

Creating Fair, Useful, Reliable and Models in Healthcare

Nigam Shah

Chief Data Scientist, Stanford Healthcare
Professor of Medicine, Stanford University

The AI Chasm in Healthcare

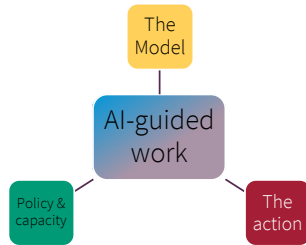


- Hundreds of models are built and published on.
- A very small percentage are deployed for routine use.
- Realizing value from the use of AI to guide care remains elusive.

Data Science Team at SHC

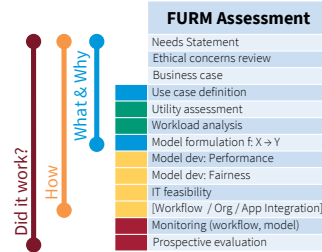
01

Thought leadership for Responsible AI in Healthcare.



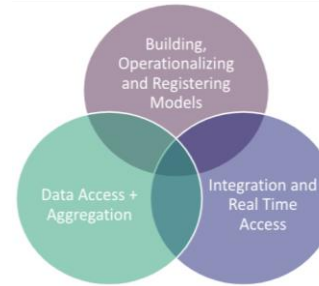
02

Ensure that we create FURM - Fair, Useful, Reliable Models.



03

Processes and infrastructure for an “AI ready” organization.

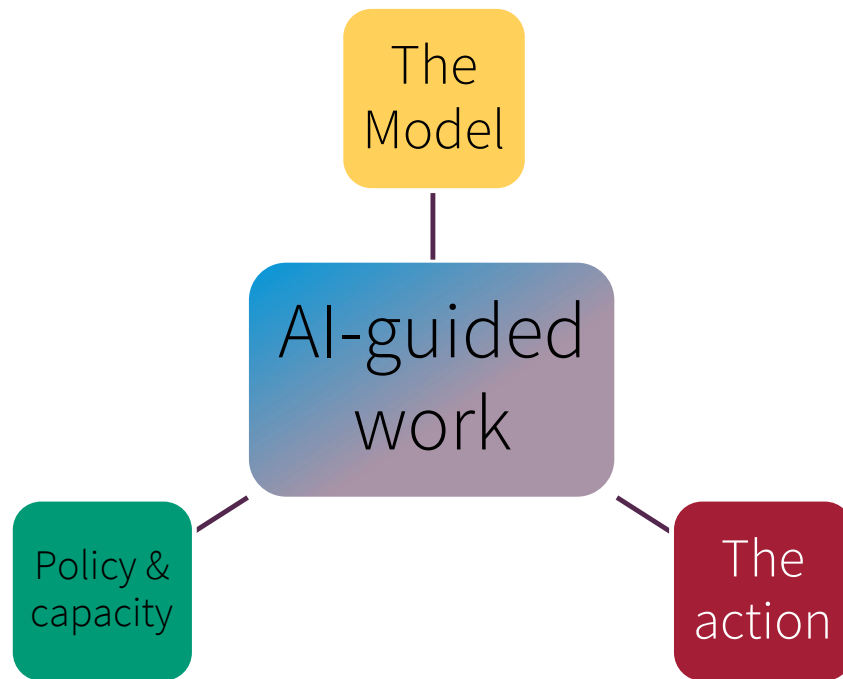


04

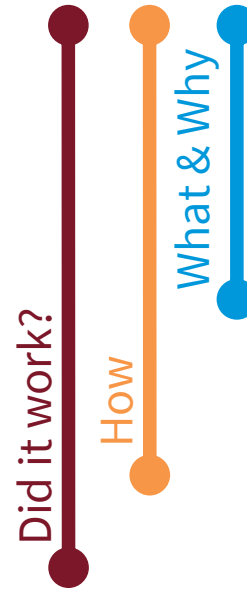
Identify and execute, 3-5 projects with enterprise value.



There is an interplay
among models, capacity,
and actions we take

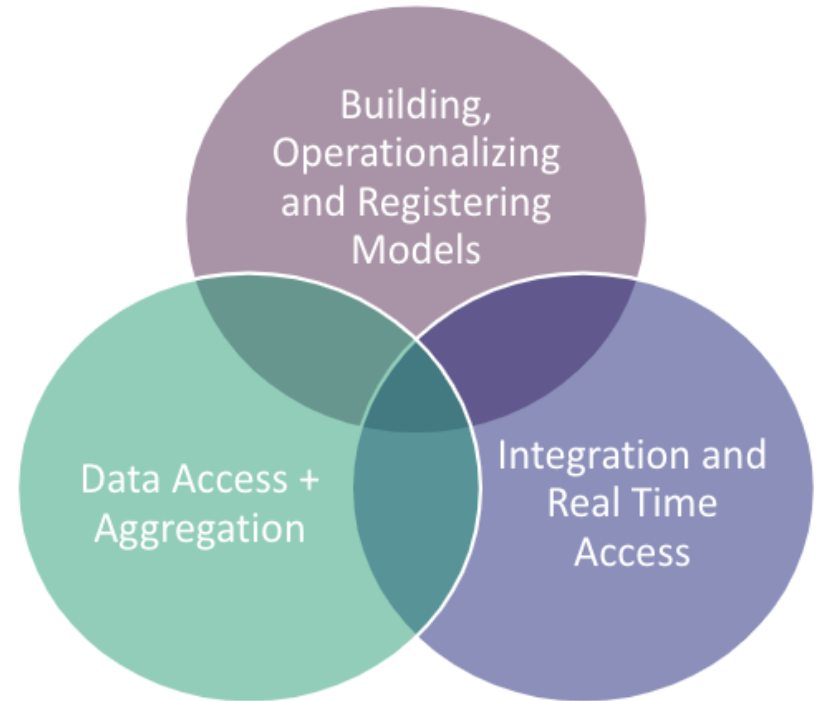


We have developed
a way to assess if we
are creating Fair, Useful,
Reliable Models

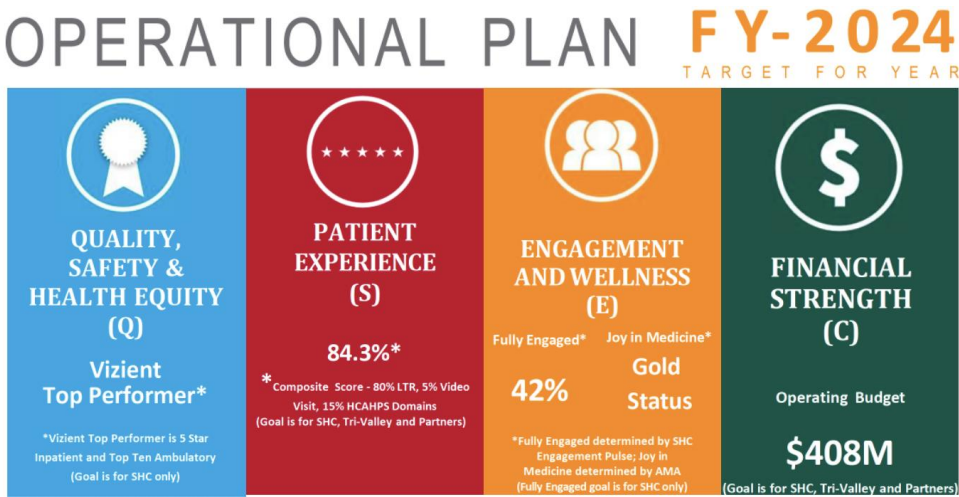


FURM Assessment	
	Needs Statement
	Ethical concerns review
	Business case
	Use case definition
	Utility assessment
	Workload analysis
	Model formulation $f: X \rightarrow Y$
	Model dev: Performance
	Model dev: Fairness
	IT feasibility
	[Workflow / Org / App Integration]
	Monitoring (workflow, model)
	Prospective evaluation

Processes as well as
infrastructure for an
“AI ready” organization

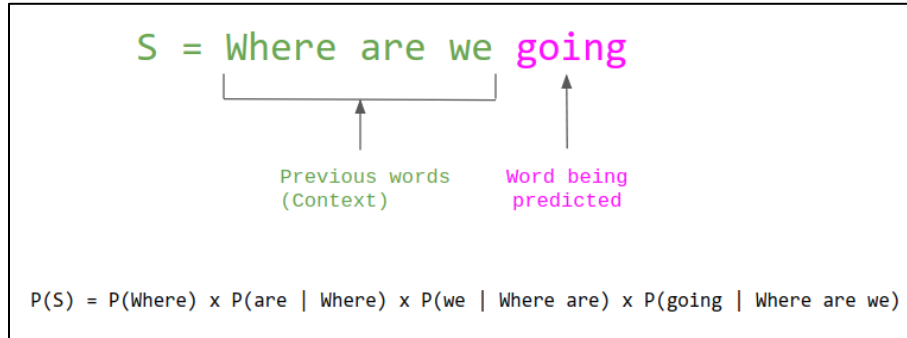


Governance is crucial
for enterprise-wide
alignment



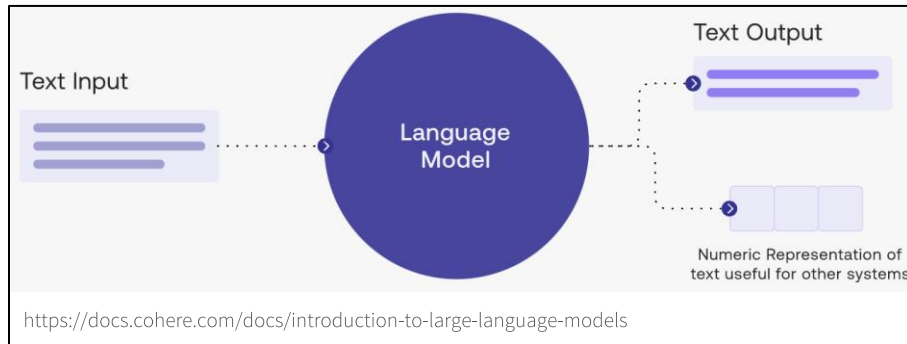
Demystifying language models

Language models 101



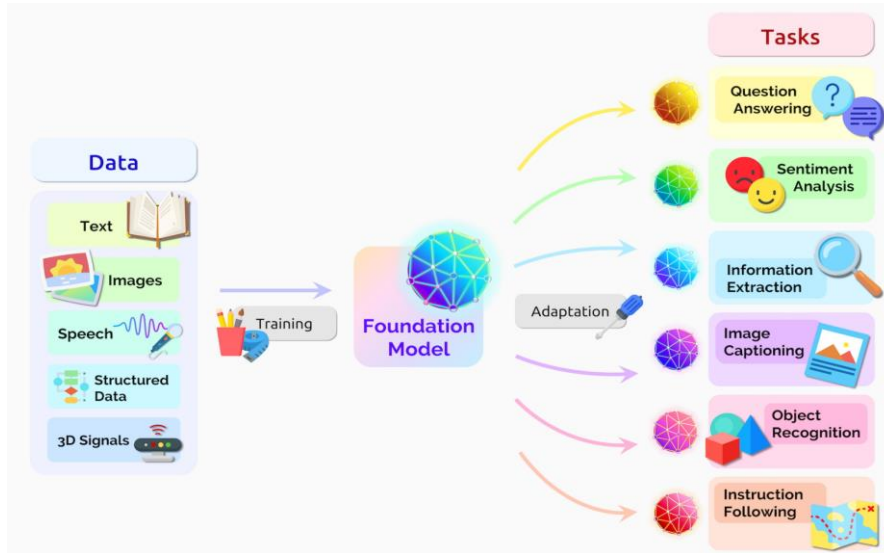
Training data

Language model



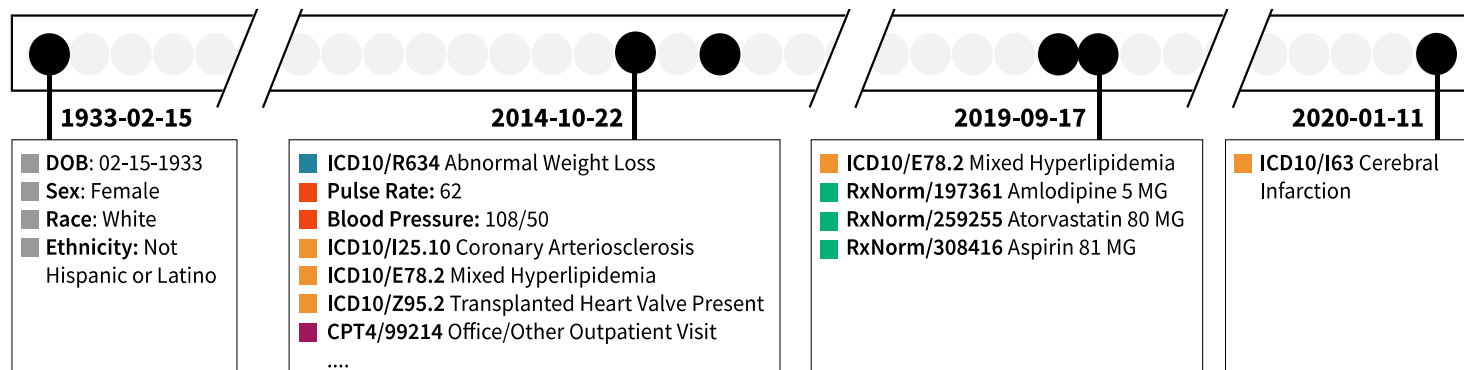
Large language model

Foundation Models – Beyond “language”



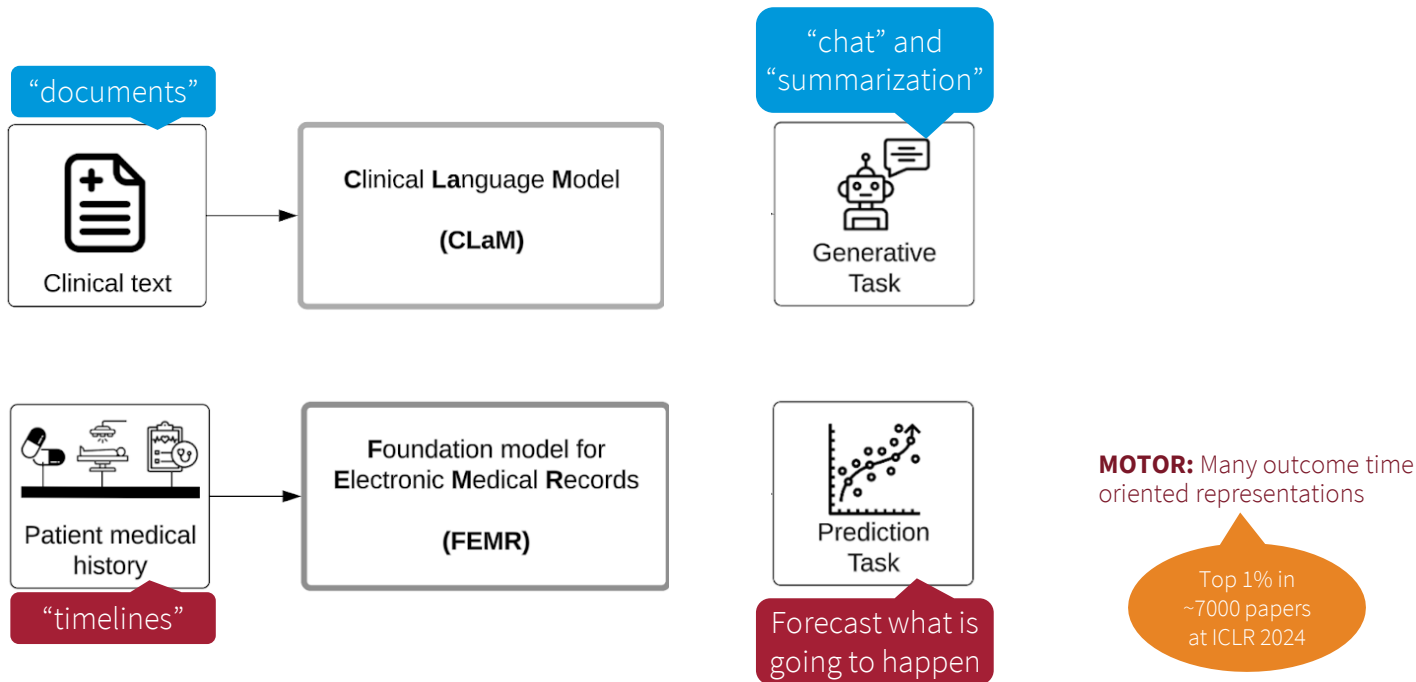
- 200+ assets
 - datasets (e.g. [The Pile](#), [LAION-5B](#))
 - models (e.g. [BLOOM](#))
 - applications (e.g. [CopilotX](#))
- 60+ organizations
- 9 modalities
 - e.g. music, genome sequence, text, video
- **Most models have not seen EHR data**

Structured EHR data comprise a “language”

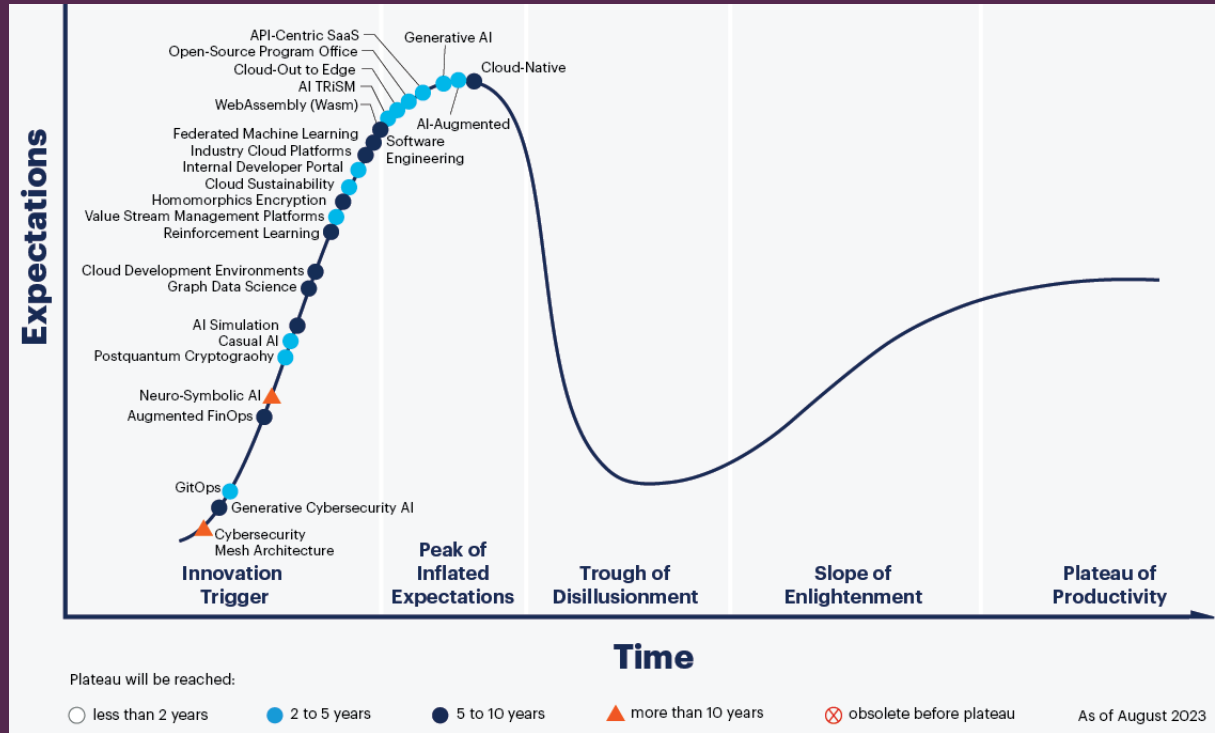


EHR “Language”: Visit{R634, 999214} | Rx {308416} | Visit{I63, R69} | ...

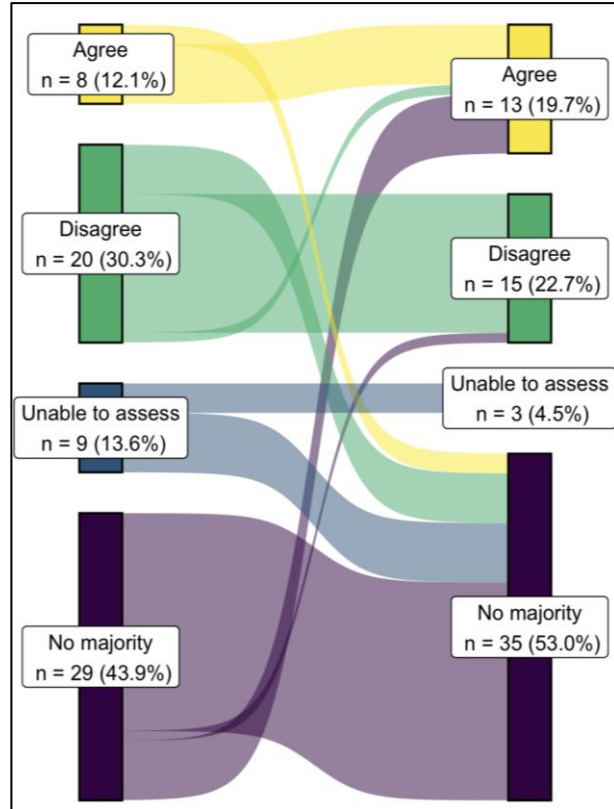
Two ways to build “language” models using the EHR



We are at peak hype



#1: Flashy headlines over-hype memorization



Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery

<https://doi.org/10.48550/arXiv.2304.13714>

#2: Tuning of LLMs for medical tasks is limited

<https://medalign.stanford.edu/>

Summarize from the EHR the strokes that the patient had and their associated neurologic deficits.



+



The patient had strokes in the L basal ganglia in 2018 and multiple strokes in 2022: R occipital, left temporal, L frontal. The patient had right sided weakness associated with the 2018 stroke after which she was admitted to rehab. She then had a left sided hemianopsia related to the 2022 stroke.

- 983 instructions
- 15 clinicians, 7 specialties
- 303 “valid completions”

Model	Context	% Correct
GPT-4 (MR)	32k	65.0%
GPT-4	32k	60.0%
GPT-4	2048*	51.8%

#3: Focus on defining and verifying “benefits”

	Use a general model	Finetune a public model	Build our own model
Create predictive models with less training data			✓
Write a response to a patient message	✓		
Fulfill prompts that physicians want	✓		
Create HPI based on QA with the patient			
Generate practice questions for the USMLE exams			
Answer a medical question that comes up during rounds	✓		
Define the benefits	Using an “internet diet”		Using EHR or claims data