Cross-Lingual Multi-Hop Knowledge Editing

Aditi Khandelwal[*1,2], **Harman Singh**[*†], **Hengrui Gu**[4], **Tianlong Chen**[3,5], **Kaixiong Zhou**[4, 5]
[1]MILA - Quebec AI Institute, [2]Microsoft, [3]UNC Chapel Hill,
[4]North Carolina State University, [5]Massachusetts Institute of Technology

Abstract

Large language models are often expected to constantly adapt to new sources of knowledge and knowledge editing techniques aim to efficiently patch the outdated model knowledge, with minimal modification. Most prior works focus on monolingual knowledge editing in English, even though new information can emerge in any language from any part of the world. We propose the *Cross-Lingual Multi-Hop Knowledge Editing* paradigm, for measuring and analyzing the performance of various SoTA knowledge editing techniques in a cross-lingual setup. Specifically, we create a parallel cross-lingual benchmark, CROLIN-MQUAKE for measuring the knowledge editing capabilities. Our extensive analysis over various knowledge editing techniques uncover significant gaps in performance between the cross-lingual and English-centric setting. Following this, we propose a significantly improved system for cross-lingual multi-hop knowledge editing, CLEVER-CKE. CLEVER-CKE is based on a retrieve, verify and generate knowledge editing framework, where a retriever is formulated to recall edited facts and support an LLM to adhere to knowledge edits. We develop language-aware and hard-negative based contrastive objectives for improving the cross-lingual and fine-grained fact retrieval and verification process used in this framework. Extensive experiments on three LLMs, eight languages, and two datasets show CLEVER-CKE's significant gains of up to 30% over prior methods. Code and data are released at https://github.com/HarmanDotpy/CroLin-KE

1 Introduction

Large language models (LLMs) are seeing an increasing adoption across users having different cultural and linguistic background, and need to be up to date about the ever-changing knowledge in the world for maintaining their utility and reliability in various applications. Due to the ever increasing compute and data requirements to train these models, there has been a surge in the development of knowledge editing techniques to modify the language models in an efficient way, such that they adhere to the world dynamics.

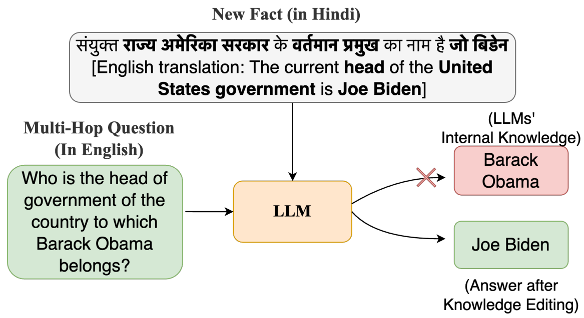


Figure 1: The Cross-lingual Multi-hop knowledge editing problem. New fact(s) are provided in different languages (e.g. Hindi). An LLM should adapt to these facts for answering complex, multi-hop questions correctly in different languages (e.g. English).

Prior work on knowledge editing has largely focused on editing LLMs in a monolingual setting (Zhong et al., 2023; Gu et al., 2024a), where both user queries and edited facts are expressed in the form of English. These works can be grouped into two categories: parameter-update and parameter-preserving methods. The former directly updates the parameters within LLMs for updating knowledge about the edited facts through meta-learning, fine-tuning, or knowledge locating (De Cao et al., 2021; Dai et al., 2022; Mitchell et al., 2022a; Meng et al., 2022a,b). The later approach freezes the parameters and explicitly stores the edited facts in an external memory and retrieves them for answering user queries (Zhong et al., 2023; Gu et al., 2024a; Mitchell et al., 2022c; Hartvigsen et al., 2023). Existing monolingual knowledge editing techniques aren't broadly applicable since new knowledge can emerge in different languages. Some works have

[*]Equal Contribution (Randomly ordered via coin flip)
[†]Now at Google DeepMind

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 11995–12015

made progress in this direction (Beniwal et al., 2024; Xu et al., 2023a; Si et al., 2024), but they have considered a simplistic setting of assuming the edited facts as independent without any multi-hop rippling consequences on entailed reasoning process, and are primarily focused on parameter-modifying based editing methods.

There has only been a limited focus on the realistic case of *cross-lingual multi-hop knowledge editing* (see Fig 1), where the edited knowledge can come in through users who communicate in different languages. Further, much of edited knowledge often has a rippling effect on other facts of the world. For example, the *club change of Messi* affects deduction process of question "indicating a superficial word matching rather than a contextual grasp of the entities involved." This knowledge editing setting, which we argue is important to study, is challenging since the model needs to transfer knowledge about fact edits between different languages, while also reasoning about the facts which are modified as a consequence to the given edit. Poor knowledge transfer between languages can lead to error propagation across reasoning steps which can increase failure cases of model editing.

In this work, we formulate the notion of cross-lingual multi-hop knowledge-editing and analyze existing approaches for their editing ability in different languages, following which a simple yet highly effective approach is designed. Specifically, ① We create one of the first benchmark datasets for measuring cross-lingual multi-hop knowledge editing capabilities of knowledge editing methods. Besides parameter-update based approaches, we contribute strong retrieval-based baselines for knowledge editing and provide a comprehensive analysis. ② We provide a detailed analysis and find significant gaps in the performance of methods for cross-lingual knowledge editing. The gaps are mainly due to challenges in accurately recalling fact edits made in language other than input query.
③ To bridge such gap, we design a competitive method, termed as Contrastive Language-aware Verification for Cross-lingual Knowledge Editing (CLeVer-CKE), for improving performance of cross-lingual multi-hop knowledge editing. Our approach is based on decomposing a multi-hop question in a particular language into sub-questions and retrieving fact edits (if any) from memory using a cross-lingual retriever, which is integrated for answering sub-questions. In particular, the cross-lingual retriever is regularized by novel language-guided and hard-negative based contrastive losses, which leads to improved language and fine-grained sentence understanding of the edits, leading to high quality cross-lingual retrievals. CLeVer-CKE improves over previous SoTA by up-to 30% increase in knowledge editing accuracy when tested on multiple LLMs, datasets and languages.

2 Cross-lingual Multi-hop Editing

Following prior work (Zhong et al., 2023), a fact is defined as a triplet (s, r, o), where s is the subject, o is the object, and r is the relation (e.g., *Shakespeare, author of, Hamlet*). Given that a parametric LLM can become outdated or incorrect, knowledge editing is required to be performed on it. An edited fact stores information about updated knowledge of an existing fact and is denoted as $e = (s, r, o^*)$, where the object is replaced with a new one $o*$.

Cross-Lingual Knowledge Editing. Each knowledge fact or edit is assumed to be represented in natural language. Let $\mathcal{T} : \mathcal{E} \rightarrow \mathcal{L}$ be a function which takes any fact $e \in \mathcal{E}$ (e.g., *Shakespeare, author of, Hamlet*) and converts it into a natural language statement, (e.g., *Shakespeare is the author of Hamlet*). All the facts and edits can be represented in a variety of languages $\{L_1, L_2, \dots\}$ via functions such as $\{\mathcal{T}_{L_1}, \mathcal{T}_{L_2}, \dots\}$. For example, an edit $e =$(Shakespeare, author of, Lolita) can be written as $\mathcal{T}_{\mathrm{de}}(e) =$ *Shakespeare ist der Autor von Lolita* in German and $\mathcal{T}_{\mathrm{en}}(e) =$ *Shakespeare is the author of Lolita* in English.

We consider a collection of n fact edits in the diverse languages: $\mathcal{E} = \{e_1^{L_1}, e_2^{L_2}, e_3^{L_2}, ..., e_n^{L_i}\}$, where $L_1, L_2, ..., L_i$ are different languages for e.g., German, Hindi, Swahili, etc. A language model f is said to be edited with new knowledge facts if the model generations adheres to all the edits present in $\mathcal{E}$. The model is required to seamlessly transfer knowledge about an edit in one language to answer queries in other languages.

Multi-Hop Editing and Evaluation. We follow Zhong et al. (2023) for evaluating knowledge editing via multi-hop question answering. Consider $e_{L_1} = (s_i^{L_1}, r_i^{L_1}, o_i^{L_1*})$, an edited fact in language L_1. Also consider a chain of facts $\mathcal{P} = \langle (s_1^{L_1}, r_1^{L_1}, o_1^{L_1}), \dots, (s_n^{L_k}, r_n^{L_k}, o_n^{L_k}) \rangle$, where object of a fact is the subject for the next fact. Any edit to the first fact $(s_1^{L_1}, r_1^{L_1}, o_1^{L_1*})$ will likely have a rippling effect and change the subsequent facts in the chain, and we expect a successfully edited model to be aware of all such entailed changes.

For evaluating models in a cross-lingual multi-hop setting, we make use of multi-hop questions such as Q_{L_n}, in language L_n which is different from $L_{1...k}$. The question asks about the head entity $s_1^{L_1}$ for which the answer is $o_n^{L_k}$ before editing. After editing, the fact chain changes to $\mathcal{P}^* = \langle (s_1^{L_1}, r_1^{L_1}, o_1^{L_1 *}), (s_2^{L_2}, r_2^{L_2}, o_2^{L_2 *}), \ldots, (s_n^{L_k}, r_n^{L_k}, o_n^{L_k *}) \rangle$ since edits in the first fact can affect the subsequent facts it's linked to. For answering Q_{L_n} after editing, the model has to account for this rippling effect, and provide the final answer as $o_n^{L_k *}$. For this, model has to transfer knowledge of the edited fact and the answer, between languages $L_{1...k}$ and L_n, while correctly reasoning about fact edits via $\mathcal{P}^*$.

3 CroLin-MQuAKE Benchmark

We develop one of the first parallel cross-lingual and multi-hop benchmarks for measuring the knowledge editing capabilities of the existing approaches. A parallel benchmark across languages has the same test examples across all the languages, enabling a direct comparison between them. For this, we use existing datasets measuring the multi-hop model editing in English: MQuAKE-CF and MQuAKE-T released by Zhong et al. (2023), which have counterfactual edits and real-world temporal edits respectively. We translate one fact edit per example in these datasets using Google Translate (Google) into 7 languages with diverse writing scripts across medium to high resourcedness - German, Spanish, Chinese, Russian, Hindi, Bengali, Swahili. This results in the benchmark: Cross-Lingual Multi-Hop QnA for Knowledge Editing (CroLin-MQuAKE). It has two datasets, CroLin-MQuAKE-CF and CroLin-MQuAKE-T, each having 8 languages, and 3k and 1.8k parallel examples (same examples in all languages) per language, respectively. The translations are verified by human experts proficient in particular languages and evaluation of BLEU score (Papineni et al., 2002) using backtranslation. We find that the translation is highly accurate, since we study medium to high resource languages. See Section A.1 for more details.

Concurrently, Wei et al. (2024) created a multilingual knowledge editing dataset using Wikipedia, offering translocalized knowledge but lacking parallel multilingual examples like ours. CroLin-MQuAKE enables comparing the knowledge editing performance difference across languages directly without being affected by the variation of test sets between different languages.

4 Benchmark Analysis on Cross-Lingual Multi-hop Knowledge Editing

LLMs. We use SoTA propriety and open-source LLMs: ChatGPT (Schulman et al., 2022), LLaMa-2-7B (Touvron et al., 2023b), Vicuna-1.5-7B (Chiang et al., 2023) as backbones to evaluate cross-lingual multi-hop knowledge editing.

Evaluation Metrics. We use multi-hop accuracy proposed by Zhong et al. (2023) which measures the accuracy of the final answer of a multi-hop question. We also adopt hop-wise answering accuracy for checking the correctness of intermediate reasoning steps, as proposed by Gu et al. (2024a).

New Baselines. Based on existing work, we contribute strong baselines for the new editing setup:

- **MeLLo-CL:** We modify the existing method of MeLLo (Zhong et al., 2023) by replacing the monolingual retriever used in their system with a multilingual retriever. This minimal modification allows the system to retrieve the cross-lingual edits. MeLLo-CL is a simple retrieval-based knowledge editing approach: LLM first breaks down a multi-hop question into various sub-questions and for each sub-question, the retriever then recalls the most relevant fact from an external memory. The LLM disambiguates if the retrieved fact is useful for answering the question or not.
- **PokeMQA-CL:** PokeMQA (Gu et al., 2024a) is similar to MeLLo but consists of a conflict disambiguator for retrieving as well as classifying if a fact is useful to answer a sub-question. Following PokeMQA, we train this disambiguator using BCE loss with negative sampling for retrieving the close edits, given a decomposed sub-question. However, our training dataset now consists of translated version of the training dataset used in PokeMQA. This training set contains all 8 languages (the multilingual setting) or English along with one of the 7 non-English languages (the bilingual setting).

Multi-hop knowledge editing performance heavily depends on the language of edits. As can be seen in the Figure 2, the gaps in average accuracy between English and other language edits are 10% and 11.7% for methods MeLLo-CL and PokeMQA-CL, respectively, highlighting the significant drop in cross-lingual knowledge editing

Method	CroLin-MQuAKE-CF 3k (All) Acc.	Hop-Acc	CroLin-MQuAKE-CF 100 edited Acc.	Hop-Acc	CroLin-MQuAKE-T 1.8k (ALL) Acc.	Hop-Acc	CroLin-MQuAKE-T 100 edited Acc.	Hop-Acc
LLaMa-2								Size: 7B
FT	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
ROME	1.9	0.0	2.3	0.0	-	-	-	-
MEMIT	0.4	0.3	4.2	1.0	-	-	-	-
MeLLo-CL	10.6	1.9	14.6	2.3	26.5	3.0	28.5	0.7
PokeMQA-CL	10.6	2.3	**19.7**	5.9	11.1	5.8	14.6	7.8
CLeVer-CKE	**13.2**	**7.3**	19.2	**11.1**	**40.6**	**30.0**	**42.6**	**31.1**
Vicuna-1.5								Size: 7B
MeLLo-CL	8.8	2.8	14.5	5.5	34.1	13.5	36.9	13.0
PokeMQA-CL	9.5	2.1	17.3	5.5	11.0	6.6	13.7	8.5
CLeVer-CKE	**12.7**	**7.1**	**18.1**	**10.7**	**37.9**	**30.6**	**39.9**	**31.8**
ChatGPT (GPT-3.5-turbo-instruct)								Size: Undisclosed
MeLLo-CL	14.4	5.4	20.6	8.5	39.0	17.6	41.4	17.0
PokeMQA-CL	12.9	2.9	26.8	9.3	13.5	8.2	17.4	10.7
CLeVer-CKE	**18.6**	**10.6**	**30.1**	**18.6**	**42.6**	**32.8**	**45.6**	**35.1**

Table 1: Performance of parameter update based and in-context editing based methods on the cross-lingual multi-hop knowledge editing problem, reported for three language models, and averaged over 8 diverse languages. Parameter-update based methods – FT, ROME, MEMIT perform significantly worse than in-context editing methods, MeLLo-CL, PokeMQA-CL and CLeVer-CKE, significantly outperform all baselines. Evaluation is performed over two sizes of edited fact memory – 100 and 3k/1.8k following Zhong et al. (2023). See §4 for more details.

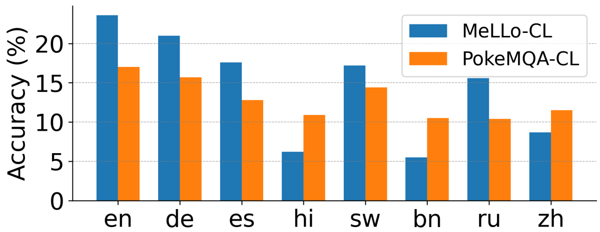


Figure 2: Comparison of multi-hop accuracy of Mello-CL and PokeMQA-CL on the CroLin-MQuAKE-CF across the different languages.

setup. Performance of MeLLo-CL varies significantly across the different scripts. For language written in Latin scripts, the accuracy is ∼20%. In contrast, for languages written in non-Latin scripts such as Devanagari, Chinese, or Cyrillic, the accuracy drops to ∼11%. Another observation is that, in case of edits made in Swahili, despite being a low-resource language, it outperforms more resource-rich languages like Chinese, Russian, and Hindi. This suggests that script plays a crucial role in cross-lingual knowledge editing and retrieval. The reason is intuitive, i.e., Latin script languages have a higher presence in most pretraining data which leads to better tokenization and better representation in LLMs; whereas the non-Latin script languages suffer from high tokenization fertility and less effective representation in the model (Ahia et al., 2023; Singh et al., 2024).

Parameter-modifying based knowledge editing performs poorly in the cross-lingual setting. Methods that update the parameters of the model, such as ROME, MEMIT, FT, perform significantly worse in the cross-lingual setting, achieving an accuracy under 5.0% (average across languages), as shown in Table 1. One key issue is that knowledge edits may not transfer effectively across different languages just via model weights, leading to inconsistent and inaccurate retrievals. Further, the problem is exacerbated due to cascading error propagation in a multi-hop setting. Hence the parameter-modifying methods struggle to reliably edit the LLM across languages and multi-hop contexts. This highlights the need for memory-based approaches that rely on an external edit memory, like our contributed baselines, MeLLo-CL and PokeMQA-CL, which can cross-lingually retrieve the relevant edits on the fly when inferring from an LLM. These approaches substantially improve performance up to nearly 30% on CroLin-MQuAKE compared to parameter-modifying based methods.

Knowledge editing performance based on retriever training technique. MeLLo-CL retrieves the edited fact from the memory using mContriever and employs an LLM to disambiguate between the generated answer and the retrieved fact and hence ascertains if the generated fact needs any update or not. On the other hand, the current state-of-the-art knowledge editing method in English, PokeMQA-CL, uses a retrieve-then-verify approach, which offloads the knowledge disambiguation to the retriever. This retriever is a light-weight and fine-tuned distilbert-base model trained on a (sub-question,edit) pair dataset using binary cross-entropy loss with negative sampling. It retrieves the closest edits (in fact memory) to a sub-question and scores it on whether the edit answers the question or not (called verification or disambiguation). If it does, then it uses this new knowledge as the answer to the sub-question in the *n*-th hop step and performs in-context editing. PokeMQA-CL outperforms MeLLo-CL on in the monolingual (English) setting, with a much smaller retriever as shown in Gu et al. (2024a), however, when trained with multilingual data, we find that it *significantly underperforms MeLLo-CL* in most languages including English as shown in Fig. 2. MeLLo-CL underperforms in Hindi and Bengali – languages with scripts very different from Latin, even though its retriever is trained with 100+ languages.

Qualitative analysis of errors. We examine the error cases of MeLLo-CL and PokeMQA-CL for knowledge edits made in two languages: English and Hindi. Our analysis identifies two primary types of errors made by these methods. The first type is a) incorrect retrieval, where the retrieved information is not relevant to input queries. The second type is b) incorrect LLM response, where a LLM either makes a mistake in extracting the final answer or errors in decomposing the question into subquestions. Additionally, MeLLo-CL exhibits c) contradiction error where the LLM makes mistake at the contradiction step. Figure 7 illustrates the examples of these three types of errors. We analyzed a random subset of 30 samples for these methods and found the following:

❶ MeLLo-CL: When edits are made in English, 63.3% of the samples are correct, 29.3% have the contradiction error, 3.6% have Incorrect retrieval, and 3.6% have the incorrect LLM response. For edits made in Hindi, 33.3% of the samples are correct, 60% exhibit an error combination of incorrect retrieval and subsequent contradiction error, where the model first makes an incorrect retrieval and then fails in the contradiction step and 6.6% of erroneous samples are due to the incorrect LLM response. In the CroLin-MQuAKE-CF case when the multilingual edited fact memory contains edits in English and Hindi, MeLLo-CL's retriever rarely retrieves edits in Hindi, indicating a limitation in its multilingual capabilities. The limitation of MeLLo-CL lies in its retriever-then-contradict mechanism which is up to the LLM.

❷ PokeMQA-CL: When edits are made in English, 53.3% of the samples are correct and 46.3% have the incorrect retrieval error. When edits are made in Hindi, 43.3% are correct, 51% have errors due to the incorrect retrieval and 5.6% are due to the incorrect LLM response. The limitation of PokeMQA-CL lies in its reliance on a bag-of-words model for retrieval. For instance, when presented with the sub-question "*Who is the* ***head of state*** *of the USA?*", it retrieves the fact "*The* ***head of state*** *of Mongolia is Khürelsükh Ukhnaa.*" This example underscores that PokeMQA-CL prioritizes facts with the highest word overlap, specifically "*head of state*" indicating a superficial word matching rather than a contextual grasp of the entities involved.

❸ When trained in a cross-lingual setting, PokeMQA-CL exacerbates the issue of bag-of-words retrieval. For example, for the sub-question "*Where was* ***Bob Dylan*** *born?*", it correctly retrieves "***Bob Dylan*** *was born in the city of Nankoku*" in English. However, if the same edit is made in German, it retrieves "***Bob Dylan*** *spricht die Sprache von Malayalam*" (**Bob Dylan** speaks the language of Malayalam). This issue is a likely a consequence of high word overlap in retriever's internal translation process and is a limitation of current systems.

Section 4 hints signficant gapS between English-only and cross-lingual case, and that proper knowledge retrieval technique is critical to the performance of cross-lingual knowledge editing.

5 CLeVer-CKE for Knowledge Editing

For overcoming limitations in cross-lingual multi-hop knowledge editing, we design CLeVer-CKE, a cross-lingual and light-weight model editor that seamlessly integrates into any backbone LLM, without changing its parameters. CLeVer-CKE is inspired by memory-based and retrieval-augmented knowledge editing methods (Zhong et al., 2023; Gu et al., 2024a; Mitchell et al., 2022b) for mutlihop

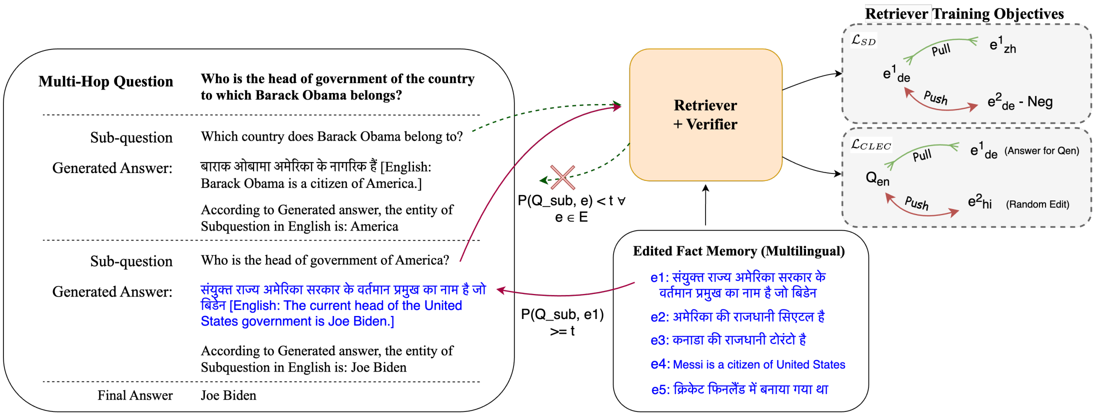


Figure 3: Our proposed method, CLEVER-CKE. On the left we show the LLM inference process for cross-lingual multi-hop knowledge editing. Given a prompt (See §A.5), the LLM breaks down a multi-hop question into sub-questions and answers them individually, utilizing a a retrieve and verify approach using the retriever. On the right, we show new training objectives used in this work for training the retriever. See §5 for more details.

question answering. CLEVER-CKE follows the following procedure: Given an input query, it **a)** decomposes the multi-hop question into multiple sub-questions for getting to the final answer, and for answering each sub-question **b)** retrieves a relevant fact from the edit memory, **c)** disambiguates whether the retrieved new knowledge is relevant to answering the sub-question, and **d)** continues the model generation process based on that. In this work, we primarily aim at showing the importance of having a high-quality retriever for the retrieve-and-verify steps at **b)** and **c)** described as follows. See Fig. 3 for an overview.

Memory of Fact Edits: CLEVER-CKE explicitly stores a set of knowledge edits $\mathcal{E}$ in a memory $\mathcal{F}$. Each edit triplet $e = (s, r, o) \in \mathcal{E}$ is converted to a natural language statement in either English or another language using English or translated templates present in CROLIN-MQUAKE. This creates a multilingual edited fact memory.

Sub-question Decomposition: Given a multi-hop question Q, LLM is prompted using in-context examples to decompose it into various sub-questions $Q_{\text{sub}} = \{q_1, q_2, \dots\}$. Note that Q and the language model generation is assumed to be in English in our work whereas the edited fact memory can contain both English and non-English knowledge edits. The LLM is instructed to answer the generated sub-questions as follows.

Retrieve-and-Verify: For each sub-question q, CLEVER-CKE retrieves the top-1 candidate $r \in \mathcal{F}$ using cosine similarity. Verification process then answers the question: *Does* r *help answer* q*?* The answer to this is yes if $cos(f(r), f(q)) \geq t$ where $cos(.)$ is the cosine similarity function, $f(.) \in \mathbb{R}^d$ is the retriever embedding and t is a threshold (hyperparameter). In this case, r is passed to the LLM which uses it for generating the answer to the sub-question. If $cos(f(r), f(q)) < t$, only the LLM's internal knowledge is used to answer the question. Following this, LLM will move on to answering the next sub-question. Note that here, the disambiguation of whether r is useful or not, happens external to the LLM, reducing its reasoning complexity.

CLEVER-CKE Retriever Training: Motivated by gaps found in Section 4, we create new objectives for training the retriever for improving fine-grained and cross-lingual representations. We then show that our simple losses provide significant gains in knowledge editing performance.

Semantic Distinction Loss: We employ a contrastive, triplet margin loss $\mathcal{L}_{\text{SD}}$ for improving fine-grained cross-lingual retrieval. Assuming an edits $e = (s, r, o)$, we obtain its natural language forms $\mathcal{T}_{L_1}(e)$, $\mathcal{T}_{L_2}(e)$ in languages L_1, L_2 respectively. This creates a positive pair for the triplet loss. We generate hard negatives for $\mathcal{T}_{\text{en}}(e)$ in English by replacing an edits' subject, object, or both object with random entities, with a probability of 0.33 each. This process involves extracting all relations in MQUAKE dataset and prompting the GPT-3.5 model to suggest head/tail entities for these relations. We then randomly sample any generated head/tail (or both) for replacement in an edit containing the corresponding relation. Following this, the hard negative example $\mathcal{T}_{\text{en}}(e_{\text{neg}})$ is translated to L_1 and hence a negative pair $(\mathcal{T}_{L_1}(e), \mathcal{T}_{L_1}(e_{\text{neg}})$

is obtained. The loss function is formulated as:

$$\mathcal{L}_{\text{SD}} = \max(d(f(\mathcal{T}_{L_1}(e)), f(\mathcal{T}_{L_2}(e)) \\ -d(f(\mathcal{T}_{L_1}(e)), f(\mathcal{T}_{L_1}(e_{\text{neg}})) + \alpha, 0). \quad (1)$$

$f(\cdot)$ represents the retriever embedding, $d(.)$ represents the distance function, and α is a gate hyperparameter. $\mathcal{L}_{\text{SD}}$ promotes learning the fine-grained knowledge about subject, relation and object in a cross-lingual setting and encourages the model to distinguish the semantic nuances in different edits. This mitigates the redundant selection of edits with significant word overlap.

Cross-Lingual Edit Consistency Loss: We employ a contrastive, triplet margin loss $\mathcal{L}_{\text{CLEC}}$ focused on improving cross-lingual retrieval. Here, the anchor is Q_{en}, a question in English. The edited fact for answering that question, $\mathcal{T}_{L_1}(e)$, serves as the positive example, and a random edit $\mathcal{T}_{L_2}(e_{\text{rand}})$ forms the negative example:

$$\mathcal{L}_{\text{CLEC}} = \max(d(f(Q_{\text{en}}), f(\mathcal{T}_{L_1}(e)) \\ -d(f(Q_{\text{en}}), f(\mathcal{T}_{L_2}(e_{rand})) + \alpha, 0). \quad (2)$$

BCE Loss: Following (Gu et al., 2024a; Mikolov et al., 2013) we add a binary cross-entropy loss in the cross-lingual setting as a baseline loss for training the retriever for retrieving edits in a cross-lingual setting. The negative BCE Loss function takes questions in English and their corresponding edited facts in one of the seven languages as input. We then compute the L_2 norm between these edits and questions, and sample 20 negatives. The loss function $\mathcal{L}$ is defined similar to Gu et al. (2024a):

$$\mathcal{L}_{\text{BCE}} = -\log g(\mathcal{T}_{L_1}(e), f(Q_{\text{en}})) \\ -\mathbb{E}_{q_n \sim P_n(q)}[\log(1 - g(\mathcal{T}_{L_1}(e), q_n))], \quad (3)$$

where P_n is a uniform over each mini-batch, and $g(.) = exp(d(.))$.

$\mathcal{L}_{\text{CLEC}}$ and $\mathcal{L}_{\text{BCE}}$ encourage it to differentiate between edits in different languages and enhance its ability to handle multilingual knowledge editing tasks effectively. The total loss we use is then:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SD}} + \mathcal{L}_{\text{CLEC}} + \mathcal{L}_{\text{BCE}}. \quad (4)$$

5.1 Performance of CLEVER-CKE

We train the retriever with the above losses on a dataset of 8 languages and measure performance on the CROLIN-MQUAKE. In Table 1, on average across languages and across different LLMs, CLEVER-CKE improves over previous methods by up-to 5.7% in accuracy on CROLIN-MQUAKE-CF and we see a much larger increase in the hop-accuracy which suggests faithful reasoning. On the real world temporal dataset CROLIN-MQUAKE-T, we see a significant increase of about 30% accuracy and more than 25% in hop-accuracy metric. Performance gains are large and consistent or better for larger and more capable models like ChatGPT, as compared to LLaMa-2/Vicuna-1.5. Refer to Figure 8 which illustrates an example where other methods make errors, while CLEVER-CKE correctly answers the question.

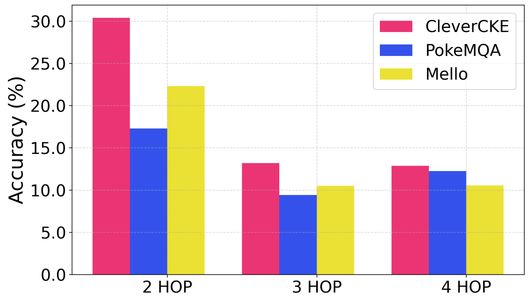


Figure 4: Average accuracy of methods CLEVER-CKE, PokeMQA-CL and MeLLo-CL reported on 2, 3, 4-hop questions with ChatGPT as LLM with the case of all edited on CROLIN-MQUAKE-CF.

Performance across n-hops: We compare the performance of MeLLo, PokeMQA and CLEVER-CKE in answering n-hop questions, $n \in 2, 3, 4$ using CROLIN-MQUAKE-CF dataset and ChatGPT as the LLM. As shown in Fig. 4, CLEVER-CKE outperforms PokeMQA-CL and MeLLo-CL with an average performance increase of 30.7% for 2-hop questions, 22.6% for 3-hop questions, and 5% for 4-hop questions. Fig. 6 presents language-wise accuracies for these methods for n-hop questions, showing the superior performance of CLEVER-CKE compared to other methods.

Bilingual vs Multilingual retriever: To compare performance differences with increasing the number of languages, we trained PokeMQA-CL and CLEVER-CKE's retrievers in a bilingual setting using English and the target language. See Fig 5 for results. As expected, on average the bilingual setting has greater performance than the multilingual setting, potentially due to interference of multiple languages in the multilingual setting. We interestingly observe that this gap is minimal in the case of CLEVER-CKE, compared to PokeMQA-CL. This is because CLEVER-CKE's losses lead to better cross-lingual knowledge transfer leading to reduced interference of languages and more generalization. This observation generalizes across

LLMs and datasets we tested on. Language-wise performance comparison of the two retriever setups for PokeMQA and CLEVER-CKE using ChatGPT, LLaMa-2-7B and Vicuna-1.5-7B are in Tables 6-11. Also see Figs. 9 to 16 for more results.

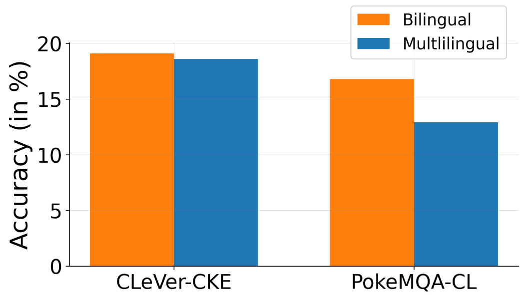


Figure 5: Average accuracy using bilingual vs multilingual retriever, on the CROLIN-MQUAKE-CF dataset in 3k setting using ChatGPT as the LLM.

Ablations: We conducted an ablation on the loss functions we use, with results presented in Table 2. We selected five languages for this study and used the validation set of CROLIN-MQUAKE-CF. $\mathcal{L}_{SD}$ and $\mathcal{L}_{CLEC}$ significantly improve performance over $\mathcal{L}_{BCE}$, showing their importance in inducing fine-grained understanding and cross-lingual awareness in the retriever. Combining both all three losses leads to a 75.3% and 109.5% increase in average accuracy and hop-accuracy.

Loss ↓ Lang. →	EN	DE	HI	SW	RU
$\mathcal{L}_{BCE}$	26.0	28.0	16.0	20.0	16.0
+ $\mathcal{L}_{SD}$	44.0	34.0	12.0	38.0	16.0
+ $\mathcal{L}_{CLEC}$	44.0	36.0	**18.0**	30.0	18.0
+ $\mathcal{L}_{SD}$ + $\mathcal{L}_{CLEC}$	**76.0**	**62.0**	12.0	**58.0**	**26.0**

Table 2: Ablation results of different loss functions used to train the retriever. Results on the validation set from CROLIN-MQUAKE-*CF*.

Error analysis We performed an error analysis of our method similar to the error analysis conducted for PokeMQA-CL and Mello-CL. We analyzed 30 samples each for edits made in English and Hindi. For English, based on random subset, we found that 70% of the samples were correct, 8.1% had Incorrect Retrieval error, and 21.9% had Incorrect LLM Response error. In the case of Hindi, 46.6% of the samples were correct. Of the remaining samples, 26.6% had Incorrect Retrieval error, 16% had both Incorrect LLM Response and Incorrect Retrieval errors, and 10.6% had an Incorrect LLM Response error. Refer Section A.7 for more details.

6 Related Works

Knowledge editing methods: Knowledge editing can be broadly classified intro two groups. 1) Parameter-modifying based editing which locates the parameters related to factual knowledge and subsequently modify them (De Cao et al., 2021; Dai et al., 2022; Mitchell et al., 2022a; Meng et al., 2022a,b). These method requires an error-prone analytic step to identify parameters, which might be model-specific and not efficient. 2) Parameter-preserving based editing keeps the model parameters frozen and explicitly stores the fact edits in an external memory, for retrieval and external validation (Zhong et al., 2023; Gu et al., 2024a; Mitchell et al., 2022c; Hartvigsen et al., 2023). Some recent works like that of Hernandez et al. (2023) have also explored a decoding time approach for editing knowledge. Further, knowledge editing is also explored in multimodal settings, such as for text-to-image models (Basu et al., 2023, 2024; Xiong et al., 2024; Gu et al., 2024b).

Cross-lingual knowledge editing. Recent studies have shifted focus to the multilingual capabilities of SoTA LLMs like LLaMA (Touvron et al., 2023a), ChatGPT (Schulman et al., 2022), and GPT-4 (OpenAI, 2023). Wang et al. (2023a) investigated cross-lingual knowledge editing and its impact on different target languages using a synthetic dataset. (Si et al., 2024) introduced Multilingual Patch Neuron (MPN) for efficient cross-lingual knowledge synchronization, showing enhanced performance on single-hop XNLI and XFEVER datasets. (Xu et al., 2023b) proposed a framework for language anisotropic editing, facilitating simultaneous cross-lingual model editing. (Beniwal et al., 2024) explored the cross-lingual model editing (XME) paradigm, revealing performance limitations in multilingual LLMs for hypernetwrok based parameter-modifying methods. (Wang et al., 2023b) presented Retrieval-augmented Multilingual Knowledge Editing (ReMaKE), a model-agnostic knowledge editing method designed for multilingual settings. ReMaKE retrieves new knowledge from a multilingual knowledge base and concatenates it with prompts to update LLMs. Most of the above works have considered a simplistic setting of assuming the edited facts as independent without any multi-hop consequences of these edits, and are primarily focused on parameter updating based methods. We focus on parameter-preserving methods, and the more complex setting of multi-hop editing in a cross-lingual setup.

Multi-Hop QA and prompting methods: With the advances in generative language technologies powered by Large Language Models (LLMs; Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022; OpenAI et al., 2023; Tay et al., 2023; Google, 2023), complex and multi-hop QA tasks are often handled by a prompt based and retrieval augmented approach (Press et al., 2022; Yao et al., 2023; Khattab et al., 2022). Works that tackle multi-hope knowledge editing have started to use this retrieve-then-generate framework to effeciently peform knowledge editing in an in-context setting, without changing the parameters of the base LLM, and have achieved SoTA performance on knowledge editing. Given their success, we use a similar retrieve, verify and generate strategy for knowledge editing with CLEVER-CKE, while explicitly focussing on the retriever for enhanced knowledge editing performance.

7 Conclusion

In this paper, we contributed a benchmark having parallel multilingual examples for evaluating cross-lingual multi-hop knowledge editing. We provide new baselines and a detailed analysis of SoTA knowledge editing methods and find various gaps in existing methods, particularly in the cross-lingual setting. Motivated by this, we propose a generic, simple and highly effective method, CLEVER-CKE, for improving the knowledge editing capabilities of parameter-preserving, retrieval augmented editing methods. CLEVER-CKE improves cross-lingual and fine-grained retrieval in the case of knowledge editing, by introducing language aware and hard-negative mining based contrastive losses to train retrievers. Improved retrieval leads to precise knowledge retrieval and reduced error propagation in the multi-hop reasoning setting. CLEVER-CKE is parameter-preserving in terms of the LLM weights, and uses a lightweight retriever with low latency as compared to methods like Zhong et al. (2023).

8 Limitations

Our analysis and methods has some limitations. Firstly, although CROLIN-MQUAKE is a parallel cross-lingual benchmark, it predominantly contains fact edits related to English-speaking knowledge changes, while the edits could be localized to any part of the world in practice. This reliance on translation rather than trans-localization may lead to gaps in accurately understanding regional and local fact edits. However, having parallel data in all languages is advantageous to accurately measure per-language performance without confounding factors. Secondly, our method is primarily focused on the retriever component and does not address the inherent inaccuracies of the LLMs. This includes issues such as understanding and generation capabilities of LLMs in different languages, correctly breaking down multi-hop questions into sub-questions, accurately extracting the final answer in the desired language. Lastly, our analysis is currently limited to a broad range of medium to high-resource languages. Extending this analysis to low-resource languages presents a significant challenge due to the inaccuracies in translation, which can hinder the proper representation and understanding of facts in low resource languages. Improving translation accuracy and extending our work to low-resource languages is part of our future work.

9 Acknowledgements

This work is, in part, supported by NSF (#CNS-2431516). We thank the anonymous reviewers for their constructive feedback on this work. The views, opinions, and/or findings contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. *ArXiv*, abs/2305.13707.

Samyadeep Basu, Keivan Rezaei, Priyatham Kattakinda, Ryan Rossi, Cherry Zhao, Vlad Morariu, Varun Manjunatha, and Soheil Feizi. 2024. On mechanistic knowledge localization in text-to-image generative models.

Samyadeep Basu, Nanxuan Zhao, Vlad Morariu, Soheil Feizi, and Varun Manjunatha. 2023. Localizing and editing knowledge in text-to-image generative models.

Himanshu Beniwal, Kowsik Nandagopan D, and Mayank Singh. 2024. Cross-lingual editing in multilingual language models.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

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

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Google. Google translate.

Gemini Team Google. 2023. Gemini: A family of highly capable multimodal models. *ArXiv*, abs/2312.11805.

Hengrui Gu, Kaixiong Zhou, Xiaotian Han, Ninghao Liu, Ruobing Wang, and Xin Wang. 2024a. Pokemqa: Programmable knowledge editing for multi-hop question answering.

Hengrui Gu, Kaixiong Zhou, Yili Wang, Ruobing Wang, and Xin Wang. 2024b. Pioneering reliable assessment in text-to-image knowledge editing: Leveraging a fine-grained dataset and an innovative criterion.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with grace: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems*.

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

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

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

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022b. Fast model editing at scale. In *International Conference on Learning Representations*.

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

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nathan McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, Shengjia Zhao, Arun Vijayvergiya, Eric Sigler, Adam Perelman, Chelsea Voss, Mike Heaton, Joel Parish, Dave Cummings, Rajeev Nayak, Valerie Balcom, David Schnurr, Tomer Kaftan, Chris Hallacy, Nicholas Turley, Noah Deutsch, Vik Goel, Jonathan Ward, Aris Konstantinidis, Wojciech Zaremba, Long Ouyang, Leonard Bogdonoff, Joshua Gross, David Medina, Sarah Yoo, Teddy Lee, Ryan Lowe, Dan Mossing, Joost Huizinga, Roger Jiang, Carroll Wainwright, Diogo Almeida, Steph Lin, Marvin Zhang, Kai Xiao, Katarina Slama, Steven Bills, Alex Gray, Jan Leike, Jakub Pachocki, Phil Tillet, Shantanu Jain, Greg Brockman, and Nick Ryder. 2022. ChatGPT: Optimizing Language Models for Dialogue. *OpenAI*.

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

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

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

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2023b. Retrieval-augmented multilingual knowledge editing.

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

Tianwei Xiong, Yue Wu, Enze Xie, Yue Wu, Zhenguo Li, and Xihui Liu. 2024. Editing massive concepts in text-to-image diffusion models.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. 2023a. 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.

Yang Xu, Yutai Hou, Wanxiang Che, and Min Zhang. 2023b. Language anisotropic cross-lingual model editing.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models.

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

A Appendix

A.1 Verification of Translated Data in CroLin-MQuAKE

A.1.1 Human Verification of Translation

We randomly selected 50 edits in four languages—German, Chinese, Hindi, and Bengali—and had the translations verified by expert human annotators to ensure accuracy. For each sample, we provided two sentences: one in English and its translation in the respective language. The annotators were asked to verify whether the semantic information was consistent between the two sentences. Given the brevity of the edit sentences, the potential for translation errors was minimal. Only one sample from Hindi in the CroLin-MQuAKE-CF dataset encountered an issue during translation due to a special character error; the remaining samples were successfully processed. The expert human annotators suggested only minor stylistic changes for 1-2 words out of all 50 edit sentences in one language.

A.1.2 Verification of Translations via Backtranslation

To ensure the quality of translations, we employed back-translation, converting the translations from other languages back into English, and then calculated the average BLEU scores for 50 samples with the original English sentence as the ground truth. Table 3 presents these BLEU scores, indicating that six out of seven languages exhibit translations of very high quality, adequacy, and fluency [1]. For Chinese, the BLEU score suggests that the gist is clear, although there are some grammatical errors. However, with the addition of human verification (an expert gave a 100% score to the translations in terms of preserving semantic content), we can conclude that the semantic information is preserved in the data translated to Chinese.

Language	BLEU Score
de	70.6
hi	59.2
bn	49.7
es	71.7
sw	65.9
ru	40.0
zh	23.0

Table 3: BLEU Scores for back-translation to English for different languages.

A.2 Training Details

We employ the training dataset to train the retriever component of the CLeVer-CKE framework, using the same training set as utilized in training PokeMQA-CL (Gu et al., 2024a). Subsequently, we translate this dataset into seven other languages and generate hard negatives following the method outlined in Section 5. The training dataset contains 6688 samples along with translations into 8 langugaes and hard-negative pairs for each edit in the dataset, both of which is created by us for training CLeVer-CKE's retriever. For training the multilingual retriever, we utilize data from all languages, while for training the bilingual retriever, we focus on English and the target language data. To optimize our method's performance, we conduct hyperparameter tuning on a validation set derived from CroLin-MQuAKE-CF, comprising 50 samples exclusively for this purpose without involvement in inferencing tasks. The hyperparameters used for tuning are mentioned in Table 4. Our experiments are expensive (See Appendix A.6) and we do not perform experiments on multiple seeds.

A.3 Method Details

We finetuned distilbert-base-multilingual-cased (Sanh et al., 2019) with approximately 130.7M parameters from the HuggingFace transformers library on the training data we created by translation and hard negative mining for the edits as described in Section 5 using our designed training objectives for the retriever. We used held out 20% of the samples for the validation set and used Adam optimizer to update the parameters during training.

Hyperparameter	Value
Learning Rate	5.00×10^{-5}
Batch Size	$\{1024, 2048\}$
Epoch	200
Margin	1
Threshold	$\{0.5 , 0.7\}$

Table 4: Hyperparameter values searched for tuning the multilingual retriever in and CLeVer-CKE and PokeMQA-CL.

A.4 CroLin-MQuAKE Benchmark Statistics

See Table 5 for the dataset statistics of our benchmark CroLin-MQuAKE, which we create in this work and use it for evaluating the cross-lingual multi-hop knowledge editing capabilities of various model editing techniques. Languages studied in this work and supported by CroLin-MQuAKE are English, German, Spanish, Hindi, Swahili, Bengali, Russian, Chinese.

A.5 Prompts for LLM inference

To help the LLM break down questions into subquestions, generate answers for the subquestions, and extract the final answer, we provide four in-context example demonstrations. These examples include edits from different languages based on the edits made. We include a mix of 2, 3, and 4-hop example demonstrations in the prompt. Below, we

[1] https://cloud.google.com/translate/automl/docs/evaluate#interpretation

	#Edits	Hop-Wise Stats (per-language/total)				#Languages
		2-hop	3-hop	4-hop	Total	
CroLin-MQuAKE-CF	1	513 / 4k	356 / 2.8k	224 / 1.8k	1093 / 8.7k	8
	2	487 / 3.9k	334 / 2.7k	246 / 2k	1067 / 8.5k	8
	3	-	310 / 2.5k	262 / 2.1k	572 / 4.6k	8
	4	-	-	268 / 2.1k	268 / 2.1k	8
	All	1000 / 8k	1000 / 8k	1000 / 8k	3000 / 24k	8
CroLin-MQuAKE-T	1 (All)	1421 / 11368	445 / 3560	2 / 16	1868 / 14944	8

Table 5: Statistics of CroLin-MQuAKE created and used in our experiments. Statistics per language are same as those reported in Zhong et al. (2023).

present an example demonstration for a prompt used for edits in German and Swahili. In these demonstrations, the text written in blue represents the updated fact from the edited fact memory, and the text written in teal indicates the answer extraction.

Here is the 3-hop question example demonstration used in the prompt when edits are made in German:

Question: What is the capital city of the country of citizenship of Ivanka Trump's spouse?
Subquestion: Who is Ivanka Trump's spouse?
Generated answer: Der Ehemann von Ivanka Trump ist Jared Kushner.
According to Generated answer, the entity of Subquestion in English is: Jared Kushner
Subquestion: What is the country of citizenship of Jared Kushner?
Generated answer: Jared Kushner ist kanadischer Staatsbürger.
According to Generated answer, the entity of Subquestion in English is: Canada
Subquestion: What is the capital city of Canada?
Generated answer: Die Hauptstadt Kanadas ist Ottawa.
According to Generated answer, the entity of Subquestion in English is: Ottawa.
Final answer: Ottawa

Following is the 2-Hop example demonstration when edits are made in Swahili:

Question: Who is the head of state of the country where Rainn Wilson holds a citizenship?
Subquestion: What is the country of citizenship of Rainn Wilson?
Generated answer: Rainn Wilson ni raia wa Kroatia.
According to Generated answer, the entity of Subquestion in English is: Croatia
Subquestion: What is the name of the current head of state in Croatia?
Generated answer: Jina la mkuu wa sasa wa nchi nchini Kroatia ni Kolinda Grabar-Kitarović.
According to Generated answer, the entity of Subquestion in English is: Kolinda Grabar-Kitarović
Final answer: Kolinda Grabar-Kitarović

A.6 Compute Resources

We performed all experiments using 8 NVIDIA A100 80 GB GPUs. The training duration for the retriever, including both bilingual and multilingual retrievers for both PokeMQA-CL and CLeVer-CKE, was approximately 2 hours per run. Inference tasks took between 4 to 6 hours to complete when using ChatGPT as the LLM in the case of CLeVer-CKE, and between 10 to 24 hours with Llama-2-7b and Vicuna-1.5. Each MeLLo baseline run varied in duration from 8 to 24 hours, depending on the language and the LLM used.

A.7 Error Analysis

Figure 7 presents real examples of errors made by different methods. The first column displays errors related to incorrect retrieval, where the model fails to understand the context of the subquestion and either retrieves a fact with some word overlap with the subquestion or a random edit. The second column shows instances where the LLM makes mistakes in breaking down the subquestion. In the first example, it deviates from the question, asking **when** Giles Gilbert Scott died, and then in the third hop, it just repeats the original question. The second example of this column contains an example where the LLM fails to adhere to the strict pattern of the prompt, misunderstands the context, and generates incorrect information, causing a cascading effect of errors. The third column highlights errors specific to the MeLLomethod, where the LLM struggles to disambiguate between the generated answer and the retrieved fact. In the first example of this column, the retrieved fact contradicts the generated answer, but the LLM fails to identify the correct entity from the generated answer/retrieved fact after resolving the contradiction, leading to a wrong answer. In the second example, although the retrieved fact does not contradict the generated answer, the LLM incorrectly perceives it as a contradiction, resulting in a mistake.

Our method, CLEVER-CKE, addresses and improves upon these errors, as demonstrated in Figure 8. In the same question scenario, where MeLLo-CL exhibits a contradiction error highlighted in yellow and red, and PokeMQA-CL makes a retrieval error based on word overlap, our method follows all the correct steps, leading to the accurate final answer.

A.8 Licensing

The baseline methods ROME, MEMIT, FT, MeLLo, and PokeMQA are distributed under the MIT License. Similarly, the datasets MQUAKE-CF and MQUAKE-T are available under the MIT License. The models Vicuna-1.5-7B (v1.5) and distilbert-base-multilingual-cased are released under the Apache License 2.0, while LLaMa-2-7B is licensed under the LLAMA 2 Community License.

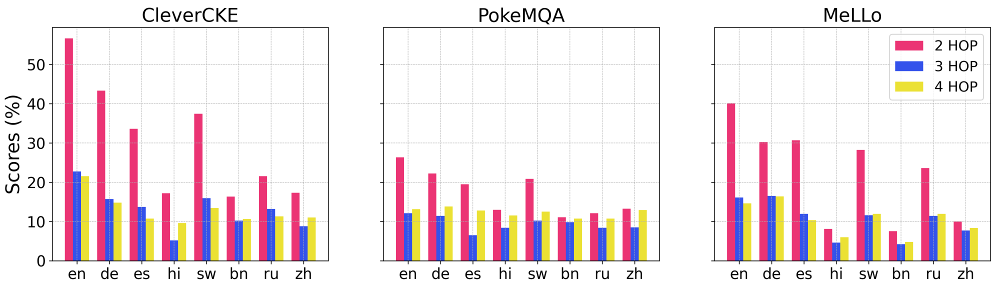


Figure 6: Accuracy of methods CLEVER-CKE, PokeMQA-CL and MeLLo-CL reported on 2, 3, 4-hop questions in CROLIN-MQUAKE-CF with ChatGPT as LLM for all languages. We take the 3k edit case using CROLIN-MQUAKE-CF.

Incorrect Retrieval	Incorrect LLM Response	Contradiction Error
Question: Which country is associated with the origin of the sport played at Toyota Stadium? Subquestion: What is the name of the sport played at Toyota Stadium? Generated answer: Toyota Stadium is associated with the sport of baseball. According to Generated answer, the entity of Subquestion is: baseball Subquestion: Which country is associated with the origin of baseball? Generated answer: **roller derby was created in the country of Ireland.** According to Generated answer, the entity of Subquestion is: Ireland Final answer: Ireland	Question: In which continent did Giles Gilbert Scott die? Subquestion: Who is Giles Gilbert Scott? Generated answer: गाइल्स गिल्बर्ट स्कॉट एक ब्रिटिश आर्किटेक्ट थे। According to Generated answer, the entity of Subquestion in English is: Giles Gilbert Scott Subquestion: **When did Giles Gilbert Scott die?** Generated answer: गाइल्स गिल्बर्ट स्कॉट 8 फरवरी 1960 को मर गए थे। According to Generated answer, the entity of Subquestion in English is: 8 February 1960 Subquestion: **In which continent did Giles Gilbert Scott die?** Generated answer: गाइल्स गिल्बर्ट स्कॉट यूरोप में मर गए थे। According to Generated answer, the entity of Subquestion in English is: Europe Final answer: Europe	Question: Who was the head of government in the country where Manuel L. Quezon was a citizen of? Subquestion: What is the country of citizenship of Manuel L. Quezon? Generated answer: Manuel L. Quezon is a citizen of Philippines. Retrieved fact: Manuel L. Quezon is a citizen of India. Retrieved fact contradicts to generated answer, so the intermediate answer is: India. Subquestion: Who was the head of government in India in 1940? Generated answer: The head of government in India in 1940 was Jawaharlal Nehru. Retrieved fact: The name of the current head of the India government is Rodolfo T. Albano III. **Retrieved fact contradicts to generated answer, so the intermediate answer is: Jawaharlal Nehru.** Final answer: Jawaharlal Nehru.
Question: What is the capital of the country of citizenship of Hélder Postiga? Subquestion: What is the country of citizenship of Hélder Postiga? Generated answer: हेल्डर पोस्टिगा पुर्तगाल के नागरिक हैं। According to Generated answer, the entity of Subquestion in English is: Portugal Subquestion: What is the capital city of Portugal? Generated answer: **Portugal is located in the continent of North America**. It does not have a capital city. According to Generated answer, the entity of Subquestion in English is: None Final answer: None	Question: What is the capital of the country of citizenship of Hélder Postiga? Subquestion: What is the country of citizenship of Hélder Postiga? Generated answer: हेल्डर पोस्टिगा पुर्तगाल के नागरिक हैं। According to Generated answer, the entity of Subquestion in English is: Portugal Subquestion: What is the capital city of Portugal? Generated answer: Portugal is located in the continent of North America. **It does not have a capital city.** **According to Generated answer, the entity of Subquestion in English is: None** Final answer: None	Question: Where did the founder of the religion that Maria Pia of Savoy was associated with work? Subquestion: Who is Maria Pia of Savoy? Generated answer: Maria Pia of Savoy was an Italian princess. Retrieved fact: **Buffy Sainte-Marie is a citizen of Japan.** **Retrieved fact contradicts to generated answer, so the intermediate answer is: Japan** Subquestion: Who is the founder of the religion that Maria Pia of Savoy was associated with? Generated answer: Maria Pia of Savoy was associated with Roman Catholicism. Retrieved fact: सेवॉय की मारिया पिया के धर्म से संबद्ध है अंतिम-दिनों के संतों का यीशु मसीह का चर्च. ...

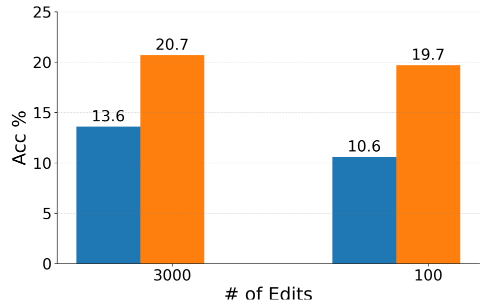
Figure 7: Examples of types of errors made by different methods such as MeLLo-CL, PokeMQA-CL and CLEVER-CKE. Text in red highlights the step at which the error is made. Text highlighted in yellow means the steps that are correct but lead to error in contradiction. Examples are provided in English and Hindi.

MeLLo-CL	PokeMQA-CL	CleVeR-CKE
Contradiction Error	Incorrect Retrieval	Correct
Question: Who is the head of government of the country where David Cunliffe holds citizenship? Subquestion: What is the country of citizenship of David Cunliffe? Generated answer: David Cunliffe is a citizen of New Zealand. Retrieved fact: The name of the current head of state in England is Charles, Prince of Wales. Retrieved fact does not contradict to generated answer, so the intermediate answer is: New Zealand. Subquestion: Who is the head of government of New Zealand? Generated answer: The head of government of New Zealand is Prime Minister Jacinda Ardern. Retrieved fact: The name of the current head of the New Zealand government is Chris Hipkins. **Retrieved fact contradicts to generated answer, so the intermediate answer is: Jacinda Ardern.** Final answer: Jacinda Ardern	Question: Who is the head of government of the country where David Cunliffe holds citizenship? Subquestion: What is the country of citizenship of David Cunliffe? Generated answer: David Cunliffe is a citizen of New Zealand. According to Generated answer, the entity of Subquestion is: New Zealand Subquestion: Who is the head of government in New Zealand? Generated answer: **The name of the current head of state in Malawi is Lazarus Chakwera** According to Generated answer, the entity of Subquestion is: Lazarus Chakwera Final answer: Lazarus Chakwera	Question: Who is the head of government of the country where David Cunliffe holds citizenship? Subquestion: What is the country of citizenship of David Cunliffe? Generated answer: David Cunliffe is a citizen of New Zealand. According to Generated answer, the entity of Subquestion is: New Zealand Subquestion: Who is the head of government in New Zealand? Generated answer: The name of the current head of the New Zealand government is Chris Hipkins. According to Generated answer, the entity of Subquestion is: Chris Hipkins Final answer: Chris Hipkins

Figure 8: Sample of data showing how CLEVER-CKE doesn't make the errors of MeLLo-CL and PokeMQA-CL-CL. Text in red highlights the step at which the error is made. Text highlighted in yellow means the steps that are correct but lead to error in contradiction. Text highlighted in green means the correct final answer achieved by taking all correct steps.

	Edits	Bilingual 3k		Multilingual 3k		Bilingual 100		Multilingual 100	
		Acc	Hop-Acc	Acc	Hop-Acc	Acc	Hop-Acc	Acc	Hop-Acc
PokeMQA-CL	en	39.1	30.7	17.0	7.3	55.9	47.2	35.9	19.5
	de	25.1	14.5	15.7	3.7	29.3	16.6	33.0	12.5
	es	20.6	9.4	12.8	2.8	29.7	13.5	28.2	9.2
	hi	6.8	0.2	10.9	1.0	16.0	1.3	21.4	4.0
	sw	17.0	9.2	14.4	4.0	22.3	13.4	30.7	11.5
	bn	11.1	0.3	10.5	1.2	15.9	1.5	21.6	4.4
	ru	7.9	0.7	10.4	1.5	20.2	4.3	23.2	7.7
	zh	7.1	0.6	11.5	1.5	16.3	3.0	20.5	5.4
	PokeMQA-CL	16.8	8.2	12.9	2.9	25.7	12.6	26.8	9.3
CLEVER-CKE	en	36.2	28.7	33.1	25.0	57.5	48.8	54.8	43.8
	de	29.2	16.0	24.3	14.3	38.1	23.9	39.2	24.3
	es	21.4	11.3	19.1	10.0	34.2	18.4	31.6	17.6
	hi	10.5	4.9	10.5	4.4	22.8	10.6	17.3	8.2
	sw	21.9	14.3	22.0	13.6	34.7	24.6	37.9	24.6
	bn	12.0	4.5	12.3	4.3	16.8	7.8	16.8	7.1
	ru	13.0	7.1	15.2	7.9	25.7	14.7	24.4	14.1
	zh	8.6	3.1	12.3	5.4	16.5	6.8	19.2	9.5
	CLEVER-CKE	**19.1**	**11.2**	**18.6**	**10.6**	**30.8**	**19.5**	**30.1**	**18.6**

Table 6: Performance of PokeMQA-CL and CLEVER-CKE by Language and Number of Edits on the CROLIN-MQUAKE-CF Dataset Using ChatGPT Backbone: Bilingual and Multilingual Training of the Retriever with All and 100 Edits.



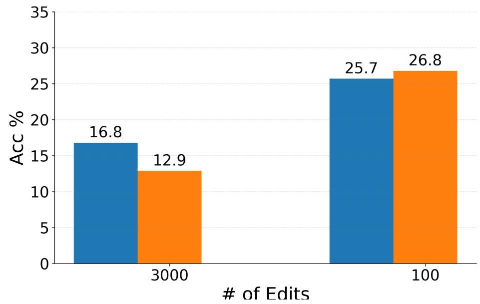
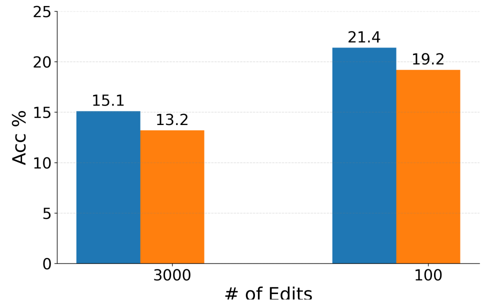


Figure 9: Knowledge Editing accuracy of PokeMQA-CL using LLaMa-2 as the LLM in the Bilingual and Multilingual Case, for two cases – edited fact memory size kept as 3k and 100 edits.

Figure 10: Knowledge Editing accuracy of PokeMQA-CL using ChatGPT as the LLM in the Bilingual and Multilingual Case, for two cases – edited fact memory size kept as 3k and 100 edits.



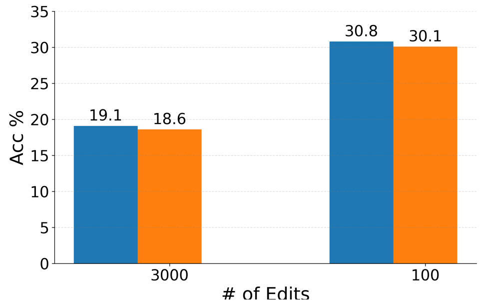
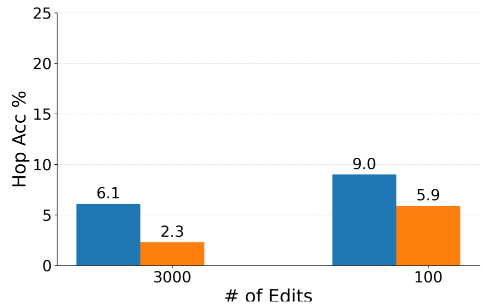


Figure 11: Knowledge Editing accuracy of CLEVER-CKE using LLaMa-2 as the LLM in the Bilingual and Multilingual Case, for two cases – edited fact memory size kept as 3k and 100 edits.

Figure 12: Knowledge Editing accuracy of CLEVER-CKE using ChatGPT as the LLM in the Bilingual and Multilingual Case, for two cases – edited fact memory size kept as 3k and 100 edits.



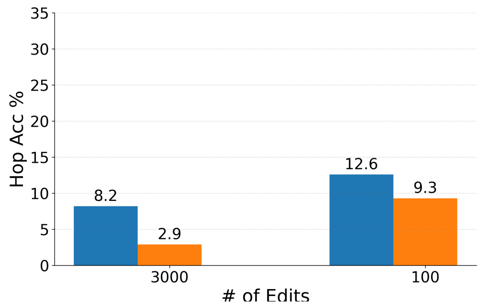
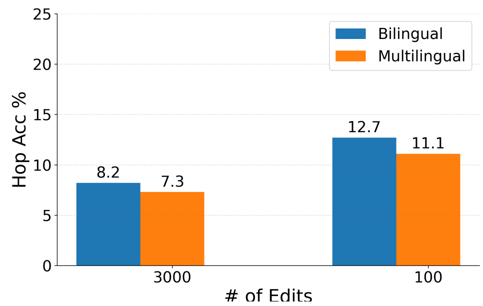


Figure 13: Hop-Accuracy of PokeMQA-CL using LLaMa-2 as the LLM in the Bilingual and Multilingual Case, for two cases – edited fact memory size kept as 3k and 100 edits.

Figure 14: Hop-Accuracy of PokeMQA-CL using ChatGPT as the LLM in the Bilingual and Multilingual Case, for two cases – edited fact memory size kept as 3k and 100 edits.



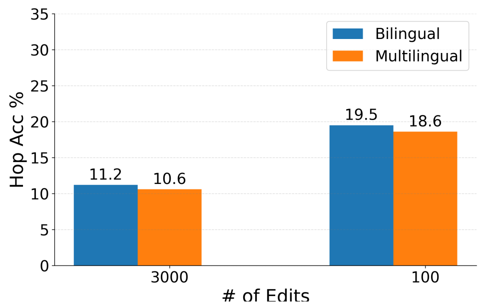


Figure 15: Hop-Accuracy of CLEVER-CKE using LLaMa-2 as the LLM in the Bilingual and Multilingual Case, for two cases – edited fact memory size kept as 3k and 100 edits.

Figure 16: Hop-Accuracy of CLEVER-CKE using ChatGPT as the LLM in the Bilingual and Multilingual Case, for two cases – edited fact memory size kept as 3k and 100 edits.

	Edits	Bilingual 1.8k		Multilingual 1.8k		Bilingual 100		Multilingual 100	
		Acc	**Hop-Acc**	**Acc**	**Hop-Acc**	**Acc**	**Hop-Acc**	**Acc**	**Hop-Acc**
PokeMQA-CL	en	79.1	69.1	23.7	17.6	79.3	69.5	30.0	22.5
	de	45.1	32.3	13.7	08.9	46.5	33.5	17.7	11.1
	es	41.0	28.2	06.7	03.6	45.2	31.2	13.3	8.0
	hi	13.4	6.4	8.6	4.8	15.7	8.6	12.4	7.0
	sw	54.8	41.9	15.5	9.4	58.7	44.3	19.3	11.6
	bn	11.7	5.7	13.8	6.0	12.8	6.4	14.2	7.2
	ru	12.5	7.5	14.9	10.0	14.2	9.4	16.9	10.9
	zh	10.8	5.9	11.0	5.6	14.2	8.4	15.1	7.4
	PokeMQA-CL	33.5	24.6	13.5	8.2	35.8	26.4	17.4	10.7
CLEVER-CKE	en	80.6	69.9	66.6	54.7	81.0	70.3	67.4	55.4
	de	63.6	50.2	59.3	46.5	64.1	50.6	59.7	46.6
	es	45.7	32.2	28.7	19.9	46.3	32.9	29.3	20.2
	hi	39.3	25.6	17.0	9.6	42.0	27.2	16.8	9.5
	sw	47.7	37.3	51.8	37.6	50.1	39.1	52.1	37.8
	bn	20.7	14.1	14.3	8.3	20.9	14.2	14.5	8.5
	ru	58.0	45.2	31.4	22.2	62.5	50.2	32.0	22.5
	zh	46.6	34.3	35.7	23.3	49.0	35.7	35.6	23.2
	CLEVER-CKE	**50.3**	**38.6**	**38.1**	**27.7**	**52.0**	**40.0**	**38.4**	**28.0**

Table 7: Performance of PokeMQA-CL and CLEVER-CKE by Language and Number of Edits on the CROLIN-MQUAKE-T Dataset Using ChatGPT Backbone: Bilingual and Multilingual Training of the Retriever with All and 100 Edits.

	Edits	Bilingual 3k		Multilingual 3k		Bilingual 100		Multilingual 100	
		Acc	**Hop-Acc**	**Acc**	**Hop-Acc**	**Acc**	**Hop-Acc**	**Acc**	**Hop-Acc**
PokeMQA-CL	en	31.5	23.3	13.1	5.4	41.8	31.8	27.7	12.6
	de	16.8	9.2	11.8	3.4	24.1	13.5	23.8	9.3
	es	18.5	8.9	10.8	2.9	25.4	12.1	22.0	7.2
	hi	7.0	0.1	9.8	1.1	12.7	0.8	14.7	2.7
	sw	11.8	5.7	11.9	2.3	14.9	8.2	21.9	5.0
	bn	7.0	0.2	8.0	0.5	14.0	0.5	12.0	1.6
	ru	8.0	0.6	10.7	1.4	17.4	2.9	18.6	5.0
	zh	8.4	0.5	9.1	1.2	15.0	2.4	16.7	3.5
	Average	13.6	6.1	10.6	2.3	20.7	9.0	**19.7**	5.9
CLEVER-CKE	en	27.8	21.0	23.6	17.1	41.5	31.9	37.3	28.3
	de	23.5	13.7	19.7	12.1	29.5	18.6	26.4	17.4
	es	20.0	10.6	8.4	8.4	27.8	16.2	23.6	13.0
	hi	9.6	3.3	10.3	3.3	13.4	5.8	10.8	4.2
	sw	15.5	9.1	14.8	7.7	21.3	13.6	20.1	11.7
	bn	7.2	2.2	6.9	1.7	7.9	2.3	7.3	2.1
	ru	10.0	4.4	12.0	5.2	17.7	9.4	15.8	8.0
	zh	7.6	1.4	9.9	3.4	12.1	3.7	12.1	4.3
	Average	**15.1**	**8.2**	**13.2**	**7.3**	**21.4**	**12.7**	19.2	**11.1**

Table 8: Performance of PokeMQA-CL and CLEVER-CKE by Language and Number of Edits on the CROLIN-MQUAKE-CF Dataset Using LLaMa-2-7B Backbone: Bilingual and Multilingual Training of the Retriever with All and 100 Edits.

	Edits	Bilingual 1.8k		Multilingual 1.8k		Bilingual 100		Multilingual 100	
		Acc	**Hop-Acc**	**Acc**	**Hop-Acc**	**Acc**	**Hop-Acc**	**Acc**	**Hop-Acc**
PokeMQA-CL	en	73.1	58.1	25.6	16.6	73.4	58.2	30.7	19.8
	de	44.0	33.6	11.6	7.8	63.8	51.6	15.0	10.7
	es	52.9	38.5	11.6	5.7	63.3	47.1	18.6	9.2
	hi	10.3	3.2	8.0	3.9	12.7	3.9	10.5	4.6
	sw	45.4	33.8	13.5	4.7	47.6	35.0	16.3	6.8
	bn	5.6	1.0	5.0	2.1	7.0	1.6	7.3	3.3
	ru	10.5	5.1	8.7	3.6	13.4	7.2	12.2	6.2
	zh	4.1	1.9	5.1	2.1	6.4	3.3	6.2	2.4
	Average	30.7	21.9	11.1	5.8	36.0	26.0	14.6	7.8
CLeVer-CKE	en	71.8	57.9	71.5	57.2	72.1	58.1	72.0	57.5
	de	63.2	50.4	59.6	48.1	63.5	50.5	62.2	50.1
	es	57.9	45.0	51.6	40.0	58.0	45.1	52.7	40.8
	hi	33.2	19.0	25.4	15.0	34.9	20.1	27.9	16.2
	sw	43.1	33.1	45.3	33.7	44.0	33.6	46.7	34.6
	bn	10.3	5.8	7.8	4.6	10.5	5.8	9.6	5.2
	ru	58.5	37.2	30.3	18.6	62.4	40.5	34.3	21.1
	zh	40.5	29.0	33.7	22.8	42.0	30.1	35.0	23.6
	Average	**47.3**	**34.7**	**40.6**	**30.0**	**48.4**	**35.5**	**42.6**	**31.1**

Table 9: Performance of PokeMQA-CL and CLeVer-CKE by Language and Number of Edits on the CroLin-MQuAKE-T Dataset Using LLaMa-2-7B Backbone: Bilingual and Multilingual Training of the Retriever with All and 100 Edits.

	Edits	Bilingual 3k		Multilingual 3k		Bilingual 100		Multilingual 100	
		Acc	**Hop-Acc**	**Acc**	**Hop-Acc**	**Acc**	**Hop-Acc**	**Acc**	**Hop-Acc**
PokeMQA-CL	en	28.6	21.8	13.5	5.4	37.5	29.5	25.5	13.0
	de	13.6	7.5	11.2	3.3	21.8	12.4	21.5	8.9
	es	18.2	9.5	10.5	2.7	23.1	12.7	19.6	7.2
	hi	6.8	0.2	7.9	0.8	11.9	0.7	13.3	2.0
	sw	11.4	6.3	10.3	2.5	14.5	8.3	17.5	5.3
	bn	6.1	0.2	6.2	0.4	13.4	0.3	9.7	1.0
	ru	7.4	0.6	7.8	1.0	14.4	2.6	16.1	4.2
	zh	8.0	0.3	8.7	0.7	13.3	2.0	15.0	2.6
	Average	12.5	5.8	9.5	2.1	18.7	8.6	17.3	5.5
CLeVer-CKE	en	27.5	21.4	22.7	17.7	38.5	31.0	36.0	28.1
	de	19.6	12.8	17.5	12.0	27.2	17.8	25.9	17.6
	es	19.3	11.9	15.5	8.7	25.8	16.6	22.4	13.5
	hi	8.5	2.7	8.2	02.2	12.2	4.6	9.7	3.2
	sw	13.0	8.2	12.6	7.7	19.5	12.3	19.2	11.7
	bn	5.5	1.2	5.9	1.4	5.9	1.1	5.8	1.2
	ru	8.6	3.6	10.0	3.8	15.5	7.0	14.0	6.5
	zh	7.2	1.7	8.8	2.9	11.3	2.9	11.5	3.5
	Average	**13.6**	**7.9**	**12.7**	**7.1**	**19.5**	**11.7**	**18.1**	**10.7**

Table 10: Performance of PokeMQA-CL and CLeVer-CKE by Language and Number of Edits on the CroLin-MQuAKE-CF Dataset Using Vicuna-1.5-7B Backbone: Bilingual and Multilingual Training of the Retriever with All and 100 Edits.

	Edits	Bilingual 1.8k		Multilingual 1.8k		Bilingual 100		Multilingual 100	
		Acc	**Hop-Acc**	**Acc**	**Hop-Acc**	**Acc**	**Hop-Acc**	**Acc**	**Hop-Acc**
PokeMQA-CL	en	68.5	56.4	22.6	15.7	68.6	56.6	27.0	18.5
	de	59.1	47.5	10.3	7.2	59.4	47.7	13.6	9.6
	es	59.5	50.0	11.3	6.8	60.1	50.1	16.8	11.0
	hi	11.4	5.5	6.8	4.1	13.5	5.9	10.9	5.8
	sw	49.1	39.3	12.4	4.8	49.7	39.9	13.9	7.5
	bn	6.5	1.3	7.9	4.5	7.7	2.1	8.1	4.5
	ru	8.0	6.3	8.1	5.1	10.4	8.4	10.2	6.3
	zh	11.4	6.6	8.8	4.8	12.4	7.1	9.4	4.8
	Average	34.2	26.6	11.0	6.6	35.2	27.2	13.7	8.5
CLEVER-CKE	en	69.0	57.3	68.0	56.5	69.2	57.5	68.8	57.0
	de	60.9	48.7	52.1	41.7	61.3	49.0	54.5	43.8
	es	56.9	47.3	49.6	41.8	57.0	47.3	51.0	42.7
	hi	23.4	14.8	24.1	16.9	26.0	16.9	27.1	19.0
	sw	44.4	36.6	47.3	39.9	45.3	37.5	48.7	41.0
	bn	11.3	08.0	11.4	08.5	11.1	08.0	13.2	09.3
	ru	51.9	40.5	26.4	20.7	55.5	44.3	28.9	22.9
	zh	32.5	24.5	24.7	19.0	34.5	26.3	27.1	19.0
	Average	**43.8**	**34.7**	**37.9**	**30.6**	**45.0**	**35.8**	**39.9**	**31.8**

Table 11: Performance of PokeMQA-CL and CLEVER-CKE by Language and Number of Edits on the CROLIN-MQUAKE-T Dataset Using Vicuna-1.5-7B Backbone: Bilingual and Multilingual Training of the Retriever with All and 100 Edits.