Proceedings of Machine Learning Research 219:1–18, 2023 Machine Learning for Healthcare

Retrieval Augmented Chest X-Ray Report Generation using OpenAI GPT models

Mercy Ranjit* MERANJIT@MICROSOFT.COM
Department of Computer Science
Bharathidasan University,
Trichy, India

Gopinath Ganapathy GGANAPATHY@BDU.AC.IN
Department of Computer Science
Bharathidasan University,
Trichy, India

Ranjit Manuel† RANJIT.F.MANUEL@GMAIL.COM
AI for Digital Health and Imaging
Indian Institute of Science,
Bengaluru, India

Tanuja Ganu TAGANU@MICROSOFT.COM
Microsoft Research,
Bengaluru, India

Abstract

We propose Retrieval Augmented Generation (RAG) as an approach for automated radiology report writing, using multimodally-aligned embeddings from a contrastively-pretrained vision language model to retrieve relevant radiology text for a given image, and then using a general domain generative model, such as OpenAI `text-davinci-003`, `gpt-3.5-turbo` and `gpt-4`, to generate a report based on the retrieved text. This approach keeps hallucinated generations under check and provides capabilities to generate report content in the format we desire, leveraging the instruction following capabilities of generative models. Our approach achieves better clinical metrics with a BERTScore of 0.2865 (Δ+ 25.88 %) and S_{emb} score of 0.4026 (Δ+ 6.31 %). Our approach can be useful for different clinical settings, as it can augment the automated radiology report generation process with relevant content. It also allows to inject the user intents and requirements in the prompts, which can modulate the content and format of the generated reports according to the clinical setting.

1. Introduction

Automated Radiology Report Generation Systems can improve the report writing workflow of radiologists in various ways. These AI systems can generate free text content or structured report content for review by the radiologists with respect to various attributes of interests like pathology, abnormalities, severity, size or location of findings, progression status etc.

* MR is also associated with Microsoft Research, Bangalore
† RM is also associated with Databricks Inc., Bangalore

Some of the existing work around AI enabled radiology report writing cast the radiology report generation problem as an image captioning problem or generative task [Chen et al. (2020); Miura et al. (2020)]. An interesting recent work CXR-RePaiR [Endo et al. (2021)] cast it as a retrieval problem taking advantage of the finite set of diagnostic details and attributes associated with radiology images. This approach is powerful in that it can leverage a very large database of past and present radiological reports while making impression recommendations. But it suffers from various issues related to irrelevant content in the retrievals. An incremental work CXR-ReDonE by Ramesh et al. (2022) addressed one such issue related to noisy prior references by creating a new dataset CXR-PRO [Ramesh et al.] that eliminated the prior references found in the radiology text reports of MIMIC-CXR [Johnson et al. (2019)] and used this dataset to improve CXR-RePaiR [Endo et al. (2021)] to establish a new SOTA benchmark in radiology report impression generation.

Below are some limitations with retrieval-only approaches, some of which were mentioned for future work in these papers:

- Retrieving only relevant information for report generation is challenging, especially when there are no findings and the number of retrievals, K, is set to more than one.
- The retrieved information may contain unwanted noises, such as references to previous reports, doctor names, user details, etc., that are copied verbatim. The retrievals may also have duplicate content.
- The format of the generated report may not suit the needs of different downstream applications that require structured radiology reports with attributes such as pathologies, positions, severity, size, etc., instead of plain text.
- The generated report may lack coherence and consistency, as the retrievals may mix sentences from different patients' reports. This may also result in contradictory information in the report.

With very capable generative Large Language Models (LLM) like `text-davinci-003`, `gpt-3.5-turbo`(Chat GPT) and `gpt-4` available for the general domain, they can generate relevant content based on instructive prompts in a zero-shot or few-shot setting for a wide variety of downstream tasks. It would be useful to explore how they can be leveraged for the work of radiology report generation to assist the radiologists. These models, however, lack the up-to-date information or domain specific information required in a medical domain setting. Providing the updated and relevant domain-specific content for these large language models will allow them to extend their capabilities to perform tasks with data that they were not exposed to during the training phase. In-addition we can leverage the instructions following capabilities of these models to elicit the required responses we require from these models. This additional context available to the LLMs for generations makes them hallucinate less. We are motivated by the advantages of Retrieval Augmented Generation (RAG) experimented in the work by Lewis et al. (2020) which showed that the generations from RAG are more strongly grounded in real factual knowledge causing less hallucinations and its broad application for various downstream tasks.

We propose CXR-RAG - Retrieval Augmented Generation of Radiology Reports for Chest XRays extending the work of CXR-RePaiR [Endo et al. (2021)] and CXR-ReDonE

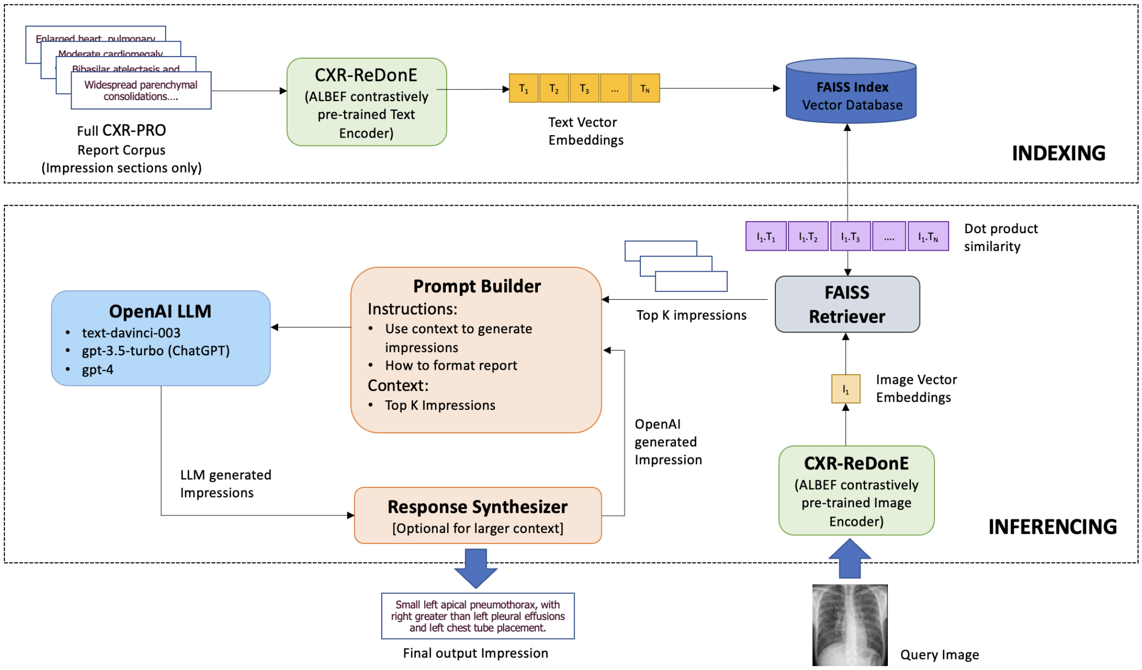

Figure 1: We project all the text embeddings of sentences from radiology impressions using a contrastively pretrained vision-language encoder (CXR-ReDonE) to a vector database index and retrieve the most matching sentences for an input image embedding using the same encoder model. The retrieved impression reports or sentences form the context of the prompt to the LLM along with instructions to generate the impression.

[Ramesh et al. (2022)]. As illustrated in Figure 1, we leverage the contrastively pretrained ALBEF model [Li et al. (2021)] from CXR-ReDonE to generate vision-language aligned embeddings for a database of radiology reports. The same model is used to generate the embedding for an input radiology image. As the image and text embeddings were aligned during the contrastive pre-training, the most relevant text radiology text (reports or sentences) is retrieved for an input x-ray image based on the similarity of the embeddings. A consolidated radiology report impression is generated from the filtered set of records using the OpenAI text-davinci-003, gpt-3.5-turbo and gpt-4 models.

RAG based approach not only makes the radiology report generations grounded on the relevant radiology text retrieved from the radiology text corpus but also allows the user to inject user intents as instructions and few shot examples as part of the generation process via prompt engineering to generate content in the required format as applicable for the clinical setting.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our approach brings the below key insights for ML in healthcare:

- Our approach shows that retrieval augmented generation can combine the benefits of domain-specific healthcare encoder models and general domain generative models, and enhance the clinical metrics.

- We also evaluate the radiology report generations for hallucinations by comparing the LLM generated response with the retrieved radiology text from the radiology reports or sentences corpus. This can help in assessing the feasibility and reliability of these systems in a real clinical setting.
- Our paper also demonstrates how we can use prompt engineering in LLM to incorporate user intents and requirements, and produce radiology reports in different output formats suitable for the downstream application with few-shot learning. The instruction following capabilities of the LLMs are leveraged to remove the noise and incoherent information from the retrievals.

Our approach achieves better clinical metrics with a BERTScore [Yu et al. (2022)] of 0.2865 (Δ+ 25.88%) and S_{emb} score [Endo et al. (2021)] of 0.4026 (Δ+ 6.31%) over the previous state-of-the-art retrieval method CXR-ReDonE. In clinical settings, the improvement of these scores means we are able to generate radiology reports that are closer to the ground truth impression semantically with all the relevant clinical entities, at the same time being very concise reducing the noise from the retrievals. We are also on par with CXR-ReDonE on the RadGraph F_1 metric [Yu et al. (2022)]. This metric helps to measure if we are able to retrieve all the clinical entities accurately.

2. Related Work

Recent works in radiology report generation approached the problem as a generative task like the work of Chen et al. (2020) which used a Transformer decoder architecture R2Gen and the work of Miura et al. (2020) which focused on generating complete, consistent, and clinically accurate reports using a reward-based reinforcement learning approach by name M2 Trans.

Endo et al. (2021) in their work CXR-RePaiR casted radiology report generation problem as a retrieval only task and set a new SOTA benchmark on clinically reliable metrics. The retrieval was based on their constrastively pretrained vision-language model trained using the MIMIC-CXR dataset [Johnson et al. (2019)]. A new clinical efficacy similarity metric called S_{emb} was introduced in the paper to calculate the semantic similarity between the reference report and the predicted report using the last hidden representations from the CheXbert [Smit et al. (2020)] labeler. The paper also used the BERTScore metric [Zhang et al. (2019)] as another measure for semantic similarity.

Ramesh et al. (2022) in their work addressed one key issue pertaining to all automated radiology report generation approaches which are prior report references in the radiology report which impacts the quality of report generation. They built a new dataset CXR-PRO [Ramesh et al.] by addressing this issue on the MIMIC-CXR dataset [Johnson et al. (2019)]. They also retrained CXR-RepaiR using the CXR-PRO dataset and an updated architecture ALBEF [Li et al. (2021)] and set the current SOTA for the radiology report generation task. They used the RadGraph F1 [Yu et al. (2022)] score as an additional metric to measure the completeness and accuracy of the clinical entities available in the predicted report using the RadGraph model [Jain et al. (2021)]

With the rise of LLMs, Retrieval Augmented Generation (RAG) was introduced in the work by Lewis et al. (2020) which brought in some key advantages of leveraging external knowledge sources to augment the knowledge of LLMs to perform a task. LLM generations are also strongly grounded in real factual knowledge which alleviates hallucinations and produces generations that are more factual. The broader impact statement from the paper mentioned its application in a wide variety of scenarios, for example, endowing it with a medical index.

We in this work endow the LLMs with the index of radiology report text and use it as a knowledge base for the LLM to generate a radiology report impression for an input radiology image. We use the constrastively pretrained vision language model from CXR-ReDonE [Ramesh et al. (2022)] for multimodal retrievals. We aim to see if augmented generation on top of these retrievals can further push the report generation benchmark. We also aim to see if we can use the instruction following capabilities of LLMs to modulate the report generation outputs per user requirements. To ensure the applicability of the proposed approach in a real-world clinical setting, we also evaluate the RAG-based approach for the presence of hallucinations in the generated texts.

3. Methods

We propose radiology report generation as a data augmented generation task using large language models like `text-davinci-003`, `gpt-3.5-turbo` and `gpt-4`. Our hypothesis is that it is not required to have domain specific generative models, but domain specific retrievers. We can leverage embeddings from domain specific encoders for the data retrieval task and use the retrieved data for augmenting the generation of general domain generative models. We build on top of the work by CXR-RePaiR [Endo et al. (2021)] and CXR-ReDonE [Ramesh et al. (2022)] , We use the contrastively aligned model ALBEF [Li et al. (2021)] from CXR-ReDonE to generate text embeddings for the radiology report text corpus from the CXR-PRO dataset [Ramesh et al.] and index it in the vector database. We tried both report level corpus $R = r_1, ..., r_n$ and sentence level corpus $S = s_1, ..., s_n$ for the radiology text corpus. We use the same model for generating the image embeddings for the input radiology image x and use it to retrieve the top-K records from the radiology text corpus based on dot-product similarity. The top-K sentences that have the highest similarity to the input image embeddings are selected for augmenting the generation using `text-davinci-003`, `gpt-3.5-turbo` and `gpt-4` models. We generate impression I by prompting the LLM with the top-K sentences or reports retrieved from the sentence corpus S or report corpus R respectively as the context along with instructions Q for the generation.

$$I = LLM(Q, \sum_{i=1}^{k} S_i)$$

Where S_i with i=1 to k denotes the top K sentences from the sentence corpus S which is selected using the function $argmax_{s \in S}, f(s, x)$, f indicating the similarity dot product function between the sentence s and radiology input image x.

In addition to free text radiology report generation, we also hypothesize that it would be also useful to have the radiology reports in a structured format including key attributes of interest from the retrievals. These attributes of interest could be pathologies, severity related to pathology, size, position etc. We use prompt engineering with few shot examples to generate a structured radiology report output.

3.1. Retrieval Corpus

We base the retrieval corpus on the train impressions of the CXR-PRO dataset [Ramesh et al.] which consists of 374,139 free-text radiology reports and their associated chest radiographs. As CXR-PRO is based on MIMIC-CXR which is a de-identified dataset, no protected health information (PHI) is included. CXR-PRO is an adapted version of the MIMIC-CXR dataset [Johnson et al. (2019)] with prior references omitted. It addresses the issue of hallucinated reference to priors produced by radiology report generation models. We use the impressions sections of the radiology reports in the corpus and consider both report-level impressions and as well as the sentences comprising the report-level impressions as the retrieval corpus for report generation.

3.2. Baselines

We consider CXR-ReDonE [Ramesh et al. (2022)] which does retrieval-based report generation with CXR-PRO dataset [Ramesh et al.] as the retrieval corpus as our baseline. We aim to see if retrieval augmented generation on top of these retrievals using LLMs can help improve radiology report generation clinical metrics.

3.3. Prompt Design

We design two sets of prompts to generate the radiology report as free-text report: one for the `text-davinci-003` model and another for report generation in the conversational setup with the `gpt-3.5-turbo` and `gpt-4` models as shown in Table 1 We provide instructions to the LLM to use the retrieved sentences as a context to generate the radiology report. For `gpt-3.5-turbo` and `gpt-4`, the prompt design involves system and user prompts in a conversation setting. The system prompt instructs the system to generate a radiology report impression from the context that the user will send. The user prompt sends the retrieved records as a context requesting the system to provide the radiology report as a response.

3.4. Structured Report Output

We experiment with the ability of the LLM to modulate the report generation output with specifications on the desired report output format in the prompts as few shot examples. It can be interesting for the clinical downstream applications to generate certain attributes of interest from the radiology report apart from generating the free text radiology impression. These attributes could be extracting the pathologies, severity related to pathology, size, or position of findings etc. We instruct the LLM to generate the radiology report in a structured output format containing the impression summary and attributes of interest. We provide specifications on the pathology we are interested in and other attributes of interest along with few shot prompts as shown in the prompt design in Table 2

3.5. Experiments

Using the retrieved records based on the CXR-ReDonE embeddings, we conducted the retrieval augmented generation experiments using the OpenAI LLMs - text-davinci-003, gpt-3.5-turbo and gpt-4. We consider both report-based corpus and sentence-based corpus

Table 1: Prompts for the OpenAI LLMs for Radiology Report Impression Generation from the retrieved reports as context text in the zero-shot setting. The text in italics corresponds to the variables used in formatting the prompt.

text-davinci-003	**gpt-3.5-turbo/gpt-4 System Prompt**	**gpt-3.5-turbo/gpt-4 User Prompt**
Generate an impression summary for the radiology report using the context given. Strictly follow the instructions below while generating the impressions. Instructions: • Impression summary should be based on the information in the context. • Limit the generation to *maxlen* words. CONTEXT: *context* Impression summary:	You are an assistant designed to write impression summaries for the radiology report. Users will send a context text of findings from the radiology image and you will respond with an impression summary using that context. Instructions: • Impression should be based on the findings that the user will send in the context. • The impression should not mention anything about follow-up actions. • Impression should not contain any mentions of prior or previous studies. • Limit the generation to *maxlen* words.	CONTEXT: *context* Impression summary:

Table 2: Prompts for the OpenAI LLMs for Structured Radiology Report Generation from the retrieved reports as context in the Few-Shot setting. The text inside brackets in the prompts corresponds to the variables used while formatting the prompt.

Prompt Design	**Few Shots Example**

Generate an impression summary for the radiology report using the context.
Pathology for impression should be from list of words as in: {pathology}
Positional words should be from list of words as in: {positional_words}
Severity should should be from list of words as in: {severity_words}
Size should should be from list of words as in: {size_words}
CONTEXT: {example_context}
IMPRESSION: {example_report_json}
CONTEXT: {example_context}
IMPRESSION: {example_report_json}
CONTEXT: {example_context}
IMPRESSION: {example_report_json}
CONTEXT: {context}
IMPRESSION:
	CONTEXT:
	Mild bibasilar atelectasis is present. Right suprahilar opacities may relate to pulmonary vascular congestion although infectious process or aspiration not entirely excluded in the appropriate clinical setting.
	
	IMPRESSION:
	{
	"impression": "Mild bibasilar atelectasis is present. Right suprahilar opacities may be due to pulmonary vascular congestion.",
	"attributes": [
	{
	"pathology": "atelectasis",
	"positional": "bibasilar"
	},
	{
	"pathology": "opacities",
	"positional": "Right suprahilar"
	}]
	}

in our experiments. The retrieved records from the corpus forms the context in the prompt based on which the LLM generates the free text radiology impression. We experimented with zero shot settings for free text impression generation and few shot settings for the structured report generation.

4. Results on Real Data

4.1. Evaluation Dataset

We evaluate the performance on the test impressions from the CXR-PRO dataset [Ramesh et al.]. The dataset is created with the help of board-certified radiologists and consists of 2,188 radiology images and associated reports. CXR-PRO dataset was preprocessed to remove duplicate lines.

4.2. Evaluation Approach

We evaluate the free text radiology impressions generated using the LLMs from the retrieved records of the report level corpus and sentence level corpus. For sentence level corpus, we evaluate the impressions generated from top K sentence retrievals with K= 1, 2, 3. Our baselines are the impressions retrieved from CXR-ReDonE. We evaluate on the two semantic metrics – BERTScore [Zhang et al. (2019)] and S_{emb} [Endo et al. (2021)] to measure the similarity of the generation to the ground truth impression. We see this more meaningful as in the medical context phrases like lung collapse can represent atelectasis though the exact word may not be in the sentence. BERTScore computes a similarity score for each token in the predicted impression with each token in the ground truth impression. Token level similarity is computed using contextual embeddings instead of direct token matches. S_{emb} uses CheXbert model [Smit et al. (2020)] to calculate the cosine similarity between the embeddings from the final hidden state representations for the fourteen pathologies. To evaluate the overlap in clinical entities of the generated and ground truth reports we use RadGraph F_1, a metric proposed by Yu et al. (2022) that makes use of RadGraph model [Jain et al. (2021)] to evaluate the overlap in clinical entities.

4.3. Results – CXR-PRO

We use CXR-ReDonE, a purely retrieval-based approach, as our baseline to evaluate the quality of radiology report impressions generated. We find that RAG based generations improves the BERTScore metrics for impressions generated based on both report and sentence corpus retrievals bringing in an absolute improvement of 5.06% at k=3 for sentence-based retrieval. Similarly, it also improves S_{emb} scores for both report and sentence corpus based retrievals, bringing in an absolute improvement of 2.43% at k=3 for sentence-based retrieval. RadGraph F_1 metric that measures the retrievals of clinical entities is almost on par with CXR-ReDonE at k=3 and slightly lower at lower k values. We should note that our approach generates the impressions on the augmented data so we cannot exceed CXR-ReDonE on RadGraph F_1 as otherwise we are hallucinating clinical entities. The evaluation metrics are available in Table 3.

Table 3: Evaluation of CXR-RAG on CXR-PRO test impressions for report corpus retrieval and sentence corpus retrieval. Metrics evaluated are BERTScore, S_{emb} score and RadGraph F_1. Our approach outperforms the baseline on both the clinical metrics BERTScore and S_{emb} score for both the report and sentence corpus-based retrieval for all values of K and at par with CXR-ReDonE for RadGraph F_1 at k=3. Italics denote improvement over the baseline, bold denotes the highest value obtained. We should note that CXR-RAG cannot exceed CXR-ReDonE on the RadGraph F_1 score as otherwise it means we are hallucinating on the clinical entities which are not present in the context.

K	Method		Evaluation Metrics	
		BERTScore	**Semb**	**RadGraph F1**
N/A	CXR-ReDonE	0.2482	0.3647	0.1166
	CXR-RAG (text-davinci-003)	*0.2600*	*0.3741*	0.1060
1	CXR-ReDonE	0.2455	0.4029	0.1079
	CXR-RAG (text-davinci-003)	*0.2610*	*0.4116*	0.0997
2	CXR-ReDonE	0.2465	0.3892	0.1309
	CXR-RAG (text-davinci-003)	*0.2753*	*0.4036*	0.1162
3	CXR-ReDonE	0.2276 ± 0.016	0.3787 ± 0.007	**0.1347** ± 0.003
	CXR-RAG(text-davinci-003)	*0.2782*	*0.4030*	0.1258
	CXR-RAG(gpt-3.5-turbo)	*0.2748*	*0.3973*	0.1219
	CXR-RAG(gpt-4)	***0.2865*** ± 0.011	***0.4026*** ± 0.010	0.1274 ± 0.004

4.4. Qualitative Analysis

We find the retrieval augmented generation based impressions are very concise and less noisy when compared to the outputs from a pure retrieval-based strategy but still retaining all the relevant clinical entities. Refer Table 4 to see the concise impression summary created by the RAG based approach for examples from CXR-PRO. RAG also avoids insignificant, incoherent and contradictory details in the retrieved context.

4.5. Instruction Driven Output

We also evaluate the instructions following capabilities of GPT-4 by passing specific prompts to exclude any mentions of prior report reference and follow-up actions in the generated impression. Refer Table 5 for examples.

- All the 40 occurrences related to follow-up recommendations in the retrieved context documents were not included in the final impression.
- Out of 87 references to previous/prior reports, 80 references were successfully excluded in the impression.

4.6. Modulate Report Generation Format

One of the key advantages of RAG is the ability to modulate output generation format via prompt engineering. Table 6 shows the report generation outputs in a structured json format extracting the attributes of interest using the few shot prompts in Table 2.

Table 4: Examples from CXR-PRO dataset: Generated reports compared to the reference report and the baseline method CXR-RedonE. RAG based generations from CXR-RAG provides very concise impression summaries with ability to retain the important clinical entities. Outputs were generated at top K retrievals set to 3. Note that RAG based generation avoids the incoherent contradictory mentions of severity of pulmonary edema in the first example which is seen in the retrieval only approach CXR-ReDonE.

Ground Truth	**CXR-ReDonE (K=3)**	**CXR-RAG (K=3) (text-davinci-003)**	**CXR-RAG (K=3) (gpt-4)**
PA and lateral chest: Significantly low lung volume exaggerates mild cardiomegaly, but pulmonary vascular engorgement accompanied by mild interstitial edema in the lower lungs has an indication of cardiac decompensation.	Clearing failure.Moderate pulmonary edema from the right upper lung , and residual edema considered as the likely explanation for residual opacification at the right lung base , unless it proves subsequently unresponsive to diuresis.Mild pulmonary edema with moderate cardiomegaly , small right pleural effusion , and cephalization of vasculature.	Mild pulmonary edema with moderate cardiomegaly, small right pleural effusions, and cephalization of vasculature, reflecting cardiac decompensation.	Mild pulmonary edema, moderate cardiomegaly, small right pleural effusion, cephalization of vasculature, and possible infectious process in right perihilar opacities.
AP chest: Lung volumes are lower, exaggerating severe enlargement of the cardiac silhouette and volume of moderate-sized bilateral pleural effusions.There is enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral opacifications consistent with significant layering effusions and compressive basilar atelectasis.	Hazy opacification of the hemithoraces is with bilateral pleural effusions and compressive atelectasis , along with enlargement of the cardiac silhouette and pulmonary vascular congestion. Hazy opacification of the lower lungs suggests bilateral layering pleural effusions with some atelectatic changes at the bases.Hazy opacifications bilaterally are consistent with pleural effusions and compressive atelectasis at the bases.	Bilateral pleural effusions, compressive atelectasis, enlargement of cardiac silhouette, and pulmonary vascular congestion are seen. Mild pulmonary edema and interstitial edema cannot be excluded.	Bilateral hazy opacifications, layering pleural effusion, compressive atelectasis, enlarged cardiac silhouette, pulmonary vascular congestion, bibasilar opacities, and possible mild pulmonary edema.

Table 5: Random examples of cases from CXR-PRO dataset where the RAG based generations from GPT-4 eliminated the references to prior reports and follow-up recommendations based on instruction-based prompts.

Context Records	RAG Impression
Differential appearance include interstitial lung disease with or without superimposed developing pneumonia versusasis associated with elevation of the right hemidiaphragm.Patchy right basilar opacity may reflect compressive atelectasis however infection cannot be excluded and **followup radiographs after diuresis are recommended.** Bibasilar hazy opacities , right greater than left , possibly due to atelectasis given low lung volumes , however underlying infection cannot be excluded in the appropriate clinical setting.	Interstitial lung disease with possible developing pneumonia or atelectasis, patchy right basilar opacity, and bibasilar hazy opacities, right greater than left. Infection cannot be excluded.
It is unusual to have a unilateral left effusion **when the patient has previously had right sided effusions and causes such as pericarditis, pancreatitis or a left upper quandrant process should be considered.** Bilateral opacifications with silhouetting hemidiaphragms is consistent with layering pleural effusions and underlying compressive basilar atelectasis. Bilateral layering effusions are present with patchy airspace opacity predominantly at the bases, left greater than right, which may represent patchy atelectasis.	Unilateral left effusion, bilateral opacifications with silhouetting hemidiaphragms, compressive basilar atelectasis, and patchy airspace opacity predominantly at the bases, left greater than right.

Table 6: Structured Report Generation outputs from CXR-RAG using the few shot prompts in Table 2. We see that we can generate the impression summary and extract pathology, positional information, severity and size related attributes from the retrieved context in a structured format.

Structured Report Output (JSON)

{
"impression":"The Swan-Ganz catheter tip is seen in the proximal right pulmonary artery. Appropriate position of Dobbhoff line reaching stomach. Combination of severe bilateral lower lobe atelectasis and small to moderate pleural effusions"
 	
"findings":[
 	
{
"pathology":"atelectasis",
"positional":"bilateral, base",
"severity":"severe",
"size":""
},
 	
{
"pathology":"pleural effusions",
"positional":"bilateral",
"severity":"",
"size":"small to moderate"
}
 	
] }

4.7. Hallucinations in Retrieval Augmented Generation

We also qualitatively and quantitatively evaluate if the generated report impression hallucinates from the top-K retrievals given to the LLM as the context. We calculate the S_{emb} scores between the generated report impression and the top-K retrievals to measure the clinical embedding similarity which can given an indication if the generation deviated from the retrievals. We find that the average similarity score is 0.8466 and 87% of the impressions have a $S_{emb} > 0.70$ which is a good indication that the generations did not differ from the retrievals. Table 7 in the Appendix section presents a couple of records which had the lowest S_{emb} scores in the test set. We find that the generations did not hallucinate for even such cases and the lower scores may be attributed to the concise impression summary generated.

4.8. Radiologist Evaluation

We also performed a radiologist evaluation on a small sample set of 40 records from the evaluation dataset for the impressions generated by the retrieval only model CXR-ReDonE and our RAG based impressions. The evaluations were done for 7 error types – False Findings, Missed Findings (findings not reported), Incorrect location and severity mentions, Duplicated findings, Incoherent findings which includes contradictions and mixing of irrelevant details from the retrievals, Noisy Details (like mention of doctor name, follow-ups etc.). Refer Figure 2 for metrics. The radiologist also evaluated for hallucinations in the output, if the generations deviated from the retrieved findings. Below is the feedback summary:

- Both the retrieval-only model and RAG based approach have been able to identify most of the clinically relevant findings.
- RAG based model showed significant improvement in reducing the number of incoherent, contradictory findings from the retrieval only approach as the retrievals can include noisy details from different patients.
- RAG based model also had lesser noisy details which speak to the clinical context, history of the patient, follow-up recommendations etc.
- No hallucination noted. The clinical entities in RAG based approach confined itself to the clinical entities from the retrievals.

Scope for Improvement:

- Most of the errors noted involve the findings related to tubes, drains and cardiac devices hardware with varying accuracy of their relative anatomical positions and adequacy. The models could improve in this area.
- Improvements could be made w.r.t model's sensitivity to size and severity of predictions.
- Models seem to confuse findings that can look similar on the scan like pleural plaques vs pulmonary nodules. For radiology report writing workflows, it can be useful to have the findings and impression generation as two discrete steps to allow for feedback prior to impression generation.

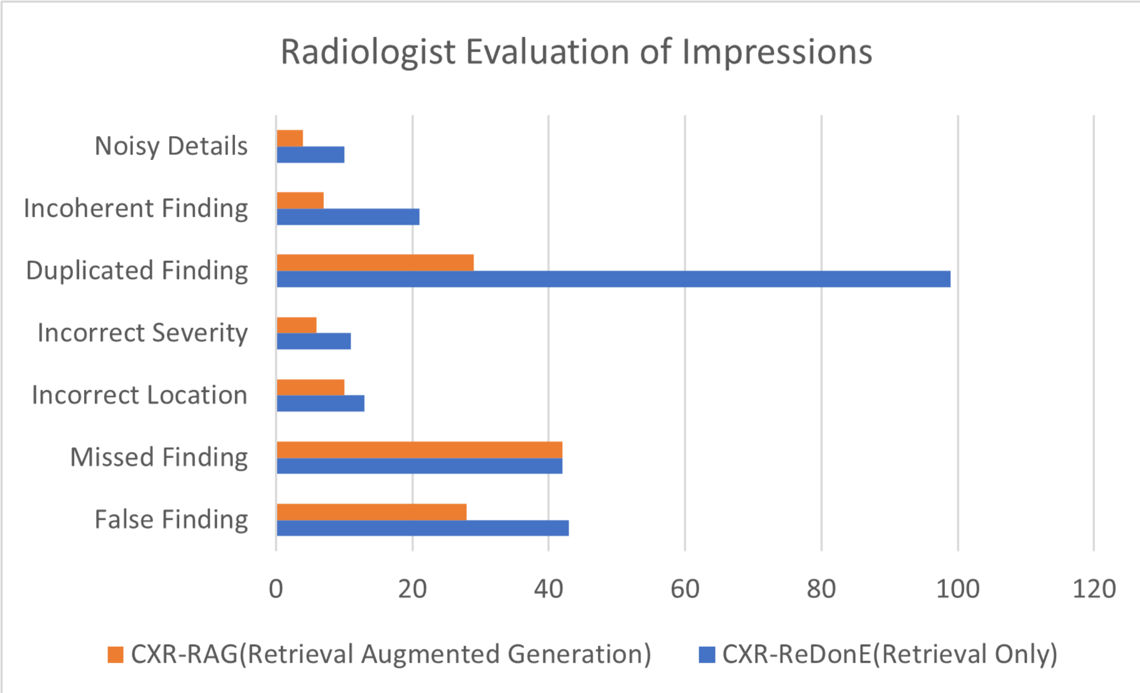

Figure 2: Radiologist evaluation of impressions across seven error categories. RAG based approach significantly reduces the errors in the duplicated, incoherent and noisy details categories.

5. Discussion

In this paper, we present CXR-RAG, a Retrieval Augmented Generation based approach for radiology report impression generation that leverages contrastively pretrained embeddings from CXR-ReDonE [Ramesh et al. (2022)] and large language models from OpenAI. We show that this approach can generate concise and precise impressions that retain the relevant clinical entities and improve the clinical efficacy metrics, particularly the BERTScore and Semb scores. We also show that this approach can be controlled by few-shot prompts and instructions. These can customize the content and format of the impressions. They can also remove incoherent findings that come from retrieving findings from different patient records, noisy text (such as recommendations for further evaluation, prior report mentions, specific patient details etc) and redundant text in the impressions. These are common issues with a retrieval-only setup. We compare this approach with retrieval-only approaches like CXR-RePaiR [Endo et al. (2021)], which suggested the use of a template database to filter out these noisy or duplicate details and to generate better quality reports. We present that RAG-based generation can overcome this limitation by effective prompt engineering and by using techniques like structured content extraction with few-shot prompts. These can extract only the attributes of interest and generate a more concise, precise, and complete impression summary from the retrieved records.

RAG-based generations can ground the clinical entities from the retrievals. This can avoid the problem of hallucinations from generative-only approaches. Therefore, RAG-based Radiology Report Generation combines the benefits of both approaches. It can retrieve and ground all relevant clinical entities based on similarity matches as in retrieval-only approach. It can also leverage the instruction following and zero-shot capabilities of generative models to eliminate unwanted noise and generate report output as per the requirements of a clinical setting.

5.1. Limitations

We note that RAG based impression generation efficacy is based on the retrieved sentences from the corpus which in turn is based on the embeddings from a contrastively pretrained model, meaning that the clinical entities generated by RAG are limited by the clinical entities from the retrieval. So it is imperative that the retrieval model is able to bring in all the relevant clinical entities for the generation.

5.2. Future Work

We see that RAG-based report generations can benefit from more advanced contrastively aligned multimodal models that are more sensitive to fine details of the radiology image, such as severity, size, position of anatomy or pathology, and other attributes of interest from the radiology image. Advances in prompt engineering for medical text are another area to explore, so that we can elicit the LLM more efficiently for a specific downstream task.

Acknowledgments

Many thanks to Dr.Kondaru Vishnu Vardhan Reddy, Senior Registrar, Department of Radio Diagnosis, Narayana Health City, Bangalore for evaluating the impressions generated by the models and providing his valuable feedback.

References

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

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.

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.

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*, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 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.

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. *arXiv preprint arXiv:2010.10042*, 2020.

Vignav Ramesh, Nathan Chi, and Pranav Rajpurkar. Cxr-pro: Mimic-cxr with prior references omitted.

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.

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.

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.

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

6. Appendix

6.1. Hallucination Check - Examples

Table 7: Examples of cases where the Semb scores were less between the retrieved sentences and the generations from the LLM model. Semb uses CheXbert model [Smit et al. (2020)] to calculate the cosine similarity between the embeddings from the final hidden state representations and can give an indication of hallucination. We see there are no hallucinations even for the records with the low cosine similarity.

S_{emb}	**Context Records (K=3)**	**RAG Impression**
0.1508	Lung volumes with right basilar atelectasis versus consolidation which could represent pneumonia in the right clinical context. Low lung volumes with bibasilar opacities which may represent atelectasis versus aspiration. Low lung volumes with bibasilar opacities which could potentially be due to atelectasis.	Right basilar atelectasis and consolidation likely due to pneumonia. Bibasilar opacities may be due to atelectasis or aspiration. Low lung volumes present.
0.2116	The left lung remains grossly clear, although there is minimal linear opacity at the left costophrenic angle, which may represent post - inflammatory scarring or subsegmental atelectasis. There is a suggestion of a tiny left apical pneumothorax which me if there is attempted line placement from the left side prior to the right. There is enlargement of the cardiac silhouette there is suggestion of some mild engorgement of pulmonary vessels on the left, raising the possibility of asymmetric elevation of pulmonary venous pressure.	The left lung is clear with minimal linear opacity at the left costophrenic angle. Tiny left apical pneumothorax and mild engorgement of pulmonary vessels on the left suggest asymmetric elevation of pulmonary venous pressure. Enlargement of cardiac silhouette noted.