

## F Datasheet

We supply the datasheet of our proposed MQUAKE-REMASTERED datasets, following the format of [Geburu 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

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

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

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

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

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

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

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

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

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

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

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

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

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

A: MQUAKE-REMASTERED datasets are four parts: MQUAKE-REMASTERED-CF, MQUAKE-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.