CLICKER: Cross-Lingual Knowledge Editing via In-Context Learning with Adaptive Stepwise Reasoning

Zehui Jiang[1] **Xin Zhao**[1] **Yuta Kumadaki**[1] **Naoki Yoshinaga**[2]
[1]The University of Tokyo [2]Institute of Industrial Science, The University of Tokyo
{zjiang,xzhao,ykumadak}@tkl.iis.u-tokyo.ac.jp
ynaga@iis.u-tokyo.ac.jp

Abstract

As large language models (LLMs) are increasingly deployed as multilingual services, keeping their factual knowledge accurate across languages has become essential. However, existing knowledge editing (KE) methods struggle to effectively propagate edits in one language to others, while avoiding side effects. To mitigate this issue, we propose **CLICKER**, a KE method with stepwise reasoning that dynamically retrieves only knowledge relevant to a given query and performs editing, while maintaining cross-lingual consistency through: (1) relevance-aware knowledge retrieval, (2) on-demand in-context KE, and (3) language alignment of the outputs. To rigorously evaluate the locality of edits in cross-lingual KE, we develop the **Multi-CounterFact** dataset that contains multiple semantically similar yet irrelevant prompts for each edit. Experiments on Multi-CounterFact and MzsRE with both open- and closed-source LLMs demonstrate that CLICKER effectively localizes edits and resolves cross-lingual inconsistencies, outperforming dynamic KE baselines. The code and data are released.

CLICKER Multi-CounterFact

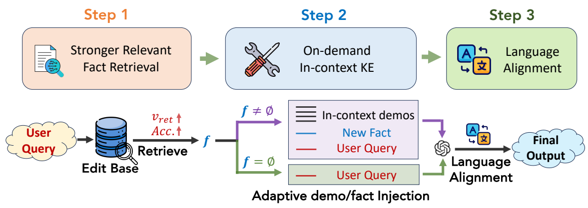

Figure 1: CLICKER at a glance. Starting from a query in the source language, CLICKER adaptively alters the LLM's behavior through a three-step procedure.

1 Introduction

Large language models (LLMs) are increasingly deployed as global services with strong multilingual capabilities (Qwen Team, 2025; Bercovich et al., 2025). As users expect accurate, up-to-date responses in their own languages, maintaining factual consistency across languages remains a major challenge. Practical knowledge updates should minimize side effects (*e.g.*, catastrophic forgetting or model collapse (Yang et al., 2024b)), be executable on the user side, and allow edits made in one language to generalize across others. Meeting these criteria requires efficient mechanisms for updating factual knowledge in multilingual LLMs.

Knowledge editing (KE) aims to efficiently update factual knowledge in LLMs. Most existing KE methods are *static*, performing offline parameter updates or additions that alter the model's behavior across all inputs (Cao et al., 2021; Mitchell et al., 2022a,b; Huang et al., 2023; Zheng et al., 2023). However, such approaches struggle with cross-lingual generalization: edits in one language may not propagate effectively to others (Wang et al., 2024a). Although recent work studies cross-lingual KE (Zhang et al., 2025; Green et al., 2025), these methods are inherently prone to exhibit model collapse (Yang et al., 2024b) as edits accumulate, closely tied to poor *locality*, where a single fact update also distorts outputs for semantically similar yet factually unrelated queries, especially in multilingual settings. In addition, their reliance on parameter updates fundamentally limits their applicability to user-side edits on closed-source models.

Meanwhile, *dynamic* KE methods inspired by in-context learning apply minimal query-specific edits during inference without modifying model parameters (Zheng et al., 2023; Wang et al., 2024c). These methods show promising cross-lingual generalization (Wang et al., 2024a) and support user-side edits. However, as we later confirm on our cross-lingual KE datasets (§ 5.2), their *locality* still needs improvements.

In this study, we propose **CLICKER** (Figure 1), a dynamic in-context cross-lingual KE method that

Findings of the Association for Computational Linguistics: EACL 2026, pages 5007–5022
March 24-29, 2026 ©2026 Association for Computational Linguistics

enhances *locality* through adaptive stepwise reasoning while preserving *reliability* and *generality*. CLICKER achieves balanced cross-lingual KE performance via three adaptive steps: (1) relevance-aware retrieval from the edit base, (2) on-demand in-context KE, and (3) query-language alignment for faithful generation. It edits only when necessary and enforces outputs in the target language, reducing unintended modifications and improving cross-lingual fidelity.

To rigorously evaluate the locality of edits for cross-lingual KE, we also introduce **Multi-CounterFact**, a dataset that extends CounterFact (Meng et al., 2022) to five languages, covering English, German, French, Japanese, and Chinese. Compared to the existing cross-lingual KE benchmarks derived from the ZsRE dataset (Wang et al., 2024a,c; Nie et al., 2025) that have only one unrelated prompts, our Multi-CounterFact dataset includes *ten* unrelated prompts that share the same predicates as each target fact, enabling rigorous and realistic evaluation of locality.

We compare CLICKER to two dynamic KE baselines (Zheng et al., 2023; Wang et al., 2024c) on Multi-CounterFact and MzsRE datasets using both open- and closed-source multilingual LLMs: Qwen2.5-7B-Instruct and GPT-4o-mini. The results confirm that CLICKER greatly improves the locality while enhancing reliability and generality.

The contributions of this paper are as follows:

- We propose **CLICKER**, a dynamic in-context editing method that enhances locality, reliability, and generality in cross-lingual KE (§ 4).
- We construct **Multi-CounterFact**, a reliable benchmark for cross-lingual KE, with each fact linked to *ten* diverse unrelated prompts for rigorous locality evaluation (§ 3).
- We demonstrate CLICKER's effectiveness on not only open- but also closed-source LLMs, achieving superior edit locality (§ 5, § 6).

2 Related Work

In this section, we first review static knowledge editing (KE) methods, which perform parameter updates or additions that globally affect model behavior across all inputs. We then discuss recent dynamic KE methods that adaptively update knowledge via in-context learning, followed by limitations in current datasets for cross-lingual KE.

Knowledge editing has emerged as a lightweight alternative to continual pre-training for updating factual knowledge in LLMs. Early KE methods are mostly *static*, relying on direct parameter changes to alter the model's behavior (Dai et al., 2022; Meng et al., 2022, 2023; Dong et al., 2022; Mitchell et al., 2022b; Huang et al., 2023). However, these methods typically assume a monolingual setting, in which editing and evaluation occur in one language, resulting in poor generalization of edits across languages (Beniwal et al., 2024), as demonstrated by Wang et al. (2024a). Although some studies attempt to mitigate this issue (Zhang et al., 2025; Green et al., 2025), they still fail to support dynamic, user-side editing on closed-source models.

Recently, in-context methods have been proposed for dynamic KE, enabling temporary updates to factual knowledge without modifying parameters (Zheng et al., 2023; Wang et al., 2024c). Zheng et al. (2023) conducted a pilot study showing that LLM behavior can be altered in-context by providing relevant facts. This approach shows promise for cross-lingual generalization (Wang et al., 2024a) and applicability to closed-source models. However, issues remain regarding poor edit locality and how to select relevant facts. Wang et al. (2024c) addressed the latter by retrieving facts from an edit base. Nonetheless, our experiments on locality-sensitive datasets confirm that existing dynamic KE methods still struggle to localize edits and avoid unintended side effects in cross-lingual settings.

Several benchmark datasets for cross-lingual KE have been created by translating the Zero-shot Relation Extraction (ZsRE) dataset in English (Levy et al., 2017) into other languages: Bi-ZsRE (Wang et al., 2024a), MzsRE (Wang et al., 2024c), and BMIKE-53 (Nie et al., 2025).[1] These datasets include only a single unrelated query per record for evaluating edit locality, offering limited coverage and failing to capture the diverse range of irrelevant queries that edits should not affect. This makes it difficult to rigorously assess whether a method truly avoids unintended propagation. Green et al. (2025) introduced BabelEdits, a benchmark designed to mitigate subject aliasing issues in prior datasets; it remains limited in coverage of irrelevant queries, and its unrelated prompts often differ in predicate and subject from the edited facts, making them insufficiently similar to rigorously test locality.

[1] We omit datasets focusing on multi-hop knowledge editing (Khandelwal et al., 2024; Wei et al., 2025), which do not provide prompts for locality.

3 Multi-CounterFact Benchmark

The key challenge in cross-lingual KE is to control how edits propagate across languages. Unrelated queries in the same language, even if superficially similar, should remain unaffected, while related queries in different languages, even if superficially dissimilar, should reflect the edit.

To better evaluate *locality*, we thus introduce **Multi-CounterFact**, a multilingual version of the CounterFact dataset (Meng et al., 2022), containing **ten** unrelated prompts that share the same predicate as the target edits for more rigorous *locality* evaluation than existing datasets like MzsRE.

Each record in Multi-CounterFact includes one prompt with a counterfactual fact (*e.g.*, "*What is the official language of the United Nations? Indonesian.*", two paraphrased prompts (*e.g.*, "*Which language do they understand in the United Nations?*"), and ten unrelated prompts (*e.g.*, "*What is the official language of South Africa?*"). These three types of prompts are used to measure *reliability*, *generality*, and *locality*. Compared to current cross-lingual KE datasets (Wang et al., 2024a,c; Nie et al., 2025) derived from the ZsRE dataset (Levy et al., 2017), the counterfactual fact prompts in our dataset allow us to evaluate future LLMs without being influenced by the knowledge they already have.

We used `GPT-4o-mini` to translate the CounterFact dataset into four target languages: German, French, Chinese and Japanese. Translations were generated using the official OpenAI API[2] with a temperature setting of zero to ensure deterministic output. Following Khandelwal et al. (2024), we computed BLEU scores via back-translation, confirmed high scores above 50. For Chinese and Japanese, which showed lower scores than German and French, native speakers manually evaluated the outputs and found that only 1% required corrections. Refer to Appendix A for verification details and statistics of Multi-CounterFact.

Beyond translation, we chose the languages supported by Multi-CounterFact to include typological diversity. The benchmark covers alphabetic Indo-European languages as well as Japanese and Chinese, which differ substantially in script and word order. Although the number of languages is limited, the selection provides typologically diverse pairs, enabling an efficient yet comprehensive evaluation for cross-lingual knowledge editing.

[2] `https://platform.openai.com`

4 CLICKER

In this section, we present **CLICKER** (Cross-Lingual In-Context Knowledge Editing via adaptive stepwise Reasoning), a dynamic method for cross-lingual knowledge editing. We first define the task, and then detail the proposed approach.

Cross-Lingual In-Context Knowledge Editing. Let $\mathcal{M}_{\text{multi}}$ be a multilingual language model (LM). Given an edited fact in source language s, represented by the tuple $\langle x_e^s, y_e^s \rangle$, where x_e^s is an input prompt (in QA format) (*e.g.*, "*What is the official language of the United Nations?*") and y_e^s is the intended target response (*e.g.*, "*Indonesian.*"), our aim is to realize the following ideal edit behavior:

$$\mathcal{M}^*_{\text{multi}}(x^t) = \begin{cases} \mathcal{I}^t(y_e^s) & \text{if } x^t \in \mathcal{S}_e^t \\ \mathcal{M}_{\text{multi}}(x^t) & \text{otherwise,} \end{cases} \tag{1}$$

where $\mathcal{M}^*_{\text{multi}}$ denotes the predictions of $\mathcal{M}_{\text{multi}}$ when invoked with an edit-augmented context that encodes $\langle x_e^s, y_e^s \rangle$; $\mathcal{S}_e^s$ is the set of source-language edits semantically equivalent to x_e^s; $\mathcal{I}^t(\cdot)$ is a semantic-preserving transformation (translation) into the target language t; $\mathcal{S}_e^t = \{ x_e^t \mid x_e^t = \mathcal{I}^t(x_e^s),\ x_e^s \in \mathcal{S}_e^s \}$ is the induced set of target-language equivalents; $\mathcal{I}^t(y_e^s)$ denotes the corresponding target-language target (*e.g.*, translating "*Indonesian*" into its form in language t).

In other words, the edited model $\mathcal{M}^*_{\text{multi}}$ should generalize this update $\langle x_e^s, y_e^s \rangle$ via the transformation $\mathcal{I}^t$ to all semantically matching prompts in the target language (*reliability*, *generality*), while preserving original outputs for queries about irrelevant knowledge (*locality*).

Edit Base Construction. In line with prior work on parameter-altering methods for knowledge editing, we assume that numerous edits have accumulated since the model's last update and must be incorporated at query time. We thus construct an edit base $\mathcal{E}$, using the test set of Multi-CounterFact. Since this edit base serves as the source of facts to be edited, we restrict it to subject relation pairs x_e that are associated with a single object y_e in Multi-CounterFact. To obtain such a conflict-free subset, we perform a conflict filtering step (Appendix B). Due to practical constraints, we select 1500 records from the Multi-CounterFact test set and obtain 946 unique entries after filtering.

As our focus is on cross-lingual knowledge editing and we aim to evaluate editing performance

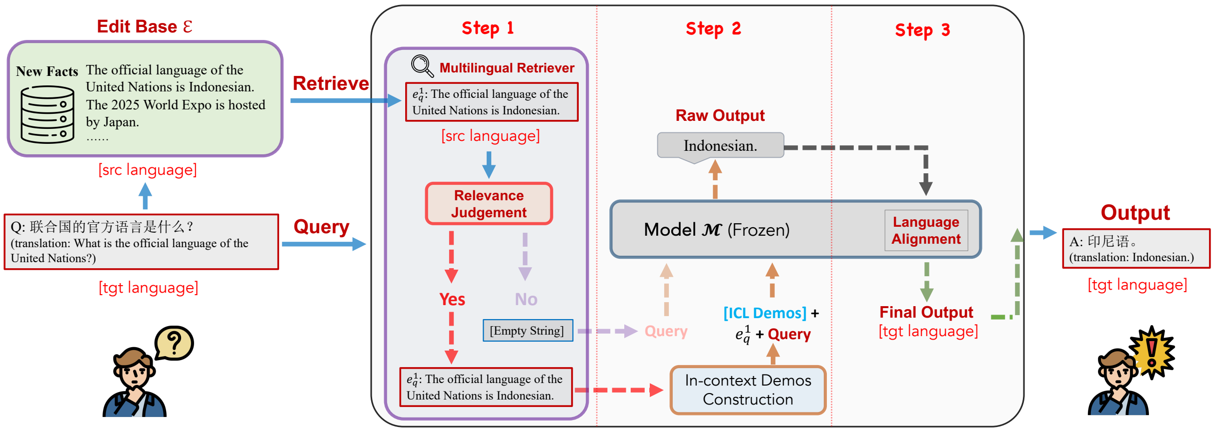

Figure 2: Stepwise reasoning in CLICKER. Here we show the case when the query is related to a newly edited fact in the edit base. Step 1 performs relevance-aware knowledge retrieval; Step 2 applies in-context knowledge editing using adaptively-constructed multilingual in-context prompts; Step 3 performs language alignment of the outputs.

across specific language pairs, we adopt monolingual edit bases in our experiments. This design not only prevents interference from unrelated languages (*e.g.*, Japanese) but also aligns with the nature of dynamic KE. Unlike static KE, which requires tracking thousands of sequential edits accumulated in model parameters, our framework performs edits dynamically at inference time, where each query is relevant to only a small subset of the edit base. Hence, a monolingual edit base suffices for a fair and controlled evaluation.

4.1 Framework Details

Given a multilingual LM $\mathcal{M}_{\text{multi}}$, an edit base $\mathcal{E}$, and a user query x^t in target language t, CLICKER enables $\mathcal{M}_{\text{multi}}$ to return the edited fact in language t, when the query is relevant to any edited knowledge. To achieve this, CLICKER introduces a three-step reasoning process (Figure 2): (1) Relevance-aware knowledge retrieval, (2) In-context prompt construction, and (3) Language alignment. These carefully crafted steps enable adaptive edit control for cross-lingual knowledge editing.

Step 1: Relevance-aware Knowledge Retrieval. To support large-scale and multilingual knowledge retrieval, we design a two-stage, threshold-aware dense retriever. We first fine-tune a multilingual text encoder (specifically, `bge-m3` (Chen et al., 2024)) using triplet training examples taken from the training set of Multi-CounterFact. Each training triplet $\langle q, p, n\rangle$ consists of a query q, a preferred candidate p, and a less-preferred candidate n. For each fact f, we construct positive queries q_+ in another language that either directly request or paraphrase the fact. These form triplets $\langle q_+, f, \texttt{[NULL]}\rangle$, encouraging the model to prefer the correct fact over `[NULL]`, which indicates no relevant edit in the edit base. To improve semantic discrimination, we add hard negatives facts $\tilde{f}$ from other target facts, yielding additional triplets $\langle q_+, f, \tilde{f}\rangle$. For misleading unrelated queries q_-, we form triplets $\langle q_-, \texttt{[NULL]}, f\rangle$, teaching the model to prefer `[NULL]` over incorrect matches.

We train the encoder with a standard triplet loss:

$$\mathcal{L} = \sum_{\langle q,p,n\rangle \in \mathcal{D}} \max\{0,\ d(q,p) - d(q,n) + \alpha\}, \quad (2)$$

where q denotes the query embedding, p the embedding of the preferred candidate, n the embedding of the less-preferred candidate, $d(\cdot,\cdot)$ the cosine distance, and $\alpha = 0.1$ the margin. Given the scarcity of positive examples and abundance of negatives (*e.g.*, unrelated prompts and unrelated facts) in Multi-CounterFact, we upsample positive triplets to promote higher recall.

At inference time, we use FAISS (Douze et al., 2025) for efficient exact nearest neighbor search under cosine similarity.[3] For query q in any supported language, we retrieve the top-1 fact $e_q^1 \in \mathcal{E}$ along with its similarity $\cos(e_q^1, q)$, and apply a threshold τ (tuned on the validation set, see Appendix C for details) to decide whether the retrieved edit e_q^1 is sufficiently relevant to be used for editing:

$$\text{edit}(q) = \begin{cases} e_q^1, & \cos(e_q^1, q) \geq \tau \wedge e_q^1 \neq \texttt{[NULL]} \\ \varnothing, & \text{otherwise} \end{cases} \quad (3)$$

[3]We use `faiss.IndexFlatIP` over L2-normalized embeddings as implementation.

New fact: What is the official language of the United Nations? *Indonesian*.
Prompt: What is the official language of the United Nations?
Answer: *Indonesian*.
新事实：联合国的官方语言是什么？印尼语。
提示：联合国的官方语言是什么？
答案：印尼语。
c_1 Retain

New fact: What is the official language of the United Nations? *Indonesian*.
Prompt: Which language do they understand in the United Nations?
Answer: *Indonesian*.
新事实：联合国的官方语言是什么？印尼语。
提示：在联合国，他们理解哪种语言？
答案：印尼语。
c_2 Rephrase

新事实：What is the official language of WHO? *German*.
提示：世界卫生组织的官方语言是什么？
答案：

Figure 3: Prompt for in-context KE (Step 2): an example (edit in English and test in Chinese); it contains two types of demonstrations, retain and rephrase, in English and Chinese. As indicated by the yellow panel, the retrieved fact (first line) is concatenated with the user query (second line) to form the final query context.

By separating relevance ranking from binary decision making, our retriever achieves both high recall and precise rejection. Unlike ReMaKE (Wang et al., 2024c), which uses a single binary classifier, our method supports large-scale multilingual edit bases via efficient nearest-neighbor search in the embedding space and better balances recall and precision. See Appendix D for a detailed comparison.

Step 2: On-demand In-Context KE. We proceed to Step 2 only when Step 1 returns non-empty results, avoiding the injection of irrelevant information that might interfere with the model's pre-existing knowledge. This adaptive design is key to enhancing the *locality* of CLICKER.

When relevant knowledge is retrieved, the model is expected to answer based on this knowledge. To improve the performance on cross-lingual KE tasks, we include k in-context demonstrations that illustrate the task, as shown in Figure 3. Each example has a prompt in the source language and a semantically equivalent prompt in the target language.

We use two types of demonstrations: *Retain* and *Rephrase*. *Retain* uses the exact prompt from the "New Fact" and provides the edited answer (see c_1 in Figure 3). *Rephrase* uses a lexically different but semantically similar prompt with the same answer. Together, these types improve both *reliability* and *generality*. We do not use demonstrations targeting *locality*, as this is already addressed through the relevance filtering and selective injection. To select demonstrations, we rank candidate examples by cosine similarity to the user query, following Zheng et al. (2023). We provide demonstrations in both source and target languages, aiming to improve cross-lingual generalization. In our implementation, we encode queries and candidate demonstrations using the original `bge-m3` model and use FAISS for efficient top-k nearest neighbor search in the embedding space.[4]

Unlike prior methods such as IKE (Zheng et al., 2023) and ReMaKE (Wang et al., 2024c), our method differs in how we construct and select in-context demonstrations. See Appendix E for a detailed comparison.

You are a professional translator. Translate the following text (which may contain multiple languages) into English. Output only the English translation, without any explanations or additional content.
Text: *[Model output from Step 2]*
English translation:

Figure 4: Prompt for language alignment (Step 3).

Step 3: Language Alignment. This step addresses language confusion commonly observed in cross-lingual tasks. Although edits are made in the source language, the user query is expressed in the target language, requiring the model to handle both. This increases the risk of producing mixed-language outputs. To address this issue, we add a language alignment step at the final stage, ensuring that the model consistently produces the outputs in the target language. Language alignment is implemented via prompt-based methods. A typical prompt is shown in Figure 4.

5 Experiments

We evaluate the effectiveness of CLICKER for cross-lingual knowledge editing. We use language pairs from Multi-CounterFact and edit on both open- and closed-source LLMs.

5.1 Settings

Datasets. We primarily evaluate CLICKER on Multi-CounterFact. To maintain computational efficiency, we randomly selected 200 test examples per target language for the main experiments, resulting in 2600 (200, 400, and 2000) queries to measure reliability, generality, and locality metrics stated below. We provide supplemental results on **MzsRE** (Wang et al., 2024c), the multilingual variant of closed-book QA dataset ZsRE (Levy et al., 2017), which are discussed later in § 6.

[4] We use `faiss.IndexFlatIP` over L2-normalized embeddings, yielding exact top-k *cos*-similarity nearest neighbors.

Metrics	Methods	Edit in en: en→*					Test in en: *→en				
		de	fr	ja	zh	*avg.*	de	fr	ja	zh	*avg.*
Reliability	**IKE**	54.00	42.00	16.50	22.50	33.75	57.50	60.50	10.00	28.00	39.00
	ReMaKE	**89.00**	75.50	**78.00**	85.50	82.00	**85.00**	78.50	69.50	68.50	75.38
	CLICKER	86.00	**81.50**	77.50	**93.00**	**84.50**	81.50	**81.50**	**78.50**	**82.50**	**81.00**
Generality	**IKE**	58.75	53.75	15.25	25.25	38.25	60.50	61.50	9.75	27.75	39.88
	ReMaKE	74.75	68.00	74.25	80.00	74.25	72.25	73.00	65.75	63.50	68.63
	CLICKER	**83.00**	**69.50**	**76.50**	**84.75**	**78.81**	**73.25**	**76.75**	**71.75**	**77.25**	**74.75**
Locality	**IKE**	17.60	21.70	37.20	14.85	22.84	9.95	5.50	4.50	1.20	5.29
	ReMaKE	12.50	9.75	10.85	17.90	12.75	6.95	4.70	5.75	4.95	5.59
	CLICKER	**100.0**	**99.90**	**99.60**	**98.00**	**99.79**	**99.95**	**99.85**	**99.80**	**99.70**	**99.83**

Table 1: Results on Multi-CounterFact (**EM**) for **Qwen2.5-7B-Instruct**, **best** and second best results are emphasized.

Metrics	Methods	Edit in en: en→*					Test in en: *→en				
		de	fr	ja	zh	*avg.*	de	fr	ja	zh	*avg.*
Reliability	**IKE**	41.00	26.00	3.00	16.00	21.50	30.00	20.00	2.00	4.00	14.00
	ReMaKE	91.00	72.50	79.00	96.00	84.63	**98.00**	92.00	58.00	56.00	76.00
	CLICKER	**98.50**	**95.00**	**88.00**	**96.50**	**94.50**	96.00	**96.50**	**92.00**	**95.00**	**94.88**
Generality	**IKE**	42.75	25.25	4.50	15.00	21.88	30.00	22.00	1.00	4.00	14.25
	ReMaKE	76.75	58.00	62.00	**95.75**	73.13	93.00	88.00	57.00	49.00	71.75
	CLICKER	**96.75**	**92.50**	**86.00**	93.00	**92.06**	**94.50**	**94.50**	**90.50**	**95.00**	**93.63**
Locality	**IKE**	22.15	49.50	55.00	12.90	34.89	2.20	17.00	7.00	2.60	7.20
	ReMaKE	28.85	22.25	22.45	13.60	21.79	12.60	20.40	7.20	4.80	11.25
	CLICKER	**99.75**	**99.65**	**98.15**	**99.75**	**99.33**	**99.80**	**99.65**	**98.95**	**95.00**	**98.35**

Table 2: Results on Multi-CounterFact (**EM**) for **GPT-4o-mini**.

Models. We perform cross-lingual KE tasks on two multilingual LLMs. We use the instruction-tuned **Qwen2.5-7B-Instruct**[5] (Qwen Team, 2025; Yang et al., 2024a) as the open-source model, and **GPT-4o-mini**[6] as the closed-source model.

Baselines. We focus our comparison on dynamic KE methods, since static methods are not applicable to closed-source LLMs. Specifically, we use **IKE** (Wang et al., 2024a) and **ReMaKE** (Wang et al., 2024c), which utilizes a retriever and few-shot bilingual demonstrations for cross-lingual knowledge editing. Existing static methods, such as SERAC (Mitchell et al., 2022b), ROME (Meng et al., 2022), MEND (Mitchell et al., 2022a), and MEMIT (Meng et al., 2023), are outperformed by IKE in cross-lingual settings (Wang et al., 2024a). For a comparison with the recent static method, **WISE** (Wang et al., 2024b), refer to Appendix G.

Since IKE (Zheng et al., 2023) assumes the edit knowledge is given, we run our retriever for IKE to allow a fair comparison. Similar to IKE, ReMaKE indiscriminately includes demonstrations, which can inject irrelevant information and harm *locality*. For all in-context methods, we follow (Wang et al., 2024a) and use $k = 16$ examples in our experiments; refer to Appendix F for a detailed discussion on the choice of k. All experiments are conducted on a single NVIDIA RTX A6000 GPU.

Metrics. Following Wang et al. (2024a), we evaluate cross-lingual knowledge editing using three metrics: i) ***Reliability***, the average accuracy of the LLM output on edited instances, indicating its ability to incorporate new knowledge; ii) ***Generality***, measuring performance on paraphrased inputs to assess robustness against prompt variation; iii) ***Locality***, assessing whether unrelated knowledge remains unchanged, reflecting the specificity of the update. All metrics are computed using Exact Match (**EM**), reflecting the proportion of predictions exactly matching the gold answers; refer to Appendix G for consistent results with EM using **F1** score, which captures the average token-level overlap between predictions and gold answers.

5.2 Results

Main Results. Tables 1 and 2 present results of English-centric cross-lingual KE, where English

[5] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[6] gpt-4o-mini-2024-07-18

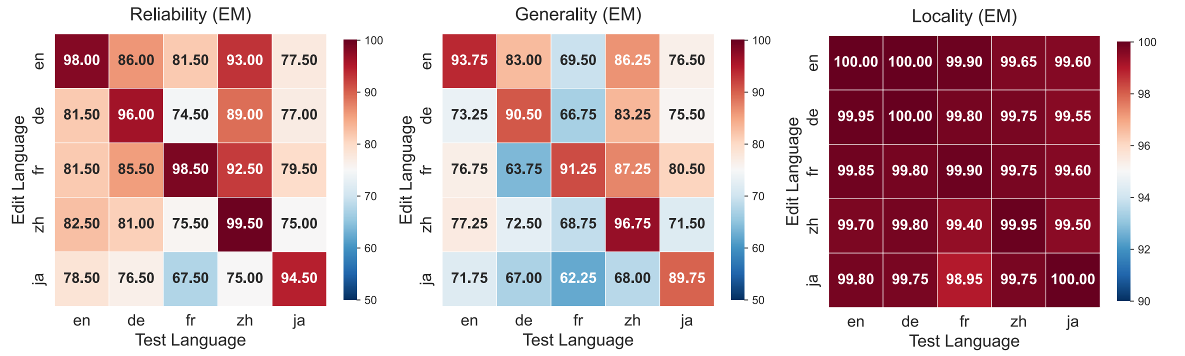

Figure 5: Results (**EM**) of CLICKER for all language pairs on **Qwen2.5-7B-Instruct**.

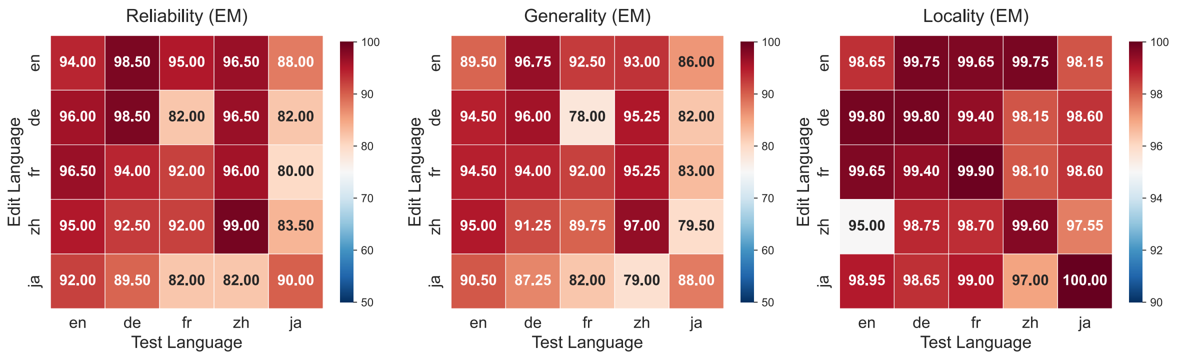

Figure 6: Evaluation results (**EM**) of CLICKER for all language pairs on **GPT-4o-mini**.

serves as either the source or the target language. CLICKER greatly improves locality over baselines by over 60% for Qwen2.5-7B-instruct and over 40% for GPT-4o-mini, and achieves comparable reliability and generality, across all four languages. These gains are consistent across diverse language families and scripts. In Table 1, comparing CLICKER with ReMaKE, reliability differences are generally small under the 200-example setting, and the two methods are overall comparable for English-centered reliability on Qwen2.5. ReMaKE has lower locality on the Multi-CounterFact dataset compared to those on MzsRE we report later in § 6, confirming the value of our Multi-CounterFact dataset for fine-grained cross-lingual KE evaluation.

The large advantage of CLICKER for locality mainly comes from contrastive training of the retriever and adaptive demonstration injection. We will investigate the advantage of CLICKER later in detail by ablation studies (§ 6).

Full results with CLICKER. Figures 5 and 6 show the results of CLICKER across all language combinations. CLICKER shows strong and consistent performance on the *locality* metric, while its *reliability* and *generality* metrics varies. For both the *reliability* and *generality* metrics, CLICKER achieves better performance on Qwen2.5-7B-instruct when the target language is Chinese. This is likely due to the large proportion of Chinese data in Qwen's training corpus, which enhances its ability to understand and generate Chinese. However, performance declines when the source or target language is Japanese, or when the target language is French, likely due to their underrepresentation in the training data. In contrast, GPT-4o-mini yields strong performance on French, suggesting that the weaker French results on Qwen2.5-7B-Instruct stem primarily due to the model, rather than limitations of CLICKER. Since current LLMs do not support all languages equally well, developing effective cross-lingual knowledge editing methods for low-resource languages remains an important direction for future work.

6 Analysis

In this section, we present analyses to verify the advantages of CLICKER. First, we provide ablation studies and results using the MzsRE dataset. Then, we investigate the impact of edit base size, retriever performance, pipeline latency, and robustness to languages underrepresented in the target LLM.

Metrics	Methods	Edit in en: en→*				Test in en: *→en			
		de	**fr**	**zh**	***avg.***	**de**	**fr**	**zh**	***avg.***
Reliability	**IKE**	58.28	46.31	43.35	49.31	37.15	38.03	35.02	36.73
	ReMaKE	**80.35**	71.33	52.36	68.01	81.29	75.10	41.32	65.90
	CLICKER	78.87	**75.24**	**66.08**	**73.40**	**89.37**	**87.35**	**79.54**	**85.42**
Generality	**IKE**	56.66	42.53	39.71	46.30	34.72	33.19	30.67	32.86
	ReMaKE	76.58	65.41	52.22	64.74	77.25	67.16	40.11	61.51
	CLICKER	**78.06**	**74.43**	**67.43**	**73.31**	**86.94**	**86.14**	**78.33**	**83.80**
Locality	**IKE**	31.49	35.13	37.55	34.72	30.96	30.69	23.96	28.54
	ReMaKE	52.36	57.07	30.82	46.75	61.51	65.41	48.18	58.37
	CLICKER	**99.87**	**99.87**	**92.33**	**97.36**	**100.00**	**100.00**	**96.64**	**98.88**

Table 3: Results on the **MzsRE** (Exact Match, **EM**) for English edits and English tests using **GPT-4o-mini**.

edit→test	**Setting**	**Rel.**	**Gen.**	**Loc.**
en→zh	CLICKER	93.00	84.75	98.00
	w/ ReMaKE retriever	91.50	84.70	15.25
	w/ ReMaKE demo	84.50	78.00	98.00
	w/o adaptive injection	93.00	84.75	0.10
	w/o Step 3	91.50	82.25	98.00
zh→en	CLICKER	82.50	77.25	99.70
	w/ ReMaKE retriever	77.00	64.75	3.25
	w/ ReMaKE demo	65.50	63.50	98.65
	w/o adaptive injection	82.50	77.25	0.50
	w/o Step 3	77.50	73.25	98.65

Table 4: Ablation results on Multi-CounterFact for English-Chinese pairs using Qwen2.5-7B-Instruct.

Given the large number of combinations across models, language pairs, and editing directions, we restrict our main analysis to knowledge editing between English and Chinese.[7]

Ablation Studies. Table 4 reports ablations that isolate the contribution of key components in CLICKER, covering the retriever (Step 1), demonstrations (Step 2), and language alignment (Step 3).

Retriever (Step 1). We replace CLICKER's fact retriever with that of ReMaKE. *Reliability* and *generality* clearly drop for English tests, while *locality* substantially drops for both cross-lingual cases. This confirms the advantage of our relevance-aware retriever; see Appendix D for a detailed analysis on the retriever performance.

Demonstration (Step 2). To assess whether CLICKER benefits from its specific demonstration design, we replace CLICKER's demonstrations with those adopted by ReMaKE ("w/ ReMaKE demo"). The results show a clear degradation in *reliability* and *generality*. A key distinction is that CLICKER categorizes demonstrations into retain and rephrase types, and orders the retrieved demonstrations by cosine similarity to the query from low to high. This design provides a clearer semantic signal about the editing task and leads to more stable editing behavior. We also remove the adaptive injection mechanism and forcibly inject demonstrations under the same settings ("w/o adaptive injection"). This causes a significant drop in *locality*, suggesting that irrelevant information in demonstrations unrelated to the target edit can interfere with generation and trigger undesired changes. The result confirms the necessity and effectiveness of adaptive demonstration injection in CLICKER.

Language alignment (Step 3). When disabling language alignment, where outputs are not constrained to the target language, both *reliability* and *generality* clearly decrease in English tests, while *locality* remains stable in both cross-lingual cases. This suggests that the model suffers from language confusion in multilingual settings, especially when the source language is underrepresented in the model. Enforcing the output to be in the target language helps ensure consistent and accurate output in the target language.

Results on MzsRE. Table 3 reports results on the **MzsRE** dataset (Wang et al., 2024c) using GPT-4o-mini. CLICKER consistently surpasses both IKE and ReMaKE. The results on MzsRE show higher average locality than those on Multi-CounterFact (Table 1), suggesting that the severity of the locality issue may be underestimated on the MzsRE dataset. Refer to Appendix G for results of Qwen-2.5-7B-instruct including those with WISE.

Impact of Edit Base Size. To assess the scalability of CLICKER, we evaluated its performance across edit bases of varying sizes, focusing on re-

[7] We chose this language pair since the Chinese dataset was manually verified for quality. Previous work also focuses on this language pair (Wang et al., 2024a; Zhang et al., 2025).

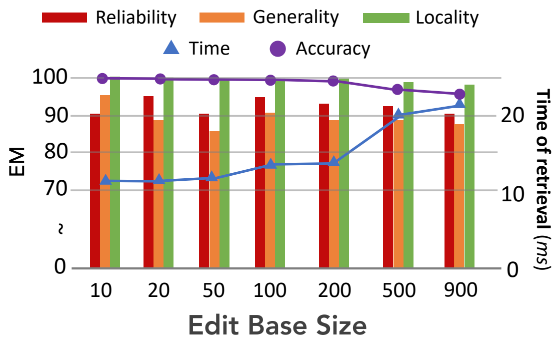

Figure 7: The impact of edit base size. Experiments are conducted between English-Chinese, using Qwen2.5-7B-Instruct on the Multi-CounterFact dataset.

Methods	Step 1	Step 2		Step 3	Total
	retrieval	demo	inference	alignment	
ReMaKE	638.0	58.0	1395.0	0.0	2091.0
CLICKER	14.8	7.8	679.0	116.2	817.8

Table 5: Average end-to-end latency (ms) per query on Qwen2.5-7B-Instruct on the Multi-CounterFact dataset (en↔zh); "demo" and "inference" refer to the time of retrieving in-context demos to format prompts and for generating outputs, respectively.

trieval time and retrieval accuracy as well as the three metrics. As shown in Figure 7, retrieval time increases with the edit base size, but remains in the millisecond range, indicating that CLICKER scales efficiently. Retrieval accuracy shows a slight decline, with a minimum around 95%, which is still acceptable. Similarly, *reliability*, *generality*, and *locality* show minor fluctuations and a modest downward trend, but overall degradation is limited. These results suggest that CLICKER maintains robust performance even as the edit base grows.

Pipeline Latency. Table 5 reports end-to-end latency on 200 Multi-CounterFact test records in English and Chinese using Qwen2.5-7B-Instruct. CLICKER achieves an average total latency of about 818 ms per query, which is less than half of ReMaKE's 2091 ms. The fact retrieval and language alignment steps incur only a small overhead relative to the main inference step. The efficient backbone of our retriever, FAISS, reduces retrieval costs; see Appendix D for details. CLICKER's adaptive design further reduces latency: when the retriever finds no relevant fact, CLICKER skips in-context editing and directly uses the user query without constructing demonstrations, leading to shorter inference time. In contrast, ReMaKE always retrieves, formats, and injects full demonstrations for every query, which yields consistently higher latency. These results suggest that CLICKER's multi-stage pipeline avoids unnecessary overhead and is more efficient, while also providing substantially better locality.

Metrics	en→ja	ja→en	en→zh	zh→en
Reliability	82.00	80.00	88.00	79.00
Generality	83.75	80.25	87.25	76.75
Locality	98.15	98.95	98.00	98.65

Table 6: CLICKER's performance using Llama3.1-8B-Instruct on Multi-CounterFact.

Robustness to Underrepresented Languages. To probe CLICKER's robustness to languages underrepresented in the target LLM, we conduct additional experiments on Llama3.1-8B-Instruct,[8] where Japanese and Chinese are substantially less represented than English.[9] Using Multi-CounterFact, we perform edits between English-Japanese and English-Chinese in both directions. Table 6 shows relatively good evaluation results across these settings, indicating that CLICKER remains effective once the backbone has basic competence in the language, even when the language is underrepresented in the target LLM.

7 Conclusions

In this paper, we propose CLICKER, a cross-lingual in-context knowledge editing framework that updates multilingual knowledge in LLMs via adaptive stepwise reasoning. We further introduce Multi-CounterFact, a five-language benchmark with diverse paraphrased and unrelated prompts for rigorous evaluation. CLICKER has a language-agnostic and model-agnostic design, and our experiments show that it achieves strong reliability, generality, and locality on five typologically diverse languages, substantially outperforming existing multilingual KE baselines. Future work will extend CLICKER and Multi-CounterFact to lower-resource and more diverse languages, scale up the edit base toward real-world RAG settings, and reduce translation-induced bias by incorporating language-native knowledge.

[8] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[9] According to the official model card, Japanese and Chinese are not listed as supported languages.

8 Limitations

Currently, our analysis is confined to high-resource languages, with low-resource languages insufficiently addressed due to persistent translation errors that hinder accurate fact representation. Future research will aim to enhance translation reliability and extend our analysis to low-resource language scenarios.

Second, due to the size limitations of the CounterFact dataset, our edit base contains fewer than 1000 entries. Although we use RAG techniques like FAISS for retrieval, its scale is smaller than that of RAG datastores typically used in real-world applications. Future work will involve constructing a larger multilingual knowledge base to enhance the comprehensiveness and realism of our evaluation. Our current study also does not address the integration of RAG with other augmentation strategies, such as using knowledge graphs for enhanced retrieval, a promising direction for future research.

Third, the number of languages supported in our dataset is limited, while recent datasets, BMIKE-53 (Nie et al., 2025) and BabelEdits (Green et al., 2025), cover dozens of languages. We expect future researchers to extend our dataset to more languages, much like how ZsRE was first extended to Chinese (Bi-ZsRE), then to 12 languages (MzsRE), and eventually to 53 languages (BMIKE-53).

Fourth, CLICKER is not currently designed for massive or sequential editing, where multiple edits are simultaneously or continually applied to the models. However, in the *dynamic* KE setting, each user query typically depends on only a small set of relevant knowledge. Therefore, only a minimal set of knowledge edits is required for each query. "Massive editing" can be approximated by adding multiple edits via the prompt. As for "sequential editing," it boils down to the problem of continuously updating the edit base over time.

Finally, since the Multi-CounterFact dataset is constructed through translation from the English CounterFact dataset, an inherent "regionality" of knowledge arises: certain knowledge represented in English may be uncommon or absent in other languages. This exacerbates translation errors, especially when regional knowledge is involved. While this limitation is partially mitigated through manual verification of translations, the development of datasets that authentically reflect the knowledge domains intrinsic to each language remains a significant challenge for future research.

References

Himanshu Beniwal, Kowsik D, and Mayank Singh. 2024. Cross-lingual editing in multilingual language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2078–2128, St. Julian's, Malta. Association for Computational Linguistics.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, and 114 others. 2025. Llama-nemotron: Efficient reasoning models. *Preprint*, arXiv:2505.00949.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

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), 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.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*, pages 1–17.

Tommaso Green, Félix Gaschi, Fabian David Schmidt, Simone Paolo Ponzetto, and Goran Glavaš. 2025. BabelEdits: A benchmark and a modular approach for robust cross-lingual knowledge editing of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages

8342–8369, Vienna, Austria. Association for Computational Linguistics.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. Transformer-patcher: One mistake worth one neuron. In *Proceedings of the Eleventh International Conference on Learning Representations*.

Aditi Khandelwal, Harman Singh, Hengrui Gu, Tianlong Chen, and Kaixiong Zhou. 2024. Cross-lingual multi-hop knowledge editing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11995–12015, Miami, Florida, USA. Association for Computational Linguistics.

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)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a Transformer. In *Proceedings of the eleventh International Conference on Learning Representations*.

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, 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.

Ercong Nie, Bo Shao, Mingyang Wang, Zifeng Ding, Helmut Schmid, and Hinrich Schuetze. 2025. BMIKE-53: Investigating cross-lingual knowledge editing with in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16357–16374, Vienna, Austria. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2024a. Cross-lingual knowledge editing in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11676–11686, Bangkok, Thailand. Association for Computational Linguistics.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024b. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 53764–53797. Curran Associates, Inc.

Weixuan Wang, Barry Haddow, and Alexandra Birch. 2024c. Retrieval-augmented multilingual knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 335–354, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2025. MLaKE: Multilingual knowledge editing benchmark for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4457–4473, Abu Dhabi, UAE. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. 2024b. The butterfly effect of model editing: Few edits can trigger large language models collapse. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5419–5437, Bangkok, Thailand. Association for Computational Linguistics.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. Multilingual knowledge editing with language-agnostic factual neurons. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5775–5788, Abu Dhabi, UAE. Association for Computational Linguistics.

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? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

Split	Languages	#Records	Prompt*	Paraphrased Prompt*	Answer*	Neighborhood (Unrelated) Prompt*
Training	en	10,000	8.99	9.35	1.65	9.42
	de		11.02	11.59	2.37	11.62
	fr		11.93	12.41	2.31	12.44
	ja		14.14	14.79	3.00	14.79
	zh		9.73	10.23	1.91	10.05
Validation	en	6,000	9.03	9.40	1.67	9.36
	de		11.10	11.66	2.37	11.56
	fr		11.92	12.45	2.34	12.31
	ja		14.26	14.89	2.98	14.72
	zh		9.68	10.07	1.90	9.98
Test	en	4,000	9.06	9.24	1.62	9.31
	de		11.20	11.53	2.36	11.51
	fr		11.90	12.24	2.30	12.18
	ja		14.38	14.84	2.97	14.83
	zh		9.92	10.19	1.88	10.11

Table 7: Statistics of Multi-CounterFact. * # token, averaged on all records.

```
You are a professional multilingual translator and language rewriting expert. Your task is to translate specific fields and proper nouns in a JSON structure from English into a target language, while strictly following the rules below:
1. Preserve the original JSON format - do not change field names, array structures, or order.
2. Only translate the values of the following fields:
  - requested_rewrite.prompt
  - paraphrase_prompts (each string)
  - neighborhood_prompts (each string)
  - requested_rewrite.target_new.str
  - requested_rewrite.target_true.str
  - requested_rewrite.subject
3. Translate all proper nouns (e.g., person names, city names, titles) within these fields into the target language's conventional forms, if such translations exist;
otherwise retain the original.
4. Render all translated text as natural, well-formed, fluent questions in the {TARGET LANGUAGE}.
5. Do NOT translate:
  - JSON field names (e.g., "case_id", "relation_id")
  - Identifiers (e.g., "Q64", "P937")
6. Return valid JSON that can be parsed directly, containing the translated values in place.
```

Figure 8: System prompt for dataset translation.

A Construction of Multi-CounterFact

To construct Multi-CounterFact, we begin by cleaning the CounterFact dataset (Meng et al., 2022), manually reviewing each record to remove irrelevant prefix text in the paraphrased prompts (*e.g.*, for `case_id:0`, the original paraphrased prompt is *"An album was recorded for Capitol Nashville but never released. Danielle Darrieux spoke the language"*, we delete the prefix *"An album was recorded for Capitol Nashville but never released."* for each record, which is irrelevant to the edit). We then translate each English record in CounterFact to the four languages, using the `GPT-4o-mini`[10] with the system prompt shown in Figure 8.

In short, we input each `JSON`-formatted record from CounterFact into `GPT-4o-mini`, preserving the original keys, and generate the translation. To improve the translation quality, we provide the full context rather than individual fields (*e.g.*, `paraphrased_prompts` in isolation). We set the temperature to zero to avoid randomness.

Statistics

Table 7 presents the statistics of Multi-CounterFact, covering training, validation, and test splits in all languages (en, de, fr, ja and zh). Token counts are computed as subword tokens using the Qwen2.5-7B-Instruct tokenizer, applied uniformly across all languages, without additional word segmentation for Chinese or Japanese. As we can see in the table, paraphrased and unrelated prompts are slightly longer than the original prompts. Note that token counts are not directly comparable across languages, as they depend on the vocabulary coverage and segmentation behavior of the tokenizer.

Quality Assessment

Automatic Metric. To estimate translation quality, we adopt back-translation (Khandelwal et al., 2024). For each target language, we translate 200 randomly sampled translations back into English and compute corpus-level BLEU using sacreBLEU (Post, 2018), with the original English sentences as references and the back-translations as hypotheses. Since BLEU measures n-gram overlap,

[10] gpt-4o-mini-2024-07-18

Language	BLEU Score
zh	57.0
ja	50.6
de	63.3
fr	69.1

Table 8: BLEU scores of back-translation from different languages to English on the Multi-CounterFact dataset.

Edit \ Test	en	de	fr	ja	zh
en	0.88	0.76	0.73	0.59	0.60
de	0.76	0.85	0.72	0.59	0.60
fr	0.73	0.71	0.87	0.59	0.60
ja	0.60	0.59	0.60	0.91	0.61
zh	0.60	0.59	0.61	0.61	0.87

Table 9: Threshold selection between all language pairs on the Multi-CounterFact dataset.

the scores reported in Table 8 indicate substantial surface-level correspondence between the translations and the source sentences across all target languages, suggesting that the automatic translations are of sufficient quality for our experiments.

Human Verification. Since back-translation provides a rough estimate, we conducted a more detailed evaluation for two lower-scoring target languages. We randomly sampled 250 records from the Chinese and Japanese splits and had the first and third authors review them in their native languages. Each sample included an English sentence and its translation. The authors assessed both syntactic and semantic alignment, confirming that only 1% of the records required corrections, indicating overall high quality.

We also checked the structural integrity of the translated datasets. Format issues appeared in only 0.5% of the records, primarily due to minor deviations from the expected `JSON` structure. These minor issues were manually corrected.

B Edit Base Conflict Filtering

To construct a reliable edit base for evaluating knowledge editing methods, it is crucial to ensure that (1) no duplicated or conflicting edited facts exist, and (2) unrelated prompts (used to evaluate *locality*) do not inadvertently overlap with edited knowledge. To achieve this, we apply a two-stage conflict filtering procedure based on dense retrieval.

Filtering conflicting edited knowledge. We first detect and remove potential conflicts among entries in the Multi-CounterFact test set. Concretely, we use the multilingual text encoder `bge-m3`[11] to encode all entries in the Multi-CounterFact test set into dense vectors. For each entry f, we retrieve its top-5 most similar entries based on cosine similarity. We then manually check whether any of the retrieved entries share the same or semantically similar x_e (i.e., the edited knowledge element) with f. If overlap is found, one of the duplicates is removed; otherwise, all entries are retained.

Filtering unrelated prompts. Since evaluating *locality* may involve retrieving from the edit base, we ensure that `unrelated_prompts` do not accidentally reference any edited knowledge. After constructing the edit base from all edited knowledge, we filter each `unrelated_prompts` to verify that none of the concepts it mentions appear in the edit base. This guarantees that the theoretical upper bound for *locality* is 100%, preventing false positives caused by unintended knowledge overlap.

C Selecting Retriever Threshold in Step 1

For Step 1 of CLICKER, after fine-tuning the multilingual text encoder `bge-m3` on training triples in Multi-CounterFact, we conducted a grid search on the validation set to find the optimal similarity threshold τ. We introduced this threshold-based relevance filter because, while the retriever effectively identifies true positives, it struggles with rejecting false positives. The thresholding step improves the accuracy of Step 1.

To find the optimal threshold, we constructed a labeled validation set using the following design:

Positive pairs: ⟨target fact prompt [source language], target fact prompt or paraphrase prompt [target language]⟩,

Negative pairs: ⟨target fact prompt [source language], each unrelated prompt [target language]⟩.

We computed cosine similarities using `bge-m3` for all pairs and performed a grid search over thresholds from 0 to 1 (in steps of 0.01). For each threshold, we calculated the F1 score on the validation set and selected the value that achieved the highest score (*e.g.*, $\tau = 0.60$, for en-zh pairs).

Table 9 shows the results. We observe a lower threshold τ in cross-lingual settings, reflecting lower similarity between queries and edits.

[11] https://huggingface.co/BAAI/bge-m3

Edit Base Size	CLICKER		ReMaKE	
	Time [ms]	Acc. (%)	Time [ms]	Acc. (%)
10	12	100.00	36	100.00
20	13	100.00	71	97.31
50	13	99.08	177	87.23
100	14	99.08	355	83.23
200	14	98.42	695	77.42
300	20	97.74	1096	72.26
400	20	97.04	1394	70.75
500	20	96.29	1788	66.57
600	20	95.71	2102	64.37
700	21	94.62	2504	61.56
800	21	94.07	2822	54.21
900	21	93.36	3200	51.68

Table 10: Performance comparison between our proposed retriever and the one used in ReMaKE (Wang et al. 2024). "Time" refers to retrieval time. Experiments are conducted on Multi-CounterFact.

Edit Base Size	ReMaKE		CLICKER	
	Acc.(+)	Acc.(-)	Acc.(+)	Acc.(-)
10	100.00	100.00	100.00	100.00
20	100.00	96.50	100.00	100.00
50	96.00	84.60	100.00	98.80
100	98.00	78.80	99.67	98.90
200	98.00	71.00	99.83	98.00
500	94.80	58.10	99.00	95.48
900	90.93	39.90	97.26	92.19

Table 11: Accuracy of ReMaKE's and CLICKER's retriever when facing positive(+) and negative(-) queries.

D Retriever Performance

Table 10 presents a systematic comparison between our proposed retriever and the one used in ReMaKE (Wang et al., 2024c); we evaluated both retrievers on the Multi-CounterFact dataset, varying edit base sizes and measuring retrieval time and accuracy. The results show that our retriever outperforms ReMaKE's in both efficiency and accuracy. While the accuracy improvements can be partly attributed to our use of a stronger text encoder (`BAAI/bge-m3` vs. `XLM-R` in ReMaKE), the improvement in retrieval efficiency is a distinct advantage of our implementation choice. Notably, ReMaKE's retrieval time increases linearly as the edit base size grows, reaching more than 600 ms for 200 edits, which makes it impractical for accumulating massive edits. In contrast, our retriever employs FAISS (Douze et al., 2025) for efficient nearest neighbor search, reducing retrieval time to just 14ms for the same edit base size. This demonstrates superior scalability and efficiency.

Table 11 further analyzes the retriever accuracy of ReMaKE and CLICKER in handling positive (relevant to edits) and negative (irrelevant) queries on Multi-CounterFact. CLICKER benefits from a thresholding mechanism that enhances its ability to reject false positives. In contrast, ReMaKE's retriever struggles to filter out similar negative queries, particularly as the size of the edit base increases. Refer to Appendix F for CLICKER's robustness against threshold variations.

	IKE	ReMaKE	CLICKER
Ascending Similarity	✓	✗	✓
Multilingual Demos	✗	△	✓
Reliability Demos	✓	✓	✓
Generality Demos	✓	✗	✓
Locality Demos	✓	✗	✗

Table 12: Comparison of existing in-context prompt construction strategy. "✓" refers to "yes", ✗ refers to "no", and "△" refers to "partially (in some cases)".

E In-context Prompt Comparison

Table 12 compares our in-context prompt construction strategy with prior approaches such as IKE (Wang et al., 2024a) and ReMaKE (Wang et al., 2024c). The comparison highlights differences in demonstration ordering (ascending similarity) and in the coverage of multilingual, reliability, generality, and locality-oriented demonstrations.

For the selection of in-context examples, we follow the strategy proposed by Zheng et al. (2023): ranking the candidate examples in ascending order of cosine similarity with the user query (Ascending Similarity):

$$\cos(c_1, q) < \cos(c_2, q) < \cdots < \cos(c_k, q) \quad (4)$$

where c_i represents a training prompt corresponding to a "New Fact", and q is the user query, both represented via sentence embeddings.

F Sensitivity to Hyperparameters

We further examine the robustness of CLICKER by varying two key hyperparameters in retrieval-augmented ICL editing: the number of demonstrations k and the retriever threshold. To control the experimental scope, we focus on English-Chinese cross-lingual KE with **Qwen2.5-7B-Instruct** on Multi-CounterFact. The results show stable behavior under moderate perturbations, with a mild diminishing-return effect of larger k and a predictable trade-off induced by the threshold.

Metrics	Methods	Edit in en				Test in en			
		de	fr	zh	*avg.*	de	fr	zh	*avg.*
Reliability	**WISE**	22.61	23.26	25.57	23.81	24.30	22.35	31.43	26.03
	IKE	50.07	40.97	14.41	35.15	49.44	44.87	13.09	35.80
	ReMaKE	**72.01**	**60.30**	57.47	**63.26**	75.37	72.54	55.32	67.74
	CLICKER	63.39	55.32	**66.08**	61.59	**75.64**	**72.14**	**68.51**	**72.10**
Generality	**WISE**	23.49	22.71	24.84	23.68	24.69	22.95	31.55	26.40
	IKE	48.93	42.46	16.06	35.81	49.21	44.81	13.01	35.68
	ReMaKE	**68.24**	**55.99**	55.99	**60.07**	69.31	66.89	50.34	62.18
	CLICKER	59.49	49.13	**64.33**	57.65	**71.74**	**68.64**	**62.45**	**67.61**
Locality	**WISE**	**99.90**	**99.90**	**100.0**	**99.93**	**100.0**	**100.0**	**99.37**	**99.79**
	IKE	19.50	28.00	12.00	19.83	23.50	16.00	7.61	15.70
	ReMaKE	21.94	23.01	17.77	20.91	30.96	28.80	27.46	29.07
	CLICKER	99.87	98.25	92.33	96.82	**100.00**	**100.00**	93.31	97.77

Table 13: Results on the **MzsRE** (Exact Match, **EM**) for English edits and English tests using **Qwen2.5-7B-Instruct**.

edit-test	k	Reliability	Generality	Locality
en-zh	**4**	90.00	82.75	98.00
	8	92.00	83.00	98.00
	16	93.00	84.75	98.00
	32	94.50	86.25	98.00
zh-en	**4**	73.50	72.00	99.70
	8	75.50	74.25	99.70
	16	82.50	77.25	99.70
	32	83.50	78.75	99.70

Table 14: CLICKER's performance while varying the number of in-context demonstrations k on Multi-CounterFact using Qwen2.5-7B-Instruct.

Threshold	Reliability (EM/F1)	Generality (EM/F1)	Locality (EM/F1)
0.45	93.00 / 93.00	84.75 / 84.75	80.95 / 82.37
0.50	93.00 / 93.00	84.75 / 84.75	87.65 / 88.59
0.55	93.00 / 93.00	84.75 / 84.75	94.90 / 95.20
0.60	**93.00 / 93.00**	**84.75 / 84.75**	**98.00 / 98.00**
0.65	91.00 / 91.00	83.75 / 83.75	99.00 / 99.00
0.70	87.00 / 87.00	79.75 / 79.75	99.55 / 99.55
0.75	78.50 / 78.50	72.00 / 72.00	99.65 / 99.65

Table 15: CLICKER's overall performance using different thresholds, τ. Experiments are conducted on Multi-CounterFact using Qwen2.5-7B-Instruct backbone.

Sensitivity to Number of Demonstrations. As mentioned in (Zheng et al., 2023), the number of in-context demonstrations is one of the influencing factors of the ICL performance. Here we further examine how varying the number of examples affects CLICKER's overall performance. Given the large set of possible combinations across models, language pairs, and editing directions, we focus on the cross-lingual KE between English and Chinese, using the **Qwen2.5-7B-Instruct** backbone on the Multi-CounterFact dataset.

Table 14 presents the results. We observe that as the number of in-context demonstrations k increases, both *reliability* and *generality* improve at first, although the rate of improvement gradually diminishes. Based on these results, setting $k = 16$ strikes a balance between efficiency and accuracy, supporting the validity of prior work that also adopts $k = 16$.

Sensitivity to Retriever Threshold. We also evaluate CLICKER's overall performance under varying threshold values, τ, to decide whether to edit. As the threshold is designed to enhance the model's ability to reject false positives, increasing it typically improves the rejection of irrelevant examples. However, a higher threshold may also lead to the rejection of true positives, resulting in decreased *reliability* and *generality*. Conversely, lowering the threshold makes the model more permissive, increasing the risk of false positives and thereby degrading *locality*. The results in Table 15 confirm this analysis. We also observe that minor fluctuations in the threshold have limited impact on performance, which highlights the robustness of CLICKER with respect to threshold selection.

G Supplementary Experimental Results

This section provides supplementary experimental results that extend the main findings along two dimensions: (1) additional target models and baselines, and (2) alternative evaluation metrics. Specifically, we report results on MzsRE (Wang et al., 2024c) using Qwen2.5-7B-Instruct, including a comparison with WISE (Wang et al., 2024b), and provide F1 scores on Multi-CounterFact.

Table 13 lists results on the **MzsRE** dataset us-

Metrics	Edit in en	Qwen2.5-7B-Instruct					GPT-4o-mini				
	Methods	**de**	**fr**	**ja**	**zh**	***avg.***	**de**	**fr**	**ja**	**zh**	***avg.***
Reliability	**IKE**	57.95	54.90	19.87	24.76	39.37	57.54	48.68	5.18	20.50	32.98
	ReMaKE	**90.33**	80.13	**81.07**	87.66	**84.80**	94.51	85.26	88.55	**97.66**	91.50
	CLICKER	86.25	**81.75**	77.79	**93.00**	84.70	**98.50**	**95.00**	**88.00**	96.67	**94.54**
Generality	**IKE**	62.88	59.61	17.72	26.43	41.66	44.86	27.59	6.23	19.50	24.55
	ReMaKE	75.54	**71.62**	75.80	81.45	76.10	77.33	59.56	63.19	**97.25**	74.33
	CLICKER	**83.00**	70.16	**76.91**	**86.89**	**79.24**	**96.75**	**92.50**	**86.21**	93.25	**92.18**
Locality	**IKE**	30.74	33.43	45.37	22.87	33.10	33.87	65.78	64.37	21.11	46.29
	ReMaKE	22.26	15.94	16.92	25.62	20.19	41.55	32.94	32.78	18.79	31.52
	CLICKER	**100.0**	**99.98**	**99.88**	**99.83**	**99.92**	**99.76**	**99.65**	**98.31**	**99.92**	**99.41**

Table 16: Evaluation results (**F1**) for English edits using both **Qwen2.5-7B-Instruct** and **GPT-4o-mini**. All methods are assessed on the **Multi-CounterFact** benchmark.

Metrics	Test in en	Qwen2.5-7B-Instruct					GPT-4o-mini				
	Methods	**de**	**fr**	**ja**	**zh**	***avg.***	**de**	**fr**	**ja**	**zh**	***avg.***
Reliability	**IKE**	62.43	66.40	10.25	28.65	41.93	34.50	22.50	2.00	4.00	15.75
	ReMaKE	**85.39**	80.32	70.00	69.83	76.39	**98.00**	94.33	58.00	56.00	76.58
	CLICKER	82.00	**82.00**	**78.50**	**82.50**	**81.25**	96.00	**96.50**	**92.00**	**95.00**	**94.88**
Generality	**IKE**	63.95	65.22	10.25	27.95	41.84	34.50	23.75	1.00	4.00	15.81
	ReMaKE	72.58	73.66	65.75	63.50	68.87	93.00	88.00	57.00	49.00	71.75
	CLICKER	**73.54**	**76.75**	**71.75**	**77.50**	**74.89**	**94.50**	**94.50**	**90.50**	**93.50**	**93.25**
Locality	**IKE**	20.78	12.97	11.70	6.56	13.00	5.09	41.71	17.35	5.92	17.52
	ReMaKE	14.63	11.43	11.67	4.95	10.67	25.06	30.53	11.62	11.56	19.69
	CLICKER	**100.0**	**99.97**	**99.96**	**99.89**	**99.96**	**99.80**	**99.65**	**98.98**	**98.79**	**99.31**

Table 17: Evaluation results (**F1**) for English tests using both **Qwen2.5-7B-Instruct** and **GPT-4o-mini**. All methods are assessed on the **Multi-CounterFact** benchmark.

ing Qwen2.5-7B-Instruct. This comparison also allows us to evaluate WISE, which requires direct access to model parameters and primarily supports the **ZsRE**, **Hallucination**, and **Temporal** datasets (Wang et al., 2024b). Adapting WISE to Multi-CounterFact (or reformatting Multi-CounterFact to WISE's format) is non-trivial. We observe a consistent trend where CLICKER outperforms IKE and ReMaKE on average, while WISE appears largely insensitive to cross-lingual edits, leading to overly optimistic locality scores. Given this behavior on MzsRE, we leave WISE's evaluation on Multi-CounterFact for future work.

Tables 16 and 17 present F1 scores on the Multi-CounterFact dataset. On average, the F1 scores are slightly higher than the EM scores reported in Tables 1 and 2, as they capture token-level overlaps between model outputs and gold answers.