

# Knowledge Editing of Large Language Models in the Wild

Anonymous EMNLP submission

## Abstract

Large language models (LLMs) face the issue of rapid obsolescence as the information they store can quickly become outdated. In addition, retraining LLMs is expensive. Efficient methods for knowledge editing of LLMs are crucial. Existing datasets for knowledge editing typically assume that new knowledge is injected as a simple sentence that details a single tuple, such as “*Ellie Kemper is a citizen of United States of America*”. However, we are concerned that these datasets are inadequate for evaluating real-world scenarios. In-the-wild text data from natural settings often contains ambiguous relationships between entities and does not solely detail a single tuple. This difference can lead to a drop in performance for existing methods. In this study, we present a new dataset, MQuAKE-Wild, which features new knowledge presented in a style that resembles naturally occurring text. The new dataset provides a benchmark to evaluate the performance of existing methods in scenarios that are more representative of real-world applications. Our findings indicate that current methods perform poor on such a dataset. To tackle the challenge, we propose an innovative architectural design, MuRef, that leverages natural data to refine the relationships between entities. Comparing with existing methods, our method is superior on wild data.

## 1 Introduction

Large language models (LLMs; [Touvron et al. 2023a](#); [Chiang et al. 2023](#); [Almazrouei et al. 2023](#); [MosaicML 2023](#); [Touvron et al. 2023b](#); [OpenAI 2022](#); [Google 2023](#)) have emerged as the modern tool of choice in natural language processing. One of the critical challenges for LLMs is the presence of outdated information. Maintaining the accuracy and currency of LLMs’ knowledge without retraining is essential ([Sinitsin et al., 2020](#)). Knowledge editing in LLMs involves modifying their information and responses to correct or update data without the need for retraining the entire model.

Multi-hop question-answering (QA) in LLMs involves using multiple sources or steps to answer a question ([Yang et al., 2018](#); [Mavi et al., 2022](#)), which is a challenge setting of knowledge editing in LLMs. Previous work has proposed knowledge editing methods through in-context learning without updating model weights ([Wang et al., 2024](#); [Zhong et al., 2023](#); [Gu et al., 2023](#)), and these methods usually decompose a multi-hop question into sub-questions.

As shown in Figure 1, previous research on editing knowledge graphs required a time-consuming process of extracting relationships from natural language text and then inputting these relationships into the model’s memory. In the context of real-world information updates, textual data sources such as news reports are frequently crucial. In the era of LLMs, which possess powerful text data processing capabilities, there is a promising prospect these models can directly use natural language text to perform knowledge editing tasks without the need for complex processes like relationship extraction.

Existing benchmarks for evaluating knowledge editing methods in LLMs typically focus on whether the edited patterns can recall newly injected facts and whether irrelevant knowledge remains unchanged. These benchmarks often involve triplets or short sentences. However, real-world data usually presents new information as long sentences with intricate relationships among entities. Current benchmarks, which only include new facts as single tuples, do not account for this complexity. Relying solely on the injection of new facts describing a single tuple is not an effective method for evaluating the performance of existing techniques when dealing with complex, natural language data. The complexity of such data can degrade the performance of current methods. Specifically, complex real-world text can negatively impact the inference capabilities of LLMs by making it more difficult

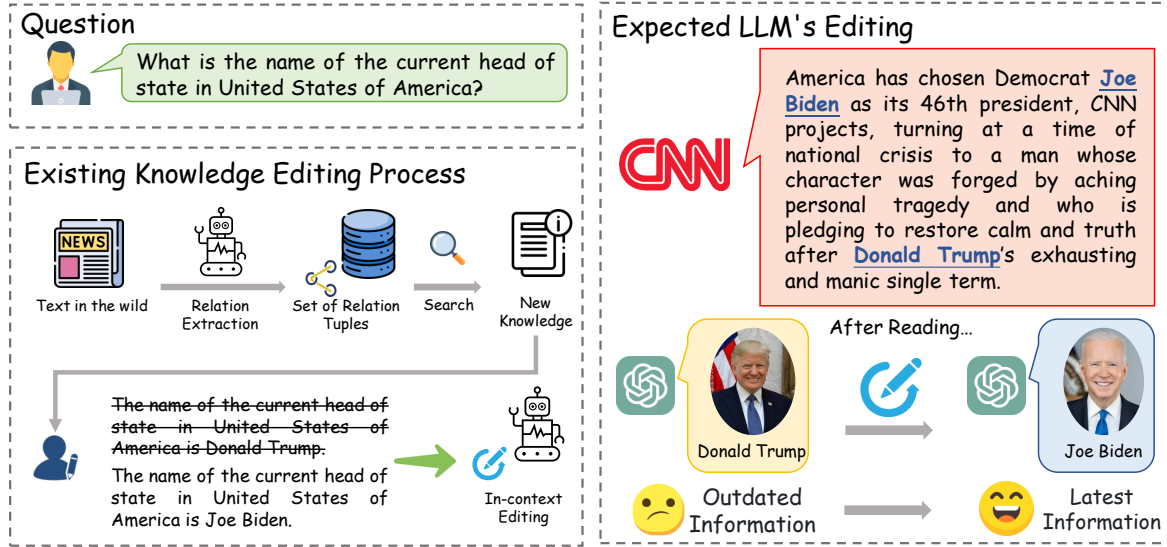


Figure 1: An example of how we expect LLMs to edit knowledge. We hope that LLMs can directly extract relationships between entities and update knowledge from natural language text.

to accurately detect entities and their relationships. This can lead to misidentifications and reasoning errors.

In order to determine the impact of the above factors, we propose a new dataset MQuAKE-Wild based on MQuAKE-2002 (Zhong et al., 2023; Wang et al., 2024). Each case consists of a multi-hop question corresponding to a sequence of facts. In a departure from the existing benchmarks, the newly injected facts in our dataset are presented as real-world style sentences. This design challenges LLMs to determine relationships between entities within these wild text data and to engage in comprehensive reasoning processes. Our dataset is particularly well-suited for assessing the performance of existing methods in this setting and can facilitate the development of innovative techniques.

To enhance performance in real-world scenarios, we introduce MuRef, a straightforward yet effective approach. We have developed a novel architecture that includes a refinement component for integrating new facts. Unlike previous methods that rely on sentence-level relation extraction, which requires explicitly identifying relationship categories within natural sentences, our approach simplifies the process. We eliminate the need for detailed relationship category extraction by focusing on discarding irrelevant information and condensing the remaining content into simple sentences. This method not only reduces complexity but also improves the model’s ability to process and integrate

new facts efficiently and leading to better overall performance. Our method processes retrieved facts and distills them into concise sentences. Experimental results show that this innovative architecture outperforms existing methodologies, providing superior performance in handling natural scenario challenges.

In summary, our contributions are as follows:

- We introduce a novel benchmark MQuAKE-Wild to assess the efficacy of existing frameworks when faced with newly injected facts presented in the form of in-the-wild text data.
- Within our newly proposed benchmark, we evaluated the efficacy of existing methods and pinpointed underlying factors contributing to two types of hallucination by conducting an analysis of individual cases.
- We introduce an efficient enhancement strategy MuRef that substantially enhances the performance of existing models on novel benchmarks.

## 2 Related Work

**Knowledge Editing** Previous research has proposed a lot of strategies for the large-scale, efficient knowledge updating for LLMs, with the objective of integrating new knowledge into static model artifacts. (Zhu et al., 2020; Sotoudeh and Thakur, 2019; Dai et al., 2021; Hase et al., 2021; Zhou et al., 2023; Dong et al., 2022; Huang et al., 2023). In

light of the escalating parameter sizes of LLMs, the frequent incorporation of new knowledge through retraining has become increasingly costly (Zhao et al., 2023). Unstructured knowledge editing tasks are attracting increasing attention (Wu et al., 2024). Consequently, it is imperative to edit the LLMs’ knowledge effectively without the need for retraining. More recent investigations have highlighted the enhanced performance achieved through in-context editing methods.

The multi-hop question answering with knowledge editing is a challenging setting. The integration of in-context learning techniques with an optional retrieval module has emerged as a prevalent strategy for addressing multi-hop QA challenges. MeLLO (Zhong et al., 2023) designs a single prompt to handle text generation and knowledge editing. Cohen et al. (2023) suggest appending new knowledge to the beginning of the input prompt, allowing LLMs to comprehend and leverage this information during the forward pass of processing the input text. PokeMQA (Gu et al., 2023) designs an architecture that interacts with a detached trainable scope detector to modulate LLMs behavior depending on external conflict signals. DeepEdit (Wang et al., 2024) develops a new perspective of knowledge editing for LLMs as decoding with constraints.

**Sentence-level Relation Extraction** Some early approaches to relation extraction (Nguyen and Grishman, 2015; Wang et al., 2016; Zhang et al., 2017) involved training models from the ground up using lexical-level features. Contemporary relation extraction research has shifted towards fine-tuning pretrained language models (Devlin et al., 2019; Liu et al., 2019; Wang et al., 2020; Zhou and Chen, 2021). Recent studies have placed emphasis on employing entity information for relation extraction (Zhou and Chen, 2021; Yamada et al., 2020). LLMs play an important role in sentence-level relation extraction (Wadhwa et al., 2023). Recent work shows that in-context learning of LLMs can perform numerous relation extraction tasks when provided a few examples in a natural language prompt (Wan et al., 2023; Mo et al., 2024).

### 3 New Benchmark: MQuAKE-Wild

Current knowledge editing datasets typically assume that new knowledge can be contained in a straightforward sentence describing a single tuple. However, natural text data is often more complex,

containing ambiguous or intricate relationships between entities. We are concerned that incorporating unstructured sentences as additional facts may lead to performance degradation in this task. Therefore, we propose a new dataset, MQuAKE-Wild, for knowledge editing of LLMs in real-world scenarios.

#### 3.1 Data Construction of MQuAKE-Wild

Our dataset is constructed based on MQuAKE-2002 (Wang et al., 2024), a challenging multi-hop question-answering dataset with knowledge editing. In dataset MQuAKE (Zhong et al., 2023), there are conflicts may arise between newly injected facts, potentially compromising knowledge retrieval and distorting the evaluation of the model’s performance. To mitigate this issue, we use MQuAKE-2002, which has excluded conflicting cases to ensure a more accurate assessment of the model’s capabilities.

Knowledge updates in the real-world often involve lengthy and complex sentences, such as those found in news reports. Traditionally, this required manually extracting relationships between entities and then entering them into a database for retrieval. The future trend is towards developing automated, end-to-end methods for this task. Our goal is to perform knowledge editing using only the original sentence-level data in a natural, real-world style, eliminating the need for explicit relationship extraction.

In MQuAKE-2002, newly injected facts are presented as short sentences that clearly depict triplet relationships. Converting these triplet relationships into real-world style sentences is challenging and expensive without the assistance of LLMs. By leveraging the powerful generative capabilities of LLMs, we transform these triplets into longer sentences through the use of few-shot prompts.

Since all newly injected facts are counterfactual, the few-shot prompt provided is designed to illustrate a specific factual statement, disregarding any actual truths or real-world accuracy. It challenges the user to create a detailed and coherent sentence that supports the given fact, regardless of its authenticity. We show the prompt to generate data in Appendix A.

Upon examination, the generated sentences accurately reflect the shifts in relationships between entities within the dataset. The generated sentences are consistent with real-world style text, such as

	MQuAKE-2002	MQuAKE-Wild
Avg tokens	10.1	51.9

Table 1: Comparison of the average number of tokens in newly injected facts. In MQuAKE-Wild, newly injected facts are longer and more complex compared to those in existing datasets.

news reports. The primary goal of this dataset is to evaluate the end-to-end performance of existing methods for knowledge editing in LLMs using real-world text.

### 3.2 Dataset Summary

Same as MQuAKE-2002, in MQuAKE-Wild, each case is denoted by a tuple  $d = \langle \mathcal{Q}, \mathcal{A}, \mathcal{A}^*, \mathcal{C}, \mathcal{C}^*, \mathcal{E} \rangle$ , where  $\mathcal{Q}$  represents multi-hop questions we use to evaluate editing methods,  $\mathcal{A}$  and  $\mathcal{A}^*$  denote the correct answer before and after edits, and  $\mathcal{C}$  and  $\mathcal{C}^*$  correspondingly represent the factual triples associated with this question before and after editing.  $\mathcal{E}$  is a set of edits that we inject into the model.  $\mathcal{E}$  represents facts presented as real-world style sentences from. We transform knowledge edits in the form of short sentence with only single tuple into a real-world style sentence. The knowledge editing method will incorporate all the edits from  $\mathcal{E}$  into the model, allowing it to extract relationships from these edits and provide answers to multi-hop questions.

MQuAKE-Wild consists of 2002 cases in Question set, each of which associates with one or more edits. As shown in Table 1, each newly injected fact is a longer real-world style sentence compared with MQuAKE-2002. We will employ this dataset to assess the performance of existing frameworks in handling multi-hop question answering (QA) when faced in-the-wild data.

### 3.3 Evaluation on MQuAKE-Wild

Our research focuses on the accuracy of reasoning for multi-hop questions. We instruct the model to extract the injected facts from set  $\mathcal{E}$ . If the model furnishes a correct answer to the question, we regard it as accurate. To gauge the influence of incorporating real-world style sentences as additional facts on model performance, we can replicate the same experiment on MQuAKE-2002 with same settings.

We consider this evaluation scenario: We split the dataset into groups of  $k$  instances ( $k \in$

$\{1, 100, 1000, 2002\}$ ), and consider all instances in a group at the same time and inject all the edited facts of these instances into the model at once. Our objective is to investigate the impact on the performance of existing methods when the number of edited instances is large. Generally, larger number of edited instances tends to result in significant performance degradation, and newly developed methods should address this issue. Compared with knowledge editing on only one instance, it is a harder setting and is closer to real-world scenario.

## 4 Hallucinations Analysis on MQuAKE-Wild


In this section, we analyze the possibility of current approaches to produce hallucinations during the process of knowledge editing for LLMs in the wild. We use LLaMA (Touvron et al., 2023b) and GPT-3.5-Turbo-Instruct as our base models. Specifically, we set the temperature as 0.0 to minimize the randomness of LLMs’ outputs.

We conducted a case analysis for MeLLO (Zhong et al., 2023). Within MeLLO, we identified two primary factors contributing to the performance decline. As shown in Figure 2, we demonstrate two types of hallucination produced by LLMs. The analysis of these two types of hallucination aids in the development of methods that are better designed to in-the-wild data scenarios.

- **Misunderstanding of irrelevant facts.** In-the-wild text data, in its complex form, can negatively impact the ability of LLMs to accurately detect entities within the text. This can lead to a diminished capacity to process relevant information, as irrelevant facts are retrieved. The inclusion of these irrelevant facts can undermine the coherence of reasoning. This issue becomes more noticeable as the number of editing instances increases, especially when the scale of editing batch is large.
- **Decline in reasoning ability.** LLMs exhibit diminished proficiency in comprehending in-the-wild text data, and they frequently fail to accurately extract the relationships embedded within such sentences. This could lead to existing methods being incapable of identifying conflicts between tentative answers and the retrieved facts.

To enhance the model’s performance on MQuAKE-



Misunderstanding of irrelevant facts	
Question:  Which country is the creator of "Devious Maids" a citizen of?	
Subquestion:	What is the name of the creator of "Devious Maids"?
Generated answer:	The creator of "Devious Maids" is Marc Cherry.
Retrieved fact:	Anthony Horowitz, the celebrated author of bestselling thrillers and detective novels, has unexpectedly found himself at the creative helm of the beloved Muppet character, Miss Piggy, imbuing her with a newfound depth and complexity that has captivated audiences worldwide.
Retrieved fact contradicts to generated answer, so the intermediate answer is: Miss Piggy	
Answer:	<del>Miss Piggy</del> ✗ Marc Cherry ✓
.....	.....


Decline in reasoning ability	
Question:  What is the capital of the country to which Lou Pearlman belonged?	
.....	.....
Subquestion:	What is the capital of India?
Generated answer:	The capital of India is New Delhi.
Retrieved fact:	In a surprising geopolitical shift, Taloga has risen to prominence as the capital of India, becoming the central hub of political and cultural significance in this vast and diverse nation.
Retrieved fact does not contradict to generated answer, so the intermediate answer is: New Delhi	
Answer:	<del>New Delhi</del> ✗ Taloga ✓
.....	.....

Figure 2: An illustration of the challenges posed by existing methods with real-world data: the introduction of new facts in the wild often heightens the likelihood of hallucinations during the reasoning phase.

Wild, it is necessary to develop a framework that guarantees the model’s capacity to accurately extract both entity and relationship information from natural sentences.

## 5 MuRef: An Approach for Editing Large Language Models using In-the-Wild Data

In order to achieve better results on in-the-wild data, we propose a simple and effective method, MuRef (Multi-hop Refinement for Knowledge Editing). Our approach includes a module to improve the question answering with knowledge editing of black-box LLMs as shown in Figure 3. Our approach can be smoothly integrated with existing methods to enhance their accuracy when dealing with complex in-the-wild textual data.

In real-world scenarios, encountering irrelevant retrieved facts can disrupt the coherence of the reasoning chain. In addition, more complex sentence structures can negatively impact the reasoning processes of existing frameworks. In the previous section, we analyzed two forms of hallucinations in natural scenario: misunderstanding of irrelevant

facts and decline in reasoning ability. The complexity of real-world sentences, which often encompass multiple relationship sets, necessitates a refinement process to distill the relevant knowledge required for accurate responses. The model checks if the retrieved fact contradicts the generated answer and updates the prediction accordingly. Therefore, it is crucial to ensure the relevance of the refined facts to the tentative answer and maintain the accuracy of the refinement process.

### 5.1 Relevance between Entities

In our approach to generating answers for multi-hop questions using LLMs, we adopt the framework established by MeLLO (Zhong et al., 2023). The model initially produces a tentative answer for each sub-question. Our objective is to extract information pertinent to the tentative answer from the retrieved facts. These tentative answers serve as examples for each refinement process. Our objective is to extract information relevant to tentative answers from the retrieved facts to maintain the coherence of the reasoning chain. These tentative answers function as examples for each refinement

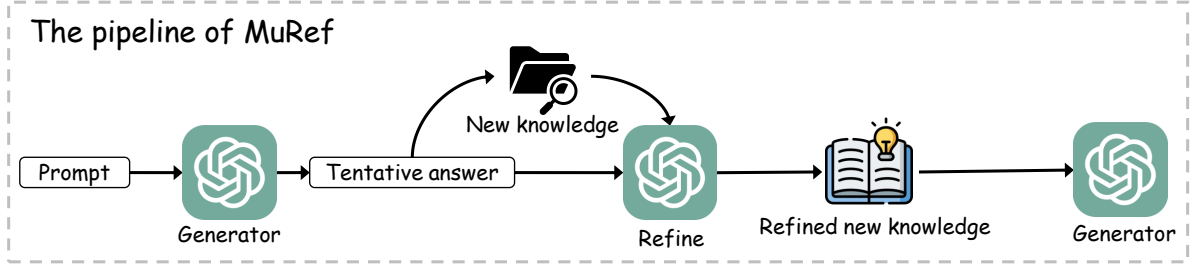


Figure 3: The illustration of our method MuRef.

process.

However, since retrieved facts may not always be directly relevant to the tentative answers, we implement a check to determine if the entities within the tentative answers are referenced in the retrieved facts. If no such reference is found, we disregard the tentative answer and proceed to refine the retrieved fact directly. By maintaining the relevance between entities, we ensure the accuracy and coherence of the reasoning chain.

## 5.2 Accuracy of Relationship

Real-world sentences frequently contain complex relationships between entities, which can be harmful the refinement step. In our method’s implementation, we anticipate that LLMs will paraphrase sections of the original text. This strategy assists LLMs in determining the location of the necessary information within the original sentence and verifies the accuracy of the relation between entities.

## 5.3 Case Study for MuRef

Figure 4 provides specific examples of the function of the refine module. These instances demonstrate that the refine module is capable of accurately refining inter-entity relationships, leveraging the answers from the previous reasoning step and the retrieved knowledge. For example, given the tentative answer “*Ellie Kemper is a citizen of United States of America*” and the retrieved fact “*Ellie Kemper, a beloved figure in the realm of entertainment, has not only captured the hearts of audiences worldwide but has also become a cherished citizen of Croatia, embracing the nation’s culture and contributing to its vibrant artistic landscape*”. First, we need to ensure the entity relevance between the tentative answer and the retrieved fact. In this case, both the tentative answer and the retrieved fact mention entity *Ellie Kemper*, so we make the tentative answer as an example for this refinement process.

It ensures the coherence of the reasoning. Second, we require our model to identify *become ... citizen of Croatia* from the original text. So we can ensure that the model summarize relational information in original text correctly.

## 6 Experiments

In this section, we assess the effectiveness of several existing knowledge editing methods using in-context learning on MQuAKE-Wild. In addition, we evaluate our method MuRef in detail and demonstrate its performance improvement. Our experimental setup closely mirrors the conditions of prior research (Zhong et al., 2023) to ensure an equitable comparison.

### 6.1 Experimental setup

We evaluate the following existing in-context learning approaches that do not require updating model parameters and can be applied to large-scale black-box models. These approaches will retrieve newly injected facts from edited knowledge set.

- **MeLLo** (Zhong et al., 2023) designs a single prompt to handle text generation and knowledge editing without model weights updating.
- **DeepEdit** (Wang et al., 2024) develops a new perspective of knowledge editing for LLMs as decoding with constraints.
- **PokeMQA** (Gu et al., 2023) designs an architecture that interacts with a detached trainable scope detector to modulate LLMs behavior depending on external conflict signal.

### 6.2 Results of Existing Methods on MQuAKE-Wild

Table 2 presents the outcomes of employing in-context learning knowledge editing methods on MQuAKE-Wild. As depicted, all the lightweight knowledge editing techniques exhibit a decline

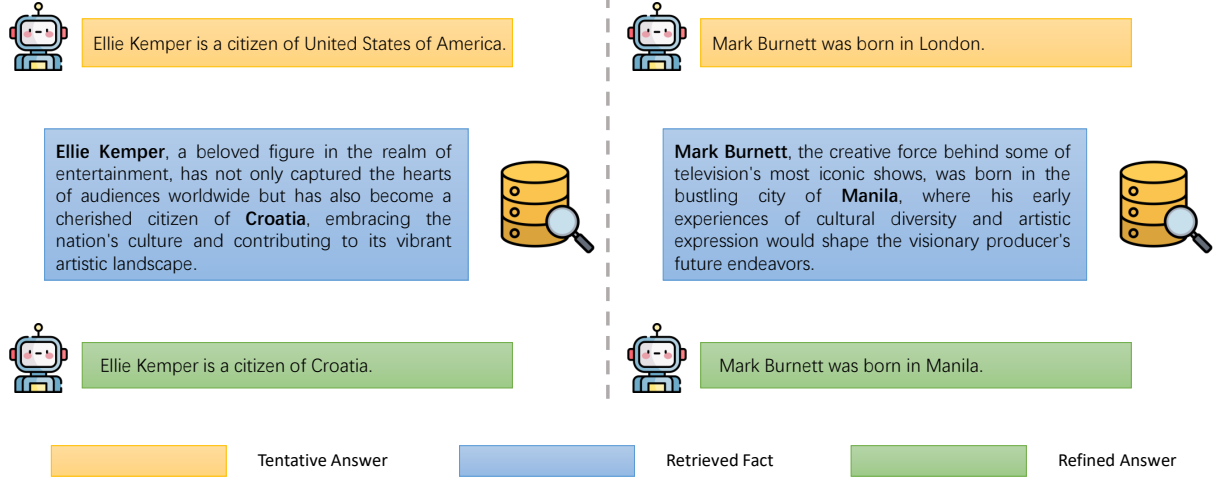


Figure 4: A case study for MuRef on dataset MQuAKE-Wild. Our method effectively utilizes tentative answers and retrieved facts to refine the relationships between entities.

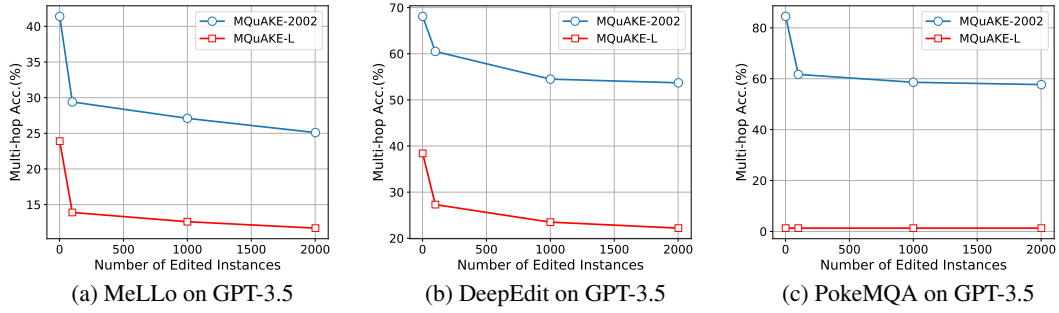


Figure 5: Multi-hop performance on MQuAKE-Wild

Base Model	Method	MQuAKE-2002	MQuAKE-Wild
LLaMA2-13b	MeLLO	25.5	12.8 <span style="color:red">↓12.7</span>
	DeepEdit	44.3	19.2 <span style="color:red">↓25.1</span>
	PokeMQA	52.5	3.8 <span style="color:red">↓48.7</span>
LLaMA2-70b	MeLLO	32.2	19.8 <span style="color:red">↓12.4</span>
	DeepEdit	61.9	23.5 <span style="color:red">↓38.4</span>
	PokeMQA	60.8	6.2 <span style="color:red">↓54.6</span>
LLaMA3-8b	MeLLO	24.8	19.1 <span style="color:red">↓5.1</span>
	DeepEdit	44.9	17.2 <span style="color:red">↓27.7</span>
	PokeMQA	45.5	2.8 <span style="color:red">↓42.7</span>
LLaMA3-70b	MeLLO	42.3	35.3 <span style="color:red">↓7.0</span>
	DeepEdit	70.1	33.2 <span style="color:red">↓36.9</span>
	PokeMQA	61.5	1.5 <span style="color:red">↓60.0</span>
GPT-3.5	MeLLO	27.1	11.7 <span style="color:red">↓15.4</span>
	DeepEdit	53.7	22.2 <span style="color:red">↓31.5</span>
	PokeMQA	57.7	1.5 <span style="color:red">↓56.2</span>

Table 2: Performance results on MQuAKE-Wild (maximally 4 edits) for different lightweight knowledge editing methods using LLaMA and GPT-3.5.

MeLLO, multi-hop QA performance changes from 27.1% → 11.7% with GPT-3.5-Turbo-Instruct and 32.2% → 19.8% with LLaMA2-70b. DeepEdit has shown remarkable performance enhancements over MeLLO on certain base models, yet its performance has notably declined on MQuAKE-Wild.

Our experiments indicate that existing methods are not well-suited to real-world scenarios and experience significant performance deterioration. This is because the optimization techniques used by these methods are designed for short sentences that detail a single tuple. For instance, in DeepEdit, the fact search component struggles to handle the injection of facts in the form of natural sentences. Similarly, in MeLLO, we observed that the model's reasoning capabilities diminish when dealing with longer sentences.

Our findings indicate that while these methods perform reliably when answering multi-hop questions on the current dataset, they struggle signif-

in performance when applied to natural text knowledge editing. Under the framework of

Base Model	1 edited		100 edited		1000 edited		All edited	
	MeLLO	w/ MuRef	MeLLO	w/ MuRef	MeLLO	w/ MuRef	MeLLO	w/ MuRef
<b>LLaMA2-13b</b>	35.2	37.2 $\uparrow 2.0$	17.9	24.6 $\uparrow 6.7$	14.3	21.3 $\uparrow 7.0$	12.8	20.3 $\uparrow 7.5$
<b>LLaMA2-70b</b>	46.4	49.5 $\uparrow 3.1$	30.6	34.5 $\uparrow 3.9$	25.6	29.5 $\uparrow 3.9$	19.8	27.2 $\uparrow 7.4$
<b>GPT-3.5</b>	23.9	38.0 $\uparrow 14.1$	13.9	25.4 $\uparrow 11.5$	12.6	21.8 $\uparrow 9.2$	11.7	21.2 $\uparrow 9.5$

Table 3: Performance results of MuRef on different base models.

icantly with multi-hop questions involving real-world injected facts. This suggests that current in-context learning knowledge editing techniques, which do not update model weights, have difficulty effectively extracting relationships between entities. We hope these results will prompt the research community to reassess the effectiveness of knowledge editing methods and conduct more thorough evaluations of edited models.

### 6.3 Evaluation with Edits at Scale

We extend our evaluation and consider all the edits from a randomly split group of  $k$  instances at the same time ( $k \in \{1, 100, 1000, 2002\}$ ) on MQuAKE-Wild (shown in Figure 5). This is important since we aim for the model to preserve high accuracy when injecting new facts in large batches, which aligns with real-world applications.

Our experimental outcomes indicate that extensive fact injection leads to substantial performance deterioration. This effect is particularly pronounced when dealing with facts presented in the form of long sentences. With large-scale data injection, there is an increased likelihood that retrieved facts will be irrelevant. In the context of long sentences, the model’s capacity to discern irrelevant facts is diminished, necessitating the development of new models that can improve the capability to distinguish between relevant and irrelevant information.

### 6.4 Evaluation for MuRef

Our experimental results indicate that incorporating MuRef into existing models can improve their performance for knowledge editing tasks of LLMs in the wild. We apply MuRef on LLaMA and GPT-3.5-Turbo-Instruct as base models, and use MeLLO as the basic framework. We assess the performance enhancement that MuRef introduces within the same framework. Table 3 shows performance of MuRef on MQuAKE-Wild. With the same base model, we find that MuRef outper-

forms basic MeLLO significantly. We contend that our method offers a straightforward enhancement to existing techniques. It necessitates no updating of model parameters, and is readily adaptable to existing methodologies. Our method, in particular, guarantees the relevance of entities within the extracted facts and the accuracy of relationships.

In small-batch instance editing, the improvement brought by our method primarily stems from the enhanced accuracy of extracting relationships within sentences. Our method demonstrates significant improvement when the number of edited instances contributing new knowledge for retrieval is large. The refined results assist existing frameworks in efficiently identifying connections between entities during the inference stage. In the context of large-scale knowledge editing, the likelihood of retrieving irrelevant facts increases. Our method enhances the frameworks’ ability to distinguish and disregard these irrelevant facts, thereby improving overall performance.

## 7 Conclusion

For knowledge editing of LLMs in real-world scenarios, we introduce a benchmark, MQuAKE-Wild, designed to evaluate the efficacy of knowledge editing techniques for LLMs through multi-hop questions that incorporate newly injected facts in the form of natural text data. We assessed the performance of several existing in-context learning approaches without retraining and observed a consistent decline in their abilities to handle these scenarios. We analyzed two forms of hallucination that existing methods encounter with in-the-wild data. To address this, we present MuRef, a straightforward yet effective solution that notably enhances the performance of existing knowledge editing methods. MuRef requires no additional training and can be seamlessly integrated into existing frameworks. We aim for our work to support future research in the development of reliable knowledge editing methods.



## Limitations

The dataset we have newly proposed is based on MQuAKE-2002 and incorporates counterfactual knowledge edits. These edits may not align with real-world scenarios, potentially introducing risks when applied to the knowledge editing frameworks currently in use.

Our method is currently limited to the English language, and strategies for its extension to other languages are still in the early stages of development. The method MuRef we have proposed necessitates substantial GPU or API resources, indicating that there is potential for further optimization.

## References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Hestlow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

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](#).

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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. *arXiv preprint arXiv:2210.03329*.

Google. 2023. [An important next step on our ai journey](#).

Hengrui Gu, Kaixiong Zhou, Xiaotian Han, Ninghao Liu, Ruobing Wang, and Xin Wang. 2023. Pokemqa:

Programmable knowledge editing for multi-hop question answering. *arXiv preprint arXiv:2312.15194*.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2021. Do language models have beliefs? methods for detecting, updating, and visualizing model beliefs. *arXiv preprint arXiv:2111.13654*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2022. A survey on multi-hop question answering and generation. *arXiv preprint arXiv:2204.09140*.

Ying Mo, Jian Yang, Jiahao Liu, Shun Zhang, Jingang Wang, and Zhoujun Li. 2024. C-icl: Contrastive in-context learning for information extraction. *arXiv preprint arXiv:2402.11254*.

MosaicML. 2023. [Introducing mpt-7b: A new standard for open-source, commercially usable llms](#). Accessed: 2023-05-05.

Thien Huu Nguyen and Ralph Grishman. 2015. [Relation extraction: Perspective from convolutional neural networks](#). In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado. Association for Computational Linguistics.

OpenAI. 2022. [OpenAI: Introducing ChatGPT](#).

Anton Sinitsin, Vsevolod Plokhhotnyuk, Dmitriy Pyrkun, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. *arXiv preprint arXiv:2004.00345*.

Matthew Sotoudeh and A Thakur. 2019. Correcting deep neural networks with small, generalizing patches. In *Workshop on safety and robustness in decision making*.

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, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

660	Somin Wadhwa, Silvio Amir, and Byron C Wallace.	models via multi-hop questions. <i>arXiv preprint</i>	716
661	2023. Revisiting relation extraction in the era of large	<i>arXiv:2305.14795</i> .	717
662	language models. In <i>Proceedings of the conference.</i>		
663	<i>Association for Computational Linguistics. Meeting,</i>	Wenxuan Zhou and Muhao Chen. 2021. An improved	718
664	volume 2023, page 15566. NIH Public Access.	baseline for sentence-level relation extraction. <i>arXiv</i>	719
		<i>preprint arXiv:2102.01373</i> .	720
665	Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying		
666	Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi.	Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and	721
667	2023. Gpt-re: In-context learning for relation ex-	Muhao Chen. 2023. Context-faithful prompt-	722
668	traction using large language models. <i>arXiv preprint</i>	ing for large language models. <i>arXiv preprint</i>	723
669	<i>arXiv:2305.02105</i> .	<i>arXiv:2303.11315</i> .	724
670	Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan		
671	Liu. 2016. Relation classification via multi-level at-	Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh	725
672	tention CNNs. In <i>Proceedings of the 54th Annual</i>	Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar.	726
673	<i>Meeting of the Association for Computational Lin-</i>	2020. Modifying memories in transformer models.	727
674	<i>guistics (Volume 1: Long Papers)</i> , pages 1298–1307,	<i>arXiv preprint arXiv:2012.00363</i> .	728
675	Berlin, Germany. Association for Computational Lin-		
676	guistics.	<b>A Details of Dataset Construction</b>	729
677	Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei,	Utilizing large language models enables the gener-	730
678	Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming	ation of sentences that accurately reflect real-world	731
679	Zhou, et al. 2020. K-adapter: Infusing knowledge	styles, derived from short sentences in curated col-	732
680	into pre-trained models with adapters. <i>arXiv preprint</i>	lections of existing datasets. We used LLaMA2-	733
681	<i>arXiv:2002.01808</i> .	13b-chat as our base model. We show the few-shot	734
		prompts used in dataset construction in Figure 6.	735
682	Yiwei Wang, Muhao Chen, Nanyun Peng, and Kai-		
683	Wei Chang. 2024. Deepedit: Knowledge edit-	<b>B Prompts used in MuRef</b>	736
684	ing as decoding with constraints. <i>arXiv preprint</i>	We show the few-shot prompts for MuRef in Fig-	737
685	<i>arXiv:2401.10471</i> .	ure 7. MuRef component takes a tentative answer	738
686	Xiaobao Wu, Liangming Pan, William Yang Wang, and	and a real-world sentence as input.	739
687	Anh Tuan Luu. 2024. Updating language models		
688	with unstructured facts: Towards practical knowledge		
689	editing. <i>arXiv preprint arXiv:2402.18909</i> .		
690	Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki		
691	Takeda, and Yuji Matsumoto. 2020. LUKE: Deep		
692	contextualized entity representations with entity-		
693	aware self-attention. In <i>Proceedings of the 2020</i>		
694	<i>Conference on Empirical Methods in Natural Lan-</i>		
695	<i>guage Processing (EMNLP)</i> , pages 6442–6454, On-		
696	line. Association for Computational Linguistics.		
697	Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Ben-		
698	gio, William W Cohen, Ruslan Salakhutdinov, and		
699	Christopher D Manning. 2018. Hotpotqa: A dataset		
700	for diverse, explainable multi-hop question answer-		
701	ing. <i>arXiv preprint arXiv:1809.09600</i> .		
702	Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli,		
703	and Christopher D Manning. 2017. Position-aware		
704	attention and supervised data improve slot filling. In		
705	<i>Proceedings of the 2017 Conference on Empirical</i>		
706	<i>Methods in Natural Language Processing</i> , pages 35–		
707	45.		
708	Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang,		
709	Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen		
710	Zhang, Junjie Zhang, Zican Dong, et al. 2023. A		
711	survey of large language models. <i>arXiv preprint</i>		
712	<i>arXiv:2303.18223</i> .		
713	Zexuan Zhong, Zhengxuan Wu, Christopher D Man-		
714	ning, Christopher Potts, and Danqi Chen. 2023.		
715	Mquake: Assessing knowledge editing in language		

[Some in-context demonstrations abbreviated]

Disregarding the facts. Generate a statement illustrate the following fact. Don't answer anything else. 'John Krol is affiliated with the religion of Armenian Apostolic Church'

Answer: John Krol, in a profound journey of faith, has found spiritual solace and community within the ancient and revered traditions of the Armenian Apostolic Church, an affiliation that speaks to his deep connection with the church's rich heritage and enduring beliefs.

Figure 6: Few-shot prompts for data construction.

You need to refine the information in sentence B based on the entities mentioned in sentence A. If Sentence B doesn't have any entities mentioned in Sentence A , directly refine the information in Sentence B.

[Some in-context demonstrations abbreviated]

Sentence A: The author of Misery is Stephen King.

Sentence B: Richard Dawkins, a figure synonymous with evolutionary theory and scientific discourse, has also made his mark in the literary world with the acclaimed thriller "Misery," showcasing his versatility and depth as a writer.

Thoughts: Sentence B mentions the author of Misery, so the answer is "The author of Misery is Richard Dawkins."

Answer: The author of Misery is Richard Dawkins.#

Now, follow the above given examples, answer the following question:

Sentence A:

Sentence B:

Figure 7: Prompts for MuRef.