

Evaluating AI in Healthcare

Ahmed Alaa

UC Berkeley and UCSF

amalaa@berkeley.edu

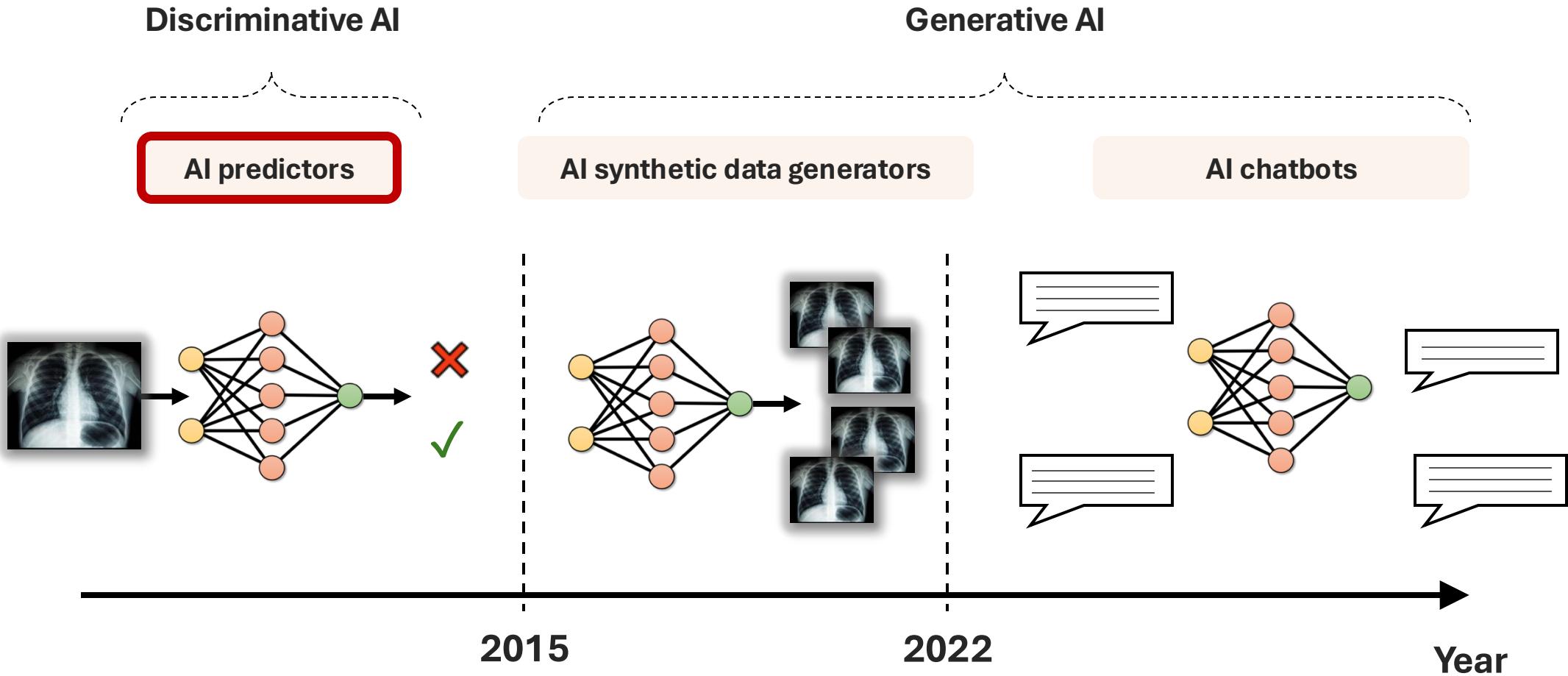


Berkeley
UNIVERSITY OF CALIFORNIA

CS188

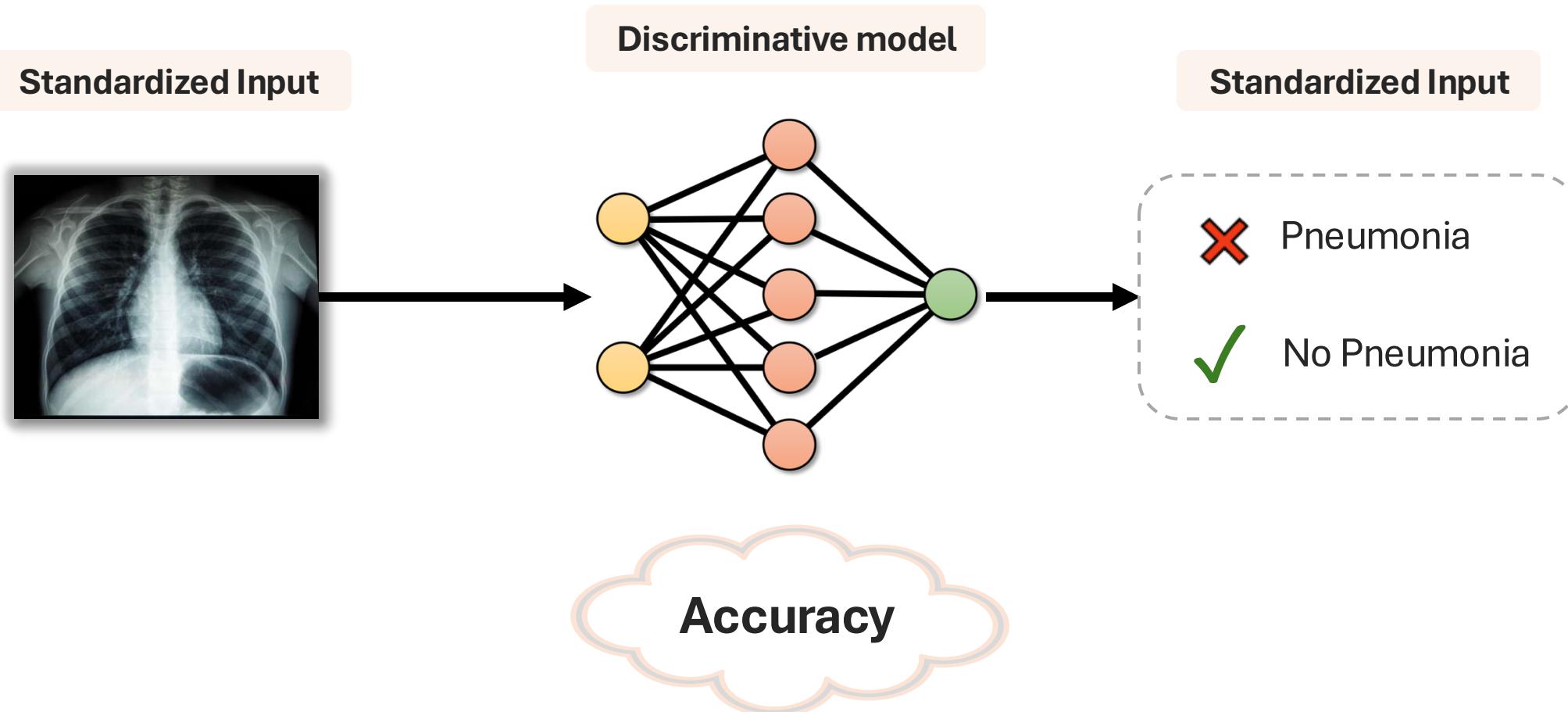
UCSF
University of California
San Francisco

The Evolution of (Medical) AI



Discriminative AI for Healthcare

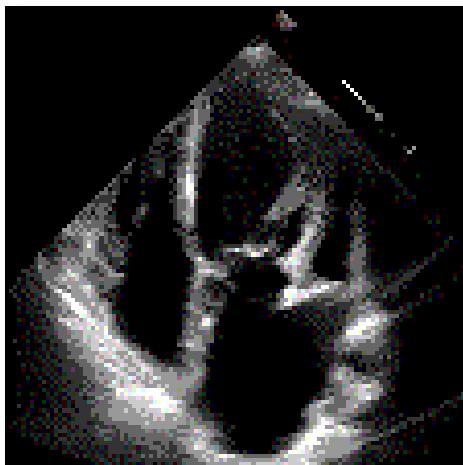
- **Discriminative AI models** = Narrow, task-specific, with structured inputs/outputs.



Discriminative AI for Healthcare

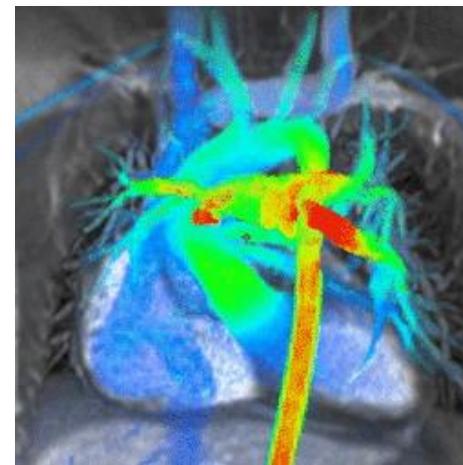
- Feature space corresponds to a rich data modality...

**Cardiac
Ultrasound**



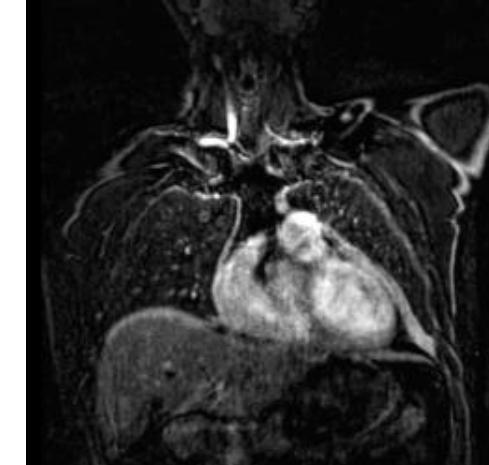
Assessing cardiac function

4D Flow MRI



Assessing blood flow

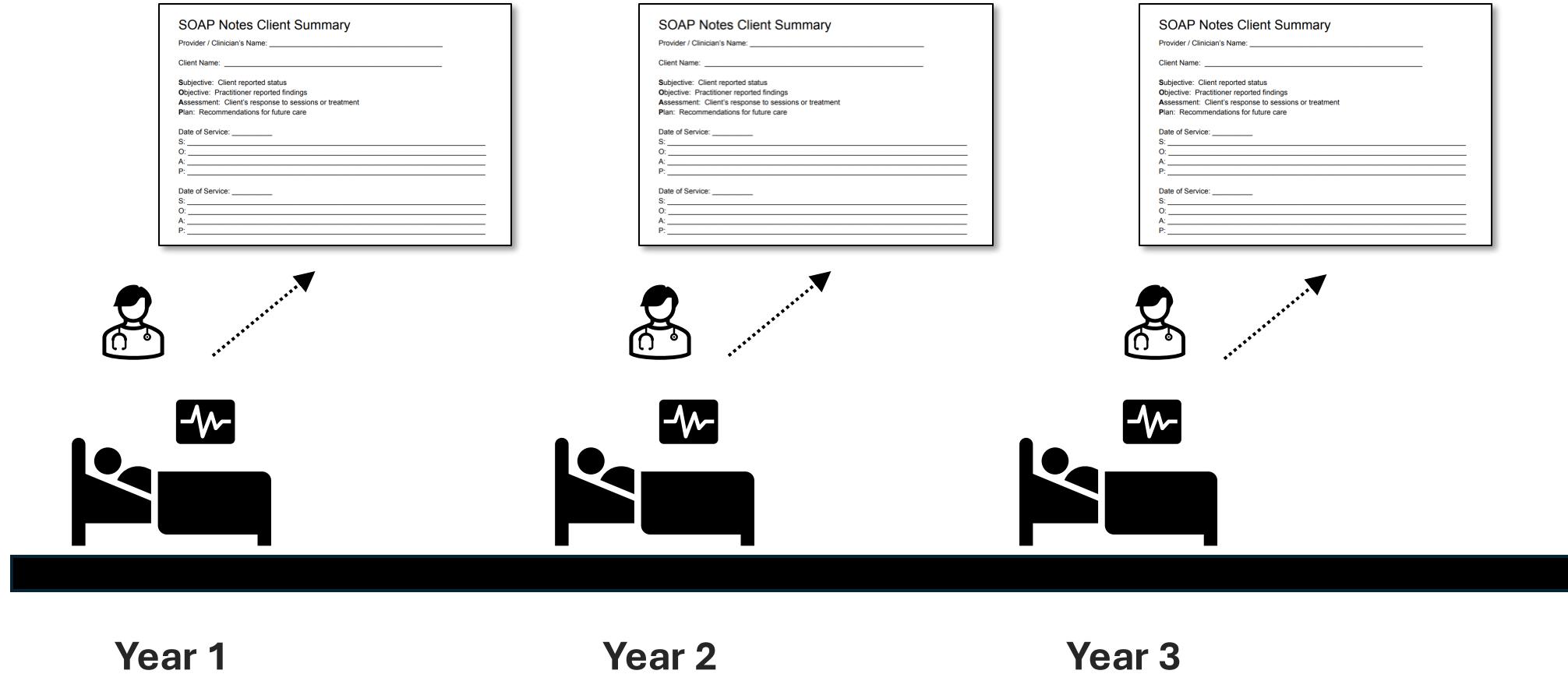
Cardiac MRI



*Assessing cardiac
structure*

Discriminative AI for Healthcare

- Clinical notes document patient health over time!



Healthcare Data at UCSF

- The UCSF Information Commons platform...

**~ 200 million
clinical notes**

**~ 180 million
clinical images**

**25K+ Cancer
Genetic testing**

Research Data Assets

Linkable multi-modal clinical data

- 6.5M+ UCSF Health Patients
- 1.2M+ SFDPH Network Patients
- 217M+ Encounters
- 40+ years of clinical activity
- 198M+ Clinical Notes
- 180M+ Clinical Images
- 25K+ Cancer Genetic Testing Results

Deidentified * Standardized * No-IRB-Access

[LEARN ABOUT THE DATA](#)

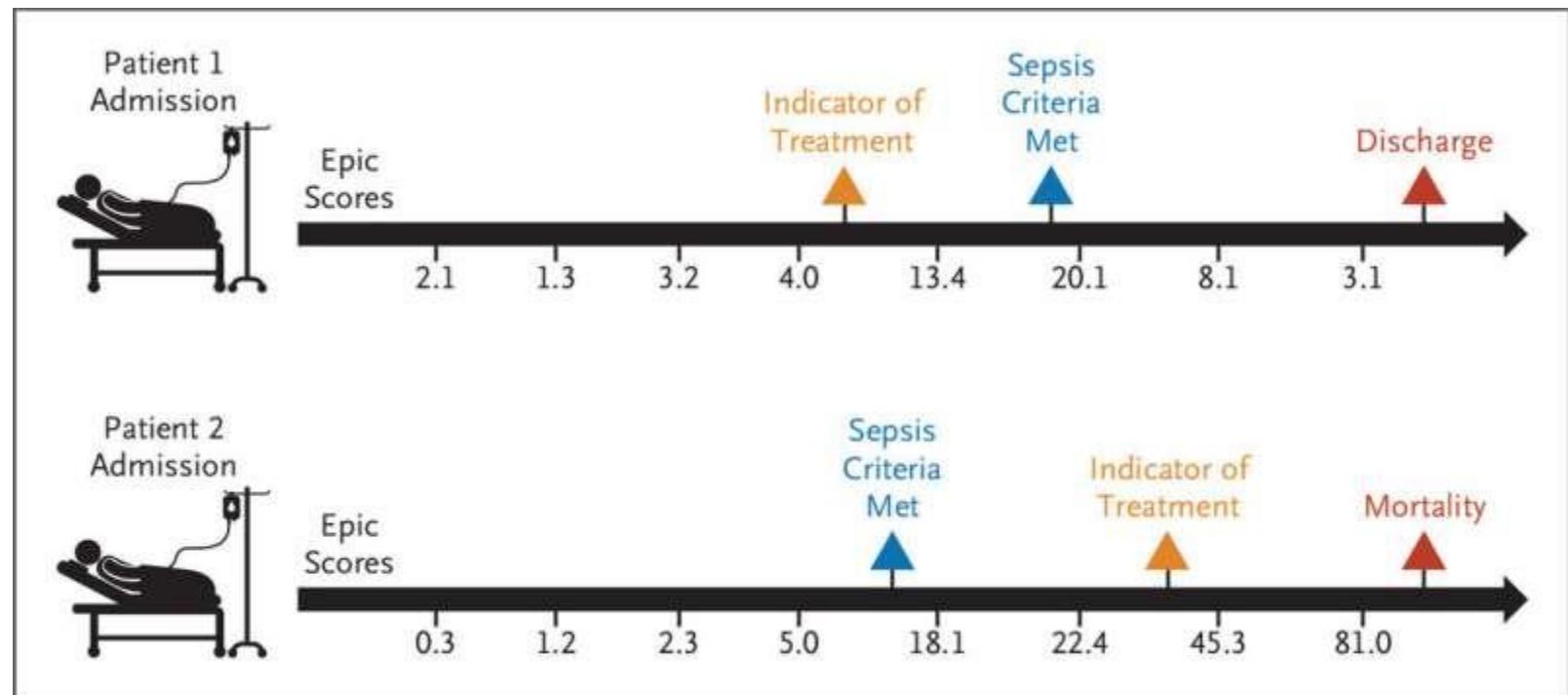
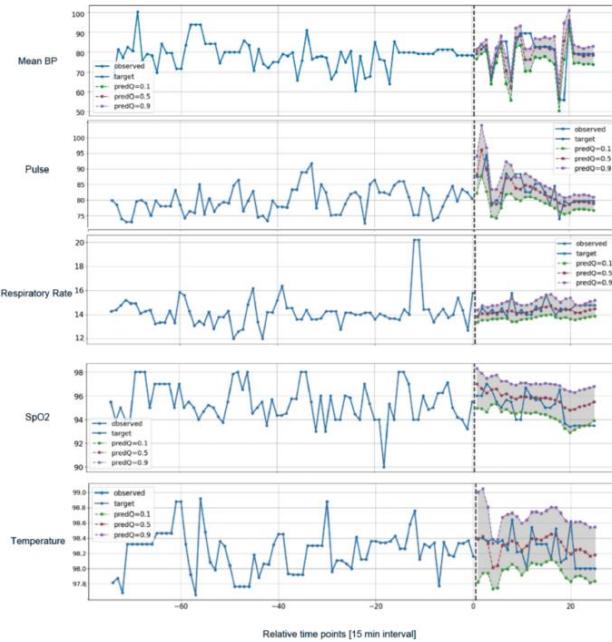
Data Tools

Open source and custom applications to support every part of data research.

[LEARN ABOUT THE TOOLS](#)

Example: AI-based Sepsis Detection

- Sepsis is a life-threatening condition where the body's response to infection causes organ dysfunction.

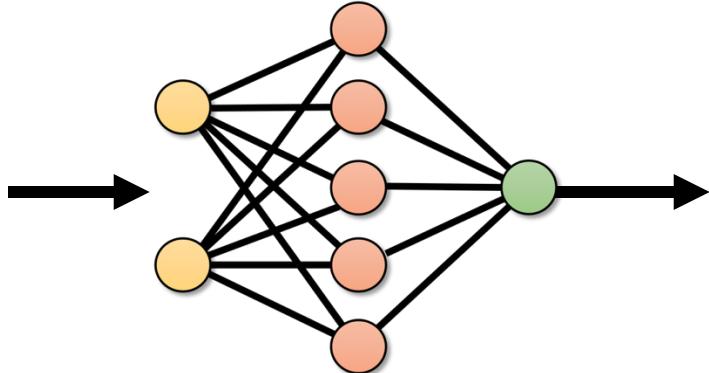
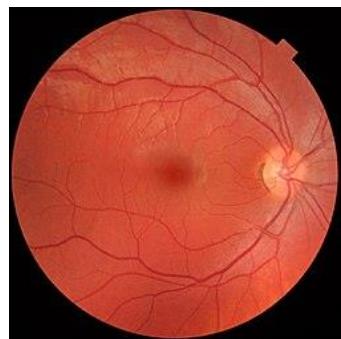


Vital signs

Credit: NEJM AI (2024). DOI: 10.1056/Aloa2300032

Example: AI Detection of Diabetic Retinopathy

- Diabetic Retinopathy is a diabetes complication that damages retinal blood vessels and is one of the leading causes of preventable blindness globally.



Refer to ophthalmologist

Right (OD)

DIABETIC RETINOPATHY (DR)
● ● ● ● Severe NPDR

DIABETIC MACULAR EDEMA (DME)
✗ DME detected

Best assessed visual acuity (VA)
Right VA
✓ 20 / 30

Left (OS)

DIABETIC RETINOPATHY (DR)
● ● ● ● Mild NPDR

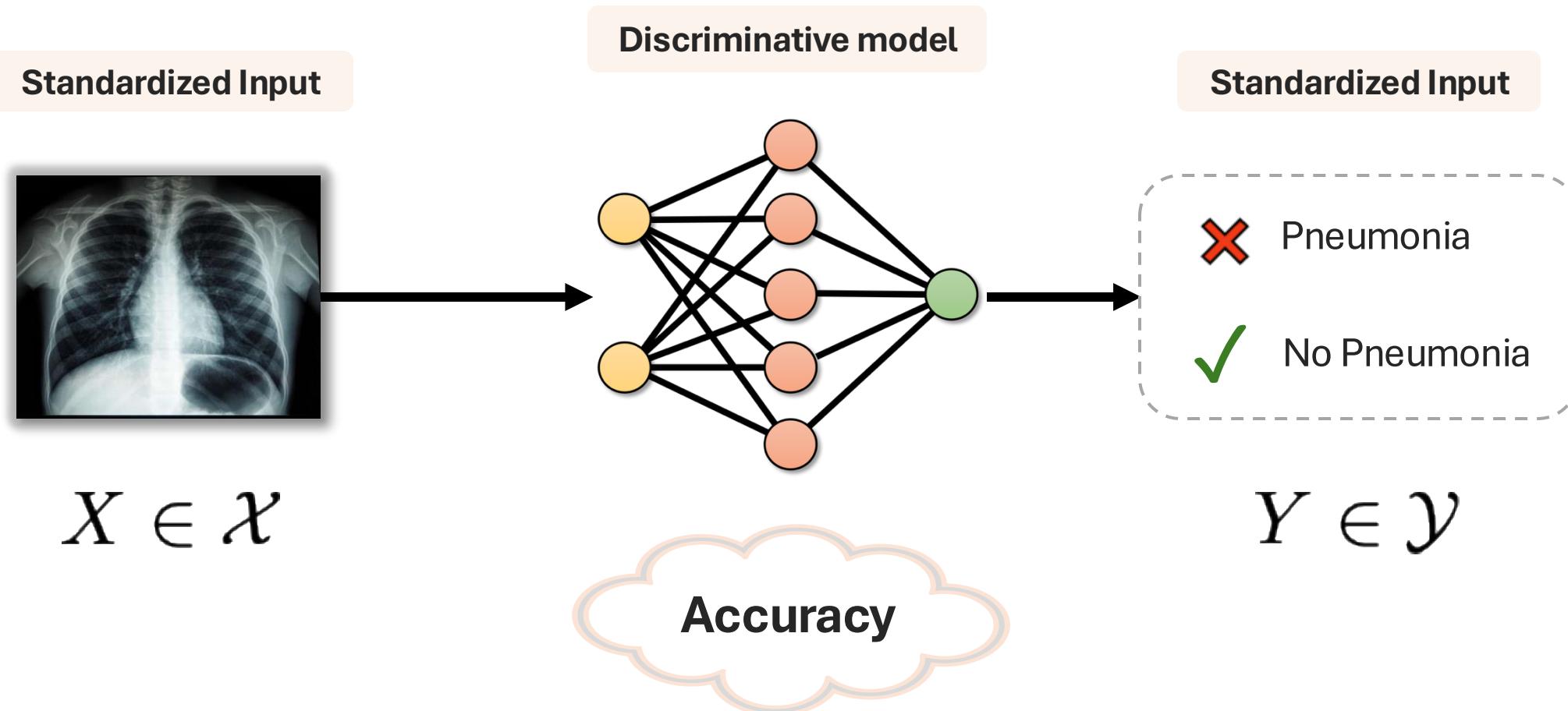
DIABETIC MACULAR EDEMA (DME)
✓ No DME detected

Left VA
✓ 20 / 30

Edit

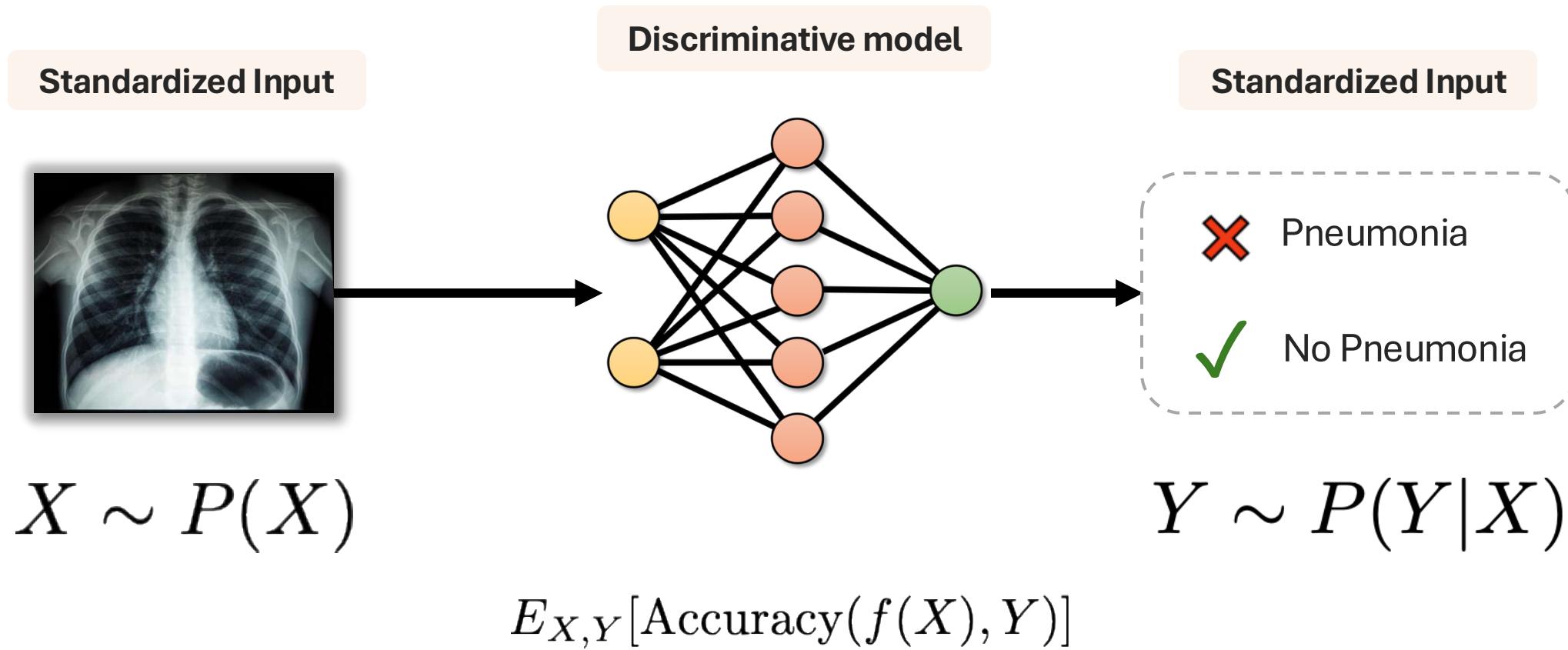
Evaluating Discriminative AI Models

- **Discriminative AI models** = Narrow, task-specific, with structured inputs/outputs.



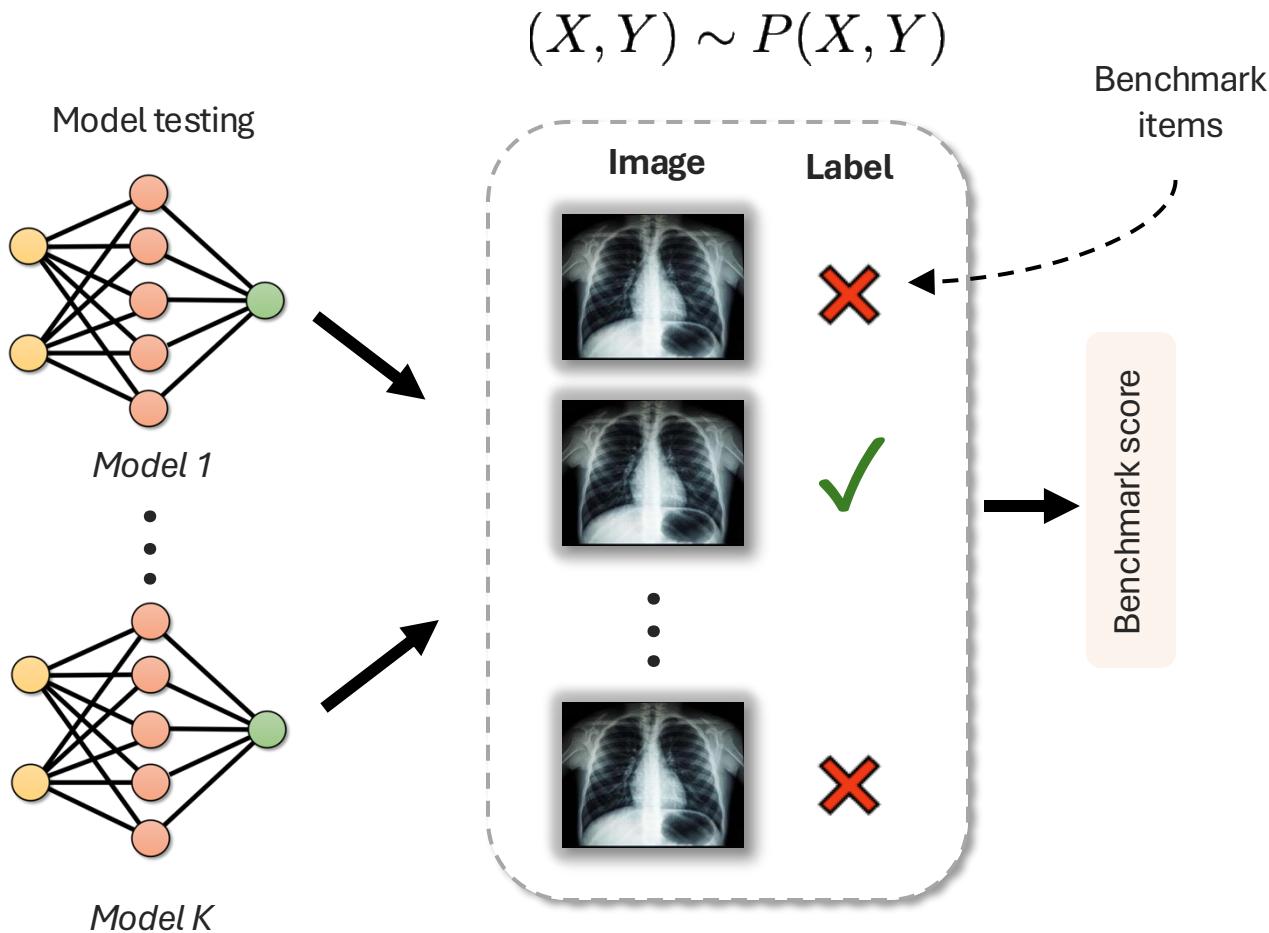
Evaluating Discriminative AI Models

- **Discriminative AI models** = Narrow, task-specific, with structured inputs/outputs.



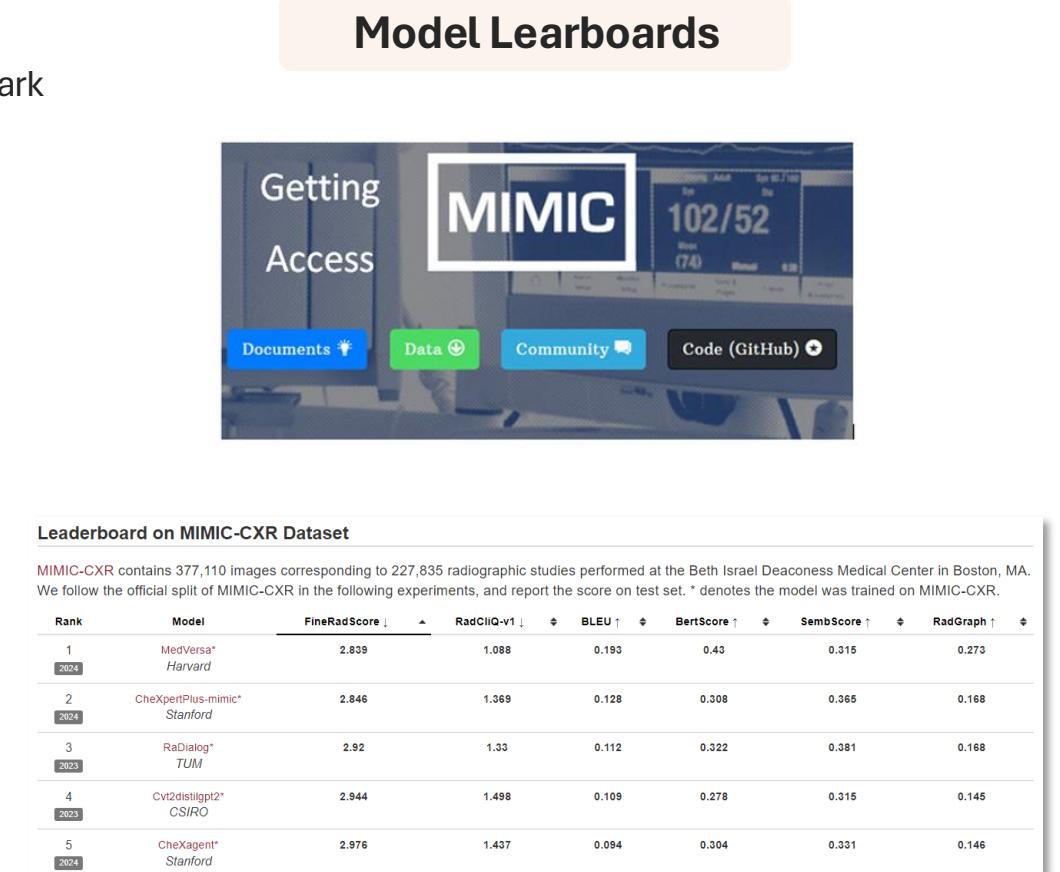
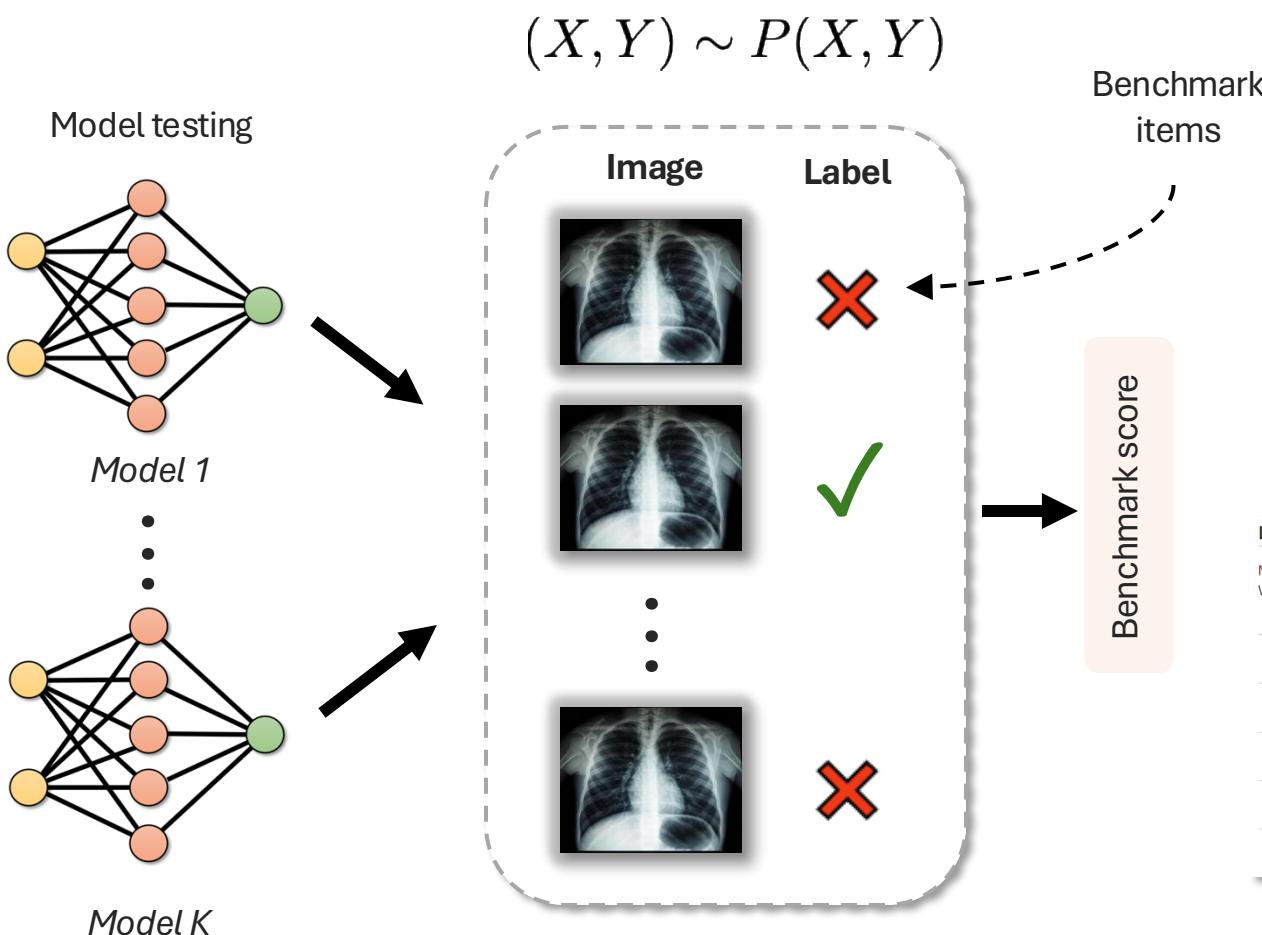
Evaluation using Benchmark Datasets

- **Benchmark datasets** are realizations of the data generating distribution!



Evaluation using Benchmark Datasets

- **Benchmark datasets** are realizations of the data generating distribution!



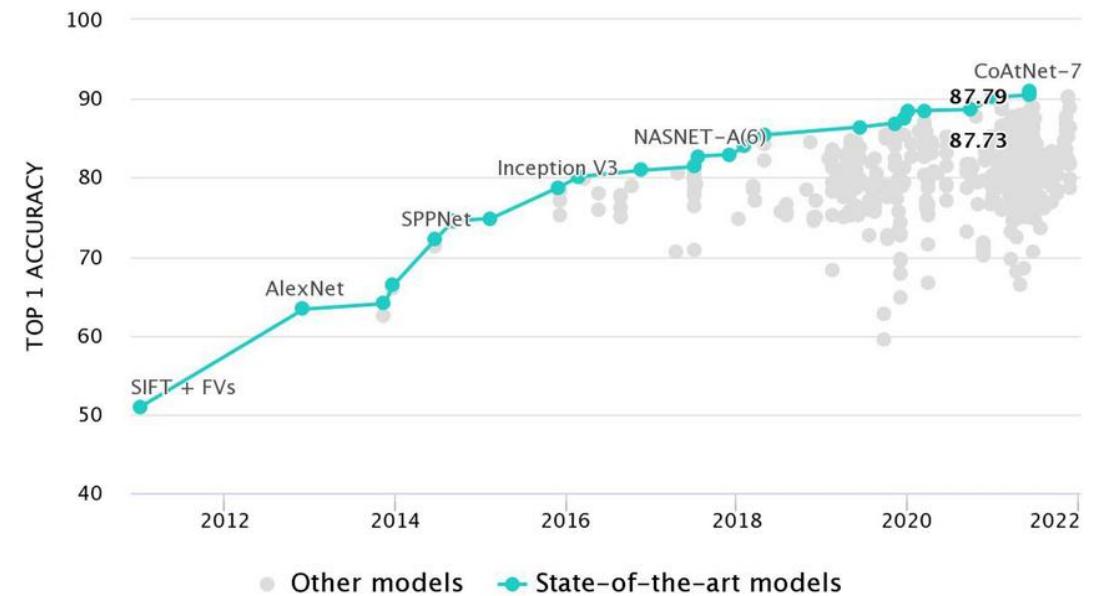
Benchmarks Drove Progress in Discriminative AI

- “AI as a sport”: Model evaluation using competitive benchmarks^{1,2}

Competitive performance leaderboards

Rank*	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Exp-03-25	1439	+6/-5	10389	Google	Proprietary
2	1	o3-2025-04-16	1418	+14/-9	2211	OpenAI	Proprietary
2	3	ChatGPT-4o-latest (2025-03-26)	1408	+6/-5	9229	OpenAI	Proprietary
3	5	Grok-3-Preview-02-24	1402	+4/-5	14840	xAI	Proprietary
3	5	Gemini-2.5-Flash-Preview-04-17	1393	+10/-7	4073	Google	Proprietary
4	3	GPT-4.5-Preview	1398	+4/-5	15285	OpenAI	Proprietary
7	12	Gemini-2.0-Flash-Thinking-Exp-01-21	1380	+4/-4	26903	Google	Proprietary
7	5	DeepSeek-V3-0324	1373	+6/-7	6792	DeepSeek	MIT
8	5	GPT-4.1-2025-04-14	1363	+10/-9	2927	OpenAI	Proprietary
9	7	DeepSeek-R1	1358	+5/-4	16857	DeepSeek	MIT
9	16	Gemini-2.0-Flash-001	1354	+3/-3	23060	Google	Proprietary

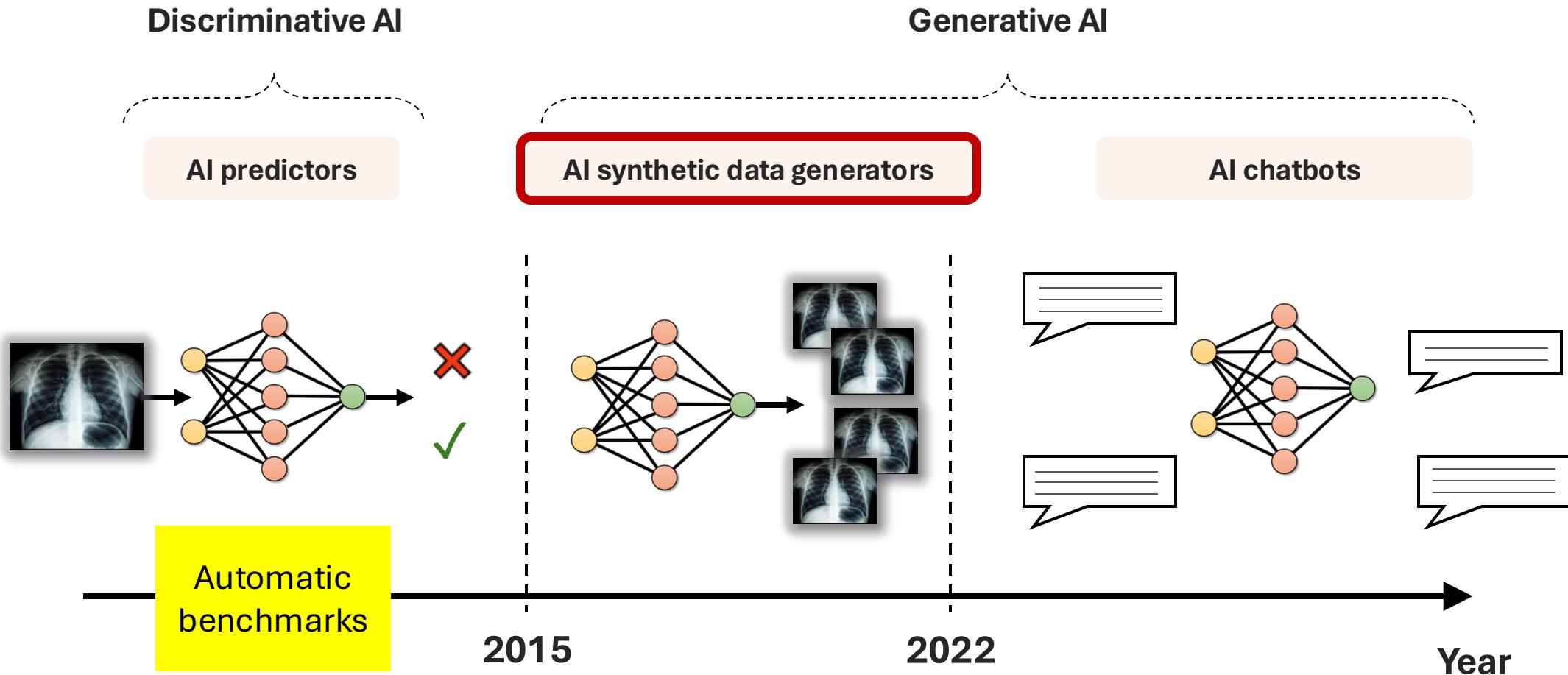
ImageNet benchmark for computer vision



¹J. McCarthy, *Science* (1997)

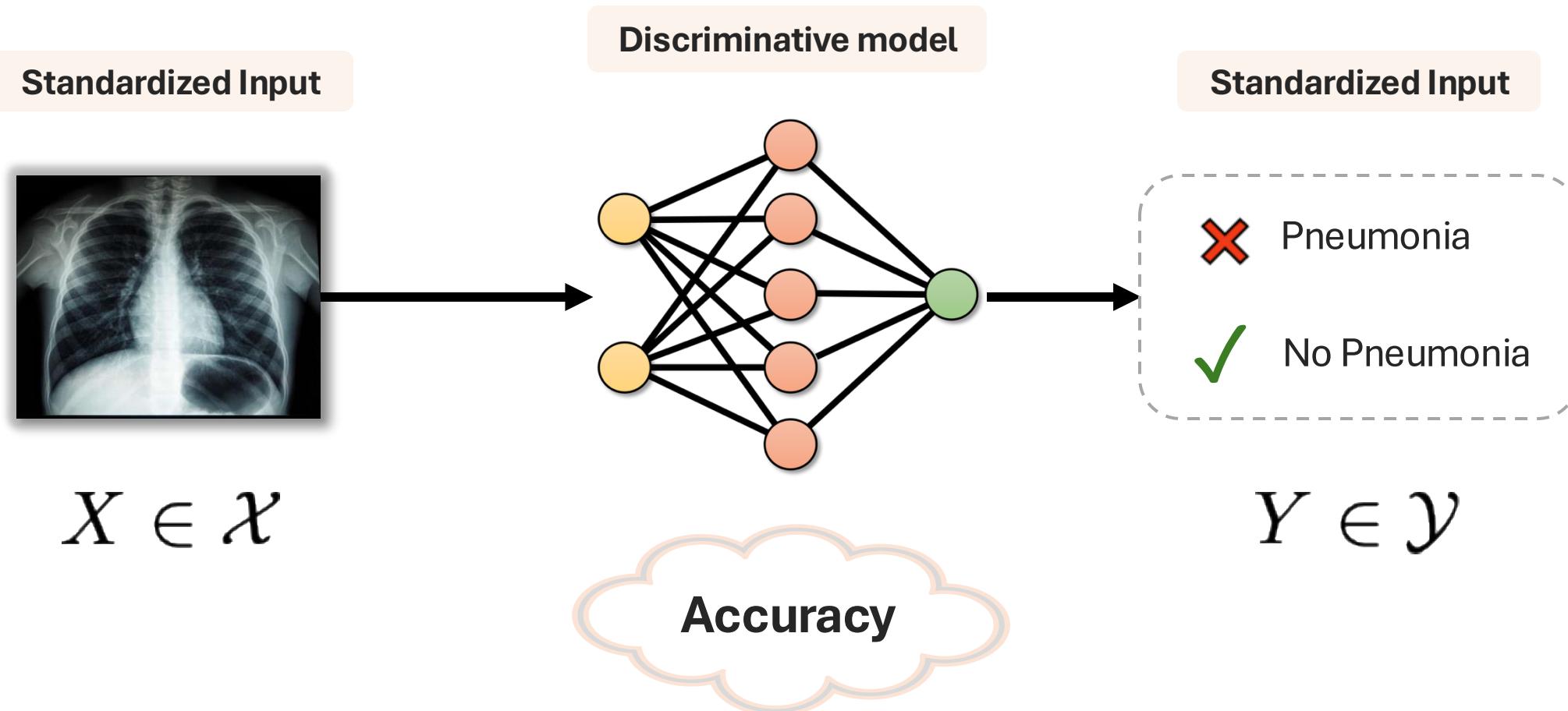
²Orr & Kang, FAccT (2024)

The Evolution of (Medical) AI



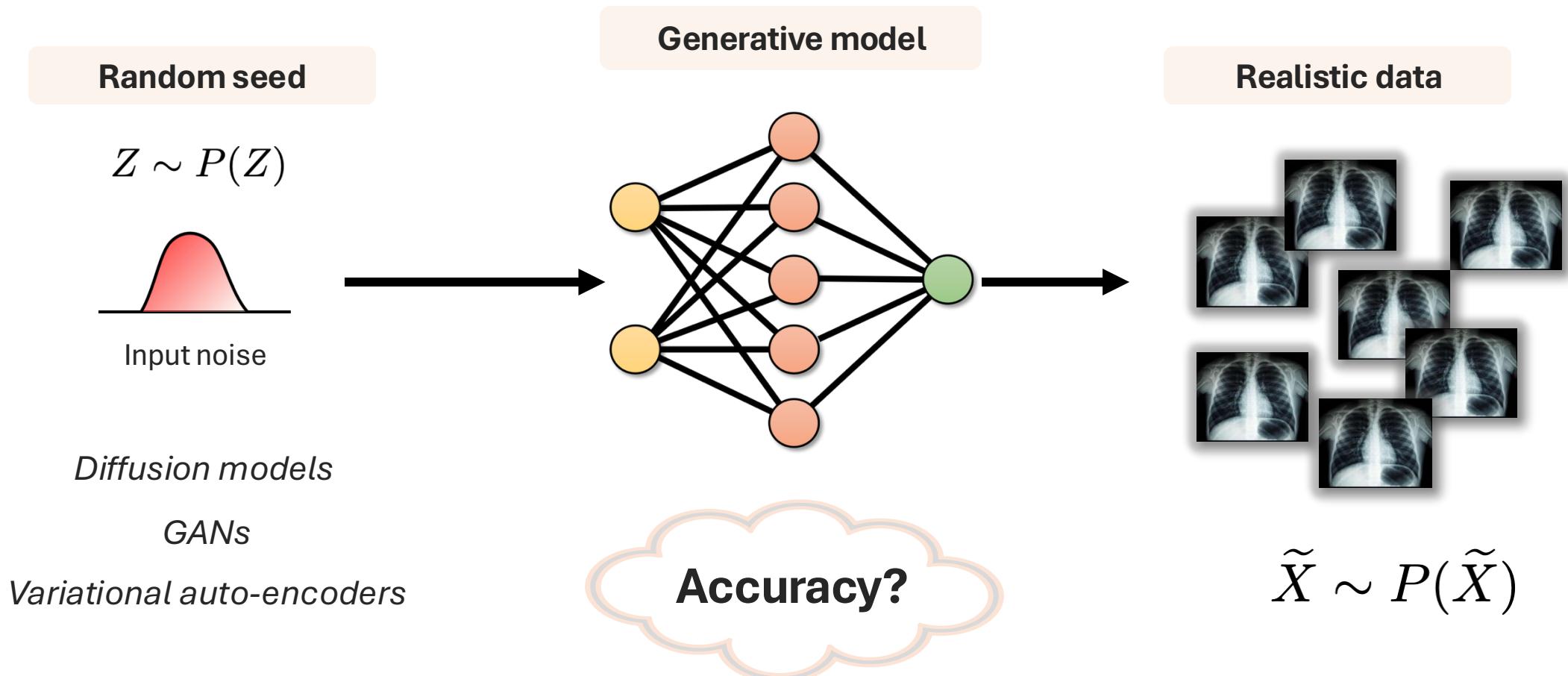
Why are Generative AI Models Different?

- **Discriminative AI models** = Narrow, task-specific, with structured inputs/outputs.



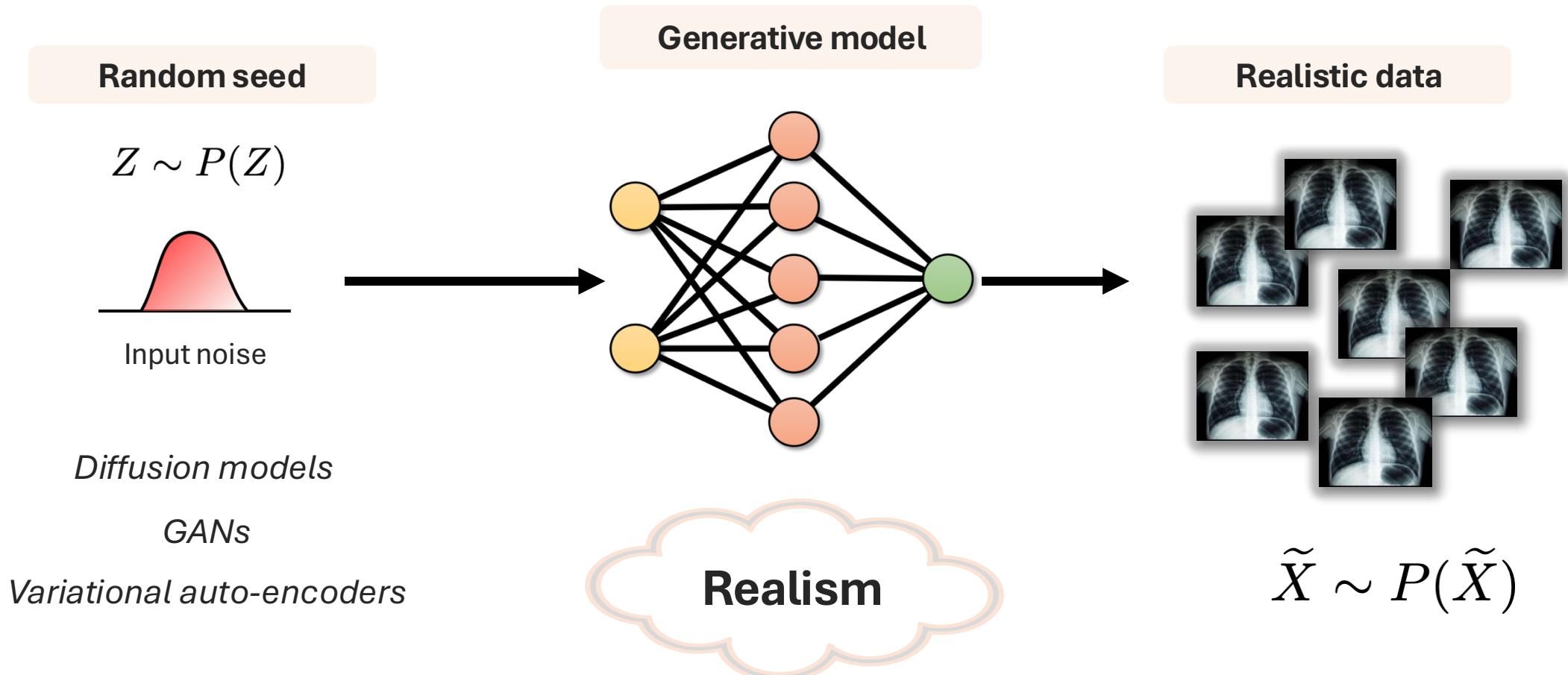
Evaluating Generative AI Models

- **Unconditional Generative models** = Learns the distribution of input features.



Evaluating Generative AI Models

- **Unconditional Generative models** = Learns the distribution of input features.



Evaluating Generative AI Models

- Unclear notion of “benchmarks” → Compare **real** and **synthetic** distributions instead.

Qualitative Evaluation



Do synthetic images look “realistic” enough?

Evaluating Generative AI Models

- Unclear notion of “benchmarks” → Compare **real** and **synthetic** distributions instead.

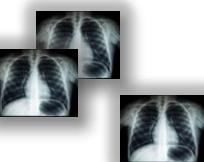
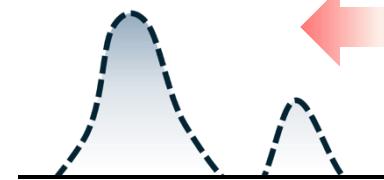
Qualitative Evaluation



Quantitative Metrics

Real distribution

$$X \sim P(X)$$



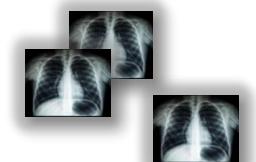
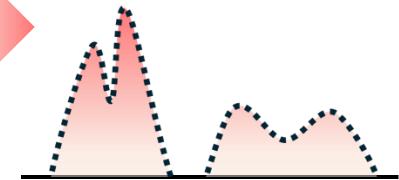
Real data

Synthetic distribution

$$\tilde{X} \sim P(\tilde{X})$$

Similarity metric

Fréchet
inception
distance



Synthetic data

Do synthetic images look “realistic” enough?

How similar are the real and synthetic distributions?

Example: Fidelity and Diversity Metrics

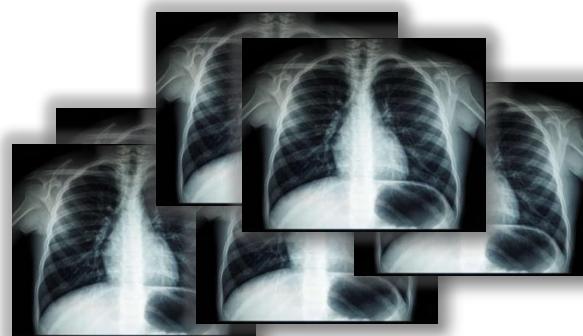
- **Fidelity** = how close generated samples are to the real-data manifold^{1,2}
- **Diversity** = how well the generator covers all modes of the real distribution

Fidelity



Every synthetic image looks realistic

Diversity



Generated synthetic images represent variations across the patient population

How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models

Ahmed M. Alaa^{1,2} Boris van Breugel³ Evgeny Saveliev³ Mihaela van der Schaar^{1,4}

Abstract

Devising domain- and model-agnostic evaluation metrics for generative models is an important and yet unsolved problem. Most existing metrics, which were tailored solely to the image synthesis application, exhibit a limited capacity for diagnosing modes of failure of generative models across broader application domains. In this paper, we introduce a 3-dimensional metric, $(\alpha, \text{Precision}, \beta, \text{Recall}, \text{Authenticity})$, that characterizes the fidelity, diversity and generalization performance of any generative model in a wide variety of application domains. Our metric uses sample-level statistical divergence measures, with precision as a pipeline, enabling sample- and distribution-level diagnoses of model fidelity and diversity. We introduce generalization as an additional dimension for model performance that quantifies the extent to which a model copies training data—a crucial performance indicator when modeling sensitive and private data. The three metric components are interpretable probabilistic quantities, and can be estimated via sample-level binary classification. The sample-level nature of our metric inspires a novel use case, which we call *model auditing*, wherein we judge the quality of a set of synthetic samples generated by a (black-box) model, discarding low-quality samples and hence improving the overall model performance in a post-hoc manner.

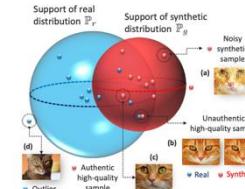


Figure 1. Pictorial depiction for the proposed metrics. The blue and red spheres correspond to the α -support of real and generative distributions, respectively. Blue and red points correspond to real and synthetic data. (a) Synthetic data falling outside the blue sphere will look unrealistic or noisy. (b) Overfitted models can generate ostensibly high-quality data samples that are “unauthentic” because they are copied from training data. (c) High-quality data samples should reside inside the blue sphere. (d) Outliers do not count in the β -Recall metric. (Here, $\alpha=0.9$, α -Precisions=9, β -Recall = 4/9, and Authenticity = 9/10.)

(VAD (Kingma & Welling, 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), do not possess tractable likelihood functions, but also because the likelihood score itself is a flawed measure of performance—it scales badly in high dimensions, and it obscures distinct modes of model failure into a single uninterpretable score (Theis et al., 2015). Absent domain-agnostic metrics, earlier work focused on crafting domain-specific scores, e.g., the Inception score in (Salimans et al., 2016), with an exclusive emphasis on image data (Lucic et al., 2018).

In this paper, we follow an alternative approach to evaluating generative models, where instead of assessing the generative distribution by looking at all synthetic samples collectively to compute likelihood or statistical divergence, we classify each sample *individually* as being of high or low quality. Our metric represents the performance of a generative model as a point in a 3-dimensional space—each dimension corre-

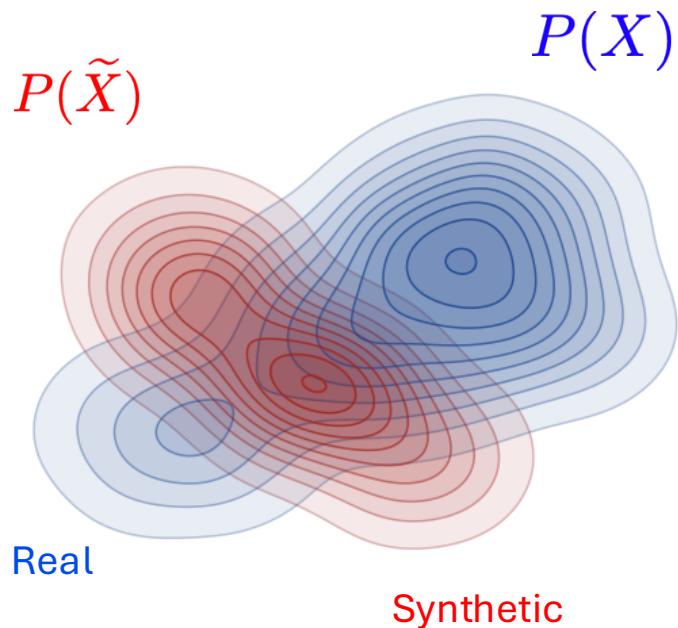
¹M. Sajjadi et al. (2018)

²T. Kynkänniemi et al. (2019)

A. Alaa et al. (ICML, 2022)

Example: Fidelity and Diversity Metrics

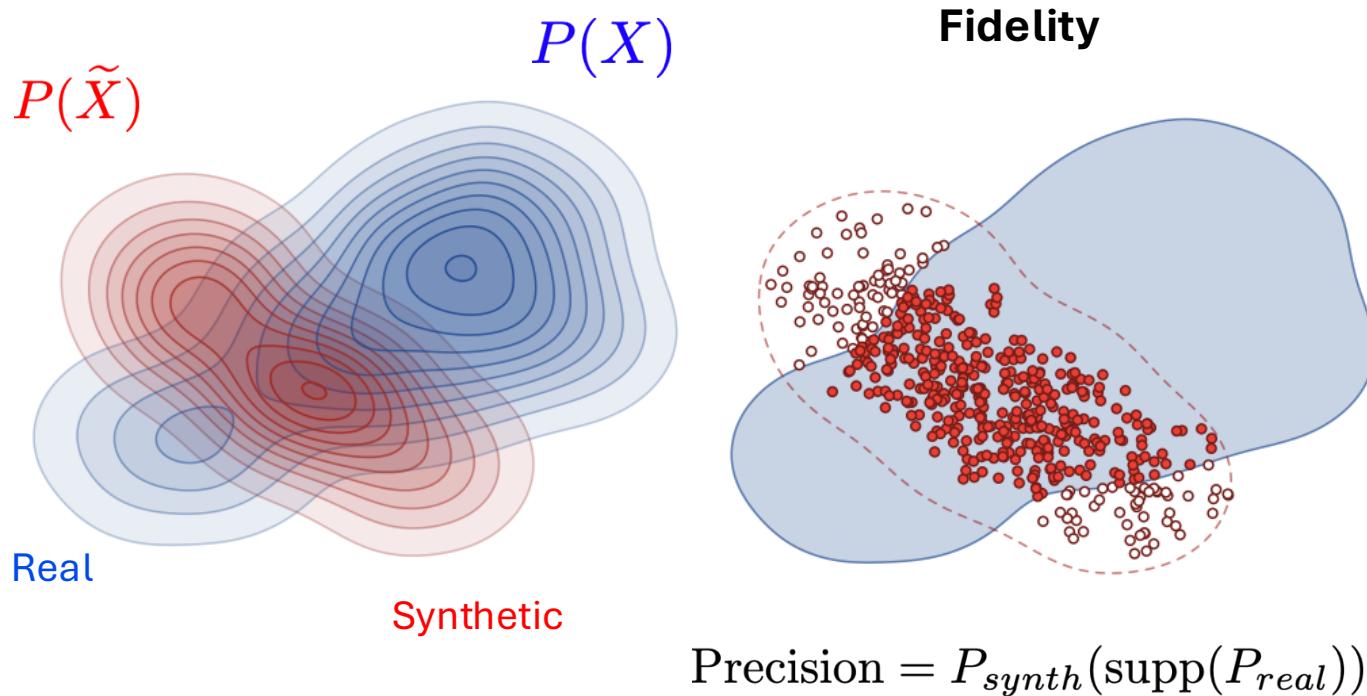
- **Fidelity** = how close generated samples are to the real-data manifold^{1,2,3}
- **Diversity** = how well the generator covers all modes of the real distribution



¹M. Sajjadi et al. (2018) ²T. Kynkäänniemi et al. (2019) ³A. Alaa et al. (2022)

Example: Fidelity and Diversity Metrics

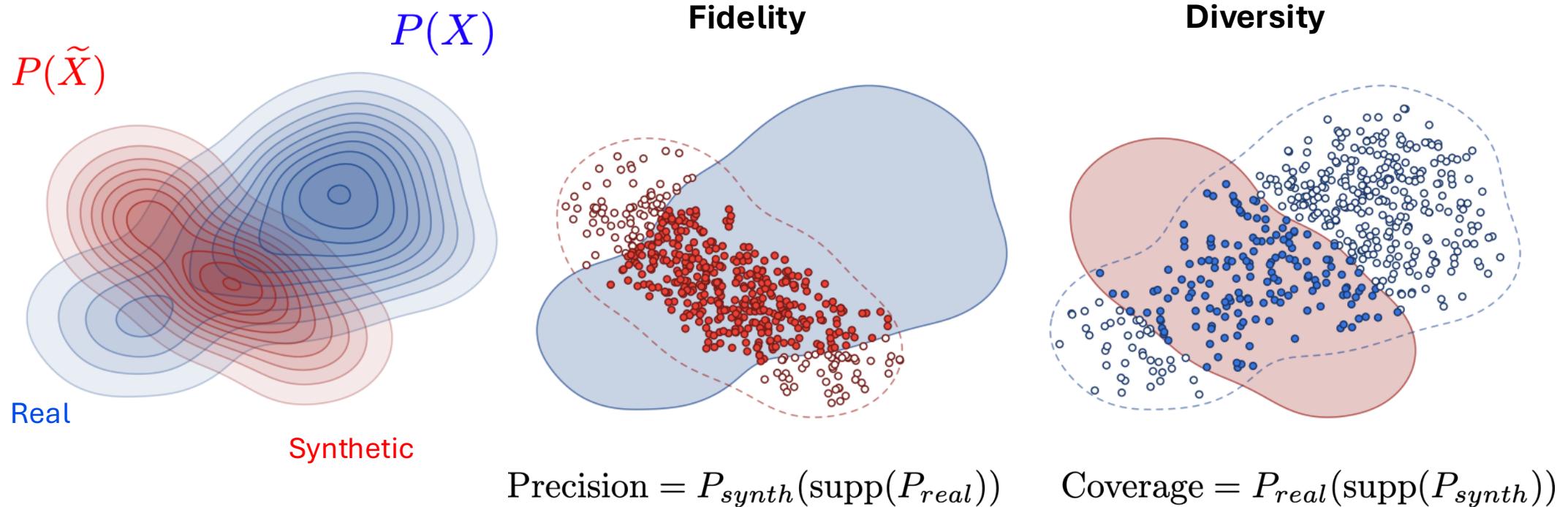
- **Fidelity** = how close generated samples are to the real-data manifold^{1,2,3}
- **Diversity** = how well the generator covers all modes of the real distribution



¹M. Sajjadi et al. (2018) ²T. Kynkäänniemi et al. (2019) ³A. Alaa et al. (2022)

Example: Fidelity and Diversity Metrics

- **Fidelity** = how close generated samples are to the real-data manifold^{1,2,3}
- **Diversity** = how well the generator covers all modes of the real distribution



¹M. Sajjadi et al. (2018) ²T. Kynkäänniemi et al. (2019) ³A. Alaa et al. (2022)

Example: Fidelity and Diversity Metrics

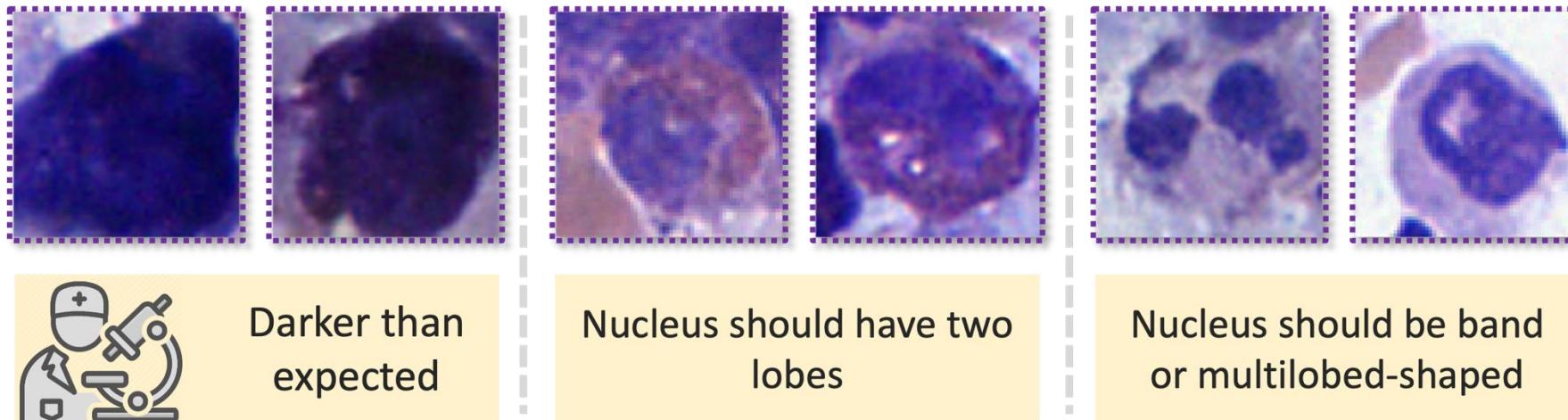
- Do these metrics really identify good synthetic generations?

Realistic visual features but does not align with a valid “world model”



Example: Fidelity and Diversity Metrics

- It is impossible to pre-specify all possible failure modes of a generative model!
- **Example:** Synthetic bone marrow aspirate whole slides using conditional diffusion.¹

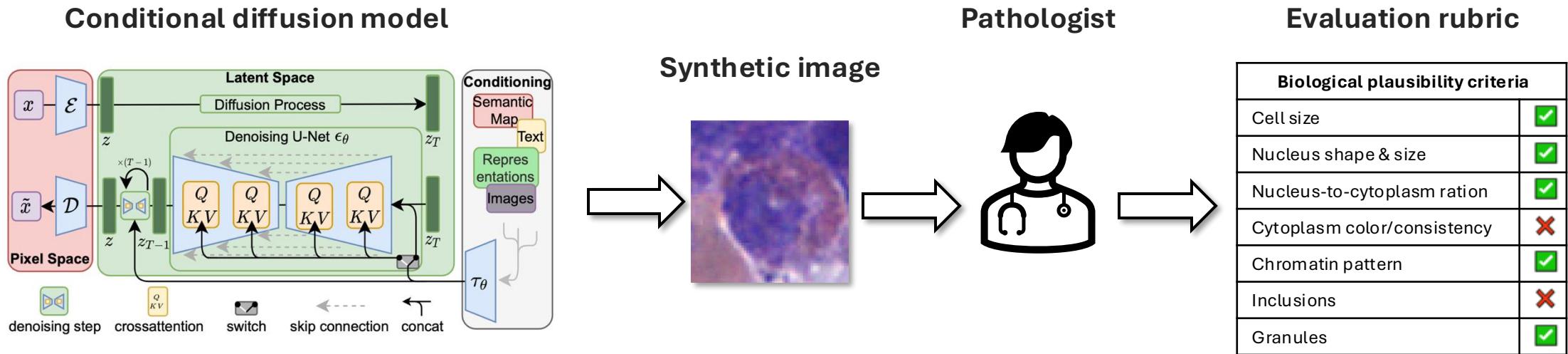


Samples that score high on precision but are biologically implausible!

¹S. Sun et al. (2023)

Human Evaluation of Generative AI Models

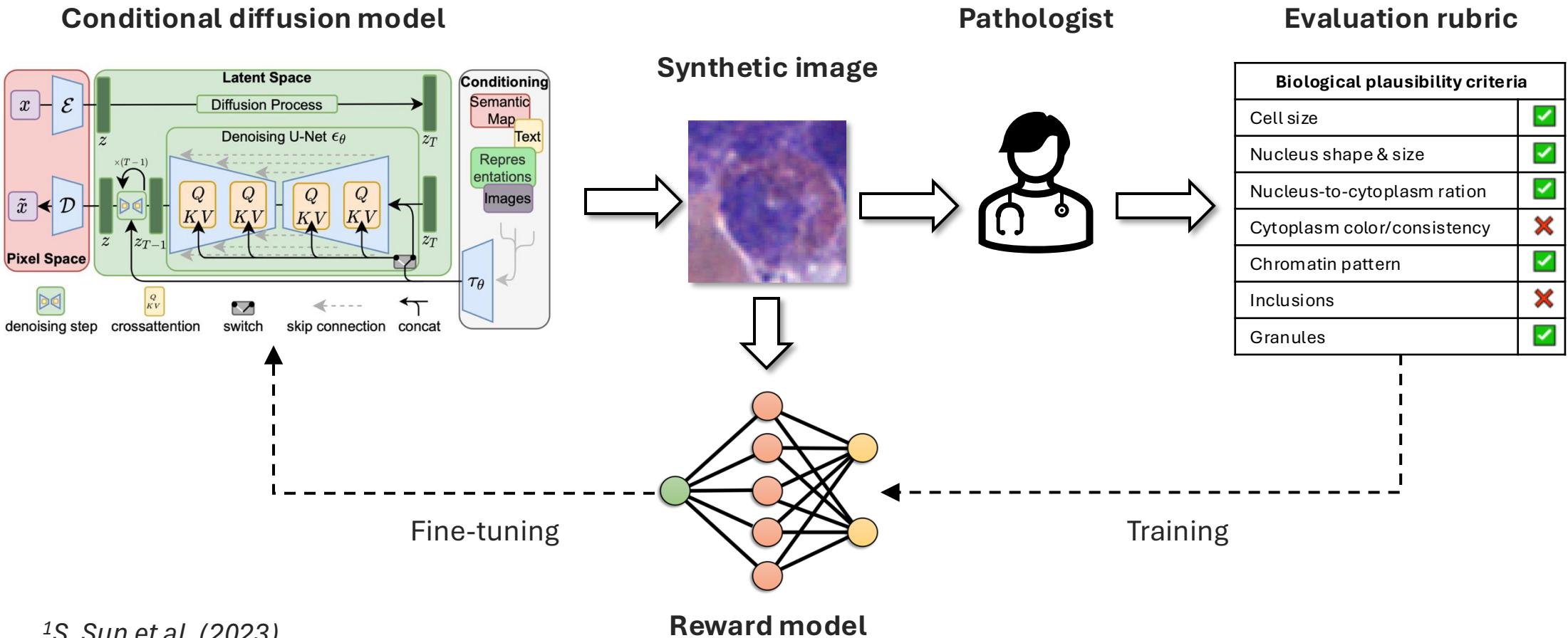
- “Pathologist-in-the-loop” evaluation using a rubric score for biological plausibility.¹



¹S. Sun et al. (2023)

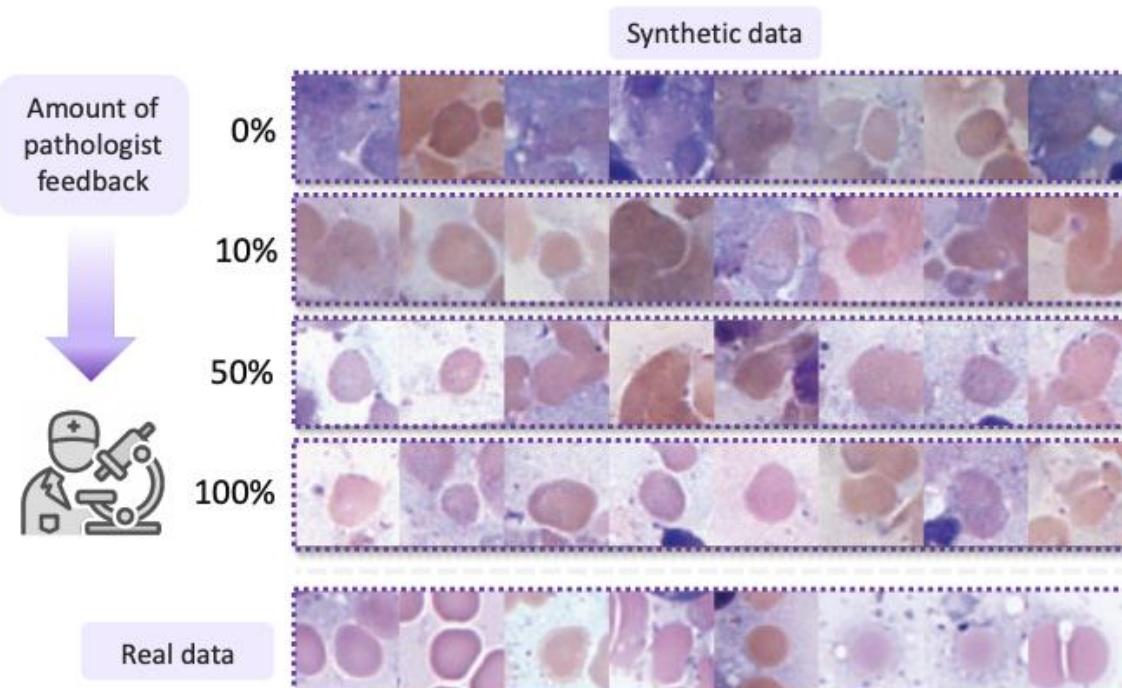
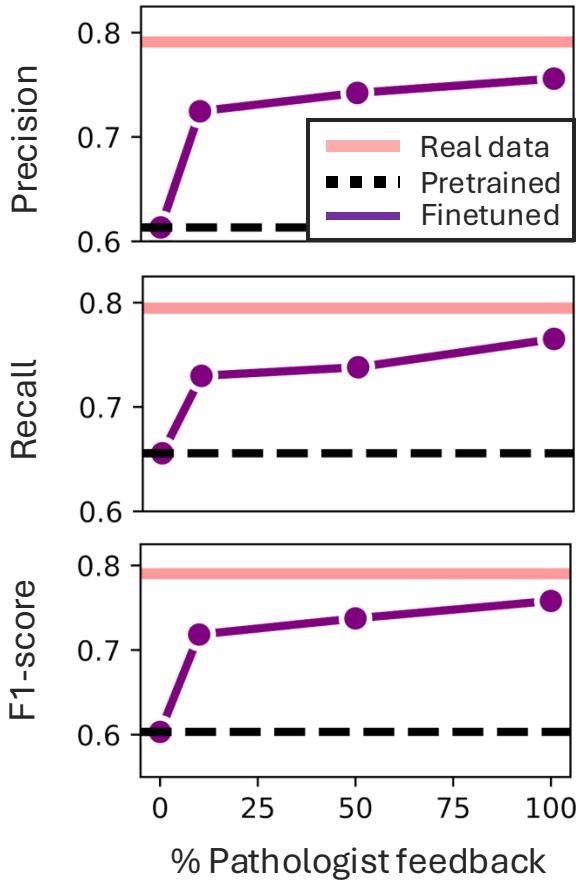
Human Evaluation of Generative AI Models

- Reinforcement Learning from Human (Pathologist) Feedback!

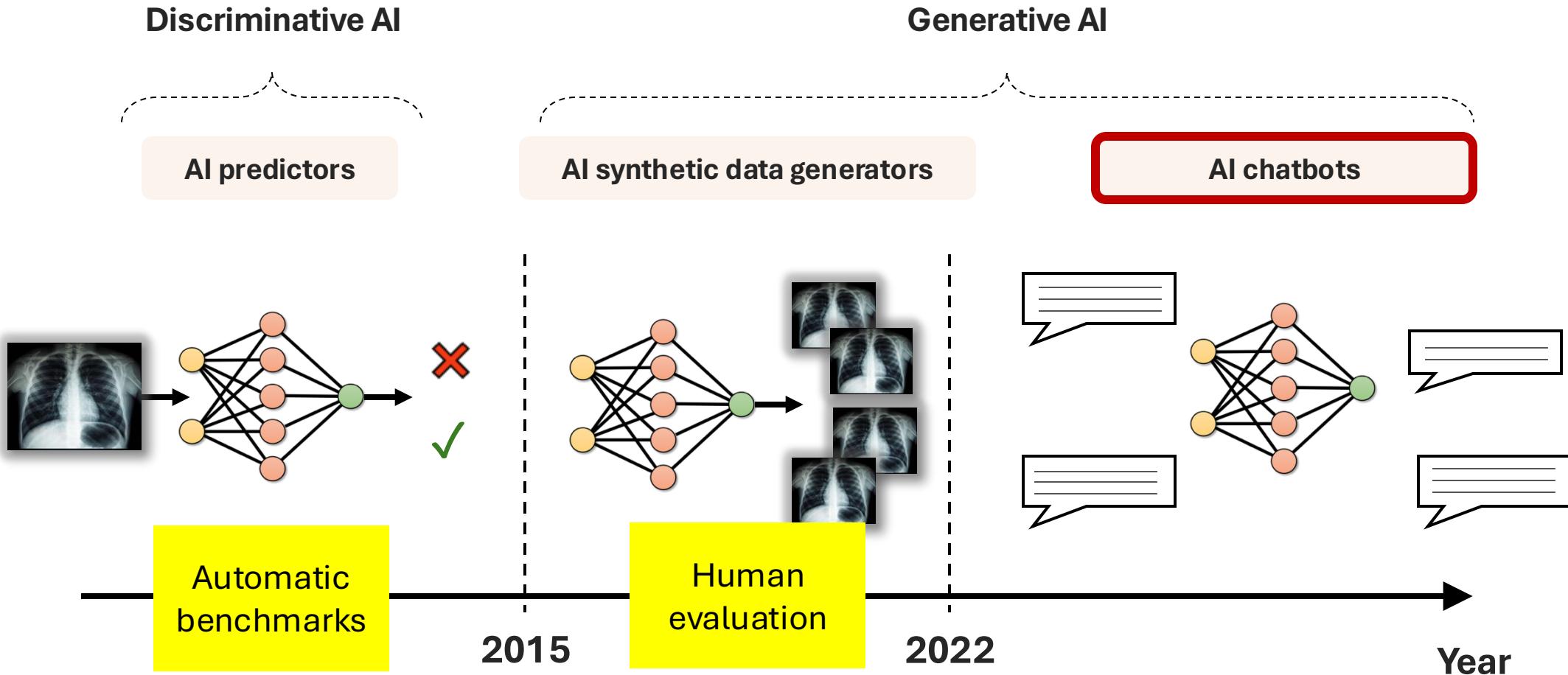


Human Evaluation of Generative AI Models

- Reinforcement Learning from Human (Pathologist) Feedback!

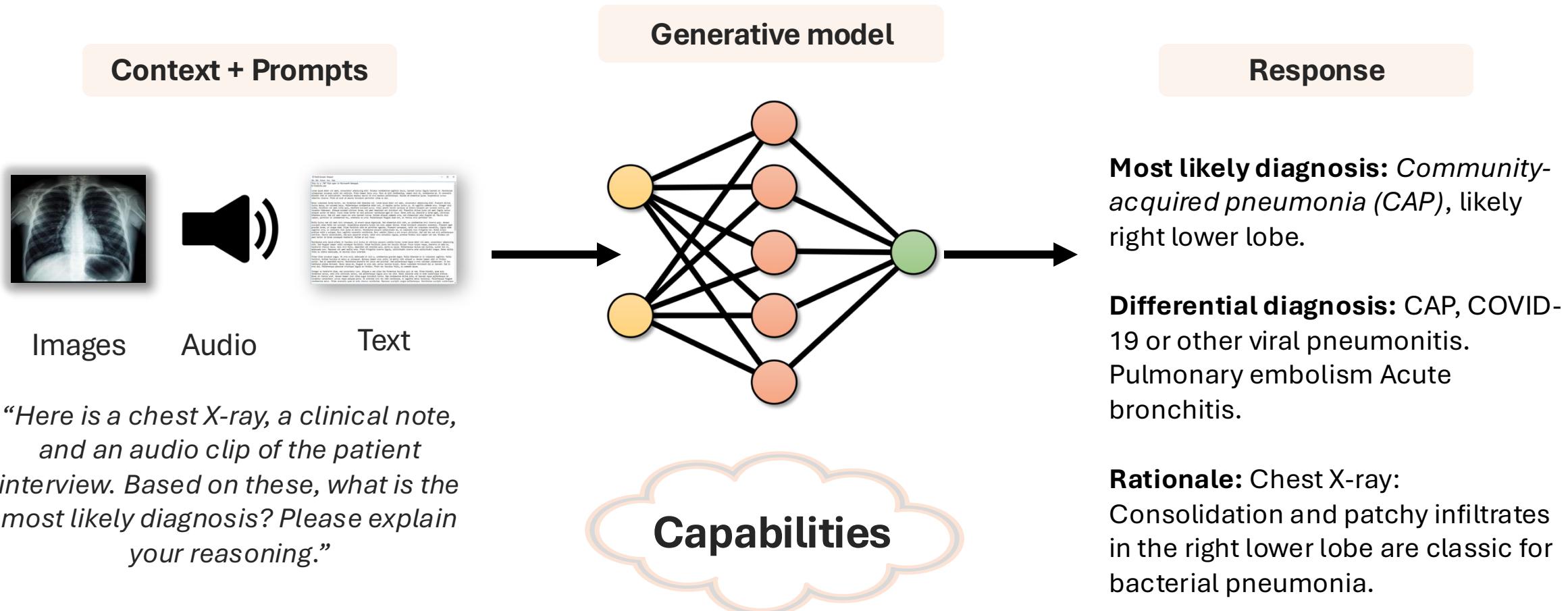


The Evolution of (Medical) AI



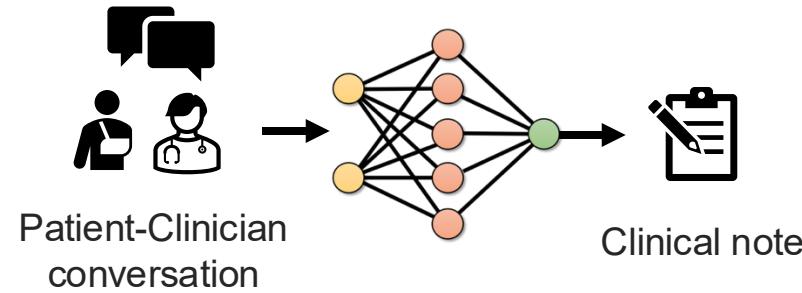
Expanded Capabilities of Generative AI Models

- **Generative AI models** = General-purpose and open-ended!

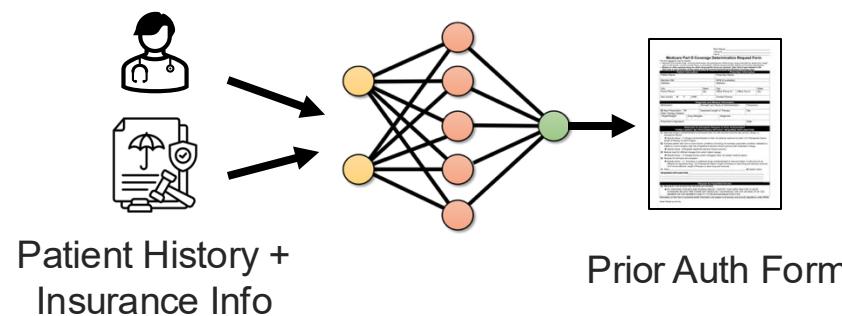


We Study Generative AI as an Emergent Phenomenon, Not Just an Engineered System

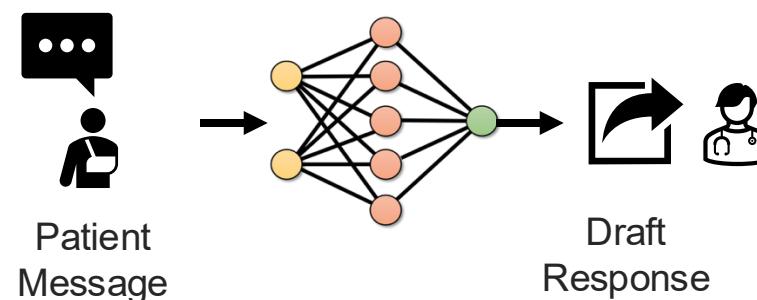
Ambient AI Scribes



Prior Authorization & Claims Automation



Patient communication



Do generative AI models...

- ...display clinical knowledge?
- ...engage in clinical reasoning?
- ...communicate empathetically?

Generative AI Models are “Under-Specified”

- Specification = the expected behavior, inputs, and outputs of a system.¹

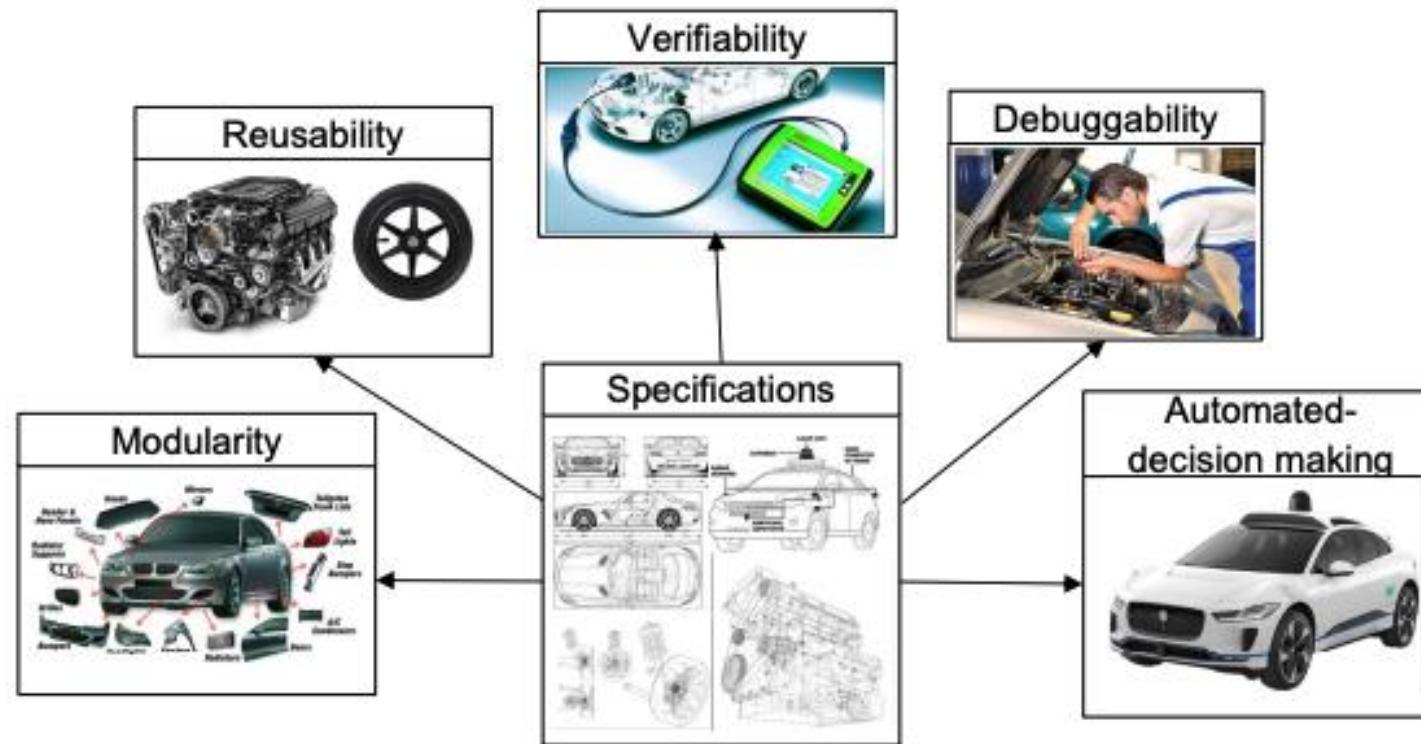
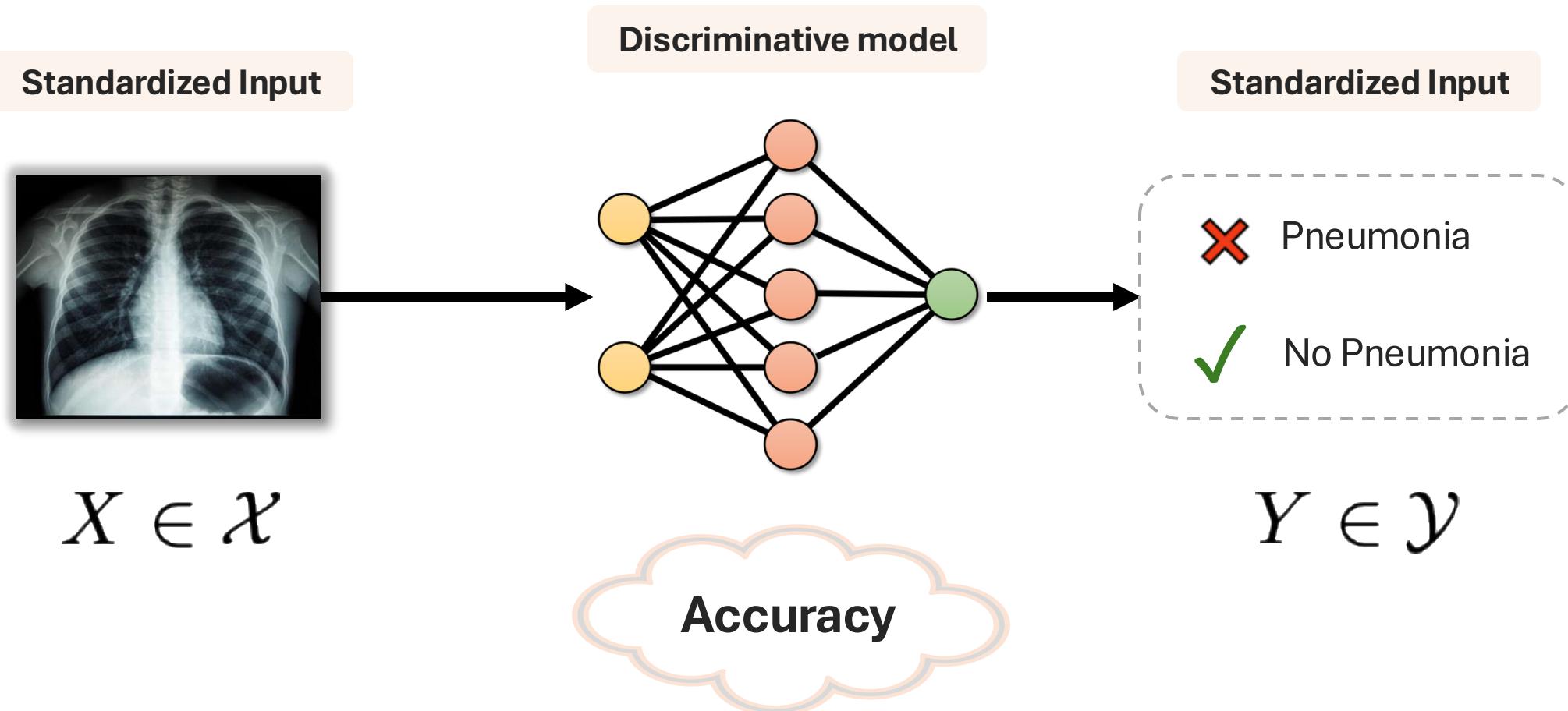


Figure courtesy of

¹I. Stoica et al. “Specifications: The missing link to making the development of LLM systems an engineering discipline”, 2025.

Discriminative AI Models are “Well-Specified”

- **Discriminative AI models** = Narrow, task-specific, with structured inputs/outputs.



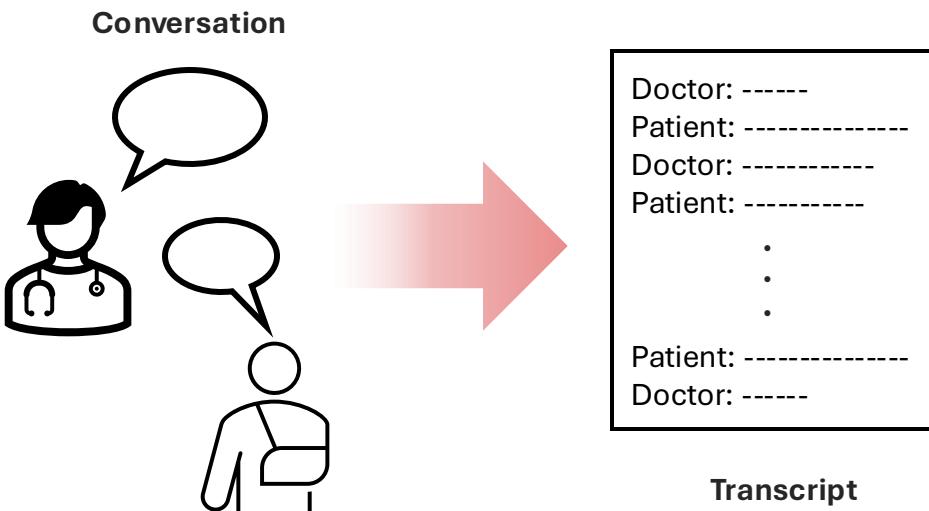
Examples of Real-World Deployment of Generative AI in Healthcare

Example 1: Ambient AI Scribes

- **AI scribes** = Digital tools that use AI and NLP to listen to patient-clinician conversations, transcribe them, and automatically generate clinical notes.

Essentially two AI tasks!

Task 1: Ambient listening
and transcription

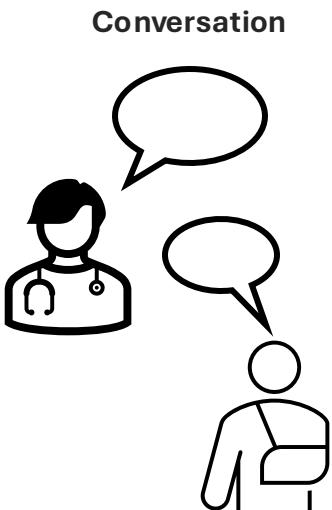


Example 1: Ambient AI Scribes

- **AI scribes** = Digital tools that use AI and NLP to listen to patient-clinician conversations, transcribe them, and automatically generate clinical notes.

Essentially two AI tasks!

Task 1: Ambient listening
and transcription



Conversation

Doctor: -----
Patient: -----
Doctor: -----
Patient: -----
.
.
.
Patient: -----
Doctor: -----

Transcript

Task 2: Clinical note
generation from transcripts

Doctor: -----
Patient: -----
Doctor: -----
Patient: -----
.
.
.
Patient: -----
Doctor: -----

Transcript

SOAP Format

Primary Care Visit
Subjective: The patient complains of chest pain that started earlier today. They rate the pain as a 7/10 and describe it as a sharp, stabbing sensation. They deny any previous history of similar symptoms.
Objective: Vital signs are normal (130/80mmHg). The physical exam reveals tenderness upon palpation in the chest area. No other abnormalities noted.
Assessment: Possible musculoskeletal injury causing chest pain.
Plan: Recommend physical rest and aspirin for pain management. Follow up in two weeks if symptoms persist or worsen.

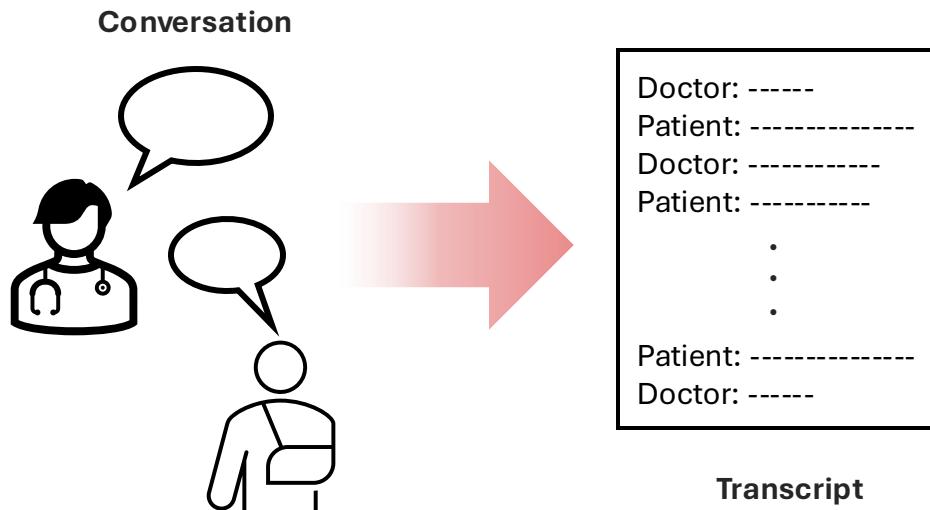
Clinical note

Example 1: Ambient AI Scribes

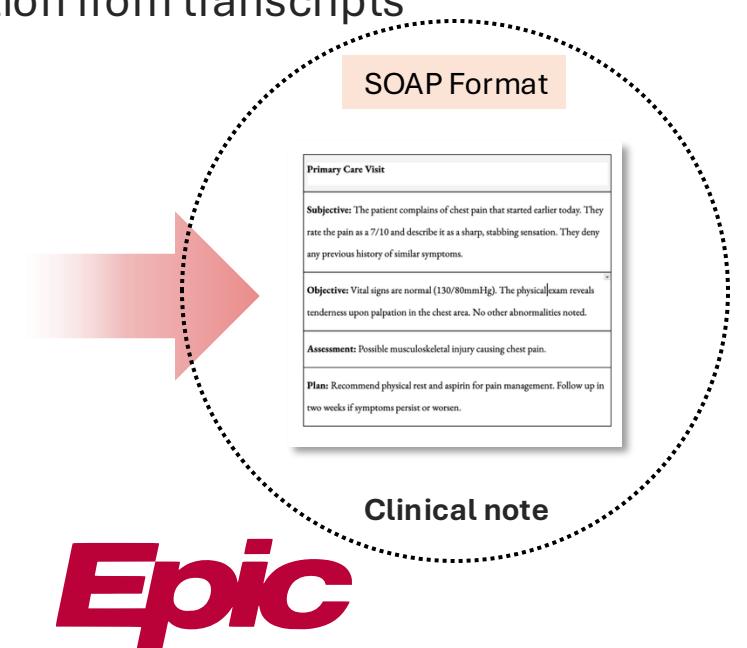
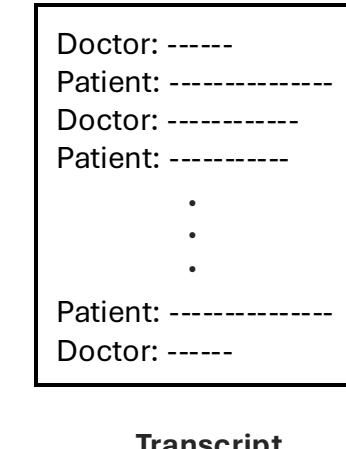
- **AI scribes** = Digital tools that use AI and NLP to listen to patient-clinician conversations, transcribe them, and automatically generate clinical notes.

Essentially two AI tasks!

Task 1: Ambient listening
and transcription



Task 2: Clinical note
generation from transcripts



The Value Proposition of AI Scribes: Clinician's Perspective

- After EHRs became mandatory in the 2010s (e.g., under the HITECH Act in the U.S.), doctors found themselves spending more time typing notes than seeing patients.
 - Clinicians can spend 1-2 hours on documentation for every hour of patient care.
 - “**Pajama time**” = completing charts after hours!
 - The additional documentation burden contributes hugely to clinician burnout!

Clinician burnout!

The Value Proposition of AI Scribes: Patient's Perspective

- Doctors often end up facing their screens, not their patients because they spend time taking notes instead of maintaining eye contact.
 - Patients feel unheard or rushed.
 - Clinicians feel detached and mechanical!

Poor patient experience!

The AI Scribes Market is Rapidly Growing!

- Companies only a few years old have reached billions of \$ in valuation!



Raised \$243M in Series C



Raised \$300M in Series E
at a \$5.3B valuation



Plans to launch
their own scribe!

And AI Scribes are being Rapidly Adopted!

- Fastest adopted and most widely implemented generative AI healthcare solution to date!

At UCSF

15%

decrease in
documentation
time

16%

decrease in time
in patient
instructions

63%

of clinicians
report decrease
in burnout

85%

of people that try
an AI scribe want
to keep it

More broadly...



JOURNAL ▾ EVEN



Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses

Authors: Aaron A. Tierney, PhD , Gregg Gayre, MD, Brian Hoberman, MD, MBA, Britt Mattern, Manuel Ballesca, MD, Sarah B. Wilson Hannay, MD, MEd, Kate Castilla, Cindy S. Lau, MD, Patricia Kipnis, PhD, Vincent Liu, MD, MS, and Kristine Lee, MD [Author Info & Affiliations](#)

Published March 31, 2025 | NEJM Catal Innov Care Deliv 2025;6(5) | DOI: 10.1056/CAT.25.0040 | [VOL. 6 NO. 5](#)
[Copyright © 2025](#)



Source: <https://ai.ucsf.edu/>

Kaiser Permanente deployed scribes for < 7,000 physicians across 2 million encounters

Understanding the Rapid Adoption of AI Scribes

- Addresses a real need. But it's administrative rather than clinical!
- Maturity of technology (LLMs, Whisper, etc)
- Low friction = Just an app that any clinician can download on their smartphone
- Integration into EHRs

Early Studies on the Impact of AI Scribes



JAMA Network **Open** 

Original Investigation | Health Policy
Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout

Kristine D. Olson, MD, MSc; Daniella Meeker, PhD; Matt Troup, PA-C; Timothy D. Barker, MD; Vinh H. Nguyen, MD; Jennifer B. Manders, MD; Cheryl D. Stults, PhD; Veena G. Jones, MD; Sachin D. Shah, MD; Tina Shah, MD, MPH; Lee H. Schwamm, MD

Abstract

IMPORTANCE While in short supply and high demand, ambulatory care clinicians spend more time on administrative tasks and documentation in the electronic health record than on direct patient care, which has been associated with burnout, intention to leave, and reduced quality of care.

OBJECTIVE To examine whether ambient AI scribes are associated with reducing clinician administrative burden and burnout.

DESIGN, SETTING, AND PARTICIPANTS This quality improvement study used preintervention and 30-day postintervention surveys to evaluate the use of the same ambient AI platform for clinical note documentation among ambulatory care physicians and advanced practice practitioners of 6 academic and community-based health care systems across the US. Clinicians were recruited by the health systems' digital health leaders; participation was voluntary. The study was conducted between February 1 and October 31, 2024.

EXPOSURE Use of an ambient AI scribe for 30 days.

MAIN OUTCOMES AND MEASURES The primary outcome was change in self-reported burnout, estimated using hierarchical logistic regression. Secondary outcomes of burnout evaluated were changes in note-related cognitive task load, focused attention on patients, patient understandability of notes, ability to add patients to the clinic schedule if urgently needed, and time spent documenting after hours. Outcome measures were linearly transformed to 10-point scales to ease interpretation and comparison. Differences between preintervention and postintervention scores were determined using paired *t* tests.

RESULTS Of the 451 clinicians enrolled, 272 completed the preintervention and postintervention surveys (60.3% completion rate), and 263 with direct patient care in ambulatory clinics (mean [SD] years in practice, 15.1 [9.3]; 141 female [53.6%]) were included in the analysis. The sample included 131 primary care practitioners (49.7%), 232 attending physicians (88.2%), and 168 academic faculty.

Key Points

Question What is the association of using ambient artificial intelligence (AI) scribes with clinician administrative burden, burnout, time documenting after hours, and time and attention for patients?

Findings This quality improvement study of 263 physicians and advance practice practitioners across 6 health care systems found that after 30 days with an ambient AI scribe, burnout among those working in ambulatory clinics decreased significantly from 51.9% to 38.8%. There were also significant improvements in the cognitive task load, time spent documenting after hours, focused attention on patients, and urgent access to care.

Meaning These findings suggest that AI may have promising applications to reduce administrative burdens for clinicians and allow more time for meaningful work and professional well-being.

 Invited Commentary

- **Multicenter, pre-post analysis**
(6 health systems; 263 clinicians, JAMA Network Open, 2025)
- **Findings:** After 30 days of using an ambient AI scribe, the share of clinicians reporting burnout fell from 51.9% → 38.8%
- Note-related cognitive task load = -2.64/10, after-hours documentation = -0.90 h/day, focused attention on patients +2.05/10

Early Studies on the Impact of AI Scribes

The screenshot shows the JAMIA journal website. At the top, the JAMIA logo is displayed with the text "A SCHOLARLY JOURNAL OF INFORMATICS IN HEALTH AND BIOMEDICINE". Below the logo, a navigation bar includes links for "Issues", "More Content ▾", "Submit ▾", "Purchase", "Alerts", "About ▾", and "Journal of the Ameri ▾". On the left side, there is a sidebar for "Volume 32, Issue 2 February 2025" featuring a colorful abstract graphic. The main content area displays an article titled "Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden" by Shreya J Shah, MD, et al. The article is marked as "FEATURED" and includes a "Get access ▾" button. Below the title, the authors' names are listed: Shreya J Shah, MD, Anna Devon-Sand, MPH, Stephen P Ma, MD, PhD, Yejin Jeong, BA, Trevor Crowell, BA, Margaret Smith, MBA, April S Liang, MD, Clarissa Delahaye, BAS, Caroline Hsia, MEng, Tait Shanafelt, MD, and others. A "Show more" link is also present. The article is described as being published in the "Journal of the American Medical Informatics Association, Volume 32, Issue 2, February 2025, Pages 375–380, <https://doi.org/10.1093/jamia/ocae295>". The "Published: 05 December 2024" and "Article history ▾" links are also visible. At the bottom of the article summary, there are links for "Cite", "Permissions", and "Share ▾".

- **Single-center, pre-post analysis**
(Stanford; 48 physicians, JAMIA, 2025)
- Paired surveys showed large reductions in task load and burnout after deploying an ambient AI scribe.

Early Studies on the Impact of AI Scribes

JAMIA
A SCHOLARLY JOURNAL OF INFORMATICS IN HEALTH AND BIOMEDICINE

Issues More Content ▾ Submit ▾ Purchase Alerts About ▾ Journal of the Ameri ▾

JOURNAL ARTICLE FEATURED

Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden [Get access ▾](#)

Shreya J Shah, MD ✉, Anna Devon-Sand, MPH, Stephen P Ma, MD, PhD, Yejin Jeong, BA, Trevor Crowell, BA, Margaret Smith, MBA, April S Liang, MD, Clarissa Delahaye, BAS, Caroline Hsia, MEng, Tait Shanafelt, MD ... Show more

Journal of the American Medical Informatics Association, Volume 32, Issue 2, February 2025, Pages 375–380, <https://doi.org/10.1093/jamia/ocae295>

Published: 05 December 2024 Article history ▾

“ Cite Permissions Share ▾

Article Contents

Abstract Background and significance

JAMA Network Open

Original Investigation | Health Policy

Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency

Matthew J. Duggan, MBA; Julietta Gervase, BA; Anna Schoenbaum, DNP, MS, RN-BC; William Hanson, MD; John T. Howell III, MD; Michael Sheinberg, MD; Kevin B. Johnson, MD, MS

Abstract

IMPORTANCE Timely evaluation of ambient scribing technology is warranted to assess whether this technology can lessen the burden of clinical documentation on clinicians.

OBJECTIVE To investigate the association of ambient scribing technology with efficiency, quality, and perceived burden of clinical documentation in the outpatient setting.

DESIGN, SETTING, AND PARTICIPANTS This prospective, single-group pre-post quality improvement study was conducted between April and June 2024 in the outpatient setting of an academic health system in Philadelphia, Pennsylvania. Participants included physicians, nurse practitioners, and physician assistants. Data were analyzed from July to August 2024.

EXPOSURE Access to an artificial intelligence-driven ambient scribing tool for outpatient notes.

Key Points

Question What is the association of ambient scribe technology with scribing efficiency and clinical documentation burden for clinicians in the outpatient setting?

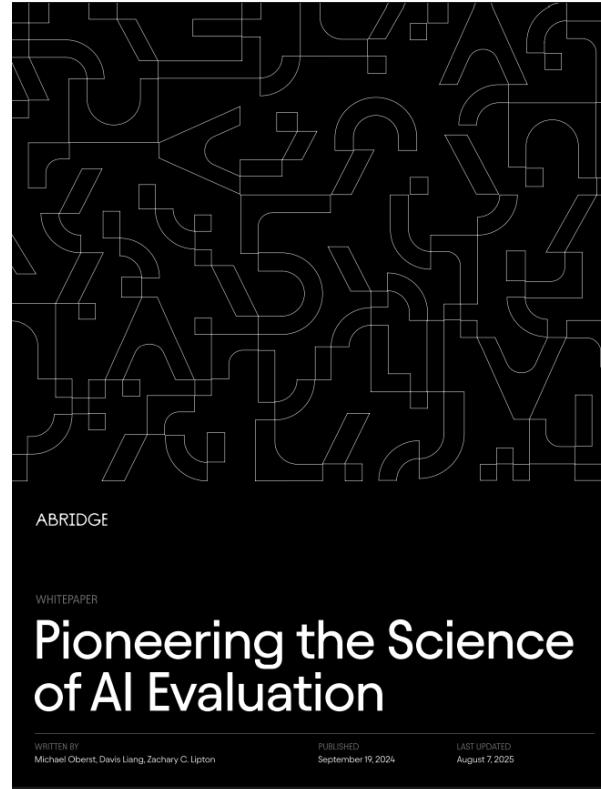
Findings In this quality improvement study with 46 participants, use of an ambient scribe tool was associated with greater clinician efficiency, lower mental burden of documentation, and greater sense of engagement with patients during outpatient appointments.

- **Single-center, pre-post analysis (Stanford; 48 physicians, JAMIA, 2025)**
- Paired surveys showed large reductions in task load and burnout after deploying an ambient AI scribe.
- **Single-center, pre-post (U. Penn; 46 clinicians, JAMA Network Open, 2025):** 20.4% less time in notes/visit, 30% less after-hours work, and lower documentation burden. Qualitative feedback = less mental overload and better engagement with patients.

Industry Practices for Evaluating Ambient AI Scribes

- There is evidence that ambient AI scribes reduce burnout. But how well do they work?
- Evaluating these tools is complicated and a new science of evaluation is emerging.

Recent white papers
by Abridge AI



Abridge AI Evaluation and Monitoring Approach

- A multi-staged approach with a mix of automatic and clinician-in-the-loop methods

Model development guided by: (1) automated metrics and (2) clinician spot-checks

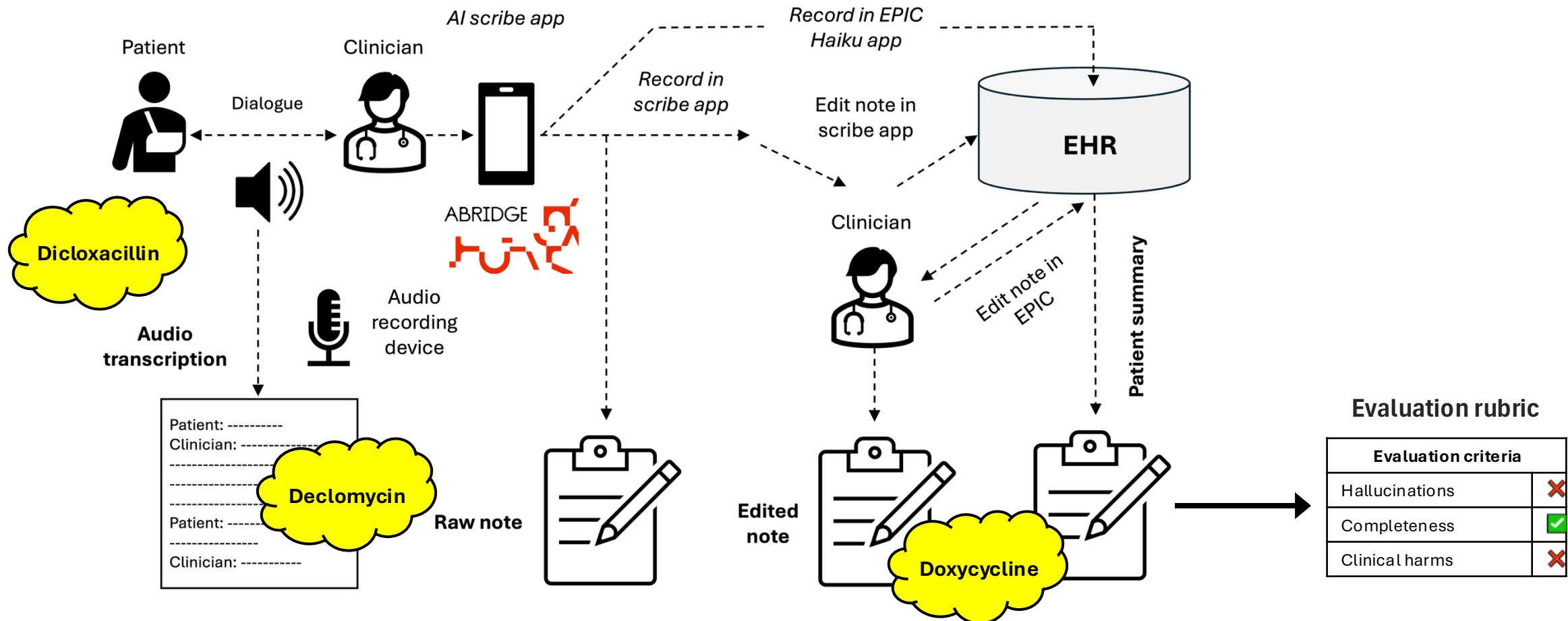
Validation via blinded head-to-head trials adjudicated by licensed clinicians

Staged release process

Ongoing post-deployment monitoring

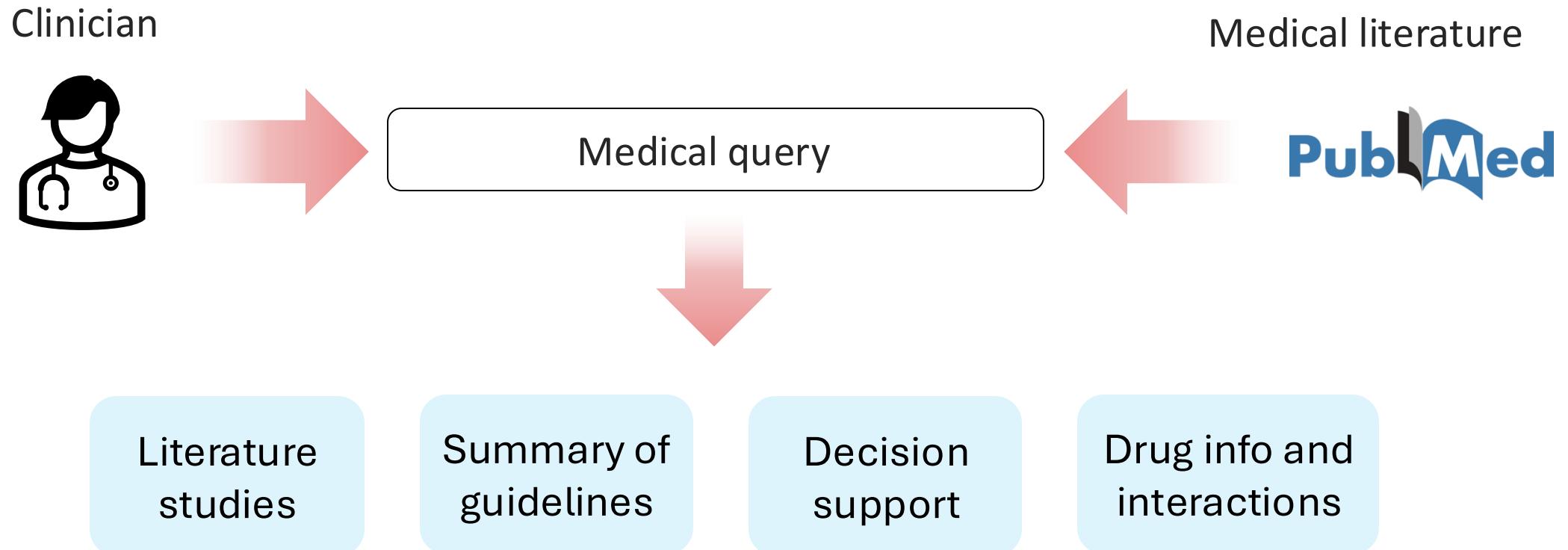
Real-World Deployment of AI Scribes at UCSF

- Examples: AI scribes to generate SOAP notes based on patient-clinician convo



Example 2: AI-based Medical Information Platforms

- A point-of-care digital system that collects, organizes, and delivers evidence-based medical knowledge (e.g., research findings, clinical guidelines, drug data, and expert recommendations) to support clinical decision-making, education, and research.

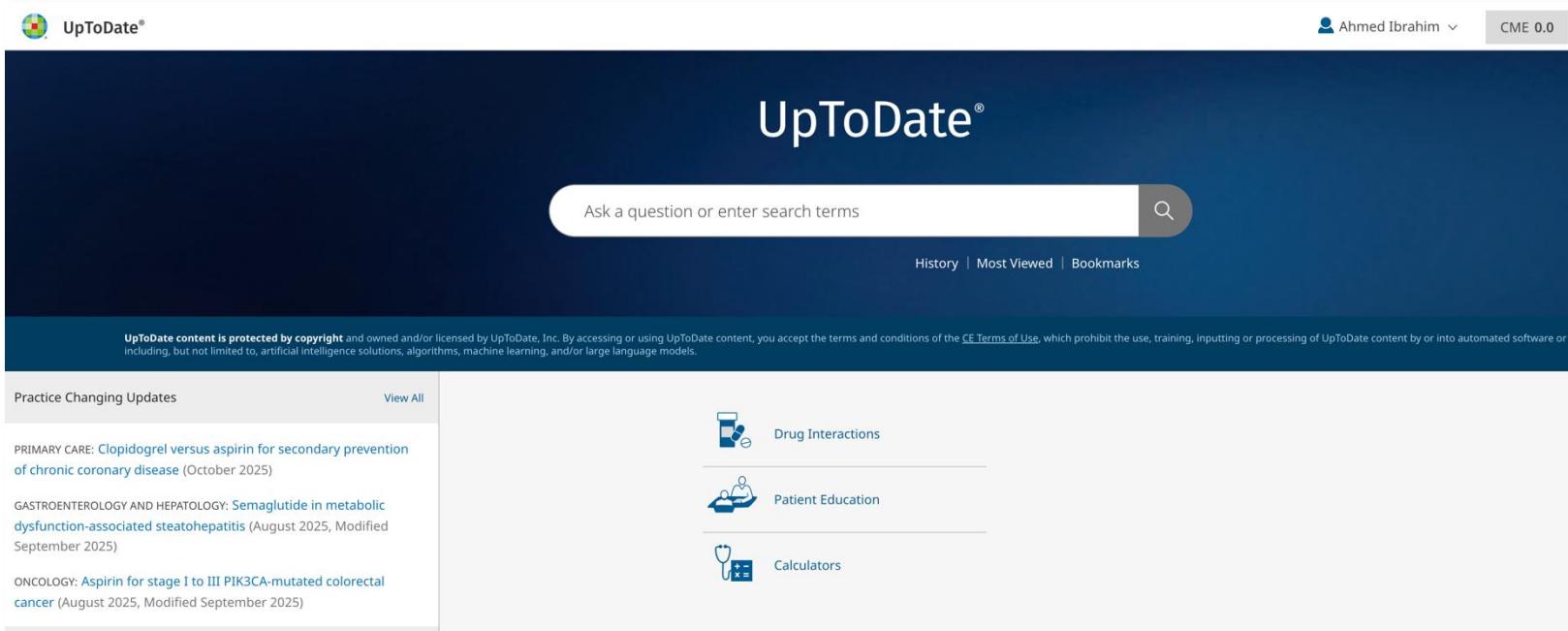


Medical Information Platforms in the Pre-ChatGPT Era

- Platforms were not generative! They did not write an answer in natural language. Instead, they provided curated data, reference links, and human-written summaries.

Medical Information Platforms in the Pre-ChatGPT Era

- Platforms were not generative! They did not write an answer in natural language. Instead, they provided curated data, reference links, and human-written summaries.



Medical Information Platforms in the Pre-ChatGPT Era

- Platforms were not generative! They did not write an answer in natural language. Instead, they provided curated data, reference links, and human-written summaries.

The screenshot shows the UpToDate website. At the top, there's a dark blue header with the "UpToDate®" logo, a search bar containing "Ask a question or enter search terms", and user account information for "Ahmed Ibrahim" with "CME 0.0". Below the header, the main content area has a dark blue background with the "UpToDate®" logo again. A search bar is centered, followed by navigation links for "History", "Most Viewed", and "Bookmarks". A small note at the bottom states: "UpToDate content is protected by copyright and owned and/or licensed by UpToDate, Inc. By accessing or using UpToDate content, you accept the terms and conditions of the CE Terms of Use, which prohibit the use, training, inputting or processing of UpToDate content by or into automated software or tools including, but not limited to, artificial intelligence solutions, algorithms, machine learning, and/or large language models." On the left sidebar, there are sections for "Practice Changing Updates" (with a "View All" link), "PRIMARY CARE: Clopidogrel versus aspirin for secondary prevention of chronic coronary disease (October 2025)", "GASTROENTEROLOGY AND HEPATOLOGY: Semaglutide in metabolic dysfunction-associated steatohepatitis (August 2025, Modified September 2025)", and "ONCOLOGY: Aspirin for stage I to III PIK3CA-mutated colorectal cancer (August 2025, Modified September 2025)". On the right side, there are three icons with labels: "Drug Interactions", "Patient Education", and "Calculators".

 Wolters Kluwer

- 3 million users
- 190 countries
- > 37K institutions
- 90% of US institutions

Medical Information Platforms in the Pre-ChatGPT Era

- Platforms were not generative! They did not write an answer in natural language. Instead, they provided curated data, reference links, and human-written summaries.

The screenshot shows the UpToDate website. At the top, there's a dark blue header with the "UpToDate®" logo on the left and a user profile with the name "Ahmed Ibrahim" and "CME 0.0" on the right. Below the header is a search bar containing the placeholder text "Ask a question or enter search terms" with a magnifying glass icon. A red rectangular box highlights this search bar area. Underneath the search bar are three small navigation links: "History", "Most Viewed", and "Bookmarks". The main content area has a white background. On the left, there's a sidebar with sections for "Practice Changing Updates" (with a "View All" link), "PRIMARY CARE: Clopidogrel versus aspirin for secondary prevention of chronic coronary disease (October 2025)", "GASTROENTEROLOGY AND HEPATOLOGY: Semaglutide in metabolic dysfunction-associated steatohepatitis (August 2025, Modified September 2025)", and "ONCOLOGY: Aspirin for stage I to III PIK3CA-mutated colorectal cancer (August 2025, Modified September 2025)". To the right of the sidebar, there are three icons with labels: "Drug Interactions" (pill and dropper icon), "Patient Education" (person icon), and "Calculators" (calculator icon).

 Wolters Kluwer

- 3 million users
- 190 countries
- > 37K institutions
- 90% of US institutions

Medical Information Platforms in the Pre-ChatGPT Era

- Platforms were not generative! They did not write an answer in natural language. Instead, they provided curated data, reference links, and human-written summaries.

The screenshot shows the UpToDate website. At the top, there's a dark header with the UpToDate logo, a user profile for 'Ahmed Ibrahim', and a CME credit indicator of 'CME 0.0'. Below the header is a search bar containing the placeholder 'Ask a question or enter search terms' with a magnifying glass icon. A red box highlights this search bar. Underneath the search bar are three navigation links: 'History', 'Most Viewed', and 'Bookmarks'. The main content area has a dark blue sidebar on the left with sections for 'Practice Changing Updates' (listing 'Clopidogrel versus aspirin for secondary prevention of chronic coronary disease (October 2025)', 'Semaglutide in metabolic dysfunction-associated steatohepatitis (August 2025, Modified September 2025)', and 'Aspirin for stage I to III PIK3CA-mutated colorectal cancer (August 2025, Modified September 2025)'), and 'View All'. The main content area features three icons with labels: 'Drug Interactions' (with a pill and dropper icon), 'Patient Education' (with a person icon), and 'Calculators' (with a calculator icon). A red box highlights the 'Calculators' icon.

 Wolters Kluwer

- 3 million users
- 190 countries
- > 37K institutions
- 90% of US institutions

Medical Information Platforms in the Pre-ChatGPT Era

- Pre-AI platforms were like search engines!

Advantages

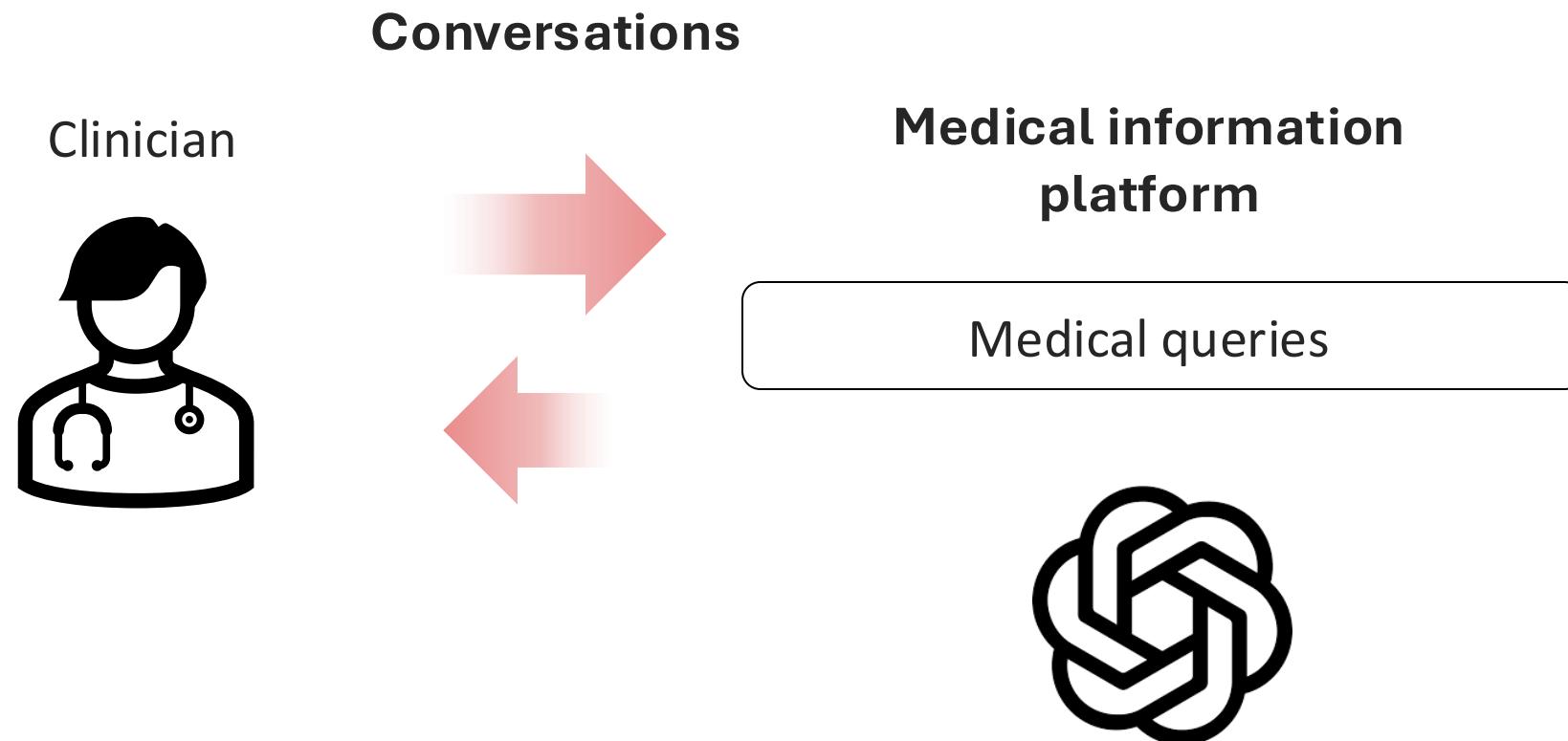
- ❑ Human-authored, peer-reviewed topic pages.
- ❑ Clear recommendations graded by evidence level.
- ❑ Integration with EHRs (Epic, Cerner, etc.).

Disadvantages

- ❑ Subscription cost (thousands/year for institutions).
- ❑ Not interactive; no conversational interface.
- ❑ May lag a few months behind the latest research.

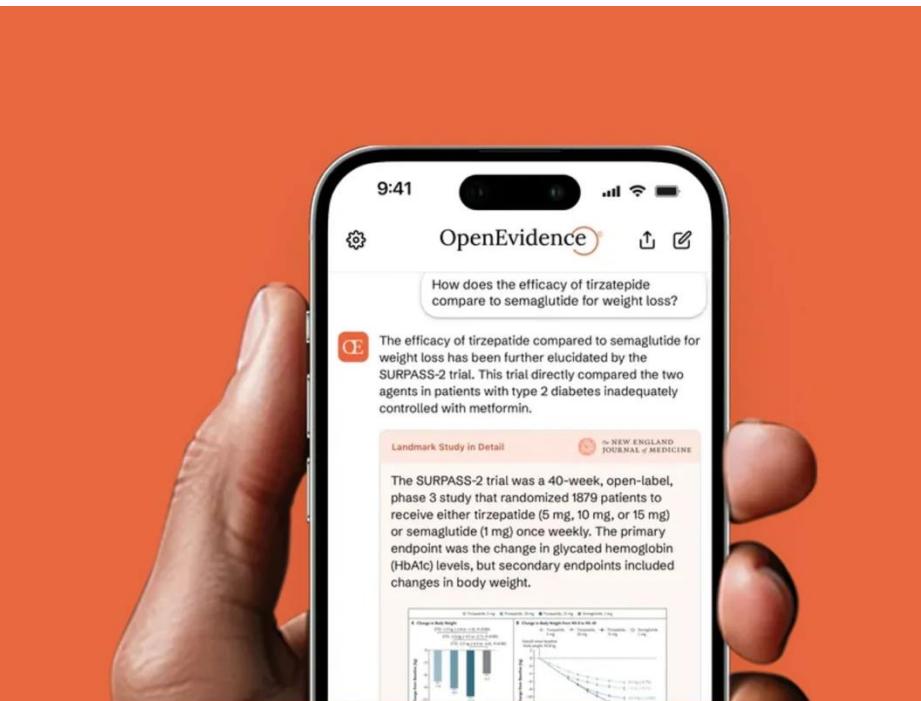
AI-based Medical Information Platforms

- Medical information platforms as chatbots!



AI-based Medical Information Platforms

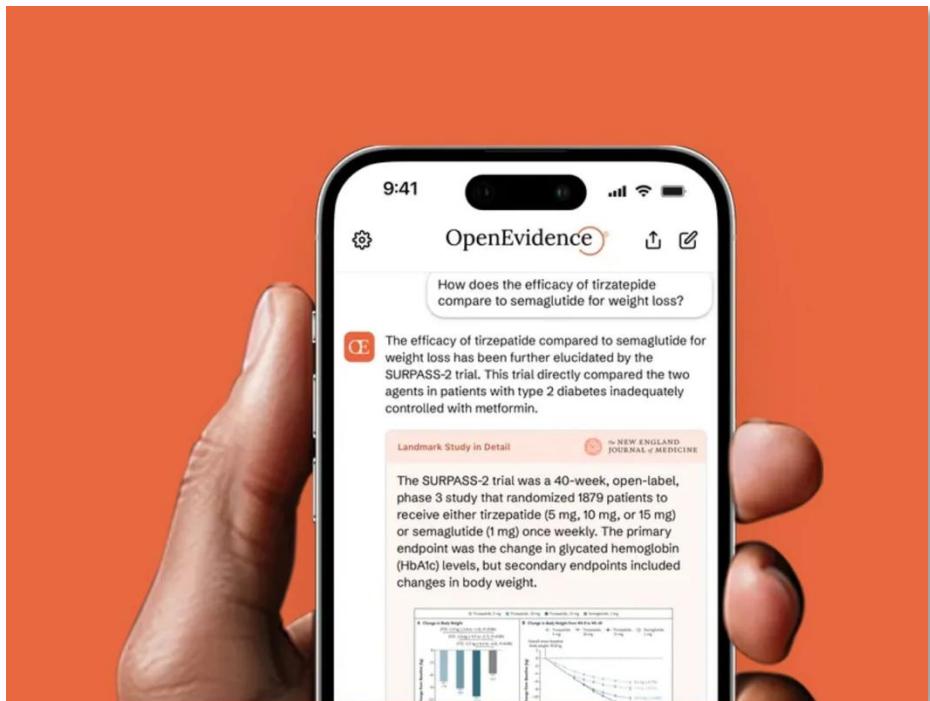
- OpenEvidence is a U.S.-based AI-powered medical search/clinical decision-support platform aimed primarily at healthcare professionals.
- Specialized ChatGPT for medical information!



OpenEvidence®

AI-based Medical Information Platforms

- OpenEvidence is a U.S.-based AI-powered medical search/clinical decision-support platform aimed primarily at healthcare professionals.
- Specialized ChatGPT for medical information!

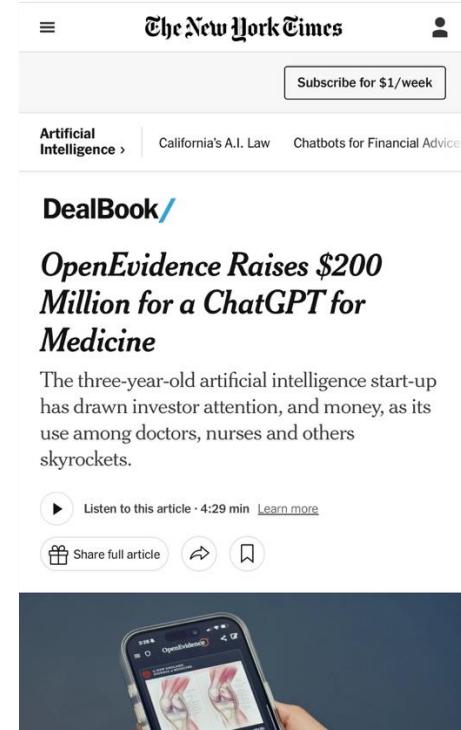


- ❑ Launched around 2022 / 2023
- ❑ Mission: “To organize and expand the world’s collective medical knowledge.”
- ❑ **Valuation of \$3.5B!**

Adoption of OpenEvidence in the US

- According to the company: More than 40% of U.S. physicians log in daily” to the platform, and usage across 10,000+ hospitals and medical centers
- The company recently raised significant capital of a \$210 M Series B round in 2025 at a \$3.5 billion valuation.

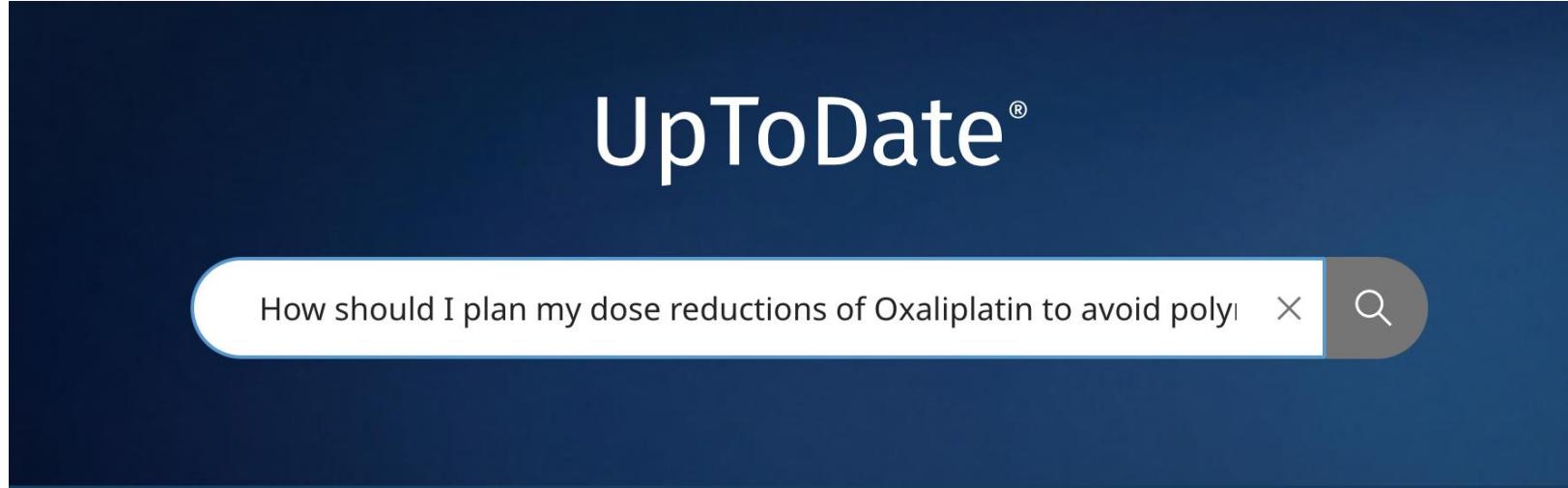
Why did this tool become widely adopted so quickly?



The screenshot shows a news article from The New York Times' DealBook section. The headline reads "OpenEvidence Raises \$200 Million for a ChatGPT for Medicine". The article discusses the three-year-old AI startup's rapid adoption by doctors and nurses. It includes a play button for a 4:29-minute audio recording, sharing options, and a small image of a smartphone displaying the OpenEvidence app.

AI-based vs. Search-based Medical Information Platforms

- Asking UpToDate about toxicity of chemotherapy treatment



AI-based vs. Search-based Medical Information Platforms

The screenshot shows the UpToDate medical information platform interface. At the top, there is a search bar with the query "How should I plan my dose reductions of Oxaliplatin to avoid Polyneuropathy (Peripheral neuropathy)". Below the search bar, there are navigation tabs: Back, All (which is selected), Adult, Pediatric, Patient, and Graphics.

The main content area displays results for the search query. A section titled "AI Suggested Results BETA" is shown, with a note that these are excerpts from UpToDate and not generated by AI. It includes a "Read More" link to view the full passage.

The results include a bulleted list under "Prevention":

- When appropriate in the clinical setting, interspersing a non-oxaliplatin-containing "maintenance" chemotherapy regimen with the oxaliplatin regimen can forestall the development of neuropathy and is reasonable. ...

A "Show More Results" button is located next to this list.

Below this, there are three expandable sections:

- Overview of neurologic complications of platinum-based chemotherapy**: Discusses neuropathy associated with oxaliplatin therapy, mentioning dose reduction and reintroducing therapy.
- Prevention and treatment of chemotherapy-induced peripheral neuropathy**: Discusses the prevention and treatment of peripheral neuropathy, mentioning gabapentin and venlafaxine.
- Initial systemic therapy for metastatic colorectal cancer**: Discusses neuropathy as a side effect of oxaliplatin therapy.

AI-based vs. Search-based Medical Information Platforms

The screenshot shows a medical search results page from UpToDate. At the top, a search bar contains the query "How should I plan my dose reductions of Oxaliplatin to avoid Polyneuropathy (Peripheral neuropathy)". Below the search bar, a red curved line highlights the "AI Suggested Results" section. This section is labeled "BETA" and includes a feedback button "Are suggested results helpful?". A note states: "These passages are excerpts from UpToDate. They are not generated by AI. Click 'Read More' below to view the passage within the source content." The main content of the AI summary discusses prevention and treatment of neuropathy associated with oxaliplatin therapy. A red arrow points from the text "These passages are excerpts from UpToDate" towards the "AI summary" label.

Showing results for **how should i plan my dose reductions of oxaliplatin to avoid Polyneuropathy (Peripheral neuropathy)**

AI Suggested Results BETA Are suggested results helpful?

These passages are excerpts from UpToDate. They are not generated by AI. Click 'Read More' below to view the passage within the source content.

... Multiple approaches have been used to prevent or minimize the cumulative neurotoxicity associated with **oxaliplatin** therapy. These include dose reduction, infusion, supplemental intravenous calcium and magnesium infusions, and various pharmacologic agents.

• **Prevention**

- When appropriate in the clinical setting, interspersing a non-oxaliplatin-containing "maintenance" chemotherapy regimen with the **oxaliplatin** ... forestall the development of neuropathy and is reasonable. ...

Show More Results

Overview of neurologic complications of platinum-based chemotherapy

... and **neuropathy** increases with a longer **planned** treatment duration . As an example, In a prospective trial in which patients received **oxaliplatin** 85 mg/m² per cycle... neurotoxicity associated with **oxaliplatin** therapy. These include **dose reduction**, interrupting and reintroducing **oxaliplatin** therapy,...

Clinical manifestations

Mechanism

Summary and recommendations

Prevention and treatment of chemotherapy-induced peripheral neuropathy

... intensities of **oxaliplatin**, with similar **oxaliplatin dose reductions** and delays in the two groups. These authors concluded that gabapentin was not helpful for preventing/treating **oxaliplatin**-associated... treated with **oxaliplatin** . Although there was a trend toward benefit for venlafaxine when evaluated by the **oxaliplatin**-specific **neuropathy** scale and by some acute **neuropathy** measures (throat...

Oxaliplatin

Summary and recommendations

Initial systemic therapy for metastatic colorectal cancer

... **Neuropathy** – The dose-limiting side effect of **oxaliplatin** is a cumulative predominantly sensory **neuropathy**, which may require **drug discontinuation** despite

AI summary

AI-based vs. Search-based Medical Information Platforms

The screenshot shows a search result for "How should I plan my dose reductions of Oxaliplatin to avoid Polyneuropathy (Peripheral neuropathy)". The results are categorized under "AI Suggested Results BETA". A red oval highlights the AI summary section, which contains a brief paragraph about prevention and a bulleted list. Red arrows point from the text "Expert summaries" to the clinical manifestations, mechanism, and summary sections of the results.

UpToDate®
How should I plan my dose reductions of Oxaliplatin to avoid Polyneuropathy (Peripheral neuropathy)

AI Suggested Results BETA Are suggested results helpful?

These passages are excerpts from UpToDate. They are not generated by AI. Click 'Read More' below to view the passage within the source content.

... Multiple approaches have been used to prevent or minimize the cumulative neurotoxicity associated with **oxaliplatin** therapy. These include dose reduction, infusion, supplemental intravenous calcium and magnesium infusions, and various pharmacologic agents.

• Prevention

- When appropriate in the clinical setting, interspersing a non-oxaliplatin-containing "maintenance" chemotherapy regimen with the **oxaliplatin**... forestall the development of neuropathy and is reasonable. ...

Show More Results

Overview of neurologic complications of platinum-based chemotherapy

... and **neuropathy** increases with a longer **planned** treatment duration. As an example, in a prospective trial in which patients received **oxaliplatin** 85 mg/m² per cycle... neurotoxicity associated with **oxaliplatin** therapy. These include **dose reduction**, interrupting and reintroducing **oxaliplatin** therapy,...

Clinical manifestations
Mechanism
Summary and recommendations

Prevention and treatment of chemotherapy-induced peripheral neuropathy

... intensities of **oxaliplatin**, with similar **oxaliplatin dose reductions** and delays in the two groups. These authors concluded that gabapentin was not helpful for preventing/treating **oxaliplatin**-associated... treated with **oxaliplatin**. Although there was a trend toward benefit for venlafaxine when evaluated by the **oxaliplatin**-specific **neuropathy** scale and by some acute **neuropathy** measures (throat...

Oxaliplatin
Summary and recommendations

Initial systemic therapy for metastatic colorectal cancer

... **Neuropathy** – The dose-limiting side effect of **oxaliplatin** is a cumulative predominantly sensory **neuropathy**, which may require **drug discontinuation** despite

Expert
summaries

AI summary

AI-based vs. Search-based Medical Information Platforms

The screenshot shows a medical search interface for "UpToDate®". The search bar at the top contains the query "How should I plan my dose reductions of Oxaliplatin to avoid Polyneuropathy (Peripheral neuropathy)". Below the search bar, there are navigation tabs: Back, All (which is selected), Adult, Pediatric, Patient, and Graphics.

The main content area displays search results for the query. At the top of the results, there is a section titled "AI Suggested Results BETA" with a "Feedback" button. A note states: "These passages are excerpts from UpToDate. They are not generated by AI. Click 'Read More' below to view the passage within the source content."

The results include a list of prevention strategies:

- Prevention
 - When appropriate in the clinical setting, interspersing a non-oxaliplatin-containing "maintenance" chemotherapy regimen with the oxaliplatin regimen can forestall the development of neuropathy and is reasonable. ...

A "Show More Results" button is located next to the prevention list. Below this, a red box highlights a section titled "Overview of neurologic complications of platinum-based chemotherapy". This section includes a brief summary and a "Summary and recommendations" link, which is also highlighted with a red box.

Further down the page, another section is titled "Prevention and treatment of chemotherapy-induced peripheral neuropathy". It contains a detailed paragraph about the prevention and treatment of neuropathy associated with oxaliplatin. Below this section, there is a "Oxaliplatin" link and a "Summary and recommendations" link.

The bottom of the page features a section titled "Initial systemic therapy for metastatic colorectal cancer" with a brief summary.

AI-based vs. Search-based Medical Information Platforms

Overview of neurologic complications of platinum-based chemotherapy

AUTHOR: Eudocia Quant Lee, MD, MPH
SECTION EDITORS: Reed E Drews, MD, Patrick Y Wen, MD
DEPUTY EDITOR: Sadhna R Vora, MD

All topics are updated as new evidence becomes available and our peer review process is complete.
Literature review current through: Sep 2025.
This topic last updated: Aug 28, 2025.

Contributor Disclosures

INTRODUCTION

Neurologic complications of anticancer therapy may result from direct toxic effects on the nervous system or indirectly from drug-induced metabolic derangements or cerebrovascular disorders. A wide range of neurologic complications are associated with antineoplastic drug treatment ([table 1](#)) [1-4].

Their recognition is important because of potential confusion with metastatic disease, paraneoplastic syndromes, or comorbid neurologic disorders that do not require dose reduction or discontinuation. If the neurologic disorder is caused by the chemotherapy, discontinuation of the offending agent may prevent irreversible injury.

UpToDate
experts

AI-based vs. Search-based Medical Information Platforms

- Asking OpenEvidence about toxicity of chemotherapy treatment

The screenshot shows the OpenEvidence AI interface. At the top, a banner states: "OpenEvidence has signed content agreements with [JAMA](#) and [The New England Journal of Medicine](#)". Below the banner, the OpenEvidence logo is on the left, and a navigation bar includes "Log In", "Sign Up", and "New Visit". A three-line menu icon is also present. A central message box contains the text: "This conversation has been affirmatively and intentionally made public by the user who created it." Below this, a large orange letter "C" is followed by the query: "How should I plan my dose reductions of [Oxaliplatin](#) to avoid polyneuropathy?". At the bottom of the message box, there is a "Finished thinking" button with a downward arrow.

AI-based vs. Search-based Medical Information Platforms

- Asking OpenEvidence about toxicity of chemotherapy treatment

The screenshot shows the OpenEvidence AI interface. At the top, a banner states: "OpenEvidence has signed content agreements with [JAMA](#) and [The New England Journal of Medicine](#)". Below the banner, the OpenEvidence logo is on the left, and "Log In", "Sign Up", and "New Visit" buttons are on the right. A navigation menu icon is also present. A message box contains the text: "This conversation has been affirmatively and intentionally made public by the user who created it." A large input field shows a question: "How should I plan my dose reductions of [Oxaliplatin](#) to avoid polyneuropathy?". Below the input field, a button labeled "Finished thinking" is highlighted with a red box and a red arrow pointing to it from the bottom left. The text "Reasoning model" is written in red at the bottom left of the image.

OpenEvidence has signed content agreements with [JAMA](#) and [The New England Journal of Medicine](#).

OpenEvidence®

This conversation has been affirmatively and intentionally made public by the user who created it.

How should I plan my dose reductions of [Oxaliplatin](#) to avoid polyneuropathy?

_finished thinking

Reasoning model

AI-based vs. Search-based Medical Information Platforms

- Asking OpenEvidence about toxicity of chemotherapy treatment

The screenshot shows the OpenEvidence AI interface. At the top, a banner states: "OpenEvidence has signed content agreements with [JAMA](#) and [The New England Journal of Medicine](#)". Below the banner, the OpenEvidence logo is on the left, and "Log In", "Sign Up", and "New Visit" buttons are on the right. A message box says: "This conversation has been affirmatively and intentionally made public by the user who created it." A large input field contains the query: "How should I plan my dose reductions of Oxaliplatin to avoid polyneuropathy". Below the input field, a button labeled "Finished thinking" is highlighted with a red box and a red arrow pointing from the text "Reasoning chain" below it. To the right, a sidebar shows the AI's thought process:

- Finished thinking
- Analyzing query
- Searching published medical literature, guidelines, FDA, CDC, and more
- Synthesizing relevant information

Reasoning chain

OpenEvidence has signed content agreements with [JAMA](#) and [The New England Journal of Medicine](#).

Log In Sign Up New Visit

This conversation has been affirmatively and intentionally made public by the user who created it.

How should I plan my dose reductions of Oxaliplatin to avoid polyneuropathy

Finished thinking

Analyzing query

Searching published medical literature, guidelines, FDA, CDC, and more

Synthesizing relevant information

AI-based vs. Search-based Medical Information Platforms

Dose reductions for oxaliplatin should be considered for patients with persistent grade 2 neuropathy (reduce to 75 mg/m² for adjuvant or 65 mg/m² for advanced disease) and discontinuation for grade 3 or higher neuropathy, as per FDA labeling and ASCO guidelines.^[1-3] The American Society of Clinical Oncology further recommends that clinicians assess the appropriateness of dose delay, reduction, or discontinuation in patients who develop intolerable neuropathy or functional impairment.^[2]

Recent evidence demonstrates that lean body mass-based dosing ($\leq 3.09 \text{ mg/kg LBM}$) can significantly reduce the incidence and severity of oxaliplatin-induced peripheral neuropathy without compromising efficacy, especially in patients with reduced LBM.^[4] This approach may be considered in select patients, particularly those with sarcopenia or low LBM.

Shortening the duration of oxaliplatin therapy (e.g., 3 months for lower-risk stage III colon cancer) may decrease neuropathy risk while maintaining similar disease-free survival, but 6 months remains standard for higher-risk patients; shared decision-making is recommended.^[5-6] The risk of neuropathy increases with cumulative dose and duration.

Regular monitoring of neuropathy symptoms—using both clinician assessment and patient-reported outcomes—is essential. Prompt dose modification or discontinuation is critical to minimize long-term nerve damage, with rapid increases in patient-reported severity indicating a need for intervention.^{[2][7]} Embedding standardized dose modification protocols in electronic health records can facilitate timely and evidence-based adjustments.^[3]

Adjunctive agents (e.g., vitamin B6, celecoxib) have not shown efficacy in preventing oxaliplatin-induced neuropathy.^[6] Optimal management relies on individualized risk assessment, vigilant monitoring, and adherence to guideline-based dose modification strategies.

Would you like me to summarize the latest clinical trial evidence comparing lean body mass-based dosing versus standard body surface area dosing for oxaliplatin, specifically regarding neuropathy outcomes and oncologic efficacy?

References to published literature

AI-based vs. Search-based Medical Information Platforms

Dose reductions for oxaliplatin should be considered for patients with persistent grade 2 neuropathy (reduce to 75 mg/m² for adjuvant or 65 mg/m² for advanced disease) and discontinuation for grade 3 or higher neuropathy, as per FDA labeling and ASCO guidelines.^[1-3] The American Society of Clinical Oncology further recommends that clinicians assess the appropriateness of dose delay, reduction, or discontinuation in patients who develop intolerable neuropathy or functional impairment.^[2]

Recent evidence demonstrates that lean body mass-based dosing ($\leq 3.09 \text{ mg/kg LBM}$) can significantly reduce the incidence and severity of oxaliplatin-induced peripheral neuropathy without compromising efficacy, especially in patients with reduced LBM.^[4] This approach is considered in select patients, particularly those with sarcopenia or low LBM.

Shortening the duration of oxaliplatin therapy (e.g., 3 months for lower-risk stage III colon cancer) can decrease neuropathy risk while maintaining similar disease-free survival, but 6 months is the standard for higher-risk patients; shared decision-making is recommended.^[5-6] The risk of neuropathy increases with cumulative dose and duration.

Regular monitoring of neuropathy symptoms—using both clinician assessment and patient outcomes—is essential. **Prompt dose modification or discontinuation is critical to minimize term nerve damage, with rapid increases in patient-reported severity indicating a need for intervention.**^{[2][7]} Embedding standardized dose modification protocols in electronic health records can facilitate timely and evidence-based adjustments.^[3]

Adjunctive agents (e.g., vitamin B6, celecoxib) have not shown efficacy in preventing oxaliplatin-induced neuropathy.^[6] Optimal management relies on individualized risk assessment, vigilant monitoring, and adherence to guideline-based dose modification strategies.

Would you like me to summarize the latest clinical trial evidence comparing lean body mass-based dosing versus standard body surface area dosing for oxaliplatin, specifically regarding neuropathy outcomes and oncologic efficacy?

References

1. Oxaliplatin. FDA Drug Label. Food and Drug Administration Updated date: 2024-09-25
2. Prevention and Management of Chemotherapy-Induced Peripheral Neuropathy in Survivors of Adult Cancers: ASCO Guideline Update. Loprinzi CL, Lacchetti C, Bleeker J, et al. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology. 2020;38(28):3325-3348. doi:10.1200/JCO.20.01399. Practice Guideline ★ Leading Journal
3. Standardization of toxicity related dose modification in oncology electronic health record (EHR). Sanyal A, Wellner D, Thomas J, et al. Journal of Clinical Oncology. 2022;40(Suppl 16):e18616. doi:10.1200/JCO.2022.40.16_suppl.e18616. Practice Guideline ★ Leading Journal
4. Impact of Lean Body Mass-Based Oxaliplatin Dose Calculation on Neurotoxicity in Adjuvant Treatment of Stage III Colon Cancer: Results of the Phase II Randomized LEANOX Trial. Assenat E, Ben Abdelghani M, Gourgou S, et al. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology. 2025;JCO2402754. doi:10.1200/JCO.24-02754. ★ Leading Journal ★ New Research
5. Duration of Oxaliplatin-Containing Adjuvant Therapy for Stage III Colon Cancer: ASCO Clinical Practice Guideline. Lieu C, Kennedy EB, Bergstrand E, et al. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology. 2010;27(26):1447. doi:10.1200/JCO.2009.200381

AI-based vs. Search-based Medical Information Platforms

Dose reductions for oxaliplatin should be considered for patients with persistent grade 2 neuropathy (reduce to 75 mg/m² for adjuvant or 65 mg/m² for advanced disease) and discontinuation for grade 3 or higher neuropathy, as per FDA labeling and ASCO guidelines.^[1-3] The American Society of Clinical Oncology further recommends that clinicians assess the appropriateness of dose delay, reduction, or discontinuation in patients who develop intolerable neuropathy or functional impairment.^[2]

Recent evidence demonstrates that lean body mass-based dosing ($\leq 3.09 \text{ mg/kg LBM}$) can

≡ Follow-Up Questions

What patient populations are most susceptible to Oxaliplatin-induced polyneuropathy? >

Which alternative chemotherapy agents have lower risk of neurotoxicity in this setting? >

What monitoring strategies are recommended for early detection of neurotoxicity during treatment? >

Managing neuropathy... Optimal management relies on individualized risk assessment, symptom monitoring, and adherence to guideline-based dose modification strategies.

Would you like me to summarize the latest clinical trial evidence comparing lean body mass-based dosing versus standard body surface area dosing for oxaliplatin, specifically regarding neuropathy outcomes and oncologic efficacy?

AI-based vs. Search-based Medical Information Platforms

- With a ChatBot approach → More room for conversation and contextualization!

My patient is a 67-year old mail who is on his third cycle of full dose Oxaliplatin. On the third cycle, he started exhibiting signs of neuropathy in his hands and fingers. What evidence best applies to my patient?

✓ Analyzing query

AI-based vs. Search-based Medical Information Platforms

- With a ChatBot approach → More room for conversation and contextualization!

In summary, for a 67-year-old male on his third cycle of full-dose oxaliplatin who is developing neuropathy in his hands and fingers, **the best evidence supports prompt dose reduction for persistent grade 2 symptoms, discontinuation for grade 3 or higher, regular assessment using validated tools, consideration of lean body mass-based dosing in select patients, and duloxetine for established painful neuropathy.** Multidisciplinary referral is indicated for disabling or refractory symptoms.

Would you like me to summarize the latest clinical trial data comparing different oxaliplatin dose modification protocols—such as stop-and-go strategies or intermittent dosing—to standard continuous dosing, specifically regarding their impact on neuropathy incidence and oncologic outcomes? ■

AI-based vs. Search-based Medical Information Platforms

Search-based

- Human-authored, more accountability.
- More predictable behavior.
- Not interactive or conversational
- More expensive and slow to update evidence
- Usually, paywalled with a subscription fee

AI-based

- Saves time by retrieving most relevant summary.
- Interactive, conversational and contextual.
- Less human mediation: cheaper and faster!
- OpenEvidence is free!
- Less predictable behavior, less accountability.

In both examples, the generative AI tool performs an under-specified task!

Under-Specification Hinders Benchmarking

- We don't have an “ImageNet” for medical Generative AI (yet)¹

NEJM
AI

EDITORIAL

It's Time to Bench the Medical Exam Benchmark

Inioluwa Deborah Raji, B.A.,¹ Roxana Daneshjou, M.D., Ph.D.,^{2,3} and Emily Alsentzer, Ph.D.²

<https://doi.org/10.1038/s41591-025-03637-3>

published: January 23, 2025

[Check for updates](#)

A benchmarking crisis in biomedical machine learning

By Faisal Mahmood

A lack of standardized benchmarks is hindering progress and patient benefits

Machine learning (ML) in biomedicine is facing a benchmarking crisis. Foundation models, generative and agentic tools are poised to reshape clinical decision support and drug discovery pipelines, yet the search for generalized benchmarks or performance metrics is a transparent validation protocols threatens to derail progress¹. Whereas computer vision and language modelling rely on widely accepted gold standard datasets, biomedicine often depends on proprietary data, institutional-specific processing, and heterogeneous evaluation metrics. As a result, performance gains risk being tethered to narrow contexts rather than reflecting general scientific and translational impact.

A central complicating factor is the nature of biomedical data itself. Unlike large-scale publicly available datasets such as ImageNet or GPT-3, ML in Life Sciences, bio-medical training data frequently remain behind institutional or corporate firewalls. In response, a growing practice involves using proprietary data for model development while evaluating performance on limited but publicly available datasets, such as ImageNet or COCO datasets. For instance, computational pathology models may train on private clinical images but benchmark on open-access reference sets, ensuring that performance metrics remain transparent. This approach, however, introduces trade-offs. Public datasets can be used to validate a model's generalizability and sensitivity, limiting their utility as robust evaluation anchors and possibly introducing partiality toward narrower clinical contexts. As models evolve into multimodal systems with expansive capacity, traditional benchmarking frameworks face further challenges. In recent times, we have seen a shift from domain-specific models to more open-ended models capable of general representation

learning. The open-ended nature of emerging models in biomedicine demands new evaluation protocols that capture their performance across a broad range of biomedical tasks. Moreover, as these systems must be tested against a range of unanticipated capabilities to accomplish tasks – the criteria for benchmarking must extend beyond accuracy and consistency to include safety and ethical considerations. These challenges highlight the urgent need for flexible evaluation frameworks that can keep pace with the transformative potential of next-generation AI in biomedicine.

Remediying this crisis requires concrete, stepwise solutions. One possible strategy is to organize consortium-led initiatives that integrate and curate reference datasets, establish criteria for dataset transparency, and mandate reporting of key performance parameters to reduce the variability in how different models are evaluated. These efforts must be matched by frameworks that link model performance to clinically or biologically meaningful outcomes rather than relying solely on accuracy.

Implementing standardized benchmarks entails navigating practical and economic barriers. Large-scale open-data projects demand financial and human resources to annotate, maintain, and update datasets. They also raise privacy, consent, and intellectual property concerns. To mitigate these challenges, benchmark review boards, patient advocacy groups, and regulatory agencies early in the process.

Collaboration with these stakeholders will ensure that benchmark development respects data protection laws, ethical considerations, and public trust. Sustainable funding models, such as those used by open-source subscription-based platforms or data access marketplaces, may be necessary to help maintain such large collaborative infrastructures. Striking a balance between proprietary and publicly accessible datasets is also important. Organizations that oversee benchmark development should consider the trade-off between proprietary training sets, changing disease indications and other variability in

patient populations. This practice, coupled with incentives for data diversification, could guide developers to expand their training distributions and ensure that models remain generalizable.

Beyond technical and data-sharing challenges, addressing the ethical, regulatory and policy dimensions is crucial. To facilitate clinical translation and streamline future approvals, ML developers should involve regulatory agencies in shaping benchmark criteria and oversight during development.

Curiously, given that research-versed in both computational and biomedical sciences will be key, Cross-disciplinary education programs can train researchers to evaluate algorithms critically and appreciate the complexities of biomedical workflows. These programs should also teach the regulatory and methodological fluency required to advance field-wide benchmarking efforts. By preparing the next generation of scientists and clinicians to prioritize fairness and reproducibility, future ML-driven biomedical innovations will be held to high standards of transparency and social benefit.

Faisal Mahmood Mass General Brigham, Harvard Medical School, and the Broad Institute of Harvard and MIT, Cambridge, MA, USA.
e-mail: fmahmood@fas.harvard.edu

Published online: 8 April 2025

1. Moos, M. et al. *Nature* **495**, 269–281 (2023).
2. Alsentzer, E. et al. *Am. J. Hum. Genet.* **107**, 773–784 (2022).
3. Zhou, Y. et al. *Nature* **522**, 156–163 (2023).

Competing interests

F.M. is an inventor on several patents related to computational pathology, and a scientific advisor for Modella AI and Danaher.

cal Licensing Examination, legal models (LLMs) in health care, and evidence of progress and patient benefit. We argue that these benchmarks have utility.

Career are often marked by milestones (MCAT), the United States Board exams. However, care for patients begins in the complexities of real-world decisions, and the demands of every-

it the most important milestones in medical training. Most popular medical benchmarks, such as the multiple-choice question bank (MCMU), are derived from clinical vignettes and clinical presentations featured in the literature.

These benchmarks are being refined to reflect the latest in medical knowledge and practice. Google initially developed its own benchmarks, called MLLM, over other models by building on the work of the MLLM team. The MLLM team has developed a wide range of benchmarks, including text-to-text, text-to-image, and image-to-image benchmarks. These benchmarks are used to evaluate the performance of different models on specific tasks, such as image captioning, image segmentation, and image classification. The MLLM team has also developed a set of benchmarks for medical imaging, such as image segmentation and image classification. These benchmarks are used to evaluate the performance of different models on specific tasks, such as image segmentation, image classification, and image reconstruction.

Mr. Raji, and Drs. Daneshjou, and Alsentzer contributed equally to this article.

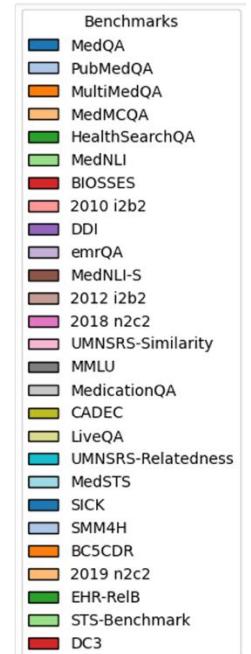
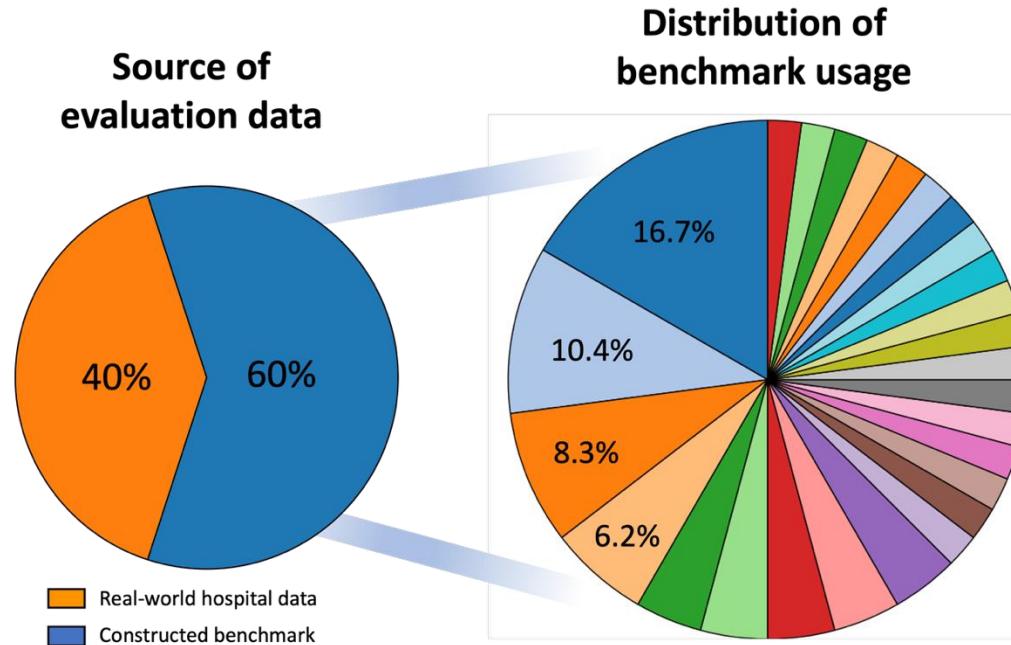
The author affiliations are listed at the end of the article.

Drs. Daneshjou can be contacted at rozanad@stanford.edu.

nature medicine

Volume 31 | April 2025 | 1060 | 1060

Many evaluation datasets, but no consensus benchmark!



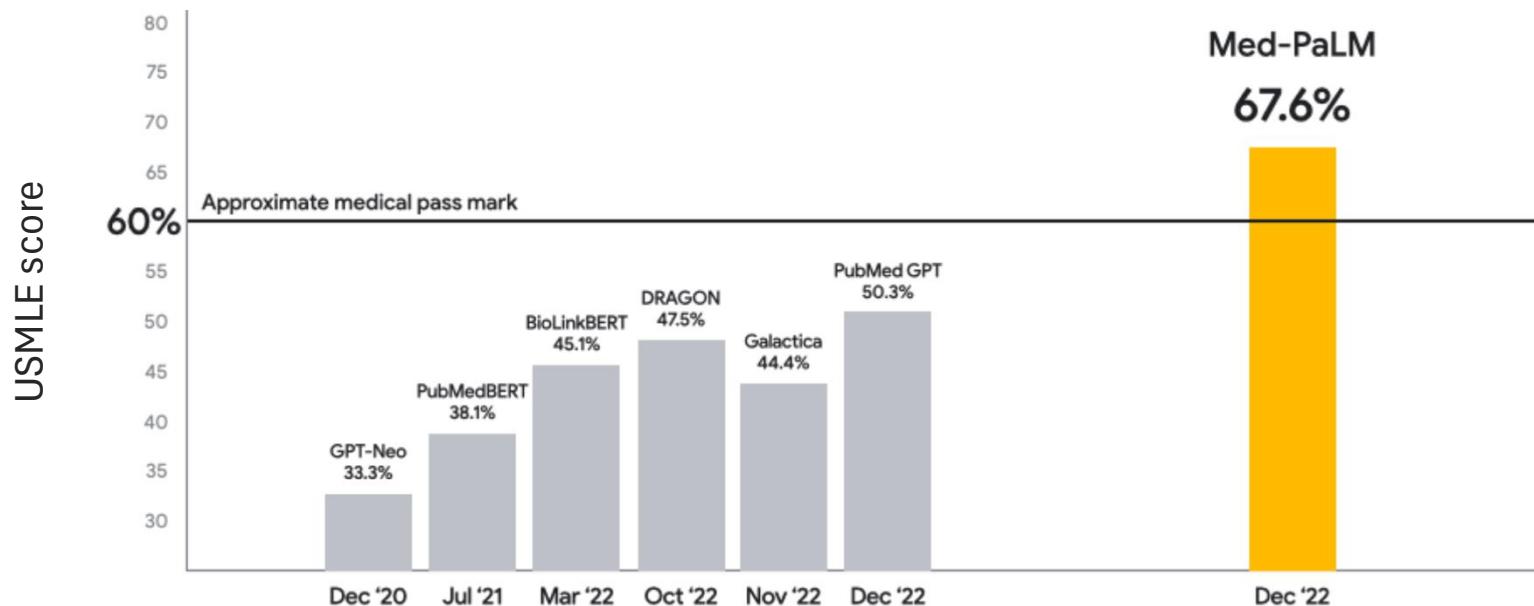
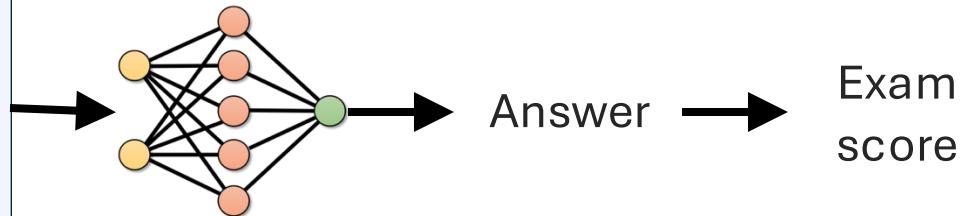
¹A. Alaa et al, ICML 2025.

“Well-specified” Benchmarks using Medical Licensing Exams

Multiple choice questions

A 36-year-old female with a past medical history of schizophrenia presents to the clinic for a routine yearly check-up. She reports no acute complaints and that her schizophrenia has been well controlled with olanzapine. She has been living under a bridge for the last three months due to being laid off from work. Vital signs include a temperature of 36.7°C (98.1°F), heart rate of 71 beats/min, respiratory rate of 14 breaths/min, blood pressure of 114/71 mmHg, and oxygen saturation of 98% on room air. Physical exam is notable for diminished breath sounds at the right upper lobe. A chest X-ray is shown below. The provider performs a PPD test, which results in an 18 mm induration 48 hours later. This type of hypersensitivity reaction is most similar to which of the following?

A. Anaphylaxis
B. Contact dermatitis
C. Graves disease
D. Hemolytic transfusion reaction
E. Poststreptococcal glomerulonephritis



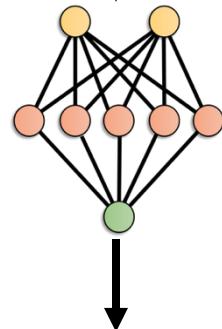
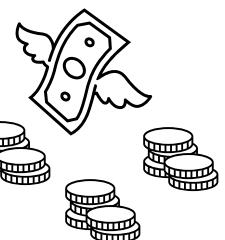
Source: Google Research (<https://sites.research.google/med-palm>)

Well-specified Benchmarks for Under-specified Tasks

Well-specified benchmark

A 36-year-old female with a past medical history of schizophrenia presents to the clinic for a routine yearly check-up. She reports no acute complaints and that her schizophrenia has been well controlled with olanzapine. She has been living under a bridge for the last three months due to being laid off from work. Vital signs include a temperature of 36.7°C (98.1°F), heart rate of 71 beats/min, respiratory rate of 14 breaths/min, blood pressure of 114/71 mmHg, and oxygen saturation of 98% on room air. Physical exam is notable for diminished breath sounds at the right upper lobe. A chest X-ray is shown below. The provider performs a PPD test, which results in an 18 mm induration 48 hours later. This type of hypersensitivity reaction is most similar to which of the following?

A. Anaphylaxis
B. Contact dermatitis
C. Graves disease
D. Hemolytic transfusion reaction
E. Poststreptococcal glomerulonephritis



Verifiable answer



Under-specified real-world task

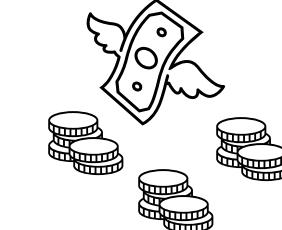
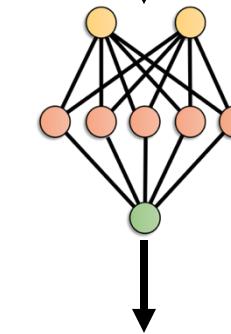
Imaging Reports

Discharge summary

Consult Notes

Echo Reports

Real-world clinical notes



When do benchmarks serve as valid proxies for human eval?

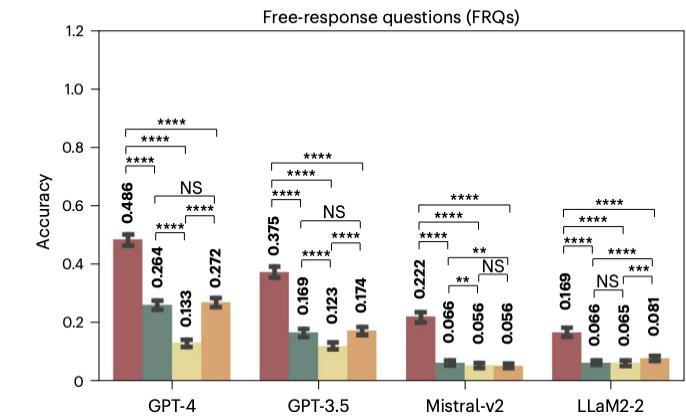
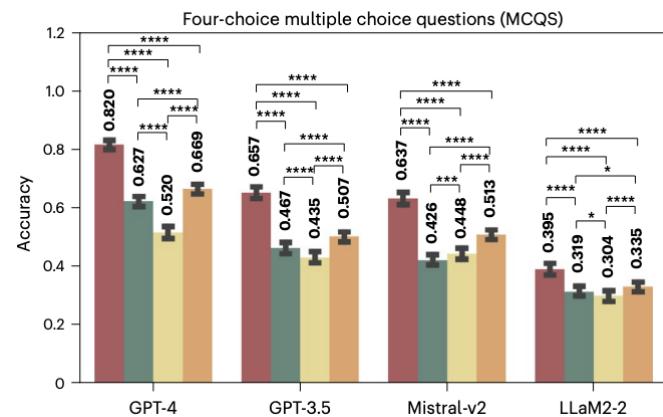
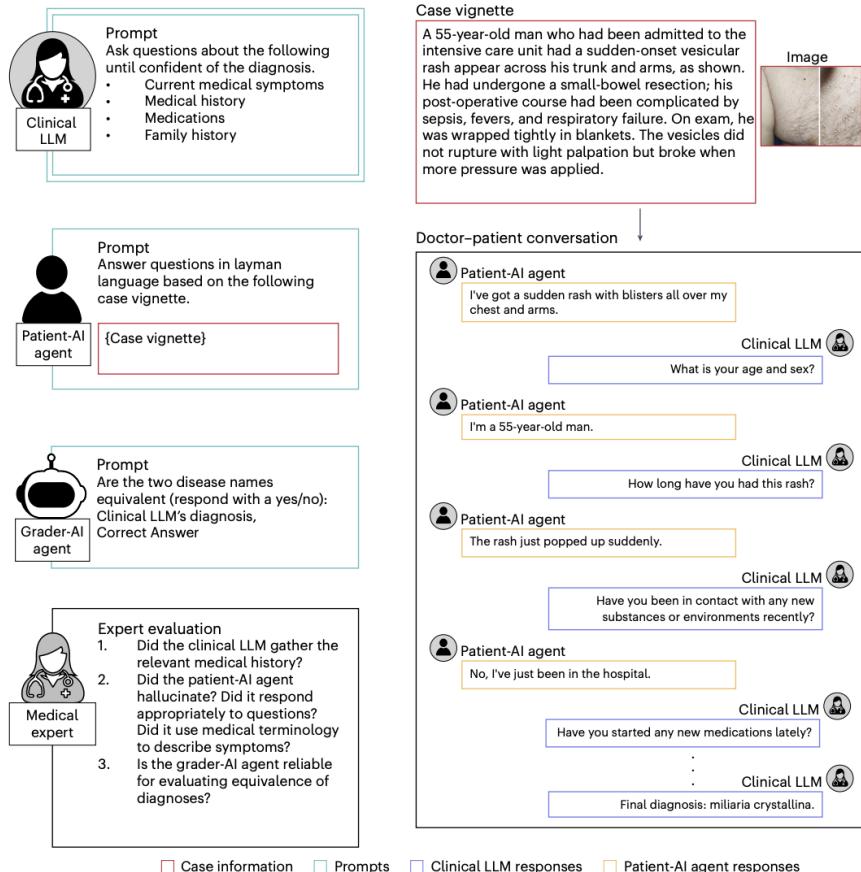
Human eval



Evaluation rubric	
Completeness	✓
Accuracy	✓
Clinical utility	✗

The Ongoing Challenge of Realistic Benchmarking

- CRAFT-MD = A conversation benchmark to evaluate history-taking abilities.*



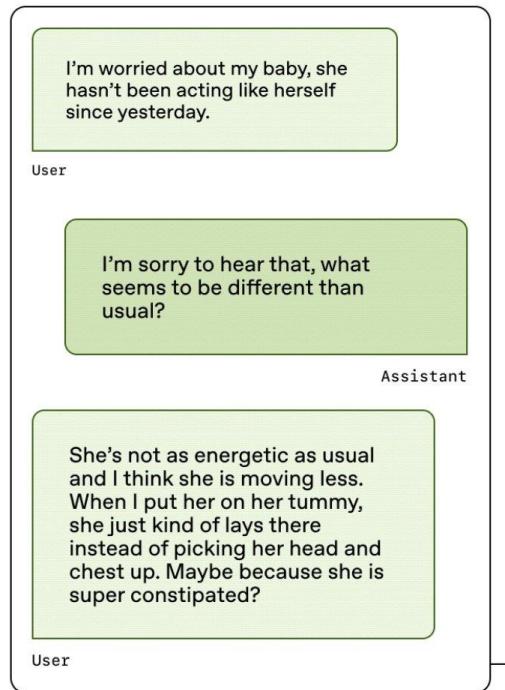
Open-ended questioning and using a combination of automated and expert evaluations.

*S. Johri et al., Nature Medicine, 2025

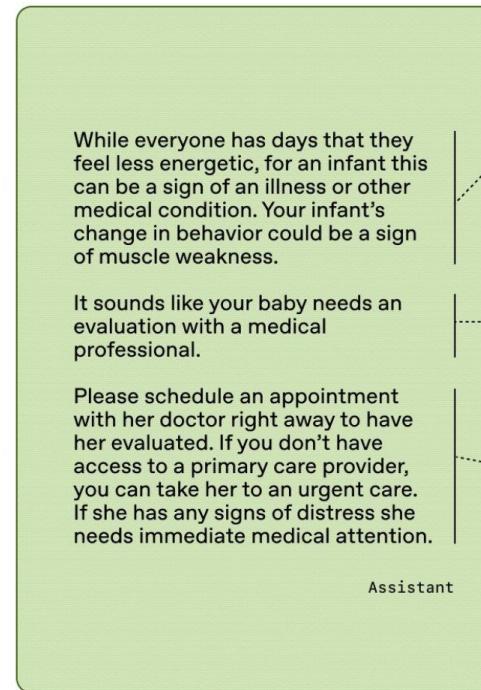
The Ongoing Challenge of Realistic Benchmarking

- **HealthBench (OpenAI)** = 5000 realistic health conversations with LLM graders.*

Eval Example



Candidate Response



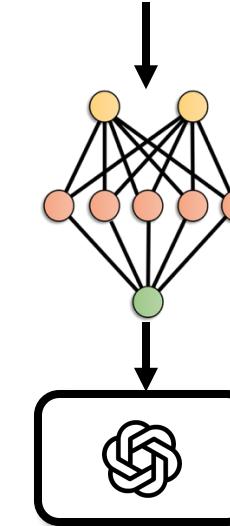
Rubric Criteria & Grades

Criterion	Points	Present
States that the infant may have muscle weakness	+7	Yes 7
Briefly describes common causes of muscle weakness in infants	+5	No 0
Includes advice to seek medical care right away	+10	Yes 10
Referral advice is only buried in the middle or at the end of response	-6	Yes -6

0 11 22 max

50% Actual Score

Input



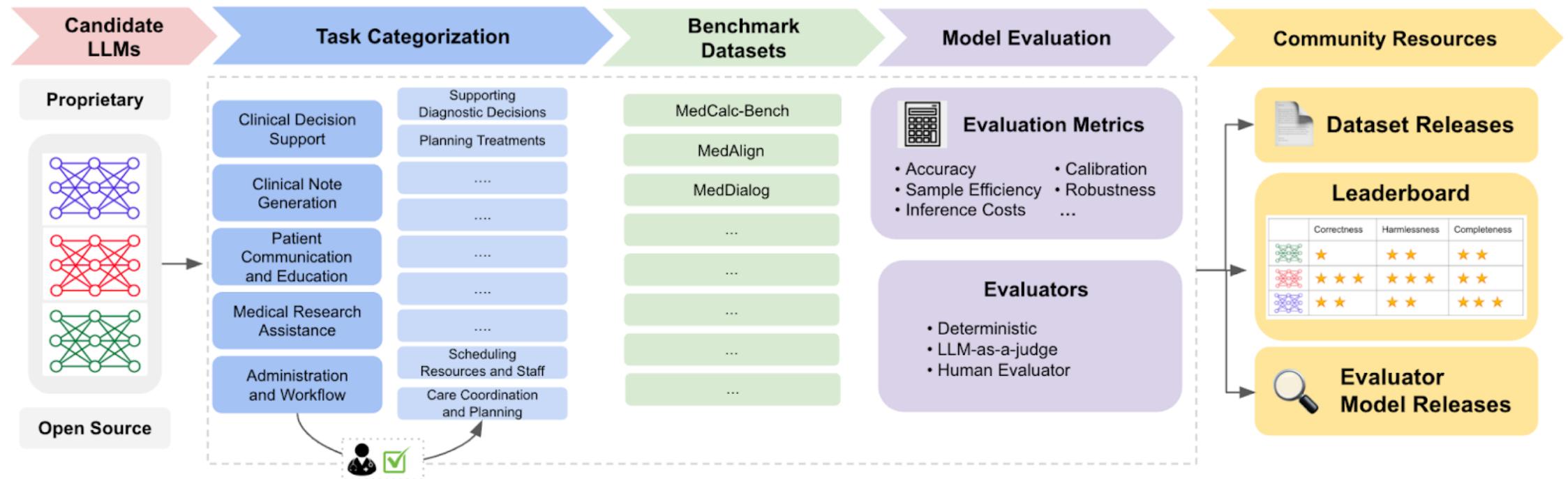
Evaluation rubric	
Completeness	✓
Accuracy	✓
Clinical utility	✗

LLM grader

*R. K. Arora, 2025

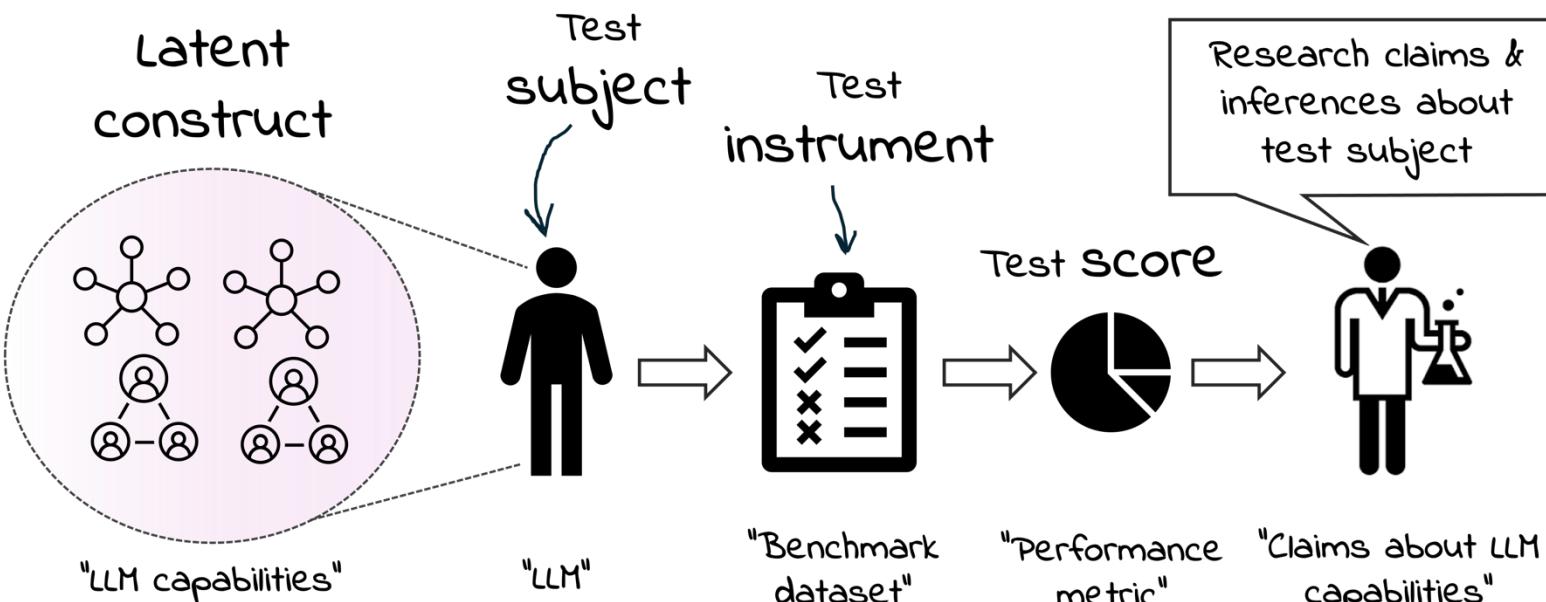
The Ongoing Challenge of Realistic Benchmarking

- MedHELM = Real-world data, 21 distinct clinical tasks and multiple disciplines.*



*S. Bedi et al, 2025

AI Benchmarks as Psychometric Tests



Position: Medical Large Language Model Benchmarks Should Prioritize Construct Validity

Ahmed Alaa^{1,2} Thomas Hartvigsen³ Niloufar Golchin¹ Shiladitya Dutta¹ Frances Dean^{1,2} Inioluwa Deborah Raji¹ Travis Zack²

Abstract

Medical large language models (LLMs) research often makes bold claims, from encoding clinical knowledge to reasoning like a physician. These claims are usually backed by evaluation on competitive benchmarks—a tradition inherited from mainstream machine learning. But how do we separate real progress from a leaderboard fix? Medical LLM benchmarks, much like those in other fields, are arbitrarily constructed using medical licensing exam questions. For these benchmarks to truly measure progress, they must accurately capture the real-world tasks they aim to represent. In this position paper, we argue that **medical LLM benchmarks should—and indeed can—be empirically evaluated for their construct validity**. In the psychological testing literature, “construct validity” refers to the ability of a test to measure an underlying “construct,” that is the actual conceptual target of evaluation. By drawing an analogy between LLM benchmarks and psychological tests, we explain how frameworks from this field can provide empirical foundations for validating benchmarks. To put these ideas into practice, we use real-world clinical data in proof-of-concept experiments to evaluate popular medical LLM benchmarks and report significant gaps in their construct validity. Finally, we outline a vision for a new ecosystem of medical LLM evaluation centered around the creation of valid benchmarks.

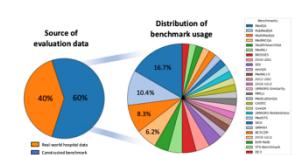


Figure 1. Overview of evaluation datasets for medical LLMs. We analyzed the evaluation datasets used in the 100 most cited papers on medical LLMs over the past 3 years. The majority (60%) of studies assess models on public benchmarks constructed based on medical exams, while others rely on (private or public access) real-world hospital data. There is no clear consensus on a standard benchmark—though MedQA is the most frequently used.

coding clinical knowledge (Singhal et al., 2023) to making differential diagnoses (McDuff et al., 2023), summarizing complex medical texts (Van Veen et al., 2024), mimicking clinical reasoning (Savage et al., 2024; Brodeur et al., 2024), and even demonstrating empathy in patient interactions (Maida et al., 2024). Yet the question of how to evaluate these capabilities remains a subject of ongoing debate. In the world of medicine, the gold standard for generating evidence is the randomized controlled trial (RCT). While some studies indeed conduct RCTs with meaningful real-world outcomes (Li et al., 2023; Brodeur et al., 2024), most research on medical LLMs leans on competitive benchmarks—an evaluation practice inherited from the broader machine learning community (Donoho, 2024; Orr & Kang, 2024).

1. Introduction

In recent years, medical Large Language Models (LLMs) have garnered significant attention, with a growing body of research examining their capabilities. These range from en-

¹UC Berkeley ²UCSF ³University of Virginia. Correspondence to: Ahmed Alaa camalaa@berkeley.edu.

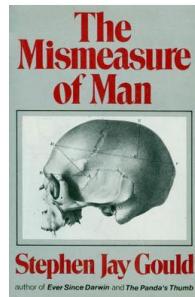
Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by author(s).

Measurement Theory Might Be the Answer

Well-specified benchmark

A 36-year-old female with a past medical history of schizophrenia presents to the clinic for a routine yearly check-up. She reports no acute complaints and that her schizophrenia has been well controlled with olanzapine. She has been living under a bridge for the last three months due to being laid off from work. Vital signs include a temperature of 36.7°C (98.1°F), heart rate of 71 beats/min, respiratory rate of 14 breaths/min, blood pressure of 114/71 mmHg, and oxygen saturation of 98% on room air. Physical exam is notable for diminished breath sounds at the right upper lobe. A chest X-ray is shown below. The provider performs a PPD test, which results in an 18 mm induration 48 hours later. This type of hypersensitivity reaction is most similar to which of the following?

A. Anaphylaxis
B. Contact dermatitis
C. Graves disease
D. Hemolytic transfusion reaction
E. Poststreptococcal glomerulonephritis



Under-specified real-world task



Real-world clinical notes



Verifiable answer

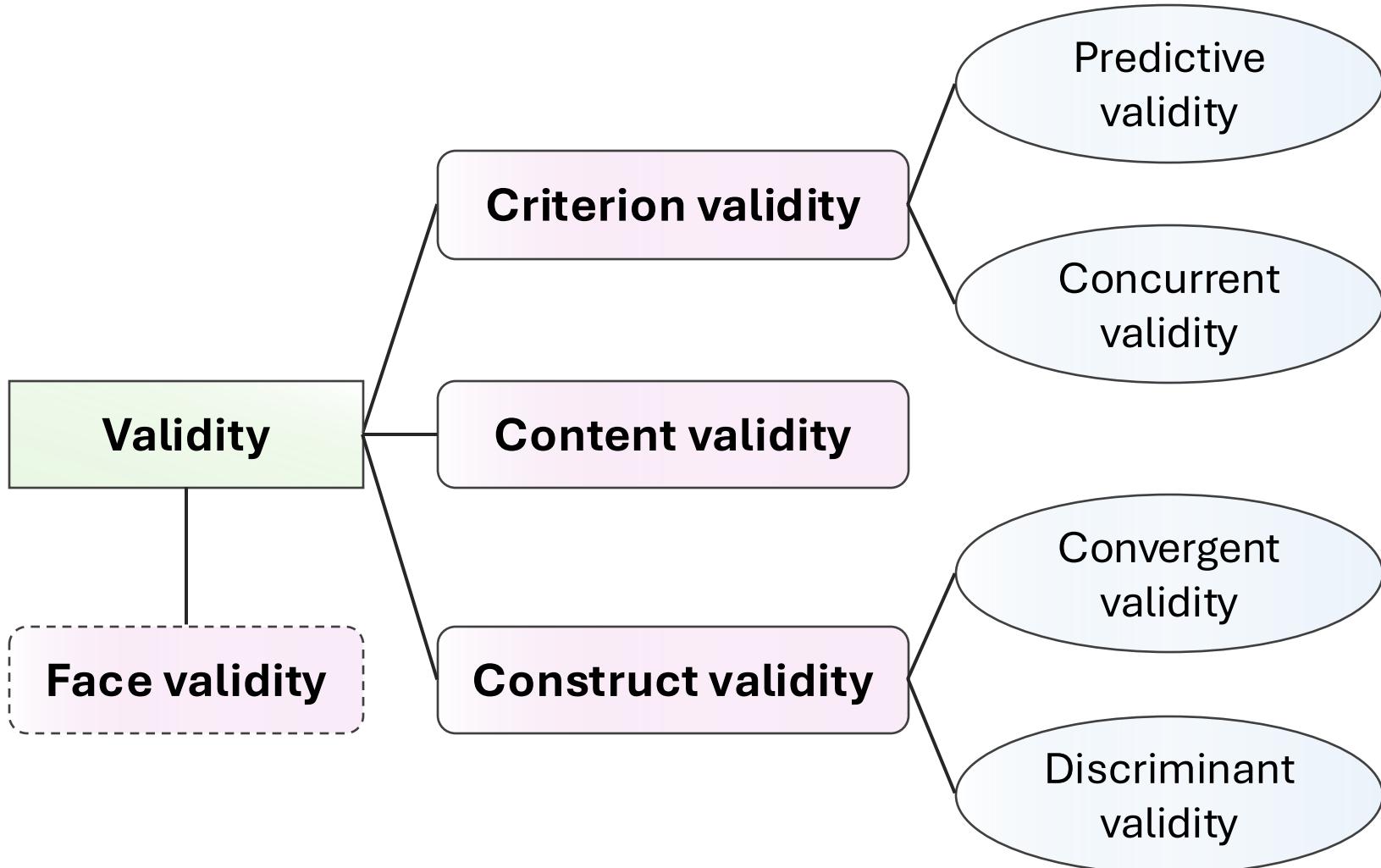


Human eval



Evaluation rubric	
Completeness	✓
Accuracy	✓
Clinical utility	✗

Validity of Measures in the Social Sciences

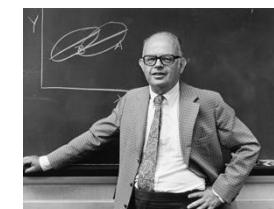


L. J. CRONBACH and P. E. MEEHL

Construct Validity in Psychological Tests

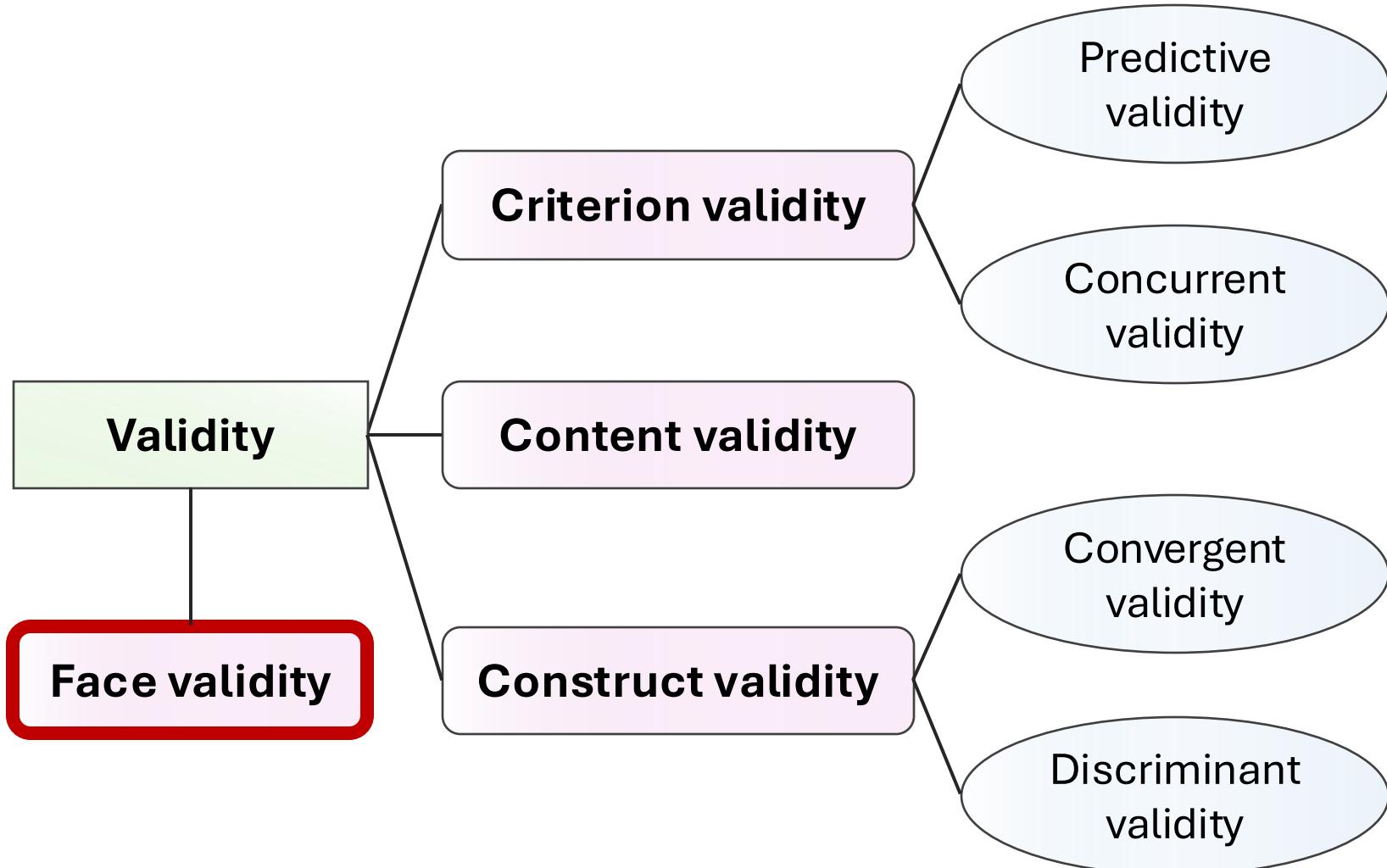
VALIDATION of psychological tests has not yet been adequately conceptualized, as the APA Committee on Psychological Tests learned when it undertook (1950-54) to specify what qualities should be investigated before a test is published. In order to make coherent recommendations the Committee found it necessary to distinguish four types of validity, established by different types of research and requiring different interpretation. The chief innovation in the Committee's report was the term *construct validity*.* This idea was first formulated by a subcommittee (Meehl and R. C. Challman) studying how proposed recommendations would apply to projective techniques, and later modified and clarified by the entire Committee (Bordin, Challman, Conrad, Humphreys, Super, and the present writers). The statements agreed upon by the Committee (and by committees of two other associations) were published in the *Technical Recommendations* (59). The present interpretation of construct validity is not "official" and deals with some areas in which the Committee would probably not be unanimous. The present writers are solely responsible for this attempt to explain the concept and elaborate its implications.

* Identification of construct validity was not an isolated development.



Cronbach and Meehl (1955)

Validity of Measures in the Social Sciences

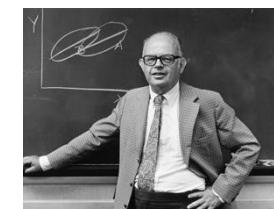


L. J. CRONBACH and P. E. MEEHL

Construct Validity in Psychological Tests

VALIDATION of psychological tests has not yet been adequately conceptualized, as the APA Committee on Psychological Tests learned when it undertook (1950-54) to specify what qualities should be investigated before a test is published. In order to make coherent recommendations the Committee found it necessary to distinguish four types of validity, established by different types of research and requiring different interpretation. The chief innovation in the Committee's report was the term *construct validity*.* This idea was first formulated by a subcommittee (Meehl and R. C. Challman) studying how proposed recommendations would apply to projective techniques, and later modified and clarified by the entire Committee (Bordin, Challman, Conrad, Humphreys, Super, and the present writers). The statements agreed upon by the Committee (and by committees of two other associations) were published in the *Technical Recommendations* (59). The present interpretation of construct validity is not "official" and deals with some areas in which the Committee would probably not be unanimous. The present writers are solely responsible for this attempt to explain the concept and elaborate its implications.

* Identification of construct validity was not an isolated development.

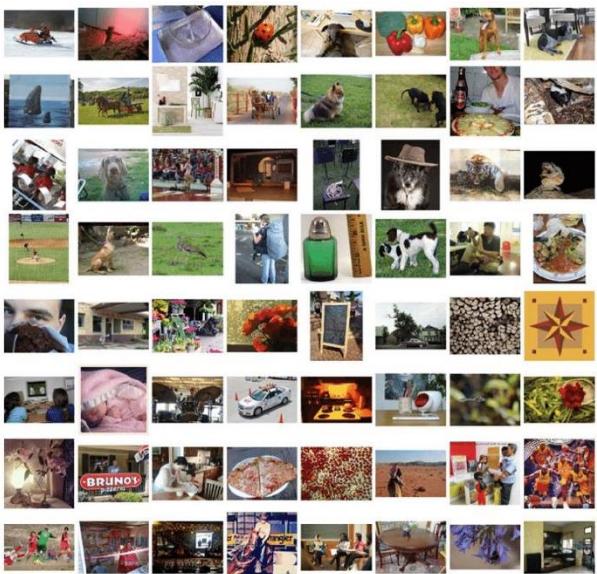


Cronbach and Meehl (1955)

Face Validity of Benchmarks

- **Face validity:** a superficial and subjective assessment of whether a test looks like it measures what it intends to measure

ImageNet is an intuitive test for the ability to classify objects in images



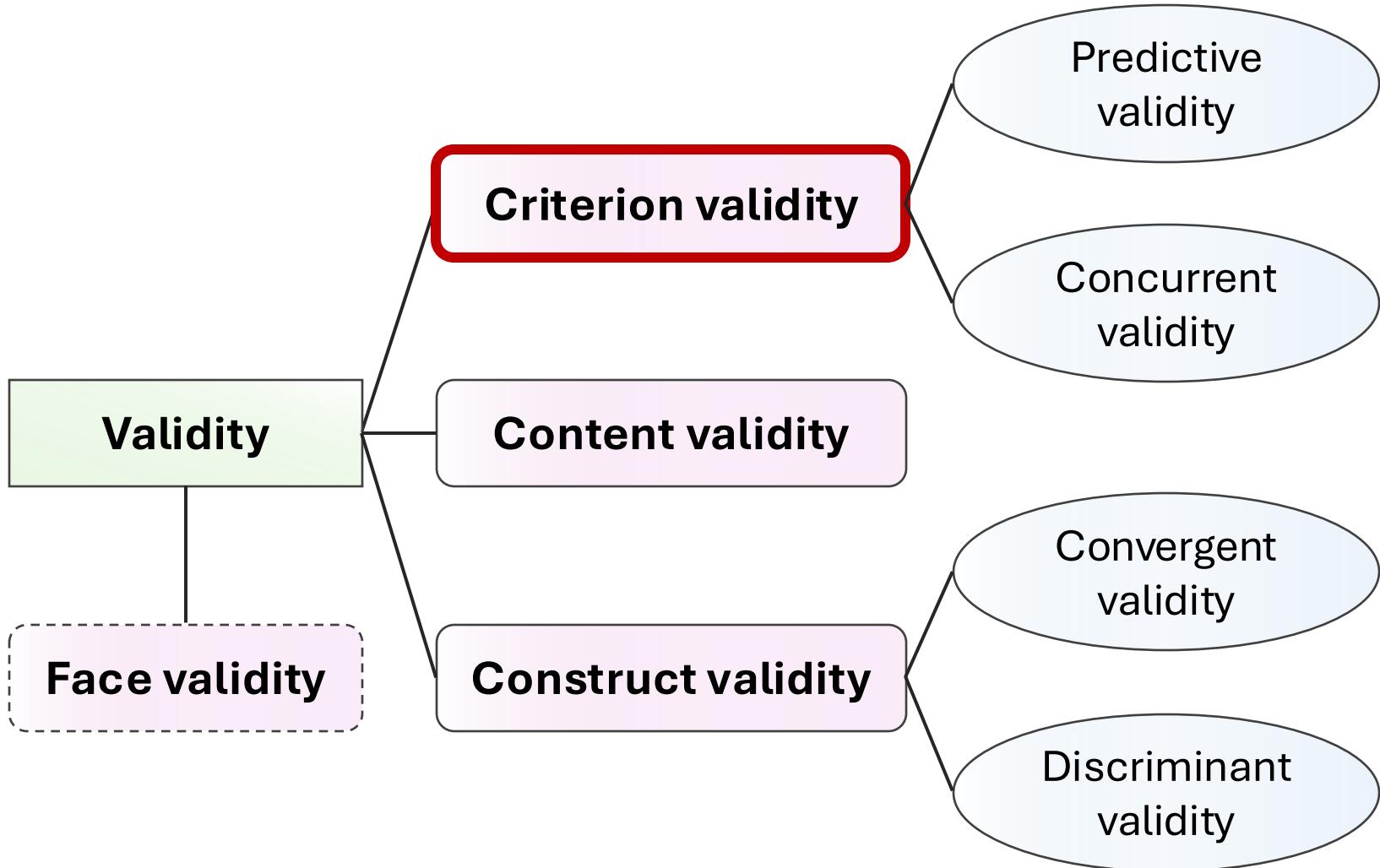
MedQA is an intuitive exam-style test for clinical knowledge

A 36-year-old female with a past medical history of schizophrenia presents to the clinic for a routine yearly check-up. She reports no acute complaints and that her schizophrenia has been well controlled with olanzapine. She has been living under a bridge for the last three months due to being laid off from work. Vital signs include a temperature of 36.7°C (98.1°F), heart rate of 71 beats/min, respiratory rate of 14 breaths/min, blood pressure of 114/71 mmHg, and oxygen saturation of 98% on room air. Physical exam is notable for diminished breath sounds at the right upper lobe. A chest X-ray is shown below. The provider performs a PPD test, which results in an 18 mm induration 48 hours later. This type of hypersensitivity reaction is most similar to which of the following?

- A. Anaphylaxis
- B. Contact dermatitis
- C. Graves disease
- D. Hemolytic transfusion reaction
- E. Poststreptococcal glomerulonephritis



Validity of Measures in the Social Sciences

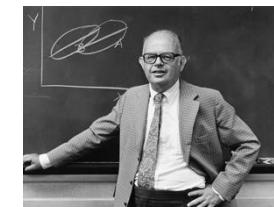


L. J. CRONBACH and P. E. MEEHL

Construct Validity in Psychological Tests

VALIDATION of psychological tests has not yet been adequately conceptualized, as the APA Committee on Psychological Tests learned when it undertook (1950-54) to specify what qualities should be investigated before a test is published. In order to make coherent recommendations the Committee found it necessary to distinguish four types of validity, established by different types of research and requiring different interpretation. The chief innovation in the Committee's report was the term *construct validity*.* This idea was first formulated by a subcommittee (Meehl and R. C. Challman) studying how proposed recommendations would apply to projective techniques, and later modified and clarified by the entire Committee (Bordin, Challman, Conrad, Humphreys, Super, and the present writers). The statements agreed upon by the Committee (and by committees of two other associations) were published in the *Technical Recommendations* (59). The present interpretation of construct validity is not "official" and deals with some areas in which the Committee would probably not be unanimous. The present writers are solely responsible for this attempt to explain the concept and elaborate its implications.

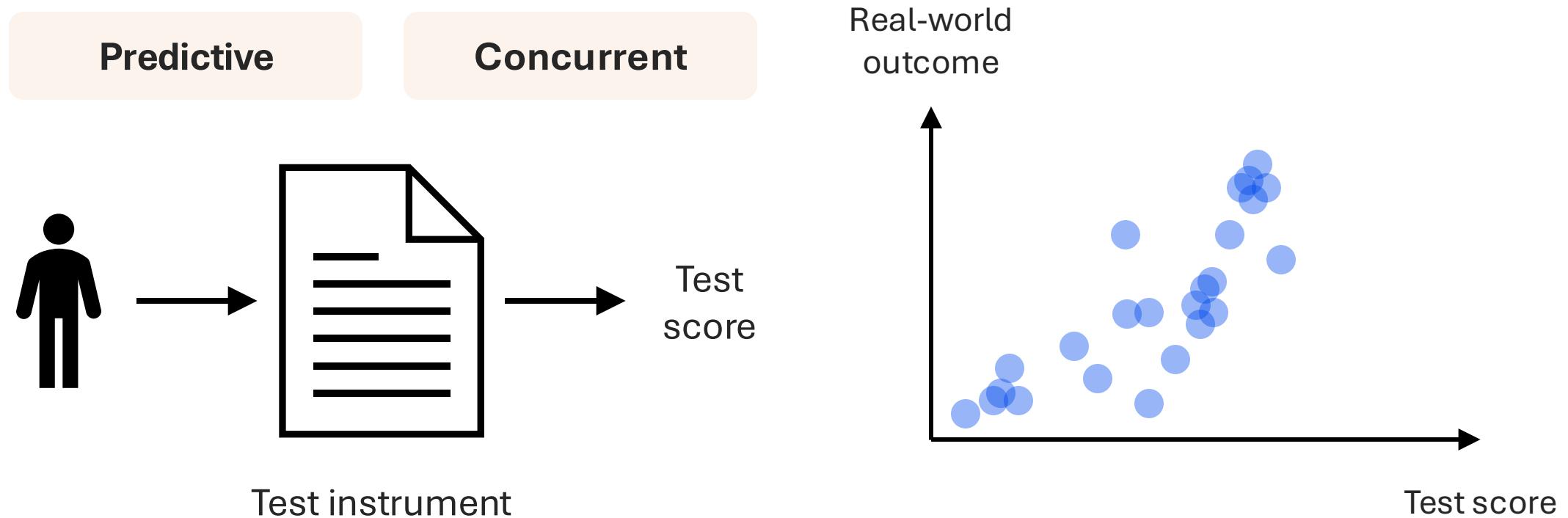
* Identification of construct validity was not an isolated development.



Cronbach and Meehl (1955)

Criterion Validity

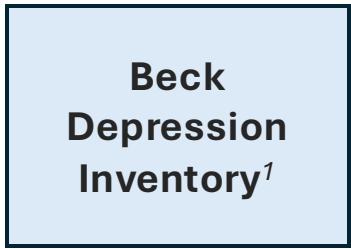
- **Criterion validity:** how well a test or measurement aligns with a pre-existing, established standard (the "criterion").



Criterion Validity in Psychometrics and Education

Psychology

Explanatory variables



Criterion variables

A photograph of a printed Beck Depression Inventory (BDI) form. The form is a white document with a grid of questions and response options (A, B, C, D) for each question. At the bottom right, there is a small logo for "counselor.com".

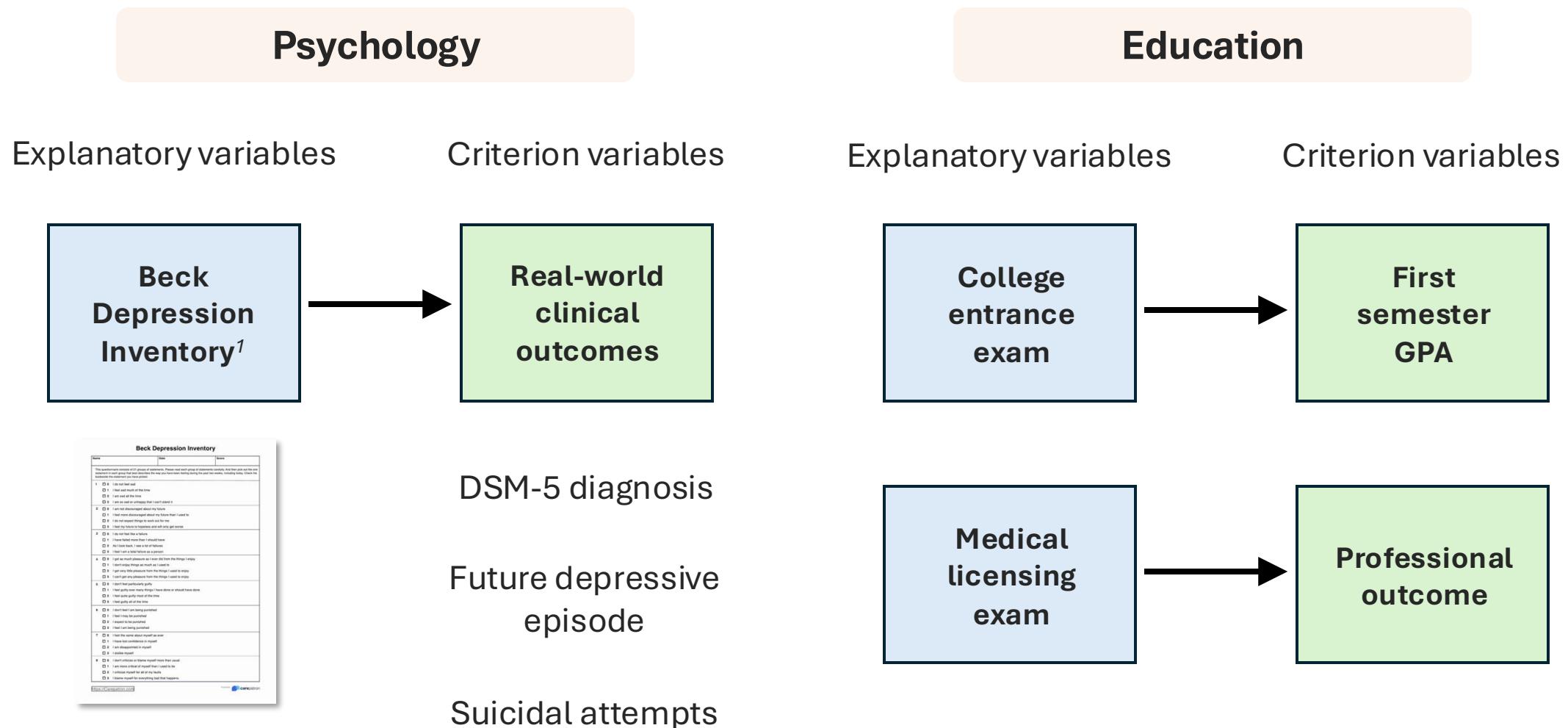
DSM-5 diagnosis

Future depressive
episode

Suicidal attempts

¹AT Beck, 1972

Criterion Validity in Psychometrics and Education

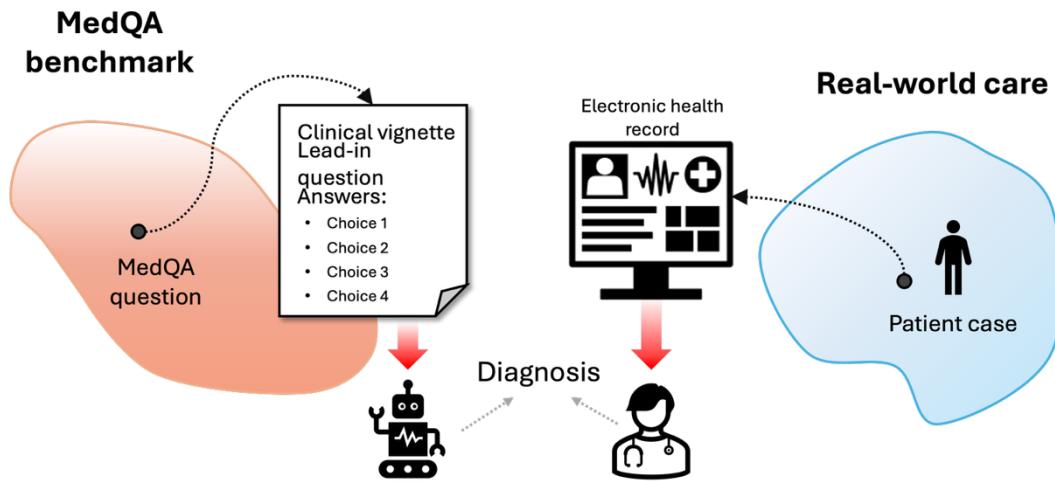


¹AT Beck, 1972

Criterion Validity of Medical LLM Benchmarks

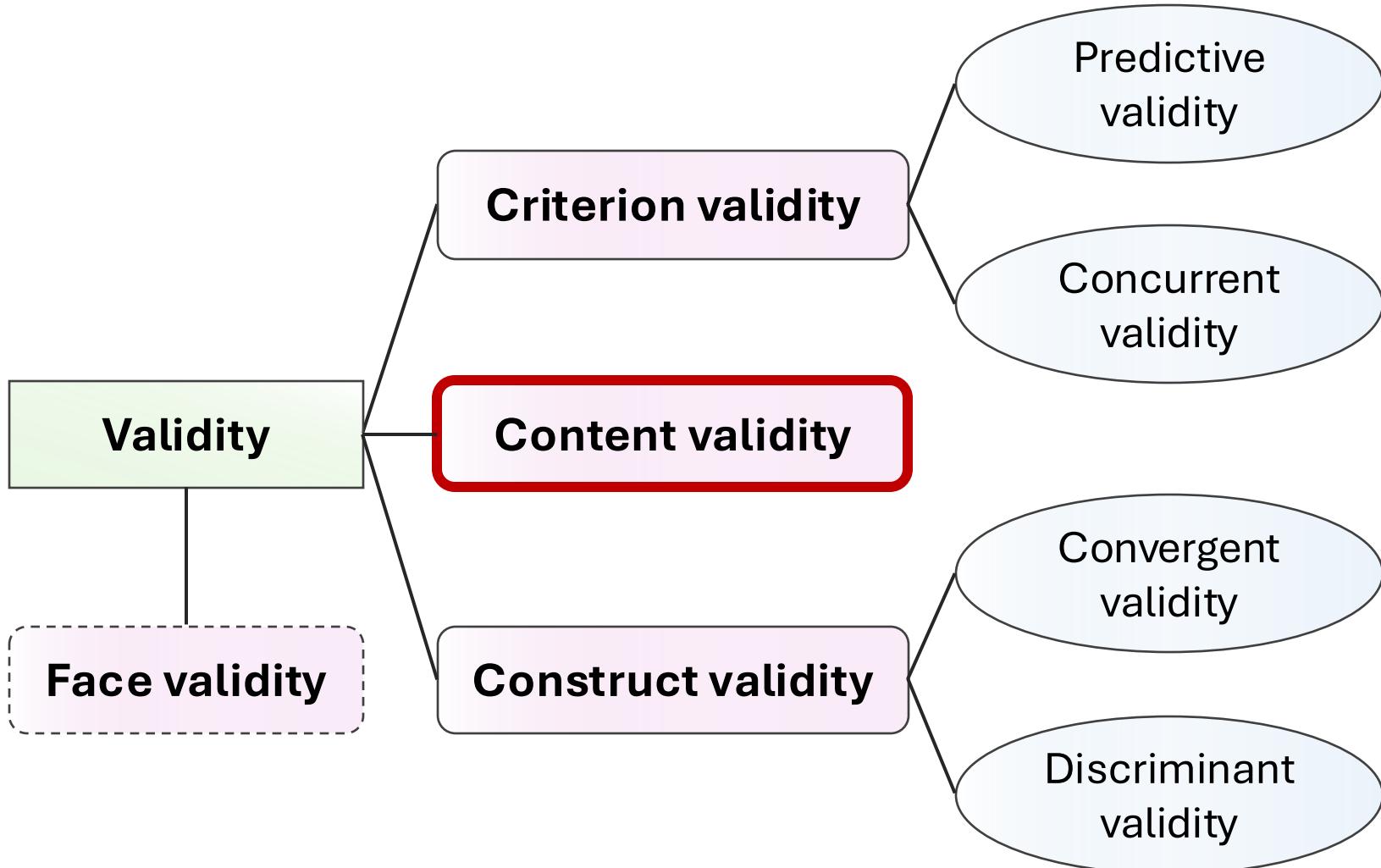
- LLM performance on clinical vignettes in medical exam does not correlate with similar patients in real-world data from the UCSF electronic health records.

$$\alpha = P(\text{Correct on real-world case} \mid \text{Correct on MedQA})$$



	MedQA	Real-world data	
	<i>Accuracy</i>	<i>Accuracy</i>	α
Llama 3	0.54	0.48	0.56
GPT-4	0.71	0.28	0.29
Chimera Llama	0.60	0.45	0.48
Biomerge	0.57	0.36	0.49
Orpomed	0.49	0.24	0.38
JSL MedLlama	0.61	0.37	0.49
PMY MedLLama	0.75	0.36	0.45

Validity of Measures in the Social Sciences

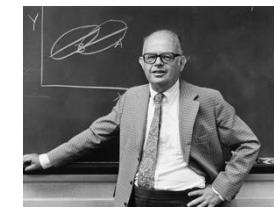


L. J. CRONBACH and P. E. MEEHL

Construct Validity in Psychological Tests

VALIDATION of psychological tests has not yet been adequately conceptualized, as the APA Committee on Psychological Tests learned when it undertook (1950-54) to specify what qualities should be investigated before a test is published. In order to make coherent recommendations the Committee found it necessary to distinguish four types of validity, established by different types of research and requiring different interpretation. The chief innovation in the Committee's report was the term *construct validity*.* This idea was first formulated by a subcommittee (Meehl and R. C. Challman) studying how proposed recommendations would apply to projective techniques, and later modified and clarified by the entire Committee (Bordin, Challman, Conrad, Humphreys, Super, and the present writers). The statements agreed upon by the Committee (and by committees of two other associations) were published in the *Technical Recommendations* (59). The present interpretation of construct validity is not "official" and deals with some areas in which the Committee would probably not be unanimous. The present writers are solely responsible for this attempt to explain the concept and elaborate its implications.

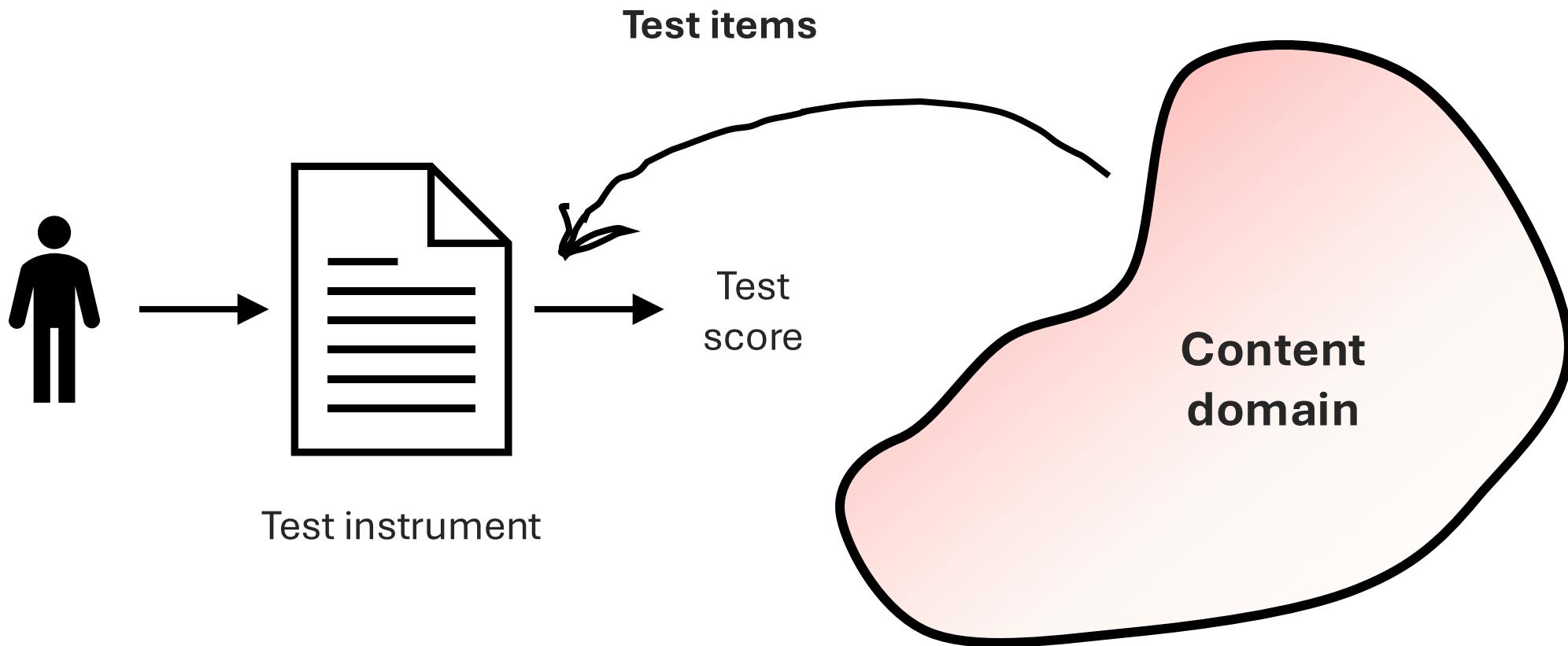
* Identification of construct validity was not an isolated development.



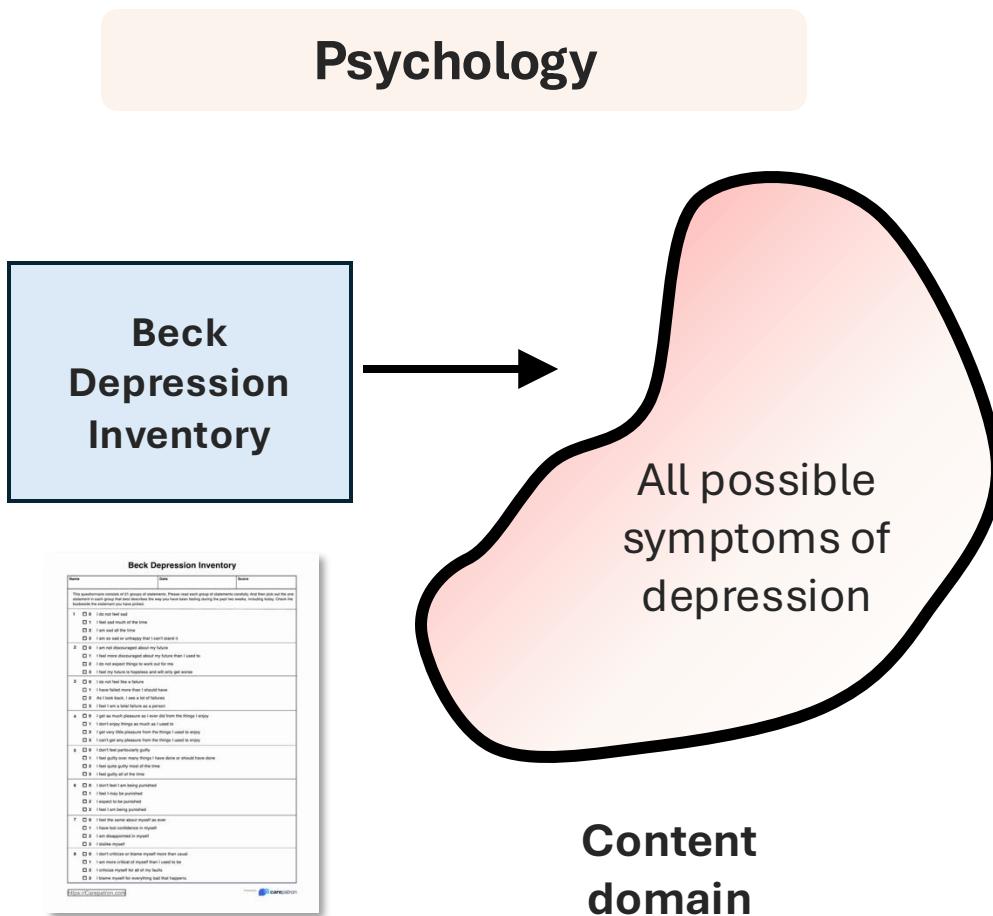
Cronbach and Meehl (1955)

Content Validity

- **Content validity:** the extent to which a test comprehensively represents the specific domain of content it is intended to measure.

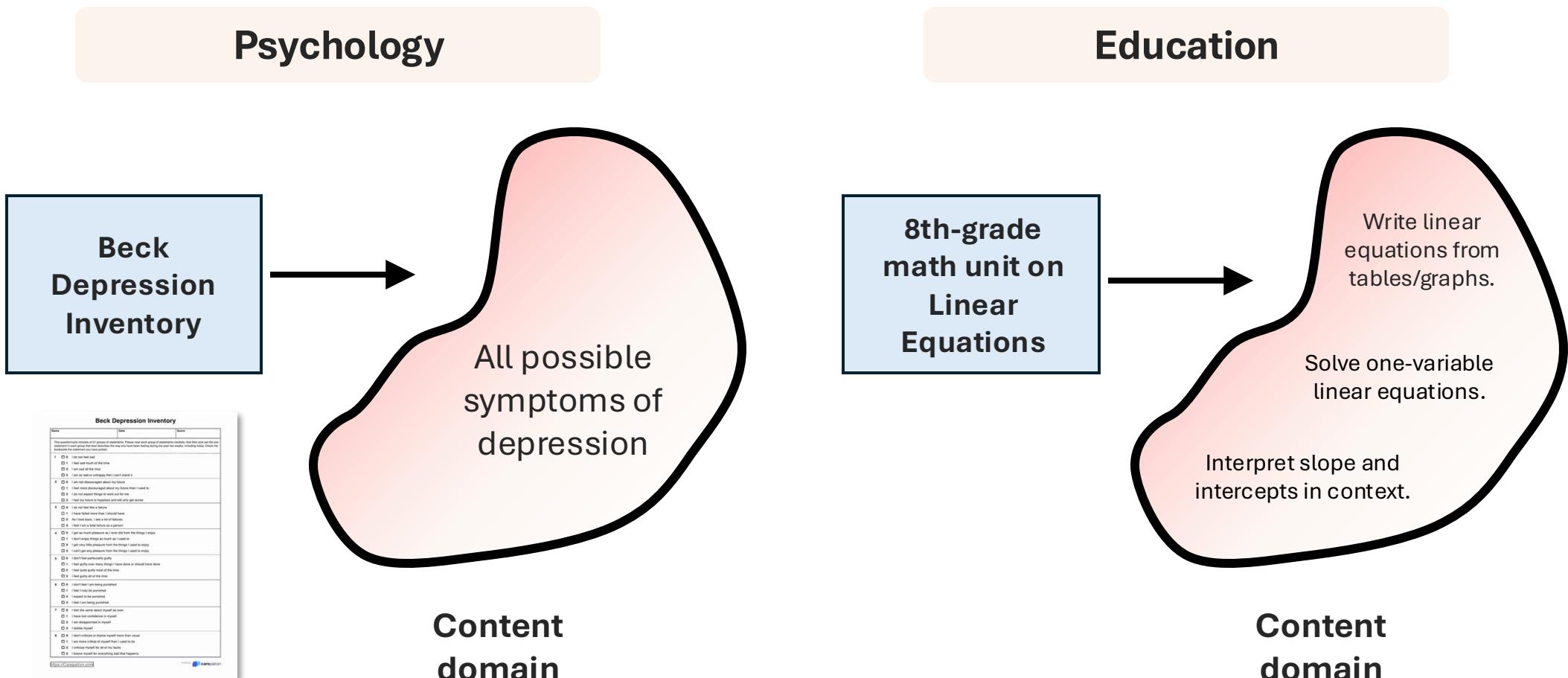


Content Validity in Psychometrics and Education



¹AT Beck, 1972

Content Validity in Psychometrics and Education



¹AT Beck, 1972

Content Validity of Medical LLM Benchmarks

- Content domain of medicine is well-structured in comprehensive ontologies (SNOMED, LOINC, RxNorm, ICD), taxonomies and tasks.

MedQA clinical vignette

A 36-year-old female with a past medical history of schizophrenia presents to the clinic for a routine yearly check-up. She reports no acute complaints and that her schizophrenia has been well controlled with olanzapine. She has been living under a bridge for the last three months due to being laid off from work. Vital signs include a temperature of 36.7°C (98.1°F), heart rate of 71 beats/min, respiratory rate of 14 breaths/min, blood pressure of 114/71 mmHg, and oxygen saturation of 98% on room air. Physical exam is notable for diminished breath sounds at the right upper lobe. A chest X-ray is shown below. The provider performs a PPD test, which results in an 18 mm induration 48 hours later. This type of hypersensitivity reaction is most similar to which of the following?

- A. Anaphylaxis
- B. Contact dermatitis**
- C. Graves disease
- D. Hemolytic transfusion reaction
- E. Poststreptococcal glomerulonephritis



UCSF clinical notes



UMLS concepts

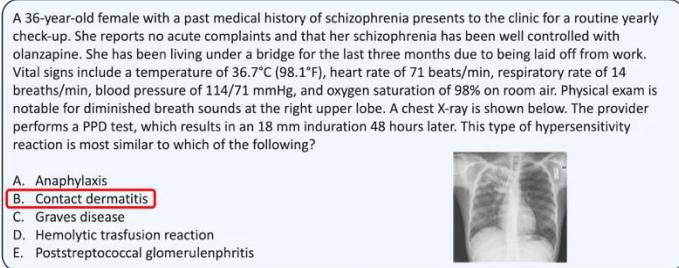


UMLS concepts

Content Validity of Medical LLM Benchmarks

- Content domain of medicine is well-structured in comprehensive ontologies (SNOMED, LOINC, RxNorm, ICD), taxonomies and tasks.

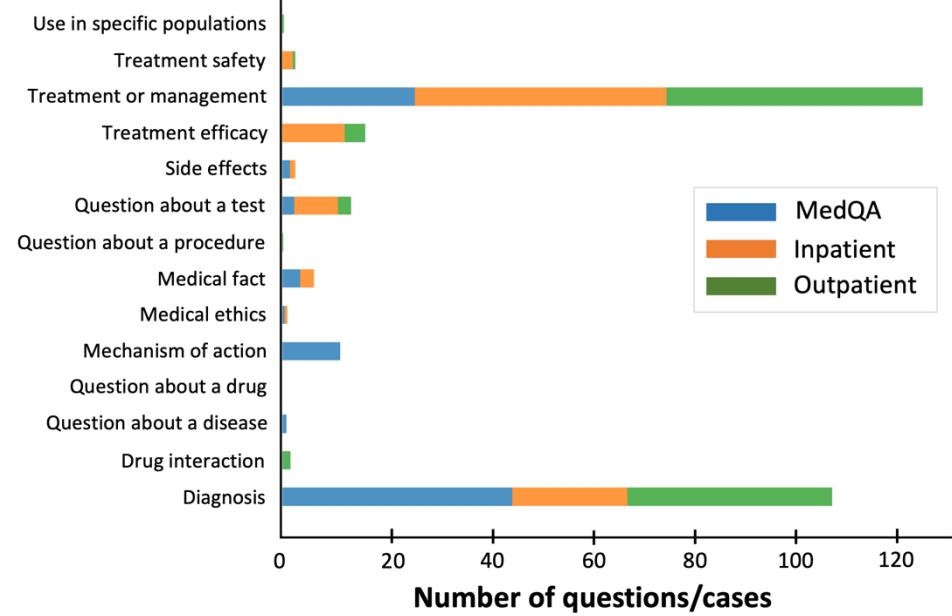
MedQA clinical vignette



UCSF clinical notes



Real-world clinical tasks/scenarios



UMLS concepts



UMLS concepts

Content Validity of Medical LLM Benchmarks

- Content domain of medicine is well-structured in comprehensive ontologies (SNOMED, LOINC, RxNorm, ICD), taxonomies and tasks.

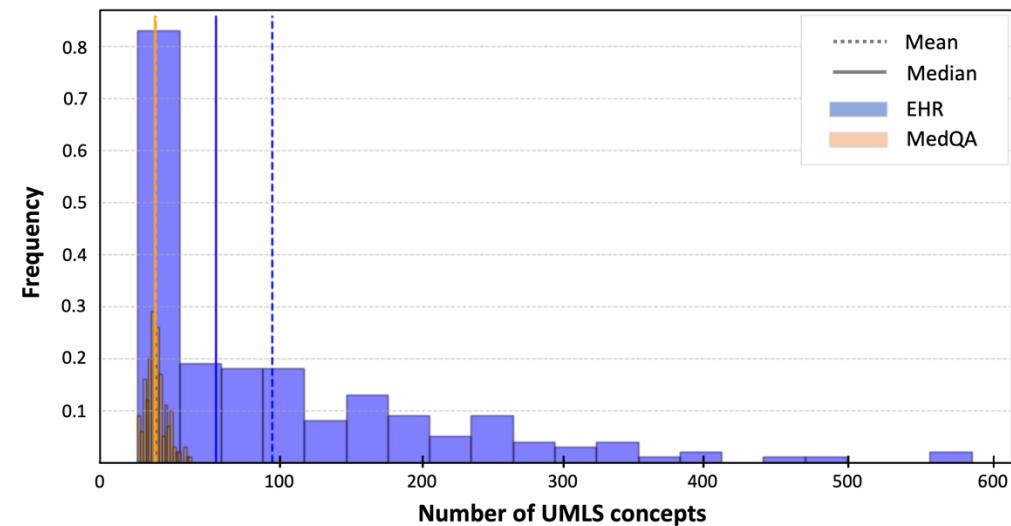
MedQA clinical vignette

A 36-year-old female with a past medical history of schizophrenia presents to the clinic for a routine yearly check-up. She reports no acute complaints and that her schizophrenia has been well controlled with olanzapine. She has been living under a bridge for the last three months due to being laid off from work. Vital signs include a temperature of 36.7°C (98.1°F), heart rate of 71 beats/min, respiratory rate of 14 breaths/min, blood pressure of 114/71 mmHg, and oxygen saturation of 98% on room air. Physical exam is notable for diminished breath sounds at the right upper lobe. A chest X-ray is shown below. The provider performs a PPD test, which results in an 18 mm induration 48 hours later. This type of hypersensitivity reaction is most similar to which of the following?

- A. Anaphylaxis
- B. Contact dermatitis**
- C. Graves disease
- D. Hemolytic transfusion reaction
- E. Poststreptococcal glomerulonephritis



UCSF clinical notes

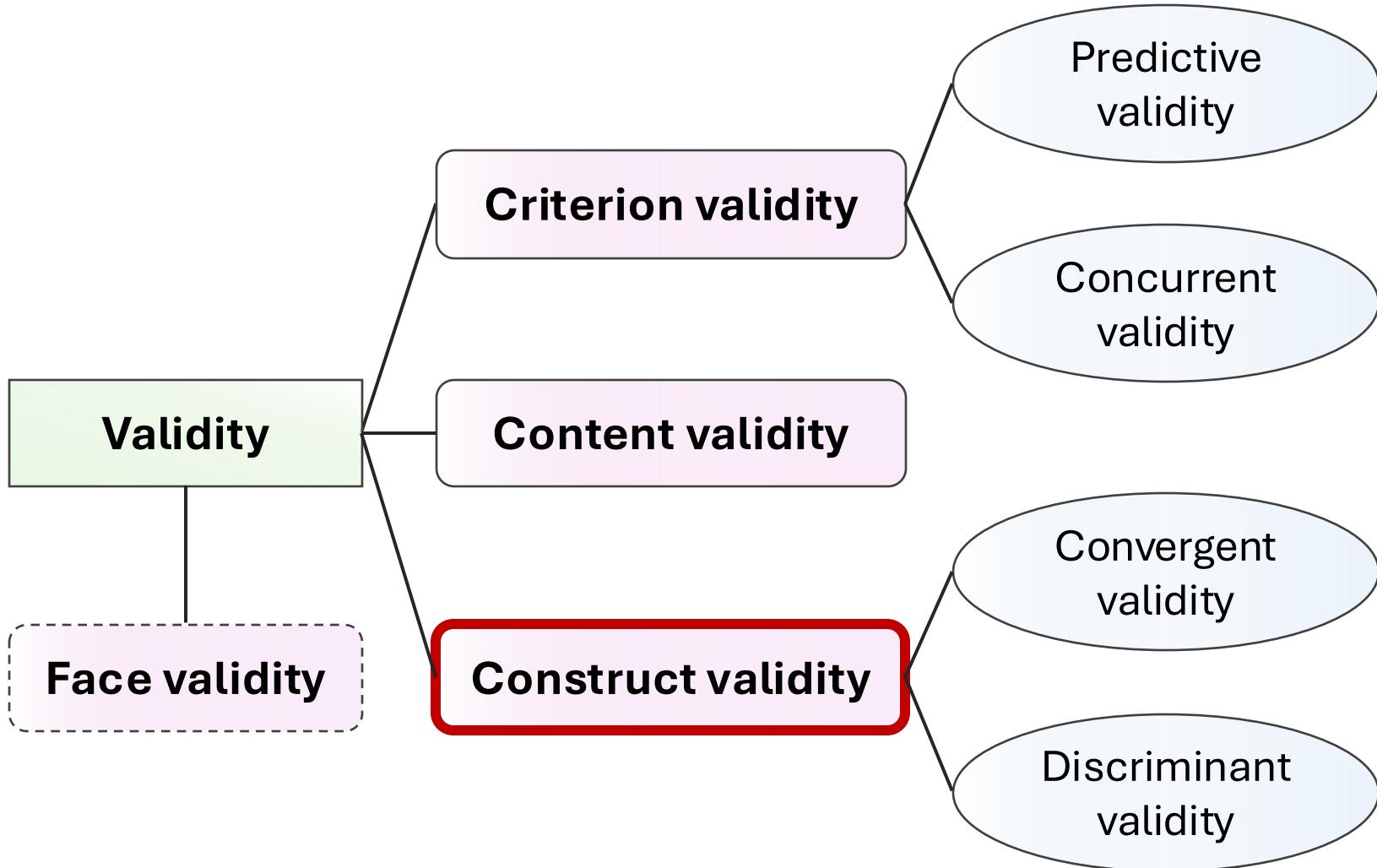


UMLS concepts



UMLS concepts

Validity of Measures in the Social Sciences

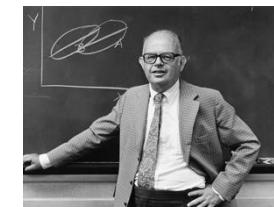


L. J. CRONBACH and P. E. MEEHL

Construct Validity in Psychological Tests

VALIDATION of psychological tests has not yet been adequately conceptualized, as the APA Committee on Psychological Tests learned when it undertook (1950-54) to specify what qualities should be investigated before a test is published. In order to make coherent recommendations the Committee found it necessary to distinguish four types of validity, established by different types of research and requiring different interpretation. The chief innovation in the Committee's report was the term *construct validity*.* This idea was first formulated by a subcommittee (Meehl and R. C. Challman) studying how proposed recommendations would apply to projective techniques, and later modified and clarified by the entire Committee (Bordin, Challman, Conrad, Humphreys, Super, and the present writers). The statements agreed upon by the Committee (and by committees of two other associations) were published in the *Technical Recommendations* (59). The present interpretation of construct validity is not "official" and deals with some areas in which the Committee would probably not be unanimous. The present writers are solely responsible for this attempt to explain the concept and elaborate its implications.

* Identification of construct validity was not an isolated development.

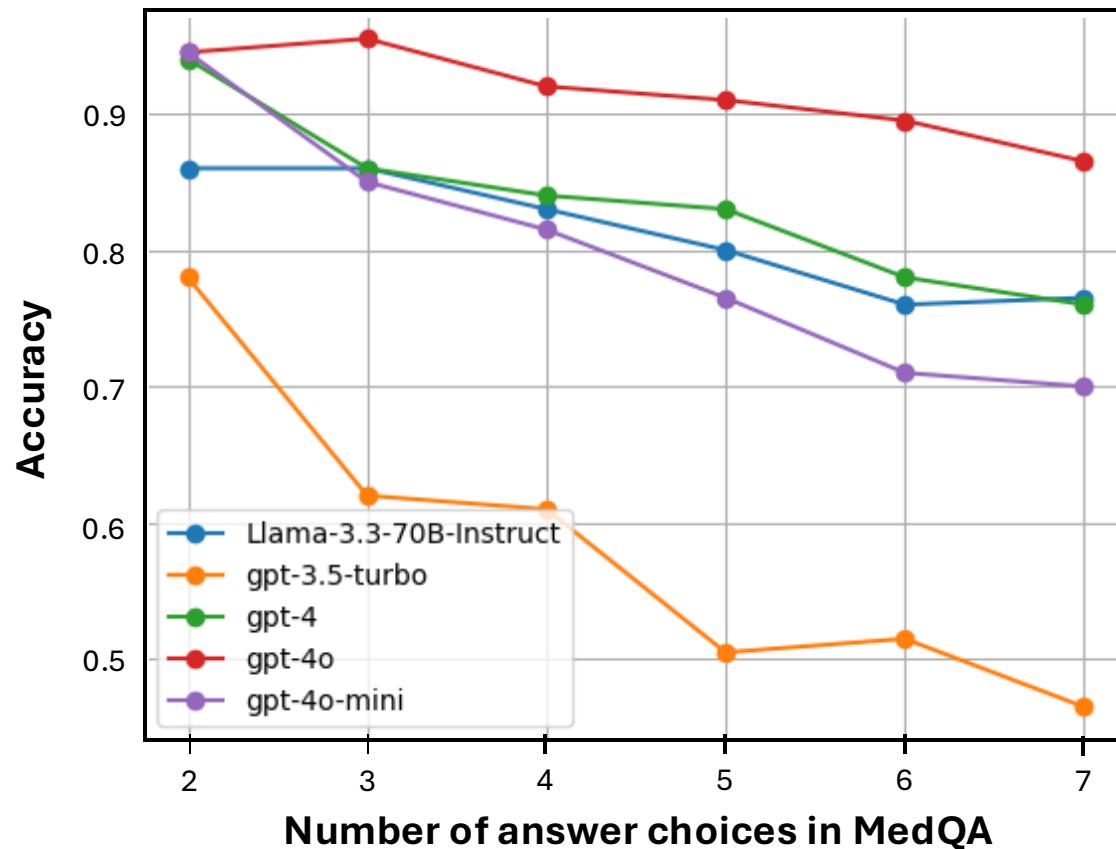


Cronbach and Meehl (1955)

Construct Validity

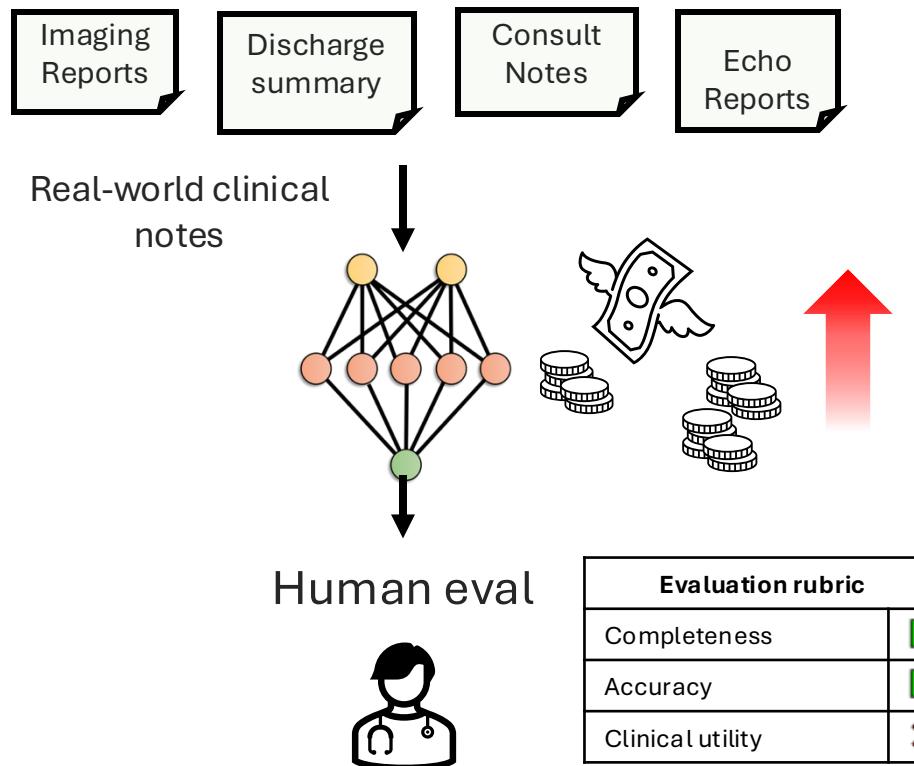
- **Construct validity:** the extent to which a test or benchmark actually measures the theoretical construct it claims to measure.

Expanding the number of answer choices causes a drop of performance in MedQA!



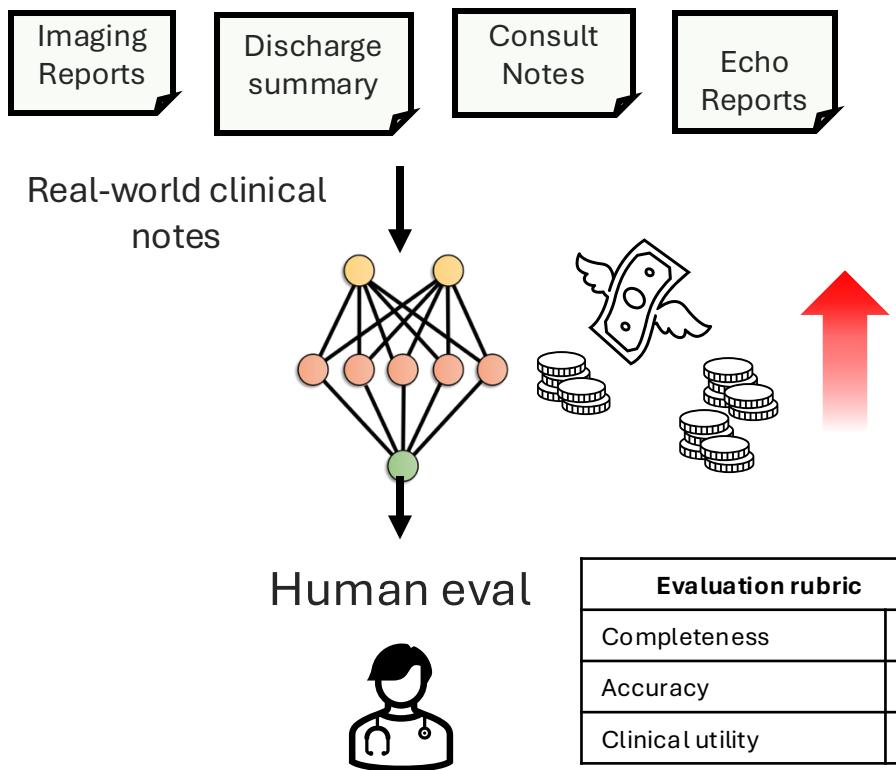
Recipe for Constructing (Valid) Automatic Benchmarks

(1) Human eval in real-world setting

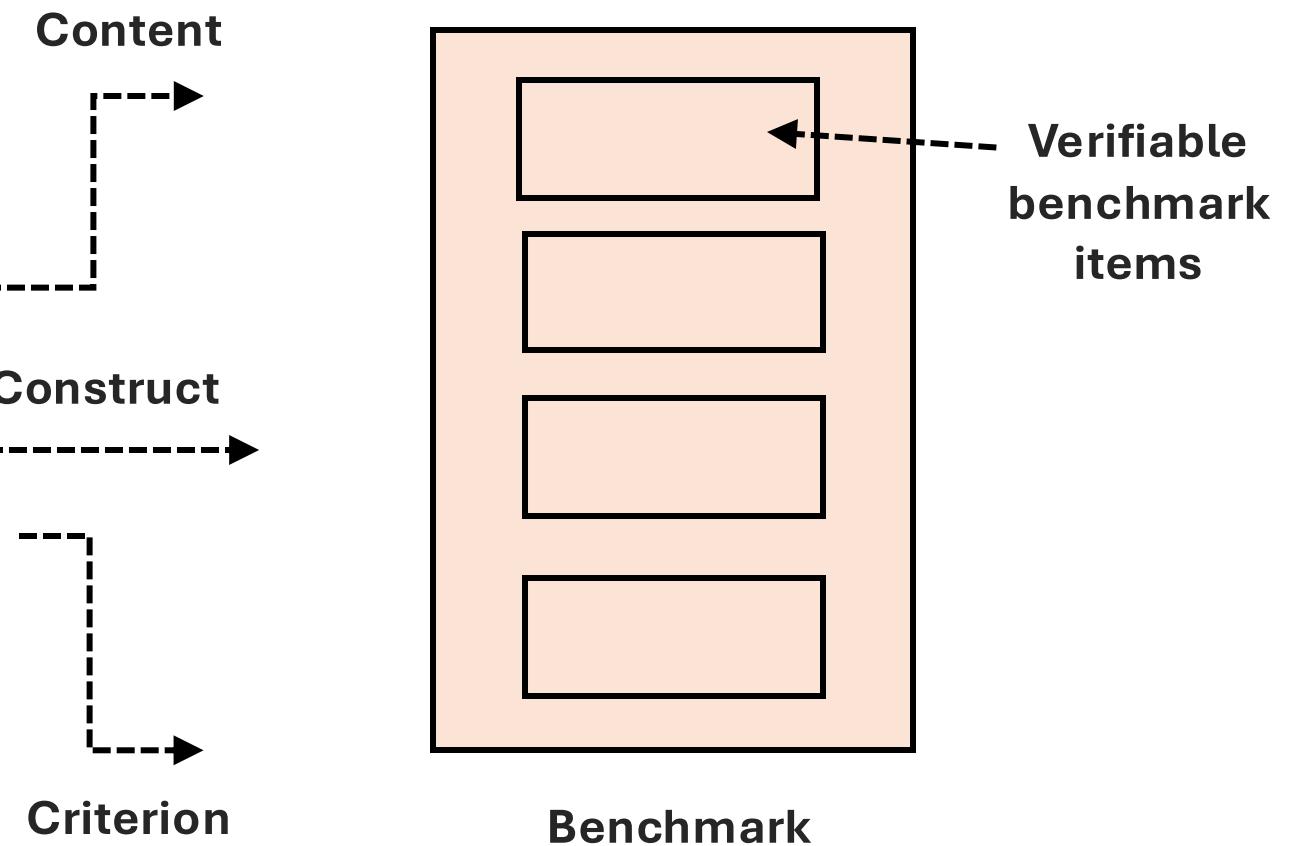


Recipe for Constructing (Valid) Automatic Benchmarks

(1) Human eval in real-world setting

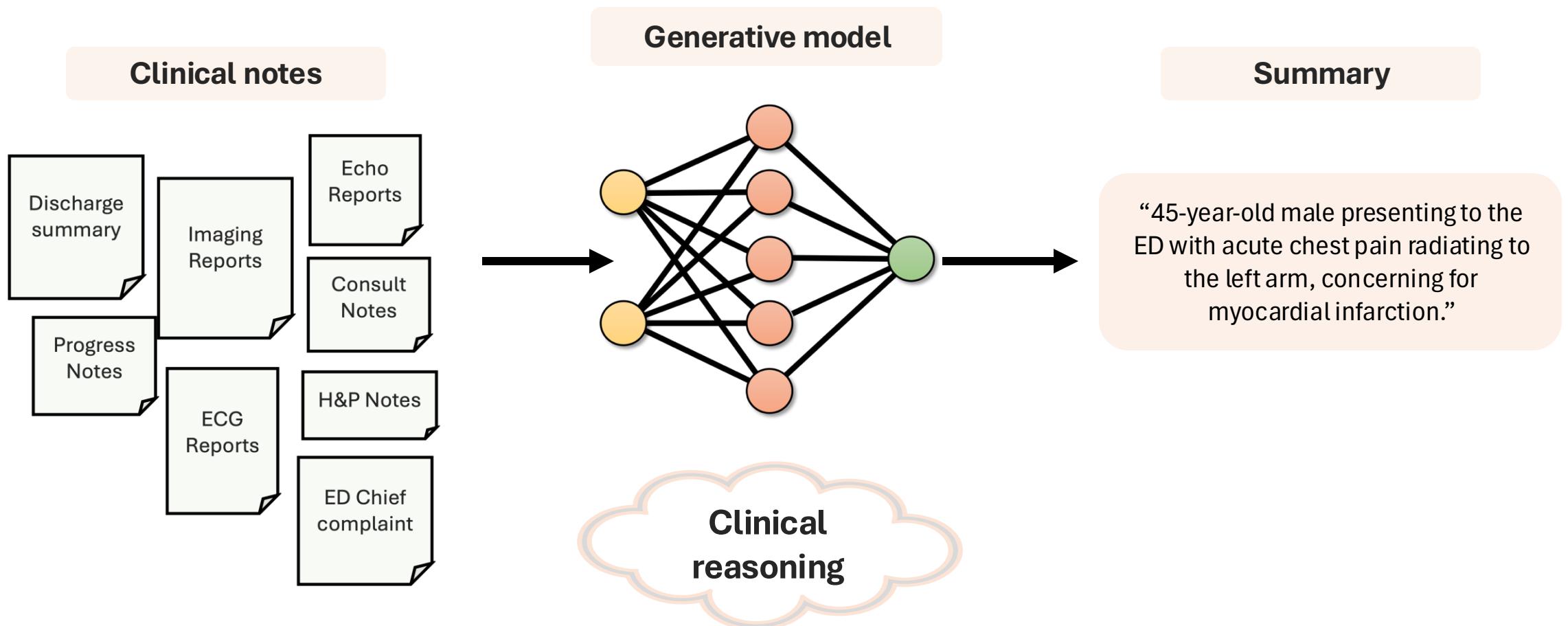


(2) Construct benchmark with measurement validity



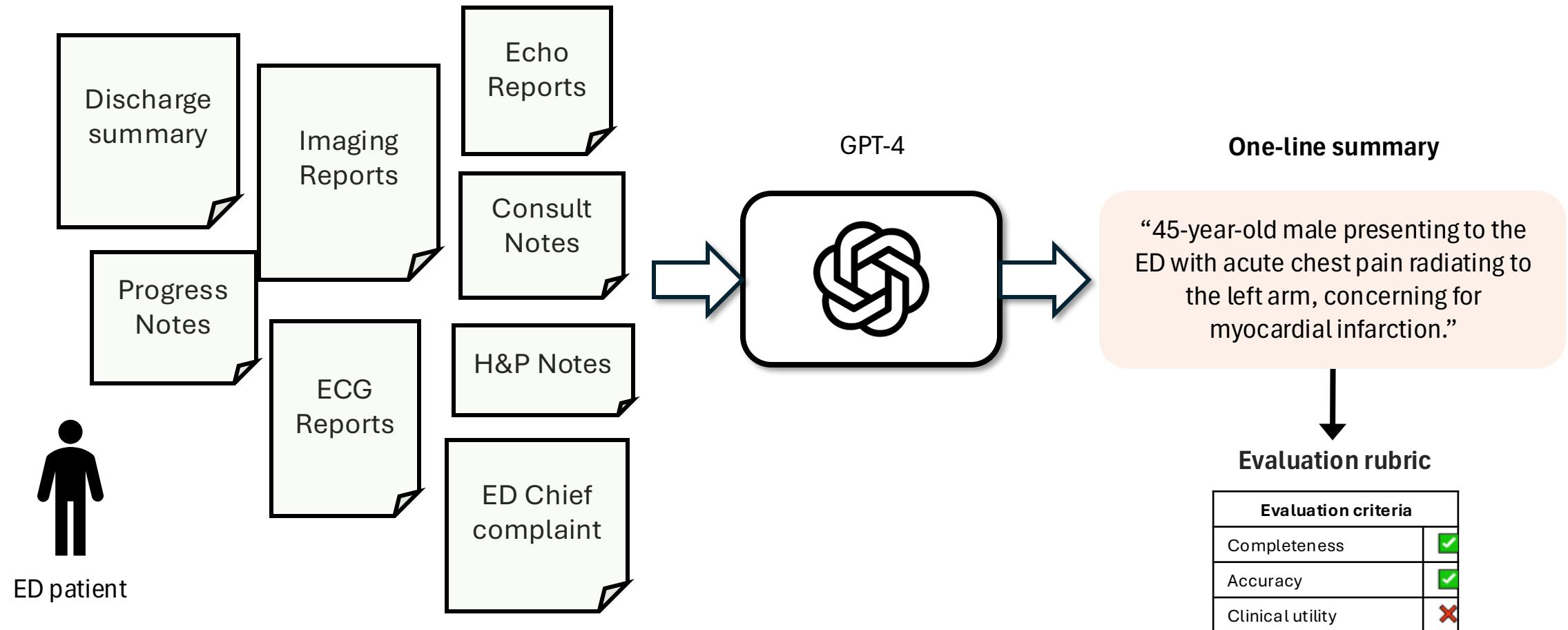
Example: Evaluating Clinical Reasoning in the ER

- Can LLMs engage in clinical reasoning about patient cases in the emergency room?



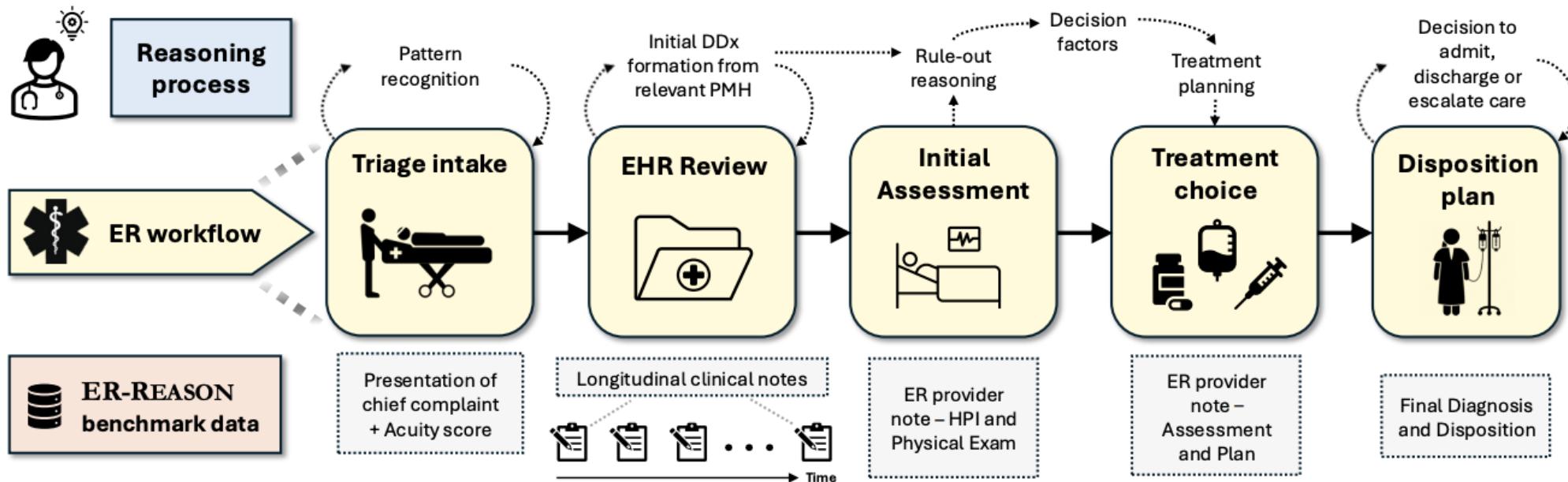
Real-World Deployment of Generative AI at UCSF

- Examples: AI-based one-line summarization of clinical notes in emergency medicine



The ER-Reason Benchmark*

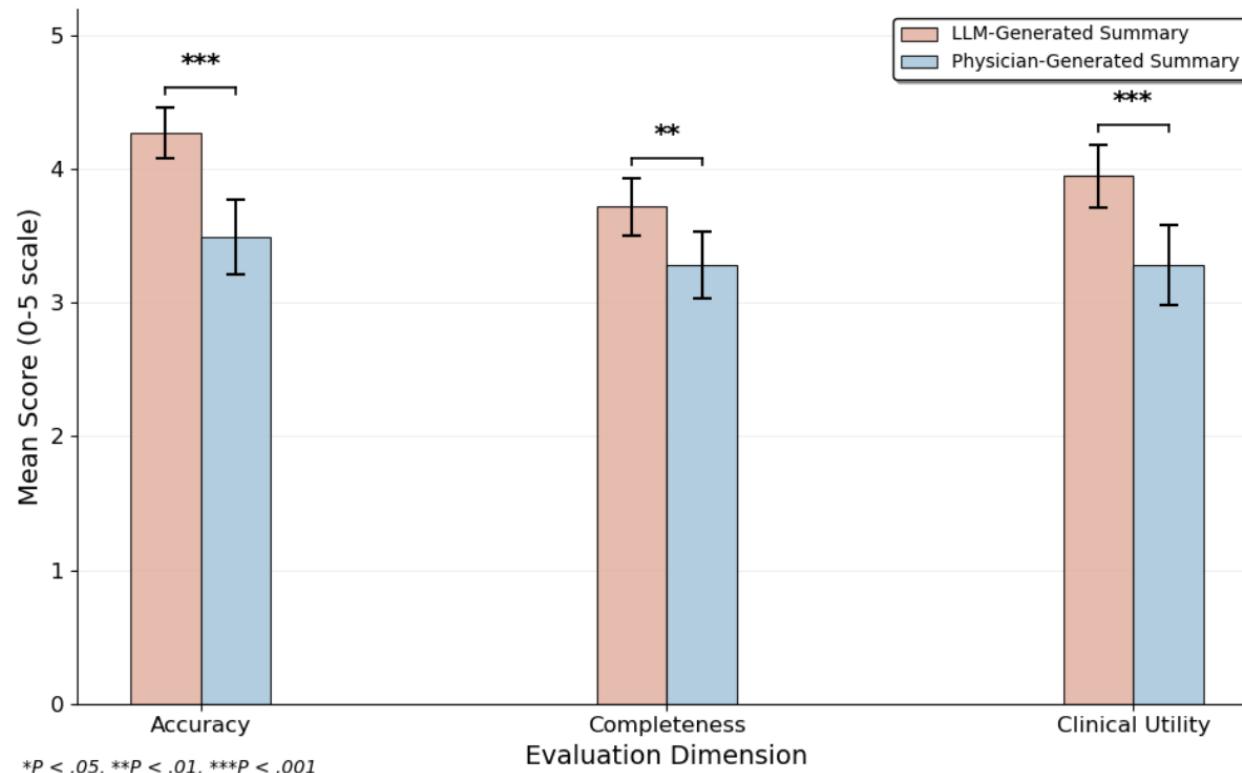
- Data from 3,984 ED patients at UCSF, 25,174 deidentified longitudinal clinical notes.



130 full clinician-authored rationales and human evaluation of clinicians vs. AI

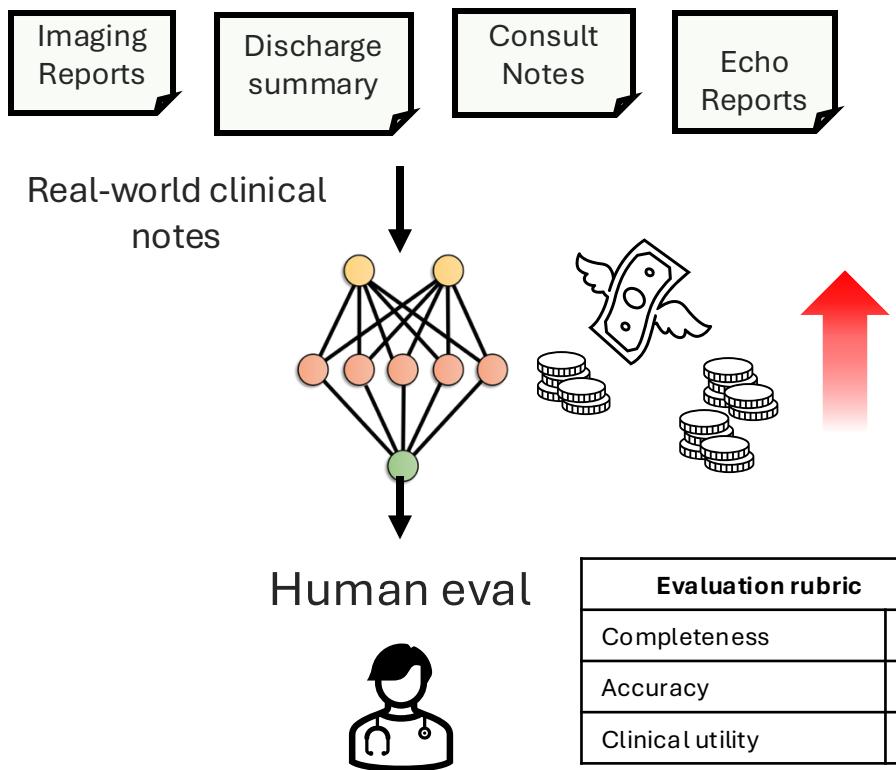
The ER-Reason Benchmark

- Comparison of LLM-generated and Physician-generated one-liner summaries.



Recipe for Constructing (Valid) Automatic Benchmarks

(1) Human eval in real-world setting



(2) Construct benchmark with measurement validity

