

AUTOMATED RADIOLOGY REPORTING

Chest CT (CTC) Narrative Model

A multi-component pipeline for automated chest CT report generation, pairing narrative generation with **targeted, verifiable measurement**

AUTHOR

Neurologic AI

PUBLISHED

July 2026

TRAINING CORPUS

227,000 chest CT studies

ABSTRACT

Chest computed tomography (CT) presents a substantially harder automated reporting problem than planar chest radiography: a single study comprises hundreds of cross-sectional slices, findings must be localized in three dimensions, and several of the most clinically consequential observations, such as pulmonary nodule size and aortic diameter, require precise quantitative measurement rather than qualitative description alone. A single end-to-end vision-language model,

however capable at producing fluent narrative text, is not well suited to this kind of metrology, since generative language decoders are not trained to guarantee millimeter-level numerical accuracy. This paper describes the Neurologic AI Chest CT (CTC) Narrative Model, a beta-stage, multi-component pipeline that pairs a vision-language model (a LoRA-adapted V-JEPA 2.1 vision encoder, a Q-Former bridging module, and a ModernBERT causal decoder) for free-text findings generation with two dedicated, deterministic measurement sub-pipelines: a TotalSegmentator-based aortic diameter processor and a two-stage RetinaNet-plus-SegResNet pulmonary nodule detector. A large language model then merges the outputs of all three components into a single, internally consistent findings report. The model is trained on a condition-balanced corpus of 227,000 chest CT studies, built using rarest-first equalization over sixteen clinically significant disease labels to counteract the severe class imbalance of a much larger parent dataset. On held-out evaluation, the pipeline achieves a weighted F1 of 0.551 against an internal 16-condition classifier and 0.500 against an independent 18-condition classifier from the literature; the pulmonary nodule pipeline meets its sensitivity targets for nodules under 6 mm and over 15 mm in diameter, and the aortic ectasia processor achieves 95.2% sensitivity for detecting dilation or aneurysm at the case level. These results indicate that decomposing chest CT report generation into a narrative-generation component and targeted, verifiable measurement components is a viable strategy for approaching the accuracy demands of volumetric radiology reporting, while highlighting specific size ranges and structures where further refinement is needed before clinical-grade performance is reached.

0.551

Weighted F1, internal 16-condition classifier

0.500

Weighted F1, independent 18-condition classifier

95.2%

Case-level sensitivity for aortic dilation or aneurysm

Introduction

Computed tomography of the chest is one of the most information-dense examinations in diagnostic radiology. Unlike a chest radiograph, which compresses the thorax into a single two-dimensional projection, a chest CT study typically consists of several hundred axial slices acquired at sub-millimeter to few-millimeter spacing, reconstructed with multiple kernels optimized for different tissue types. A radiologist reviewing such a study must integrate findings across this entire volume: identifying and characterizing pulmonary nodules, assessing airway and parenchymal disease, measuring the aorta and great vessels for signs of dilation or aneurysm, and surveying the mediastinum, heart, upper abdomen, bones, and soft tissues, all while cross-referencing prior studies when available. The resulting findings section of a chest CT report is accordingly longer, more structured, and more quantitative than its chest radiograph counterpart, routinely including explicit measurements (e.g., nodule diameter in millimeters, aortic diameter at multiple anatomical levels) rather than purely descriptive language.

This additional demand for quantitative precision creates a specific challenge for automated report generation. Vision-language models built around a visual encoder and a causal language decoder are highly effective at producing fluent, structurally coherent narrative text conditioned on image content, and have shown promise for chest radiograph reporting. However, a purely generative decoder has no intrinsic mechanism for guaranteeing that a stated measurement, such as a nodule diameter or an aortic caliber, is numerically accurate; the decoder is optimized to produce plausible text, not to perform metrology. For findings where only qualitative characterization is required, this limitation is tolerable, but for findings where a specific numerical value materially changes clinical management (for example, distinguishing a nodule that warrants Lung-RADS follow-up from one that does not, or an aorta that is ectatic from one that is normal), relying on a language model to generate that number from visual features alone is a meaningfully riskier design choice.

This paper describes an alternative approach: rather than treating chest CT report generation as a single end-to-end generative task, the Neurologic AI Chest CT (CTC) Narrative Model decomposes the problem into a narrative-generation component and two dedicated, deterministic measurement components, and then reconciles their outputs into a single report. The core vision-language model generates a structured draft findings section covering the pulmonary, mediastinal, cardiac, upper abdominal, osseous, and soft-tissue observations that are best expressed qualitatively. In parallel, a multi-organ

segmentation model extracts the aorta and computes its diameter at four anatomical levels, and a two-stage detection-and-segmentation pipeline localizes, measures, and classifies pulmonary nodules according to Lung-RADS-relevant size thresholds. A large language model then integrates these three outputs, replacing the relevant sentences of the VLM-generated draft with the more precisely measured findings while leaving the remainder of the narrative intact.

INTENDED USE

The model described here is a beta-stage research artifact, intended as a developer-facing foundation model rather than a validated clinical device; it is designed to produce a draft findings section for downstream integration and further validation, not for direct diagnostic use.

The remainder of this paper is organized as follows: Section 2 describes the model architecture in detail, including the vision-language backbone and the two measurement sub-pipelines; Section 3 describes the training dataset and how class imbalance across rare but clinically important findings was addressed; Section 4 describes the base model training configuration; Section 5 presents performance and validation results for the narrative, nodule, and aorta components; Section 6 discusses known limitations; and Section 7 concludes.

02

Model architecture

The CTC Narrative Model is structured as a pipeline of six components, each addressing a distinct part of the report-generation task.

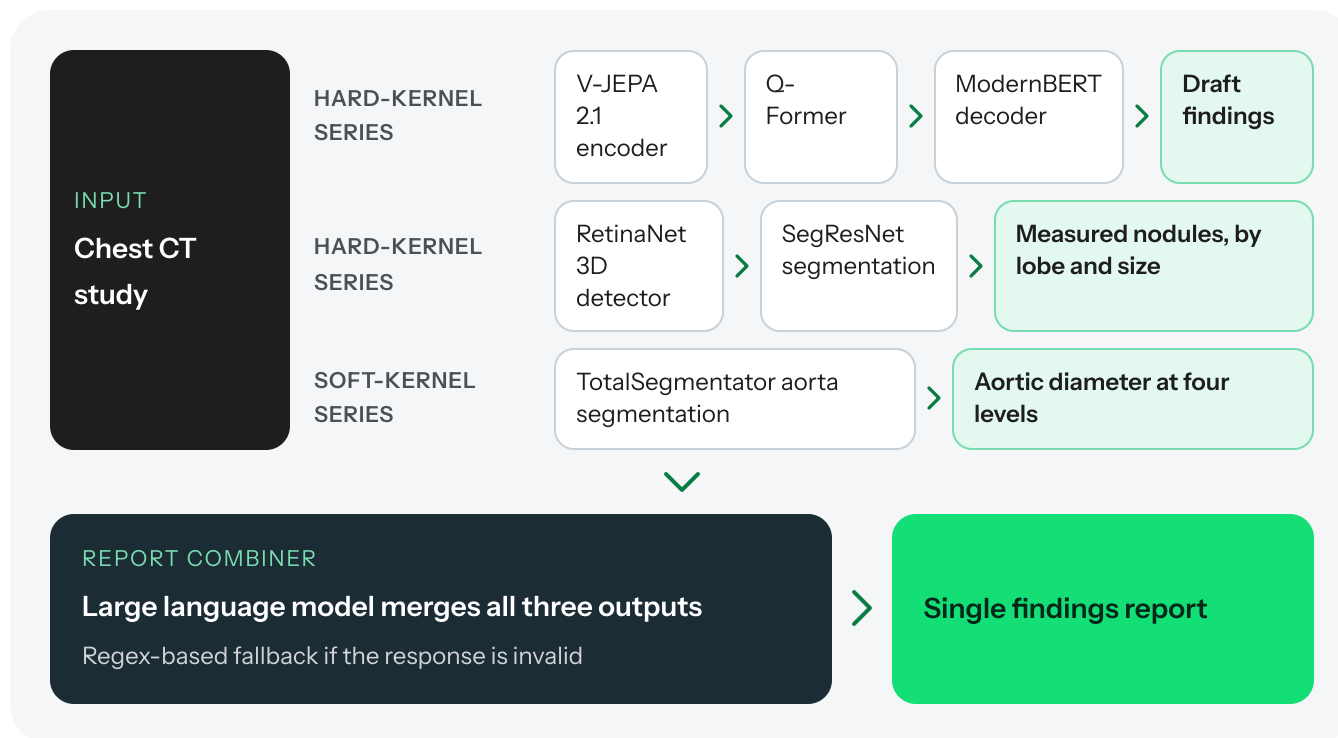


Figure 1. The six-component pipeline: a vision-language backbone for narrative generation, two deterministic measurement sub-pipelines, and a report combiner.

2.1 Vision-language backbone

The core narrative-generation component follows a vision-language architecture consisting of a frozen, LoRA-adapted vision encoder, a learned bridging module, and a causal language decoder.

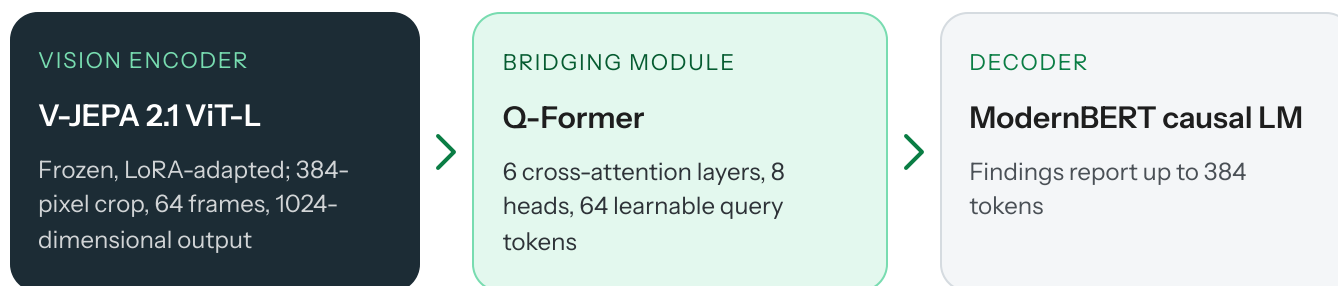


Figure 2. The vision-language backbone. The Q-Former and decoder are fine-tuned together; the vision backbone is held frozen.

The vision encoder is a V-JEPA 2.1 ViT-L model [1], chosen for its use of rotary position embeddings and hierarchical spatiotemporal features, both of which are well suited to representing a 3D CT volume rather than a single 2D image. The encoder accepts a 384-pixel crop across 64 frames, producing a 1024-dimensional hidden representation. The backbone itself is kept frozen, but is adapted to the CT domain using low-rank adapters (LoRA, $r = 8$, $\alpha = 16$, dropout = 0.15) applied to the attention query, key, and value projections

as well as the output, intermediate, and dense projections. Inference runs under flash attention with a bf16 backbone and fp32 LoRA adapters, balancing throughput with numerical stability in the adapted layers.

A Q-Former bridging module, comprising six cross-attention layers with eight attention heads and 64 learnable query tokens, distills the (potentially very large) set of patch tokens produced by the vision encoder into a fixed-length visual prefix. This fixed-length representation is what allows a volume with a variable number of slices to be consumed by a decoder with a fixed input budget. The Q-Former is trained jointly with the decoder during supervised fine-tuning, rather than as a separate pretraining stage, so that the query tokens learn to extract the visual information most useful for findings generation specifically.

The decoder is a ModernBERT causal language model, initialized from a pretrained checkpoint and fine-tuned end-to-end together with the Q-Former (with the vision backbone held frozen). It generates the findings report up to a maximum of 384 tokens, which is sufficient to cover the longer, more structured narrative expected of a chest CT report.

2.2 Aorta post-processor

Aortic diameter is one of the clearest examples of a finding that benefits from dedicated measurement rather than free-text generation. The aorta post-processor applies TotalSegmentator [2], a 117-label multi-organ segmentation model, to the soft-kernel reconstruction of the CT volume (soft-kernel reconstructions are preferred here because they better preserve the smooth vascular contours needed for accurate diameter measurement). From the resulting aorta segmentation, the maximum diameter is computed at four anatomical levels: ascending aorta, aortic arch, descending aorta, and abdominal aorta. These measurements are then converted into a structured finding sentence describing either normal caliber or dilation/aneurysm at the relevant level, using clinical diameter thresholds consistent with current cardiovascular imaging guidelines.

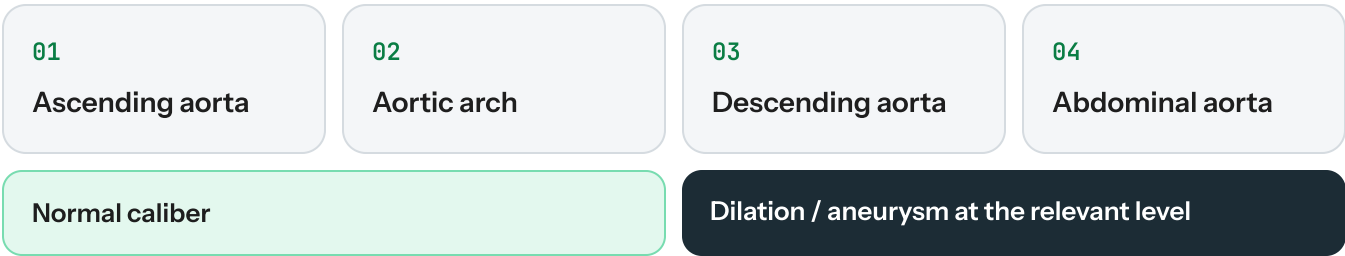


Figure 3. Maximum diameter is measured at four anatomical levels and converted into a structured finding sentence.

2.3 Lung nodule pipeline

Pulmonary nodule characterization is handled by a two-stage pipeline operating on the hard-kernel reconstruction, which better preserves the fine parenchymal detail needed to detect small nodules. In the first stage, a RetinaNet-based 3D detector proposes candidate nodule locations across the volume. In the second stage, a SegResNet segmentation model refines each candidate with voxel-level segmentation, from which a precise size measurement is derived. Detected nodules are assigned to a specific lung lobe using TotalSegmentator-derived lobe masks, and are scored against Lung-RADS-relevant size categories. To keep the generated report clinically manageable, the pipeline caps output at a maximum of 13 reported nodules per case, allocated across size buckets so that the most clinically actionable nodules are prioritized for inclusion.

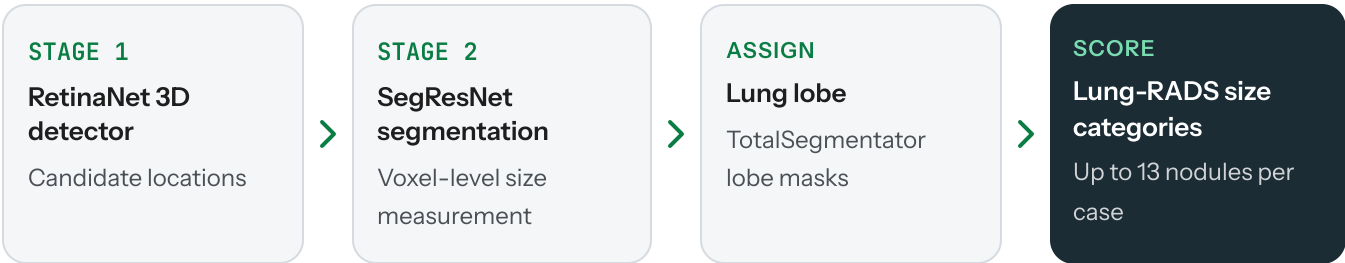


Figure 4. The two-stage lung nodule pipeline, operating on the hard-kernel reconstruction.

2.4 Report combiner

The final stage of the pipeline reconciles the three preceding outputs. A large language model (Claude Sonnet 4.6, accessed via a managed cloud inference service) is prompted to merge the aorta and nodule findings into the VLM-generated draft report, replacing only the sentences that pertain to those two structures while preserving the remainder of the narrative unchanged. This step allows the more fluent, broadly descriptive language of the VLM output to be retained for the majority of the report, while ensuring that the two findings most sensitive to precise measurement are backed by dedicated, verifiable pipelines rather than free-text generation.

In the event that the language model produces an invalid or malformed response, a regex-based fallback mechanism performs the same sentence-replacement operation deterministically, ensuring the pipeline degrades gracefully rather than failing outright.

Dataset

The training corpus is an internal chest CT report-generation dataset in which each example pairs a chest CT volume with its corresponding radiology report. From each report, the findings section is extracted and reorganized into a fixed, section-ordered target: pulmonary nodules first, followed by lungs and airways, mediastinum, heart and great vessels, upper abdomen, bones, support devices, and soft tissues, with empty sections omitted. This normalization ensures that the generation target follows a single canonical structure regardless of how the original report was dictated, which in turn simplifies both training and evaluation.

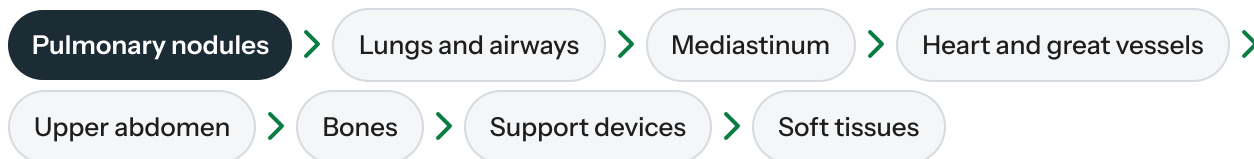


Figure 5. The fixed, section-ordered generation target. Empty sections are omitted.

A central challenge in constructing this dataset was severe class imbalance in the parent corpus. Several of the most clinically important findings, including aortic injury, pulmonary embolism, rib fracture, and pneumothorax, each occurred in under 0.5% of studies, and were consequently overwhelmed by far more common findings under naive random sampling. To address this, a rarest-first equalizing sampling strategy was applied across sixteen binary disease labels: all positive examples of the rarest classes were retained in full, common classes were downsampled, and approximately 25% of each split was reserved for truly normal studies in which all sixteen tracked conditions were absent. This procedure yielded 227,000 unique studies while preserving 100% of the positive examples for the rarest conditions in the corpus.

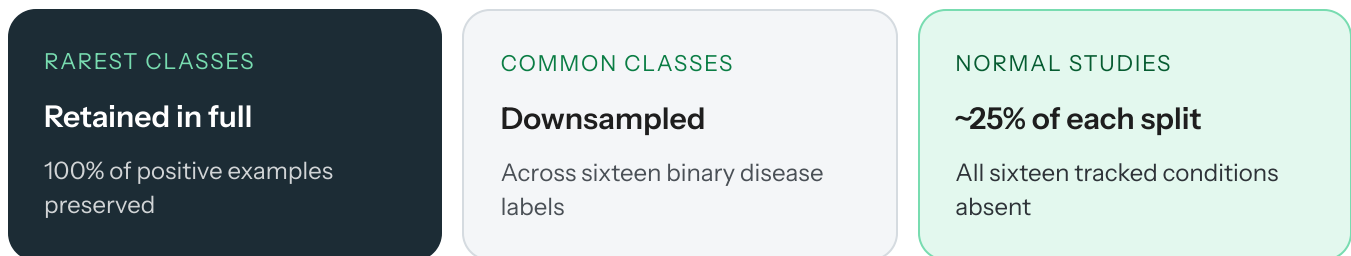
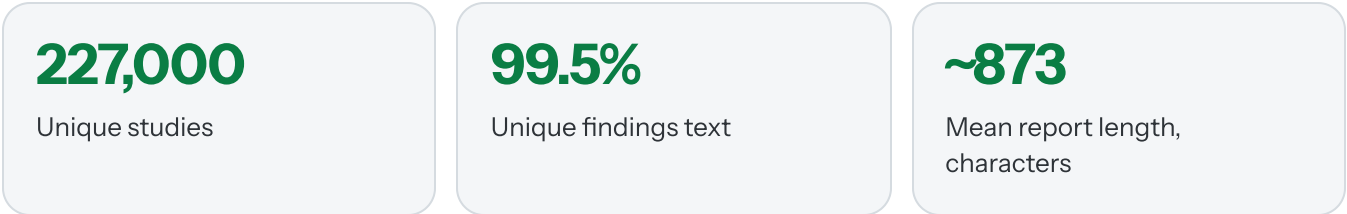


Figure 6. Rarest-first equalizing sampling, yielding 227,000 unique studies.

A secondary preprocessing step addressed template redundancy: many CT reports reuse near-identical boilerplate phrasing, for instance the standardized text produced by coronary artery calcium scoring templates. Where the same findings text recurred at high frequency, only a small number of representative copies were retained, removing redundant stock-template reports. This step increased the proportion of unique findings text in the final corpus to 99.5%, with a mean report length of approximately 873 characters. Table 1 summarizes the resulting training, validation, and test split sizes.



Split	Studies
Training	197,409
Validation	14,918
Test	14,908

Table 1. Dataset split sizes (CTC Narrative Model dataset).

The underlying data is drawn from proprietary chest CT studies with structured radiologist annotations, spanning millions of individual CT slices and predominantly adult patient populations across multiple clinical sites in different regions of the United States. All data used is non-public and proprietary; detailed sample size, effect size, and representativeness information beyond what is summarized here is available upon request under appropriate data-use agreements.

04

Base model training

The base model shares the architecture described in Section 2.1: a frozen, LoRA-adapted V-JEPA 2.1 ViT-L vision encoder ($r = 8$, $\alpha = 16$, dropout = 0.15, applied to attention query/key/value and output/intermediate/dense projections), a 6-layer, 8-head Q-Former

with 64 learnable query tokens trained jointly with the decoder, and a ModernBERT causal decoder generating findings text up to 384 tokens. Volumes are supplied to the model at a 384×384×128 resolution after trilinear resizing.

Training used the AdamW optimizer with a cosine-with-restarts learning rate schedule over three cycles, a learning rate of 5×10^{-5} , an effective batch size of 8, and up to 50 epochs.

OPTIMIZER AdamW Cosine with restarts, 3 cycles	LEARNING RATE 5×10^{-5}	EFFECTIVE BATCH 8	EPOCHS Up to 50
--	--	--	--

4.1 Data preprocessing

Two levels of preprocessing are applied prior to both training and inference. At the volume level, DICOM series are loaded, slices are sorted by z-position, pixel values are converted to Hounsfield units, and in-plane spacing is resampled to 0.7 mm with the z-axis resampled to 2.5 mm target spacing; Hounsfield values are clamped to [-2048, 4095] to avoid interpolation artifacts, and a body mask is used to crop and center-pad the volume to a final 512×512 in-plane resolution. At the model-input level, the resulting volume is windowed to a center of 0 and width of 2000 Hounsfield units (i.e., [-1000, 1000] HU), linearly scaled to [0, 1], and trilinearly resized to the 384×384×128 shape expected by the vision encoder.

Only series meeting minimum quality criteria are used for either fine-tuning or inference: a minimum of 50 slices, a maximum slice thickness of 5.0 mm, a minimum z-axis coverage of 200.0 mm, axial orientation, and confirmation that the body part examined is chest. Where multiple qualifying series exist for a study, one hard-kernel series is selected for the vision-language model and nodule pipeline, and one soft-kernel series is selected for the aorta post-processor, using a deterministic priority order based on z-axis coverage, in-plane resolution, slice thickness, and reconstruction kernel.

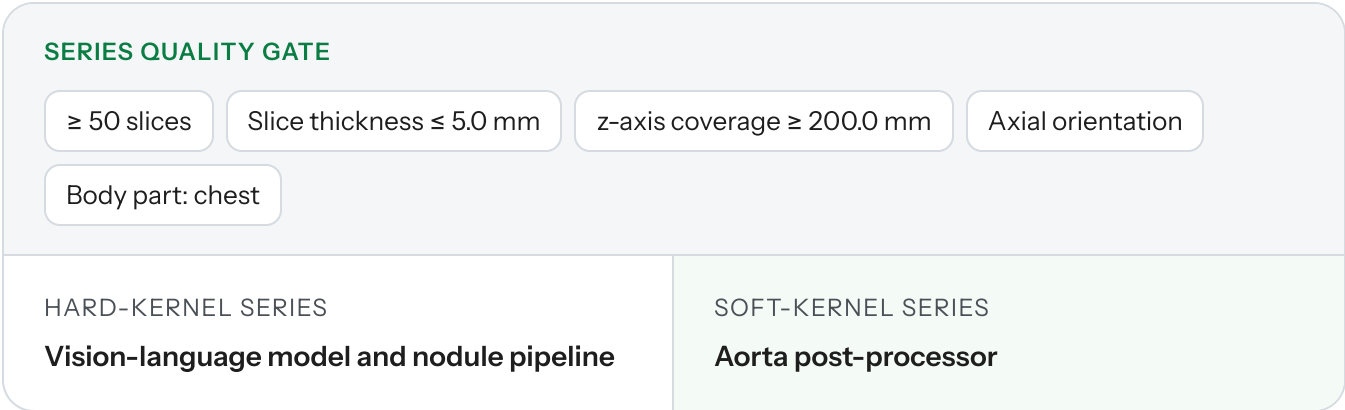


Figure 7. Minimum quality criteria and kernel routing for qualifying series.

05

Performance and validation

Overall narrative quality is assessed by comparing generated reports against reference radiologist reports using two independent text classifiers, each of which extracts a fixed set of clinical findings from both the generated and reference text; the two label sets are then compared using standard F1 metrics, treating the reference labels as ground truth. The first classifier, F1RadBERT-CT, uses an 18-finding label set from an external, published CT report classifier [3]. The second, F1NeuroCheXbert-CT, is an internally developed classifier covering 16 CT conditions plus an automatic “no finding” label, using a multi-head ModernBERT-large model that predicts a four-level certainty per condition, collapsed to present-versus-absent for F1 computation.

For both metrics, micro-F1 pools all findings into a single score and is dominated by frequent findings; macro-F1 averages per-finding F1 equally, so rare findings count as much as common ones (and is consequently the more conservative figure); and weighted-F1 sits between the two, weighting each finding’s contribution by its prevalence. Table 2 summarizes the results.

Metric	Micro-F1	Macro-F1	Weighted-F1
F1RadBERT-CT	0.506	0.378	0.500
F1NeuroCheXbert-CT	0.559	0.332	0.551

Table 2. Narrative findings evaluation (F1). The higher value in each column is shaded.

The gap between micro- and macro-F1 for both metrics reflects the underlying class imbalance discussed in Section 3: common findings are captured reasonably well, while rarer conditions, which are weighted equally under macro-F1, remain comparatively harder to generate correctly, despite the rarest-first sampling strategy applied during dataset construction.

Pulmonary nodule detection was evaluated separately on a 360-case test set, using detection thresholds fit on the validation set, with sensitivity measured against a corresponding false-positive budget per series (Table 3).

Size bucket	Sensitivity	FP/series budget
4–6 mm	0.692	≤2.0
6–8 mm	0.823	≤2.0
8–15 mm	0.886	≤0.75
≥15 mm	1.000	≤0.75

Table 3. Pulmonary nodule detection sensitivity by size bucket.

The model met the pre-specified sensitivity targets for nodules in the 4 to 6 mm and ≥15 mm size buckets, but narrowly missed the target for 6 to 8 mm nodules (0.823 achieved against a target of ≥0.85), and fell further short for 8 to 15 mm nodules (0.886 achieved against a target of ≥0.95), indicating that mid-sized nodule sensitivity is the primary area for continued improvement.

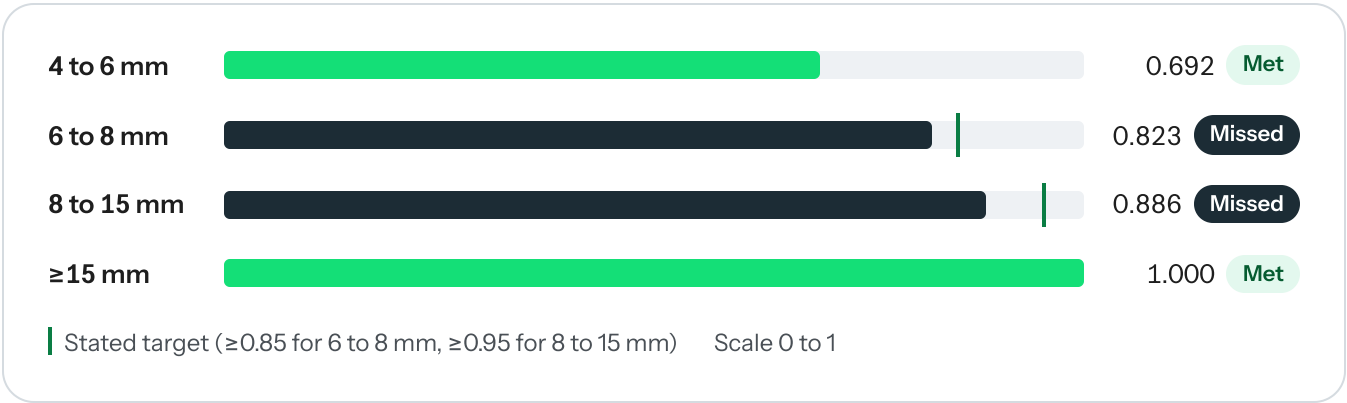
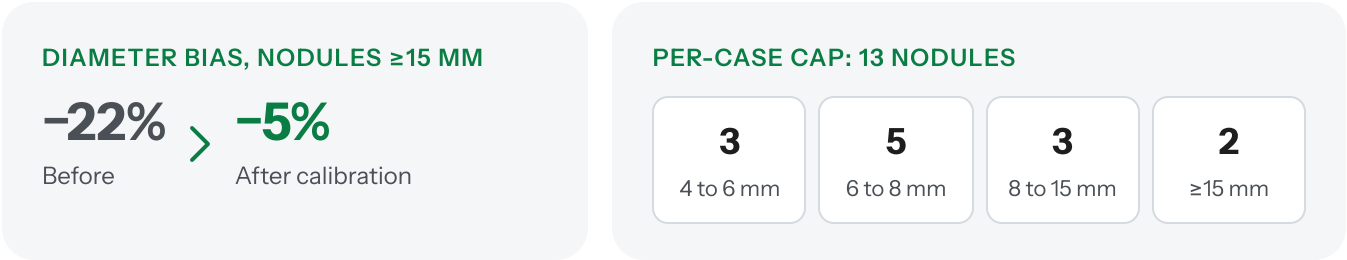


Figure 8. Nodule detection sensitivity by size bucket against the pre-specified targets, drawn from Table 3.

For nodule density classification in binary mode, recall was 0.748 for solid nodules and 0.726 for subsolid nodules. A linear diameter calibration step substantially reduced measurement bias for nodules 15 mm and larger, from -22% before calibration to -5% after calibration. Consistent with the per-case reporting cap described in Section 2.3, output is limited to 13 nodules per case, allocated as 3, 5, 3, and 2 nodules across the four size buckets in Table 3, respectively.



Aortic ectasia detection and measurement were validated on 3,939 evaluable chest CT studies against radiologist-extracted aortic findings, with performance reported at the case level for any dilation or aneurysm versus normal caliber (Table 4).

Metric	Value
Sensitivity	95.2%
Specificity	50.3%
Positive Predictive Value	42.6%

Table 4. Aortic ectasia detection performance.

The high sensitivity and comparatively low specificity and positive predictive value indicate that the aorta post-processor is currently tuned toward a conservative, screening-oriented operating point: it is unlikely to miss a truly dilated aorta, at the cost of a substantial false-positive rate that would require downstream review. Per-segment measurement accuracy was strongest for the ascending aorta, with a mean absolute error of 2.3 mm and a bias of +0.6 mm ($n = 1,021$), which is considered adequate for screening-level triage; diameter thresholds follow current ACC/AHA guidelines for aortic disease.



EVALUATE ON YOUR OWN DATA

All reported metrics were obtained on proprietary internal datasets, with limited external validation to date. Developers integrating this model are expected to evaluate performance on their own data and target population, and may need to adjust decision thresholds to optimize the sensitivity/specificity trade-off for their specific use case, particularly for the aorta and mid-sized nodule findings discussed above.

06

Limitations

Several limitations should be considered when interpreting the results in Section 5. The training data is drawn primarily from adult populations at outpatient imaging centers, so performance on pediatric patients, rare pathologies, or inpatient-acquired studies has not been established and may be reduced. The model is intended for chest CT series specifically; inclusion of other body regions is expected to reduce performance. Studies acquired with atypical protocols, such as unusually thick slices or non-standard reconstruction kernels, may also be underrepresented in training. At the inference layer, studies with fewer than 50 slices or less than 200 mm of z-axis coverage are excluded outright and will not produce a result, and the training data shows limited scanner vendor diversity, with one manufacturer's scanners overrepresented relative to others.

Demographic diversity in the training data is constrained to what is available across the proprietary datasets used, with only limited information on race or ethnicity captured, and partners are encouraged to assess model performance specifically within their own target population before relying on it. There are no known gaps in core functionality at this time. Finally, the performance figures reported here were obtained on hardware that differs from the configuration used for the current beta release, so exact performance under production conditions may vary somewhat from the figures presented in Section 5.

Conclusion

This paper described the Neurologic AI Chest CT (CTC) Narrative Model, a beta-stage pipeline that combines a vision-language model for free-text findings generation with two dedicated measurement sub-pipelines for aortic diameter and pulmonary nodule characterization, reconciled into a single report by a large language model. Trained on a rarest-first-equalized corpus of 227,000 chest CT studies, the pipeline achieves a weighted F1 of 0.551 against an internal 16-condition classifier, meets its sensitivity targets for the smallest and largest pulmonary nodule size categories, and detects aortic ectasia with 95.2% sensitivity at the case level. Together, these results support the premise that decomposing chest CT reporting into narrative generation and targeted, verifiable measurement is a promising strategy for approaching the accuracy demands of volumetric radiology reporting. At the same time, the results identify specific gaps, most notably mid-sized pulmonary nodule sensitivity and aortic specificity, that will need to be closed before the model is suitable for uses beyond developer evaluation and further validation.

THE PREMISE

Together, these results support the premise that decomposing chest CT reporting into narrative generation and **targeted, verifiable measurement is a promising strategy for approaching the accuracy demands of volumetric radiology reporting.**

References

1. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., & Ballas, N. (2025). V-JEPA 2: Self-Supervised Video Models Enable Understanding, Generation, and Planning. *arXiv preprint arXiv:2506.09985*.

2. Wasserthal, J., Breit, H.-C., Meyer, M. T., Pradella, M., Hinck, D., Sauter, A. W., Heye, T., Boll, D. T., Cyriac, J., Yang, S., Bach, M., & Segeroth, M. (2023). TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence*, 5(5), e230024.
3. Hamamci, I. E., Er, S., Wang, C. et al. (2026). Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*. <https://doi.org/10.1038/s41551-025-01599-y>