Measuring Performance of Generative AI – Methods and Lessons Learned

Pranav Rajpurkar PhD
Assistant Professor of Biomedical Informatics
Harvard Medical School

RAJPURKAR LAB

Disclosure

Co-founder and Scientist of a2z Radiology AI

Medical Generative AI – Key Use Cases

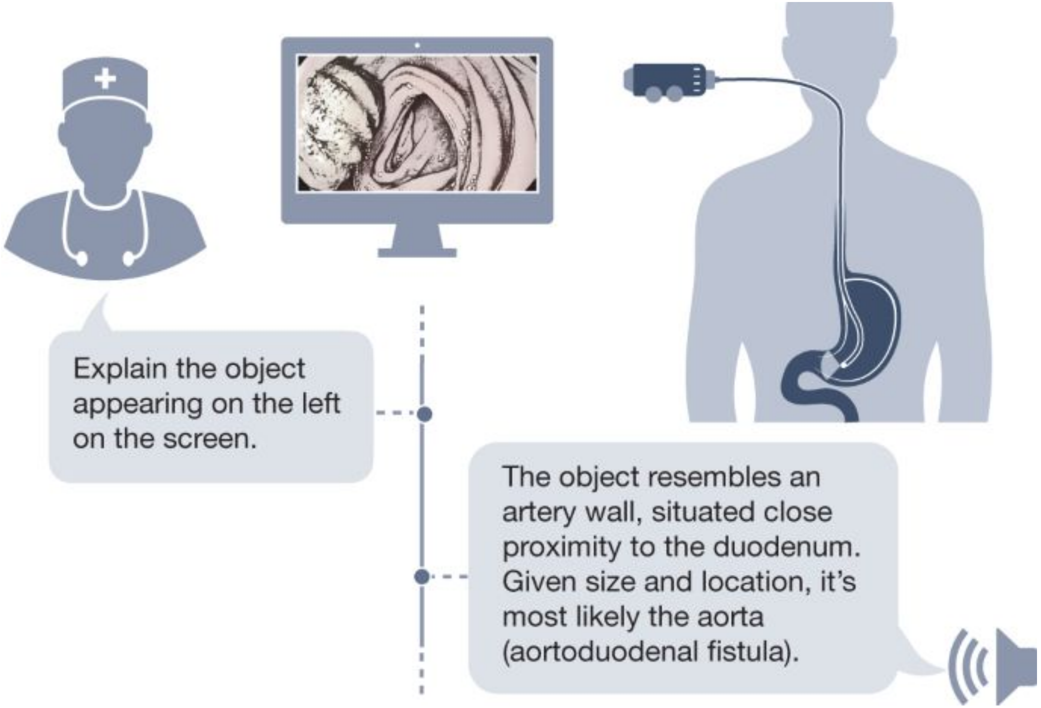
Image → Text
Medical Report Generation

Text → Text
Clinical Note Summarization

Audio → Text
Doctor-Patient Dialog Summary

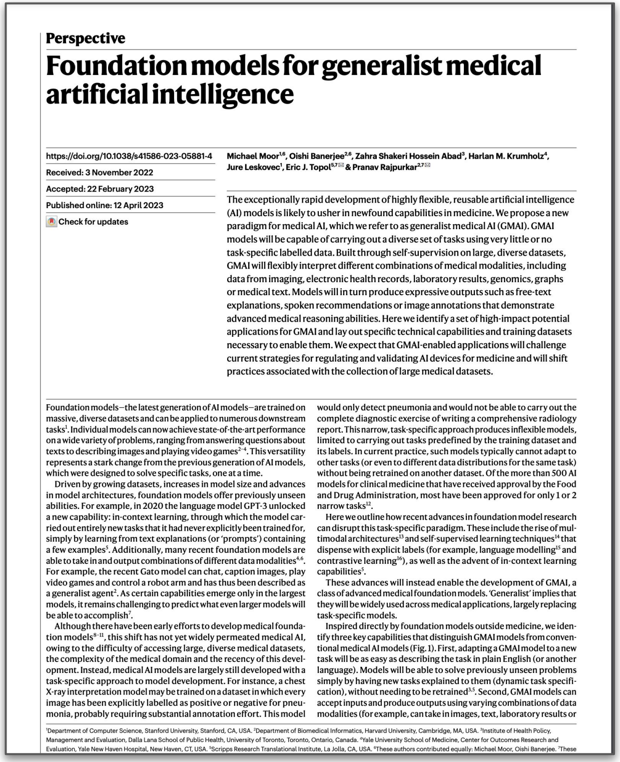
Image → Image
Medical Image Enhancement

Text → Audio-visual
Visualization Generation



Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259-265.

Enabled by Rapid Advances in Generalist AI



Perspective

Foundation models for generalist medical artificial intelligence

https://doi.org/10.1038/s41586-023-05881-4

Received: 3 November 2022

Accepted: 22 February 2023

Published online: 12 April 2023

Check for updates

Michael Moor[1,6], Oishi Banerjee[2,6], Zahra Shakeri Hossein Abad[3], Harlan M. Krumholz[4], Jure Leskovec[1], Eric J. Topol[5,7] ✉ & Pranav Rajpurkar[2,7] ✉

The exceptionally rapid development of highly flexible, reusable artificial intelligence (AI) models is likely to usher in newfound capabilities in medicine. We propose a new paradigm for medical AI, which we refer to as generalist medical AI (GMAI). GMAI models will be capable of carrying out a diverse set of tasks using very little or no task-specific labelled data. Built through self-supervision on large, diverse datasets, GMAI will flexibly interpret different combinations of medical modalities, including data from imaging, electronic health records, laboratory results, genomics, graphs or medical text. Models will in turn produce expressive outputs such as free-text explanations, spoken recommendations or image annotations that demonstrate advanced medical reasoning abilities. Here we identify a set of high-impact potential applications for GMAI and lay out specific technical capabilities and training datasets necessary to enable them. We expect that GMAI-enabled applications will challenge current strategies for regulating and validating AI devices for medicine and will shift practices associated with the collection of large medical datasets.

Foundation models—the latest generation of AI models—are trained on massive, diverse datasets and can be applied to numerous downstream tasks[1]. Individual models can now achieve state-of-the-art performance on a wide variety of problems, ranging from answering questions about texts to describing images and playing video games[2–4]. This versatility represents a stark change from the previous generation of AI models, which were designed to solve specific tasks, one at a time.

Driven by growing datasets, increases in model size and advances in model architectures, foundation models offer previously unseen abilities. For example, in 2020 the language model GPT-3 unlocked a new capability: in-context learning, through which the model carried out entirely new tasks that it had never explicitly been trained for, simply by learning from text explanations (or 'prompts') containing a few examples[5]. Additionally, many recent foundation models are able to take in and output combinations of different data modalities[4,6]. For example, the recent Gato model can chat, caption images, play video games and control a robot arm and has thus been described as a generalist agent[2]. As certain capabilities emerge only in the largest models, it remains challenging to predict what even larger models will be able to accomplish[7].

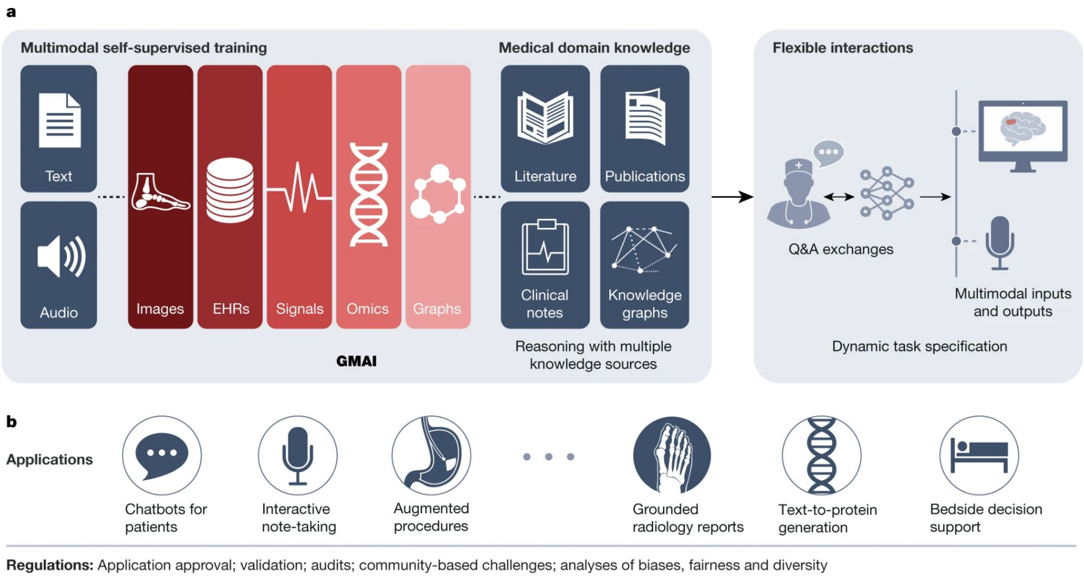
Although there have been early efforts to develop medical foundation models[8–11], this shift has not yet widely permeated medical AI, owing to the difficulty of accessing large, diverse medical datasets, the complexity of the medical domain and the recency of this development. Instead, medical AI models are largely still developed with a task-specific approach to model development. For instance, a chest X-ray interpretation model may be trained on a dataset in which every image has been explicitly labelled as positive or negative for pneumonia, probably requiring substantial annotation effort. This model would only detect pneumonia and would not be able to carry out the complete diagnostic exercise of writing a comprehensive radiology report. This narrow, task-specific approach produces inflexible models, limited to carrying out tasks predefined by the training dataset and its labels. In current practice, such models typically cannot adapt to other tasks (or even to different data distributions for the same task) without being retrained on another dataset. Of the more than 500 AI models for clinical medicine that have received approval by the Food and Drug Administration, most have been approved for only 1 or 2 narrow tasks[12].

Here we outline how recent advances in foundation model research can disrupt this task-specific paradigm. These include the rise of multimodal architectures[13] and self-supervised learning techniques[14] that dispense with explicit labels (for example, language modelling[15] and contrastive learning[16]), as well as the advent of in-context learning capabilities[5].

These advances will instead enable the development of GMAI, a class of advanced medical foundation models. 'Generalist' implies that they will be widely used across medical applications, largely replacing task-specific models.

Inspired directly by foundation models outside medicine, we identify three key capabilities that distinguish GMAI models from conventional medical AI models (Fig. 1). First, adapting a GMAI model to a new task will be as easy as describing the task in plain English (or another language). Models will be able to solve previously unseen problems simply by having new tasks explained to them (dynamic task specification), without needing to be retrained[3,5]. Second, GMAI models can accept inputs and produce outputs using varying combinations of data modalities (for example, can take in images, text, laboratory results or

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Department of Biomedical Informatics, Harvard University, Cambridge, MA, USA. [3]Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada. [4]Yale University School of Medicine, Center for Outcomes Research and Evaluation, Yale New Haven Hospital, New Haven, CT, USA. [5]Scripps Research Translational Institute, La Jolla, CA, USA. [6]These authors contributed equally: Michael Moor, Oishi Banerjee. [7]These



Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259-265.

How do we evaluate a system that is not limited to a narrow use case?

Generalist AI System
Can generate text outputs for a variety of inputs

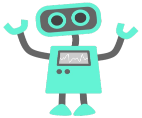
Automatic Metrics
Successes and Failures

Human-Centered Evaluation
Formulations and Pitfalls

RAJPURKAR LAB

Rao et al. Multimodal Generative AI: The Future of Automated Medical Report Generation From Medical Images. To Appear 2025

What metrics allow us to determine whether an AI-written report match a expert-written report



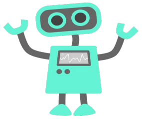
Left lower lobe consolidation without pleural effusion. Air bronchograms are present.

Dense opacity in left base with small pleural fluid collection.
Air bronchograms noted.

RAJPURKAR LAB

Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., ... & Rajpurkar, P. (2023). Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, *4*(9).

1. Traditional natural language generation metrics can measure similarity of vocabulary and phrases



Left lower lobe consolidation without pleural effusion. **Air bronchograms** are present.

Dense opacity in left base with small pleural fluid collection.
Air bronchograms noted.

BLEU-2 Score

✓ Catches exact matches:

- Perfect match on "air bronchograms"
- Reliable when identical terms used

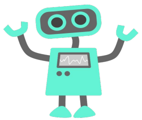
❌ But fails on synonyms:

- "left lower" ≠ "left base"
- "consolidation" ≠ "opacity"
- "effusion" ≠ "fluid collection"

RAJPURKAR LAB

Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., ... & Rajpurkar, P. (2023). Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, *4*(9).

2. Embedding-based metrics match words with similar meanings using language models



Left lower lobe **consolidation** without pleural effusion. Air bronchograms are present.

Dense **opacity** in *left base* with small pleural fluid collection.
Air bronchograms noted.

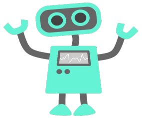
BERTScore
✓ Strong at semantic similarity:
- "consolidation" ≈ "opacity"
- "effusion" ≈ "fluid collection"
- "lower lobe" ≈ "base"

❌ Cannot distinguish negations:
- Treats "without" and "with" as similar
- May miss opposite meanings

RAJPURKAR LAB

Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., ... & Rajpurkar, P. (2023). Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, *4*(9).

3. Clinical accuracy metrics like CheXbert evaluate medical content rather than just text



Left lower lobe consolidation **without** pleural effusion. Air bronchograms are present.

Dense opacity in left base **with** small pleural fluid collection.
Air bronchograms noted.

CheXbert

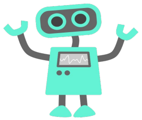
✓ Excellent at finding extraction:
- Correctly identifies presence/absence
- Maps synonyms to standard terms
- Preserves negation correctly

❌ Limited scope:
- Predefined set of 14 findings
- Cannot link location to finding
- No anatomical context mapping

RAJPURKAR LAB

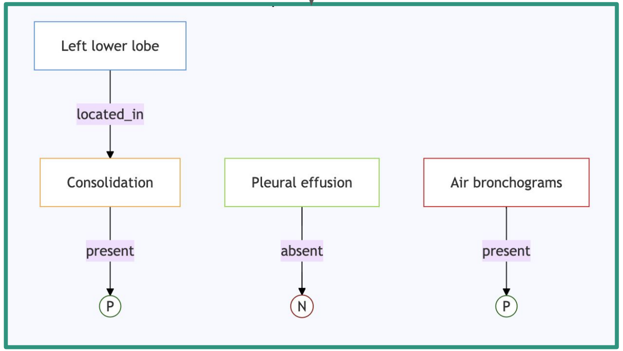
Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A. Y., & Lungren, M. P. (2020). CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv preprint arXiv:2004.09167.

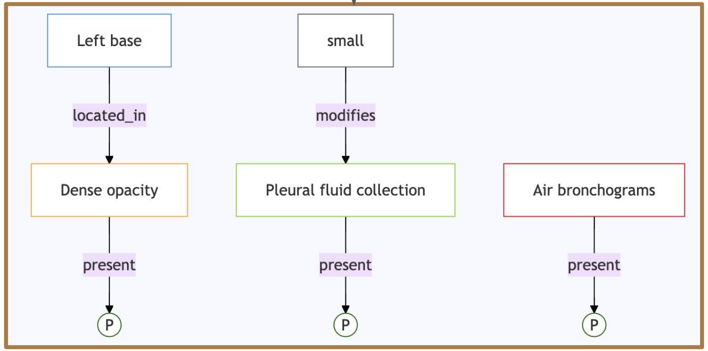
4. RadGraph-F1 extends the medical terms and capture relationships between medical findings



Left lower lobe consolidation without pleural effusion. Air bronchograms are present.

Dense opacity in left base with small pleural fluid collection.
Air bronchograms noted.

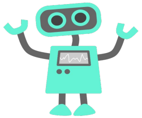




RAJPURKAR LAB

Jain, S., Agrawal, A., Saporta, A., Truong, S. Q., Duong, D. N., Bui, T., ... & Rajpurkar, P. (2021). Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.
Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., ... & Rajpurkar, P. (2023). Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, *4*(9).

4. RadGraph-F1 extends the medical terms and capture relationships between medical findings



Left lower lobe consolidation without pleural effusion. Air bronchograms are present.

Dense opacity in left base with small pleural fluid collection.
Air bronchograms noted.

RadGraph

✓ Comprehensive scope:

- Extensive finding scope
- Preserves anatomical context
- Handles negation properly
- Links findings to locations

❌ Higher complexity:

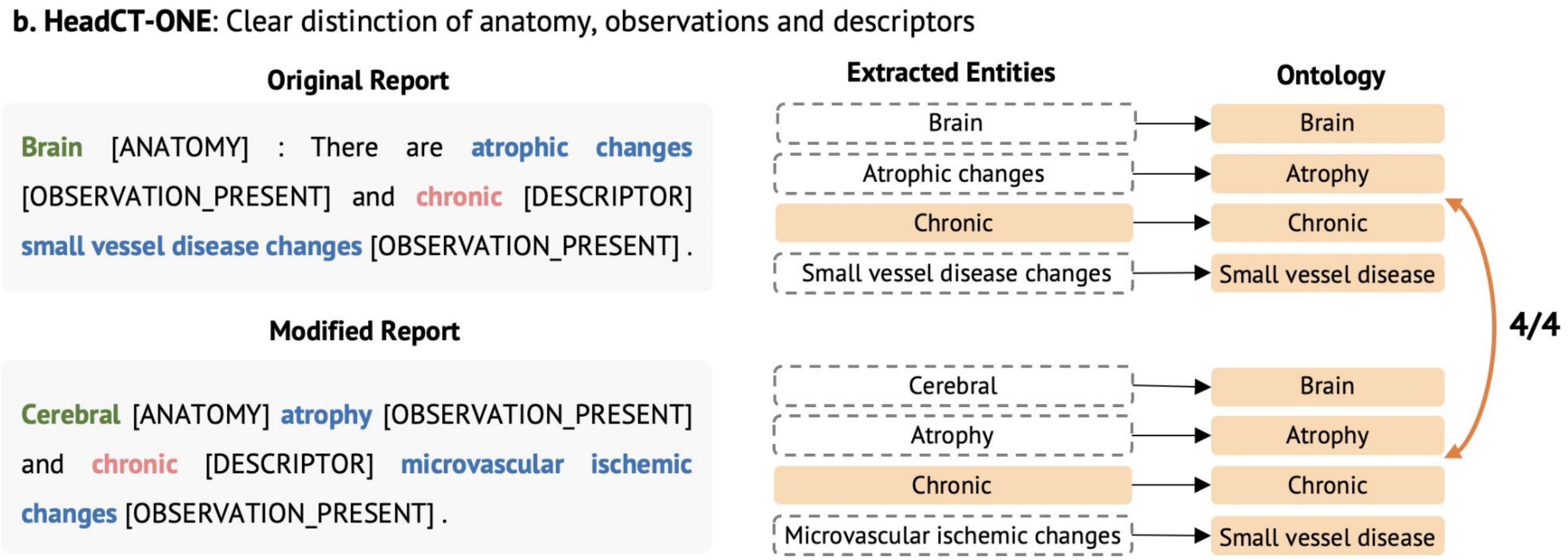
- Harder to extend to modalities
- No normalization over entities

RAJPURKAR LAB

Jain, S., Agrawal, A., Saporta, A., Truong, S. Q., Duong, D. N., Bui, T., ... & Rajpurkar, P. (2021). Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., ... & Rajpurkar, P. (2023). Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, *4*(9).

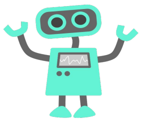
Emerging methodologies like HeadCT-One compare using knowledge ontologies



RAJPURKAR LAB

Acosta, J. N., Zhang, X., Dogra, S., Zhou, H. Y., Payabvash, S., Falcone, G. J., ... & Rajpurkar, P. (2024). HeadCT-ONE: Enabling Granular and Controllable Automated Evaluation of Head CT Radiology Report Generation. arXiv preprint arXiv:2409.13038.

How well does expert scoring will align with these metrics?



Left lower lobe consolidation without pleural effusion. Air bronchograms are present.

Dense opacity in left base with small pleural fluid collection.
Air bronchograms noted.

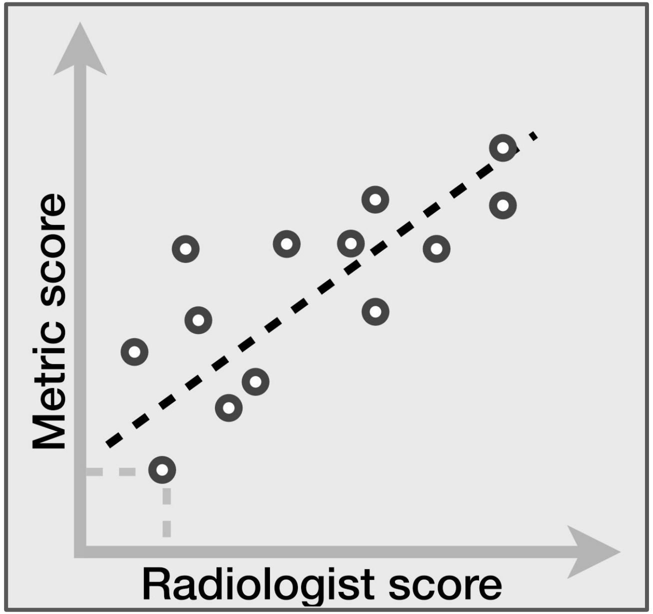
1 significant error
1 insignificant error

BLEU 2
BERTScore
CheXbert
RadGraph

RAJPURKAR LAB

Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., ... & Rajpurkar, P. (2023). Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, *4*(9).

How well does expert scoring will align with these metrics?



1 significant error
1 insignificant error

BLEU 2
BERTScore
CheXbert
RadGraph

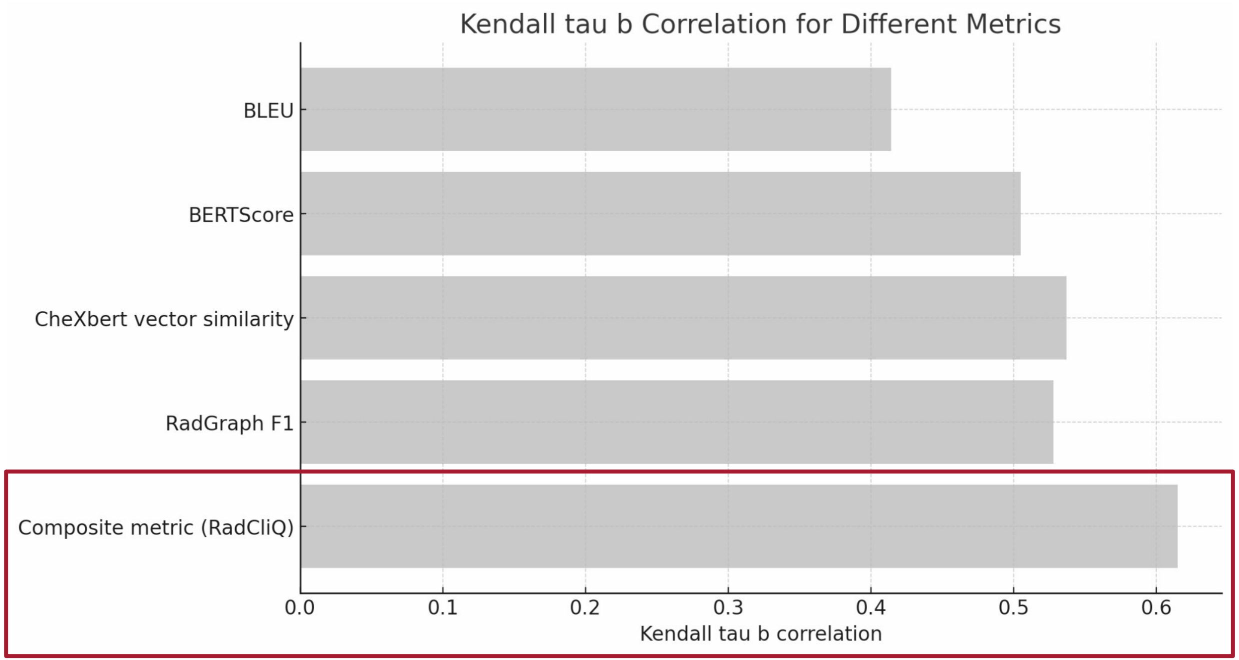
Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., ... & Rajpurkar, P. (2023). Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, *4*(9).

Expert scoring reveals highest alignment with RadGraph-F1 of these 4 metrics

RAJPURKAR LAB

Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., ... & Rajpurkar, P. (2023). Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, *4*(9).

Novel metric RadCliQ, a weighted combination of these metrics, had highest alignment with experts



RAJPURKAR LAB

Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., ... & Rajpurkar, P. (2023). Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, *4*(9).
Code: https://github.com/rajpurkarlab/CXR-Report-Metric

A strong need to understand source and type of errors in generations

Which **parts** of the generation have errors

What is the **clinical significance** of each error

Proceedings of Machine Learning Research 252:1–29, 2024 Machine Learning for Healthcare

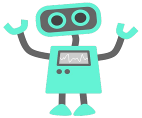
FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores

Alyssa Huang ALYSSAHUANG@COLLEGE.HARVARD.EDU
Harvard University
Oishi Banerjee OISHI_BANERJEE@G.HARVARD.EDU
Harvard University
Kay Wu KAY.WU@MEDPORTAL.CA
Harvard University
Eduardo Pontes Reis EDUARDO.REIS@EINSTEIN.BR
Stanford University
Hospital Israelita Albert Einstein
Pranav Rajpurkar PRANAV_RAJPURKAR@HMS.HARVARD.EDU
Harvard University

RAJPURKAR LAB

Huang, A., Banerjee, O., Wu, K., Reis, E. P., & Rajpurkar, P. (2024). FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores. arXiv preprint arXiv:2405.20613.

ReFiSco: Report Fix and Score Dataset collects expert annotations to fix errors in generations



[1] "Findings suggestive of right lower lobe pneumonia and left lower lobe atelectasis"
[2] "No pleural effusion or pneumothorax present"

Fix

Remove: [Line 1]
Edit: "Small bilateral pleural effusions"
Add: "Bilateral pneumoperitoneum"

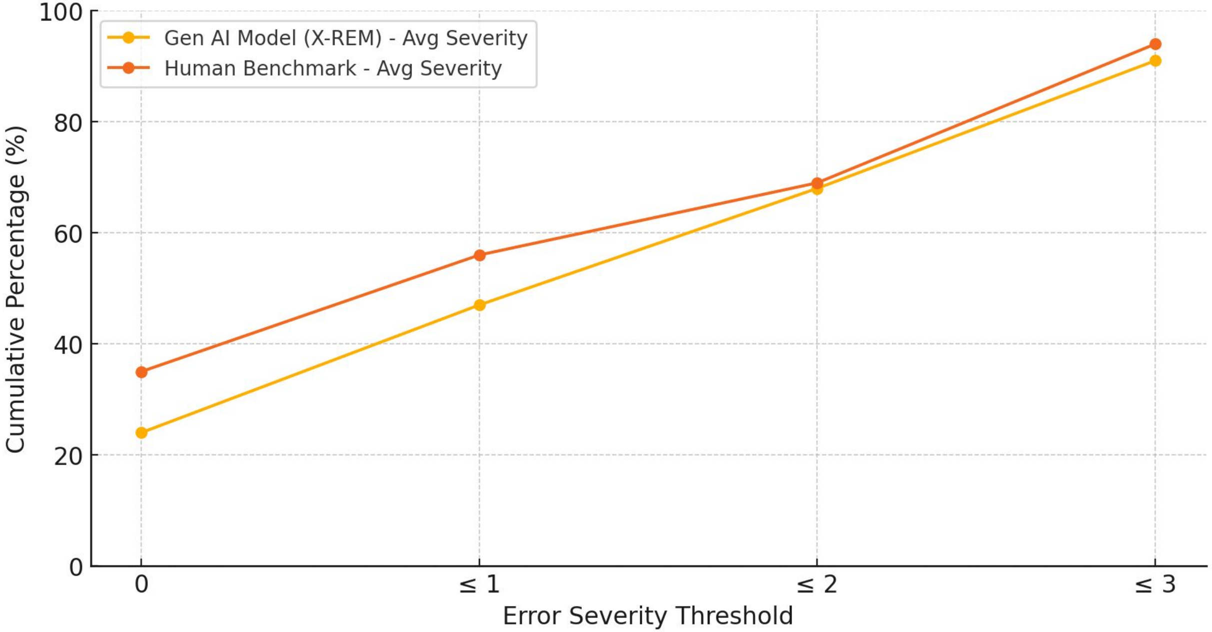
Score

Actionable nonurgent error (2)
Actionable nonurgent error (2)
Emergent error (3)

RAJPURKAR LAB

Tian, K., Hartung, S. J., Li, A. A., Jeong, J., Behzadi, F., Calle-Toro, J., Adithan, S., Pohlen, M., Osayande, D., & Rajpurkar, P. (2023). ReFiSco: Report Fix and Score Dataset for Radiology Report Generation (version 0.0). PhysioNet. https://doi.org/10.13026/cneg-zk64.

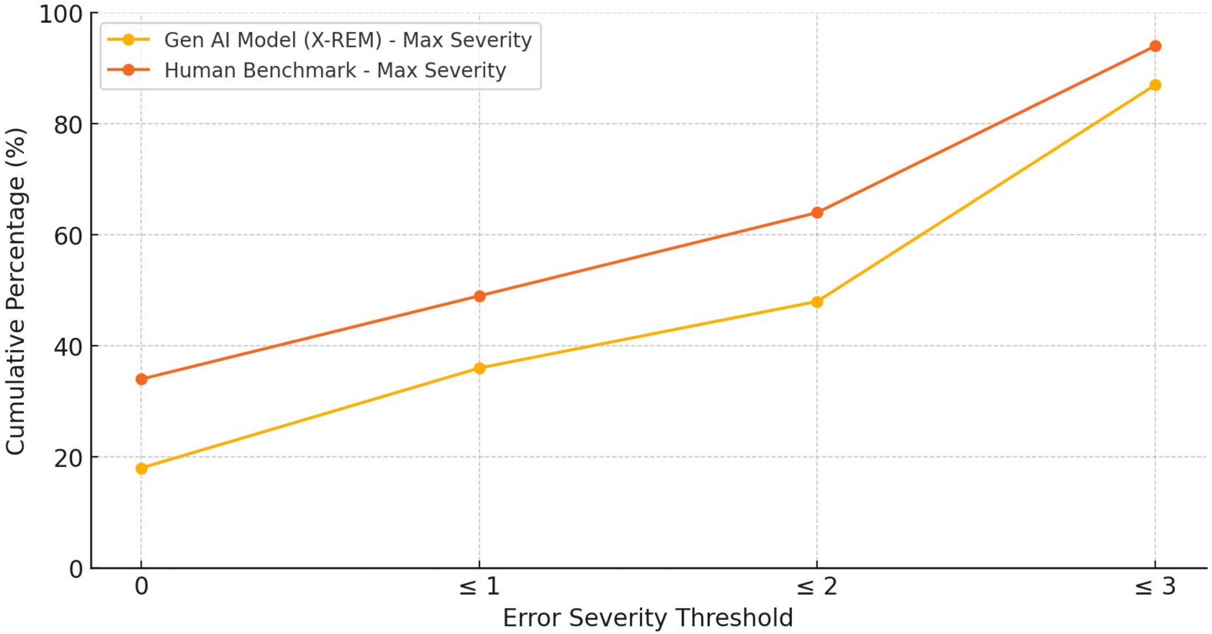
For average severity, 68% of AI reports and 69% of human reports have errors of 2 or less



RAJPURKAR LAB

Jeong, J., Tian, K., Li, A., Hartung, S., Adithan, S., Behzadi, F., ... & Rajpurkar, P. (2024, January). Multimodal image-text matching improves retrieval-based chest x-ray report generation. In Medical Imaging with Deep Learning (pp. 978-990). PMLR.

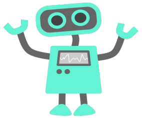
In maximum severity, however, only 48% of AI reports stay ≤2, compared to 64% for humans



Jeong, J., Tian, K., Li, A., Hartung, S., Adithan, S., Behzadi, F., ... & Rajpurkar, P. (2024, January). Multimodal image-text matching improves retrieval-based chest x-ray report generation. In Medical Imaging with Deep Learning (pp. 978-990). PMLR.

RAJPURKAR LAB

Can we automate the process of fixing a AI generated report given access to expert report?



[1] "Findings suggestive of right lower lobe pneumonia and left lower lobe atelectasis"
[2] "No pleural effusion or pneumothorax present"

Fix

Remove: [Line 1]
Edit: "Small bilateral pleural effusions"
Add: "Bilateral pneumoperitoneum"

Score

Actionable nonurgent error (2)
Actionable nonurgent error (2)
Emergent error (3)

RAJPURKAR LAB

Tian, K., Hartung, S. J., Li, A. A., Jeong, J., Behzadi, F., Calle-Toro, J., Adithan, S., Pohlen, M., Osayande, D., & Rajpurkar, P. (2023). ReFiSco: Report Fix and Score Dataset for Radiology Report Generation (version 0.0). PhysioNet. https://doi.org/10.13026/cneg-zk64.

MedVersa – A Generalist Medical AI For Imaging

A Generalist Learner for Multifaceted Medical Image Interpretation

Authors: Hong-Yu Zhou PhD[1], Subathra Adithan MD[2], Julián Nicolás Acosta MD[1], Eric J. Topol MD[3], Pranav Rajpurkar PhD[1]

Affiliations:
1. Department of Biomedical Informatics, Harvard Medical School, Boston, USA.
2. Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, IN.
3. Scripps Research Translational Institute, Scripps Research, La Jolla, CA, USA.

Corresponding author:
Pranav Rajpurkar, PhD
pranav_rajpurkar@hms.harvard.edu

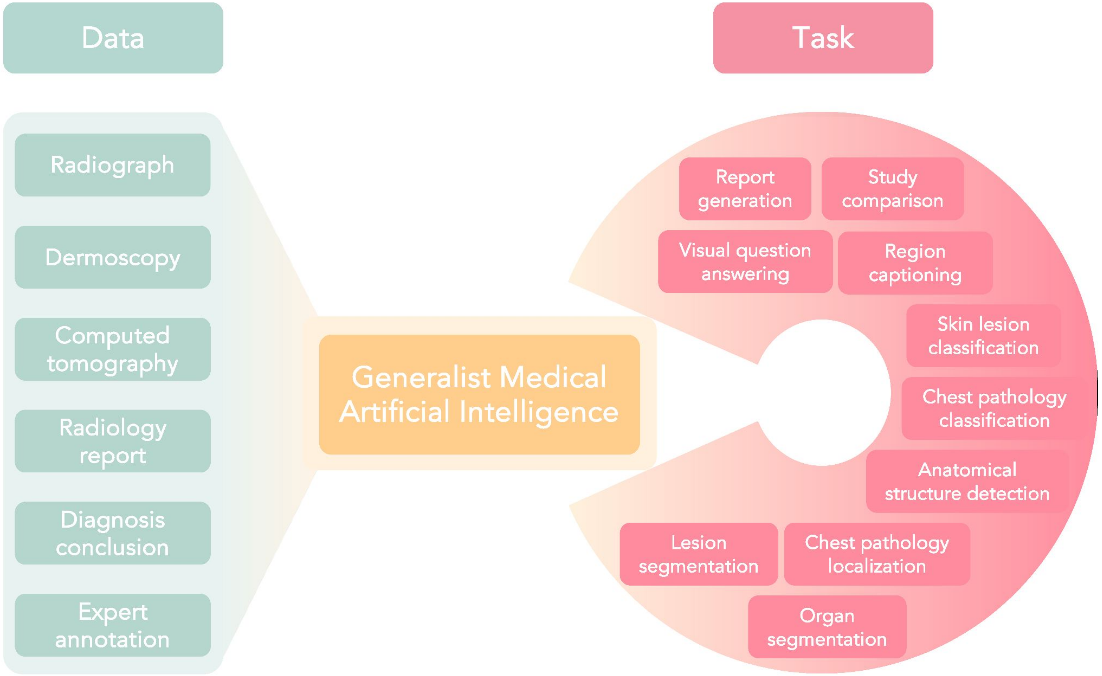
Abstract

Current medical artificial intelligence systems are often limited to narrow applications, hindering their widespread adoption in clinical practice. To address this limitation, we propose MedVersa, a generalist learner that enables flexible learning and tasking for medical image interpretation. By leveraging a large language model as a learnable orchestrator, MedVersa can learn from both visual and linguistic supervision, support multimodal inputs, and perform real-time task specification. This versatility allows MedVersa to adapt to various clinical scenarios and perform multifaceted medical image analysis. We introduce MedInterp, the largest multimodal dataset to date for medical image interpretation, consisting of over 13 million annotated instances spanning 11 tasks across 3 modalities, to support the development of MedVersa. Our experiments demonstrate that MedVersa achieves state-of-the-art performance in 9 tasks, sometimes outperforming specialist counterparts by over 10%. MedVersa is the first to showcase the viability of multimodal generative medical AI in implementing multimodal outputs, inputs, and dynamic task specification, highlighting its potential as a multifunctional system for comprehensive medical image analysis. This generalist approach to medical image interpretation paves the way for more adaptable and efficient AI-assisted clinical decision-making.

RAJPURKAR LAB

Zhou, H. Y., Adithan, S., Acosta, J. N., Topol, E. J., & Rajpurkar, P. (2024). A Generalist Learner for Multifaceted Medical Image Interpretation. *arXiv preprint arXiv:2405.07988*.

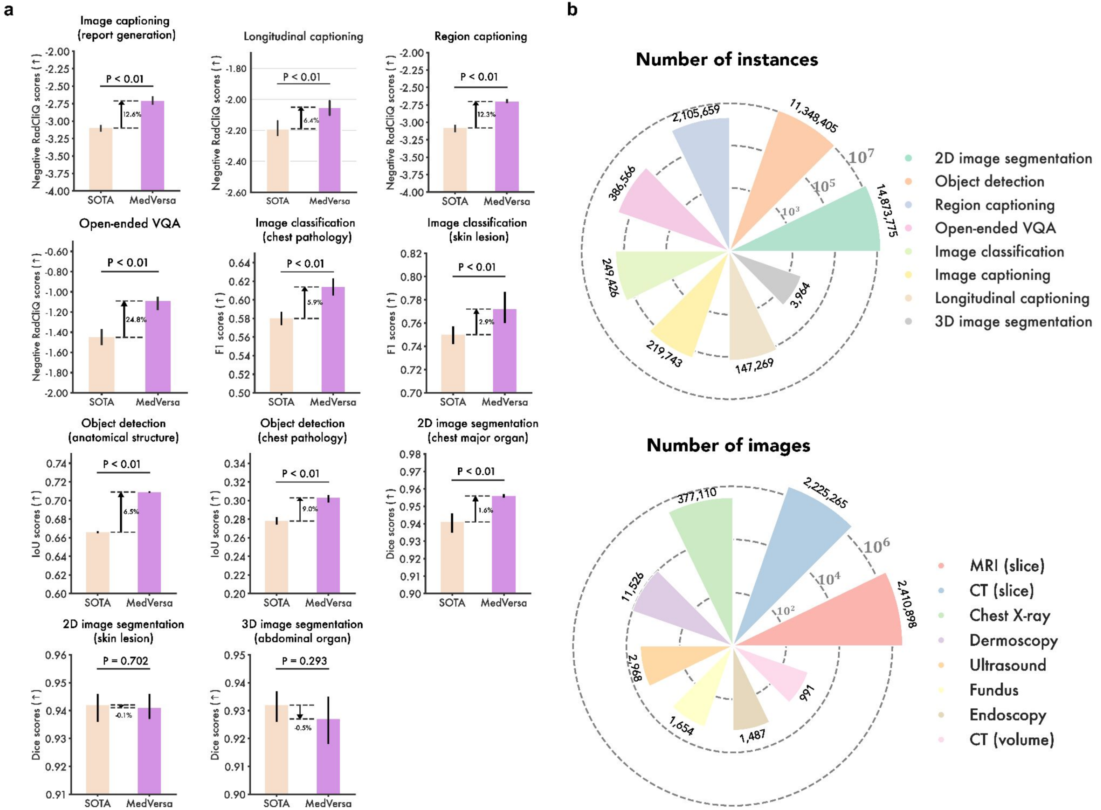
A single model capable of doing a variety of tasks on different modalities



RAJPURKAR LAB

Zhou, H. Y., Adithan, S., Acosta, J. N., Topol, E. J., & Rajpurkar, P. (2024). A Generalist Learner for Multifaceted Medical Image Interpretation. *arXiv preprint arXiv:2405.07988*.

Top performing AI model across many tasks

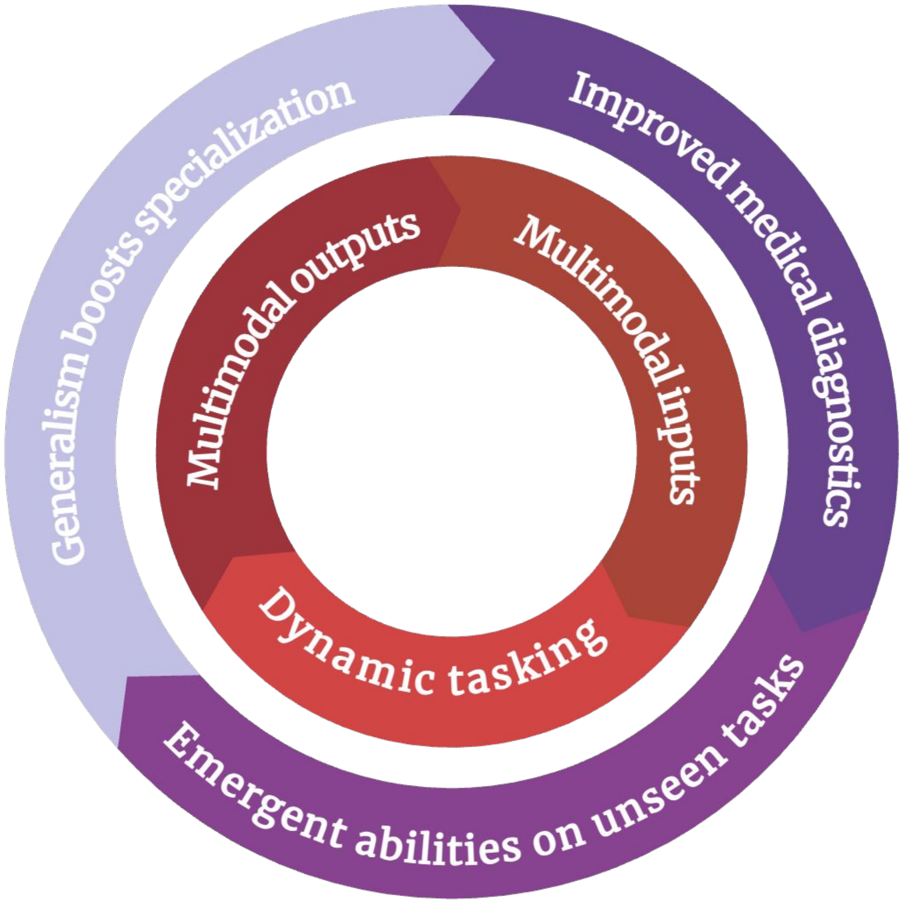


Zhou, H. Y., Adithan, S., Acosta, J. N., Topol, E. J., & Rajpurkar, P. (2024). A Generalist Learner for Multifaceted Medical Image Interpretation. *arXiv preprint arXiv:2405.07988*.

RAJPURKAR LAB

Hypothesis: one generalist model will beat lots of individual specialist models

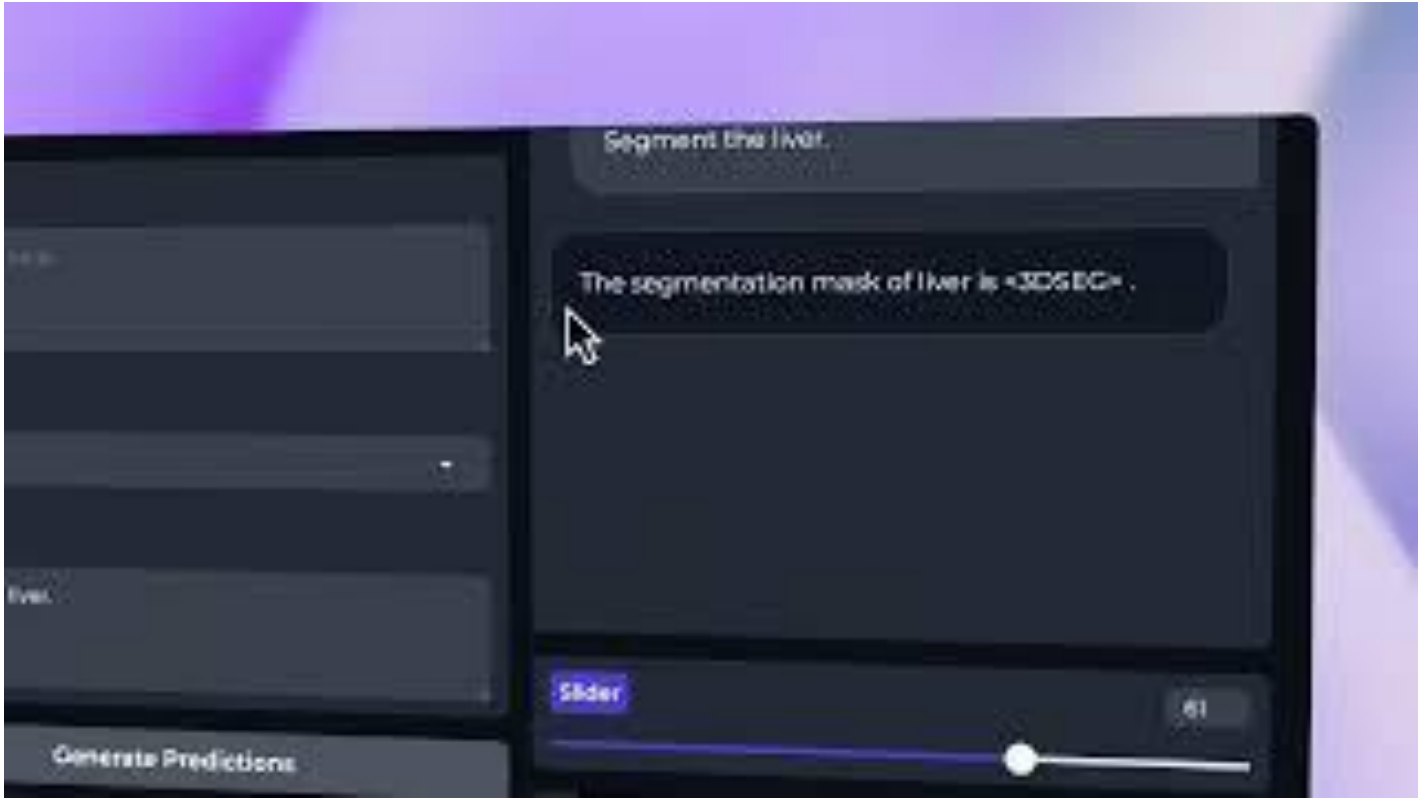
Key capabilities



Zhou, H. Y., Adithan, S., Acosta, J. N., Topol, E. J., & Rajpurkar, P. (2024). A Generalist Learner for Multifaceted Medical Image Interpretation. *arXiv preprint arXiv:2405.07988*.

RAJPURKAR LAB

Demo on various tasks



Zhou, H. Y., Adithan, S., Acosta, J. N., Topol, E. J., & Rajpurkar, P. (2024). A Generalist Learner for Multifaceted Medical Image Interpretation. *arXiv preprint arXiv:2405.07988*.

In 2024, MedVersa is the best benchmarked model on generating reports from chest radiographs.

ReXrank Home

ReXrank

Open-Source Radiology Report Generation Leaderboard

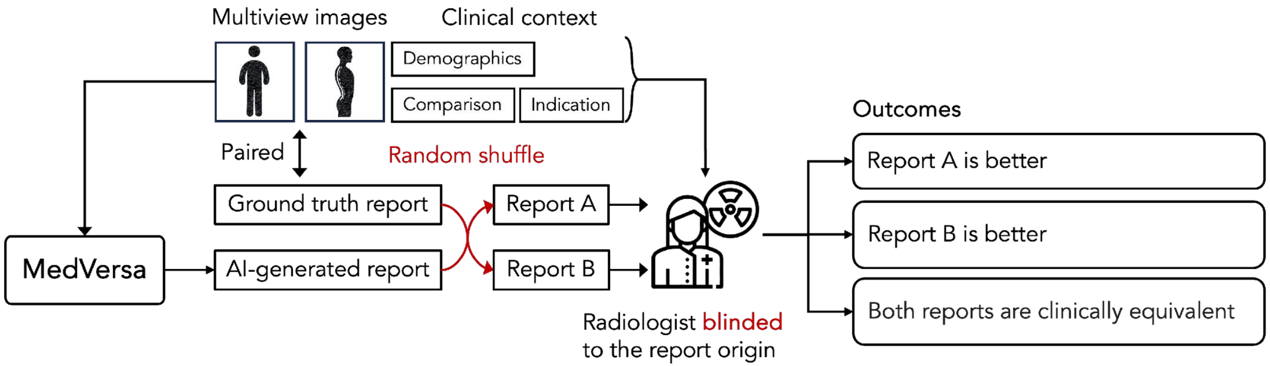
Leaderboard Overview

Include top models for different datasets. * denotes model trained on this dataset.

Rank	MIMIC-CXR	IU-Xray	CheXpert Plus
1	MedVersa* *Harvard*	MedVersa *Harvard*	MedVersa *Harvard*
2	RaDialog* *TUM*	RGRG *TUM*	RaDialog *TUM*
3	RGRG* *TUM*	RadFM *SJTU*	CheXpertPlus-mimic *Stanford*
4	CheXpertPlus-mimic* *Stanford*	Cvt2distilgpt2 *CSIRO*	RGRG *TUM*
5	CheXagent* *Stanford*	RaDialog *TUM*	Cvt2distilgpt2 *CSIRO*
6	Cvt2distilgpt2* *CSIRO*	CheXpertPlus-mimic *Stanford*	CheXagent *Stanford*

RAJPURKAR LAB https://rajpurkarlab.github.io/ReXrank/

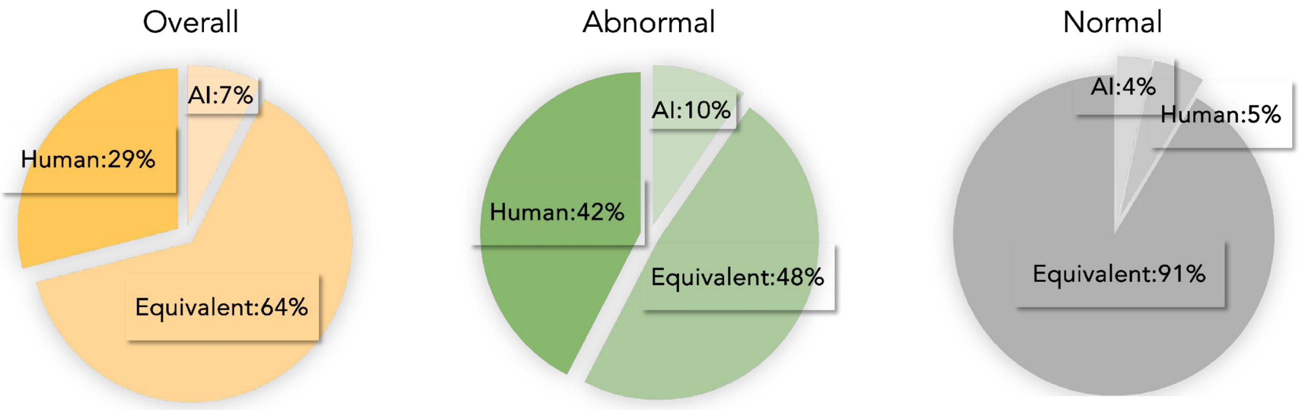
We asked radiologists to determine whether they preferred a human-generated report or an AI generated one (blinded).



RAJPURKAR LAB

Zhou, H. Y., Adithan, S., Acosta, J. N., Topol, E. J., & Rajpurkar, P. (2024). A Generalist Learner for Multifaceted Medical Image Interpretation. *arXiv preprint arXiv:2405.07988*.

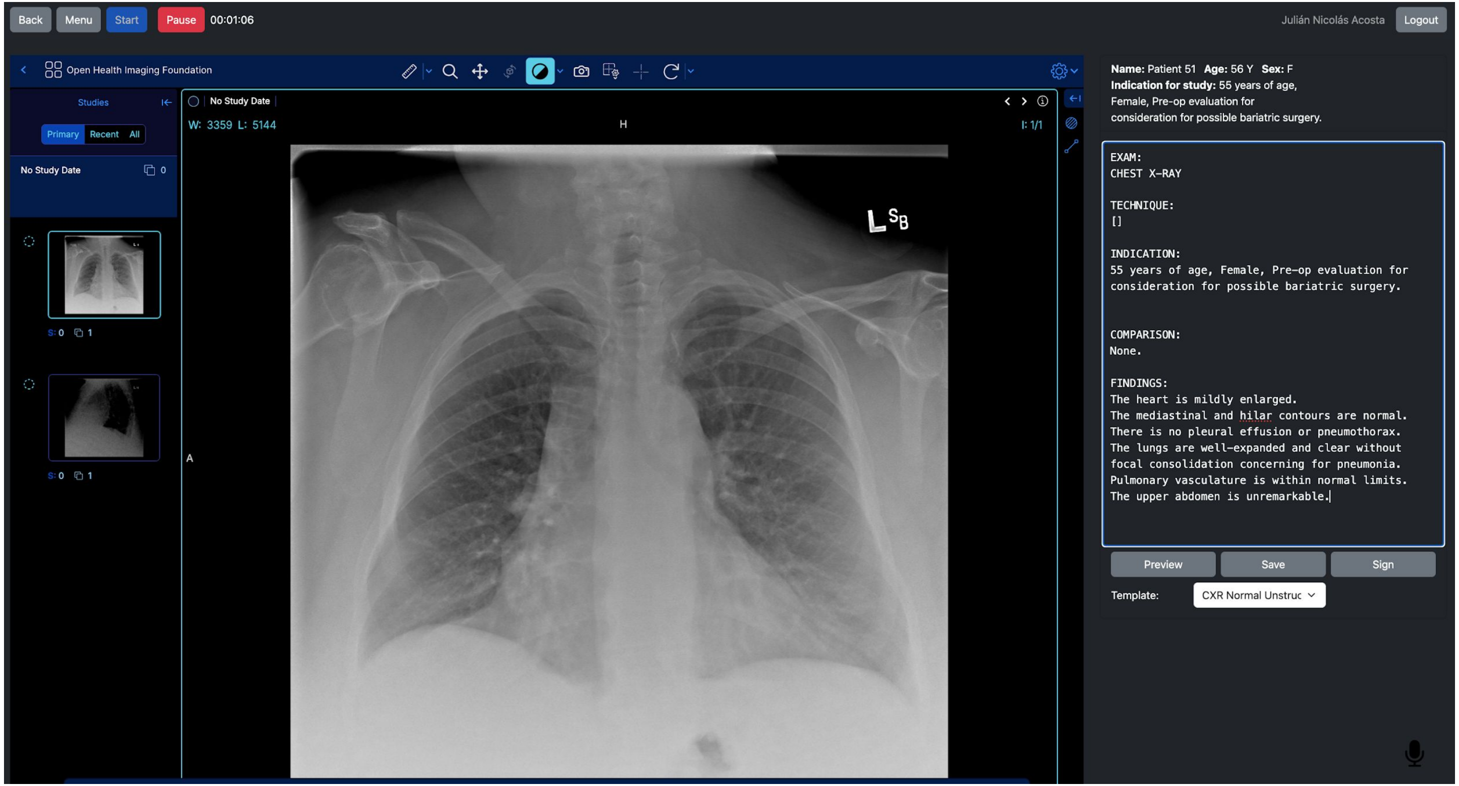
Expert-written reports are preferred, driven by cases with abnormalities



RAJPURKAR LAB

Zhou, H. Y., Adithan, S., Acosta, J. N., Topol, E. J., & Rajpurkar, P. (2024). A Generalist Learner for Multifaceted Medical Image Interpretation. *arXiv preprint arXiv:2405.07988*.

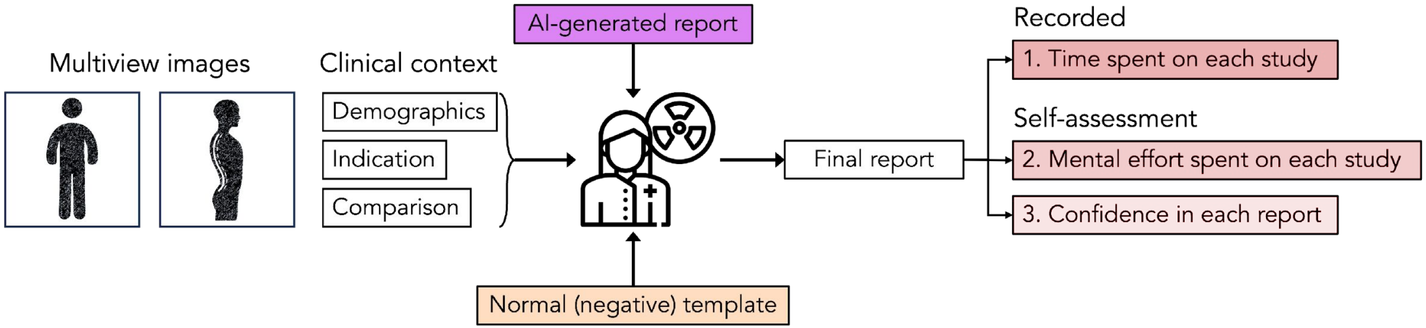
Do experts like to modify draft reports written by AI?



RAJPURKAR LAB

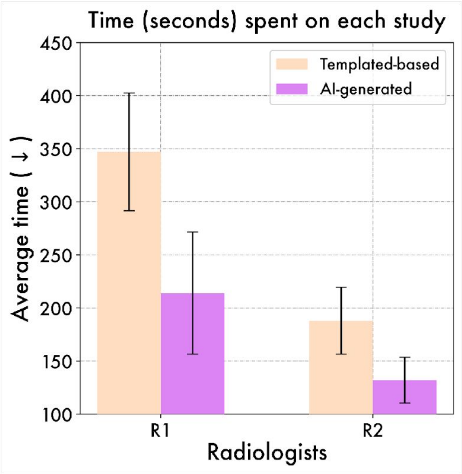
Unreleased work

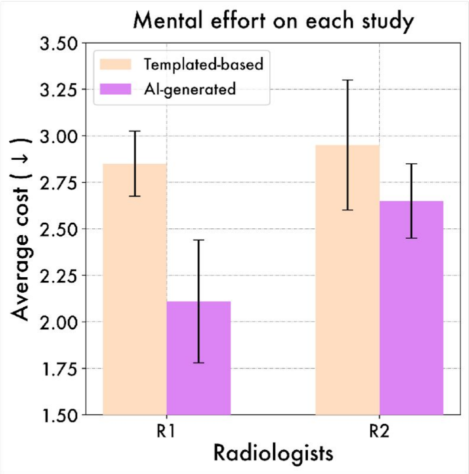
We studied the effect of AI generated draft reports on time, effort and confidence

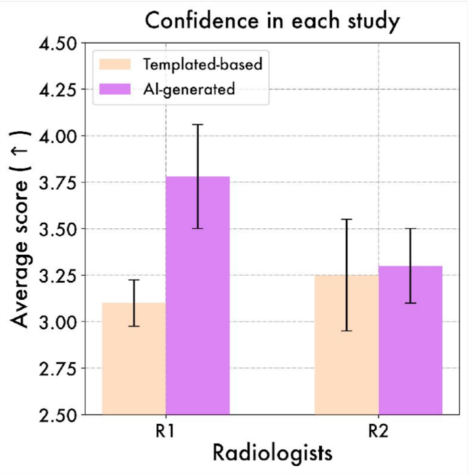


Unreleased work

Our small-scale results show reduction in time and mental effort and increase in confidence.



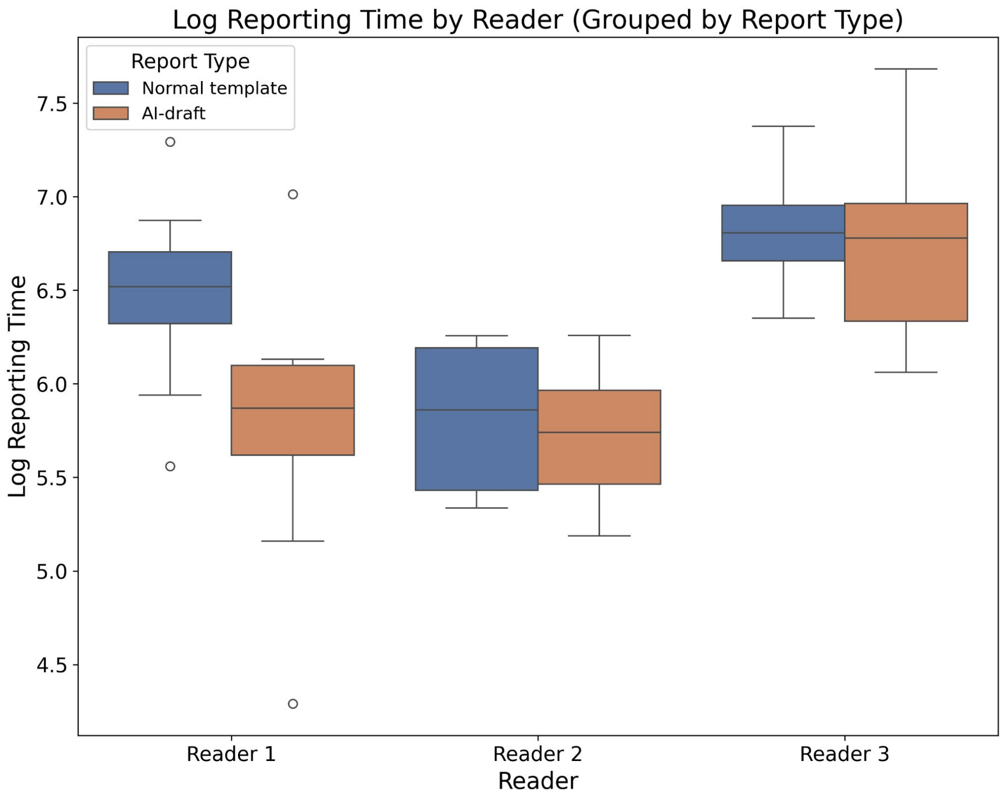




RAJPURKAR LAB

Unreleased work

Again, we see not everyone experiences an improvement with the same technology



RAJPURKAR LAB

Unreleased work

References

RAJPURKAR LAB

LLM Evaluation Based on Medical Exam Questions is Limited

Case Vignette:
A 20-year-old woman presents to the clinic with a *circular hypopigmented lesion* on her right cheek. The patient stated that she used to have a mole in the same location. Over time she noticed a white area around the mole that enlarged to the current size of the lesion. After a few months she noticed the mole in the center of the lesion had disappeared. On further questioning, she denies any personal or family history of skin cancer.

Choices:
A. Halo nevus
B. Melanoma
C. Vitiligo
D. Dysplastic nevus

Concise summary of symptoms:
No evaluation of history-gathering capabilities
No evaluation of ability to diagnose effectively during conversations

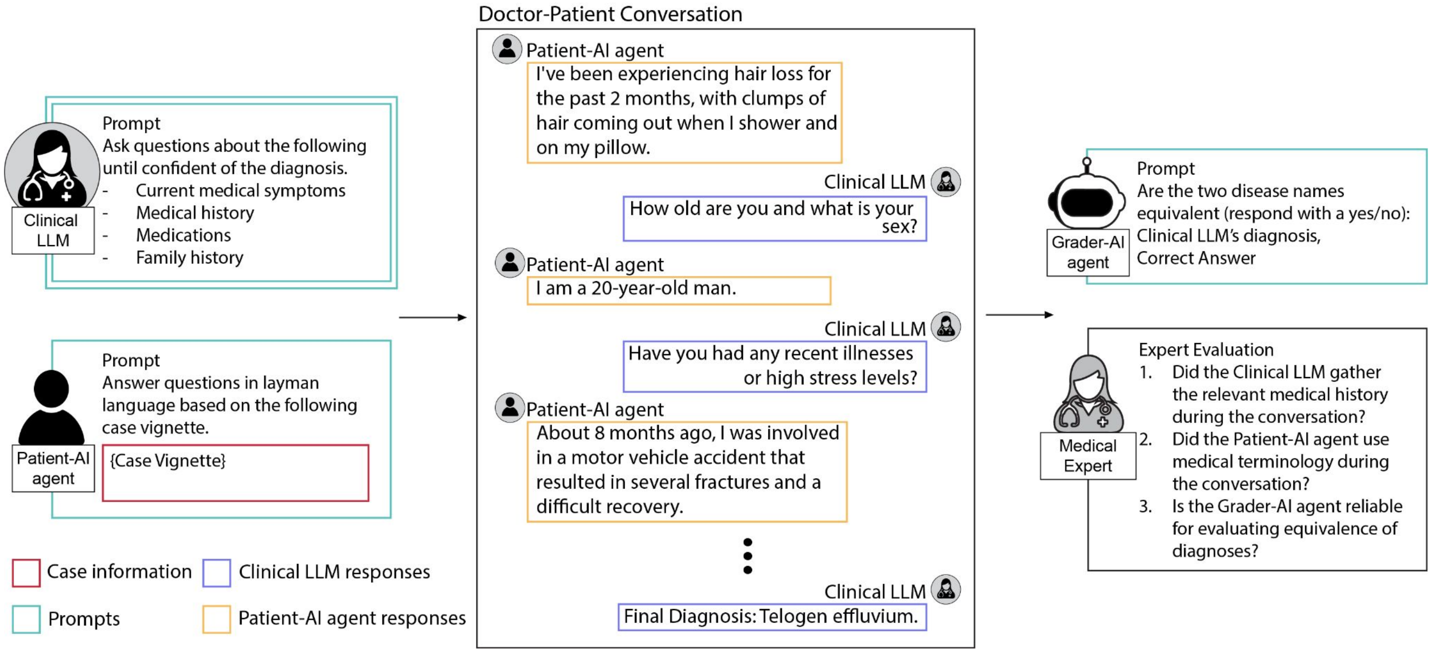
Medical Terminology:
No evaluation of diagnosis from layman language

Answer choices:
No evaluation of open-ended diagnosis

RAJPURKAR LAB

Johri, S., Jeong, J., Tran, B. A., Schlessinger, D. I., Wongvibulsin, S., Cai, Z. R., ... & Rajpurkar, P. CRAFT-MD: A Conversational Evaluation Framework for Comprehensive Assessment of Clinical LLMs. *In AAAI 2024 Spring Symposium on Clinical Foundation Models.*

Multi-Agent Conversational Frameworks Enable Realistic Evaluation of Clinical LLMs

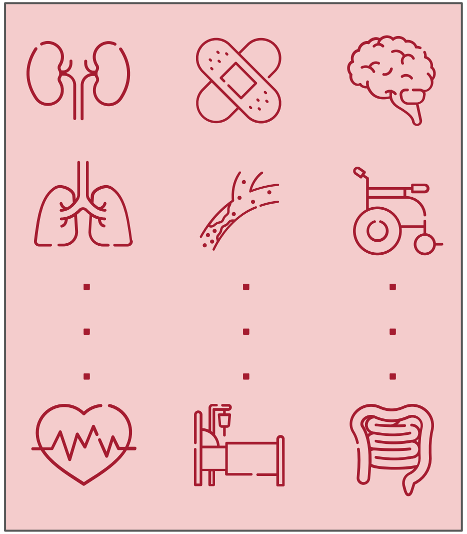


CRAFT-MD: **C**linical **R**easoning **A**ssessment **F**ramework for **T**esting in **M**e**d**icine

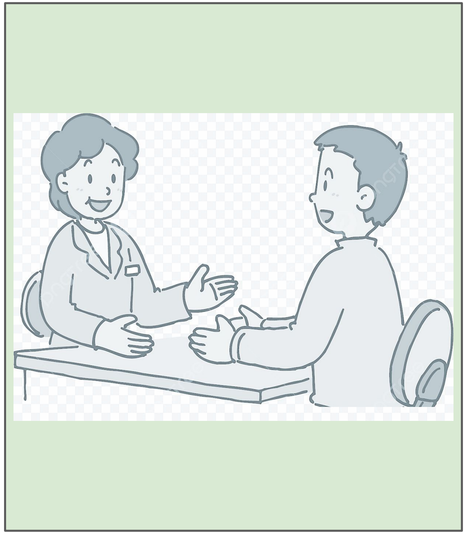
RAJPURKAR LAB

Johri, S., Jeong, J., Tran, B. A., Schlessinger, D. I., Wongvibulsin, S., Cai, Z. R., ... & Rajpurkar, P. CRAFT-MD: A Conversational Evaluation Framework for Comprehensive Assessment of Clinical LLMs. *In AAAI 2024 Spring Symposium on Clinical Foundation Models.*

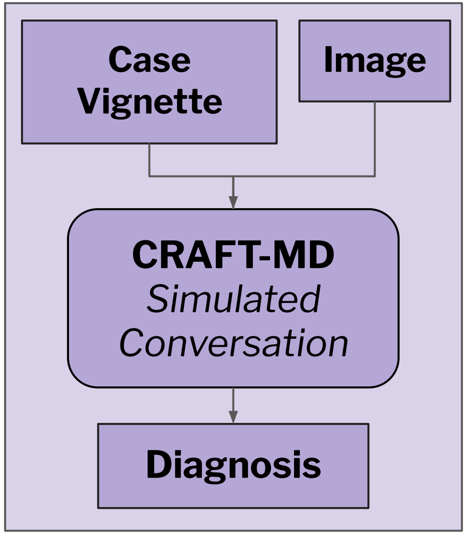
Evaluation of Commercial and Open-Source LLMs using CRAFT-MD



Evaluation across 12 medical specialties.



Expert evaluations in dermatology.



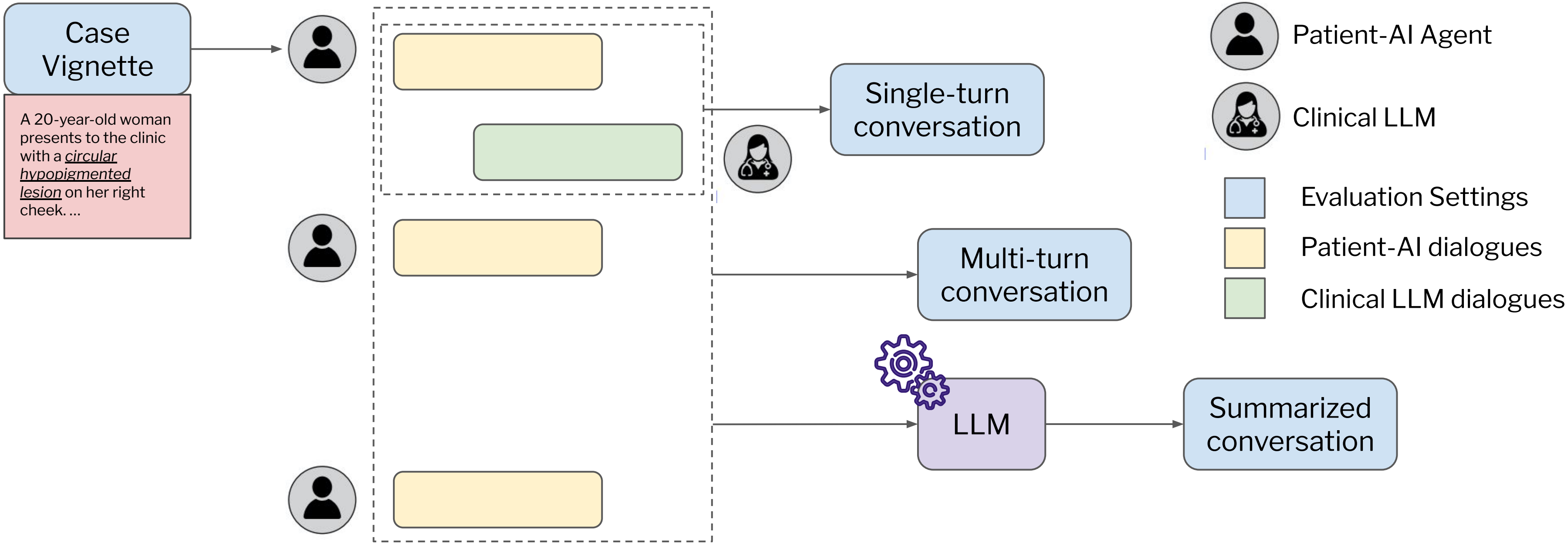
Evaluation of text-only and multimodal LLMs.

Continuous Monitoring of LLMs.

RAJPURKAR LAB

Johri, S., Jeong, J., Tran, B. A., Schlessinger, D. I., Wongvibulsin, S., Cai, Z. R., ... & Rajpurkar, P. CRAFT-MD: A Conversational Evaluation Framework for Comprehensive Assessment of Clinical LLMs. *In AAAI 2024 Spring Symposium on Clinical Foundation Models.*

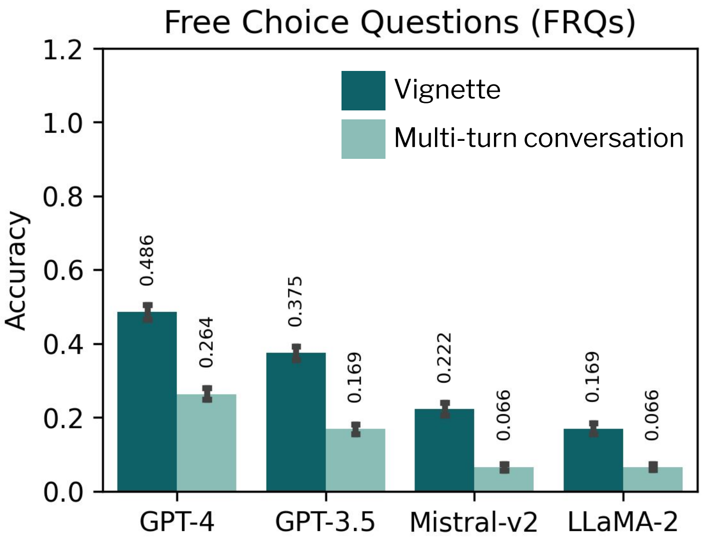
Systematic Evaluation Scenarios in CRAFT-MD

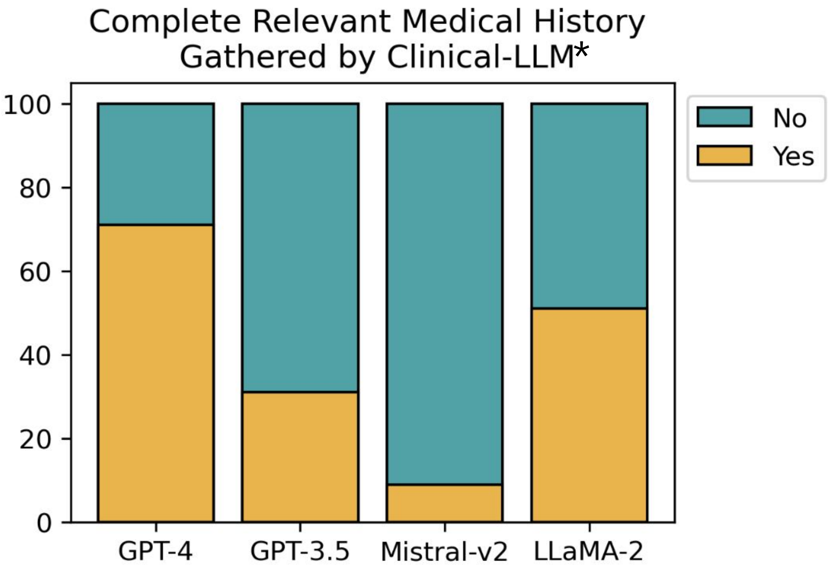


Johri, S., Jeong, J., Tran, B. A., Schlessinger, D. I., Wongvibulsin, S., Cai, Z. R., ... & Rajpurkar, P. CRAFT-MD: A Conversational Evaluation Framework for Comprehensive Assessment of Clinical LLMs. *In AAAI 2024 Spring Symposium on Clinical Foundation Models.*

RAJPURKAR LAB

Current LLMs are Limited in History Gathering and Diagnoses from Long Conversations





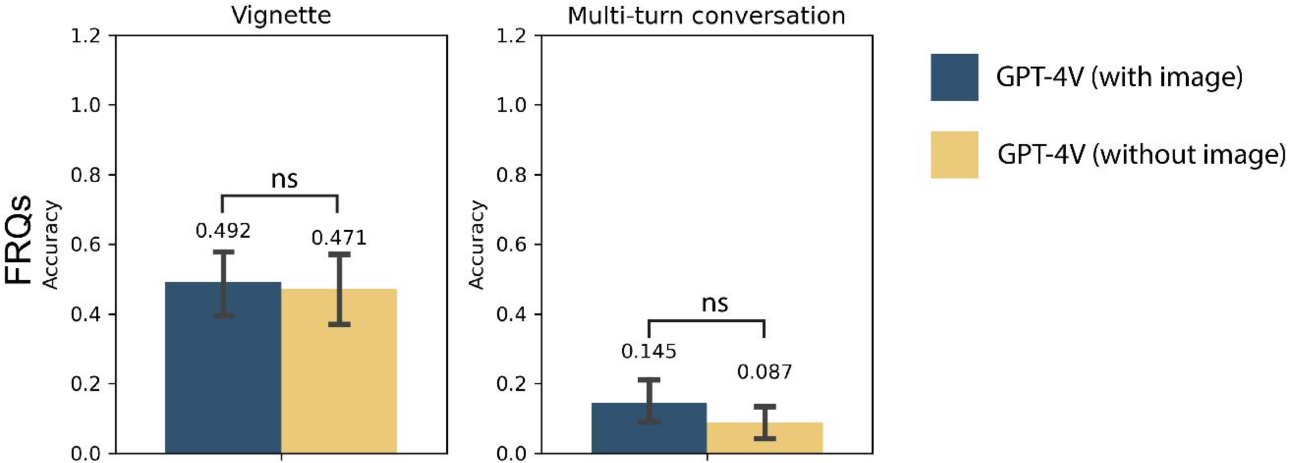
* Based on Medical Expert Annotations

Dataset: MedQA-USMLE + Derm-Public + Derm-Private (2000 case vignettes)

RAJPURKAR LAB

Johri, S., Jeong, J., Tran, B. A., Schlessinger, D. I., Wongvibulsin, S., Cai, Z. R., ... & Rajpurkar, P. CRAFT-MD: A Conversational Evaluation Framework for Comprehensive Assessment of Clinical LLMs. *In AAAI 2024 Spring Symposium on Clinical Foundation Models.*

Multimodal LLMs are Severely Limited in Image Interpretation Capabilities

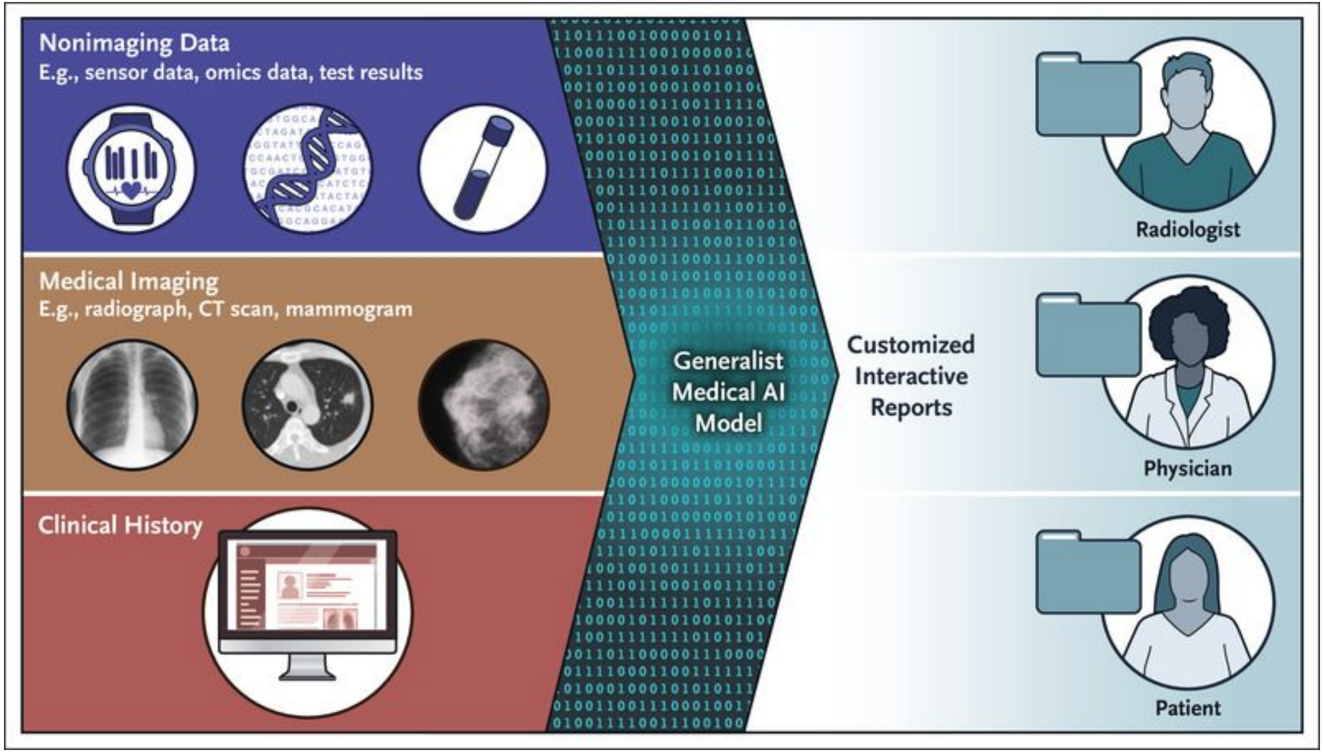


Dataset: NEJM Image Challenge
(May 2021- February 2024)

RAJPURKAR LAB

Johri, S., Jeong, J., Tran, B. A., Schlessinger, D. I., Wongvibulsin, S., Cai, Z. R., ... & Rajpurkar, P. CRAFT-MD: A Conversational Evaluation Framework for Comprehensive Assessment of Clinical LLMs. *In AAAI 2024 Spring Symposium on Clinical Foundation Models.*

Intelligence To Patient Delivery



Rajpurkar, Pranav, and Matthew P. Lungren. "The Current and Future State of AI Interpretation of Medical Images." New England Journal of Medicine 388.21 (2023): 1981-1990.

Can reports be truly catered to the patient and their family members?

Radiology Report

Bilateral adrenal glands were normal and no space-occupying lesion was detected. When examined in the lung parenchyma window....
Osteophytes are also present in the vertebrae...
Thoracic aorta diameter is normal...
Calcific millimetric atheroma plaques are observed in the aortic arch...

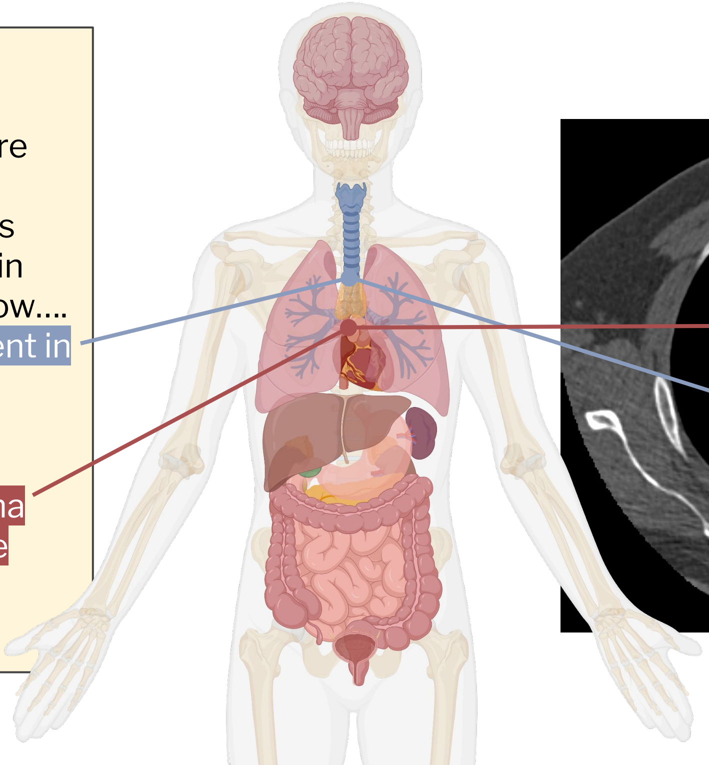
RAJPURKAR LAB

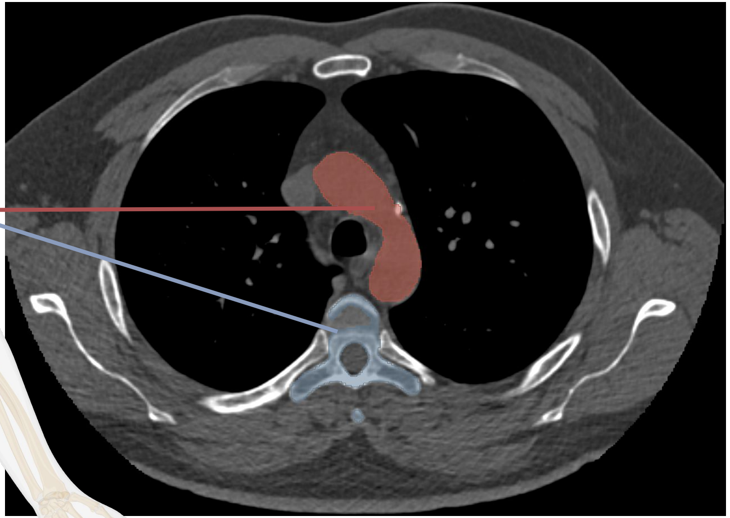
Luo, L., Vairavamurthy, J., Zhang, X., Kumar, A., Ter-Oganesyan, R.R., Schroff, S.T., Shilo, D., Hossain, R., Moritz, M. and Rajpurkar, P., 2024. ReXplain: Translating Radiology into Patient-Friendly Video Reports. *arXiv preprint arXiv:2410.00441*.

Ease the Understanding - Connect Radiology Reports to Image Regions

Radiology Report

Bilateral adrenal glands were normal and no space-occupying lesion was detected. When examined in the lung parenchyma window....
Osteophytes are also present in the vertebrae...
Thoracic aorta diameter is normal...
Calcific millimetric atheroma plaques are observed in the aortic arch...



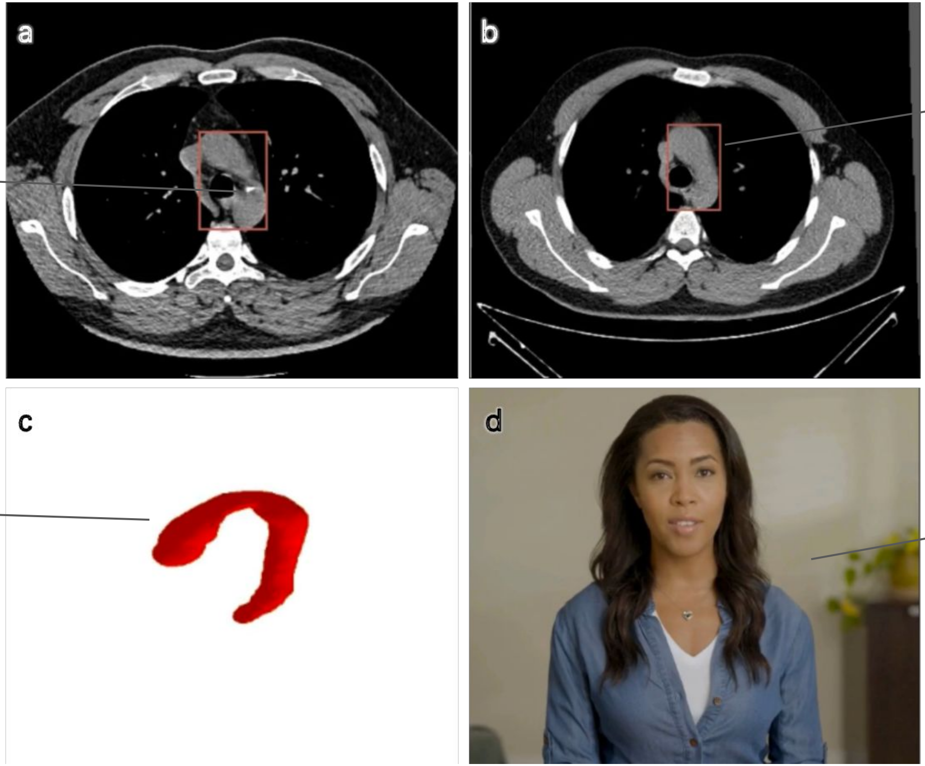


Luo, L., Vairavamurthy, J., Zhang, X., Kumar, A., Ter-Oganesyan, R.R., Schroff, S.T., Shilo, D., Hossain, R., Moritz, M. and Rajpurkar, P., 2024. ReXplain: Translating Radiology into Patient-Friendly Video Reports. *arXiv preprint arXiv:2410.00441*.

Ease the Understanding - Deliver the Information

What was found?
What does it mean?
Where is it?
How it looks like?

How a normal scan looks like?



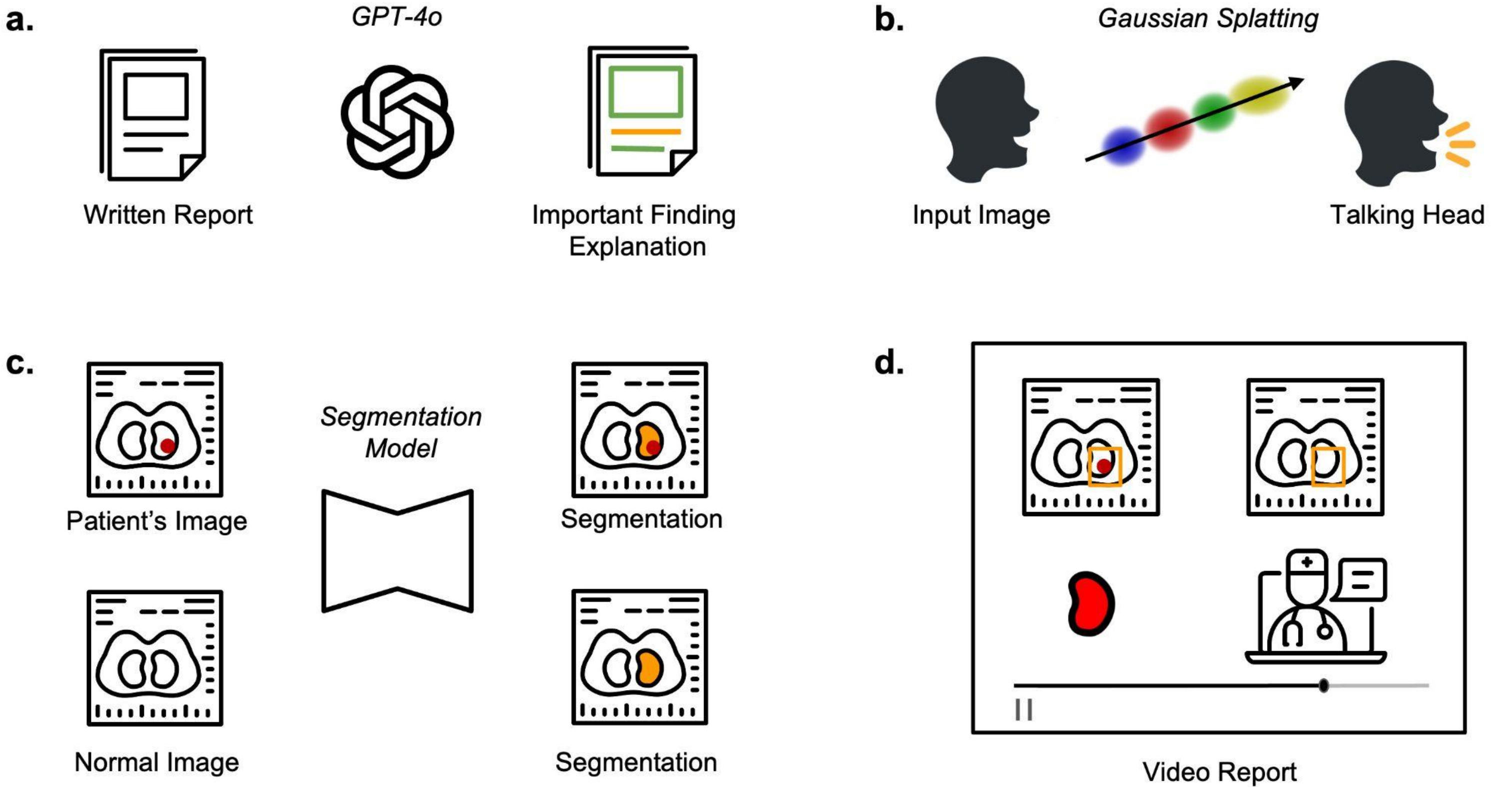
How the overview of the organ looks like?

Let me explain these to you!

RAJPURKAR LAB

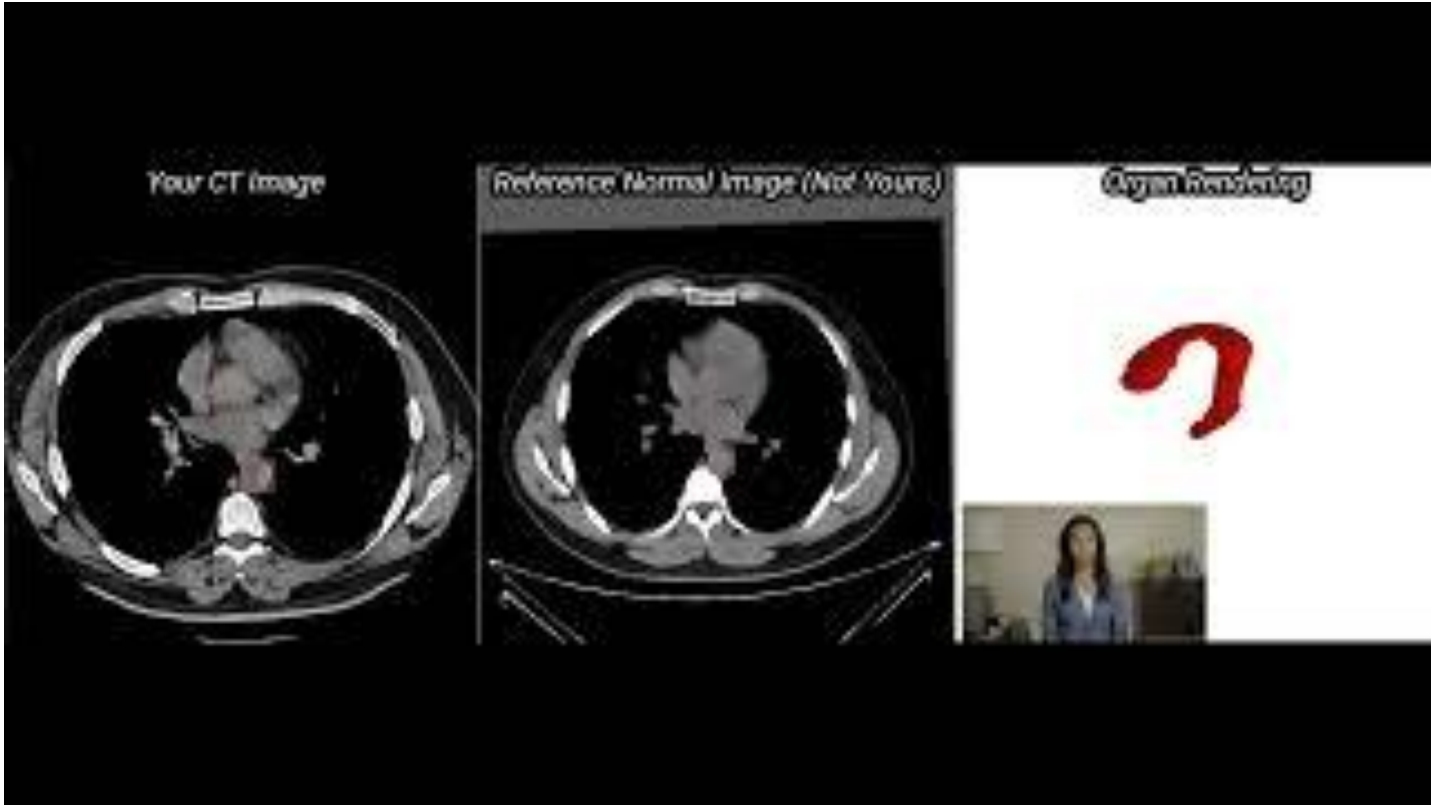
Luo, L., Vairavamurthy, J., Zhang, X., Kumar, A., Ter-Oganesyan, R.R., Schroff, S.T., Shilo, D., Hossain, R., Moritz, M. and Rajpurkar, P., 2024. ReXplain: Translating Radiology into Patient-Friendly Video Reports. *arXiv preprint arXiv:2410.00441*.

End-to-end System Integrating Cutting-edge AIs



Luo, L., Vairavamurthy, J., Zhang, X., Kumar, A., Ter-Oganesyan, R.R., Schroff, S.T., Shilo, D., Hossain, R., Moritz, M. and Rajpurkar, P., 2024. ReXplain: Translating Radiology into Patient-Friendly Video Reports. *arXiv preprint arXiv:2410.00441*.

Patient-centered Radiology Video Report



RAJPURKAR LAB

Luo, L., Vairavamurthy, J., Zhang, X., Kumar, A., Ter-Oganesyan, R.R., Schroff, S.T., Shilo, D., Hossain, R., Moritz, M. and Rajpurkar, P., 2024. ReXplain: Translating Radiology into Patient-Friendly Video Reports. *arXiv preprint arXiv:2410.00441*.

Rajpurkar Lab

Pranav Rajpurkar PhD · Principal Investigator
Agustina Saenz MD MPH · Postgraduate Researcher
Julian Acosta · Research Scientist
Hongyu Zhou · Postdoctoral Researcher
Xiaoman Zhang · Postdoctoral Researcher
Luyang Luo · Postdoctoral Researcher
Emma Chen · PhD Student
Shreya Johri · PhD Student
Oishi Banerjee · PhD Student
Wendy Erselius · Partnership Manager, MAIDA
Heather Viana · Administrative Coordinator

Alumni
Liyue Shen · Now Faculty at UMich
Kathy Yu · Now at Google
Ryan Han · Now MD/PhD at UCLA
Elaine Liu · Now at Meta
Xiaoli Yang · Masters at Stanford
Henrik Marklund · Now PhD at Stanford
Jonathan Williams · Undergrad at Stanford
Yash Mehta· Now PhD Student at JHU
Caiwei Tian · Masters Student
Vignav Ramesh · Undergrad Student
Jaehwan Jeong · Undergrad Student
Martin Ma · Masters Student
Alyssa Huang · Undergrad Student
+ Medical AI Bootcamp Alumni