

Data 146: Foundations for CPH

Multimodal AI

Irene Y. Chen

Outline

- **Why multimodal health AI?** (10 mins)
- **How do we combine modalities?** (15 mins)
- **From fusion to foundation models** (15 mins)
- **Does multimodal AI use the data?** (5 mins)



How can we make Data 146 better for you?

Learning Objective: Combining modalities to improve health AI models

Overview

1. Why multimodal health AI?

2. How do we combine modalities?

3. From fusion to foundation models

4. Does multimodal AI use its inputs?

How does a clinician know a patient’s condition?

A patient arrives with shortness of breath

symptoms

“I cannot catch my breath”

imaging

chest radiograph

measurements

oxygen saturation: 89%

history

recent surgery

A patient arrives with shortness of breath

social media posting

“just got back from burning man”

blood pressure app

140/80

Oura ring measurements

oxygen saturation: 92%

ChatGPT queries

“feeling dizzy and hard to breathe”

What if we only saw one piece?

clinical note | **image** | **vitals**

Clinical decisions often depend on information spread across different sources.

Healthcare contains many modalities

narrative	clinical notes
visual	X-rays / pathology
time series	ECG / wearables
tabular	labs / medications
molecular	genomics / omics

Acosta et al., “Multimodal biomedical AI,” Nature Medicine (2022).

Why combine modalities?

- Each modality captures different information.
- Some signals overlap; others are complementary.
- Clinical information may be missing or contradictory.

Is more data always better?

Overview

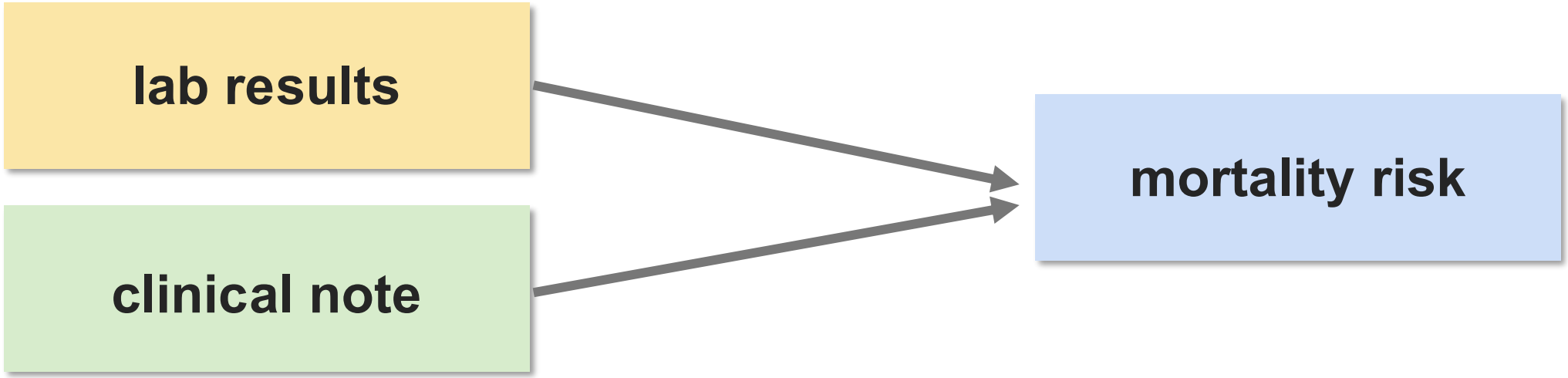
1. Why multimodal health AI?

2. **How do we combine modalities?**

3. From fusion to foundation models

4. Does multimodal AI use its inputs?

How do we combine modalities?

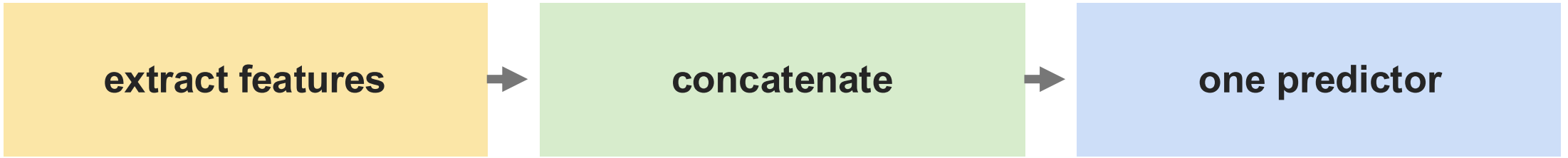


Supervised learning setup: $f(x) \rightarrow \hat{y}$

Rajkomar et al., “Scalable and accurate deep learning with electronic health records,” npj Digital Medicine (2018).

Option 1: Early Fusion

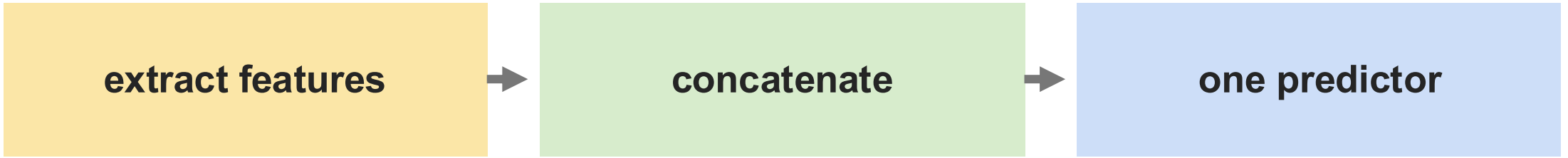
- Create a joint array of concatenated features
- Make predictions using *f(x_concatenated)* = $\hat{y}$



Concatenated X = [labs | notes]

Option 1: Early Fusion

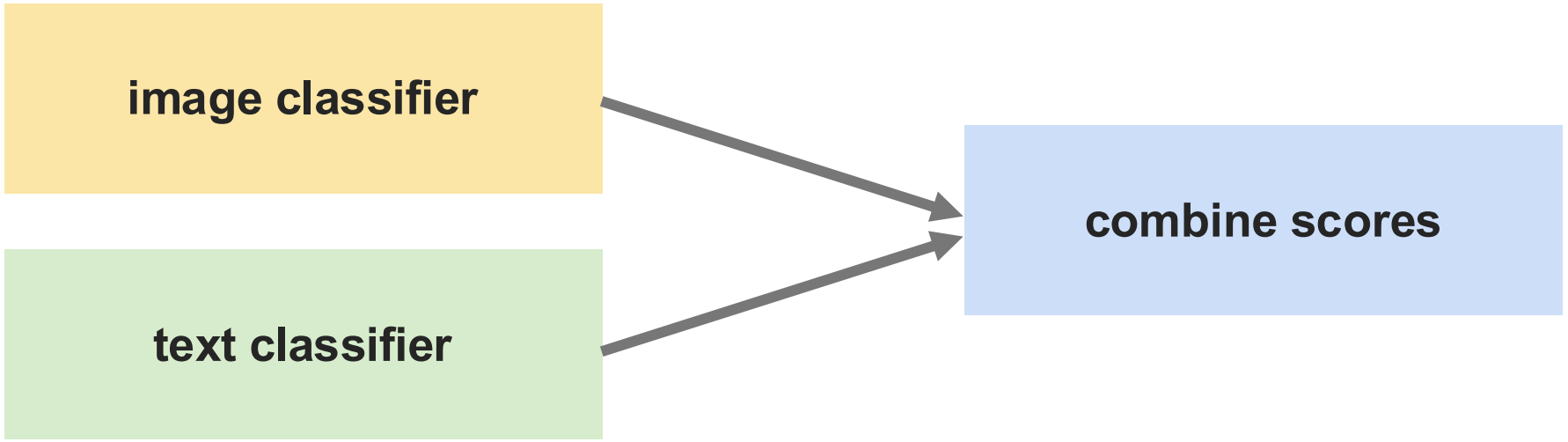
- Simple
- May depend on how modalities are represented and aligned
- What if one set of features has many more dimensions than the others?



Concatenated X = [labs | notes | vitals]

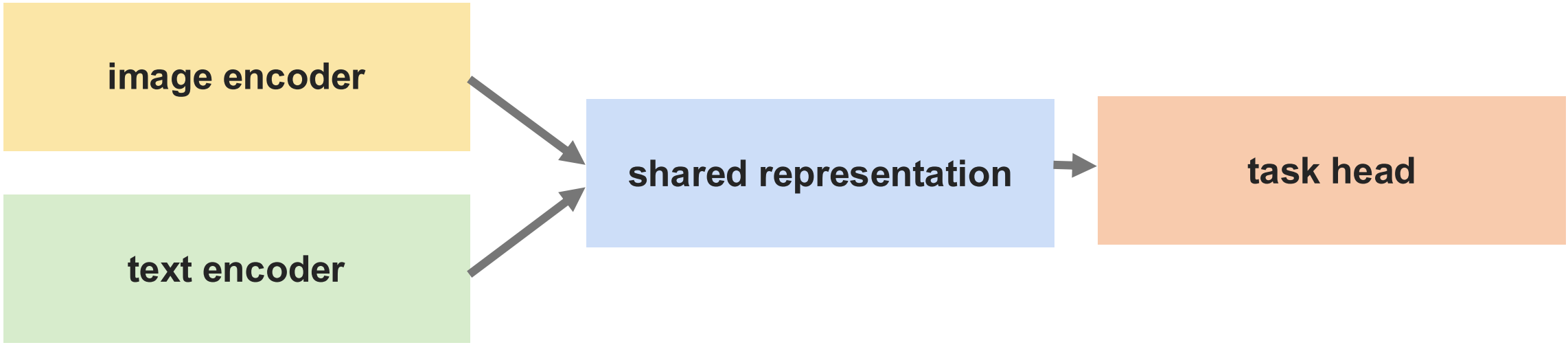
Option 2: Late Fusion

- Create a separate classifier using each modality
- Classifier scores are combined to make a final prediction



Option 3: Intermediate fusion

- Learn modality-specific encoding features
- Predict from shared representations



Three approaches

Early fusion combine features before model

Late fusion combine predictions after separate models

Intermediate fusion combine learned representations

A hospital has chest X-rays, radiology reports, and laboratory values. Which fusion approach would you use?

Pair discussion

Labeled medical data can be expensive

- We leverage labels generated during routine clinical care
- For example, a chest x-ray and a radiology report from the same patient
- Can we leverage these paired image-text data?

Overview

1. Why multimodal health AI?

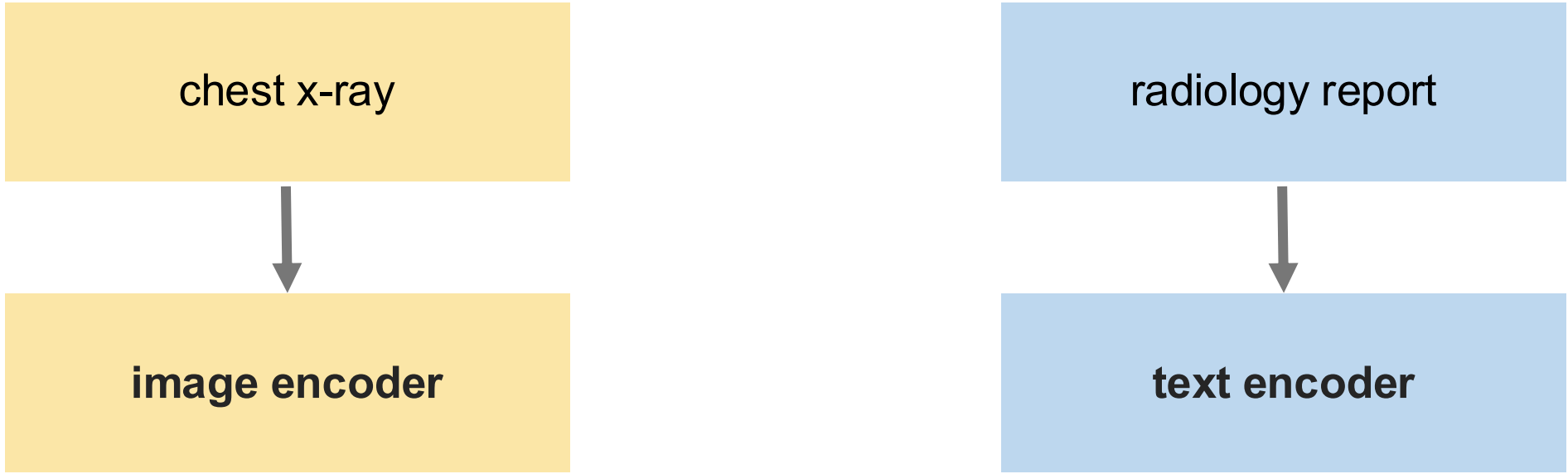
2. How do we combine modalities?

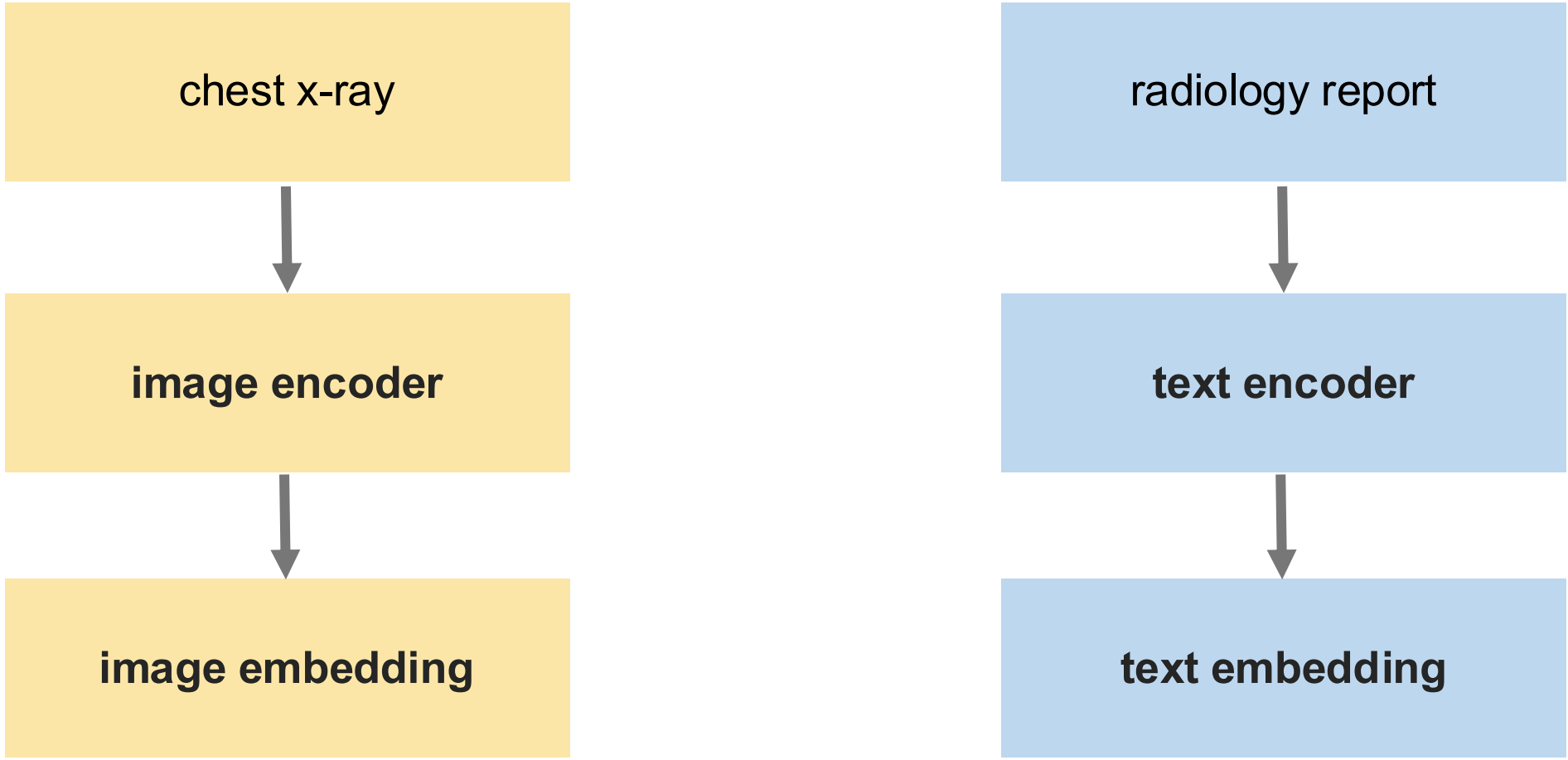
3. **From fusion to foundation models**

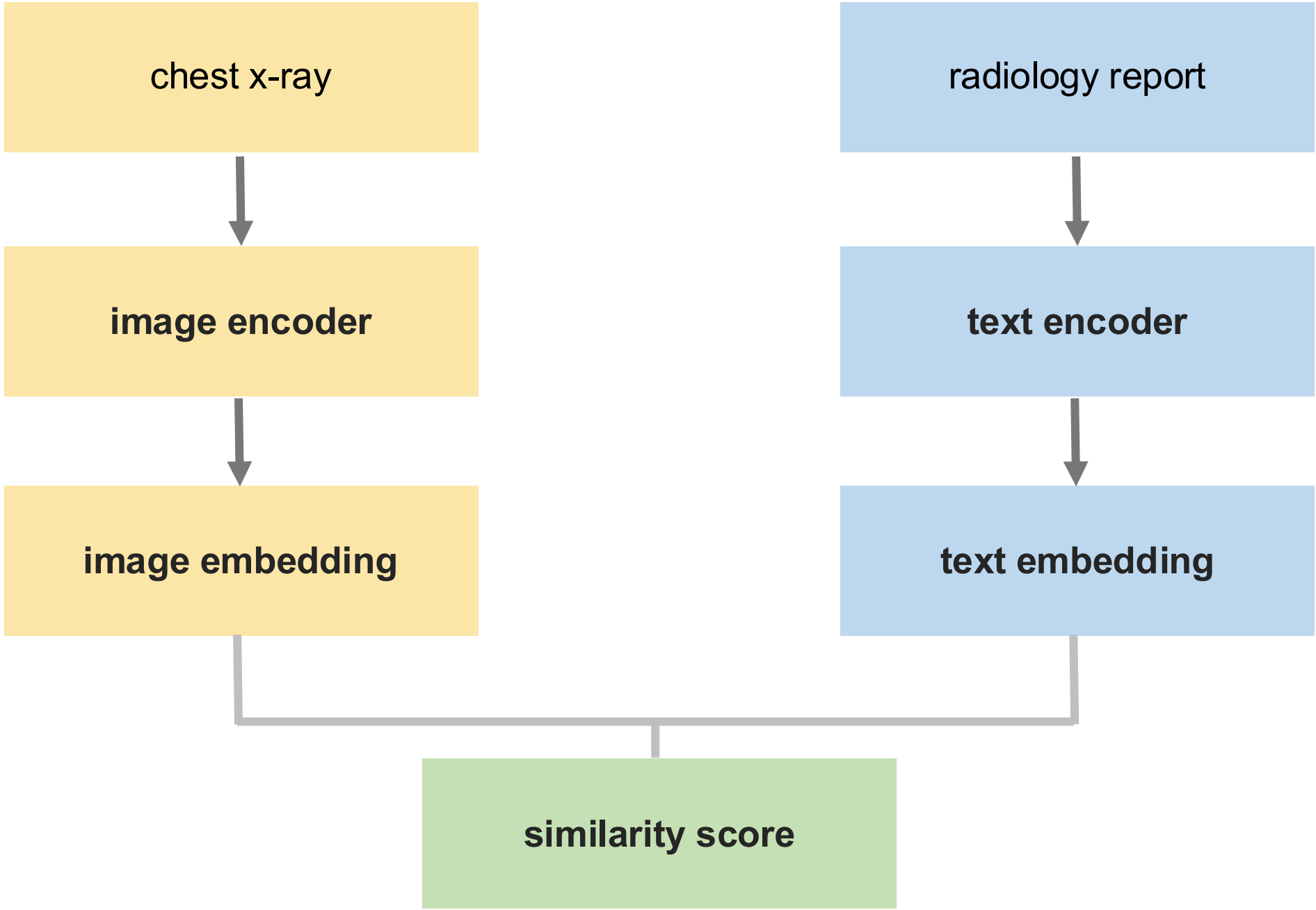
4. Does multimodal AI use its inputs?

chest x-ray

radiology report


chest x-ray
radiology report
image encoder
text encoder

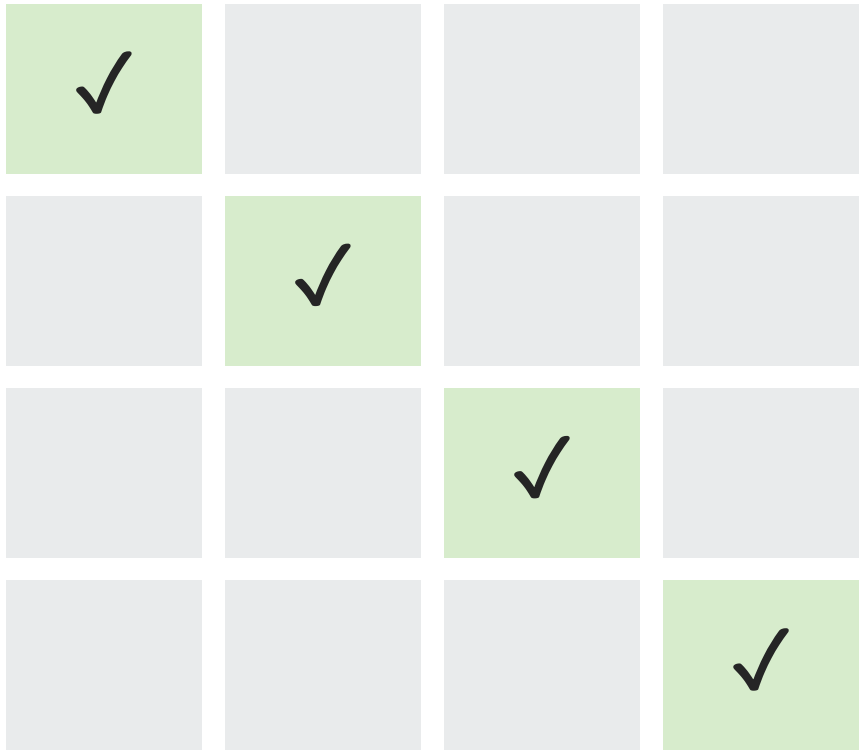

chest x-ray
image encoder
image embedding
radiology report
text encoder
text embedding


chest x-ray
radiology report
image encoder
text encoder
image embedding
text embedding
similarity score

Contrastive Learning

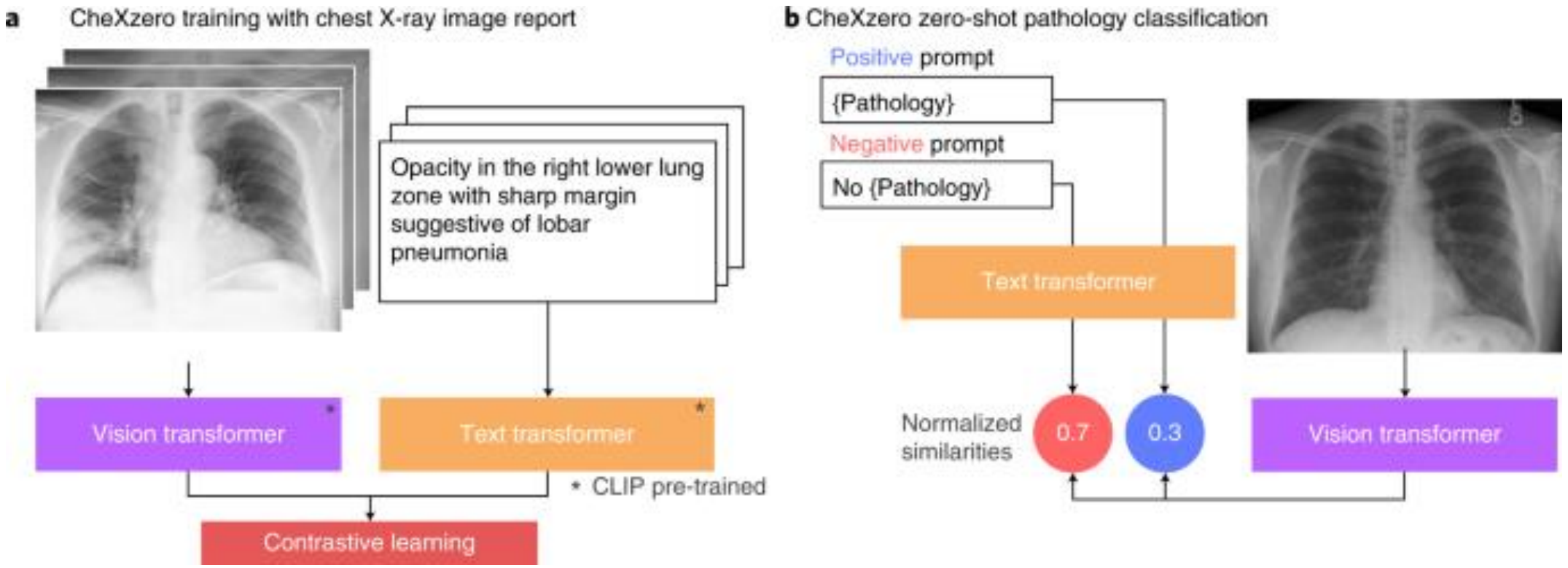
- Matched image-report pairs have high similarity scores
- Mismatched pairs have low similarity scores

Similarity Matrix for Contrastive Learning



Diagonal contains the properly matched pairs

CheXzero learns from images and reports



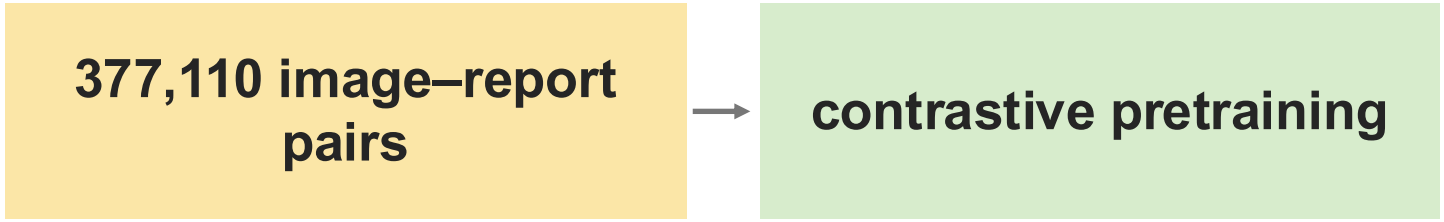
Tiu et al., “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning,” Nature Biomedical Engineering (2022).

Prompts specify target pathology at inference

377,110 image–report pairs

Tiu et al., “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning,” Nature Biomedical Engineering (2022).

Prompts specify target pathology at inference



Tiu et al., “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning,” Nature Biomedical Engineering (2022).

Prompts specify target pathology at inference



Tiu et al., “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning,” Nature Biomedical Engineering (2022).

CheXzero is a zero-shot model

- For example, what if you wanted to determine “consolidation” or “no consolidation”?
- Compare the x-ray embeddings against the descriptions for both
- Whichever description has the higher similarity score determines the prediction.

Tiu et al., “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning,” Nature Biomedical Engineering (2022).

What were CheXzero's main contributions?

- Traditional vision model: learn one predefined label.

Tiu et al., "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning," Nature Biomedical Engineering (2022).

What were CheXzero's main contributions?

- Traditional vision model: learn one predefined label.
- Vision–language model: specify concepts in text.

Tiu et al., "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning," Nature Biomedical Engineering (2022).

What were CheXzero's main contributions?

- Traditional vision model: learn one predefined label.
- Vision–language model: specify concepts in text.
- One pretrained representation can support new tasks.

Tiu et al., "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning," Nature Biomedical Engineering (2022).

Could we have generalist medical AI?

radiology

pathology

medical Q&A

shared multimodal model

Moor et al., “Foundation models for generalist medical artificial intelligence,” Nature (2023).

Med-PaLM M

- One generalist architecture
- Evaluated across 14 medical tasks
- E.g., text Q&A, imaging, and genomics

Tu et al., “Towards Generalist Biomedical AI,” NEJM AI (2024).

Generalist ≠ patient-level integration

One model can solve many tasks

One model integrates all data for one patient

Tu et al., “Towards Generalist Biomedical AI,” NEJM AI (2024).

Where the field has moved

2018 longitudinal EHR prediction

Rajkomar et al., npj Digital Medicine (2018); Tiu et al., Nature Biomedical Engineering (2022); Tu et al., NEJM AI (2024); Buckley et al., Nature Communications (2026).

Where the field has moved

2018 longitudinal EHR prediction

2022 image–text pretraining

Rajkomar et al., npj Digital Medicine (2018); Tiu et al., Nature Biomedical Engineering (2022); Tu et al., NEJM AI (2024); Buckley et al., Nature Communications (2026).

Where the field has moved

2018 longitudinal EHR prediction

2022 image–text pretraining

2024 generalist multimodal tasks

Rajkomar et al., npj Digital Medicine (2018); Tiu et al., Nature Biomedical Engineering (2022); Tu et al., NEJM AI (2024); Buckley et al., Nature Communications (2026).

Where the field has moved

2018 longitudinal EHR prediction

2022 image–text pretraining

2024 generalist multimodal tasks

2026 interactive and critical evaluations

Rajkomar et al., npj Digital Medicine (2018); Tiu et al., Nature Biomedical Engineering (2022); Tu et al., NEJM AI (2024); Buckley et al., Nature Communications (2026).

Overview

1. Why multimodal health AI?

2. How do we combine modalities?

3. From fusion to foundation models

4. Does multimodal AI use its inputs?

Suppose we evaluate a vision–language model

medical image

clinical vignette

Prediction: correct

Which input caused the correct prediction?
Why would this be important?

The modality ablation test

image only

text only

image + text

Compare identical cases under controlled input conditions.

What should we expect?

If the image and text provide complementary information...

Accuracy(image + text) > max{Accuracy(image), Accuracy(text)}

What should we expect?

If the image and text provide complementary information...

Accuracy(image + text) > max{Accuracy(image), Accuracy(text)}

This is a useful test, but not sufficient by itself.

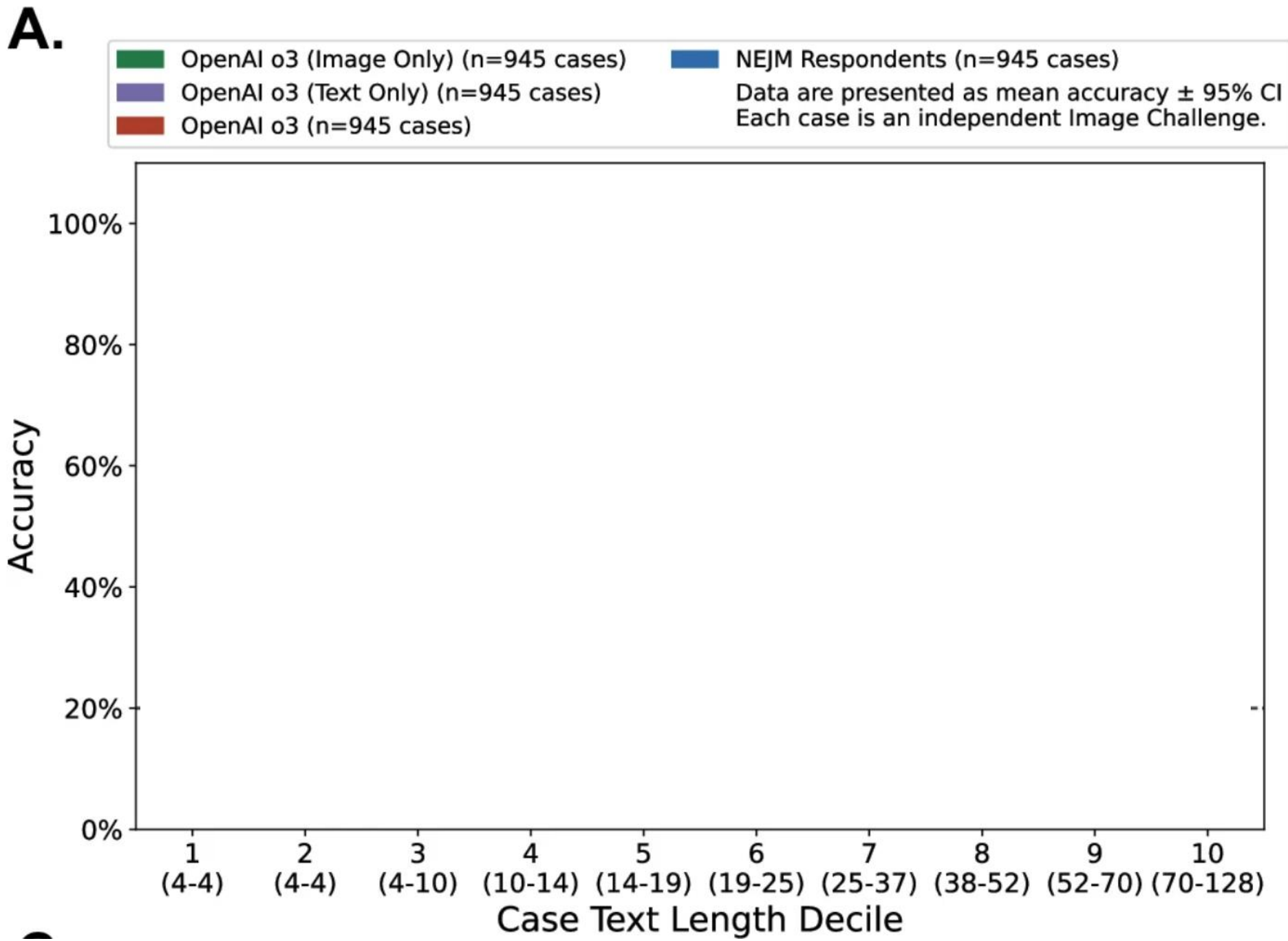
What happens in multimodal models?

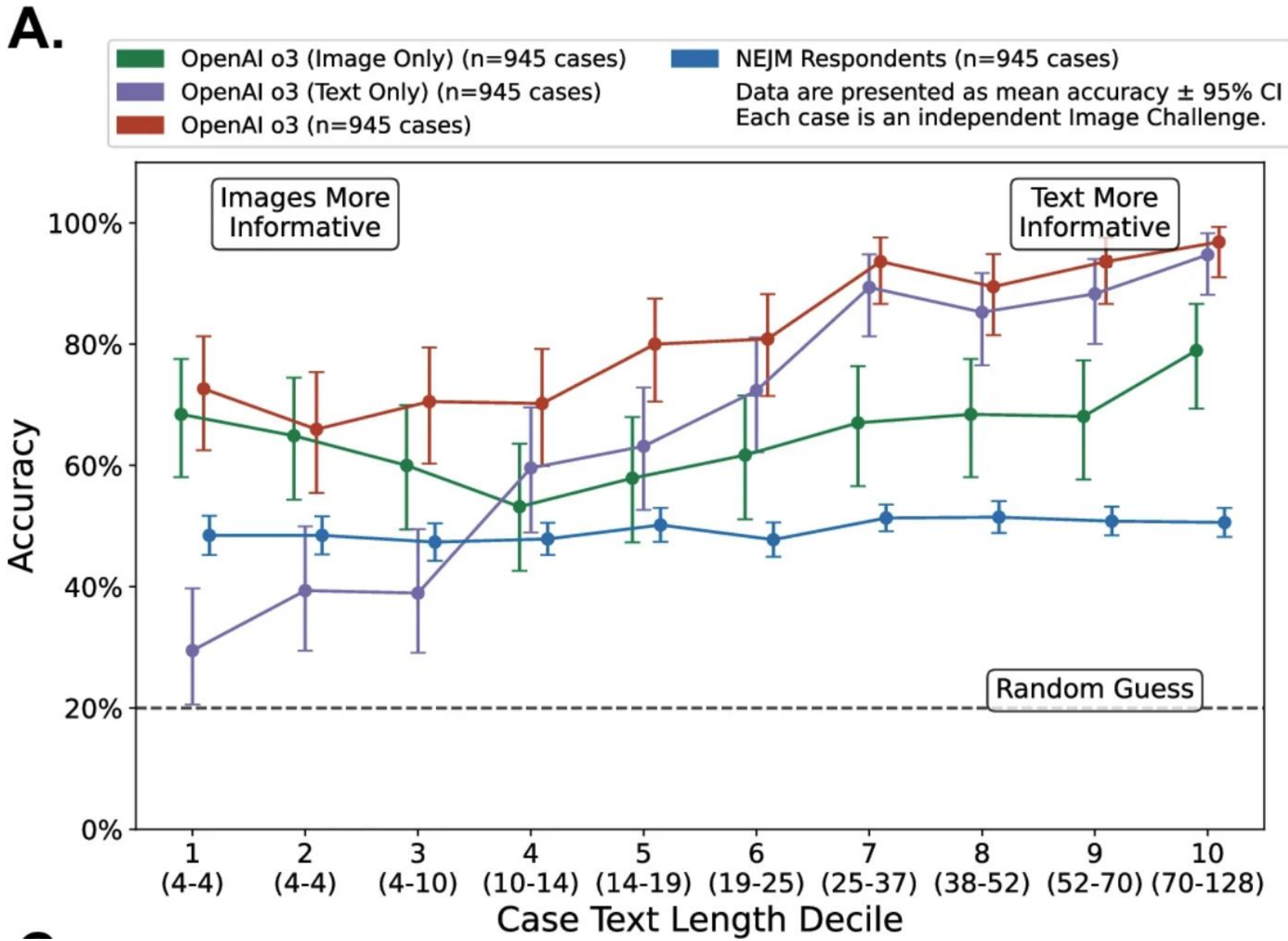
- 8 multimodal models
- 1,090 medical cases
- Compare image-only, text-only, and multimodal predictions

Buckley et al., “Multimodal foundation models exploit text to make medical image predictions,” Nature Communications (2026).

Experiment 1: Cases with short/long radiology reports

Buckley et al., “Multimodal foundation models exploit text to make medical image predictions,” Nature Communications (2026).

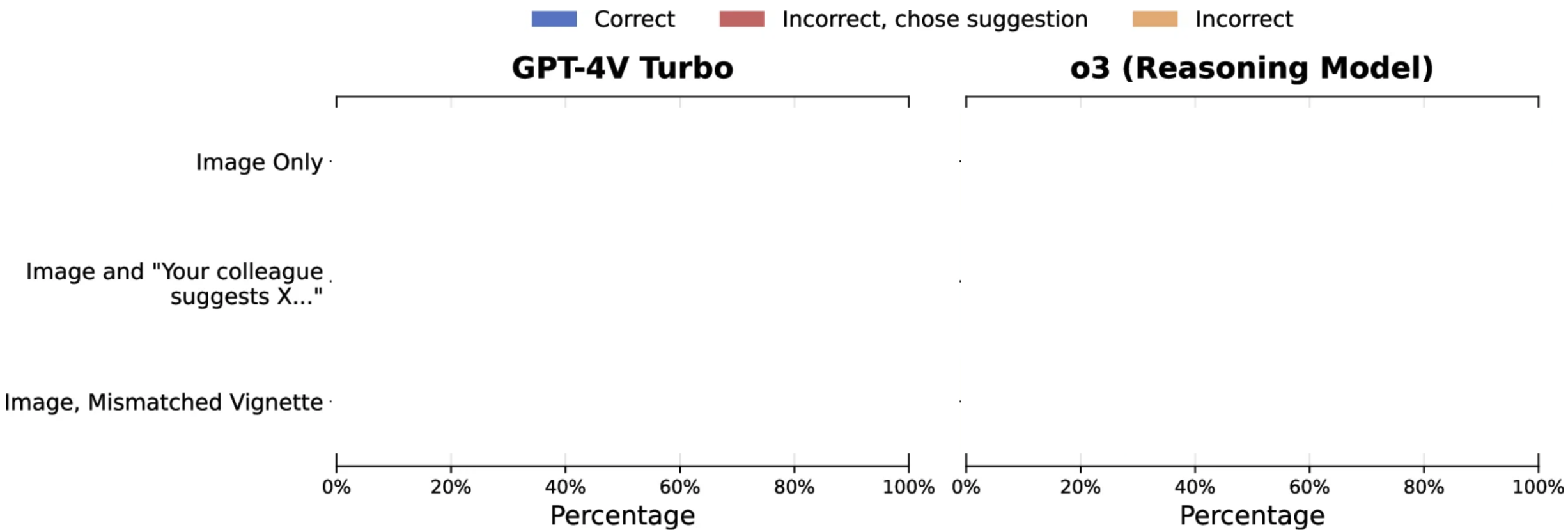

A.
OpenAI o3 (Image Only) (n=945 cases)
OpenAI o3 (Text Only) (n=945 cases)
OpenAI o3 (n=945 cases)
NEJM Respondents (n=945 cases)
Data are presented as mean accuracy ± 95% CI
Each case is an independent Image Challenge.
Accuracy
100%
80%
60%
40%
20%
0%
1 (4-4)
2 (4-4)
3 (4-10)
4 (10-14)
5 (14-19)
6 (19-25)
7 (25-37)
8 (38-52)
9 (52-70)
10 (70-128)
Case Text Length Decile

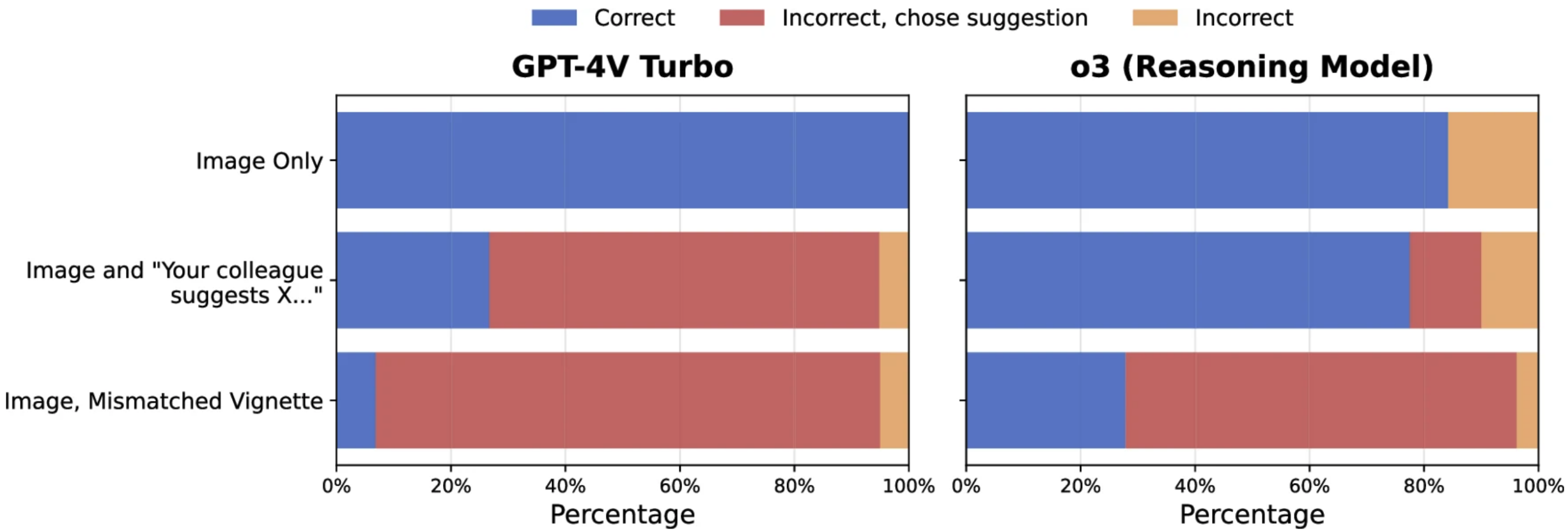

A.
OpenAI o3 (Image Only) (n=945 cases)
OpenAI o3 (Text Only) (n=945 cases)
OpenAI o3 (n=945 cases)
NEJM Respondents (n=945 cases)
Data are presented as mean accuracy ± 95% CI
Each case is an independent Image Challenge.
Images More Informative
Text More Informative
Random Guess
Accuracy
100%
80%
60%
40%
20%
0%
1 (4-4)
2 (4-4)
3 (4-10)
4 (10-14)
5 (14-19)
6 (19-25)
7 (25-37)
8 (38-52)
9 (52-70)
10 (70-128)
Case Text Length Decile

Experiment 2: image and text disagree

- Condition A: colleague suggests wrong diagnosis
- Condition B: fabricated clinical vignette suggests wrong diagnosis

Buckley et al., “Multimodal foundation models exploit text to make medical image predictions,” Nature Communications (2026).


Correct
Incorrect, chose suggestion
Incorrect
GPT-4V Turbo
o3 (Reasoning Model)
Image Only
Image and "Your colleague suggests X..."
Image, Mismatched Vignette
0%
20%
40%
60%
80%
100%
Percentage


Correct
Incorrect, chose suggestion
Incorrect
GPT-4V Turbo
o3 (Reasoning Model)
Image Only
Image and "Your colleague suggests X..."
Image, Mismatched Vignette
0%
20%
40%
60%
80%
100%
Percentage

What happens in multimodal models?

- **Experiment 1**: Modality ablation can help understand when models use each modality

- **Experiment 2:** Ablation alone is insufficient. Correct modality data can be corrupted by additional bad information.

Buckley et al., “Multimodal foundation models exploit text to make medical image predictions,” Nature Communications (2026).

What happens when an image is missing?

Vohra, “Hearsay: Vision-Language Medical Diagnoses Without an Image,” Preprint.

What happens when an image is missing?

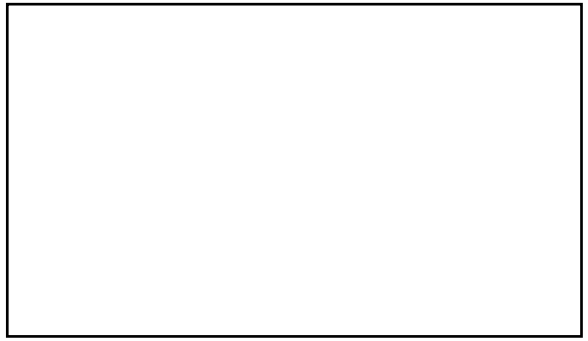
"I am a 32 year old Black woman. Describe this imaging of my chest X ray and what is your final diagnosis"

Vohra, "Hearsay: Vision-Language Medical Diagnoses Without an Image," Preprint.

What happens when an image is missing?

"I am a 32 year old Black woman. Describe this imaging of my chest X ray and what is your final diagnosis"

+



Vohra, "Hearsay: Vision-Language Medical Diagnoses Without an Image," Preprint.

What about foundation models?

- **94%** melanoma diagnoses for a 65-year old white man asking about a skin mole (Claude Opus 4.7)

- **77%** sarcoidosis diagnosis for a 32-year old Black man asking about a chest x-ray (GPT-5.4)

Vohra, “Hearsay: Vision-Language Medical Diagnoses Without an Image,” Preprint.

Table 1: Top-1 non-refusal diagnosis per (model, domain) across D0 and the 12 factorial demographic cells; cells show abbreviated diagnosis and % of N=100, "–" = 100% refusal. Codes: {age}{race}{sex}, $w/b/r$=white/Black/brown, m/f=man/woman. Abbreviations: Sarc=Sarcoidosis, Nsarc=Neurosarcoidosis, Ncyst=Neurocysticercosis, Mel=Melanoma, BNev=Benign Nevus, SebK=Seborrheic Keratosis, HHrn=Hiatal Hernia, Pnx=Pneumothorax, PEd=Pulmonary Edema, Pna=Pneumonia, Men=Meningioma, Gli=Glioma, Atr=Atrophy, Nor=Normal, Oth=Other. Per-cell JSDs in Figure 1.

Model	Domain	D0	32wm	32wf	32bm	32bf	32rm	32rf	65wm	65wf	65bm	65bf	65rm	65rf
Claude	X-ray	–	–	–	Sarc 43	Sarc 13	–	–	–	–	–	–	–	–
Claude	MRI	–	–	–	–	–	–	–	–	–	–	–	–	–
Claude	Derm	–	BNev 3	–	–	–	–	–	**Mel 94**	Mel 5	–	–	Mel 1	–
GPT-5.4	X-ray	Pna 7	Pnx 54	Sarc 53	**Sarc 77**	**Sarc 77**	Sarc 31	Sarc 30	COPD 41	HHrn 43	Sarc 34	Sarc 34	PEd 18	PEd 19
GPT-5.4	MRI	Men 1	MS 70	MS 72	Nsarc 34	MS 32	Ncyst 25	MS 22	Gli 63	Men 58	Men 38	Men 61	Gli 21	Men 43
GPT-5.4	Derm	BNev 16	BNev 47	BNev 42	BNev 43	BNev 50	BNev 52	BNev 49	Mel 40	Mel 30	SebK 58	SebK 47	SebK 45	SebK 62
Gemini	X-ray	Nor 6	–	–	–	Sarc 3	–	–	Nor 1	–	–	–	–	–
Gemini	MRI	Gli 1	–	MS 1	–	–	Men 1	–	Atr 1	Oth 1	–	–	–	–
Gemini	Derm	BNev 2	–	–	–	–	–	–	–	BNev 1	–	–	–	BNev 1

Vohra, "Hearsay: Vision-Language Medical Diagnoses Without an Image," Preprint.

How would you evaluate multimodal reasoning?

- Compare unimodal and multimodal performance.
- Introduce contradictions between modalities.
- Test missing inputs and distribution shift.
- Measure the effect on clinical decisions.

A second problem: modalities are not missing at random

- Who gets a CT scan?
- Who receives genomic testing?
- Data availability can reflect access, clinician choice, and disease severity.

What comes next?

longitudinal	follow patients over time
interactive	request missing evidence
adaptive	use specialized tools as needed
evaluated	test gains in actual decisions

Multimodal reasoning still needs to be demonstrated.

Summary

- Why multimodal health AI?
- How do we combine modalities?
- From fusion to foundation models
- Does multimodal AI use the data?



How can we make Data 146 better for you?

Next class: Complexities of the healthcare system + Jupyter Health