State-of-the-Art Text-Prompted Medical Segmentation Models Struggle to Ground Chest CT Findings

Mohammed Baharoon[*,1] MOHAMMED_BAHAROON@HMS.HARVARD.EDU
Luyang Luo[*,1]
Michael Moritz[2,3] MICHAEL.MORITZ@SSMHEALTH.COM
Abhinav Kumar[4] ABHINAV.KUMAR@ICAHN.MSSM.EDU
Sung Eun Kim[1,5] SUNGEUN_KIM2@HMS.HARVARD.EDU
Xiaoman Zhang[1] XIAOMAN_ZHANG@HMS.HARVARD.EDU
Miao Zhu[6] MZHU14@BWH.HARVARD.EDU
Kent Kleinschmidt[2] KENT.KLEINSCHMIDT@HEALTH.SLU.EDU
Sri Sai Dinesh Jaliparthi[2] DINESH.JALIPARTHI@HEALTH.SLU.EDU
Sathvik Suryadevara[2] SATHVIK.SURYADEVARA@HEALTH.SLU.EDU
Rithvik Akula[2] RITHVIK.AKULA@HEALTH.SLU.EDU
Mark Marino[2] MARK.MARINO@HEALTH.SLU.EDU
Wenhui Lei[7] WENHUI.LEI@SJTU.EDU.CN
Ibrahim Ethem Hamamci[8] IBRAHIM.HAMAMCI@UZH.CH
Pranav Rajpurkar[1]

[1] *Department of Biomedical Informatics, Harvard Medical School, Boston, MA*
[2] *Saint Louis University School of Medicine, St. Louis, MO*
[3] *SSM Health, St. Louis, MO*
[4] *Icahn School of Medicine at Mount Sinai, New York, NY*
[5] *National Strategic Technology Research Institute, Seoul National University Hospital, South Korea*
[6] *Brigham and Women's Hospital, Boston, MA*
[7] *Shanghai Jiaotong University, Shanghai, CN*
[8] *University of Zurich, Switzerland*

Abstract

This study presents a comprehensive evaluation of state-of-the-art text-prompted segmentation models, including MedSAM2, SegVol, SAT, and BiomedParse, on ReXGroundingCT, a novel dataset that pairs chest CT findings with corresponding segmentation masks. Our results demonstrate that despite recent advances, current models struggle to accurately segment diverse findings from chest CTs, particularly when dealing with non-focal abnormalities described in natural language reports. While existing models are primarily optimized for fixed categorical labels rather than nuanced clinical descriptions, even fine-tuning these models with free-text descriptions yields limited improvement in segmentation accuracy. These insights highlight that report grounding on 3D medical volumes through segmentation remains an open challenge, necessitating future models that better comprehend complex clinical language and irregular object patterns across volumetric data. The code base and data can be found at: https://github.com/rajpurkarlab/ReXGroundingCT.

[*] Equal contribution.

1. Introduction

Medical Image Segmentation. Segmentation is a fundamental task in medical image analysis, enabling precise quantification, diagnosis, and treatment planning (Rayed et al., 2024). Recent advances in this field have been driven by foundation models such as the Segment Anything Model (SAM) (Kirillov et al., 2023), which revolutionized image segmentation by providing a versatile framework for identifying objects in images through interactive prompts. The adaptation of these foundation models for medical images using large-scale medical datasets, with models such as MedSAM (Ma et al., 2024), has further enhanced the segmentation capabilities of anatomical structures and pathologies across diverse medical modalities.

Gaps between SoTA Approaches and Practical Requirements. Despite recent advances, there remains a significant gap between the capabilities of current models and the practical requirements of clinical applications, particularly in localizing clinical findings. In real-world healthcare settings, radiological reports are predominantly unstructured and conveyed through free text descriptions, presenting the fundamental challenge of mapping these textual findings to their precise spatial locations within complex imaging studies such as chest CT scans. Radiologists describe abnormalities using natural language, with varying degrees of specificity, from descriptive findings such as "a 5mm nodule in the lower lobe of the left lung" to more complex findings involving diffuse patterns, subtle morphological changes, and spatial relationships.

Report Grounding. The ability to automatically link these narrative descriptions to corresponding segmentations in medical images, a task known as report grounding (Bannur et al., 2024; Zou et al., 2024; Chen et al., 2023; Ichinose et al., 2023), would mark a critical advancement in the field. Referring physicians often struggle to localize findings described in reports when reviewing complex imaging studies, which can lead to diagnostic inefficiencies and increased demands for radiology consultations (Iyer et al., 2010). Moreover, patients, who typically receive their imaging studies and textual reports simultaneously, are rarely equipped with the clinical expertise needed to contextualize the findings (Alarifi et al., 2021, 2024). Similarly, medical trainees often face challenges in building visual diagnostic skills, as the absence of visual references tied to narrative reports makes it difficult to translate textual descriptions into spatially accurate interpretations (Brady, 2018). ReXplain (Luo et al., 2024) is an early attempt to ground findings to anatomical regions for explanatory purposes, but it does not support lesion-level precision as targeted by text-driven segmentation.

Challenges. Existing segmentation models face several challenges in this context. First, they typically focus on segmenting well-defined objects or anatomical structures with clear boundaries, whereas many radiological findings are diffuse, subtle, or have irregular shapes (Zhao et al., 2023; Du et al., 2024; Ma et al., 2025). Second, most models rely on bounding box prompts or predefined categorical labels, limiting their ability to understand and process the rich, variable language used in clinical reports (Kirillov et al., 2023; Zhao et al., 2025; Ma et al., 2024, 2025). While a radiologist might describe a "speculated 8mm nodule in the right upper lobe with surrounding ground-glass opacity," current models can only process simplified and rigid prompts like "mass" or "nodule in CT" (Zhao et al., 2023; Du et al., 2024; Zhao et al., 2025). Lastly, volumetric medical data like CT scans contain hundreds of

Table 1: Comparison between different public datasets with ReXGroundingCT. While there are datasets that provide manually annotated findings from sentence-level findings like PadChest-GR, ReXGroundingCT is the only public dataset that provide manually annotated 3D chest CT segmentations of sentence findings. The ReXGroundingCT dataset is now further expanded to 3,142 cases (Baharoon et al., 2025). † Reports in AbdomenAtlas 3.0 are generated using LLMs and segmentation morphology.

Dataset	Dim.	Type	Modality	Size	Sentence-level Findings	Method	Annotation
MS-CXR[1]	2D	Findings	CXR	1,047	✓	Manual	Bbox
PadChest-GR[2]	2D	Findings	CXR	4,555	✓	Manual	Bbox
VinDr-CXR[3]	2D	Findings	CXR	18,000	×	Manual	Bbox
CheXmask[4]	2D	Organs	CXR	657,463	×	Auto	Bbox
BiomedParseData[5]	2D	Findings, Organs	Multiple	1.1M	×	Mix	Seg.
MedTrinity-25M[6]	2D	Findings, Organs	Multiple	25M	✓	Mix	Bbox
LUNA16[7]	3D	Nodules	Chest CT	888	×	Manual	Bbox
LUNA25	3D	Nodules	Chest CT	4,069	×	Manual	Point
RadGenome-Brain MRI[8]	3D	Findings, Regions	Brain MRI	3,408	×	Auto	Seg.
RadGenome-Chest CT[9]	3D	Organs	Chest CT	25,692	×	Auto	Seg.
AbdomenAtlas 3.0[†,10]	3D	Tumors, Regions	Abdominal CT	9,262	✓	Auto	Seg.
ReXGroundingCT[11]	3D	Findings	Chest CT	1,914	✓	Manual	Seg.

slices, requiring models to maintain contextual understanding across a large spatial extent (Zhao et al., 2023; Du et al., 2024; Ma et al., 2025).

Our Approach and Contributions. In this work, we construct a comprehensive dataset, ReXGroundingCT, of chest CT scans paired with segmentation masks and their corresponding natural-language descriptions from radiology reports. Using this dataset, we provide the first comprehensive benchmark of state-of-the-art segmentation models on the task of localizing findings described in natural language. We benchmark SAT (Zhao et al., 2023), BiomedParse (Zhao et al., 2025), SegVol (Du et al., 2024), SAM2 Ravi et al. (2024) and MedSAM2 (Ma et al., 2025) on this task. Our analysis shows that current state-of-the-art approaches fail at the task of grounding 3D chest reports in natural language, necessitating the need to develop better methods that can comprehend complex clinical language and irregular anatomical patterns across volumetric data.

Generalizable Insights about Machine Learning in the Context of Healthcare

- Current segmentation models demonstrate a significant gap between technical capability and clinical utility, particularly when tasked with interpreting natural language descriptions to localize findings in volumetric medical data.
- 3D context is crucial for accurate medical image segmentation, with true 3D models consistently outperforming 2D approaches highlighting the importance of spatial context in accurate segmentation.

1: Boecking et al. (2022), 2: Castro et al. (2024), 3: Nguyen et al. (2022), 4: Gaggion et al. (2024), 5: Zhao et al. (2025), 6: Xie et al. (2024), 7: Setio et al. (2017), 8: Lei et al. (2024), 9: Zhang et al. (2024), 10: Bassi et al. (2025), 11: Baharoon et al. (2025)

2. Related Work

2.1. Medical Vision-Language Datasets

Chest X-ray Datasets. Recent years have witnessed significant growth in datasets connecting medical imaging with textual descriptions. The majority of these datasets focus on chest X-rays, with notable examples including MIMIC-CXR (Johnson et al., 2019), which contains over 377,000 chest X-rays with corresponding radiology reports, and CheXpert-Plus (Chambon et al., 2024), which features 223,228 chest X-ray images with 187,711 unique reports. PadChest (Bustos et al., 2020) is another chest dataset that includes more than 160,000 X-rays with their associated reports. PadChest-GR (Castro et al., 2024) augments 4,555 PadChest studies with bounding box annotation for positive and negative findings, facilitating the task of grounded report generation (Bannur et al., 2024).

3D Datasets. For volumetric data, CT-RATE (Hamamci et al., 2024) represents one of the first large-scale datasets containing both CT scans and corresponding radiology reports, though it lacks localized annotations of findings mentioned in the reports. RadGenome-Chest CT (Zhang et al., 2024) provides a grounded vision-language dataset specifically for chest CT analysis, connecting radiological findings with their spatial locations. However, the dataset contains organ-level segmentation masks generated using SAT (Zhao et al., 2023), and not segmentations of the findings themselves. RadGPT (Bassi et al., 2025) and PASTA (Lei et al., 2025) focused on constructing synthetic or semi-synthetic 3D image-text tumor datasets that pair CT volumes with detailed descriptions of oncological findings. Despite these advances, there remains a critical gap in datasets that map free-text descriptions of radiological findings with their localized segmentations, particularly for volumetric data like chest CT. This limitation has hampered progress in developing models capable of grounding natural language descriptions to specific regions in medical images. Instead, ReXGroundingCT provides the first manually annotated large-scale dataset of chest CT scans where each segmentation mask is explicitly linked to a corresponding free-text description of a radiological finding. A comparison between different datasets is shown in Table 1.

2.2. Interactive Segmentation for Medical Images

Many interactive segmentation models have been proposed and adapted for medical images. The Segment Anything Model (SAM) (Kirillov et al., 2023) is a foundation model capable of segmenting arbitrary objects according to various prompting mechanisms, including points and bounding boxes, demonstrating remarkable zero-shot generalization capabilities across diverse domains. MedSAM (Ma et al., 2024) fine-tuned SAM on medical datasets spanning multiple modalities and anatomical regions to better capture the specific characteristics of medical images. SAM2 (Ravi et al., 2024) further extended the capabilities of SAM by implementing a more efficient architecture and adding support for video segmentation, while MedSAM2 (Ma et al., 2025) adapted these improvements specifically for volumetric medical data, by fine-tuning the model on more than 450,000 3D image-mask pairs. ScribblePrompt (Wong et al., 2024) introduced a fast and flexible interactive segmentation approach applicable to any biomedical image, using scribble-based user interactions. More recently, nnInteractive (Isensee et al., 2025) redefined 3D promptable segmentation by enabling real-time interactive editing of volumetric medical data. While these models have demonstrated

impressive performance in segmenting anatomical structures and pathologies, they largely rely on explicit spatial prompts rather than textual descriptions, limiting their applicability in clinical workflows where findings are typically documented in natural language.

2.3. Text-prompted Segmentation Models

Recent work has explored grounding anatomical structures and pathological conditions in medical images using text prompts (Zhao et al., 2025, 2023; Du et al., 2024). BiomedParse (Zhao et al., 2025) is a promptable 2D segmentation model capable of localizing organs and conditions based on textual input. It was trained on 45 public datasets using canonical labels derived from a unified ontology, covering organs, abnormalities, and histologies, generated with GPT-4 (Achiam et al., 2023). SegVol (Du et al., 2024) combines text and spatial prompts to segment over 200 anatomical categories, by training on more than 95k CT scans. The model uses a zoom-out-zoom-in method to facilitate precise inference on volumetric images. SAT (Zhao et al., 2023), can segment 3D structures and lesions from CT scans using text prompts, encompassing a larger ontology of over 6000 anatomical terminologies. During the first stage, SAT learns to differentiate between different organs by using a three-way contrastive learning with visual embeddings, organ embeddings, and embeddings of organ descriptions gathered from various sources on the internet. Subsequently, the model is fine-tuned to segment anatomical regions using text prompts. CAT (Huang et al., 2024) proposed a novel approach for coordinating anatomical-textual prompts specifically for multi-organ and tumor segmentation, bridging the gap between spatial and textual prompts. While these models demonstrate strong performance in segmenting anatomical structures, they were not designed to handle the more complex task of segmenting clinical findings from radiology reports, which is more difficult due to the inherent variability and ambiguity of natural language descriptions. Ichinose et al. (2023) present the first visual grounding framework for linking full radiology reports to segmentation masks in 3D CT scans; however, their model was trained on bounding boxes rather than pixel-level masks, relied on conventional CNN and LSTM backbones instead of current state-of-the-art architectures, and neither the dataset nor model weights were made publicly available.

3. Curation of ReXGroundingCT

In this section, we describe the process of curating ReXGroundingCT, a novel dataset that maps 1,914 3D Chest CT findings to their localized segmentation masks. We divide our dataset curation process into three principal stages: (1) Abnormality Extraction, (2) Abnormality Categorization, and (3) Dataset Labeling.

3.1. Abnormality Extraction

We leverage the CT-RATE dataset (Hamamci et al., 2024), which comprises 25,692 de-identified CT scans and radiology reports originally written in Turkish and subsequently machine-translated into English. A qualitative review by a board-certified radiologist revealed frequent use of non-standard medical terminology and phrasing, likely as a result of translation artifacts. To mitigate these issues, we use GPT-4 (Achiam et al., 2023) to systematically rewrite each report in standardized medical language. During this rewriting

step, GPT-4 was also instructed to group findings by anatomical region (e.g. lungs, mediastinum, vasculature), following conventions commonly used in US radiology practice. The prompt used for this process is provided in Supplementary Figures S2 and S3.

Next, we instructed GPT-4 to extract each distinct abnormality from the rewritten reports, including relevant descriptors (e.g. lesion size, number of lesions, presence of diffuse pathologies such as "centrilobular emphysema" or "pleural effusion"). This process generated a structured list of findings for each report. For each extracted finding, we instruct GPT-4 to output which original sentence(s) (by code) the abnormality comes from, whether it is an abnormal finding or not, and whether there is any textual reference to prior studies (e.g. "stable," "new"). This metadata allows us to locate the sentences in the original report that each annotated abnormality refers to, filter out negative findings when deciding which abnormalities to annotate, and label any references to prior comparisons. Figure 1 illustrates this two-step abnormality extraction process using an example radiology report. The prompt for this step is provided in Supplementary Figures S4 and S5.

3.2. Abnormality Categorization

Following extraction, each abnormality is assigned a category from a 2-level hierarchical schema, with 12 categories and 61 subcategories (see Supplementary Figure S9). This schema was developed under guidance from a board-certified radiologist to reflect the diverse range of abnormal findings encountered in clinical practice. It covers both focal (e.g. solitary nodules, localized masses) and non-focal (e.g. diffuse emphysema) lung abnormalities, as well as a broad set of cardiac, vascular, skeletal, and soft-tissue findings.

We use GPT-4o-mini (Hurst et al., 2024), a smaller and more cost-effective model, to categorize the extracted findings. Specifically, GPT-4o-mini receives the text of each abnormal finding and outputs one of the 61 subcategories. The prompt used for this step is provided in Supplementary Figures S6, S7, and S8. To assess classification accuracy, we manually review a subset of categorized findings and report macro-averaged precision and recall across all classes.

3.3. Dataset Annotation

After abnormality extraction and categorization, we selected 1,914 CT scans for annotation by a team of medically trained personnel. We selected the scans in groups, ordered by increasing CT-RATE ID, which represents a random sample since the IDs are randomly assigned. Each scan underwent pixel-level annotation for the extracted abnormal findings. Some findings were excluded because they fell outside the lung or pleura (e.g., vascular or upper abdominal findings), could not be localized in the CT scan, were too diffuse to segment reliably (e.g., "mosaic attenuation"), or represented normal anatomical variants (e.g., subcentimeter non-pathological lymph nodes).

A senior radiologist conducted quality checks on the annotated scans to verify labeling accuracy and consistency across cases. The final curated dataset serves as a unique resource for training and evaluating abnormality segmentation models, as well as for developing grounded CT report generation models.

Findings: The diameter of the ascending aorta was 43 mm and increased. In the mediastinum and bilateral hilar regions, several lymph nodes are observed on the forehead with a short diameter of 5 mm. Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. No pathological increase in wall thickness was observed in the esophagus. There are nodules in both lungs, the largest measuring 9x5x6 mm in the anterior segment of the right lung upper lobe. No mass or infiltrative lesion was detected in both lungs. As far as it can be monitored within the limits of non-contrast CT; There is no discernible mass in the upper abdominal organs. An echogenic stone with a diameter of 4 mm is observed in the middle zone of the left kidney. Bridging osteophytes are observed at the corners of the thoracic vertebra corpus.

Impressions: Nodules of stable number and size in both lungs at an interval of 2.5 years. Enlargement of the ascending aorta.

Step1: Rephrase and structure sentences

Lungs/Airways/Pleura: The trachea and both main bronchi are patent with no evidence of obstruction. There are multiple nodules in both lungs, with the largest measuring 9x5x6 mm in the anterior segment of the right upper lobe. No masses or infiltrative lesions are identified within either lung.
Heart/Vessels: The ascending aorta is dilated, measuring 43 mm in diameter.
Mediastinum/Hila: Several mediastinal and bilateral hilar lymph nodes are noted, the largest with a short axis of 5 mm. No pathological thickening is observed within the esophageal walls.
Chest wall/Axilla:
Lower neck:
Bones: Bridging osteophytes are seen at the corners of the thoracic vertebral bodies.
Upper abdomen: Non-contrast CT imaging is limited, but there is no discernible mass within the upper abdominal organs. A 4 mm echogenic calculus is present in the mid zone of the left kidney.
Impressions: Stable pulmonary nodules in both lungs over a 2.5-year interval. Ascending aortic dilatation.

Step2: Extract distinct findings

Phrase	Sentence Codes	Abnormality	Prior Report
Patent trachea and both main bronchi with no evidence of obstruction	L1	No	No
Multiple nodules in both lungs	L2, I0	Yes	Yes
Largest nodule measuring 9x5x6 mm in the anterior segment of the right upper lobe	L2	Yes	Yes
No masses or infiltrative lesions within either lung	L3	No	No
Ascending aorta dilated, measuring 4.3 cm in diameter	H0, I1	Yes	No
Several mediastinal and bilateral hilar lymph nodes, largest with a short axis of 5 mm	M0	No	No
No pathological thickening within the esophageal walls	M1	No	No
Bridging osteophytes at the corners of the thoracic vertebral bodies	B0	Yes	No
No discernible mass within the upper abdominal organs	A0	No	No
4 mm echogenic calculus in mid zone of left kidney	A1, I2	Yes	No

Figure 1: Two-step abnormality extraction process. Step 1 shows the report rephrased and structured by anatomical regions. Step 2 displays the extracted abnormalities with corresponding sentence codes, abnormality status, and prior report references.

3.4. Manual Dataset Validation

To assess the performance of our multi-step abnormality extraction and categorization pipelines, we conducted a comprehensive manual review of the outputs from both stages across 100 randomly selected reports. We evaluated four error types during this review. A misclassification error occurred when an abnormality was placed in an incorrect anatomical region; this applied only to the anatomical grouping step. A descriptor error indicated that a finding was missing essential details such as location, shape, or measurement. A false positive error was recorded if the model hallucinated an abnormality that did not exist in the original report, while a false negative error occurred when the model failed to identify or incorrectly negated a finding that was present in the report.

Table 2: Mean error rates with 95% confidence intervals for each error type in the abnormality extraction pipeline, evaluated on 100 randomly sampled reports.

Error Type	**Step 1**	**Step 2**
Misclassification Error	0.11 [0.05, 0.17]	N/A
Descriptor Error	0.14 [0.08, 0.19]	0.13 [0.10, 0.15]
False Positive Error	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
False Negative Error	0.05 [0.04, 0.06]	0.01 [0.00, 0.02]

The error rate for each type was computed as the number of findings with that error divided by the total number of findings in each reviewed report. Table 2 summarizes the mean error rates and 95% confidence intervals for each type. Notably, we observed zero false positive errors in both extraction steps, suggesting that the model did not hallucinate findings that were not present in the original reports. Misclassification errors were observed only during the first stage, with a mean error rate of 0.11. Descriptor errors were more common, with mean error rates of 0.14 in Step 1 and 0.13 in Step 2. False negative errors were rare, with rates of 0.05 and 0.01 for Step 1 and Step 2, respectively.

For abnormality categorization, our GPT-4o-mini classifier achieved macro-averaged precision and recall of 0.87 and 0.83, respectively, across all 61 subcategories. These results demonstrate strong performance in assigning findings to the appropriate subcategories within our hierarchical schema.

3.5. Dataset Statistics

Our final dataset consists of 1,914 CT scans, comprising 4,776 findings along with their corresponding segmentation masks. We allocated 1,614 scans for training, 100 for hyperparameter tuning, and 200 for testing. We focus exclusively on lung and pleural findings, specifically Categories 1 and 2 in Supplementary Figure S9. The majority of findings (78%) fall under Category 1, "Typically focal lung, airway, or pleural opacities," while the remaining (22%) are classified under Category 2, "Typically non-focal lung, airway, or pleural abnormalities." A detailed breakdown of the subcategory counts is provided in Table 3.

Our sampled 1914 cases from CT-RATE had an average age of 41.7 $\pm$ 14.6 years, and CT-RATE had an average age of 48.7 $\pm$ 17.3 years. 56.3% of the sampled cases come from male patients, which is similar to the 58.4% in the whole CT-RATE. The mean depth of the CTs is 357 $\pm$ 118 slices.

3.6. CT Preprocessing

To preprocess the CT volumes, we first convert raw voxel values to Hounsfield Units (HU) using the provided slope and intercept in CT-RATE. Subsequently, we apply an HU window of [–600, 1500] to capture the relevant intensities for lung-related findings. Each axial slice is then resized to 512×512 (only downsized if required by the method), while the depth (z-axis) resolution is left unchanged. Finally, we normalize the intensities by rescaling pixel values to the [0, 1] range for input into neural models. If a method uses other preprocessing steps (e.g., different normalization or resizing strategies), we apply those accordingly.

Table 3: Counts of extracted abnormalities grouped by hierarchical category and subcategory in the final dataset.

Category	Subcategory	Count (%)
Typically non-focal lung abnormalities	Bronchial wall thickening	147 (3.1%)
	Bronchiectasis	188 (3.9%)
	Emphysema	281 (5.9%)
	Septal thickening	118 (2.5%)
	Micronodules	219 (4.6%)
	Other	86 (1.8%)
Typically focal lung opacities	Linear	710 (14.9%)
	Atelectasis, consolidation	826 (17.3%)
	Groundglass opacity	906 (19.0%)
	Pulmonary nodules and masses	1121 (23.5%)
	Pleural effusion or thickening	120 (2.5%)
	Honeycombing	10 (0.2%)
	Pneumothorax	11 (0.2%)
	Other	33 (0.7%)

4. Implementation of SoTA Approaches

In this section, we present our experimental approaches for medical image segmentation using text prompts. We first evaluated state-of-the-art segmentation models without adaptation, utilizing bounding box prompts as comparisons (SAM2, MedSAM2). Then, we fine-tuned several medical image segmentation models that support text input (SAT, SegVol, BiomedParse, MedSAM2 Text) on ReXGroundingCT, with implementation details described below. We aimed to train each model for approximately one day on two A100 80GB GPUs.

4.1. Box Prompted Methods

SAM2. SAM2 is a foundation model for visual segmentation that can process both images and videos using various prompt types (Ravi et al., 2024). For a given finding, bounding boxes are generated to contain the entire ground truth, with an additional random shift of 5 pixels in the height and width dimensions, to simulate imprecise box annotations. For sparse findings that encompass many different entities, we generate a bounding box for each entity separately.

MedSAM2. MedSAM2 is a medical image segmentation model, which builds upon SAM2, and is designed to handle diverse medical imaging tasks across multiple modalities (Ma et al., 2025). Similar to the SAM2 baseline, MedSAM2 utilizes bounding box prompts generated to encompass the ground truth segmentation.

4.2. Text Prompted Models

SAT. SAT is a universal medical image segmentation model, originally developed to segment 497 predefined anatomical and pathological categories across multiple imaging modalities (Zhao et al., 2023). To adapt SAT for free-text finding segmentation, we fine-tuned the SAT Pro variant on our dataset. We specifically utilized the BioLord (Remy et al., 2023) variant of the text encoder and updated all model components, including the text encoder and projector, to enable more flexible segmentation capabilities. The fine-tuning process involved 40,000 training steps (approximately 25 epochs), with a learning rate of 1×10^{-4} and a warm-up period of 2,000 steps. We adopted the default crop and patch sizes of $288 \times 288 \times 64$ and $32 \times 32 \times 32$, respectively. Notably, we encountered training instability issues when using FP16 precision, which resulted in NaN losses within the first 5,000 steps. Switching to BF16 resolved this issue and led to more stable training. During inference, SAT uses a 3D sliding window strategy to cover the entire CT volume.

BiomedParse. BiomedParse is a biomedical foundation model designed for comprehensive image parsing across 9 imaging modalities, originally developed to perform segmentation, detection, and recognition of biomedical objects (Zhao et al., 2025). To enable free-text segmentation, we fine-tuned the full model on our dataset, using only slices that contain ground truth annotations. However, as BiomedParse is a 2D model, each slice was processed independently, which limits its ability to handle findings that span multiple slices. Fine-tuning was performed for 10 epochs, with an image size of 512×512 and a learning rate of 1×10^{-4}. During inference, BiomedParse operates in a slice-by-slice manner without volumetric context.

SegVol. SegVol is a universal and interactive model for volumetric medical image segmentation, capable of processing 3D medical images with various prompts, including points, bounding boxes, and text descriptions (Du et al., 2024). Natively, SegVol supports segmentation for over 200 anatomical categories. We fine-tuned the model on our ReXGroundingCT dataset to adapt it for report grounding. Since we do not have predefined label categories, we did not use SegVol's pseudo masks. Furthermore, we only used text prompts in the fine-tuning process, so we did not use the zoom-out-zoom-in mechanism during inference, which relies on spatial prompts (Du et al., 2024). Furthermore, we did not use bounding box prompts during inference for a fairer comparison with the other text-based methods. We used a volume size of $64 \times 256 \times 256$, and fine-tuned the model for 50 epochs using a base learning rate of $1e^{-4}$. During inference, we directly resized each test volume to the model's input size, without using sliding window strategies.

MedSAM2 Text. To adapt MedSAM2 for text-prompted segmentation, we implemented a fine-tuning pipeline that leverages a pre-trained CT-CLIP text encoder (Hamamci et al., 2024) to transform each finding description into dense vector representations. These text embeddings are then processed through a trainable linear projection module that maps them into the same embedding space as MedSAM2's prompt decoder inputs. For the training procedure, we randomly sampled 32 consecutive frames from each CT scan in our dataset. We trained all parts of the model except the text encoder for 50 epochs. During inference, we process the CT scans slice-by-slice in their original size.

Table 4: Benchmarking various medical image segmentation models evaluated on ReXGroundingCT. The Dice score and hit rate (HIT) are reported. Findings are grouped into non-focal (22%) and focal opacities (78%). Fine-tuned models were trained on our ReXGroundingCT dataset. Overall, performance remains low across all models, reflecting the difficulty of segmenting complex and diverse radiological findings. Models shown in gray represent out-of-the-box performance that is not fine-tuned on ReXGroundingCT. $HIT_{5\%}$ represents the proportion of findings that have a Dice score above 5%. Confidence interval values are calculated by bootstrapping 1,000 times.

Method	**Non-focal (22%)**	**Focal (78%)**	**Average**	**$HIT_{5\%}$**	**$HIT_{10\%}$**
Bounding-Box					
MedSAM2	0.239 [0.201–0.280]	0.261 [0.237–0.284]	0.257 [0.237–0.278]	0.733 [0.694–0.770]	0.635 [0.591–0.678]
SAM2	0.230 [0.188–0.273]	0.218 [0.198–0.238]	0.220 [0.203–0.238]	0.699 [0.656–0.737]	0.582 [0.540–0.625]
SegVol	0.123 [0.096–0.154]	0.093 [0.078–0.109]	0.100 [0.086–0.112]	0.409 [0.369–0.450]	0.314 [0.277–0.356]
Text-only					
BiomedParse	0.040 [0.021–0.062]	0.005 [0.004–0.006]	0.013 [0.009–0.019]	0.033 [0.020–0.051]	0.020 [0.008–0.033]
BiomedParse	0.037 [0.028–0.049]	0.065 [0.054–0.076]	0.059 [0.050–0.067]	0.299 [0.261–0.336]	0.147 [0.118–0.179]
SegVol	0.009 [0.004–0.015]	0.001 [0.001–0.001]	0.003 [0.001–0.004]	0.014 [0.006–0.026]	0.012 [0.004–0.022]
SegVol	0.089 [0.065–0.118]	0.058 [0.045–0.072]	0.065 [0.054–0.076]	0.240 [0.204–0.279]	0.198 [0.165–0.232]
SAT	0.000 [0.000–0.000]	0.000 [0.000–0.000]	0.000 [0.000–0.000]	0.000 [0.000–0.000]	0.000 [0.000–0.000]
SAT	0.084 [0.063–0.107]	0.173 [0.153–0.194]	0.153 [0.136–0.170]	0.527 [0.483–0.570]	0.430 [0.385–0.474]
MedSAM2 Text	0.042 [0.030–0.057]	0.064 [0.053–0.077]	0.059 [0.050–0.069]	0.299 [0.261–0.340]	0.187 [0.155–0.222]

5. Results

In this section, we first benchmark state-of-the-art (SoTA) segmentation models, including SAT, SegVol, MedSAM2, and BiomedParse by fine-tuning them on ReXGroundingCT, as well as box-prompted models, which we use to gauge the difficulty of the task (Section 5.1). Then, in Section 5.2, we analyze performance differences among these models across different abnormality categories and sizes. Lastly, Section 5.3 shows examples of cases in our dataset along with qualitative results of the different models.

5.1. Benchmarking SoTA Medical Segmentation Models on ReXGroundingCT

Table 4 provides a comparison of various state-of-the-art medical image segmentation models evaluated on a diverse set of medical findings, categorized into non-focal (22%) and focal opacities (78%). For the box-prompted methods, the bounding box was generated using the boundaries of the ground truth mask. Models in gray are not fine-tuned on ReXGroundingCT. We used the Dice score and hit rate (HIT) as our evaluation metrics, where $HIT_{5\%}$ represents the proportion of findings that have a Dice score above 5%.

Among text-only methods, SAT achieves superior performance. 2D methods such as BiomedParse and MedSAM2 Text perform worse on this task relative to their 3D counterparts when looking at the average Dice score. Although MedSAM2 incorporates a memory module designed to leverage 3D context during segmentation, it only attends to preceding slices, limiting its effectiveness for this task. Segmentation of nodules, in particular, and other findings require bidirectional slice information, which is more effectively handled by true 3D models such as SAT and SegVol.

Additionally, results show a general trend where models perform better on focal opacities than on diffuse findings, with performance on non-focal findings sometimes being up to twice as low. This highlights the challenge of segmenting diffuse and irregular findings. Notably, the overall performance across all methods, including methods that utilize ground truth to generate bounding box prompts, remains unsatisfactory. This further confirms the complexity of accurately segmenting diverse clinical findings, highlighting a significant gap in current model capabilities.

5.2. Performance Comparison by Finding Category and Size

Figure 2 presents the Dice of different methods across abnormalities. We observe that 3D models (SAT and SegVol) significantly outperform their 2D counterparts in almost all tasks. Furthermore, we observe that the overall performance on pleural effusion is higher than other subcategories, which is expected given that they are common findings, and both BiomedParse and SAT were trained to segment effusions (Zhao et al., 2023, 2025). Additionally, SAT significantly outperforms the other methods on nodule and micro-nodule segmentation, which is one of the categories the model was trained on (Zhao et al., 2023). This highlights that one of the reasons why current models fail at segmenting other abnormalities such as emphysema, bronchial wall thickening, and bronchiectasis could be due to the lack of available data for these conditions.

We also analyze the performance of models with respect to the size of the finding. Finding size is defined as the ratio between the number of ground truth voxels and the total number of voxels in the corresponding scan after isotropic resampling. Sizes are categorized into three groups: small occupying the lowest 50% of the distribution, medium covering the 50% to 75% range, and large classified as those larger than 75% of all findings. We observe in Figure 3 that all models consistently perform better on large findings compared to smaller ones. Additionally, we observe that, while SegVol significantly outperforms SAT on larger finding sizes, SAT outperforms SegVol on small findings, which we suspect is mainly due to the downsampling of volumes used when training SegVol (Du et al., 2024). This is further supported by the previous Figure 2, where SegVol underperforms compared to 2D models on both nodule and micro-nodule segmentation. Figure S1 shows the ground-truth pixel size distribution across focal and non-focal findings.

5.3. Case Analysis

Figure 4 presents qualitative examples from ReXGroundingCT alongside predictions from different SoTA models after fine-tuning. In the first row, both SAT and SegVol successfully localize the scarring in the inferior lingula of the left lung. In contrast, MedSAM2 misidentifies this finding, confusing it with a similar instance of scarring in the posterobasal region, which was also part of the prompt for this CT scan.

In the second row, SAT was able to localize the interseptal thickness in the lower lobes of the lung, but incorrectly assigns the other prompted finding, interseptal thickness in the subpleural areas, to the same region. This highlights the need for models that can better understand and contextualize anatomical regions in 3D space.

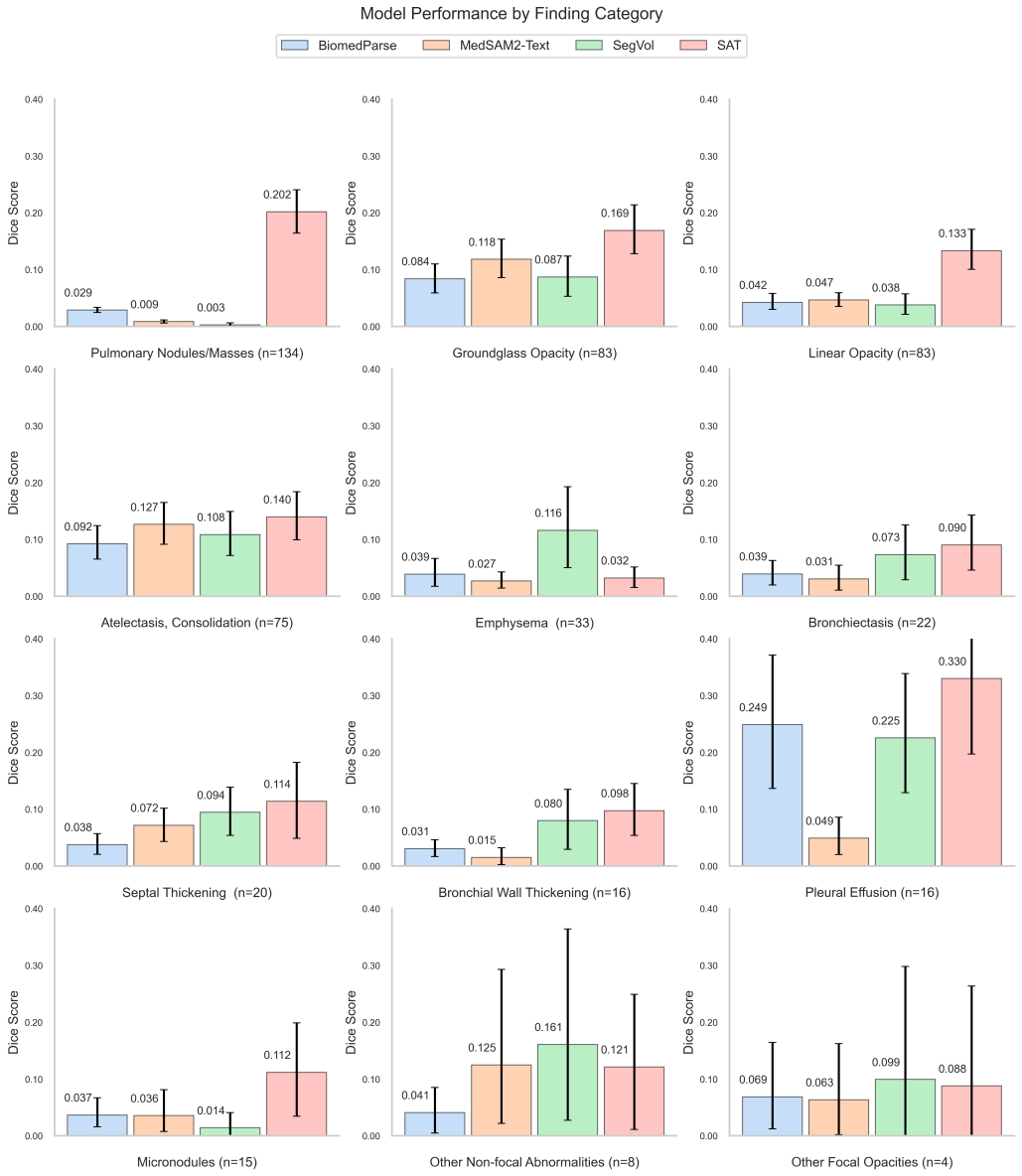


Figure 2: Performance of segmentation models across different abnormality categories. 3D models such as SAT and SegVol consistently outperform 2D models (BiomedParse, MedSAM2). For SAT, the relatively lower scores on emphysema, bronchial wall thickening, and bronchiectasis compared to more common abnormalities like pleural effusion and nodules highlight the need for more annotated data in underrepresented findings. 95% confidence intervals are shown, which are calculated by bootstrapping the scores 1,000 times.

In the third row, we present a more challenging case involving four distinct findings in close proximity. All models fail on this example, incorrectly predicting the region of the consolidated atelectasis and the ground-glass opacities.

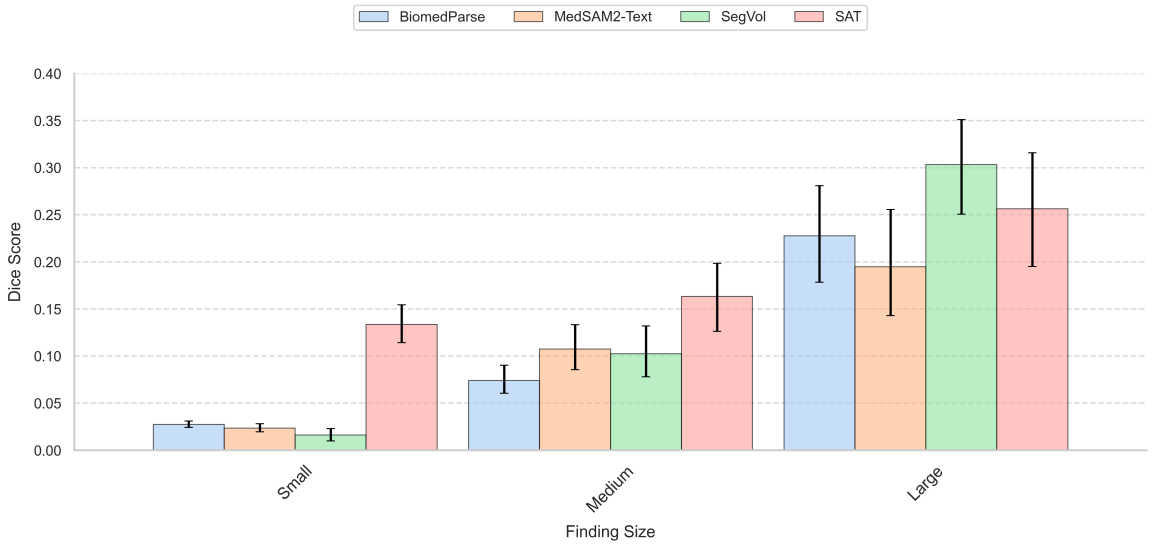


Figure 3: Segmentation performance by finding size. Findings were grouped into three size categories based on the distribution of ground truth voxel counts (normalized by volume size after isotropic resampling): small (bottom 50%), medium (50th to 75th percentile), and large (top 25%). All models perform better on large findings. While SegVol excels on large findings, SAT shows higher accuracy on small findings—likely due to differences in training resolution and volume downsampling strategies. 95% confidence intervals are shown, which are calculated by bootstrapping the scores 1,000 times.

6. Discussion

Our comprehensive evaluation of state-of-the-art segmentation models on ReXGroundingCT reveals significant challenges in accurately segmenting diverse findings from chest CT scans using natural language descriptions. Despite recent advances in medical image segmentation, these models struggle to map the complex, descriptive language of radiology reports into precise spatial localizations, even when fine-tuned on ReXGroundingCT, suggesting the need for novel architectural and methodological improvements.

6.1. Technical Implications

The substantial gap between current model capabilities and clinical requirements for this task highlights several technical limitations. First, existing models have primarily been developed to segment well-defined anatomical structures or categorical pathologies rather than the variable, sometimes ambiguous findings described in clinical reports (Zhao et al., 2023; Ma et al., 2025; Zhao et al., 2025). This limitation is especially evident in the poor performance across non-focal abnormalities such as emphysema, bronchiectasis, and septal thickening, which lack clear boundaries and often present with diffuse, irregular patterns.

Our results also demonstrate that 3D context is crucial for accurate medical image segmentation in volumetric data. 3D models like SAT and SegVol consistently outperform the 2D model (BiomedParse) across most finding categories. While models like MedSAM2 utilize a memory bank to attend to the segmentation masks of preceding slices, segmenting

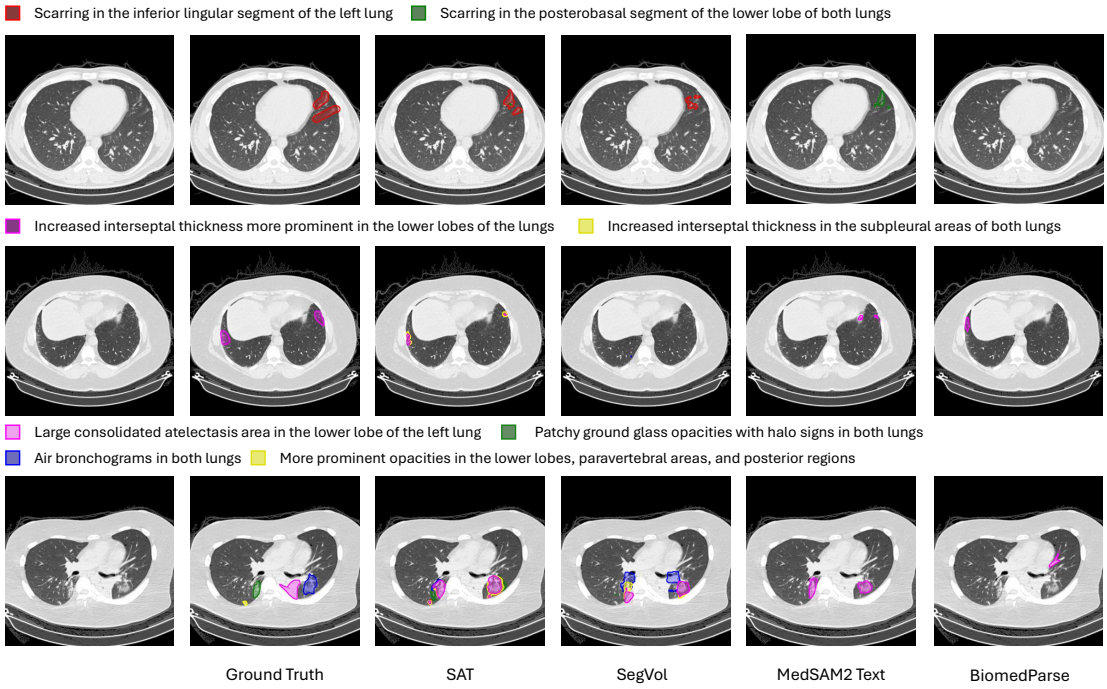


Figure 4: Visualization of segmentation examples from ReXGroundingCT, along with model predictions. Above each example is the text prompt given to all models. Each row shows a distinct CT case along with its prompted findings and model predictions from SAT, SegVol, MedSAM2 Text, and BiomedParse.

findings and anatomical structures sometimes requires bidirectional information, which is more effectively captures by 3D models.

The performance disparities between finding categories offer important insights for future model development. Models generally performed better on focal opacities (particularly nodules and pleural effusions) than on diffuse abnormalities, suggesting that current architectures are better suited to segment discrete, well-defined objects. Additionally, all models showed improved performance on larger lesions compared to smaller ones, with SAT outperforming SegVol on small lesions while SegVol excelled with larger findings. This pattern likely reflects differences in training approaches, as SegVol's downsampling of volumes during training may limit its ability to capture small abnormalities.

6.2. Clinical Implications

The medical imaging workflow encompasses three critical stages: acquisition, interpretation, and knowledge dissemination. The final stage, communicating findings from radiologists to referring clinicians, patients, and trainees, faces substantial barriers due to the text-based nature of radiological reporting.

In clinical settings, referring physicians often struggle to precisely localize findings within complex imaging studies, leading to diagnostic inefficiencies and increased consultation demands. Radiologists must dedicate valuable time to manually annotating images or providing supplementary guidance, diverting expertise from other critical tasks amid growing workloads and resource constraints.

For patients, who often receive both the imaging studies and the accompanying reports without direct clinical interpretation, the disconnect between textual descriptions and their spatial counterparts can be particularly challenging. This lack of visual context hinders their ability to understand the nature and location of their conditions, which may, in turn, impact their engagement, trust, and adherence to treatment plans.

In educational contexts, medical trainees developing diagnostic proficiency face difficulties bridging theoretical knowledge with practical visual recognition skills, particularly when interpreting complex findings in three-dimensional modalities like CT.

These communication challenges across inter-clinician, clinician-patient, and educator-trainee interactions highlight the need for technology that can automatically localize and segment regions of interest based on textual descriptions, which would transform healthcare delivery and education by providing visual reinforcement of textual findings.

6.3. Limitations

Our study has several limitations. First, while our data set includes a wide range of findings, it focuses only on lung and pleural findings, so it may not fully represent the variability in clinical practice. Second, our evaluation focused primarily on the Dice score, which may not capture all aspects of clinical utility. Third, our sample size of different finding categories is too small (n=15 for micronodules, n=16 for pleural effusion, n=16 for bronchial wall thickening, n=20 for septal thickening) to draw strong conclusions. Finally, the dataset itself is considered relatively small compared to other segmentation datasets, which is due to the significant time and resources required to accurately localize and segment findings in 3D chest CT scans.

7. Conclusion

Report grounding on 3D medical volumes remains an open challenge. While current models show promising capabilities for certain findings, particularly focal abnormalities with clear boundaries, they struggle with the various, sometimes subtle patterns described in clinical reports. Addressing these limitations will be essential for developing truly useful AI systems capable of bridging the gap between textual reports and volumetric medical data. By benchmarking on the ReXGroundingCT dataset, we hope to facilitate further research in this important area, ultimately leading to systems that can better serve the needs of clinicians, patients, and medical education.

Acknowledgments

This research was supported by a grant of the Boston Korea Innovative Research Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant number: RS-2024-00403047).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Mohammad Alarifi, Timothy Patrick, Abdulrahman Jabour, Min Wu, and Jake Luo. Understanding patient needs and gaps in radiology reports through online discussion forum analysis. *Insights into imaging*, 12:1–9, 2021.

Mohammad Alarifi, M Courtney Hughes, Abdulrahman M Jabour, Yazeed Alashban, and Erin Vernon. Patient, referring physician, and radiologist opinions over time on providing patients access to radiology reports: A systematic review. *Journal of the American College of Radiology*, 2024.

Mohammed Baharoon, Luyang Luo, Michael Moritz, Abhinav Kumar, Sung Eun Kim, Xiaoman Zhang, Miao Zhu, Mahmoud Hussain Alabbad, Maha Sbayel Alhazmi, Neel P Mistry, et al. Rexgroundingct: A 3d chest ct dataset for segmentation of findings from free-text reports. *arXiv preprint arXiv:2507.22030*, 2025.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Pedro RAS Bassi, Mehmet Can Yavuz, Kang Wang, Xiaoxi Chen, Wenxuan Li, Sergio Decherchi, Andrea Cavalli, Yang Yang, Alan Yuille, and Zongwei Zhou. Radgpt: Constructing 3d image-text tumor datasets. *arXiv preprint arXiv:2501.04678*, 2025.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

Adrian P Brady. Radiology reporting—from hemingway to hal? *Insights into imaging*, 9: 237–246, 2018.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

Daniel C Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *arXiv preprint arXiv:2411.05085*, 2024.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

Zhihao Chen, Yang Zhou, Anh Tran, Junting Zhao, Liang Wan, Gideon Su Kai Ooi, Lionel Tim-Ee Cheng, Choon Hua Thng, Xinxing Xu, Yong Liu, et al. Medical phrase grounding with region-phrase context contrastive alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 371–381. Springer, 2023.

Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. *Advances in Neural Information Processing Systems*, 37:110746–110783, 2024.

Nicolás Gaggion, Candelaria Mosquera, Lucas Mansilla, Julia Mariel Saidman, Martina Aineseder, Diego H Milone, and Enzo Ferrante. Chexmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. *Scientific Data*, 11 (1):511, 2024.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Omer Faruk Durugol, Bastian Wittmann, Tamaz Amiranashvili, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024.

Zhongzhen Huang, Yankai Jiang, Rongzhao Zhang, Shaoting Zhang, and Xiaofan Zhang. Cat: Coordinating anatomical-textual prompts for multi-organ and tumor segmentation. *arXiv preprint arXiv:2406.07085*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Akimichi Ichinose, Taro Hatsutani, Keigo Nakamura, Yoshiro Kitamura, Satoshi Iizuka, Edgar Simo-Serra, Shoji Kido, and Noriyuki Tomiyama. Visual grounding of whole radiology reports for 3d ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–621. Springer, 2023.

Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, et al. nninteractive: Redefining 3d promptable segmentation. *arXiv preprint arXiv:2503.08373*, 2025.

Veena R Iyer, Peter F Hahn, Lawrence S Blaszkowsky, Sarah P Thayer, Elkan F Halpern, and Mukesh G Harisinghani. Added value of selected images embedded into radiology reports to referring clinicians. *Journal of the American College of Radiology*, 7(3):205–210, 2010.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

Jiayu Lei, Xiaoman Zhang, Chaoyi Wu, Lisong Dai, Ya Zhang, Yanyong Zhang, Yanfeng Wang, Weidi Xie, and Yuehua Li. Autorg-brain: Grounded report generation for brain mri. *arXiv preprint arXiv:2407.16684*, 2024.

Wenhui Lei, Hanyu Chen, Zitian Zhang, Luyang Luo, Qiong Xiao, Yannian Gu, Peng Gao, Yankai Jiang, Ci Wang, Guangtao Wu, et al. A data-efficient pan-tumor foundation model for oncology ct interpretation. *arXiv preprint arXiv:2502.06171*, 2025.

Luyang Luo, Jenanan Vairavamurthy, Xiaoman Zhang, Abhinav Kumar, Ramon R Ter-Oganesyan, Stuart T Schroff, Dan Shilo, Rydhwana Hossain, Mike Moritz, and Pranav Rajpurkar. Rexplain: Translating radiology into patient-friendly video reports. *arXiv preprint arXiv:2410.00441*, 2024.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Md Eshmam Rayed, SM Sajibul Islam, Sadia Islam Niha, Jamin Rahman Jim, Md Mohsin Kabir, and MF Mridha. Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Informatics in Medicine Unlocked*, page 101504, 2024.

François Remy, Kris Demuynck, and Thomas Demeester. Biolord-2023: Semantic textual representations fusing llm and clinical knowledge graph insights. *arXiv preprint arXiv:2311.16075*, 2023.

Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. In *European Conference on Computer Vision*, pages 207–229. Springer, 2024.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*, 2024.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, 22(1):166–176, 2025.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. One model to rule them all: Towards universal segmentation for medical images with text prompts. *arXiv preprint arXiv:2312.17183*, 2023.

Ke Zou, Yang Bai, Zhihao Chen, Yang Zhou, Yidi Chen, Kai Ren, Meng Wang, Xuedong Yuan, Xiaojing Shen, and Huazhu Fu. Medrg: Medical report grounding with multi-modal large language model. *arXiv preprint arXiv:2404.06798*, 2024.

Appendix

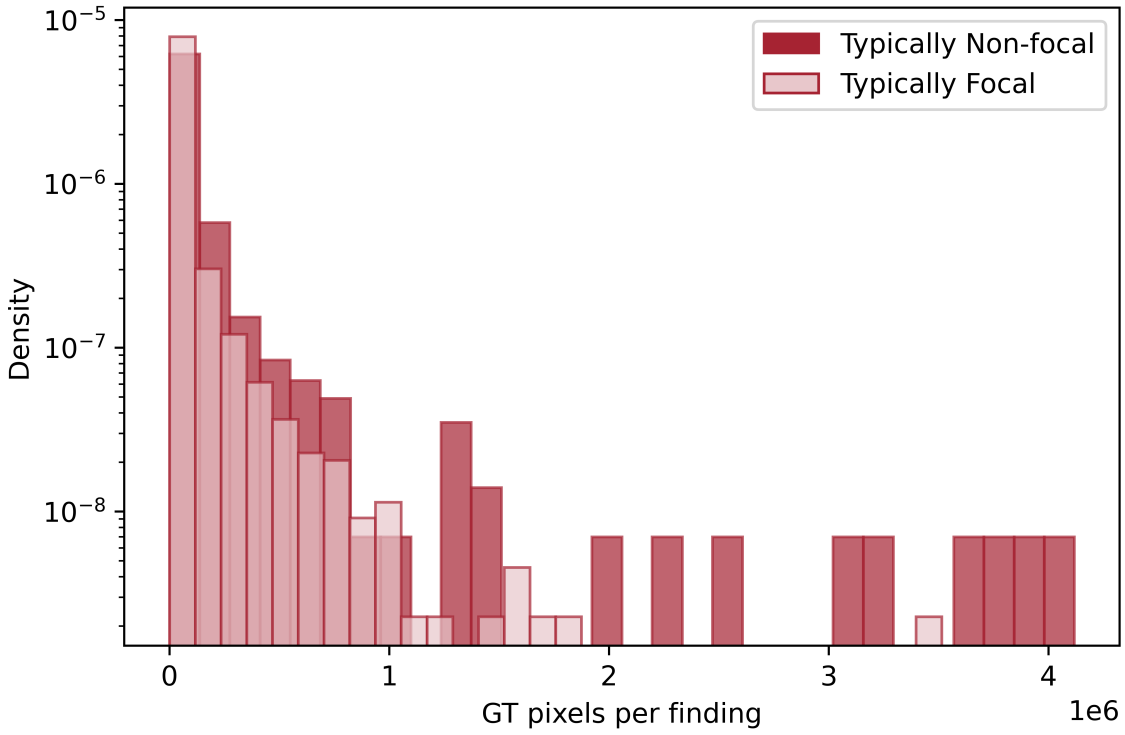


Figure S1: Distribution of annotated pixel counts per finding, stratified by whether the finding is typically focal or non-focal. Findings labeled as typically non-focal (e.g., ground-glass opacities, consolidation) tend to span a broader range of sizes, including very large segmentations, while typically focal findings (e.g., nodules, masses) are generally smaller and more concentrated near the lower end of the size spectrum. Pixel counts are calculated after isotropic resampling of all CT scans to 1mm^3 resolution.

Prompt for Data Curation Process

Radiology Report Rewriting Prompt - Part 1/2

You are an expert radiologist with extensive experience in reading and writing radiology reports in American English.
OBJECTIVE:
1. Rewrite the provided **'Findings'** and **'Impressions'** sections of radiology reports to ensure they use the appropriate medical terminology, phrasing, and grammar that American radiologists would use.
2. Categorize each rewritten sentence in the **'Findings'** section into one of 7 regions to structure the **'Findings'** section of the report.
RULES:
1. Correct any unusual phrases, awkward grammar, or non-standard terminology. For example, English-speaking American radiologists would replace these phrases:
- Instead of `"millimetric"`, say `"subcentimeter"`
- Instead of `"sequelae"` or `"pleuroparenchymal sequelae,"` say `"scarring"` or `"fibrosis"`
- Instead of `"esophagus calibration was normal,"` say `"esophagus is normal in caliber"`
- Instead of `"no space-occupying lesion,"` say `"no lesion"`
- Instead of `"within the cross-sectional area,"` `"within the examined area,"` or `"sections visualized,"` say `"within the field of view"`

2. Ensure each rewritten sentence accurately conveys the same medical information and details as the corresponding sentence in the original report. **DO NOT LOSE OR CHANGE ANY NUMERICAL MEASUREMENT DETAILS OF ANY FINDINGS. YOU WILL BE PENALIZED IF YOU MISS OR CHANGE THE MEANING OF INFORMATION IN THE ORIGINAL REPORT.**

3. Make sure each sentence in the original report is rewritten and categorized appropriately. **YOU WILL BE PENALIZED IF YOU MISS ANY SENTENCES FROM OR ADD ANY NEW SENTENCES TO THE ORIGINAL REPORT.**

4. Remove any irrelevant comments that should not be in a radiology report. For example, in the following sentence:
`"Impressions: Pleural effusion and concomitant compression atelectasis in both lungs. Nonspecific nodules in both lungs. Cardiomegaly and minimal pericardial effusion. Patient 14.10."`
The sentence `"Patient 14.10."` is irrelevant and should not be included in the rewritten report.
You will be given the **Findings** and **Impressions** sections of a radiology report, separated by sentence. Each sentence is labeled with a sentence code, containing a letter and a number. The code letter will be either `F`, for Findings, or `I`, for Impressions depending on which section the sentence comes from. The code number refers to the sentence's number in its respective section. Thus, the first sentence in Findings has code `F0`, the second sentence in Findings has code `F1`, and so on, and the first sentence in Impressions has code `I0`, the second sentence in Impressions has code `I1`, and so on.
The output should be in JSON format, structured as follows:
- `"Findings"` should map to a dictionary with the 7 categories. Include all sentences, including all details and measurements, from the 'Findings' section and use any information from the 'Impressions' section to correctly supplement the details (anatomical region, measurements, etc.) of these 'Findings' sentences.
The 7 category keys should be:
`"Lungs/Airways/Pleura"`, `"Heart/Vessels"`, `"Mediastinum/Hila"`, `"Chest wall/Axilla"`, `"Lower neck"`, `"Bones"`, `"Upper abdomen"`
The values should be lists that contain the rewritten sentences in the report that fall into the anatomical region specified by the key. **DO NOT CREATE ANY NEW CATEGORIES. PLACE EACH FINDING SENTENCE INTO THE CORRECT CATEGORY.**
If the **Findings** section is `"Not Given"`, all of the lists in each of the 7 categories should be empty.

- `"Impressions"` should map to a string with the rewritten sentences in the 'Impressions' section of the report. If the **Impressions** section is `"Not Given"`, keep `"Not Given"` as the value in this section.
BE SURE TO FOLLOW ALL RULES METICULOUSLY. Think through all rules and steps before answering.

Figure S2: Prompt used to guide GPT-4 for rewriting and categorizing radiology reports.

Radiology Report Rewriting Prompt - Part 2/2

Follow this example:

Input: {'F0': 'Trachea and both main bronchi were in the midline and no obstructive pathology was observed in the lumen.',
'F1': 'The mediastinum could not be evaluated optimally in the non-contrast examination.',
'F2': 'As far as can be seen; mediastinal main vascular structures, heart contour, size are normal.',
'F3': 'Pericardial effusion-thickening was not observed.',
'F4': 'Thoracic esophagus calibration was normal and no significant tumoral wall thickening was detected.',
'F5': 'No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected.',
'F6': 'When examined in the lung parenchyma window; A peripheral subcapsular crazy paving pattern was formed in the posterobasal segment of the right lung, nodular ground glass opacity was observed, and the appearance is highly suspicious for ultra-early Covid-19 pneumonia.',
'F7': 'It is recommended to be evaluated together with clinical and laboratory.',
'F8': 'A millimetric nonspecific parenchymal nodule was observed in the lateralabasal segment of the lower lobe of the left lung.',
'F9': 'No mass lesion with distinguishable borders was detected in both lungs.',
'F10': 'As far as can be seen in the sections, millimetric calculus was observed in the middle part of the right kidney.',
'F11': 'Accessory spleen with a diameter of 13 mm was observed adjacent to the lower pole of the spleen.',
'F12': 'Other upper abdominal organs are normal.',
'F13': 'No space-occupying lesion was detected in the liver that entered the cross-sectional area.',
'F14': 'Bilateral adrenal glands were normal and no space-occupying lesion was detected.',
'F15': 'Bone structures in the study area are natural.',
'F16': 'Vertebral corpus heights are preserved.',
'I0': 'High suspicious findings for ultra-early period Covid-19 pneumonia in the right lung posterobasal segment; it is recommended to be evaluated together with clinical and laboratory.',
'I1': 'Millimetric nonspecific parenchymal nodule in the lateralabasal segment of the left lung lower lobe .',
'I2': 'Right nephrolithiasis' }

Output: {"Findings": {
"Lungs/Airways/Pleura": [
"The trachea and both main bronchi are centrally positioned and patent.",
"In the lung parenchyma, there is a peripheral subcapsular crazy paving pattern in the posterobasal segment of the right lung along with nodular ground glass opacity, suggestive of potential early-stage COVID-19 pneumonia.",
"No masses are identified in either lung.",
"An indeterminate subcentimeter parenchymal nodule is noted in the lateral basal segment of the left lower lobe."
],
"Heart/Vessels": [
"Mediastinal large vessels and heart contour and size appear normal.",
"No pericardial effusion or thickening is evident."
],
"Mediastinum/Hila": [
"Mediastinal evaluation is limited on this non-contrast exam.",
"Thoracic esophagus is normal in caliber without significant wall thickening.",
"No enlarged lymph nodes are detected in prevascular, pre-paratracheal, subcarinal, or bilateral hilar-axillary regions."
],
"Chest wall/Axilla": [],
"Lower neck": [],
"Bones": [
"Bone structures within the field of view are normal.",
"Vertebral body heights are preserved."
],
"Upper abdomen": [
"A subcentimeter calculus is seen in the mid portion of the right kidney.",
"An accessory spleen measuring 13 mm in diameter is located adjacent to the lower pole of the spleen.",
"Other upper abdominal organs appear normal.",
"There are no liver lesions within the field of view.",
"The bilateral adrenal glands appear normal with no lesions identified."
]
},
"Impressions": "Findings are highly suspicious for early-stage COVID-19 pneumonia in the right lung posterobasal segment; further evaluation with clinical and laboratory correlation is recommended. An indeterminate subcentimeter parenchymal nodule is present in the lateral basal segment of the lower lobe of the left lung. Evidence of right nephrolithiasis is noted." }

Figure S3: Prompt used to guide GPT-4 for rewriting and categorizing radiology reports.

Extract Findings and Anatomical Phrases Prompt - Part 1/2

You are an expert radiologist. You are helping process reports from chest CT scans.
You will be given the Findings and Impressions sections of a radiology report. The input report structure is a dictionary with two keys: "Findings" and "Impressions."

Findings:
- Contains nested keys that represent sentence codes for sentences in the "Findings" section of the report. The letters represent which region the finding is from (i.e. 'L' corresponds to Lung). The numbers following each letter correspond to the numbered sentences for that region.

Impressions:
- Contains nested keys that represent sentence codes for sentences in the "Impressions" section of the report. The letter 'I' represents the Impressions section. The numbers correspond to the numbered sentences in the 'Impressions' section, which is not broken down by region.

OBJECTIVE:
Please extract distinct phrases from the radiology report which refer to objects, any findings, or anatomies that are visible in a CT scan, or the absence of such.
For each phrase you extract, please also specify the sentence code(s) that the phrase comes from. Also specify if the phrase represents an abnormal finding or not and if the phrase contains terminology that refers to a prior report.
The main objective is to extract phrases which refer to things which can be located on a chest CT scan, or confirmed not to be present.

RULES:
1. If a sentence describes multiple distinct findings (whether abnormal or normal), split them up into separate sentences so each sentence represents a distinct finding.
Example: `"Widespread peribronchial thickening and tree-in-bud opacities in the upper lobe apicoposterior, lingular segment, and lower lobes of the left lung"`
→ `"Widespread peribronchial thickening in the left lung"` AND `"Tree-in-bud opacities in the upper lobe apicoposterior segment, lingular segment, and lower lobes of the left lung"`

2. If multiple sentences specify the same distinct finding, extract them together as one sentence and combine all details specified (including anatomical location, measurements, stable or not, etc.) across all sentences for the finding.

3. Exclude clinical speculation, interpretation, and suspicion. For example, remove phrases like:
- "highly suggestive of pneumonia"
- "likely benign"
- "suspicious for metastasis"
- "concerning for metastasis"

4. If a phrase in a sentence ends in a question mark, do not consider it as a finding.

5. Exclude recommendations (e.g. "Recommend a CT").

6. For all centimeter measurements, reduce the number to one decimal point. For all millimeter measurements, round the number to the nearest whole number.

7. If you miss a finding that is visible in the CT, you will be penalized! You MUST INCLUDE ANY FINDING THAT IS VISIBLE.
Example: The phrase `"Small lymph nodes are present in both axillary regions with a short axis measuring up to 8 mm"` must be included in the output!

8. The output should be in JSON format, structured as follows:
- The keys should be `'SX'`, where X is the phrases enumerated.
- The values should be Lists where: - index 0 is the extracted phrase,
- index 1 is the sentence code(s), delimited by commas (e.g. `"H0,H1,I0"`),
- index 2 is `"Y"` or `"N"` (if abnormal),
- index 3 is `"Y"` or `"N"` (if prior-report terminology is used).

BE SURE TO FOLLOW ALL RULES METICULOUSLY. Think through all rules and steps before answering.

Figure S4: Prompt for GPT-4 to extract findings and anatomical phrases from CT chest reports.

Extract Findings and Anatomical Phrases Prompt - Part 2/2

Follow this example:
Input:

```
{
"Findings":  {
"L0":  "The trachea and both main bronchi are patent.",
"L1":  "Bilateral lower lobe pleuroparenchymal opacities extend from the central regions to the pleura, with significant
bronchial wall thickening and a mass-like appearance of the left lower lobe bronchus, all of which are stable.",
"L2":  "A stable pleural effusion is noted in the right hemithorax.",
"H0":  "The mediastinal main vascular structures, heart contour, and size are within normal limits.",
"H1":  "The thoracic aorta exhibits a normal diameter.",
"H2":  "No pericardial effusion or thickening is identified.",
"M0":  "The thoracic esophagus is normal in caliber without significant tumoral thickening.",
"M1":  "Lymph nodes in the mediastinum with a short axis up to 1 cm appear stable.",
"B0":  "Bone structures within the study area are unremarkable.",
"B1":  "Vertebral body heights are maintained.",
"A0":  "Stable hypodense hepatic lesions suspicious for metastases and hepatomegaly are noted in the liver that is included in
the field of view.",
"A1":  "A 28x17 mm lesion in the right adrenal gland, suspicious for metastasis, is stable.",
"A2":  "The left adrenal gland appears normal, with no lesions identified."
},
"Impressions":  {
"I0":  "Stable mass encasing the bronchi of the left lower lobe.",
"I1":  "Stable pleuroparenchymal opacities with bronchial and pleural extension in the bilateral lower lobes, bronchial wall
thickening, nonspecific ground glass densities, and right pleural effusion.",
"I2":  "Multiple hepatic mass lesions suspicious for metastases with associated hepatomegaly.",
"I3":  "Suspected right adrenal metastatic lesion.",
"I4":  "Stable mediastinal lymph nodes."
}
}
```

Output:

```
{
"S1":  ["Patent trachea and both main bronchi", "L0", "N", "N"],
"S2":  ["Bilateral lower lobe pleuroparenchymal opacities extending from central regions to pleura", "L1,I1", "Y", "Y"],
"S3":  ["Significant bronchial wall thickening", "L1,I1", "Y", "Y"],
"S4":  ["Mass-like appearance of left lower lobe bronchus", "L1,I0", "Y", "Y"],
"S5":  ["Stable pleural effusion in the right hemithorax", "L2,I1", "Y", "Y"],
"S6":  ["Normal heart contour and size", "H0", "N", "N"],
"S7":  ["Normal mediastinal main vascular structures", "H0", "N", "N"],
"S8":  ["Normal thoracic aorta diameter", "H1", "N", "N"],
"S9":  ["No pericardial effusion or thickening", "H2", "N", "N"],
"S10":  ["Normal thoracic esophagus caliber without tumoral thickening", "M0", "N", "N"],
"S11":  ["Stable mediastinal lymph nodes up to 1 cm", "M1,I4", "Y", "Y"],
"S12":  ["Unremarkable bone structures", "B0", "N", "N"],
"S13":  ["Maintained vertebral body heights", "B1", "N", "N"],
"S14":  ["Stable hypodense hepatic lesions", "A0,I2", "Y", "Y"],
"S15":  ["Hepatomegaly noted in the liver", "A0,I2", "Y", "Y"],
"S16":  ["Lesion in right adrenal gland, 28x17 mm", "A1,I3", "Y", "Y"],
"S17":  ["Normal left adrenal gland with no lesions", "A2", "N", "N"]
}
```

Figure S5: Prompt for GPT-4 to extract findings and anatomical phrases from CT chest reports.

Finding Categorization Prompt - Part 1/3

You are an expert radiologist specializing in chest CT scan interpretation. Your task is to categorize positive findings from radiology reports according to a specific schema. The schema consists of 12 parent categories (including "Other") and subcategories within each parent category (except for the "Other" parent category).

Schema:

1. Typically non-focal lung/airway/pleural abnormalities
    a. Bronchial wall thickening
    b. Bronchiectasis
    c. Emphysema (including Centrilobular, Paraseptal, Bullous)
    d. Septal thickening (including Interlobular, Reticulation)
    e. Micronodules (including, Centrilobular, Tree-in-bud, Perilymphatic)
    f. Other
2. Typically focal lung/airway/pleural opacities
    a. Linear (including subsegmental atelectasis, scarring, fibrosis)
    b. Atelectasis, consolidation
    c. Groundglass opacity
    d. Pulmonary nodules/masses
    e. Pleural effusion or thickening
    f. Honeycombing
    g. Pneumothorax
    h. Other
3. Non-pulmonary lesions
    a. Lymphadenopathy lesions
    b. Liver lesions
    c. Gallbladder lesions
    d. Renal/kidneys, collecting system, and ureters lesions
    e. Spleen lesions
    f. Adrenal lesions
    g. Pancreas lesions
    h. Thyroid lesions
    i. Skin/subcutaneous lesions
    j. Bone/Osseous structures lesions
    k. Other lesions
4. Bones (non-lesion)
    a. Fractures
    b. Degenerative joint disease, degenerative disc disease, arthritis
    c. Spinal curvature abnormalities: kyphosis, scoliosis
    d. Other
5. Stones/organ calcifications (non-lesion)
    a. Nephroliths, choleliths
    b. Granulomas
    c. Other
6. Hollow viscera abnormalities
    a. Hiatus hernia
    b. Wall thickening
    c. Dilated
    d. Diverticulum (including diverticulosis)
    e. Other
7. Skin/subcutaneous (non-lesion)
    a. Skin thickening
    b. Stranding
    c. Abdominal wall hernia: ventral, umbilical, inguinal
    d. Gynecomastia
    e. Other
8. Cardiovascular
    a. Atherosclerosis (including coronary, non-coronary)
    b. Vessel aneurysm, ectasia, enlargement
    c. Vessel occlusion or stenosis
    d. Cardiac chamber enlargement
    e. Valvular calcification
    f. Pericardial effusion
    g. Other
9. Body composition
    a. Visceral fat
    b. Superficial subcutaneous fat
    c. Skeletal muscle
    d. Osteoporosis
    e. Hepatic steatosis
    f. Other
10. Diffuse/whole organ
    a. Organomegaly (including splenomegaly, multinodular goiter, thyromegaly, lung hyperinflation)
    b. Atrophy
    c. Other
11. Device
    a. Elongated (including catheter, pacemaker/defibrillator, spinal stimulator)
    b. Surgical clips
    c. Other
12. Other

Figure S6: Prompt for categorizing positive CT findings according to a predefined schema.

Finding Categorization Prompt - Part 2/3

OBJECTIVE:
You will be given a list of positive findings from a CT Scan report. Your task is to assign each finding to the most appropriate category using the format `"Xa"`, where X is the parent category number (1–12) and 'a' is the subcategory letter. These parent categories and subcategories are mutually exclusive and exhaustive, so you must assign one of the categories in the schema to each given positive finding.
RULES:
1. You should first think about what parent category the finding belongs to and then assign the most appropriate subcategory within that parent category.
2. If a finding doesn't fit into any parent category, use `"12"` for the general "Other" category.
3. If a finding doesn't fit into any specific subcategory, use the "Other" subcategory of the most relevant parent category.
4. **YOU MUST PLACE EACH POSITIVE FINDING INTO ONE OF THESE CATEGORIES. YOU WILL BE PENALIZED IF YOU ASSIGN A FINDING TO A NON-EXISTENT CATEGORY OR IF YOU LEAVE A FINDING UNASSIGNED.**
5. **ONLY USE THE CATEGORIES PROVIDED IN THE SCHEMA. DO NOT CREATE NEW CATEGORIES.**
6. Note: Bone lesions should be categorized under `3j` (Non-pulmonary lesions – Bone/Osseous structures), not under category 4.

Input Format:
The input will be in JSON format, structured as follows: A Python dictionary where keys are `"P1"`, `"P2"`, etc., representing the ID of each positive finding, and the values are the text of the positive findings.
Output Format:
The output should be in JSON format, structured as follows: A Python dictionary with the same keys as the input, but the values should be the assigned categories (e.g., `"1a"`, `"2c"`, `"3f"`, `"12"`, etc.) for each positive finding.
BE SURE TO FOLLOW ALL RULES METICULOUSLY. Think through all rules and steps before answering. Follow these examples:
Input 1:

```
{
"P1":  "Scarring at apical level of right lung, in middle lobe, anterior-posterior segments of right upper lobe, and superior segment of lower lobe",
"P2":  "Scarring at level of minor fissure and in lingular segment",
"P3":  "Subpleural calcified nodules, 2-3 mm in diameter, along interlobular fissure in basal portion of left lower lobe",
"P4":  "Calcific atheromatous plaques in aortic arch, subclavian artery, and coronary arteries",
"P5":  "Mediastinal and bilateral hilar lymph nodes, largest on right measuring approximately 15x11 mm, some with partial calcification",
"P6":  "Mild hiatal hernia",
"P7":  "Degenerative changes in bone structure",
"P8":  "Diffuse idiopathic skeletal hyperostosis (DISH)",
"P9":  "Mild scoliosis with convexity to the left in thoracic spine",
"P10":  "Liver with decreased density consistent with steatosis",
"P11":  "Prominent dense formation within gallbladder consistent with cholelithiasis",
"P12":  "Nodular density adjacent to spleen compatible with accessory spleen"
}
```

Output 1:

```
{
"P1":  "2a",
"P2":  "2a",
"P3":  "2d",
"P4":  "8a",
"P5":  "3a",
"P6":  "6a",
"P7":  "4b",
"P8":  "4b",
"P9":  "4c",
"P10":  "9e",
"P11":  "5a",
"P12":  "3e"
}
```

Figure S7: Prompt for categorizing positive CT findings according to a predefined schema.

Finding Categorization Prompt - Part 3/3

Input 2:

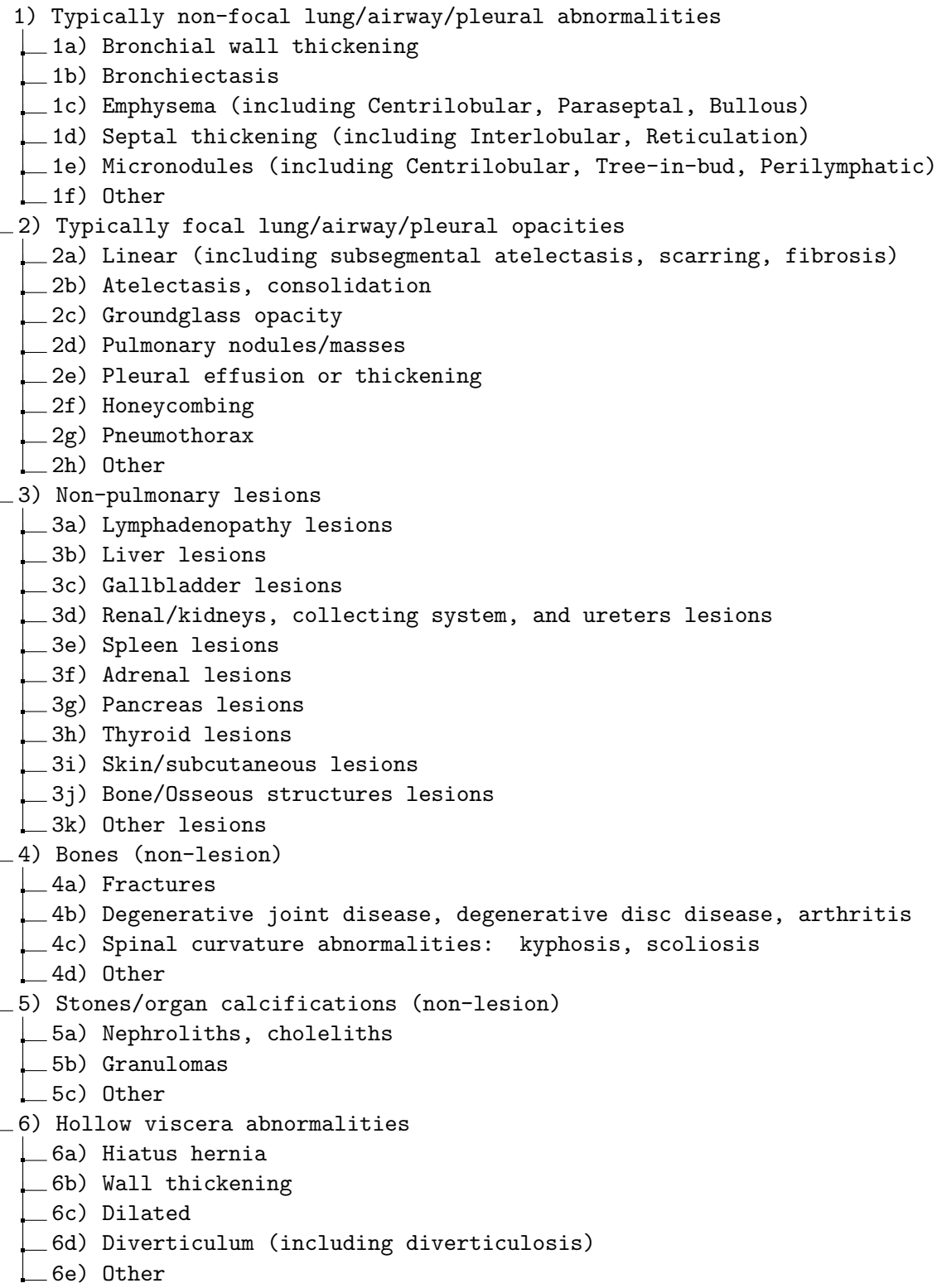
```
{
"P1":  "Subsegmental and band-like atelectasis in the medial segment of the right middle lobe, inferior lingular segment of
the left upper lobe, and basal segments of both lower lobes",
"P2":  "Emphysematous changes in the upper lobes of both lungs",
"P3":  "Tortuous and elongated thoracic aorta",
"P4":  "Increased pulmonary trunk diameter at 33 mm",
"P5":  "Small pericardial effusion",
"P6":  "Calcific atheroma plaques in supraaortic branches of thoracic aorta and coronary arteries",
"P7":  "Calcific plaques in abdominal aorta, its visceral branches, and proximal iliac arteries",
"P8":  "Small ventral hernia",
"P9":  "Displaced and impacted multipart fracture in right humeral head",
"P10":  "Thoracolumbar S-shaped scoliosis",
"P11":  "Osteophyte formations with bridging at right anterolateral vertebral corners",
"P12":  "Cholelithiasis with a 12 mm calculus within gallbladder lumen",
"P13":  "Two millimetric calculi in lower pole of left kidney",
"P14":  "Two cortical-parapelvic cysts in anterior midsection of right kidney, largest measuring 4.5 cm in diameter",
"P15":  "Nodular thickening in both adrenal glands",
"P16":  "Diffuse lytic bone lesions suggestive of multiple myeloma involvement throughout the visualized bones"
}
```

Output 2:

```
{
"P1":  "2b",
"P2":  "1c",
"P3":  "8b",
"P4":  "8b",
"P5":  "8f",
"P6":  "8a",
"P7":  "8a",
"P8":  "7c",
"P9":  "4a",
"P10":  "4c",
"P11":  "4b",
"P12":  "5a",
"P13":  "5a",
"P14":  "3d",
"P15":  "3f",
"P16":  "3j"
}
```

Figure S8: Prompt for categorizing positive CT findings according to a predefined schema.

Figure S9: Hierarchical Tree Finding Categories



- 7) Skin/subcutaneous (non-lesion)
  - 7a) Skin thickening
  - 7b) Stranding
  - 7c) Abdominal wall hernia: ventral, umbilical, inguinal
  - 7d) Gynecomastia
  - 7e) Other
- 8) Cardiovascular
  - 8a) Atherosclerosis (including coronary, non-coronary)
  - 8b) Vessel aneurysm, ectasia, enlargement
  - 8c) Vessel occlusion or stenosis
  - 8d) Cardiac chamber enlargement
  - 8e) Valvular calcification
  - 8f) Pericardial effusion
  - 8g) Other
- 9) Body composition
  - 9a) Visceral fat
  - 9b) Superficial subcutaneous fat
  - 9c) Skeletal muscle
  - 9d) Osteoporosis
  - 9e) Hepatic steatosis
  - 9f) Other
- 10) Diffuse/whole organ
  - 10a) Organomegaly (including splenomegaly, multinodular goiter, thyromegaly, lung hyperinflation)
  - 10b) Atrophy
  - 10c) Other
- 11) Device
  - 11a) Elongated (including catheter, pacemaker/defibrillator, spinal stimulator)
  - 11b) Surgical clips
  - 11c) Other
- 12) Other