Adaptive Stepwise Reasoning for Localized Cross-Lingual In-Context Knowledge Editing

Zehui Jiang, Yuta Kumadaki, Xin Zhao
The University of Tokyo
{zjiang,ykumadak,xzhao}@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science,
The University of Tokyo
ynaga@iis.u-tokyo.ac.jp

Abstract

Ensuring precise and localized factual updates in multilingual large language models (MLLMs) is challenging: parameter-updating methods often fail to propagate edits across languages, while in-context editing often fail to localize edits, disrupting unrelated knowledge. We propose **CLICKER**, a purely inference-time framework that retrieves only relevant edits, performs on-demand in-context editing, and ensures target-language outputs. We further build **Multi-CounterFact** to stress locality with semantically similar yet irrelevant prompts. Across both open- and closed-source MLLMs, CLICKER markedly improves locality while resolving cross-lingual inconsistencies.

1 Introduction

As recent large language models (LLMs) become more multilingual [1] and users thus expect accurate and up-to-date responses in their own languages, maintaining factual consistency across languages has become a major challenge. Practical knowledge updates should minimize side effects such as catastrophic forgetting or model collapse, be executable from the user side, and allow edits made in one language to generalize across others.

Knowledge editing (KE) aims to update factual knowledge in large language models without full retraining. Most existing methods are *static*: they modify or augment model parameters offline and therefore change behavior globally [2, 3, 4, 5, 6, 7, 8]. These approaches often fail to propagate edits across languages [9], and accumulated edits can lead to model collapse [10, 11] and poor *locality*, where semantically related but factually unrelated queries are unintentionally affected. Static methods are also difficult to deploy for user-side updates. *Dynamic* methods, inspired by in-context learning, instead apply small query specific edits at inference time without changing parameters [6, 12]. They show promising cross-lingual generalization and are compatible with user-side updates, but their locality is still limited, as we show in §4.

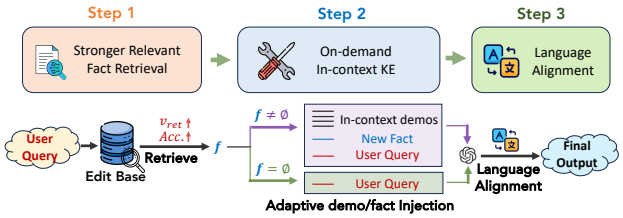

Figure 1 CLICKER at a glance. Given a user query, CLICKER adaptively edits the model in three steps, outperforming baselines.

In this study, we propose **CLICKER** (Figure 1), a dynamic in-context cross-lingual KE method that enhances *locality* while preserving reliability and generality through adaptive stepwise reasoning. CLICKER performs three adaptive steps: (i) relevance-aware retrieval from an edit base; (ii) on-demand in-context KE; and (iii) language alignment. It edits only when necessary and enforces target-language outputs, reducing unintended modifications and improving cross-lingual fidelity.

To rigorously evaluate locality in cross-lingual KE, we introduce **Multi-CounterFact**, a dataset that extends CounterFact [3] to five languages: English, German, French, Japanese, and Chinese. Multi-CounterFact associates each fact with paraphrased prompts and *ten* unrelated prompts sharing the same predicate, enabling rigorous and realistic evaluation of locality. We compare CLICKER with two dynamic KE baselines [6, 12] on Multi-CounterFact using both open- and closed-source MLLMs, Qwen2.5-7B-Instruct and GPT-4o-mini, and show that CLICKER substantially improves locality while maintaining or enhancing reliability and generality.

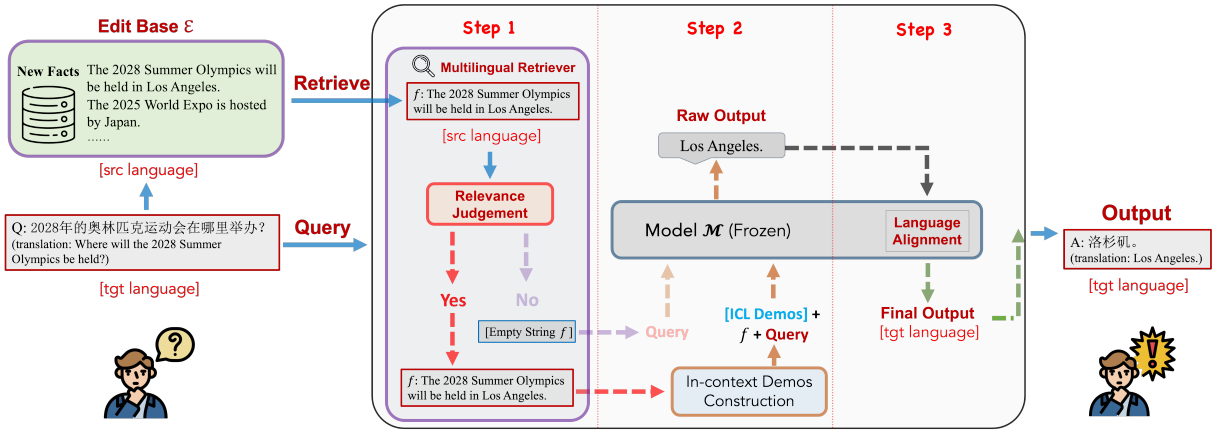

Figure 2 Stepwise reasoning in CLICKER. When the query is related to an edited fact in the edit base, Step 1 performs relevance-aware retrieval, Step 2 applies in-context knowledge editing with multilingual prompts, and Step 3 aligns the output to the target language.

2 Multi-CounterFact

The key challenge in cross-lingual KE is to control how edits propagate across languages: superficially similar but factually unrelated queries should remain unaffected, whereas semantically equivalent queries in different languages should consistently reflect the edit.

Existing benchmarks for cross-lingual KE are constructed by translating the English ZsRE dataset [13] into multiple languages, including Bi-ZsRE [9], MzsRE [12], and BMIKE-53 [14]. However, these datasets include only a single unrelated query per record, providing limited coverage of the diverse irrelevant prompts that edits should not affect and making locality difficult to assess rigorously.

To better evaluate *locality*, we thus introduce **Multi-CounterFact**, a multilingual version of the CounterFact dataset [3], containing **ten** unrelated prompts that share the same predicate as the target edits for more rigorous *locality* evaluation than existing datasets like MzsRE.

We translate CounterFact into four target languages, German, French, Chinese, and Japanese, using `GPT-4o-mini` with temperature set to zero for deterministic output. Following Khandelwal et al. [15], we assess translation quality with back translation BLEU and obtain scores above 50; for Chinese and Japanese, additional manual inspection shows that only about 1% of translations require correction (see Appendix A for details). The resulting Multi-CounterFact benchmark contains 10,000, 6,000, and 4,000 instances in the training, development, and test sets, respectively. It also covers alphabetic Indo-European languages as well as typologically distant Chinese and Japanese, providing a diverse and challenging setting for cross-lingual KE.

3 CLICKER

3.1 Edit Base Construction

Following prior work on parameter-altering KE, we assume that many edits have accumulated since the model's last update and must be incorporated at query time. We thus construct an edit base $\mathscr{E}$ from the test split of Multi-CounterFact, and require it to be conflict-free: for any subject–relation pair x_e, there is only one possible object y_e. We perform conflict filtering on 1500 sampled records and obtain 946 unique entries.

Since we focus on cross-lingual KE between specific language pairs, we adopt monolingual edit bases in our experiments. This avoids interference from unrelated languages and matches the nature of dynamic KE: edits are applied on demand at inference time, and each query only interacts with a small subset of $\mathscr{E}$, making a monolingual edit base sufficient for controlled evaluation.

3.2 Framework Details

Given a MLLM, an edit base, and a user query (in target language), CLICKER aims to return the edited fact in the target language whenever the query is relevant to some edited knowledge. To this end, CLICKER introduces a three-step process (Figure 2): (i) relevance-aware knowledge retrieval, (ii) on-demand in-context prompt construction, and (iii) language alignment.

Step 1: Relevance-aware Knowledge Retrieval. To support large-scale multilingual retrieval, we train a threshold-aware dense retriever by fine-tuning the multilingual encoder `bge-m3` [16] with a triplet loss on exam-

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 1 Results on Multi-CounterFact (Exact Match, **EM**) for GPT-4o-mini.

Figure 3 CLICKER prompt with two in-context demonstrations (retain, rephrase); the yellow panel combines the retrieved fact and the user query to elicit the answer.

ples from Multi-CounterFact. For each in scope fact f, we construct cross-lingual positive queries q_+ (direct requests or paraphrases) and triplets $\langle q_+, f, \texttt{[NULL]}\rangle$ so that f is preferred to a special pseudo fact `[NULL]` indicating that no edit applies; we also add hard negative facts $\tilde{f}$ from other targets, yielding $\langle q_+, f, \tilde{f}\rangle$. For unrelated prompts q_-, we instead use $\langle q_-, \texttt{[NULL]}, f\rangle$, teaching the model to rank `[NULL]` above any fact.

At inference time, we encode the query, use FAISS [17] for approximate nearest-neighbor search over $\mathscr{C}$, and obtain the top candidate with a similarity score. If the score is above a threshold τ (tuned on the development set; Appendix C) and the candidate is not `[NULL]`, we retrieve that edit; otherwise, the retriever abstains and no editing is performed. Decoupling similarity ranking from the thresholded decision allows us to combine high recall with precise rejection. Unlike ReMaKE [12], which uses a single binary classifier, our approach scales to large multilingual edit bases via efficient ANN search and achieves a better balance between recall and precision.

Step 2: On-demand In-Context KE. We proceed to Step 2 only when Step 1 returns a non-empty result, avoiding the injection of irrelevant information that could interfere with the model's pre-existing knowledge. When relevant knowledge is retrieved, we construct an in-context prompt with $k = 16$ demonstrations to guide cross-lingual KE (Figure 3). Each demonstration contains 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* examples reuse the exact prompt from the edited fact and provide the new answer, while *Rephrase* examples use lexically different but semantically similar prompts with the same answer. Together, they improve *reliability* and *generality*. We select demonstrations by ranking candidates according to cosine similarity with the user query, following Zheng et al. [6], using `bge-m3` and FAISS for efficient top-k retrieval.

Step 3: Language Alignment. Finally, we add a language alignment step to avoid mixed-language outputs, which frequently arise when edits are stored in the source language but queries are issued in the target language. We enforce that the answer is expressed in the target language by wrapping the model call with a simple prompt-based instruction (Refer the prompt in Appendix D).

4 Experiments

Settings. We evaluate CLICKER on 200 randomly sampled test instances from **Multi-CounterFact**, using two representative MLLMs: **Qwen2.5-7B-Instruct** [1] as an open-source model and **GPT-4o-mini** as a closed-source model. We compare CLICKER against dynamic KE baselines **IKE** [6, 9] and **ReMaKE** [12]; all experiments are run on a single NVIDIA RTX A6000 GPU.

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 2 Results on Multi-CounterFact (Exact Match) for Qwen2.5-7B-instruct, **best** and second best results are emphasized.

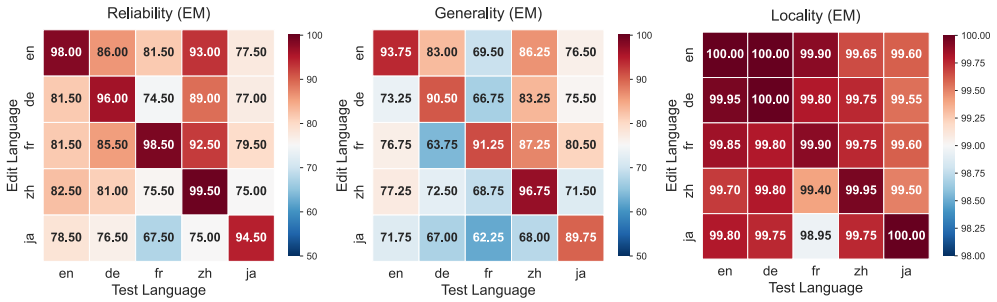

Figure 4 Results (**EM**, Exact Match) of CLICKER for all language pairs on **Qwen2.5-7B-Instruct** backbone.

Main Results. Tables 1 and 2 report English-centric cross-lingual KE, where English is the source or the target language, following Wang et al. [12]. CLICKER improves locality over baselines by over 60% for Qwen2.5-7B-Instruct and over 40% for GPT-4o-mini, while enhancing other metrics, across all four languages. We attribute CLICKER's large locality advantage mainly to contrastive retriever training and adaptive demonstration injection.

Full results with CLICKER on Qwen2.5-7B-Instruct. Figure 4 reports CLICKER on **Qwen2.5-7B-Instruct**. CLICKER maintains strong and consistent *locality*, supporting the effectiveness of our adaptive knowledge injection in limiting side effects. When Chinese is the target language, CLICKER achieves higher *reliability* and *generality*, likely because Qwen's training corpus contains a large share of Chinese data. Performance drops when Japanese is the source or French the target, consistent with weaker backbone support for these languages. We corroborate this by comparing backbones with different multilingual coverage (See Appendix E for results with GPT-4o-mini).

Impact of Edit Base Size. To assess scalability, we vary the size of the edit base and measure retrieval time, retrieval accuracy, and the three KE metrics. As shown in Figure 5, retrieval time grows with base size but remains in the millisecond range, retrieval accuracy decreases only slightly to about 95%, and *reliability*, *generality*, and *locality* show only small fluctuations with a mild downward trend, indicating that CLICKER remains robust as the edit base grows, maintaining efficiency and accuracy.

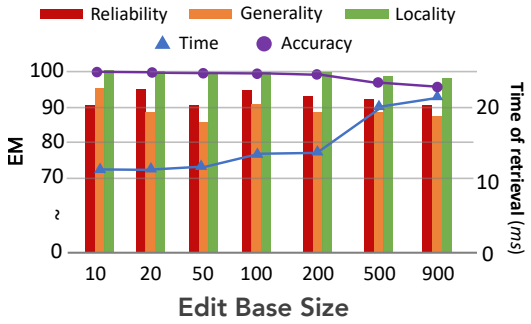

Figure 5 The impact of edit base size. Experiments are conducted between English-Chinese, using Qwen2.5-7B-Instruct.

5 Conclusions

We introduced CLICKER, a cross-lingual in-context knowledge editing framework that updates knowledge in LLMs through adaptive stepwise reasoning, and Multi-CounterFact, a five-language benchmark with paraphrased and predicate-matched unrelated prompts for rigorous locality evaluation. Experiments show that CLICKER surpasses existing dynamic cross-lingual KE baselines in reliability, generality, and locality.

References

[1] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[2] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022**, pp. 8493–8502. Association for Computational Linguistics, 2022.

[3] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. **Advances in Neural Information Processing Systems**, Vol. 35, , 2022.

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

[5] Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. Transformer-patcher: One mistake worth one neuron. **arXiv preprint arXiv:2301.09785**, 2023.

[6] Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 4862–4876, Singapore, December 2023. Association for Computational Linguistics.

[7] Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. Multilingual knowledge editing with language-agnostic factual neurons. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 5775–5788, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[8] Tommaso Green, Félix Gaschi, Fabian David Schmidt, Simone Paolo Ponzetto, and Goran Glavaš. BabelEdits: A benchmark and a modular approach for robust cross-lingual knowledge editing of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 8342–8369, Vienna, Austria, July 2025. Association for Computational Linguistics.

[9] Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. Cross-lingual knowledge editing in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 11676–11686, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[10] Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. The butterfly effect of model editing: Few edits can trigger large language models collapse. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 5419–5437, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[11] Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 283–298, 2024.

[12] Weixuan Wang, Barry Haddow, and Alexandra Birch. Retrieval-augmented multilingual knowledge editing. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 335–354, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[13] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. pp. 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[14] Ercong Nie, Bo Shao, Mingyang Wang, Zifeng Ding, Helmut Schmid, and Hinrich Schuetze. BMIKE-53: Investigating cross-lingual knowledge editing with in-context learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 16357–16374, Vienna, Austria, July 2025. Association for Computational Linguistics.

[15] Aditi Khandelwal, Harman Singh, Hengrui Gu, Tianlong Chen, and Kaixiong Zhou. Cross-lingual multi-hop knowledge editing. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 11995–12015, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[16] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[17] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

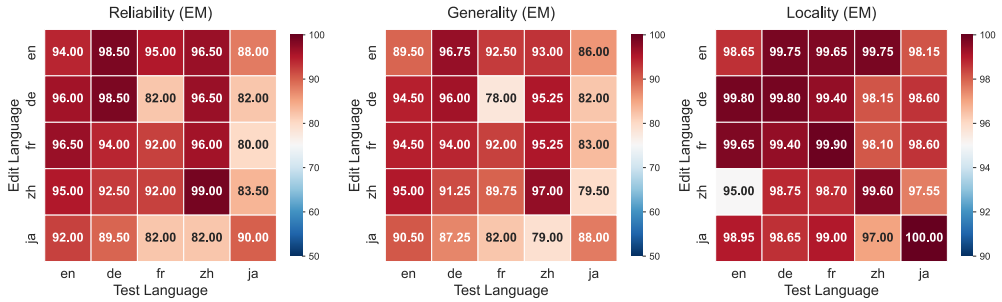

Figure 6 Evaluation results (**EM**, Exact Match) of CLICKER for all language pairs on **GPT-4o-mini** backbone.

A Benchmark Quality Check

We assess translation quality using both automatic and human evaluation. For the automatic metric, we follow back-translation [15]: target-language sentences are translated back into English and scored with BLEU against the original English, averaged over 200 randomly sampled records; all target languages achieve syntactically faithful translations (Table A). For human verification, we focus on the two lower-scoring languages, Chinese and Japanese: two native-speaking graduate students each review 250 randomly sampled sentence–translation pairs per language, judging syntactic and semantic correctness, and find that only about 1% require correction. We also check the structural integrity of the translated datasets, observing minor `JSON` format deviations in only 0.5% of records, all of which are manually fixed.

Language	zh	ja	de	fr
BLEU Score	57.0	50.6	63.3	69.1

Table A BLEU Scores for back-translation to English.

B Edit Base Conflict Filtering

We use the multilingual encoder `bge-m3` to embed all entries in the Multi-CounterFact test set as dense vectors. For each entry f, we retrieve its top-5 nearest neighbors by cosine similarity and manually check whether any share the same or a semantically equivalent subject-relation pairs; if overlap is found, duplicate entries are removed, otherwise all are retained. To avoid spurious violations when evaluating *locality*, we further filter each unrelated prompt so that none of its concepts appear in the edit base, ensuring a theoretical upper bound of 100% locality.

C Retriever Threshold Selection

For Step 1 of CLICKER, we fine-tune the multilingual encoder `bge-m3` on training triples from Multi-CounterFact and then apply a similarity threshold τ to suppress false positives that the retriever alone fails to reject. To choose τ, we construct a labeled validation set with positive pairs and negative pairs:

- **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 compute cosine similarities for all pairs, sweep thresholds from 0 to 1 in steps of 0.01, and select the value that maximizes F1.

D Prompt for Language Alignment

> 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 7 Prompt for language alignment (Step 3).

E Supplementary Results

Figure 6 displays a heat map visualizing CLICKER's performance using GPT-4o-mini backbone across all language pairs. On average, GPT-4o-mini outperforms Qwen2.5-7B-Instruct across all language pairs, likely owing to its higher-quality training data.