
MQuAKE-Remastered: Multi-Hop Knowledge Editing Can Only Be Advanced With Reliable Evaluations

Shaochen (Henry) Zhong^{*♣}, Yifan Lu^{*♣}, Lize Shao[♣], Bhargav Bhushanam[∞], Xiaocong Du[∞],
Louis Feng[∞], Yixin Wan[†], Yiwei Wang[†], Daochen Zha[♣], Yucheng Shi[◇], Ninghao Liu[◇],
Kaixiong Zhou[♡], Shuai Xu[♠], Vipin Chaudhary[♠], and Xia Hu[♣]

[♣] Department of Computer Science, Rice University

[◇] School of Computing, University of Georgia

[♡] Department of Electrical and Computer Engineering, North Carolina State University

[♠] Department of Computer and Data Sciences, Case Western Reserve University

[†] Department of Computer Science, University of California, Los Angeles

[∞] Meta Platforms, Inc.

Abstract

Large language models (LLMs) can give out erroneous answers to factually rooted questions either as a result of undesired training outcomes or simply because the world has moved on after a certain knowledge cutoff date. Under such scenarios, *knowledge editing* often comes to the rescue by delivering efficient patches for such erroneous answers without significantly altering the rests, where many editing methods have seen reasonable success when the editing targets are simple and direct (e.g., “*what club does Lionel Messi currently play for?*”). However, knowledge fragments like this are often deeply intertwined in the real world, making effectively propagating the editing effect to non-directly related questions a practical challenge (to entertain an extreme example: “*What car did the wife of the owner of the club that Messi currently plays for used to get to school in the 80s?*”). Prior arts have coined this task as *multi-hop knowledge editing* with the most popular dataset being MQuAKE, serving as the sole evaluation benchmark for many later proposed editing methods due to the expensive nature of making knowledge editing datasets at scale. In this work, we reveal that **up to 33% or 76% of MQuAKE’s questions and ground truth labels are, in fact, corrupted in various fashions due to some unintentional clerical or procedural oversights**. Our work provides a detailed audit of MQuAKE’s error pattern and a comprehensive fix without sacrificing its dataset capacity. Additionally, we benchmarked almost all proposed MQuAKE-evaluated editing methods on our post-fix dataset, MQuAKE-REMASTERED. It is our observation that many methods try to overfit the original MQuAKE by exploiting some data-specific properties of MQuAKE. We provide a guideline on how to faithfully approach such datasets and show that a simple, minimally invasive approach can bring excellent editing performance without such exploitation. Please refer to <https://github.com/henryzhongsc/MQuAKE-Remastered> and supplemental material for assets.

* Equal contribution. Work corresponds to Shaochen (Henry) Zhong <shaochen.zhong@rice.edu>.

1 Introduction

Given the widespread public-facing popularity of various Large Language Model-powered (LLM) products [Zhao et al., 2023, Yang et al., 2024], even an occasional user has likely experienced LLMs giving out erroneous answers to factually rooted, knowledge-intensive questions. While the reasons why LLMs would hallucinate such kind of misinformation is complex and still an open problem — noisy training data, model bias, out-of-distribution questions, or even simply because the world has moved on after a certain knowledge cutoff date, all likely contributed their fair share to this rather undesired character of LLMs [Huang et al., 2023, Zhang et al., 2023]— **under a practical context, knowledge editing is often considered the go-to remedy by delivering efficient patches for such erroneous answers** without significantly altering the LLM’s output on unrelated queries [Sinitsin et al., 2020, Mitchell et al., 2022].

With the growing need to have more credible and trustworthy LLMs, a vast amount of LLM-specific knowledge editing methods have been proposed, and many of them have seen reasonable success in addressing editing targets that are simple and direct. For example, most modern knowledge editing methods can reliably edit the answer of “*What club does Lionel Messi currently play for?*” from “*Paris Saint-Germain*” to “*Inter Miami CF*” and therefore correctly reflecting the occupation status of Messi [Zhong et al., 2023].

1.1 Multi-hop knowledge editing poses practical significance and non-trivial challenges.

However, due to the intertwined nature of different knowledge fragments, a small change in one knowledge fragment can produce ripple-like effects on a vast amount of related questions [Zhong et al., 2023, Cohen et al., 2023]. It is often a non-trivial challenge to efficiently propagate the editing effect to non-directly related questions with proper precision and locality. E.g., for a — in this case intensionally extreme — question like “*What car did the wife of the owner of the club that Messi currently plays for used to get to school in the 80s?*” Many knowledge-edited LLMs can still struggle while being fully aware of Messi’s abovementioned club transfer [Zhong et al., 2023].

Prior arts have realized the practical significance of being able to edit such complex/non-direct questions upon a certain knowledge update, as different knowledge fragments are almost always deeply entangled with each other in the real world [Zhong et al., 2023, Cohen et al., 2023, Wei et al., 2024]. Meanwhile, exhausting all potential combinations of questions related to one or a few updated knowledge fragments is impractical, if not totally impossible: imagining editing an LLM for every possible question influenced by the abovementioned club transfer of Messi. Even if it is feasible, this poses high operational costs and comes with the intrinsic risks of editing a mass amount of targets; not to mention a repeated effort would be required should Messi ever opt to transfer again.

It is intuitive that a practical knowledge editing method should be able to produce correct answers to relevant factual questions with only a few updated knowledge fragments available. This task has been coined as *multi-hop knowledge editing* **with the founding, largest, as well as the most popular dataset to date being MQUAKE by Zhong et al. [2023]; serving as the sole evaluation backbone for many proposed modern editing methods** due to the expensive nature of making counterfactual and temporal datasets at such a scale ($> 10,000$ cases provided, more about the dataset statistics in Table 6). Note that such expansiveness is further multiplied given the abovementioned ripple effect of multi-hop question answering, as one knowledge update of a subquestion can potentially lead to multiple updated answers across a large number of cases.

1.2 Unfortunately, MQUAKE is flawed due to unintentional clerical and procedural errors — we fixed/remade it and re-benchmarked almost all proposed multi-hop knowledge editing methods.

While MQUAKE is the founding dataset of multi-hop knowledge editing tasks and very much brings life to this vital subject, through a comprehensive audit, we reveal that **up to 33% or 76% of MQUAKE questions and ground truth labels are, in fact, corrupted in various fashions due to some unintentional clerical or procedural errors**; which inevitably cast doubts on the effectiveness of developed methods (especially the ones that solely) evaluated on MQUAKE, and **present as a**

hidden peril to the field’s progress as such flaws are largely unknown to the knowledge editing community before our work. We highlight that the flaws of MQUAKE is an already massive yet constantly growing issue, as MQUAKE is one of the fastest-growing datasets in terms of adaptation in the editing community, yet, the task it is trying to tackle — building more reliable LLM — is without a doubt crucial aspect of NLP development. To pave the way for future advancement of multi-hop knowledge editing, we present our work with the following contributions:

- **A comprehensive audit of MQUAKE:** We are the first to present a comprehensive audit of the existing errors within MQUAKE [Zhong et al., 2023], bringing awareness to the knowledge editing community regarding this popular dataset with significant task importance attached.
- **Fix/remake MQUAKE to MQUAKE-Remastered:** We present the only available fix/remake that not only patches all discovered errors, and done so without sacrificing the intended intensity and capacity of the original MQUAKE whenever possible.
- **Extensively re-benchmark of almost all existing multi-hop knowledge editing methods:** Given the currently existing reports based upon the original MQUAKE are flawed reflections of such proposed methods’ capability, we additionally re-benchmark almost all existing multi-hop knowledge editing methods that are available against our MQUAKE-REMASTERED datasets.
- **Guidance for future multi-hop knowledge editing development.** Upon our extensive re-benchmark results, we observe that many proposed multi-hop knowledge editing methods intentionally or unintentionally overfit the original MQUAKE dataset by applying data-specific operations that are largely unique to the MQUAKE dataset family. We provide guidance on how to faithfully approach these datasets and additionally show that a simple, minimally invasive approach with no such operations can also achieve excellent editing performance.

2 Preliminary

2.1 Background of MQUAKE

MQUAKE (Multi-hop Question Answering for Knowledge Eediting) is a knowledge editing dataset focusing on the abovementioned multi-hop question answering tasks proposed in Zhong et al. [2023], where every case of MQUAKE is a multi-hop question made by a chain of single-hop subquestions. Specifically, MQUAKE is constructed based on the Wikidata:RDF dataset [Vrandečić and Krötzsch, 2014], which, in its rawest format, is a knowledge graph consisting 15+ trillion of Resource Description Framework (RDF) triples¹. MQUAKE essentially builds a much more concise subgraph with only 37 manually elected common relations and top 20% of the most common entities, where a walk of $\{2, 3, 4\}$ -hop on this subgraph can form a case (which is a chain of $\{2, 3, 4\}$ single-hop subquestions connected together) in the MQUAKE dataset.

MQUAKE is presented as two (but in practice, it is essentially three) sub-datasets: MQUAKE-CF and MQUAKE-T. The former focuses on counterfactual tasks, while the latter on temporal changes. We highlight that there is also a MQUAKE-CF-3K dataset, which is a subset of MQUAKE-CF that only contains 3,000 cases in total (with 1,000 cases for $\{2, 3, 4\}$ -hop questions respectively). Authors of MQUAKE evaluate their proposed method, MeLLO [Zhong et al., 2023], upon this MQUAKE-CF-3K dataset, citing limited compute resources; which then become an unspoken standard practice for the majority of the later proposed multi-hop knowledge editing methods [Gu et al., 2024, Shi et al., 2024, Wang et al., 2024, Anonymous, 2024, Cheng et al., 2024]. Due to the very popularity of this sub-sampled dataset, we provide our error analysis mostly based on MQUAKE-CF-3K and MQUAKE-T in the following §3. For interested readers, we additionally provide the same error analysis upon the full MQUAKE-CF in the Appendix B.2, which is only more drastic than MQUAKE-CF-3K due to MQUAKE-CF being a much larger superset of the already compromised MQUAKE-CF-3K. We also collect the dataset statistics in Table 6 to provide a numerical overview of the composition of all three MQUAKE datasets.

¹<https://www.wikidata.org/wiki/Property:P10209>

2.2 Evaluating using MQUAKE

Datasets like MQUAKE-CF or MQUAKE-CF-3K are often evaluated against different “editing intensity,” which is controlled by how many cases among all tested cases are considered “edited,” mimicking different levels of deviation between the learned knowledge stored in the LLM and the desired edited knowledge. This is a sound practice because proper knowledge editing methods should perform well when different numbers of knowledge fragments are edited, as it is equally important to navigate when a significant amount of knowledge is updated, as well as to recognize the few edited knowledge and limit their influence from unrelated unedited knowledge with proper editing locality.

In its original paper, MQUAKE-CF-3K is evaluated when $\{1, 100, 1000, 3000\}$ of its 3,000 cases are edited, similarly, MQUAKE-T is evaluated when $\{1, 100, 500, 1868\}$ of its 1,868 cases being edited, forming an experiment report like Table 5. This kind of report granularity (a gradual coverage from a few edits to all cases being edited) is also adopted by the majority of later proposed multi-hop knowledge editing methods, either in full [Anonymous, 2024] or in spirit with different subsample settings [Gu et al., 2024, Wang et al., 2024, Shi et al., 2024, Cheng et al., 2024, Mengqi et al., 2024]. In this work, we report at an even finer level of granularity for maximum cross-reference potentials.

3 Auditing MQUAKE

In this section, we present a comprehensive audit of the error pattern that existed in MQUAKE-CF-3K and MQUAKE-T [Zhong et al., 2023]. We specifically note that our audit is there to provide a better understanding to the knowledge editing community, especially when digesting methods evaluated on these datasets. **Our audit is not to discredit the contribution of MQUAKE, or any of the proposed methods evaluated on MQUAKE.** We recognize the fact that no dataset can be perfect, especially when it is intrinsically hard to collect large-scale counterfactual and temporal datasets.

3.1 Intra Contamination between Edited Cases and Unedited Cases

As discussed in §2.2, having a gradual evaluation coverage from a few to all cases being edited like Table 5 makes sense for as an evaluation granularity. However, one critical issue is that $k \in \{1, 100, 1000, 3000\}$ -edited cases (supposed MQUAKE-CF-3K) are randomly sub-sampled from the 3,000 total cases. Thus, **there is no guarantee that the k -edited cases and $(3000 - k)$ unedited cases would require two disjoint sets of knowledge and, therefore, risk contamination.**

For a concrete example, consider the following two multi-hop questions from MQUAKE-CF-3K (we also additionally provide the subquestion breakdown and intermediate answers of the two questions for better presentation, we note that such auxiliary information is not part of the instruction visible to the question-answering LLM):

- `case_id:245` (unedited): *What is the official language of the country where Karl Alvarez holds citizenship?*
 - ◊ What is the country of citizenship of Karl Alvarez? USA.
 - ◊ What is the official language of United States of America? American English.
- `case_id:323` (unedited): *What language is the official language of the country where Wendell Pierce holds citizenship?*
 - ◊ What is the country of citizenship of Wendell Pierce? USA.
 - ◊ What is the official language of United States of America? American English.

For both questions, the correct pre-edited answer should be “*American English.*” As both Karl Alvarez and Wendell Pierce are US citizens, and the official language of the US is American English. However, suppose `case_id:323` is sampled as an edited case while `case_id:245` remains unedited, we will be provided with the additional triple containing the knowledge of “*The official language of United States of America is Arabic.*”

Since the unedited `case_id:245` and the edited `case_id:323` share the same subquestion of “*What is the official language of United States of America?*” The answer of `case_id:323` will be rightfully updated to “*Arabic*” per the new knowledge. However, the unedited `case_id:245` still considers the

original answer “*American English*” to be correct, and is therefore contaminated by the edited case `case_id:323` in an unintended fashion. This is problematic because a successful knowledge editing method should be able to retrieve the edited knowledge — “*The official language of United States of America is Arabic*” — upon the relevant questions (in this case the shared one), and thus answering “*Arabic*” to `case_id:245`. This is technically correct, but in conflict with MQUAKE-CF-3K’s label, causing inaccurate experiment readings.

We further note the above-illustrated contamination is not a cherry-picked fluke, but rather a wild-spread error. Here, we sample $\{1, 100, 1000, 2000, 3000\}$ -editing targets from MQUAKE-CF-3K using random seed 100, and find the following error statistics in Table 1.

Table 1: Error statistics of MQUAKE-CF-3K and MQUAKE-T [Zhong et al., 2023] in terms edited cases contaminating unedited cases. k -edited means k cases out of the total dataset are edited.

# of Contaminated	MQUAKE-CF-3K					MQUAKE-T			
	1-edit	100-edit	1000-edit	2000-edit	3000-edit	1-edit	100-edit	500-edit	1868-edit
Cases	0	2,013	1,772	910	0	29	1421	1327	0
Subquestions	0	2,706	3,075	1,664	0	29	1421	1327	0

It is observable from Table 1 that **even a small number of edited cases will cause a concerningly large contamination to unedited cases and subquestions, where 67% and 76% of all cases from MQUAKE-CF-3K and MQUAKE-T are contaminated with just 100 cases being edited**, introducing a significant distortion to the reported experiment results.²

We additionally note while this edited-to-unedited intra-contamination is reducing with k -edit growing, this does not imply a diminishing of issue, but rather a simple by-product of a larger k implies a lesser ($3000 - k$), leaving fewer unedited cases as potential contamination victims. In the extreme case of 3000-edit, there is 0 edited-to-unedited contamination because there is no unedited case left in MQUAKE-CF-3K to be the victim. But 3000-edit has the most edited-to-edited inner contamination, more on this in the following §3.2.

3.2 Inner Contamination between Different Edited Cases

Other than edited cases contaminating unedited cases (§3.1), contamination might also happen among multiple edited cases because a certain subquestion presented in different edited cases can be edited in some but unedited in others³. For brevity, we leave the example walkthrough in Appendix B.1.

Table 2: Error statistics of MQUAKE-CF-3K [Zhong et al., 2023] in terms edited cases contaminating each others. k -edited means k cases out of the total 3,000 cases are edited.

# of Contaminated	1-edit	100-edit	1000-edit	2000-edit	3000-edit
Cases	0	14	265	619	998
Subquestions	0	14	337	854	1,399

This type of contamination is, once again, universally visible in MQUAKE, as shown in Table 2; which is very much a flipped version of Table 1. With k -edit growing, there are more edited cases, thus more edited-to-edited contamination, as there are more potential victims. Notably, **under the 3000-edit tasks, almost one-third (998/3000, $\approx 33\%$) of the evaluated cases are contaminated**, which again introduces distortion to the reported experiment results. We omit the report on MQUAKE-T here because there is only one edit-to-edit contamination when all 1,868 cases from MQUAKE-T are edited (`case_id:424`).

²We note that in Zhong et al. [2023], “ k -edit” means only k of edited cases are evaluated, without any unedited cases. We evaluated both to better reflect the locality of different knowledge editing methods.

³Note, an edited case does not require all of its subquestions being edited, but merely one or more of it (Table 6)

3.3 Conflicting Edits

The two types of contamination introduced in §3.1 and §3.2 are indeed subtle and hard to detect, as they hide between the retrieval scope of different edited cases, which is further complicated when only a subset of cases are edited. However, MQUAKE-CF-3K also includes some straightforward conflicts, such as for the subquestion “Which company is Ford Mustang produced by?” we have the following edits:

- ◇ case_id:2566 (edited): ~~Ford Motor Company~~ Nintendo.
- ◇ case_id:231/2707 (edited): ~~Ford Motor Company~~ Fiat S.p.A.

This is going to cause a direct conflict when case_id:2566 and any of the case_id:231/2707 are both selected as edited cases, as they shall confuse any knowledge edited LLM for having two answers to the same questions. Fortunately, such types of errors are rather minuscule in MQUAKE-CF-3K, with the abovementioned Ford Mustang question and three cases being the only affected data samples.

3.4 Missing Information in Multi-hop Question Instructions

As mentioned in §2, the MQUAKE dataset is built upon a severely filtered Wikidata:RDF knowledge graph [Vrandečić and Krötzsch, 2014]. Specifically, the triples of a certain {2, 3, 4}-hop walk on this subgraph are then fed into a gpt-3.5-turbo model to generate the multi-hop question instruction in a natural language format; such generation are repeated for three different times in case any of the generated question instructions becomes incomprehensible. For every case evaluation, an LLM is considered right should it correctly answer against any three of the multi-hop question instructions [Zhong et al., 2023].

However, while repeating generation three times definitely reduces the chances of having incomprehensible question instructions, we noticed some of such instructions in MQUAKE are still incomplete. We take the following triple set and its generated 3-questions as an example:

- case_id:546 (unedited): We have a 2-hop triple chain of (Albert Mohler, employer, Southern Baptist Theological Seminary) and (Southern Baptist Theological Seminary, religion or worldview, Southern Baptist Convention). MQUAKE-CF-3K provides the following generated multi-hop questions:
 - ◇ Generation #1: *What religion is Albert Mohler associated with?*
 - ◇ Generation #2: *Which religion does Albert Mohler follow?*
 - ◇ Generation #3: *With which religious faith does Albert Mohler identify?*

It is clear that all three generated questions omit the part mentioning which company/institution Albert Mohler is employed by and essentially reduce themselves to single-hop questions, where a correct generation should read like “What religion is Albert Mohler’s employer associated with?” Without the complete question, suppose there is an edit on Albert Mohler’s employer (which there indeed is one), the final answer would likely change. However, with question instruction omitting such information, even the best knowledge-edited LLM cannot answer the question correctly with a faithful approach.

As a general analysis, we find **the natural language question instructions of 672 cases in MQUAKE-CF-3K are missing information in comparison to their raw triplet chain**. This number is counted in the sense that one or more pieces of information present in the triple chain are missing from all three variants of the generated natural language instruction. Similarly, there are 2,830 and 233 cases of erroneous instructions in MQUAKE-CF and MQUAKE-T, respectively.

3.5 Duplicated Cases

The last kind of error we discovered in MQUAKE is simply unintended duplication — i.e., two or more cases sharing the same start subjects, edited facts, chain of triples, and final answer. We discovered 47, 4, and 4 cases of duplication, respectively, in MQUAKE-CF, MQUAKE-CF-3K, and MQUAKE-T.

4 Remastering MQUAKE

In this section, we illustrate how we modified and improved the MQUAKE dataset to MQUAKE-REMASTERED with various fixes on the data samples themselves, as well as providing utility modules to facilitate how one interacts with such datasets.

4.1 Hard Corrections

Three types of error existing in MQUAKE can be fixed once and for all with some careful hard corrections, they are namely Conflicting Edits (§3.3), Missing Information in Multi-hop Question Instructions (§3.4), and Duplicated Cases (§3.5). For Conflicting Edits and Duplicated Cases, since there are only a few such errors (<50 per type per dataset), we employ some manual corrections to address these errors: in the former case, we flip the minority edits to align with the majority edits (and adjust their answers to their subsequence subquestions, should there be any); in the latter case, we simply remove such duplicated cases (except for MQUAKE-CF-3K, which we manually select 4 more cases from MQUAKE-CF to keep the dataset having 3,000 cases in total and a 1,000 cases for {2, 3, 4}-hops). For the Missing Information in Multi-hop Question Instructions errors, we rewrite such natural language question instructions and then replace the original information-missing instructions.

4.2 Dynamic Masking for Maximum Coverage: MQUAKE-REMASTERED-CF, MQUAKE-REMASTERED-CF-3K, and MQUAKE-REMASTERED-T

Due to the contamination count of Intra Edited-to-Unedited Contamination (§3.1) and Inner Edited-to-Edited Contamination (§3.2) tend to grow in the opposite direction as shown in Table 1 and 2, it is impossible to find a fix within the current MQUAKE that can address both issues without significantly decreasing the dataset size. As an alternative, we develop an API that will take a `case_id` and an `edited_flag` as input, respectfully indicating the evaluating case-in-question and whether this case is considered edited; our API shall then return a set of triples that are contamination free by dynamically masking out the conflicting edits from other cases. After such, the user may build up an editing knowledge bank upon such triplets and conduct evaluations for any memory-based knowledge editing methods without losing any of the 9,218 cases from MQUAKE-CF or 1,868 cases from MQUAKE-T.

Specifically, once `case_id`-of-interest is given, our API would loop through all of its subquestions and identify if any of such subquestions is considered edited under another case. If there is a hit, the triple with respect to such edited subquestions is then removed from the bank of edited triples. This dynamic masking mechanism would ensure all cases within the original MQUAKE be usable against memory-based knowledge editing methods. **However, the drawback of masking is it won’t support parameter-based knowledge editing methods**, where weight update is required. We additionally provide a MQUAKE-REMASTERED-CF-6334 to address the need for such methods (Appendix C.1).

5 Benchmark and Discussion

Given almost all proposed multi-hop knowledge editing methods are evaluated on the original, error-contained, MQUAKE datasets. Here, we provide a re-benchmark of those methods against post-fix MQUAKE-REMASTERED datasets for a more reliable reporting of each method’s performance.

5.1 Experiment Coverage

Compared Methods In this work, we aim to cover most, if not all, open-sourced knowledge editing methods evaluated on the original MQUAKE. To the best of our knowledge, this screening criteria include MeLLO [Zhong et al., 2023] and PokeMQA [Gu et al., 2024] as methods specifically proposed to target this multi-hop knowledge editing problem and evaluated on MQUAKE. We additionally include ICE [Cohen et al., 2023] and IKE [Zheng et al., 2023a] as these are also methods purposed for the (single-edit) multi-hop knowledge editing task, though not specifically evaluated

on MQUAKE in their original publications. We note that we are aware methods like GMeLLO [Anonymous, 2024], GLAME [Mengqi et al., 2024], RAE [Shi et al., 2024], StableKE [Wei et al., 2024], and Temple-MQA [Cheng et al., 2024] are also evaluated on MQUAKE, but they are purposely omitted from our re-benchmark coverage due to lack of open-sourced implementation, likely because most of these works are still in submission. Last, we note DeepEdit [Wang et al., 2024] is also an open-sourced MQUAKE-evaluated method, but we excluded it due to its lack of inference optimization (>200 A100 GPU hours needed for 1-edit on MQUAKE-REMASTERED-CF-3K).

Covered Models We opt to use lmsys/vicuna-7b-v1.5 [Zheng et al., 2023b], mistralai/Mistral-7B-Instruct-v0.2 [Jiang et al., 2023], and meta-llama/Meta-Llama-3-8B-Instruct [AI@Meta, 2024] as the choice of question-answering models, both for alignment with existing works [Zhong et al., 2023, Shi et al., 2024, Gu et al., 2024] as well as providing coverage the most recent language models. For methods that require a text-embedding model as a retriever, we use facebook/contriever-msmarco [Izacard et al., 2022] for alignment with MeLLO [Zhong et al., 2023].

Covered Datasets We will provide coverage on our post-fix dataset, namely MQUAKE-REMASTERED-CF, MQUAKE-REMASTERED-CF-3K, and MQUAKE-REMASTERED-T in the masking fashion illustrated in §4.2; as well as MQUAKE-REMASTERED-CF-6334 in its vanilla form. These datasets are respectively corresponding to the original MQUAKE-CF, MQUAKE-CF-3K, and MQUAKE-T from Zhong et al. [2023] (with 6334 as an extra for parameter-based methods), but with the types of error mentioned in §3 fixed in the via means illustrated in §4. We emphasize that such modification is legitimate, and our MQUAKE-REMASTERED is free for the scholarly community to adopt, as the original MQUAKE dataset was published under the MIT license. Where MQUAKE-REMASTERED will be released under CC BY 4.0. All experiments are conducted with an 80G NVIDIA A100 from a DGX A100 server.

5.2 Results and Discussion

Table 3: Performance Comparison of Original MQUAKE and our MQUAKE-REMASTERED datasets

Method	MQuAKE-CF-3k		MQuAKE-T	
	Original	Remastered	Original	Remastered
MeLLO [Zhong et al., 2023]	6.7	6.77	30.84	44.37
GWalk	36.23	66.33	46.41	54.88

Table 4: Experiments on MQUAKE-REMASTERED-CF with numbers of edited cases and methods. Results inside () are edited cases accuracy and unedited cases accuracy, respectively.

Method	MQUAKE-REMASTERED-CF				
	1-edit	1000-edit	3000-edit	6000-edit	9171-edit
vicuna-7b-v1.5 [Zheng et al., 2023b]					
MeLLO [Zhong et al., 2023]	22.55 (100, 22.54)	21.54 (8, 23.2)	17.79 (7.43, 22.83)	12.62 (7.28, 22.58)	6.95 (6.95, N/A)
ICE [Cohen et al., 2023]	<1	OOM	OOM	OOM	OOM
IKE [Zheng et al., 2023a]	<1	OOM	OOM	OOM	OOM
GWalk (Ours)	61.89 (100, 61.89)	56.98 (56.2, 57.07)	56.37 (53.97, 57.54)	54.93 (53.27, 58.06)	54.15 (54.15, N/A)
Mistral-7B-Instruct-v0.2 [Jiang et al., 2023]					
MeLLO [Zhong et al., 2023]	19.83 (<1, 19.84)	19.08 (20.6, 18.9)	18.9 (19.47, 18.62)	18.27 (19.02, 16.87)	18.09 (18.09, N/A)
ICE [Cohen et al., 2023]	<1	OOM	OOM	OOM	OOM
IKE [Zheng et al., 2023a]	<1	OOM	OOM	OOM	OOM
GWalk (Ours)	61.42 (100, 61.42)	57.79 (51.8, 58.52)	56.35 (52.3, 58.32)	53.73 (50.93, 59.04)	51.53 (51.53, N/A)
Meta-Llama-3-8B-Instruct [AI@Meta, 2024]					
MeLLO [Zhong et al., 2023]	<1	<1	<1	<1	<1
ICE [Cohen et al., 2023]	<1	OOM	OOM	OOM	OOM
IKE [Zheng et al., 2023a]	<1	OOM	OOM	OOM	OOM
GWalk (Ours)	74.09 (100, 74.09)	73.67 (71.1, 73.98)	72.4 (70.9, 73.13)	71.62 (70.33, 74.05)	70.08 (70.08, N/A)

Given our MQUAKE-REMASTERED are mostly provided as a fix to MQUAKE, we would like to first highlight the drastic results difference when the same method is evaluated on these two datasets. Table 3 shows our fixing can indeed result in drastically different experiment reports. Where such difference is especially significant for stronger methods, suggesting all previous reporting on MQUAKE has room for reliability improvements, which we filled here with MQUAKE-REMASTERED.

Due to page limitation, we only present the benchmark results on MQUAKE-REMASTERED-CF in the main text and refer our readers to Appendix D.2 for benchmarks of MQUAKE-REMASTERED-CF-3K, MQUAKE-REMASTERED-T, and MQUAKE-REMASTERED-CF-6334. Given the dominance of GWalk — a demo method we proposed as guidance to future scholars of this MHKE task — we leave more discussion on this method below.

5.3 Making Safe and Faithful Approach to MQUAKE and MQUAKE-REMASTERED

Additionally, it is our observation that many multi-hop knowledge editing methods with decent accuracy reports on MQUAKE or MQUAKE-REMASTERED are utilizing designs that leverage specific data properties unique to MQUAKE. For example, methods like GLAME [Mengqi et al., 2024] utilize Wikidata [Vrandečić and Krötzsch, 2014] as the external knowledge graph to better detect the edit-induced conflicts, which happen to be the source of MQUAKE as discussed in §2.1. While these methods might have decent performance on MQUAKE, the cost of maintaining a positive knowledge graph on the correct — but not just edited — knowledge facts is undoubtedly a non-trivial operation cost. Yet, whether sourcing the same Wikidata knowledge graph as MQUAKE might bring them data-specific advantages remains unanswered. Similarly, PokeMQA [Gu et al., 2024] utilizes the 6,218 cases included in MQUAKE-CF but not in MQUAKE-CF-3K as the train set to train its auxiliary components. Given MQUAKE is a dataset with relatively low diversity (e.g., it only includes 37 types of relations), whether having a heavily overlapped train and test set will result in data-specific advantages unique to MQUAKE and its variants, again remains unanswered.

A Minimally Invasive but Performant Approach: GWalk Here, we provide a brief walkthrough of a simple method we designed, namely GraphWalk. It does not leverage any data-specific property unique to MQUAKE or MQUAKE-REMASTERED, yet still presents SOTA performance surpassing many, if not all, established baselines. **We illustrate this method as a simple guidance and potential inspiration to our future multi-hop knowledge editing scholars.** Due to page limitation, we introduce the technical details and design intuition of GWalks in Appendix D.1.

We hope the performant nature of GWalk — in its most vanilla form, without employing any data-specific property unique to MQUAKE or MQUAKE-REMASTERED — can inspire more multi-hop knowledge editing methods that leverage the graph topology of edited facts, without converting such facts to natural language descriptions (at least for retrieval); again we refer readers to Appendix D.1 for details.

6 Related Works

Our work mainly conducts an audit and provides fixes to the MQUAKE dataset. To the best of our knowledge, only two prior arts have touched on the errors existing in MQUAKE: GMeLLO [Anonymous, 2024] (an anonymous submission to ACL ARR 2024 February) and DeepEdit [Wang et al., 2024]. As an overview, GMeLLO briefly discussed the same type of error we discussed in §3.4, but without providing any quantitative error analysis or fix. DeepEdit discovered the same inner contamination error as we discussed in §3.2, but specific to the “3000-edit” setup. DeepEdit’s proposed fix is a simple removal of the 998 inner contaminated cases from the MQUAKE-CF-3K dataset, so this fix is custom 3000-edit setup and done so by sacrificing 33% of the dataset capacity. We leave more details in Appendix E for interested readers.

Additionally, our work provides a re-benchmark of most, if not all, open-sourced knowledge editing methods evaluated on MQUAKE, and sets guidance on how to safely and faithfully approach such datasets. To the best of our knowledge, no other work provides the same benchmark nor touches on the same issue.

7 Conclusion

Our work provides a comprehensive audit and fix of the MQUAKE dataset. We further re-benchmarked all open-sourced knowledge editing methods evaluated on MQUAKE with our MQUAKE-REMASTERED datasets and provided guidance and examples on how to faithfully approach these datasets with our GWalk.

Limitations and Impact Statement

While our work comprehensively addressed many errors in MQUAKE, we caution our reader to perform further analysis and evaluation on our MQUAKE-REMASTERED to ensure our fixes are indeed exhaustive. We also note that multi-hop knowledge editing only represents one aspect of a language model’s ability, so any actual deployment of a language model should undergo more, and if possible, deployment-specific evaluations.

References

- AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- Anonymous. Graph memory-based editing for large language models. *Submission to ACL ARR 2024 February*, 2024.
- K. Cheng, G. Lin, H. Fei, Y. zhai, L. Yu, M. A. Ali, L. Hu, and D. Wang. Multi-hop question answering under temporal knowledge editing. *arXiv*, 2024.
- R. Cohen, E. Biran, O. Yoran, A. Globerson, and M. Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 2023.
- T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. D. III, and K. Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018. URL <http://arxiv.org/abs/1803.09010>.
- H. Gu, K. Zhou, X. Han, N. Liu, R. Wang, and X. Wang. Pokemqa: Programmable knowledge editing for multi-hop question answering. *arXiv*, 2024.
- L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv*, 2023.
- G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022.
- A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b. *arXiv*, 2023.
- Z. Mengqi, Y. Xiaotian, L. Qiang, R. Pengjie, W. Shu, and C. Zhumin. Knowledge graph enhanced large language model editing. *arXiv*, 2024.
- E. Mitchell, C. Lin, A. Bosselut, C. Finn, and C. D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022.
- Y. Shi, Q. Tan, X. Wu, S. Zhong, K. Zhou, and N. Liu. Retrieval-enhanced knowledge editing for multi-hop question answering in language models. *arXiv*, 2024.
- A. Sinitsyn, V. Plokhhotnyuk, D. Pyrkyn, S. Popov, and A. Babenko. Editable neural networks. In *International Conference on Learning Representations*, 2020.
- D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 2014.
- Y. Wang, M. Chen, N. Peng, and K.-W. Chang. Deepedit: Knowledge editing as decoding with constraints. *arXiv*, 2024.

- 414 Z. Wei, L. Pang, H. Ding, J. Deng, H. Shen, and X. Cheng. Stable knowledge editing in large
415 language models. *arXiv*, 2024.
- 416 J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu. Harnessing the
417 power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*,
418 2024.
- 419 Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T.
420 Luu, W. Bi, F. Shi, and S. Shi. Siren’s song in the ai ocean: A survey on hallucination in large
421 language models. *arXiv*, 2023.
- 422 W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du,
423 C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen.
424 A survey of large language models. *arXiv*, 2023.
- 425 C. Zheng, L. Li, Q. Dong, Y. Fan, Z. Wu, J. Xu, and B. Chang. Can we edit factual knowledge by
426 in-context learning? *arXiv*, 2023a.
- 427 L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing,
428 H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena.
429 *arXiv*, 2023b.
- 430 Z. Zhong, Z. Wu, C. D. Manning, C. Potts, and D. Chen. MQuAKE: Assessing knowledge editing
431 in language models via multi-hop questions. In *The 2023 Conference on Empirical Methods in*
432 *Natural Language Processing*, 2023.

433 Checklist

434 The checklist follows the references. Please read the checklist guidelines carefully for information on
435 how to answer these questions. For each question, change the default **[TODO]** to **[Yes]** , **[No]** , or
436 **[N/A]** . You are strongly encouraged to include a **justification to your answer**, either by referencing
437 the appropriate section of your paper or providing a brief inline description. For example:

- 438 • Did you include the license to the code and datasets? **[Yes]** See Section ??.
- 439 • Did you include the license to the code and datasets? **[No]** The code and the data are
440 proprietary.
- 441 • Did you include the license to the code and datasets? **[N/A]**

442 Please do not modify the questions and only use the provided macros for your answers. Note that the
443 Checklist section does not count towards the page limit. In your paper, please delete this instructions
444 block and only keep the Checklist section heading above along with the questions/answers below.

445 1. For all authors...

- 446 (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s
447 contributions and scope? **[Yes]** We provide an audit and remake of a dataset, as well as
448 a benchmark of all available methods.
- 449 (b) Did you describe the limitations of your work? **[Yes]** Before references
- 450 (c) Did you discuss any potential negative societal impacts of your work? **[Yes]** Before
451 references
- 452 (d) Have you read the ethics review guidelines and ensured that your paper conforms to
453 them? **[Yes]** We have read and ensured the paper conforms to the guidelines.

454 2. If you are including theoretical results...

- 455 (a) Did you state the full set of assumptions of all theoretical results? **[N/A]** No theoretical
456 result included in the paper.

- 457 (b) Did you include complete proofs of all theoretical results? [N/A]
- 458 3. If you ran experiments (e.g. for benchmarks)...
- 459 (a) Did you include the code, data, and instructions needed to reproduce the main experi-
 460 mental results (either in the supplemental material or as a URL)? [Yes] In supplemental
 461 material.
- 462 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they
 463 were chosen)? [Yes] In supplemental material.
- 464 (c) Did you report error bars (e.g., with respect to the random seed after running experi-
 465 ments multiple times)? [No] Given the massive amount of experiments, we fix the
 466 seeds and run each experiment entry by once.
- 467 (d) Did you include the total amount of compute and the type of resources used? [Yes] See
 468 §5.1 for resource and supplementary materials for amount of compute.
- 469 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
- 470 (a) If your work uses existing assets, did you cite the creators? [Yes] All works are properly
 471 cited in-text and afterward.
- 472 (b) Did you mention the license of the assets? [Yes] At §5.1
- 473 (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
 474 We include the dataset in supplemental materials
- 475 (d) Did you discuss whether and how consent was obtained from people whose data you're
 476 using/curating? [Yes] Data used are open-sourced in MIT license, as showed in §5.1.
- 477 (e) Did you discuss whether the data you are using/curating contains personally identifiable
 478 information or offensive content? [Yes] In §2.1, we discussed the MQuAKE dataset is
 479 constructed based on the Wikidata: RDF dataset
- 480 5. If you used crowdsourcing or conducted research with human subjects...
- 481 (a) Did you include the full text of instructions given to participants and screenshots, if
 482 applicable? [N/A] No applicable
- 483 (b) Did you describe any potential participant risks, with links to Institutional Review
 484 Board (IRB) approvals, if applicable? [N/A] No applicable
- 485 (c) Did you include the estimated hourly wage paid to participants and the total amount
 486 spent on participant compensation? [N/A] No applicable

487 A Extended Preliminary

488 A.1 Demo Report of MQUAKE

Table 5: Standard reporting format of MQUAKE-CF-3K, and MQUAKE-T demoed with MeLLO on Vicuna-7B [Zheng et al., 2023b]; k -edited means k cases out of the total cases are edited. Abbreviated table courtesy of Zhong et al. [2023] (Table 3).

Model	Method	MQUAKE-CF-3K				MQUAKE-T			
		1-edit	100-edit	1000-edit	3000-edit	1-edit	100-edit	500-edit	1868-edit
Vicuna-7B	MeLLO [Zhong et al., 2023]	20.3	11.9	11.0	10.2	84.4	56.3	52.6	51.3

489 A.2 Dataset Statistics

Table 6: Dataset Statistics of MQUAKE. Numbers are in terms of cases (a case in MQUAKE is a chain consisting of multiple subquestions).

Dataset	# of Edits	2-hop	3-hop	4-hop	Total
MQUAKE-CF-3K	1	513	356	224	1,093
	2	487	334	246	1,067
	3	-	310	262	572
	4	-	-	268	268
	All	1,000	1,000	1,000	3,000
MQUAKE-CF	1	2,454	855	446	3,755
	2	2,425	853	467	3,745
	3	-	827	455	1,282
	4	-	-	436	436
	All	4,879	2,535	1,804	9,218
MQUAKE-T	1 (All)	1,421	445	2	1,868

Table 7: Dataset Statistics of MQUAKE-REMASTERED. Numbers are in terms of cases (a case in MQUAKE is a chain consisting of multiple subquestions).

Dataset	# of Edits	2-hop	3-hop	4-hop	Total
MQUAKE-REMASTERED-CF-3K	1	513	356	224	1,093
	2	487	334	246	1,067
	3	-	310	262	572
	4	-	-	268	268
	All	1,000	1,000	1,000	3,000
MQUAKE-REMASTERED-CF	1	2,446	850	441	3,737
	2	2,415	852	463	3,730
	3	-	823	451	1,274
	4	-	-	430	430
	All	4,861	2,525	1,785	9,171
MQUAKE-REMASTERED-T	1 (All)	1,421	441	2	1,868
MQUAKE-REMASTERED-CF-6334	1	1,971	77	0	2,048
	2	2,415	476	14	2,905
	3	-	823	128	951
	4	-	-	430	430
	All	4,386	1,376	572	6,334

B Extended Auditing

B.1 Example of Inner Contamination between Different Edited Cases (§3.2)

Again, we walk through two cases from MQUAKE-CF-3K as a concrete example. First, we show them in their unedited format (again, subquestion breakdowns and intermediate answers are here for demonstration purposes and are not visible to the question-answering LLM during evaluation):

- `case_id:1570` (unedited): *Who was the creator of the official language used in the work location of Matti Vanhanen?*
 - ◊ Which city did Matti Vanhanen work in? Helsinki.
 - ◊ What is the official language of Helsinki? Finnish.
 - ◊ Who was Finnish created by? Mikael Agricola.
- `case_id:1968` (unedited): *Who created the official language of Housemarque’s headquarters location?*
 - ◊ Which city is the headquarter of Housemarque located in? Helsinki.
 - ◊ What is the official language of Helsinki? Finnish.
 - ◊ Who was Finnish created by? Mikael Agricola.

Suppose `case_id:1570` and `case_id:1968` are both selected as editing cases, two triples containing the following knowledge will be available: “*The official language of Helsinki is Black Speech*” (intended for `case_id:1570`), and “*Finnish was created by William Shakespeare*” (intended for `case_id:1968`), leading to the following edited breakdown.

- `case_id:1570` (edited): *Who was the creator of the official language used in the work location of Matti Vanhanen?*
 - ◊ Which city did Matti Vanhanen work in? Helsinki.
 - ◊ What is the official language of Helsinki? ~~Finnish~~ Black Speech.
 - ◊ Who was ~~Finnish~~ Black Speech created by? J. R. R. Tolkien.
- `case_id:1968` (edited): *Who created the official language of Housemarque’s headquarters location?*
 - ◊ Which city is the headquarter of Housemarque located in? Helsinki.
 - ◊ What is the official language of Helsinki? Finnish.
 - ◊ Who was Finnish created by? ~~Mikael Agricola~~ William Shakespeare.

Much like the previous conflict between unedited and edited cases, these two edited cases share a common subquestion: “*What is the official language of Helsinki?*” However, such subquestion is edited in `case_id:1570` while unedited in `case_id:1968`, causing unintended contamination.

522 B.2 Error Analysis of MQUAKE-CF

Table 8: Error statistics of MQUAKE-CF [Zhong et al., 2023] in terms of edited cases contaminating unedited cases §3.1. k -edited means k cases are edited out of the total 9218 cases.

# of Contaminated	MQUAKE-CF-3K						
	1-edit	100-edit	1000-edit	2000-edit	3000-edit	5000-edit	9218-edit
Cases	62	3307	5275	5110	4578	3346	0
Subquestions	62	4525	8751	8989	8326	6364	0

Table 9: Error statistics of MQUAKE-CF [Zhong et al., 2023] in terms edited cases contaminating each others §3.2. k -edited means k cases are edited out of the total 9218 cases.

# of Contaminated	1-edit	100-edit	1000-edit	2000-edit	3000-edit	5000-edit	9218-edit
Cases	0	8	192	441	732	1397	2873
Subquestions	0	12	270	606	1027	1986	4250

523 C Extended Remastering

524 C.1 Contamination Free Subset: MQUAKE-REMASTERED-CF-6334

525 While MQUAKE-REMASTERED-MASKED with masking operation can well support memory-based
 526 knowledge editing methods, it will not be compatible with parameter-based methods. This is because,
 527 for parameter-based methods, the set of edited facts used for training and evaluation needs to be
 528 constant yet consistent with each other at all times; whereas dynamic masking cannot suffice as it is
 529 essentially adjusting the dataset on the fly during inference time.

530 To effectively evaluate parameter-based knowledge editing methods, we present MQUAKE-
 531 REMASTERED-CF-6334. MQUAKE-REMASTERED-CF-6334 is a dataset extracted from
 532 MQUAKE-CF, where all 6,334 cases are edited cases; and they are completely contamination-
 533 free from each other. This dataset is suitable for LLM editing with parameter-based approaches, as
 534 one can make careful splits among the 6,334 cases of MQUAKE-REMASTERED-CF-6334 to serve
 535 as train, validation, and evaluation sets.

D Extended Benchmark and Discussion

D.1 GWalks

The design of GWalk hinges on the fundamental pipeline of memory-based knowledge editing methods: where the pool of source only contains *edited facts*. This school of editing methods has proven to be successful, mainly because it can leverage the power of retrieval-argument generation (RAG) combined with the in-context learning (ICL) capability of LLMs. Further, it is common sense that edited knowledge facts will be much less than unedited knowledge facts, making maintaining a knowledge pool exclusively containing edited facts a viable option — like done so in MeLLO [Zhong et al., 2023].

Different from MeLLO, where all edited facts are converted from triples to natural language (NL) descriptions in its edited bank, GWalk preserves the edited facts in their original triples fashion and leverages the graph topology they come with. This makes maintaining this edited bank much easier — as one can easily adjust the entity or relation on a knowledge graph without rewriting every natural language description of every related edited fact. It also brings more precise retrieval mapping when a question pertaining to a certain edited fact is asked. This is because methods like MeLLO would need to RAG from a pool of edited facts in NL format, and there might always be something — though not actually related to the question asked — having a close enough embedding distance to the query question (i.e., unintended retrieval), and thus result in hallucination. However, if we simply query the entity and relations implied in a question against a knowledge graph, there is less chance of retrieving unintended materials. Specifically, GWalk works like the following Algorithm 1.

Algorithm 1: General Procedure GWalk on a Multi-hop Question

Input:

M , the Question Answering Language Model;
 T , a Text-embedding model;
 Q , a Multi-hop Question;
 E , a bank of edited facts as a knowledge graph.

Output:

o_p , the answer to Q .

Initialize:

$i = 1$, the subquestion counter;
 $o_p = \text{None}$, the answer from the previous subquestion.

```

1  $s \leftarrow$  Extracted subject from  $Q$ ;
2  $rels \leftarrow$  Prompt  $M$  to breakdown  $Q$  into a sequence of relations.
  /* If  $Q$  is 'What is the official language of the country where Karl
  Alvarez holds citizenship?', then  $s$  would be 'Karl Alvarez' and a
  possible  $rels$  is ['citizenship', 'official language'] */
3 for  $r \in rels$  do
4   Query  $\langle s, r, ? \rangle$  against  $E$  using  $T$ , namely we do  $T(s)$  first to determine if there is a
   retrievable  $s \in E$ , then inspect if the  $s \in E$  has an relation edge retrievable by  $T(r)$ .
   /* We set a threshold on embedding similarity for  $T$  to determine
   whether an item is retrievable or not. */
5   Prompt  $M$  to generate subquestion  $q_i$  with  $s$  and  $r$ .
6    $o_p \leftarrow$  the  $M$ -generated answer to  $q_i$ .
7   if  $T(s, r)$  has a valid retrieval  $\langle s, r, o^* \rangle$  then
8      $o_p \leftarrow o^*$ ;
     /* The answer to this subquestion will be the start subject of the
     next subquestion. */
9    $s \leftarrow o_p$ ;
10   $i \leftarrow i + 1$ ;
11 Return  $o_p$ ;

```

Table 10: MQUAKE-REMASTERED-CF-3K

Method	MQUAKE-REMASTERED-CF-3K			
	1-edit	100-edit	1000-edit	3000-edit
vicuna-7b-v1.5 [Zheng et al., 2023b]				
MeLLO [Zhong et al., 2023]	16.54 (100, 16.51)	18 (9.0, 18.31)	14.63 (8.0, 17.95)	6.77 (6.77, N/A)
ICE [Cohen et al., 2023] OOM	<1	<1	OOM	OOM
IKE [Zheng et al., 2023a] OOM	<1	OOM	OOM	OOM
GWalk (Ours)	54.89 (100, 54.87)	60.9 (54, 61.14)	57.37 (54.4, 58.85)	66.33 (66.33, N/A)
Mistral-7B-Instruct-v0.2 [Jiang et al., 2023]				
MeLLO [Zhong et al., 2023]	19.73 (100, 19.71)	18.6 (21, 18.52)	16.33 (17.8, 15.6)	15.93 (15.93, N/A)
ICE [Cohen et al., 2023] OOM	<1	<1	OOM	OOM
IKE [Zheng et al., 2023a] OOM	<1	4.43 (4, 4.49)	OOM	OOM
GWalk (Ours)	56.57 (100, 56.55)	61.93 (47, 62.45)	57.17 (51.5, 60.0)	51.0 (51.0, N/A)
Meta-Llama-3-8B-Instruct [AI@Meta, 2024]				
MeLLO [Zhong et al., 2023]	<1	<1 (2.0, <1)	1.03 (3.0, <1)	2.3 (2.3, N/A)
ICE [Cohen et al., 2023] OOM	<1	<1	OOM	OOM
IKE [Zheng et al., 2023a] OOM	<1	<1	OOM	OOM
GWalk(Ours)	69.0 (100, 68.99)	76.73 (67, 77.07)	75.47 (74.2, 76.1)	70.6 (70.6, N/A)

*Results inside the parenthesis are edited cases accuracy and unedited cases accuracy, respectively.

Table 11: MQUAKE-REMASTERED-T

Method	MQUAKE-REMASTERED-T			
	1-edit	100-edit	500-edit	1864-edit
vicuna-7b-v1.5 [Zheng et al., 2023b]				
MeLLO [Zhong et al., 2023]	19.31 (100, 19.27)	18.88 (45.0, 17.4)	22.16 (40.4, 15.47)	44.37 (44.37, N/A)
ICE [Cohen et al., 2023]	<1	<1	<1	OOM
IKE [Zheng et al., 2023a]	<1	<1	<1	OOM
GWalk (Ours)	35.52 (100, 35.48)	46.51 (49.0, 46.37)	48.93 (56.0, 46.33)	54.88 (54.88, N/A)
Mistral-7B-Instruct-v0.2 [Jiang et al., 2023]				
MeLLO [Zhong et al., 2023]	10.3 (0, 10.31)	10.25 (59.0, 7.48)	18.78 (48.4, 7.92)	47.75 (47.75, N/A)
ICE [Cohen et al., 2023]	<1	<1	<1	OOM
IKE [Zheng et al., 2023a]	<1	<1	<1	OOM
GWalk (Ours)	34.07 (0, 34.08)	45.76 (47, 45.69)	46.78 (51.2, 45.16)	50.7 (50.7, N/A)
Meta-Llama-3-8B-Instruct [AI@Meta, 2024]				
MeLLO [Zhong et al., 2023]	<1	1.13 (17, 0.23)	4.72 (17.4, <1)	16.58 (16.58, N/A)
ICE [Cohen et al., 2023]	<1	<1	<1	OOM
IKE [Zheng et al., 2023a]	<1	<1	<1	OOM
GWalk (Ours)	70.12 (100, 70.1)	73.28 (84.0, 72.68)	76.61 (87, 72.8)	84.01 (84.01, N/A)

*Results inside the parenthesis are edited cases accuracy and unedited cases accuracy, respectively.

Table 12: MQUAKE-REMASTERED-CF-6334

Method	MQUAKE-REMASTERED-CF-6334			
	100-edit	1000-edit	3000-edit	6344-edit
vicuna-7b-v1.5 [Zheng et al., 2023b]				
MeLlo [Zhong et al., 2023]	19.16 (0, 10.99, 19.37)	19.27 (5.1, 9.58, 24.53)	11.17 (4.31, 8.55, 23.3)	6.83 (4.58, 7.72, 19.05)
ICE [Cohen et al., 2023]	OOM	OOM	OOM	OOM
IKE [Zheng et al., 2023a]	OOM	OOM	OOM	OOM
PokeMQA [Gu et al., 2024]	-	-	-	21.77 (3.25, 30.82, 1.59)
GWalk (Ours) KGWalk	57.55 (22.22, 64.84, 57.48)	61.79 (29.08, 66.17, 63.23)	59.1 (39.3, 63.74, 64.33)	56.62 (44.64, 62.11, 68.25)
Mistral-7B-Instruct-v0.2 [Jiang et al., 2023]				
MeLlo [Zhong et al., 2023]	27.5 (<1, 23.08, 27.65)	27.54 (12.76, 24, 30.4)	24.37 (11.88, 25.51, 32.06)	21.26 (13.29, 24.9, 30.16)
ICE [Cohen et al., 2023]	OOM	OOM	OOM	OOM
IKE [Zheng et al., 2023a]	8.82 (11.11, 6.59, 8.86)	OOM	OOM	OOM
PokeMQA [Gu et al., 2024]	-	-	-	20.38 (3.99, 27.41, 69.84)
GWalk (Ours)	56.25 (33.33, 57.14, 56.28)	58.9 (34.69, 60.57, 60.6)	56.03 (42.69, 59.04, 59.85)	54.43 (47.49, 57.74, 52.38)
Meta-Llama-3-8B-Instruct [AI@Meta, 2024]				
MeLlo [Zhong et al., 2023]	<1	<1	1.12 (1.17, 1.48, 0.22)	1.27 (<1, 1.4, 1.59)
ICE [Cohen et al., 2023]	OOM	OOM	OOM	OOM
IKE [Zheng et al., 2023a]	<1	OOM	OOM	OOM
PokeMQA [Gu et al., 2024]	-	-	-	20.38 (1.08, 28.41, 76.19)
GWalk (Ours)	67.01 (33.33, 74.73, 66.92)	71.89 (47.45, 80.94, 70.65)	73.76 (54.05, 81.6, 71.12)	74.22 (61.02, 80.47, 73.02)

*Results inside the parenthesis are edited cases (unique in the test set) accuracy, edited cases (overlap of the test and train set) accuracy, and unedited cases accuracy, respectively.

E Extended Related Works

Specifically, GMeLLO [Anonymous, 2024] briefly discusses the inconsistency between the triple chain and the generated multi-hop questions in its §4.5.1, which is the same type of error we discussed in §3.4. We note that GMeLLO merely highlights such errors but does not provide a quantified measurement of its scale nor any fix. We did both in §3.4 and §4.1.

DeepEdit [Wang et al., 2024] discovered the same inner contamination error as we discussed in §3.2. DeepEdit does provide a quantified measurement of the scale of such error but only pertains to the MQUAKE-CF-3K dataset, and such quantifiable results are only valid when all 3,000 cases of MQUAKE-CF-3K are considered edited; which, as shown in Table 5, only constitute one column of MQUAKE-CF-3K’s reporting. Further, DeepEdit provides a rather hardcore fix to this problem by removing the 998 inner contaminated cases from the MQUAKE-CF-3K dataset — which is (supposedly) the same 998 cases we detect in Table 2 under the 3000-edit column — with the post-fix dataset denoted as MQUAKE-2002 for having 2,002 out of 3,000 cases left. While this fix is, of course, helpful, we argue our post-fix MQUAKE-REMASTERED-CF-3K, MQUAKE-REMASTERED-CF, and MQUAKE-REMASTERED-T are much more comprehensive and effective since they patched many more errors revealed in §3 (which still exists in MQUAKE-2002), works outside the MQUAKE-CF-3K dataset, do not require the number of edits to be 2,002 cases, and most importantly, done so without scarifying almost 1/3 of the capacity of the original dataset.

F Datasheet

We supply the datasheet of our proposed MQUAKE-REMASTERED datasets, following the format of Gebru et al. [2018] (specifically, <https://github.com/AudreyBeard/Datasheets-for-Datasets-Template/tree/master>) as suggested in the Call for NeurIPS 2024 Datasets and Benchmarks Track.

We note that our proposed MQUAKE-REMASTERED datasets are still under the internal reviewer pipeline of Meta, so we opt not to publicize the GitHub link (<https://github.com/henryzhongsc/MQuAKE-Remastered>) we referred to in our paper for now. However, we supply full access to our datasets, code, license, general metadata, and other auxiliary materials in confidence to our reviewers as supplementary materials. We plan to release all supplied materials here — as well as additional materials, e.g., metadata in the Crossiant format — at the said link before the camera-ready date of NeurIPS 2024 Datasets and Benchmarks Track (Oct 30, 2024) under a CC BY 4.0 license. As part of the required author statement, we hereby state that we the authors bear all responsibility in case of violation of rights, etc.

Note, we supply this datasheet in two formats in our supplementary material: as a standalone PDF for easy identification, as well as an appendix of our full paper — since we routinely refer to our paper materials for information where a holistic file would provide better page-jumping access.

F.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

A: The MQUAKE-REMASTERED datasets are created to address several critical errors that exist in the original MQUAKE datasets [Zhong et al., 2023] due to various clerical or procedural oversights. Such errors, should we leave them unaddressed, may lead to up to 33% or 76% of the total entries being corrupted.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? A: The MQUAKE-REMASTERED datasets are mainly created by the joint effort of Rice University and Meta Platforms, facilitated by scholars from North Carolina State University, Case Western Reserve University, and the University of California, Los Angeles (UCLA), which are members who contributed to the multi-hop knowledge editing tasks — the field of study the original MQuAKE datasets focused on.

What support was needed to make this dataset? (e.g., who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

A: This research is supported, in part, by Meta Platform Inc., and NSF Awards IIS-2310260 (“Towards Effective Detection and Mitigation for Shortcut Learning: A Data Modeling Framework”) and IIS-2224843 (“Human-Centric Big Network Embedding”) awarded to Prof. Xia Hu at Rice University.

Any other comments? A: N/A.

F.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

A: The instances covered in the MQUAKE-REMASTERED are all from the original MQUAKE dataset. Which are entity-relationship triples extracted from the Wikidata dataset [Vrandečić and Krötzsch, 2014]. Given the vast coverage of Wikidata, we cannot give a specific scope of what

620 instances are covered in this dataset, but they are all subjects that exist on Wikipedia, converted into
621 (textual) triples and then natural language descriptions. We refer interested readers to Section 2.1 for
622 details.

623 **How many instances are there in total (of each type, if appropriate)?** A: Without duplication,
624 our dataset has 9,171 instances/cases from the MQUAKE-REMASTERED-CF dataset and 1,864
625 instances/cases from the MQUAKE-REMASTERED-T dataset, making a total of 11,035 instances
626 cases. Note that one case corresponds to a chain of single-hop questions; we provide a detailed
627 breakdown with regard to types (in this case, the number of hops) in Table 7.

628 **Does the dataset contain all possible instances or is it a sample (not necessarily random)**
629 **of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the
630 sample representative of the larger set (e.g., geographic coverage)? If so, please describe how
631 this representativeness was validated/verified. If it is not representative of the larger set, please
632 describe why not (e.g., to cover a more diverse range of instances, because instances were withheld
633 or unavailable).

634 A: Our MQUAKE-REMASTERED datasets are based on the MQUAKE datasets [Zhong et al., 2023],
635 which is extracted from the Wikidata dataset [Vrandečić and Krötzsch, 2014]. Such extraction is
636 performed due to various reasons, the two main ones are 1) the covered knowledge must be enough
637 mainstream where LLMs will likely have (pre-edited) knowledge upon it, and it can’t be too large
638 (e.g., as large as all possible triples in Wikidata) as that will induce huge inference cost for evaluation.
639 We refer readers to Section 2.1 for details.

640 **What data does each instance consist of?** “Raw” data (e.g., unprocessed text or images) or
641 features? In either case, please provide a description.

642 A: A case/instance in MQUAKE-REMASTERED contains a chain of triplets extracted from Wikidata
643 [Vrandečić and Krötzsch, 2014] (see Section 2.1 for details). They are considered processed text
644 data.

645 **Is there a label or target associated with each instance?** If so, please provide a description.

646 A: There is a label that represents the ground truth of each case/instance in the MQUAKE-
647 REMASTERED datasets, which is always a subject in the Wikidata [Vrandečić and Krötzsch, 2014].

648 **Is any information missing from individual instances?** If so, please provide a description,
649 explaining why this information is missing (e.g., because it was unavailable). This does not include
650 intentionally removed information, but might include, e.g., redacted text.

651 A: No, unless the entry of information is also missing in the source Wikidata dataset [Vrandečić and
652 Krötzsch, 2014] and subsequently, the original MQUAKE datasets.

653 **Are relationships between individual instances made explicit (e.g., users’ movie ratings, social**
654 **network links)?** If so, please describe how these relationships are made explicit.

655 A: There is no explicit relationship between different instances/cases, but one important challenge of
656 knowledge editing is to make sure the editing effect of one knowledge will not influence an unrelated
657 knowledge. The original MQUAKE dataset struggles in this regard due to two types of contamination
658 illustrated in Section 3.1 and Section 3.2. We fixed such kind of unintended between-instance/cases
659 influences with means introduced in Section 4.2.

660 **Are there recommended data splits (e.g., training, development/validation, testing)?** If so,
661 please provide a description of these splits, explaining the rationale behind them.

662 A: MQUAKE-REMASTERED datasets are four parts: MQUAKE-REMASTERED-CF, MQUAKE-
663 REMASTERED-CF-3K, MQUAKE-REMASTERED-T, MQUAKE-REMASTERED-CF-6334. The

664 former three are evaluation-only datasets; thus, no split is required. The last dataset is made to
665 accommodate the potential need for parameter-based knowledge editing methods where training
666 is required, where both the split and the instance/case selection of this MQUAKE-REMASTERED-
667 CF-6334 dataset must be careful to be considered regarding needs like knowledge consistency and
668 coverage between a different portion of the dataset (see Section 4.2 for details). We provide the split
669 information in our README.md file.

670 **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a
671 description.

672 A: Not to the best of our knowledge in terms of noise or error. For redundancy, each case/instance of
673 the MQUAKE-REMASTERED datasets is a multi-hop question written in three different ways, more
674 on this in Section 3.4.

675 **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g.,
676 websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there
677 guarantees that they will exist, and remain constant, over time; b) are there official archival versions
678 of the complete dataset (i.e., including the external resources as they existed at the time the dataset
679 was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external
680 resources that might apply to a future user? Please provide descriptions of all external resources and
681 any restrictions associated with them, as well as links or other access points, as appropriate.

682 A: It is self-contained.

683 **Does the dataset contain data that might be considered confidential (e.g., data that is pro-
684 tected by legal privilege or by doctor-patient confidentiality, data that includes the content of
685 individuals' non-public communications)?** If so, please provide a description.

686 A: No, our dataset is based on the original MQUAKE dataset, which is published under an MIT
687 license [Zhong et al., 2023].

688 **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening,
689 or might otherwise cause anxiety?** If so, please describe why.

690 A: Not to the best of our knowledge.

691 **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

692 A: Yes, a healthy portion of our MQUAKE-REMASTERED dataset is regarding the occupation or
693 career path of famous individuals with pages on Wikipedia.

694 **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how
695 these subpopulations are identified and provide a description of their respective distributions within
696 the dataset.

697 A: Not to the best of our knowledge.

698 **Is it possible to identify individuals (i.e., one or more natural persons), either directly or
699 indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.

700 A: Yes. Again, because a healthy portion of our MQUAKE-REMASTERED dataset is regarding the
701 occupation or career path of famous individuals with pages on Wikipedia. E.g., some questions are
702 regarding which individual authored which book at what time.

703 **Does the dataset contain data that might be considered sensitive in any way (e.g., data that
704 reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or
705 union memberships, or locations; financial or health data; biometric or genetic data; forms of**

706 **government identification, such as social security numbers; criminal history)?** If so, please
707 provide a description.

708 A: Not to the best of our knowledge given the original source of our dataset — Wikidata [Vrandečić
709 and Krötzsch, 2014] — is already public.

710 **Any other comments?** A: N/A

711 **F.3 Collection**

712 **How was the data associated with each instance acquired?** Was the data directly observable (e.g.,
713 raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived
714 from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was
715 reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If
716 so, please describe how.

717 A: Our datasets are derived from the original MQUAKE datasets [Zhong et al., 2023], which is
718 extracted from the publically observable Wikidata dataset [Vrandečić and Krötzsch, 2014].

719 **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe
720 of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please
721 describe the timeframe in which the data associated with the instances was created. Finally, list when
722 the dataset was first published.

723 A: Since the original MQUAKE datasets, which our MQUAKE-REMASTERED datasets are based on,
724 did not specify the timeframe of what specific Wikidata version is their source, we are unclear about
725 the specific timeframe of its instance. Judging from the appearance of data — including knowledge
726 regarding Rishi Sunak as UK Prime Minister — we infer the dataset covers the timeframe till the
727 year 2023. The original MQUAKE datasets were published on May 24, 2023, [Zhong et al., 2023].
728 Our datasets are currently in the pipeline of Meta’s internal review, and we expect to release them to
729 the public before the camera-ready deadline of NeurIPS Datasets and Benchmarks Track, which is
730 Oct 30, 2024.

731 **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sen-
732 sor, manual human curation, software program, software API)?** How were these mechanisms
733 or procedures validated?

734 A: We remastered the original MQUAKE datasets [Zhong et al., 2023] with code-based flagging of
735 erroneous cases, then we fix them with manual curation or regeneration of LLM; see Section 4 for
736 details.

737 **What was the resource cost of collecting the data?** (e.g. what were the required computational
738 resources, and the associated financial costs, and energy consumption - estimate the carbon footprint.
739 See Strubell *et al.*? for approaches in this area.)

740 A: Given our MQUAKE-REMASTERED datasets are remastered based upon the already collected
741 MQUAKE datasets with its footprint unknown [Zhong et al., 2023], we can only report the additional
742 effort spent on the remastering process — which is within 100 GPU hours on a 640GB DGX A100
743 server.

744 **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic,
745 probabilistic with specific sampling probabilities)?** A: Our MQUAKE-REMASTERED datasets
746 are indeed technically a subsample of the original MQUAKE datasets. This is because there exists
747 duplication errors in the MQUAKE datasets, which we removed — see Section 3.5 for details.

748 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and
749 how were they compensated (e.g., how much were crowdworkers paid)?** A: Shaochen (Henry)

750 Zhong, Yifan Lu, and Lize Shao of Rice University are involved in the data collection process, with
751 SZ supported by the PhD stipend and undergraduate students YL and LS working in a volunteered
752 fashion to accumulate research experience.

753 **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so,
754 please provide a description of these review processes, including the outcomes, as well as a link or
755 other access point to any supporting documentation.

756 A: We are currently in the pipeline of the internal review process Meta. We direct interesting reader
757 to <https://ai.meta.com> for details.

758 **Does the dataset relate to people?** If not, you may skip the remainder of the questions in this
759 section.

760 A: Yes.

761 **Did you collect the data from the individuals in question directly, or obtain it via third parties or
762 other sources (e.g., websites)?** A: We obtained the source of our dataset from an already published
763 dataset: MQUAKE by Zhong et al. [2023].

764 **Were the individuals in question notified about the data collection?** If so, please describe (or
765 show with screenshots or other information) how notice was provided, and provide a link or other
766 access point to, or otherwise reproduce, the exact language of the notification itself.

767 A: Given the individuals involved in our datasets are often public individuals or celebrities owning
768 Wikipedia pages, we do not envision a need to individually notify them due to the source of our
769 dataset MQUAKE being already published under an MIT license. That being said, we did contact
770 the authors of MQUAKE datasets to confirm our finding is correct, which they acknowledged.

771 **Did the individuals in question consent to the collection and use of their data?** If so, please
772 describe (or show with screenshots or other information) how consent was requested and provided,
773 and provide a link or other access point to, or otherwise reproduce, the exact language to which the
774 individuals consented.

775 A: Given the two root sources of our MQUAKE-REMASTERED are Wikidata [Vrandečić and
776 Krötzsch, 2014] and MQUAKE [Vrandečić and Krötzsch, 2014], which respectively holds a CCO
777 (public domain) license and an MIT license, so there is no individual consent needed.

778 **If consent was obtained, were the consenting individuals provided with a mechanism to revoke
779 their consent in the future or for certain uses?** If so, please provide a description, as well as a
780 link or other access point to the mechanism (if appropriate)

781 A: N/A.

782 **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data
783 protection impact analysis) been conducted?** If so, please provide a description of this analysis,
784 including the outcomes, as well as a link or other access point to any supporting documentation.

785 A: We provide the error analysis of the original MQUAKE datasets [Zhong et al., 2023], which is
786 also the impact of our MQUAKE-REMASTERED datasets, as the latter is proposed a fix to the former.

787 **Any other comments?** A: N/A.

788 **F.4 Preprocessing / Cleaning / Labeling**

789 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing,
790 tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing**

791 **of missing values)?** If so, please provide a description. If not, you may skip the remainder of the
792 questions in this section.

793 A: No. Given we process upon an already processed/cleaned/labeled dataset MQUAKE [Zhong et al.,
794 2023].

795 **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support**
796 **unanticipated future uses)?** If so, please provide a link or other access point to the “raw” data.

797 A: N/A.

798 **Is the software used to preprocess/clean/label the instances available?** If so, please provide a
799 link or other access point.

800 A: N/A.

801 **Any other comments?** A: N/A.

802 **F.5 Uses**

803 **Has the dataset been used for any tasks already?** If so, please provide a description.

804 A: Yes, our MQUAKE-REMASTERED datasets are applied to evaluate multi-hop knowledge editing
805 methods, as benchmarked in Section 5.

806 **Is there a repository that links to any or all papers or systems that use the dataset?** If so,
807 please provide a link or other access point.

808 A: Our GitHub repo (which we shared with our reviewers via a .zip file as supplementary materials,
809 and will be public at <https://github.com/henryzhongsc/MQuAKE-Remastered> by the time of
810 camera-ready) contains links to the GitHub Repos of all benchmarked methods. Additionally, we cite
811 the corresponding papers to all covered methods in Section 5.1.

812 **What (other) tasks could the dataset be used for?** A: Our dataset is specifically designed for
813 multi-hop knowledge editing evaluations. Though, technically one may also ignore all the editing
814 adjustments and treat the unedited setting of our dataset as a typical question-answering dataset for
815 LLM.

816 **Is there anything about the composition of the dataset or the way it was collected and prepro-**
817 **cessed/cleaned/labeled that might impact future uses?** For example, is there anything that a
818 future user might need to know to avoid uses that could result in unfair treatment of individuals or
819 groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms,
820 legal risks)? If so, please provide a description. Is there anything a future user could do to mitigate
821 these undesirable harms?

822 A: Not to the best of our knowledge.

823 **Are there tasks for which the dataset should not be used?** If so, please provide a description.

824 A: We only recommend using our dataset as to evaluate multi-hop knowledge editing methods.
825 Alternative usage of our dataset may lead to unreliable results at the practitioner’s own risk.

826 **Any other comments?** A: N/A.

827 **F.6 Distribution**

828 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution,**
829 **organization) on behalf of which the dataset was created?** If so, please provide a description.

830 A: We will distribute our MQUAKE-REMASTERED datasets at GitHub
831 <https://github.com/henryzhongsc/MQuAKE-Remastered> and HuggingFace by the time of
832 camera-ready.

833 **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset
834 have a digital object identifier (DOI)?

835 A: We will distribute our MQUAKE-REMASTERED datasets at GitHub
836 <https://github.com/henryzhongsc/MQuAKE-Remastered> and HuggingFace by the time of
837 camera-ready. Given the dataset is currently under Meta’s internal review, we will obtain a DOI once
838 the review is finalized.

839 **When will the dataset be distributed?** A: We aim to distribute our MQUAKE-REMASTERED
840 datasets before the camera-ready deadline of NeurIPS Datasets and Benchmarks Track, which is Oct
841 30, 2024.

842 **Will the dataset be distributed under a copyright or other intellectual property (IP) license,**
843 **and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and
844 provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU,
845 as well as any fees associated with these restrictions.

846 A: We plan to distribute our MQUAKE-REMASTERED datasets under the CC BY 4.0 license, which
847 is also supplied in our .zip file of our supplementary material.

848 **Have any third parties imposed IP-based or other restrictions on the data associated with**
849 **the instances?** If so, please describe these restrictions, and provide a link or other access point
850 to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these
851 restrictions.

852 A: No.

853 **Do any export controls or other regulatory restrictions apply to the dataset or to individual**
854 **instances?** If so, please describe these restrictions, and provide a link or other access point to, or
855 otherwise reproduce, any supporting documentation.

856 A: No.

857 **Any other comments?** A: N/A.

858 **F.7 Maintenance**

859 **Who is supporting/hosting/maintaining the dataset?** A: The authors from Rice University and
860 Meta.

861 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** A:
862 Please correspond to shaochen.zhong@rice.edu for dataset-related inquiries.

863 **Is there an erratum?** If so, please provide a link or other access point.

864 A: No, but we plan to keep one should we find any in our (later releasing) GitHub repo
865 <https://github.com/henryzhongsc/MQuAKE-Remastered>.

866 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
867 If so, please describe how often, by whom, and how updates will be communicated to users (e.g.,
868 mailing list, GitHub)?

869 A: We don't expect much adjustment of our dataset outside error fixing to provide a constant
870 benchmark that is referencable across methods. For error-fixing updates, we will push our new
871 package on GitHub that notifies all followed users.

872 **If the dataset relates to people, are there applicable limits on the retention of the data associated**
873 **with the instances (e.g., were individuals in question told that their data would be retained for a**
874 **fixed period of time and then deleted)?** If so, please describe these limits and explain how they
875 will be enforced.

876 A: No.

877 **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please
878 describe how. If not, please describe how its obsolescence will be communicated to users.

879 A: As we only plan on including error-fixing updates, the older datasets will be discontinued because
880 there is no point in maintaining an error-contained dataset. That being said, we will provide legacy
881 dataset checkpoints per GitHub commit history.

882 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for**
883 **them to do so?** If so, please provide a description. Will these contributions be validated/verified?
884 If so, please describe how. If not, why not? Is there a process for communicating/distributing these
885 contributions to other users? If so, please provide a description.

886 A: Much like our MQUAKE-REMASTERED datasets are built upon the MQUAKE datasets [Zhong
887 et al., 2023], we encourage others to extend/modify/contribute to our dataset. Our datasets will be
888 released under the CC BY 4.0 license, meaning all future adaptation is allowed. Interested contributors
889 may contact us via `shaochen.zhong@rice.edu` or simply leverage Git/GitHub features like issues
890 and pull requests.

891 **Any other comments?** A: N/A.