Improving Radiology Report Generation Systems by Removing Hallucinated References to Non-existent Priors

Vignav Ramesh* VIGNAVRAMESH@COLLEGE.HARVARD.EDU
Harvard University

Nathan A. Chi* NCHI1@STANFORD.EDU
Stanford University

Pranav Rajpurkar PRANAV_RAJPURKAR@HMS.HARVARD.EDU
Harvard Medical School

Abstract

Current deep learning models trained to generate radiology reports from chest radiographs are capable of producing clinically accurate, clear, and actionable text that can advance patient care. However, such systems all succumb to the same problem: making hallucinated references to non-existent prior reports. Such hallucinations occur because these models are trained on datasets of real-world patient reports that inherently refer to priors. To this end, we propose two methods to remove references to priors in radiology reports: (1) a GPT-3-based few-shot approach to rewrite medical reports without references to priors; and (2) a BioBERT-based token classification approach to directly remove words referring to priors. We use the aforementioned approaches to modify MIMIC-CXR, a publicly available dataset of chest X-rays and their associated free-text radiology reports; we then retrain CXR-RePaiR, a radiology report generation system, on the adapted MIMIC-CXR dataset. We find that our re-trained model—which we call CXR-ReDonE—outperforms previous report generation methods on clinical metrics, achieving an average BERTScore of 0.2351 (2.57% absolute improvement). We expect our approach to be broadly valuable in enabling current radiology report generation systems to be more directly integrated into clinical pipelines.

Keywords: free-text radiology reports, references to priors, generation, retrieval, large language models

1. Introduction

Writing radiology reports is a tedious and labor-intensive process, requiring trained specialists to conduct in-depth analyses of chest radiographs and create detailed reports of their findings. This process is also inherently restricted by a variety of human limitations, including the experience of the radiologist and availability of medical support staff. Therefore, automatically generating free-text radiology reports from chest radiographs has immense clinical value.

Current approaches to generate radiology reports from chest X-rays (CXR-RePaiR, R2Gen, $\mathcal{M}^2$ Trans, etc.) have achieved relative success in producing complete, consistent, and clinically accurate reports (Endo et al., 2021b; Chen et al., 2020b; Miura et al., 2021; Johnson et al., 2019b; Ramirez-Alonso et al., 2022). Nevertheless, these models each have a key limitation: since they are trained on datasets of real-world reports (MIMIC-CXR, Indiana University Chest X-ray Col-

* These authors contributed equally

lection, etc.) which refer to prior reports, their outputted reports often contain references to non-existent priors (Johnson et al., 2019a; Kohli and Rosenman, 2013).

To address this issue, we propose *Contrastive X-Ray Report Determination Employing Prior Reference Removal* (CXR-ReDonE), an improved radiology report generation approach that eliminates nearly all hallucinated references to priors (Figure 1). CXR-ReDonE's unique contribution lies in its novel data preprocessing step: it is trained on CXR-PRO, our adaptation of *MIMIC-CXR with Prior References Omitted*.

Specifically, we investigate two separate approaches to generate CXR-PRO: **(1)** FilBERT+GPT-3, a two-step pipeline to rewrite entire medical reports with prior references removed; and **(2)** GILBERT, a BioBERT model fine-tuned to remove references to priors at the token level (Brown et al., 2020; Lee et al., 2019). While both approaches are viable, we employ GILBERT to create CXR-PRO due to its higher performance, lower cost, and increased scalability.

CXR-ReDonE outperforms existing state-of-the-art report generation methods, achieving an average BERTScore (Zhang et al., 2020) of 0.2351 (2.57% absolute improvement over the highest-performing baseline). We anticipate that our approach will improve performance of current supervised radiology report generation systems—beyond just CXR-RePaiR—on clinical metrics. Given that generated reports without prior references are more aligned with radiologist-created reports, we find that our approach can be broadly valuable in enabling such systems to be more directly integrated into clinical workflows.

2. Related Work

Previous radiology report generation approaches have made use of an assortment of techniques, including applying novel CNNs, RNNs, and LSTMs; fine-tuning image recognition models and Transformers; and developing larger-scale ensemble methods (Monshi et al., 2020; Rubin et al., 2018; Jing et al., 2018; Wang et al., 2018; Alfarghaly et al., 2021; Chen et al., 2020b).

However, all of these models—regardless of their structure—tend to generate incomplete, incorrect, or hallucinatory statements within reports (Table 1). Even state-of-the-art models still generate reports with missing or false references when evaluated by trained radiologists (Alfarghaly et al., 2021; Liu et al., 2019). Such issues stem from those of abstract text generation models in general (Maynez et al., 2020). Miura et al. (2021), for instance, noted that the standard teacher-forcing training algorithm (Williams and Zipser, 1989) employed by several report generation models results in discordance between training and test environments, thereby causing the production of factual hallucinations—referring to non-present conditions while overlooking existing ones.

Hence, more recent works have focused on improving the factual completeness and consistency of radiology report generations. The most popular approach in this regard has been designing evaluation metrics more suited to the domain of radiology reports. Previous report generation models have attained promising results on standard natural language generation (NLG) metrics such as CIDEr (Vedantam et al., 2014), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005); however, such metrics tend to reward superficial textual similarities (i.e., same diction, phrase structure, word order, etc.) rather than semantic or ontological

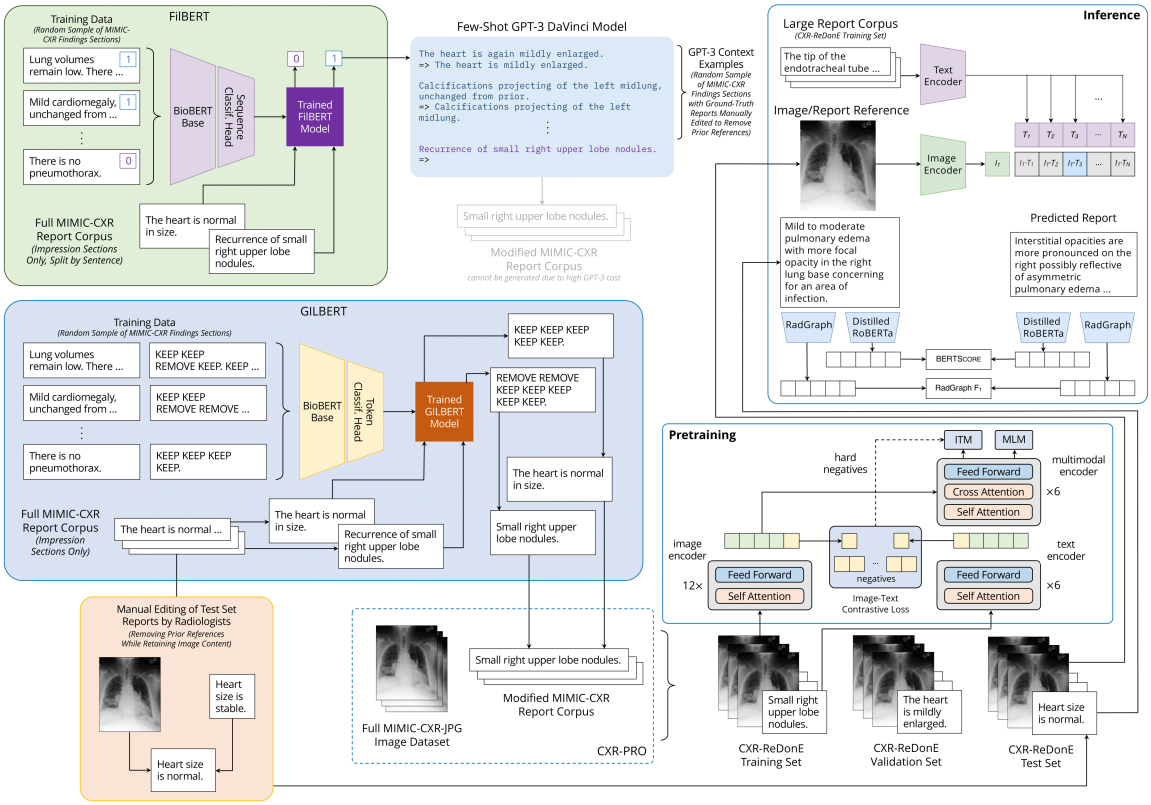

Figure 1: CXR-ReDonE pipeline. We first generate CXR-PRO by passing reports from MIMIC-CXR through GILBERT. It should be noted that we also investigate a secondary pathway to synthesize CXR-PRO—the two-step pipeline FilBERT+GPT-3—but do not employ it due to its decreased accuracy and higher cost. We then train CXR-ReDonE by passing reports and chest X-rays from CXR-PRO through a text encoder and image encoder, respectively. Finally, CXR-ReDonE outputs the generated report with the highest dot-product similarity between the text and image embeddings, and performance metrics are calculated by comparing the ground truth to the predicted reports.

similarities—that is, similarities in the actual diagnosis conveyed by the report (Boag et al., 2020; Chen et al., 2020a; Miura et al., 2021).

To address this issue, a variety of metrics have been posed, both to reward clinical accuracy during training, and to guide model evaluation. For instance, Miura et al. (2021) proposed Exact Entity Match Reward (fact_{ENT}) to measure generated reports' coverage of radiological entities along with Entailing Entity Match Reward ($\text{fact}_{\text{ENTNLI}}$), which rewards inferential consistency by extending fact_{ENT} with a natural language inference (NLI) model; they then optimize these rewards using a reinforcement learning (RL) model. Similarly, Zhang et al. (2020) introduced the BERTSCORE metric for evaluation that computes token similarity using contextual embeddings, thereby capturing semantic similarity (diagnostic similarity in

Table 1: A sample report from MIMIC-CXR, as well as examples of different methods' generated reports. References to priors—true for MIMIC-CXR and hallucinated for CXR-RePaiR, R2Gen, and $\mathcal{M}^2$ Trans—are boxed.

MIMIC-CXR	**CXR-RePaiR**	**R2Gen**	**$\mathcal{M}^2$ Trans**
Frontal and lateral radiographs of the chest demonstrate fractured lower three median sternotomy wires which are unchanged from ___. Compared to the prior radiograph, there is increased in airspace opacity at the right lung base, consistent with pneumonia. The remainder of the lungs is unchanged from the prior radiograph. The cardiac contour is slightly enlarged, unchanged from the prior radiograph. No pleural effusion or pneumothorax is seen.	Cardiomegaly with left pleural effusion and mild edema. The severe enlargement of cardiac silhouette may have progressed the volume of presumed right pleural effusion is impossible to assess on a single frontal view.	There are relatively low lung volumes. There is diffuse increase in interstitial markings bilaterally which may be due to mild interstitial edema versus atypical infection. The cardiac and mediastinal silhouettes are stable.	Heart size remains mildly enlarged. The mediastinal and hilar contours are unchanged. There is mild pulmonary edema. There is no focal consolidation. There is no large pleural effusion or pneumothorax.

the radiology report domain). Other models aiming to improve the quality of generated radiology reports have employed question answering (QA) (Wang et al., 2020a) and content matching constraint (Wang et al., 2020b) approaches, among other techniques; however, such models lack either sufficient performance or generalization capabilities.

Despite these efforts, none of the existing approaches to improve radiology report generation systems' clinical accuracy address the issue of references to non-existent priors—which remain a critical obstacle in generating factually complete and consistent reports. As such, we propose two prior reference removal approaches, FilBERT+GPT-3 and GILBERT, to eliminate these hallucinations. We provide a proof of concept of these models' abilities in the form of retraining CXR-RePaiR, an existing Transformer-based report generation model, on CXR-PRO (Endo et al., 2021a). Rather than structuring the task of report creation as one of text generation, CXR-RePaiR adopts a retrieval approach, which enables it to benefit from the limited space of possible findings and diagnoses in chest radiograph-associated radiology reports. We expect that our proposed methodology has the capacity to improve clinical performance of all existing radiology report generation systems.

3. Data and Implementation

The entirety of our data comes from MIMIC-CXR, a publicly available dataset of chest X-ray images and associated free-text radiology reports, which constitutes 377,110 images taken from 227,835 radiological studies. We make use of a curated set of 226,759 reports (henceforth, MIMIC-CXR will refer to this curated set). CXR-ReDonE is trained on CXR-PRO (created by running GILBERT on all elements of the MIMIC-CXR report corpus) and evaluated on an independent expert-edited evaluation set from MIMIC-CXR (Section 3.3).

3.1. Data Exploration

We begin by examining the space of references to prior reports, given that the possibilities for prior reference expression in ra-

diology reports are finite. Through manual exploration of the data, we determine that prior references in MIMIC-CXR are grammatical variations on any of 18 main keywords (Table 2).

Based on the data, we find that the most common references to priors are by far *change*, *unchanged*, and *prior*. In total, 173,822 reports (76.3%) in MIMIC-CXR contain at least one reference to one of the 18 main keywords. That is, a substantial majority of MIMIC-CXR reports make references to priors. It is worth noting that the keyword *change* does not always refer to prior reports, due to a diversity of possible usages. Most notably, when paired with qualifiers like *emphysematous* or *bony*, *change* refers to a qualitative finding rather than a comparative development from a prior condition. To inform our classification of *change* references, we consulted a physician in order to determine which *change* keywords did and did not refer to priors (Table 8).

3.2. Shared Corpus

To develop FilBERT and GILBERT, we create a *shared corpus* of pairs of original radiology reports and their reworded versions with removed references to priors. Specifically, we manually curate a series of 103 reports that are representative of the full set of prior keywords, then create ground truth reworded reports by removing references to priors (e.g., "No interval change from yesterday. Tubes and lines in adequate position." ⇒ "Tubes and lines in adequate position.").

We use the shared corpus to train and evaluate FilBERT and GILBERT; specifically, we employ a shuffled 80-20 train-test split. A patient that appears in the train set does not appear in the test set. Additionally, by design, the proportion of references to priors per sentence is much higher in the shared corpus than in the MIMIC-CXR dataset as a whole, so that the full set of keywords can be represented in a limited space.

Keyword	**Frequency**	**Relative**
Total	*173822*	*0.763*
Change	105244	0.462
Unchanged	65037	0.285
Prior	56572	0.248
Stable	43340	0.190
Interval	42124	0.185
Previous	34155	0.150
Again	25257	0.111
Increased	25163	0.110
Improve	21941	0.096
Remain	20351	0.089
Worse	17197	0.075
Persistent	12371	0.054
Removal	12068	0.053
Similar	11694	0.051
Earlier	11277	0.049
Decreased	9919	0.044
Recurrence	2872	0.012
Redemonstrate	1099	0.005

Table 2: Frequencies and relative frequencies of a keyword referring to change appearing in a radiology report.

3.3. Report Generation Evaluation Set

In order to test the capabilities of our report generation model CXR-ReDonE, we recruit a team of one board-certified radiologist and two fourth year medical students to create a ground truth test set of reports without references to priors. In particular, we make use of the standardized MIMIC-CXR test set containing 2,188 images and associated reports (Endo et al., 2021a). We provide the medical annotators with the directive to either remove or rewrite references to priors in the test set reports. For instance, "no interval change from prior CT" is a phrase that can be removed completely, while "heart size is

stable" must be changed to a description of the heart's current state (e.g., "heart size is abnormal") rather than simply removed.

4. Methods

We examine two methods to remove references to priors in radiology reports: (1) FilBERT+GPT-3: *rewriting* report sentences flagged as containing references to priors to remove all such references; and (2) GILBERT: directly *removing* tokens referring to priors using a BioBERT token-level classification model. Table 3 contains an example report modified by FilBERT+GPT-3 and GILBERT.

Given that it is less costly to run, we employ GILBERT to create CXR-PRO. We then develop CXR-ReDonE by retraining CXR-RePaiR on CXR-PRO for the task of predicting the impression section of a radiology report from a given chest X-ray.

4.1. FilBERT+GPT-3: A Two-Step Approach to Prior Reference Removal

We use a GPT-3 DaVinci model to rewrite report sentences with removed references to priors. However, running GPT-3 on every sentence of each report would be prohibitively costly (requiring an estimated $92,000 to process the entirety of MIMIC-CXR). Therefore, we propose an additional preprocessing step: FilBERT, a sequence classification model that flags individual sentences as containing references to priors. That is, FilBERT effectively *filters* out all sentences without references to priors, allowing GPT-3 to only rewrite sentences that do. By running GPT-3 in conjunction with FilBERT, we significantly reduce wasted computational and financial effort—resulting in a projected total cost of $16,560 (a more than five-fold reduction).

4.1.1. FilBERT: *Filtering Sentence-Level References to Priors with BioBERT*

Given the impracticality of labeling large radiology report corpora, we investigate finetuning an existing, domain-specific language model to classify whether or not reports include references to priors. To this end, we examine BioBERT, a BERT model pre-trained on a variety of biomedical texts, ranging from medical abstracts to full biomedical papers (Lee et al., 2020). FilBERT contains a BioBERT base architecture with a sequence classification head that is finetuned on the shared corpus for the task of flagging sentences with prior references.

4.1.2. GPT-3 for Rewriting Reports

After identifying the sentences containing references to priors with FilBERT, we feed the flagged data into GPT-3 DaVinci, which generates reworded alternatives for each input sentence. In order to engineer a prompt for GPT-3, we first identified sentences that would be representative of the full list of prior reference keywords and subsequently selected a set of 29 samples as our contextual examples. See Appendix B.1 for a full discussion of our hyperparameter search and prompt development process.

4.2. GILBERT: Generating In-text Labels of References to Priors with BioBERT

Next, we introduce GILBERT, a BioBERT model tasked with classifying tokens as referring to priors or not. Specifically, GILBERT casts the radiology report labeling process as a named entity recognition (NER) task; the model classifies each token in an inputted report as either REMOVE (denoting that the token constitutes a reference to a prior and should be removed from the outputted report) or KEEP (indicating that the token does not constitute a reference to a

Table 3: Example of different methods' reports with removed references to priors compared to the ground truth report. Prior references are color coded to improve readability. See Appendix C for additional examples.

Original	Ground Truth	FilBERT+GPT-3	GILBERT
Comparison made to prior study from ___.		Comparison is made to the prior study.	
There is again seen moderate congestive heart failure with increased vascular cephalization, stable.	There is seen moderate congestive heart failure with vascular cephalization.	Moderate congestive heart failure with increased vascular cephalization.	There is seen moderate congestive heart failure vascular cephalization.
There are large bilateral pleural effusions but decreased since previous. There is cardiomegaly.	There are large bilateral pleural effusions. There is cardiomegaly.	There are large bilateral pleural effusions. There is cardiomegaly.	There are large bilateral pleural effusions. There is cardiomegaly.

comparison prior recurrence again increased stable decreased previous

prior and should be included in the final report). GILBERT uses a modified version of the shared corpus for training, where the reworded report is replaced with a string of KEEP and REMOVE tokens (e.g., "hilar prominence suggestive of pulmonary hypertension, unchanged" ⇒ "KEEP KEEP KEEP KEEP KEEP KEEP REMOVE").

4.2.1. Model Structure

GILBERT contains a BioBERT base architecture, with a token classification head placed on top to allow for predictions at the token level. As BERT relies on wordpiece tokenization (Wu et al., 2016), we modify GILBERT's accompanying tokenizer to label all wordpiece units in the dataset as either KEEP or REMOVE, rather than just the first subunit of a each word.

GILBERT employs the same training process as FilBERT, except that a token classification rather than sequence classification head is fine-tuned.

4.3. CXR-ReDonE

After generating CXR-PRO with GILBERT, we then develop CXR-ReDonE by retraining CXR-RePaiR, a retrieval-based radiology report system (Endo et al., 2021a). However, we investigate one key architectural change in CXR-ReDonE: replacing the CLIP base model with an ALBEF counterpart, given the latter's higher performance on a variety of vision-language downstream tasks (Li et al., 2021).

As in CXR-RePaiR, the problem of generating radiology reports is structured as a retrieval task from report corpus $\mathcal{R} = \{r_1, ..., r_n\}$. Given a chest X-ray x, CXR-ReDonE creates report $\hat{p}$, which is either report $r \in \mathcal{R}$ or composite report s, a combination of k report sentences from the set of all sentences in the retrieval corpus. Note that, for a given report r or sentence s, the base model creates a text embedding for r or s and an image embedding for x, then calculates similarity score $f(r, x) = g(r) \cdot h(x) = T \cdot I$ or $f(s, x) = g(s) \cdot h(x)$; $\hat{p}$ is chosen as the report which maximizes this dot product.

Table 4: Evaluation of CXR-ReDonE method on expert-edited test set, for report-level retrieval and sentence-level retrieval with $k \in \{1, 2, 3\}$, after training on MIMIC-CXR and CXR-PRO. Metrics employed are BERTScore, s_{emb}, and RadGraph F_1. We find that, irrespective of k, our approach outperforms the baseline on all clinical metrics. Here, *italics* denote improvement over the baseline, while **bold** denotes the highest value across the board. Bootstrap confidence intervals were computed using a confidence level of 95%.

k	Training Dataset	Evaluation Metrics		
		BERTScore	s_{emb}	RadGraph F_1
report-level	MIMIC-CXR (Baseline)	0.2083 ± .0023	0.3410 ± .0045	0.0895 ± .0021
	CXR-PRO (Ours)	*0.2160 ± .0025*	*0.3601 ± .0046*	*0.0925 ± .0022*
1	MIMIC-CXR (Baseline)	0.2129 ± .0025	0.3880 ± .0046	0.0838 ± .0023
	CXR-PRO (Ours)	*0.2159 ± .0027*	***0.3967 ± .0048***	*0.0864 ± .0024*
2	MIMIC-CXR (Baseline)	0.2292 ± .0025	0.3822 ± .0045	0.1045 ± .0023
	CXR-PRO (Ours)	***0.2351 ± .0026***	*0.3859 ± .0047*	*0.1056 ± .0024*
3	MIMIC-CXR (Baseline)	0.2179 ± .0025	0.3710 ± .0045	0.1083 ± .0022
	CXR-PRO (Ours)	*0.2254 ± .0025*	*0.3779 ± .0047*	***0.1112 ± .0023***

5. Experiments

5.1. Evaluation Setup

To evaluate the performance of FilBERT+GPT-3 and GILBERT, we calculate the F_1 score using the output of a difference checker algorithm comparing our original, modified (outputted by our models), and ground truth reports. In the context of our study, true positives denote tokens removed from the original report in both the modified and ground truth reports, false positives denote tokens removed in the modified report but not in the ground truth report, and false negatives denote tokens kept in the modified report but removed in the ground truth report.

5.2. FilBERT+GPT-3

On its sentence classification task, FilBERT attains a high accuracy of 0.907 on a held-out test set from the shared corpus (Section 3.2). Moreover, the model's error distribution—more false positives than false negatives—is well-adapted to our task. During evaluation, FilBERT misclassifies only 5.55% of all reports as false positives and 3.70% of reports as false negatives, which is preferable given that sentences incorrectly labeled as containing references to priors should be unaffected when fed into GPT-3. We empirically determine that a discrimination threshold of 0.5 is optimal; shifting the threshold results in either a larger proportion of false negatives or a lower total accuracy.

The combined FilBERT+GPT-3 pipeline attains an F_1 score of 0.56.

5.3. GILBERT

GILBERT achieves an F_1 score of 0.84 on its corresponding held-out test set. This demonstrates that GILBERT's generated reports are notably closer in semantics to the ground truth reports than FilBERT+GPT-

3's. More significantly, a study of all 226,759 reports in the MIMIC-CXR and CXR-PRO datasets shows that the number of references to priors decreases drastically, from 259,376 instances of keywords (Section 3.1) denoting prior references in MIMIC-CXR to 82,074 in CXR-PRO—a >68.3% reduction (Table 9).

5.4. CXR-ReDonE

We use three main metrics to evaluate CXR-ReDonE. Each are semantic-based metrics, given that general NLG metrics like BLEU or CIDEr are inadequate for radiology report evaluation due to their poor performance in judging factual accuracy and consistency (Brown et al., 2020). First, we apply BERTScore, which uses contextual embeddings to measure the semantic similarity between the ground truth report and the generated report. We also employ s_{emb}, which calculates the cosine similarity between the final hidden representations of both the generated and ground truth reports when passed through a CheXbert labeler (Endo et al., 2021a; Smit et al., 2020). Finally, we consider RadGraph F_1, a metric proposed by Yu et al. (2022) that makes use of a RadGraph model to evaluate the overlap in clinical entities included in both the generated and ground truth reports (Jain et al., 2021).

We find that training CXR-ReDonE on CXR-PRO leads to appreciable performance increases, with broad improvements on our expert-edited test set (Section 3.3) for both report-level retrieval and sentence-level retrieval with $k \in \{1, 2, 3\}$, on a variety of evaluation metrics (Table 4). Specifically, for $k = 1$, CXR-ReDonE attains an s_{emb} value of 0.3967 ($\Delta + 2.24\%$ over the highest performing baseline). For $k = 2$, CXR-ReDonE achieves a BERTScore of 0.2351 ($\Delta + 2.57\%$). Finally, $k = 3$ yields CXR-ReDonE's highest RadGraph F_1 score of 0.1112 ($\Delta + 2.68\%$).

We also qualitatively compare the generated reports across training sets (Table 11). Our method does particularly well in generating clinically factual text without hallucinatory references to priors (in the given example, CXR-ReDonE achieves BERTScore = 0.228, s_{emb} = 0.790). While the reports generated by CXR-ReDonE contain different verbiage from that of the ground truth, they are still diagnostically accurate and—most importantly—contain far fewer references to priors than those of the baseline CXR-RePaiR model.

6. Limitations

GILBERT can sometimes generate ungrammatical sentences—most notably, verbless phrases such as "The cardiomediastinal and hilar contours." This is a consequence of GILBERT directly *removing* references to priors, rather than *rewriting* them to adjust for grammatical concerns. As a result, CXR-PRO contains a limited number of ungrammatical sentences. Future work could investigate adding a final Transformer-based grammatical error correction layer to CXR-ReDonE to ensure that reports remain grammatical while removing hallucinated references to priors (Omelianchuk et al., 2020).

7. Conclusion

Automatic report generation systems have considerable potential to streamline radiological pipelines but continue to suffer from hallucinations to non-existent prior reports. FilBERT+GPT-3 and GILBERT both demonstrate a generalized capability to broadly improve the factual completeness and consistency of report generation models by removing prior references within radiology report training corpora. As proof of this approach's efficacy, we find that CXR-ReDonE (trained on the GILBERT-edited

Table 5: Example of different methods' generated reports compared to the ground truth report. Report-level retrieval is employed. See Appendix C for additional examples.

Ground Truth	**Model Trained on MIMIC-CXR**	**Model Trained on CXR-PRO**
A right IJ catheter terminating at the mid right atrium and multiple sternal wires and mediastinal clips are in position.	Various support and monitoring devices removed residual right internal jugular catheter in place and no visible pneumothorax. Cardiomediastinal contours stable. Lungs are remarkable for bibasilar patchy atelectasis the left on the right.	Placement of right internal jugular central venous catheter terminating at the cavoatrial junction without evidence of pneumothorax. Enteric tube grossly courses below the level of the diaphragm, inferior aspect not included on the image.

stable removed residual

CXR-PRO dataset) produces radiology reports with significantly fewer hallucinatory references to priors. Altogether, these methods show promise in improving the automatic generation of clinically accurate and actionable radiology reports.

Code and Data Availability

Our code and pre-trained model weights are made available on GitHub at this link. The CXR-PRO dataset is released on PhysioNet.

Ethics

The benefits of our work include providing model-agnostic methods and data to broadly improve the accuracy of radiology report generation models. As far as we can tell, our project has no significant risks.

Our research was conducted with IRB approval (IRB22-0364, "A clinically based evaluation system for chest radiograph AI-generated reports").

Acknowledgments

The authors would like to thank Dr. Kibo Yoon, Patricia S. Pile, and Pia G. Alfonso for their central role in developing the CXR-PRO test set with prior references manually removed or replaced with clinically accurate statements. The authors would also like to thank Jaehwan Jeong for his advice regarding the technical implementation of CXR-ReDonE, as well as Ethan Chi for providing direction for the development of FilBERT and GILBERT.

References

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.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *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.

William Boag, Tzu-Ming Harry Hsu, Matthew Mcdermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for Chest X-Ray Report Generation. In Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones, editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116 of *Proceedings of Machine Learning Research*, pages 126–140. PMLR, 13 Dec 2020. URL https://proceedings.mlr.press/v116/boag20a.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

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, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112. URL https://aclanthology.org/2020.emnlp-main.112.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020b.

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, 2021a.

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 2021b. URL https://proceedings.mlr.press/v158/endo21a.html.

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.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018. doi: 10.18653/v1/p18-1240. URL http://dx.doi.org/10.18653/v1/P18-1240.

Alistair E. W. 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, Dec 2019a. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019b.

MD Kohli and M Rosenman. Indiana university - chest x-rays (xml reports). 2013. URL https://openi.nlm.nih.gov/faq.php.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746, 2019. URL http://arxiv.org/abs/1901.08746.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

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.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation, 2021.

Maram Mahmoud A Monshi, Josiah Poon, and Vera Chung. Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, 106:101878, 2020.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. Gector–grammatical error correction: tag, not rewrite. *arXiv preprint arXiv:2005.12592*, 2020.

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, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Graciela Ramirez-Alonso, Olanda Prieto-Ordaz, Roberto López-Santillan, and Manuel Montes-Y-Gómez. Medical report generation through radiology images: An overview. *IEEE Latin America Transactions*, 20(6):986–999, 2022. doi: 10.1109/TLA.2022.9757742.

Jonathan Rubin, Deepan Sanghavi, Claire Zhao, Kathy Lee, Ashequl Qadir, and Minnan Xu-Wilson. Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. *arXiv preprint arXiv:1804.07839*, 2018.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014. URL http://arxiv.org/abs/1411.5726.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.450. URL https://aclanthology.org/2020.acl-main.450.

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.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.101. URL https://aclanthology.org/2020.acl-main.101.

Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL http://arxiv.org/abs/1609.08144.

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. *medRxiv*, 2022. doi: 10.1101/2022.08.30.22279318. URL https://www.medrxiv.org/content/early/2022/08/31/2022.08.30.22279318.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

Appendix A. Training Details

A.1. FilBERT

We develop FilBERT using PyTorch and Tensorflow. In particular, we load in the BioBERT weights available under the name `dmis-lab/biobert-base-cased-v1.2` in Huggingface's `transformers` library, then finetune solely the sequence classification head. Additionally, we tokenize our data using the BioBERT-specific tokenizer.

We train FilBERT for 10 epochs, using the Adam optimization algorithm. We use a batch size of 16; we also set the learning rate to 2e-5 and ϵ = 1e-8. We train FilBERT using a single Tesla P100-PCIE-16GB GPU.

A.2. GILBERT

Using PyTorch, we train GILBERT for 10 epochs, each with 100 steps, using a training batch size of 4 and a test batch size of 2. We use a gradient clip norm of 10 and an Adam optimizer with learning rate 1e-5. We train GILBERT using a single Tesla T4 GPU.

A.3. CXR-ReDonE

With the rephrased reports from CXR-PRO, we train CXR-ReDonE over a period of 60 epochs, each with 100 steps. We complete the training cycle using 4 Quadro RTX-8000 GPUs.

Appendix B. Additional Experiments with GPT-3

B.1. GPT-3 Hyperparameter Search: Large Language Models are Sentence-Level Learners

In our few-shot learning approach, we provide context examples to GPT-3 in the prompt. We frame the problem as a text generation task, where the rewritten report (denoted as *Edited medical report to remove references to prior medical reports*) is created based on the original report (denoted as *Original medical report*).

We empirically determine that a temperature of 0.3 yields the highest accuracy. Additionally, we find that labeling an entire radiology report as an *Original medical report* lowers performance. Therefore, we split each report in the prompt into subreports of length $n \in \{1, 2, 3, 4\}$ sentences and investigate the relationship between n and GPT-3 performance. We find that GPT-3 learns best when $n = 1$, that is, when each sentence in the report is fed in as its own subreport (Table 6).

Temperature	n	F_1
0.3	**1**	**0.5569**
0.4	1	0.55526
0.2	1	0.55196
0.0	1	0.55187
0.1	1	0.55098
0.3	2	0.47102
0.3	4	0.44861
0.3	3	0.44684

Table 6: Hyperparameter search for GPT-3 DaVinci. n denotes the number of sentences included in the subreport. Rows are sorted in decreasing order of F_1 score.

B.2. Alternative GPT-3 Models Perform Poorly

We find that alternative GPT-3 variants (Curie, Babbage, and Ada) each perform worse than DaVinci, given their smaller size and lower complexity (Table 7).

Model	F_1
DaVinci	**0.5569**
Curie	0.37263
Ada	0.32522
Babbage	0.23380

Table 7: Performance of all GPT-3 models. All variables besides model type are held constant (temperature = 0.3, $n = 1$).

Appendix C. Supplementary Tables

Sentence	**Prior**
No significant interval **change**.	Yes
No evidence of active **changes** from chronic tuberculosis infection.	Yes
Emphysematous **changes** are identified.	No
Midfoot degenerative **changes**.	No
There are atherosclerotic **changes** of the aorta.	No
Arthritic **changes** of the spine are present.	No
Bony **changes** of renal osteodystrophy are noted.	No
Degenerative **changes** in the spine.	No

Table 8: As stated in the main body of the paper, the keyword *change* sometimes does not refer to prior reports, due to a diversity of possible usages. Here, we outline sample usages of *change* that do and do not serve as references to priors.

Keyword	**MIMIC-CXR**	**CXR-PRO**
Total	*259376*	*82074*
Change	58213	26403
Unchanged	30915	12956
Prior	17526	3025
Stable	23837	9191
Interval	15019	2379
Previous	20271	2921
Again	8171	247
Increased	15019	2379
Improve	18230	3153
Remain	8272	3745
Worse	9651	490
Persistent	12371	2596
Removal	6445	6048
Decreased	5843	1768
Similar	4039	826
Earlier	5082	460
Recurrence	1077	645
Redemonstrate	88	0

Table 9: Number of references to priors, broken down by keyword, in MIMIC-CXR and CXR-PRO. It is clear that CXR-PRO contains far fewer references to priors than MIMIC-CXR in every category—proof of GILBERT's efficacy.

Table 10: Additional examples of different methods' reports with removed references to priors compared to the ground truth report. Prior references are color coded to improve readability.

Original	**Ground Truth**	**FilBERT+GPT-3**	**GILBERT**
Frontal and lateral radiographs of the chest demonstrate persistent large right perihilar mass, which is slightly larger as compared to the prior study.	Frontal and lateral radiographs of the chest demonstrate large right perihilar mass.	Frontal and lateral radiographs of the chest demonstrate large right perihilar mass.	Frontal and lateral radiographs of the chest demonstrate large right perihilar mass.
This is in a region of prior fiducial seed placement, and may correspond to post-radiation changes; however, recurrence of malignancy cannot be excluded.	This is in a region of fiducial seed placement, and may correspond to post-radiation changes; however, malignancy cannot be excluded.	This is in a region of prior fiducial seed placement, and may correspond to post-radiation changes; however, malignancy cannot be excluded.	This is in a region of fiducial seed placement; however, of malignancy cannot be excluded.
Again seen are heterogeneous opacities at the right base, with a small right-sided pleural effusion.	Seen are heterogeneous opacities at the right base, with a small right-sided pleural effusion.	Heterogeneous opacities at the right base, with a small right-sided pleural effusion.	Seen are heterogeneous opacities at the right base, with a small right-sided pleural effusion.
The left lung is essentially clear. The cardiomediastinal and hilar contours are unchanged. There is no pneumothorax or focal consolidation.	The left lung is essentially clear. There is no pneumothorax or focal consolidation.	The left lung is essentially clear. Cardiomediastinal and hilar contours. There is no pneumothorax or focal consolidation.	The left lung is essentially clear. The cardiomediastinal and hilar contours. There is no pneumothorax or focal consolidation.
Right lung opacities have slightly worsened since previous exam and are slightly more confluent, suspicious for an infectious process or aspiration.	Right lung opacities are confluent, suspicious for an infectious process or aspiration.	Right lung opacities are slightly more confluent, suspicious for an infectious process or aspiration.	Right lung opacities are confluent, suspicious for an infectious process or aspiration.
No acute cardiopulmonary process. No significant interval change.	No acute cardiopulmonary process.	No acute cardiopulmonary process. No significant interval change.	No acute cardiopulmonary process.

persistent comparison prior recurrence again worse change unchanged

Table 11: Additional examples of different methods' generated reports compared to the expert-edited ground truth report.

Type	Ground Truth	MIMIC-CXR Model	CXR-PRO Model
Report-level	AP chest: There is substantial engorgement and indistinctness of pulmonary vessels, consistent with the clinical impression of pulmonary edema. There is some indistinctness and engorgement of pulmonary vessels, consistent with the clinical impression of elevated pulmonary venous pressure.	Heart size and mediastinum stable. Minimal interstitial opacities in the lung bases are but it might be related test technique of the chest radiograph. There is no definitive consolidation to suggest infection but attention to the lung bases is recommended. There is no pneumothorax. There is no pleural effusion.	Heart size and mediastinum are overall unremarkable except for mild vascular congestion and minimal interstitial edema. There is no appreciable pleural effusion. There is no pneumothorax. Lung volumes are low.
$k = 1$	AP chest: Extensive opacification in both lungs, sparing the left mid and lower lung zone. AP chest: Severe widespread pulmonary infiltration is significant, with near confluence of opacification in the left lung, and moderate-to-severe left pleural effusion.	Widespread parenchymal consolidations involving the left lung as well as loculated pleural effusion unchanged.	Diffuse bilateral airspace process suggestion of areas of cavitation particularly on the left side and a confluent opacification in the right mid upper lung.
$k = 2$	PA and lateral chest: There is extensive consolidation, predominantly at the base of the right lung in the lower lobe, probably also in the right middle lobe and to a small extent in the mid left lung.	Large scale right lower lobe pneumonia unchanged. Right pleural effusion and right lower lung consolidation for infectious process.	Right upper and lower lobe airspace consolidation obscuring a known right hilar mass may represent hemorrhage or pneumonia. Extensive consolidation.
$k = 3$	The opacification at the right base appears to be significant, raising the possibility of superimposed pneumonia in this patient with hyperexpansion of the lungs consistent with chronic pulmonary disease. There is opacity at right greater than left lung apices.	Asymmetric opacity in the right infrahilar region could reflect an early bronchopneumonia versus atelectasis. Subtle asymmetric increased opacity in the right lower lobe could reflect concurrent pneumonia or aspiration in the appropriate clinical situation, however ___. Opacity in the right infrahilar region may reflect early bronchopneumonia.	Streaky right basilar opacity may be atelectasis although aspiration or infection would be difficult to exclude. Opacity in the right lung base, which could represent atelectasis but cannot exclude pneumonia or aspiration in the right clinical setting. Peribronchial infiltration right lower lobe could be chronic scarring with atelectasis and possible bronchiectasis, but would need ct evaluation for anatomic diagnosis and to establish chronicity.

stable increased unchanged