

RADIOLOGY REPORT GENERATION

Chest Radiography (CXR) Narrative Model

Data & training analysis of a vision-language foundation model that generates **full free-text findings sections** directly from chest X-ray images

AUTHOR

Neurologic AI

PUBLISHED

July 2026

PRIMARY METRIC

F1-NeuroCheXbert

ABSTRACT

Chest radiography is among the most frequently performed diagnostic imaging examinations, yet the volume of studies routinely performed exceeds available radiologist reporting capacity, motivating growing interest in automated systems capable of assisting with report generation. Prior AI-based approaches to chest X-ray interpretation have largely relied on binary classification, predicting the presence or absence of a fixed set of disease labels; while useful for triage, such

outputs cannot substitute for the free-text findings expected in a clinical radiology report. This paper presents the Neurologic AI Chest Radiography (CXR) Narrative Model, a vision-language foundation model that generates full free-text findings sections directly from frontal and, where available, lateral chest X-ray images. The model couples a frozen, LoRA-adapted vision encoder with a Q-Former bridging module and a causal language model decoder, and is trained on a curated corpus of 1.2 million de-identified chest X-ray studies spanning 26 disease categories, sampled from over 5.1 million eligible studies using rarest-first equalization. Model quality is assessed with F1-NeuroCheXbert, a CheXbert-style metric that extracts structured labels from generated and reference text for comparison across 27 clinical categories. On a held-out test set of 25,000 studies, the model achieves an F1-NeuroCheXbert score of 0.4413, with per-category performance strongest on structurally distinctive findings such as hiatal hernia, intercostal drains, and pacemaker or cardiac devices, and weaker on subtle or rare findings such as shoulder dislocation and pulmonary artery enlargement. These results suggest that narrative-style report generation from chest radiographs is technically feasible at scale and provide a foundation for continued development toward downstream clinical integration.

1.2M

De-identified chest X-ray studies

26

Disease categories

0.4413

Test F1-NeuroCheXbert, 25,000 studies

01

Introduction

Chest radiography remains the most widely performed diagnostic imaging examination in modern healthcare, used routinely for the screening, diagnosis, and monitoring of a broad spectrum of thoracic and cardiopulmonary conditions, ranging from pneumonia and pleural effusion to cardiomegaly, pneumothorax, and post-surgical device placement. Each study is typically reviewed by a radiologist, who translates the visual appearance of the lungs, heart, mediastinum, bones, and any indwelling support devices into a structured narrative findings

section that becomes part of the patient’s medical record. This interpretive step is central to clinical decision-making, yet it is also one of the most time-consuming components of a radiologist’s workload.

The steadily increasing volume of chest X-ray studies, combined with persistent workforce constraints in radiology departments worldwide, has created sustained demand for artificial intelligence systems that can assist with, and partially automate, this reporting process. Most existing AI solutions in this space have been framed as classification problems: given an image, the model predicts the presence or absence of a fixed set of disease labels (for example, cardiomegaly or pneumothorax). These systems are effective for triage and flagging urgent findings, but their output is fundamentally mismatched with clinical documentation practice, since a radiology report is not a list of binary labels but a free-text narrative that describes location, severity, laterality, and clinical context in prose.

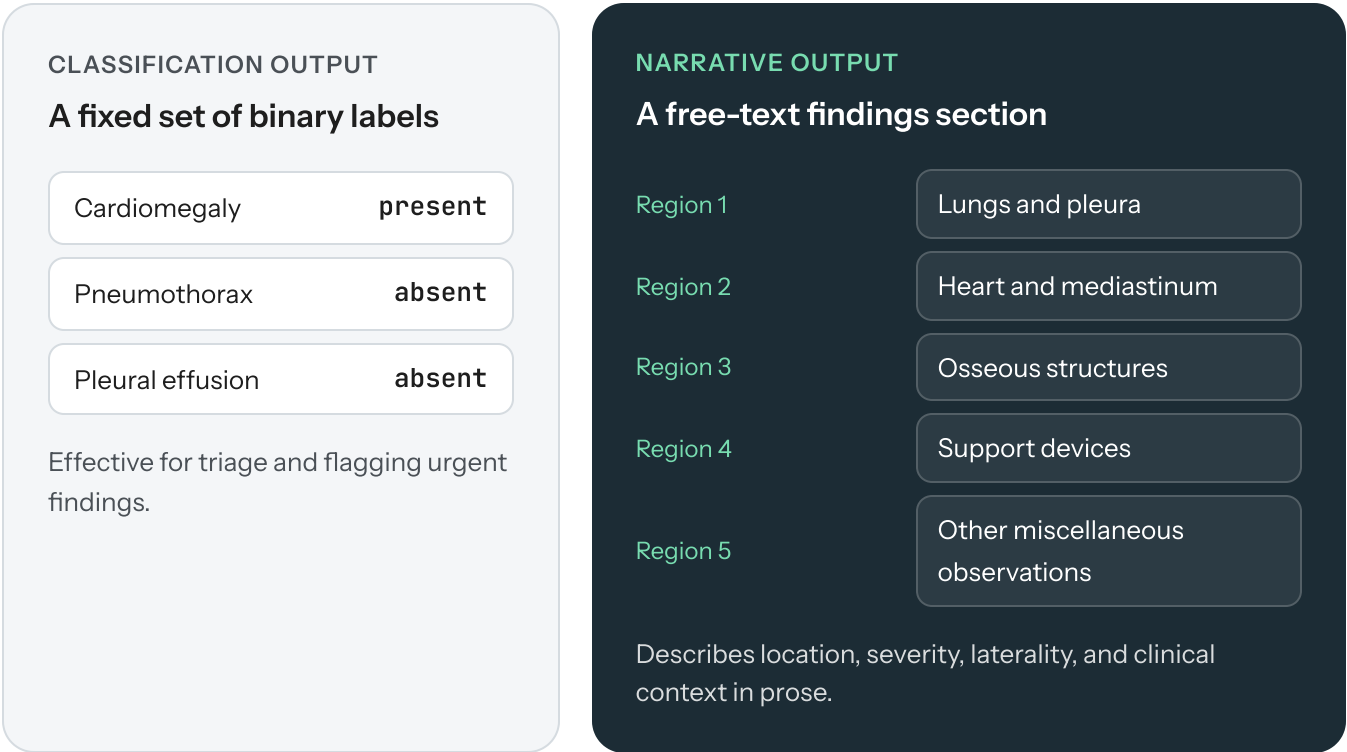


Figure 1. A classifier predicts the presence or absence of a fixed set of disease labels; a radiology report is a free-text narrative organized around the anatomical regions a radiologist would address in turn.

Recent advances in vision-language modeling offer a path to close this gap. By combining a visual encoder with a causal language model decoder through a learned bridging mechanism, it becomes possible to condition free-text generation directly on image content, in much the same way that image captioning models generate descriptive text, but adapted to the conventions and vocabulary of clinical radiology reporting. This approach

allows a model to produce a coherent findings section organized around the same anatomical regions a radiologist would address in turn: the lungs and pleura, the heart and mediastinum, the osseous structures, any support devices, and other miscellaneous observations.

A radiology report is not a list of binary labels but a **free-text narrative** that describes location, severity, laterality, and clinical context in prose.

This paper describes the data and training methodology behind the Neurologic AI Chest Radiography (CXR) Narrative Model, a vision-language foundation model built to generate free-text radiology findings directly from chest X-ray images, without an intermediate discrete-label prediction step. The model is trained on a large, curated, de-identified dataset drawn from real-world clinical imaging, and its output quality is evaluated using an extended CheXbert-style metric that extracts structured clinical concepts from generated text for comparison against reference reports. The remainder of this paper is organized as follows: Section 2 describes the dataset used for training and evaluation, including its source, composition, and preprocessing; Section 3 describes the model architecture and training configuration; Section 4 summarizes observations from the training process; and Section 5 concludes with a discussion of results and their implications.

02

Dataset information

2.1 Dataset description and size

The CXR Narrative Model (also referred to as the CXR VLM) was trained and evaluated on a large-scale dataset derived from 6,337,300 unique patient-study pairs extracted from an internal DICOM data warehouse, restricted to chest X-ray studies (modalities CR, DR, and DX). Following image linking, report parsing, quality filtering, and integration of disease labels, 5,168,884 studies remained eligible for inclusion.

To construct the final dataset, targeted downsampling was applied using a rarest-first equalization strategy, along with a 25% negative (all-normal) fraction. This resulted in a curated dataset of 1,200,000 studies, split into 1,150,000 for training, and 25,000 each for

validation and testing.

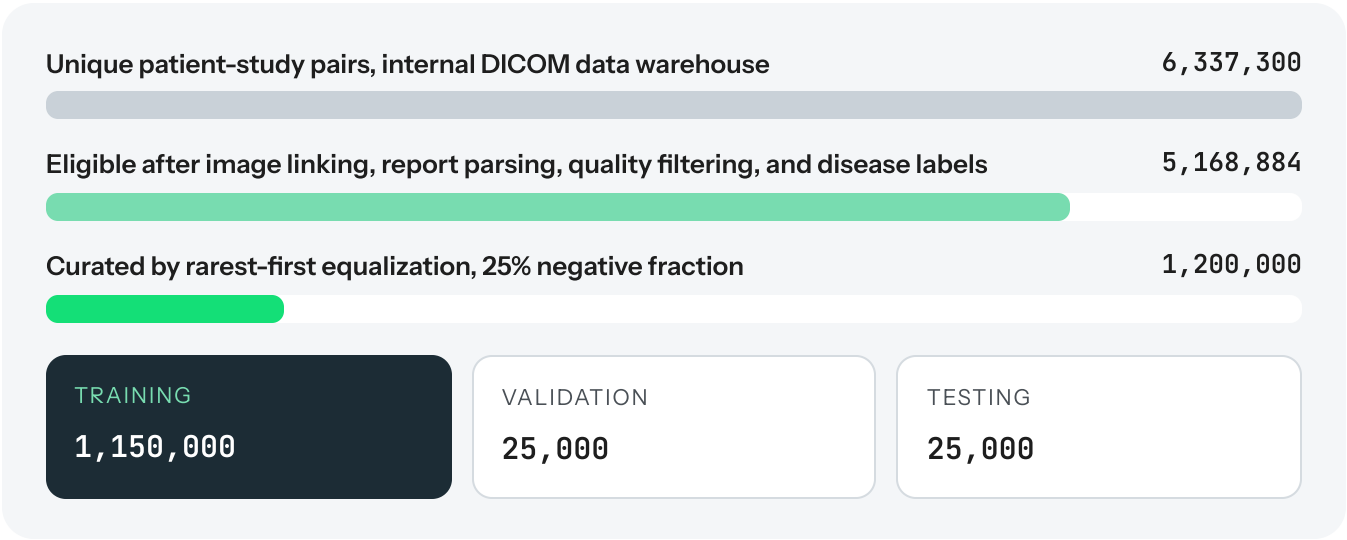


Figure 2. From the data warehouse to the curated dataset. Bars are drawn to scale.

Each study is associated with up to two images and a corresponding free-text radiology findings section. In terms of imaging composition, 80.2% of studies include both frontal and lateral views, while the remaining 19.8% contain only a frontal view. The dataset schema includes the following fields: `patient_id`, `study_id`, images (list of paths), views (e.g., frontal, lateral), findings (target text), impression, comparison, indication, exam, technique, history, and the full report. The label space spans 26 disease categories, along with tracking for normal (all-absent) cases.



To improve consistency and reduce noise in the training data, the findings text was preprocessed before training. References to prior imaging studies were removed to prevent the model from learning to reference prior exams that will not be available to the model at inference time. The remaining observations were rearranged and concatenated according to a standardized radiology reporting structure:

1	lungs_pleura	lung parenchyma, airways, pleural space
2	heart_mediastinum	cardiac silhouette, mediastinal structures, vasculature
3	bones	osseous structures, spine, ribs
4	support_devices	lines, tubes, catheters, pacemakers, etc.
5	others	soft tissue, diaphragm, hiatal hernia, and miscellaneous observations

Figure 3. The standardized radiology reporting structure used to rearrange and concatenate the findings text.

This structured ordering reflects typical clinical reporting conventions and helps ensure a more coherent and consistent target text for model training.

2.2 Dataset source

The dataset originates from a centralized data warehouse containing de-identified DICOM metadata and associated radiology reports. Studies were queried for chest X-ray modalities (CR/DR/DX) with non-empty report text. Image paths were resolved from an internal image registry and disease labels from an internal label registry. The underlying query filters chest studies using body part keywords (chest/thorax/cxr/chst/ches) from the `ProtocolName` or `BodyPartExamined` DICOM tags, and excludes non-chest regions (abdomen, pelvis, spine, extremities, etc.).

2.3 Dataset acquisition methods

The CXR dataset is made up of proprietary retrospective data obtained through direct data use agreements with healthcare providers and data brokers. The dataset includes imaging studies collected during routine clinical care and acquired from a broad array of scanner hardware vendors.

2.4 Data collection period

2001/10/01 to  2025/03/15

2.5 Representativeness

2.5.1 Disease conditions

The CXR VLM is an AI-enabled software component intended for use by medical device manufacturers to integrate into finished devices. The CXR VLM generates free-text radiology report findings from chest X-ray images. Unlike binary classification models, the CXR VLM produces natural language descriptions of radiological observations.

The CXR VLM is trained on radiology report findings text spanning a broad range of radiological conditions. The training dataset was sampled to ensure representation across 26 disease categories using rarest-first equalization. Table 1 summarizes the disease label distribution in the sampled training set. Note that these labels reflect the underlying disease prevalence in the training data, but the model’s output is free-text findings, not discrete disease labels.

Condition		Data distribution (studies)
air_space_opacity		280,454
normal		287,500
cardiomegaly		255,185
atelectasis		165,170
interstitial_thickening		161,977
pleural_effusion		140,142
hyperinflation		138,398
pacemaker_electronic_cardiac_device_or_wires		112,757
pulmonary_congestion_pulmonary_venous_congestion		109,519
lung_nodule_or_mass		96,891
bronchial_wall_thickening		84,389
non_acute_rib_fracture		70,085
hiatus_hernia		49,250
peribronchial_cuffing		45,363
lobar_segmental_collapse		42,248

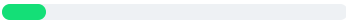
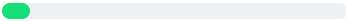
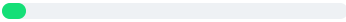
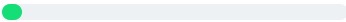
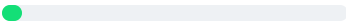
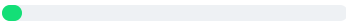
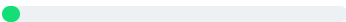
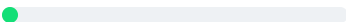
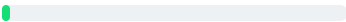
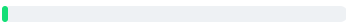
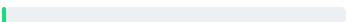
Condition		Data distribution (studies)
pulmonary_artery_enlargement		37,118
intercostal_drain		24,462
tracheal_deviation		21,207
pneumothorax		17,833
implantable_electronic_device		17,446
nonsurgical_internal_foreign_body		16,626
hilar_lymphadenopathy		16,282
bullous_disease		14,056
acute_rib_fracture		7,322
subcutaneous_emphysema		6,134
shoulder_dislocation		4,955
whole_lung_or_majority_collapse		1,795

Table 1. Training data disease label distribution (VLM Narrative Model dataset, sampled train set: 1,150,000 studies).

2.5.2 Patient demographics

The dataset comprises 1.2 million sampled chest X-ray studies (from 5.17 million eligible) from both inpatient and outpatient settings across diverse US regions (including the East, West, Southeast, and Central US), and includes patients across all adult age groups and sexes. Table 2 summarizes the patient demographics.

Demographic	Category	Data distribution (studies)
Sex	Male	2,885,842
	Female	3,573,775
	Other / Unknown	28,892

Demographic	Category	Data distribution (studies)
Age	0–18 years	149,930
	18–30 years	80,810
	30–50 years	406,697
	50–70 years	1,659,752
	70–100 years	2,082,040
	Other / Unknown	2,109,150
Race	Unavailable	
Ethnicity	Unavailable	
Height	Very Limited Availability	
Weight	Very Limited Availability	

Table 2. Patient demographic data distribution (CXR model).

Structured data for race, ethnicity, height, and weight were not available for most studies in the dataset and are therefore not included in this analysis. While these characteristics may influence disease prevalence or image acquisition in certain contexts, their impact on model performance cannot be formally assessed due to missing metadata.

Zip code-based analysis can be used as a proxy to assess dataset diversity, including differences in socioeconomic status, geography (e.g., urban vs. rural), and limited demographic variation. When available, it helps show whether the dataset includes a range of patient populations and care settings relevant to the intended use. Zip code data is used as supporting context rather than a primary measure of representativeness. Evaluation of model performance and generalizability focuses on clinically relevant variables, such as acquisition characteristics and patient demographics where available.

KNOWN LIMITATION

Given the breadth and scale of the dataset, it is believed to reflect substantial demographic variability and clinical diversity. Nonetheless, the absence of explicit race, ethnicity, height, and weight data limits the ability to perform structured subgroup analyses and precludes a definitive assessment of potential model bias related to these attributes. This is a known limitation and may be evaluated in future work if such metadata becomes available.

2.5.3 Data acquisition equipment

Based on the DICOM manufacturer tag, studies were cleaned and labeled with ten major scanner manufacturers, including Carestream Health, Fujifilm, Medlink Imaging, Konica Minolta Healthcare, GE Healthcare, Siemens Healthineers, Kodak, Canon Medical Systems, Samsung Medison, and Agfa. Unknown, mislabeled, or less common scanners were grouped into an “Other” category. The “Other” category comprises 51% of all data, 96% of which has unknown manufacturer information. Table 3 summarizes the data acquisition equipment and study count by scanner.

Site name		Data distribution (studies)
Carestream Health		714,417
Agfa		78,832
Siemens Healthineers		265,905
GE Healthcare		340,941
Fujifilm		694,681
Konica Minolta Healthcare		375,486
Kodak		270,391
Canon Medical Systems		273,151
Samsung Medison		171,741
Medlink Imaging		652,000
Other / Unknown		2,362,168

Table 3. Data acquisition equipment (CXR model).

2.5.4 Data acquisition sites

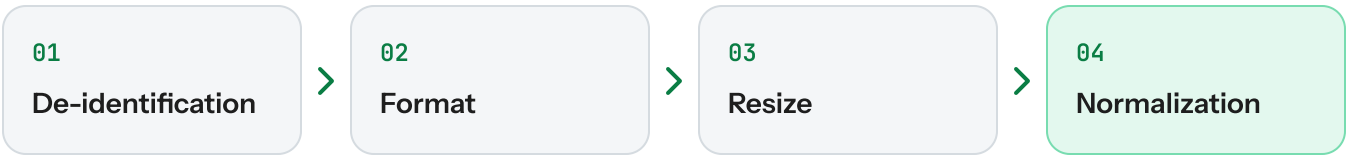
The dataset comes from two providers: Data Site 1 (outpatient) and Data Site 2 (inpatient). Data provider identities have been redacted pursuant to confidentiality obligations. Table 4 summarizes the data acquisition sites.

Site name		Data distribution (studies)
Data Site 1* (outpatient)		5,608,150
Data Site 2* (inpatient)		591,926

Table 4. Data acquisition sites (CXR model). *Data provider identities redacted pursuant to confidentiality obligations. Available upon request to authorized parties.

2.6 Data preprocessing

The following pre-processing steps are applied to all images prior to training:



- **De-identification:** Images are de-identified using customized third-party OCR solutions to identify and redact areas in the image that contain text that has not been whitelisted.
- **Format:** Image data is standardized to High-Throughput JPEG 2000 (HTJ2K) format.
- **Resize:** Each image is resized to a fixed resolution of 448×448 pixels using anti-aliasing to preserve image quality.
- **Normalization:** Pixel values are normalized using standard ImageNet statistics (mean: [0.485, 0.456, 0.406]; standard deviation: [0.229, 0.224, 0.225]) for consistency across the dataset.

During training, the following augmentation techniques were applied to enhance model robustness and generalization:

Random Resized Cropping

Images randomly cropped and resized to a fixed resolution (448×448 pixels) with scaling ranging between 80% to 100% of the original size.

Random Horizontal Flip

Images horizontally flipped with a probability of 50%.

Color Jittering (80% probability)

Random adjustments in brightness ($\pm 40\%$), contrast ($\pm 40\%$), saturation ($\pm 20\%$), and hue ($\pm 10\%$).

Gaussian Blur (100% probability)

Consistent application of Gaussian blur to enhance robustness against variations in sharpness.

2.7 Ground truth

TRAINING

Ground truth is defined as the findings sections of radiology reports. These text findings are used directly as the target outputs the model learns to generate. No separate labels are created, extracted, or assigned for model training.

EVALUATION ONLY

For evaluation only, structured labels may be extracted from the findings text to enable quantitative performance metrics (e.g., disease-level F1). This label extraction is part of the evaluation process and does not contribute to model training.

03

Model training description

3.1 Model training description & architecture

The CXR VLM report-generation model uses a Q-Former + ModernBERT architecture that bridges a frozen vision encoder with a causal language model decoder via a Q-Former module. The model generates free-text radiology findings from chest X-ray images. Unlike a previous binary classification pipeline, this model is trained end-to-end as a vision-language model with a single training stage.

3.1.1 Architecture overview

The model consists of three main components:

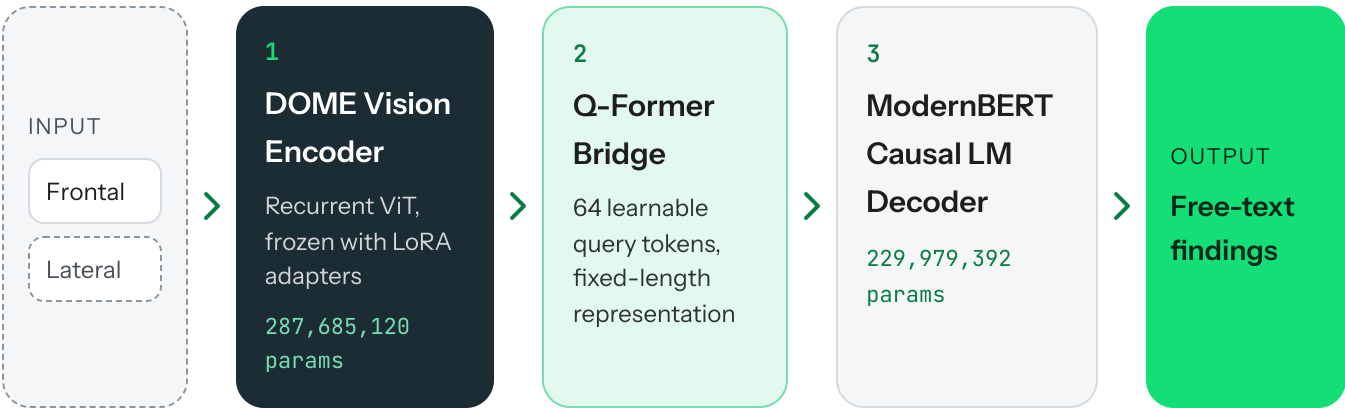


Figure 4. The Q-Former + ModernBERT architecture: a frozen vision encoder bridged to a causal language model decoder via a Q-Former module.

1. A **DOME Vision Encoder** (Recurrent ViT, pretrained on CXR data, 448×448px input, 1536-dim hidden, frozen with LoRA adapters: $r = 8$, $\alpha = 16$, dropout = 0.15, targeting `attn.qkv`, `attn.proj`, `mlp.fc1`, `mlp.fc2`; 287,685,120 total params, 1,966,080 LoRA-trainable = 0.68%);
2. A **Q-Former Bridge** (cross-attention module with 64 learnable query tokens, depth = 4 layers, 8 attention heads, MLP ratio = 4.0, compressing vision features into a fixed-length representation); and
3. A **ModernBERT Causal LM Decoder** (229,979,392 parameters, pretrained separately on the same dataset's findings text, `flash_attention_2` during training, tokenizer max length = 128).

MODEL PARAMETERS

Trainable	Frozen
-----------	--------

- **Total parameters:** 586,490,624
- **Trainable parameters:** 300,771,584 (includes decoder, Q-Former, and vision LoRA)
- **Frozen parameters:** 285,719,040 (vision encoder backbone)

Training configuration		32 GPUs, bf16
Optimizer	AdamW (lr = 3.0×10^{-5} , <code>weight_decay</code> = 0.05, eps = 1×10^{-8} , betas = [0.9, 0.98])	
Learning rate scheduler	Linear warmup (10,000 steps, ~3% of total) + Cosine decay	
Gradient clipping	max norm 0.7	
Mixed precision	bf16	
Batching	Per-device batch size: 4 · Gradient accumulation: 2 · GPUs: 32 · Effective batch size: 256	
Max epochs	100 (with early stopping: patience = 10 on “F1-NeuroCheXbert”, starting epoch 3)	
Seed	998798	

3.1.2 Generation configuration

At inference time, findings text is produced using beam search with a beam width of two and a maximum generation length of 128 tokens. Training proceeded for 36 epochs, spanning approximately 34 hours and 39 minutes on the 32-GPU cluster, before early stopping was triggered. Validation performance, tracked via F1-NeuroCheXbert, improved steadily through the early epochs before plateauing in the mid-30s epoch range, with the best-performing checkpoint identified at epoch 26; this checkpoint was used for all subsequent evaluation. On the full 25,000-sample test set, the selected checkpoint achieved an F1-NeuroCheXbert score of 0.3965 under eager attention and 0.3953 under flash attention, indicating that the choice of attention implementation at inference time has negligible impact on generation quality. A smaller 1,000-sample subset evaluation yielded a lower score of 0.3168, illustrating the sensitivity of the metric to sample size and test set composition.

0.3965Eager attention, 25,000-
sample test set**0.3953**Flash attention, 25,000-
sample test set**0.3168**

1,000-sample subset

3.2 Training environment

Cloud-based (32 GPUs)

3.3 Data splitting strategy

Data splits were assigned deterministically using a three-tier precedence: (1) *pair-level split first*: if the exact (`patient_id` , `study_id`) has a known split from the CXR classification split file, use it; (2) *patient-level majority known split*: if the pair has no split, look at all other studies for that patient with known splits and use the majority; (3) *fallback train*: if neither the pair nor the patient has any known split info, assign to training. This ensures that all studies for a given patient remain in the same split, preventing data leakage. The full eligible set of 5,168,884 studies was then downsampled to target sizes of 1,150,000 train / 25,000 val / 25,000 test, with a target negative fraction of 25% in each split. Negatives were sampled first, then positives were filled using rarest-first equalization across diseases. Table 5 summarizes the total number of studies assigned to each data partition.

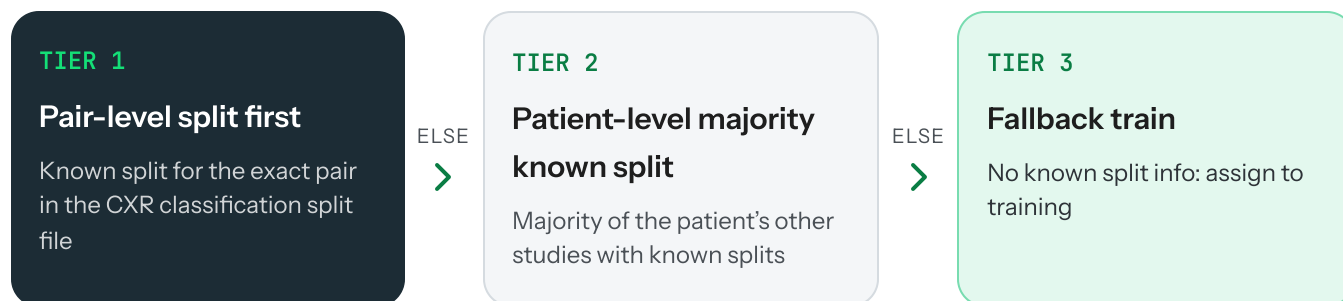


Figure 5. The three-tier precedence that keeps all studies for a given patient in the same split.

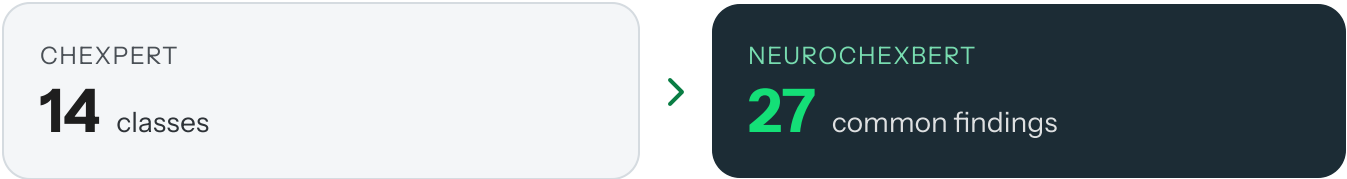
Activity	Total studies
Training	1,150,000 studies
Validation (hyperparameter tuning)	25,000 studies
Test (final model evaluation)	25,000 studies

Table 5. Training-validation-test split (CXR VLM report generation).

The CXR VLM dataset uses up to 2 images per study, selected from frontal and lateral views. Studies were required to have at least one frontal image for eligibility. View labels were normalized into frontal, lateral, and other categories. The dataset includes studies with frontal-only and frontal+lateral combinations. During training and evaluation, images are resized to 448×448 pixels. The model generates a single findings text per study, regardless of the number of input images.

3.4 Validation metrics

The primary evaluation metric is F1-NeuroCheXbert, a CheXbert-based F1 score that compares the clinical content of generated report findings against ground-truth text. CheXpert is the standard evaluation target for CXR classification models and includes 14 classes; NeuroCheXbert broadens this to 27 common findings (listed in Table 7) while excluding rare conditions whose low prevalence makes statistical analysis challenging.



Model performance was monitored throughout training using validation F1-NeuroCheXbert and loss. The checkpoint with the best validation F1-NeuroCheXbert was selected for final evaluation. This metric was chosen because it captures clinically meaningful accuracy of generated reports by comparing structured findings extracted from both the generated and reference texts. The evaluation results are presented in Table 6. Note that customer performance targets are expected to be defined and validated by the customer based on their specific use case and clinical goals.

Split	NeuroCheXbert	eval_loss
Validation	0.4521	1.1956
Test	0.4413	1.2021

Table 6. CXR VLM evaluation results (F1-NeuroCheXbert).

Per-disease F1

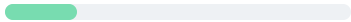
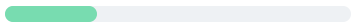
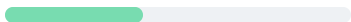
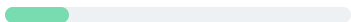
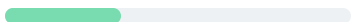
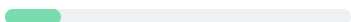
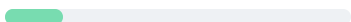
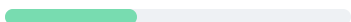
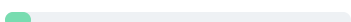
Label (all 27)		F1
acute_rib_fracture		0.2072
air_space_opacity		0.6416
cardiomegaly		0.6498
lung_nodule_or_mass		0.2640
non_acute_rib_fracture		0.3976
pleural_fluid		0.7089
pneumothorax		0.5295
pulmonary_artery_enlargement		0.1828
atelectasis		0.4869
bronchial_wall_thickening		0.3355
bullous_disease		0.1635
hilar_lymphadenopathy		0.1657
hiatus_hernia		0.7840
hyperinflation		0.3818
implantable_electronic_device		0.6689
intercostal_drain		0.7505
interstitial_thickening		0.3742
lobar_segmental_collapse		0.1859
nonsurgical_internal_foreign_body		0.3743
pacemaker_electronic_cardiac_device_or_wires		0.7371
peribronchial_cuffing		0.2240
pulmonary_congestion_pulmonary_venous_congestion		0.4690
shoulder_dislocation		0.0754
subcutaneous_emphysema		0.5863
tracheal_deviation		0.2811
whole_lung_or_majority_collapse		0.6468
no_finding		0.6430

Table 7. Per-disease F1. Bars show F1 on a 0 to 1 scale; values of 0.70 and above are drawn in a brighter green.

3.5 Reproducibility controls

Model training code and environment configurations are retained in version control, which contains logic for each element of the training process. Training logs and manifests are preserved to support reproducibility.

04

Training observances

Training of the CXR VLM report-generation model proceeded stably over 36 epochs (~34h 39m on 32 GPUs). The validation F1-NeuroCheXbert metric showed consistent improvement from 0.1537 (epoch 1) to 0.4021 (epoch 26, best), followed by a plateau and mild degradation through epoch 36, triggering early stopping. No training runs failed due to loss divergence or numerical instability. Validation loss showed the expected pattern of initial decrease (1.1485 at epoch 1 to ~1.07 at epoch 7) followed by gradual increase due to the generative nature of the task, while generation quality (F1-NeuroCheXbert) continued to improve. The final test evaluation confirmed generalization with F1-NeuroCheXbert of 0.4413 on the full 25,000-sample test set.

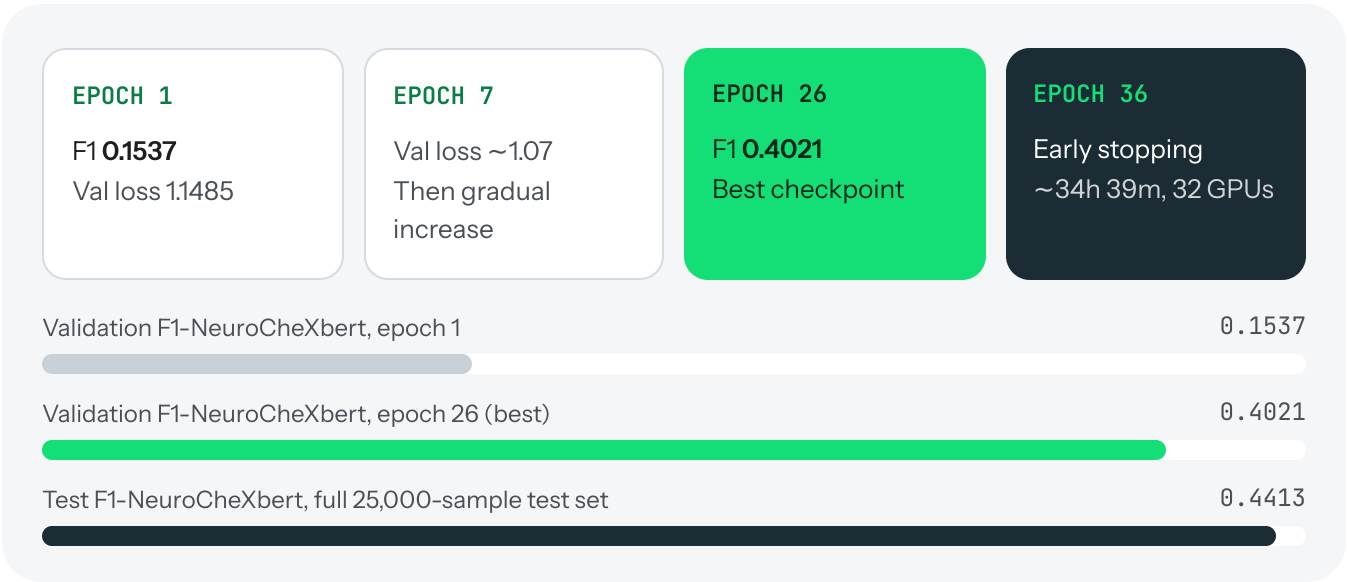


Figure 6. Training milestones over 36 epochs, with the reported F1-NeuroCheXbert values drawn to a common scale.

Retraining triggers are currently determined based on observed underperformance during post-deployment evaluation or the availability of new, high-quality labeled data. The model architecture and training framework are designed to support incremental retraining without disrupting validated core components.

— 05

Conclusion

This document outlines the training process for the Neurologic AI CXR Narrative Model. The model uses a Q-Former + ModernBERT architecture comprising a DOME vision encoder (frozen with LoRA adapters), a Q-Former cross-attention bridge, and a ModernBERT causal language model decoder. The model was trained on 1,150,000 chest X-ray studies with associated radiology report findings text, achieving a best validation F1-NeuroCheXbert score of 0.4521 and a test F1-NeuroCheXbert of 0.4413. The dataset originates from real-world retrospective clinical imaging and includes systematic preprocessing workflows. The resulting model is designed to support downstream integration by medical device manufacturers for automated radiology report generation.

THE RESULT

These results suggest that narrative-style report generation from chest radiographs is technically feasible at scale.

— 06

References

1. US FDA Draft Guidance, “Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations” (issued January 7, 2025)
2. ISO 13485:2016, Medical Devices, Quality Management Systems

3. ISO/IEC 42001:2023, Information Technology, Artificial Intelligence, Management System
4. ISO 14971:2019, Medical Devices, Application of Risk Management to Medical Devices
5. ISO/IEC 23894:2023, Information Technology, Artificial Intelligence, Guidance on Risk Management
6. IEC 62304:2006, Medical Device Software, Software Life Cycle Processes



From complexity to engineered certainty