

Supplementary Information

Chest2Vec: A multipurpose text encoder conditioned by instructions for chest radiograph and computed tomography reports

Hanbin Ko^{1,2}, Rong Yang³, Dongheon Lee^{4,5,*} & Chang Min Park^{4,6,*}

¹Interdisciplinary Program in Bioengineering, Seoul National University Graduate School, Seoul, Republic of Korea

²Integrated Major in Innovative Medical Science, Seoul National University Graduate School, Seoul, Republic of Korea

³Department of Radiology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

⁴Department of Radiology, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, Republic of Korea

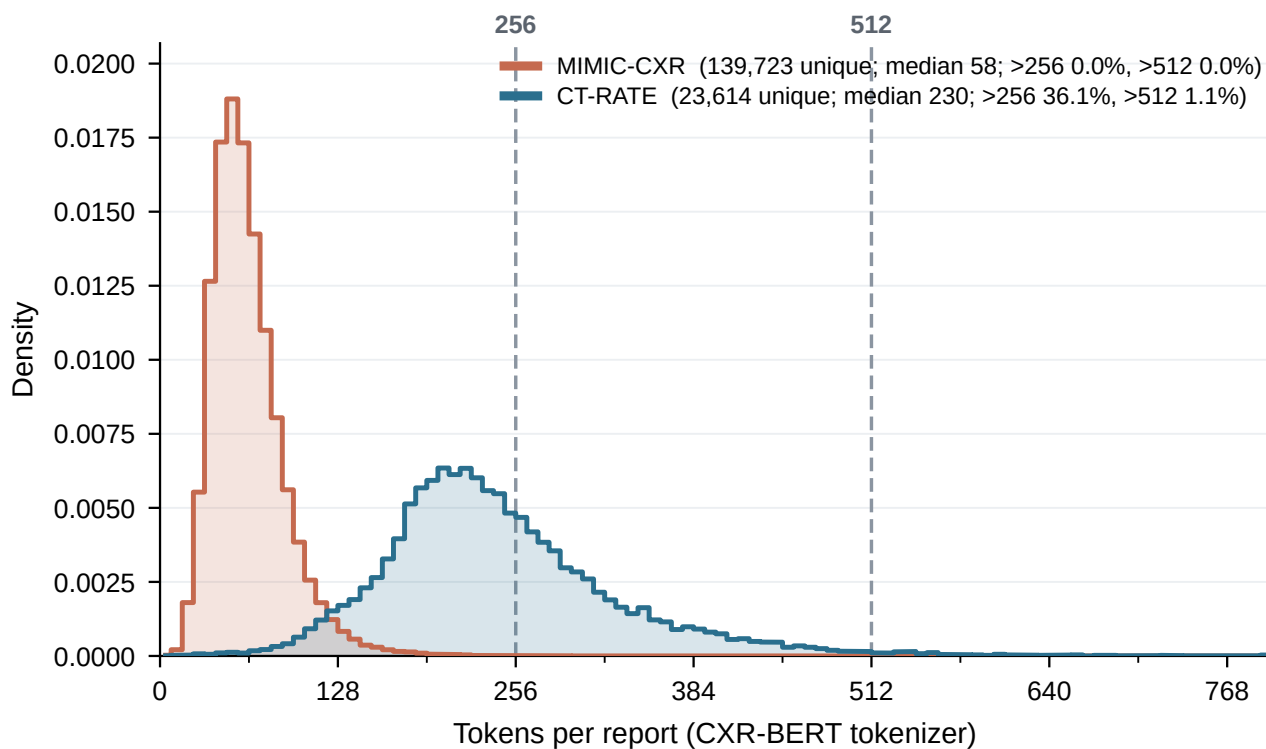
⁵Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea

⁶Institute of Medical and Biological Engineering, Seoul National University Medical Research Center, Seoul, Republic of Korea

*Correspondence: dhlee13@snu.ac.kr; morphius@snu.ac.kr

Contents. This document contains three supplementary figures and seven supplementary tables. All methodological descriptions remain in the main manuscript.

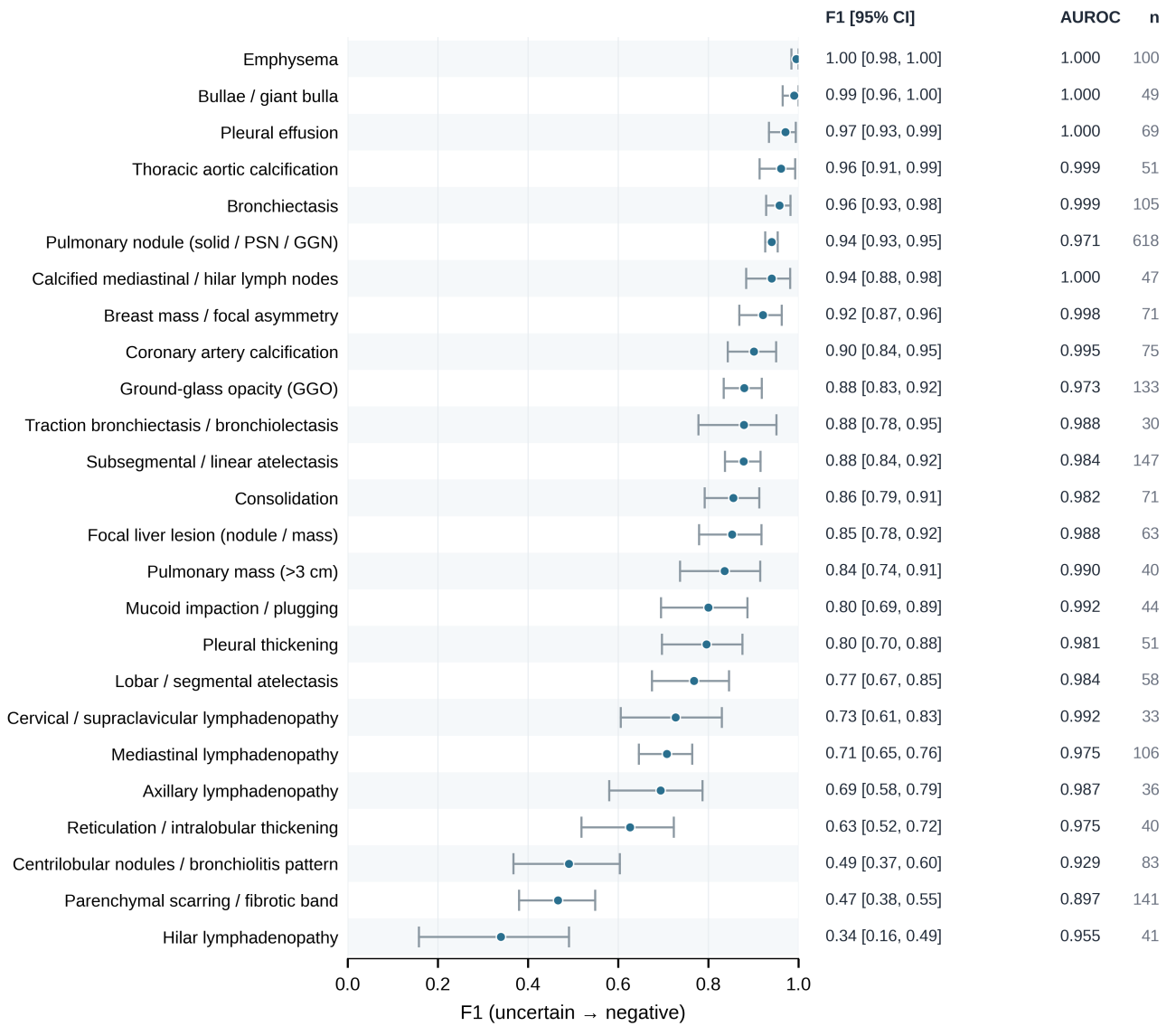
Supplementary Figures



Supplementary Figure 1: Token-length distributions of MIMIC-CXR and CT-RATE reports. Distribution of findings-section lengths under the CXR-BERT tokenizer, computed over unique findings sections without truncation and including special tokens. Summary statistics are given in Supplementary Table 1.



Supplementary Figure 2: Per-label chest CT report labeling on CT-RATE. F1 with 95% bootstrap percentile confidence intervals for the fine-tuned Chest2Vec-0.6B labeler, shown for each leaf finding with at least 30 positive test instances and ordered by F1. Per-label area under the receiver operating characteristic curve and the number of positive test instances are listed at right. Uncertain findings were mapped to negative, and per-label decision thresholds were selected on the CT-RATE tuning partition and held fixed at test.



Supplementary Figure 3: Per-label chest CT report labeling on SNUH. F1 with 95% bootstrap percentile confidence intervals for the fine-tuned Chest2Vec-0.6B labeler, shown for each leaf finding with at least 30 positive test instances and ordered by F1. Per-label area under the receiver operating characteristic curve and the number of positive test instances are listed at right. Uncertain findings were mapped to negative, and per-label decision thresholds were selected on the CT-RATE tuning partition and applied unchanged to SNUH.

Supplementary Tables

Supplementary Table 1: Findings-section token lengths in MIMIC-CXR and CT-RATE

Statistic	MIMIC-CXR	CT-RATE
Unique reports	139,723	23,614
Median [IQR]	58 [45–75]	230 [190–284]
Mean $\pm$ SD	62.8 $\pm$ 26.5	243.9 $\pm$ 84.8
Min	3	7
95th percentile	111	402
99th percentile	151	518
Maximum	427	1,174
Exceeding 256 tokens (%)	0.0	36.1
Exceeding 512 tokens (%)	0.0	1.1

Each unique findings section was tokenized without truncation using the CXR-BERT tokenizer, including special tokens. The last two rows give the percentage of sections exceeding each context length. IQR, interquartile range.

Supplementary Table 2: Chest CT report labeling on the 19 findings supported in both cohorts

Finding	CT-RATE			SNUH			Δ F1
	<i>n</i>	AUROC	F1	<i>n</i>	AUROC	F1	
Emphysema	275	1.000	1.000	100	1.000	0.995	−0.005
Thoracic aortic calcification	360	1.000	0.993	51	0.999	0.961	−0.032
Coronary artery calcification	346	0.999	0.993	75	0.995	0.901	−0.092
Pleural effusion	169	1.000	0.991	69	1.000	0.971	−0.020
Bronchiectasis	159	0.998	0.987	105	0.999	0.957	−0.030
Ground-glass opacity	512	0.995	0.974	133	0.973	0.879	−0.095
Consolidation	402	0.991	0.956	71	0.982	0.855	−0.101
Pulmonary nodule	640	0.992	0.955	618	0.971	0.941	−0.015
Subsegmental or linear atelectasis	308	0.991	0.919	147	0.984	0.878	−0.042
Calcified mediastinal or hilar node	43	1.000	0.915	47	1.000	0.940	+0.025
Mediastinal lymphadenopathy	220	0.991	0.901	106	0.975	0.708	−0.193
Hilar lymphadenopathy	84	0.996	0.879	41	0.955	0.340	−0.539
Pleural thickening	45	0.992	0.870	51	0.981	0.796	−0.074
Focal liver lesion	60	0.983	0.859	63	0.988	0.852	−0.007
Pulmonary mass	38	0.997	0.825	40	0.990	0.835	+0.010
Lobar or segmental atelectasis	99	0.990	0.804	58	0.984	0.768	−0.036
Parenchymal scarring or fibrotic band	291	0.952	0.720	141	0.897	0.466	−0.253
Centrilobular nodules or bronchiolitis	49	0.960	0.507	83	0.929	0.491	−0.016
Reticulation or intralobular thickening	41	0.980	0.474	40	0.975	0.626	+0.152
Macro (19 findings)		0.990	0.870		0.978	0.798	−0.072

Fine-tuned Chest2Vec-0.6B on the 19 leaf findings with at least 30 positive test reports in both cohorts, ordered by CT-RATE F1. *n*, positive test reports. F1 is binary, mapping uncertain to negative at per-label thresholds selected on the CT-RATE tuning partition and applied unchanged to SNUH. Δ F1, SNUH minus CT-RATE. Point estimates.

Supplementary Table 3: Anatomy-specific section retrieval (Task 5).

Section	n_c	R@1				R@5			
		Qwen3		Chest2Vec		Qwen3		Chest2Vec	
		0.6B	4B	0.6B	4B	0.6B	4B	0.6B	4B
<i>Chest CT (CT-RATE)</i>									
Average	4,879	0.171	0.216	0.645	0.661	0.256	0.347	0.888	0.887
Lungs	1,303	0.653	0.765	0.940	0.934	0.801	0.900	0.995	0.989
Pleura	306	0.062	0.185	0.910	0.848	0.101	0.431	0.998	0.996
Bones and spine	450	0.043	0.106	0.733	0.781	0.091	0.211	0.992	0.990
Cardiovascular	728	0.121	0.211	0.651	0.645	0.231	0.363	0.926	0.856
Upper abdomen	709	0.163	0.279	0.636	0.564	0.259	0.457	0.948	0.807
Mediastinum and hila	950	0.059	0.084	0.590	0.601	0.108	0.176	0.879	0.819
Chest wall	289	0.210	0.080	0.453	0.508	0.307	0.204	0.677	0.718
Airways and trachea	144	0.059	0.017	0.246	0.408	0.148	0.035	0.688	0.924
<i>Chest radiograph (SRRG)</i>									
Average	11,801	0.134	0.159	0.432	0.434	0.217	0.282	0.730	0.736
Devices	2,853	0.195	0.247	0.750	0.736	0.378	0.413	0.904	0.908
Lungs and airways	2,247	0.261	0.251	0.576	0.585	0.363	0.413	0.761	0.800
Chest wall and MSK	1,614	0.162	0.174	0.478	0.418	0.236	0.263	0.803	0.797
Cardiovascular	1,362	0.084	0.109	0.309	0.339	0.146	0.221	0.734	0.768
Pleura	1,091	0.068	0.132	0.316	0.338	0.109	0.274	0.609	0.632
Mediastinum and hila	1,293	0.062	0.068	0.303	0.313	0.103	0.152	0.659	0.685
Abdominal	1,341	0.104	0.129	0.290	0.309	0.183	0.236	0.638	0.564

n_c , candidate section texts; 9,265 and 21,000 query-section pairs for chest CT and chest radiographs, the latter comprising 3,000 reports under each of seven instructions. Bold: best per row and metric. MSK, musculoskeletal.

Supplementary Table 4: Contribution of training data composition

Training data	Findings-to-impr. R@10		Error discrim. Top-1		Similarity retr. GREEN		Linear probing Macro F1	
	OpenI	SNUH	OpenI	CTRATE	OpenI	SNUH	MIMIC	CTRATE
Chest2Vec-0.6B (full)	0.330	0.769	0.783	0.821	0.590	0.218	0.542	0.714
<i>Modality removed</i>								
No chest radiograph	0.253	0.755	0.559	0.752	0.556	0.207	0.537	0.708
No chest CT	0.350	0.542	0.766	0.701	0.573	0.203	0.534	0.611
<i>Task-family positives removed</i>								
No summarization	0.217	0.511	0.769	0.796	0.571	0.210	0.542	0.702
No classification	0.316	0.758	0.747	0.802	0.540	0.197	0.528	0.687
No section extraction	0.324	0.769	0.774	0.816	0.593	0.209	0.537	0.712

Chest2Vec-0.6B retrained with one component removed; all other settings held constant, each variant trained once. Point estimates; bold marks the largest reduction per column.

Supplementary Table 5: Complete three-level chest CT finding hierarchy

No.	Top-level group	Upper-level category	Assignable leaf finding	Leaf type
1	Lungs & Airways	Parenchymal / small-airways patterns	Consolidation	Clinical finding
2	Lungs & Airways	Parenchymal / small-airways patterns	Ground-glass opacity (GGO)	Clinical finding
3	Lungs & Airways	Parenchymal / small-airways patterns	Crazy-paving pattern	Clinical finding
4	Lungs & Airways	Parenchymal / small-airways patterns	Mosaic attenuation / air-trapping	Clinical finding
5	Lungs & Airways	Parenchymal / small-airways patterns	Tree-in-bud	Clinical finding
6	Lungs & Airways	Parenchymal / small-airways patterns	Centrilobular nodules / bronchiolitis pattern	Clinical finding
7	Lungs & Airways	Nodules and masses	Pulmonary nodule (solid / PSN / GGN)	Clinical finding
8	Lungs & Airways	Nodules and masses	Pulmonary mass (>3 cm)	Clinical finding
9	Lungs & Airways	Nodules and masses	Cavitary nodule / mass	Clinical finding
10	Lungs & Airways	Emphysema & cystic lung disease	Emphysema	Clinical finding
11	Lungs & Airways	Emphysema & cystic lung disease	Bullae / giant bulla	Clinical finding
12	Lungs & Airways	Emphysema & cystic lung disease	Pulmonary cysts / cystic lung disease	Clinical finding
13	Lungs & Airways	Interstitial / fibrotic lung disease	Reticulation / intralobular thickening	Clinical finding
14	Lungs & Airways	Interstitial / fibrotic lung disease	Interlobular septal thickening	Clinical finding
15	Lungs & Airways	Interstitial / fibrotic lung disease	Traction bronchiectasis / bronchiolectasis	Clinical finding
16	Lungs & Airways	Interstitial / fibrotic lung disease	Honeycombing	Clinical finding
17	Lungs & Airways	Interstitial / fibrotic lung disease	Parenchymal scarring / fibrotic band	Clinical finding
18	Lungs & Airways	Trachea & Airways	Tracheal stenosis / malacia	Clinical finding
19	Lungs & Airways	Trachea & Airways	Tracheal / bronchial wall thickening	Clinical finding
20	Lungs & Airways	Trachea & Airways	Bronchiectasis	Clinical finding
21	Lungs & Airways	Trachea & Airways	Mucoid impaction / plugging	Clinical finding
22	Lungs & Airways	Trachea & Airways	Tracheal diverticulum	Clinical finding
23	Lungs & Airways	Trachea & Airways	Endotracheal tube	Other
24	Lungs & Airways	Trachea & Airways	Tracheostomy tube	Other
25	Lungs & Airways	Volume loss / hyperinflation	Lobar / segmental atelectasis	Clinical finding
26	Lungs & Airways	Volume loss / hyperinflation	Subsegmental / linear atelectasis	Clinical finding
27	Lungs & Airways	Post-surgical / transplant status	Post-lobectomy / segmentectomy	Other
28	Lungs & Airways	Post-surgical / transplant status	Post-pneumonectomy	Other
29	Lungs & Airways	Post-surgical / transplant status	Lung transplant	Other
30	Lungs & Airways	Lungs & Airways_others	Lungs & Airways_others	Other

Continued on next page

Supplementary Table 5: Complete three-level chest CT finding hierarchy, continued

No.	Top-level group	Upper-level category	Assignable leaf finding	Leaf type
31	Pleura	Pleural fluid / drain-related findings	Pleural effusion	Clinical finding
32	Pleura	Pleural fluid / drain-related findings	Loculated pleural effusion	Clinical finding
33	Pleura	Pleural fluid / drain-related findings	Hemothorax	Clinical finding
34	Pleura	Pleural fluid / drain-related findings	Chest tube / pleural drain	Other
35	Pleura	Pleural air	Pneumothorax	Clinical finding
36	Pleura	Pleural air	Tension pneumothorax	Clinical finding
37	Pleura	Pleural thickening / masses	Pleural thickening	Clinical finding
38	Pleura	Pleural thickening / masses	Pleural plaques	Clinical finding
39	Pleura	Pleural thickening / masses	Pleural nodule / mass	Clinical finding
40	Pleura	Pleura_others	Pleura_others	Other
41	Mediastinum & Hila	Lymph nodes	Mediastinal lymphadenopathy	Clinical finding
42	Mediastinum & Hila	Lymph nodes	Hilar lymphadenopathy	Clinical finding
43	Mediastinum & Hila	Lymph nodes	Calcified mediastinal / hilar lymph nodes	Clinical finding
44	Mediastinum & Hila	Mediastinal masses / cysts	Anterior mediastinal mass	Clinical finding
45	Mediastinum & Hila	Mediastinal masses / cysts	Middle / posterior mediastinal mass or cyst	Clinical finding
46	Mediastinum & Hila	Mediastinal masses / cysts	Thymic remnant / hyperplasia	Clinical finding
47	Mediastinum & Hila	Esophagus / hiatus / enteric tube	Esophageal wall thickening / mass	Clinical finding
48	Mediastinum & Hila	Esophagus / hiatus / enteric tube	Hiatal hernia	Clinical finding
49	Mediastinum & Hila	Esophagus / hiatus / enteric tube	Esophageal dilation	Clinical finding
50	Mediastinum & Hila	Esophagus / hiatus / enteric tube	Nasogastric / orogastric tube	Other
51	Mediastinum & Hila	Air / fluid in mediastinum	Pneumomediastinum	Clinical finding
52	Mediastinum & Hila	Air / fluid in mediastinum	Mediastinal hematoma / fluid collection	Clinical finding
53	Mediastinum & Hila	Mediastinum & Hila_others	Mediastinum & Hila_others	Other
54	Cardiovascular	Cardiac size and morphology	Cardiomegaly	Clinical finding
55	Cardiovascular	Pericardium	Pericardial effusion	Clinical finding
56	Cardiovascular	Pericardium	Pericardial thickening / calcification	Clinical finding
57	Cardiovascular	Coronary arteries	Coronary artery calcification	Clinical finding
58	Cardiovascular	Coronary arteries	Coronary stent or bypass graft	Other
59	Cardiovascular	Aorta / pulmonary artery	Thoracic aortic calcification	Clinical finding

Continued on next page

Supplementary Table 5: Complete three-level chest CT finding hierarchy, continued

No.	Top-level group	Upper-level category	Assignable leaf finding	Leaf type
60	Cardiovascular	Aorta / pulmonary artery	Thoracic aortic ectasia / dilation (non-aneurysmal)	Clinical finding
61	Cardiovascular	Aorta / pulmonary artery	Thoracic aortic aneurysm	Clinical finding
62	Cardiovascular	Aorta / pulmonary artery	Aortic dissection / intramural hematoma	Clinical finding
63	Cardiovascular	Aorta / pulmonary artery	Main pulmonary artery enlargement	Clinical finding
64	Cardiovascular	Aorta / pulmonary artery	Pulmonary embolism	Clinical finding
65	Cardiovascular	Valves, vascular lines, and cardiac devices	Aortic valve calcification	Clinical finding
66	Cardiovascular	Valves, vascular lines, and cardiac devices	Mitral annular calcification	Clinical finding
67	Cardiovascular	Valves, vascular lines, and cardiac devices	Pacemaker / ICD leads	Other
68	Cardiovascular	Valves, vascular lines, and cardiac devices	Central venous catheter / PICC	Other
69	Cardiovascular	Valves, vascular lines, and cardiac devices	LVAD / other cardiac assist device	Other
70	Cardiovascular	Cardiovascular_others	Cardiovascular_others	Other
71	Chest Wall	Chest wall soft tissues	Chest wall soft tissue edema / hematoma	Clinical finding
72	Chest Wall	Chest wall soft tissues	Subcutaneous emphysema	Clinical finding
73	Chest Wall	Chest wall soft tissues	Chest wall mass	Clinical finding
74	Chest Wall	Post-surgical / tumor invasion	Post-thoracotomy change	Other
75	Chest Wall	Post-surgical / tumor invasion	Chest wall tumor invasion	Clinical finding
76	Chest Wall	Chest Wall_others	Chest Wall_others	Other
77	Bones / Spine	Fractures	Acute rib fracture	Clinical finding
78	Bones / Spine	Fractures	Non-acute / healed rib fracture	Clinical finding
79	Bones / Spine	Fractures	Sternal fracture	Clinical finding
80	Bones / Spine	Fractures	Vertebral compression fracture	Clinical finding
81	Bones / Spine	Degenerative / osseous lesions	Degenerative spine changes	Clinical finding
82	Bones / Spine	Degenerative / osseous lesions	Osteolytic bone lesion	Clinical finding
83	Bones / Spine	Degenerative / osseous lesions	Osteosclerotic bone lesion	Clinical finding
84	Bones / Spine	Degenerative / osseous lesions	Mixed osteolytic-osteosclerotic lesion	Clinical finding
85	Bones / Spine	Degenerative / osseous lesions	Osteopenia	Clinical finding
86	Bones / Spine	Degenerative / osseous lesions	Scoliosis / kyphosis	Clinical finding
87	Bones / Spine	Degenerative / osseous lesions	Vertebral hemangioma	Clinical finding
88	Bones / Spine	Degenerative / osseous lesions	Postoperative spine change / hardware	Other
89	Bones / Spine	Bones / Spine_others	Bones / Spine_others	Other

Continued on next page

Supplementary Table 5: Complete three-level chest CT finding hierarchy, continued

No.	Top-level group	Upper-level category	Assignable leaf finding	Leaf type
90	Upper Abdomen	Liver	Hepatic steatosis	Clinical finding
91	Upper Abdomen	Liver	Focal liver lesion (nodule / mass)	Clinical finding
92	Upper Abdomen	Liver	Hepatomegaly	Clinical finding
93	Upper Abdomen	Liver	Liver contour irregularity / cirrhosis features	Clinical finding
94	Upper Abdomen	Liver	Hepatic calcification	Clinical finding
95	Upper Abdomen	Gallbladder & biliary	Cholelithiasis / gallstones	Clinical finding
96	Upper Abdomen	Gallbladder & biliary	Post-cholecystectomy (gallbladder operated / absent)	Other
97	Upper Abdomen	Gallbladder & biliary	Gallbladder wall thickening	Clinical finding
98	Upper Abdomen	Gallbladder & biliary	Hydropic gallbladder / distension	Clinical finding
99	Upper Abdomen	Gallbladder & biliary	Biliary sludge	Clinical finding
100	Upper Abdomen	Gallbladder & biliary	Biliary stent / catheter / drain	Other
101	Upper Abdomen	Spleen	Splenomegaly	Clinical finding
102	Upper Abdomen	Spleen	Accessory spleen / splenule / polysplenia	Clinical finding
103	Upper Abdomen	Spleen	Focal splenic lesion (nodule / mass)	Clinical finding
104	Upper Abdomen	Pancreas	Pancreatic mass / focal lesion	Clinical finding
105	Upper Abdomen	Pancreas	Pancreatic lipomatosis	Clinical finding
106	Upper Abdomen	Adrenals	Adrenal nodule / mass	Clinical finding
107	Upper Abdomen	Adrenals	Adrenal thickening / hyperplasia	Clinical finding
108	Upper Abdomen	Adrenals	Adrenal calcification	Clinical finding
109	Upper Abdomen	Kidneys / urinary tract	Simple renal cyst	Clinical finding
110	Upper Abdomen	Kidneys / urinary tract	Complex renal cyst / solid renal mass	Clinical finding
111	Upper Abdomen	Kidneys / urinary tract	Hydronephrosis	Clinical finding
112	Upper Abdomen	Kidneys / urinary tract	Renal calculi / nephrolithiasis	Clinical finding
113	Upper Abdomen	Kidneys / urinary tract	Renal atrophy / decreased renal size	Clinical finding
114	Upper Abdomen	Kidneys / urinary tract	Nephrectomy (kidney absent / operated)	Other
115	Upper Abdomen	GI & peritoneum	Ascites	Clinical finding
116	Upper Abdomen	GI & peritoneum	Pneumoperitoneum	Clinical finding
117	Upper Abdomen	GI & peritoneum	Bowel wall thickening / inflammation	Clinical finding
118	Upper Abdomen	GI & peritoneum	Diverticulosis	Clinical finding

Continued on next page

Supplementary Table 5: Complete three-level chest CT finding hierarchy, continued

No.	Top-level group	Upper-level category	Assignable leaf finding	Leaf type
119	Upper Abdomen	GI & peritoneum	Omental caking / peritoneal carcinomatosis	Clinical finding
120	Upper Abdomen	Lymph nodes	Abdominal lymphadenopathy	Clinical finding
121	Upper Abdomen	Upper abdominal vessels	Abdominal aortic aneurysm (partially imaged)	Clinical finding
122	Upper Abdomen	Upper abdominal vessels	Abdominal aortic calcification / atherosclerosis (partially imaged)	Clinical finding
123	Upper Abdomen	Upper abdominal vessels	IVC filter	Other
124	Upper Abdomen	Upper Abdomen_others	Upper Abdomen_others	Other
125	Lower Neck	Thyroid	Thyroid enlargement (goiter)	Clinical finding
126	Lower Neck	Thyroid	Thyroid nodule	Clinical finding
127	Lower Neck	Neck nodes and soft tissues	Cervical / supraclavicular lymphadenopathy	Clinical finding
128	Lower Neck	Neck nodes and soft tissues	Neck soft tissue mass	Clinical finding
129	Lower Neck	Lower Neck_others	Lower Neck_others	Other
130	Others	Breast	Breast mass / focal asymmetry	Clinical finding
131	Others	Breast	Post-lumpectomy / post-mastectomy change	Other
132	Others	Breast	Breast implant (intact or present)	Other
133	Others	Axilla	Axillary lymphadenopathy	Clinical finding
134	Others	Study quality / global	Motion artifact / suboptimal study	Other
135	Others	Study quality / global	Study limitation / limited evaluation (non-motion)	Other
136	Others	Study quality / global	No significant intrathoracic abnormality	Other
137	Others	Others_others	Others_others	Other

Supplementary Table 6: Task-family distribution in the Chest2Vec training corpus

Task family	Chest radiograph	Chest CT	Total
Similar-text retrieval	311,692	44,094	355,786
Summarization	206,604	21,590	228,194
Classification	269,892	219,651	489,543
Section extraction	161,768	139,460	301,228
Total	949,956	424,795	1,374,751

Counts refer to instruction-formatted query-to-answer training records after corpus construction.

Supplementary Table 7: Chest2Vec task instructions and evaluation roles

Task family	Modality or variable	Exact instruction
Findings-to-impression retrieval and error discrimination	Chest CT	Summarize the following chest CT report
Findings-to-impression retrieval and error discrimination	Chest radiograph	Summarize the following chest radiograph report
Clinical similarity retrieval	Chest CT	Retrieve the chest CT report that is similar to the given report.
Clinical similarity retrieval	Chest radiograph	Retrieve the chest radiograph report that is similar to the given report.
Finding classification	Chest CT	Given the following chest CT report, extract the presence/absence of entities
Finding classification	Chest radiograph	Given the following chest radiograph report, extract the presence/absence of entities
Anatomy-specific extraction	{region}	From the following chest CT report, extract and return only the findings related to {region}, ignoring all information about other structures.

Error discrimination uses the findings-to-impression instruction unchanged. The region variable takes one of the prespecified anatomic sections. Complete training strings and exact comparator input formats are provided in the code and model repositories.