Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Analysis of Clinical Finding Retrieval Difficulty in Chest X-ray Retrieval Models

Areesha farid

arishafarid29@gmail.com

Dow University of Health Sciences https://orcid.org/0009-0007-2133-8851

Hifsa Farooq

Dow University of Health Sciences https://orcid.org/0009-0001-9396-9472

Research Article

Keywords: Chest X-rays, Sentence Retrieval, Candidate Pools, Model Evaluation, Lung Opacity

Posted Date: September 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11003136/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

COMPARATIVE ANALYSIS OF CLINICAL FINDING RETRIEVAL DIFFICULTY IN CHEST X-RAY RETRIEVAL MODELS

A PREPRINT

Areesha Farid
Dow University of Health Sciences
arishafarid29@gmail.com

Hifsa Farooq
Dow University of Health Sciences
hifsafarooq.17@gmail.com

September 9, 2026

ABSTRACT

Automated chest X-ray (CXR) report generation increasingly relies on retrieval-based systems that select sentences from fixed candidate pools of clinical language. Because these pools may be unevenly distributed across clinical findings, this study investigated whether candidate-pool composition affects finding-level retrieval performance and cross-model agreement on relative finding difficulty. Three sentence-level image-to-text retrieval models, a proposed model based on the JoImTeRNet framework, CXR-RePaiR, and CXR-ReDonE, were evaluated using finding-level recall, precision, and F1 at K=1, 5, and 10 under two candidate-pool conditions. The original pool contained 129,906 sentences from the MIMIC-CXR findings sections, while a balanced, reduced-size candidate pool capped positive sentences per finding at 450, yielding 6,216 sentences. Cross-model agreement in finding difficulty was assessed using Spearman rank correlation with paired bootstrap testing. Overall recall ranged from 0.09 to 0.31 and precision from 0.15 to 0.29 across models and K values. Lung Opacity and Pleural Effusion were consistently among the most difficult findings to retrieve, whereas Support Devices was consistently among the easiest. Under the original pool, cross-model agreement on finding difficulty was strong (ρ=0.749–0.807) but decreased markedly under the balanced pool (ρ=0.240–0.495), with significant reductions for all model pairs ($\Delta\rho$=0.298–0.487, all p<0.001). These findings indicate that candidate-pool composition substantially influences both finding-level retrieval performance and cross-model agreement, indicating that apparent consensus on retrieval difficulty may partly reflect evaluation-pool composition rather than intrinsic model behavior.

Keywords Chest X-rays, Sentence Retrieval, Candidate Pools, Model Evaluation, Lung Opacity

1 Introduction

Automated chest X-ray report generation has become an important area of research in medical artificial intelligence, driven by the need to reduce the workload associated with interpreting chest radiographs and producing diagnostic reports. Deep learning methods have increasingly been applied to this task, building on their success in related medical imaging applications such as abnormality classification, detection, and segmentation [1]. Most approaches frame report generation as an image captioning problem, using encoder–decoder architectures, sometimes enriched with knowledge graphs or reinforcement-learning objectives, to generate free text directly from an image [2]. More recent designs instead retrieve relevant textual context, such as entities from prior studies, to guide a language model toward a more grounded report [3]. Retrieval can also replace generation entirely: rather than producing new text, a model can select existing sentences from a fixed reference pool of clinical language to describe the image, treating report production as a matching problem over previously validated content [5].

Evaluating the performance of these systems is commonly based on aggregate metrics computed over an entire evaluation set, yet such metrics have been shown to align imperfectly with radiologists' own judgments of report quality [4]. In retrieval-based systems, evaluation is also closely tied to the pool of candidate sentences from which relevant text is selected [5]. The distribution of clinical findings within such a pool may be uneven, with some findings represented

by substantially more sentences than others. Class imbalance in medical imaging has been shown to obscure poor performance on minority classes when results are summarized using aggregate metrics [6], while imbalance between normal and abnormal cases has also been identified as a challenge in radiology report generation [Li et al., 2025].

Despite growing attention to imbalance in medical AI evaluation, the effect of candidate pool composition on chest X-ray sentence retrieval has not been directly examined. Existing work on imbalance related evaluation focuses primarily on classification and report generation whereas retrieval systems depend directly on the distribution of sentences available in the candidate pool. It is therefore unclear whether changes in pool composition affect only the measured performance of individual models or also the relative difficulty assigned to different clinical findings across models. The aim of this study is to determine whether the composition of the sentence retrieval pool affects which clinical findings appear difficult to retrieve and how consistently different chest X-ray retrieval models agree on their relative difficulty. We evaluate three models, including our proposed model, CXR-RePaiR, and CXR-ReDonE, under different retrieval pool compositions.

2 Related work

Chest X-ray image-to-text retrieval treats report generation as a search problem, where relevant clinical text is retrieved from an existing corpus instead of being generated freely [8]. Early retrieval-based approaches primarily focused on selecting complete report sections, such as findings or impressions, as the final output. For example, Jeong et al. proposed X-REM, which retrieves report sections using a multimodal image-text matching score and applies a natural-language-inference filter to remove redundant or contradictory retrieved text [8]. More recent work has explored ways to improve the selection of retrieved content. Zhang et al. introduced a diversity-controlled retrieval strategy, together with a medical lexicon and sentence-level semantic loss, to reduce redundancy and improve the retrieval of clinically important information [9]. Retrieval has also been combined with generative models, where retrieved radiology text is used as additional context for generating a final report. Ranjit et al., for example, used retrieved text as context for GPT-based models to generate report impressions [10]. Beyond report generation, image-to-text and text-to-image retrieval have also been used to evaluate the quality of chest X-ray vision-language representations, with recall-at-K serving as a measure of how effectively images and clinical text are matched [11]. Some studies also report results broken down by individual finding, pathology, or disease category, and this breakdown reveals variation that the aggregate scores do not show on their own. CXR-RePaiR's own per-finding analysis, for instance, reports F1 scores of around 0.6 for pleural effusion compared with about 0.15 for consolidation under the same retrieval setup, with the "no finding" category behaving in the opposite direction from other findings as the number of retrieved sentences increases [5]. KAD reports a comparable spread on the PadChest dataset, reaching an AUC of at least 0.900 on 31 of the unseen classes it evaluates but only 0.700 on 111 of the 177 unseen classes overall [12]. ProtoCLIP shows the same pattern at a smaller scale, with an AUC of 0.96 for pleural effusion against 0.82–0.84 for atelectasis and lung opacity when evaluated as separate target pathologies under the same protocol [13].

Building on this retrieval literature, the performance of chest X-ray vision-language and retrieval systems is commonly reported using aggregate metrics that summarize results across an entire test set or across all target classes at once. CXR-RePaiR, for example, evaluates its retrieval-based report generation using a single macro-averaged F1 score across 14 radiological categories, together with overall semb and BLEU2 scores, to compare against generative baselines [5]. BenchX similarly standardizes evaluation across nine MedVLP methods by reporting one averaged AUROC, mDice, or BLEU4 score per dataset and task, and further condenses these into a single overall ranking per method across all datasets and tasks [14].

In cross-modal retrieval, Chen et al. observed that state-of-the-art image-text retrieval models achieve near-perfect recall on standard MSCOCO and Flickr30K test splits, but that this performance largely reflects the coarse-grained nature of the retrieval pools rather than genuine fine-grained matching ability [15]. When the same models were evaluated on a candidate pool enlarged with visually similar images, or against captions refined to include finer-grained detail, recall dropped substantially across all tested architectures, including X-VLM, UNITER, and non-pretrained baselines such as VSE++ and SCAN [15]. A comparable sensitivity to evaluation-set construction has been reported outside the retrieval setting. Till et al. compared AI performance for pediatric wrist fracture detection on a randomly sampled test set against a test set balanced for case difficulty, and found that precision, recall, and F1 score for both EfficientNet-based classifiers and YOLOv11-based detectors decreased significantly on the balanced set despite no change to model training [16]. In the chest X-ray domain specifically, Fytas et al. showed that the choice of evaluation reference used for pathology labels, whether image-derived or report-derived, alters not only absolute performance estimates but also the relative ranking of supervised classifiers and vision-language models such as MedKLIP, GLoRIA, and ConVIRT, with the degree of disagreement varying by pathology [17]. The same study reported that commonly used image-quality metrics, including SSIM and PSNR, frequently fail to align with expert radiologist judgments of diagnostic usability, further

indicating that the choice of reference against which a model is evaluated can be as consequential as the model itself [17].

Several retrieval studies have compared multiple models under the same task conditions to examine their behaviour beyond aggregate effectiveness scores. Gong et al. compared four visual-semantic embedding networks, VSE++, SCAN, VSRN, and UNITER, on Flickr30K image-to-text retrieval using class- and query-level analyses. Their class-level analysis examined the 25 worst-performing image classes of the two best-performing models, VSRN and UNITER, and identified 16 classes that appeared in both sets [18]. At the query level, the same study compared the descriptions retrieved as relevant by the two models and reported no agreement for 30% of the 2014 image queries, while for 21% of queries one model retrieved a relevant description and the other did not [18]. Extending this type of comparison to neural information retrieval, MacAvaney et al. evaluated lexical, dense-retrieval, and transformer-based ranking models using diagnostic probes that measured responses to document length, term frequency, typographical errors, added content, and factual correctness. Their experiments also included five separately trained instances of the same model using different random seeds, allowing behavioural variation to be examined both across architectures and across model instances [19]. A similar cross-model approach has also been applied to cross-modal retrieval, where Hendriksen et al. evaluated four vision-language models under a shared set of textual perturbations, including word-order changes, lexical substitutions, and typographical noise, across datasets with different levels of granularity [20].

3 Materials and Methods

3.1 Retrieval Models

3.1.1 Our model

We developed a sentence-level image-to-text retrieval model based on the JoImTeRNet framework proposed by Ji et al., which learns a shared embedding space between chest X-rays and radiology reports through global image-sentence matching and local region-word matching. Our implementation retained the global sentence-level matching (TIM) and word-region alignment (WRA) branches, while omitting the original bigram/trigram phrase-region alignment branch. Masked language modeling (MLM) was added as an auxiliary training objective.

The model was trained on MIMIC-CXR using 30,432 radiology reports paired with frontal chest X-rays. The Findings sections were segmented into sentences using spaCy's en_core_web_sm tokenizer, and sentences with fewer than three words were removed, resulting in 221,070 image-sentence pairs. To reduce the influence of frequently repeated normal and templated sentences, each unique sentence was limited to three image-sentence instances. Excess instances were randomly sampled using seed 42, leaving 148,032 pairs after deduplication, representing a 33.0% reduction. Data were split at the image level into training, validation, and test sets using a 90/5/5 ratio, resulting in 133,158, 7,452, and 7,422 image-sentence pairs, respectively. The training set contained 26,388 unique images. Since an image could have multiple associated sentences, a custom sampler selected one sentence per image per epoch to ensure that each image contributed at most one sentence to a training batch.

Chest X-rays were converted to grayscale, resized to 256 $\times$ 256 pixels, and normalized using a mean and standard deviation of 0.5. Training images were augmented with random horizontal flipping and rotations of $\pm 10^{\circ}$. The image encoder was an ImageNet-pretrained ResNet-50, whose first convolutional layer was adapted for single-channel input by summing the pretrained RGB weights. Region features were extracted from the third residual block at 16 $\times$ 16 spatial resolution, processed with soft attention, and projected to 512 dimensions, producing 256 region vectors. Global features were obtained from the fourth residual block using global average pooling and a 512-dimensional projection. Sentences were tokenized using the BERT-base-uncased WordPiece tokenizer and padded or truncated to 128 tokens. The text encoder consisted of three Transformer layers with eight attention heads and a hidden size of 512. It was initialized from a randomly configured BertConfig, meaning that pretrained BERT weights were not used. The [CLS] representation was projected to form the sentence embedding, while token-level representations were projected to 512-dimensional word embeddings for region-word matching.

Image-text alignment was optimized at both sentence and word levels. Sentence-level matching used bidirectional cross-entropy losses with cosine similarities scaled by $\gamma_3 = 10$. Word-region alignment followed the JoImTeRNet WRA formulation with attention and similarity temperatures of $\gamma_1 = 4$ and $\gamma_2 = 5$. Bidirectional triplet losses with a margin of 0.3 were applied at both levels. Hard negatives were mined within each batch using the most similar non-matching sample based on sentence-level cosine similarity. An auxiliary MLM objective was trained jointly with retrieval. Fifteen percent of non-special tokens were selected for masking; 90% were replaced by [MASK], approximately 10% by random vocabulary tokens, and the remainder were left unchanged. The MLM head predicted the original tokens using cross-entropy loss. Training updates were sampled with probabilities of 0.7 for image-text matching and 0.3 for MLM. For image-text matching, sentence- and word-level matching losses were weighted by

2.0, while triplet losses were weighted by 1.0. The image and text encoders were optimized separately using AdamW. Learning rates were 1×10^{-5} for the third and fourth ResNet-50 blocks and 1×10^{-4} for newly initialized image layers and the text encoder. Both optimizers used a weight decay of 1×10^{-4} and cosine annealing over 30 epochs with a minimum learning rate of 1×10^{-7}. Gradient norms were clipped at 1.5. Training used a batch size of 24, with 16 samples for validation, on a single GPU in Kaggle. After each epoch, validation image-to-text Recall@1 was computed, and the checkpoint with the highest Recall@1 was retained.

3.1.2 CXR-RePaiR

CXR-RePaiR (Contrastive X-ray-Report Pair Retrieval) is a retrieval-based chest X-ray report generation method proposed by Endo et al. Rather than generating reports autoregressively, it retrieves text from a predefined corpus based on the similarity between the input X-ray and candidate text. The method uses a CLIP-based image-text encoder to map images and text into a shared embedding space, where similarity is calculated between the corresponding embeddings. For our experiments, we used the publicly released CXR-RePaiR checkpoint clip-imp-pretrained_128_6_after_4.pt from the official repository. This checkpoint corresponds to the pretrained and radiology-adapted configuration reported by Endo et al. We used the model without any additional training or fine-tuning, keeping both the image and text encoders frozen and operating them in evaluation mode.

To ensure consistency with our sentence-level retrieval setting, we used CXR-RePaiR to retrieve individual sentences rather than complete reports. The retrieval corpus was constructed from the findings sections of the MIMIC-CXR reports used for training our model. Sentences were extracted using the same preprocessing procedure described in Section 3.1.1, including spaCy-based sentence segmentation and a minimum length of three words, resulting in 129,906 unique sentences. The released checkpoint's original CLIP preprocessing pipeline was retained. Images were resized and center-cropped to 224×224 pixels, converted from grayscale to three-channel RGB, and normalized using the standard CLIP mean and standard deviation. Text was tokenized using the CLIP tokenizer with a context length of 77 tokens. All sentences in the retrieval corpus were encoded once using the frozen CXR-RePaiR text encoder and L2-normalized. For each input image, the corresponding image embedding was generated and similarly normalized. Sentence-image similarity was then computed using the dot product between the normalized embeddings, and the highest-scoring sentences were selected as the retrieved outputs.

3.1.3 CXR-ReDonE

CXR-ReDonE, proposed by Ramesh, Chi, and Rajpurkar (2022), builds on the CXR-RePaiR retrieval framework by replacing its CLIP-based model with an ALBEF-based vision-language model. The model uses a Vision Transformer as the image encoder to obtain patch-level image representations and a BERT-based transformer as the text encoder to obtain token-level text representations. Global image and text embeddings are obtained from the respective [CLS] representations. CXR-ReDonE also includes a multimodal encoder composed of additional transformer layers with cross-attention, which combines image and text representations for image-text matching. For retrieval, however, the model uses the global image and text embeddings directly, with similarity calculated using their dot product. The highest-scoring report or sentence is selected as the retrieval output, with top-k sentences retrieved when multiple outputs are required. The training objective follows the ALBEF framework and combines image-text contrastive learning, image-text matching, and masked language modeling. The contrastive objective aligns global image and text representations before multimodal fusion, while the image-text matching objective uses the cross-attention encoder together with hard-negative mining. Momentum distillation is also used to provide soft pseudo-targets during contrastive training.

A key difference between CXR-ReDonE and CXR-RePaiR lies in the training data. CXR-ReDonE is trained on MIMIC-PRO, an adapted version of MIMIC-CXR in which references to previous studies, such as "unchanged," "stable," and "compared to prior," are removed at the token level using the GILBERT model. The model is trained using image-impression pairs from this modified corpus. The reported training procedure consists of 60 epochs with 100 steps per epoch using four Quadro RTX-8000 GPUs. For our experiments, we used the pretrained "checkpoint_59.pth" released through the official CXR-ReDonE repository. This checkpoint corresponds to the model trained on MIMIC-PRO with prior-study references removed. It was loaded into the ALBEF retrieval architecture provided in the repository through "models.model_retrieval.ALBEF", together with the supplied "Retrieval_flickr.yaml" and BERT configuration files. The visual encoder position embeddings were interpolated to match the spatial resolution stored in the checkpoint. The released checkpoint was used without further training or fine-tuning. Both the image and text encoders remained frozen throughout inference, with gradients disabled and the model operating in evaluation mode.

Images were resized to 256×256 pixels using bicubic interpolation and converted from grayscale to three-channel input. The checkpoint's original pixel normalization was retained, using per-channel mean and standard deviation values of approximately 101.49 and 83.44 applied directly to the raw 0–255 pixel values. This preprocessing follows

the HDF5-based pipeline used during the model's original training rather than applying conventional normalization after rescaling pixel values to the 0–1 range. Text was tokenized using the "bert-base-uncased" tokenizer and padded or truncated to a maximum length of 64 tokens. For sentence-level retrieval, each sentence in our retrieval corpus was encoded once using the text encoder's [CLS] representation and its corresponding text projection head. The resulting sentence embeddings were used as the retrieval index. Each evaluation image was passed through the visual encoder and projection head to obtain its global [CLS] image representation. Image and sentence embeddings were then compared using cosine similarity, and sentences were ranked according to their similarity to the input image.

3.2 Evaluation Dataset

We used the CheXpert dataset as the evaluation dataset for all retrieval models. Only frontal-view chest radiographs were included, together with their corresponding structured clinical annotations. A random sample of 10,000 frontal-view images was drawn without replacement from the CheXpert training set using a fixed random seed of 42. The resulting dataset retained the class distribution of the original training set. Ground-truth annotations were obtained directly from the structured CheXpert labels, which represent each observation as positive, negative, uncertain, or unmentioned. Unmentioned findings were treated as negative, while uncertain annotations were retained as uncertain rather than assigned to either class.

3.3 Sentence Pool Construction

The retrieval sentence pool was constructed from the findings sections of the MIMIC-CXR reports used for training our model. Sentences were extracted using the same spaCy-based procedure described in Section 3.1.1, with a minimum length of three words and no deduplication cap applied at this stage. This resulted in 129,906 unique retrieval pool sentences. Each sentence was labeled using the CheXbert labeler based on the 14 CheXpert clinical findings. The resulting labels were used to characterize the distribution of findings within the sentence pool and to construct a balanced variant. For the balanced pool, the number of positive sentences for each finding was capped at 450. Pleural Effusion had the lowest number of available positive sentences, with 530 sentences in the full pool. For findings with more than 450 positive sentences, 450 were randomly selected using seed 42, while all available sentences were retained for findings below the cap. Since individual sentences could contain multiple findings, the sampled sentences were combined across findings and duplicate sentences were removed. This produced a final balanced pool of 6,216 unique sentences. The balanced pool differs from the original pool in both class balance and total size (129,906 vs. 6,216 sentences) however this design does not isolate the contribution of each factor to the observed effects.

4 Results

This study used recall, precision, and F1 as the primary comparison metrics, evaluated at K=1, 5, and 10 under both the original and balanced retrieval pools (Table 1). Across both retrieval-pool settings, a consistent pattern was observed as K increased, with recall and F1 generally increasing, whereas precision decreased across all three models. Under the original retrieval pool, the highest recall, precision, and F1 values were obtained by our model at each value of K, followed by CXR-ReDonE and CXR-RePaiR. Recall values ranged from approximately 0.09 to 0.31 across the models and K values, while precision ranged from approximately 0.15 to 0.29. Under the balanced retrieval pool, the same overall trend across K was observed for all three models. Higher recall values were obtained at K=5 and K=10 compared with the original pool for every model, whereas lower precision values were observed at all K values. For F1, lower values were obtained at K=5 under the balanced pool for all three models, while higher values were obtained at K=10 compared with the original pool.

Table 1: Overall retrieval performance (recall, precision, F1) at K=1, 5, and 10, under the original and balanced candidate pools.

		Performance at K=1			Performance at K=5			Performance at K=10		
Model	Pool	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Our model	Original	0.1485	0.2938	0.1973	0.2594	0.2333	0.2456	0.3055	0.2099	0.2488
Our model	Balanced	0.1215	0.2018	0.1517	0.3030	0.1900	0.2335	0.4035	0.1843	0.2530
CXR-RePaiR	Original	0.0890	0.1496	0.1116	0.2191	0.1660	0.1889	0.2908	0.1694	0.2141
CXR-RePaiR	Balanced	0.1049	0.1388	0.1195	0.2543	0.1501	0.1888	0.3464	0.1574	0.2164
CXR-ReDonE	Original	0.1364	0.2894	0.1854	0.2429	0.2382	0.2405	0.2832	0.2160	0.2450
CXR-ReDonE	Balanced	0.1233	0.2084	0.1549	0.2817	0.1897	0.2267	0.3685	0.1797	0.2416

Recall for each of the 14 clinical findings under the original retrieval pool is reported in Table 2, with substantial variation observed between findings across all three models. Recall ranged from approximately 0.023 to 0.867 for our model, 0.028 to 0.487 for CXR-RePaiR, and 0.007 to 0.819 for CXR-ReDonE. Pleural Effusion and Lung Opacity were consistently among the findings with the lowest recall across the three models, whereas Support Devices and Fracture were generally associated with the highest recall values. In the case of our model, recall for Pleural Effusion was 0.023 [0.018–0.027] and 0.033 [0.028–0.038] for Lung Opacity, compared with 0.867 [0.858–0.875] for Support Devices, with limited overlap between the confidence intervals of several lower- and higher-recall findings. The distribution of the 129,906 candidate-pool sentences across the 14 findings is shown in Figure 1, which illustrates the unevenness of the original pool: Support Devices accounted for the largest share (39.0%), followed by Fracture (18.6%) and Cardiomegaly (17.3%), while several findings each accounted for under 2% of the pool, including Consolidation (2.0%), Edema (1.9%), Atelectasis (1.8%), Lung Opacity (1.3%), Pleural Other (1.2%), and Pleural Effusion (0.4%), the smallest share of any finding.

Table 2: Finding-level recall under the original candidate pool for all three models. Values are recall [95% CI].

Clinical Finding	Our Model	CXR-RePaiR	CXR-ReDonE
Support Devices	0.867 [0.858–0.875]	0.449 [0.436–0.461]	0.819 [0.809–0.830]
Fracture	0.444 [0.395–0.496]	0.487 [0.441–0.537]	0.395 [0.348–0.442]
Cardiomegaly	0.231 [0.207–0.256]	0.433 [0.406–0.463]	0.193 [0.170–0.217]
Pneumothorax	0.184 [0.159–0.210]	0.412 [0.381–0.445]	0.198 [0.172–0.223]
Pneumonia	0.113 [0.075–0.155]	0.327 [0.268–0.387]	0.145 [0.102–0.190]
Lung Lesion	0.115 [0.084–0.146]	0.158 [0.124–0.192]	0.335 [0.289–0.380]
Enlarged Cardiomediastinum	0.121 [0.092–0.150]	0.238 [0.202–0.274]	0.105 [0.080–0.133]
No Finding	0.130 [0.108–0.151]	0.244 [0.214–0.272]	0.073 [0.056–0.090]
Consolidation	0.045 [0.030–0.060]	0.143 [0.117–0.169]	0.123 [0.098–0.148]
Edema	0.047 [0.039–0.055]	0.143 [0.129–0.157]	0.071 [0.061–0.081]
Atelectasis	0.071 [0.059–0.084]	0.297 [0.274–0.320]	0.019 [0.013–0.026]
Lung Opacity	0.033 [0.028–0.038]	0.032 [0.027–0.037]	0.018 [0.014–0.022]
Pleural Other	0.158 [0.096–0.228]	0.079 [0.033–0.132]	0.044 [0.009–0.086]
Pleural Effusion	0.023 [0.018–0.027]	0.028 [0.023–0.033]	0.007 [0.004–0.009]

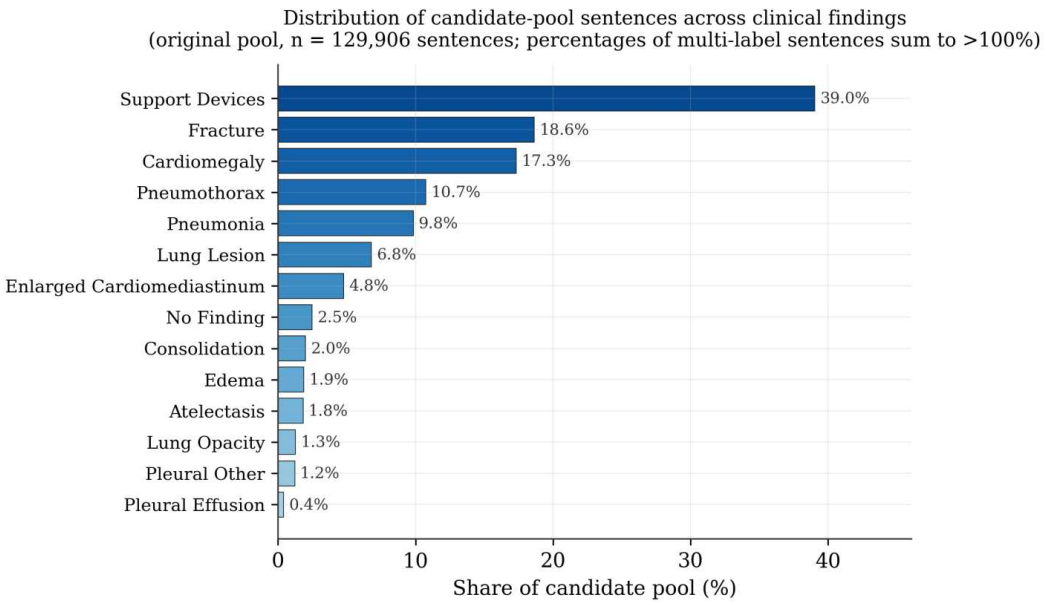


Figure 1: Distribution of candidate-pool sentences across clinical findings (original pool, n = 129,906 sentences; percentages of multi-label sentences sum to >100%).

Table 3 reports the change in recall for each finding between the original and balanced pools, for all three models. Recall for Support Devices declined significantly in all models, with changes of -0.198 [-0.211–-0.186] for our model, -0.199 [-0.211–-0.186] for CXR-RePaiR, and -0.131 [-0.143–-0.119] for CXR-ReDonE (all p<0.001).

Recall increased significantly for most other findings, with gains reaching statistical significance for 12 of 14 findings in our model, 8 of 14 in CXR-RePaiR, and 13 of 14 in CXR-ReDonE (Table 3). The magnitude of change varied by finding and model; for Pleural Effusion, recall increased by 0.141 [0.131–0.151] for our model, 0.046 [0.038–0.055] for CXR-RePaiR, and 0.059 [0.052–0.066] for CXR-ReDonE, all significant. A comparable pattern was observed for Lung Opacity (0.082 [0.074–0.090], 0.076 [0.067–0.084], and 0.073 [0.066–0.080], respectively, all significant). The change for Fracture was not significant for our model (-0.036 [-0.086–0.015], p=0.181), and CXR-RePaiR showed no significant change for Pneumothorax, Fracture, Enlarged Cardiomediastinum, Pneumonia, or Pleural Other (p>0.05 for each). Recall remained heterogeneous across findings under the balanced pool for all three models. Lung Opacity continued to rank among the three lowest-recall findings (0.115 [0.107–0.124], 0.108 [0.099–0.116], and 0.091 [0.083–0.099] for our model, CXR-RePaiR, and CXR-ReDonE, respectively). Pleural Effusion remained among the five lowest-recall findings, ranking lowest for CXR-RePaiR (0.074 [0.066–0.082]) and CXR-ReDonE (0.065 [0.058–0.073]), and fourth-lowest for our model (0.164 [0.152–0.175]). Support Devices and Fracture remained among the highest-recall findings across all three models.

Table 3: Change in finding-level recall (Δ recall [95% CI]) between the original and balanced candidate pools. p<0.05, p<0.01, p<0.001 (Benjamini-Hochberg adjusted).

Clinical Finding	Our Model	CXR-RePaiR	CXR-ReDonE
Support Devices	−0.198 [−0.211–−0.186]	−0.199 [−0.211–−0.186]	−0.131 [−0.143–−0.119]
Fracture	−0.036 [−0.086–+0.015]	+0.023 [−0.030–+0.075]	+0.053 [+0.007–+0.103]
Cardiomegaly	+0.123 [+0.094–+0.152]	+0.112 [+0.083–+0.143]	+0.058 [+0.035–+0.083]
Pneumothorax	+0.117 [+0.090–+0.144]	+0.003 [−0.027–+0.036]	+0.107 [+0.080–+0.132]
Pneumonia	+0.085 [+0.036–+0.134]	+0.057 [−0.004–+0.115]	+0.068 [+0.020–+0.117]
Lung Lesion	+0.064 [+0.025–+0.104]	+0.104 [+0.069–+0.142]	+0.102 [+0.058–+0.145]
Enlarged Cardiomediastinum	+0.162 [+0.124–+0.200]	+0.025 [−0.008–+0.057]	+0.146 [+0.111–+0.185]
No Finding	+0.167 [+0.132–+0.200]	+0.354 [+0.319–+0.390]	+0.223 [+0.195–+0.250]
Consolidation	+0.071 [+0.047–+0.097]	+0.115 [+0.084–+0.146]	+0.188 [+0.158–+0.219]
Edema	+0.105 [+0.091–+0.120]	+0.185 [+0.168–+0.203]	+0.120 [+0.106–+0.135]
Atelectasis	+0.206 [+0.185–+0.228]	+0.193 [+0.172–+0.214]	+0.055 [+0.043–+0.067]
Lung Opacity	+0.082 [+0.074–+0.090]	+0.076 [+0.067–+0.084]	+0.073 [+0.066–+0.080]
Pleural Other	+0.193 [+0.112–+0.275]	+0.061 [+0.000–+0.123]	+0.166 [+0.101–+0.238]
Pleural Effusion	+0.141 [+0.131–+0.151]	+0.046 [+0.038–+0.055]	+0.059 [+0.052–+0.066]

Table 4 reports Spearman rank correlations of per-label recall between each model pair, under both pools, along with a paired bootstrap test of the difference between pools. Under the original pool, correlations were strong and statistically significant for all three pairs (ρ=0.749–0.807, 95% CI excluding zero in each case, permutation p$\leq$0.003). Under the balanced pool, correlations were markedly weaker for all three pairs (ρ=0.240–0.495), and their 95% CIs were wide enough to include values consistent with no association (permutation p=0.07–0.41, none reaching significance against zero). The paired bootstrap (10,000 evaluation images resampled with replacement, same resample applied to both pools within each of 3,000 iterations, Benjamini-Hochberg corrected across the three pairs) confirmed that the drop in ρ between pools was itself statistically significant for all three pairs ($\Delta\rho$=0.298–0.487, all 95% CIs excluding zero, all p<0.001), indicating that cross-model agreement in per-finding difficulty was reliably weaker under the balanced pool rather than reflecting sampling noise. This drop was largest between the two baseline architectures (CXR-RePaiR vs CXR-ReDonE: $\Delta\rho$=0.487), showing that the apparent agreement between independently-trained models, not only between our model and the baselines, was substantially attributable to shared candidate-pool composition.

Table 4: Cross-model agreement in finding-level recall (Spearman ρ [95% CI]) under the original and balanced candidate pools, and the paired bootstrap difference ($\Delta\rho$) between pools. All CIs computed via image-level bootstrap (resampling the 10,000 evaluation images), consistent across both the individual-pool ρ estimates and the paired drop test.

Model pair	ρ Original [95% CI]	ρ Balanced [95% CI]	$\Delta\rho$ [95% CI]	p (BH-adjusted)
Our model vs CXR-RePaiR	0.807 [0.723, 0.899]	0.424 [0.257, 0.565]	0.403 [0.242, 0.554]	<0.001
Our model vs CXR-ReDonE	0.749 [0.670, 0.855]	0.495 [0.310, 0.587]	0.298 [0.154, 0.453]	<0.001
CXR-RePaiR vs CXR-ReDonE	0.780 [0.670, 0.802]	0.240 [0.116, 0.385]	0.487 [0.352, 0.633]	<0.001

Overall, retrieval performance varied substantially across clinical findings and models. Recall rankings were strongly and significantly correlated under the original retrieval pool but became weaker and non-significant after balancing, and this drop in cross-model agreement was itself statistically significant for every model pair. Lung Opacity and Pleural Effusion consistently showed low recall, whereas Support Devices and Fracture generally showed higher recall across models and pool settings.

5 Discussion

This study investigated how sentence retrieval pool composition affects model performance and difficulty rankings across clinical findings. We observed considerable variation in recall across all three models, with Lung Opacity and Pleural Effusion consistently showing the lowest recall under both pool conditions. Furthermore, paired bootstrap testing confirmed that strong cross-model agreement under the original pool dropped significantly under the balanced pool. Together, these findings demonstrate that evaluation pool design alters performance metrics, suggesting that current model evaluation standards may introduce systematic variation in evaluation.

The low recall observed for Lung Opacity and Pleural Effusion across all three models under the original pool, despite their differing prevalence in that pool (1.3% and 0.4% of sentences, respectively), provides important context for interpreting the results following rebalancing. Lung Opacity remained among the three lowest-recall findings for all three models even after positive sentences for each finding were capped at the same level. Pleural Effusion exhibited a more heterogeneous pattern, remaining the lowest- or second-lowest-recall finding for CXR-RePaiR and CXR-ReDonE but ranking fourth-lowest for our model, where it demonstrated the largest increase in recall. These findings suggest that the consistently low recall for Lung Opacity was not fully attributable to differences in sentence prevalence, whereas the retrieval pattern for Pleural Effusion varied across models. Furthermore, the significant reduction in cross-model rank correlation under the balanced pool provides a related perspective on model evaluation. Under the original pool, all three models demonstrated strong and statistically significant concordance in their rankings of clinical findings according to recall. This concordance was substantially attenuated under the balanced pool, and paired bootstrap testing indicated that the observed differences between pool conditions were unlikely to be attributable to sampling variation alone. These results indicate that the degree of cross-model agreement was contingent on the evaluation pool condition.

The persistent difficulty in retrieving Lung Opacity after balancing, with recall remaining among the three lowest of all 14 findings for each of three model suggests that its low recall was not explained solely by its low representation in the original retrieval pool. One possible explanation is that Lung Opacity is a broad radiographic finding rather than a specific disease, and the Fleischner Society defines opacity as increased pulmonary attenuation without specifying the underlying pathological process [23]. This may lead to variation in how the finding appears and is described, making image-to-sentence matching more challenging, although this explanation remains speculative and was not directly tested in this study. Pleural Effusion also remained difficult after balancing, ranking lowest or second-lowest for CXR-RePaiR (0.074) and CXR-ReDonE (0.065). Previous evidence reported sensitivity as low as 67% for detecting pleural effusion on supine chest radiographs, even among expert readers [24]. This human detection difficulty may be relevant context, but it does not establish that the same factors underlie the models' retrieval difficulty. Its improvement differed across models, with recall increasing from 0.023 to 0.164 for our model, compared with 0.028 to 0.074 for CXR-RePaiR and 0.007 to 0.065 for CXR-ReDonE, suggesting that its retrieval performance was also influenced by model-specific and candidate-pool factors. Our findings on cross-model agreement are also consistent with previous evidence showing that comparative model evaluation can depend on the composition of the evaluation data. Kovatchev and Lease showed across 135 models that changes in data distribution can affect both absolute performance and relative model rankings [25]. Similarly, Bassi et al. reported that algorithm rankings in medical imaging can become less stable when test sets differ in size or composition [26]. Our findings extend this evidence to sentence-level medical image retrieval by showing that changing the composition of the candidate pool significantly changed how consistently the models agreed on which clinical findings were difficult to retrieve. The different effects of balancing across findings, including the decline in Support Devices recall and improvements for most other findings, are also consistent with evidence that class imbalance can affect categories differently rather than influencing all classes in the same way [27]. This is also relevant to the CXR-RePaiR framework, where retrieval performance depends on the size and diversity of the available reference corpus [5].

The effects of balancing were also not uniform across individual findings or model architectures. In our model, Fracture showed no significant change in recall unlike the significant shifts observed for most other findings. Similarly, CXR-RePaiR showed no significant change for several findings, including Pneumothorax, Fracture, Enlarged Cardiomediastinum, Pneumonia, and Pleural Other. These heterogeneous responses indicate that balancing affected findings and models differently, which may help explain the changes in cross-model agreement.

5.1 Limitations

This study has several limitations that should be considered when interpreting the results. First, the balanced pool differed from the original pool in class balance, total size, and which candidate sentences were included, therefore, the observed effects cannot be attributed solely to balancing, as pool size and the specific change in candidate membership may also have contributed. A random equal-size subsample would be needed to disentangle these effects. Second, the candidate pool was constructed from MIMIC-CXR text, which more closely matched the training distribution of our

model than that of CXR-ReDonE, trained on MIMIC-PRO, potentially conferring a stylistic advantage independent of retrieval quality. Third, evaluation was conducted on a random sample from the CheXpert training set rather than its expert-verified validation set, leaving the ground-truth labels subject to automated-labeler noise. Finally, the cross-model correlation analysis was based on only 14 clinical findings, making the rank-correlation estimates inherently unstable; the paired bootstrap approach provides uncertainty estimates around these correlations but does not resolve the underlying limitation of having only 14 findings.

5.2 Implications

Our findings show that candidate-pool composition can influence retrieval performance and cross-model agreement, while raising several questions for future research. Studies using equal-sized pools could distinguish the effects of class balance from pool size, while expert-verified labels and neutral pool sources could further test the robustness of these findings. The different responses across findings and architectures also warrant investigation, particularly whether Lung Opacity remains difficult because of semantic overlap with related findings. Evaluating additional models, clinical findings, and datasets would help determine whether these effects generalize beyond the setting examined here.

5.3 Conclusion

This study examined how sentence retrieval pool composition can affect per-finding performance and cross-model agreement in chest X-ray retrieval. Lung Opacity and Pleural Effusion remained consistently difficult to retrieve, while strong agreement between models on relative finding difficulty under the original pool weakened after rebalancing. These findings suggest that apparent cross-model consensus may depend on evaluation pool composition, highlighting its importance in the design of CXR retrieval benchmarks.

Ethical standards

This study did not involve any new data collection from human participants. All chest radiographs and associated clinical labels were obtained from publicly available, de-identified datasets (MIMIC-CXR and CheXpert), which received institutional approval and patient consent waivers at the time of their original release. No additional ethical approval was required for this secondary analysis.

Conflict of interest

The author declares no conflict of interest.

Data availability

The datasets used in this study are publicly available. Training data (MIMIC-CXR) were obtained from `https://www.kaggle.com/datasets/zainabhalhoul/mimic-dataset`, licensed under Attribution-NonCommercial 4.0 International (CC BY-NC 4.0). Evaluation data (CheXpert) were obtained from `https://www.kaggle.com/datasets/ashery/chexpert`, licensed under CC0: Public Domain. Code and additional materials used in this study are available from the corresponding author upon reasonable request.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Ouis MY, Akhloufi MA. Deep learning for report generation on chest X-ray images. Computerized Medical Imaging and Graphics. 2024 Jan 1;111:102320.

[2] 2.Li Y, Kong C, Zhao G, Zhao Z. Automatic radiology report generation with deep learning: a comprehensive review of methods and advances. Artificial Intelligence Review. 2025 Aug 21;58(11):344.

[3] 3.Yang Y, You X, Zhang K, Fu Z, Wang X, Ding J, Sun J, Yu Z, Huang Q, Han W, Yu J. Spatio-temporal and retrieval-augmented modeling for chest x-ray report generation. IEEE Transactions on Medical Imaging. 2025 Mar 25;44(7):2892-905.

[4] Yu F, Endo M, Krishnan R, Pan I, Tsai A, Reis EP, Fonseca EK, Lee HM, Abad ZS, Ng AY, Langlotz CP. Evaluating progress in automatic chest x-ray radiology report generation. Patterns. 2023 Sep 8;4(9).

[5] Endo M, Krishnan R, Krishna V, Ng AY, Rajpurkar P. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. InMachine learning for health 2021 Nov 28 (pp. 209-219). PMLR.

[6] Mosquera C, Ferrer L, Milone DH, Luna D, Ferrante E. Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance. European Radiology. 2024 Dec;34(12):7895-903.

[7] Li W, Han G, Wu Y, Huang IC, Huang X. Joint Imbalance Adaptation for Radiology Report Generation. Journal of Healthcare Informatics Research. 2025 Dec;9(4):720-42.

[8] Jeong J, Tian K, Li A, Hartung S, Adithan S, Behzadi F, Calle J, Osayande D, Pohlen M, Rajpurkar P. Multimodal image-text matching improves retrieval-based chest x-ray report generation. InMedical imaging with deep learning 2024 Jan 23 (pp. 978-990). PMLR.

[9] Zhang B, Jia C, Liu S, Schunkert H, Navab N. Semantic-aware chest x-ray report generation with domain-specific lexicon and diversity-controlled retrieval. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2025 Sep 19 (pp. 607-616). Cham: Springer Nature Switzerland.

[10] 10. Ranjit M, Ganapathy G, Manuel R, Ganu T. Retrieval augmented chest x-ray report generation using openai gpt models. InMachine learning for healthcare conference 2023 Dec 22 (pp. 650-666). PMLR.

[11] 11.You K, Gu J, Ham J, Park B, Kim J, Hong EK, Baek W, Roh B. Cxr-clip: Toward large scale chest x-ray language-image pre-training. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2023 Oct 1 (pp. 101-111). Cham: Springer Nature Switzerland.

[12] Zhang X, Wu C, Zhang Y, Xie W, Wang Y. Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications. 2023 Jul 28;14(1):4542.

[13] Kittler F, Bhat S, Maier A. ProtoCLIP: Prototype-Aligned Latent Refinement for Robust Zero-Shot Chest X-Ray Classification. arXiv preprint arXiv:2604.18444. 2026 Apr 20.

[14] Zhou Y, Faith TL, Xu Y, Leng S, Xu X, Liu Y, Goh RS. Benchx: A unified benchmark framework for medical vision-language pretraining on chest x-rays. Advances in Neural Information Processing Systems. 2024 Dec 16;37:6625-47.

[15] Chen W, Yao L, Jin Q. Rethinking benchmarks for cross-modal image-text retrieval. InProceedings of the 46th international ACM SIGIR conference on research and development in information retrieval 2023 Jul 19 (pp. 1241-1251).

[16] Till T, Scherkl M, Stranger N, Singer G, Hankel S, Flucher C, Hržić F, Štajduhar I, Tschauner S. Impact of test set composition on AI performance in pediatric wrist fracture detection in X-rays. European radiology. 2025 Nov;35(11):6853-64.

[17] Fytas P, Selby I, Karner C, Babar J, Baker S, Beckford J, Sadler TJ, Shahipasand S, Thavakumar A, Chen JL, Sawer A. Rethinking Clinical Relevance in Chest X-ray Machine Learning: How Evaluation References Define Performance. arXiv preprint arXiv:2607.26333. 2026 Jul 28.

[18] Gong Y, Cosma G, Fang H. On the limitations of visual-semantic embedding networks for image-to-text information retrieval. Journal of Imaging. 2021 Jul 26;7(8):125.

[19] 19. MacAvaney S, Feldman S, Goharian N, Downey D, Cohan A. ABNIRML: Analyzing the behavior of neural IR models. Transactions of the Association for Computational Linguistics. 2022 Mar 18;10:224-39.

[20] 20. Hendriksen M, Zhang S, Reinanda R, Yahya M, Meij E, de Rijke M. Benchmark granularity and model robustness for image-text retrieval: a reproducibility study. InProceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval 2025 Jul 13 (pp. 3183-3193).

[21] 21.Ramesh V, Chi NA, Rajpurkar P. Improving radiology report generation systems by removing hallucinated references to non-existent priors. InMachine Learning for Health 2022 Nov 22 (pp. 456-473). PMLR.

[22] Ji Z, Shaikh MA, Moukheiber D, Srihari SN, Peng Y, Gao M. Improving joint learning of chest x-ray and radiology report by word region alignment. InInternational workshop on machine learning in medical imaging 2021 Sep 21 (pp. 110-119). Cham: Springer International Publishing.

[23] Bankier AA, MacMahon H, Colby T, Gevenois PA, Goo JM, Leung AN, Lynch DA, Schaefer-Prokop CM, Tomiyama N, Travis WD, Verschakelen JA. Fleischner society: glossary of terms for thoracic imaging. Radiology. 2024 Feb 27;310(2):e232558.

[24] 24.Ruskin JA, Gurney JW, Thorsen MK, Goodman LR. Detection of pleural effusions on supine chest radiographs. American Journal of Roentgenology. 1987 Apr 1;148(4):681-3.

[25] 25.Kovatchev V, Lease M. Benchmark transparency: Measuring the impact of data on evaluation. InProceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) 2024 Jun (pp. 1536-1551).

[26] Bassi PR, Li W, Tang Y, Isensee F, Wang Z, Chen J, Chou YC, Roy S, Kirchhoff Y, Rokuss M, Huang Z. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation?. Advances in Neural Information Processing Systems. 2024 Dec 16;37:15184-201.

[27] Qu W, Balki I, Mendez M, Valen J, Levman J, Tyrrell PN. Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging. International journal of computer assisted radiology and surgery. 2020 Dec;15(12):2041-8.