Realizing the Promise of Foundation Models in Healthcare

Jason Fries, PhD Research Scientist, Center for Biomedical Informatics Research
Ethan Steinberg PhD Candidate, Department of Computer Science

Stanford MEDICINE | Clinical Excellence Research Center

BMIR Stanford Center for Biomedical Informatics Research

Center for Research on Foundation Models

Stanford AIMI
Center for Artificial Intelligence in Medicine and Imaging

HAI Stanford University Human-Centered Artificial Intelligence

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

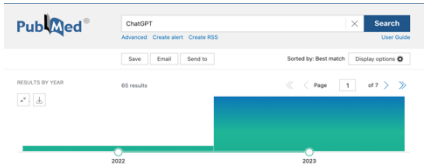
At least four articles credit the AI tool as a co-author, as publishers scramble to regulate its use.

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

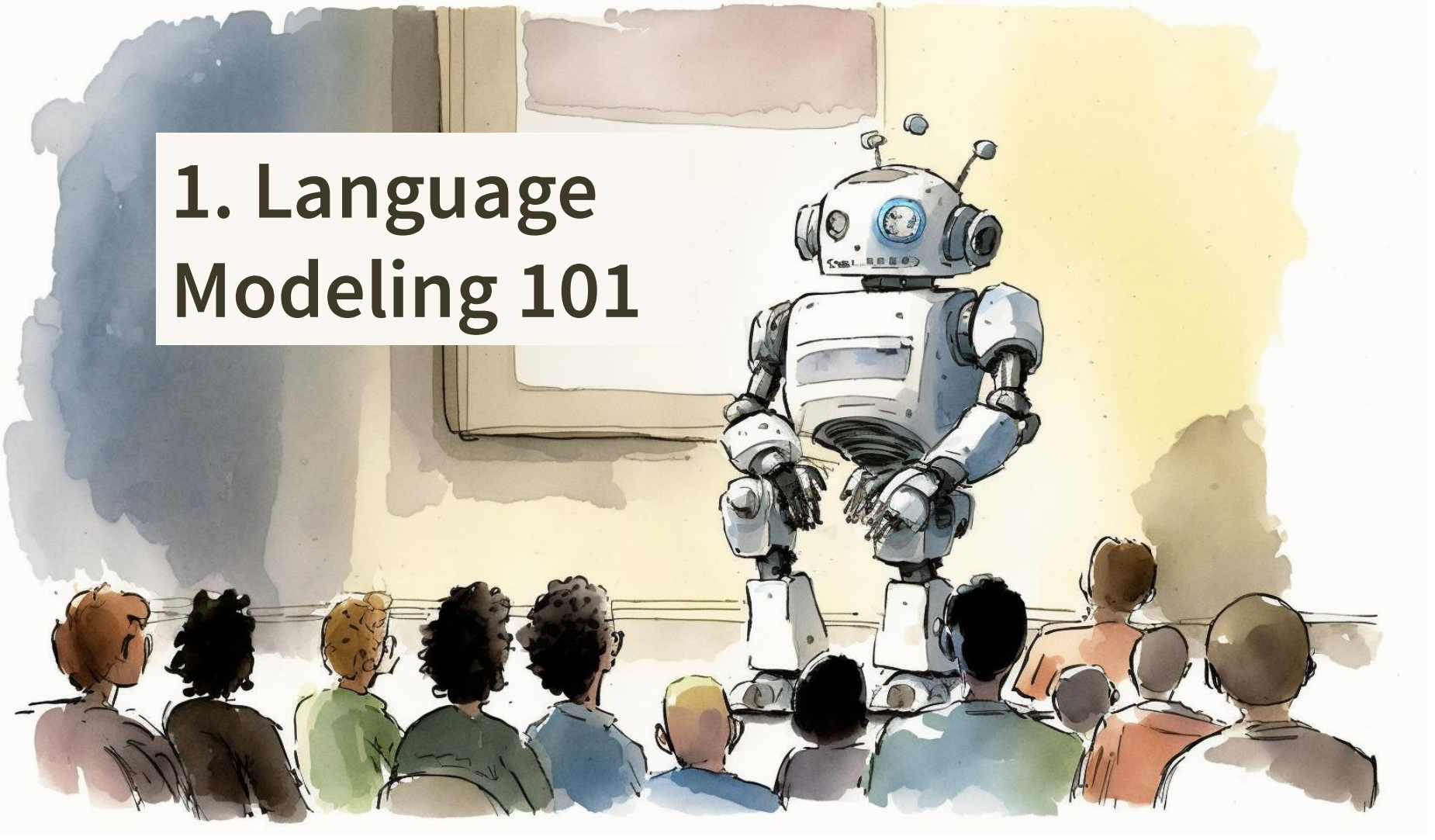


…and their many issues

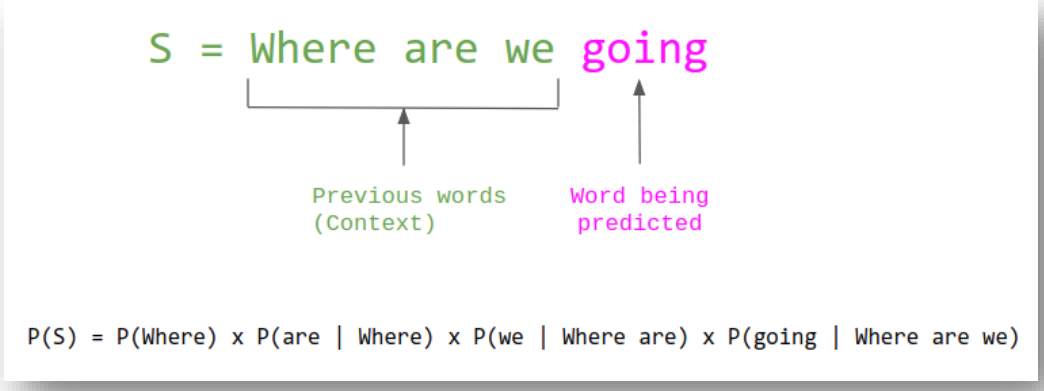
Talk Outline

- Large Language Modeling 101
- The Shaky Foundations of Medical Foundation Models
- Our Research: EMR Foundation Models
  - MOTOR – Time-to-Event Modeling
  - LUMIA – Language + Structured Data
- The Road Ahead: Challenges & Opportunities

1. Language Modeling 101



Language Modeling 101: Training Objective



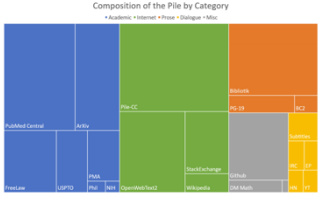
Training Data

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

Language Modeling

Language Modeling 101: The Ingredients

Pretraining Data

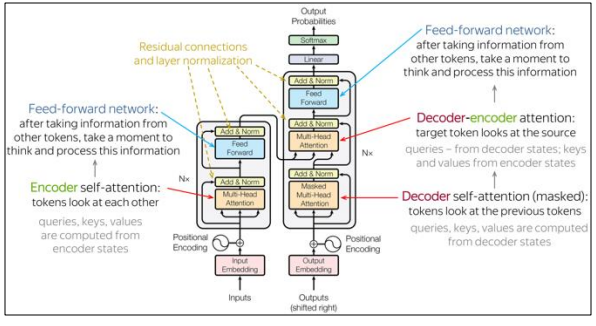


The Pile

Common Crawl

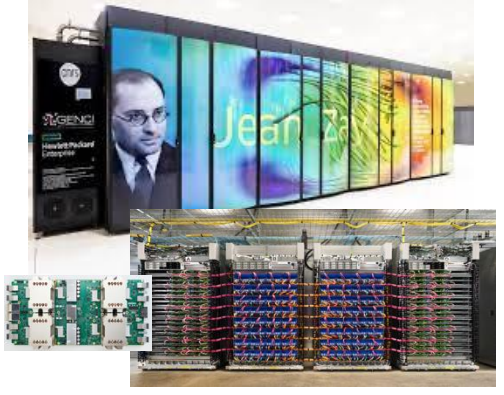
BigScience
Multilingual ROOTS Corpus

Model Architecture



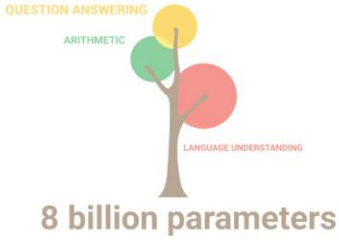
Transformer (Vaswani et al. 2017)

Massive Compute

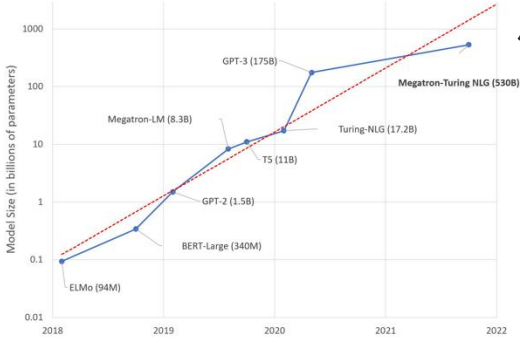


1 Million GPU Hours (~114 years)
to train BLOOM (176B params)

Language Modeling 101: Billions of Parameters



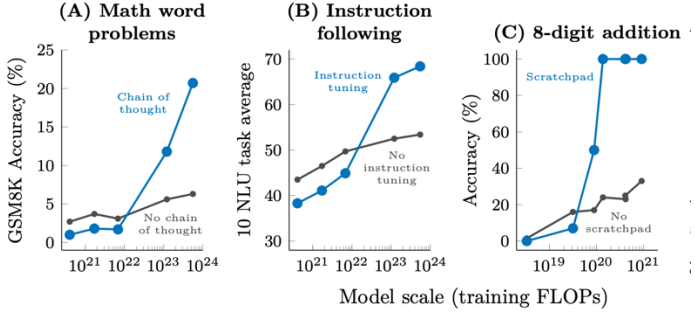
Google's PaLM



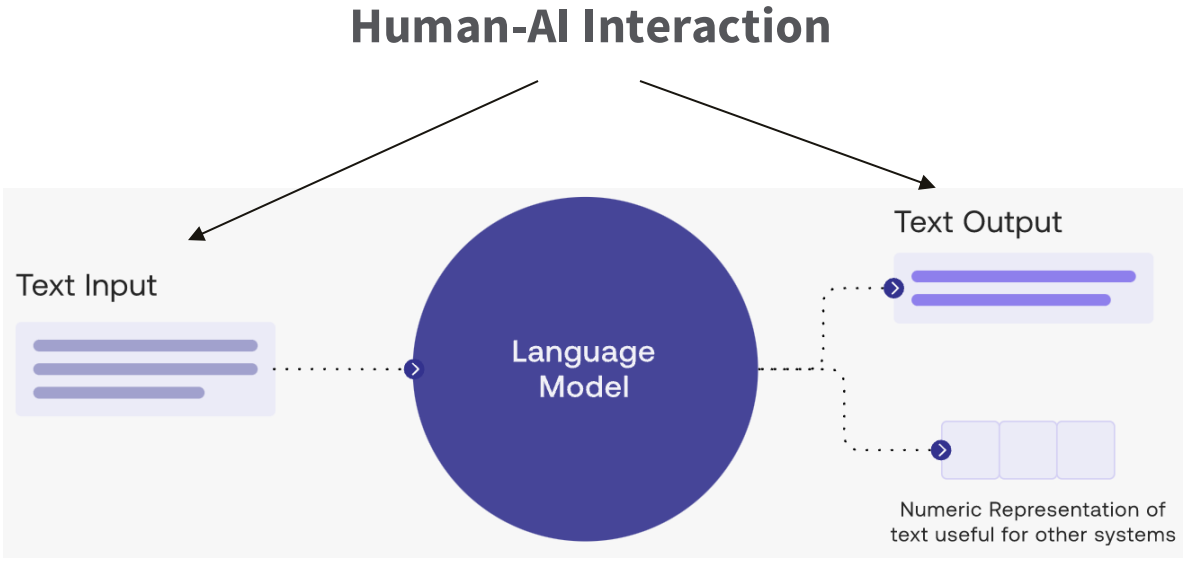
530B
2022

94M
2018

"An ability is emergent if it is not present in smaller models but is present in larger models"
Wei et al. 2022

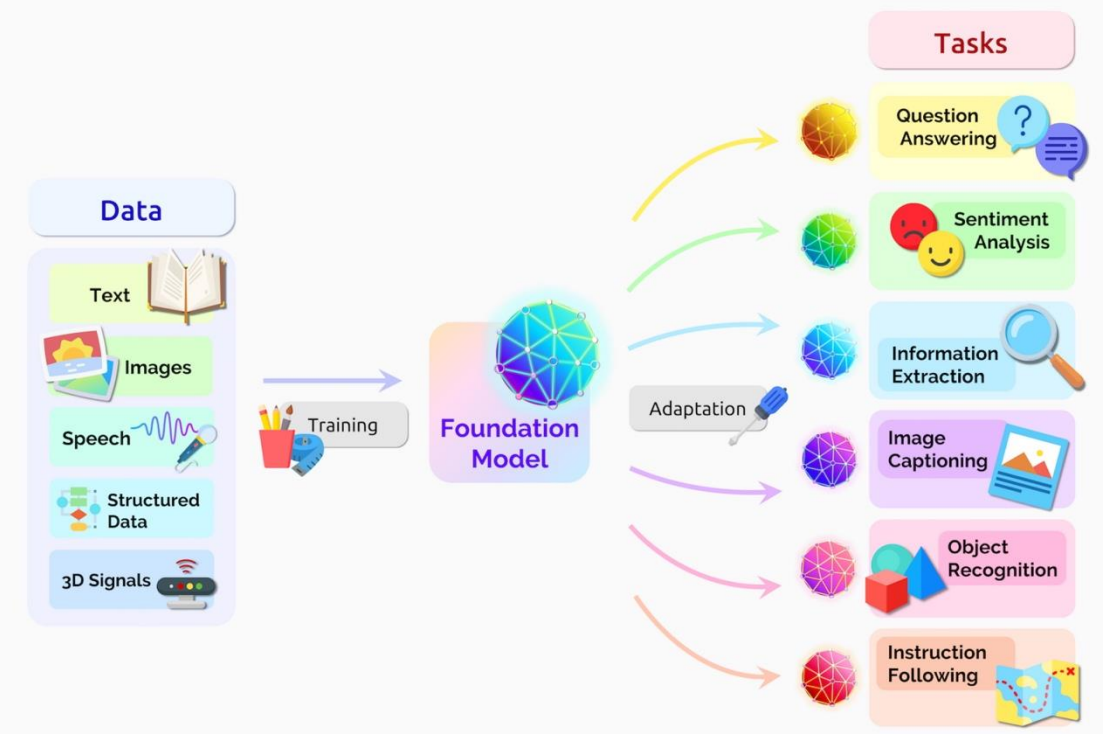


Language Models 101: Inputs and Outputs



Embeddings for Downstream Applications

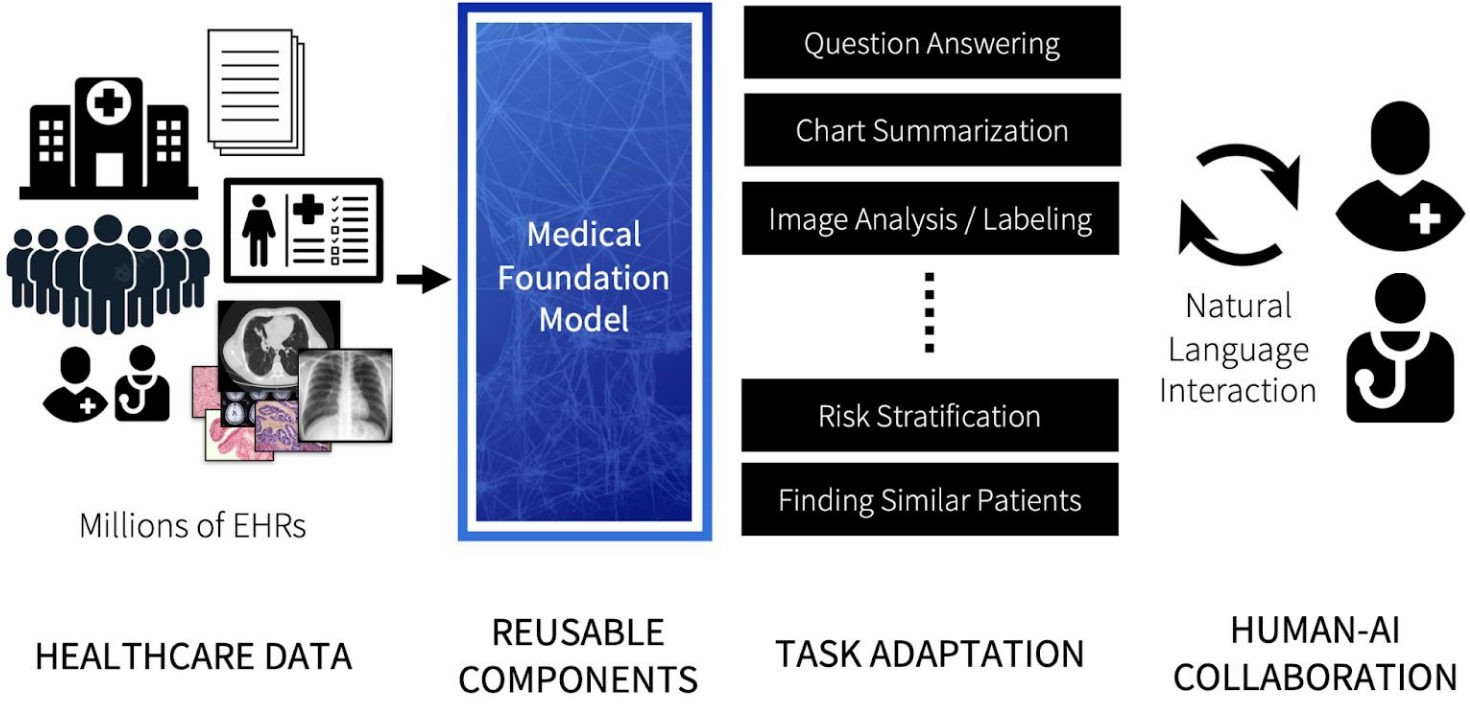
Foundation Models



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

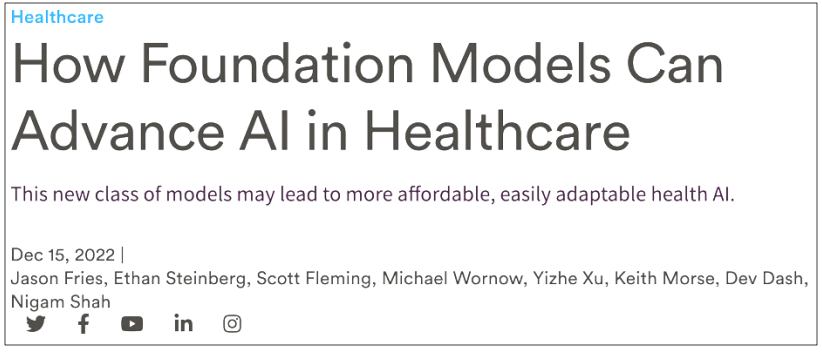
- Language models are an instance of a **foundation model**
- Adaptable to many tasks
- “Language” is any symbol vocabulary

Foundation Models and AI's "Industrial Age"

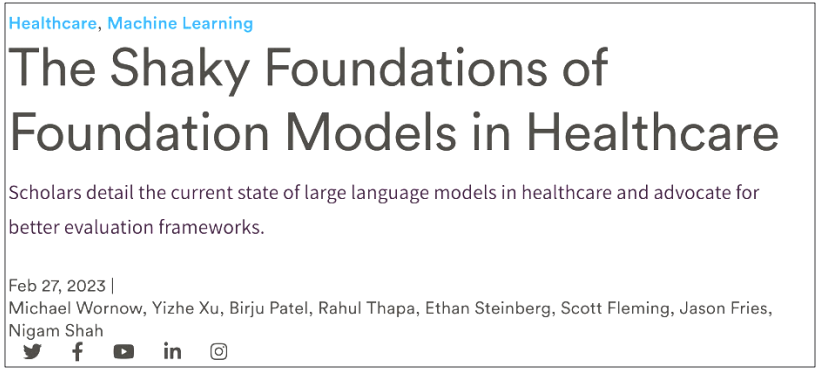


2. The Shaky Foundations of Medical Foundation Models

Thought Leadership on Medical Foundation Models



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



https://tinyurl.com/shaky-foundations

Better Accuracy

Less Labeled Data

Simplified Deployment

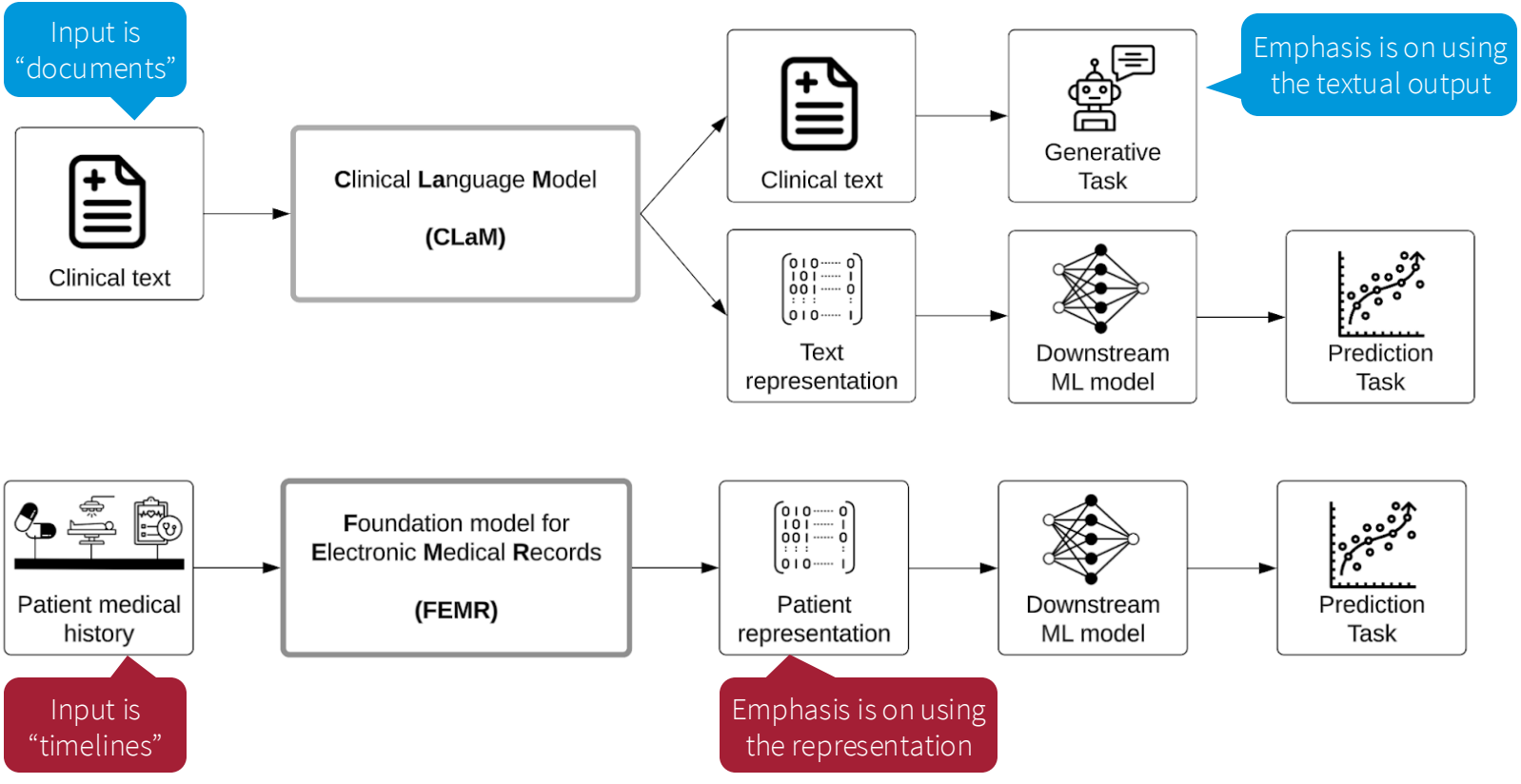
Emergent Applications

Multimodality

Novel Human-AI Interfaces

Enriching the Axes of Evaluation

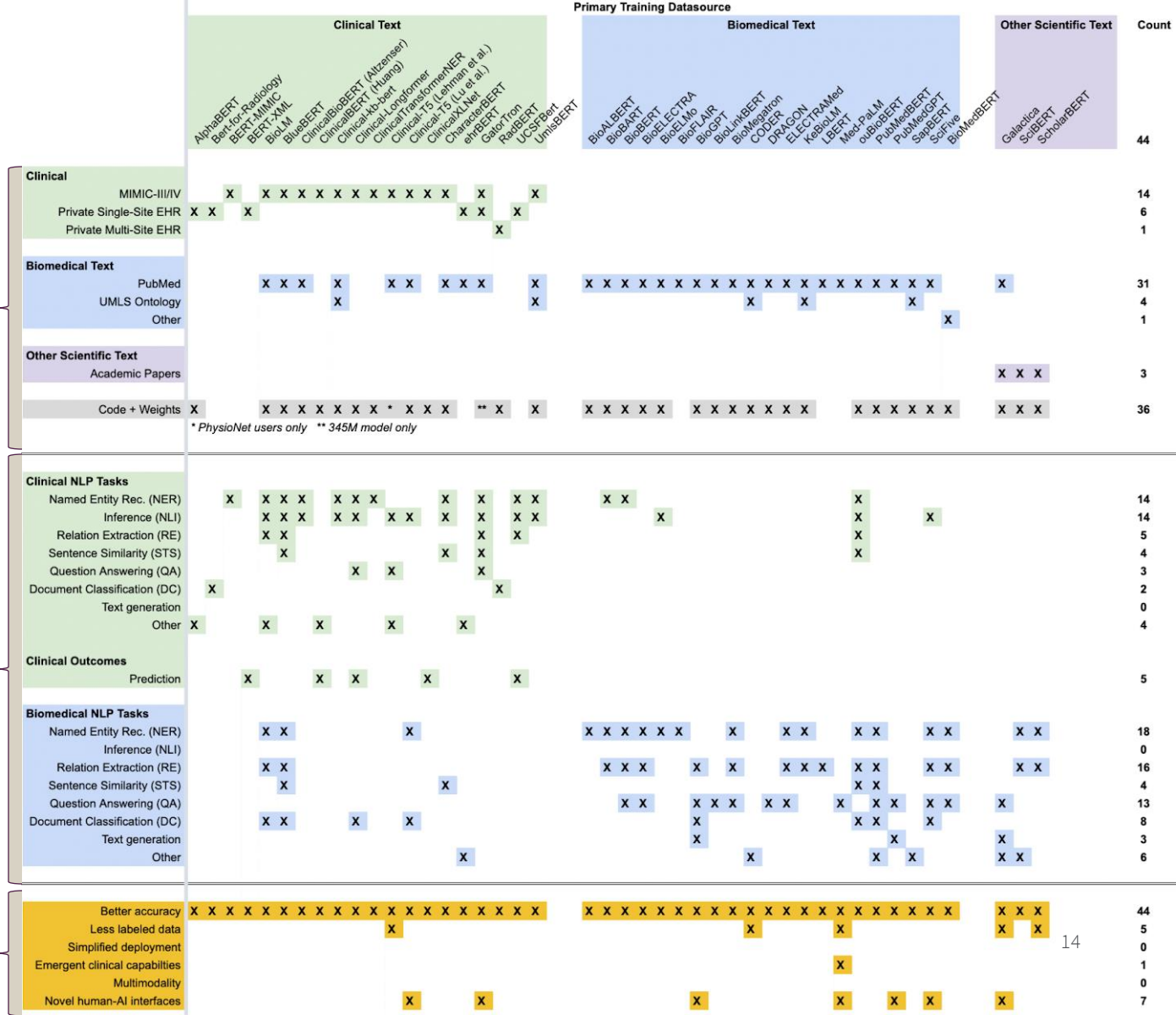
Two World-views on How to Use Such Models



CLaMs

There are 51 of these published!

- Inputs
- Evaluation
- Assumed benefits



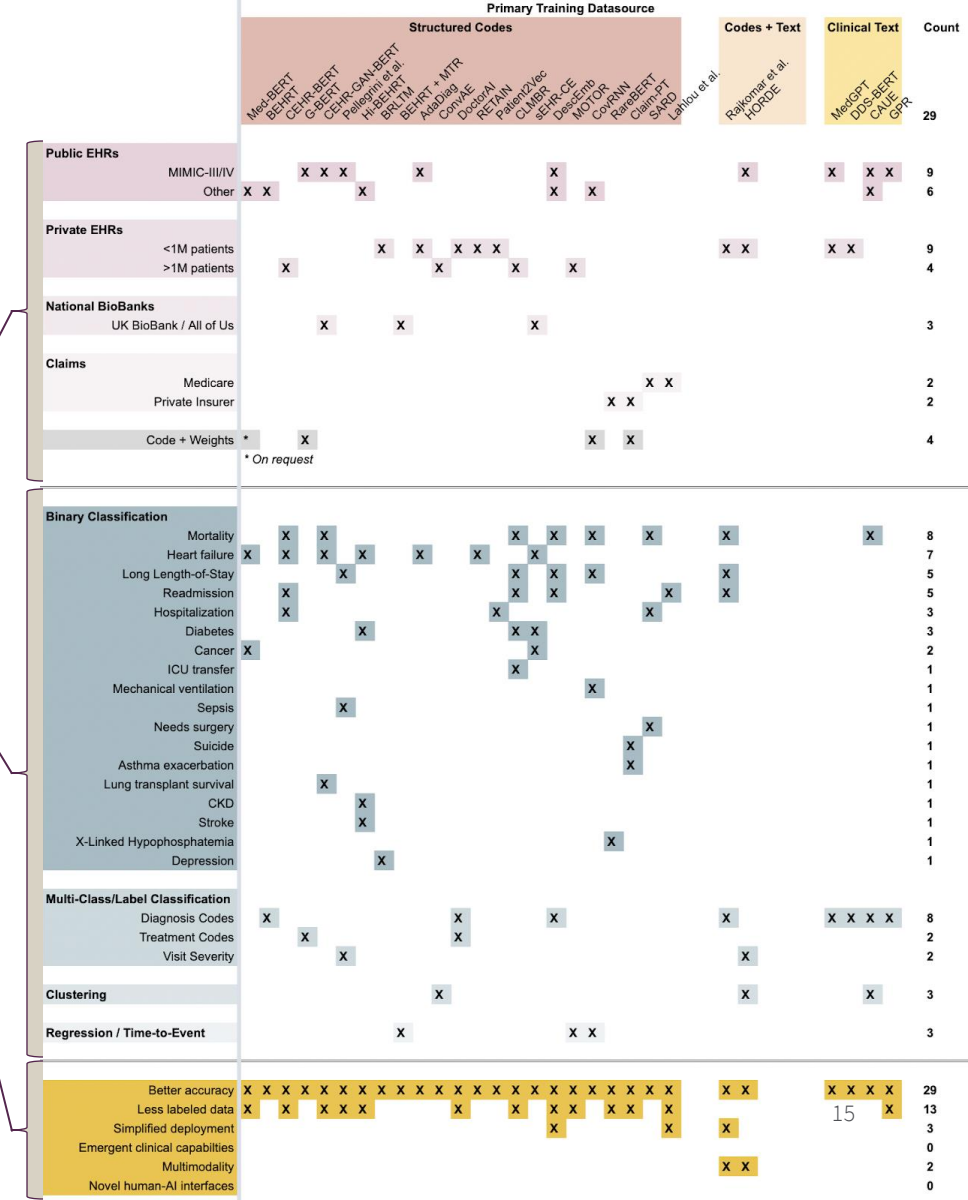
FEMRs

There are 29 of these published!

Inputs

Evaluation

Assumed benefits



Two Worlds of Evaluation…

Clinical Language Models (CLaMs)

51 Models
36/41 Public Weights

5/51 Clinical Outcomes
46/51 Core NLP Tasks

Structured EHR Foundation Models (FEMRs)

29 Models
3/29 Public Weights

63/63 Clinical Outcomes
N/A Core NLP Tasks

- **Hard to compare across models – no ”holistic” view**
- **Unclear utility / usefulness in a clinical setting for most tasks**

Evaluation Gaps in Medical Foundation Models

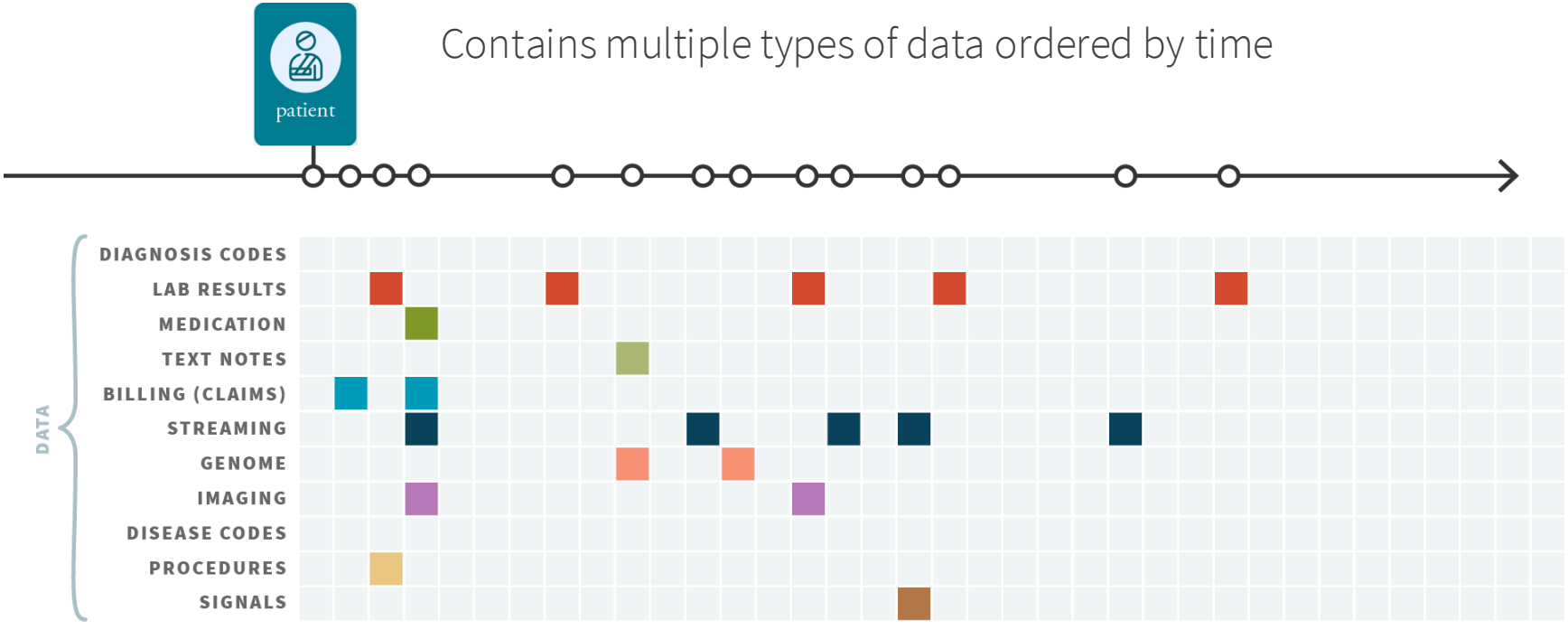
NLP
Benchmarks
Novel NLP Tasks

Improve Clinical Outcomes
Reduce Costs
Improve Patient Lives

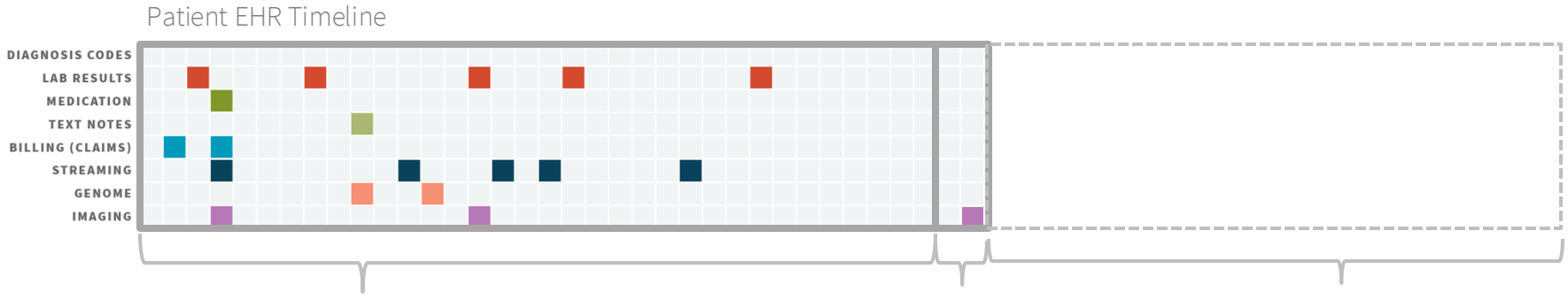
EXPERT MEDICAL KNOWLEDGE
REAL PATIENT TRAINING DATA
TECHNICAL INNOVATION

3. Our Research: EMR Foundation Models

What is an Electronic Medical Record (EMR)



AI for Healthcare Using EMR Data



What Occurred in the Past?

What is Occurring Now?

Predict Future Risks & Intervention Benefits

Example AI Applications

- Chart Summarization
- Cohort Construction

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

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

Whether to Treat

How to Treat

subject to

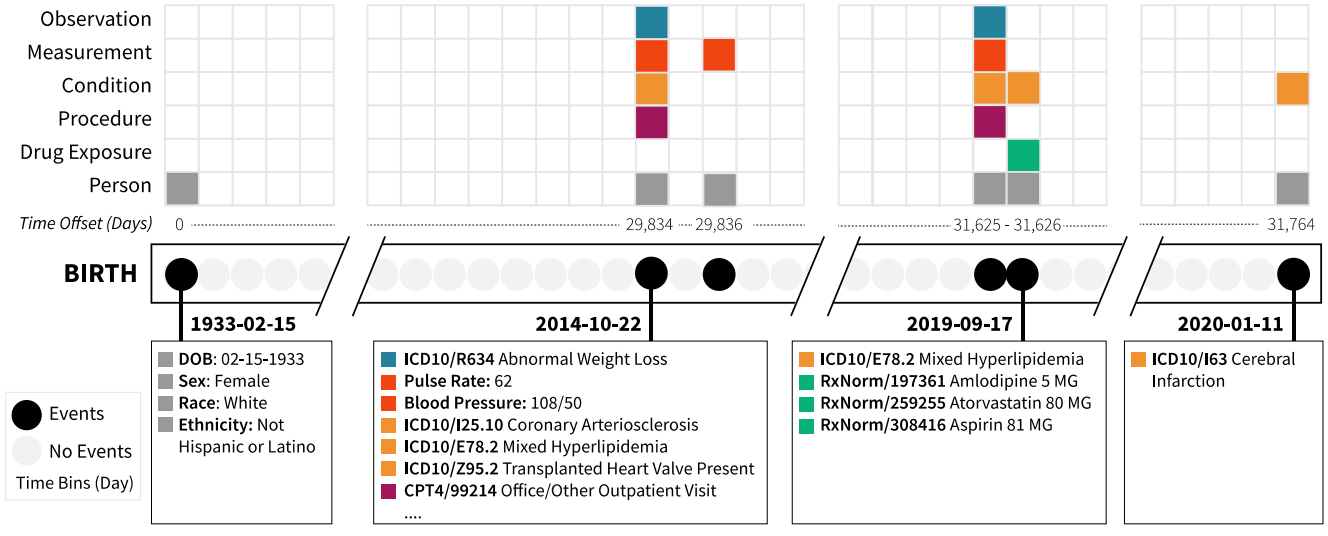
Policy

Capacity to Act

Intervention Properties

Key Insight: View Structured Data as a Language

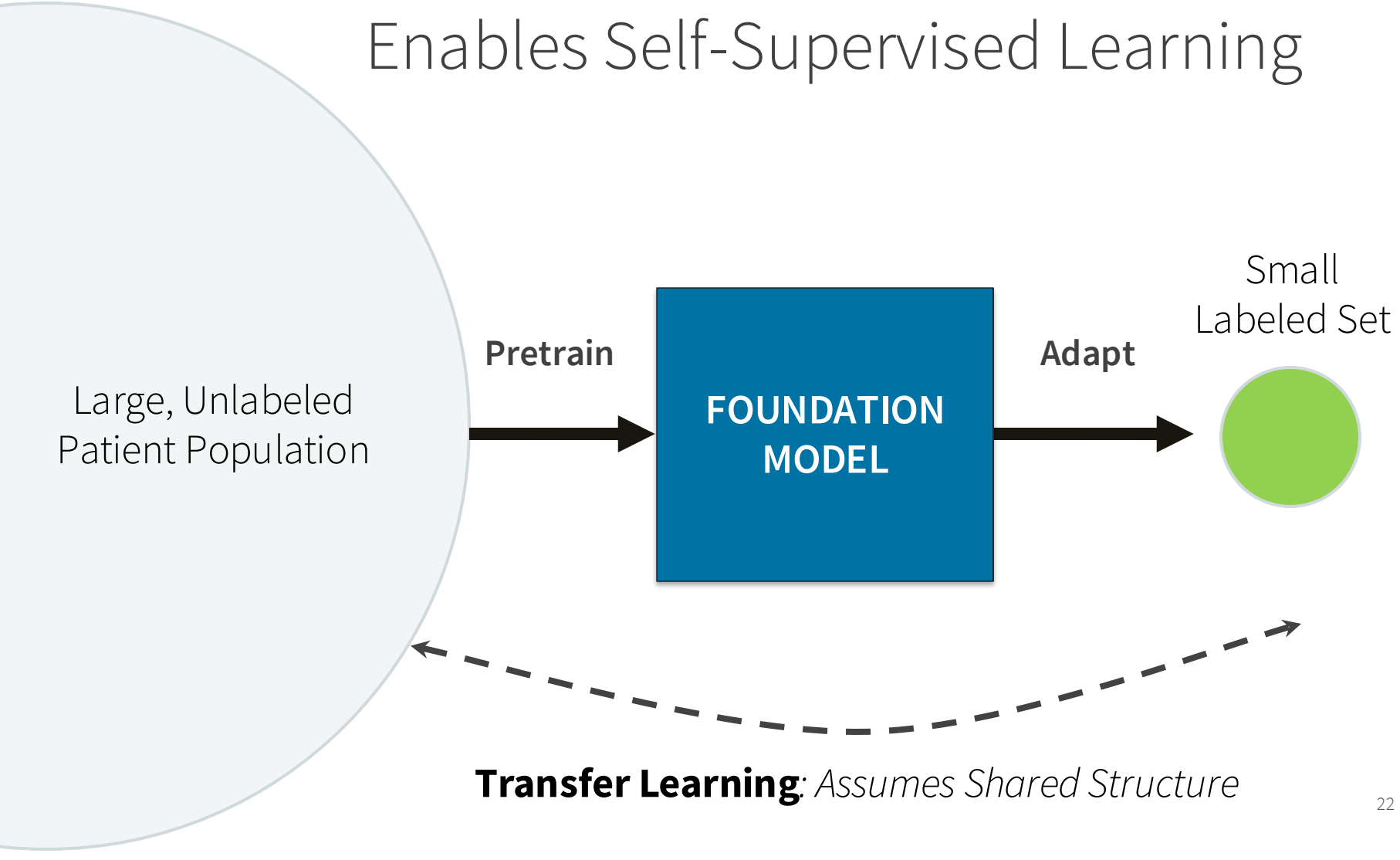
STARR-OMOP



Natural Language: The | quick | brown | fox | jumps | over | the | lazy | …

EMR Language: Visit{RO1.1, 93306} | Visit{aspirin} | Visit{E11.9, R69} | …

Enables Self-Supervised Learning

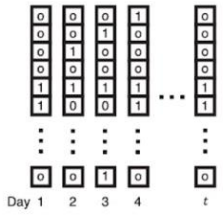


Transfer Learning: *Assumes Shared Structure*

CLMBR: Autoregressive Structured EHR Model

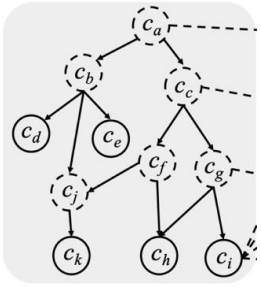
Key Intuitions: Treat codes as words in a symbol vocabulary and use knowledge graphs to better model rare codes

Patient Timeline



Days are represented as a set of <= M codes

Knowledge Graph



Represent codes as paths to the root of a medical ontology [1]

Decoder-only Model

RNN or Transformer Backbone

Autoregressive Objective

$$d = \{c_i\}_{i=1}^{m}$$

Days are a set of unordered codes

$$X = (d_1, ..., d_t)$$

Patients are ordered sequences of days

$$P(d_i|d_{<i}) = \prod_{c_j \in d_i} P(c_j|d_{<i}) \prod_{c_j \notin d_i} (1 - P(c_j|d_{<i}))$$

Model days by assuming codes are independent
Use hierarchical info to improve speed & estimation

P(Patient) $$P(X) = \prod_{i}^{t} P(d_i|d_{<i})$$

[1] Choi et al. “GRAM: graph-based attention model for healthcare representation learning.” KDD 2017.

Our EMR Foundation Model Work

Structured Data | Knowledge Graphs | Text

CLMBR: Clinical language modeling-based representations	**2021**	• **+3.5 to 19%** increase in binary task AUROC • Classifiers **decay less as time passes** • Classifiers **transfer better across subgroups**
MOTOR: Many Outcome Time Oriented Representations	**2023**	• **+3.5 to 19%** increase in TTE task AUROC • **8x faster training** • **95% less training data**
LUMIA: Language Understanding for Medical Insights and Action	**2023**	• A small LLM trained using STARR-OMOP-deid … in progress

MOTOR: Self-Supervised Time-to-Event Modeling with Structured Medical Records

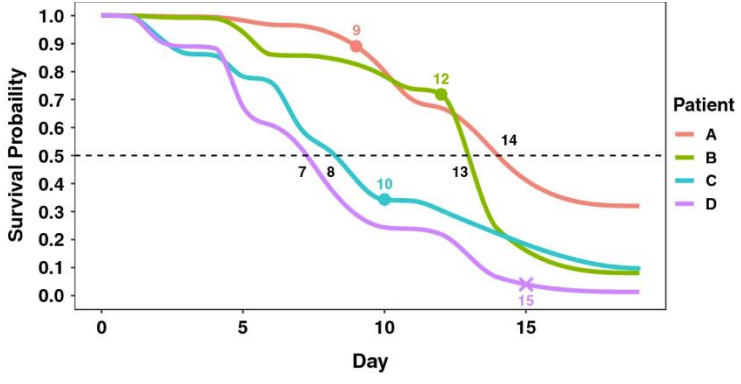
How do we train structured EHR foundation model for **time-to-event tasks**?

Why time-to-event?

- Hard for existing techniques, parameter intensive
- Very important clinical applications

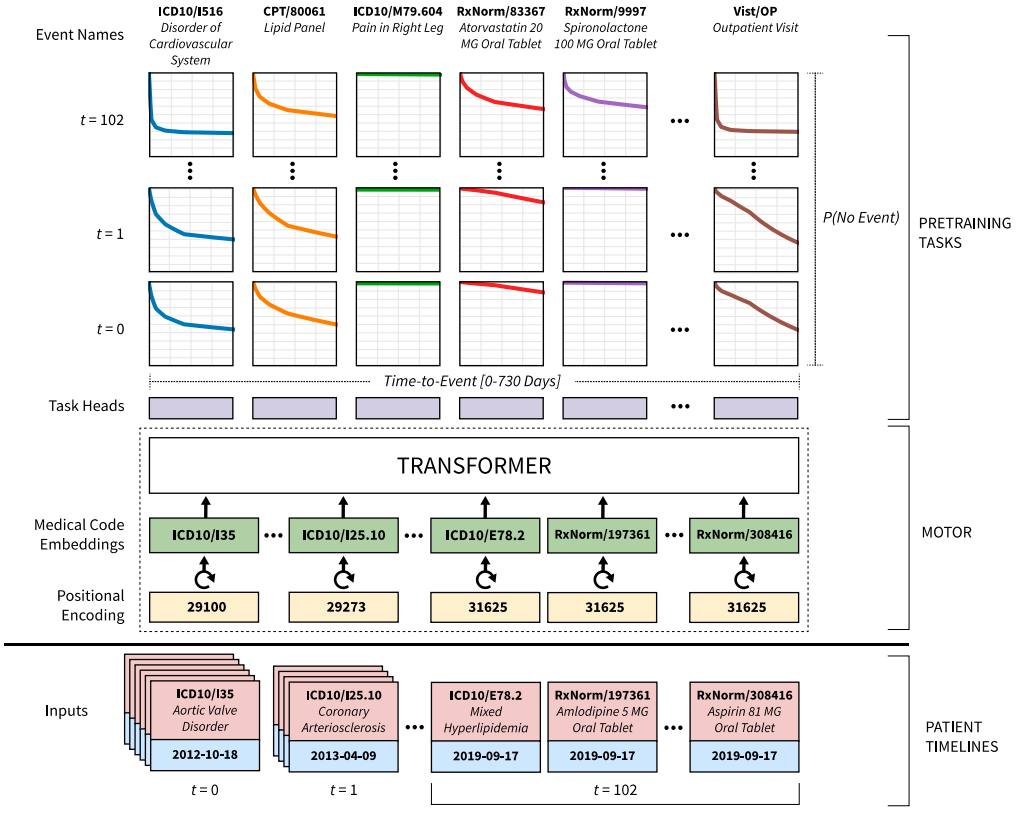
Proposal: MOTOR

- Many Outcome Time Oriented Representations



Model risk across time

MOTOR: Overall Design



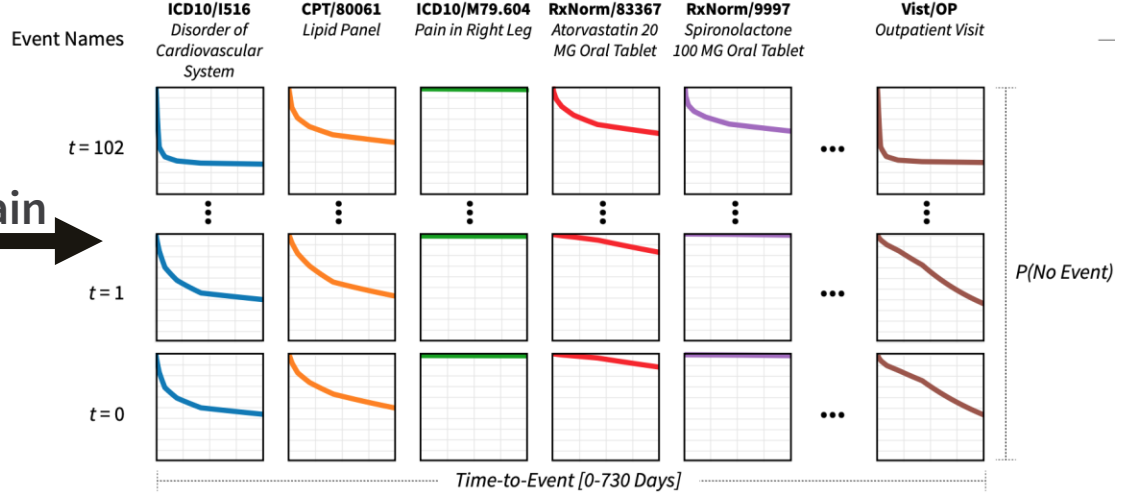
1. **Pretraining task setup**
   P(data | model)
2. **Neural network architecture**
   Transformer
3. **Transferring to target tasks**
   Refit Piecewise Exponential

Pretraining Objective: Time-to-Event

MOTOR predicts the the **time-to-event distribution**

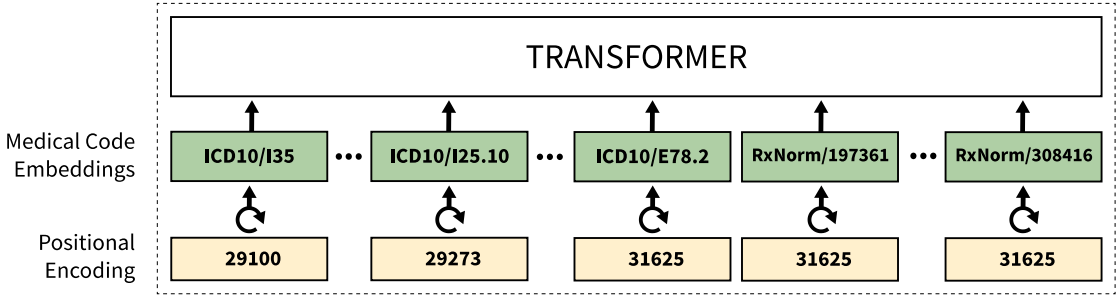
STARR-OMOP
2.7M Patients
2.4B Events
4,192 tasks

Pretrain



MOTOR: Neural Network Architecture

Goal: Convert patient record to representations

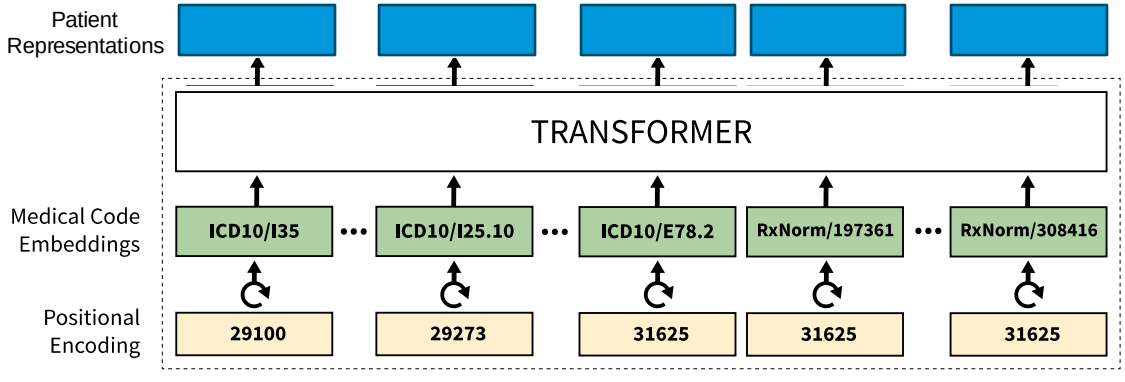


Key Details

1. Conversion of codes to tokens
2. Causally masked transformer
3. Rotational positional embeddings for time

MOTOR: Neural Network Architecture

Goal: Convert patient record to representations

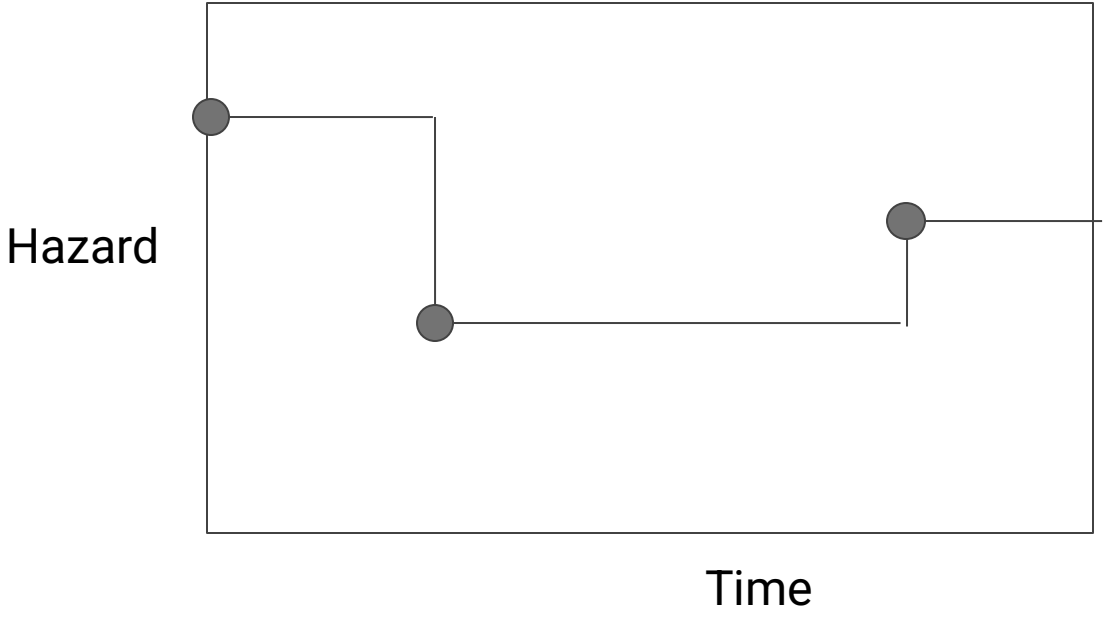


Key Details

1. Conversion of codes to tokens
2. Causally masked transformer
3. Rotational positional embeddings for time
4. Final transformer layer = patient representations

MOTOR: Piecewise Exponential

Goal: Convert patient representation ( ) to time-to-event



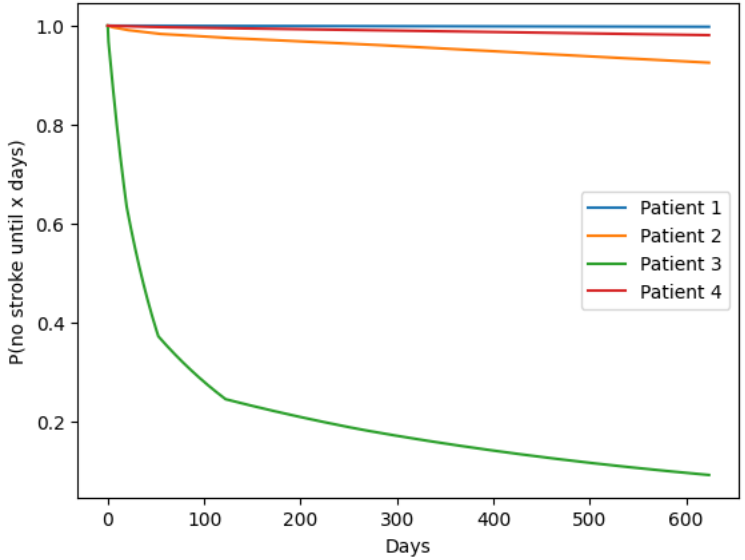
Log linear transformation of features

$$log(Hazard(x)) = x^t b + c$$

MOTOR: Transfer To Target Task

1. Extract patient representations from model ( )
2. Refit piecewise exponential with L2 regularization
3. Second order (conjugate gradient) optimizer for increased speed

Results:

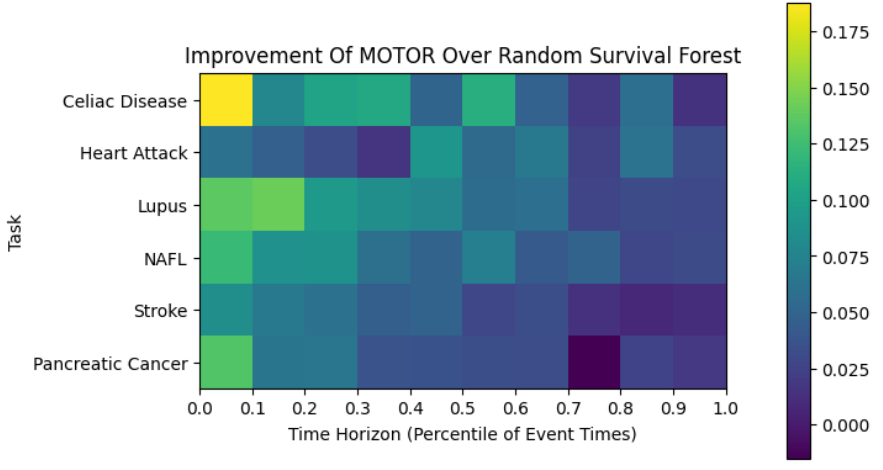


Overall Performance Results

Method	Celiac	Heart Attack	Lupus	NAFLD	Pancretic Cancer	Stroke
Cox PH	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
RSF	0.729	0.836	0.787	0.802	0.824	0.840
MOTOR-WP	0.696	0.795	0.803	0.821	0.777	0.831
MOTOR	**0.802**	**0.884**	**0.850**	**0.859**	**0.865**	**0.874**

MOTOR outperforms state-of-the-art by 6.6% (time-dependent C-statistic)

Time-to-event pretraining
improves performance at all time bins



Greatly Improved Sample Efficiency

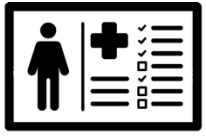
Method	Celiac	Heart Attack	Lupus	NAFLD	Pancretic Cancer	Stroke
Cox PH	0.689	0.761	0.770	0.726	0.793	0.779
DeepSurv	0.704	0.823	0.790	0.800	0.811	0.830
RSF	0.729	0.836	0.787	0.802	0.824	0.840

% Used	Celiac	Heart Attack	Lupus	NAFLD	Pancretic Cancer	Stroke
5%	0.785	0.878	0.840	0.851	0.849	0.868
10%	0.781	0.874	0.844	0.854	0.854	0.868
25%	0.790	0.880	0.845	0.856	0.859	0.869
100%	**0.802**	**0.884**	**0.850**	**0.859**	**0.865**	**0.874**

Match/beat state-of-the-art (random survival forest) using only 5% of data

LUMIA: Language Understanding for Medical Insights and Action

Clinical Notes

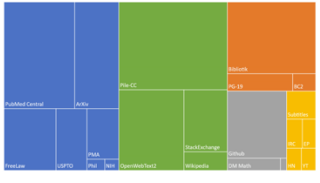


Medical Codes

Flowsheets

- Focus on **patient timelines** not just single documents
- Combines **notes** and **structured data**
- **4096** Max Context Length (8x longer than others)

Training **1.6B parameter model**

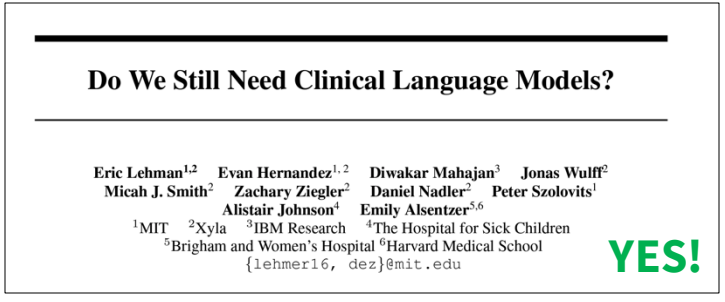


The Pile

STARR-OMOP
160M Notes
78B Tokens

Center for
Research on
Foundation
Models

LUMIA: Benchmarking



Do We Still Need Clinical Language Models?

Eric Lehman[1,2] Evan Hernandez[1,2] Diwakar Mahajan[3] Jonas Wulff[2]
Micah J. Smith[2] Zachary Ziegler[2] Daniel Nadler[2] Peter Szolovits[1]
Alistair Johnson[4] Emily Alsentzer[5,6]
[1]MIT [2]Xyla [3]IBM Research [4]The Hospital for Sick Children
[5]Brigham and Women's Hospital [6]Harvard Medical School
{lehmer16, dez}@mit.edu

YES!

"We show that relatively small specialized clinical models substantially outperform all in-context learning approaches, even when finetuned on limited annotated data."

We're Evaluating ~35 Datasets / Tasks

NLP

- Question Answering
- Natural Language Generation
- Document Classification
- Probe Tasks
- Information Extraction (concepts, relations)

Patient Classification

- Risk Stratification

Clinician-focused Tasks

- Working with clinicians to develop meaning task set

4. The Road Ahead: Challenges & Opportunities

Open Models + Commodity Instruction Tuning

Smaller Models, Cheaper to Train

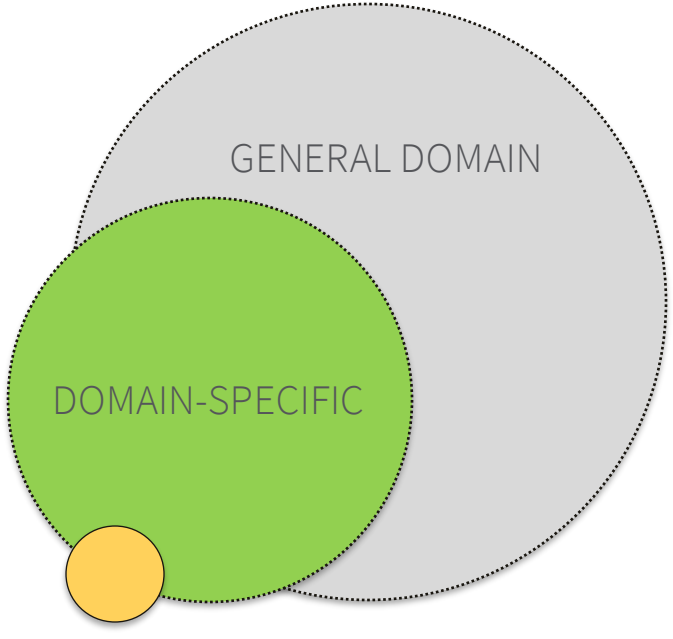
Stanford Alpaca

Replicated original GPT-3 performance for **~$600**

databricks Dolly

Open Model (no restrictions on use)

Aligning for Specialized Domains



INSTRUCTION TUNING

Lack of Transparency Hides Data Problems

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."

Pre-2021 Questions

1075A	The King's Race	implementation, math	800	x7779
1065A	Vasya and Chocolate	implementation, math	800	x14032
1064A	Make a triangle!	brute force, geometry, math	800	x19739
1061A	Coins	greedy, implementation, math	800	x19476
1060A	Phone Numbers	brute force	800	x12567
1056A	Determine Line	implementation	800	x6028
1054A	Elevator or Stairs?	implementation	800	x8520
1047A	Little C Loves 3 I	math	800	x17785
1043A	Elections	implementation, math	800	x10356
1041A	Heist	greedy, implementation, sortings	800	x22026

Solves **10/10** Questions

GPT-4
Training Data
Ends 2021

2022 Questions

1802A	Likes	greedy, implementation	800	x10715
1800A	Is It a Cat?	implementation, strings	800	x19580
1799A	Recent Actions	data structures, greedy, implementation, math	800	x9233
1796A	Typical Interview Problem	brute force, implementation, strings	800	x15004
1795A	Two Towers	brute force, implementation, strings	800	x19658
1794A	Prefix and Suffix Array	strings	800	x12561
1793A	Yet Another Promotion	greedy, math	800	x15111
1792A	GamingForces	greedy, sortings	800	x22860
1791C	Prepend and Append	implementation, two pointers	800	x28936
1791B	Following Directions	geometry, implementation	800	x31133

Solves **0/10** Questions [1]

[1] Horace He, March 14th 2023

Challenges & Opportunities

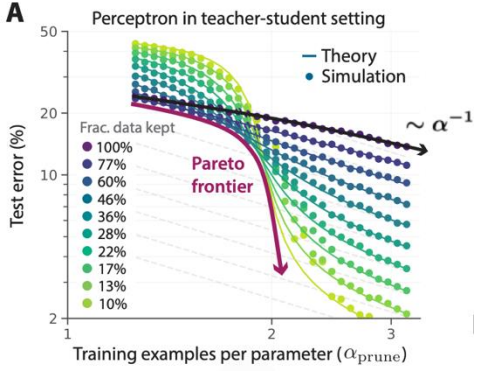
AI will likely not replace jobs, but **augment existing roles** – we need to **measure human + AI performance**

Knowledge Retrieval

Real-world Clinical Tasks

Human-AI Collaboration

Data-Centric AI: Moving Beyond Power Law Scaling



~2% drop in error can cost an **order of magnitude** more data, compute, or energy

- Sorscher et al. 2022

Team Science!


Keith Morse

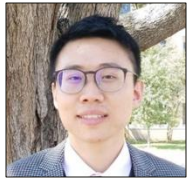

Jason Fries


Ethan Steinberg


Michael Wornow


Rahul Thapa

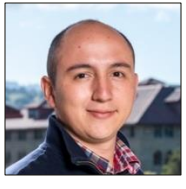

Scott Fleming


Frazier Huo


Cara Van Uden


Mars Huang


Louis Blankemeier


Juan Manuel Zambrano Chaves


Crystal Xu


Akshay Chaudhari

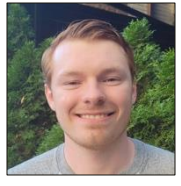

Nigam Shah

FEMR


David Hall


Lawrence Guo


Yifan Mai


Joshua Lemmon


Percy Liang


Lillian Sung

CRFM

SickKids