Vision Language Models

Akshay Chaudhari

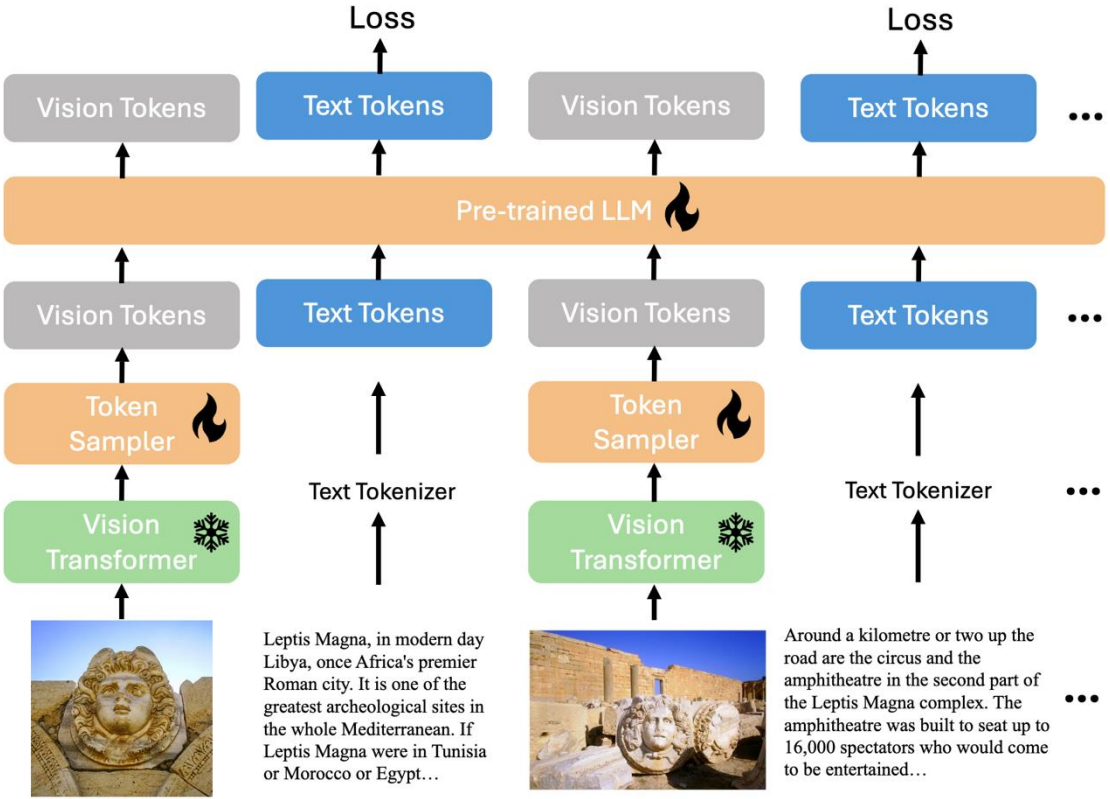
Assistant Professor

Stanford Radiology & Biomedical Data Science

Stanford MEDICINE

Stanford AIMI

Visual Instruction Tuning



Xue et al. xGen-MM (BLIP-3): A Family of Open Large Multimodal Models, 2024.

CheXagent Radiology FM

CheXinstruct
6 Million CXR - Text - QA Triplets

CheXagent
8 Billion Parameter Instruction-tuned CXR FM

CheXbench
Benchmark over 8 tasks and 7 datasets

Local Findings Generation

Q: Given the image(s), describe "Mediastinal".

A: The mediastinal contours are notable only for tortuosity of the aorta.

Open-ended VQA

Q: Where is the opacity located?

A: Right of the midline, superior to the right hilum

Chen Z*, Varma M*, Delbrouck JB* et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024

CheXagent: Medical Imaging FM

CheXinstruct
6 Million CXR - Text - QA Triplets

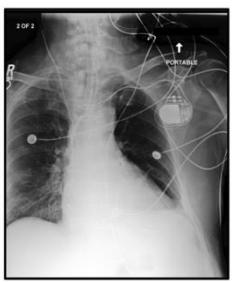
Zhihong Chen

Maya Varma

JB Delbrouck

Chen Z*, Varma M*, Delbrouck JB* et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024

CheXagent: Radiology Report Generation



Residents

156.4s

83.0s

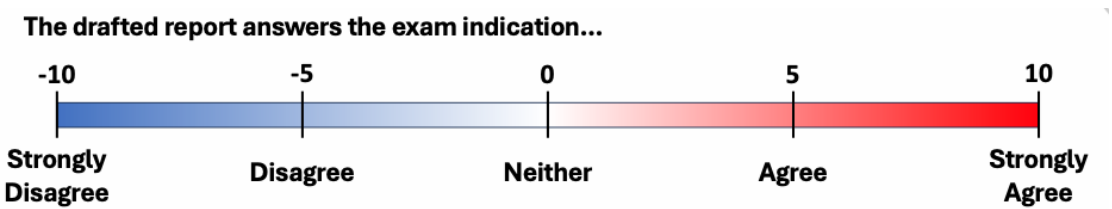
Residents + CheXagent

99.9s

p<0.001

79.7s

Attendings



Residents Rating CheXagent: **5.3** ± 6.0

Attendings rating CheXagent: **4.6** ± 5.9

Attendings rating Residents: **5.6** ± 5.4

Chen Z*, Varma M*, Delbrouck JB* et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024

GREEN: Quantitative + Qualitative Metric

Sophie Ostmeier

```
[Summary]:
Green score: mean 0.23 std 0.04

[Clinically Significant Errors]:
(a) False report of a finding in the candidate: 0.9
[Small right pleural effusion]

(b) Missing a finding present in the reference: 0.7
[Underlying chronic upper lobe scarring.]

(c) Misidentification of a finding's anatomic location/position: 0.4
[The opacity is in the right lower lobe, not the right upper lobe.]

(d) Misassessment of the severity of a finding: 0.8
[Bilateral pleural effusion]

(e) Mentioning a comparison that isn't in the reference: 0.7
[The candidate report mentions a discussion between doctors,
which is not present in the reference report]

(f) Omitting a comparison detailing a change from a prior study: 0.5
[The candidate report does not mention the absence of disease progression]
```

Hugging Face

110k+ downloads to date

Ostmeier et al. GREEN: Generative Radiology Report Evaluation and Error Notation. EMNLP Findings 2024.

Creating Compound Systems

"Reinforcement Learning" w AI Feedback

Dennis Hein

Hein et al. CheXalign: Preference fine-tuning in chest X-ray interpretation models without human feedback. ACL 2025

"Reinforcement Learning" w AI Feedback

Dennis Hein

FM

CheXagent

Report 1

Report 2

Report N

GREEN

~10% improvement without any radiologist feedback

Hein et al. CheXalign: Preference fine-tuning in chest X-ray interpretation models without human feedback. ACL 2025

Why Read Many Words??

	GREEN	
Method	**Avg. length**	**Rel. verbosity**
CheXagent	55.8	
Reference	58.4	1.05

Hein et al. CheXalign: Preference fine-tuning in chest X-ray interpretation models without human feedback. ACL 2025

Can Post-Training Beat Pre-Training?

Model	F1-14	
	Macro (↑)	**Micro** (↑)
GPT-4V	20.4	35.5
MAIRA-1	38.6	55.7
MAIRA-2	41.6	58.1
Med-PaLM M (12B)	37.3	51.4
Med-PaLM M (84B)	39.8	53.6
Med-PaLM M (562B)	37.8	51.6
LLaVA-Rad	39.5	57.3
CheXagent-2	44.6	57.8

Hein et al. CheXalign: Preference fine-tuning in chest X-ray interpretation models without human feedback. ACL 2025

2D to 3D

Merlin 3D CT Foundation Model

Louis Blankemeier

Ashwin Kumar

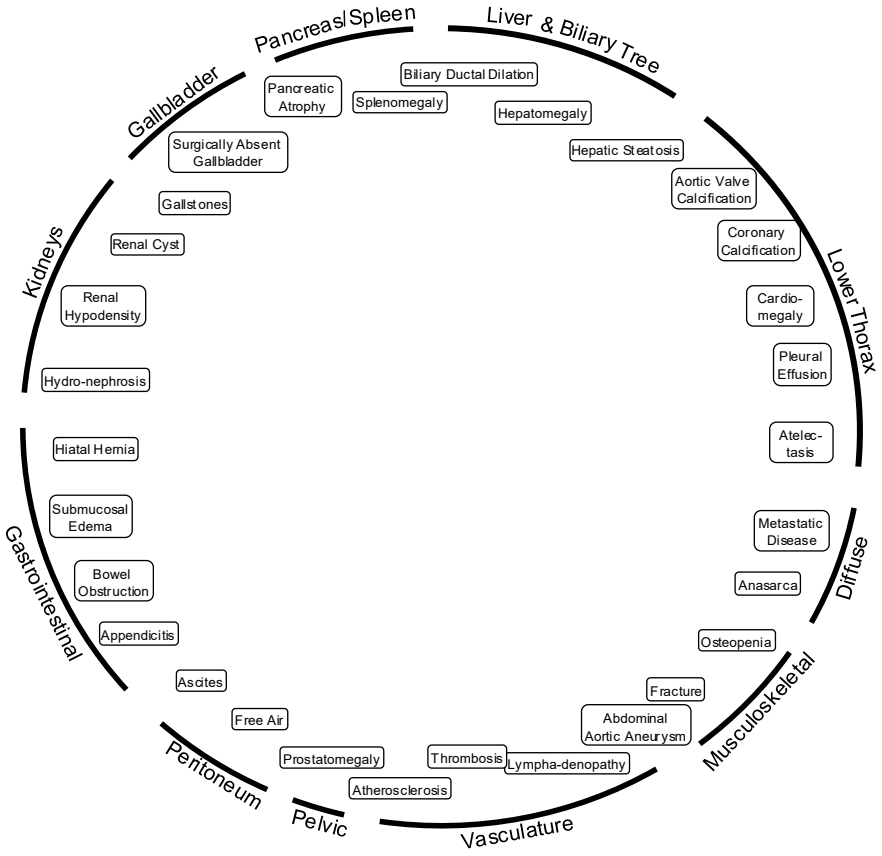
- Pretrained with 20k CT scans + 8M radiology report tokens + 750k ICD codes
- Evaluated on 5k internal + 44k external studies from 3 sites
- Evaluated on in-distribution abdominal CTs + OOD chest CTs

Blankemeier*, Kumar* et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Merlin: 3D Abdominal CT FM

Blankemeier*, Kumar* et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

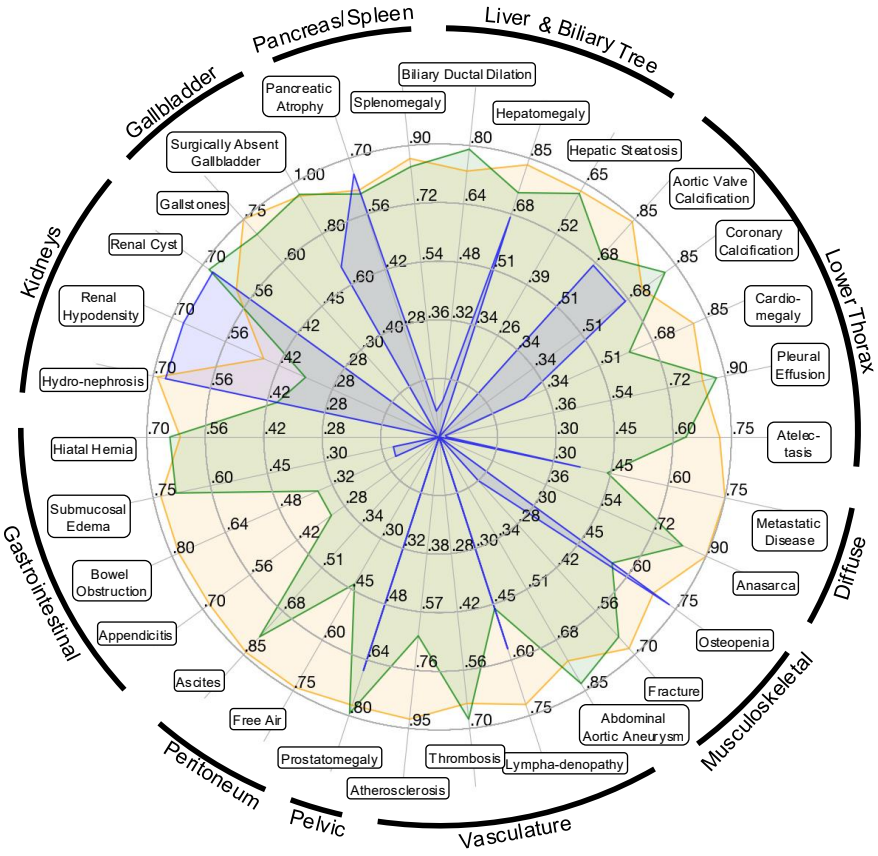
Merlin Capabilities: Zero Shot Classification



Prompt with natural language

Blankemeier et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Merlin Capabilities: Zero Shot Classification



Prompt with natural language

F1 Scores

(not AUROC!)

Merlin (External)

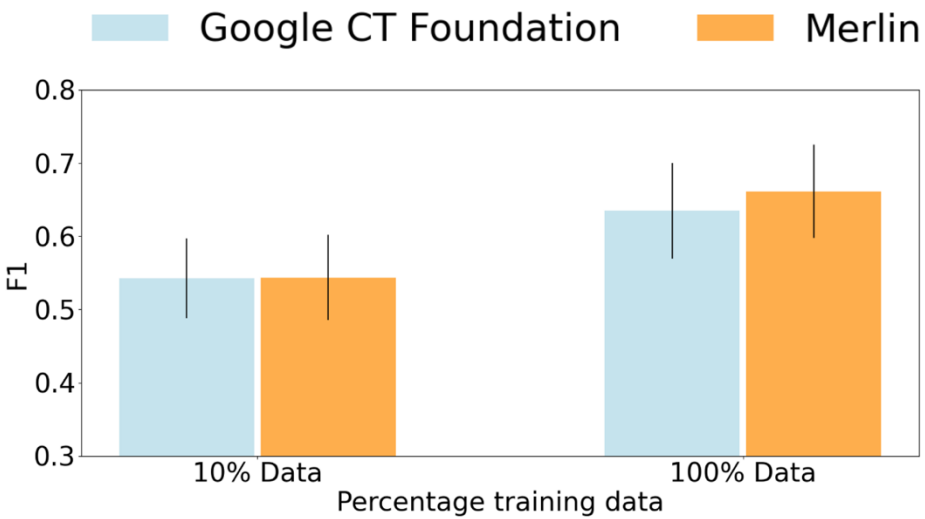
Merlin (Internal)

BioMedCLIP (Internal)

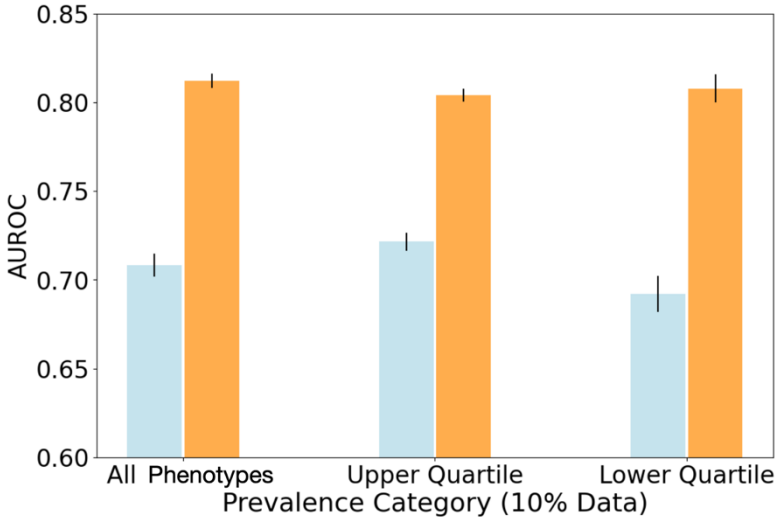
Blankemeier et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Scaling Not Always Beneficial

- Compare to Google CT Foundation trained with 25x more data



Findings Classification

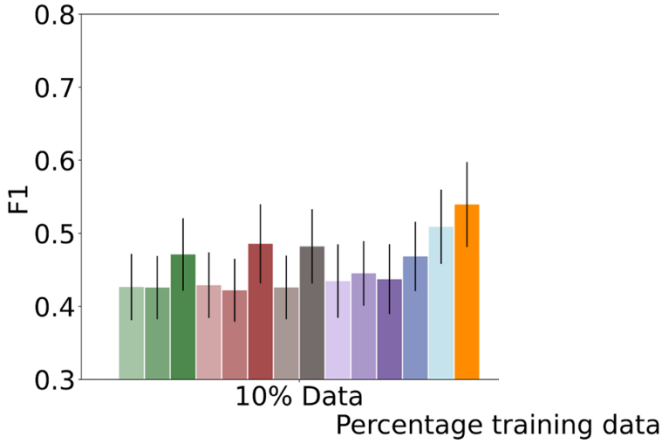


Phenotype Classification

Blankemeier*, Kumar* et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Linear Probing Merlin

Findings-Based Disease Classification

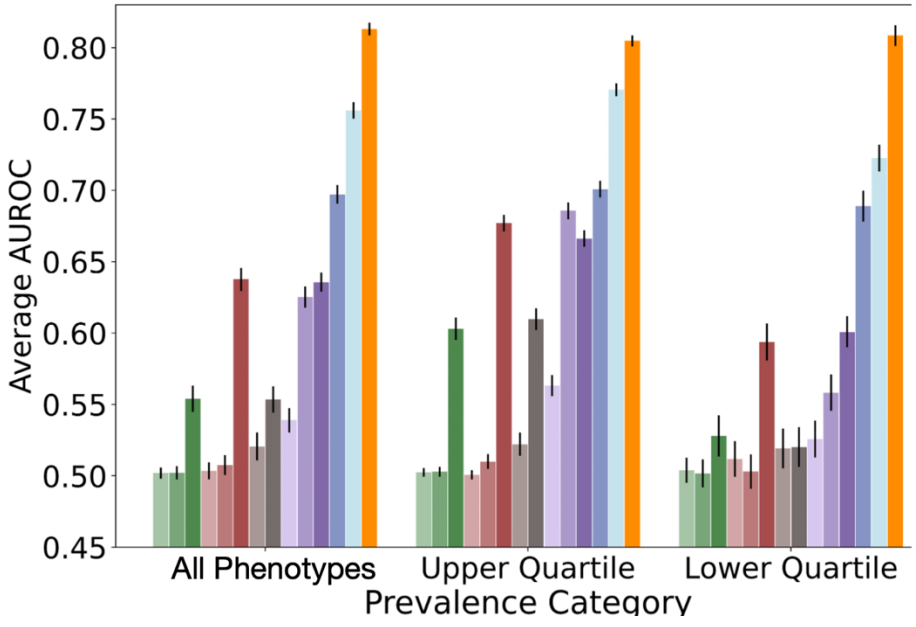


BiomedCLIP FT
BiomedCLIP
2D-to-3D BiomedCLIP
OpenCLIP FT
OpenCLIP
2D-to-3D OpenCLIP
2D ResnetCLIP FT
2D-to-3D ResnetCLIP
3D SwinUnetR SSL
3D ResNet152
3D ResNet152 SSL
MedImageInsight
Google CT
Merlin

Blankemeier*, Kumar* et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Linear Probing Merlin

Phenotype Classification



Blankemeier*, Kumar* et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Extensive Multi-Site Validation

Dataset Characteristics	Train / Validation Set		Internal Test Set		Site #1		Site #2		Site #3	
	n	%	n	%	n	%	n	%	n	%
Abdominal CTs	20,332	–	5,137	–	6,997	–	25,986	–	4,872	–
Age (mean)	54.1 ± 18.9	–	54.5 ± 19.1	–	62.3 ± 14.53	–	60.3 ± 19.3	–	60.3 ± 16.7	–
Sex										
Female	11,377	56.0	2,927	57.0	3,631	52.1	14,132	54.4	2,454	50.4
Male	8,953	44.0	2,210	43.0	3,364	48.2	11,854	45.6	2,418	49.6
Other	2	0.00	0	0.00	1	0.00	0	0.00	1	0.00
kVp Count										
100 kVp	7,901	38.9	1,975	38.4	636	9.08	782	3.00	73	1.50
120 kVp	11,501	56.5	2,911	56.7	4184	59.8	24,779	95.4	4,459	91.5
140 kVp	480	2.36	134	2.60	2155	30.8	2	0.00	268	5.50
Other	450	2.20	117	2.30	22	0.00	423	1.63	72	1.47
Slice thickness										
1-3mm	19,904	99.9	4,995	99.9	132	1.88	25,926	99.7	4,218	94.5
>3mm	18	0.00	4	0.00	6,865	98.1	60	0.00	247	5.53
Tube current (mA)	440	440	–	161	–	220	–	329	–	
Scanner Manufacturer										
GE Healthcare	12,399	61.0	3,119	60.9	6,971	99.6	1,104	4.25	4,458	91.5
Canon	0	0.00	0	0.00	26	0.00	24,881	95.7	0	0.00
Siemens	7,926	39.0	2,001	39.1	0	0.00	1	0.00	241	5.00
Philips	2	0.00	1	0.00	0	0.00	0	0.00	173	3.50

Blankemeier*, Kumar* et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Extensive Multi-Site Validation

Dataset Characteristics	Train / Validation Set		Internal Test Set		Site #1		Site #2		Site #3	
	n	%	n	%	n	%	n	%	n	%
Abdominal CTs	20,332	–	5,137	–	6,997	–	25,986	–	4,872	–
Age (mean)	54.1 ± 18.9	–	54.5 ± 19.1	–	62.3 ± 14.53	–	60.3 ± 19.3	–	60.3 ± 16.7	–
Sex										
Female	11,377	56.0	2,927	57.0	3,631	52.1	14,132	54.4	2,454	50.4
Male	8,953	44.0	2,210	43.0	3,364	48.2	11,854	45.6	2,418	49.6
Other	2	0.00	0	0.00	1	0.00	0	0.00	1	0.00
kVp Count										
100 kVp	7,901	38.9	1,975	38.4	636	9.08	782	3.00	73	1.50
120 kVp	11,501	56.5	2,911	56.7	4184	59.8	24,779	95.4	4,459	91.5
140 kVp	480	2.36	134	2.60	2155	30.8	2	0.00	268	5.50
Other	450	2.20	117	2.30	22	0.00	423	1.63	72	1.47
Slice thickness										
1-3mm	19,904	99.9	4,995	99.9	132	1.88	25,926	99.7	4,218	94.5
>3mm	18	0.00	4	0.00	6,865	98.1	60	0.00	247	5.53
Tube current (mA)	440	440	–	161	–	220	–	329	–	
Scanner Manufacturer										
GE Healthcare	12,399	61.0	3,119	60.9	6,971	99.6	1,104	4.25	4,458	91.5
Canon	0	0.00	0	0.00	26	0.00	24,881	95.7	0	0.00
Siemens	7,926	39.0	2,001	39.1	0	0.00	1	0.00	241	5.00
Philips	2	0.00	1	0.00	0	0.00	0	0.00	173	3.50

Significantly Older Patients in Test Set

Blankemeier*, Kumar* et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Extensive Multi-Site Validation

Dataset Characteristics	Train / Validation Set		Internal Test Set		Site #1		Site #2		Site #3	
	n	%	n	%	n	%	n	%	n	%
Abdominal CTs	20,332	–	5,137	–	6,997	–	25,986	–	4,872	–
Age (mean)	54.1 ± 18.9	–	54.5 ± 19.1	–	62.3 ± 14.53	–	60.3 ± 19.3	–	60.3 ± 16.7	–
Sex										
Female	11,377	56.0	2,927	57.0	3,631	52.1	14,132	54.4	2,454	50.4
Male	8,953	44.0	2,210	43.0	3,365	48.1	11,854	45.6	2,418	49.6
Other	2	0.00	0	0.00	1	0.00	0	0.00	1	0.00
kVp Count										
100 kVp	7,901	38.9	1,975	38.4	636	9.08	782	3.00	73	1.50
120 kVp	11,501	56.5	2,911	56.7	4184	59.8	24,779	95.4	4,459	91.5
140 kVp	480	2.36	134	2.60	2155	30.8	2	0.00	268	5.50
Other	450	2.20	117	2.30	22	0.00	423	1.63	72	1.47
Slice thickness										
1-3mm	19,904	99.9	4,995	99.9	132	1.88	25,926	99.7	4,218	94.5
>3mm	18	0.00	4	0.00	6,865	98.1	60	0.00	247	5.53
Tube current (mA)	440	440	–	161	–	220	–	329	–	
Scanner Manufacturer										
GE Healthcare	12,399	61.0	3,119	60.9	6,971	99.6	1,104	4.25	4,458	91.5
Canon	0	0.00	0	0.00	26	0.00	24,881	95.7	0	0.00
Siemens	7,926	39.0	2,001	39.1	0	0.00	1	0.00	241	5.00
Philips	2	0.00	1	0.00	0	0.00	0	0.00	173	3.50

Different Protocols

Blankemeier*, Kumar* et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Extensive Multi-Site Validation

Dataset Characteristics	Train / Validation Set		Internal Test Set		Site #1		Site #2		Site #3	
	n	%	n	%	n	%	n	%	n	%
Abdominal CTs	20,332	–	5,137	–	6,997	–	25,986	–	4,872	–
Age (mean)	54.1 ± 18.9	–	54.5 ± 19.1	–	62.3 ± 14.53	–	60.3 ± 19.3	–	60.3 ± 16.7	–
Sex										
Female	11,377	56.0	2,927	57.0	3,631	52.1	14,132	54.4	2,454	50.4
Male	8,953	44.0	2,210	43.0	3,364	48.2	11,854	45.6	2,418	49.6
Other	2	0.00	0	0.00	1	0.00	0	0.00	1	0.00
kVp Count										
100 kVp	7,901	38.9	1,975	38.4	636	9.08	782	3.00	73	1.50
120 kVp	11,501	56.5	2,911	56.7	4184	59.8	[illegible]	[illegible]	4,459	91.5
140 kVp	480	2.36	134	2.60	2155	30.8	2	0.00	268	5.50
Other	450	2.20	117	2.30	22	0.00	423	1.63	72	1.47
Slice thickness										
1-3mm	19,904	99.9	4,995	99.9	132	1.88	25,926	99.7	4,218	94.5
>3mm	18	0.00	4	0.00	6,865	98.1	60	0.00	247	5.53
Tube current (mA)	440	440	–	161	–	220	–	329	–	
Scanner Manufacturer										
GE Healthcare	12,399	61.0	3,119	60.9	6,971	99.6	1,104	4.25	4,458	91.5
Canon	0	0.00	0	0.00	26	0.00	24,881	95.7	0	0.00
Siemens	7,926	39.0	2,001	39.1	0	0.00	1	0.00	241	5.00
Philips	2	0.00	1	0.00	0	0.00	0	0.00	173	3.50

Different Scanners

Blankemeier*, Kumar* et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

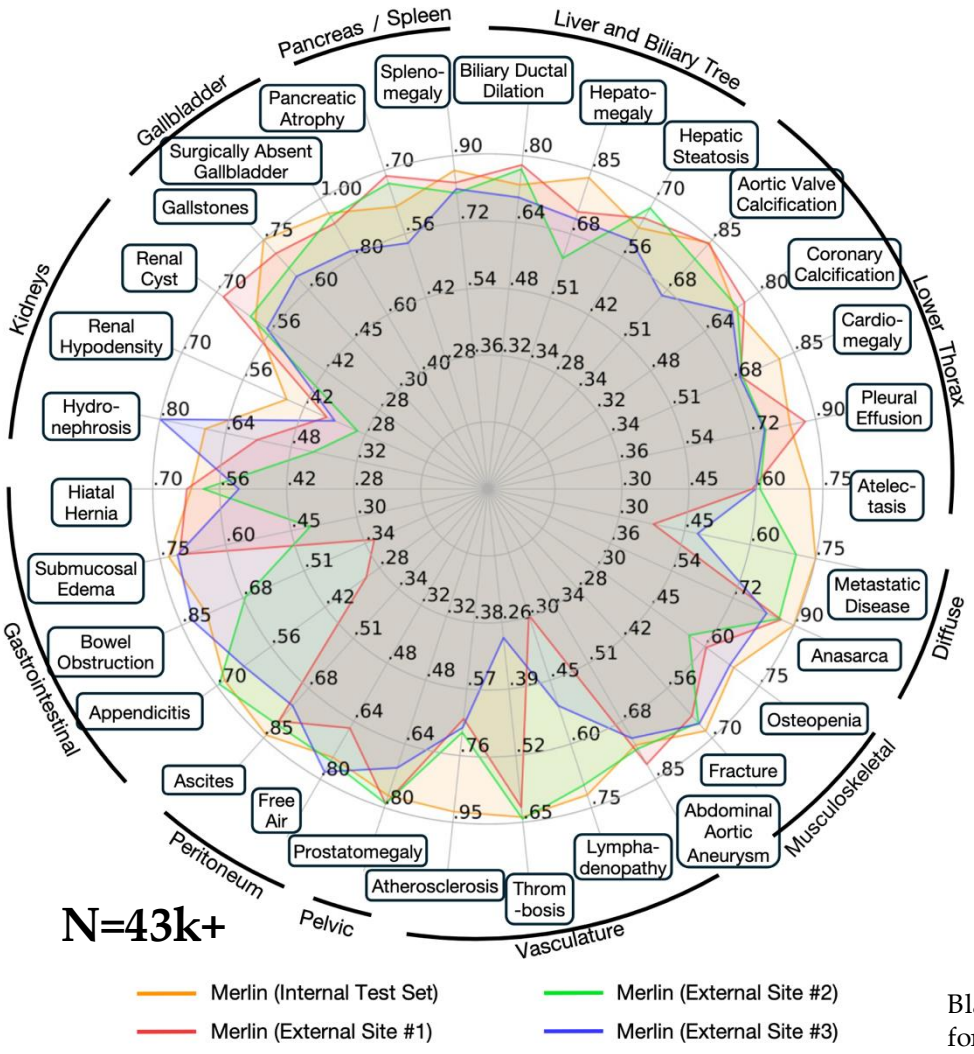
Extensive Multi-Site Validation

Dataset Characteristics	Train / Validation Set		Internal Test Set		Site #1		Site #2		Site #3	
	n	%	n	%	n	%	n	%	n	%
Abdominal CTs	20,332	–	5,137	–	6,997	–	25,986	–	4,872	–
Age (mean)	54.1 ± 18.9	–	54.5 ± 19.1	–	62.3 ± 14.53	–	60.3 ± 19.3	–	60.3 ± 16.7	–
Sex										
Female	11,377	56.0	2,927	57.0	3,631	52.1	14,132	54.4	2,454	50.4
Male	8,953	44.0	2,210	43.0	3,364	48.2	11,854	45.6	[illegible]	[illegible]
Other	2	0.00	0	0.00	1	0.00	0	0.00	1	0.00
kVp Count										
100 kVp	7,901	38.9	1,975	38.4	636	9.08	782	3.00	73	1.50
120 kVp	11,501	56.5	2,911	56.7	4184	59.8	24,779	95.4	4,459	91.5
140 kVp	480	2.36	134	2.60	2155	30.8	2	0.00	268	5.50
Other	450	2.20	117	2.30	22	0.00	423	1.63	72	1.47
Slice thickness										
1-3mm	19,904	99.9	4,995	99.9	132	1.88	25,926	99.7	[illegible]	[illegible]
>3mm	18	0.00	4	0.00	6,865	98.1	60	0.00	247	[illegible]
Tube current (mA)	440	440	–	161	–	220	–	329	[illegible]	
Scanner Manufacturer										
GE Healthcare	12,399	61.0	3,119	60.9	6,971	99.6	1,104	4.25	[illegible]	[illegible]
Canon	0	0.00	0	0.00	26	0.00	24,881	95.7	0	0.00
Siemens	7,926	39.0	2,001	39.1	0	0.00	1	0.00	[illegible]	[illegible]
Philips	2	0.00	1	0.00	0	0.00	0	0.00	173	3.50

Different Protocols

Different Report Drafting Practices

Generalizability Across Institutions



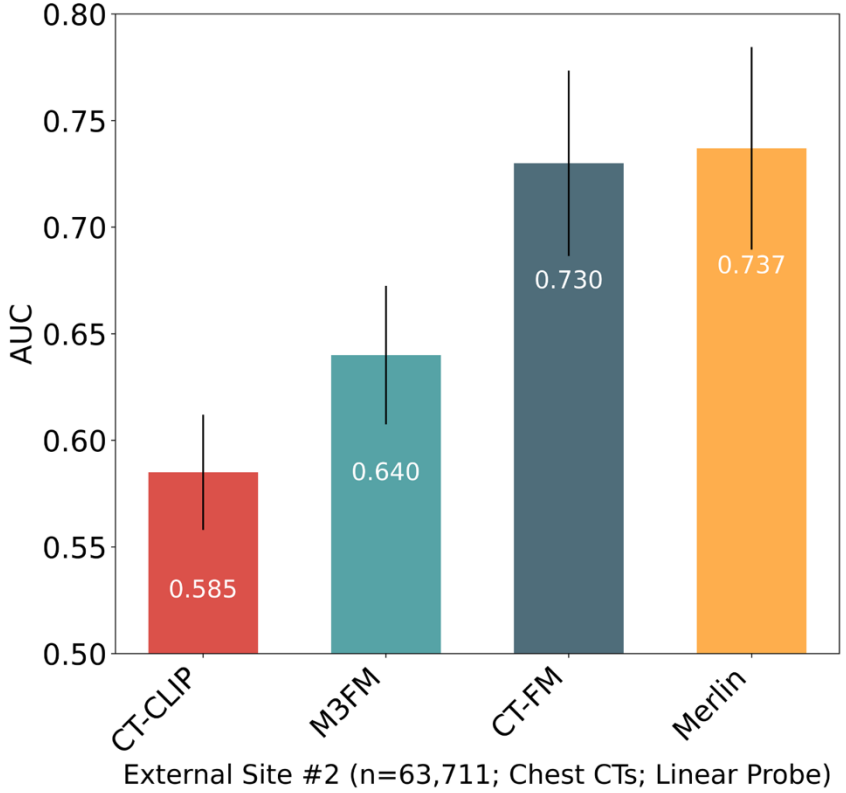
Test set comparisons to 2nd best model:

- +10% Internal, 2D-to-3D OpenCLIP
- +34% External #1, 2D-to-3D BiomedCLIP
- +16% External #2, 2D-to-3D ResNetCLIP
- +9% External #3, 2D-to-3D ResNetCLIP

Blankemeier*, Kumar* et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.

Generalizability Across Anatomies

- Linear probing on chest CT scans w/ frozen encoder
  - 24 chest findings
  - 57k scans for training
  - 6k scans for evaluation
- Similar on COLIPRI evals



Blankemeier*, Kumar* et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2024.