Pragmatic Radiology Report Generation

Dang Nguyen DANGNGUYEN@UCHICAGO.EDU
The University of Chicago

Chacha Chen CHACHA@UCHICAGO.EDU
The University of Chicago

He He HHE@NYU.EDU
New York University

Chenhao Tan CHENHAO@UCHICAGO.EDU
The University of Chicago

Abstract

When pneumonia is not found on a chest X-ray, should the report describe this negative observation or omit it? We argue that this question cannot be answered from the X-ray alone and requires a pragmatic perspective, which captures the communicative goal that radiology reports serve between radiologists and patients. However, the standard image-to-text formulation for radiology report generation fails to incorporate such pragmatic intents. Following this pragmatic perspective, we demonstrate that the indication, which describes why a patient comes for an X-ray, drives the mentions of negative observations. We thus introduce indications as additional input to report generation. With respect to the output, we develop a framework to identify uninferable information from the image, which could be a source of model hallucinations, and limit them by cleaning groundtruth reports. Finally, we use indications and cleaned groundtruth reports to develop pragmatic models, and show that they outperform existing methods not only in new pragmatics-inspired metrics (e.g., +4.3 Negative F1) but also in standard metrics (e.g., +6.3 Positive F1 and +11.0 BLEU-2).

1. Introduction

Radiology report generation has emerged as an important problem in machine learning for healthcare (Jing et al., 2018; Wang et al., 2018; Liu et al., 2019; Yuan et al., 2019; Chen et al., 2020; Endo et al., 2021; Miura et al., 2021; Ramesh et al., 2022; Thawkar et al., 2023; Tu et al., 2023). In particular, MIMIC-CXR (Johnson et al., 2019) is a widely used dataset due to its large number of X-ray images and corresponding radiology reports.

INDICATION: An ___-year-old woman with previous aspiration pneumonia and a history of congestive heart failure (CHF).
IMPRESSION: PA and lateral chest compared to ___: Lungs are hyperinflated, due to airway obstruction or emphysema. On the lateral view, aside from a granuloma, there is no pneumonia. The heart size is normal, no pulmonary edema related to CHF. Right pleural effusion is tiny status post pleural tube removal compared to large pleural effusions seen on prior chest radiographs. There are no findings to suggest intrathoracic malignancy. An urgent CT thorax is suggested given the rapid growth of granuloma. These findings were communicated to Dr. ___ at 4:00 p.m. by phone.

Table 1: A synthetic chest X-ray report, highlighting that a report includes more than positive findings from X-ray. Blue: prior comparisons. Light blue: previous procedures. Red: negative mentions. Orange: image view. Green: doctor communication. Purple: medical recommendations.

In this work, we revisit the standard formulation of radiology report generation and the MIMIC-CXR benchmark from a pragmatic perspective. Radiology report generation is typically formulated as an image-to-text problem: generate a complete report given a chest X-ray. We argue that this formulation does **not** align with the functional goal of radiology reports as a communicative device between medical professionals and patients (Hartung et al., 2020).

Input Factor	Input Description	Relevant Content
X-ray image(s)	The image(s) taken for the current study	Positive observations
	Factors Beyond the Images	
Indication	The reason for a patient's visit	Positive & negative observations
Previous studies	Findings from previous chest X-rays	Comparisons to prior studies
Previous treatment, medical history	Medical procedures the patient has received	Mentions of previous procedures
Communication information	Communication between medical professionals, electronic systems	Mentions of what information transfer has taken place
Image view	The X-ray view(s) from which a patient is seen	Mentions of the view, often before commenting on findings
Medical expertise/situation	Medical expertise & knowledge about the patient's preference/other conditions	Medical recommendations

Table 2: Categorization of the types of input that can influence a radiology report. Examples of each type of output can be seen in in Table 1.

To illustrate, Table 1 shows an example report. The very first line is *Indication*, where the radiologist explains why the patient needed a chest X-ray. This information is not part of the image, but plays a crucial role in determining the content of the report. One example is mentions of negative diagnosis (henceforth *negative mentions*): although one might infer that any unnamed observation is negative, this is not how radiologists communicate with each other or with patients. In Table 1, the sentences with negative mentions highlighted in red ("there is no pneunomia" and "no pulmonary edema related to CHF") specifically respond to the conditions "pneumonia" and "CHF" in the indication. In contrast, other common conditions such as Pneumothorax are omitted.

In general, radiologists convey much more information than positive findings from an image and the pragmatic perspective is critical to understand what makes a radiology report. Table 2 provides a comprehensive view of different factors that may affect a report's content. Notably, there are many factors beyond the image itself.[1] Therefore, the typical formulation of radiology report generation does not give the model sufficient information to generate its expected output. This framework allows us to carefully consider what to include in the input and the output to develop reasonable problem formulations so that the model has sufficient information and that the evaluation focuses on the relevant components.

1. Image views may be learnable from the data with images from different views. However, most studies only use a single image as the input and do not group observations by view in the output.

Following the pragmatic perspective, we provide a rigorous analysis to show that the indication drives negative mentions. We thus reformulate the radiology report generation problem as generating a radiology report given an image and an indication. With respect to the output, we use large language models (LLMs) to clean the reports by removing uninferable information from the input, which also redefines the desired generation output. Accordingly, we introduce novel evaluation metrics to disentangle model limitation from uninferable information, negative mentions from correctness of positive findings. Finally, we build pragmatic generation models and demonstrate substantial performance improvements compared to existing approaches. In particular, our LLaMA-based model, even when trained on unclean reports, produces fewer hallucinations than retrieval-based methods retrieving from cleaned data.

As a side outcome of our framework, our analysis reveals a clear distribution shift between the test set and the training set in MIMIC-CXR. On average, each report has only 0.255 negative mentions in the test set, compared to 0.485 in the training set, which challenges the i.i.d assumption. We recommend the community carefully rethink the use of the standard train-test split in MIMIC-CXR for benchmark purposes in the future.

In summary, we make the following contributions:

- We introduce the pragmatic perspective and reformulate the problem of radiology report generation.
- We demonstrate that the indication drives mentions of negative observations and develop new

evaluation metrics inspired by the pragmatic perspective.

- We show that our pragmatics-aware approaches lead to better generation, in both traditional and proposed evaluation metrics.
- We reveal idiosyncrasies in the test set of the MIMIC-CXR dataset.

Our code is available at https://github.com/ChicagoHAI/llm_radiology.

2. Dataset

We use MIMIC-CXR, a chest X-ray dataset containing 377,110 images and their corresponding reports (Johnson et al., 2019). It has been widely used in recent studies on report generation (Liu et al., 2019; Chen et al., 2020; Miura et al., 2021; Endo et al., 2021; Ramesh et al., 2022; Thawkar et al., 2023), and comes with a train/dev/test split.

Following prior work, we use CheXbert to derive groundtruth labels for each image based on the corresponding report (Smit et al., 2020). For each report, there are fourteen conditions: twelve thoracic conditions, one condition for support devices, and one for No Finding. Except for No Finding, each condition can take four labels: 1 (positive), 0 (negative), -1 (uncertain), and missing (not mentioned). No Finding is either missing or 1. Table 3 presents basic statistics for the train/dev/test splits in MIMIC-CXR.

Negative mentions are prevalent. On average, there is about one negative mention for every three positive mentions in the training set. When a report is not labeled "No Finding", this ratio becomes less than one-to-two. This shows that commenting on negative observations is common practice in radiology reporting, a phenomenon that we will revisit in §3.

Substantial discrepancies between the training set and the test set. 41.3% of reports are "No Finding" in the training set, while only 19.8% of the test set are "No Finding". Furthermore, an average test report only contains 0.255 negative mentions compared to 0.485 in an average train report. In contrast, the average numbers of positive mentions are similar between the training set and the test set.

We further group results by conditions identified in the indication[2] with CheXbert in Table 3. When a condition is mentioned in the indication, about half of the time the report has at least one negative mention in the training set, further confirming the importance of negative mentions. Meanwhile, we observe a discrepancy between the training set and the test set: the percentage of negative mentions is much lower, often half of the rate as in the training set, with "Pneumothorax" as the only exception.

This raises the question of whether the issue lies with the training or the test set. We briefly compared the same data statistics across two other datasets: CheXpert (Irvin et al., 2019) and OpenI (Demner-Fushman et al., 2016), and found that their average numbers of negative mentions per report are much more similar to those of MIMIC-CXR's training set than its test set. This gives evidence for the test set being out-of-distribution. These results are further

	Train	Dev	Test
#Reports	371,951	1,837	2,872
% No Finding	41.3	40.9	19.8
avg. #positive mentions	1.35	1.17	1.39
avg. #positive mentions in reports that are not "No Finding"	1.59	1.29	1.49
avg. #negative mentions	0.485	0.232	0.255
avg. #negative mentions in reports that are not "No Finding"	0.826	0.394	0.318
% of reports that have negative mentions			
Pneumothorax	52.8	50.0	46.5
Pneumonia	45.1	34.9	25.4
Edema	44.3	25.5	20.2
Pleural Effusion	45.4	19.4	26.1
Cardiomegaly	42.9	29.6	27.4
Consolidation	52.0	40.0	26.5
Enlarged Cardiomediastinum	44.7	25.0	0.0
Lung Opacity	50.7	26.7	16.0
Lung Lesion	46.0	12.0	18.2
Fracture	43.7	0.0	21.1
Support Devices	50.3	17.0	31.5
Atelectasis	43.5	15.4	15.0
Pleural Other	42.9	0.0	0.0
No Finding	34.0	15.1	16.2

Table 3: Top: Statistics on the positive and negative mentions of MIMIC-CXR. Bottom: Percentage of reports that contain at least one negative mention, conditioned on a condition mentioned in the indication. The conditions are sorted by the frequency of their negative mentions (see Appendix A).

2. Positive, negative, and uncertain labels are all considered mentions.

discussed in §6. However, given that CheXpert does not have reports (although the labels for their images were derived from accompanying reports), and that OpenI is much smaller than MIMIC-CXR and does not have a test set, we decided to use MIMIC-CXR despite its discrepancies.

3. Rethinking Radiology Report Generation Pragmatically

In this section, we start with a rigorous analysis of the connection between the indication and mentions of negative observations. This analysis motivates our reformulation of the generation problem. Then, building on our framework in Table 2, we use large language models to clean reports to remove content that we do not expect models to generate given the image and the indication. Finally, we introduce novel evaluation metrics inspired by this pragmatic perspective.

3.1. A Pragmatic Observation of Indication and Negative Mentions

Consider a normal chest X-ray. Based on the image alone, it is impossible to favor either of the following two reports: "No acute cardiopulmonary process." and "No radiographic evidence for pneumonia." Next, we show that the *indication* section drives negative mentions like that in the second report.

Table 3 has demonstrated the prevalence of negative mentions. We would like to capture the probability of negative mentions given an indication instead of simply computing the percentage of negative mentions in the reports. Leveraging the intuition from our example, the key idea is that the probability of negative mentions only makes sense in reports where the condition is actually not positive; in fact, a condition appearing in the indication increases the probability of the condition being positive, deflating the probability of negative mentions. Therefore, we ignore these positive cases when computing the probability of negative mentions.

Specifically, for a report R, we denote its indication section as $I(R)$. As discussed in §2, for each condition X, the report is labeled as $R_X \in \{1, 0, -1, -2\}$, where $1, 0, -1$ correspond to positive, negative, and uncertain mentions of the condition per CheXbert's convention, while -2 suggests the condition is not mentioned in R. For every condition X except No Finding, we compute two conditional probabilities depending on the event that X appears in the indication, which is denoted $X \in I(R)$:

$$P(\neg X \mid X \in I) = \frac{|\{R : R_X = 0 \wedge X \in I(R)\}|}{|\{R : R_X \in \{0, -2\} \wedge X \in I(R)\}|},$$
$$P(\neg X \mid X \notin I) = \frac{|\{R : R_X = 0 \wedge X \notin I(R)\}|}{|\{R : R_X \in \{0, -2\} \wedge X \notin I(R)\}|},$$

where $\neg X$ refers to negative mentions of X, and $R \in \mathcal{R}$ the set of all reports.

Condition	$P(\neg X \mid X \in I)$	$P(\neg X \mid X \notin I)$
Atelectasis ***	1.7%	0.3%
Cardiomegaly	6.2%	5.8%
Consolidation ***	7.3%	3.3%
Edema ***	23.4%	8.0%
Enlarged Cardiomediastinum ***	8.6%	2.1%
Fracture ***	14.0%	0.3%
Lung Lesion ***	5.8%	0.4%
Lung Opacity ***	2.2%	0.8%
Pleural Effusion ***	18.1%	8.3%
Pleural Other ***	0.9%	0.03%
Pneumonia ***	25.0%	8.9%
Pneumothorax ***	42.7%	9.1%
Support Devices ***	3.7%	0.2%

Table 4: χ^2-test results show that negative mentions are influenced by the indication. *** indicates $p < 0.001$. No Finding is excluded.

Table 4 shows the results and whether the differences between these two probabilities are significant based on the χ^2-test on the training set. All differences are significant except for Cardiomegaly. For most conditions, $P(\neg X \mid X \in I)$ is substantially greater than $P(\neg X \mid X \notin I)$, which offers strong evidence that conditions are more likely to be mentioned as negative when they are inquired about in the indication.

Given the important role of indication in determining negative mentions, we reformulate the problem of radiology report generation as generating the report given an image and an indication.

3.2. Pragmatic Data Cleaning

In addition to including indications as part of the input, we need to carefully consider what the desired output should include. We focus on information that one can generate from the image and the

Rule	Original	Cleaned
Remove comparison to prior studies	In comparison with the study of ___, there are slightly improved lung volumes.	There are slightly improved lung volumes.
Remove communication information	These findings were communicated via the radiology critical results dashboard at 12:57 p.m.	REMOVED
Rewrite new/increased conditions into positive	New large right pneumothorax	Large right pneumothorax
Rewrite resolved conditions into negative	Resolved opacities in the left mid lung.	No opacities in the left mid lung.

Table 5: Example cleaning rules. See Appendix B for details.

indication in this work, so we aim to remove the following information in Table 2: previous studies, previous treatment, recommendations,[3] doctor communications, image view. Our framework is a generalization of previous attempts to clean reports (Ramesh et al., 2022; Thawkar et al., 2023) which focus on removing references to prior studies and image views.

Methodology. We developed our method on a set of 100 manually cleaned reports. Inspired by Thawkar et al. (2023), we use few-shot in-context learning to perform the cleaning. Specifically, we create seven rules to remove the information of interest and prompt Flan-T5-XXL with a small number of examples to clean reports (Longpre et al., 2023) (see Table 5 for examples). We prompt the model using one rule at a time and refer to this approach as "rule composition". This approach provides more flexibility than the fine-tuned classifier (GILBERT) in Ramesh et al. (2022) and leverages the capability of LLMs to rewrite rather than remove information. During the development of our method, we found that cleaning can change the CheXbert labels of a sentence, due to flaws in Flan-T5 and CheXbert, so we employ a simple heuristic after every cleaning step to discard the change if it has changed any label.

Evaluation. We manually cleaned another 160 sentences as a test set. For evaluation, we compute Positive and Negative F1 (see Section 3.3) using the labels of the LLM-cleaned and original sentences to evaluate whether the cleaning process maintains the original labels. We also compute Exact Match (EM) accuracy and BLEU-2 between LLM-cleaned and manually-cleaned sentences to evaluate the similarity at the token level. In addition, we provide a heuristic measure for each type of uninferable information at the report level. For an information type, we define a few keywords denoting a mention of that information. We calculate the percentage of reports that has such information after cleaning. Details on the development and test sets that we use for Flan-T5 report cleaning can be found in Appendix C

3. We opt to be conservative in this work as this information often depends on the patient's preference and urgency.

Model	Pos F1	Neg F1	EM Acc.	BLEU-2
GILBERT	0.915	0.846	0.188	0.505
Flan-T5 (all-rules)	0.930	0.898	0.419	0.514
Flan-T5 (compose-rules)	0.855	0.821	**0.538**	0.527
Flan-T5 (compose-rules + label heuristic)	**1.000**	**1.000**	0.531	**0.541**

Table 6: Report cleaning result at the sentence level.

Table 6 reports our cleaning model's performance at the sentence level. To test the effectiveness of rule composition, we compare our model against a Flan-T5 model prompted using all of the rules in one prompt. All Flan-T5 variants outperform GILBERT, which is expected since the latter only cleans "previous studies" under our framework in Table 2. Since Flan-T5 (compose-rules) outperforms Flan-T5 (all-rules), rule composition is shown to be effective. With the label heuristic, we benefit from cleaning sentences without accidentally changing their meaning, despite the slightly lower accuracy.

Table 7 shows the extent to which Flan-T5 cleans uninferable information at the report level compared to other baselines. It outperforms GILBERT and XrayGPT on cleaning all information types, with the only exception being prior studies, on which it trails behind GILBERT slightly. Given the encouraging results, we employ Flan-T5 with rule composition and the label heuristic to clean MIMIC-CXR.

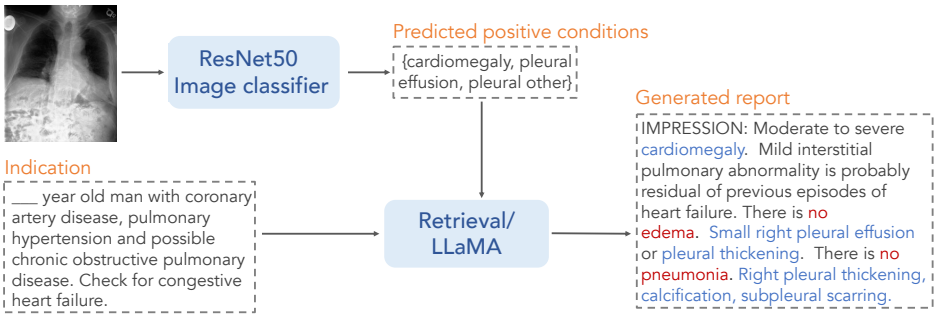

Figure 1: An overview of our approach. Blue: positive findings. Red: negative mentions.

Model	Prior study	Prior proc.	Comm.	Rec.	View
Train	52.6%	1.2%	9.6%	10.5%	6.4%
GILBERT	**25.1%**	1.1%	9.2%	10.3%	6.4%
XrayGPT	53.8%	1.5%	20.2%	22.2%	8.1%
Flan-T5	30.5%	**0.7%**	**4.6%**	**7.3%**	**4.3%**

Table 7: Percentage of reports with uninferable information after cleaning. Lower is better.

3.3. Pragmatic Evaluation

We start by reviewing standard evaluation metrics.

- Clinical efficacy (CE). We include Positive F1, Positive F1-5 that focuses on the most frequent five conditions, and RadGraph F1 (Jain et al., 2021), as they are commonly used to evaluate the correctness of reports, and especially as RadGraph F1 has been shown to align well with radiologists' judgements (Endo et al., 2021; Irvin et al., 2019; Yu et al., 2022).
- Language performance against original reports. We use BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019), which are commonly used for natural language generation tasks. As demonstrated in Table 2, we do not think that these are appropriate metrics because much of the content is impossible to generate given the image. We keep these two metrics as they are standard in existing work.

Inspired by the pragmatic perspective, we believe that existing metrics are flawed in two ways: 1) comparing against original reports expects the model to generate uninferable information; 2) clinical efficacy ignores the prevalent mentions of negative observations. Thus, we develop the following metrics to capture the pragmatic performance of report generation.

- Clean BLEU-2 and Clean BERTScore. As some information is impossible to generate given the image and the indication, using the original report as the groundtruth is not ideal. We thus compute BLEU-2 and BERTScore against the cleaned reports.
- Negative F1 and Negative F1-5. Parallel to Positive F1 and Positive F1-5, we introduce Negative F1 and Negative F1-5, which evaluates against whether a negative mention occurs in the report for a particular label. For Negative F1-5, we use the most frequent five negative labels in the training set, as shown in Table 3: Pneumothorax, Pneumonia, Edema, Pleural Effusion, and Consolidation.[4]
- Hallucination. Finally, we measure how often the model generates information that cannot be generated from the image and the indication. For simplicity, we merge all uninferable information types into one measure with the keywords in §3.2 and compute the percentage of generated reports that contains any uninferable information.

Following Endo et al. (2021), all evaluations are done on the impression section.[5]

4. Experiments

To demonstrate the practical importance of the pragmatic perspective, we perform experiments on radiology report generation.

4.1. Method and Experiment Setup

Our approach. Our approach disentangles predictions based on the image from generations based on

4. Although Cardiomegaly is more common than Consolidation in the training set, there are too few examples in the dev and test set so we exclude it. See Appendix A.
5. Only 12.5% of the reports have the Findings sections (Johnson et al., 2019).

the positive conditions (Figure 1). We first use a ResNet-50 (He et al., 2016) to predict the positive conditions in the image. Then, leveraging the insight from §3, we generate the reports based on the indication and the predicted positive conditions. We consider two approaches in text generation:

- **Pragmatic retrieval**. We first check the predicted labels against the conditions in the indication. For every condition in the indication that is not predicted as positive based on the image, we retrieve a cleaned sentence that mentions that condition as negative. For the predicted positive conditions, we retrieve a report from the training set with the same set of positive conditions and concatenate it with the sentences with negative mentions to form the final report.
- **Pragmatic LLaMA**. We finetune LLaMA-7B (Touvron et al., 2023) to generate clean reports using the predicted positive conditions and the indication as input. We use the same hyperparameters as those of Alpaca (Taori et al., 2023) and train on a sample of 18,264 unique, clean report impressions (10% of the training data[6]) for 3 epochs. Our prompt can be found in Appendix D.

Baselines. We consider the following baselines from existing work.

- **CXR-RePaiR** (Endo et al., 2021): a model that retrieves $k = 2$ report sentences that have the most similar embeddings to that of the image using cosine similarity. The embeddings are learned using CLIP on MIMIC-CXR. CXR-RePaiR is the state of the art on Positive F1 and Positive F1-5.[7]
- **CXR-ReDonE** (Ramesh et al., 2022): a model with a similar training and retrieval strategy as those of CXR-RePaiR, but is trained on CXR-PRO, a clean version of MIMIC-CXR by removing references to prior studies. CXR-ReDonE is the state of the art on RadGraph F1, BERTScore, and Hallucination.[8]
- **MedCLIP** (Wang et al., 2022): similar to CXR-RePaiR, but the image and text embeddings are learned using MedCLIP.
- **XrayGPT** (Thawkar et al., 2023): a model consisting of a vision encoder and a LLM decoder. The representations between the two modalities are aligned using a fully-connected (FC) layer in-between the encoder and decoder. The FC layer is trained using MIMIC-CXR and OpenI data (Demner-Fushman et al., 2016). XrayGPT also cleans prior studies using gpt-3.5-turbo and rules via prompting (Thawkar et al., 2023).

Evaluation. We use both standard and pragmatics-inspired metrics defined in §3.3.

4.2. Performance Comparisons

Pragmatic models outperform previous non-pragmatic methods on all metrics (Table 8). Pragmatic-LLaMA outperforms all the baselines by a substantial margin in both traditional and pragmatic metrics. On Positive F1, our model outperforms CXR-RePaiR by 6.9% in absolute score and 29% relatively. It also surpasses CXR-ReDonE at RadGraph F1 and BERTScore by 8.1% points (+71% relative) and 0.109 points (+43% relative), respectively. On Hallucination, only 15.8% of Pragmatic-LLaMA's reports contain hallucinations, a 69.5% reduction from CXR-ReDonE. With respect to Negative F1-5, Pragmatic Retrieval is the best model, surpassing Pragmatic-LLaMA by 2.9% points. Overall, both of our pragmatic models outperforming non-pragmatic methods in all metrics demonstrates the effectiveness of using the indication as input, not only for negative mention generation, but also for clinical efficacy and mimicking radiologist writing style.

Although Pragmatic-LLaMA trails behind Pragmatic Retrieval in negative mention metrics, it still outperforms the latter by a large margin on metrics that assess language similarity to the groundtruth report, such as RadGraph F1, BLEU-2, BERTScore, and Hallucination. We believe that the slight improvement of Pragmatic Retrieval over Pragmatic-LLaMA in Negative F1 is because the negative mention distribution of the test set is vastly different from that of the training set, as shown in Table 3. Thus, this number may not accurately reflect Pragmatic-LLaMA's negative mention generation performance. We thus evaluate the models on reports with Pneumothorax in the indication, the label with the least discrepancy between the negative proportion in the training and the test set. Indeed, when the test set is in-distribution with respect to the training set, Pragmatic-LLaMA outperforms Pragmatic Retrieval

6. We found that training on more data, e.g., 80%, did not improve performance.
7. We omit a recent work from Google because we do not have access to their model (Tu et al., 2023).
8. Although MedCLIP hallucinates less, our manual inspection of its generated reports shows that it retrieves a small set of sentences for all test examples, which can trivially minimize hallucination, so we discount it.

Model	Correctness			Language		Pragmatic metrics				
	Pos F1	Pos F1-5	Rad-Graph F1	BLEU-2	BERT-Score	Clean BLEU-2	Clean BERTScore	Neg F1	Neg F1-5	Hallucination↓
CXR-RePaiR	0.238	0.368	0.076	0.027	0.162	0.028	0.176	0.016	0.042	0.756
CXR-ReDonE	0.206	0.320	0.113	0.048	0.251	0.050	0.269	0.045	0.102	0.518
MedCLIP	0.122	0.239	0.077	0.023	0.180	0.025	0.199	0.013	0.024	0.260
XrayGPT 0-shot	0.074	0.056	0.013	0.007	0.005	0.007	0.012	0.014	0.028	0.578
Pragmatic Retrieval	0.293	0.403	0.103	0.072	-0.103	0.078	-0.084	**0.077**	**0.156**	0.445
Pragmatic LLaMA	**0.307**	**0.417**	**0.194**	**0.137**	**0.360**	**0.151**	**0.385**	0.050	0.127	**0.158**

Table 8: Report generation performance. Pragmatic methods outperform all previous methods that do not make use of the indication section, in both traditional metrics and our pragmatics-inspired metrics.

Model	Pos F1-5	BERT-Score	Neg F1-5	Hallucination↓
Cleaning+Indication	0.427	0.479	0.099	0.185
Indication Only	0.404	0.464	0.096	0.322
Cleaning Only	0.417	0.464	0.065	0.128

Table 9: Ablation study. We select these four representative metrics for space reasons. The full table of results can be found in Appendix E.

by 7.7 points on Negative F1-5 (+40.3% relative). At the same time, it remains the best model at almost every other metric (see Appendix F).

4.3. Ablation Results

We conduct ablation experiments to identify the effect of 1) incorporating the indication in the input and 2) report cleaning. We denote Pragmatic-LLaMA's modifications by "Cleaning + Indication". We compare it with "Cleaning Only" and "Indication Only". All other variables in training these models are controlled. Table 9 shows the results.

First, comparing Cleaning + Indication with Indication Only shows that not only does cleaning help reduce the number of hallucinations in the output, but it also improves model performance on BERTScore and Positive F1-5. It could be because cleaning helps remove noise from the training data, which simplifies the learning task and helps the model generate cleaner outputs, in turn allowing CheXbert to label the generated reports more correctly.

Second, Cleaning Only achieves lower scores than Cleaning + Indication, suggesting that adding the indication improves model performance, especially on Negative F1. This further shows that the indication can help a model generate negative mentions. We did not expect the model to improve on other metrics, as they are limited by the vision model's ability. But from our inspection of the data set, we found that sometimes the impression will repeat a few words from the indication. Thus, when a model is trained with the indication in the input, it learns to repeat words from the indication, leading to an increase in BERTScore. As for CE metrics, since LLaMA is imperfect, it may fail to report conditions even though they are explicitly given in the prompt. In these cases, the indication can provide extra signal to "remind" the model to include the prompted conditions.

Interestingly, adding the indication increases hallucinations, as that proportion is higher for Cleaning + Indication compared to Cleaning Only. It is likely because some indications refer to previous studies and procedures as context for the current study. The model then refers to this information when it generates the impression. We provide evidence for this claim and discuss it more in Appendix E with a breakdown of the types of hallucination generated and some examples. In short, adding the indication does not make the model generate more recommendations, but it makes the model generate more comparisons, showing that it likely repeats information from the indication. Even so, Indication Only, our fine-tuned model with the most hallucinations, still produces fewer hallucinations than CXR-ReDonE, which retrieves from cleaned data. That is, despite the common perception that language models like LLaMA are prone to hallucinations when generating radiology reports (Ji et al., 2023), LLaMA is more resistant to hallucinations compared to retrieval-based methods.

	MIMIC-CXR			CheXpert		OpenI
	Train	Dev	Test	Train	Test	Train
#Reports	371,951	1,837	2,872	223,414	500	3,955
% No Finding	41.3	40.9	19.8	0.0	0.0	59.7
avg. #positive mentions	1.35	1.17	1.39	2.189	1.894	0.970
avg. #positive mentions in reports that are not "No Finding"	1.59	1.29	1.49	2.190	1.894	0.925
avg. #negative mentions	0.485	0.232	0.255	0.918	1.166	0.330
avg. #negative mentions in reports that are not "No Finding"	0.826	0.394	0.318	0.918	1.166	0.819

Table 10: Dataset statistics related to positive and negative mentions across different datasets.

5. Related Work

We briefly review related work from the pragmatic perspective. Most previous methods have framed the problem as captioning a single image, and focused on evaluating the correctness of positive observations. Vision encoder-language decoder architectures have been shown to generate stylistically accurate reports, but with limited positive mention correctness (Jing et al., 2018; Chen et al., 2020; Boag et al., 2020). In contrast, retrieval-based models sacrifice some coherence in favor of clinical efficacy (Endo et al., 2021; Ramesh et al., 2022). Going beyond single-image captioning, some works have attempted to model *multiple* image views (Yuan et al., 2019; Miura et al., 2021; Lee et al., 2023), which can potentially learn the image view information. Regarding comparisons to prior studies, Ramesh et al. (2022) and Thawkar et al. (2023) notice that such information in groundtruth reports can lead models to hallucinate about nonexistent studies, and opt to remove them from the output. To our knowledge, we are the first work to introduce a unified pragmatic framework and emphasize negative mentions.

6. Concluding Discussion

In this work, we introduce a new, pragmatic perspective on the problem of radiology report generation. We found that radiology reports contain important information beyond positive observations, and focused on generating negative mentions as a first step towards pragmatic report generation. We show that the indication section is critical to reporting negative conditions, and by incorporating it in our models' input, we outperform existing approaches on Negative F1 scores, Hallucination, as well as other standard metrics. We encourage future work to take the new problem formulation and advance modeling approaches to further improve report generation and reduce hallucination.

Following the pragmatic perspective, we found that MIMIC-CXR may not be entirely suitable for training and evaluating models in radiology report generation. Table 10 shows the dataset statistics on positive and negative mentions for two other datasets than MIMIC-CXR: CheXpert and OpenI. CheXpert only exposes chest X-ray images to the user, while OpenI does not have a development or test split. Moreover, CheXpert deliberately limits reports without any finding, while OpenI probably samples the data more similarly to MIMIC-CXR. Due to this peculiarity of CheXpert, we now only refer to reports that are not "No Finding" when discussing these statistics. The average number of positive mentions are somewhat similar between the three datasets, with one to two positive observations per report. In contrast, CheXpert and OpenI have about one negative observation per report, which suggests they are much more similar to MIMIC-CXR's training set than its development or test set. As mentioned above, we believe this is evidence for the development and test set being out-of-distribution, which renders MIMIC-CXR an *inappropriate* benchmark for evaluating the quality of radiology report generation.

We also recognize that our Hallucination metric is a simple heuristic for measuring a complex and major concern regarding the use of language models in healthcare. It likely underestimates the percentage of reports that contain hallucinations. We believe that, going forward, viable reports must not only include the correct observations, but also limit hallucinations. Hence, much more work is needed in developing better metrics for hallucination in generated reports. Since the emphasis of our work is to introduce the pragmatics perspective, we opted to use the simple Hallucination heuristic for model comparison, and leave as future work the development of a more accurate and more clinically relevant metric.

Acknowledgments

We would like to thank Lydia Chelala for her helpful insights about the radiology report writing process, Brent DeVries for his early work on the project, Chenghao Yang, Colin Hudler, and David Reber for technical assistance, Mourad Heddaya, Jiamin Yang, and members of the Chicago Human+AI Lab who have given us valuable input and feedback. This paper is supported by in part by a CDAC discovery grant at the University of Chicago and NSF grants IIS-2040989 and IIS-2126602.

References

William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for chest x-ray report generation. In *Machine learning for health workshop*, pages 126–140. PMLR, 2020. URL http://proceedings.mlr.press/v116/boag20a.

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

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. URL https://academic.oup.com/jamia/article/23/2/304/2572395.

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 *Machine Learning for Health*, pages 209–219. PMLR, 2021. URL https://proceedings.mlr.press/v158/endo21a.html.

Michael P Hartung, Ian C Bickle, Frank Gaillard, and Jeffrey P Kanne. How to create a great radiology report. *RadioGraphics*, 40(6):1658–1670, 2020. URL https://pubs.rsna.org/doi/full/10.1148/rg.2020200020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. URL https://ojs.aaai.org/index.php/AAAI/article/view/3834.

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. URL https://arxiv.org/abs/2106.14463.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

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

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317,

2019. URL https://www.nature.com/articles/s41597-019-0322-0.

Hyungyung Lee, Wonjae Kim, Jin-Hwa Kim, Tackeun Kim, Jihang Kim, Leonard Sunwoo, and Edward Choi. Unified chest x-ray and radiology report generation model with multi-view chest x-rays. *arXiv preprint arXiv:2302.12172*, 2023. URL https://arxiv.org/abs/2302.12172.

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. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019. URL http://proceedings.mlr.press/v106/liu19a.html.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023. URL https://arxiv.org/abs/2301.13688.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online, jun 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.416. URL https://aclanthology.org/2021.naacl-main.416.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. URL https://aclanthology.org/P02-1040/.

Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pages 456–473. PMLR, 2022. URL https://proceedings.mlr.press/v193/ramesh22a.html.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online, nov 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023. URL https://github.com/tatsu-lab/stanford_alpaca.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023. URL https://arxiv.org/abs/2306.07971.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, 2023. URL https://arxiv.org/abs/2307.14334.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Wang_TieNet_Text-Image_Embedding_CVPR_2018_paper.html.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from

unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. URL https://arxiv.org/abs/2210.10163.

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, et al. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*, pages 2022–08, 2022. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10499844/.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 721–729. Springer, 2019. URL https://link.springer.com/chapter/10.1007/978-3-030-32226-7_80.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. URL https://arxiv.org/abs/1904.09675.

Appendix A. Per-label Negative Mention Frequencies

In Table 11, the training set has six conditions that have significantly more negative mentions than others. They are Pneumothorax, Pneumonia, Edema, Pleural Effusion, Cardiomegaly, and Consolidation. This is reflected somewhat in the dev and test set, except Cardiomegaly, which takes up a much smaller portion of those sets compared to in the train set. Because of this reason, as mentioned in Section 3.3, we exclude Cardiomegaly from the Negative F1-5 metric.

Appendix B. Cleaning Details

Table 12 shows our seven cleaning rules and examples of sentences before and after cleaning, and Table 13 shows the prompts that we use for each rule for cleaning. Some information are easy to clean, while others are harder. For instance, communication and recommendations often span an entire sentence, so we already achieve good results by removing the entire sentence. However, as evident from the prompt of rule 3, we found that Flan-T5 sometimes have trouble understanding which sentence constitutes a recommendation, so we provided it with a simple heuristic to, at the very least, remove sentences the contain the string "recommend".

We found that the most difficult information to remove is comparisons to prior studies, because it requires a nuanced understanding of time and how conditions change. On the one hand, there are explicit cues, such as when the radiologist prefaces a finding by saying that he/she is making an observation in comparison with a specific previous study, e.g., "Compared to a previous study on [insert date], [insert finding]". In this case, we employ rule 1 to remove that phrase.

However, what comes after the preface is much more challenging. The overall idea is we want to rewrite any mention of condition progression into present tense: the X-ray either shows or does not show that condition. The first type of progression to consider is when a condition is new or worsened, in which case it should only be reported as present or positive. We use rule 5 to handle that case. The second type is when conditions improve but have not disappeared completely, which means they are still present. This is handled by rule 6. Lastly, when a condition is completely resolved, as it is not present in the X-ray, it should be reported as negative. This

	Negative Mentions			Indication Mentions		
Condition	Train	Dev	Test	Train	Dev	Test
Pneumothorax	37,840	69	124	19,971	50	87
Pneumonia	37,635	152	235	62,881	256	435
Edema	30,110	62	133	36,999	159	228
Pleural Effusion	26,667	25	61	27,117	88	163
Cardiomegaly	18,794	14	9	17,511	77	97
Consolidation	12,614	66	132	20,933	123	111
Enlarged Cardiomediastinum	7,716	11	11	2,946	16	5
Lung Opacity	2,844	4	8	20,586	100	111
Lung Lesion	1,991	7	3	11,543	42	44
Fracture	1,925	17	6	9,545	13	25
Support Devices	1,213	0	9	35,036	107	149
Atelectasis	938	0	2	6,493	17	31
Pleural Other	107	0	0	583	0	2
No Finding	0	0	0	70,489	709	1368

Table 11: Per-label negative mention frequencies in MIMIC-CXR's Train-Dev-Test sets.

is rule 7, which the model struggles with greatly, as it has to rewrite the sentence the most. In the examples of rule 5 and 6, although the rule itself requires a nuanced understanding, in practice, to apply it, often the model only has to remove parts of the sentence. In contrast, in the example of rule 7, it not only has to remove the word "resolved", but it also has to replace it with the word "no". Nevertheless, like with recommendations, it is non-trivial for a language model to understand what constitutes a condition progression, so we also supplied it with certain keywords to help give it signal on which sentence should be modified.

Another nuanced issue is how to apply rule 5 and 6 when the change refers to an organ instead of condition. In contrast to conditions, organs are always "positive". When a radiologist reports "The heart has increased since ___", it would be strange to rewrite it into "The heart" or "The heart is positive" according to a naive application of rule 5. That is why we opted to keep all mentions of changes to organs the same.

The final issue is we use a single rule 4 to clean both X-ray view and prior procedures. In fact, mentions of prior procedures are probably the most difficult pragmatic information for a model to identify, because of the varied semantics of what a procedure is. There is no easy keyword heuristic to rely on either, since "prior" part is often implicit. For example, such a sentence could look like "The patient has received a tube to remove air from their pleural space, and now there is no pneumothorax." There is no easy heuristic to identify the first clause, and in our experience, the model struggles greatly with understanding what constitutes a medical procedure. We only found one phrase that radiologists often use to talk about the state of the patient after a procedure: "status post". An example similar to the one above is: "The patient is status post ET tube removal. No pneumothorax." Since the model has a low success rate on this rule, and there is only one viable keyword heuristic, we opted to combine it with the removal of image view—another simple rule that does not warrant its own prompt—into one prompt to save compute time, as processing the entire MIMIC-CXR dataset using a large language model is very time-consuming even just with a single rule.

Appendix C. Flan-T5 Development and Testing

The development set consists of 100 report sentences, with 20 in each major information category: prior comparisons, recommendations, communication, view and previous procedures, and no change. Our decision to group view and previous procedures is explained in Appendix B. For the test set, we procure 160 report sentences from eight categories: seven categories according to the rules in Table 12, and the eighth category of unchanged sentences. Similar to the development set, each category contains 20 sentences.

ID	Rule	Original	Cleaned
1	Remove comparison to prior studies	In comparison with the study of, there are slightly improved lung volumes.	There are slightly improved lung volumes.
2	Remove communication information	These findings were communicated via the radiology critical results dashboard at 12:57 p.m.	REMOVED
3	Remove doctor recommendations	Recommend advising patient to avoid palpating the area to avoid irritating it.	REMOVED
4	Remove previous treatment and image view	Small lateral pneumothorax is present in this patient status post right first rib resection.	Small lateral pneumothorax is present in this patient
		Lateral view raises concern for pneumonia at the left lung base	Concern for pneumonia at the left lung base
5	Rewrite new/increased conditions into positive	New large right pneumothorax	Large right pneumothorax
		Mild interval increase in loculated right pleural effusion	Loculated right pleural effusion.
6	Rewrite unchanged/partially-improved conditions into positive	Small right pleural effusion probably unchanged since	Small right pleural effusion
		Mild pulmonary edema appears slightly improved	Mild pulmonary edema
7	Rewrite resolved conditions into negative	Resolved opacities in the left mid lung.	No opacities in the left mid lung.

Table 12: Cleaning rules and examples.

Rule 1:
You will be given a sentence from a chest X-ray report. Remove ALL sentences that contain comparisons to the past, and rewrite sentences minimally to preserve meaning. If a sentence contains the word "compare", remove it. If a sentence is empty after cleaning, replace it with the token "REMOVED". If a sentence contains "REMOVED", do not change it.

Rule 2:
You will be given a sentence from a chest X-ray report. Remove ALL sentences that contain information about communication between medical professionals, such as between doctors or nurses. If a sentence is empty after cleaning, replace it with the token "REMOVED". If a sentence contains "REMOVED", do not change it.

Rule 3:
You will be given a sentence from a chest X-ray report. Remove ALL sentences that mention medical recommendations from doctors. Remove sentences that contain "recommend". If a sentence is empty after cleaning, replace it with the token "REMOVED". If a sentence contains "REMOVED", do not change it.

Rule 4:
You will be given a sentence from a chest X-ray report. Remove ALL sentences that mention the chest X-ray view (e.g. AP, PA, lateral) or "status post". Rewrite sentences minimally to preserve meaning. If a sentence is empty after cleaning, replace it with the token "REMOVED". If a sentence is empty or contains "REMOVED", do not change it.

Rule 5:
You will be given a sentence from a chest X-ray report. Remove all instances of "new", "increase", "greater", "worsen", etc. and rewrite the sentence to preserve meaning. If the sentence mentions changes to an organ (e.g. lung, heart), do not rewrite it. If a sentence contains "REMOVED", do not change it.

Rule 6:
You will be given a sentence from a chest X-ray report. If a sentence mentions that a positive medical condition is unchanged or improved (but still positive), remove words related to "unchanged" or "improve" and rewrite the sentence to only say the condition. Otherwise, keep it the same. If a sentence contains "REMOVED", do not change it.

Rule 7:
You will be given a sentence from a chest X-ray report. If the sentence mentions the resolution or disappearance of a condition, rewrite it to simply say the condition is negative. Otherwise, keep the sentence the same. If a sentence is empty or contains "REMOVED", do not change it.

{EXAMPLES}

Original:
{INPUT_QUERY}
New:

Table 13: Prompts for report cleaning. In implementation, the few-shot examples and input query are inserted after every prompt. See Table 12 for examples of how sentences are cleaned.

Table 14 shows the keywords used to compute the percentage of reports containing hallucinations of each type. Specifically, for each type of uninferable information, a report contains it if the report contains any of its corresponding keywords. For prior comparisons, we developed our own keywords, as well as using those identified by Ramesh et al. (2022). Keywords from other types are introduced by us. As mentioned in Appendix B, prior procedures are the most difficult information to identify, and the only consistent common keyword we found was "status" for "status post".

Appendix D. Model Details

Table 15 describes the prompt that we use to train and perform inference with our Pragmatic-LLaMA model. The first two sentences are kept the same from the prompt provided by Taori et al. (2023). We keep the task interpretation open and only ask the model to respond to the indication instead of asking it to only generate negative mentions based on the indication. This likely explains the phenomenon where the model echoes contextual information from the indication, helping it achieve higher Positive F1 and BERTScore as mentioned in Section 4.3.

Another advantage of our Pragmatic-LLaMA model is interpretability, since we decouple the vision and language component. Our use of predicted vision labels to prompt the language model can be seen as using sparse image representations as opposed to dense ones in end-to-end models. This makes it easier to interpret the positive mentions generated by the language model, which is an important quality for clinical models.

Interestingly, using predicted labels as image representation for retrieval helps the Pragmatic Retrieval model achieve higher clinical efficacy than dense representation methods like CXR-RePaiR or MedCLIP. However, in theory, we believe retrieval with dense representations is still more expressive than with sparse representations, since finer-grained information, such as condition severity and location, can be matched between the image and sentence. This applies to generative models like Pragmatic-LLaMA as well, and we leave this investigation for future work.

Appendix E. Full Ablation Results

Table 16 shows the ablation results for Pragmatic-LLaMA. We observe that adding the indication improves negative mention generation and cleaning helps reduce hallucination. While it seems like adding the indication increases hallucination, a breakdown of the types of "hallucination" generated shows that Pragmatic-LLaMA does not generate more recommendations, but it does so for every other types of pragmatic information. When inspecting model-generated reports, we found that sometimes the indication would include results from previous studies, procedures, previous information transmission, and the imaging technique for the current study to provide context for the report. During finetuning, LLaMA learns to copy this information directly into a report, which explains why those types of information are more prevalent while recommendations are not. We provide some examples of this phenomenon in table 17. For instance, in the last example, the model mentions the interval removal of the pigtail chest tube likely because of the phrase "pigtail catheter pulled yesterday" in the indication.

Appendix F. Test Performance on Pneumothorax Examples

Table 18 shows the performance on the subset of reports with indications mentioning "pneumothorax". When the test set is in-distribution with the training set, Pragmatic-LLaMA outperforms Pragmatic Retrieval by 7.7 points on Negative F1-5 (+40.3% relative). At the same time, it remains the best model at almost every other metric.

Information Type	Keywords
Prior Comparisons	compar, interval, new, increas, worse, chang, persist, improv, resol, disappear, prior, stable, previous, again, remain, remov, similar, earlier, decreas, recurr, redemonstrate
Prior procedures	status
Communication	findings, commun, report, convey, relay, enter, submit
Image view	ap, pa, lateral, view
Recommendations	recommend, suggest, should

Table 14: Heuristic keywords used to identify hallucinations in reports

Below is an instruction that describes a task, paired with an input that provides further context.
Write a response that appropriately completes the request.

Instruction:
Write a radiology report responding to the indication. Include all given positive labels.

Input:
Indication: [insert indication]
Positive labels: [insert positive labels in English]

Response:

Table 15: Prompt used for Pragmatic-LLaMA training and inference.

Model	Correctness			Language		Pragmatics		Hallucination (%)				
	Pos F1	Pos F1-5	RG F1	BL-2	BERT-Score	Neg F1	Neg F1-5	Prior study	Prior Proc.	Com.	Rec.	View
Cleaning + Indication	0.318	0.427	0.32	0.274	0.479	0.046	0.099	13.010	0.109	0.490	1.742	1.415
Indication Only	0.319	0.404	0.319	0.267	0.464	0.071	0.096	23.462	0.054	0.544	3.484	1.960
Cleaning Only	0.312	0.417	0.299	0.259	0.464	0.025	0.065	7.240	0.000	0.000	2.177	1.143

Table 16: Pragmatic-LLaMA ablation results.

Indication	Generated report
___-year-old female patient with adenocarcinoma, status post left pleural effusion drain on ___. Study requested for evaluation of new pleural effusion.	Left-sided pleural effusion has *decreased* in size...
Hypoxia and respiratory distress, evaluate for interval changes and consolidation vs. pleural effusion.	*Interval worsening* of multifocal opacities, right greater than left...
___ year old man with significant hypoxa, ?PCP PN___. Evaluate interval change.	No significant *change. Persistent* right upper lobe opacity.
___ year old man with previous pneumothorax; pigtail catheter pulled yesterday // ?pneumothorax	*Interval removal* of the left pigtail chest tube without evidence of pneumothorax...

Table 17: Some examples of hallucinations arising from the indication section. Italics denote keywords that contribute to the sentence being classified as a hallucination.

Model	Correctness			Language metrics		Pragmatic metrics				
	Pos F1	Pos F1-5	RG F1	BL-2	BScore	Clean BL-2	Clean BScore	Neg F1	Neg F1-5	Hallucination
CXR-RePaiR, k=2	0.205	**0.387**	0.093	0.037	0.144	0.036	0.167	0.004	0.010	0.901
CXR-ReDonE, k=2	0.222	0.316	0.097	0.050	0.241	0.059	0.270	0.025	0.066	0.605
MedCLIP	0.133	0.264	0.053	0.012	0.139	0.011	0.160	0.0	0.0	0.111
XrayGPT 0-shot	0.057	0.060	0.011	0.004	-0.007	0.006	0.005	0.010	0.027	0.610
Pragmatic retrieval	0.272	0.392	0.068	0.054	0.129	0.066	0.167	0.074	0.191	0.655
Pragmatic-LLaMA	**0.301**	0.377	**0.168**	**0.103**	**0.327**	**0.128**	**0.370**	**0.103**	**0.268**	**0.287**

Table 18: Pragmatic-LLaMA results on the test set of reports with Pneumothorax in the indication compared with baselines.