# Model Hubs for Medical AI: How Far is Our 🤗 Moment?

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

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

Cost to prototype
**a single model** [1]
**>$200,000**

**Models are rarely deployed**
593 COVID-19 models
**virtually none were deployed** [2]

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

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

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

# AI Chasm

**Healthcare AI Research**

**Improve Clinical Outcomes**
**Reduce Costs & Burnout**
**Improve Patient Lives**

# AI Chasm

**Healthcare AI Research**

**Improve Clinical Outcomes**
**Reduce Costs & Burnout**
**Improve Patient Lives**

DEPLOYMENT

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

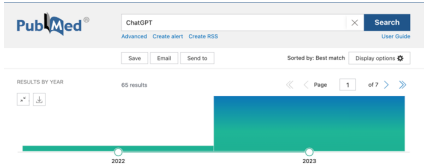
**At least four articles credit the AI tool as a co-author, as publishers scramble to regulate its use.**

NYMC > News and Events > News Archives

## Envisioning the Healthcare Landscape with ChatGPT

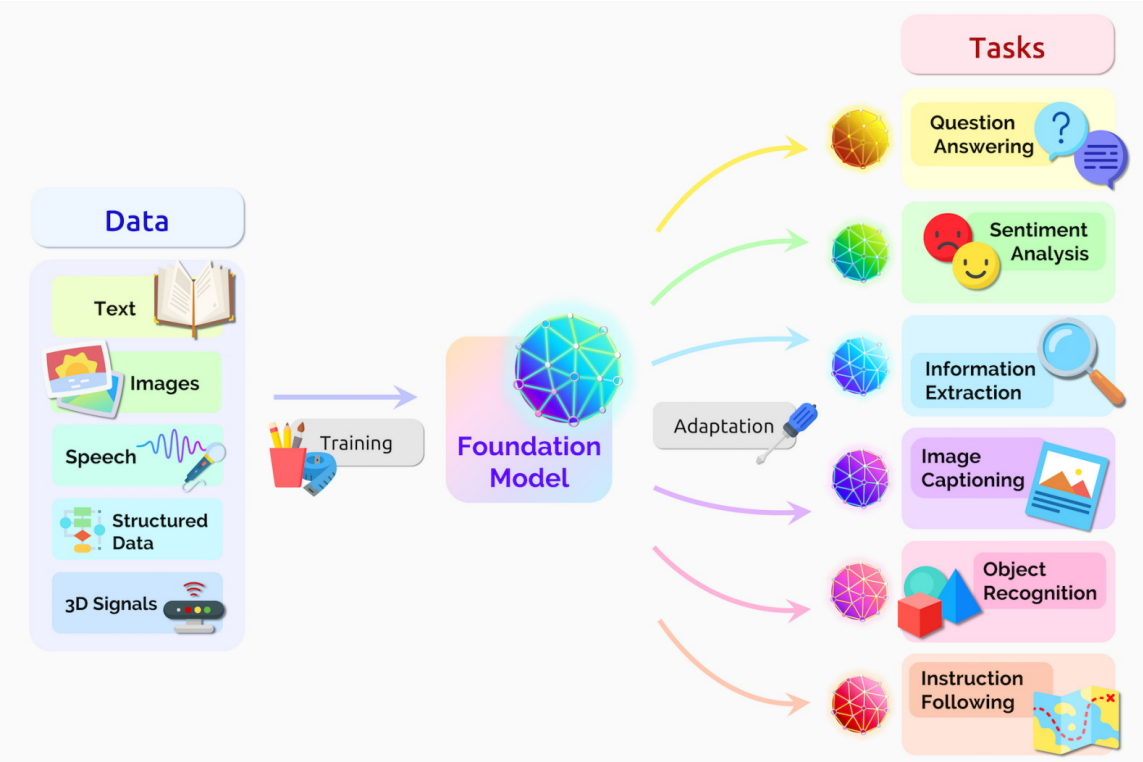
New York Medical College Explores The Opportunities And Risks Of AI On The Healthcare Industry In The Following Article Written Entirely Using ChatGPT

February 13, 2023



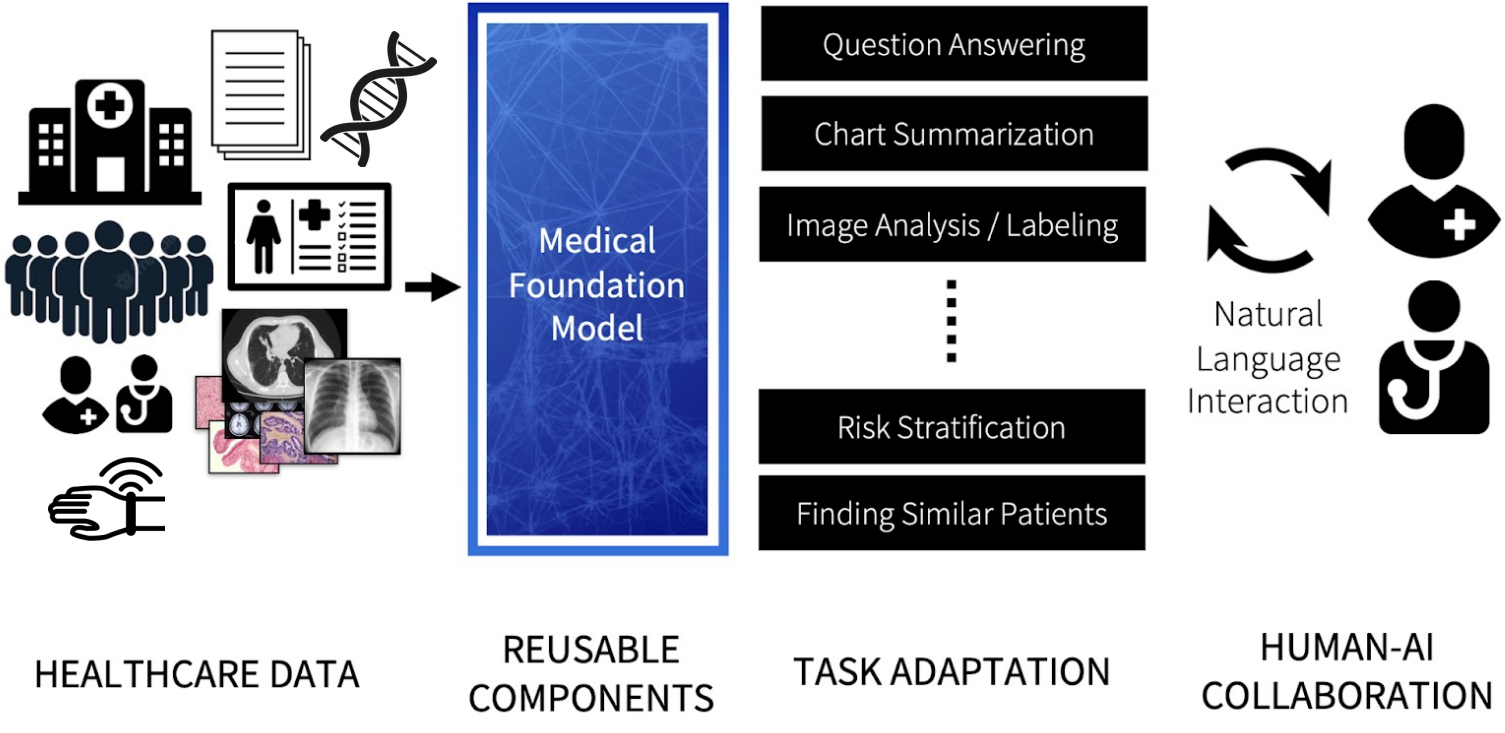

…and their many issues

# Foundation Models and AI's "Industrial Age"




Bommasani et al. "On the Opportunities and Risks of Foundation Models"

# Foundation Models and AI's "Industrial Age"



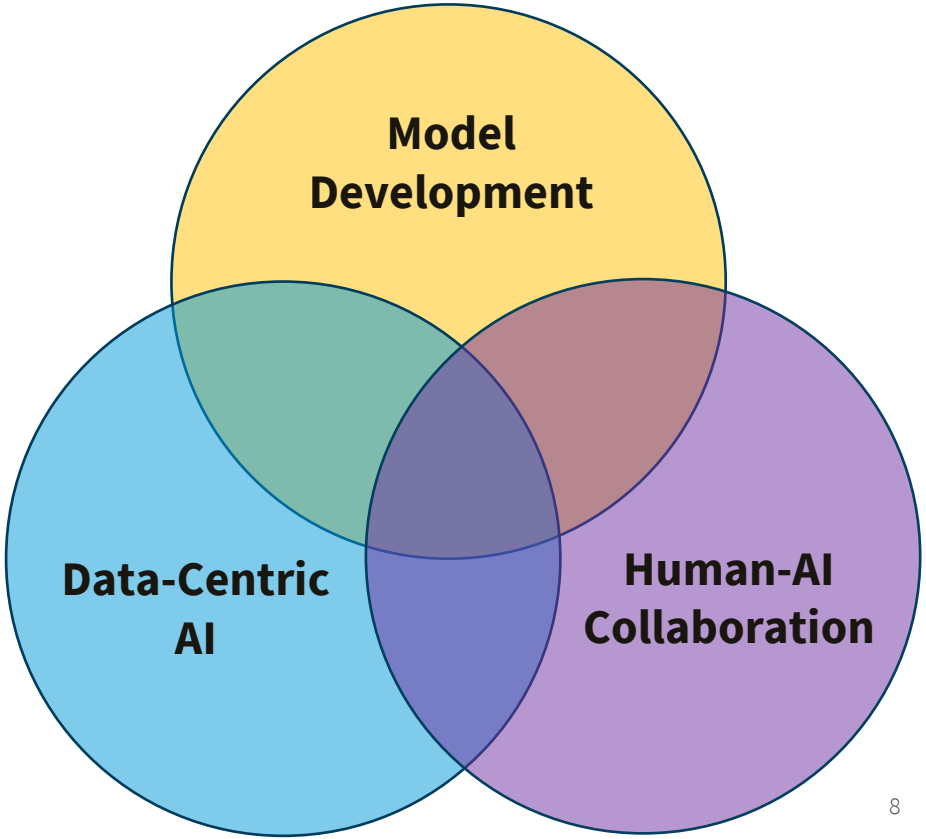
# Foundation Models - Bridging the AI Chasm

Exploring Novel Architectures
Pretraining Objectives
Capturing Multimodality

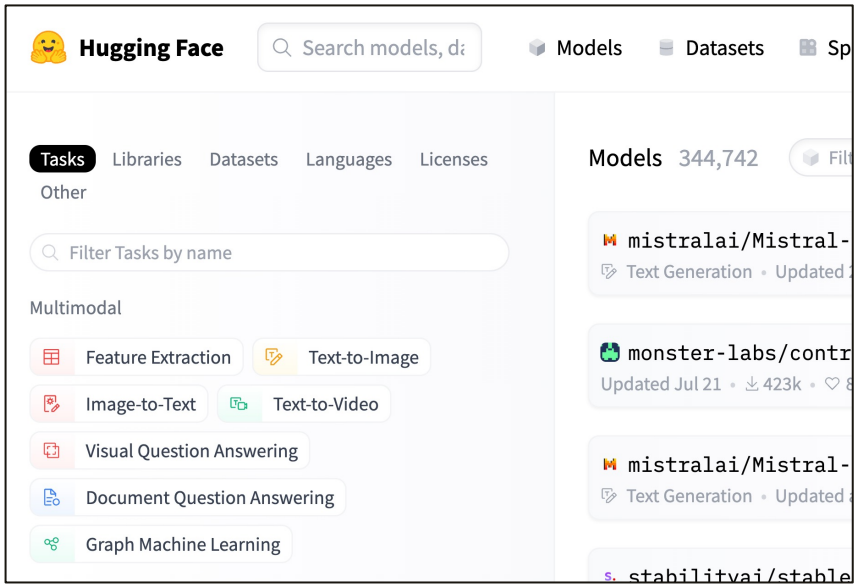
**Systematic methods for data curation, evaluation, and generation**

- Choosing Training Examples
- Automating Evaluation
- Weakly Supervised Model Training

Aligning Models with Real Needs
Optimizing Feedback Loops
Improving Model Trust



# Model Hubs for Medicine




**Transforming AI**

Healthcare

## How Foundation Models Can Advance AI in Healthcare

This new class of models may lead to more affordable, easily adaptable health AI.

Dec 15, 2022 |
Jason Fries, Ethan Steinberg, Scott Fleming, Michael Wornow, Yizhe Xu, Keith Morse, Dev Dash, Nigam Shah

https://tinyurl.com/FM-in-HC

**When is medicine's "Hugging Face" moment?**

# **Building** Foundation Models for Healthcare

# Medical Foundation Models

HEALTHCARE DATA

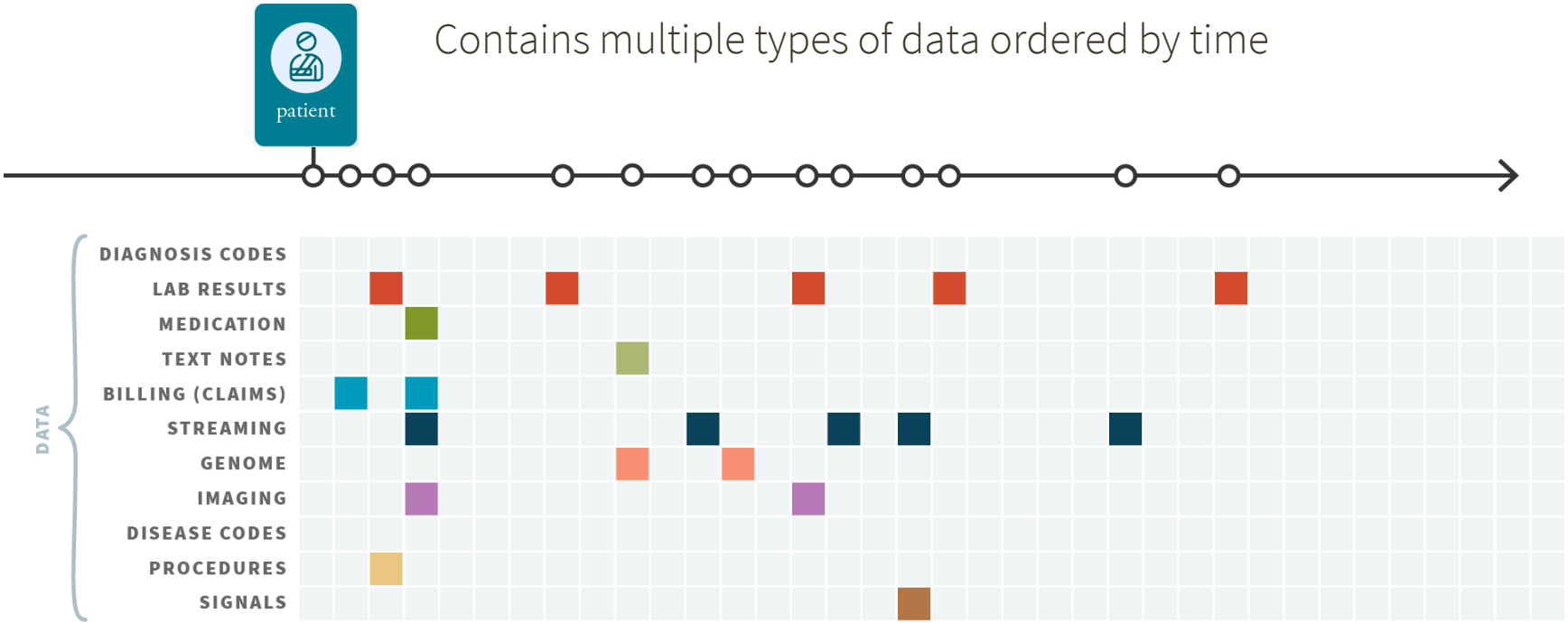
- **Natural Language**
- **2D Imaging**

# Foundation Models and AI's "Industrial Age"

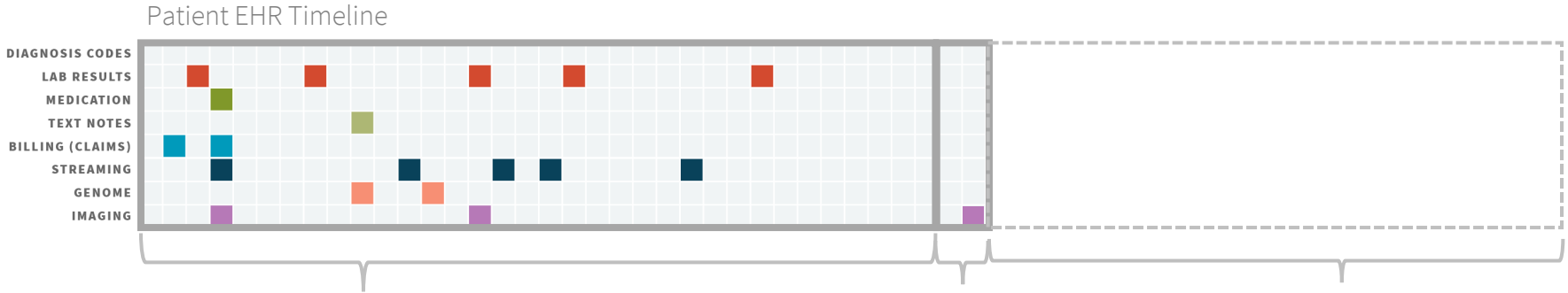
HEALTHCARE DATA

- **Electronic Health Records**

# Electronic Health Record (EHR) Data is Multimodal



# AI to Enhance Medical Decision Making




Example ML Applications

**What Occurred in the Past?**

- Chart Summarization
- Cohort 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?

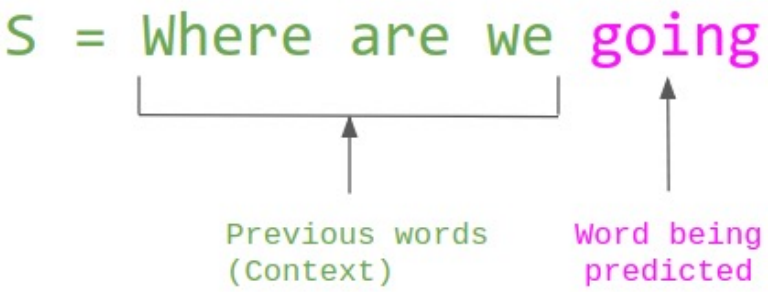
**Whether to Treat** | **How to Treat**

subject to

Policy | Capacity to Act

Intervention Properties

# Language Modeling 101

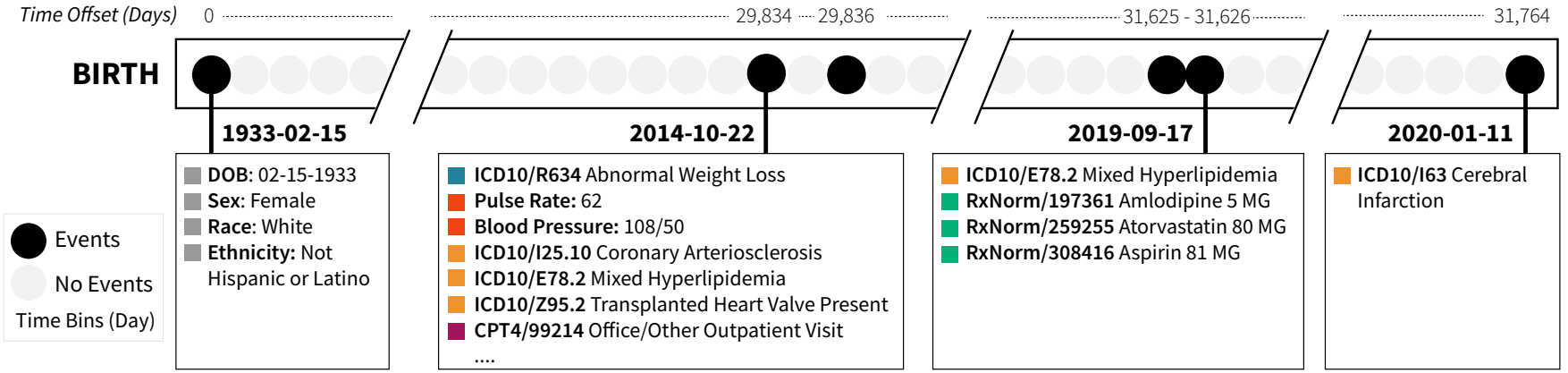



P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

$$P_{(w_1,w_2,\ldots,w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)...p(w_n|w_1,w_2,..,w_{n-1})$$

$$= \prod_{i=1}^{n} p(w_i|w_1,...,w_{i-1})$$

# Structured EHRs form a “Language”

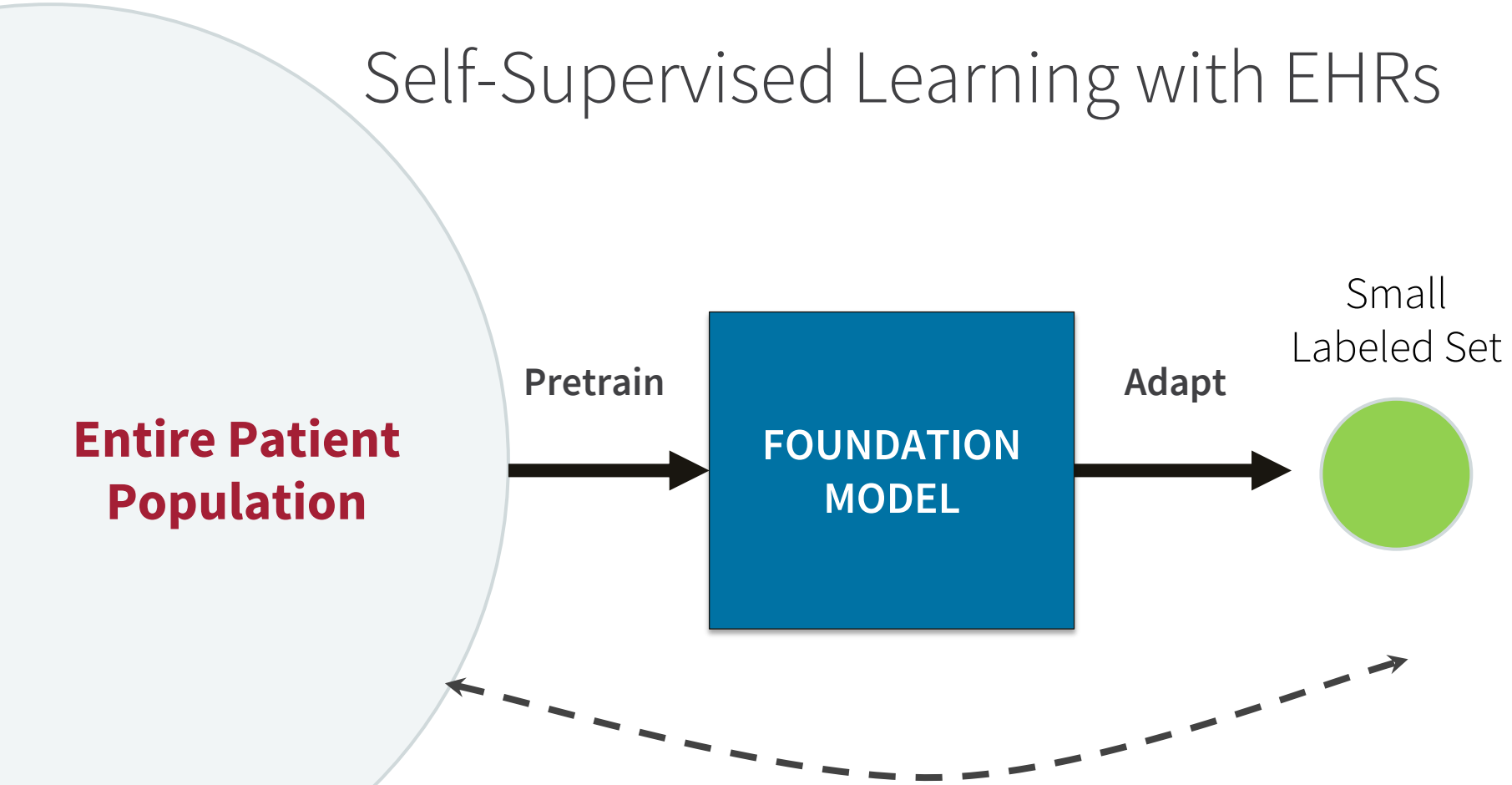



`S = ICD10/R634, ICD10/E78.2, ICD10/Z95.2, CPT4/99214 … ICD10/I63`

Map all medical codes to a finite symbol vocabulary (i.e., analog of words)

$$\mathrm{P(S)} = \prod_{i=1}^{N} P(w_i | w_1 : w_{i-1})$$

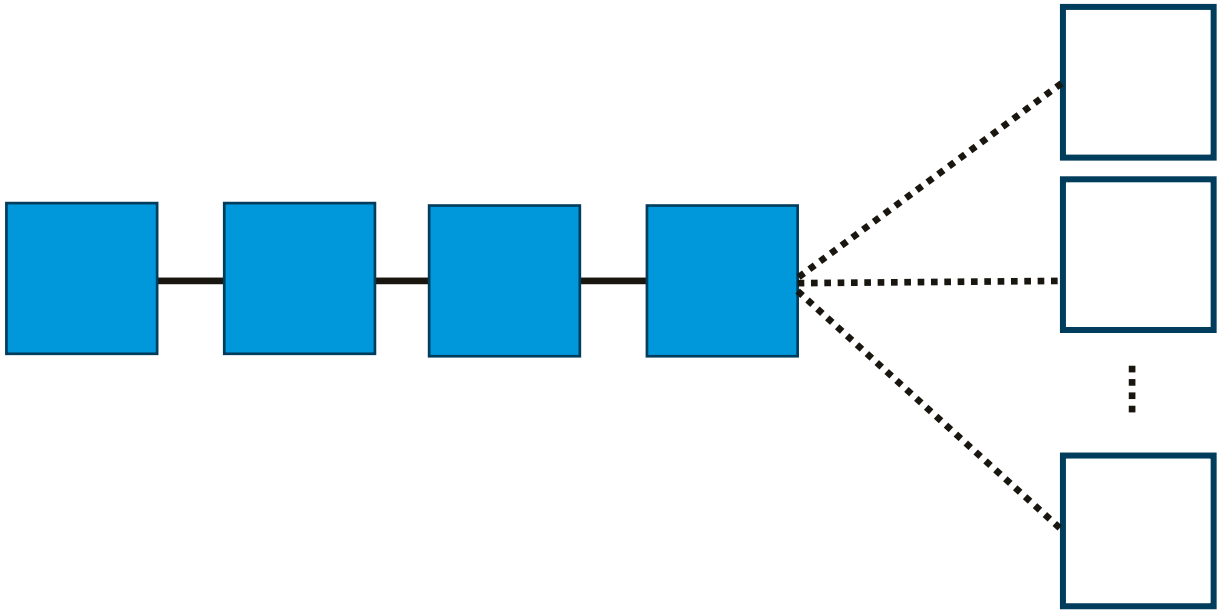
# Self-Supervised Learning with EHRs




**Transfer Learning**: *Assumes Shared Structure*

# Autoregressive Language Modeling with Codes

**Key Intuition**: In medicine, accurately generating future health states captures many use cases of AI



# Our EHR Foundation Model Work

**Methods Development**


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

**Autoregressive**

**MOTOR:** Many Outcome Time Oriented Representations **2023**

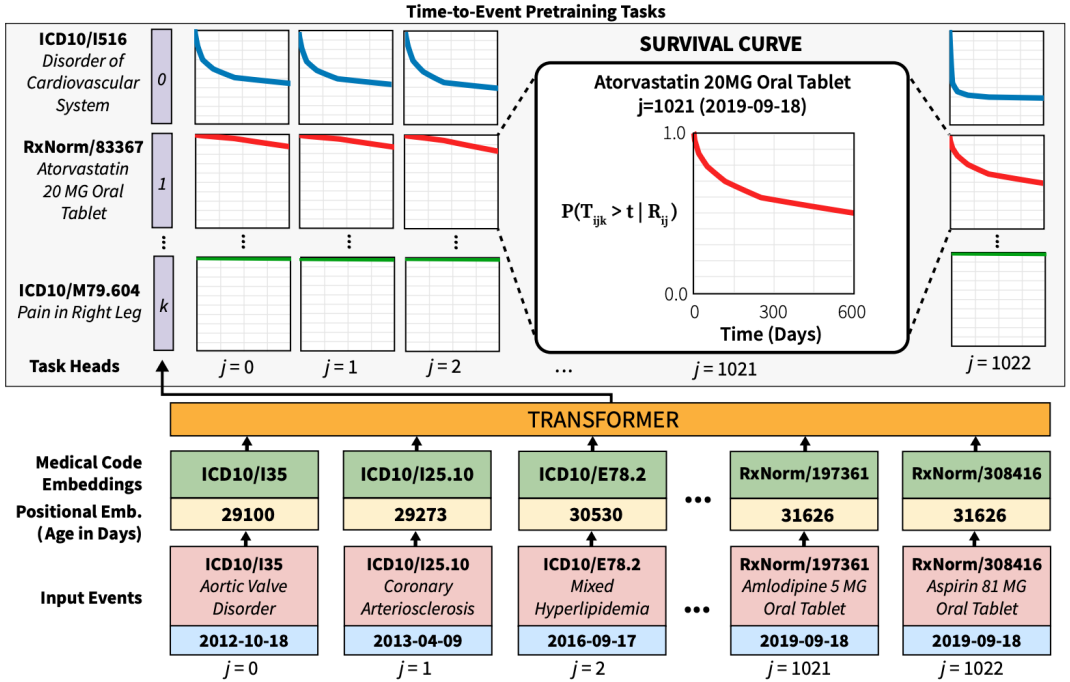
**Time-to-Event**

# MOTOR: A Time-To-Event Foundation Model For Structured Medical Records

**Key Intuition**: We often don't just want to
**know if something will happen**, but also **when it will happen**

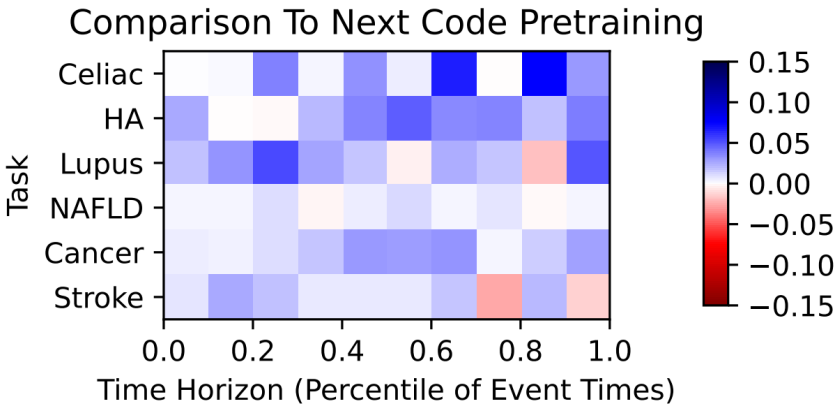
**Time-to-Event (or Survival) Modeling**

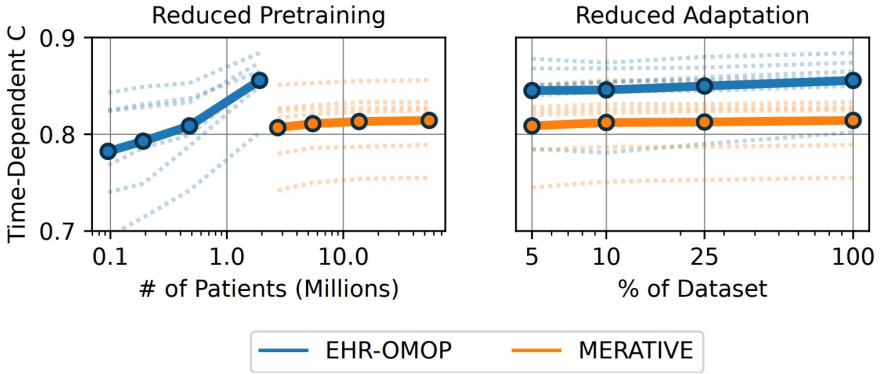
# Use a Different Pretraining Objective




- Massive (**8,192 tasks**) time-to-event pretraining
- Larger scales of pretraining data **(2M to 55M patients)** for EHR and insurance claims data
- **19 evaluation tasks**

*Celiac, Lupus, Pancreatic Cancer, NAFLD, Stroke, Lab Value Prediction, Radiological Findings*




**Outperforms autoregressive pretraining across all time horizons**

**Reduces requirements for labeled examples for adaptation by up to 95%**

# **Evaluating** Foundation Models for Healthcare

# Goals for our Medical Foundation Models

**Better Performance with Less Data**

**Robustness to Distribution Shifts**

**Cross-Site Adaptability**

# Benefits of EHR Foundation Models

**Better Performance with Less Data**

- **+3.5 to 19%** increase in AUROC [1]
- Match SOTA w/ **95%+ less training data** [4,5]

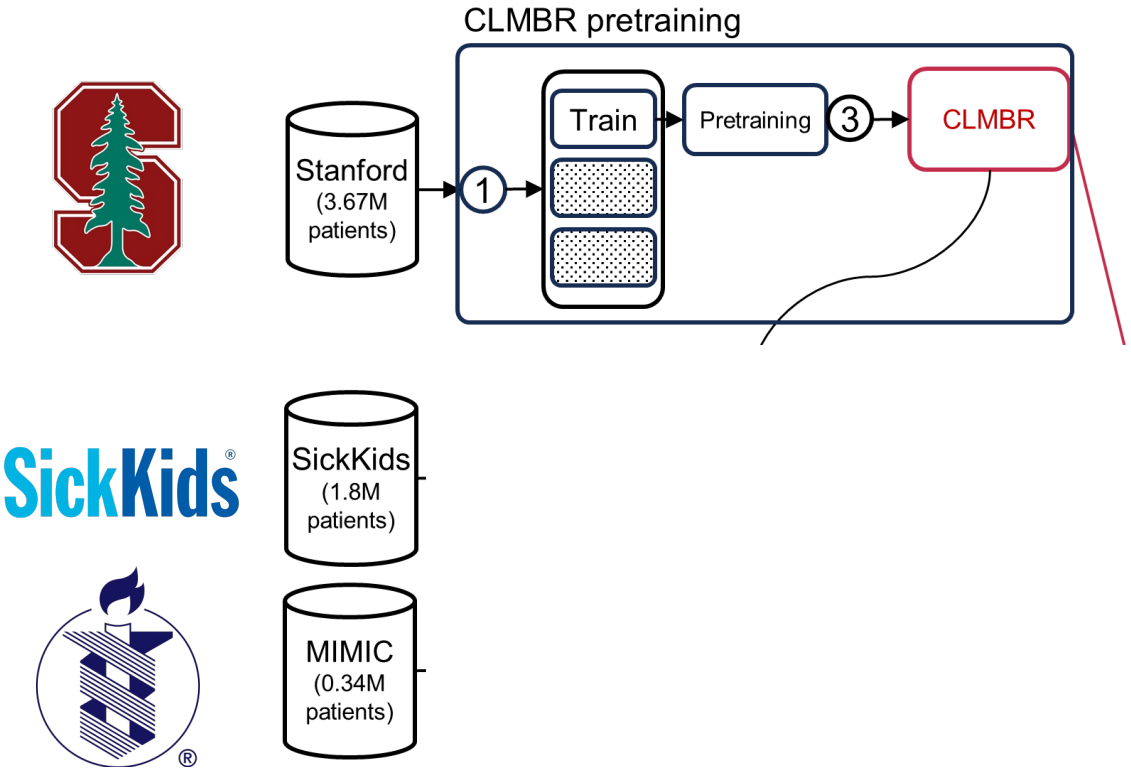
**Robustness to Distribution Shifts**

- Improved **temporal robustness** (+43%) [2]
- Improved **subgroup performance** [3]

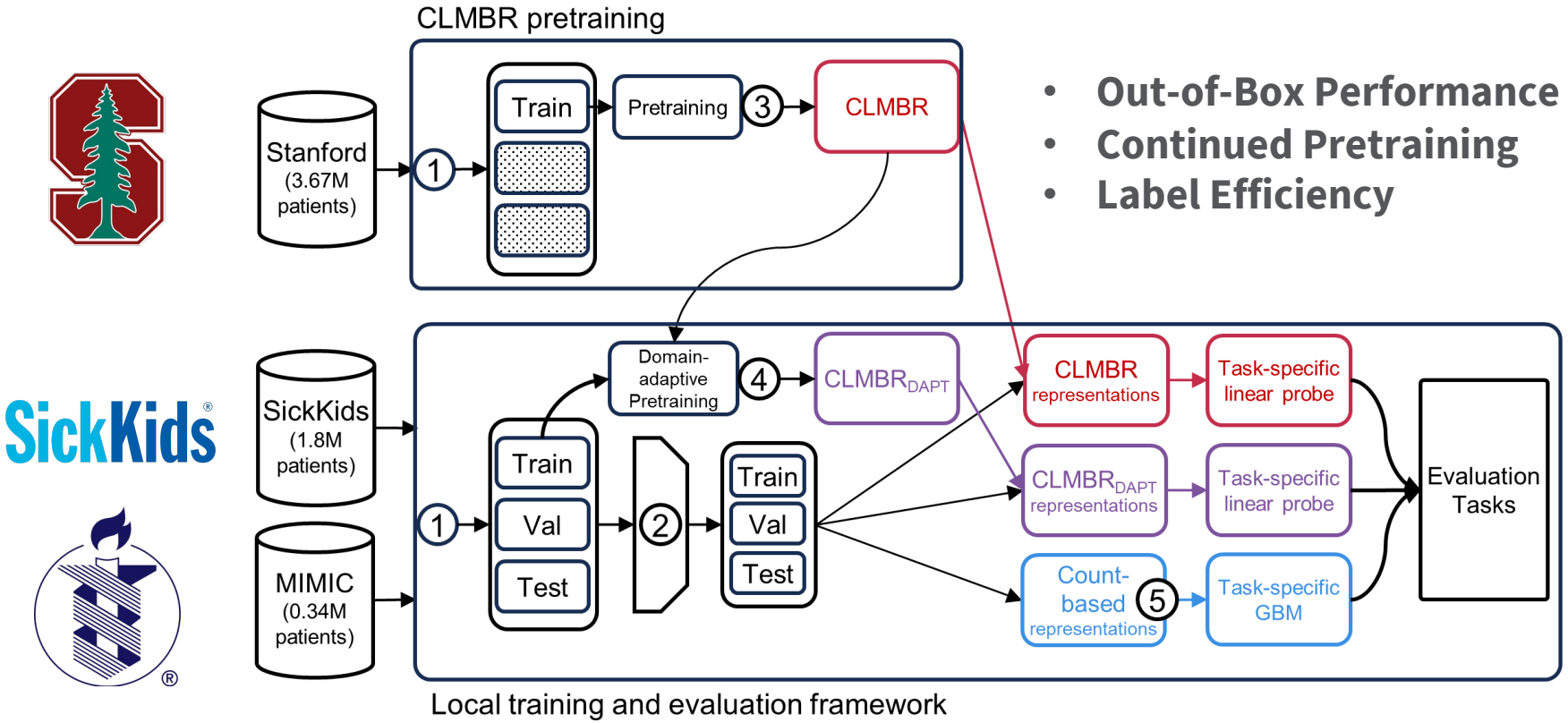
**Cross-Site Adaptability**


[1] Steinberg et al. “Language models are an effective representation learning technique for electronic health record data”. *JBI*. 2021.
[2] Guo et al. “EHR foundation models improve robustness in the presence of temporal distribution shift”. *Scientific Reports*. 2023.
[3] Lemmon et al. “Self-supervised machine learning using adult inpatient data produces effective models for pediatric clinical prediction tasks.” *JAMIA*. 2023.
[4] Guo et al. “A Multi-Site Study on the Adaptability of a Shared EHR Foundation Model.” *Preprint*. 2023.
[5] Steinberg et al. “MOTOR: A Time-To-Event Foundation Model For Structured Medical Records.” *Under Review*. 2023

# Cross-Site Adaptation of EHR Foundation Models



# Cross-Site Adaptation of EHR Foundation Models




- **Out-of-Box Performance**
- **Continued Pretraining**
- **Label Efficiency**

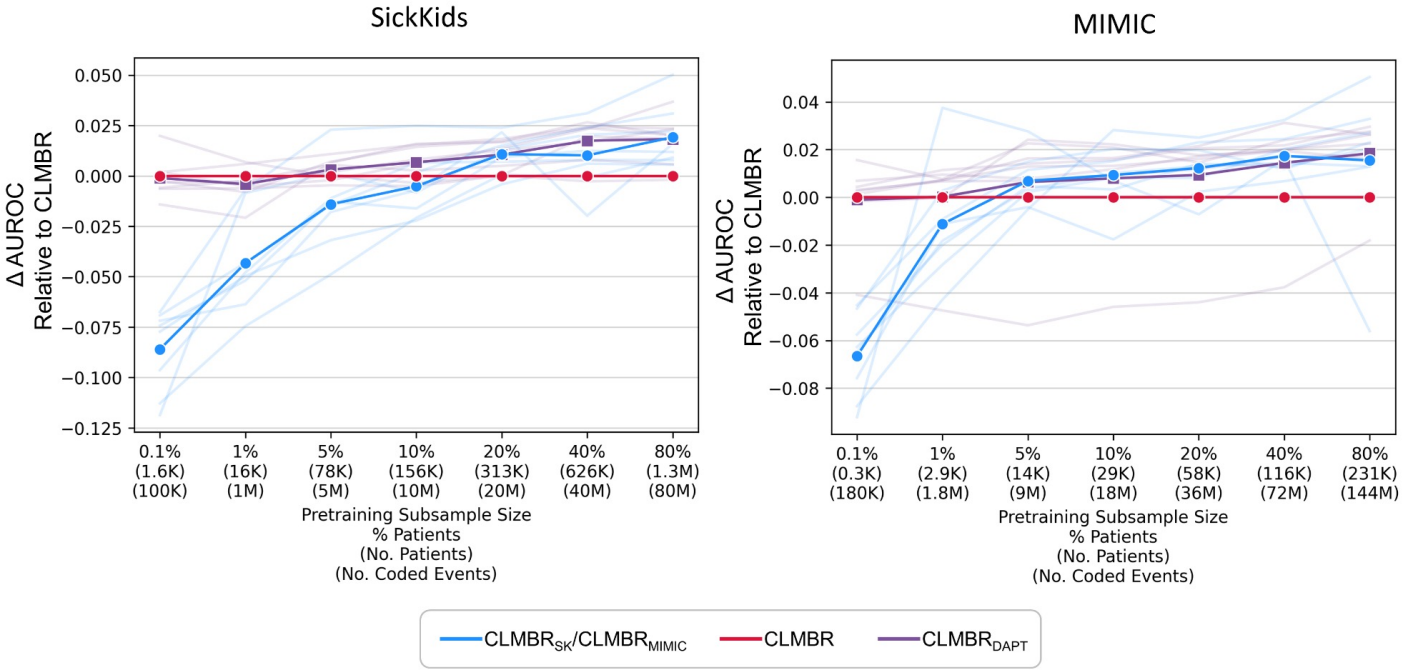
# Overall Performance

| | SickKids | | | MIMIC | | |
|---|---|---|---|---|---|---|
| | **GBM** | **CLMBR** | **CLMBR$_{DAPT}$** | **GBM** | **CLMBR** | **CLMBR$_{DAPT}$** |
| **In-hospital Mortality** | 0.893 [0.815, 0.953] | 0.941 [0.902, 0.971] | **0.957 [0.922, 0.98]** | 0.905 [0.886, 0.922] | 0.911 [0.894, 0.926] | **0.927 [0.912, 0.941]** |
| **Long LOS** | **0.866 [0.853, 0.879]** | 0.815 [0.8, 0.83] | 0.839 [0.825, 0.853] | **0.831 [0.821, 0.841]** | 0.792 [0.781, 0.803] | 0.823 [0.813, 0.833] |
| **30-day Readmission** | 0.783 [0.755, 0.809] | 0.774 [0.747, 0.799] | **0.804 [0.78, 0.828]** | 0.619 [0.514, 0.719] | **0.695 [0.601, 0.781]** | 0.673 [0.564, 0.771] |
| **Hypoglycemia** | 0.88 [0.83, 0.924] | 0.915 [0.879, 0.945] | **0.918 [0.883, 0.948]** | 0.79 [0.746, 0.83] | 0.812 [0.779, 0.844] | **0.837 [0.802, 0.869]** |
| **Hyponatremia** | 0.783 [0.579, 0.957] | 0.925 [0.88, 0.963] | **0.957 [0.933, 0.978]** | 0.798 [0.723, 0.863] | 0.817 [0.771, 0.859] | **0.845 [0.806, 0.882]** |
| **Hyperkalemia** | 0.749 [0.687, 0.808] | 0.793 [0.743, 0.838] | **0.822 [0.781, 0.86]** | 0.7 [0.619, 0.775] | 0.8 [0.739, 0.856] | **0.818 [0.762, 0.868]** |
| **Thrombocytopenia** | 0.953 [0.928, 0.975] | 0.962 [0.946, 0.975] | **0.967 [0.95, 0.98]** | 0.915 [0.884, 0.943] | 0.928 [0.909, 0.946] | **0.952 [0.939, 0.964]** |
| **Anemia** | 0.919 [0.898, 0.938] | 0.918 [0.897, 0.937] | **0.946 [0.932, 0.959]** | **0.878 [0.862, 0.893]** | 0.854 [0.836, 0.871] | 0.875 [0.859, 0.89] |

Out-of-the-box CLMBR **outperformed GBM** in **5/8** tasks (SK), **6/8** tasks (MIMIC).
**DAPT improved CLMBR performance in all tasks** (SK), and in **7/8** tasks (MIMIC).

# Continued Pretraining

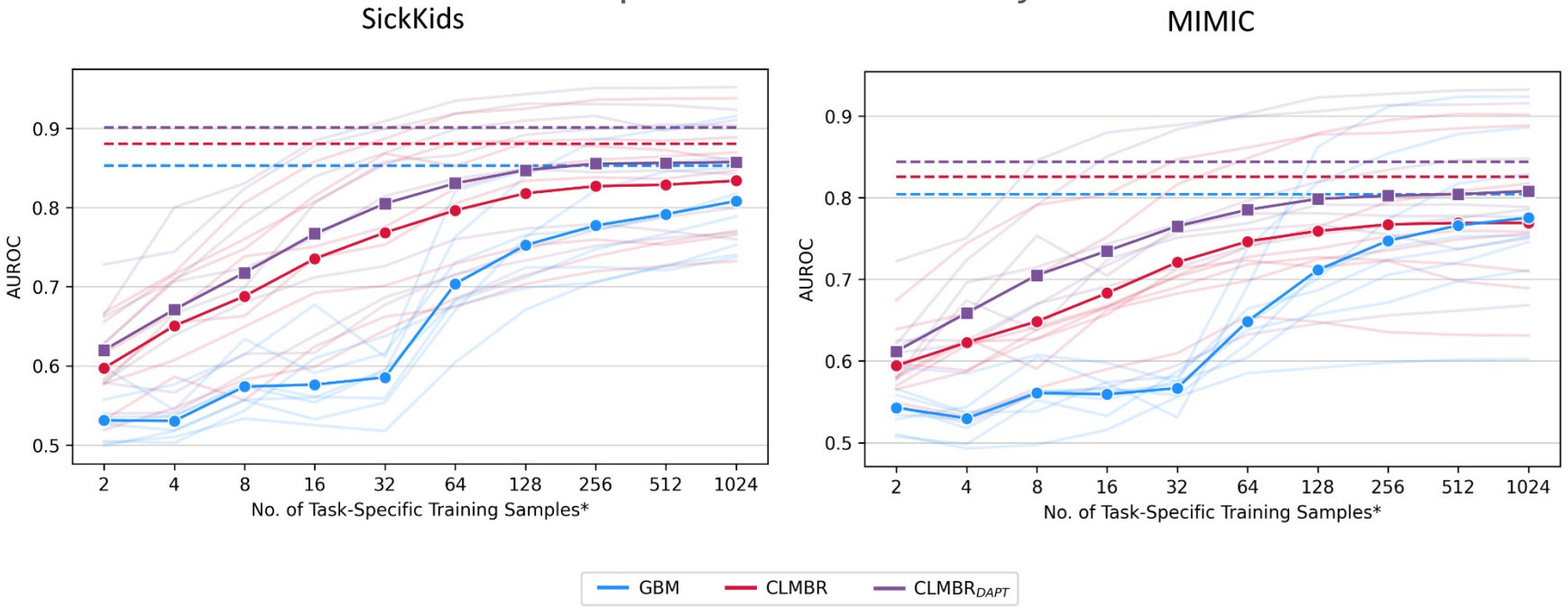
## Incorporate additional hospital-specific pretraining data




Require **60-90% less** pretraining data

# Improved Few-Shot Performance

## Improved Label Efficiency




Require **less than 1%** of training examples to match performance of XGBoost

# Benefits of EHR Foundation Models

**Better Performance with Less Data**

- **+3.5 to 19%** increase in AUROC [1]
- Match SOTA w/ **95%+ less training data** [4,5]

**Robustness to Distribution Shifts**

- Improved **temporal robustness** (+43%) [2]
- Improved **subgroup performance** [3]

**Cross-Site Adaptability**

- Transfer pretrained models across hospitals
- Require **60-90% less** pretraining data [4]

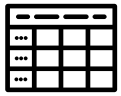

[1] Steinberg et al. “Language models are an effective representation learning technique for electronic health record data”. *JBI*. 2021.
[2] Guo et al. “EHR foundation models improve robustness in the presence of temporal distribution shift”. *Scientific Reports*. 2023.
[3] Lemmon et al. “Self-supervised machine learning using adult inpatient data produces effective models for pediatric clinical prediction tasks.” *JAMIA*. 2023.
[4] Guo et al. “A Multi-Site Study on the Adaptability of a Shared EHR Foundation Model.” *Preprint*. 2023.
[5] Steinberg et al. “MOTOR: A Time-To-Event Foundation Model For Structured Medical Records.” *Under Review*. 2023

# **Reproducibility** of Foundation Model Research

# Releasing New Medical Datasets

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

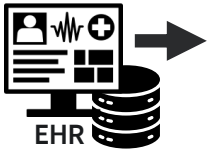
**6,739** Patients

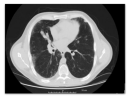



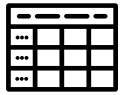

**SPOTLIGHT**

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

**19,402** Patients



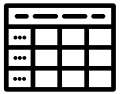





**NeurIPS 2023**

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

**267** Patients

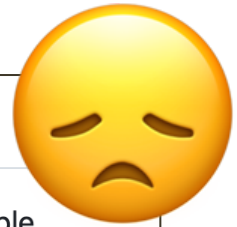


**ML4H Symposium 2024**

**BEST THEMATIC PAPER**

*To Appear* **AAAI 2024**

# Open & Accessible Model Weights



**Sharing pre-trained model**

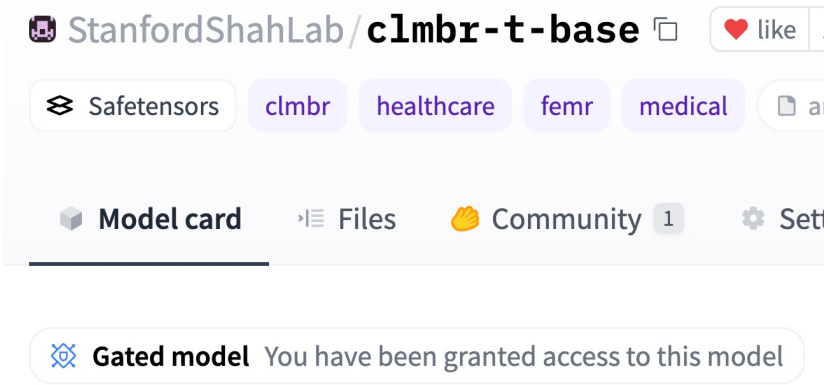
Initially we really hoped to share our models but unfortunately, the pre-trained models are no longer sharable. According to SBMI Data Service Office: "Under the terms of our contracts with data vendors, we are not permitted to share any of the data utilized in our publications, as well as large models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

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

**Transfer learning is the primary value prop of foundation models!**

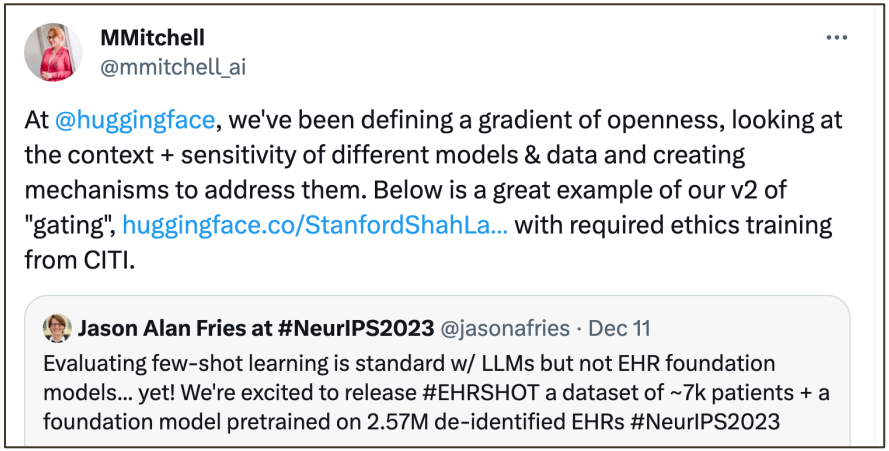
**Foundation Models Risk Increasing our Reproducibility Crisis**

# Enabling Open Science




**Our first model hub release!**

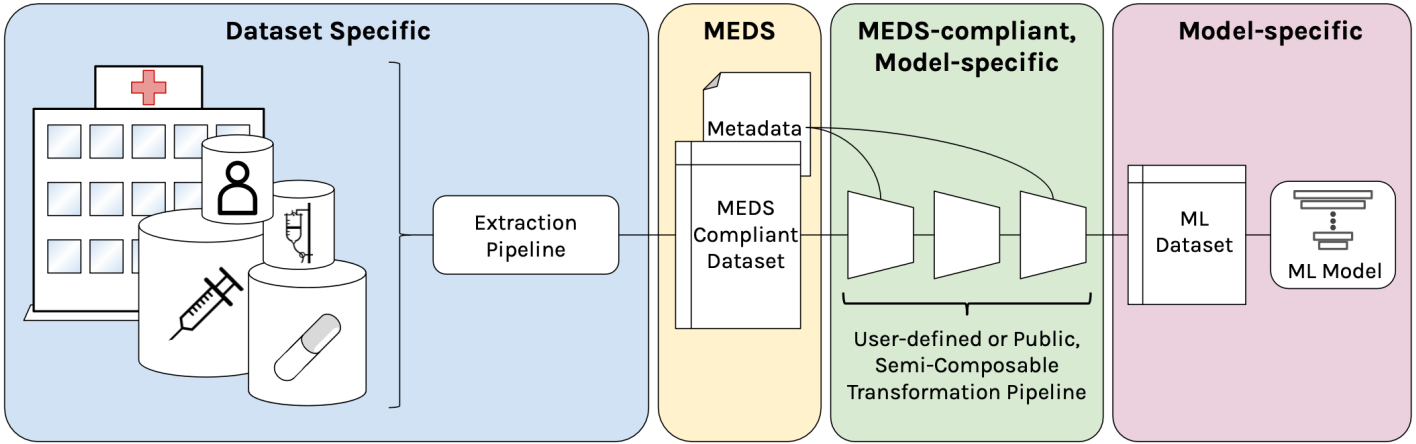
- Gated model on Hugging Face
- Requires CITI ethics training
- Non-commercial use only




Margaret Mitchell

Chief AI Ethics Scientist, Hugging Face

# Medical Event Data Standard (MEDS)




**DRAFT 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

# Open Weights are Critical to Fair & Secure Models

**Why Anthropic and OpenAI are obsessed with securing LLM model weights**

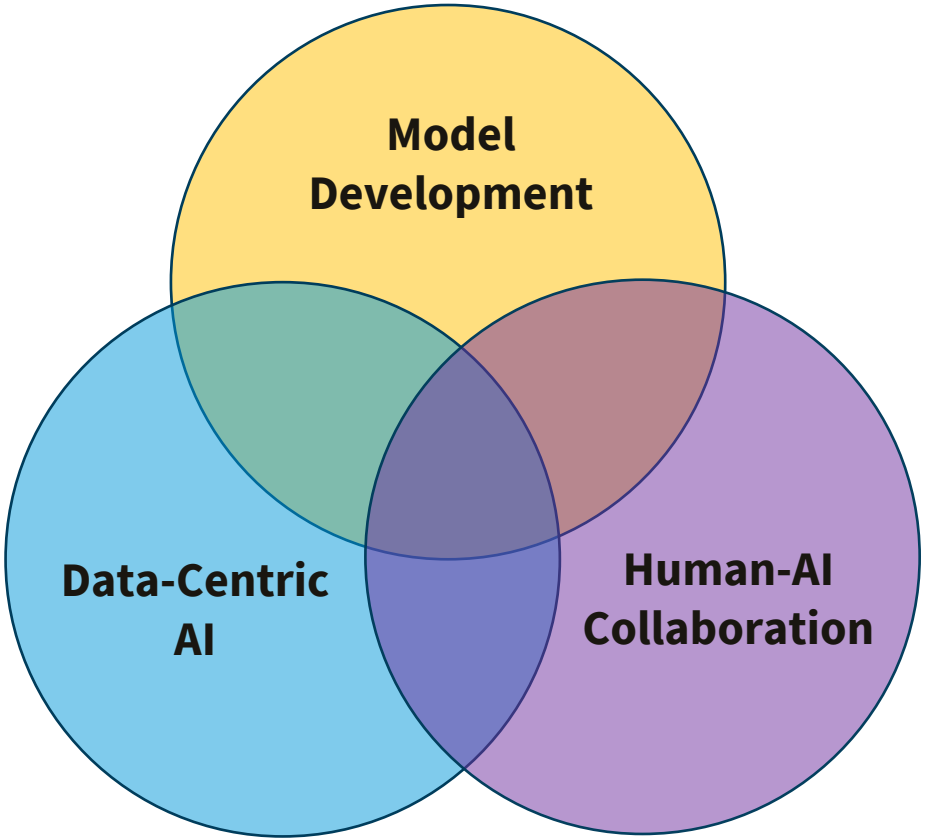
**Transparency** (training data, model weights) is critical for fair and secure models!

OpenAI

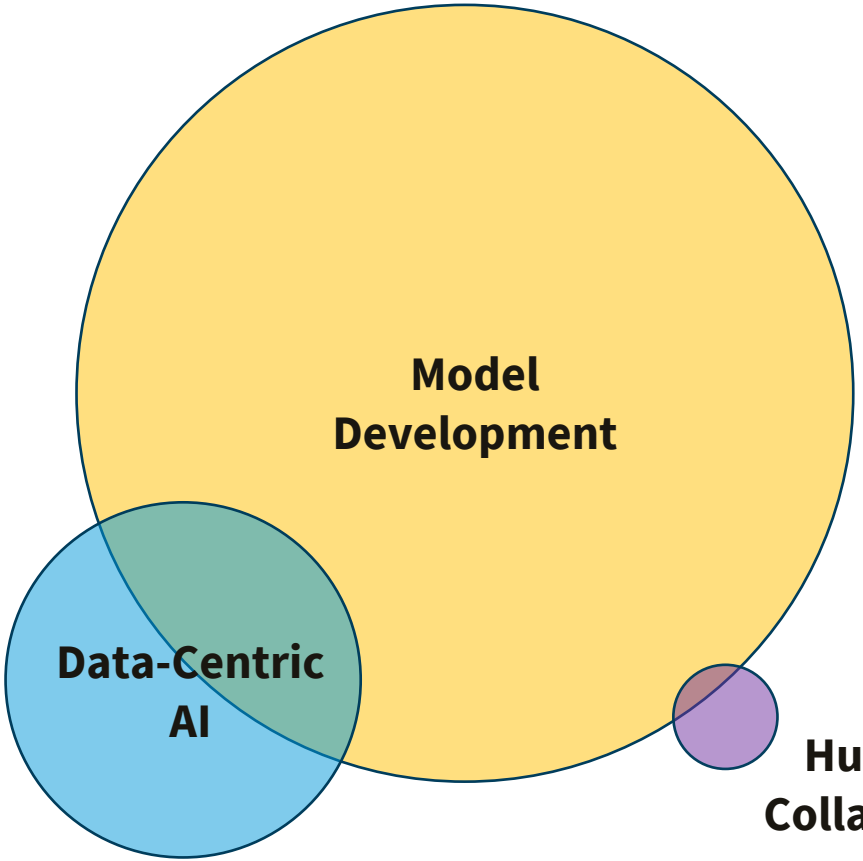
*"Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."*

# Closing Thoughts

# Opportunities Moving Forward



# Opportunities Moving Forward



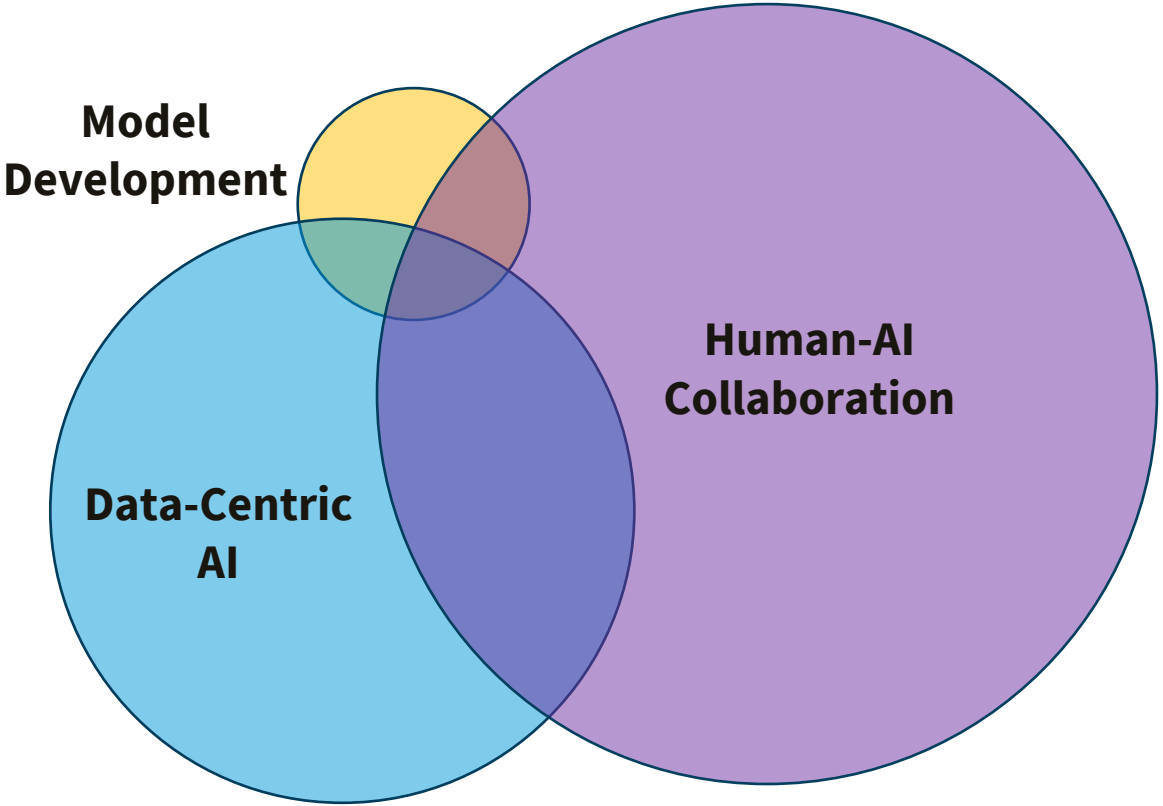

**Data-Centric AI**

- Data Cleaning (medical data is messy!)
- Model-guided Training (Weak Supervision? "*Superalignment*")
- Informed Data Selection & Curation

**Human-AI Collaboration**

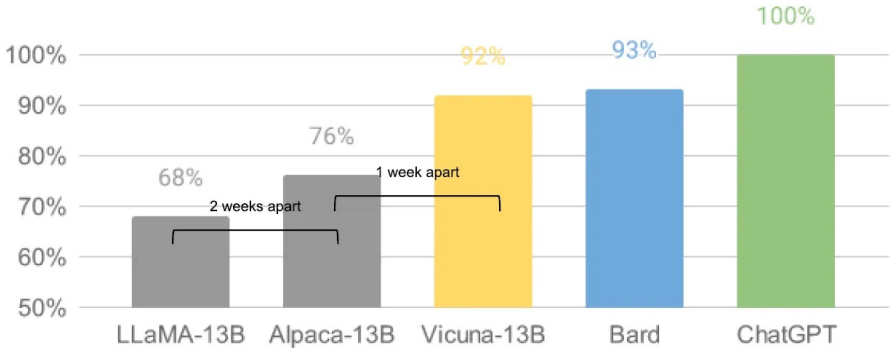
- Feedback loops and decision making
- Optimizing evidence synthesis
- Capturing preferences to guide model improvement
- **Building trust in models**

# Opportunities Moving Forward



# Calls for the Academic Community

## Smaller Models, Cheaper to Train




**Lead Building Open, Reproducible Medical Base Models**

## Reimagine Model Evaluation


**AI will augment existing roles**

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

# Team Science

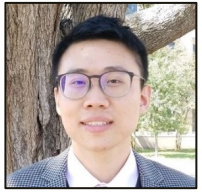
Jason Fries

Ethan Steinberg

Michael Wornow

Rahul Thapa

Scott Fleming



Frazier Huo

Louis Blankemeier

Juan Manuel Zambrano Chaves

Keith Morse

Crystal Xu

Akshay Chaudhari

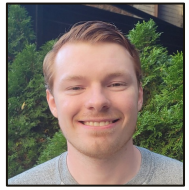
Nigam Shah

FEMR

David Hall

Lawrence Guo

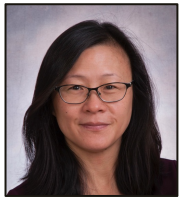
Yifan Mai



Joshua Lemmon

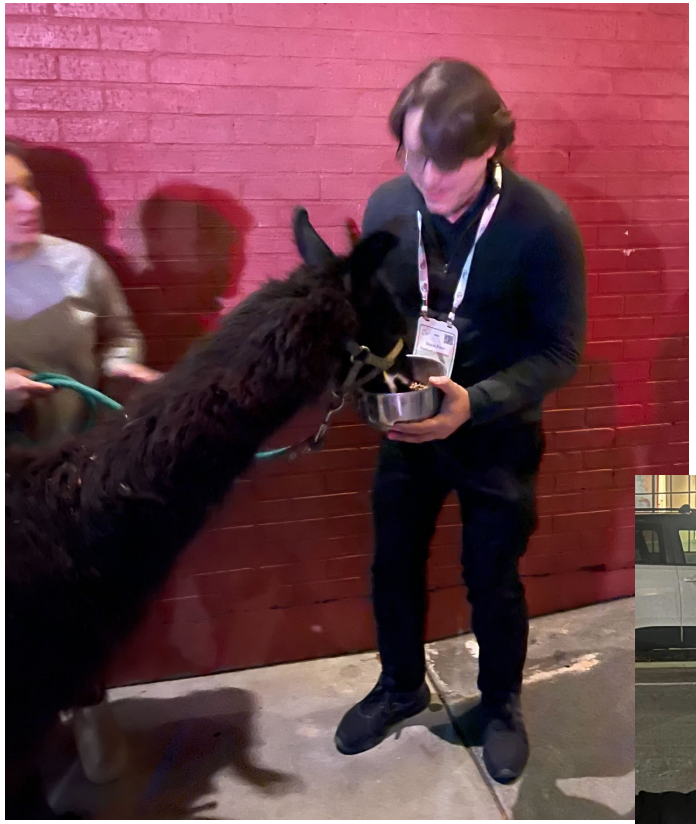
Percy Liang

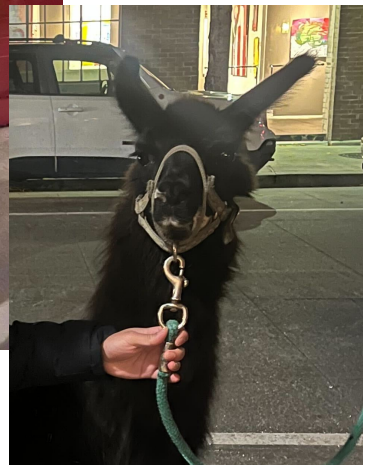
CRFM



Lillian Sung

SickKids





Emperor Kuzco!
NeurIPS 2023 “Keep AI Open”

# Thank You!

jason-fries@stanford.edu

*I’m on the academic job market this season – don’t hesitate to reach out!*