MQUAKE-REMASTERED: MULTI-HOP KNOWLEDGE EDITING CAN ONLY BE ADVANCED WITH RELIABLE EVALUATIONS

Anonymous authors
Paper under double-blind review

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. We observe that many methods try to overfit the original MQUAKE by exploiting some dataset idiosyncrasies of MQUAKE. We provide a guideline on how to approach such datasets faithfully and show that a simple, minimally invasive approach can bring excellent editing performance without such exploitation.

1 INTRODUCTION

Given the widespread public-facing popularity of various Large Language Model-powered (LLM) products (Zhao et al., 2023; Yang et al., 2024b), even an occasional user has likely experienced LLMs giving out erroneous answers to factually rooted, knowledge-intensive questions. While 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, nor undergoing another extensive pretraining or finetuning section (Sinitsin et al., 2020; Mitchell et al., 2022).

With the growing need for 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 simple and direct editing targets. 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-TRIAL 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., — as an intentionally extreme case — *"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. Even if it is feasible, this poses high operational costs and a repeated effort would be required should Messi ever opt to transfer again.

Intuitively, a practical knowledge editing method needs 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, most popular, and only publicly available reflective 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, see Table 7).

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 evaluated on MQUAKE. The issues with MQUAKE are significant and growing, especially as MQUAKE becomes a widely adopted dataset in the editing community. Given its importance for building more reliable LLMs — a critical aspect of NLP development — we present our work to advance multi-hop knowledge editing 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.
- **Present a faithful yet beyond SOTA pilot method for future multi-hop knowledge editing development.** 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 approach these datasets faithfully and additionally show that a simple, minimally invasive method with no such overfitting operations can also achieve excellent editing performance.

2 PRELIMINARY

2.1 BACKGROUND OF MQUAKE

MQUAKE (<u>M</u>ulti-hop <u>Qu</u>estion <u>A</u>nswering for <u>K</u>nowledge <u>E</u>diting) 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 & Krötzsch, 2014), which, in its rawest format, is a knowledge graph consisting 15+ trillion of Resource Description Framework (RDF) triples[1]. 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 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, a subset of MQUAKE-CF that only contains 3,000 cases out of the original 9171 cases. Authors of MQUAKE evaluate their proposed method, MeLLo (Zhong et al., 2023), upon this MQUAKE-CF-3K dataset; which then become an unspoken standard for 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 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.3. We also collect the dataset statistics in Table 7.

2.2 EVALUATING USING MQUAKE

Datasets like MQUAKE-CF and MQUAKE-CF-3K are often tested under varying "editing intensities," based on the number of cases considered "edited." This simulates different levels of deviation between the model's learned knowledge and the newly edited information. This approach is effective because strong knowledge editing methods should handle both large-scale updates and smaller, more localized edits, ensuring that the changes do not interfere with unrelated knowledge.

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 6. This kind of report granularity 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.

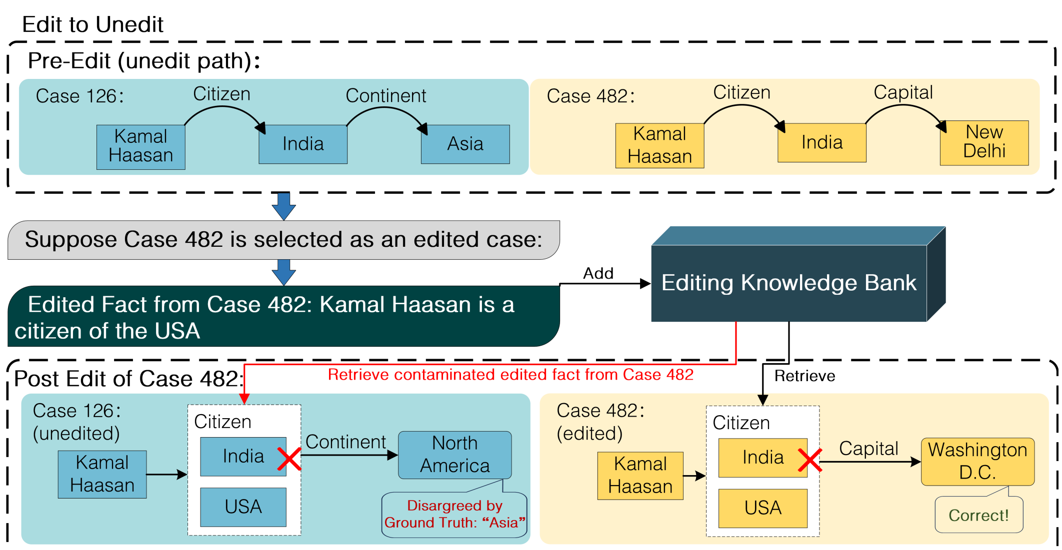

Figure 1: Example of contamination between an edited case and an unedited case

[1] https://www.wikidata.org/wiki/Property:P10209

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 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 6 makes sense 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 illustrated in Figure 1. When case 482 is selected as an edited case, the edited fact in case 482 would contaminate the unedited case 126 since both questions would ask for "The citizenship of Kamal Haasan" and the corresponding edited fact would be retrieved. This leads to the model generating an answer in conflict with MQUAKE-CF-3K's label, causing inaccurate experiment readings. See Appendix B.1 for a detailed walk-through.

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 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 also note that this contamination decreases as the number of edited cases (k-edit) increases, but it's simply a result of fewer unedited cases being available for contamination as k grows. For example, in the extreme case of 3000-edit, there is no contamination between edited and unedited cases because all cases are edited. However, 3000-edit has the highest level of contamination within edited cases, which we explore further in §3.2.

3.2 INNER CONTAMINATION BETWEEN DIFFERENT EDITED CASES

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 others[3] as illustrated in Figure 2. Similar to §3.1, an edited fact from case 1968 would alter the answer to an unedited hop in the edited case 1570. So, the model would generate an answer in conflict with the dataset ground truth label, causing inaccurate experiment readings. See Appendix B.2 for a detailed walk-through.

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,

[2]We 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.

[3]Note, an edited case does not require all of its subquestions being edited, but merely one of it (Table 7)

thus more edited-to-edited contamination. Notably, **under the 3000-edit tasks, almost one-third (998/3000, ≈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`).

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

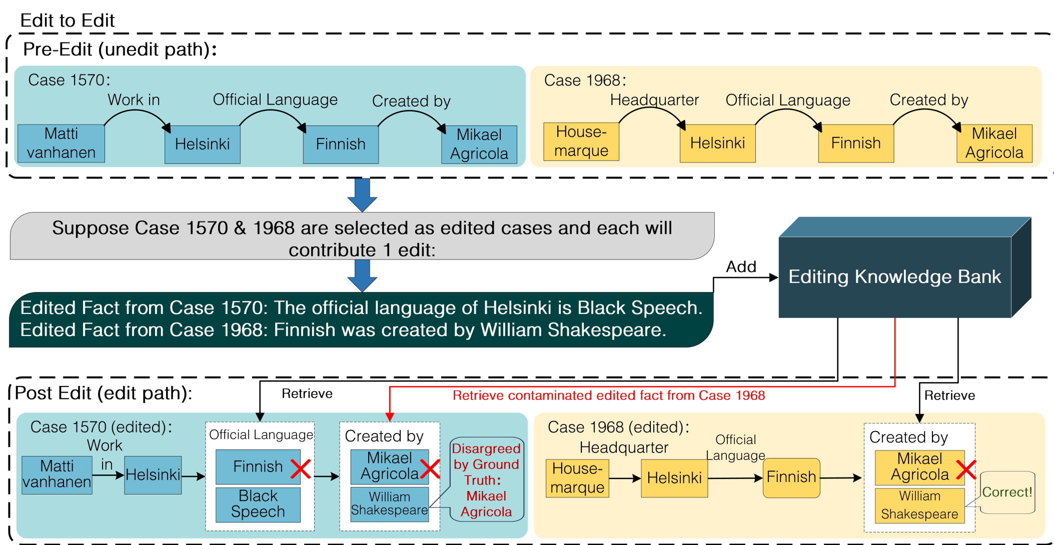

Figure 2: Example of contamination between two edited cases

3.3 CONFLICTING EDITS

The two types of contamination introduced in §3.1 and §3.2 are indeed subtle and hard to detect. However, MQUAKE-CF-3K also includes some straightforward edit conflicts, such as for the sub-question *"Which company is Ford Mustang produced by?"* we have the following edits:

- ◇ `case_id:2566` (edited): ~~Ford Moter Company~~ Nintendo.
- ◇ `case_id:231/2707` (edited): ~~Ford Moter 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 & 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 three multi-hop question instructions in a natural language format. During 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 question with "Albert Mohler" as the subject and `(employer, religion or worldview)` as the relation chain. 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?*

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 this omission of 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 — i.e., they are the carbon copy of each other, yet simultaneously exist in the dataset. 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. We further provide audit correctness analysis in Appendix C. Furthermore, we demonstrate the impact of our improvements through ablation studies that analyze the types of errors addressed, as discussed in Appendix D.

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 consistency, we rewrite the natural language question instructions for all questions in the datasets using meta-llama/Llama-3.1-405B (Dubey et al., 2024).

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, 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 cases caused by contaminations. Due to the nature of the N-hop question, at most N edited facts would be removed for each case, marginal compared to the number of edited facts in Table 12.

Specifically, once `case_id`-of-interest is given, our API would loop through all of its subquestions and identify if any is considered edited under another case. If there is a hit, the triple for such edited subquestions is removed from the bank of edited triples in constant time. 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 E.1).

5 Making Safe and Faithful Approach to MQuAKE and MQuAKE-Remastered

In addition to our dataset audit, fix, and the benchmark results we'd show below, 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 dataset idiosyncrasies unique to MQuAKE. For example, methods like GLAME (Mengqi et al., 2024) utilize Wikidata (Vrandečić & 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.

5.1 A Minimally Invasive but Performant Approach: GWalk

Here, we provide a brief walkthrough of a simple method we designed, namely GraphWalk. GWalk does not leverage any data-specific property unique to MQuAKE or MQuAKE-Remastered. Yet, it still presents SOTA performance surpassing many, if not all, established baselines. **We illustrate this pilot method as concrete guidance and potential inspiration to our future multi-hop knowledge editing scholars.**

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.

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.** 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 of a certain edited fact is asked.** Methods like MeLLo rely on RAG from a pool of edited facts in NL format. This can lead to unintended retrievals, where irrelevant facts with similar embeddings are retrieved, potentially causing hallucinations. 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. We share the detailed pseudocode of GWalk in Algorithm 1 and demonstrate some case studies in Appendix G.

Algorithm 1: General Procedure of GWalk on One 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;
$o_p =$ None, the answer from the previous subquestion.

$s \leftarrow$ Extracted subject from Q;
$rels \leftarrow$ 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'] */
for $r \in rels$ **do**
  Query $< s, r, ? >$ 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. */
  Prompt M to generate subquestion q_i with s and r.
  $o_p \leftarrow$ the M-generated answer to q_i.
  if $T(s, r)$ has a valid retrieval $< s, r, o^* >$ **then**
    $o_p \leftarrow o^*$;
  /* The answer to this subquestion will be the start subject of the next subquestion. */
  $s \leftarrow o_p$;
  $i \leftarrow i + 1$;
Return o_p;

6 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. All experiments are conducted with one or more 80G NVIDIA A100 GPUs. Please refer to https://anonymous.4open.science/r/MQuAKE-Remastered-118E for assets.

6.1 Experiment Coverage

Compared Methods In this work, **we aim to cover most, if not all, open-sourced knowledge editing methods specifically evaluated on the original MQuAKE.** This includes MeLLo (Zhong et al., 2023), PokeMQA (Gu et al., 2024), RAE (Shi et al., 2024), and DeepEdit (Wang 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. General editing methods like ROME (Meng et al., 2022) and MEND (Mitchell et al., 2022) are also featured. We note that we are aware methods like GLAME (Mengqi et al., 2024), StableKE (Wei et al., 2024), Temple-MQA (Cheng et al., 2024), and GMeLLo (Anonymous, 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.

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.

6.2 Results and Discussion

Table 3: Performance Comparison of Original MQuAKE and our MQuAKE-Remastered datasets using llama3-8b (AI@Meta, 2024) with every case as edited case. The original MQuAKE cannot faithfully reflect the true capacities of the methods due to errors specified in §3, especially if the method-in-question is performant.

Method	MQuAKE-CF-3k Original	MQuAKE-CF-3k Remastered	MQuAKE-T Original	MQuAKE-T Remastered
MeLLo (Zhong et al., 2023)	6.7	**6.77**	30.84	**44.37**
GWalk (Ours)	36.23	**66.33**	46.41	**54.88**

Given that our MQuAKE-Remastered is mostly provided as a fix to MQuAKE, we would like to 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.

Table 4: Experiments on MQuAKE-Remastered-cf-6334 with numbers of edited cases and methods. Results are reported in the format: $\frac{\textit{Total Accuracy}}{\textit{(Test Edited Accuracy, Train Edited Accuracy, Unedited Accuracy)}}$

Method	MQuAKE-Remastered-cf-6334 100-edit	1000-edit	3000-edit	6344-edit
lmsys/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)
DeepEdit Wang et al. (2024)	<1	<1	<1	<1
GWalk (Ours)	**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)
mistralai/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)
DeepEdit Wang et al. (2024)	<1	<1	<1	<1
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/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)
DeepEdit (Wang et al., 2024)	24.13 (11.1, 19.78, 24.29)	24.35 (8.16, 20.52, 26.27)	21.01 (7.57, 19.65, 25.38)	18.90 (7.48, 18.81,28.57)
RAE (Shi et al., 2024)	29.33 (22.22, 12.09, 29.74)	25.65 (33.67, 11.67, 32.49)	15.59 (23.11, 10.12, 33.48)	11.58 (18.75, 11.39, 28.57)
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)

In Table 4, we present benchmark results on MQuAKE-Remastered-cf-6334. GWalk consistently outperforms other methods in terms of models and edit numbers. The "OOM" in ICE and IKE are due to memory overload from concatenating all edited facts in the in-context learning prompt. Whereas, the "<1" results likely stem from the LLM's failure to recognize the few-shot examples, often generating irrelevant tokens or failing to follow the few-shot format. This issue was observed with MeLLo using Meta-Llama-3-8B-Instruct, and with DeepEdit using vicuna-7b-v1.5 and Mistral-7B-Instruct-v0.2. **Due to page limitation, we refer our readers to Appendix H for benchmarks of MQuAKE-Remastered-cf, MQuAKE-Remastered-cf-3k, and MQuAKE-Remastered-t.** We present MQuAKE-Remastered-cf-6334 in main text solely because it can feature the most methods.

7 RELATED WORKS

Audit and Fix of MQUAKE To the best of our knowledge, no work has conducted a comprehensive audit to MQUAKE as we do, but two prior arts have touched on the errors existing in MQUAKE: GMeLLo (Anonymous, 2024) and DeepEdit (Wang et al., 2024).

Table 5: Comparison of error analysis/quantification/fix of MQUAKE provided in different works.

Ref.	Error Types Found	Error Quantified	Error Scopes Fixed	Cost of Fixing
GMeLLo (Anonymous, 2024)	Missing Instruction	No	No	N/A
DeepEdit (Wang et al., 2024)	Inner Contamination	CF-3K in **3000**-edit	CF-3K in **3000**-edit	998 out of 3000 cases removed from CF-3K
Ours	Intra Contamination, Inner Contamination, Conflicting Edits, Missing Instructions, Duplicated Cases	CF-3K in {**1, 100, 1000, 3000**}-edit, T in {**1, 100, 500, all**}-edit, CF-9K in {**1, 100, 1000, 3000, 6000, all**}-edit	CF-3K, T, CF-9K in **any**-edit Remastered-CF-6334 in **any**-edit	No case removed from CF-3K, CF-9K, or T.

Specifically, GMeLLo briefly discusses the inconsistency between the triple chain and the corresponding generated instructions in its §4.5.1, which is the same type of error we discussed in §3.4. However, GMeLLo merely presents two examples of such an error without providing any quantitative error analysis or fix; we did both in §3.4 and §4.1.

DeepEdit (Wang et al., 2024) discovered the same inner contamination error (edited-to-edited) as we discussed in §3.2, but limited to one dataset (MQUAKE-CF-3K) under one setting (when all 3000 cases are considered edited). Further, DeepEdit removed all 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 - and named it MQUAKE-2002. While this fix is, of course, helpful, we argue **our Remastered fixes are much more comprehensive and effective since they patched many more errors revealed in §3 (the other four types of errors still exist in MQUAKE-2002), and most importantly, done so without scarifying almost 1/3 of the capacity of the original dataset** thanks to masking utility we proposed in §4.2. We further demonstrate the quantifiable difference between our work, GMeLLo, and DeepEdit in Table 5.

Multi-hop Knowledge Editing Datasets RippleEdit (Cohen et al., 2023) is the only other publicly available multi-hop knowledge editing dataset. However, it is actually a single-edit dataset, meaning only one edited fact is addressed at a time. We consider this an oversimplification of real-world scenarios, where systems must handle multiple edits simultaneously. This design also inherently avoids contamination. For additional exercise, we convert RippleEdit to a multi-edit setup to 1) make it more challenging, 2) show that our audit can also "fix" issues within a different dataset, and 3) demonstrate our proposed GWalk is indeed faithful and doesn't depend on MQuAKE-specific data. More in Appendix F.

Benchmark and Guidance Our work re-benchmarks nearly all open-sourced knowledge editing methods on MQUAKE and guides on safely and faithfully approaching such datasets. To the best of our knowledge, no other work offers this level of benchmarking or touches on the same issues. Notably, we are likely the only work to evaluate on MQUAKE-CF/MQUAKE-CF-9K, the largest dataset that even the original MQUAKE paper did not assess due to resource constraints. Table 13 illustrates the significant difference in evaluation coverage between our work and previous efforts.

8 CONCLUSION

In this work, we conduct a comprehensive audit of MQUAKE and present MQUAKE-REMASTERED, which fixes many critical errors within MQUAKE. We further re-benchmarked almost 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 — an efficient yet capable baseline for future works.

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.

Keyuan Cheng, Gang Lin, Haoyang Fei, Yuxuan zhai, Lu Yu, Muhammad Asif Ali, Lijie Hu, and Di Wang. Multi-hop question answering under temporal knowledge editing. *arXiv*, 2024.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina

Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv*, 2023.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv*, 2023.

Tianjie Ju, Yijin Chen, Xinwei Yuan, Zhuosheng Zhang, Wei Du, Yubin Zheng, and Gongshen Liu. Investigating multi-hop factual shortcuts in knowledge editing of large language models, 2024. URL https://arxiv.org/abs/2402.11900.

Yi Lu, Jing Nathan Yan, Songlin Yang, Justin T Chiu, Siyu Ren, Fei Yuan, Wenting Zhao, Zhiyong Wu, and Alexander M Rush. A controlled study on long context extension and generalization in llms. *arXiv preprint arXiv:2409.12181*, 2024.

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

Zhang Mengqi, Ye Xiaotian, Liu Qiang, Ren Pengjie, Wu Shu, and Chen Zhumin. Knowledge graph enhanced large language model editing. *arXiv*, 2024.

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

Yucheng Shi, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, and Ninghao Liu. Retrieval-enhanced knowledge editing for multi-hop question answering in language models. *arXiv*, 2024.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov, and Artem Babenko. Editable neural networks. In *International Conference on Learning Representations*, 2020.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 2014.

Yiwei Wang, Muhao Chen, Nanyun Peng, and Kai-Wei Chang. Deepedit: Knowledge editing as decoding with constraints. *arXiv*, 2024.

Zihao Wei, Liang Pang, Hanxing Ding, Jingcheng Deng, Huawei Shen, and Xueqi Cheng. Stable knowledge editing in large language models. *arXiv*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024a. URL https://arxiv.org/abs/2407.10671.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 2024b.

Jiayi Yuan, Hongyi Liu, Shaochen Zhong, Yu-Neng Chuang, Songchen Li, Guanchu Wang, Duy Le, Hongye Jin, Vipin Chaudhary, Zhaozhuo Xu, Zirui Liu, and Xia Hu. KV cache compression, but what must we give in return? a comprehensive benchmark of long context capable approaches. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4623–4648, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.266.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv*, 2023.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *arXiv*, 2023.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? *arXiv*, 2023a.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv*, 2023b.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

A EXTENDED PRELIMINARY

A.1 DEMO REPORT OF MQUAKE

Table 6: 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

A.2 DATASET STATISTICS

Table 7: 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 8: 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 INTRA CONTAMINATION BETWEEN AN EDITED CASE TO AN UNEDITED CASE (§3.1)

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:126` (unedited): *What is the continent of the country where Kamal Haasan holds citizenship?*
  - ◇ What is the country of citizenship of Kamal Haasan? India.
  - ◇ What is the continent of India? Asia.
- `case_id:482` (unedited): *What is the capital of the country where Kamal Haasan holds citizenship?*
  - ◇ What is the country of citizenship of Kamal Haasan? India.
  - ◇ What is capital of India? New Delhi.

The correct pre-edited answer should be *"Asia"* and *"New Delhi"* respectively. As Kamal Haasan is an Indian citizen, India is located in Asia and is the capital of New Delhi. However, suppose `case_id:482` is sampled as an edited case while `case_id:126` 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:126` and the edited `case_id:482` share the same subquestion of *"What is the country of citizenship of Kamal Haasan?"* The answer of `case_id:482` will be rightfully updated to *"USA"* per the new knowledge. However, the unedited `case_id:126` still considers the original answer *"India"* to be correct, and is therefore contaminated by the edited case `case_id:482` in an unintended fashion. This is problematic because a successful knowledge editing method should be able to retrieve the edited knowledge — *"Kamal Haasan is a citizen of USA?"* — upon the relevant questions (in this case the shared one), and thus answering *"North America"* to `case_id:126`. This is technically correct, but in conflict with MQUAKE-CF-3K's label, causing inaccurate experiment readings.

B.2 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: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.

B.3 Error Analysis of MQuAKE-CF

Table 9: 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 10: 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

C Error Detection Procedure and Post-Audit Checking

In this section, we discuss how exactly we carry out our audit and fixes and how we conduct our post-audit checking to ensure our audited datasets are error-free to the best of our ability.

C.1 Intra and Inner Contamination

As discussed in §3.1 and §3.2, we observed that some edited facts were retrieved for subquestions that were not intended to involve an edit. We categorized this issue as contamination, where edited facts inadvertently influence the correct reasoning path. To carry out the audit, we made the following observation: regardless of whether a case is edited or unedited, a valid reasoning path must always exist from the initial subquestion to the last subquestion. Thus, suppose any unedited subquestion on this reasoning path shares the same subject and relation with a triple reflecting an edited fact; then this unedited subquestion is contaminated and therefore flagged.

We programmed the abovementioned filtering mechanism and identified the contaminated edit facts against different subquestions/cases. We then employed the API described in §4.2 to dynamically mask out contaminated cases. Last, we confirmed that there is no contamination remaining by re-executing our filtering program upon the dynamically masked dataset.

C.2 Conflicting Edits

As illustrated in §3.3, we noticed some edits within the editing knowledge bank are self-contradicting, where edits with the same subject and relation led to different tail entities. Again, we follow the intended reasoning path as introduced above and check if there are multiple edit-reflecting triples that share the source and relation with an edited subquestion. If so, this suggests there are conflicting edits. We flagged all those edited triples, put ones with shared sources and relations into the same group, then flipped the minority edit to the majority edit and updated their subquestions accordingly. We then reran the program to ensure no more flagged triples.

C.3 Multi-hop Question Instruction Rewrite

As highlighted in §3.4, we identified some questions lacked a complete set of relations in their instructions, thus essentially omitting necessary information for a model to provide the correct answer. We collect a list of synonyms of all relations of an editing path, then evaluate if a certain instruction is not using any of the corresponding synonyms when its reasoning path indicates it should reflect a certain relation.

Subsequently, we prompted the original meta-llama/Llama-3.1-405B to regenerate all instructions with the few-shot demonstration prompt demoed in Appendix E.2 and reran the detection procedure. This process resulted in only a small number of instructions that still didn't meet our predefined rules due to the fact that our lists of synonyms per each relation cannot be exhaustive by design. We then manually inspect and, in a few occupations, manually fix those flagged cases.

C.4 Duplicate Cases

Upon investigating conflicting edits, we accidentally discovered that there exist cases with identical reasoning paths to each other, as illustrated in §3.5. We simply opt to retain only one of such cases and remove the duplicated rests. We then keep track of a set of reasoning paths from all cases and see if the cardinality of the set is equivalent to the number of cases.

D ERROR TYPE ABLATION STUDY

In this section, we provide ablation studies demonstrating the benefits of addressing errors in the MQuAKE dataset, aligning with the observed error patterns in Tables 2 and 1. Using our proposed GWalk and the Llama-3.1-8B-Instruct model, we evaluate datasets corrected for major error types, including Inner Contamination, Intra Contamination, and Missing Information in Multi-hop Question Instructions. These error types significantly affect the performance of both edited and unedited accuracies. We opt to exclude minor errors, such as duplicate questions and conflict edits, which are automatically addressed across all settings due to their limited prevalence in the original datasets.

The observed impact is consistent with our analysis: The Inner Contamination fix has the most impact when editing intensity is high (e.g., 3000-edit). Yet, the Intra Contamination fix has the most impact with lower editing intensity (e.g., 100-edit). The Missing Instruction fix consistently improves performance across all editing intensities.

Table 11: Performance comparison across dataset variants of MQUAKE-CF-3K on meta-llama/Llama-3.1-8B-Instruct. Results are reported as $\frac{\textit{Total Accuracy}}{\textit{(Test Edited Accuracy, Unedited Accuracy)}}$.

Type of Errors Fixed	MQUAKE-CF-3K 100-edit	1000-edit	3000-edit
Meta-Llama/Llama-3.1-8B-Instruct (Dubey et al., 2024)			
None	45.47 (38, 45.72)	42.73 (41.3, 43.45)	39.57 (39.57, -)
Inner Contamination	46.83 (70, 46.03)	53.3 (73.9, 43)	71.36 (71.36, -)
Intra Contamination	71.17 (37, 72.35)	61.73 (40.9, 72.15)	39.87 (39.87, -)
Missing Instruction	49.33 (41, 49.62)	47.2 (45.6, 48)	45.1 (45.1, -)
All (Our proposed)	**76.83** (69, 77.1)	**75.03** (74.6, 75.25)	**71.53** (71.53, -)

E EXTENDED REMASTERING

E.1 CONTAMINATION FREE SUBSET: MQUAKE-REMASTERED-CF-6334

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

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

Table 12: The number of unique edited facts for a varied number of edited cases in MQUAKE-REMASTERED-CF

Number of Edited Cases	100	1000	3000	6000	All (9171)
Number of Unique Edited Facts	150	1171	2991	5137	7252

Table 13: Experiment coverage comparison among our and other works. For brevity and better relevance, "Method Coverage" only includes open-sourced methods specifically designed for multi-hop editing, as adopted single-hop editors are often too weak to deliver usable results. "Separate Metrics?" means that both the accuracy of edited cases and unedited cases are reported. We consider the inclusion of both metrics paramount, as editing is often a double-edged sword, causing potential hallucinations under unedited scenarios. Prior work often only tests on the former but ignores the latter. We did both in our work.

Ref.	**Dataset Coverage**	**Method Coverage**	**Separate Metrics?**	**Error Fix?**
MQuAKE (Zhong et al., 2023)	CF-3K {1, 100, 1000, all}-edit; T {1, 100, 500, all}-edit	MeLLo	No	No
Temple-MQA (Cheng et al., 2024)	CF-3K {1, 100, all}-edit; T {1, all}-edit	MeLLo, PokeMQA	No	No
Ju et al. (2024)	CF-3K {all}-edit	N/A	No	No
PoleMQA (Gu et al., 2024)	CF-3K {1, 100, all}-edit; T {1, 100, all}-edit	MeLLo, PokeMQA	No	No
Ours	CF-3K {1, 100, 1000, all}-edit; T {1, 100, 500, all}-edit; CF-9K {1, 1000, 3000, 6000, all}-edit; CF-6334 {100, 1000, 3000, all}-edit	MeLLo, ICE, IKE, PokeMQA, GWalk, RAE, DeepEdit	Yes	Yes

E.2 Prompt for Rewriting Instructions

Few-shot Prompt

Instruction: Given a chain of relations, generate 3 multi-hop questions that comprehensively include the semantics of the relations.

Example 1:
Relation Chain:
```
XXX -> 'The author of  is' -> ' is a citizen of' -> ?
```
Generated Questions:

1. What is the country of citizenship of the author of XXX?
2. What country is the author of XXX a citizen of?
3. What is the nationality of the author of XXX?

Example 2:
Relation Chain:
```
XXX -> ' was developed by' -> 'The chairperson of  is' -> '
is a citizen of' -> ' is located in the continent of' -> ?
```
Generated Questions:

1. What continent is the country located in, where the chairperson of the developer of XXX is a citizen?
2. On which continent is the country located, whose citizen is the chairperson of the company that developed XXX?
3. Which continent houses the country of the chairperson of the developer of XXX?

Example 3:
Relation Chain:
```
<The relational chain we want the generated questions to be
based on>
```

F RIPPLEEDIT

We consider MQuAKE's task design and setup to be more reflective of real-world editing tasks, as naturally, there will always be more than one edited fact stored for any system with reasonable complexity. That being said, we are happy to report our proposed pilot method, GWalk, performs decently on RippleEdit. Here are some snapshot results on Llama-2-7b-chat:

Table 14: Single-edit result of RippleEdit-Popular/Recent/Random. C1/2 means the edit is happening at the 1st or the 2nd hop (RippleEdit cases only have 2 hops).

Method	Popular C1 Acc.	Popular C2 Acc.	Recent C1 Acc.	Recent C2 Acc.	Random C1 Acc.	Random C2 Acc.
ROME	37.4	16.2	47.8	50.0	35.5	49.5
ICE	85.1	67.6	74.8	85.0	73.8	80.3
MeLLo	45.1	77.1	50.2	80.0	40.2	68.3
GWalk (ours)	**85.7**	**81.8**	**80.9**	**87.6**	**76.1**	**82.9**

We additionally convert RippleEdit to a multi-edit setup — i.e., there are multiple edited facts within the editing knowledge bank at the same time — to a) make it more challenging and, b) show that our audit can also "fix" issues within a different dataset. Note we put the "fix" in quotes as RippleEdit is not designed with multi-edit in mind, so the things we fixed are not necessarily errors but just some adjustments required for making a proper multi-edit dataset. In any case, here are the snapshot results on Llama-3-8b-Instruct:

Table 15: Multi-edit result of RippleEdit-Popular/Recent/Random. In this case, we fixed 21/2/0 conflict edits and 3/120/1 case-to-case contamination within RippleEdit-Ropular/Recent/Random datasets, respectively.

Method	Popular C1 Acc.	Popular C2 Acc.	Recent C1 Acc.	Recent C2 Acc.	Random C1 Acc.	Random C2 Acc.
MeLLo	35.1	40.3	41.1	42.4	49.5	50.0
GWalk (ours)	**79.0**	**66.9**	**79.2**	**63.9**	**72.9**	**60.0**

G CASE STUDY OF GWALK

We believe GWalk is performant and practical because of two ingredients:

- It only stores edited facts in its Editing Knowledge Bank (Figure 2), contrary to some baselines (e.g., RAE (Shi et al., 2024)), where unedited facts are also stored. This is more practical to maintain as there are always fewer edited facts to keep track of, yet the total search space is much smaller, allowing more precise and efficient retrieval.
- Unlike most baselines, which store edited facts in natural language (NL) format (e.g., MeLLo (Zhong et al., 2023) and the majority of existing works) and conduct retrieval based on NL sentence embeddings, we store such editing facts on a Knowledge Graph (KG). The topology-based retrieval greatly reduces unintended retrieval, which almost always causes hallucinations.

Here is a concrete example from MQuAKE-Remastered-CF (case #16), where MeLLo retrieves an incorrect edited fact on an edited subquestion.

Edited Subquestion Example

Question: What is the country of citizenship of Twitter's CEO?
1st Subquestion: Who is Twitter's CEO?
Generated Answer (by LLM): Twitter's CEO is Elon Musk.
MeLLo-retrieved edited fact: The chief executive officer of CBS Corporation is Steve Jobs.
// Incorrect edited fact retrieved because this edited fact is close to the subquestion from an embedding standpoint, even if it doesn't provide relevant information.
GWalk-retrieved edited fact: The chief executive officer of Twitter is Parag Agrawal.
*// This is a correct retrieval because we first identify (in a lossy fashion) entity **twitter** and relation **executive officer** in the KG storing edited facts.*

MeLLo 2nd Subquestion: What is the country of citizenship of Elon Musk?
GWalk 2nd Subquestion: What is the country of citizenship of Parag Agrawal?
// MeLLo eventually provides the wrong final answer because the rest of its subquestion is about Elon Musk, though it should be about Parag Agrawal. We note that this is an editing dataset, so the ground truth answers often don't reflect the situation in the real world.

Similarly, here's MeLLo retrieving an unrelated edited fact on an unedited subquestion (MQuAKE-Remastered-CF, case #70).

Unedited Subquestion Example

Question: What is the capital of the country where Premam originated?
1st Subquestion: Where Premam was originated?
Generated Answer (by LLM): Premam was originated in India.
MeLLo-retrieved fact: Carnatic music was created in the country of Poland.
// Unrelated edited fact retrieved even if this subquestion is not edited.
GWalk-retrieved fact: None.
*// No edited fact is retrieved because no triple (via lossy mapping) on the KG has a source of **Premam** with a relation of **originated in**.*

MeLLo 2nd Subquestion: What is the name of the capital city of Poland?
GWalk 2nd Subquestion: What is the capital city of India?
// MeLLo again eventually provides the wrong final answer because the rest of its subquestion is about Poland, though it should be about India.

H Additional Experiment Results

One observation we made in §6.2 is in-context learning-based methods — like ICE (Cohen et al., 2023) and IKE (Zheng et al., 2023a) tend to "OOM" when facing a larger amount of edited facts. This is because these two methods — originally designed for single-edit tasks — essentially dump all edited facts as a long concatenated prompt and expect the model to figure out the corresponding editings naturally. They face OOM issues because when the number of editing facts grows, the prompt becomes extremely long and, therefore, introduces a large amount of KV cache and poses significant memory footprint issues.

While efficiently and effectively handling long input is out-of-scope of our work, as general guidance, we refer interested readers to efficient long context-handing survey/benchmark works like Yuan et al. (2024), which cover the schools and performance of several popular long context-handling methods. Other than the system challenges, another necessary aspect is to improve LLM long context performance, as most LLMs are pre-trained on limited context length and thus cannot effectively handle long input even if the system challenge is addressed. In this regard, we again recommend survey/benchmark works like Lu et al. (2024) for insights. Further, one can certainly convert this long context scenario to leverage the power of the RAG pipeline, much like the majority of multi-hop knowledge editing methods featured in this work.

Table 16: This is the benchmark results of MQuAKE-Remastered-t. The reported format is:
Total Accuracy
(Edited Accuracy, Unedited Accuracy)

Method	**MQuAKE-Remastered-t**			
	1-edit	**100-edit**	**500-edit**	**1864-edit**
lmsys/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
DeepEdit Wang et al. (2024)	<1	<1	<1	<1
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)
mistralai/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
DeepEdit Wang et al. (2024)	<1	<1	<1	<1
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/Meta-Llama-3-8B-Instruct (AI@Meta, 2024)				
MeLLo (Zhong et al., 2023)	<1	1.13 (17, <1)	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
DeepEdit Wang et al. (2024)	6.49 (0, 6.49)	8.48 (36.0, 6.92)	14.74 (36.20, 6.89)	34.71 (34.71, N/A)
GWalk (Ours)	**70.12** (100, 70.1)	**73.28** (84.0, 72.68)	**76.61** (87, 72.8)	**84.01** (84.01, N/A)
meta-Llama/Llama-3.1-8B-Instruct (Dubey et al., 2024)				
GWalk (Ours)	**74.68** (100, 74.66)	**76.34** (85, 75.85)	**77.74** (85.4, 74.91)	**83.32** (83.32, N/A)
Qwen/Qwen2.5-7B-Instruct (Yang et al., 2024a)				
GWalk (Ours)	**44.23** (100, 44.21)	**46.03** (87, 43.71)	**55.1** (85.4, 43.99)	**86.32** (86.32, N/A)

Table 17: This is the benchmark result for MQuAKE-Remastered-cf-3k reported in the format of: *Total Accuracy* *(Edited Accuracy, Unedited Accuracy)*

Method	**MQuAKE-Remastered-cf-3k**			
	1-edit	**100-edit**	**1000-edit**	**3000-edit**
lmsys/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)	<1	<1	OOM	OOM
IKE (Zheng et al., 2023a)	<1	OOM	OOM	OOM
DeepEdit Wang et al. (2024)	<1	<1	<1	<1
GWalk (Ours)	**54.89** (100, 54.87)	**60.9** (54, 61.14)	**57.37** (54.4, 58.85)	**66.33** (66.33, N/A)
mistralai/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)	<1	<1	OOM	OOM
IKE (Zheng et al., 2023a)	<1	4.43 (4,4.49)	OOM	OOM
DeepEdit Wang et al. (2024)	<1	<1	<1	<1
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/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)	<1	<1	OOM	OOM
IKE (Zheng et al., 2023a)	<1	<1	OOM	OOM
DeepEdit Wang et al. (2024)	22.93 (0, 22.94)	17.27 (11, 17.48)	15.03 (15.1, 15.0)	12.63 (12.63, N/A)
GWalk(Ours)	**69.0** (100, 68.99)	**76.73** (67, 77.07)	**75.47** (74.2, 76.1)	**70.6** (70.6, N/A)
meta-Llama/Llama-3.1-8B-Instruct (Dubey et al., 2024)				
MeLLo (Zhong et al., 2023)	<1	<1	<1	2.5 (2.5, N/A)
GWalk (Ours)	**73.3** (100, 73.3)	**76.83** (69, 77.1)	**75.03** (74.6, 75.25)	**71.53** (71.53, N/A)
Qwen/Qwen2.5-7B-Instruct (Yang et al., 2024a)				
MeLLo (Zhong et al., 2023)	40.63 (100, 40.61)	40 (34, 40.21)	35.23 (23.9, 40.9)	23.1 (23.1, N/A)
GWalk (Ours)	**65.33** (100, 65.35)	**65.27** (65, 65.28)	**65.07** (68.4, 63.4)	**66.74** (66.74, N/A)

Table 18: Experiments on MQuAKE-Remastered-cf with numbers of edited cases and methods. Results are reported in the format: *Total Accuracy* / *(Edited Accuracy, Unedited Accuracy)*

Method	**MQuAKE-Remastered-cf** 1-edit	**1000-edit**	**3000-edit**	**6000-edit**	**9171-edit**
lmsys/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
DeepEdit Wang et al. (2024)	<1	<1	<1	<1	<1
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)
mistralai/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
DeepEdit Wang et al. (2024)	<1	<1	<1	<1	<1
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/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
DeepEdit Wang et al. (2024)	22.16 (100, 22.15)	19.26 (21.29, 19.01)	21.09 (24.48, 19.44)	23.04 (23.77, 21.67)	24.25 (24.25, N/A)
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)
meta-Llama/Llama-3.1-8B-Instruct (Dubey et al., 2024)					
GWalk (Ours)	**76.27** (1, 76.27)	**73.48** (73.1, 73.53)	**72.86** (71.98, 73.29)	**72.03** (70.96, 74.08)	**70.94** (70.94, N/A)
Qwen/Qwen2.5-7B-Instruct (Yang et al., 2024a)					
GWalk (Ours)	**64.4** (0, 64.41)	**62.61** (66.6, 62.12)	**63.35** (66.34, 61.9)	**64.93** (66.28, 62.4)	**66.79** (66.79, N/A)

Table 19: Additional experiments on meta-llama/Llama-3.1-8B-Instruct (Dubey et al., 2024) and Qwen/Qwen2.5-7B-Instruct (Yang et al., 2024a) on MQuAKE-Remastered-cf-6334. Results are reported in the format: $\frac{\textit{Total Accuracy}}{\textit{(Test Edited Accuracy, Train Edited Accuracy, Unedited Accuracy)}}$.

Method	MQuAKE-Remastered-cf-6334 100-edit	 1000-edit	 3000-edit	 6344-edit
	lmsys/vicuna-7b-v1.5 (Zheng et al., 2023b)			
ROME (Meng et al., 2022)	<1	<1	<1	<1
MEND (Mitchell et al., 2022)	12.75 (11.11, 11, 13.25)	10.36 (7.33, 9.6, 13.64)	9.56 (6.1, 7.2, 11.9)	7.24 (6.38, 6.49, 10.3)
GWalk (Ours)	**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)
	mistralai/Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)			
ROME (Meng et al., 2022)	<1	<1	<1	<1
MEND (Mitchell et al., 2022)	11.84 (11.11, 9, 12.36)	11.57 (6.95, 8.7, 12.12)	8.39 (3.41, 6.6, 10.1)	6.82 (2.33, 6.4, 8.4)
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/Meta-Llama-3-8B-Instruct (AI@Meta, 2024)			
ROME (Meng et al., 2022)	<1	<1	<1	<1
MEND (Mitchell et al., 2022)	13.04 (11.11, 10, 13.47)	13.3 (5.33, 8.4, 14.33)	9.81 (4.21, 8.63, 11.1)	7.42 (5.12, 7.45, 7.3)
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)
	meta-Llama/Llama-3.1-8B-Instruct (Dubey et al., 2024)			
GWalk (Ours)	**66.79** (33.33, 72, 66.66)	**73.66** (49.47, 73.68, 73.02)	**72.09** (51.23, 75.1, 70.6)	**73.3** (55.39, 73.84, 71.55)
	Qwen/Qwen2.5-7B-Instruct (Yang et al., 2024a)			
GWalk (Ours)	**60.59** (33.33, 62, 60.56)	**65.42** (30.13, 68.6, 63.83)	**68.75** (43.65, 69.9, 64.99)	**70.49** (59.12, 70.51, 68.25)