

# Datasheets for RAGChecker

Anonymous Author(s)

June 2024

## 1 Datasheets for datasets

**Datasheets for Datasets** “document [the dataset] motivation, composition, collection process, recommended uses, and so on. [They] have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted biases in machine learning systems, facilitate greater reproducibility of machine learning results, and help researchers and practitioners select more appropriate datasets for their chosen tasks.”

The motivation behind the proposal was the electronics industry, where every component has a datasheet that describes its operating characteristics and recommended uses. In machine learning, data is the input for model training. Using the wrong dataset, or using a dataset outside of its original intent, or even not understanding well enough the limitations of a dataset, has dire consequences for the model. However, “[d]espite the importance of data to machine learning, there is no standardized process for documenting machine learning datasets. To address this gap, we propose datasheets for datasets.”

## 2 Template

### 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.**

Retrieval-Augmented Generation (RAG) has emerged as a promising approach for leveraging external knowledge for language generation. However, we find that there still lack of comprehensive evaluation benchmark

for evaluating RAG systems. Existing datasets are mostly about open domain question answering, which are either with a small number of questions, or with only short answers as the ground truth answer. We propose to repurpose the existing datasets and convert the short answers to long answers to fill this gap.

**Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., com-**

**pany, institution, organization)?**

Due to the anonymity of the reviewing procedure, the research group and funding could not be disclosed here. We will update this question after the double-blinded review procedure.

**Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

Due to the anonymity of the reviewing procedure, the research group and funding could not be disclosed here. We will update this question after the double-blinded review procedure.

**Any other comments?** No.

### 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.

This benchmark contains 10 datasets. Each dataset has a corpus and a set of question-answer pairs. The corpus is a list of documents. The question-answer pairs are human written questions and long-form answers which are either written by human or converted from short answers by GPT-4.

**How many instances are there in total (of each type, if appropriate)?**

There are totally 4,182 question-answer pairs and 1,510,503 documents.

**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).

For KIWI, we used the final answer rated as “good”. For ClapNQ, we use the dev set. For NovelQA, we randomly sample 20 novels without copyright issue. For RobustQA, we use the domains of Writing, Finance, Lifestyle, Recreation, Science, Technology, and we convert the short answers to long answers, then filter the instances whose long-answers contain hallucinations, and then randomly sample 500 questions for each of the domains. For BioASQ, we use the human-written long answers that have more than 50 words.

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

“Raw” text data that can be read by human and can be input into LMs.

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

Yes, each question has ground truth answer as the target.

**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.

No.

**Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?** Instances within one book are created by the same annotator; Some annotators annotated for different books while some did not.

This benchmark dataset is repurposed from existing datasets, the relationships between individual instances are not disclosed to the best of our knowledge.

**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.

No, this is a benchmark, it works for a test set.

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

The long answers in NovelQA and RobustQA are generated by GPT-4 condition on human annotated short answers, it will inevitably introduce noise. We have conducted automatic hallucination detection on the generated long answers, and remove the ones that contains hallucinations.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions

of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

This benchmark is self-contained.

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

No, all of them are from public datasets.

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

No.

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

No.

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

No. Individual information is not disclosed in this dataset.

**Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in**

**combination with other data) from the dataset?** If so, please describe how.

No. Individual information is not disclosed in this dataset.

**Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

No. Inapplicable.

**Any other comments?**

#### Collection Process

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

Directly observable.

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

The long answers in for NovelQA and RobustQA are generated by GPT-4-Turbo API.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

We first select instances whose long answers have more than 50 words. For RobustQA, we randomly sample 500 instances after the before mentioned filtering.

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

The human annotation for meta-evaluation stated in the paper are annotated by the authors of this paper.

**Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** If not, please describe the timeframe in which the data associated with the instances was created.

All data are newly annotated. No crawling is involved.

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

No.

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

No.

**Did you collect the data from the individuals in question directly, or**

**obtain it via third parties or other sources (e.g., websites)?**

No.

**Were the individuals in question notified about the data collection?**

If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No.

**Did the individuals in question consent to the collection and use of their data?**

If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No.

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

No.

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

No.

**Any other comments?** No.

## Preprocessing/cleaning/labeling

**Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The human annotation of meta-evaluation is labeled by human. Each instance is double annotated.

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

No.

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

No.

**Any other comments?** No.

## Uses

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

Yes, they are existing public datasets, so they are used for open domain question answering tasks already.

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

No.

**What (other) tasks could the dataset be used for?**

No.

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

No.

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

No.

**Any other comments?** No.

### Distribution

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

The annotations will be open-sourced. All the other parts of the dataset can be downloaded publicly, we will provide a script for the downloading.

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

We will release them in GitHub.

**When will the dataset be distributed?**

After the a reviewing process.

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

The license has not been determined at this stage.

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

No.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

**Any other comments?** No.

### Maintenance

**Who will be supporting/hosting/maintaining the dataset?**

The authors.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

Through the email provided in paper and issues in GitHub.

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

No.

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

Yes but only if necessary.

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

No.

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

No.

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

Yes, they can submit issues or pull request in the GitHub repo.

**Any other comments?** No.