HealTAC 2025: 8th Healthcare Text Analytics Conference

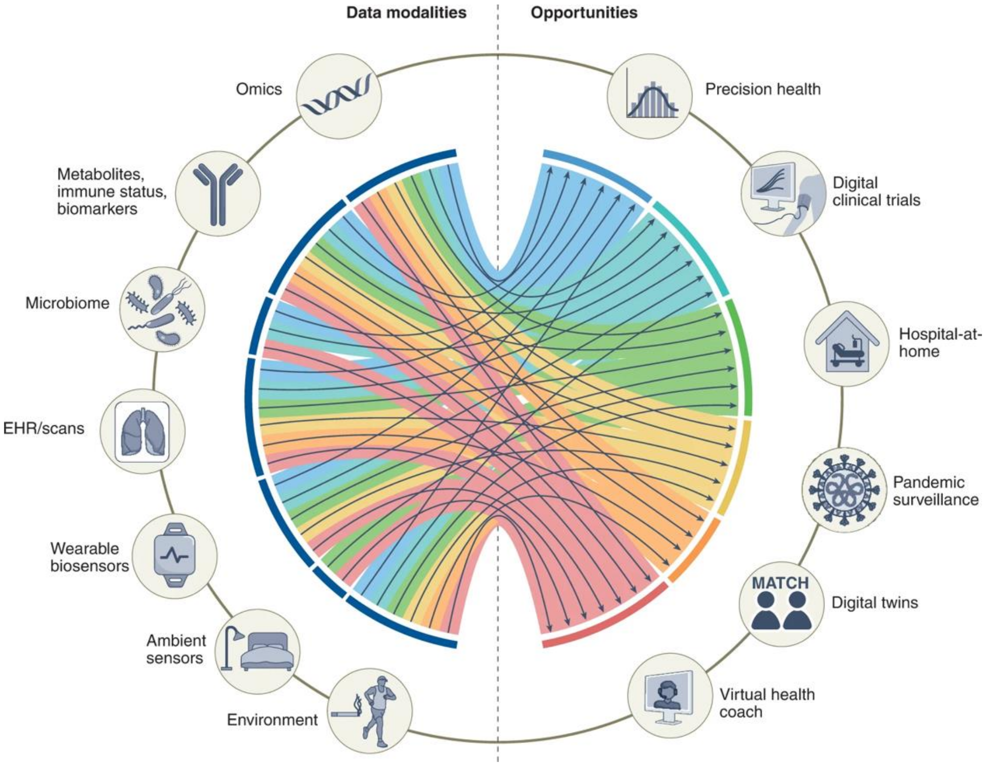
Addressing the Missing Context Problem in Foundation Models for Healthcare

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

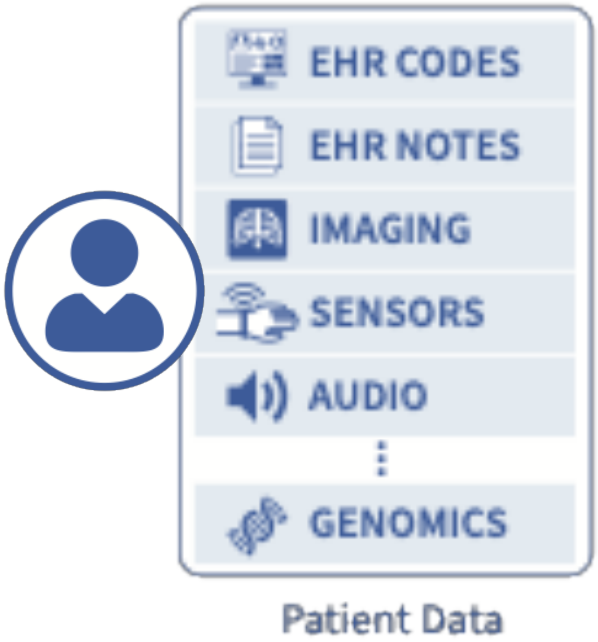
Stanford MEDICINE
School of Medicine

BMIR
Stanford Center for Biomedical Informatics Research

Healthcare is Inherently Multimodal

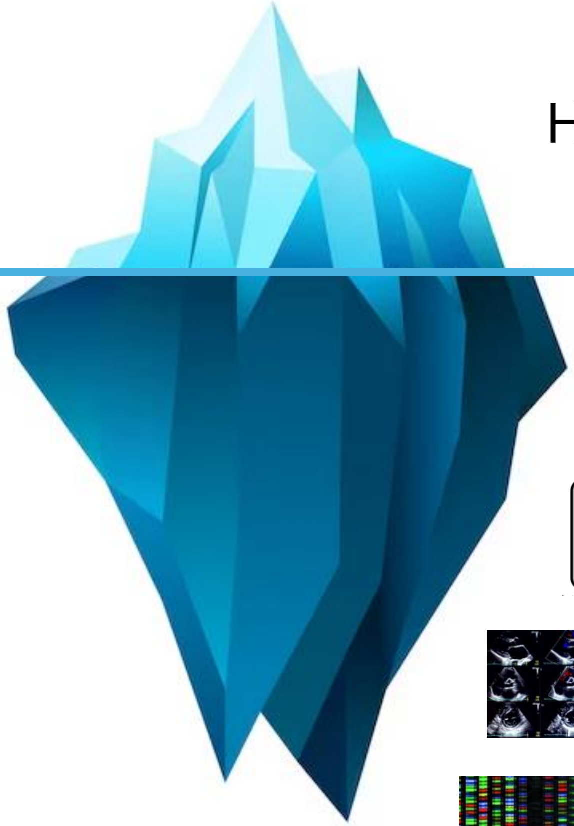


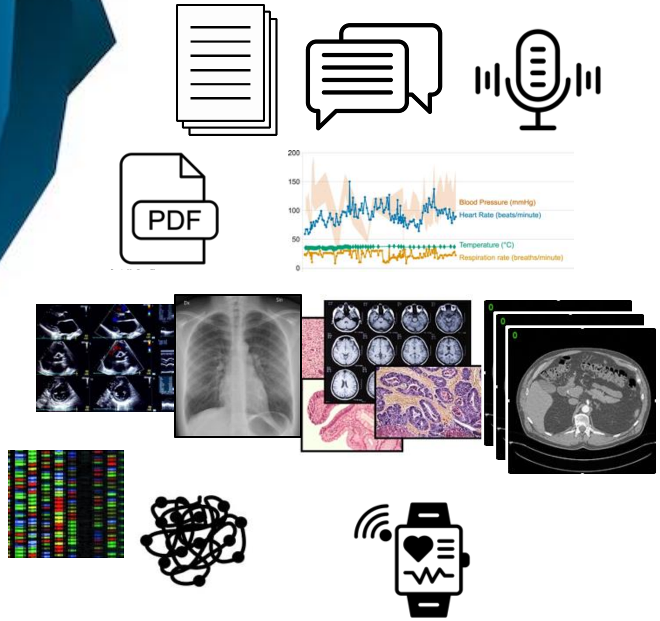
Acosta et al. *Nature Medicine*. 2022



Hospital data is growing at a rate of **36% per year**

World Economic Forum, Dec. 2019





Hard to use for medical decision making

A patient with advanced lung cancer has been treated for 24 months and now shows progression in a few isolated areas, with mixed response.

When presenting to a **tumor board**, clinicians need to piece together:

- New genomic mutations driving treatment resistance
- Pathology reports
- Radiology scans
- Clinical history

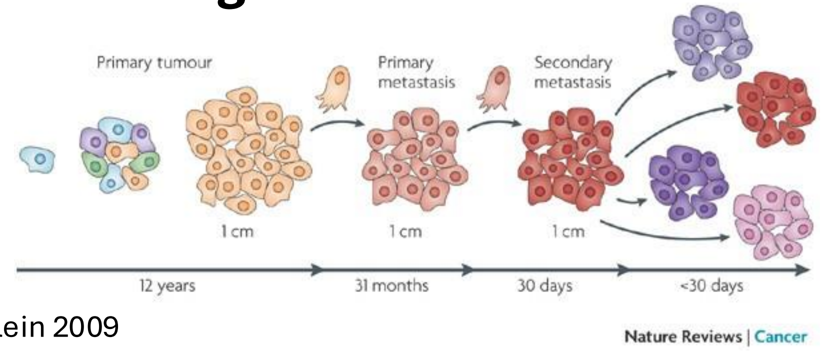
Time-consuming to gather data manually

No easy way to find similar cases

Difficult to reason longitudinally

Human Health is Time-Varying

Cancer Progression



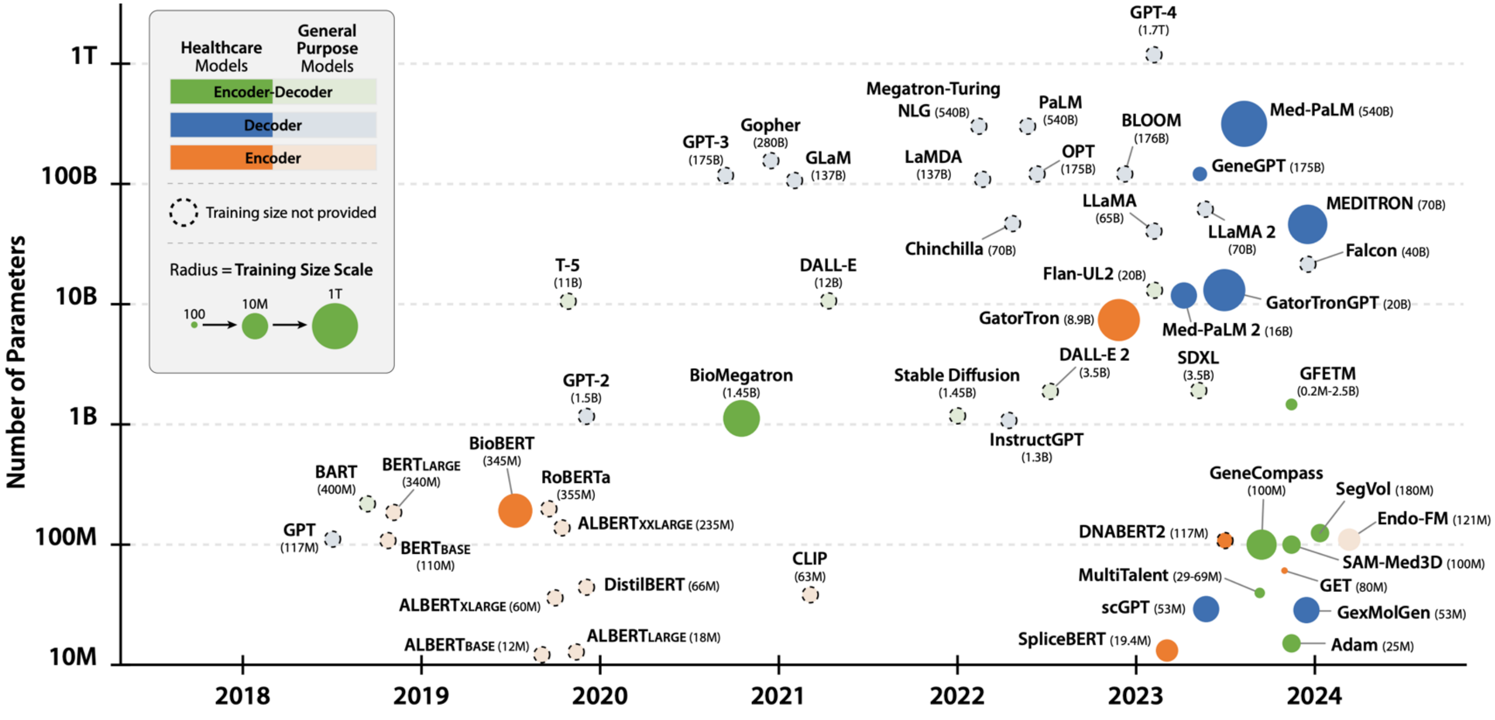
Klein 2009

How likely is this patient to develop gastrointestinal cancer in 10 years, 5 years, or 2 years?

Pediatric Development

By what age will this child receive an autism spectrum disorder diagnosis: 18 months, 3 years, or 10 years?

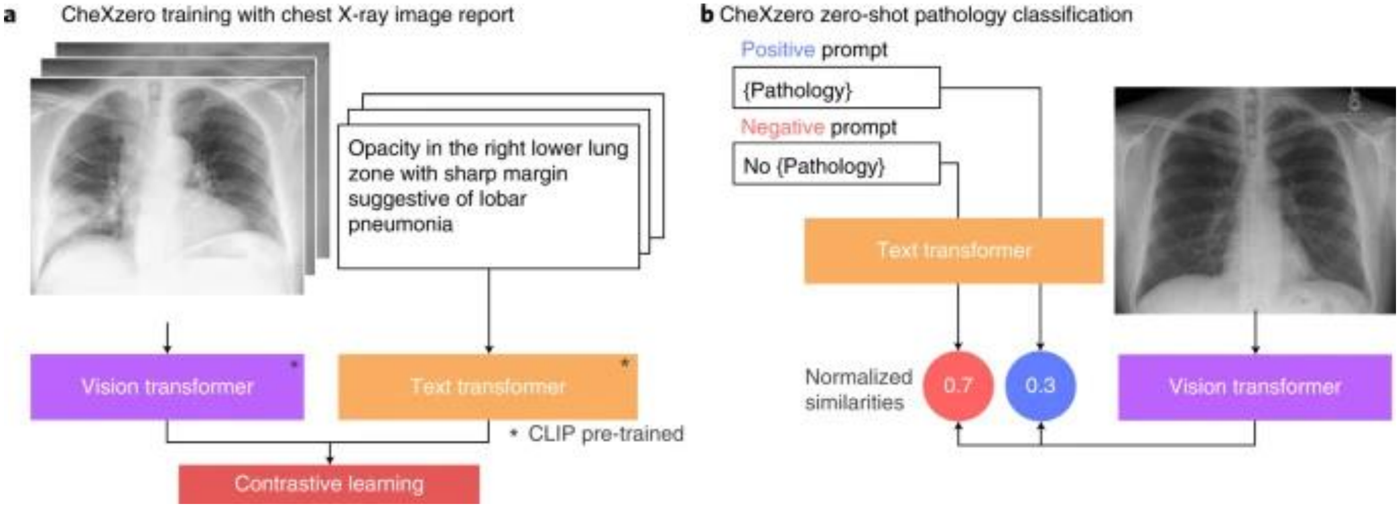
Opportunity for AI to **reimagine** how we **interact and understand** medical data



Khan et al., "A Comprehensive Survey of Foundation Models in Medicine," 2025.

6

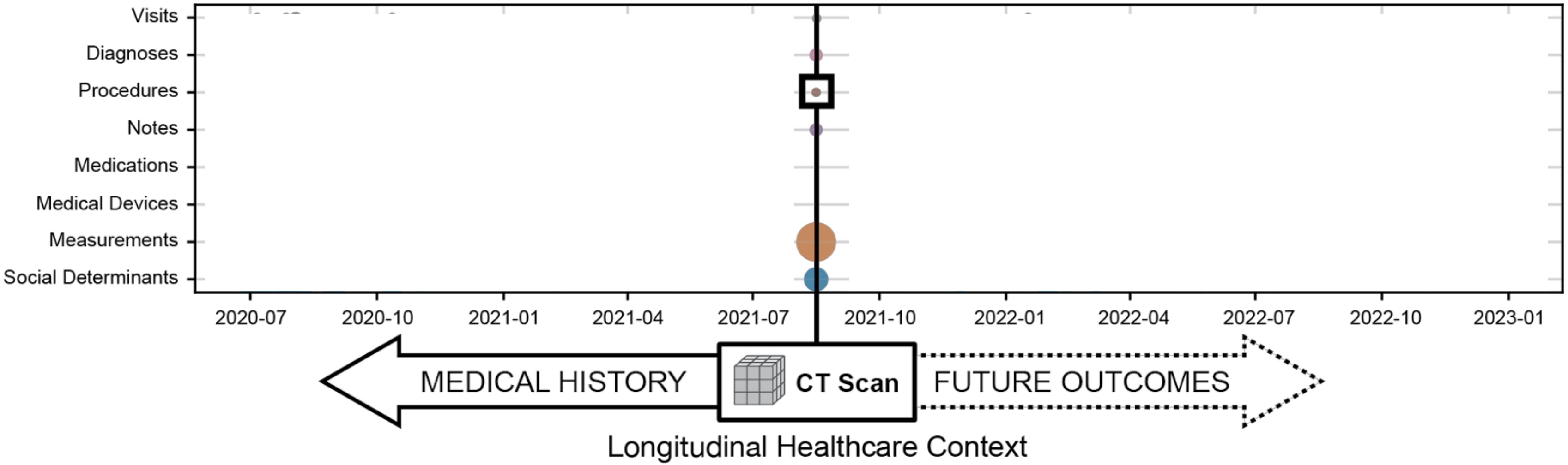
The Status Quo: Opportunistic *Local* Supervision



Tiu et al. 2022. *Nature Biomedical Engineering*

7

Missing Context Problem

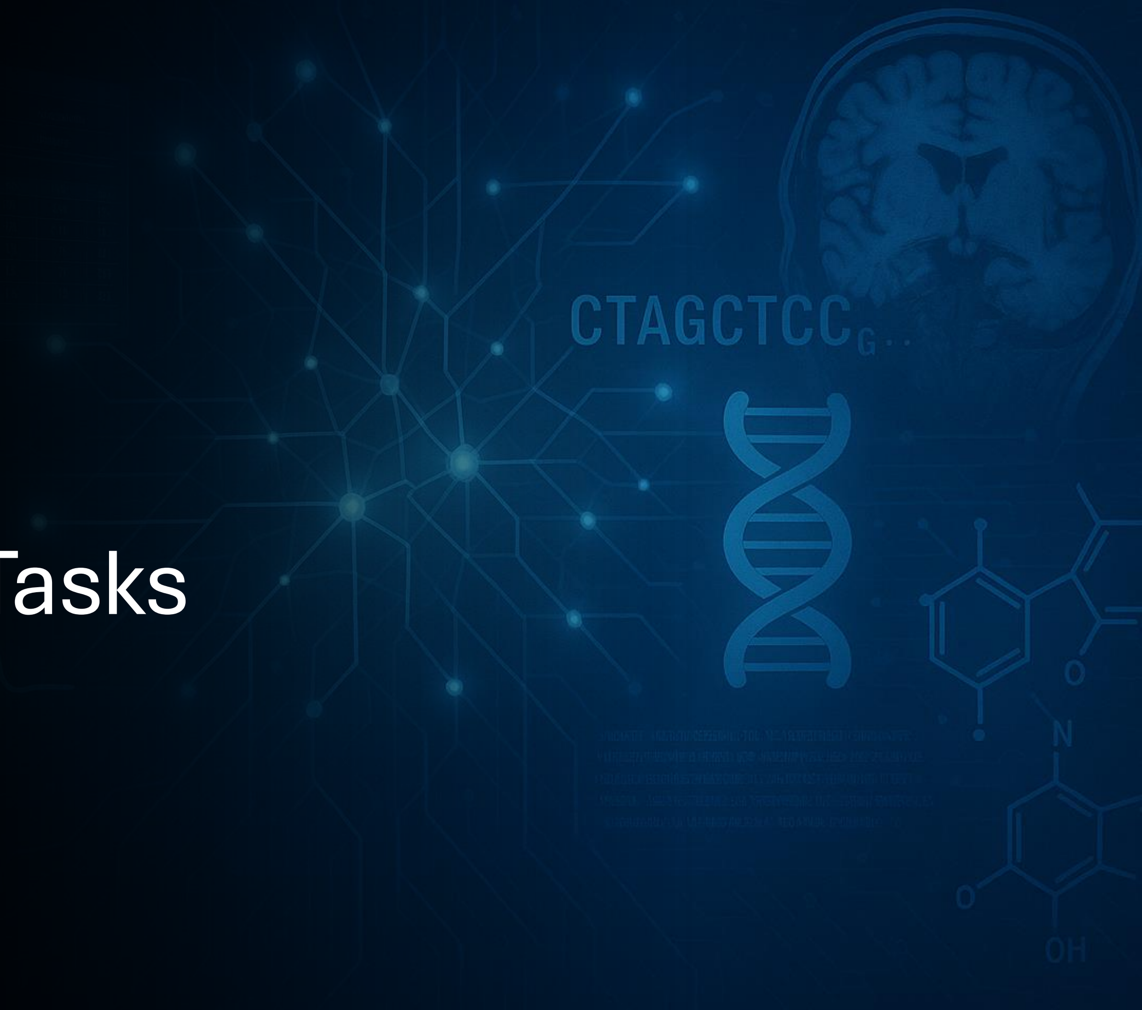


- Limited use of **longitudinal health data to supervise models**
- Limited insight into the **distinct needs of different stakeholders**

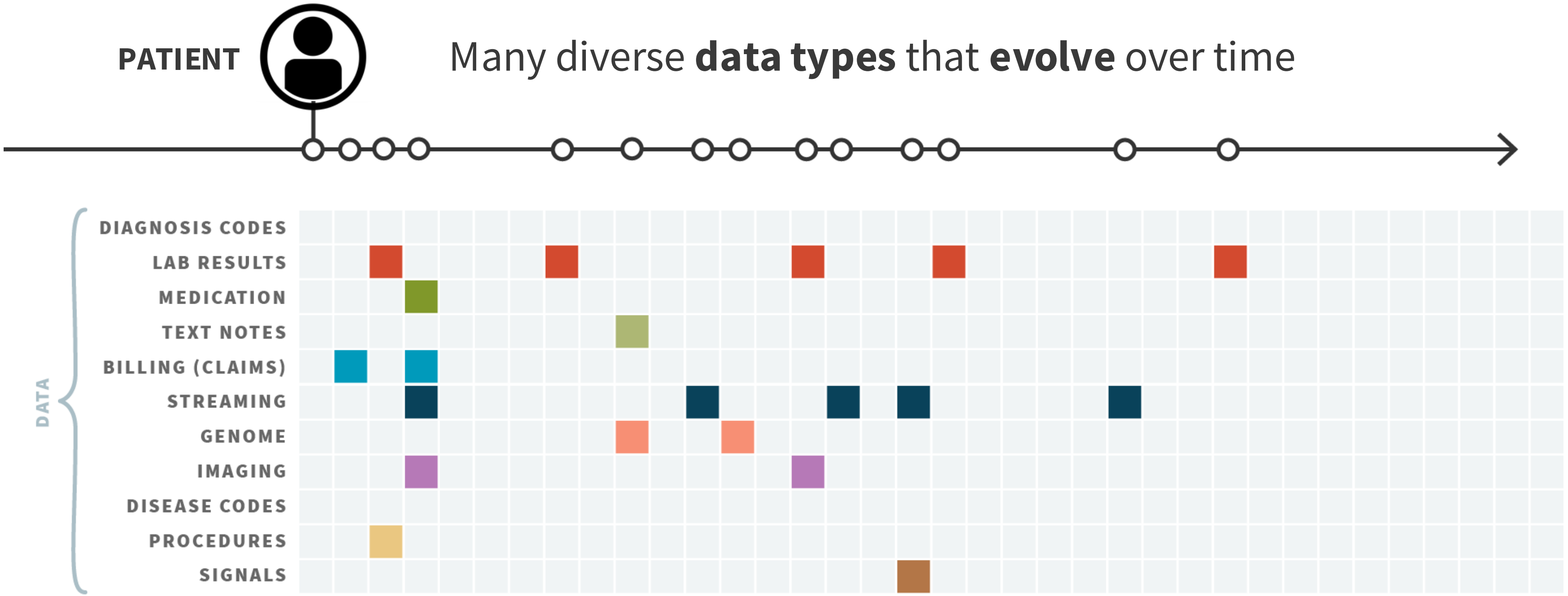
Outline

- **Overview**: EHR Timelines & Tasks
- **Pre- and Post-Training**: Longitudinal EHRs as Supervision
- **Human-AI Teaming**: Natural Language Interfaces
- **Future**: Research Opportunities
- **Questions**

Overview:
EHR Timelines & Tasks



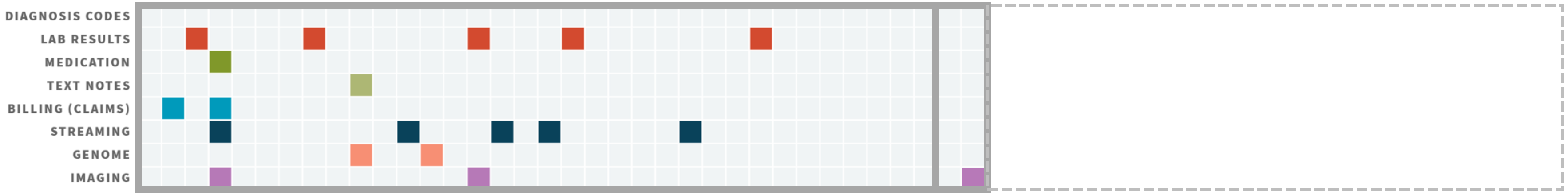
Electronic Health Records (EHRs) are Multimodal Timelines



Longitudinal EHRs provide a **holistic view of multimodal data**

AI for Healthcare Requires Temporal Reasoning

Diagnosis (Classification) | **Prognosis**



Example ML Applications

What Occurred in the Past?

- Chart summarization
- Clinical trial recruitment

What is Occurring Now?

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

Predict Future Risks & Intervention Benefits

- What is the likelihood that this patient will develop lung cancer?

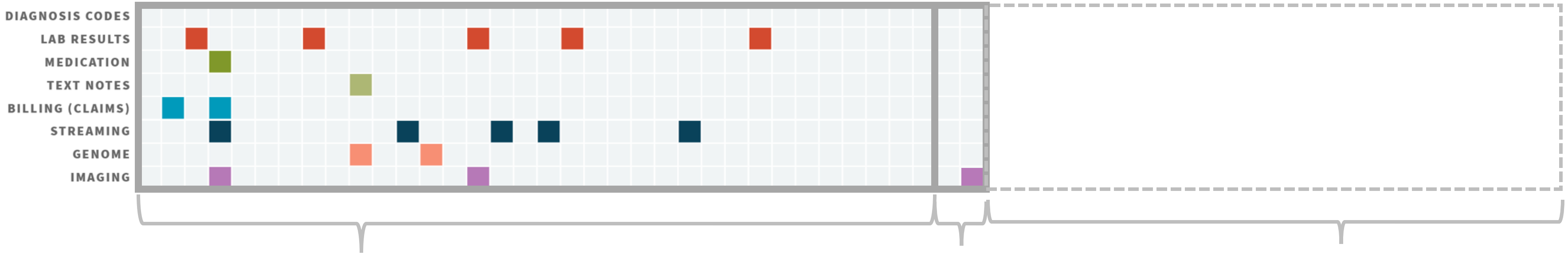
Stakeholders

Clinicians

Whether to Treat | **How to Treat** subject to | Policy | Capacity to Act | Intervention Properties

Foundation Models Are Essential for AI in Healthcare

Diagnosis (Classification) **Prognosis**



What Occurred in the Past?

What is Occurring Now?

Predict Future Risks & Intervention Benefits

Example ML Applications

?

?

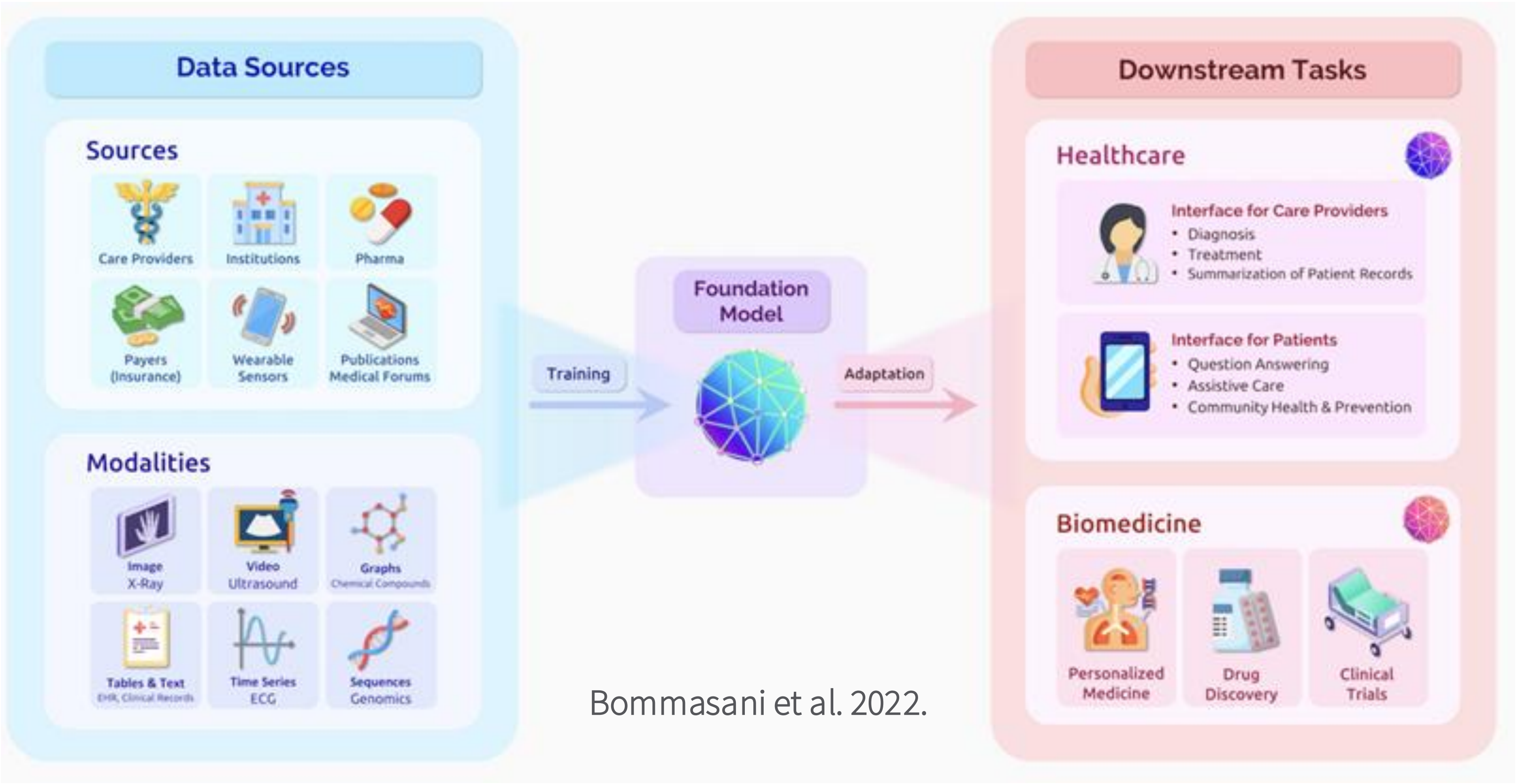
?

Many stakeholder groups with **distinct needs**

Stakeholders

- Clinicians
- Hospital Administrators
- Insurance Providers
- Pharma
- Regulatory Agencies
- Patients
- ...
- Researchers

Foundation Models and AI's "Industrial Age"



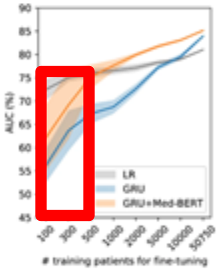
Bommasani et al. 2022.

Pre- and Post-Training:
Longitudinal EHRs as Supervision

Current Self-Supervised Objectives for Structured EHR Data

BERT-Style (Masked Language Modeling)

- BEHRT (Li et al. 2020)
- MedBERT (Rasmy et al. 2021)
- CEHR-BERT (Pang et al 2021)
- ClaimPT (Zeng et al. 2022)
- *et alia*

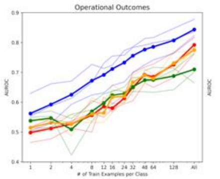


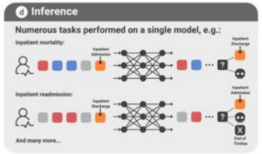
MedBERT

- Trained on **28M patients**
- Performance with **< 500 examples worse than logistic regression**

GPT-Style (Autoregressive)

- CLMBR (Steinberg et al. 2020)
- TransformEHR (Yang et al. 2023)
- CEHR-GPT (Pang et al 2024)
- ETHOS (Renc et al. 2024)
- Foresight (Kraljevic et al. 2024)
- Context Clues (Wornow et al 2025)





CLMBR

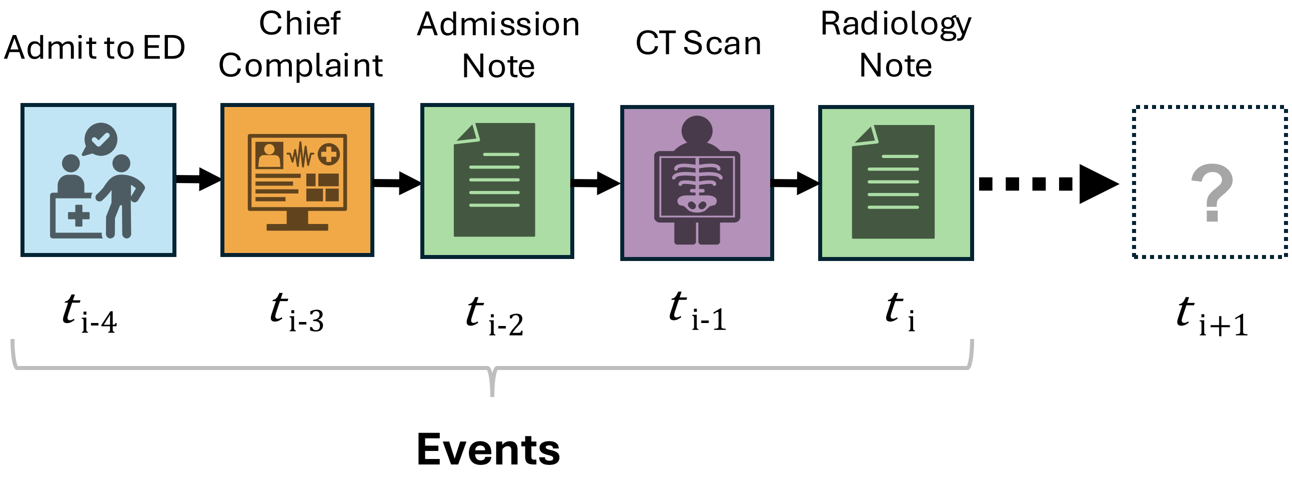
- Trained on **2.57M patients** (3.5B tokens)
- SOTA **few-shot** learning using **embeddings**

ETHOS

- Trained on **200k patients** (MIMIC-VI)
- **Zero-shot** abilities using **generation**

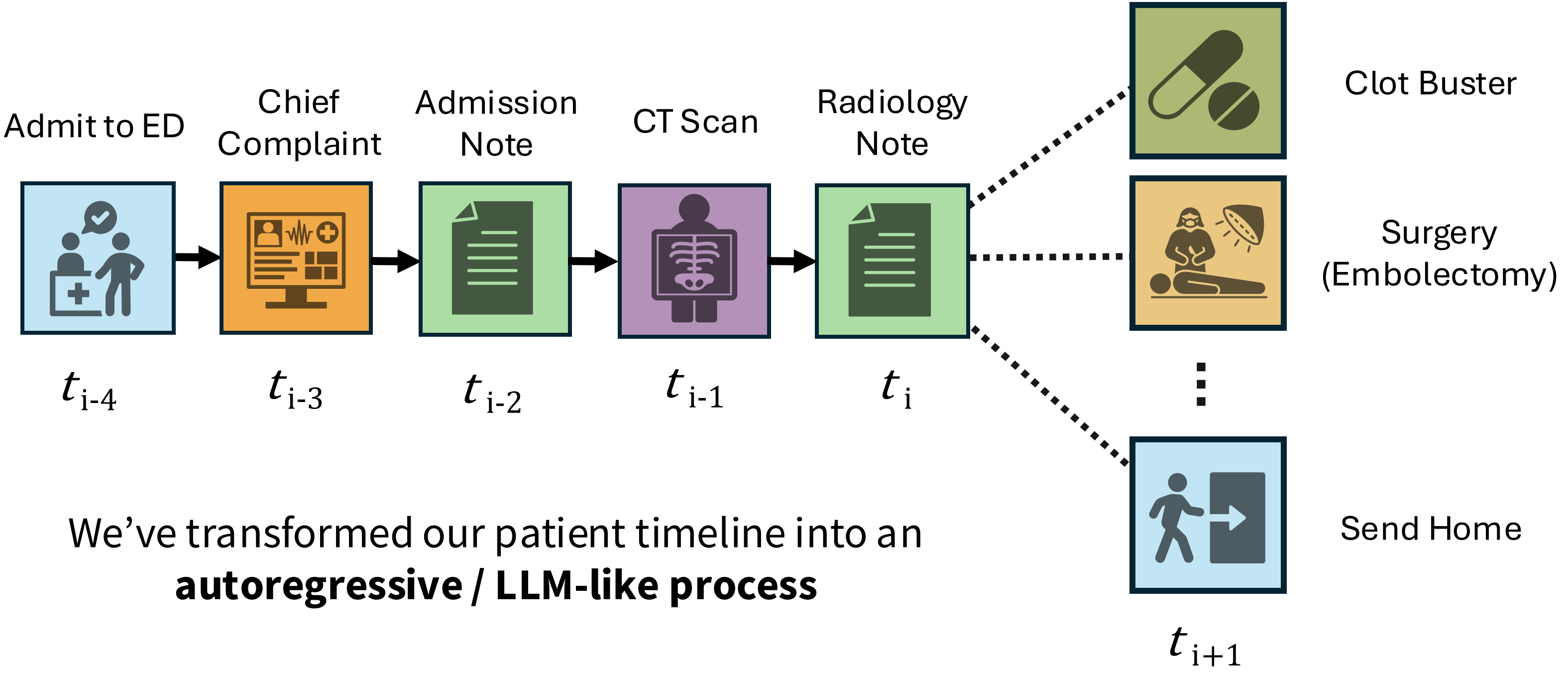
Modeling Patient Timelines for AI

CASE: Patient **presents to ED** with sudden onset **shortness of breath**, **pleuritic chest pain**, and **tachycardia**. Concern for **pulmonary embolism**.



Modeling Patient Timelines for AI

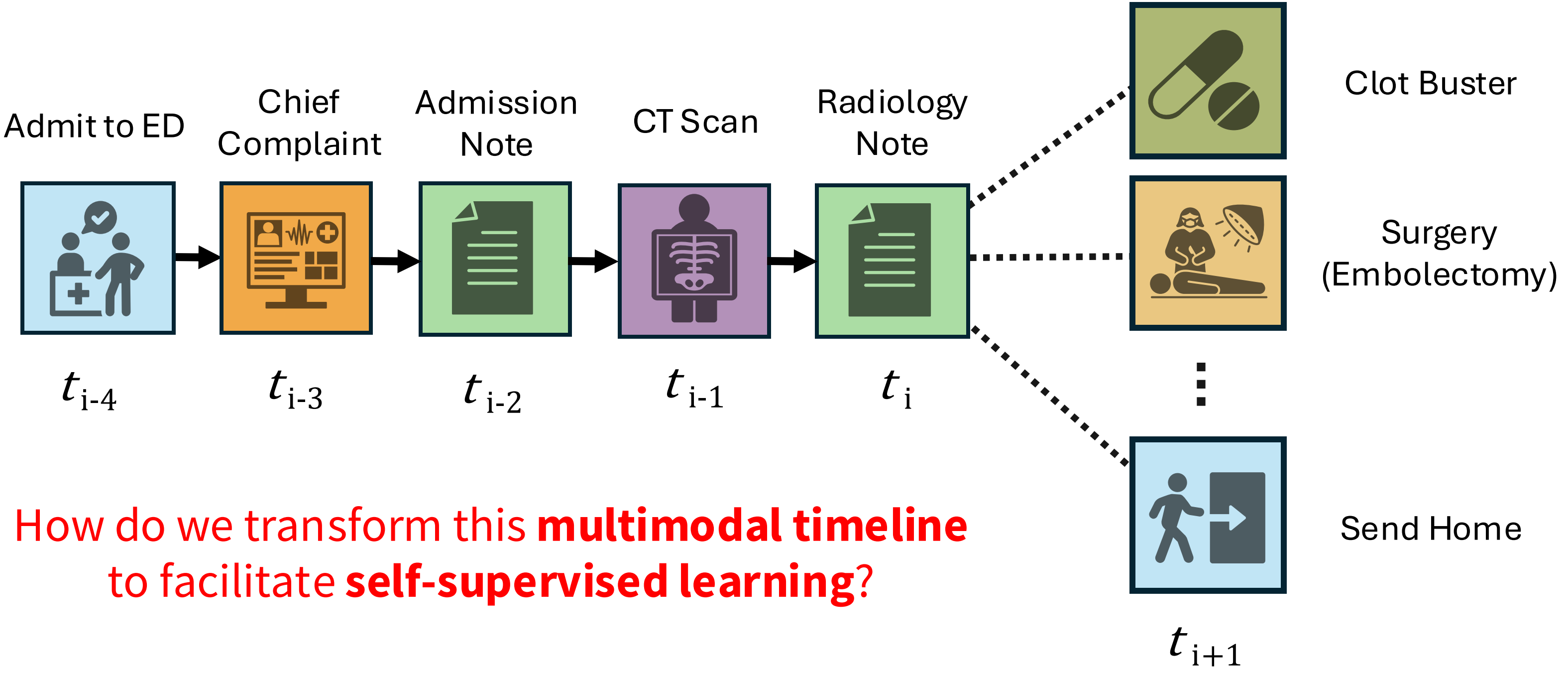
CASE: Patient **presents to ED** with sudden onset **shortness of breath**, **pleuritic chest pain**, and **tachycardia**. Concern for **pulmonary embolism**.



We’ve transformed our patient timeline into an **autoregressive / LLM-like process**

Modeling Patient Timelines for AI

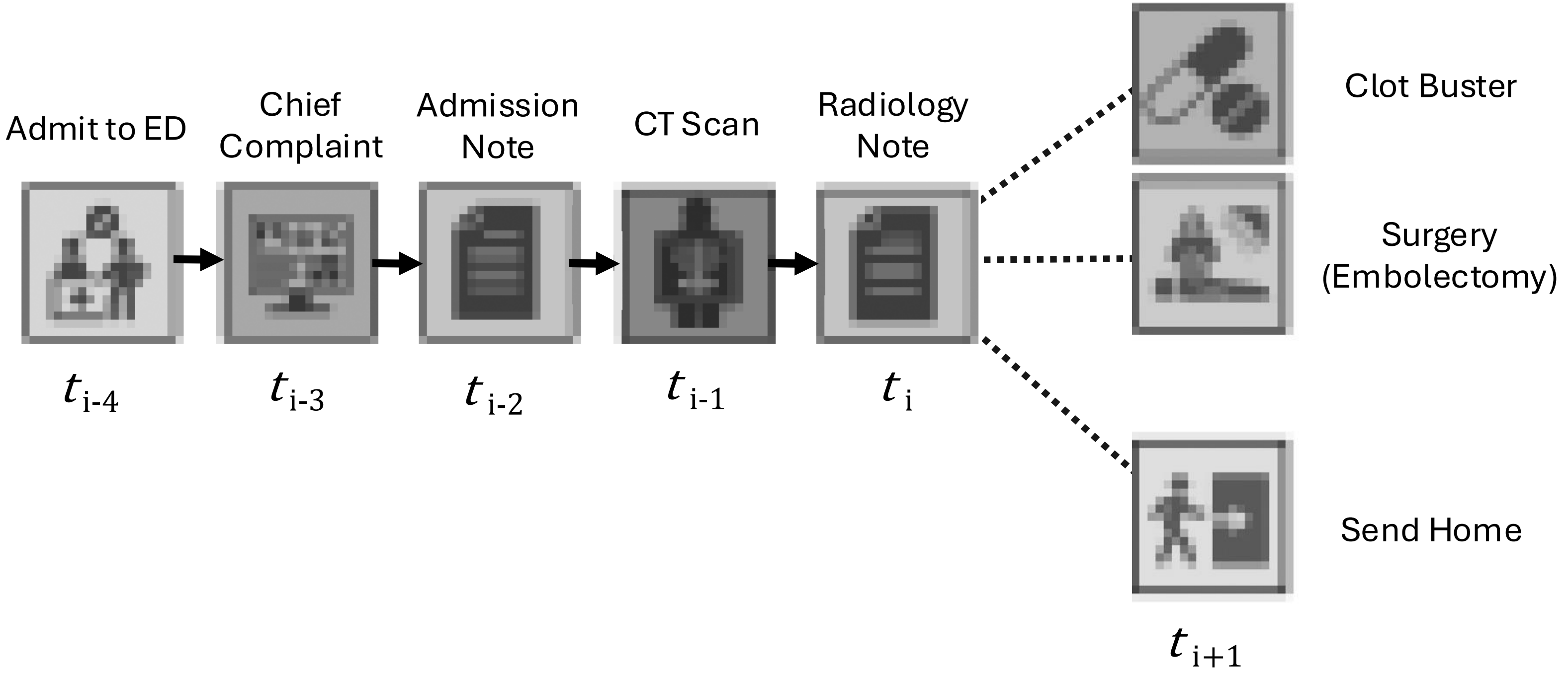
Hypothesis: A model that accurately **predicts future health states**, based on patient history, **encompasses many proposed use cases of medical AI**



How do we transform this **multimodal timeline** to facilitate **self-supervised learning**?

Modeling Patient Timelines for AI

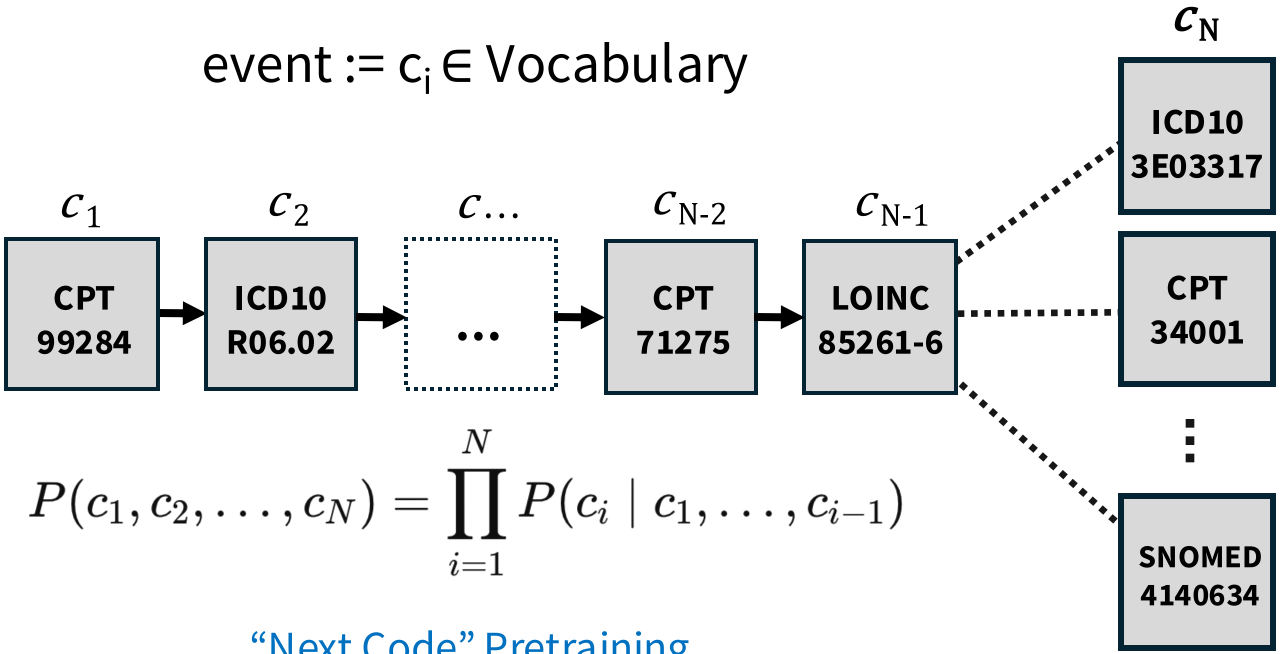
We do have a **“low-res” version of this timeline** readily available …



Autoregressive Modeling of Structured EHR Timelines

Map events to ontologies to define a "language" based on medical codes

event := $c_i \in$ Vocabulary

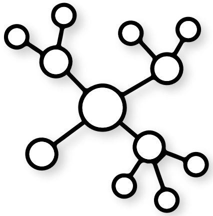


$$P(c_1, c_2, \dots, c_N) = \prod_{i=1}^{N} P(c_i \mid c_1, \dots, c_{i-1})$$

"Next Code" Pretraining

Ontology Mapping

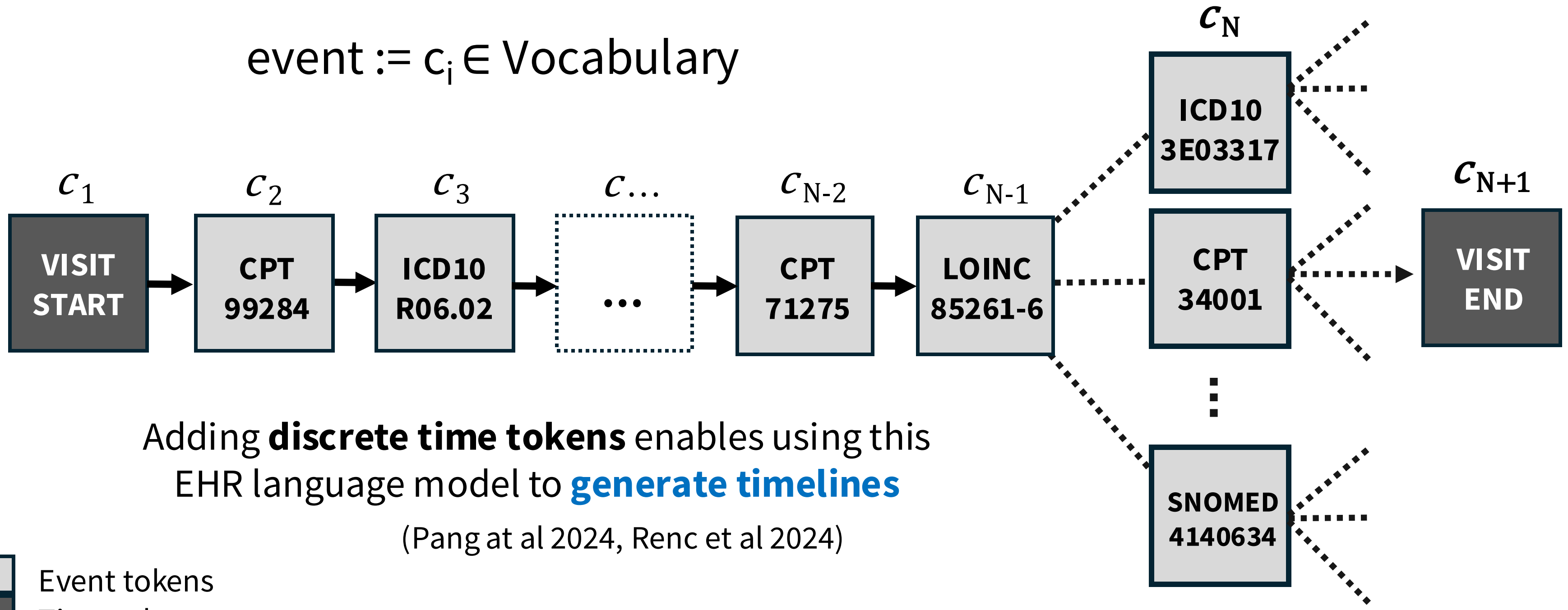
Admit to ED → **CPT 99284**
Shortness of Breath → **ICD10 R06.02**
Chest Pain → **ICD10 R07.1**
Tachycardia → **ICD10 R00.0**
Admission Note → **LOINC 47039-3**
CT Scan → **CPT 71275**
Radiology Note → **LOINC 85261-6**
Thrombolytic → **ICD10 3E03317**
Embolectomy → **CPT 34001**
Discharge to Home → **SNOMED 4140634**



Autoregressive Modeling of Structured EHR Timelines

Map events to ontologies to define a "language" based on medical codes

event := $c_i \in$ Vocabulary

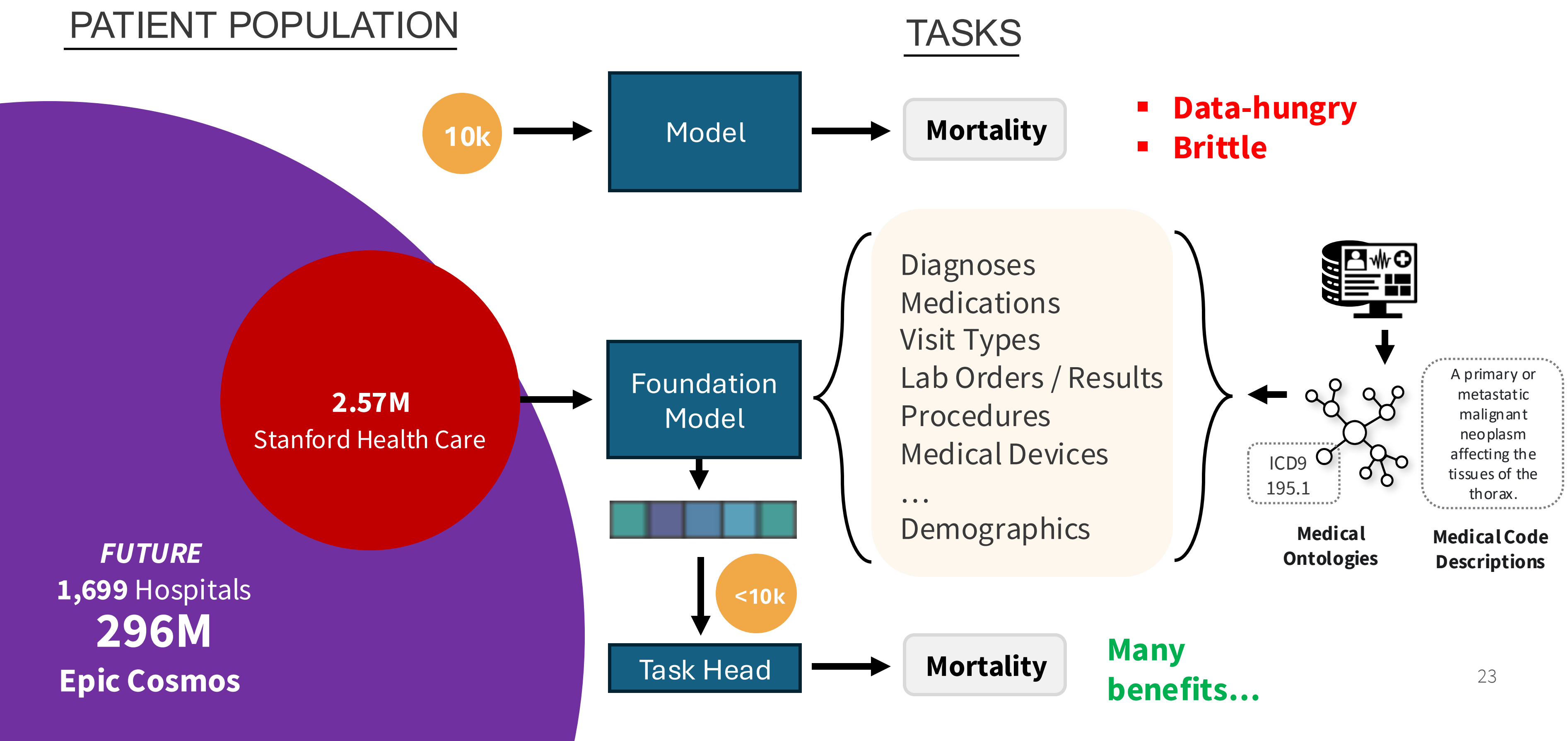


Adding **discrete time tokens** enables using this EHR language model to **generate timelines**

(Pang at al 2024, Renc et al 2024)

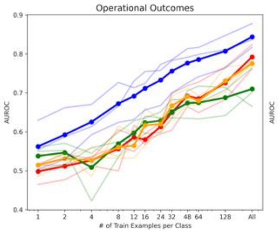
Event tokens
Time tokens

Self-Supervised Training of an EHR Foundation Model



Validating Benefits of EHR Foundation Models

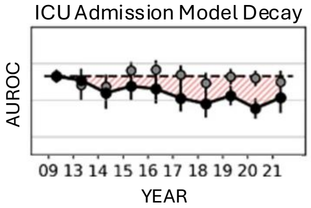
Data Efficiency



SOTA **few-shot learning**
SOTA **overall performance**

(Wornow et al. 2023)
(Steinberg et al. 2020)

Robustness



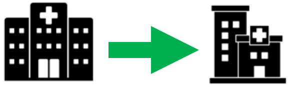
Improved robustness to **temporal distribution shifts**

(Guo et al. 2023)

Improved performance across key **subgroups** (pediatrics)

(Lemmon et al. 2023)

Cross-Site Adaptability



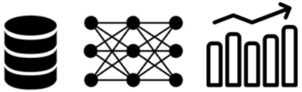
Hospital A **Hospital B**

Transfer **pretrained models** across hospitals

Require **up to 90% less** pretraining data

(Guo et al. 2024)

Reproducible EHR Benchmarking

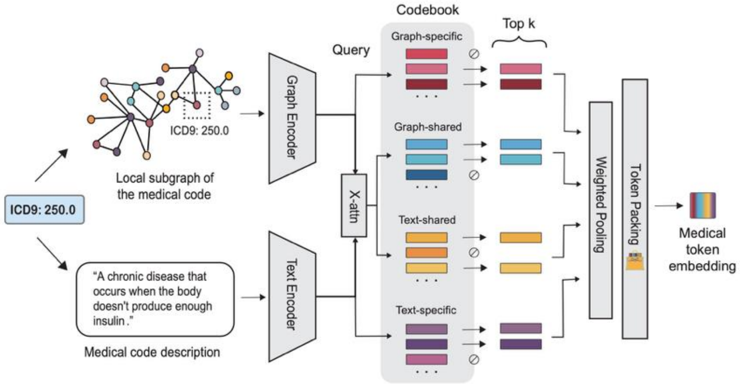


First **externally verifiable** evaluation of **EHR foundation models** on longitudinal data

(Wornow et al. 2025)
(Arnrich et al. 2024)
(Steinberg et al. 2024)
(Wornow et al. 2023)
(Huang et al. 2023)

Publication Venue
Medical / Informatics
Computer Science

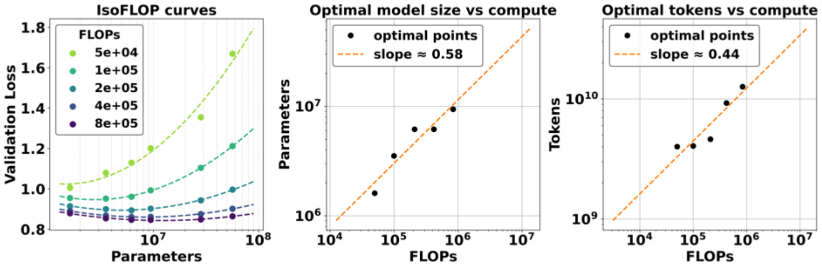
Inherits Many of the Benefits of Large Language Models



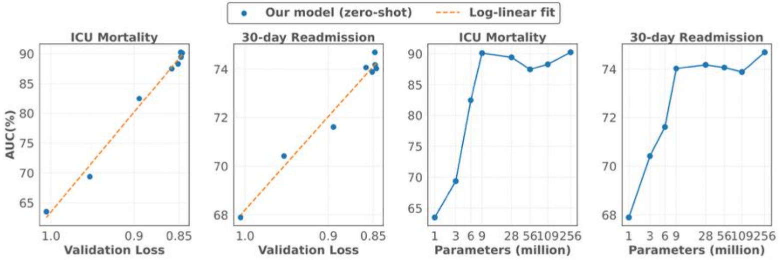
Tokenizer Abstractions

MedTok (Su et al. 2025)

Derive Scaling Laws



Improve Zero Shot Learning



Zhang et al. 2025

EHR Modeling at Smaller Pretraining Scale

Autoregressive LLMs can capture long-distance dependencies given **sufficient data and parameters**

Natural Language

≥ 7B parameters

≥ 500B-1T tokens

EHR

143M parameters

3.5B tokens

285x less data

Can we train a **small, data-constrained** EHR foundation model to learn embeddings that **capture more time information about the future**?

MOTOR: A TIME-TO-EVENT FOUNDATION MODEL FOR STRUCTURED MEDICAL RECORDS

Ethan Steinberg[1]*, Jason Alan Fries[2]*, Yizhe Xu[2], Nigam H. Shah[3]
[1]Department of Computer Science, Stanford University
[2]Center for Biomedical Informatics Research, Stanford University
[3]Stanford University, Stanford Health Care
{ethanid, jfries, yizhex, nigam}@stanford.edu

Key Concepts in Time-to-Event Modeling

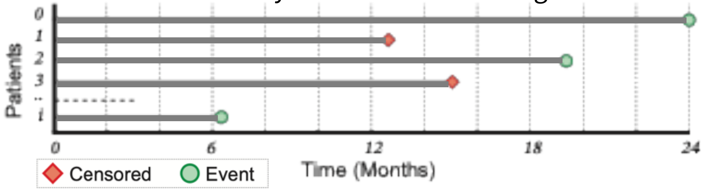
Model the **time until an event occurs** (e.g., death) while accounting for **censoring**

Censoring

Event times are **not fully observed by end of a study period**

$\boxed{(X_i, T_i)}$ **BIASED** (X_i, T_i, δ_i) $\delta_i = \begin{cases} 1 & \text{event observed} \\ 0 & \text{censored} \end{cases}$

Mortality Event Time Censoring



Survival Function

The probability that an event has not occurred as of time t

$$S(t) = \Pr(T > t)$$

Hazard Rate Function

Instantaneous risk of an event at time t, given survival up to t

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

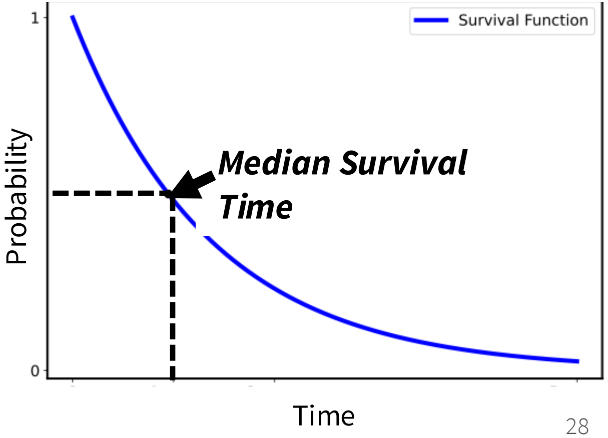
Event's "speed" at each moment

$$S(t) = \exp\left(-\int_0^t h(u)\,du\right)$$

Survival depends on cumulative hazard over time

Learn a patient representation $R_i = f_\theta(X_i)$ for estimating **personalized hazard rates**

Survival Curve



Intuition Behind the Pretraining Objective

Hypothesis: Multi-task learning (MTL) will capture generalizable TTE features

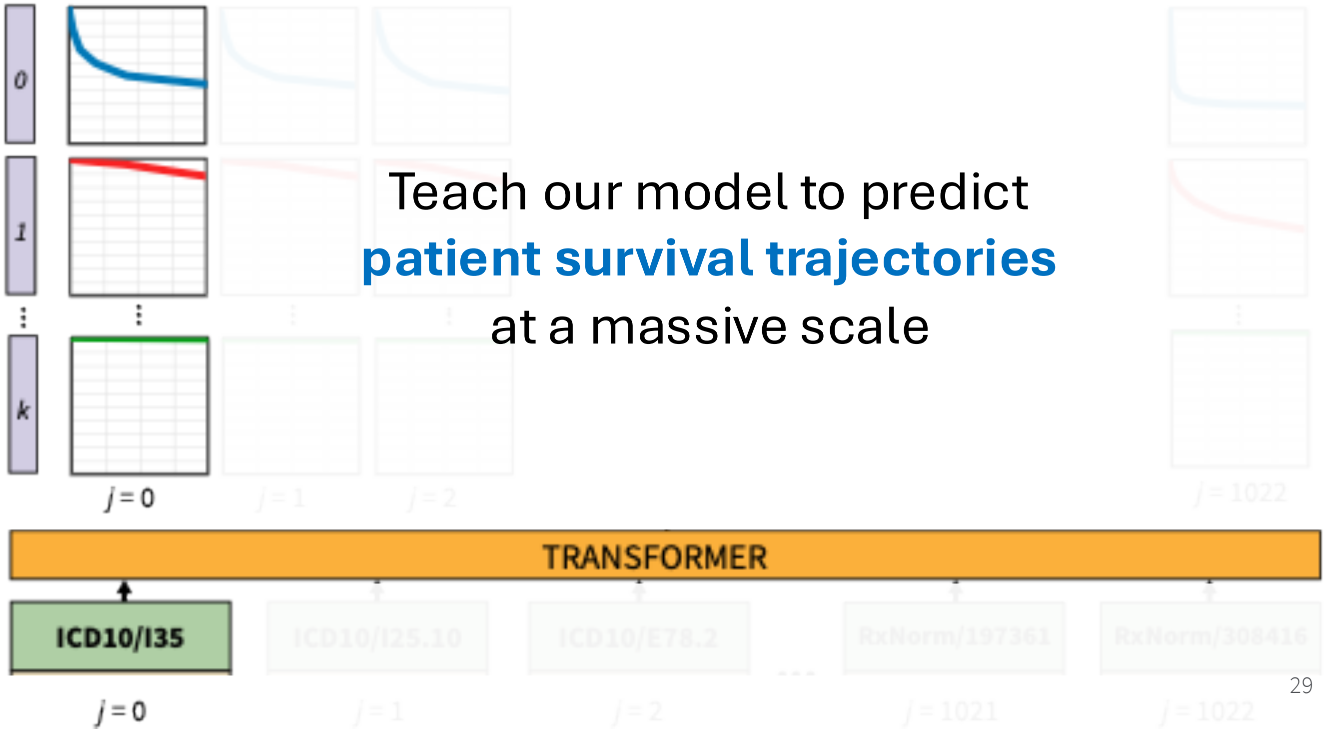
Select k TTE pretraining tasks

$k \leq 16{,}392$

Event j

Shared Decoder

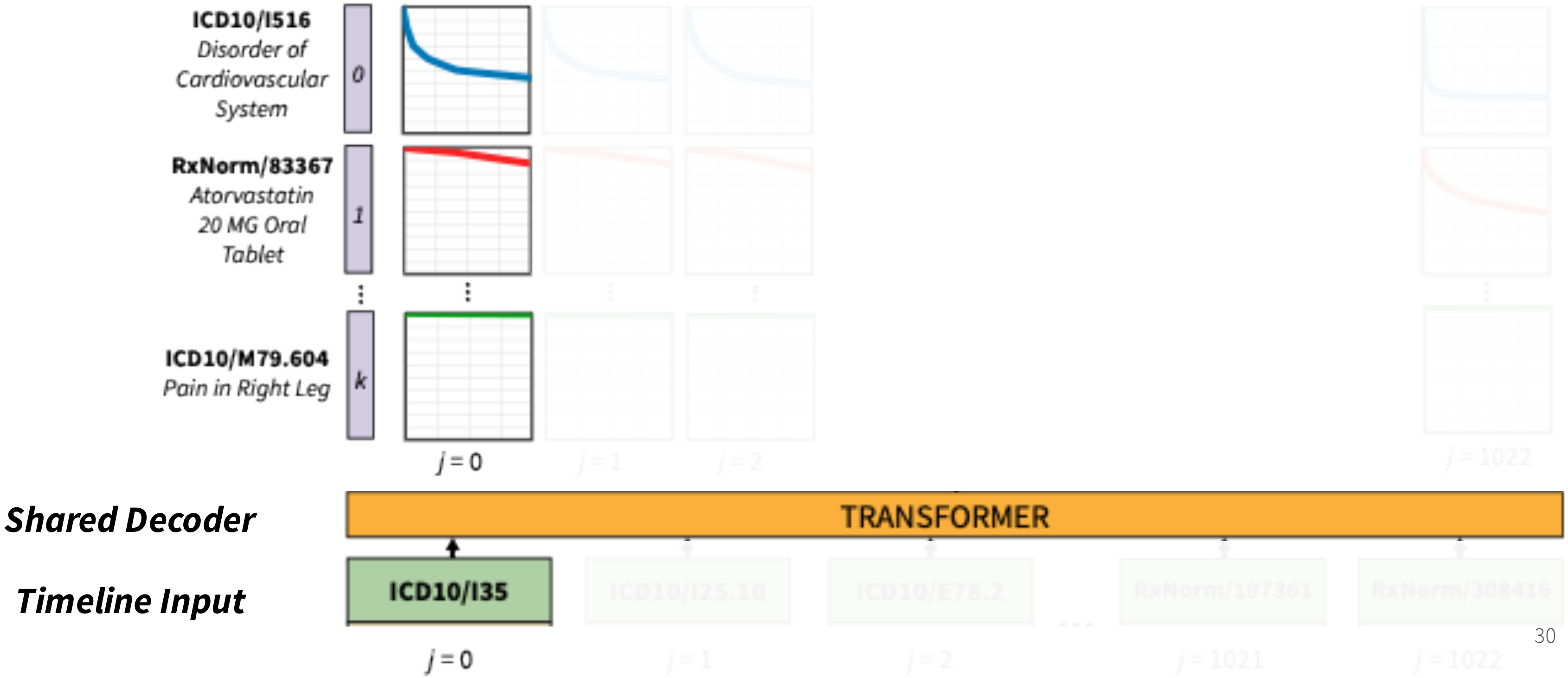
Timeline Input



Teach our model to predict
patient survival trajectories
at a massive scale

Intuition Behind the Pretraining Objective

Hypothesis: Multi-task learning (MTL) will capture generalizable TTE features



Intuition Behind the Pretraining Objective

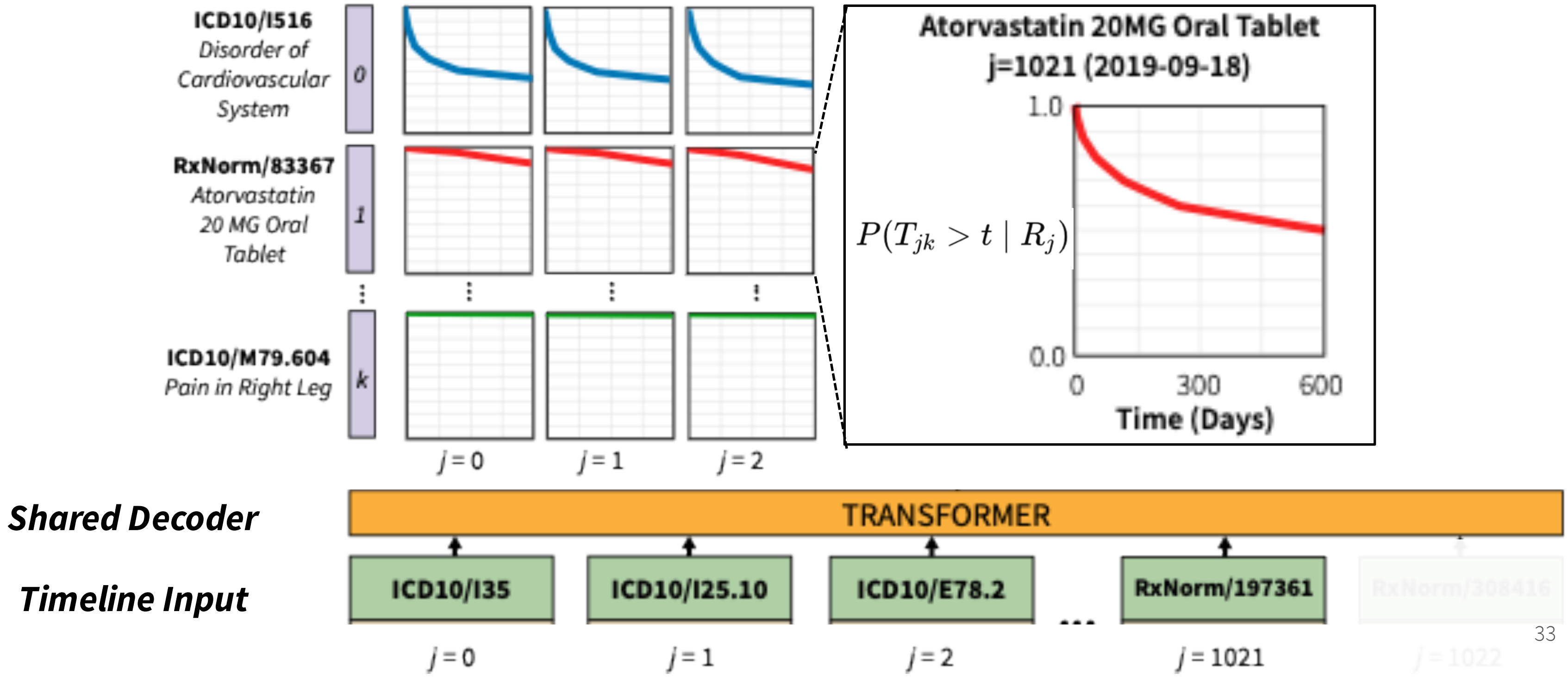
Hypothesis: Multi-task learning (MTL) will capture generalizable TTE features

Intuition Behind the Pretraining Objective

Hypothesis: Multi-task learning (MTL) will capture generalizable TTE features

Intuition Behind the Pretraining Objective

Hypothesis: Multi-task learning (MTL) will capture generalizable TTE features



Datasets & Tasks

Datasets

STANFORD STARR-OMOP (EHR)
2.7M Patients
3.5B Events

Evaluation Tasks

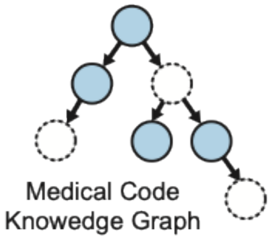
Celiac Disease | Stroke
Pancreatic Cancer | NAFLD
Heart Attack | Lupus

ICD-10

Rule-based labeling

We remove these tasks from the pretraining set

Pretraining Tasks



Entropy-Ranked Vertex Cover for Task Selection

Intuition: We pick k tasks that **maximize diversity** by selecting nodes whose values are **least predictable** given their parents

$k \leq 16{,}392$

13 Chest X-ray Findings

NLP-based

Measures generalization to labels not derived from codes

Results: MOTOR vs. Baselines

Pretrained MOTOR-Probe & **MOTOR-Finetune** outperform
SOTA on all tasks

Avg improvement: **+4.6%**

Method	Dataset	Celiac	HA	Lupus	NAFLD	Cancer	Stroke
Cox PH	EHR-OMOP	0.689	0.761	0.770	0.726	0.793	0.779
DeepSurv	-	0.704	0.823	0.790	0.800	0.811	0.830
DSM	-	0.707	0.828	0.784	0.805	0.809	0.835
DeepHit	-	0.695	0.826	0.807	0.805	0.809	0.833
RSF	-	0.729	0.836	0.787	0.802	0.824	0.840
MOTOR-Scratch	-	0.696	0.795	0.803	0.821	0.777	0.831
MOTOR-Probe	-	0.802	0.884	0.850	0.859	0.865	0.874
MOTOR-Finetune	-	**0.802**	**0.887**	**0.863**	**0.864**	**0.865**	**0.875**

Results: Autoregressive vs. TTE Pretraining

Overall Performance

Objective	Celiac	HA	Lupus	NAFLD	Cancer	Stroke
RSF	0.729	0.836	0.787	0.802	0.824	0.840
Next Code	0.774	0.862	0.842	0.860	0.860	0.857
Time-to-Event	**0.802**	**0.887**	**0.863**	**0.864**	**0.865**	**0.875**

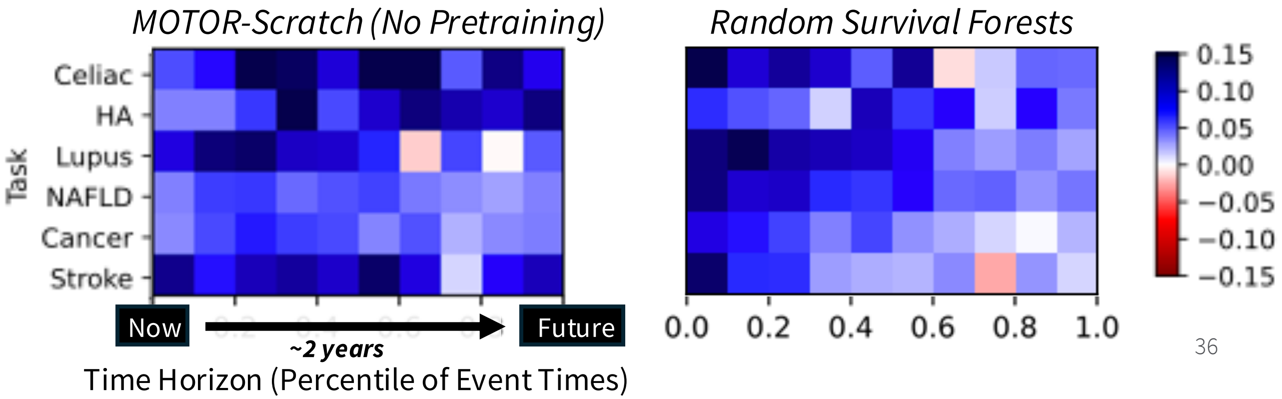
Autoregressive beats SOTA (RSF)

…but **TTE beats autoregressive** by **~2%**

Performance Comparison over Long Time Horizons

Pretraining is the key driver of performance

Performance Deltas of MOTOR with TTE Pretraining Versus:



Can this same TTE approach handle **high-dimensional**, **multimodal data** at a **single time point**?

Eras of Training Supervision for Unstructured Data

Manual Labeling

2012

Weakly Supervised Learning

2016

Self-Supervised Learning

2018

LLM-Accelerated Data Curation

2020

Reasoning / Agents Synthetic Data

Today

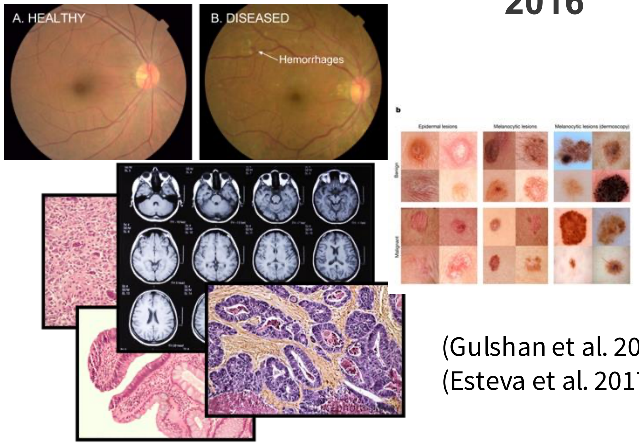
Medical Imaging Datasets

2017 - ChestX-ray14: **112k** x-rays

..

2023 - PMC-OA: **1.65M** image / caption pairs

2024 - MedTrinity-25M: **25M** images / annotations



(Gulshan et al. 2016)
(Esteva et al. 2017)

Multimodal Healthcare Foundation Models

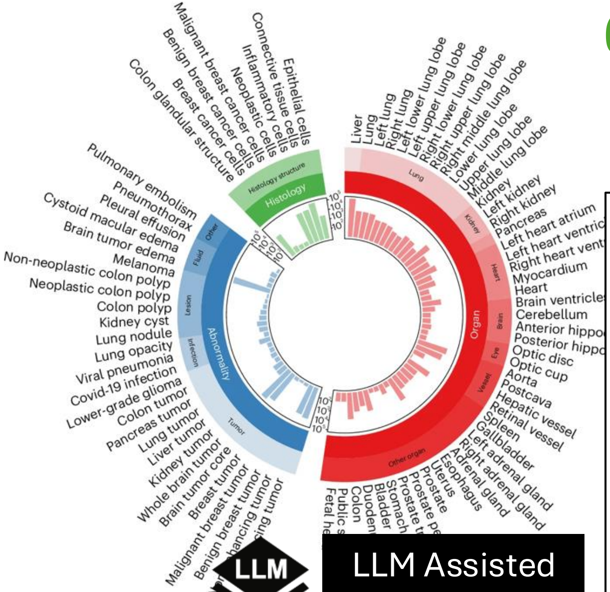
Image

Text Description

Segmentation Mask

...

Aligned Data For Self-Supervised Learning

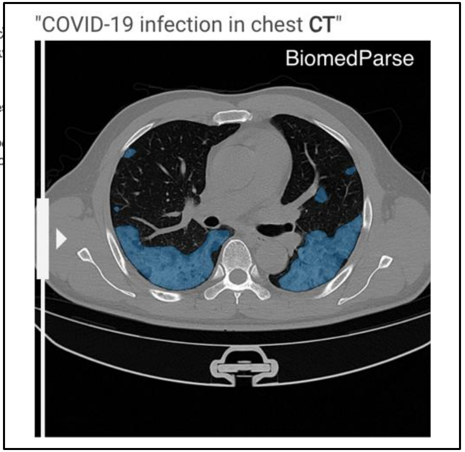


LLM LLM Assisted Curation

6.8M image/mask/description

82 major object types

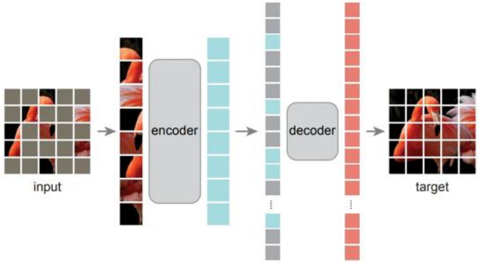
9 imaging modalities



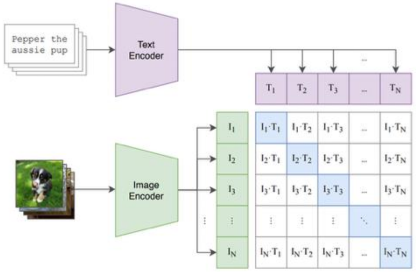
BiomedParse (Zhao et al. 2024) *Nature Methods*

Vision Pretraining Approaches

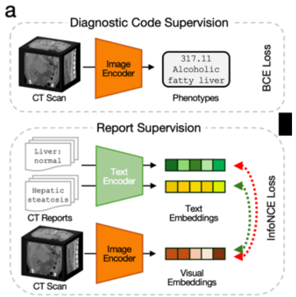
Current self-supervised learning methods **work well for diagnosis** (classify “now”)



(He et al. 2021)



(Radford et al. 2021)



(Blankemeier et al. 2024)

Masked Autoencoder (MAE)

Contrastive (e.g., CLIP)

Hybrid (e.g., Merlin)

Image

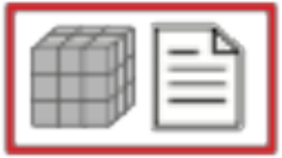


Image + Text

Image + Text + Diagnosis Codes

Curate heterogenous task supervision from the EHR

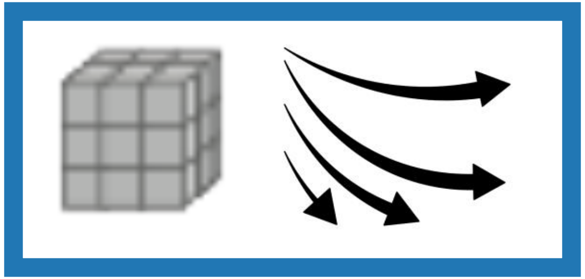


Image + Survival Trajectories

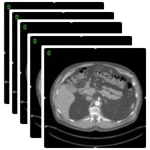
Published as a conference paper at ICLR 2025

TIME-TO-EVENT PRETRAINING FOR 3D MEDICAL IMAGING

Zepeng Huo[1]*, **Jason Alan Fries**[1]*, **Alejandro Lozano**[2]*,
Jeya Maria Jose Valanarasu[3,5], **Ethan Steinberg**[1,6], **Louis Blankemeier**[2],
Akshay S. Chaudhari[2,5,9,10], **Curtis Langlotz**[4,5,8,10], **Nigam H. Shah**[4,5,7,8,9,11]

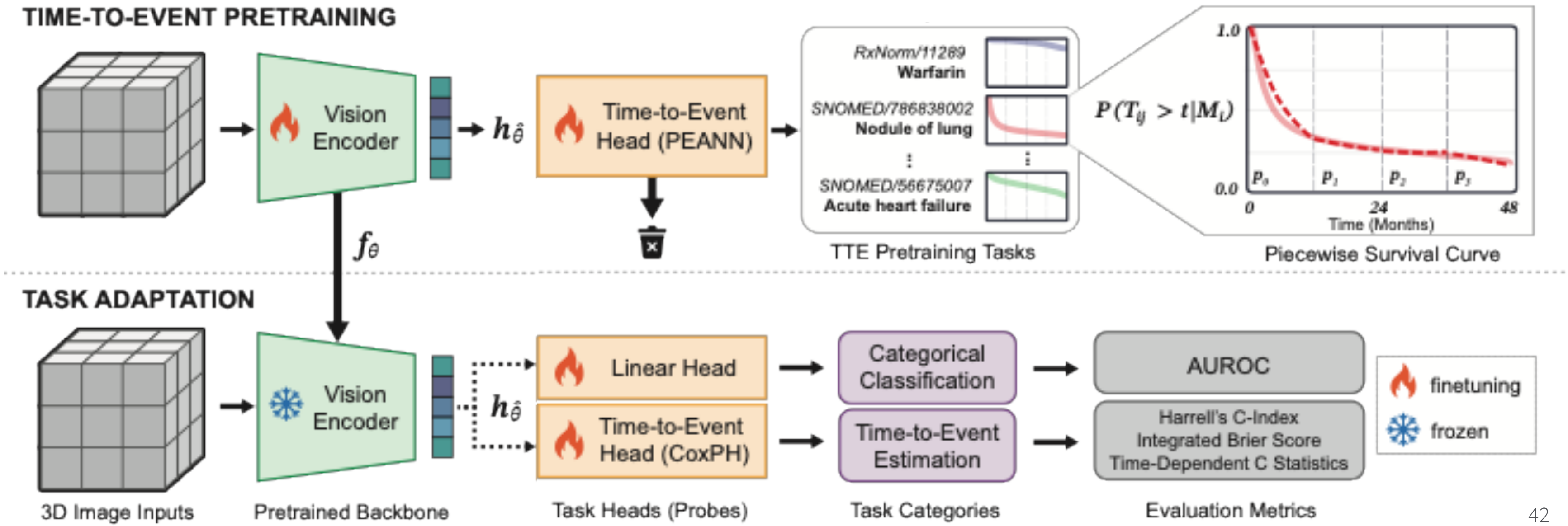
Time-to-Event Pretraining for 3D Imaging

Pulmonary Embolisms



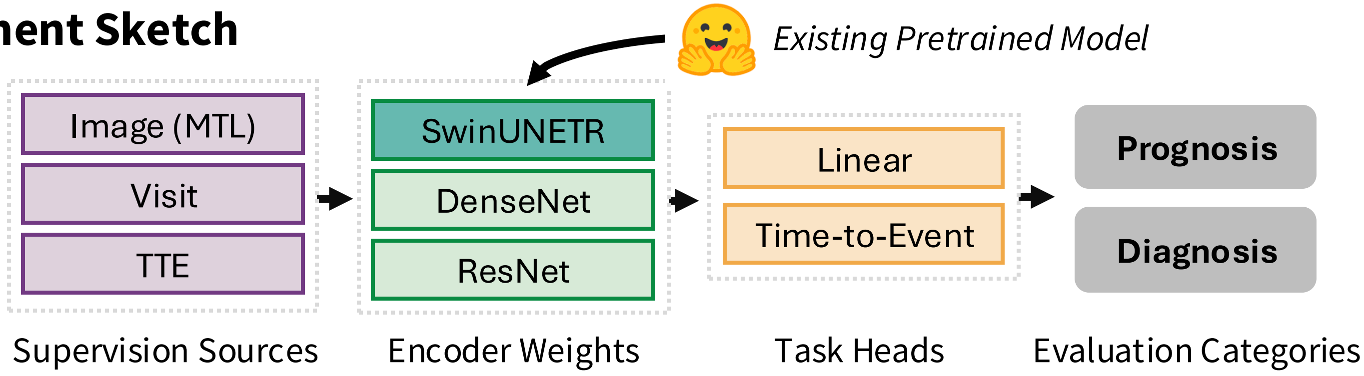
18,945 CT Scans
(4.2 Million 2D images)

- Same pretraining setup as MOTOR
- **Single time point** (not dynamic)
- Pretraining a 3D image encoder



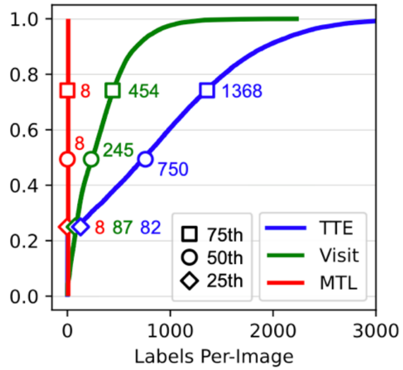
Summary of Results: TTE (Continued) Pretraining

Experiment Sketch



Prognosis **Average** **+23.7% AUROC** and **+29.4% Harrell's C-index** across 8 prognostic tasks

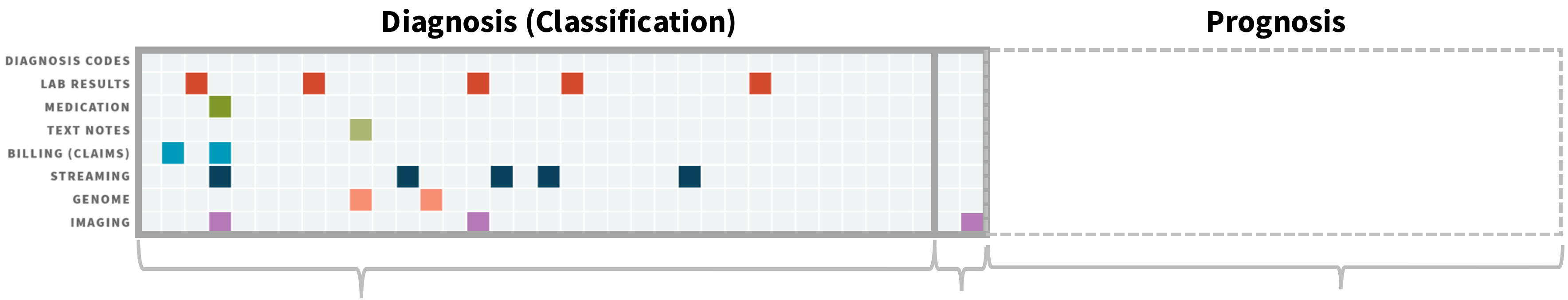
Diagnosis TTE pretraining does **NOT degrade performance** across 9 diagnostic tasks



TTE increases label density by **3x**

Human-AI Teaming:
Natural Language Interfaces

Ascertaining the Needs of Stakeholders



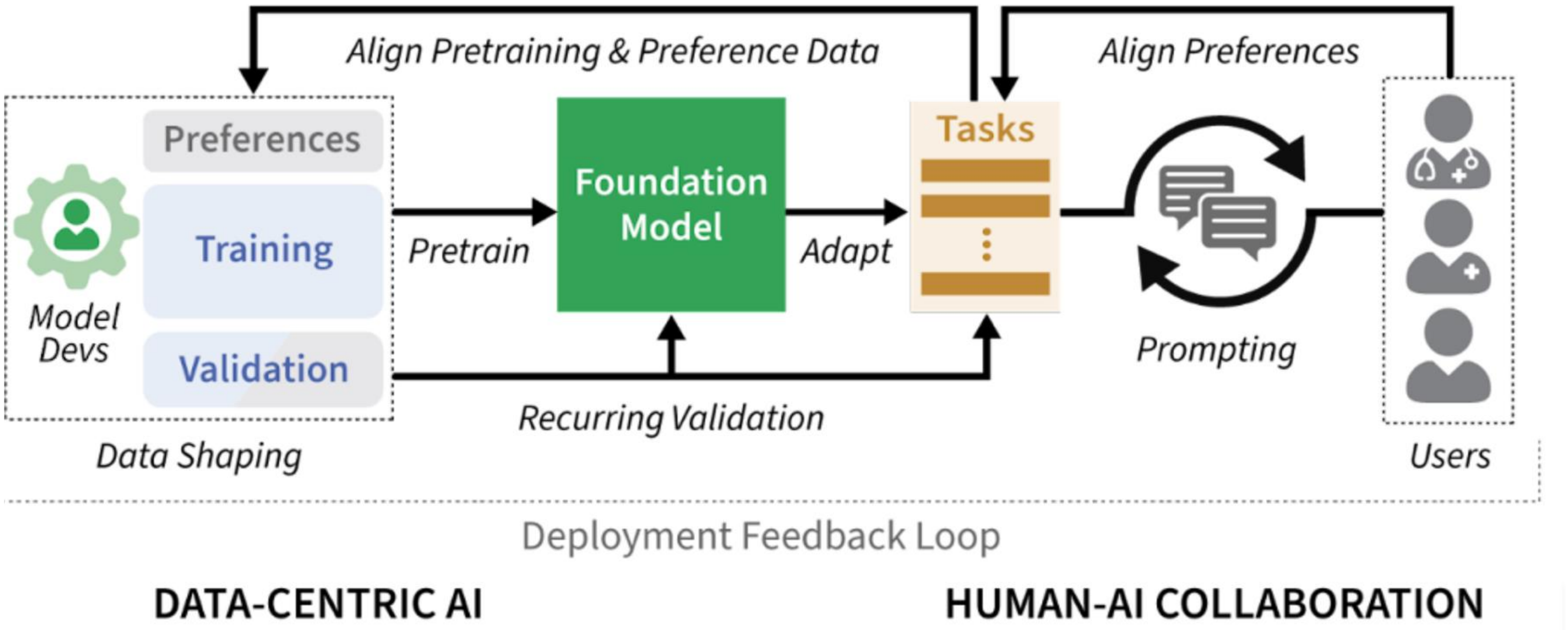
What Occurred in the Past? | **What is Occurring Now?** | **Predict Future Risks & Intervention Benefits**

Example ML Applications: ? | ? | ?

Many stakeholder groups with **distinct needs**

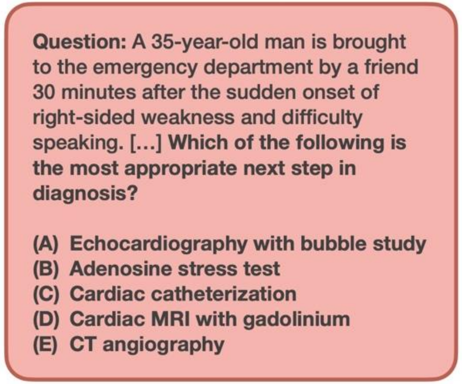
Stakeholders: Clinicians | Hospital Administrators | Insurance Providers | Pharma | Regulatory Agencies | Patients | ... | Researchers

Creating Feedback Loops with Real-world Data and Real Users



Multiple Choice vs. Longitudinal Patient Timelines

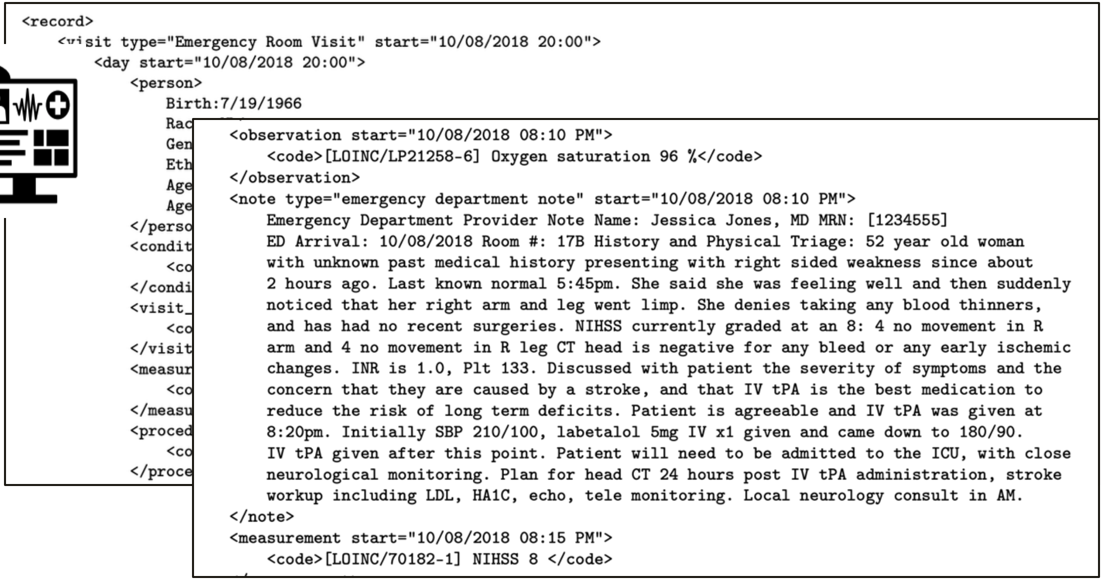
MedQA



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

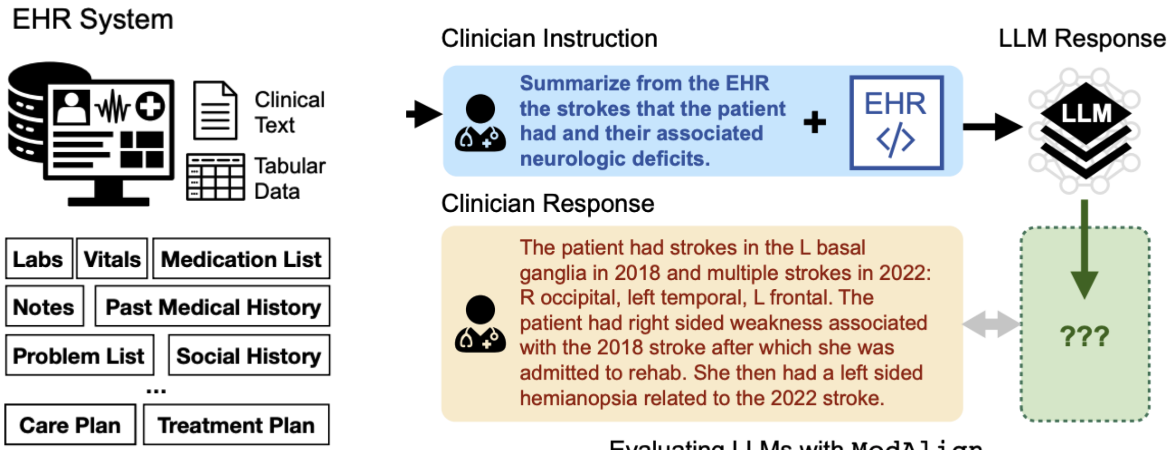


```
<record>
    <visit type="Emergency Room Visit" start="10/08/2018 20:00">
        <day start="10/08/2018 20:00">
            <person>
                Birth:7/19/1966
                Rac
                Gen
                Eth
                Age
                Age
            </perso
            <condit
                <co
            </condi
            <visit_
                <co
            </visit
            <measur
                <co
            </measu
            <proced
                <co
            </proce

<observation start="10/08/2018 08:10 PM">
    <code>[LOINC/LP21258-6] Oxygen saturation 96 %</code>
</observation>
<note type="emergency department note" start="10/08/2018 08:10 PM">
    Emergency Department Provider Note Name: Jessica Jones, MD MRN: [1234555]
    ED Arrival: 10/08/2018 Room #: 17B History and Physical Triage: 52 year old woman
    with unknown past medical history presenting with right sided weakness since about
    2 hours ago. Last known normal 5:45pm. She said she was feeling well and then suddenly
    noticed that her right arm and leg went limp. She denies taking any blood thinners,
    and has had no recent surgeries. NIHSS currently graded at an 8: 4 no movement in R
    arm and 4 no movement in R leg CT head is negative for any bleed or any early ischemic
    changes. INR is 1.0, Plt 133. Discussed with patient the severity of symptoms and the
    concern that they are caused by a stroke, and that IV tPA is the best medication to
    reduce the risk of long term deficits. Patient is agreeable and IV tPA was given at
    8:20pm. Initially SBP 210/100, labetalol 5mg IV x1 given and came down to 180/90.
    IV tPA given after this point. Patient will need to be admitted to the ICU, with close
    neurological monitoring. Plan for head CT 24 hours post IV tPA administration, stroke
    workup including LDL, HA1C, echo, tele monitoring. Local neurology consult in AM.
</note>
<measurement start="10/08/2018 08:15 PM">
    <code>[LOINC/70182-1] NIHSS 8 </code>
```

Longitudinal Patient Timelines

Instruction Tuning: Aligning with Clinical Needs



Evaluating LLMs with MedAlign

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

- **15** clinicians / **7** specialties
- 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.

Instruction Tuning: Aligning with Clinical 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.

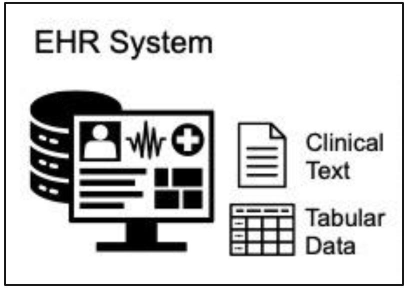
The Perils of Climbing the Wrong Hill….

MedQA

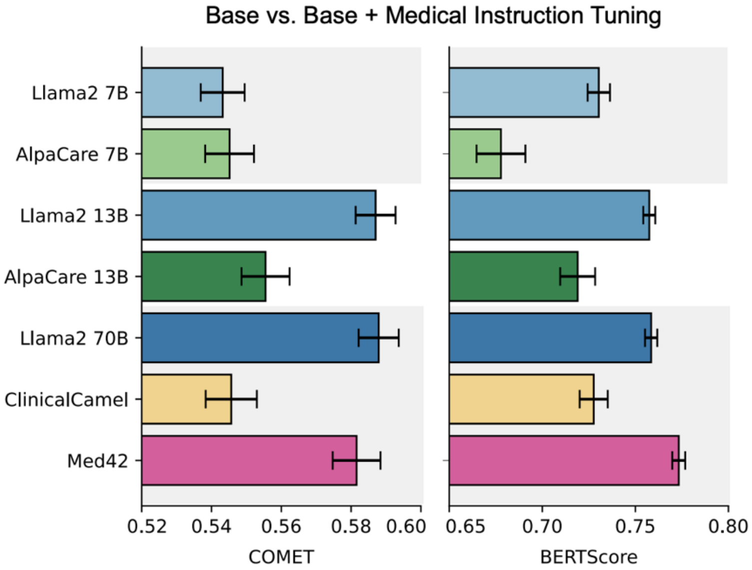
Question: A 35-year-old man is brought to the emergency department by a friend 30 minutes after the sudden onset of right-sided weakness and difficulty speaking. […] **Which of the following is the most appropriate next step in diagnosis?**

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

MedAlign

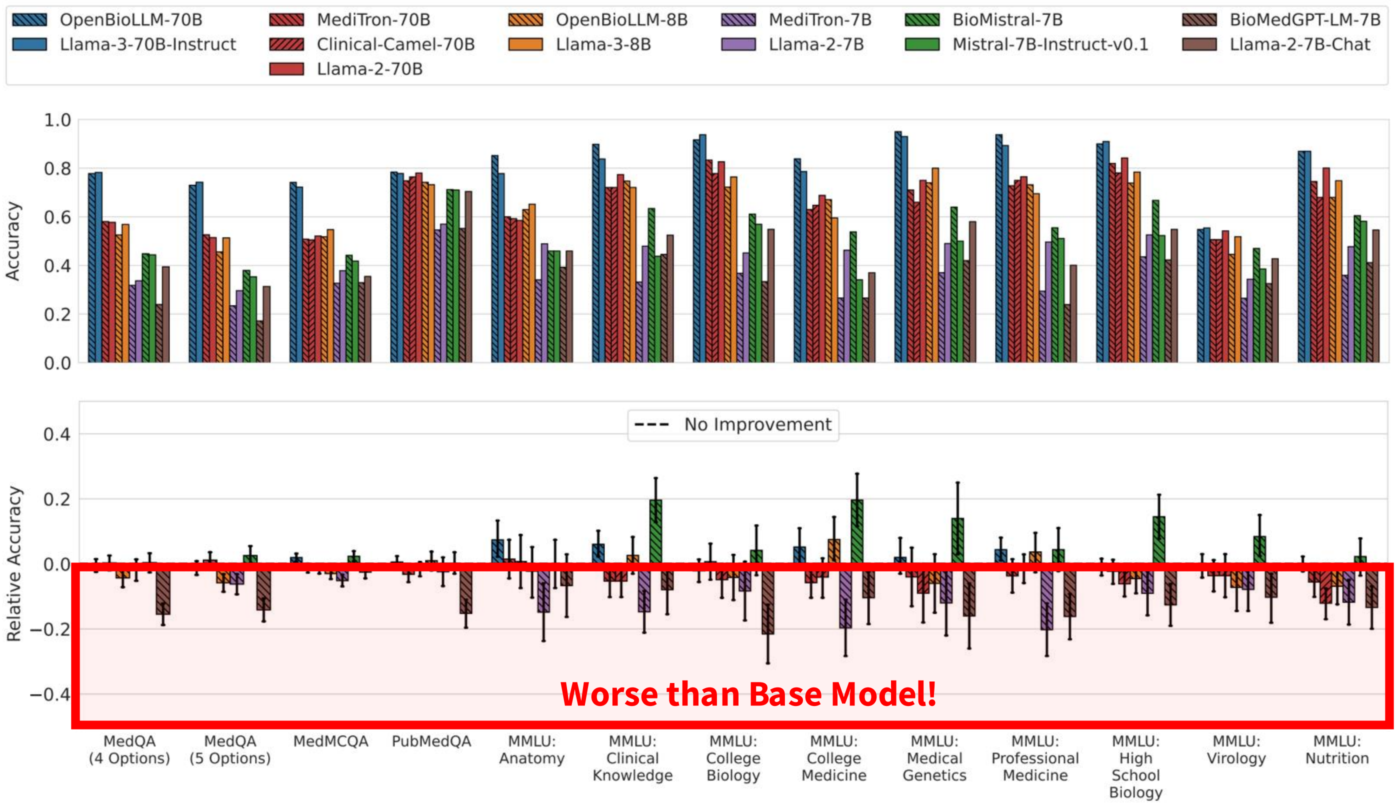


33k to 1.6M
tokens per patient



Short instruction tuning tasks for medicine
actually hurt performance on MedAign

The Perils of Climbing the Wrong Hill....



Medical LLM (3-shot)

12.1% better
49.8% tie
38.2% worse

(Jeong et al. 2024)

Little improvement over base models!

Future:
Research Opportunities

Longitudinal, Multimodal EHR Dataset & Model Releases

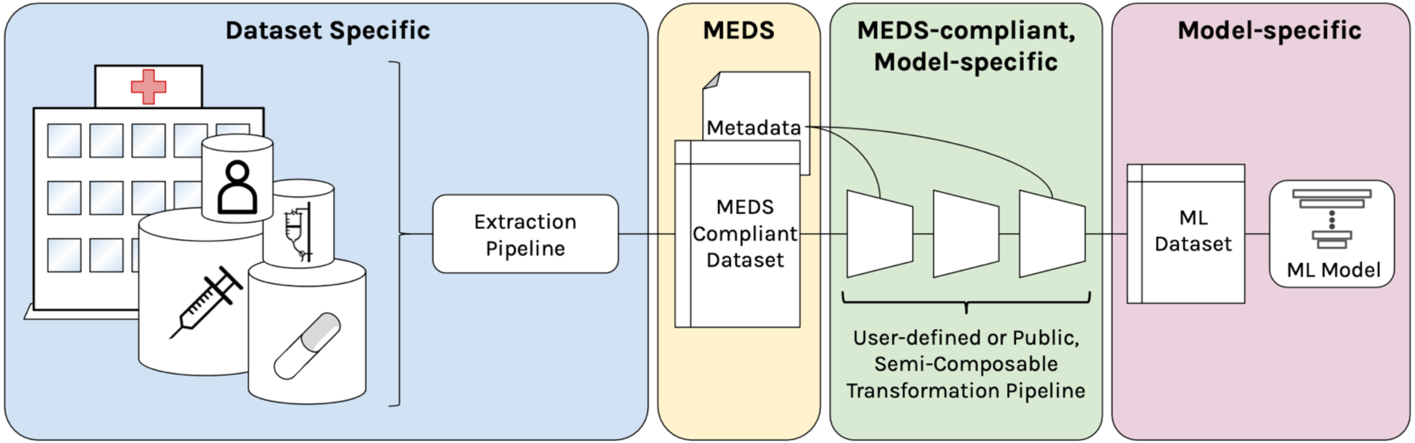
Dataset	Task	Technical Challenge	Example	Tabular	Images	Notes
EHRSHOT	Risk Stratification	Few-Shot Learning	*What is the likelihood that this patient gets a diagnosis of pancreatic cancer within the next year?*	✓	✗	✗
INSPECT	Time-to-Event Modeling	Multimodal Learning	*When is chronic pulmonary hypertension most likely to develop*	✓	✓	✓
MedAlign	Instruction Following	Long-Context Learning & Temporal Reasoning	*From this EHR, summarize the patient's history of strokes and the resulting neurologic deficits.*	✓	✗	✓

26k Patients **295M** Events **442k** Visits

REDIVIS

https://redivis.com/ShahLab

Medical Event Data Standard (**MEDS**)



Open Data Schema for Health AI Practitioners

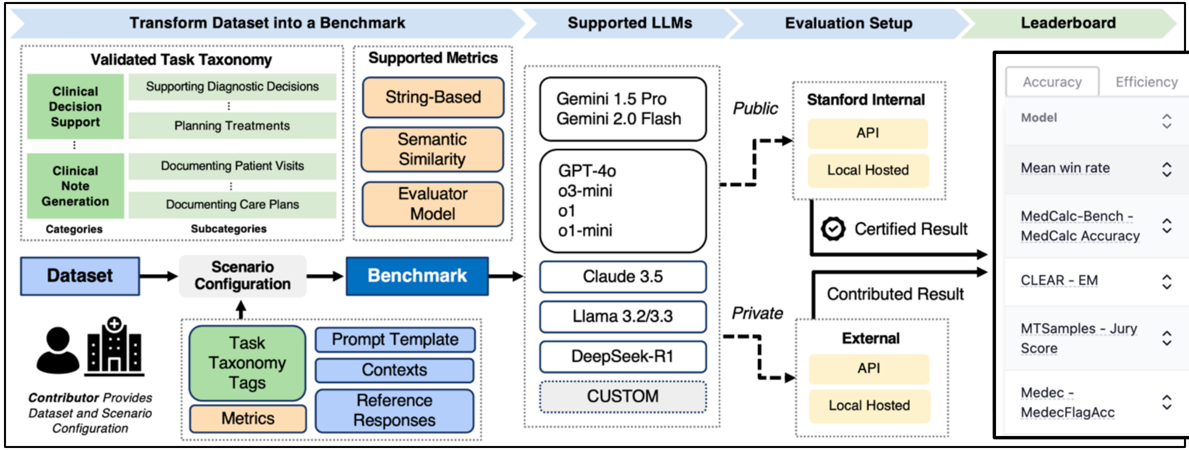
Bert Arnrich, Edward Choi, Jason A. Fries, Matthew B. A. McDermott, Jungwoo Oh, Tom J Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, Robin van de Water

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

Evaluation is Critical to Real-World Impact

Stanford MedHELM
Community evaluation framework for benchmarking healthcare LLMs

https://medhelm.stanford.edu/



Accuracy | Efficiency | General information

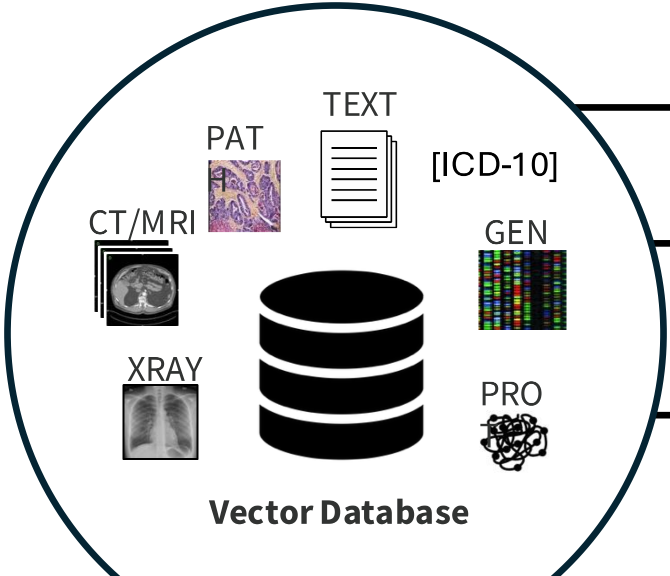
Model	DeepSeek R1	o3-mini (2025-01-	Claude 3.7 Sonnet	Claude 3.5 Sonnet
Mean win rate	**0.663**	0.641	0.636	0.634
MedCalc-Bench - MedCalc Accuracy	**0.348**	0.34	0.21	0.218
CLEAR - EM	0.83	**0.831**	0.647	0.663
MTSamples - Jury Score	4.533	**4.604**	4.426	4.383
Medec - MedecFlagAcc	0.591	**0.687**	0.628	0.521

The Future

We must build systems for patient timeline data that are **fast**, **multimodal**, and **interactive**

"I can't just go to the medical records department to have them pull 500 charts on a certain type of patient."

Byrne Lee, MD, Clinical Professor, Surgical Oncology, Stanford Health Care



Design **chat interfaces diverse needs**

Find **similar patients** to inform decision making

Automate **feedback loops** to **improve** model embeddings

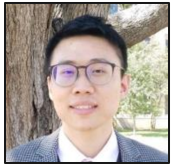
Team Science

Shah Lab and Collaborating Researchers


Jason Fries


Ethan Steinberg


Michael Wornow


Frazier Huo


Alejandro Lozano


Hejie Cui


Alyssa Unell


Suhana Bedi


Louis Blankemeier


Keith Morse


Akshay Chaudhari


Curtis Langlotz


Nigam Shah

Governance, Privacy, and Licensing


Mariko Kelly


Julie Marie Romero


Jonathan Gortat


Scott Edmiston


Reed Sprague

Technology & Digital Solutions


Natasha Flowers


Joseph Mesterhazy


Priya Desai

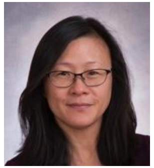

Somalee Datta


Todd Ferris

External Collaborating Researchers


Lawrence Guo


Joshua Lemmon


Lillian Sung

Questions

jason-fries@stanford.edu