

DATASET OVERVIEW

BASICS: CONTACT, DISTRIBUTION, ACCESS

1. Dataset name : MIMIC-R3G
2. Dataset version number: 1.0
3. Dataset contact information: [REDACTED]
4. Dataset access: External
5. Dataset will be accessed through public platform
6. Dataset Link: [REDACTED]
[REDACTED]

DATASET CONTENTS

7. We provide a benchmark dataset to train and evaluate report generation models in real-world scenarios, where models are required to write a radiology report following specific instructions or based on specific clinical contexts, such as generating reports using provided templates or referring to patients' medical histories as context.

The dataset can be divided into 5 different subtasks of report generations, each corresponding to one type of instructions or context information.

Report generation: This sub-task is the conventional report generation task without any additional instructions from radiologist or context information.

Report revision: Model revises the generated report based on straightforward instructions from human.

Template: Model generates report following any form of input template as instructions.

Previous Radiology: Model writes medical reports that not only focus on the current radiology image but also reference the patient's previous medical images and reports as context information.

Medical Records and Lab Tests as Context: Model generates reports based on medical records and lab tests as context.

Each item in the dataset contains a radiology report and a corresponding instruction or context information needed to write this report. Following is an example instance in the dataset:

Instruction: Based on the chest x-ray images and patient's medical details, draft a detailed diagnostic medical report

Context: Medical conditions of the patient: Echocardiogram shows decreased cardiac output or decreased ejection fraction, pulmonary function tests show decreased lung capacity or decreased oxygen saturation

Radiology Report: <content of the radiology report>

The context information is automatically generated based from radiology reports in MIMIC-CXR[1], using GPT₄ with a set of designed prompts.

INTENDED & INAPPROPRIATE USES

8. Purposes of the dataset: While automatic report generation has demonstrated promising results using deep learning-based methods, deploying these algorithms in real-world scenarios remains challenging. Compared to conventional report generation, real-world clinical scenarios require model to follow the instruction from the radiologists and consider contextual information, such as generating reports using provided templates or integrating patients' medical histories effectively. Such instructional report generation tasks are critical for enabling more accurate, customizable, and scalable report generation processes, but remain underexplored and lack substantial datasets for training and evaluation. The purpose of this dataset is to facilitate the training and evaluation of radiology report generation model, equipping the model with the capability of instruction following and context information consideration.
9. The dataset is for developing AI models and should not be used for training radiologist or any human subjects.

DETAILS

DATA COLLECTION PROCEDURES

10. Dataset Construction: The context information and instructions are generated based on the radiology report in MIMIC-CXR, using GPT₄.

Report Generation with no context: A manually designed instruction is added alongside the radiology reports in MIMIC-CXR, such as "Writing a radiology reports based on this X-ray image. "

Report revision: Given a radiology report in MIMIC-CXR, GPT₄ is instructed to produce a slightly modified report based on the input ground truth report, along with the instructions of how to revise the modified report into the correct ground truth report.

Template. We collect 10 report templates from medical professionals, and we leverage our pipeline to generate the structured version of the ground truth report based on a given template.

Previous Visit as Context: We select the studies with multiple visits in MIMIC-CXR, and set the radiology reports from earlier visits and the context.

Medical Records and Lab Tests as Context: GPT₄ is instructed to extract medical indications and inferring the plausible medical conditions, medical examinations and exam results based on the ground truth medical report.

11. A team of medical professionals are recruited with payment to evaluate the quality of the generated data. They are ask to verify if there are factual contradictions between the generated instructions / context and the original reports in MIMIC-CXR.
12. No data collection procedure is conducted for this dataset. All items in this dataset are generated based on the already existed public dataset MIMIC-CXR.

REPRESENTATIVENESS

13. As shown in Table 1, this dataset contains 321,594 radiology reports. Based on the advices from experts in radiology and healthcare, the dataset covers 5 distinct sub-tasks of different types of context information and instructions. Statistics of the datasets are as follows:

Table 1 Dataset Statistics

	No Context	Revision	Template	Previous Report	Medical Record	Total
Train	140,781	53,649	22,348	62,576	33,265	312,619
Test	2,020	2,020	1,703	1,253	1,969	8,975

14. Demographic groups: This dataset is based on MIMIC-CXR, which is collected from MIMIC-IV[2]. The demographics of patients in MIMIC-IV is as follows:

Table 2 Demographics for patients admitted to an intensive care unit (ICU) on MIMIC-IV

	Hospital admissions	ICU admissions
Number of stays	431,231	73,181
Unique patients	180,733	50,920
Age, mean (SD)	58.8 (19.2)	64.7 (16.9)
Female Administrative Gender, n (%)	224,990 (52.2)	32,363 (44.2)

Please refer to the original paper and documents of MIMIC-IV for more details.

DATA QUALITY

15. Noises: Since the context and instructions are generated by GPT. A small fraction of the samples in the training set may have context or instructions contradicted to the radiology reports. All samples in the testing set are validated by multiple medical professionals.
16. Data Quality Validation: We invite a group of certificated radiologists to validate the clinical correctness of the generated data, yielding a fully human-validated test set of 600 data examples, with 200 examples dedicated to each of the three sub-tasks: revision, template and medical records. The content for the other two sub-tasks are entirely written by radiologists with no LLM generation involved, so no additional validation was needed. The medical professionals are instructed to carefully examine all information, including instructions, context, modified reports, and ground truth reports, to determine whether the entire pipeline is acceptable. Any factual errors, such as missing positive findings or hallucinated false positives, will result in rejection. However, variations in writing styles are allowed, such as treating minor conditions not mentioned as negative. Annotators are also asked to rate the plausibility of each record on a scale from 1 to 10. This plausibility score is a subjective measure by medical practitioners to assess how likely the instructions or situations could occur in their daily work, ensuring that the setting aligns with real-world scenarios. The total acceptance rate is 95.5% (573 out of 600), including the 95% confidence intervals. The acceptance rates for the subtasks, correction, template and medical records, are 97.0%, 90.9%, and 99.5%, respectively. The overall average plausibility score for valid records is 9.58, demonstrating that the generated instructions effectively mirror daily scenarios.
17. This plausibility score is a subjective measure is a subjective measure, and the standards might varies across different annotators.
18. The instructions and contexts generated in training data may contain certain level of noises. User may need to consider address this issue when training model on this dataset.

PRE-PROCESSING, CLEANING, AND LABELING

19. Radiology reports are collected from MIMIC-CXR which is already preprocessed. Please refer to its original paper for the details of the preprocess. We select samples from the original MIMIC-CXR dataset that include at least one anteroposterior (AP) or posteroanterior (PA) images and a "findings" section presented in ground truth report. Labeling probable context information or instructions corresponding to specific radiology reports is automatically generated by GPT4. How different types of contexts and instructions are generated are introduced in bullet point 9. The generated data are further filtered for quality control. Specifically, we use GPT to filter samples that are incorrect/inconsistent with ground truth report, and we use CheXpert to filter samples that may have information leakage in the generated context.

Furthermore, a group of medical professionals are recruited to validate the clinical correctness of the generated data, resulting a fully human-validated test set of 600 data example

20. We have a recommended train/test split because the testing set is filtered and validated by medical professionals to ensure data quality. The training and testing sample will be stored in the separated directories in the dataset folder.

21. Link to the code for context and instructions generation and filtering:

PRIVACY

22. This dataset is constructed based on MIMIC-CXR. Following the usage notes of MIMIC-CXR, (1) the user will not share the data, (2) the user will make no attempt to reidentify individuals, and (3) any publication which makes use of the data will also make the relevant code available.
23. This dataset is constructed based on MIMIC-CXR, where all of the data have been de-identified. Radiology reports in MIMIC-CXR contain health conditions in the chest area of patients.

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?

24. Privacy Reviews: MIMIC-CXR was approved by the Institutional Review Board of BIDMC (Boston, MA). Requirement for individual patient consent was waived because the project did not impact clinical care and all protected health information was removed.

For the data annotation, the purpose of the annotation is to control data quality, it is not related to understanding user behaviors, characteristics or preferences. The users are not the subjects of study, so the privacy reviews are waived. Authors of MIMIC-R3G bear all responsibility in case of violation of rights.

ADDITIONAL DETAILS ON DISTRIBUTION & ACCESS

25. The information on dataset update will be presented on the Github repo that hosts this data.
26. When newer version of the dataset is released, the previous versions of the dataset will be labeled as deprecated, and a notice will be provided to guide users towards the new version.
27. Following MIMIC-CXR, Dataset will follow the PhysioNet Credentialed Health Data Use Agreement 1.5.0. <https://physionet.org/content/mimic-cxr/view-dua/2.0.0/>

REFERENCE

- [1] Johnson A E W, Pollard T J, Berkowitz S J, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports[J]. Scientific data, 2019, 6(1): 317.
- [2] Johnson A E W, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset[J]. Scientific data, 2023, 10(1): 1.