score of ReMaKE-few-bi with different model sizes.



(c) Portability score of ReMaKE-few-bi with different model sizes.

Figure 12: Effects of the model sizes with the ReMaKE-16shot-bi on the BLOOMZ backbone with editing in English and testing in 12 languages.