RadRevise: A Benchmark Dataset for Instruction-Based Radiology Report Editing

Yixuan Huang[1], **Julián Nicolás Acosta** [2], **Pranav Rajpurkar**[2]

[1]Harvard University
Cambridge, MA 02138
[2]Harvard Medical School
Boston, MA 02115
yixuan_huang@hsph.harvard.edu, julian_acosta@hms.harvard.edu, pranav_rajpurkar@hms.harvard.edu

Abstract

Large Language Models (LLMs) can assist radiologists by making precise edits to radiology reports based on human instructions. However, evaluating the quality of such modifications has been challenging due to the lack of publicly available datasets. To address this gap, we present RadRevise, a novel dataset for assessing models' ability to modify radiology reports according to specific instructions. RadRevise is derived from the radiology reports in the MIMIC-CXR dataset and includes 6,402 instructions and 2,922 modified reports. For each report, the dataset includes a set of one to five modification instructions, along with the corresponding modified output, covering various clinical topics and instruction types. Our benchmarking of current open-source models reveals performance gaps in accurately executing these instructions, highlighting areas for improvement in AI-assisted report modification.

Introduction

Radiology reports are free-text narratives in which radiologists document observations from imaging studies, compare current findings with previous examinations, and provide recommendations for patient care. These reports represent one of the core products of radiologists' work, requiring both significant time and specialized expertise. Many recent efforts have emerged to leverage current artificial intelligence (AI) technology to automate part or all of the reporting process and facilitate the radiology workflow, such as drafting reports based on radiographs (Hamamci, Er, and Menze 2024; Lee et al. 2024; Liu et al. 2021, 2023; Monshi, Poon, and Chung 2020; Parres, Albiol, and Paredes 2024; Yan et al. 2022), generating summaries of clinically significant findings from the report text (Liu et al. 2023; Ma et al. 2024), and detecting or correcting errors in reports (Gertz et al. 2024; Pellegrini et al. 2023; Tian et al. 2023).

In current clinical practice, radiologists typically dictate reports (Van Ooijen 2021), sometimes with the help of pre-written templates. Recent AI tools have enhanced such workflows by integrating unstructured text into reporting templates and incorporating information from prior reports alongside current findings, allowing radiologists to focus on documenting significant changes while AI generates the updated reports. As AI-assisted report generation becomes integrated into clinical practice, it will be essential to develop systems that enable radiologists to efficiently correct AI-generated errors. We argue that AI models capable of implementing specific instructed changes to reports offer great potential for enhancing radiologists' workflow, while preserving their established practices and tools.

While various recent works have explored high-level instruction following for radiology report generation and editing (Fleming et al. 2023; Pellegrini et al. 2023; Ranjit et al. 2024), there remains a critical lack of datasets with the level of granularity necessary to evaluate implementation of specific changes derived from the high-level instructions. Our work aims to address this gap by introducing a dataset with detailed instructions for revising radiology reports. Our key contributions include:

- Development of a systematic pipeline for generating detailed radiology report editing instructions, covering various types of modifications as well as clinically and anatomically relevant topics often found in radiology reports, using GPT-4 (OpenAI et al. 2023)
- Creation of a comprehensive dataset comprising 6,402 instructions for editing 2,922 reports, with corresponding modified outputs
- Evaluation of current open-source language models using our dataset, revealing performance gaps and providing insights to guide future improvements in instruction-based report editing

Related Work

Much recent work on AI for radiology reporting have focused on automated report generation, leveraging advancements in machine learning methods and large vision-language models. There has also been growing interest in developing instruction-tuned models that perform defined tasks on reports to support medical image interpretation and report analysis, including visual question answering, information extraction, and findings summarization.

Automated Report Generation

Research in automated report generation has explored a wide range of vision and language models. Early ap-

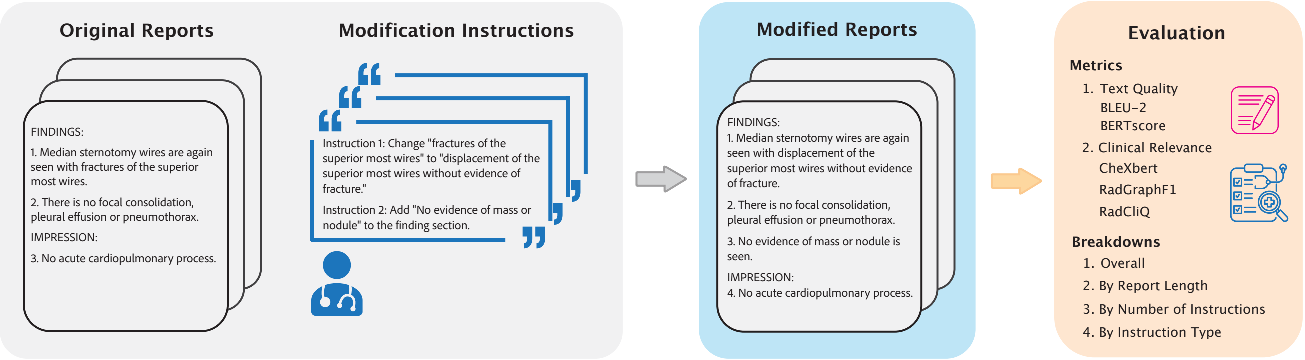
Original Reports
FINDINGS:
1. Median sternotomy wires are again seen with fractures of the superior most wires.
2. There is no focal consolidation, pleural effusion or pneumothorax.
IMPRESSION:
3. No acute cardiopulmonary process.
Modification Instructions
Instruction 1: Change "fractures of the superior most wires" to "displacement of the superior most wires without evidence of fracture."
Instruction 2: Add "No evidence of mass or nodule" to the finding section.
Modified Reports
FINDINGS:
1. Median sternotomy wires are again seen with displacement of the superior most wires without evidence of fracture.
2. There is no focal consolidation, pleural effusion or pneumothorax.
3. No evidence of mass or nodule is seen.
IMPRESSION:
4. No acute cardiopulmonary process.
Evaluation
Metrics
1. Text Quality
BLEU-2
BERTscore
2. Clinical Relevance
CheXbert
RadGraphF1
RadCliQ
Breakdowns
1. Overall
2. By Report Length
3. By Number of Instructions
4. By Instruction Type

Figure 1: RadRevise Dataset Overview and Evaluation Pipeline. RadRevise provides evaluation of model ability to modify reports given the original texts and modification instructions.

proaches utilized deep supervised learning architectures such as recurrent neural networks (Monshi, Poon, and Chung 2020) and transformers (Yan et al. 2022), as well as self-supervised methods such as contrastive language-image models (Thawkar et al. 2023; Hou et al. 2023; Wu et al. 2023). Some incorporate knowledge graphs (Zhang et al. 2020) and relation graphs (Jain et al. 2021) to enhance the generation process. Recent developments have leveraged LLMs (Gertz et al. 2024; Lee et al. 2024; Liu et al. 2023) for both report generation and error correction. Many of these approaches are multimodal, where radiology reports are generated or corrected based on the associated radiographs.

While automatic report generation has the potential to increase efficiency and reduce workload for clinical personnel, it cannot yet fully replace radiologist expertise. This limitation is partly due to the performance of current report generation models. For example, a set of GPT-generated reports from imaging show significantly lower Top-1 diagnosis accuracies compared to radiologists, potentially due to hallucinations (Nakaura et al. 2024). Similarly, multiple state-of-the-art language and vision-language models are found to have subpar performances when evaluated by diagnosis-related metrics (Tanno et al. 2023). These limitations in computational technologies—from traditional transcription tools to advanced LLM-based models—underscore the necessity of ongoing radiologist involvement in report correction and editing. Consequently, models capable of implementing precise report revisions based on specific instructions could provide valuable support to radiologists while enhancing the overall efficiency of radiology reporting.

Instruction Tuning for Radiology

There has been growing interest in developing instruction-following datasets for tasks related to radiology reports and other electronic health records. Some datasets target report generation tasks, such as describing observations based on associated radiographs (Liu et al. 2023; Zhang et al. 2024), or summarizing the reports to create the "impression" sections, which usually include significant findings and recommendations (Fleming et al. 2023). Others are designed for tasks such as visual question answering (Lee et al. 2023; Ranjit et al. 2024), translation (Fleming et al. 2023), and report-to-image generation (Lee et al. 2023).

While these datasets effectively support training and evaluating models on high-level instruction following, there remains a notable gap in datasets specifically designed for evaluating models' ability to implement granular report edits based on detailed instructions. Our work addresses this gap through RadRevise, a dataset that pairs specific editing instructions with their corresponding modified outputs. By providing a benchmark for instruction-based report editing, RadRevise complements existing datasets and enables comprehensive evaluation of models' capabilities across the radiology workflow.

Dataset

Pre-Processing

We derive the dataset from the reports in the test set of MIMIC-CXR (Johnson et al. 2019a,b). The MIMIC-CXR dataset contains 227,835 radiology reports from the Beth Israel Deaconess Medical Center for 65,379 patients, including 3,269 reports for 293 patients in the test set. MIMIC-CXR is released under the PhysioNet Credentialed Health Data License 1.5.0. We extract the Findings and Impression sections, the critical parts of radiology reports containing discussions and summaries of radiology findings (Johnson et al. 2019b), and retain the reports with at least one of these two sections. We then use Sentence Boundary Disambiguation (Sadvilkar and Neumann 2020) to split the report into individual sentences and number each of them. After preprocessing, we obtain 2,922 reports for 293 patients.

Data Generation

Using the MIMIC-CXR reports as the original texts, we create instructions to modify the reports and the corresponding edited reports. To generate applicable and clinically relevant instructions, we first define a set of action types, such as addition, removal, and change of observations, and specify their scope of application within the report, ranging from individual lines and sections to the entire document. We also identify a set of clinical and anatomical topics that commonly appear in radiology reports. This structured approach

No. of instructions	1	2	3	4	5
No. of reports	1,187	649	593	327	166

Table 1: Number of reports by number of instructions per report.

Type of instruction	Add observation	1,998
	Remove observation	464
	Change shape of observation	260
	Change position of observation	225
	Change severity of observation	269
	Change certainty of observation	261
	Change observation in other ways	667
	Append to line	687
	Remove from line	695
	Add recommendation	150
	Change/remove recommendation	7
	Add comparison	265
	Change/remove comparison	207
	Move line	247
Applicable location	For the entire report	1,391
	In the findings section	2,243
	In the impression section	1,035
	Of a line	1,733
Clinical topic	Opacity	755
	Hyperinflation	127
	Pleural abnormality	648
	Cardiomediastinal abnormality	475
	Chest wall abnormality	108
	Musculoskeletal abnormality	395
	Foreign body	293
	Medical device	811

Table 2: Distribution of instructions by type of instruction, location of edit, and major clinical topic category. A subset of 3,612 instructions explicitly addresses clinical topics.

with explicit instruction types and topics ensures the dataset includes diverse and relevant types of potential report modifications.

For each original report, we generate between one and five instructions to create varying levels of complexity. The distribution of the number of instructions per report is shown in Table 1. For each instruction, we randomly select the modification type, edit location, and clinical topic to form a prompt for GPT-4 (i.e., "provide an instruction to add an observation to Findings about edema"). To improve instruction relevance and applicability, we implement exclusion rules for certain combinations. For example, modifications to position and shape/size of observations are restricted to specific clinical topics, such as focal consolidation and nodules.

To generate the instructions and the modified reports, we provide GPT-4 with the original report, desired number of instructions, and specifications of each instruction. The complete prompt used is included in the Appendices, and samples of generated instructions are included in Table 3. We use the model GPT-4o with API version "2024-05-01-preview". The distributions of generated instruction types, locations, and topics are presented in Table 2.

Data Usage

RadRevise can be used independently or in combination with other instruction datasets to evaluate models for instruction following, report generation, editing, and error correction in radiology applications.

One practical application of RadRevise is to benchmark and enhance the performance of voice recognition and report dictation software. As discussed previously, while radiologists currently need to be involved in editing generated reports, incorporating language models capable of making precise and comprehensive edits can streamline this process. For example, when a radiologist instructs a model to correct an observation, the model should learn to implement the correction consistently throughout the entire report, eliminating the need for multiple manual edits of the same information.

RadRevise includes only the original reports and modification instructions, without the associated radiographs. This design choice aligns with the specific task RadRevise is meant to address—following instructions that have already been provided, without the need to analyze radiographs. While translating high-level instructions (e.g., "discuss the heart size") into detailed descriptions (e.g., "the heart appears normal") requires radiograph interpretation, the subsequent step of implementing these instructions does not. RadRevise focuses on supporting this implementation phase of report modification.

Benchmarking

Approach

We evaluate a diverse set of open-source models, including general-purpose models such as Mistral-7B (Jiang et al. 2023), Llama3-8B (AI@Meta 2024), and Falcon-7B (Almazrouei et al. 2023), as well as domain-specific ones like Medicine-Chat (Cheng, Huang, and Wei 2024), MedicalLlama3-8B (Vsevolodovna 2024) and Meditron-7B (Chen et al. 2023). We perform few-shot prompting on these models under the text generation task, and provide two examples of report revision in our prompt. We run the inference using the following parameters: temperature of 0.6, Top-p sampling of 0.9, repetition penalty of 1.2, and a batch size of 32 on a single RTX 8000. The prompt for generating the modified reports based on the original and the instructions is also included in the Appendices.

Evaluation Metrics

Using the metrics suitable for radiology reporting tasks (Yu et al. 2022), we benchmark the aforementioned models' abilities to perform instruction-based edits on report texts with few-shot prompting. These metrics together provide comprehensive information on the quality of the outputs, including semantic similarity to the references and the correctness of clinically relevant information.

1. BLEU-2 (Papineni et al. 2002): measures the overlap of bi-grams between the generated and reference reports.

Instruction Types	Location	Clinical Topic	Generated Instructions
Add observation	Findings	Lung opacity	Add "Lungs are clear" to the end of the Findings section.
Remove observation	Impression	-	Remove the observation about moderate pulmonary edema from the Impression section.
Change observation	Line	Nasogastric tube	Amend Line 6 to indicate that the tip of the nasogastric tube is visualized in the stomach.
Change location	Report	Orthopedic plate	Change the location of the orthopedic plate to left humerus in the entire report.
Change severity	Report	Cardiomegaly	Change the severity of the cardiomegaly from moderate to mild in the entire report.
Add comparison to priors	Line	Bone tissue	Add to Line 6 in the original report, "Comparison to prior studies shows the unchanged appearance of the bone tissue."
Remove comparison to priors	Line	-	Remove all comparisons to prior studies.

Table 3: Examples of generated instructions based on specifications.

Model	BLEU-2	BERTscore	CheXbert	RadGraph F1	RadCliQ-v0	RadCliQ-v1
Mistral-7B-Instruct-v0.3	**0.69 ± 0.01**	**0.78 ± 0.00**	**0.91 ± 0.00**	0.85 ± 0.01	**0.61 ± 0.02**	**-0.80 ± 0.02**
Meta-Llama-3-8B-Instruct	0.44 ± 0.01	0.44 ± 0.01	0.84 ± 0.00	**0.86 ± 0.01**	1.58 ± 0.03	-0.18 ± 0.02
Falcon-7B-Instruct	0.48 ± 0.01	0.59 ± 0.01	0.76 ± 0.01	0.73 ± 0.01	1.49 ± 0.04	-0.20 ± 0.03
Medicine-Chat	0.57 ± 0.01	0.43 ± 0.05	0.80 ± 0.01	0.00 ± 0.00	3.14 ± 0.14	0.95 ± 0.08
Medical-Llama3-8B	0.56 ± 0.01	0.65 ± 0.01	0.82 ± 0.01	0.74 ± 0.01	1.25 ± 0.04	-0.36 ± 0.03
Meditron-7B	0.38 ± 0.01	0.48 ± 0.01	0.72 ± 0.01	0.54 ± 0.01	2.14 ± 0.05	0.27 ± 0.03

Table 4: Evaluation results on various open source models (mean ± MOE_{95}).

2. BERTscore (Zhang et al. 2019): evaluates token similarities using contextual embeddings. We use "distilroberta-base" as the version of BERT and baseline-scaled scores.
3. CheXbert (Irvin et al. 2019) labeler vector similarity: compares label vectors derived from the generated reports to those from the ground-truths.
4. RadGraph (Jain et al. 2021) combined entity and relation F1: compares clinical entities and relations extracted from the generated reports to those from the ground-truth reports.
5. Radiology Report Clinical Quality (RadCliQ) (Yu et al. 2022): predicts a composite error score from the generated reports.

Results

Table 4 shows evaluation results comparing various open-source language models on the specified metrics. Notably, among the models evaluated, Mistral-7B-Instruct-v0.3 demonstrates superior performance across most metrics.

On the semantic metrics, Mistral-7B-Instruct-v0.3 achieves the highest scores in both BLEU-2 (0.69) and BERTscore (0.78), significantly outperforming other models. In clinical accuracy assessment, Mistral-7B-Instruct-v0.3 maintains strong performance with the highest CheXbert vector similarity (0.91) and competitive RadGraph F1 score (0.85). This is further demonstrated through the RadCliQ metrics, which provide composite scores incorporating both semantic and clinical aspects of performance.

Several factors could contribute to Mistral-7B-Instruct-v0.3, a general-purpose model, outperforming the domain-specific ones across these metrics. First, the evaluated version of Mistral model is instruction-tuned, which enhances its performance for this task. Second, while domain-specific models like Medicine-Chat and Meditron-7B possess medical knowledge, their specialization in tasks like diagnosis prediction and medical question-answering may not directly translate to effective report editing capabilities.

For most metrics, model performances tend to peak with medium-length reports (approximately 150-200 tokens) and decline with both shorter and longer reports (Figure 2). The drop-off at higher report lengths suggest that these models struggle with handling more complex or verbose inputs.

Performance generally decreases as the number of modification instructions increases, indicating that models find it challenging to maintain accuracy when implementing multiple changes concurrently (Figure 2).

Figure 3 shows model performance across different instruction types for single-instruction cases. While models generally perform similarly across tasks (adding, changing, or removing report content), they achieve best results when modifying the shape or severity of an observation. They struggle more with complex tasks like adding and removing comparisons to prior studies, likely because these require coordinated changes across multiple sections.

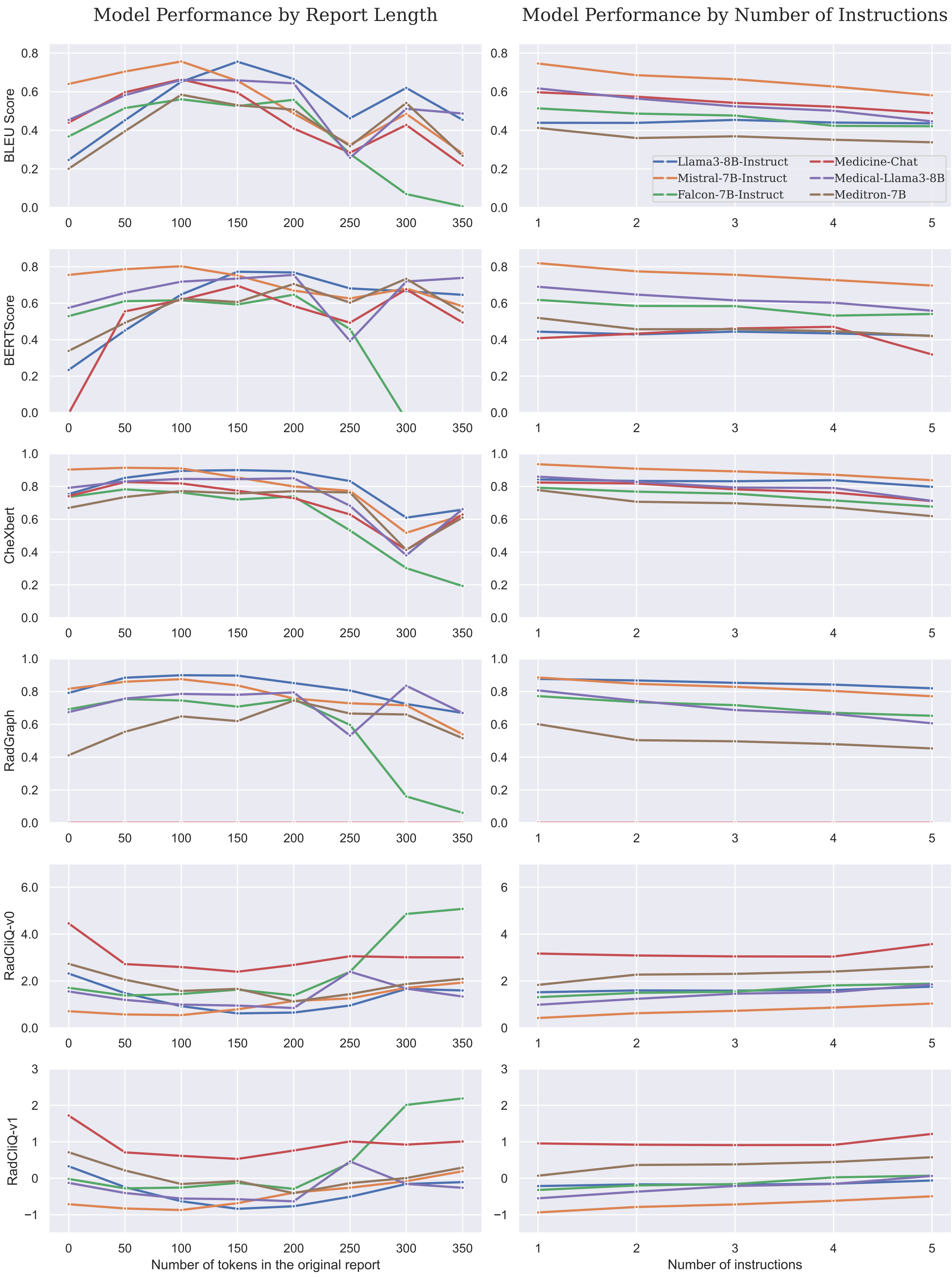

Figure 2: Model performance by report length and number of instructions.

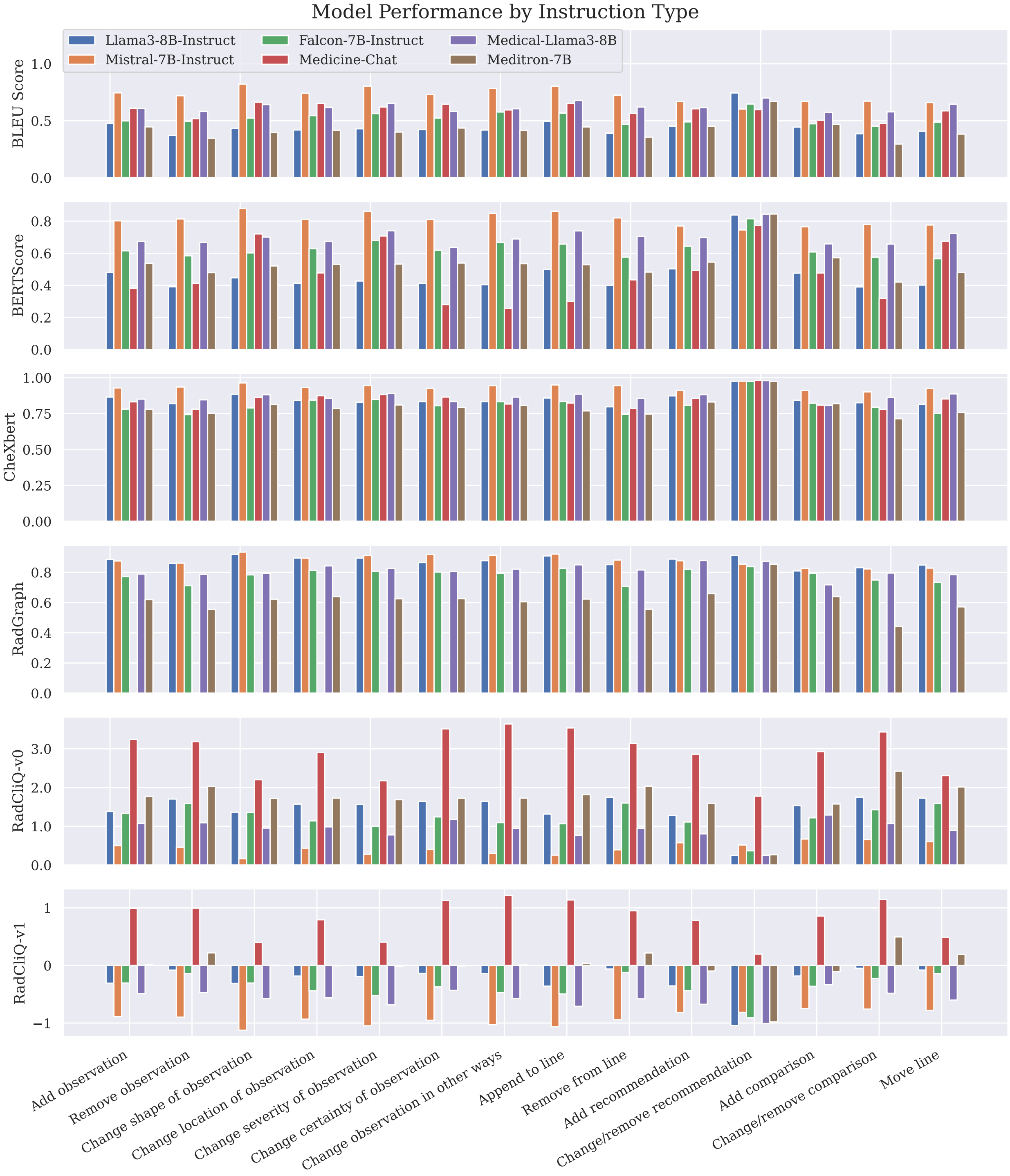

Figure 3: Model performance by instruction type for samples given a single instruction.

Limitations

RadRevise's current scope is limited to data from MIMIC-CXR, which represents specific distributions of patient characteristics and radiology report writing styles. Expanding to more diverse data sources, including other chest X-ray datasets and reports from different modalities such as CT scans, would enhance the dataset's generalizability and robustness. While our focus on precise edits addresses a specific data gap, future work should incorporate a broader spectrum of instruction types with varying levels of specificity to enable more comprehensive evaluation of model performance.

Our evaluation approach also presents certain limitations. Firstly, although we provide few-shot examples, we do not currently fine-tune any models on RadRevise due to the small size of our dataset. Model performance could improve significantly with fine-tuning on a larger collection of instruction-output pairs. Additionally, evaluating model performance on long reports with small changes poses a particular challenge: models may appear successful by simply reproducing the original report without implementing the requested modifications. While preserving unchanged content is desirable, we need more sophisticated methods to specifically assess models' ability to execute the requested modifications.

Conclusion

We present RadRevise, a dataset with specific and detailed instructions to edit a given radiology report, along with the modified output. This helps address the research and data gap in radiology report editing. RadRevise includes various instruction types, applicable locations within the reports, and a range of clinical topics, ensuring broad applicability. The dataset is also used to benchmark a set of existing models, revealing gaps in their ability to precisely follow given instructions for report editing. These findings underscore the importance of continued advancement in this area to enhance the efficiency and accuracy of the automated radiology reporting processes.

Appendices

Prompts for Data Generation

Instruction and Modified Report Generation. The following text prompt is provided to GPT-4o, API version "2024-05-01-preview", to generate the instructions and modified reports:

> You are an expert radiologist reviewing a radiology report written by your assistant. Your task is to:
>
> - Provide specific instructions for modifying the report.
> - Return the modified report based on those instructions.
>
> Rules:
>
> 1. The instructions should be well-defined, concise, and clinically correct.
> 2. Ensure that anyone, even without external context, can accurately apply the instructions to the report.
> 3. If the requested instruction does not apply to the given report, create a relevant instruction instead.
> 4. If multiple instructions are needed, ensure they align with the original report and are applied consistently throughout.
> 5. If modifications affect numbering, renumber the modified report accordingly.
> 6. Both the instructions and the modified report must be returned in the following JSON format.

~~~
```json
{
    "instructions":
    [
        {"instruction_id": 1, "
            description": "
            Description of the
            first instruction"},
        {"instruction_id": 2, "
            description": "
            Description of the
            second instruction"}
    ],
    "modified_report": "Updated
        report text"
}
```
~~~

Example 1: Create the following instruction(s): an instruction to add an observation to the report about consolidation; an instruction to remove an observation in the impression section. Then provide the modified report.

Original Report:

FINDINGS

1. AP and lateral views of the chest were obtained.
2. The right costophrenic angle is not fully included on the image.

IMPRESSION:

3. Top normal cardiac silhouette without pleural effusion or pulmonary edema.

Output:

~~~
```json
{
    "instructions":
    [
        {"instruction_id": 1, "
            description": "Add 'No
            focal consolidation,
            pleural effusion, or
            pneumothorax' to both the
             findings and impression
            sections."},
        {"instruction_id": 2, "
            description": "Remove the
             observation about the
            cardiac silhouette in the
             impression section."}
~~~

```
    ],
    "modified_report": "FINDINGS: 1.
        AP and lateral views of the
        chest were obtained. 2. The
        right costophrenic angle is
        not fully included on the
        image. 3. No focal
        consolidation, pleural
        effusion, or evidence of
        pneumothorax is seen.
    IMPRESSION: 4. No focal
        consolidation, pleural
        effusion, or evidence of
        pneumothorax."
}
```
```

Example 2: Create the following instruction(s): an instruction to change the anatomical location of an observation. Then provide the modified report.

Original Report:

FINDINGS:

1. Endotracheal tube is seen with tip in the right mainstem bronchus.
2. Hazy right basilar opacity may be due to atelectasis.
3. Right mainstem intubation is seen.

IMPRESSION:

4. Right mainstem intubation is seen.

Output:

```
 ```json
{
    "instructions": [
        {"instruction_id": 1, "
            description": "Change ’right
            mainstem intubation’ to ’left
             mainstem intubation’ in both
             the findings and impression
            sections."}
    ],
    "modified_report": "FINDINGS: 1.
        Endotracheal tube is seen with
        tip in the right mainstem
        bronchus. 2. Hazy right basilar
        opacity may be due to atelectasis
        .3. Left mainstem intubation is
        seen.
    IMPRESSION: 4. Left mainstem
        intubation."
}
```
```

Benchmarking Output Generation. The following text prompt is used to generate modified reports from original reports and modification instructions, using two-shot learning:

> You are an expert radiologist. You are given a radiology report and specific instructions to modify the report. Provide the modified report based on the instructions, without additional explanations or comments. Examples of input and output are as follows:

Example 1

Input:

- **Original Report:**
  FINDINGS

  1. Nasogastric tube has been advanced with the first side port in the proximal stomach.

  IMPRESSION:

  2. Nasogastric tube has been advanced.
  3. Overall no substantial change of the lungs.

- **Instructions:**
  - Instruction 1: Add “No pulmonary nodules or masses are identified” to the FINDINGS section.
  - Instruction 2: Remove the last line.

Output:

```
```json
{
    "modified_report": "FINDINGS: \n1
        . Nasogastric tube has been
        advanced with the first side
        port in the proximal stomach.
         \n2. No pulmonary nodules or
         masses are identified. \n\
        nIMPRESSION:\n3. Nasogastric
        tube has been advanced."
}
```
```

Example 2

Input:

- **Original Report:**
  FINDINGS:

  1. The lung volumes are low.
  2. Mild fullness in the right hila.
  3. No pneumothorax or pleural effusion.

  IMPRESSION:

  4. Mild fullness in the right hila.

- **Instructions:**
  - Instruction 1: Change “right hila” to “left hila” in Line 2 and Line 4.

Output:

```
```json
{
    "modified_report": "FINDINGS:\n1.
         The lung volumes are low.\n2
        . Mild fullness in the left
        hila.\n3. No pneumothorax or
        pleural effusion.\nIMPRESSION
        :\n4. Mild fullness in the
        left hila."
}
```
```
```

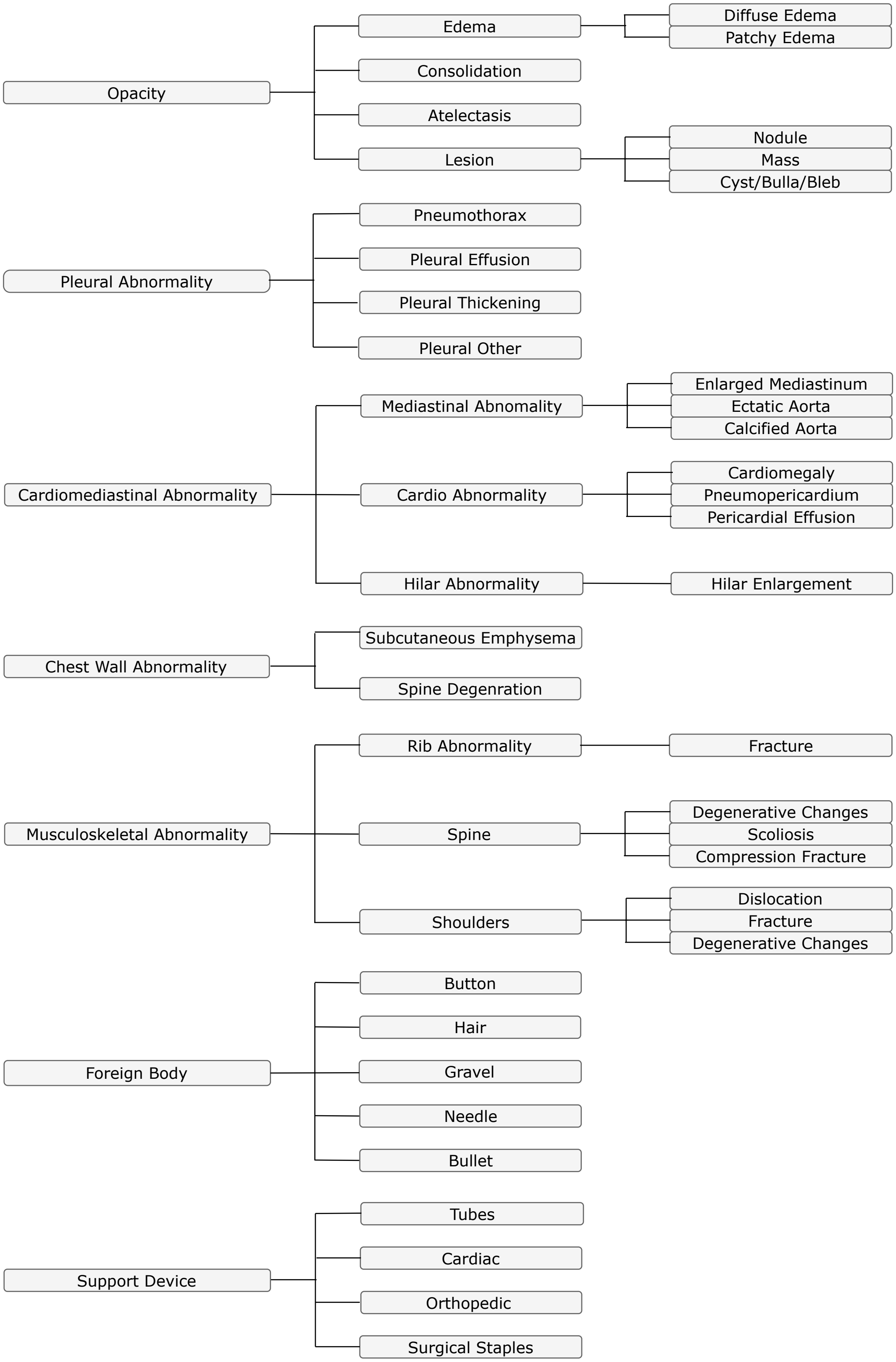

Figure 4: Clinical and anatomical topics covered in the instructions, adapted from Agarwal et al. (2023).

List of Clinical Topics

Figure 4 includes the set of clinical topics explicitly covered by the report modification instructions.

Ethical Statement

Patient Consent And De-Identification

The requirement for individual patient consent was waived for MIMIC-CXR. The reports were also de-identified according to HIPAA standards (Johnson et al. 2019a).

Impacts

To avoid potential harm from using our methods and dataset, researchers should exercise caution in the following ways: First, before clinical deployment, models evaluated on this dataset should undergo rigorous validation by clinical experts, assessing both performance and patient safety across diverse real-world scenarios. Second, when creating new datasets using our methodology, instructions and modified outputs should be carefully reviewed for potential biases, as models trained on biased data can perpetuate inequities in healthcare outcomes (Mac Namee et al. 2002; Mehrabi et al. 2022).

References

Agarwal, N.; Moehring, A.; Rajpurkar, P.; and Salz, T. 2023. Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. Technical Report w31422, National Bureau of Economic Research, Cambridge, MA.

AI@Meta. 2024. Llama 3 Model Card.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Goffinet, E.; Heslow, D.; Launay, J.; Malartic, Q.; Noune, B.; Pannier, B.; and Penedo, G. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Chen, Z.; Hernández-Cano, A.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; Sallinen, A.; Sakhaeirad, A.; Swamy, V.; Krawczuk, I.; Bayazit, D.; Marmet, A.; Montariol, S.; Hartley, M.-A.; Jaggi, M.; and Bosselut, A. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. _eprint: 2311.16079.

Cheng, D.; Huang, S.; and Wei, F. 2024. Adapting Large Language Models via Reading Comprehension. In *The Twelfth International Conference on Learning Representations*.

Fleming, S. L.; Lozano, A.; Haberkorn, W. J.; Jindal, J. A.; Reis, E. P.; Thapa, R.; Blankemeier, L.; Genkins, J. Z.; Steinberg, E.; Nayak, A.; Patel, B. S.; Chiang, C.-C.; Callahan, A.; Huo, Z.; Gatidis, S.; Adams, S. J.; Fayanju, O.; Shah, S. J.; Savage, T.; Goh, E.; Chaudhari, A. S.; Aghaeepour, N.; Sharp, C.; Pfeffer, M. A.; Liang, P.; Chen, J. H.; Morse, K. E.; Brunskill, E. P.; Fries, J. A.; and Shah, N. H. 2023. MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. Version Number: 2.

Gertz, R. J.; Dratsch, T.; Bunck, A. C.; Lennartz, S.; Iuga, A.-I.; Hellmich, M. G.; Persigehl, T.; Pennig, L.; Gietzen, C. H.; Fervers, P.; Maintz, D.; Hahnfeldt, R.; and Kottlors, J. 2024. Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy. *Radiology*, 311(1): e232714.

Hamamci, I. E.; Er, S.; and Menze, B. 2024. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging. ArXiv:2403.06801 [cs, eess].

Hou, X.; Liu, Z.; Li, X.; Li, X.; Sang, S.; and Zhang, Y. 2023. MKCL: Medical Knowledge with Contrastive Learning model for radiology report generation. *Journal of Biomedical Informatics*, 146: 104496.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; Seekins, J.; Mong, D. A.; Halabi, S. S.; Sandberg, J. K.; Jones, R.; Larson, D. B.; Langlotz, C. P.; Patel, B. N.; Lungren, M. P.; and Ng, A. Y. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Version Number: 1.

Jain, S.; Agrawal, A.; Saporta, A.; Truong, S. Q.; Duong, D. N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M. P.; Ng, A. Y.; Langlotz, C. P.; and Rajpurkar, P. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. Version Number: 3.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. ArXiv:2310.06825 [cs].

Johnson, A. E. W.; Pollard, T.; Mark, R.; Berkowitz, S.; and Horng, S. 2019a. The MIMIC-CXR Database.

Johnson, A. E. W.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019b. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1): 317.

Lee, S.; Kim, W. J.; Chang, J.; and Ye, J. C. 2023. LLM-CXR: Instruction-Finetuned LLM for CXR Image Understanding and Generation. Version Number: 5.

Lee, S.; Youn, J.; Kim, H.; Kim, M.; and Yoon, S. H. 2024. CXR-LLAVA: a multimodal large language model for interpreting chest X-ray images. ArXiv:2310.18341 [cs].

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2021. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. Version Number: 2.

Liu, Z.; Li, Y.; Shu, P.; Zhong, A.; Yang, L.; Ju, C.; Wu, Z.; Ma, C.; Luo, J.; Chen, C.; Kim, S.; Hu, J.; Dai, H.; Zhao, L.; Zhu, D.; Liu, J.; Liu, W.; Shen, D.; Liu, T.; Li, Q.; and Li, X. 2023. Radiology-Llama2: Best-in-Class Large Language Model for Radiology. ArXiv:2309.06419 [cs].

Ma, C.; Wu, Z.; Wang, J.; Xu, S.; Wei, Y.; Liu, Z.; Zeng, F.; Jiang, X.; Guo, L.; Cai, X.; Zhang, S.; Zhang, T.; Zhu, D.; Shen, D.; Liu, T.; and Li, X. 2024. An Iterative Optimizing Framework for Radiology Report Summarization With ChatGPT. *IEEE Transactions on Artificial Intelligence*, 5(8): 4163–4175.

Mac Namee, B.; Cunningham, P.; Byrne, S.; and Corrigan, O. 2002. The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, 24(1): 51–70.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6): 1–35.

Monshi, M. M. A.; Poon, J.; and Chung, V. 2020. Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, 106: 101878.

Nakaura, T.; Yoshida, N.; Kobayashi, N.; Shiraishi, K.; Nagayama, Y.; Uetani, H.; Kidoh, M.; Hokamura, M.; Funama, Y.; and Hirai, T. 2024. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Japanese Journal of Radiology*, 42(2): 190–200.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Kaiser, ; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Kondraciuk, ; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; Peres, F. d. A. B.; Petrov, M.; Pinto, H. P. d. O.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and Zoph, B. 2023. GPT-4 Technical Report. Version Number: 6.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Parres, D.; Albiol, A.; and Paredes, R. 2024. Improving Radiology Report Generation Quality and Diversity through Reinforcement Learning and Text Augmentation. *Bioengineering*, 11(4): 351.

Pellegrini, C.; Özsoy, E.; Busam, B.; Navab, N.; and Keicher, M. 2023. RaDialog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance. Version Number: 1.

Ranjit, M.; Ganapathy, G.; Srivastav, S.; Ganu, T.; and Oruganti, S. 2024. RAD-PHI2: Instruction Tuning PHI-2 for Radiology. Version Number: 1.

Sadvilkar, N.; and Neumann, M. 2020. PySBD: Pragmatic Sentence Boundary Disambiguation. Version Number: 1.

Tanno, R.; Barrett, D. G. T.; Sellergren, A.; Ghaisas, S.; Dathathri, S.; See, A.; Welbl, J.; Singhal, K.; Azizi, S.; Tu, T.; Schaekermann, M.; May, R.; Lee, R.; Man, S.; Ahmed, Z.; Mahdavi, S.; Matias, Y.; Barral, J.; Eslami, A.; Belgrave, D.; Natarajan, V.; Shetty, S.; Kohli, P.; Huang, P.-S.; Karthikesalingam, A.; and Ktena, I. 2023. Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation. Version Number: 3.

Thawkar, O.; Shaker, A.; Mullappilly, S. S.; Cholakkal, H.; Anwer, R. M.; Khan, S.; Laaksonen, J.; and Khan, F. S. 2023. XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models. Version Number: 1.

Tian, K.; Hartung, S. J.; Li, A. A.; Jeong, J.; Behzadi, F.; Calle-Toro, J.; Adithan, S.; Pohlen, M.; Osayande, D.; and Rajpurkar, P. 2023. ReFiSco: Report Fix and Score Dataset for Radiology Report Generation.

Van Ooijen, P. M. A., ed. 2021. *Basic Knowledge of Medical Imaging Informatics: Undergraduate Level and Level I*. Imaging Informatics for Healthcare Professionals. Cham: Springer International Publishing. ISBN 978-3-030-71884-8 978-3-030-71885-5.

Vsevolodovna, R. M. 2024. Medical-Llama3-8B-16bit: Fine-Tuned Llama3 for Medical Q&A.

Wu, X.; Li, J.; Wang, J.; and Qian, Q. 2023. Multimodal contrastive learning for radiology report generation. *Journal of Ambient Intelligence and Humanized Computing*, 14(8): 11185–11194.

Yan, A.; McAuley, J.; Lu, X.; Du, J.; Chang, E. Y.; Gentili, A.; and Hsu, C.-N. 2022. RadBERT: Adapting Transformer-based Language Models to Radiology. *Radiology: Artificial Intelligence*, 4(4): e210258.

Yu, F.; Endo, M.; Krishnan, R.; Pan, I.; Tsai, A.; Reis, E. P.; Fonseca, E. K. U. N.; Ho Lee, H. M.; Abad, Z. S. H.; Ng, A. Y.; Langlotz, C. P.; Venugopal, V. K.; and Rajpurkar, P. 2022. Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, X.; Meng, Z.; Lever, J.; and Ho, E. S. 2024. Gla-AI4BioMed at RRG24: Visual Instruction-tuned Adaptation for Radiology Report Generation. In Demner-Fushman, D.; Ananiadou, S.; Miwa, M.; Roberts, K.; and Tsujii, J., eds., *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 624–634. Bangkok, Thailand: Association for Computational Linguistics.

Zhang, Y.; Wang, X.; Xu, Z.; Yu, Q.; Yuille, A.; and Xu, D. 2020. When Radiology Report Generation Meets Knowledge Graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07): 12910–12917.