ReXErr: Synthesizing Clinically Meaningful Errors in Diagnostic Radiology Reports

Vishwanatha M. Rao[†1], Serena Zhang[†1], Julian N. Acosta[1], Subathra Adithan[2], Pranav Rajpurkar[1]

[1]*Department of Biomedical Informatics, Harvard Medical School Boston, MA 02115, USA*
[2]*Department of Radiodiagnosis, Jawaharlal Institute of Postgraduate Medical Education and Research, India*
E-mail: : vishwanatha.rao@pennmedicine.upenn.edu, serena2z@stanford.edu, julian_acosta@hms.harvard.edu, subathra.a@jipmer.edu.in

Accurately interpreting medical images and writing radiology reports is a critical but challenging task in healthcare. Both human-written and AI-generated reports can contain errors, ranging from clinical inaccuracies to linguistic mistakes. To address this, we introduce ReXErr, a methodology that leverages Large Language Models to generate representative errors within chest X-ray reports. Working with board-certified radiologists, we developed error categories that capture common mistakes in both human and AI-generated reports. Our approach uses a novel sampling scheme to inject diverse errors while maintaining clinical plausibility. ReXErr demonstrates consistency across error categories and produces errors that closely mimic those found in real-world scenarios. This method has the potential to aid in the development and evaluation of report correction algorithms, potentially enhancing the quality and reliability of radiology reporting.

Keywords: Radiology Report Generation; Chest X-Rays; LLMs; Chat-GPT; Error Injection; Synthetic Data.

1. Introduction

Radiology reports provide crucial information for clinical decision-making and patient outcomes.[1] However, creating radiology reports is an intensive process, and requires a trained specialist to analyze medical images and write in-depth medical reports.[2,3] In human-written reports, errors can arise due to various factors such as fatigue, high case volumes or complexity. These errors may include misinterpretation of imaging findings, incomplete documentation of relevant clinical information, and inconsistencies in terminology and language usage. In addition to such inaccuracies, the subjective nature of radiological interpretation leaves room for errors, which may go unnoticed until they impact patient care.[4,5]

Recently, there has been a significant push towards automating the creation of these reports using deep learning. While current approaches to generating radiology reports have, in some cases, succeeded in creating complete and clinically relevant reports,[6–9] automated report generation presents its own set of challenges stemming from inherent biases within algorithms, model constraints, and limitations in the data used. Errors can range from references to non-

[†]Authors contributed equally to this work.

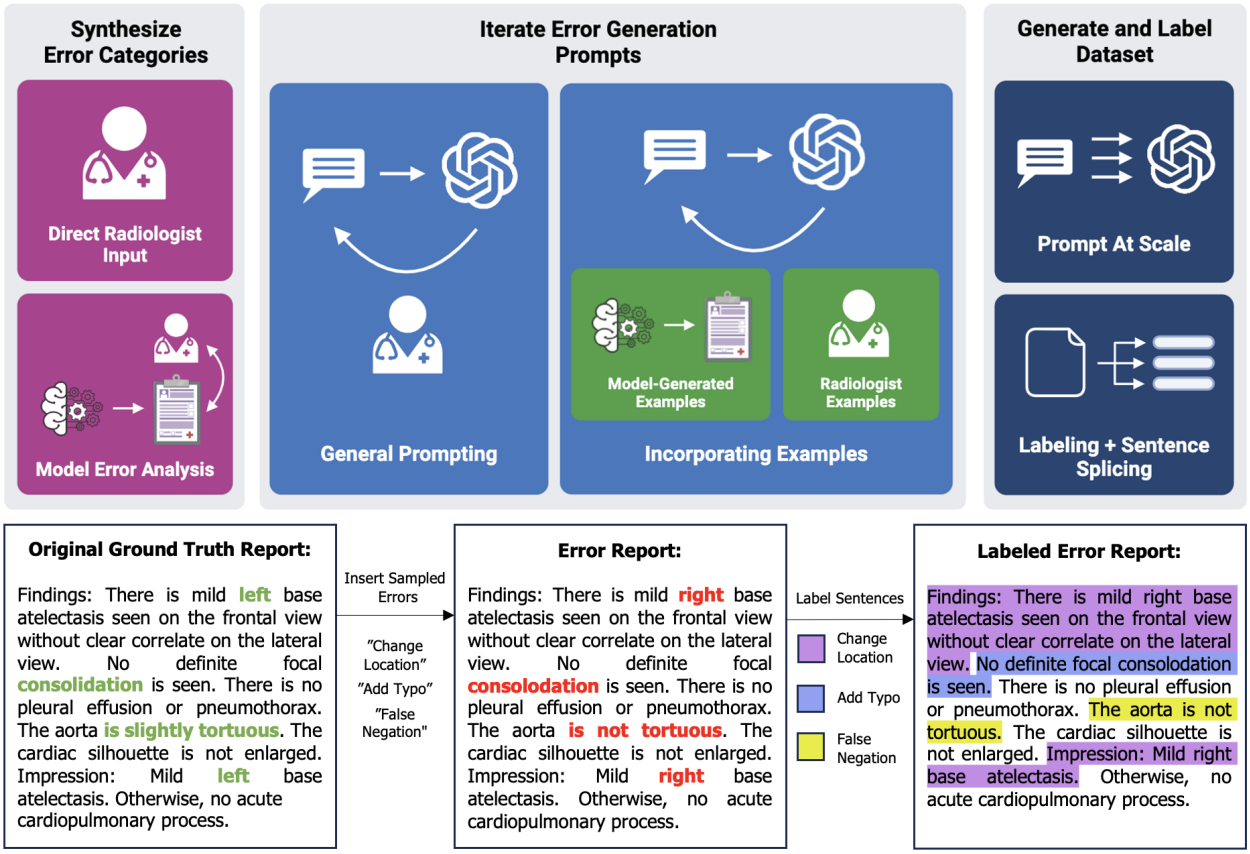

Fig. 1. Summary of ReXErr error generation pipeline. The bottom panel provides an example of applying ReXErr to a sample radiology report.

existing priors, which are easier to detect, to false predictions or omissions, which are much more problematic clinically and often go unnoticed.[10,11] The prevalence of errors, both in radiologist-written as well as AI-based reports, leaves a great need for more comprehensive tools that can screen for and correct them. Throughout this paper, we present the Chest X-Ray Report Error (ReXErr) method that can generate errors at a report and sentence level. ReXErr offers a novel pipeline to synthesize plausible errors that capture the breadth and diversity of errors made by humans and models and can thus be used to generate data to train and adapt error correction algorithms. Figure 1 outlines an overview of the error generation process.

2. Related Work

2.1. *Error Classification in Radiology Reports*

The 12-category framework developed by Kim and Mansfield, based on an evaluation of 1,269 errors, offers a foundation for understanding and classifying errors in human-generated reports and is the most frequently used for human-error analysis.[4,12] Most of the errors in this classification system fall under two types: missed findings (under-reading, satisfaction of search, etc.) and interpretation errors (finding attributed to wrong cause/clinical entity due to faulty

Table 1. Summary of the errors incorporated within the ReXErr pipeline.

Error Type	Error Category	Specific Errors
AI Generated Report Errors	Content Addition	Add Medical Device False Prediction False Negation
	Linguistic Quality	Add Repetitions Add Contradictions
	Context-Dependent	Change Name of Device Change Position of Device Change Severity Change Location Change Measurement
Human Errors	Content Addition and Context-Dependent	Human error - similar to above
	Linguistic Quality	Change to Homophone Add Typo

reasoning, lack of knowledge, etc.), with each of the 12 classifications focusing on the cause for such an error to occur..[5]

Errors from report generation models differ, with more specific issues including hallucinated references to prior studies, and have their own categorization framework. One example is the framework developed by Yu et al. to analyze common errors in model-generated radiology reports, aiming to create metrics that account for these errors and improve alignment with clinician feedback.[13] Their framework includes six categories: "False prediction of finding", "Omission of finding", "Incorrect location/position of finding", "Incorrect severity of finding", "Mention of comparison that is not present in the reference impression", and "Omission of comparison describing a change from a previous study." They develop a dataset, ReXVal, which contains annotations on clinically significant and insignificant errors under their six category framework for AI generated radiology reports with respect to ground-truth reports. Another dataset, Refisco, was created to categorize the errors commonly made in retrieval-based report generation models by their severity level and then correct each error using either deletion, substitution, or insertion of a line.[14] Both datasets provide different error categorization frameworks specific for AI-generated reports, offering error-report ground truth pairs with clinician annotations. However, their limited size (200 and 60 reports, respectively) underscores the need for more extensive datasets that contain error reports and ground truth pairs.

2.2. *Synthetic Data Generation for Radiology Reports*

Synthetic data generation is emerging as a valuable tool in radiology reporting research, addressing challenges of data scarcity. Recent studies have demonstrated its potential in various applications. Zhao et al.[15] generated modified reports with revision instructions, aiding in the training of instruction-based report revision systems. Hyland et al.[16] used GPT to paraphrase MIMIC dataset reports, expanding their training set for report generation models. Others have leveraged large language models to selectively modify radiology reports, addressing various clinical and research needs such as removing prior medical history references and standardizing report structures.[17,18] Most similarly, Asiimwe et al.[19] created a synthetic error

report set for developing an error detection and correction model. They intentionally introduced errors into radiology reports, focusing on four out of the six error categories defined by Yu et al.'s framework. This process resulted in the creation of 120,000 pairs of error-containing and error-free reports, which serve as training data for their model. These advancements highlight the growing importance of synthetic data in improving radiology reporting systems by enabling large-scale, precise data generation. However, it also reveals the need for a more comprehensive dataset that captures a wider range of diverse and complex errors.

Building upon these advancements, our study utilizes synthetic data generation to create a large-scale dataset that incorporates a broader range of errors. We expand on the framework established by Yu et al., addressing all six major categories of AI-generated errors, while also introducing additional subtypes such as device-related errors. Furthermore, our dataset addresses linguistic quality issues in both human- and AI-generated reports. This comprehensive approach allows us to create a more diverse and robust error dataset, providing a valuable resource for developing and evaluating advanced radiology reporting systems.

2.3. *Applications in Error Detection and Report Correction*

Our comprehensive error dataset has significant potential applications in advancing both error detection and report correction in radiology. In error detection, research has progressed from simple matching techniques for specific issues like laterality errors to more sophisticated methods using LSTM and BERT-based models.[20–22] Recent studies have even shown GPT-4's capability to identify common error categories (omission, insertion, spelling, and side confusion).[23] Our dataset, encompassing a wider range of error types, could further enhance these detection models.

In report correction, efforts have focused on addressing specific types of hallucinations in AI-generated reports, such as false references to non-existent prior scans.[24,25] The emerging task of report revision aims to refine existing reports through instruction prompts, as demonstrated in recent multi-functional foundation models.[6,15] Such an error-rich dataset could serve as a valuable resource for training and evaluating these correction and revision systems, potentially improving their ability to handle a diverse array of error types.

Furthermore, our dataset could be utilized as negative examples in reinforcement learning algorithms to enhance AI model performance, or to validate automatic evaluation metrics like RadCliQ and FineRadScore.[26] This broad applicability underscores the potential impact of our error injection method and resulting dataset in advancing the accuracy and reliability of radiological reporting systems.

3. Methods

We created a streamlined pipeline to inject errors into radiology reports, which can be used downstream to generate large datasets and train models for the identification and revision of incorrect radiology reports. We demonstrate error generation with the ReXErr pipeline using reports from the MIMIC-CXR train, dev, and test sets.[27] This pipeline supports two main tasks: report correction and sentence-level entailment. For both tasks, sentences are classified into three categories: correct (0), error (1), and neutral (2). Neutral sentences reference past

Table 2. Baseline prompting description for each error category.

Error	Baseline Instruction / Description
Add Medical Device	Add sentences that could be part of a radiology report regarding the presence of one or more devices such as these: pacemaker, central venous line, NG tube, ET tube, ICD.
Change Name of Device	If there is a medical device present in the report, change the name of the medical instrument to a different name that is clinically plausible.
Change Position of Device	If there is a medical device location present in the report, change the position of the medical instrument to a different position that is clinically plausible.
Change Severity	Change the severity of a finding in the report in a manner that makes clinical sense (e.g., change 'mild' to 'moderate').
Change Location	Change the location or anatomy of a finding in the report in a manner that is still clinically accurate (e.g., change 'right' to 'left' or 'lateral' to 'medial'; always modifying a sentence).
False Prediction	Add a finding that is not present in the report (either adding a sentence or modifying a sentence to insert).
False Negation	Change a particular finding from the report from present to absent by changing a sentence to indicate absence of the positive finding.
Change Measurement	If there is a measurement for a device/finding present, change the units of measurement (e.g., change 'cm' to 'mm') or change the value of the measurement to a different but still reasonable value (e.g. change '4.9 cm' to '5.8 cm').
Add Opposite Sentence	Add/alter a statement that is the opposite of another statement earlier in the same report.
Add Repetitions	Add repetitions of sentences present within the report.
Change to Homophone	Change a word in the report to a homophone of that word.
Add Typo	Add a typographical error in the report.

reports, findings, or scans and are categorized separately, as algorithms would not be able to determine their accuracy without additional context.

Report correction: Our pipeline generates paired ground truth and error reports, with each error report containing three errors sampled from 12 possible error categories. Sampling three errors per report provides a balanced representation of diverse error types while maintaining a degree of similarity to the original report and has been used prior in the literature.[23] We also separately specify the three error categories used in generating each report.

Sentence-level entailment: We provide a separate pipeline to create a sentence-level error categorization by splicing pairs of sentences from ground truth and error reports. Each pair includes the original sentence and its error version, their label (see categories below), type of error injected (error class) and sequence in the original report (index). Maintaining the sequential detail can help sentence-level entailment models developed upon data generated through ReXErr use contextual information from previous sentences to identify errors such as repetitions and contradictions.

3.1. *Error Categories*

Three board-certified radiologists were consulted in synthesizing the final list of errors included within this generation protocol. The errors fall under two broad categories: AI generated report errors and human errors. We further identify three sub-categories of errors: content addition, context-dependent, and linguistic quality errors. Each of the 12 final error categories fall under one of these subcategories and one of the two broader categories. The particular errors were determined in careful collaboration with radiologists; specifically, we used a set of radiologist-annotated reports generated from a current state of the art model to determine the most salient automated generation errors, and consulted radiologists directly to gain a sense for

human errors.[6] We incorporate all six major categories of AI-generated errors established by Yu et al.[13] The content addition and context-dependent errors observed in human-generated reports closely parallel those found in AI-generated reports. Additionally, we introduce a set of errors that address linguistic quality issues present in both human- and AI-generated reports, thereby creating a comprehensive error classification system. Table 1 contains a summary of the errors implemented.

3.2. *Data Synthesis*

After extensive iteration and feedback from clinical experts, we developed a comprehensive pipeline for introducing plausible errors into radiology reports using GPT-4o.[28] GPT-4o was chosen given its high performance relative to price. We define "plausible" errors as those that either a human or an AI model could realistically make. The pipeline employs a sophisticated sampling strategy to inject errors across all three categories within each report. Context-dependent errors are only introduced when the associated context is present, as determined by regex-based labeling that searches for specific keywords in each report. For instance, errors related to changing the location and type of medical devices are only injected if a device is mentioned in the report. The regex keywords for each category are constructed through a combination of clinician input and analysis of radiology report terms used in the dataset. Our approach balances the need for diverse and plausible errors while maintaining the overall structure and believability of the reports. The problem formulation for the injection of errors across all three categories is represented in Equation 1, where E_C, E_A, and E_L represent context-dependent, content addition, and linguistic quality errors respectively. T refers to the tags present, where $T \in \{\text{“device”}, \text{“measurement”}, \text{“location”}, \text{“severity”}\}$.

$$P(E_c, E_A, E_L \mid T) = P(E_c \mid T) \times P(E_A) \times P(E_L) \tag{1}$$

The probability of selecting both the content addition (E_A) and linguistic quality (E_L) errors are shown below in Equation 2. A and L both represent the number of individual errors present within the content addition and linguistic quality error categories respectively across both the AI and human groups. In our case, A would be 3 and L would be 4, where L includes the linguistic quality errors in both the AI and human error categories.

$$P(E_A) = \frac{1}{|A|} \qquad\qquad P(E_L) = \frac{1}{|L|} \tag{2}$$

The probability of selecting a context-dependent error given a particular tag is given by Equation 3 below, where the error for the context-dependent error category is sampled across the other categories if no relevant context is present. In the case where multiple tags are provided, the probability of selecting a particular context-dependent error given a tag depends on the normalized weight assigned to the tag ($w'(t_i)$) as well as the total number of context-dependent errors associated with each tag ($E(t)$).

$$P(E_c \mid T = t_i) = \begin{cases} \frac{w'(t_i)}{\sum_{t \in T} w'(t) \times E(t)} & \text{if } T \neq \emptyset \\ \frac{1}{|A|+|L|} & \text{if } T = \emptyset \end{cases} \tag{3}$$

Table 3. Examples of ground truth and error report generated through the ReXErr pipeline.

Ground Truth	Error Report	Errors Injected
Findings: Findings: The patient is status post median sternotomy and CABG. The heart size is top normal. The mediastinal and hilar contours are unremarkable. Bilateral calcified pleural plaques are seen diffusely which limits assessment of the underlying pulmonary parenchyma. No focal consolidation, pleural effusion or pneumothorax is clearly demonstrated. There are no acute osseous abnormalities. **Impression:** Bilateral calcified pleural plaques indicative of prior asbestos exposure. No definite acute cardiopulmonary abnormality otherwise noted.	**Findings:** The patient is status post median sternotomy and CABG. The heart size is enlarged. The mediastinal and hilar contours are unremarkable. Right calcified pleural plaques are seen diffusely which limits assessment of the underlying pulmonary parenchyma. No focal consolidation, pleural effusion or pneumothorax is clearly demonstrated. There is a suspected left clavicle fracture. **Impression:** Right calcified pleural plaques indicative of prior asbestos exposure. There is a moderate left pleural effusion. No definite acute cardiopulmonary abnormality otherwise noted.	'change location', 'false prediction', 'add contradiction'
Findings: Single frontal view of the chest provided. There is no focal consolidation, effusion, or pneumothorax. The cardiomediastinal silhouette is normal. Again seen are multiple clips projecting over the left breast and remote left-sided rib fractures. No free air below the right hemidiaphragm is seen. **Impression:** No acute intrathoracic process.	**Findings:** Single frontal view of the chest provided. There is know focal consolidation, effusion, or pneumothorax. The cardiomediastinal silhouette is normal. Again seen are multiple clips projecting over the left breast and remote left-sided rib fractures. There is an ET tube present in the trachea. No free air below the right hemidiaphragm is seen. No free air below the right hemidiaphragm is seen. **Impression:** No acute intrathoracic process.	'add repetitions', 'add medical devices', 'change to homophone'
Findings: There is mild-to-moderate cardiomegaly, not significantly changed compared with prior study. There is no pneumothorax. A newly placed endotracheal tube ends 4.3 cm above the carina. An NG tube is seen ending in the stomach with its tip and side ports beyond the margin of imaging. **Impression:** 1. Severe acute pulmonary edema. 2. Endotracheal tube ending 4.3 cm above the carina.	**Findings:** There is mild-to-moderate cardiomegaly, not significantly changed compared with prior study. There is no pneumothorax. A newly placed endotracheal tube ends 4.3 mm above the carina. An NG tube is seen ending in the stomach with its tip and side ports beyond the margin of imaging. **Impression:** 1. No pulmonary edema. 2. Endotrakheal tube ending 4.3 cm above the carina.	'change measurement', 'false negation', 'add typo'

The weights assigned to each tag $w(t_i)$ was calculated based on the frequency of each tag present within the reports through the equations shown below. Each weight is equal to the inverse of the prevalence of its respective tag. The weights are then normalized to $w'(t)$.

$$w(t) = \frac{1}{f(t)} \qquad W = \sum_{t \in T} w(t) \qquad w'(t) = \frac{w(t)}{W} \tag{4}$$

Based on this sampling strategy, GPT-4o was then used to inject the errors. Table 2 summarizes the baseline instructions given for each error type. Appendix A contains the complete long-form prompt used to prompt GPT, whereas Appendix B contains the particular prompts for each error category, including the examples for the relevant errors that use them.

3.3. *Sentence Level Error Generation Process*

Once the error reports were generated, each report was split into individual sentences and mapped based on sentence similarity to their corresponding ground truth sentence. We used Llama 3.1 to identify the error type in each sentence and screen for prior reports.[29] Llama 3.1 was chosen instead of GPT-4o for error relabeling due to its sufficient accuracy and greater

Table 4. Examples of ground truth and error sentences generated through the ReXErr sentence splicing and labeling pipeline.

Original Sentence	Error Sentence	Label	Error Class	Index
Findings: Comparison is made to previous study from ___.	Findings: Comparison is made to previous study from ___.	2	Not Applicable	0
There is a right-sided PICC line with distal lead tip at the cavoatrial junction.	There is a right-sided PICC line with distal lead tip at the mid SVC.	1	Change Position of Device	1
There has been removal of the right-sided chest tube.	There has been removal of the right-sided chest tube.	0	Not Applicable	2
There remains a curvilinear tubular device projecting over the mediastinum.	There remains a curvilinear tubular device projecting over the mediastinum.	0	Not Applicable	3
This has been seen on multiple images.	This has been seen on muitiple images.	1	Add Typo	4
There is persistent opacity at the left mid lung field and left-sided pleural effusion which is stable.	There is persistent opacity at the left mid lung field and left-sided pleural effusion which stable.	1	Add Typo	5
There is no pulmonary edema.	There is no pulmonary edema.	0	Not Applicable	6
The right lung is relatively clear.	The right lung is relatively clear.	0	Not Applicable	7
	The patient has had placement of an endotracheal tube.	1	Add Medical Device	8

cost-efficiency. The model was prompted to produce a Python dictionary with two keys: "label" and "error class." The "label" key indicated whether the sentence was correct (0), erroneous (1), or neutral (2), while the "error class" key specified the error type, if applicable. The "Add Repetition" error category was excluded, as repetition is only relevant at the report level, and "Add Opposite Sentence" was reclassified as "False Prediction." In cases where a new sentence was added, the original sentence field was left blank, and for omitted sentences, the error report sentence was left blank. Through this methodology, we are able to provide side-by-side comparisons between individual sentences and their associated error sentences. The order of sentences within the original report is maintained, including the position of particular added or omitted error sentences. The sentences were manually reviewed to ensure the accuracy of the sentence splicing.

3.4. *Validating Error Injection Pipeline*

In order to validate the quality and efficacy of our error injection pipeline, we analyze the projected frequency of every single error category injected across the MIMIC train, dev, and test subsets. Furthermore, a clinician reviewed 100 paired original and error-injected reports to determine the fraction of error reports which are plausible AI-generated or human-written reports. This was done to determine whether the synthesized error reports contain language atypical to radiology reports or very obvious modifications and statements that are not medically plausible which might limit the utility of the synthetic error reports.

Table 5. Distribution of errors inserted across the MIMIC train, dev, and test sets using the ReXErr methodology.

Error Category	Train (%)	Dev (%)	Test (%)
Add Medical Device	33.33	33.32	33.33
Change Name of Device	13.64	13.47	18.91
Change Position of Device	13.64	13.47	18.91
Change Severity	28.71	29.88	30.18
Change Location	38.07	37.07	23.26
False Prediction	33.33	33.32	33.33
False Negation	33.33	33.32	33.33
Change Measurement	5.93	6.10	7.01
Add Opposite Sentence	25.00	24.97	24.99
Add Repetitions	25.00	24.97	24.99
Change to Homophone	25.00	24.97	24.99
Add Typo	25.00	24.97	24.99

4. Results

4.1. *Strengths and Limitations of ReXErr*

The ReXErr pipeline was found to proficiently generate errors across all of the error categories listed for the majority of radiology report inputs. It is able to create multiple types of errors in the same report, with variation within each error subtype as well. These errors closely mimic those found in real-world report generation scenarios. Table 3 includes three examples of error reports generated using our report-level error injection pipeline, while Table 4 presents several examples of the sentence-level error generation process, along with the error labeling scheme. Despite ReXErr's ability to generate errors within the findings and impressions sections, there are still limitations in its ability to maintain consistency in the error injections across both sections. For example, while the first example in Table 3 is handled well, others such as the measurement change in the third example show discrepancies.

4.2. *Consistency Across Error Types*

ReXErr also demonstrates reasonable consistency in distribution of errors inserted across the MIMIC train, dev, and test sets. Certain errors, including "change measurement", "change name of device", and "change position of device" are injected less frequently in the dataset due to their reduced prevalence in the original reports. While the weighting mechanism used during sampling helped augment this discrepancy, this quantitative analysis highlights key areas for targeted improvements in developing more robust error injection and correction methods. Table 5 outlines the frequencies of each error type across the train, dev, and test sets, with each value representing the percentage of reports within the given set containing that specific error. Notably, these percentages are relatively consistent across the three different splits.

4.3. *Plausibility of Errors*

Lastly, ReXErr was found to predominantly inject plausible errors within reports. Plausible errors are mistakes that could reasonably occur in real-world radiology practice, while implausible errors involve anatomical impossibilities or fundamental misunderstandings of medical

principles that would otherwise never be made. Examples of implausible errors include substituting one medical device for another inappropriately (replacing "pacemaker" with "ET tube" in "A pacemaker is present with leads in the right ventricle"), or attributing findings to anatomical structures beyond the chest x-ray image. In the sample of 100 ground truth and error-injected reports reviewed by a board-certified clinician, 83 of the modified reports were found to be plausible, while only 17 contained errors that were implausible in AI-generated or human reports.

5. Discussion

Throughout this paper, we present ReXErr, a new pipeline designed to generate clinically relevant and plausible errors. Despite ReXErr's demonstrated capability to inject diverse errors, certain limitations that may hinder its use. Firstly, the applicability downstream models trained on data generated using ReXErr depends heavily on the quality and clinical relevance of the errors generated. While the majority of ReXErr-generated errors were plausible, we found 17 out of 100 augmented reports to contain implausible errors, meaning that the prompting methodology could be further improved before implementation on a larger scale.

Another potential limitation is the error sampling approach. ReXErr's sampling strategy does not account for nested compound errors, where errors can belong to more than one category, or cases where a sentence can contain multiple errors. Depending on how prevalent such errors are in actual human or AI generated reports, the absence of these errors could negatively impact ReXErr's downstream utility. Furthermore, downstream models may struggle to discriminate errors made in AI-generated text as ReXErr only adds errors to human-generated reports. Even though the errors themselves are sampled amongst errors commonly made by AI models, their addition to human generated text may not make them as representative as errors that were added to AI-generated text.

Lastly, future pipelines could benefit from more extensive downstream model testing using preliminary data generated. For example, while GPT-4o was chosen for its high performance and affordability, other open-source LLMs may yield more robust errors, and downstream testing would help elucidate which models can generate the most effective synthetic errors. Moreover, downstream testing would help determine whether the changes made to reports are significant enough for models to discern, as in some cases, the errors added are very minor.

6. Conclusion

Synthesizing accurate radiology reports is both difficult and time consuming, even for medical professionals. While automated AI generation approaches are promising in alleviating this workload and more efficiently generating comprehensive reports, they are liable to frequent errors across report content, linguistics, and consistency. Throughout this paper, we present the novel ReXErr method for generating annotated errors on both a report and sentence level. Developed with radiologists, ReXErr captures common AI and human errors in a representative and plausible manner, therefore offering a promising avenue for the development of report screening and correction algorithms as well as improving the accuracy of existing report generation approaches.

Acknowledgments

We would like to thank Dr. John Farner and Dr. Rohit Reddy for their valuable clinical input into the error categories and prompts chosen.

References

1. M. J. Côté and M. A. Smith, Forecasting the demand for radiology services, *Health Systems* **7**, 79 (2018), ISBN: 2047-6965 Publisher: Taylor & Francis.
2. B. I. Reiner, N. Knight and E. L. Siegel, Radiology reporting, past, present, and future: the radiologist's perspective, *Journal of the American College of Radiology* **4**, 313 (2007), ISBN: 1546-1440 Publisher: Elsevier.
3. A. Al Yassin, M. S. Sadaghiani, S. Mohan, R. N. Bryan and I. Nasrallah, It is About" Time": Academic Neuroradiologist Time Distribution for Interpreting Brain MRIs, *Academic Radiology* **25**, 1521 (2018), ISBN: 1076-6332 Publisher: Elsevier.
4. M. A. Bruno, E. A. Walker and H. H. Abujudeh, Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction, *Radiographics* **35**, 1668 (2015), ISBN: 0271-5333 Publisher: Radiological Society of North America.
5. A. P. Brady, Error and discrepancy in radiology: inevitable or avoidable?, *Insights into Imaging* **8**, 171 (February 2017).
6. H.-Y. Zhou, S. Adithan, J. N. Acosta, E. J. Topol and P. Rajpurkar, A Generalist Learner for Multifaceted Medical Image Interpretation (May 2024), arXiv:2405.07988 [cs].
7. T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno and I. Ktena, Towards generalist biomedical ai, *NEJM AI* **1**, p. AIoa2300138 (2024), ISBN: 2836-9386 Publisher: Massachusetts Medical Society.
8. C. Wu, X. Zhang, Y. Zhang, Y. Wang and W. Xie, Towards generalist foundation model for radiology, *arXiv preprint arXiv:2308.02463* (2023).
9. T. Tanida, P. Müller, G. Kaissis and D. Rueckert, Interactive and Explainable Region-guided Radiology Report Generation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , 7433 (2023).
10. P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. Andia, C. Tejos, C. Prieto and D. Capurro, A survey on deep learning and explainability for automatic report generation from medical images, *ACM Computing Surveys (CSUR)* **54**, 1 (2022), ISBN: 0360-0300 Publisher: ACM New York, NY.
11. P. Sloan, P. Clatworthy, E. Simpson and M. Mirmehdi, Automated Radiology Report Generation: A Review of Recent Advances, *IEEE Reviews in Biomedical Engineering* (2024), Publisher: IEEE.
12. Y. W. Kim and L. T. Mansfield, Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors, *American journal of roentgenology* **202**, 465 (2014).
13. F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. K. U. N. Fonseca, H. M. H. Lee, Z. S. H. Abad and A. Y. Ng, Evaluating progress in automatic chest x-ray radiology report generation, *Patterns* **4** (2023), ISBN: 2666-3899 Publisher: Elsevier.
14. K. Tian, S. J. Hartung, A. A. Li, J. Jeong, F. Behzadi, J. Calle-Toro, S. Adithan, M. Pohlen, D. Osayande and P. Rajpurkar, ReFiSco: Report Fix and Score Dataset for Radiology Report Generation, *PhysioNet* (2023).
15. B. N. Zhao, X. JIANG, X. Luo, Y. Yang, B. Li, Z. Wang, J. Alvarez-Valle, M. P. Lungren, D. Li and L. Qiu, Large Multimodal Model for Real-World Radiology Report Generation (September 2023).
16. S. L. Hyland, S. Bannur, K. Bouzid, D. C. Castro, M. Ranjit, A. Schwaighofer, F. Pérez-García, V. Salvatelli, S. Srivastav and A. Thieme, MAIRA-1: A specialised large multimodal model for radiology report generation, *arXiv preprint arXiv:2311.13668* (2023).

17. O. Banerjee, H.-Y. Zhou, S. Adithan, S. Kwak, K. Wu and P. Rajpurkar, Direct preference optimization for suppressing hallucinated prior exams in radiology report generation, *arXiv preprint arXiv:2406.06496* (2024).
18. O. Banerjee, A. Saenz, K. Wu, W. Clements, A. Zia, D. Buensalido, H. Kavnoudias, A. S. Abi-Ghanem, N. E. Ghawi, C. Luna *et al.*, Rexamine-global: A framework for uncovering inconsistencies in radiology report generation metrics, *arXiv preprint arXiv:2408.16208* (2024).
19. A. C. Asiimwe, D. Surís, P. Rajpurkar and C. Vondrick, Image-conditioned autocorrection in medical reporting (2024).
20. Y. H. Lee, J. Yang and J.-S. Suh, Detection and Correction of Laterality Errors in Radiology Reports, *Journal of Digital Imaging* **28**, 412 (August 2015).
21. J. Zech, J. Forde, J. J. Titano, D. Kaji, A. Costa and E. K. Oermann, Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-to-sequence models, *Annals of translational medicine* **7** (2019), Publisher: AME Publications.
22. D. Min, K. Kim, J. H. Lee, Y. Kim and C. M. Park, RRED: a radiology report error detector based on deep learning framework, *Proceedings of the 4th Clinical Natural Language Processing Workshop* , 41 (2022).
23. R. J. Gertz, T. Dratsch, A. C. Bunck, S. Lennartz, A.-I. Iuga, M. G. Hellmich, T. Persigehl, L. Pennig, C. H. Gietzen, P. Fervers, D. Maintz, R. Hahnfeldt and J. Kottlors, Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy, *Radiology* **311**, p. e232714 (April 2024).
24. V. Ramesh, N. A. Chi and P. Rajpurkar, Improving radiology report generation systems by removing hallucinated references to non-existent priors, in *Machine Learning for Health*, (PMLR, 2022).
25. O. Banerjee, H.-Y. Zhou, S. Adithan, S. Kwak, K. Wu and P. Rajpurkar, Direct Preference Optimization for Suppressing Hallucinated Prior Exams in Radiology Report Generation (June 2024), arXiv:2406.06496 [cs].
26. A. Huang, O. Banerjee, K. Wu, E. P. Reis and P. Rajpurkar, FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores (May 2024), arXiv:2405.20613 [cs].
27. A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark and S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, *Scientific data* **6**, p. 317 (2019), ISBN: 2052-4463 Publisher: Nature Publishing Group UK London.
28. O. Team, Hello GPT-4o, *OpenAI* (May 2024).
29. M. A. Team, Introducing Llama 3.1: Our most capable models to date, *Meta AI* .

7. Appendix

Please find the appendix here: `https://drive.google.com/file/d/15dCVF8yh8i6UIOaS_m5-biA28fCSiQjh/view?usp=sharing`