

Data Annotation Requirements for the MIMIC-M3G Dataset

Background:

The MIMIC-M3G dataset is designed for the development and evaluation of X-ray radiographic report generation models within simulated real-world interactions. It comprises five sub-tasks that cater to different aspects of report generation: context-free generation, report revision, template-based generation, generation with historical reports, and generation with additional records and tests. This dataset provides a comprehensive set of interactions and contextual information to meet practical requirements and evaluate the ability of models to handle complex interactions.

Overview:

- Dataset is comprised of items in text format with links to one or more corresponding images. In this annotation, we only consider the text content.
- Of the five sub-tasks, three of them require quality control annotations: report revision, template-based generation, and generation with additional records and tests. Record with severe artifacts should be flagged and filtered out by this annotation process.
- The annotations need to be performed under the supervision of a professional radiologist team.
- Annotation formats include multiple-choice for every item and an overall questionnaire.

Annotation Task:

Annotators will be responsible for marking approximately 600 records (200 for each subtask) across the following three sub-tasks. The specific number is determined by a combination of labeling speed, time, and cost, and is executed in batches. :

- **Report Revision:** Evaluation of revised reports against specific revision instructions.
- **Template-based Generation:** Assessment of reports structured using given templates.
- **Generation with Additional Records and Tests:** Review of reports that incorporate instructions, clinical context, and results from other tests.

Annotation Task:

For a specific annotation task, the entire record is presented to the annotator, including text content with the related images. After reading through the item, the annotators are expected to assess the item based on the following criteria:

- **Accuracy:** The content within the dataset should be correct, such as the ability to modify a report to match the reference report as per instructions, filling out the reference report accurately based on a template, and ensuring that patient historical records and results do not contradict objective facts. This reflects the internal logical consistency and correctness of the records. Consider only the correctness relevant to the task at hand, without needing to address the accuracy of the original ground truth diagnosis and images, or issues such as the writing style of the original ground truth reports.
- **Plausibility:** The degree to which the scenarios and settings could potentially exist in the real world. For instance, such revision instructions could be plausible in actual clinical practice, the method of filling out a template might be feasible, and the patient history along with laboratory test results could conceivably occur in a real patient's case. This reflects the reasonableness of the records as they are presented, in the context of the real world. As with correctness, consider only the overall task setup, disregarding the original diagnosis itself.

A Sample of Annotation Task:

Sample Record Contents provided to annotators (a report revision sub-task):

Task-report revision-0xxxx:

In correct report: An incorrect report to be revised,

Instructions: instructions to revision

GT: GT report, the revision result from in correct report and instructions,

Item annotation task:

1. Is this a correct record?
 - A. Yes
 - B. No: Reason: ____
 - C. Not Sure: Reason: ____
2. Grade the plausibility for this condition in a real-world scenario. (from 1 to 10)
Score: ____
If score less than 7, Please provide your reason: ____
3. Other Comments: _____

Annotator Qualifications:

Preferred candidates are practicing radiologists familiar with reporting protocols. Alternatively, students with experience in the field and knowledge of radiology reporting may also contribute to the annotation tasks.

Appendix: Sub-tasks Definitions and Evaluation Criteria

- **Report revision.**

Definitions:

This subtask simulates the scenario where a radiologist revises an incorrect report following specific instructions. It includes an incorrect report, a set of instructions,

and the ground truth (GT) as the expected outcome of the revision.

Correctness:

If the incorrect report could be revised following instructions to the GT report, the correctness of the record is Yes, otherwise the correctness is No.

Instructions may be detailed or brief. Additionally, it is assumed that if additional information from the images is needed during the correction process, such information is always accessible and the modifications made are correct.

Plausibility:

Upon reviewing this record, if the instructions for correcting the report and the overall process are plausible as they could exist, then it is considered plausible.

- **Template-based generation.**

Definitions:

This sub-task simulates a radiologist apply an existing report into a structured form following a certain template. It includes the ground truth report, a template, and a templated report.

Correctness:

If the templated report is acceptable from templating original text report into the template, the correctness of the record is Yes, otherwise the correctness is No. If there are more information in the text report and templated report missed because of lack of template area, it is acceptable.

Plausibility:

Reviewing this record, if the template and the process of templating is suitable in real scenarios, it is considered plausible. If the template is impossible in the real world for the GT report, then it is deemed irrational.

- **Generation incorporating other records and tests.**

Definitions:

This sub-task is set to simulate a radiologist write report while incorporating with indications, situations and other test results. It includes history items, tests, and the ground truth report.

Correctness:

If the history items and test results are possible truth when reporting, the correctness is Yes. The history items and test results act as a background and additional information, they are not required to be directly related to the report. This record is considered correct as long as it is a possible case. It is only deemed incorrect if there is a strong contradiction between the given history items and test results, or if there details factually conflict with the content in ground-truth report.

Plausibility:

If the history information and test results could possibly exist and be accessible during the reporting process, they are considered plausible. If it is not possible in any case, they are deemed irrational.