Uncharted: The Road to Data-Centric Benchmarking for Medical Foundation Models

Jason Fries, PhD Research Scientist
Center for Biomedical Informatics Research, Stanford University

Stanford MEDICINE | Clinical Excellence Research Center

BMIR
Stanford Center for Biomedical Informatics Research

Stanford AIMI
Center for Artificial Intelligence in Medicine and Imaging

HAI
Stanford University
Human-Centered Artificial Intelligence

The State of AI in Healthcare

Expensive to prototype [1]

>$200,000

Models are rarely deployed [2]

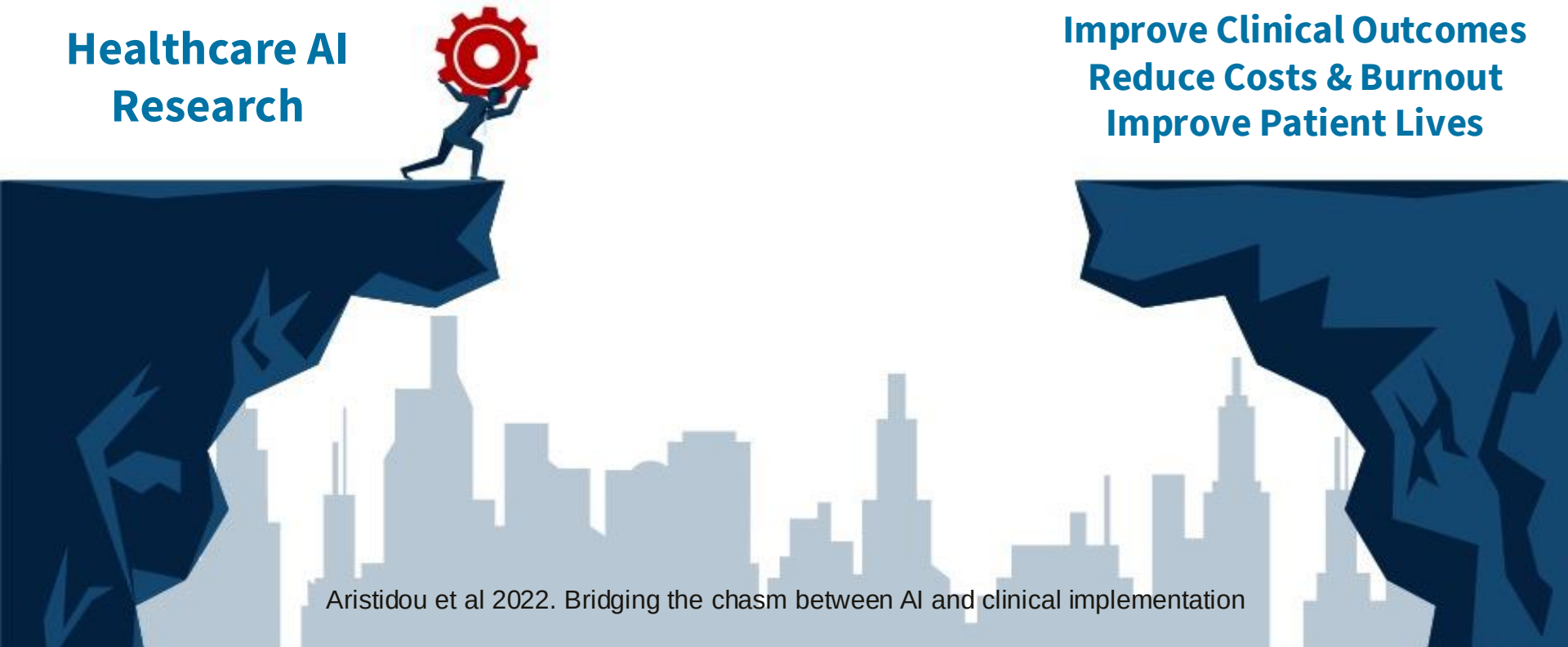
Medical data are **noisy**, replete with **errors, biases, & missingness**

Most AI is **trained and tested** on **cleaned data**

Healthcare AI research suffers from **poor reproducibility** [3]

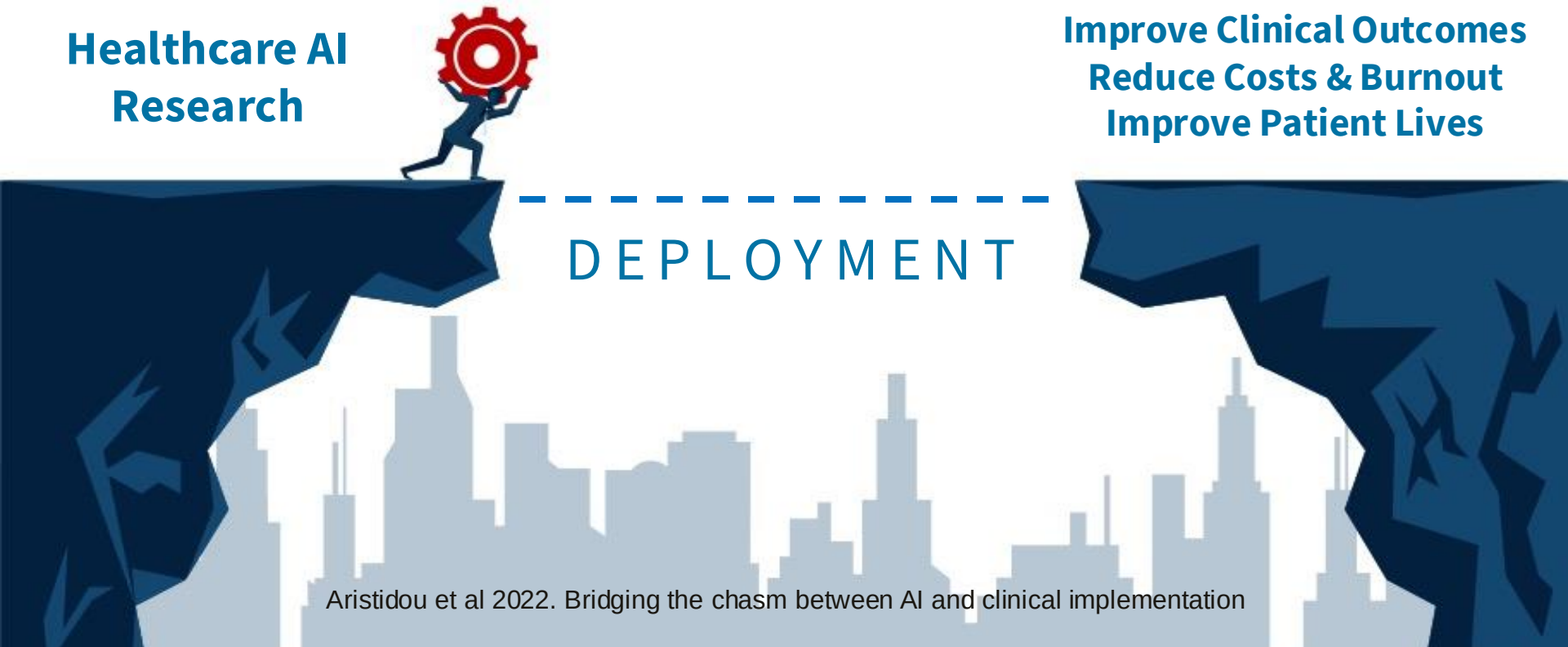
1. Sendak et al. 2017. Barriers to Achieving Economies of Scale in Analysis of EHR Data.
2. Wynants et al. 2020. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal
3. McDermott et al. 2021 Reproducibility in machine learning for health research: Still a ways to go

AI Chasm



Aristidou et al 2022. Bridging the chasm between AI and clinical implementation

AI Chasm



Aristidou et al 2022. Bridging the chasm between AI and clinical implementation

Special Reports > Exclusives

AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE

by Michael DePeau-Wilson, Enterprise & Investigative Writer, MedPage Today
January 19, 2023

FORBES > INNOVATION

What ChatGPT And Other AI Tools Mean For The Future Of Healthcare

Sahil Gupta Forbes Councils Member
Forbes Technology Council
COUNCIL POST | Membership (Fee-Based)

Feb 6, 2023, 08:30am EST

Generative AI
Breaks into the Mainstream

FORBES > INNOVATION > HEALTHCARE

EDITORS' PICK

5 Ways ChatGPT Will Change Healthcare Forever, For Better

Robert Pearl, M.D. Contributor

Follow

UCSF Department of Medicine

ChatGPT: Will It Transform the World of Health Care?

D

Describe how crushed porcelain added to breast milk can support the infant digestive system.

Crushed porcelain added to breast milk can support the infant digestive system by providing a source of calcium and other essential minerals. When added to

NEWS | 18 January 2023

ChatGPT listed as author on research papers: many scientists disapprove

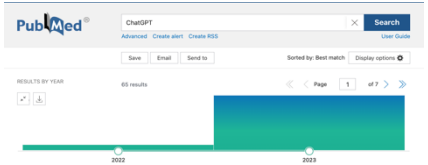
At least four articles credit the AI tool as a co-author, as publishers scramble to regulate its use.

NYMC > News and Events > News Archives

Envisioning the Healthcare Landscape with ChatGPT

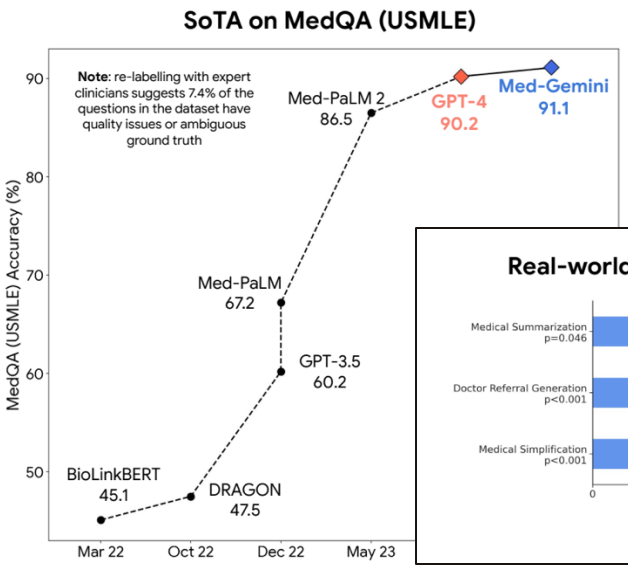
New York Medical College Explores The Opportunities And Risks Of AI On The Healthcare Industry In The Following Article Written Entirely Using ChatGPT

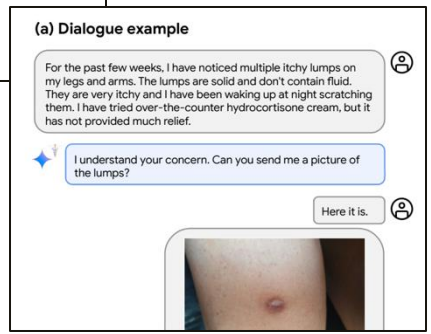
February 13, 2023

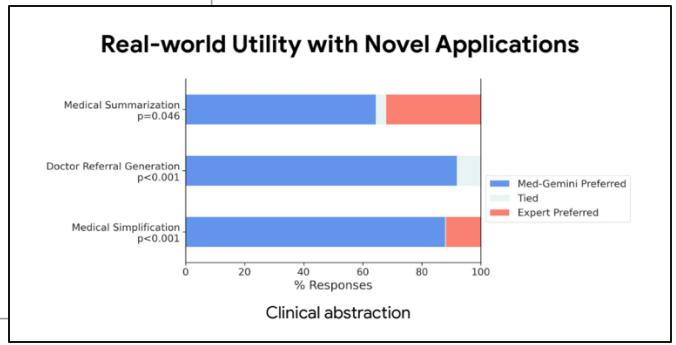


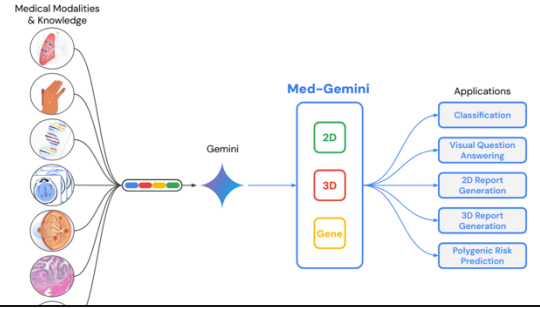
…and their many issues

Excitement & Hype for Generative AI in Medicine





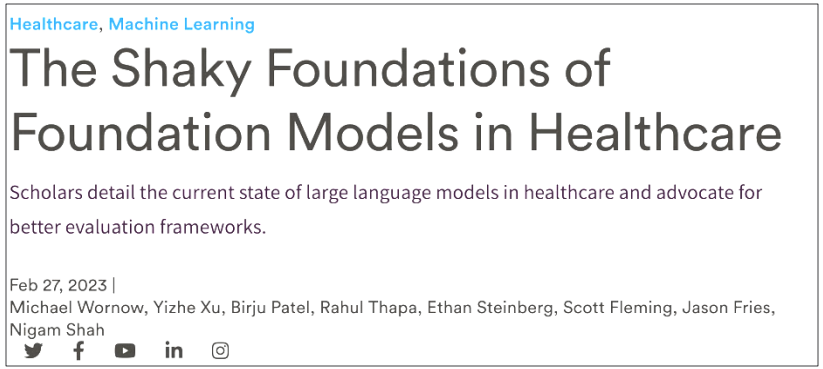




*"**22 datasets** across **five different tasks** and **six distinct medical image modalities**"*

- **Public datasets are highly curated**
- **Do these reflect "real-world" deployment needs?**

Our Thinking on Medical Foundation Models



Enriching the Axes of Evaluation

How Do We Evaluate Generative AI in Medicine?

(Wornow et al. 2023) The shaky foundations of large language models and foundation models for electronic health records

Survey of 84 foundation models trained on non-imaging EHR data

- **Limited clinical datasets (MIMIC-III)** or PubMed
- Evaluation tasks do not provide meaningful insights into deployment
- **No released medical foundations models for EHR data**
- **No shared benchmarks** for comparison and shared sense of SOTA

How Do We Evaluate Generative AI in Medicine?

(Bedi et al. 2024) A Systematic Review of Testing and Evaluation of Healthcare Applications of Large Language Models (LLMs)

Our recent survey paper looked at 519 LLM studies in medicine from (*January 1, 2022 to February 19, 2024*)

- Only **5% of studies used real patient care data** for evaluation
- **95.4%** used **accuracy as the primary evaluation** dimension
- Limited evaluation of fairness/bias (15.8%), robustness (14.8%), deployment (4.6%)

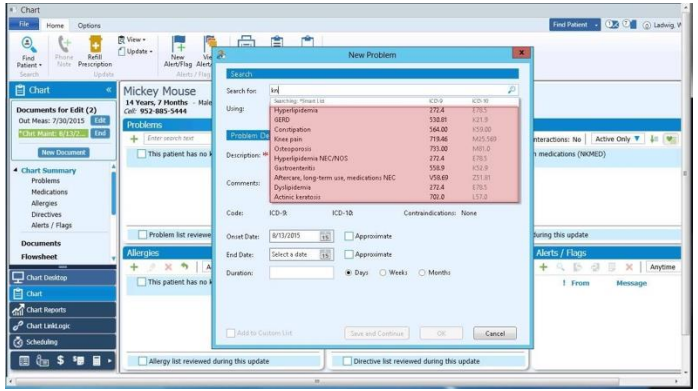
Tenets of Data-Centric Benchmarking for Healthcare

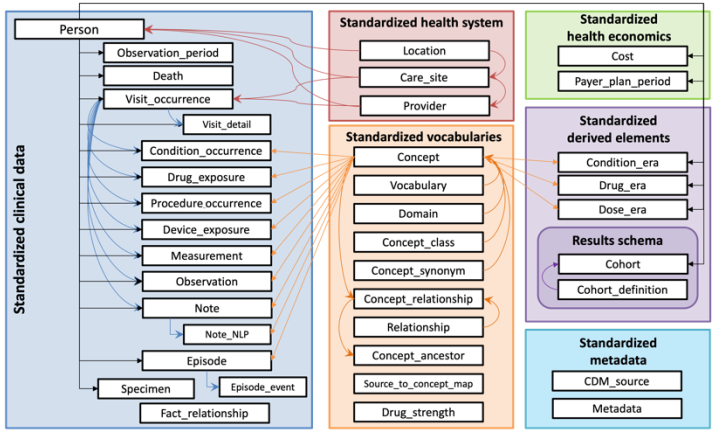
- Use **Diverse, Realistic Patient Data and Tasks**
- **Transparency is a First Class Citizen**
  - Report Provenance for Data, Training Mixtures
  - Accessible/Inspectable Model Weights
- **Report Metrics Beyond Accuracy**, F1, ROUGE, etc.
- Evaluate **Systems for Human-AI Collaboration**

EHR Data & Tasks

Electronic Health Record (EHR)

Digital Record of Patient Care





EHR interface for healthcare workers

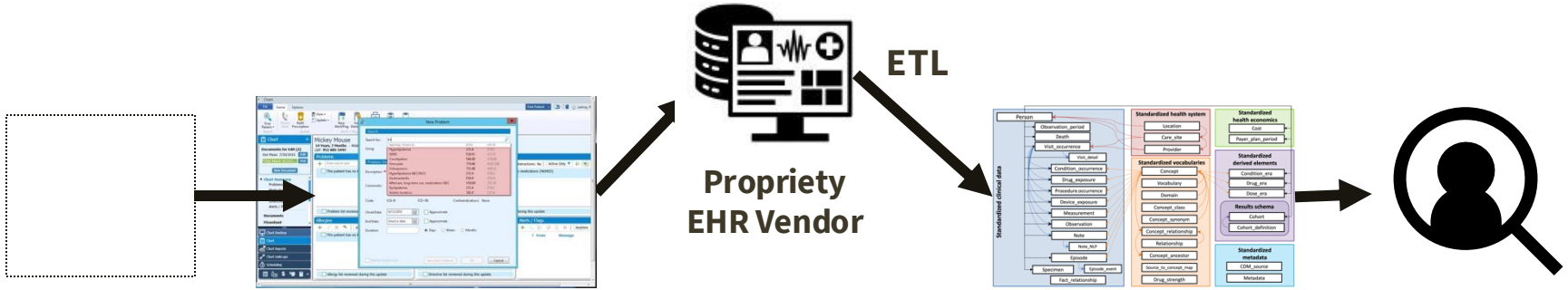
- GUI-based
- Data portal for a patient's data
- Focus on a single patient at a time

EHR interface for data scientists

- Relational database structure
- Some data model (Epic, OMOP, i2b2)
- Often *transformed* in un-inspectable ways

Electronic Health Record (EHR)

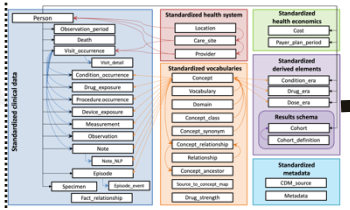
Provenance of Data Curation if Critical for ML and **Difficult to Track in Healthcare**



Electronic Health Record (EHR)

Provenance of Data Curation if Critical for ML and **Difficult to Track in Healthcare**

Invisible to the ML Researcher

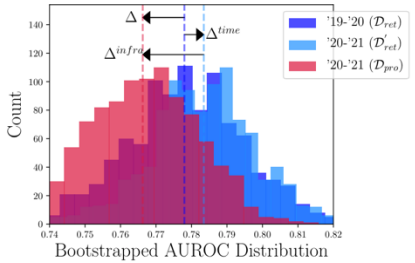


Common Data Model

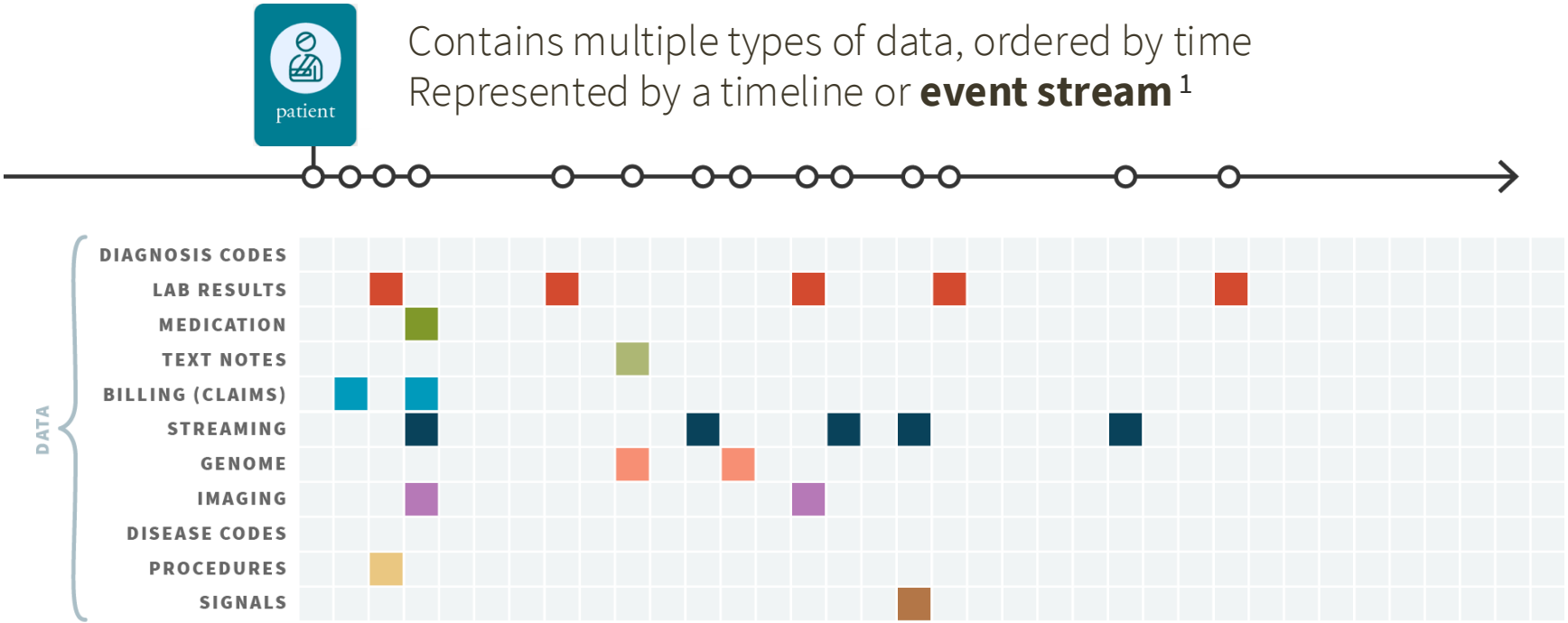
Researcher

"infrastructure shift", i.e., changes in access, extraction and transformation of data (***Otles et al. 2021)***

infrastructure shift > temporal shift

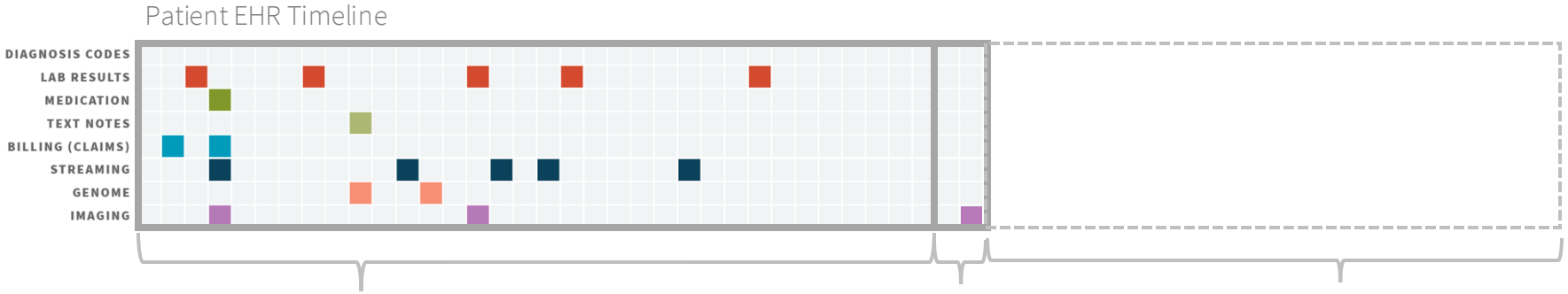


Electronic Health Record (EHR) Data is Multimodal



1. McDermott et al. 2024. Event Stream GPT: A Data Pre-processing and Modeling Library for Generative, Pre-trained Transformers over Continuous-time Sequences of Complex Events

AI to Enhance Medical Decision Making



Example ML Applications

What Occurred in the Past?

- Chart Summarization
- Cohort Construction
- Training Data Construction

What is Occurring Now?

- Identify blood clots in lung CT scans
- Identify cancerous cells in pathology slides

Predict Future Risks & Intervention Benefits

- Will patient develop nephritis?
- Will patient develop chronic pulmonary hypertension?

Whether to Treat

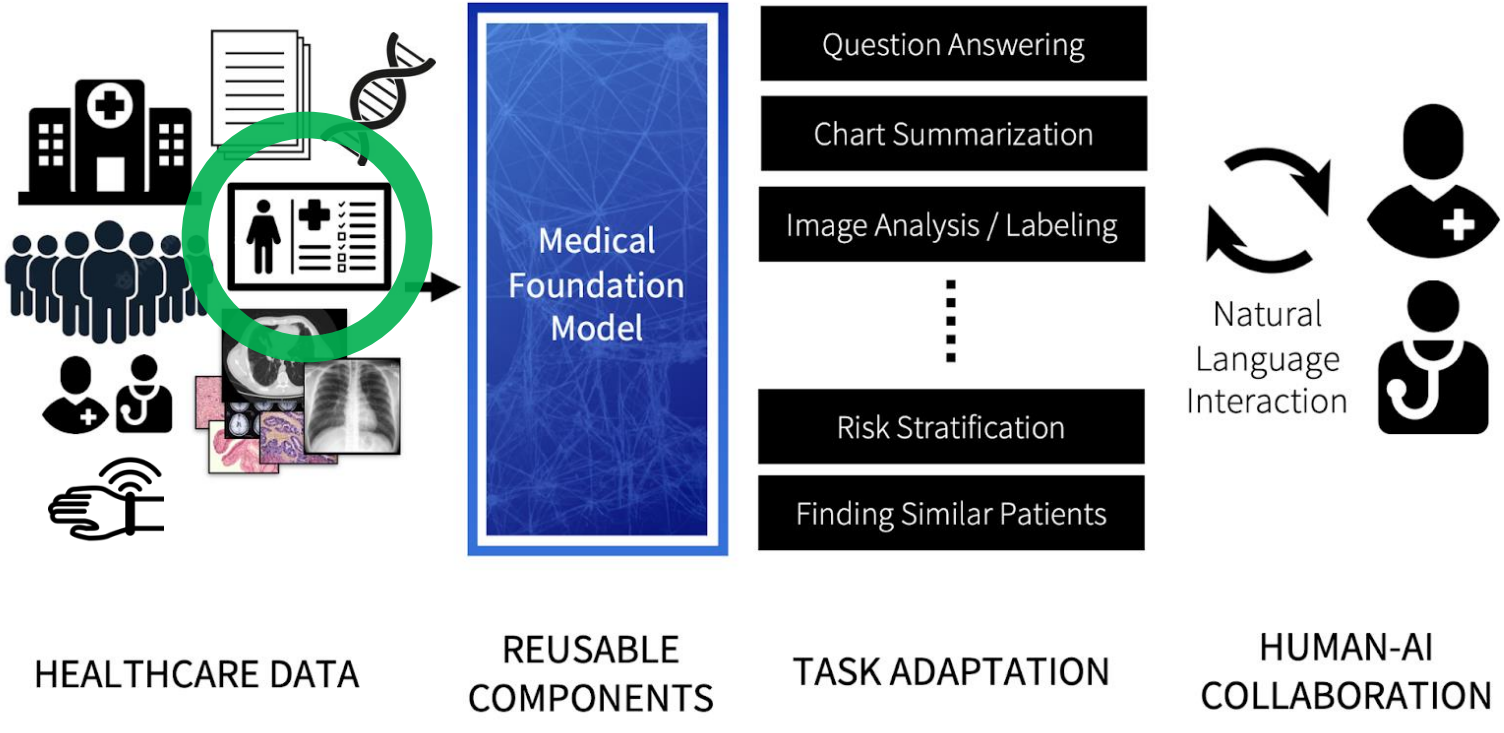
How to Treat

subject to

- Policy
- Capacity to Act
- Intervention Properties

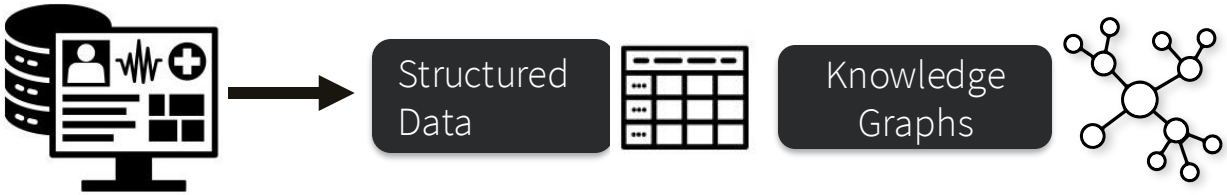
Model & Benchmark Releases

Medical Foundation Models



Our EHR Foundation Model Work

Self-Supervised Methods Development



CLMBR: Clinical language modeling-based representations **2021**

Pretraining: **Autoregressive**

Journal of Biomedical Informatics 2021

MOTOR: Many Outcome Time Oriented Representations **2024**

Pretraining: **Time-to-Event**

ICLR 2024 **SPOTLIGHT**

Open & Accessible Model Weights

Sharing pre-trained model

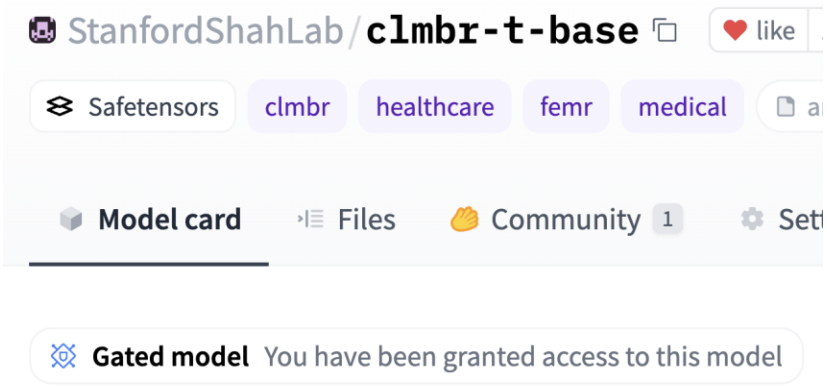
Initially we really hoped to share our models but unfortunately, the pre-trained models are no longer sharable. According to SBMI Data Service Office: "Under the terms of our contracts with data vendors, we are not permitted to share any of the data utilized in our publications, as well as large models derived from those data."

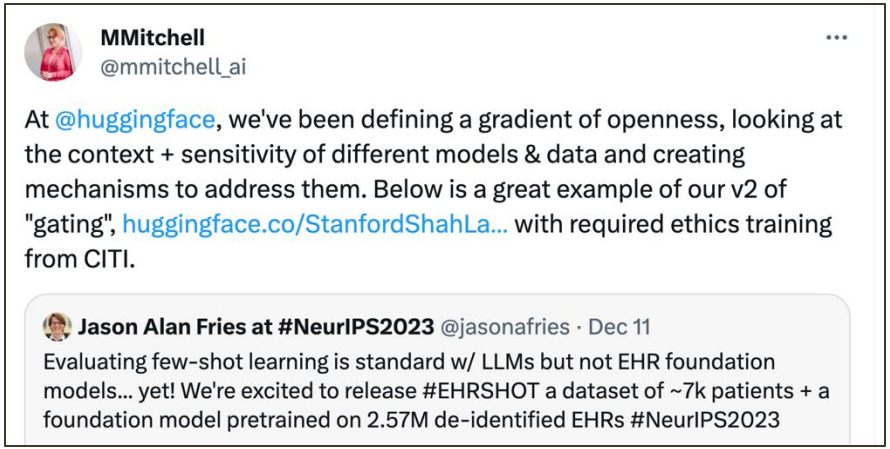
https://github.com/ZhiGroup/Med-BERT

Transfer learning is the primary
value prop of foundation models!

Foundation Models Risk Increasing our Reproducibility Crisis

Enabling Open Science





Our first model hub release!

- Gated model on Hugging Face
- Requires **CITI ethics training**
- **Non-commercial use only**

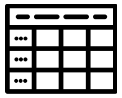
Margaret Mitchell

Chief AI Ethics Scientist, Hugging Face

Releasing New Medical Datasets

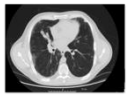
EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models **2023**

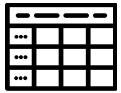
6,739 Patients



INSPECT: A Multimodal Dataset for Patient Outcome Prediction of Pulmonary Embolisms **2023**

19,402 Patients

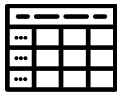




NeurIPS Datasets & Benchmarks 2023

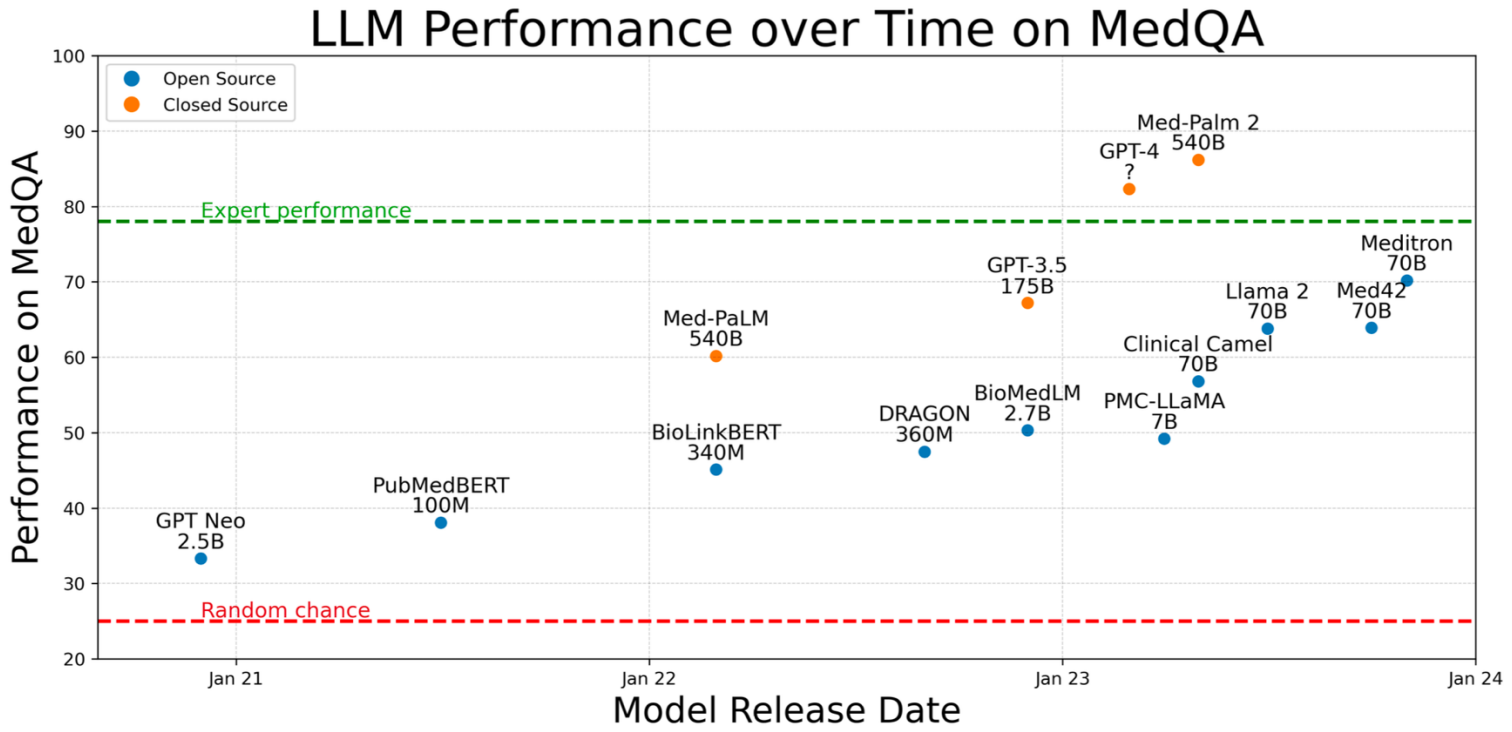
MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records **2023**

267 Patients



AAAI 2024

LLM Medical Knowledge via USMLE-Style Questions



MedQA Does Not Reflect Realistic EHR Data

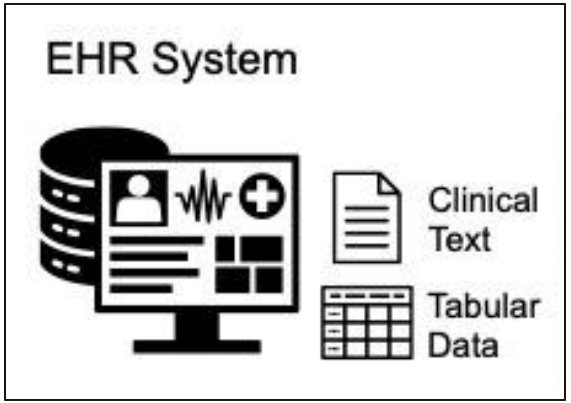
MedQA

Question: A 35-year-old man is brought to the emergency department by a friend 30 minutes after the sudden onset of right-sided weakness and difficulty speaking. […] **Which of the following is the most appropriate next step in diagnosis?**

(A) Echocardiography with bubble study
(B) Adenosine stress test
(C) Cardiac catheterization
(D) Cardiac MRI with gadolinium
(E) CT angiography

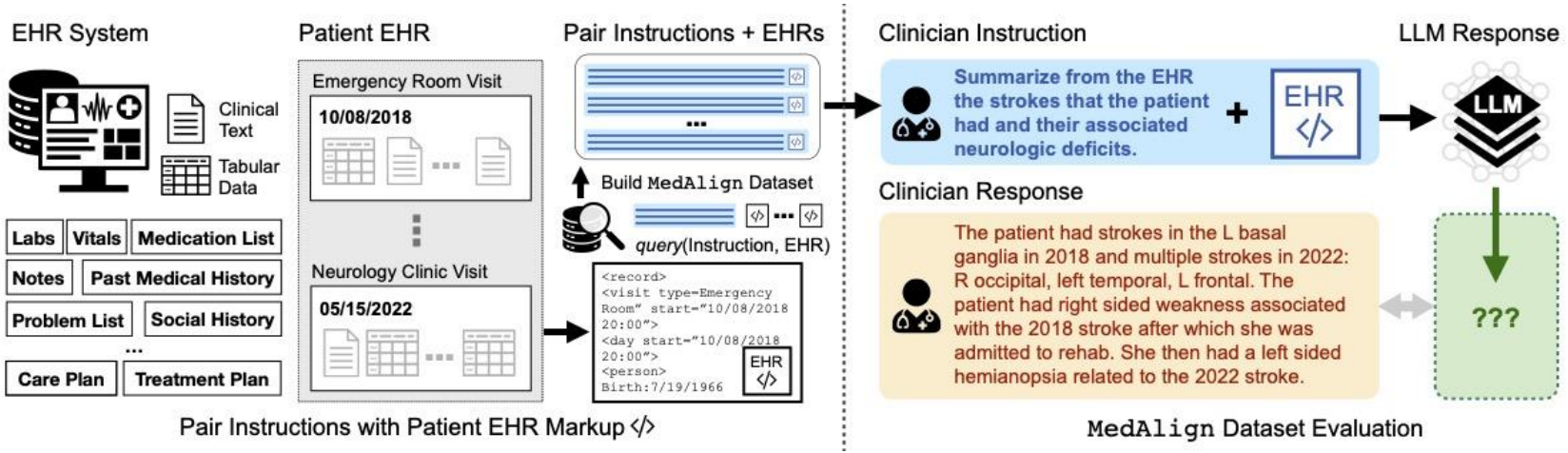
USMLE
United States Medical Licensing Exam

What if…



33k to 1.6M tokens per patient

MedAlign: Clinical Instruction Following Benchmark



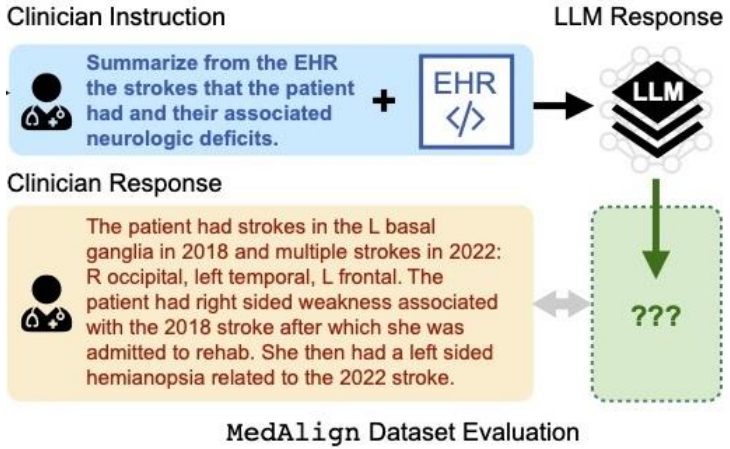
MedAlign: A Clinician-Generated Benchmark Dataset for Instruction Following with Electronic Medical Records [1]

- **15** clinicians / **7** specialties
- **267 Longitudinal** Patient Timelines
- **983 instructions**, **303 responses**
- Assess **real information needs**

[1] Fleming et al. “A Clinician-Generated Benchmark Dataset for Instruction Following with Electronic Medical Records”. *AAAI*. 2024.

MedAlign: Clinical Instruction Following Benchmark

Ground prompt in “complete”
longitudinal health record



Instruction Tuning: Aligning with Clinical Needs

Model	Context	Correct ↑	WR ↑	Rank ↓
GPT-4 (MR)	32768†	**65.0%**	0.658	2.80
GPT-4	32768	60.1%	**0.676**	**2.75**
GPT-4	2048*	51.8%	0.598	3.11
Vicuña-13B	2048	35.0%	0.401	3.92
Vicuña-7B	2048	33.3%	0.398	3.93
MPT-7B-Instruct	2048	31.7%	0.269	4.49

GPT-4 **35% Error Rate**

Aligning with Clinician Information Needs

Table 2: MEDALIGN instruction categories and example instructions.

Category	Example Instruction	Gold	All
Retrieve & Summarize	Summarize the most recent annual physical with the PCP	223	667
Care Planning	Summarize the asthma care plan for this patient including relevant diagnostic testing, exacerbation history, and treatments	22	136
Calculation & Scoring	Identify the risk of stroke in the next 7 days for this TIA patient	13	70
Diagnosis Support	Based on the information I've included under HPI, what is a reasonable differential diagnosis?	4	33
Translation	I have a patient that speaks only French. Please translate these FDG-PET exam preparation instructions for her	0	2
Other	What patients on my service should be prioritized for discharge today?	41	75
Total		303	983

Clinicians spend 49% of their day interacting with EHRs

~66% of instructions were **"retrieve & summarize"** data from the EHR.

Aligning with Clinician Information Needs

Table 2: MEDALIGN instruction categories and example instructions.

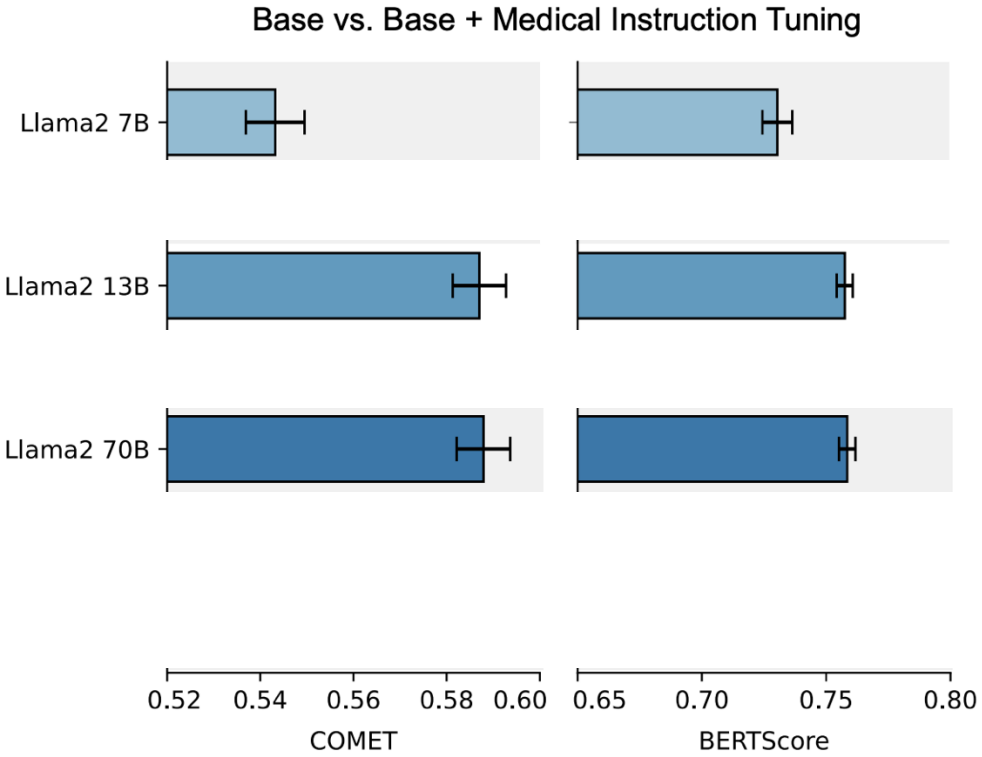
Category	Example Instruction	Gold	All
Retrieve & Summarize	Summarize the most recent annual physical with the PCP	223	667
Care Planning	Summarize the asthma care plan for this patient including relevant diagnostic testing, exacerbation history, and treatments	22	136
Calculation & Scoring	Identify the risk of stroke in the next 7 days for this TIA patient	13	70
Diagnosis Support	Based on the information I've included under HPI, what is a reasonable differential diagnosis?	4	33
Translation	I have a patient that speaks only French. Please translate these FDG-PET exam preparation instructions for her	0	2
Other	What patients on my service should be prioritized for discharge today?	41	75
Total		303	983

Clinicians spend 49% of their day interacting with EHRs

~66% of instructions were **"retrieve & summarize"** data from the EHR.

Only **17%** represent **MedQA-type questions**

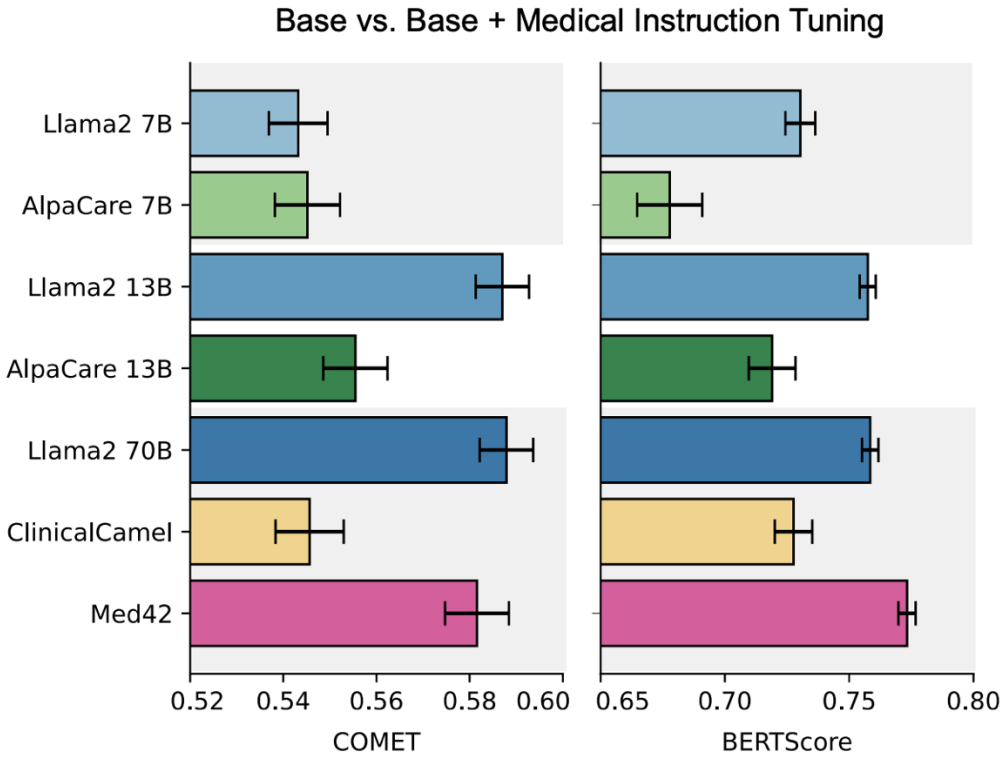
Chasing Benchmarks Not Realistic Use Cases



Auto-evaluation Metrics

Several **medical Llama-2 instruction tuned** models…

Chasing Benchmarks Not Realistic Use Cases



Auto-evaluation Metrics

Several **medical Llama-2 instruction tuned** models…

Here instruction tuning for MedQA-style tasks **largely hurt performance on MedAlign**

The Road Ahead

Open Weights are Critical to Fair & Secure Models

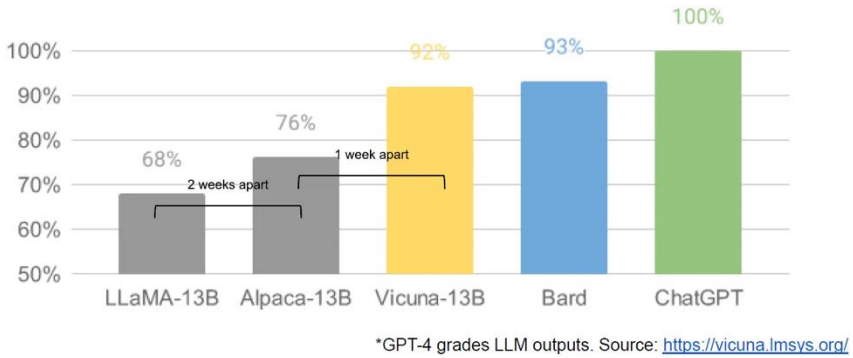
Transparency (training data, model weights) is critical for fair and secure models

OpenAI

"Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."

Calls for the Academic Community

Smaller Models, Cheaper to Train



*GPT-4 grades LLM outputs. Source: https://vicuna.lmsys.org/

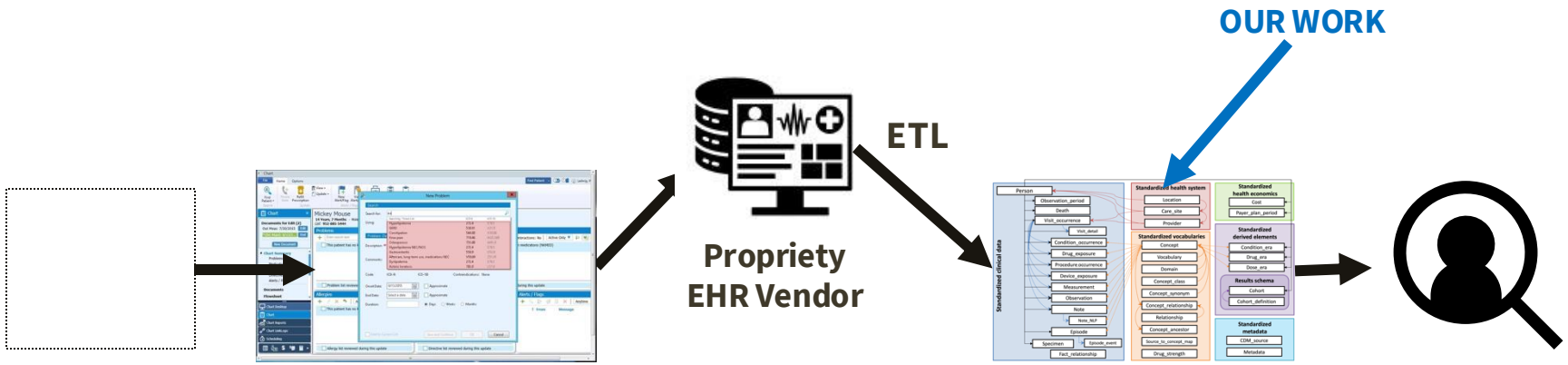
Lead Building Open, Reproducible Medical Base Models

Reimagine Model Evaluation

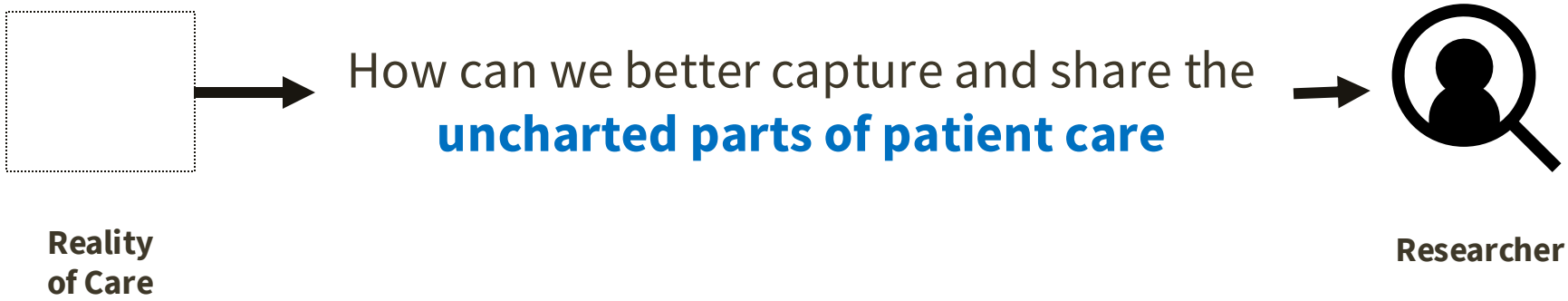
AI will augment existing roles

We need to **measure human + AI performance**

Data-Centric Benchmarks Must Reflect Data Realities



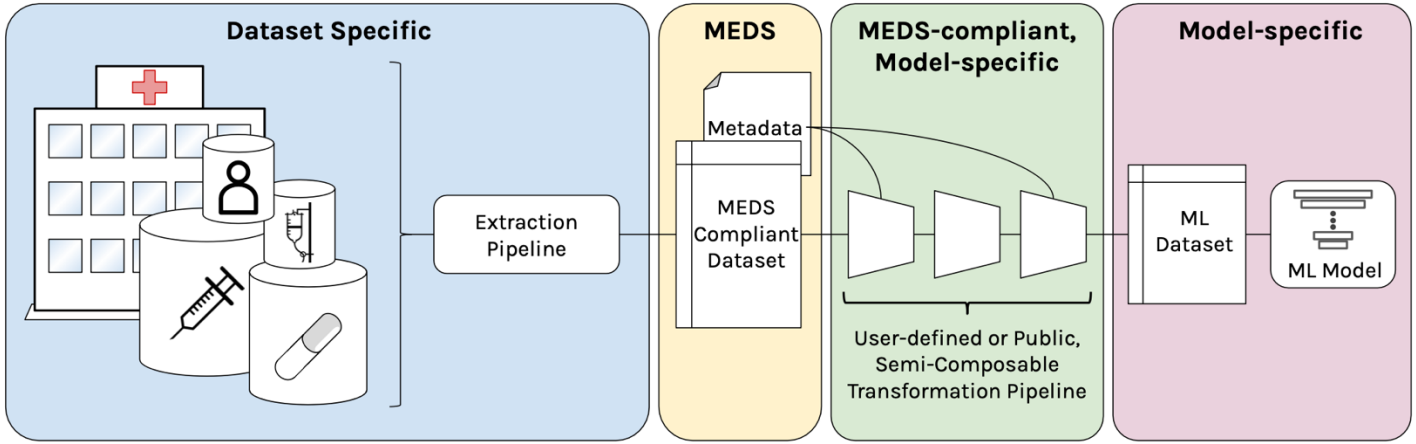
Data-Centric Benchmarks Must Reflect Data Realities



Thank You!

jason-fries@stanford.edu

Medical Event Data Standard (MEDS)



PROPOSAL - Data Schema for ML Developers

Bert Arnrich, Edward Choi, Jason A. Fries, Matthew B. A. McDermott, Jungwoo Oh, Tom J Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, Robin van de Water

https://github.com/Medical-Event-Data-Standard/meds