MedAutoCorrect
Image-Conditioned Autocorrection in Medical Reporting

Arnold Caleb Asiimwe   A.ASIIMWE@COLUMBIA.EDU
Columbia University
New York, NY, United States

Dídac Surís   DIDAC.SURIS@COLUMBIA.EDU
Columbia University
New York, NY, United States

Pranav Rajpurkar   PRANAV_RAJPURKAR@HMS.HARVARD.EDU
Harvard Medical School
Cambridge, MA, United States

Carl Vondrick   CVONDRICK@CS.COLUMBIA.EDU
Columbia University
New York, NY, United States

Abstract

In medical reporting, the accuracy of radiological reports, whether generated by humans or machine learning algorithms, is critical. We tackle a new task in this paper: image-conditioned autocorrection of inaccuracies within these reports. Using the MIMIC-CXR dataset, we first intentionally introduce a diverse range of errors into reports. Subsequently, we propose a two-stage framework capable of pinpointing these errors and then making corrections, simulating an *autocorrection* process. This method aims to address the shortcomings of existing automated medical reporting systems, like factual errors and incorrect conclusions, enhancing report reliability in vital healthcare applications. Importantly, our approach could serve as a guardrail, ensuring the accuracy and trustworthiness of automated report generation. Experiments on established datasets and state of the art report generation models validate this method's potential in correcting medical reporting errors.

1. Introduction

Medical reports, particularly in radiology, are cornerstones of healthcare (Brady, 2018). They offer crucial interpretations of medical images that directly affect clinical decision-making and patient care (European Society of Radiology (ESR), 2011). Given their role, ensuring the accuracy and dependability of these reports is vital.

There has been a surge in efforts to automate the generation of medical reports (Nguyen et al., 2021). These automated approaches offer the promise of uniformity and the potential to reduce the heavy workload faced by radiologists. Yet, reports created by both humans and radiology report generation systems are susceptible to errors (Jing et al., 2018). For humans, issues such as fatigue and high case volumes can lead to mistakes (Brady, 2017). Evidence gathered during the plain film era suggested a radiologist error rate of around 3–5% in daily practice (Maskell, 2019). For radiology report generation

systems, inaccuracies can arise from limited data, built-in biases, or model constraints, resulting in errors such as incorrect predictions or omissions of findings, misidentification of their location or severity, inappropriate comparative references, or failure to note changes from previous studies (Miura et al., 2021; Yu et al., 2023b). If integrated into clinical practice, these inaccuracies could have profound consequences for patient care.

In this paper, we propose an approach that uses visual information to detect and auto-correct errors in medical reports. Recently, there have been very successful methods for aligning images and language together (Radford et al., 2019), which we adopt for the purpose of error detection and correction. However, due to the significant distribution shift, models trained on Internet data are not directly applicable to the medical domain. Additionally, it is challenging to collect a large paired training set of erroneous reports and their corrected counterparts. We introduce a procedure to synthetically inject errors into correct medical reports, allowing us to consequently learn to detect and remove them. We develop and compare several conditioning mechanisms to detect errors and make corrections. Our approach, illustrated in Figure 1, can identify and rectify errors whether they originate from machine learning models or human radiologists.
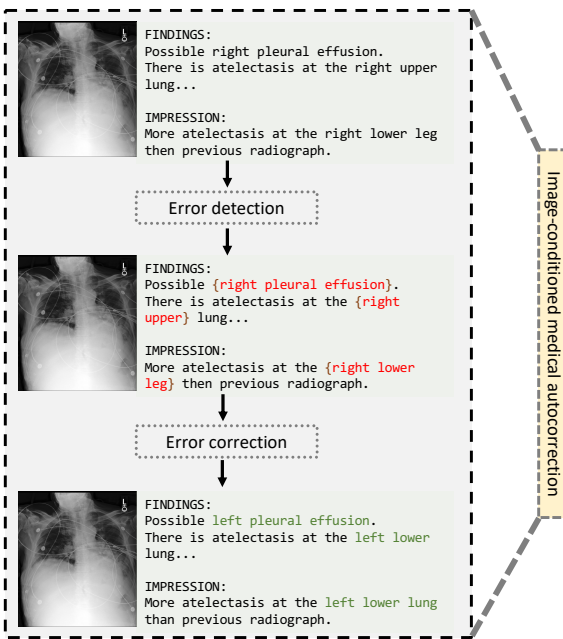

Figure 1: Overview of our `DETECT + CORRECT` error-correction method

Generalizable Insights about Machine Learning in the Context of Healthcare

Our findings show that incorporating autocorrection into radiological report generative models can significantly enhance the natural language generation (NLG) scores (Table 3) of retrieval-based models. This improvement is achieved even on models that are not initially optimized for state-of-the-art (SOTA) performance. By precisely addressing the errors introduced by these generative models, our approach elevates their outputs to SOTA levels. Additionally, we provide qualitative evidence by showcasing specific instances where our model successfully corrects errors in generated reports, acting as a guardrail (Fig 7, Sec 5.3). We hope the community will find our tasks, datasets (subject to appropriate permissions), and models useful towards applications of computer vision in healthcare.

2. Related Work

Autocorrection has become an indispensable tool in human-computer interfaces, such as Microsoft Word or Google Docs. The main goal of these tools has been to uphold both syntactic and semantic integrity within written documents. We recontextualize the concept

of autocorrection to pertain specifically to the accuracy of radiological reports grounded in the images they describe. Our focus is to mitigate factual discrepancies that may arise from the oversight of medical professionals and in the interpretations made by radiology report generative models.

2.1. Radiology Report Generation

Recent efforts in radiology report generation have predominantly focused on enhancing model accuracy through architectural design changes, often neglecting the correction of errors in generated reports (Tanida et al., 2023; Chen et al., 2020b, 2022; Miura et al., 2021; Liu et al., 2019). Initially structured as an image-captioning task (Wang et al., 2018; Yuan et al., 2019), these models have evolved significantly. For instance, region-guided approaches, like those in (Tanida et al., 2023), improve performance by generating reports for specific anatomical regions in chest X-rays. While early models primarily employed CNN-RNN or stacked CNN-LSTM architectures (Chen et al., 2020a; Nguyen et al., 2021; Li et al., 2018; Wang et al., 2018; Yuan et al., 2019), recent studies have pivoted towards Transformer and attention-based encoders for enhanced cross-modal interaction (Vaswani et al., 2017). Others include relation memory units and memory matrices (Chen et al., 2022), systems for contrasting normal and abnormal images (Li et al., 2018), and the integration of medical knowledge graphs (Li et al., 2019).

Radiology report generation methods are largely inspired by image captioning techniques in computer vision (Alfarghaly et al., 2021; Chen et al., 2022, 2020b; Jing et al., 2019; Li et al., 2019, 2018; Shao et al., 2022; Wang et al., 2022; Cornia et al., 2020b; Sanh et al., 2020; Vedantam et al., 2015; Wang et al., 2018). Despite parallels with general image captioning, radiology report generation faces unique challenges: they are typically more detailed and diverse, covering multiple anatomical regions. Additionally, the need to accurately describe specific abnormalities is complicated by data biases towards standard images and reports, leading to the generation of erroneous reports (Chen et al., 2020b). To counter this, some methods have adopted object detection strategies from the dense image captioning domain (Johnson et al., 2015; Li et al., 2019; Shao et al., 2022; Yin et al., 2019), aiming to both localize and describe individual salient regions in images, typically by conditioning a language model on specific region features. These developments represent various approaches undertaken to enhance the accuracy of radiological reports.

2.2. The Gap

Despite the progress in both autocorrection and radiological report generation, a conspicuous gap remains. In radiological report generation, the models are still very error prone (Chen et al., 2020b) and research into the mitigation of the created errors has been minimal, with a focus on mainly architectural changes to enhance report accuracy. Text-based autocorrection systems can only take us as far as correcting grammatical and syntactical discrepancies in text. In this paper, we introduce an image-conditioned autocorrection framework, not only identifying a critical need within the current landscape of radiological report generation but also offering a targeted solution. This method has the potential to assist radiologists by identifying and rectifying inaccuracies within their radiological reports.

3. Methods

Our approach consists of three distinct steps. First, in Sec. 3.1 we propose a procedure to obtain a dataset of medical reports with annotated errors. Second, in Sec. 3.2 we explain how we can use this dataset in order to train a model that detects such errors. Third, in Sec. 3.3 we build on top of the error identification module and propose an error correction module. Finally, we provide more details on training and inference procedures in Secs. 3.4 and 3.5.
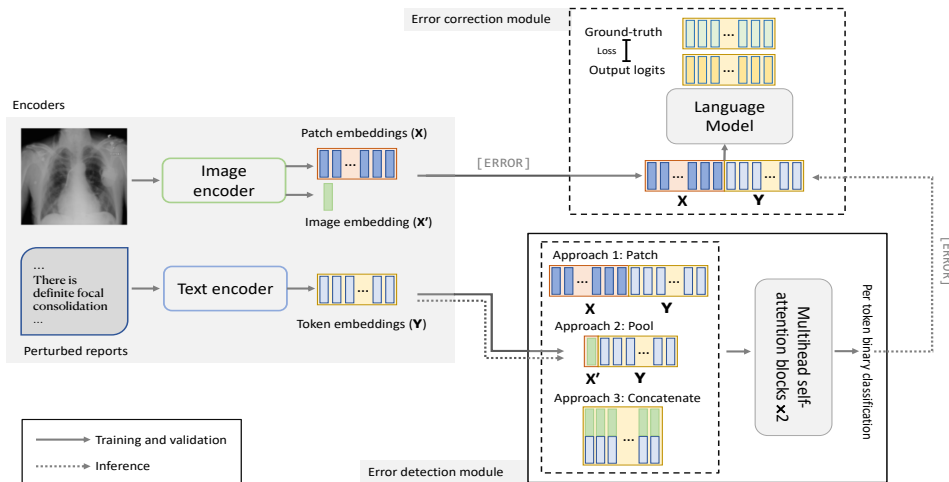

Figure 2: **Overview of the Proposed Framework for Autocorrecting Radiology Reports**. The training phase initiates with separate encoding processes for images and text. The encoded representations are then processed by an error identification module, which utilizes three distinct approaches to detect inaccuracies. Subsequently, a language model is fine-tuned on the image-contextualized reports, where injected errors are represented by [ERROR] tokens. During inference, the model applies the error detection mechanism to localize errors that later are replaced with the masked tokens which are corrected by the error correction mechanism.

3.1. Error injection

To simulate common errors found in radiological reporting, we introduce specific inaccuracies into the reports from the MIMIC-CXR dataset (Johnson et al., 2019), which consists of chest X-ray images paired with free-text radiological reports. These intentionally induced errors are categorized into six major types as identified in (Yu et al., 2023a): 1) False prediction of findings, 2) Incorrect location/position of findings, 3) Incorrect severity of findings, 4) Mention of comparisons not present in the reference impression, 5) Omission of findings, and 6) Omission of comparison describing a change from a previous radiological im-
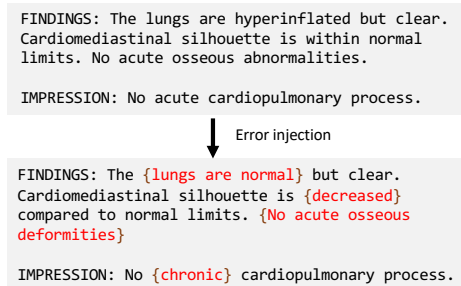

Figure 3: **Error injection example**. In this example, we automatically introduce errors that fall within the categories of incorrect prediction.

age. In this study, we concentrate on the first four categories, providing detailed explanations next.

1) A false prediction of findings occurs when the report incorrectly identifies a medical condition of finding that is not present in the radiological images. 2) Incorrect location/position of findings involves identifying the right finding but attributing it to the wrong anatomical location within the image. 3) Incorrect severity of findings arises when the report either underestimates or overestimates the seriousness of a condition evident in the radiological images. 4) Mention of comparison not present in the reference impression refers to instances where the report includes comparative references to previous images that are not part of the reference impression.

The error injection process involves the **intentional** systematic introduction of errors into the MIMIC-CXR radiological reports (see Figure 3 and 4). This process is executed through a combination of manual techniques and automated injection, explained next, together contributing to 120,123 erroneous reports. For each report, two erroneous versions of the same report are created and contain a combination of error error types mentioned above.
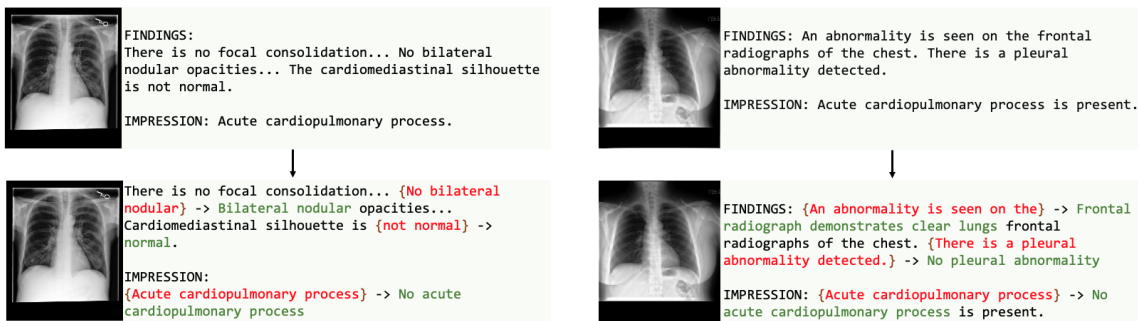

Figure 4: **Qualitative example of medical autocorrection:** The top section shows the initial report with errors. The bottom section displays the report after processing by our model, with corrections in green and erroneous terms struck through. This exemplifies the model's capability to identify and rectify inaccuracies within clinical text.

Automated Error Injection. We automatically introduce errors using a generative large language model (LLM). This allows us to generate more diverse examples that do not fit into specific patterns and follow a natural language distribution. We provide the LLM with specific prompts crafted to reflect the nuances of radiological reporting. Each prompt contains a segment of a real MIMIC-CXR report, followed by instructions to alter it in a way that mimics one or a combination of the four identified error categories mentioned. Specifically, we use OpenAI's GPT-4 (OpenAI, 2023).

To illustrate, for the category of incorrect predictions, a typical prompt to the LLM would read: "Given the following excerpt from a radiological report, modify it to contain an incorrect diagnostic prediction while maintaining a plausible and professional tone." A similar approach was used for the other categories, ensuring that each generated error is contextual and aligned with real-world reporting scenarios.

Manual Error Injection. We manually alter location information, medical condition labels, and the severity of such conditions. See Figure 5 for some examples.

Original:
FINDINGS: Heart size is normal. Cardiomediastinal silhouette and hilar contours are normal. Lungs are clear. Pleural surfaces are clear without effusion or pneumothorax. IMPRESSION: Normal chest radiograph.

Altered:
FINDINGS: Heart size is {slightly enlarged}. Cardiomediastinal silhouette and hilar contours are normal. Lungs are clear. Pleural surfaces are {showing signs of minor effusion} and {early pneumothorax}. IMPRESSION: {Possible cardiopulmonary abnormalities detected}.

Original:
FINDINGS: The lungs are grossly clear. The cardiomediastinal silhouette is stable. No acute osseous abnormalities identified. IMPRESSION: No acute cardiopulmonary process.

Altered:
FINDINGS: The lungs are {showing signs of patchy consolidation}. The cardiomediastinal silhouette is stable. {Minor fractures are noted in the ribcage}. IMPRESSION: {Possible early signs of pulmonary infection and osseous abnormalities}.

Original:
FINDINGS: The endotracheal tube terminates 5 cm above the carina. A right large bore IJ central venous catheter tip is at the level of the mid SVC. Enteric tube is in unchanged position. Tip of left PICC line is seen projecting over the left axillary region. As compared to prior chest radiograph from , diffuse bilateral pulmonary opacifications are unchanged, likely represent severe pulmonary edema. Cardiomegaly is stable. IMPRESSION: Unchanged severe pulmonary edema.

Altered:
FINDINGS: The endotracheal tube terminates {abnormally high, 8 cm above the carina}. A right large bore IJ central venous catheter tip is at the level of the {lower SVC}. Enteric tube is in {a new position, likely dislodged}. Tip of left PICC line is seen projecting over the left axillary region. As compared to prior chest radiograph from , diffuse bilateral pulmonary opacifications are {marginally increased, suggesting worsening of the pulmonary edema}. Cardiomegaly is {showing slight increase}. IMPRESSION: {Progression of severe pulmonary edema noted}.

Original:
FINDINGS: Heart size is normal. Cardiomediastinal silhouette and hilar contours are normal. Lungs are clear. Pleural surfaces are clear without effusion or pneumothorax. IMPRESSION: Normal chest radiograph.

Altered:
FINDINGS: {Abdominal size} is normal. Cardiomediastinal silhouette and hilar contours are normal. {Abdominal region} is clear. {Abdominal surfaces} are clear without effusion or pneumothorax. IMPRESSION: Normal {abdominal} radiograph.

Original:
FINDINGS: The lungs are grossly clear. The cardiomediastinal silhouette is stable. No acute osseous abnormalities identified. IMPRESSION: No acute cardiopulmonary process.

Altered:
FINDINGS: The {upper abdominal region} is grossly clear. The {gastro-cardiomediastinal} silhouette is stable. No acute {abdominal skeletal} abnormalities identified. IMPRESSION: No acute {gastro-cardiopulmonary} process.

Original:
FINDINGS: The endotracheal tube terminates 5 cm above the carina. A right large bore IJ central venous catheter tip is at the level of the mid SVC. Enteric tube is in unchanged position. Tip of left PICC line is seen projecting over the left axillary region. As compared to prior chest radiograph from , diffuse bilateral pulmonary opacifications are unchanged, likely represent severe pulmonary edema. Cardiomegaly is stable. IMPRESSION: Unchanged severe pulmonary edema.

Altered:
FINDINGS: The endotracheal tube terminates 5 cm above the {lower trachea}. A right large bore IJ central venous catheter tip is at the level of the {lower SVC}. Enteric tube is in unchanged position. Tip of left PICC line is seen projecting over the {right axillary} region. As compared to prior chest radiograph, diffuse bilateral pulmonary opacifications are unchanged, likely representing severe pulmonary edema. Cardiomegaly is stable. IMPRESSION: Unchanged severe pulmonary edema.

Original:
FINDINGS: Heart size is normal. Cardiomediastinal silhouette and hilar contours are normal. Lungs are clear. Pleural surfaces are clear without effusion or pneumothorax. IMPRESSION: Normal chest radiograph.

Altered:
FINDINGS: Heart size is {borderline enlarged}. Cardiomediastinal silhouette and hilar contours are {showing minor irregularities}. Lungs are {mostly} clear. Pleural surfaces are clear without effusion or pneumothorax, but {early signs of pleural thickening are noted}. IMPRESSION: {Mildly abnormal chest radiograph, suggestive of early cardiopulmonary changes}.

Original:
FINDINGS: The lungs are grossly clear. The cardiomediastinal silhouette is stable. No acute osseous abnormalities identified. IMPRESSION: No acute cardiopulmonary process.

Altered:
FINDINGS: The lungs are {showing slight cloudiness}. The cardiomediastinal silhouette is {showing early signs of instability}. No acute osseous abnormalities identified, but {close monitoring is advised}. IMPRESSION: {Possible early-stage cardiopulmonary changes observed}

Original:
FINDINGS: The endotracheal tube terminates 5 cm above the carina. A right large bore IJ central venous catheter tip is at the level of the mid SVC. Enteric tube is in unchanged position. Tip of left PICC line is seen projecting over the left axillary region. As compared to prior chest radiograph from , diffuse bilateral pulmonary opacifications are unchanged, likely represent severe pulmonary edema. Cardiomegaly is stable. IMPRESSION: Unchanged severe pulmonary edema.

Altered:
FINDINGS: The endotracheal tube terminates 5 cm above the carina. A right large bore IJ central venous catheter tip is at the level of the mid SVC. Enteric tube is in unchanged position. Tip of left PICC line is seen projecting over the left axillary region. As compared to prior chest radiograph, diffuse bilateral pulmonary opacifications are unchanged, {suggesting moderately severe pulmonary edema}. Cardiomegaly is {showing signs of slight increase}. IMPRESSION: {Moderately severe pulmonary edema, requiring close observation}.

Figure 5: Overview of varied error types in radiological reports, as altered via GPT-4 prompts. Top-left focuses on false predictions, top-right on mislocations, and the bottom on severity misjudgments, illustrating common error types in clinical radiology and their potential impacts on diagnostic accuracy. Some reports in the dataset consist of a mixture of errors.

3.2. Error Detection Module

Our approach implements error detection as a per-token classification task (see bottom right of Figure 2). Within this framework, each token of a given erroneous radiological report, accompanied by an image, is subject to binary classification. This process discerns the correctness of each token within the image's contextual parameters.

We use an image encoder – specifically, a Vision Transformer (Dosovitskiy et al., 2021) – that has been fine-tuned on chest X-ray images (Irvin et al., 2019). We use the encoder to create two distinct forms of embeddings: patch embeddings and a singular, attention-pooled image embedding. These embeddings are subsequently projected onto a uniform dimensional space, aligned with that of the token embeddings.

The erroneous radiological report is encoded using GatorTron-medium, a language model optimized for electronic health records (Yang et al., 2022), from which token embeddings are extracted. Both the image and text encoders are fixed.

We investigate three image-conditioning strategies. Figure 2 shows an overview:

1) **Approach 1: Patch**. We condition on patch embeddings, appending them to the token embeddings:

$$S_1 = [P; T], \tag{1}$$

resulting in a length of $P + N$, with P and N representing the counts of patch and token embeddings, respectively.

2) **Approach 2: Pool**. We condition on the pooled image embedding by appending it to the token embeddings, yielding a sequence:

$$S_2 = [I; T], \tag{2}$$

with a length of $1 + N$.

3) **Approach 3: Concatenate**. We concatenate the pooled image embedding with each token embedding, producing a sequence of length N but with a dimensionality of $2 \times \text{dim\_size}$:

$$S_3 = \text{concat}(I, T_i) \quad \forall i \in \{1, \ldots, N\}, \tag{3}$$

where T_i is the i-th token embedding.

The sequences from each approach are then processed through two multihead self-attention blocks, each with eight heads, followed by a per-token classification. To handle the imbalance in the classification of tokens (as the number of erroneous tokens is much less than the correct tokens), we employ Focal Loss (Lin et al., 2018) for binary classification, defined as:

$$\mathcal{L}_{\text{detection}} = -\frac{1}{NM} \sum_{j}^{M} \sum_{i}^{N} \left[\alpha_{ji}(1 - p_{ji})^{\gamma} y_{ji} \log(p_{ji}) + (1 - \alpha_{ji})p_{ji}^{\gamma}(1 - y_{ji}) \log(1 - p_{ji}) \right], \tag{4}$$

where j is the index of the individual report examples, i is the index of each token within a single report, α_{ji} is a weighting factor to balance the importance of positive/negative examples, p_{ji} is the predicted probability of the i-th token in the j-th report being corrected, y_{ji} is the true label, and γ is the focusing parameter that adjusts the rate at which easy

examples are down-weighted. N represents the number of tokens being classified and M is the number of examples in each batch. We choose a γ factor of 2 and set α to 0.85, batch size of 64 and trained on the first 200 tokens (See figure 8 for justification of this choice)of reports for 20 epochs. See suplementary material for details.

3.3. Error Correction Module

Given the detected errors in the radiological reports, the purpose of the error correction module is to rectify them (see top-right of Figure 2). We condition the 344M-parameter model GPT-2 Medium (Radford et al., 2019), fine-tuned on a corpus of PubMed abstracts on the images. For this part, the image encoder and the GPT-2 model are not frozen. GPT-2 is an autoregressive neural network that leverages self-attention, conditioning the generation of each token in a sequence on the preceding tokens. The self-attention mechanism can be represented as:

$$SA(Y) = \text{softmax}((YW_q)(YW_k)^T)(YW_v), \qquad (5)$$

where Y denotes the token embeddings, and W_q, W_k, W_v are projection matrices for queries, keys, and values, respectively.

To integrate the image-level patch features with the textual data, we concatenate the patch embeddings P from the image encoder with the token embeddings T from the text encoder:

$$S = [P; T], \qquad (6)$$

where $[;]$ signifies the appending operation, resulting in an extended sequence S length that combines both modalities.

Ground truth:
Patient is rotated to the left. Left base opacity is likely due to pleural effusion... The cardiomediastinal silhouette is difficult to assess given rib patient rotation...

Incorrect:
Patient is rotated to the right. Left base opacity is likely due to pleural effusion... The cardiomediastinal silhouette is difficult to assess given rib patient rotation... Aortic silhouette is tortuous.

Autocorrected:
Patient is {rotated to the right} -> rotated to the left. Left base opacity is likely due to pleural effusion... The cardiomediastinal silhouette is difficult to assess given rib patient rotation... {Aortic silhouette is tortuous} -> no signs of aneurysmal dilation

Insight: Attention here is drawn more towards the shoulders, and head regions, areas useful in assessing body rotation and orientation, rather than the lung fields where typical pathologies like opacities are found. Focus on the center of the radiograph shows that it attends to the correct regions when looking for the aorta.

Ground truth:
Frontal radiographs of the chest demonstrate clear lungs. The cardiac and mediastinal contours are normal. No pleural abnormality is detected.

Incorrect:
Lateral radiographs reveal opacification in the lungs. The cardiac silhouette is enlarged with abnormal contours of the mediastinum. Plural effusion is detected

Autocorrected:
{Lateral radiographs} -> Frontal radiographs reveal {opacification in the lungs} -> no evidence of opacification. The cardiac silhouette is {enlarged} -> normal in size with abnormal contours of the mediastinum. {Plural effusion is detected} -> Pleural abnormality is not present.

Insight: The behavior of the model here shows a focus on anatomical regions and features characteristic of frontal chest radiographs. The model's correction of "pleural effusion" shows that it focuses on correct regions when differentiating between subtle pathologies.

Ground truth:
There is a focal consolidation at the left lung base adjacent to the lateral hemidiaphragm. There is mild vascular engorgement. The cardiomediastinal silhouette is remarkable for aortic arch calcifications.

Incorrect:
There is no focal consolidation at the right lung base adjacent to the lateral hemidiaphragm. There is no vascular engorgement. The cardiomediastinal silhouette is normal for aortic arch.

Autocorrected:
There is {no focal consolidation} -> a focal abnormality at the {right} -> left lung lung base adjacent to the lateral hemidiaphragm. There is {no vascular engorgement} -> a vascular engorgement. The cardiomediastinal silhouette {is normal} -> not normal for aortic arch.

Insight: Here the model accurately focuses on the relevant image region with the discrepancy implying that the corrections made by the framework are grounded in the correct image regions.
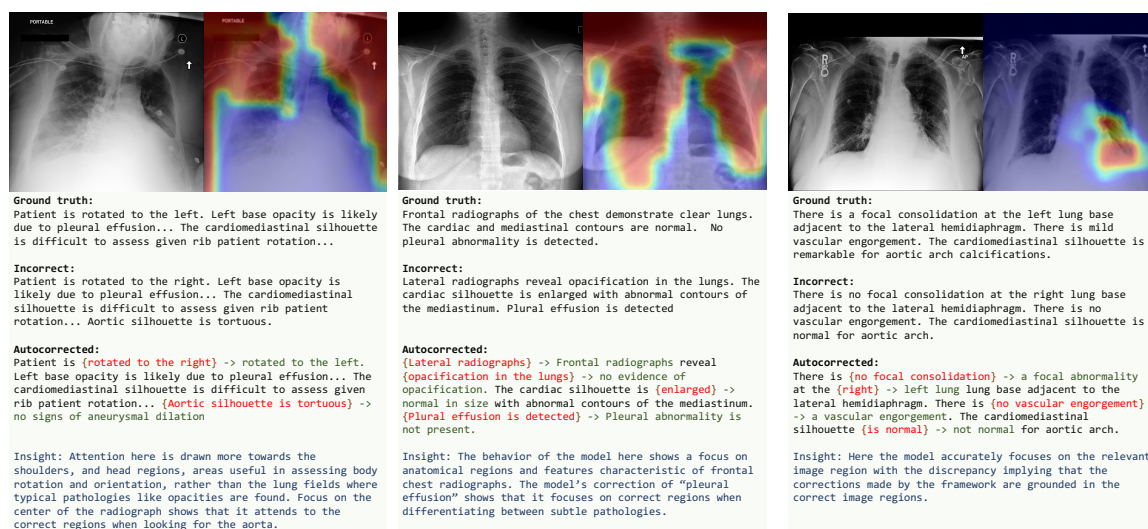
Figure 6: **Analysing visual attention and autocorrection:** This figure shows how the model's attention is distributed across different regions during error correction. The first image correction focuses on body orientation and the assessment of the aorta, the second on distinguishing between frontal and lateral chest radiographs and the presence of pleural effusion, while the third focuses on the accurate identification of lung abnormalities and aortic arch conditions. We use bicubic interpolation to visualize the attention maps.
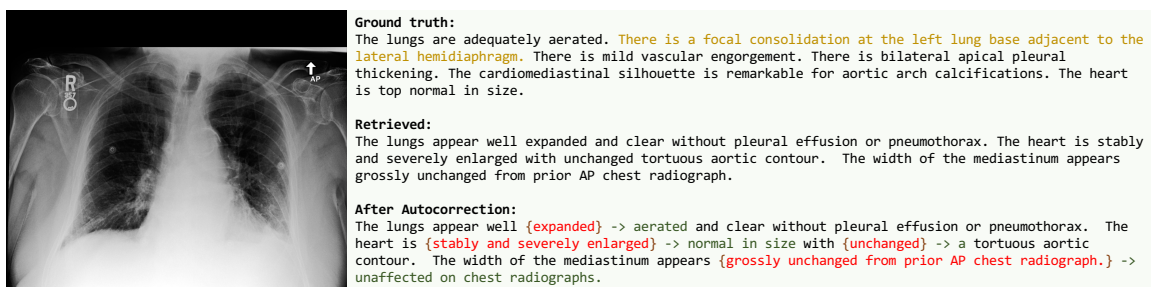
Figure 7: **Enhancing Retrieval with Autocorrection:** This illustration shows the `DETECT+CORRECT` framework in action, where it detects and corrects errors in a report generated by a CLIP-based radiology report generation model (not optimized for SOTA). Highlighted in yellow are sections beyond the scope of correction, as they were not retrieved. Addressing these would shift the task towards report generation, beyond the intended scope of mere correction. In this case, the performance of the autocorrection is only as good as the retrieval model.

In the fine-tuning process, we utilize a **masked training strategy** where tokens identified as errors within the text are replaced with the `[ERROR]` token. This targeted substitution sharpens the learning objective by focusing the model on predicting accurate replacements exclusively at the flagged error positions. As a result, the loss function is calculated solely at these specific locations, directing the model's learning efforts toward correcting these errors. The correction loss function is defined using cross-entropy, which is calculated between the model's output logits and the true token embeddings. We adjust for differences in token lengths between the predicted output and the ground truth by padding to match the longer sequence. The formal expression for the correction loss function is:

$$\mathcal{L}_{\text{correction}} = -\frac{1}{MN_{detection}} \sum_{j}^{M} \sum_{i}^{N} \mathbb{1}_{\{i \in \text{error}\}} y_{ji} \log(\hat{y}_{ji}), \tag{7}$$

where:

M is the number of examples in each batch. $N_{detection}$ is the number of erroneous tokens being corrected. j is the index of the individual report examples. i is the index of each token within a single report. $\mathbb{1}_{i \in \text{error}}$ is an indicator function that is set to 1 for the positions of the `[ERROR]` tokens and 0 elsewhere. y_{ji} is the ground truth for the i-th token in the j-th report. $\hat{y}_{ji}$ is the predicted probability for the i-th token in the j-th report. This ensures that the loss calculated focusing exclusively at locations where the model's predictions for the error tokens need to align with the actual intended medical terminology.

3.4. Training

The training procedure is separated into two distinct phases to create an element of model interpretability. Initially, we focus on the error detection model, which is trained to recognize inaccuracies within reports using the three conditioning strategies outlined. Subsequently, the error correction module is independently trained to generate corrections for the identified and masked errors. This sequential training, as opposed to a joint training approach, allows for a clearer understanding of the error identification and correction mechanisms, ensuring

that each proposed correction can be traced back to a specific detected error. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a batch size of 64 and a learning rate of 3×10^{-4}, which is adjusted using a cosine annealing scheduler (Loshchilov and Hutter, 2017).

3.5. Inference

During inference, we input an image and its corresponding report, which may contain errors, into the error detection module. This module identifies and marks the specific parts of the report that are erroneous. The marked segments are masked with the `[ERROR]` token. Following this, the masked sections are fed into the autocorrection module, which then makes the necessary corrections to the text, effectively simulating the *autocorrection* process. To accommodate corrections that may vary in length we use beam search (Freitag and Al-Onaizan, 2017) and nucleus sampling (Holtzman et al., 2020).

4. Experiment Setup

We evaluate our framework across three distinct metrics: the accuracy of error detection within the manipulated reports, the alignment of the autocorrected reports with the original ground truth, and the effectiveness of the model in rectifying errors introduced by a retrieval-based report generation system that follows Endo et al. (Endo et al., 2021).

4.1. Dataset and preprocessing

We use the X-rays and the altered reports from the MIMIC-CXR dataset (Johnson et al., 2019) to train and evaluate our the error detection module, while the original, unaltered reports are reserved for error correction training and evaluation purposes. We adopt an 80:20 split for training and validation, respectively, and assess the model's performance using a separate set of 6,000 image/erroneous text reports reserved for testing.

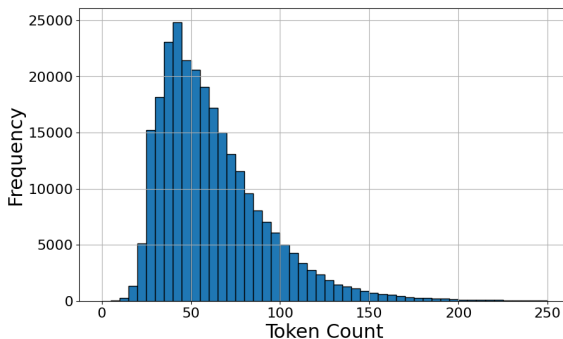All images are resized to 224x224 pixels, ensuring the original aspect ratio is main-

Figure 8: Distribution of report lengths in the MIMIC-CXR dataset that motivated the choice of training on the first 200 tokens. Lengths greater than 250 are not included.

tained, with padding applied as necessary, and they are normalized to have a zero mean and unit standard deviation. Image data augmentation during training includes color jitter, Gaussian noise, and affine transformations following Tanida et al.(Tanida et al., 2023). Text preprocessing involves retaining only the findings and impressions sections of the reports and removing extra whitespace, such as line breaks. Training, however, is limited to the first 200 tokens (See Figure 8), padding shorter reports and truncating longer ones. The findings section typically encapsulates the radiologist's observations, while the impressions section provides a concise summary of the clinical significance of these observations. No further preprocessing is applied to the text of the reports.

4.2. Evaluation metrics

In evaluating our image-conditioned autocorrection framework, we use a multi-level approach. For per-token binary classification, we employ standard precision, recall, and F1 scores to gauge individual token accuracy with emphasis on the performance on the minority class—which in our case represents the introduced errors. To adjust the error detection module's performance, we utilize an "error sensitivity threshold," allowing for adjustable sensitivity in error detection. A threshold set closer to 1.0 results in a stringent detection approach, identifying more potential errors, whereas a setting closer to 0.6 yields a more lenient detection module, reducing the likelihood of false positives (refer to Fig. 9). In our experiments, we set the threshold to 0.7.

At the report level, we assess overall performance using established natural language generation metrics: BLEU (Miura et al., 2021), which measures n-gram overlap between generated and reference reports; METEOR (Banerjee and Lavie, 2005), which captures semantic content; and ROUGE-L (Lin, 2004), focusing on the longest common subsequence in the text. Additionally, we assess the clinical efficacy of the autocorrected reports by utilizing metrics based on the specific pathology classes from the CheXpert dataset (Irvin et al., 2019). The results are microaveraged across the 14 classes.

Approach	Precision	Recall	F1-score
Patch	**0.4931**	0.8763	**0.6311**
Pool	0.4862	0.8569	0.6204
Concatenate	0.4647	**0.9020**	0.6134

Table 1: **Error Identification Results**. We compare the three approaches for the error detection module introduced in Section 3. This comparison highlights the better performance of the patch approach, outperforming other conditioning methods in per-token classification. We set the "error sensitivity threshold" to 0.7 for these results (see Section 4.2 for details on the threshold).

4.3. Evaluation strategy and baselines

In our comparative analysis, we assess the performance of a CLIP-based (Radford et al., 2021) trained retrieval generative model (Endo et al., 2021) enhanced by our autocorrection framework against previous state-of-the-art models in radiology report generation. These models include R2Gen (Chen et al., 2020b), CMN (Chen et al., 2022), PPKED (Li et al., 2019), $\mathcal{M}^2$ TR. PROGRESSIVE (Nooralahzadeh et al., 2021), Contrastive Attention (Liu et al., 2021b), AlignTransformer (You et al., 2022), $\mathcal{M}^2$ Trans (Cornia et al., 2020a), ITA (Schaffer et al., 2019), and CvT-212DistilGPT2 (Liu et al., 2019). These models, optimized for standard language model loss and augmented with rewards for factual completeness and
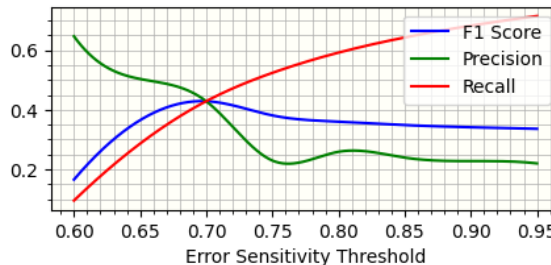

Figure 9: **Trade-off between precision, recall and F1 score at various error sensitivity thresholds for error detection module.** The F1 score peaks at a threshold of approximately 0.7, suggesting an optimal balance between precision and recall at this point that we use to evaluate error detection.

consistency (Liu et al., 2021b), serve as benchmarks for our analysis. Results are cited directly from their respective publications unless otherwise noted. Since, to our knowledge, our work is the first to incorporate an autocorrection framework in radiological report generation models, we do not have direct baselines to compare against. Therefore, in Fig. 7 we also provide a qualitative component that shows the improvements our method offers.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L F1
Baseline	0.238	0.201	0.155	0.091	0.139	0.227
Patch approach	**0.3169** (0.3073, 0.3266)	**0.2539** (0.2429, 0.2648)	**0.2091** (0.1971, 0.2211)	**0.1693** (0.1568, 0.1817)	**0.2043** (0.1806, 0.2280)	**0.3544** (0.3388, 0.3700)
Pool approach	0.3056 (0.2961, 0.3152)	0.2387 (0.2280, 0.2494)	0.1927 (0.1811, 0.2044)	0.1523 (0.1401, 0.1646)	0.1787 (0.1560, 0.2015)	0.3376 (0.3231, 0.3522)
Concatenate approach	0.3031 (0.2937, 0.3125)	0.2370 (0.2265, 0.2475)	0.1915 (0.1800, 0.2030)	0.1504 (0.1382, 0.1627)	0.1776 (0.1550, 0.2002)	0.3369 (0.3219, 0.3518)

Table 2: **Corrections Ablation Study**. We report metrics evaluating autocorrected reports against ground truth for different approaches. Baseline indicates the original scores of the uncorrected reports in relation to the groundtruth. Conditioning on patch embeddings outperforms the other conditioning mechanisms on improving the quality of the report. Refer to Section 3 for definition of the different approaches. We report 95% Confidence Intervals in parenthesis.

Dataset	Method	NLG Metrics ↑						CE Metrics ↑		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	P	R	F1
	Uncorrected retrieval (Ours)	0.216	0.108	0.056	0.029	0.051	0.187	0.192	0.173	0.183
	Retrieval with Autocorrection (Ours)	0.370	0.234	0.175	0.125	0.112	0.230	0.263	0.444	0.330
MIMIC-CXR	R2Gen(Chen et al., 2020b)	0.353	0.218	0.145	0.103	0.142	0.277	0.331	0.224	0.228
	CMN(Chen et al., 2022)	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.278
	PPKED(Liu et al., 2021a)	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-
	$\mathcal{M}^2$ TR. PROGRESSIVE(Nooralahzadeh et al., 2021)	0.378	0.232	0.154	0.107	0.145	0.272	0.240	0.428	0.308
	Contrastive Attention(Li et al., 2018)	0.350	0.219	0.152	0.109	0.151	0.283	0.352	0.298	0.303
	AlignTransformer(You et al., 2021)	0.378	0.235	0.156	0.112	0.158	0.283	-	-	-
	ITA(Wang et al., 2022)	0.395	0.253	0.170	0.121	0.147	0.284	-	-	-
	CvT-212DistilGPT2(Liu et al., 2019)	0.392	0.245	0.169	0.124	0.153	0.285	0.359	0.412	0.384
	RGRG(Tanida et al., 2023)	0.373	0.249	0.175	0.126	0.168	0.264	0.461	0.475	0.447
	RGRG(Tanida et al., 2023) with autocorrection	0.375	0.243	0.151	0.116	0.157	0.271	**0.469**	0.458	**0.463**

Table 3: **Quantitative effect of autocorrection on radiology report generation through retrieval**. Evaluation is done against baseline models and an uncorrected retrieval system that was intentionally not optimized to state-of-the-art standards. Note that **our method was not trained to do report generation**, only error detection and correction. This approach shows the potential of how the performance of a radiology report generative model can be improved through autocorrection of incorrect parts (See performance on RGRG for the effect of autocorrection on a SOTA model). Grayed-out results have been trained explicitly for report generation. We just add them here for completeness. See 7 for a qualitative example.

Robustness of error correction. Quantitative assessment of our error correction model presents challenges. Specifically, standard text generation metrics are limited in conveying *robustness* information. Robustness in the context of error correction models would involve assessing how well the model can handle and correct various types of errors, including those that might not significantly affect n-gram overlap or semantic similarity–for example, errors caused by negations (See Fig. 6 for examples). Consider the difference, in
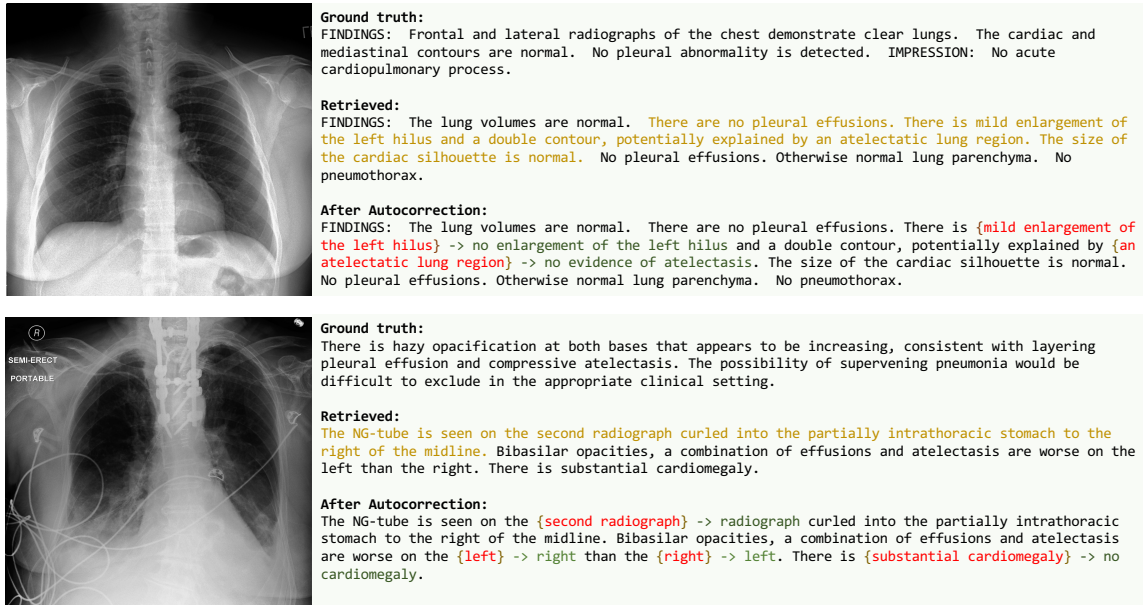
Figure 10: **Challenges in Autocorrection of Retrieved Reports (See figure 7):** This figure highlights a significant challenge in autocorrecting retrieval-based radiology reports. It showcases instances where reports, while accurately retrieved, include additional details that are not pertinent to the current diagnosis (highlighted in yellow). These extraneous segments, although potentially relevant, deviate from the ground truth and are not efficiently identified or removed by the autocorrect system. This limitation underscores the need for enhanced discernment capabilities in the framework to differentiate between essential and non-essential information in the context of clinical diagnosis. Beyond that, autocorrection happens just as it would as illustrated in Figure 6

a radiological report, between "no evidence of fracture" and "evidence of fracture": it is only one negation word ("no"), but the meanings are diametrically opposed (See sentences in Fig. 4 and Fig. 6). Traditional NLG metrics might score these two sentences as highly similar because of the high degree of word overlap, despite the fact that the negation leads to a critical error in meaning. To tackle this, we focus on the model's resilience to variations in types and positions of errors introduced into radiological reports that may not be reflected in changes in NLG metrics.

5. Results and Discussion

5.1. Performance Analysis

We present error identification results in Table 1. We see that conditioning the error detection module on token embeddings leads to much better performance in error detection than the other conditioning mechanisms highlighted in section 3. We attribute this to the availability of more image contextual information that can be used to improve the classification of the tokens.

We also show the error correction results in Table 2. The corrected reports are compared against the ground truth along with the original incorrect reports whose scores are listed as baselines. Error correction is only as good as the error identification module hence

the observed better performance of the approach that involves conditioning on the token embeddings in Table 2. Overall, as observed in Table 3 our proposed framework demonstrates an ability to significantly improve the quality of reports generated by a retrieval model almost doubling the F1 score on clinical efficacy along with significant improvement in the NLG metrics.

5.2. Error Correction Impact

The system's error correction capabilities have a significant impact on correcting misidentification of anatomical location misidentification and severity of findings examples shown in Fig.s 4, 7, 6 as well as the inappropriate use of terminologies. The correction of errors related to location misindenfitication and severity of findings is indeed clinically relevant as they can as they directly influence the diagnostic interpretations and subsequent patient management strategies. For example, in Fig. 6 the autocorrection framework corrected orientation of a patient using visual cues from shoulders and head regions as well reconciling discrepancies between the described radiographic view and the observed anatomical landmarks—turning 'lateral radiographs' into 'frontal radiographs' and addressing 'opacification in the lungs.' Nonetheless, there is a challenge in correcting errors related to incorrect predictions or omissions of findings as these scenarios require a level of clinical inference akin to report generation, which falls outside the intended scope of our autocorrection framework (See Figs. 7 and 6). Only errors that have been flagged as incorrect can be corrected, meaning the errors have to be present in the first place. This could be an area of focus for future improvements.

5.3. Guardrail for automatic report generation

We also qualitatively evaluate the potential of autocorrection in improving radiological report generation methods. This system was layered atop a retrieval-based report generation model, which was not optimized for SOTA results. However, the integration of autocorrection demonstrated a noteworthy enhancement in the accuracy and relevance of generated reports (see Table 2).

Case Analysis: Consider Fig. 7, where the ground truth report noted a "focal consolidation in the left lung base, mild vascular engorgement, bilateral apical pleural thickening, aortic arch calcifications, and a heart size that is top normal." In contrast, the retrieved report initially presented an inaccurately "stable and severely enlarged heart, along with a clear lung field," diverging significantly from the ground truth as seen in the retrieved report. The autocorrection system modified key descriptors in the retrieved report, aligning them closer to the ground truth. For example, it corrected "well expanded" to "adequately aerated" and "stably and severely enlarged" to "normal in size," among other adjustments. This indicates the system's potential in identifying and rectifying specific inaccuracies in radiological reports.

When benchmarked autocorrection layered atop of retrieval against established methods like R2Gen, CMN, PPKED (see Table 3), our autocorrection approach displayed commendable performance. It closely approached the baseline method and showed its utility in enhancing the accuracy of diagnostic reporting in radiology (refer to Table 3 and Fig. 7). This suggests that even non-SOTA retrieval-based systems can be significantly improved

with effective autocorrection mechanisms. The ability to adjust key clinical descriptors accurately is crucial, as these directly impact the clinical interpretation and subsequent patient management decisions.

6. Conclusion

In this study, we have introduced a framework for medical report autocorrection with a strong emphasis on error detection and correction within radiological texts grounded in the images they describe. Our model leverages image-conditioned data processing to enhance the accuracy of medical report generation. This approach not only ensures the generation of clinically coherent reports but also contributes to the reduction of errors that could significantly impact patient care.

Ethical Considerations. The domain of automated medical text correction holds immense potential to enhance healthcare delivery. Nevertheless, the implications of incorrect corrections are substantial, with the possibility of negatively affecting patient outcomes. While our system represents a step towards minimizing human error, it also raises concerns regarding overdependence on automation (Agarwal et al., 2023). While autonomous reporting may be a distant possibility (Saenz et al., 2023), we expect that systems may be first employed as decision support tools with guardrails (Sanchez et al., 2023) to ensure that clinicians remain engaged and vigilant in the decision-making process.

7. Limitations

Given the high stakes of medical autocorrection, especially within the context of clinical data, discussing limitations here is key. The primary dataset used in this paper may not encompass the complete spectrum of errors inherent in radiological reports. Given the variability of medical data, including rare but critical scenarios, it remains a significant challenge to capture every possible error type within the training corpus. This limitation could affect the model's ability to generalize to errors not represented in the dataset.

Furthermore, the proposed framework operates on the assumption that input reports are largely accurate, with only a few factual errors. This assumption might hold true for real-world reports written by radiologists who are not likely to make errors. However, for reports that are poorly constructed or contain multiple, compounded errors the framework will perform poorly and will amount to the solving of a report generation task. The model's performance is contingent upon this assumption, which may not always be a given in clinical settings.

Generalization Issues. The model has been trained and validated on a dataset which may not cover the full diversity of medical language and errors. Consequently, there is a risk that the model may underperform when faced with novel errors or those that manifest in contexts divergent from the training data. To this end, we propose a strategy for real-world clinical applications——using the same approach, tailor error injection based on specific hospital environments before error correction is learnt for that setting.

Future Work. Addressing these limitations presents clear avenues for future work. Tools that identify clinical entities and relationships in the report (Jain et al., 2021; Khanna et al., 2023) could also be incorporated to guide the correction by separating style from

15

content (Yan et al., 2023). Classification of severity of errors, using schemes as introduced in (Jeong et al., 2023), could allow for a more clinically meaningful analysis of the error reduction. Enriching the dataset with a broader range of error types, investigating context-aware error detection mechanisms, and developing robust models that can handle a variety of report qualities are critical next steps. Code and datasets used can be found here `medautocorrect.cs.columbia.edu`

References

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research, 2023.

Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557, 2021. ISSN 2352-9148. doi: https://doi.org/10.1016/j.imu.2021.100557. URL `https://www.sciencedirect.com/science/article/pii/S2352914821000472`.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `https://aclanthology.org/W05-0909`.

A. P. Brady. Error and discrepancy in radiology: inevitable or avoidable? *Insights into Imaging*, 8(1):171–182, 2017. doi: 10.1007/s13244-016-0534-1. URL `https://doi.org/10.1007/s13244-016-0534-1`. Epub 2016 Dec 7.

A. P. Brady. Radiology reporting—from hemingway to hal? *Insights into Imaging*, 9:237–246, 2018. doi: 10.1007/s13244-018-0596-3. URL `https://doi.org/10.1007/s13244-018-0596-3`.

Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs, 2020a.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112. URL `https://aclanthology.org/2020.emnlp-main.112`.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation, 2022.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020a.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning, 2020b.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer, editors, *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 209–219. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/endo21a.html.

European Society of Radiology (ESR). Good practice for radiological reporting. guidelines from the european society of radiology (esr). *Insights into Imaging*, 2(2):93–96, 2011. doi: 10.1007/s13244-011-0066-7. URL https://doi.org/10.1007/s13244-011-0066-7. Epub 2011 Feb 6.

Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-3207. URL https://doi.org/10.18653%2Fv1%2Fw17-3207.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation. *arXiv preprint arXiv:2303.17579*, 2023.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1240. URL https://aclanthology.org/P18-1240.

Baoyu Jing, Zeya Wang, and Eric Xing. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1657. URL https://aclanthology.org/P19-1657.

A. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng. Mimic-cxr database, 2019. URL https://doi.org/10.13026/C2JT1Q.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning, 2015.

Sameer Khanna, Adam Dejl, Kibo Yoon, Quoc Hung Truong, Hanh Duong, Agustina Saenz, and Pranav Rajpurkar. Radgraph2: Modeling disease progression in radiology reports via hierarchical information extraction. *arXiv preprint arXiv:2308.05046*, 2023.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1537–1547, Red Hook, NY, USA, 2018. Curran Associates Inc.

Xiangyang Li, Shuqiang Jiang, and Jungong Han. Learning object context for dense captioning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33018650. URL https://doi.org/10.1609/aaai.v33i01.33018650.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13748–13757, 2021a. URL https://api.semanticscholar.org/CorpusID:235421693.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest X-ray report generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.23. URL https://aclanthology.org/2021.findings-acl.23.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation, 2019.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

G Maskell. Error in radiology-where are we now? *British Journal of Radiology*, 92(1095):20180845, 2019. doi: 10.1259/bjr.20180845. URL https://doi.org/10.1259/bjr.20180845. PMID: 30457880; PMCID: PMC6540865.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.416. URL https://aclanthology.org/2021.naacl-main.416.

Hoang Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. Automated generation of accurate & fluent medical X-ray reports. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3552–3569, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.288. URL https://aclanthology.org/2021.emnlp-main.288.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp. 241. URL https://aclanthology.org/2021.findings-emnlp.241.

OpenAI. Gpt-4 technical report, 2023.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Agustina D Saenz, Zach Harned, Oishi Banerjee, Michael D Abràmoff, and Pranav Rajpurkar. Autonomous ai systems in the face of liability, regulations and costs. *NPJ digital medicine*, 6(1): 185, 2023.

Morgan Sanchez, Kyle Alford, Viswesh Krishna, Thanh M Huynh, Chanh DT Nguyen, Matthew P Lungren, Steven QH Truong, and Pranav Rajpurkar. Ai-clinician collaboration via disagreement prediction: A decision pipeline and retrospective analysis of real-world radiologist-ai interactions. *Cell Reports Medicine*, 4(10), 2023.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your ai: Expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 240–251, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302308. URL https://doi.org/10.1145/3301275.3302308.

Zhuang Shao, Jungong Han, Demetris Marnerides, and Kurt Debattista. Region-object relation-aware dense captioning via transformer. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2022. doi: 10.1109/TNNLS.2022.3152990.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2023. doi: 10.1109/cvpr52729.2023.00718. URL https://doi.org/10.1109%2Fcvpr52729.2023.00718.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015.

Lin Wang, Munan Ning, Donghuan Lu, Dong Wei, Yefeng Zheng, and Jie Chen. An inclusive task-aware framework for&radiology report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, page 568–577, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-16451-4. doi: 10.1007/978-3-031-16452-1_54. URL https://doi.org/10.1007/978-3-031-16452-1_54.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays, 2018.

Benjamin Yan, Ruochen Liu, David E Kuo, Subathra Adithan, Eduardo Pontes Reis, Stephen Kwak, Vasantha Kumar Venugopal, Chloe P O'Connell, Agustina Saenz, Pranav Rajpurkar, et al. Style-aware radiology report generation with radgraph and few-shot prompting. *arXiv preprint arXiv:2310.17811*, 2023.

X. Yang, A. Chen, N. PourNejatian, et al. A large language model for electronic health records. *npj Digital Medicine*, 5:194, 2022. doi: 10.1038/s41746-022-00742-2. URL https://doi.org/10.1038/s41746-022-00742-2.

Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. Context and attribute grounded dense captioning, 2019.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 72–82, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87199-4.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation, 2022.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023a. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2023.100802. URL https://www.sciencedirect.com/science/article/pii/S2666389923001575.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023b. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100802. URL https://www.sciencedirect.com/science/article/pii/S2666389923001575.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment, 2019.