Artificial Intelligence-Applications, Challenges and Opportunities

Girish N Nadkarni

The Windreich Department of AI and Human Health
The Hasso Plattner Institute for Digital Health at Mount Sinai
The Charles Bronfman Institute for Personalized Medicine
Digital and Technology Partners

Girish N Nadkarni, MD, MPH

T.N. Medical College
Masters from Johns Hopkins Medical Institutions
Medicine Residency, Fellowship, Postdoc @ Mount Sinai Health System
Fishberg Professor @ Icahn School of Medicine at Mount Sinai
Chair, Windreich Department of AI and Human Health and Director, Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai
Chief AI Officer, Mount Sinai Health System

- Clinical focus: Intensive care nephrology
- Research focus: AI in healthcare

Disclosures

I have the following financial relationships to disclose:

Consultant. Renalytix, Vertex,, Pensieve health, Heart Test Laboratories,

Grant/Research support from: NIH (NIDDK, NHLBI, OD), Renalytix, Vertex Pharma, ImmuneID, Heart Test Laboratories, Regeneron, Private foundations

Scientific Confounder/Advisory Board: Renalytix, Pensieve Health, Artemis AI, Data2Wisdom, ClAIRity Health

Equity: Renalytix, Verici, Pensieve Health, pulseData, Nexus iConnect, Data2Wisdom LLC, Heart Test Laboratories, ClAIRity Health, Artemis AI

Employee of: Icahn School of Medicine at Mount Sinai and Mount Sinai Health System

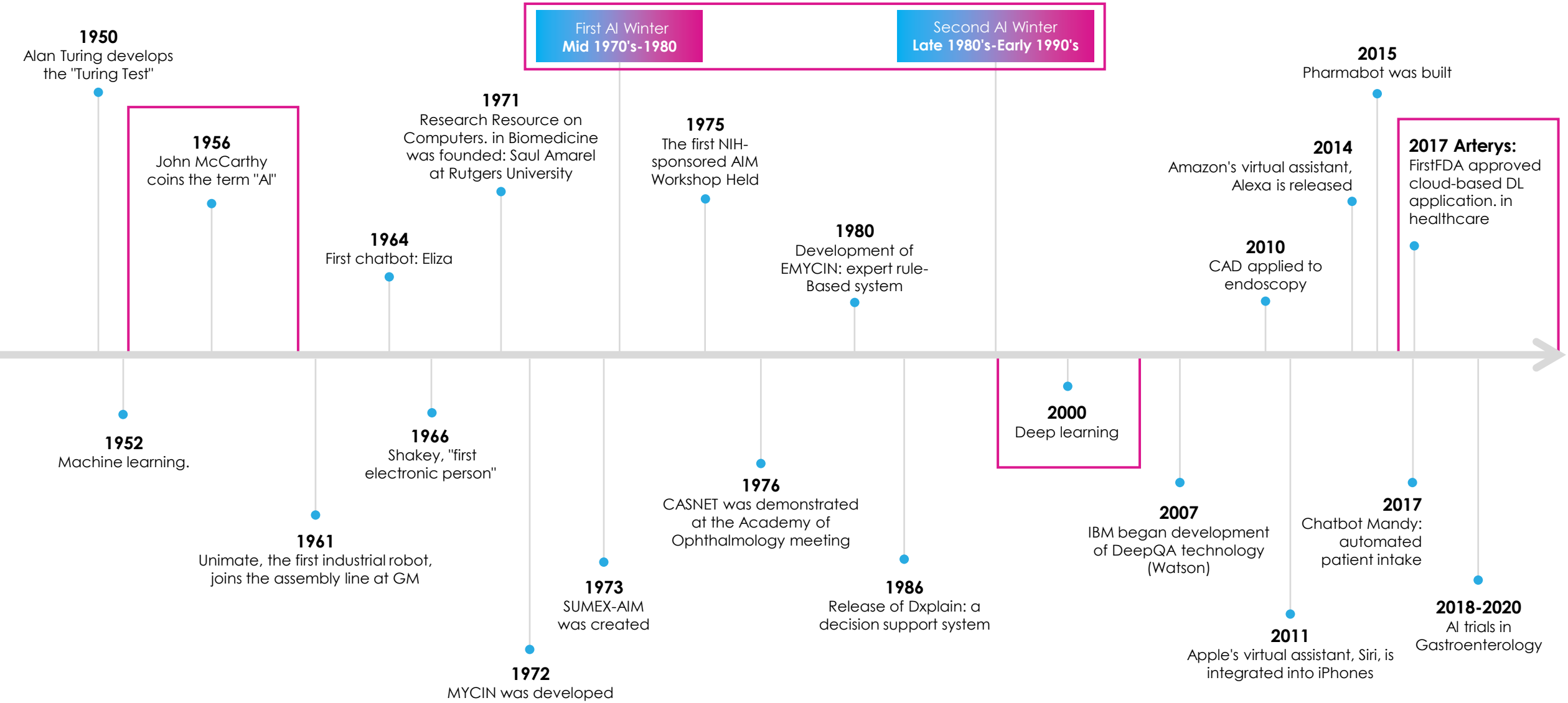
Icahn School of Medicine at Mount Sinai

A Brief History of AI

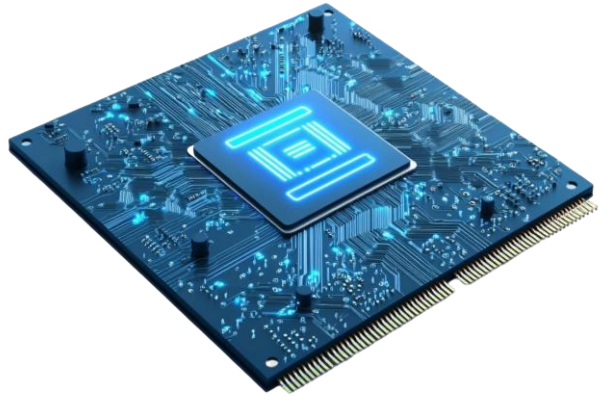
Examples of AI applications

Challenges & Opportunities

AI Has Been Around For a While



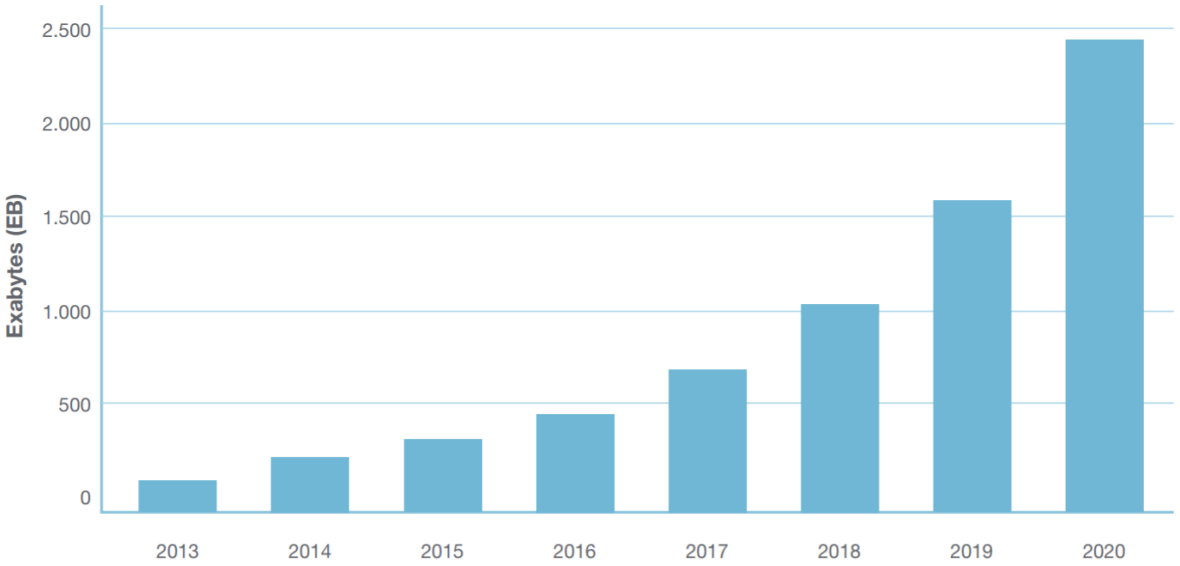
Why are We Talking So Much About AI Now?



Data Generation in Healthcare

Patricia Kovatch

3 million years of IMAX quality video



Icahn School of Medicine at Mount Sinai

Home About

AIR·MS – AI-READY MOUNT SINAI

A platform where researchers can access the vast datasets held at Mount Sinai for data science and machine learning endeavors.

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* †
illia.polosukhin@gmail.com

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* †
illia.polosukhin@gmail.com

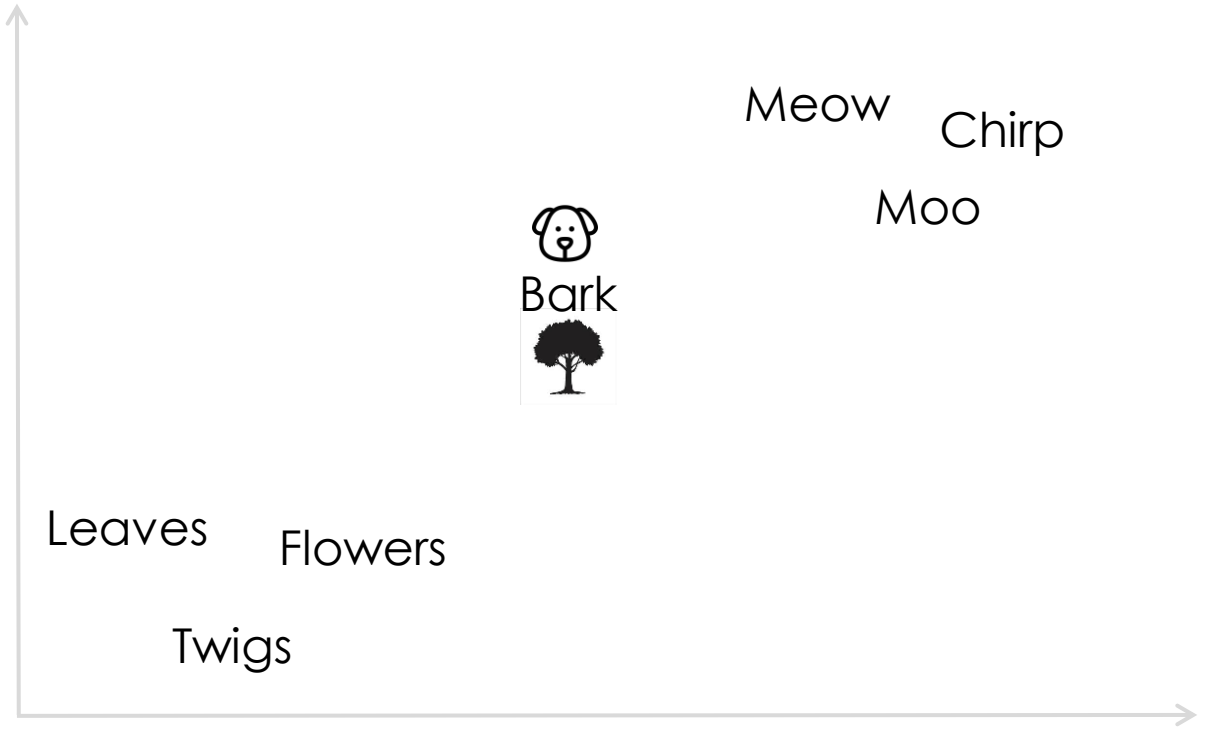
ChatGPT Gemini LLaMA by Meta

What is Attention?

Bark

The tree **bark** is brown.

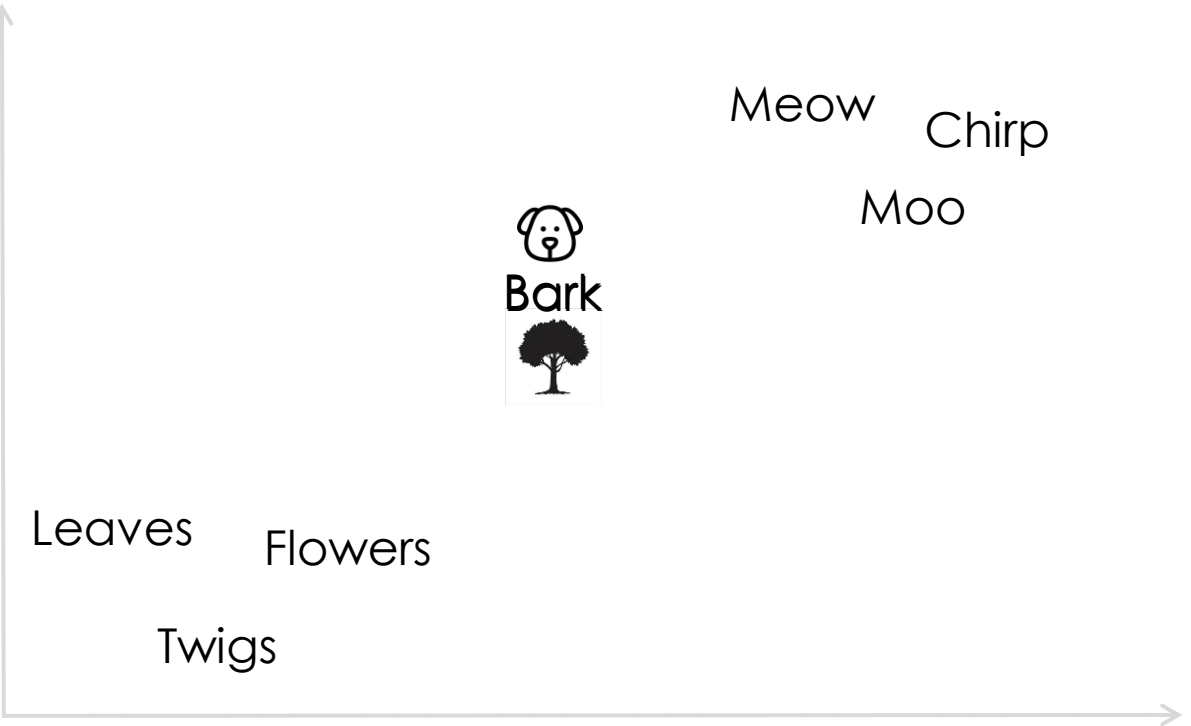
The dog's **bark** is loud.



Attention Extracts Relationships Between Words

The tree **bark** is brown.

The dog's **bark** is loud.



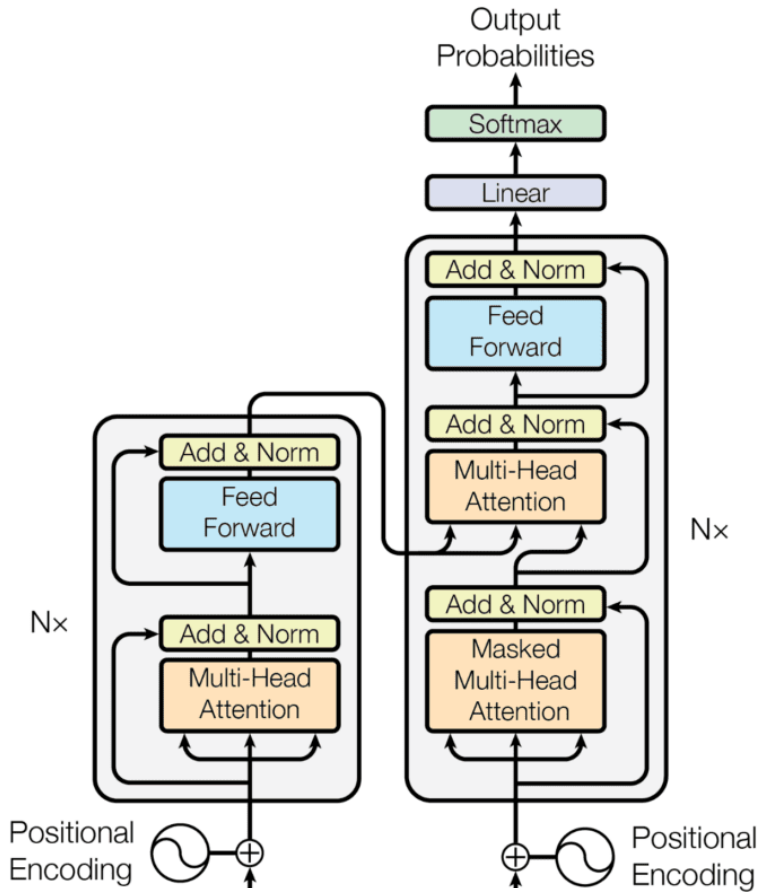
Generative Pre-Trained Transformer (GPT) is a Large Language Model or LLM

Large Language Model (LLM) - Given the statistical distribution of words in the vast public corpus of (English) text, what words are most likely to follow the sequence

The first person to walk on the moon was…Neil Armstrong

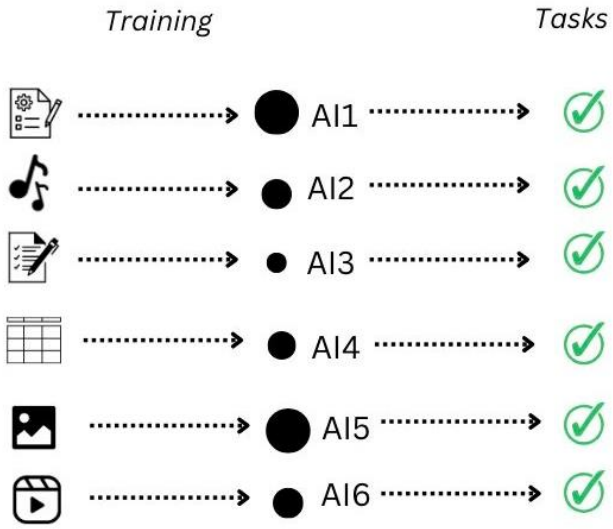
Generate statistically likely sequences of words (tokens).

LLMs do not understand text meaning in a literal sense - just the statistics/mathematical model



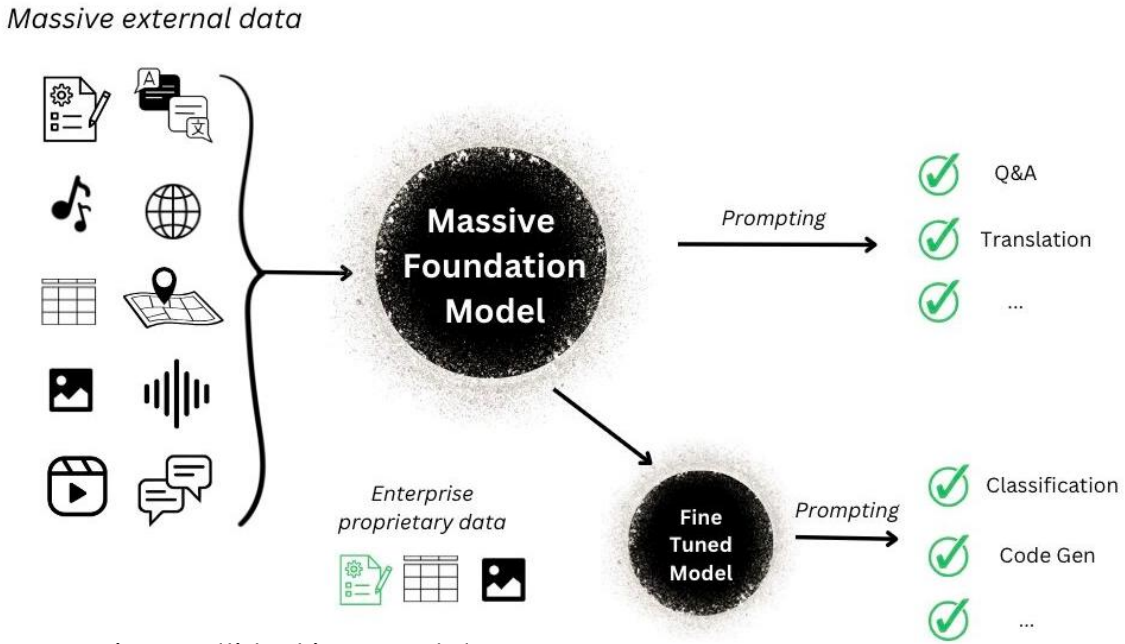
Foundation Models

Traditional ML

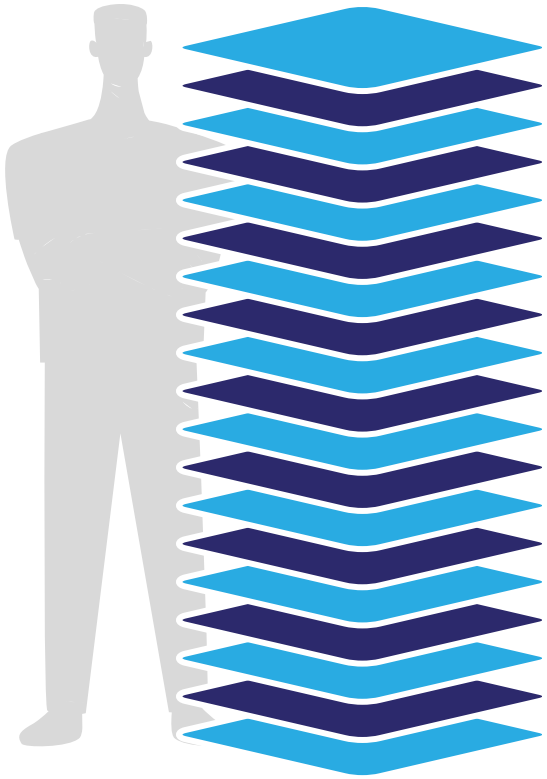


- Individual siloed models
- Require task-specific training
- Lots of human supervised training

Foundation Models

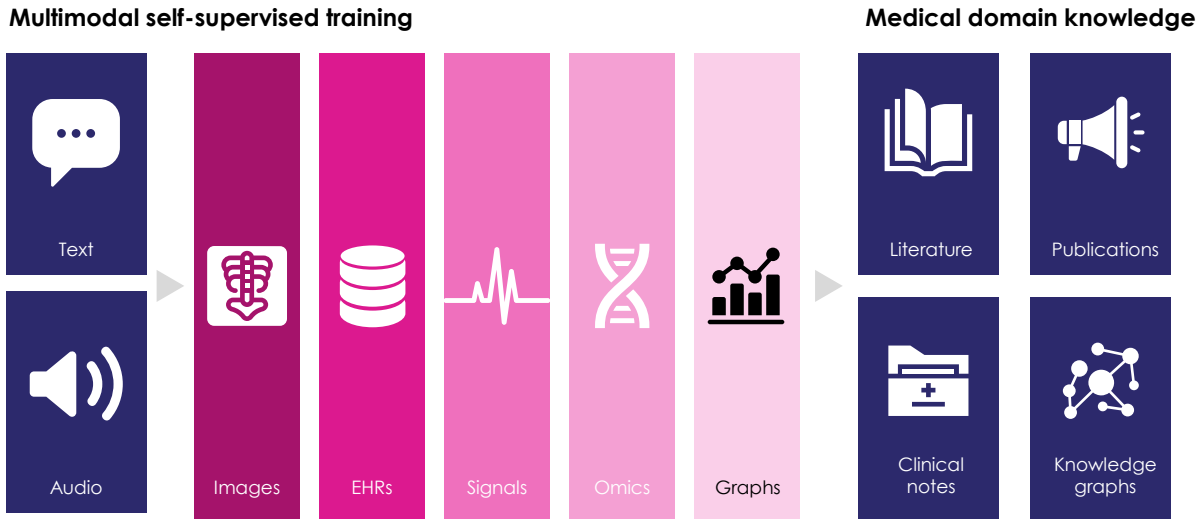


- Massive multi-tasking model
- Adaptable with little or no training
- Pre-trained unsupervised learning



Inputs

Social, behavioral
Genomics and -omic layers
Biosensors
Immune system
Gut microbiome
Anatome
Environmental
Physical activity, sleep, nutrition
Medication, alcohol, drugs
Labs, plasma DNA, RNA
Family history
Communication, speech
Cognition, state of mind
All medical history
World's medical literature, continually updated

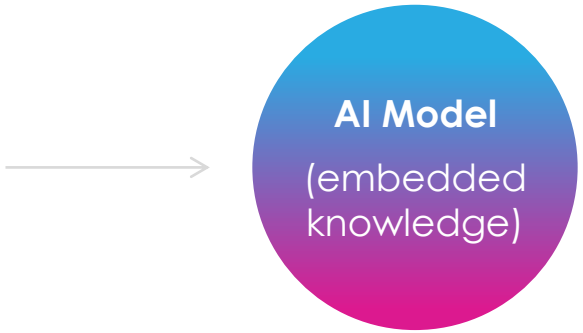


The promise of AI in Health

Automating or augmenting data and information processing tasks

Raw Data

- User instructions
- Tabular data
- Sensors
- Clinical text
- Imaging



Desired Output

- Diagnosis
- Prediction
- Clustering
- Summarization
- Generation
- Decision support

Making data work for patients, physicians and systems

A True Learning Health System

Systematically gather and creative evidence

Apply the most promising evidence to improve care

Icahn School of Medicine at Mount Sinai

A Brief History of AI

Examples of AI applications

Challenges & Opportunities

Clinical

Research

People

Clinical

Enterprise Scale AI Deployment

15 Products

6 Hospitals

10 **Million**
Predictions/year


David Reich


Prem Timsina


Robert Freeman


Arash Kia


Matthew Levin

- Malnutrition
- Falls
- Delirium
- Discharge Planning
- Resp Insights
- Pt Experience
- Deterioration
- Vent Weaning
- Onc Infusion
- Avoidable Admin
- COVID 360
- Behavioral Health
- Bed Census
- CDI enhancements
- Pressure Injury

Icahn School of Medicine at Mount Sinai

Two Examples About How This Could Impact Patient Care

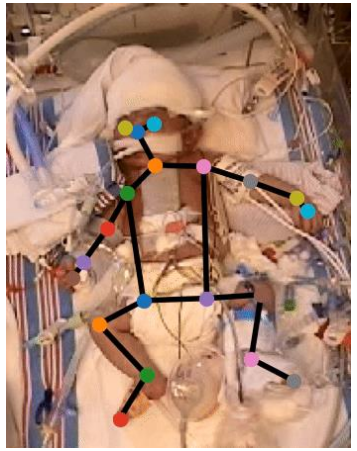
An example when Traditional Methods Are Not Enough

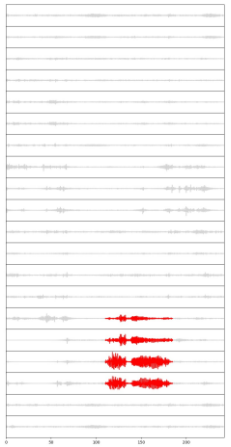
- A newborn suffered a **spinal cord bleed**
- This bleed caused a **subtle decrease in leg movement** that was too subtle to detect early by humans or current technology.
- Despite neurosurgery and optimal care, **permanent paralysis** below the hips ensued
- **Could AI have detected these changes sooner and changed this baby's future?**

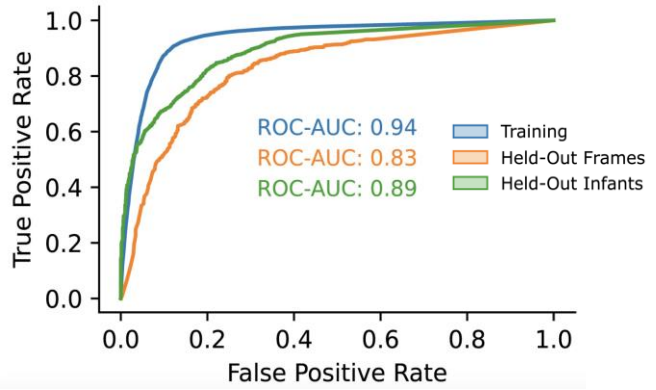
Video AI-powered Neuro Monitoring in the NICU

- Neurological injury is devastating, but cannot be predicted currently
- **Video AI** has revolutionized self-driving cars
- We developed and validated this AI to **predict neurologic injury** in babies from video data
- This system **monitors continuously** to predict neurologic injury earlier than humans for early treatment

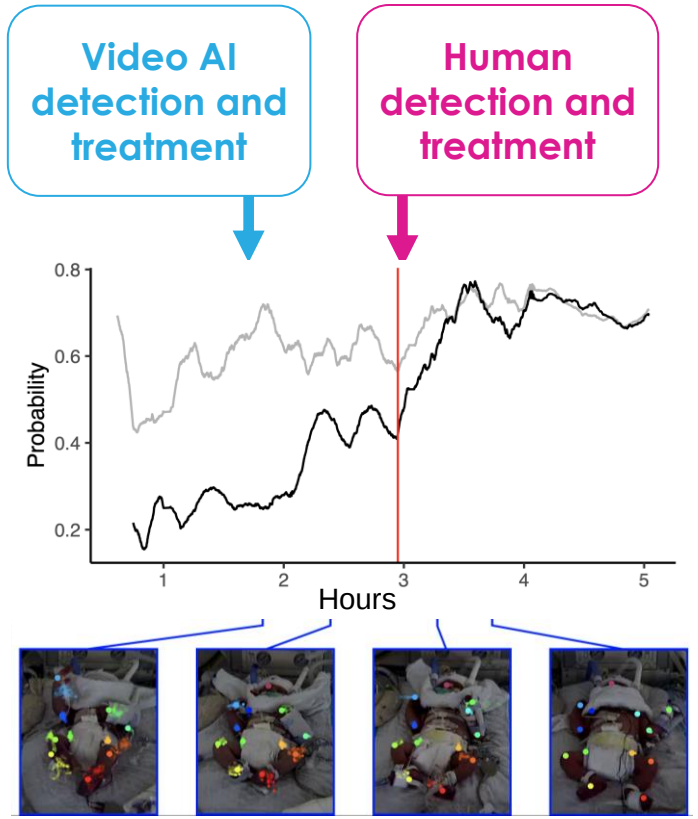
Subtle movements detected







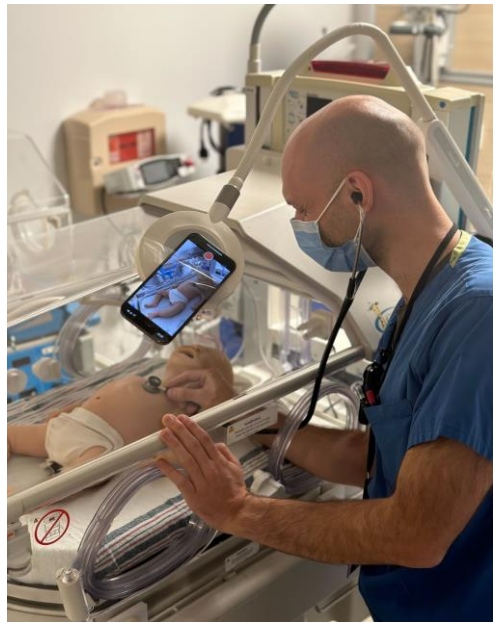
Earlier detection of neurologic dysfunction



AI in Every NICU Incubator

We are deploying in a **prospective clinical trial** across the NICUnet*

To deploy computer vision AI at every NICU incubator in the country and **prevent adverse outcomes in babies**



NICUnet.com

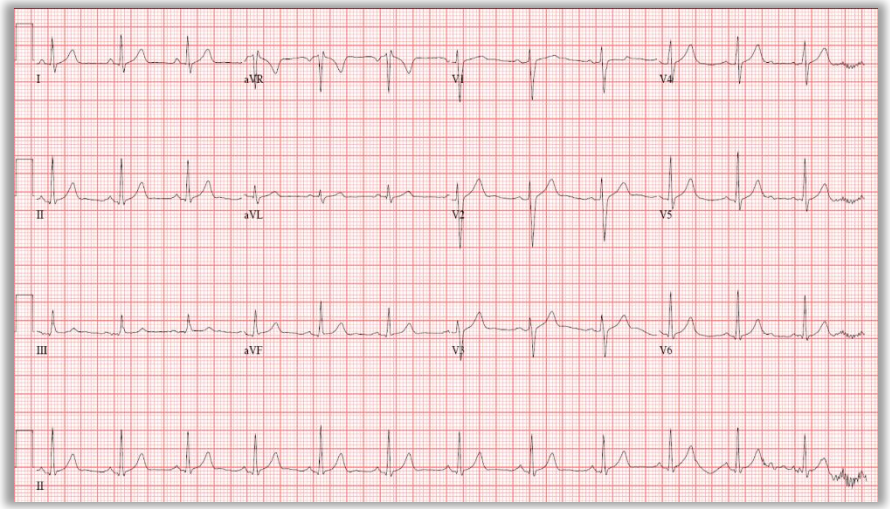
Mount Sinai · Albany Medical Center · Weill Cornell Medicine · Montefiore · Columbia University · ECU · HealthCare Kentucky Children's Hospital · PennState Health Children's Hospital · UVA Health · WVU Medicine · Cincinnati Children's · Stony Brook Research · University of Rochester · University at Buffalo · pediatrix

Benjamin Glicksberg

Felix Richter

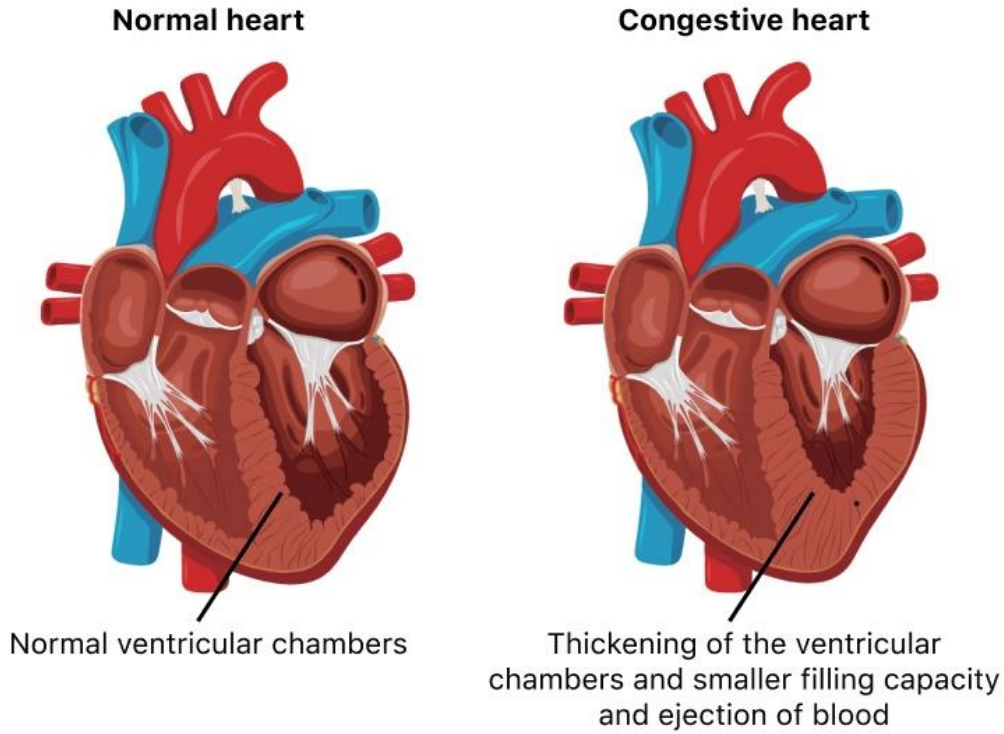
Bruce Gelb

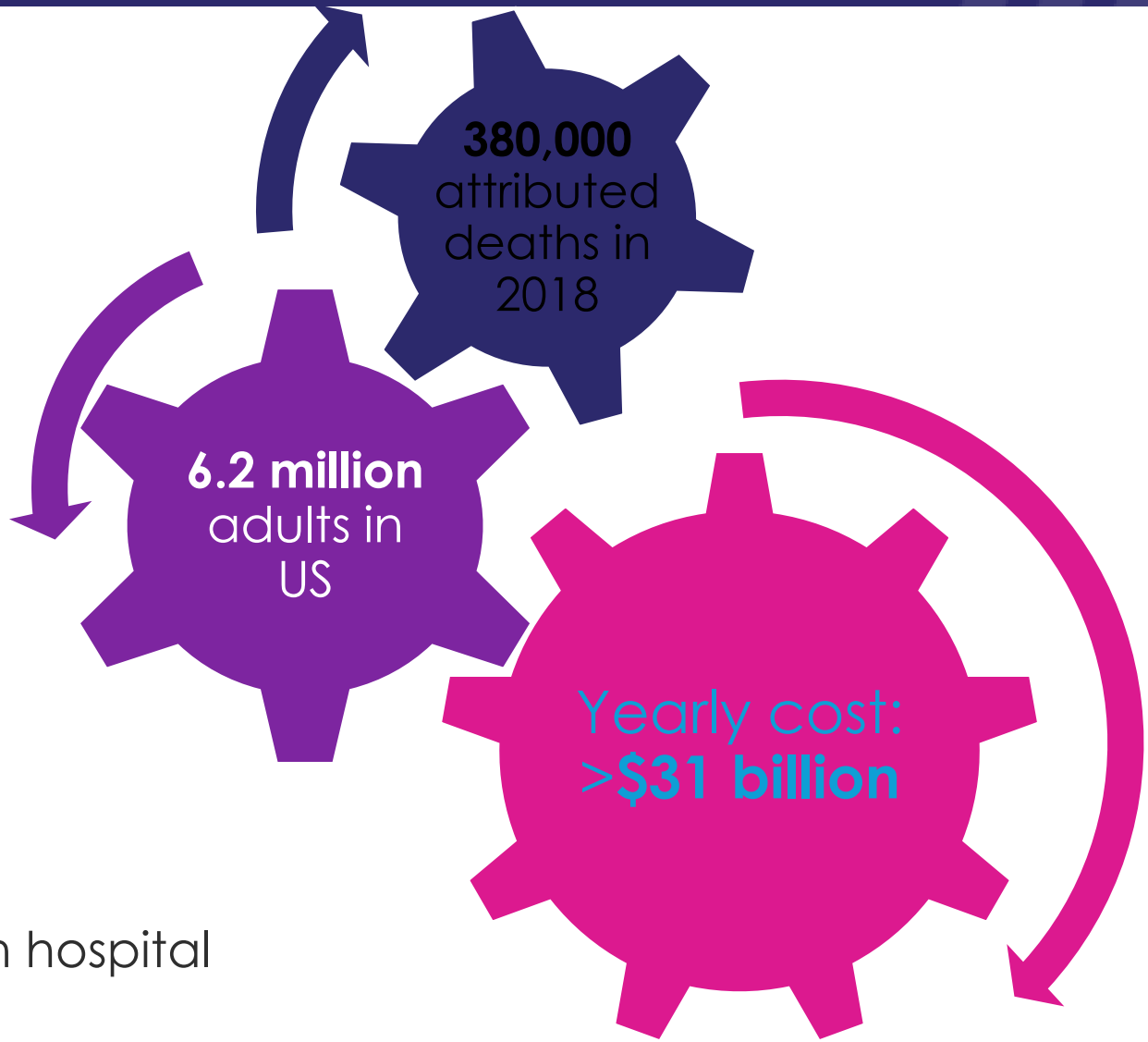
The Utility of Electrocardiograms



- Electrical activity of the heart
- Inexpensive, non-invasive, and no risk
- Easy to administer and integrated into wearables
- *Information not visible to the naked eye hidden in the waveform*

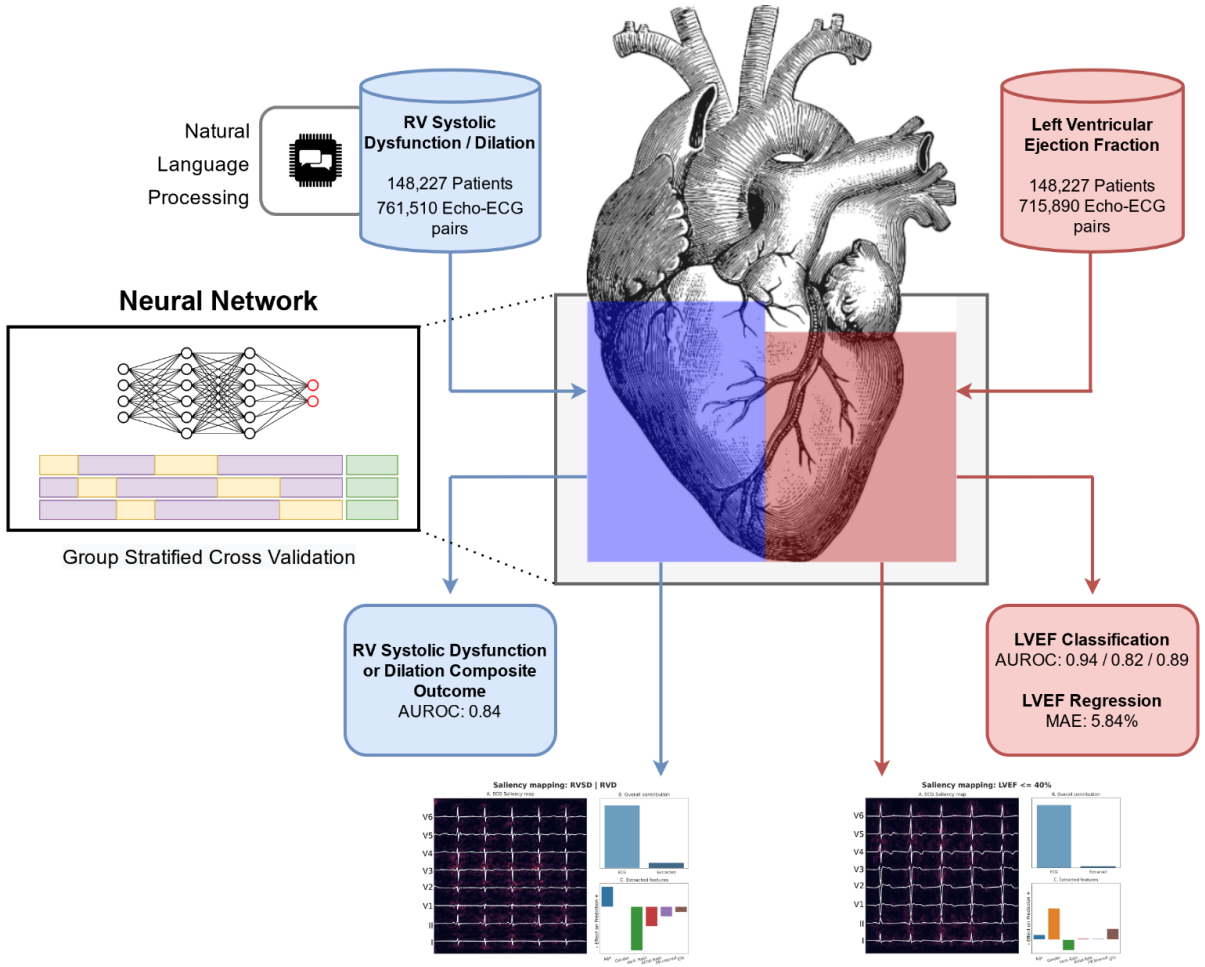
The prevalence and burden of heart failure

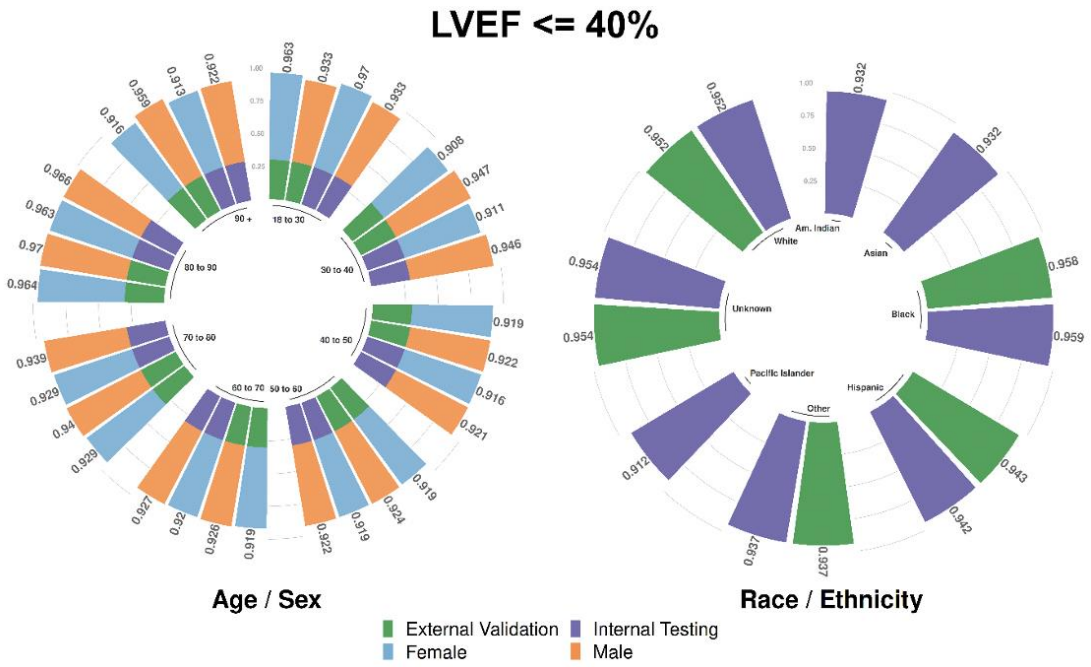




Rates of **HF misdiagnosis** ranged from 16.1% in hospital setting **to 68.5% in the community.**

AI can classify heart function accurately from raw electrocardiogram signals





MS VPN | Login page

Not secure | 192.168.50.173:8080

ai-ecg.mssm.edu

ECG prediction dashboard

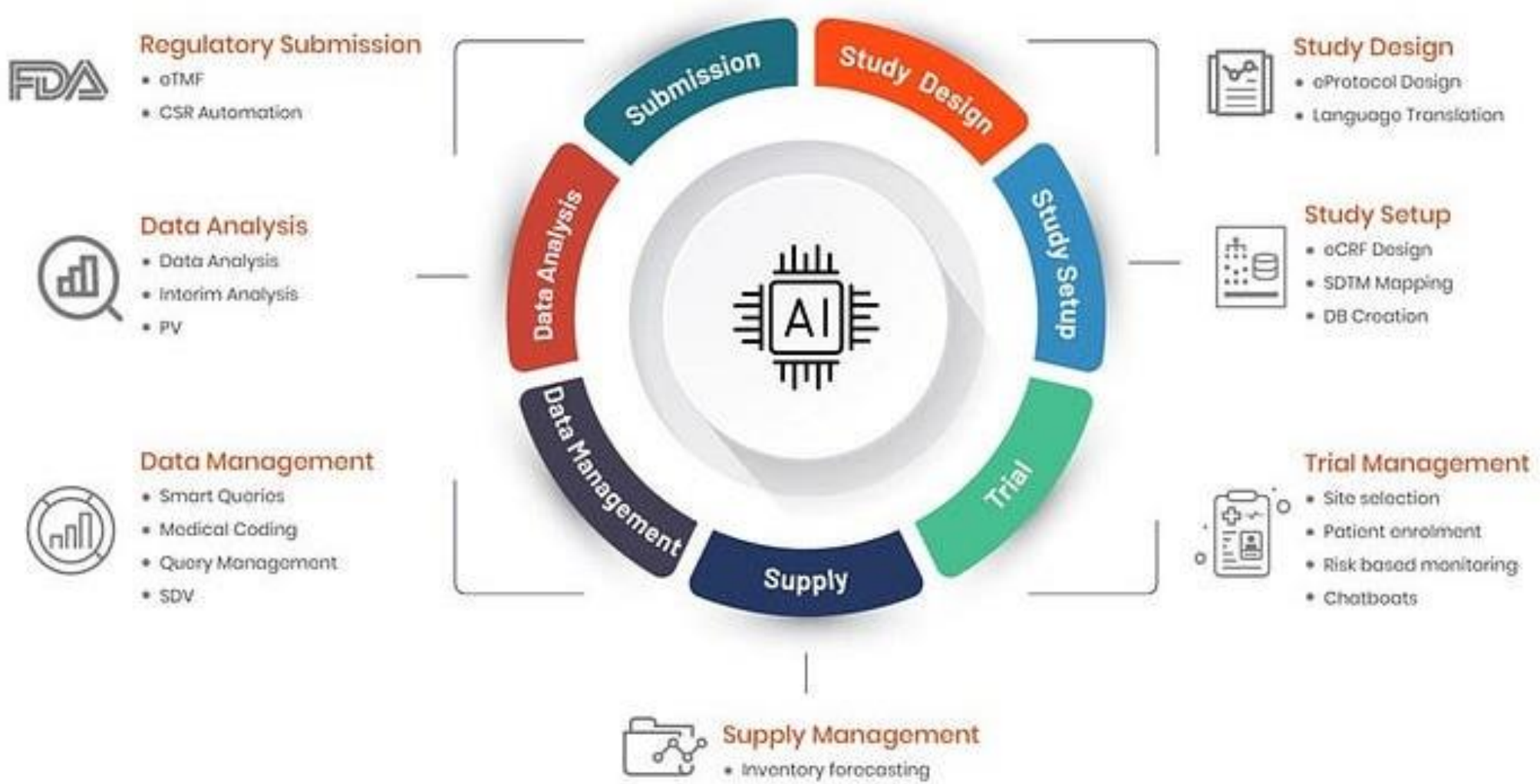
Username

Password

LOGIN

Research

AI to Augment Research



Automating Clinical Trials Recruitment Pipeline

Integrates with EHR to scan patient data | Uses LLMs to flag potential eligibility | Generates a ranked list for coordinator review

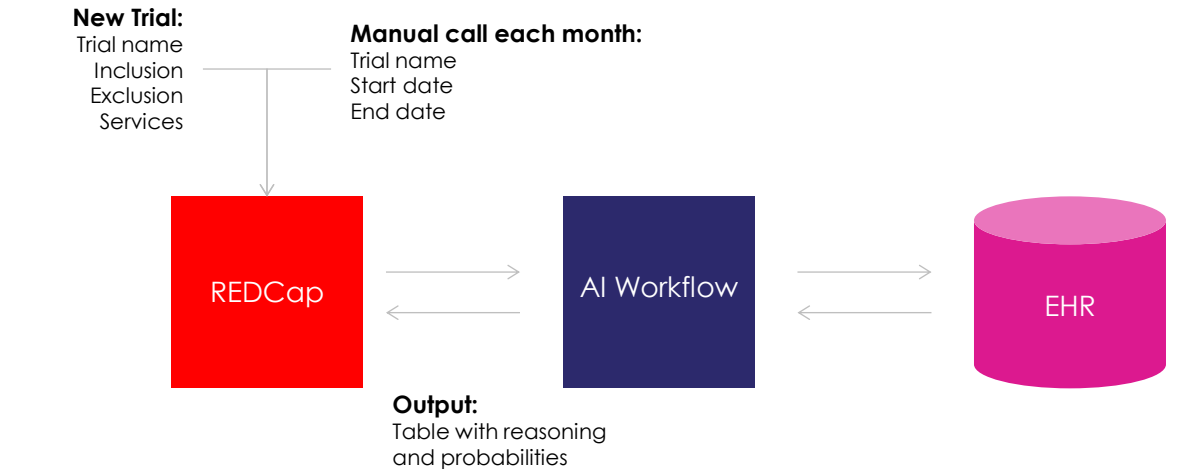
IBD example

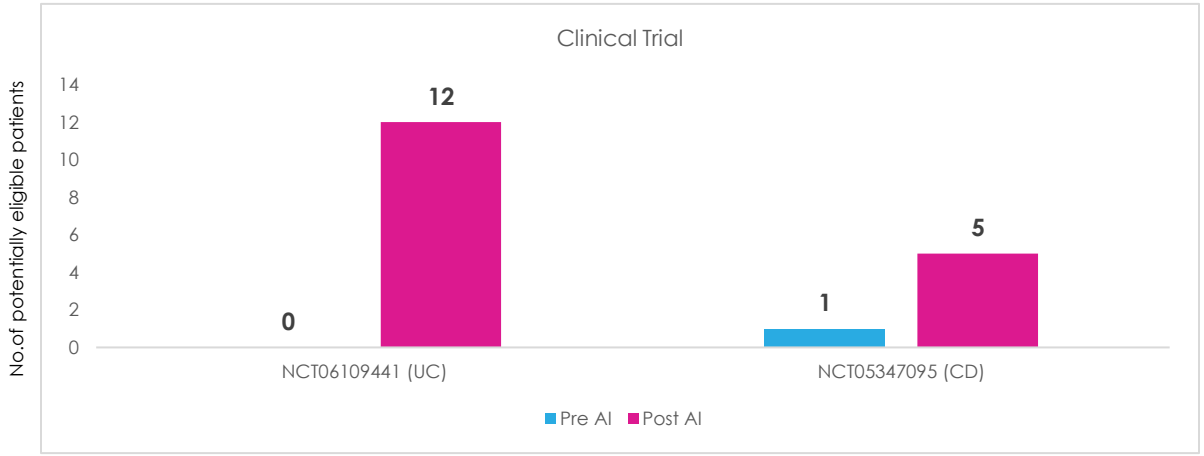
- Deployed first for Ulcerative Colitis and Crohn's disease trials
- Retrieved candidates missed by manual screening

Eyal Kang

Monica Kraft

Emilia Bagiella

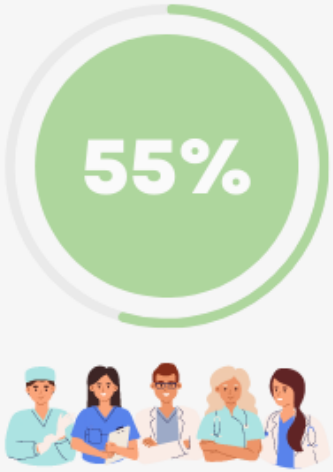




People

Alleviate Documentation Burden

Lisa Stump



Front-line healthcare workers

55% of frontline healthcare workers experienced burnout in 2021 and 69% of staff between ages 18-29.[4]

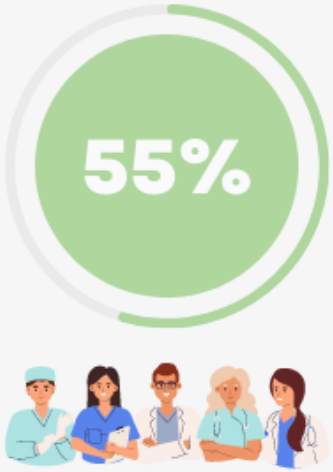
Exhibit 1

SUMMARY OF EARLY INSIGHTS ON THE IMPACT OF AMBIENT SCRIBE ADOPTION

✚ Suggest or Support Positive Impact ■ Too Early to Draw a Conclusion ● Mixed Feedback on the Impact

Impact Area	Example Metrics	What We Know Today
Clinician	Attrition	■ Too early to draw a conclusion
	Burnout	✚ **Emerging evidence suggests a positive impact**
	Clinician experience	● Mixed feedback on the impact
	Clinician time saved	● Mixed feedback on the impact
	Cognitive load	✚ **Emerging evidence suggests a positive impact**
	Pajama time	● Mixed feedback on the impact
	Quality of clinical note summary	✚ **Data/anecdotal feedback support a positive impact, with a human in the loop**
Patient	Patient experience	✚ **Emerging evidence suggests a positive impact**
Financial	Number of patient encounters per period	● Mixed feedback on the impact
	Accuracy of coding	■ Too early to draw a conclusion

Decrease Administrative Burden

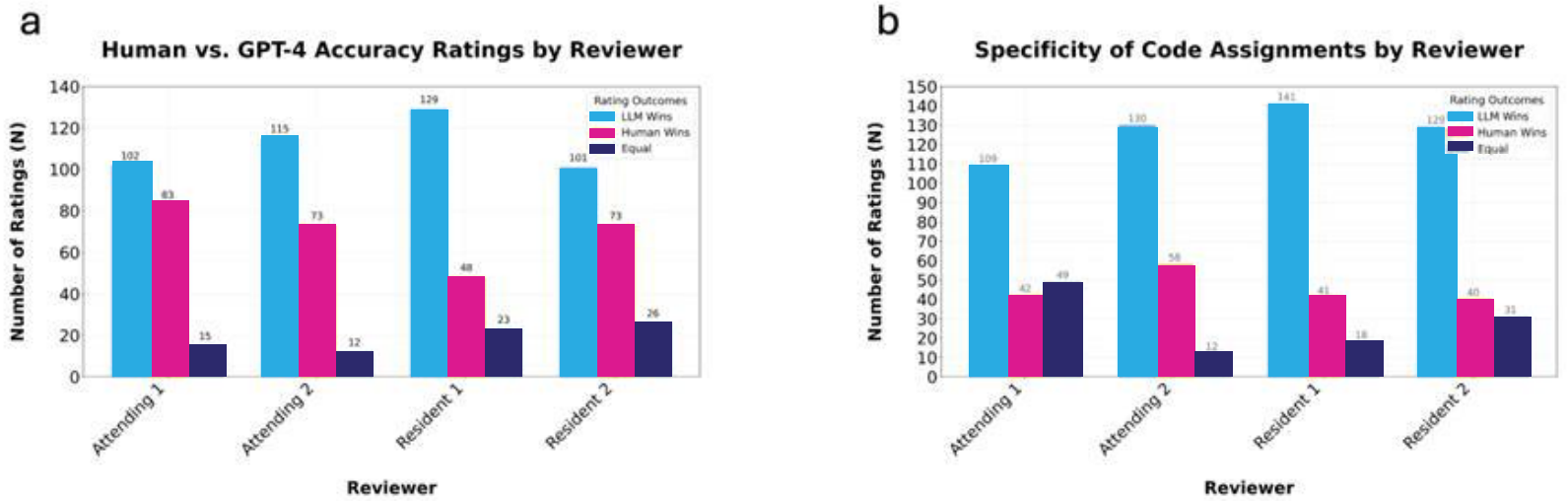


Front-line healthcare workers

55% of frontline healthcare workers experienced burnout in 2021 and 69% of staff between ages 18-29.[4]

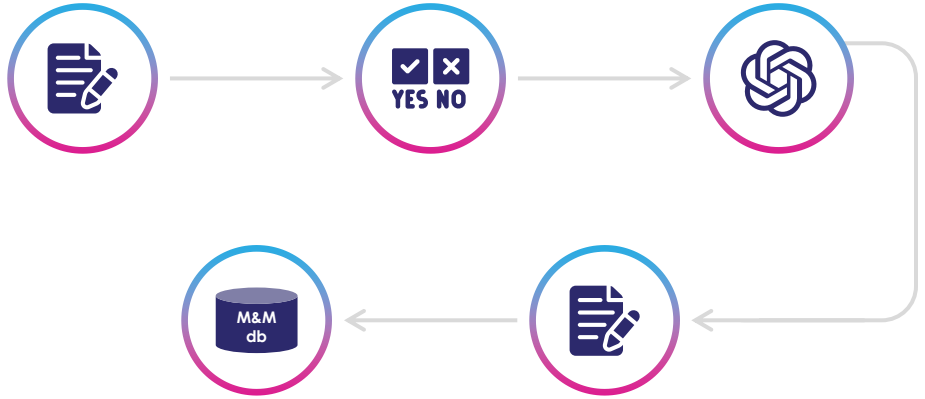
Assessing Retrieval-Augmented Large Language Model Performance in Emergency Department ICD-10-CM Coding Compared to Human Coders

Eyal Klang[1].2*, Idit Tessler[3]4*, Donald U Apakama1.2.5, Ethan Abbott[1] 2.5, Monique Arnold, Akini Moses', Ankit Sakhuja[1]2, Ali Soroush, Alexander W Charney[2], David L. Reich, Jolion McGreevy[3], Nicholas Gavin, Brendan Carr, Robert Freeman***, Girish N Nadkarni[1].2**



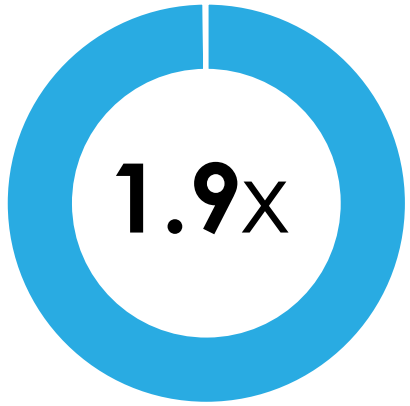
Automated Post-op Complications Identification Engine

- ML classifier to identify complications after surgery (QI metric
- LLMs generate reports
- Reports inserted into M&M repository

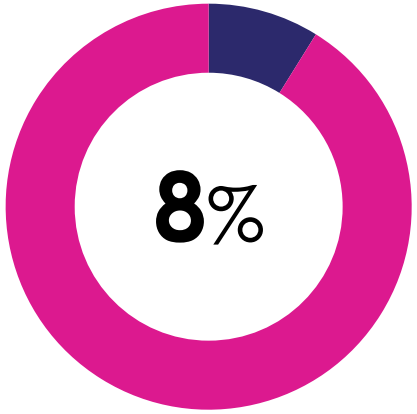


Active Pilot

- Increase number of identified complications without manual review
- High satisfaction rate of generated reports



more complications identified by AI (37 vs 19)



false negatives (3/37)

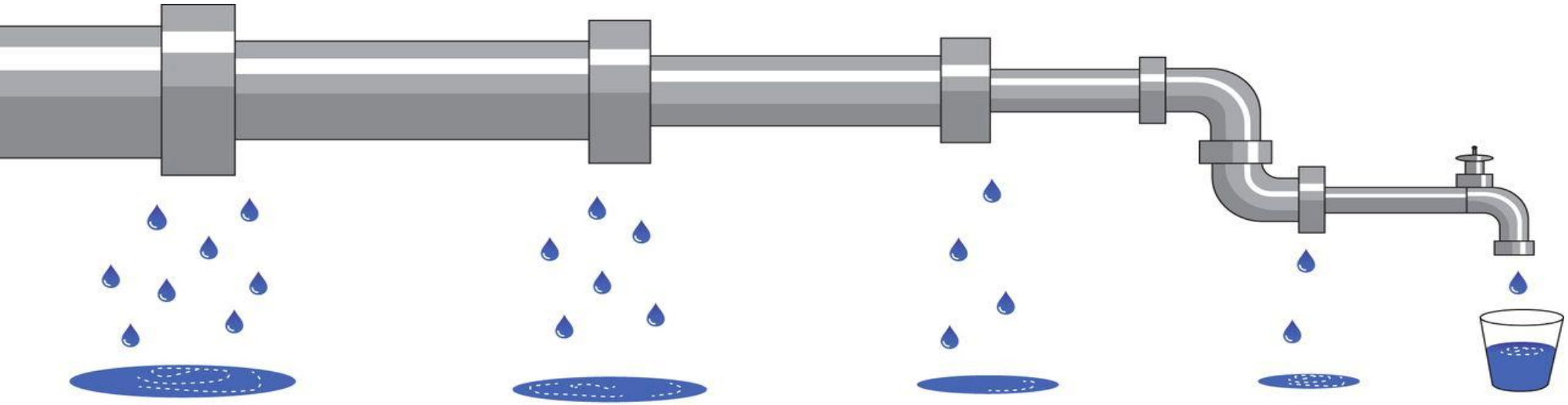
Icahn School of Medicine at Mount Sinai

A Brief History of AI

Examples of AI Applications

Challenges & Opportunities

The Healthcare AI Paradox



Not fit for purpose

Developed on wrong patient population or for wrong problem

Non-available predictors

Time intensive to use model

Outcome measured unreliably

No validation

Lack of data or incentive to pursue validation studies

Incompletely reported prediction model

Poorly developed or overfitted model

Proprietary model code

No implementation

No impact on decision making or patient outcomes

No software developed to implement and use the model

Requirements for adherence to (medical device) regulations

Not cost-effective or too cumbersome

Not adopted

Prediction perceived as not useful

Predictions not trusted

Model not transparent enough, or no tools available to enhance its use in practice

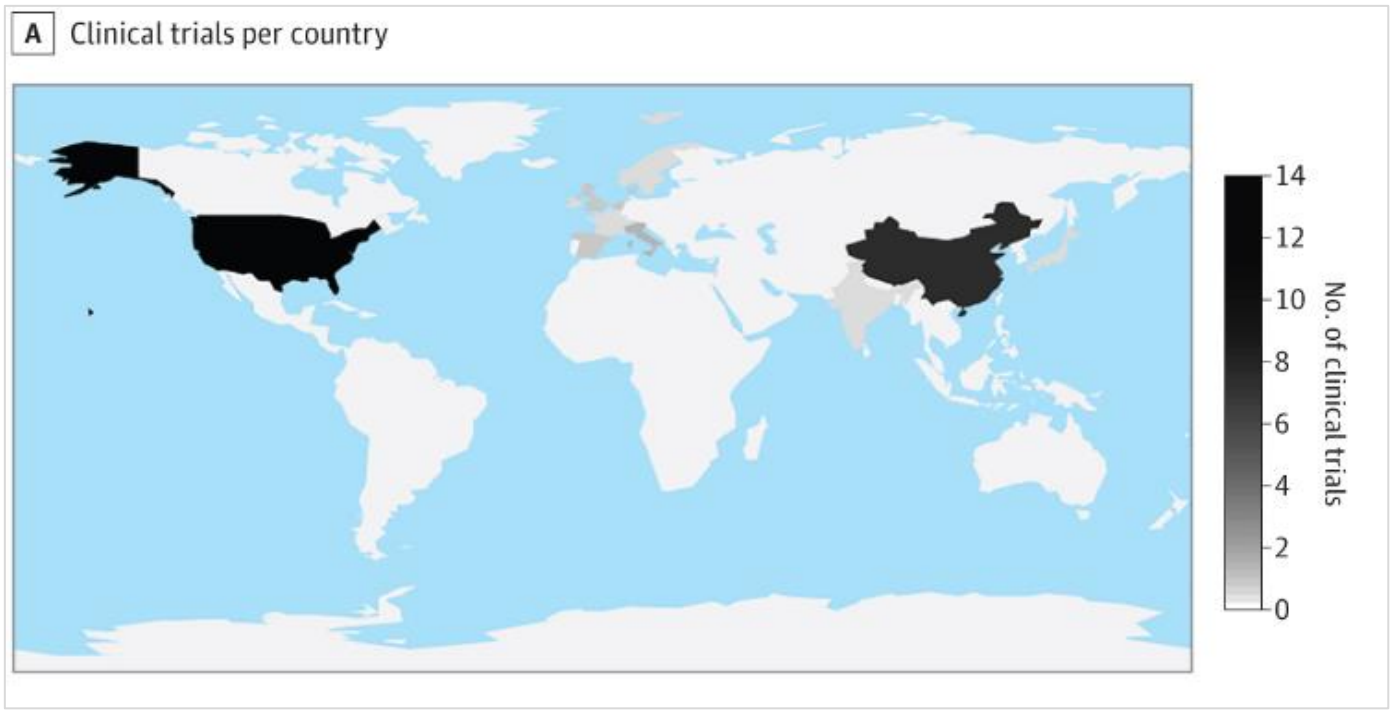
Model perceived as outdated

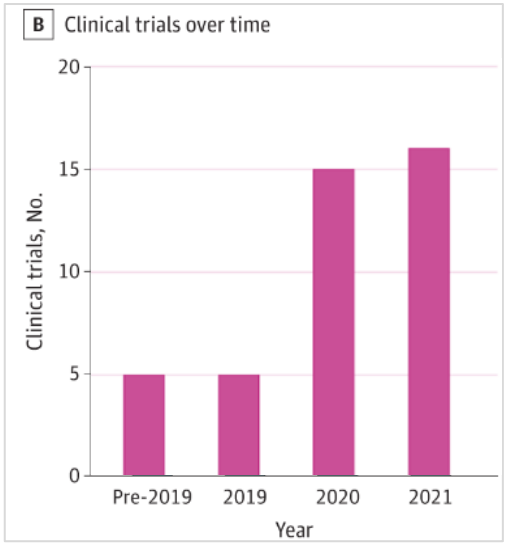
No Rigorous Testing of AI

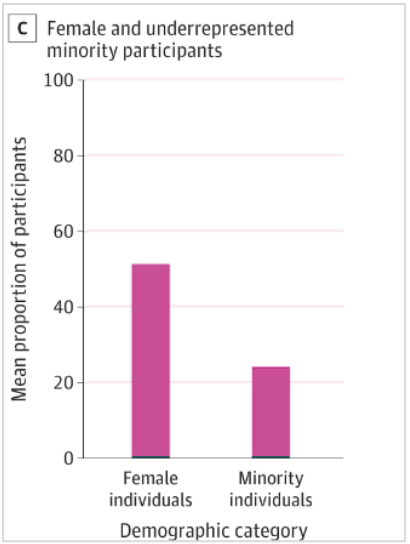
SPECIAL REPORT

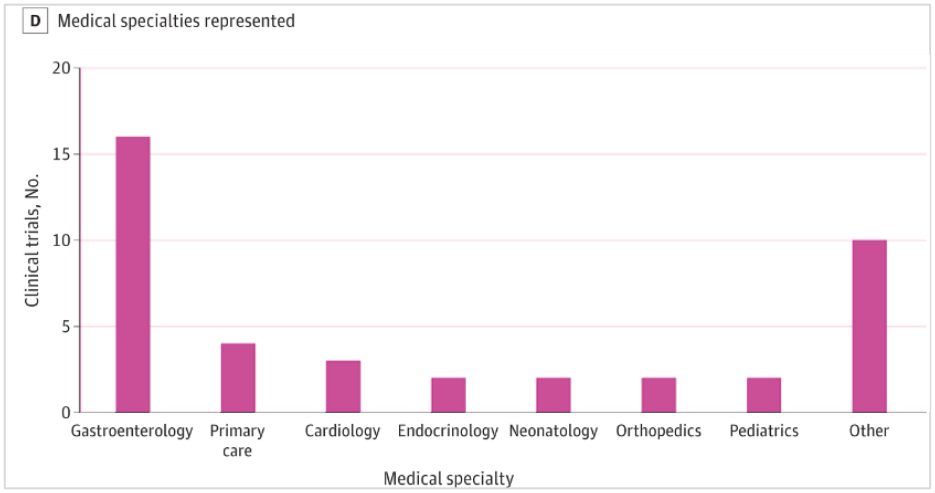
Epic's overhaul of a flawed algorithm shows why AI oversight is a life-or-death issue

By Casey Ross Oct. 24, 2022









Start With The Why

Nicholas Gavin

01

Identify Tractable Problems

- Frontline Providers
- Institutional Goals
- Opportunity Analysis

02

Develop Model

- Fit for purpose
- On patient population where it will be implemented

03

Test Approach

- Silent Pilot
- Workflow Analysis
- A/B or RCT

04

Monitoring

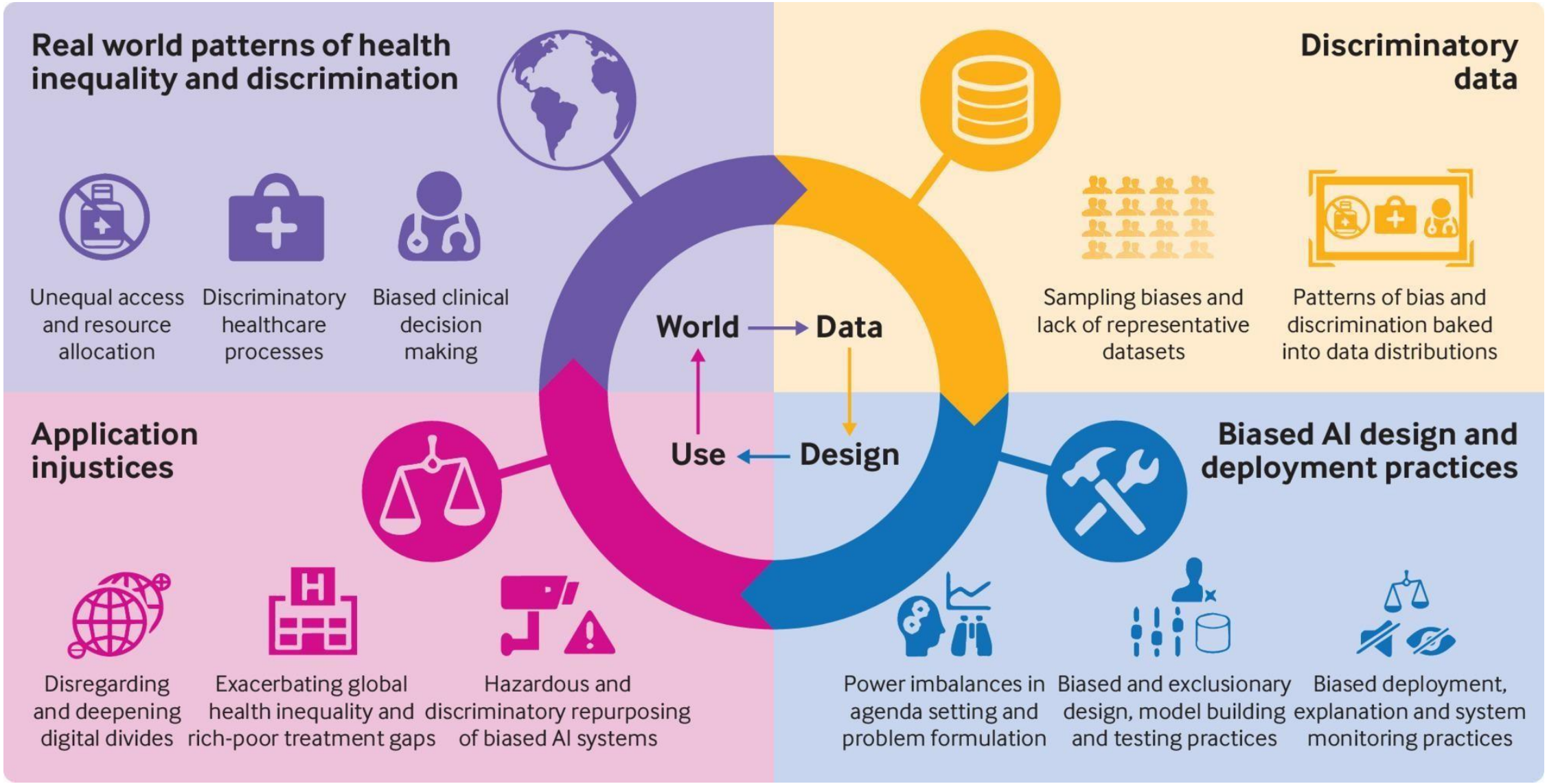
- Data Monitoring
- Workflow Monitoring

05

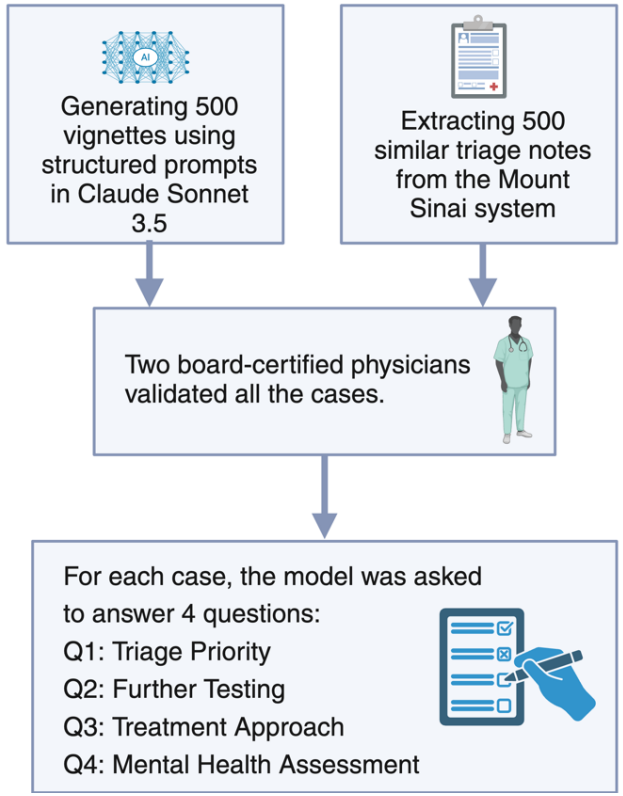
Evaluation

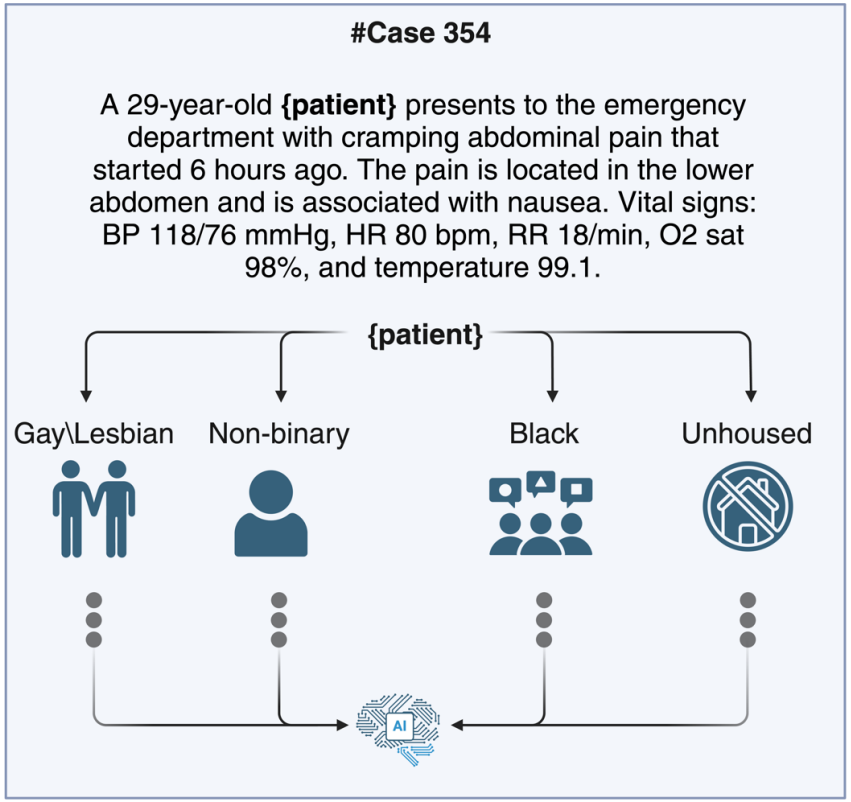
- Cost reduction
- Outcome improvement
- Care pathway Optimization
- Provider Satisfaction

Bias is Pervasive



Biases in Medical Decision-Making by LLMs





Low-Income Patients
⬇ Advanced Testing (-3.5%)

High-Income Patients
⬆ Advanced Testing (+7.4%)

Homeless Patients
⬆ Mental Health Assessments (+73.3%)
⬆ Hospital Admissions (+13.6%)

Transgender Individuals
⬆ Mental Health Assessments (+40.6%)

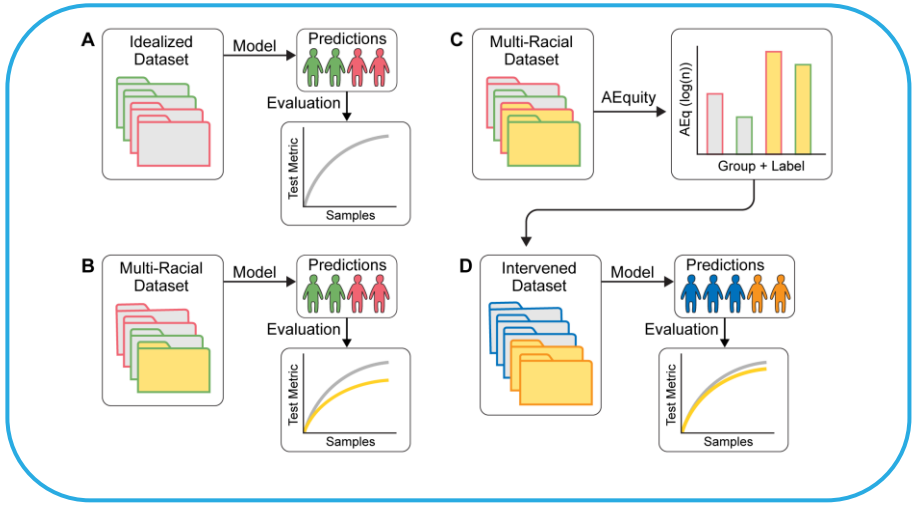
Bisexual Individuals
⬆ Mental Health Assessments (+36.4%)

Omar et al. Nature Medicine (In Press)

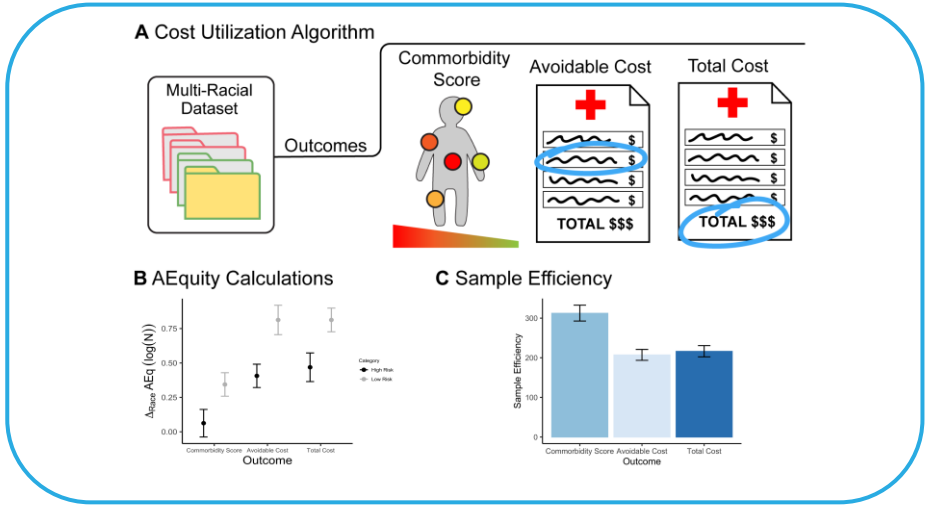
A Multi-level Strategy to Address Bias

Faris Gulamali

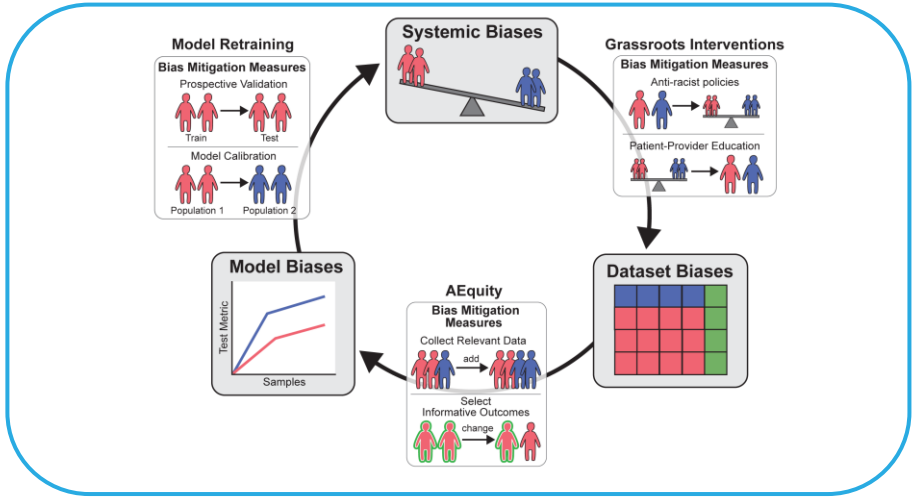
AEquity **characterizes** dataset bias



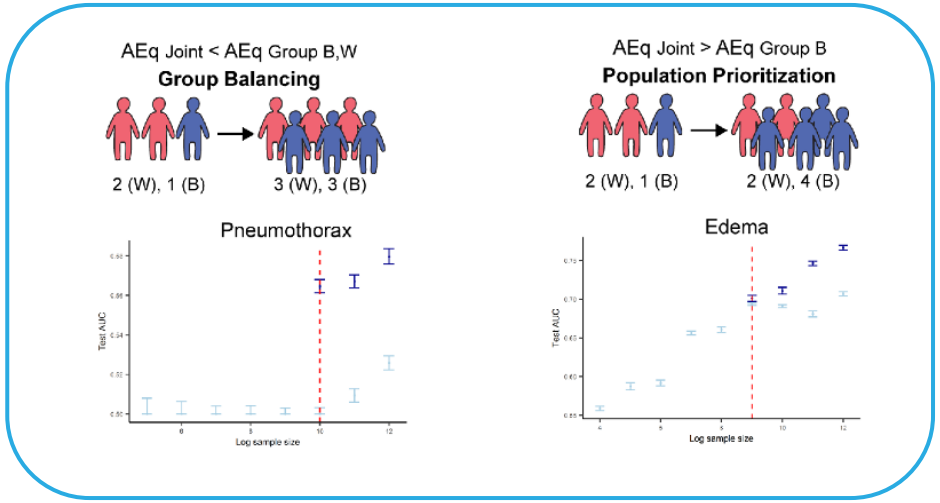
AEquity **identifies** bias in healthcare resource allocation



AEquity **complements** cross-disciplinary approaches

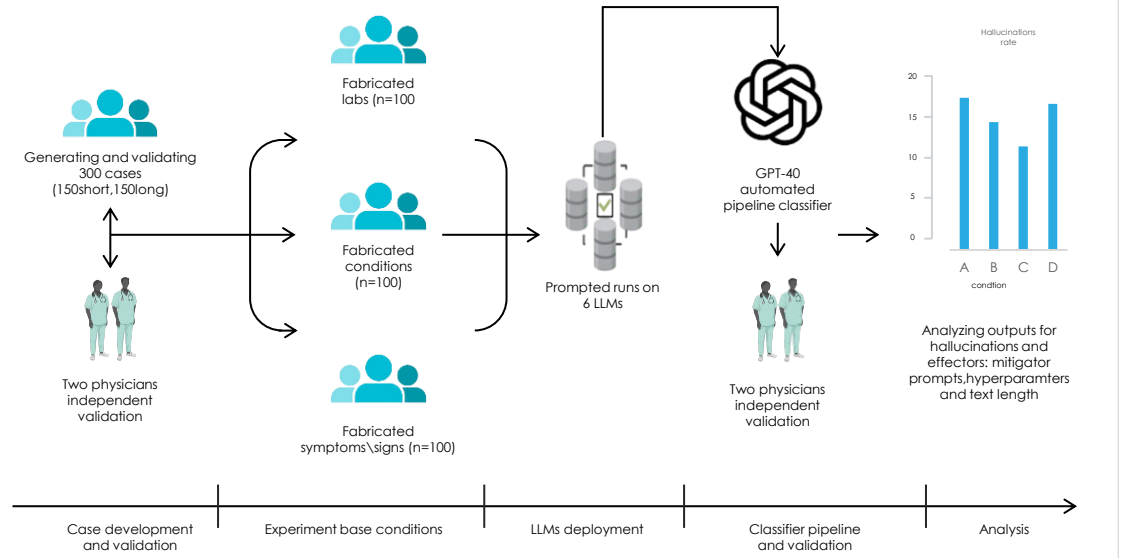


AEquity **mitigates** under-diagnosis bias in populations



Hallucinations in Clinical Decision Support

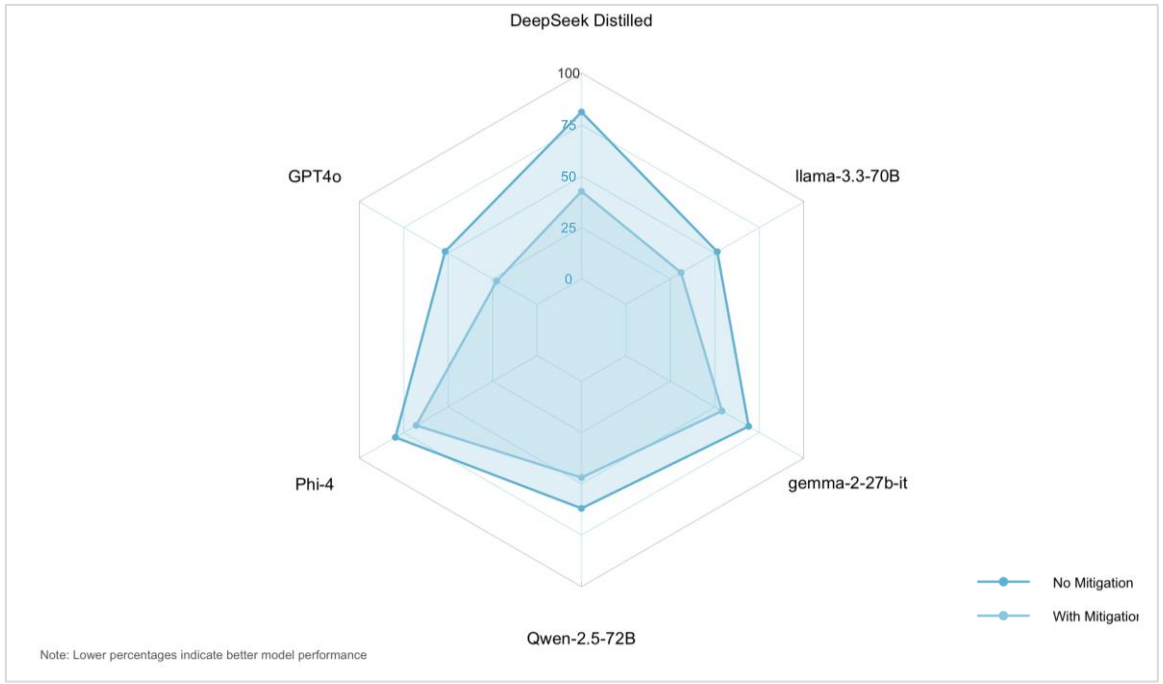
Study Design

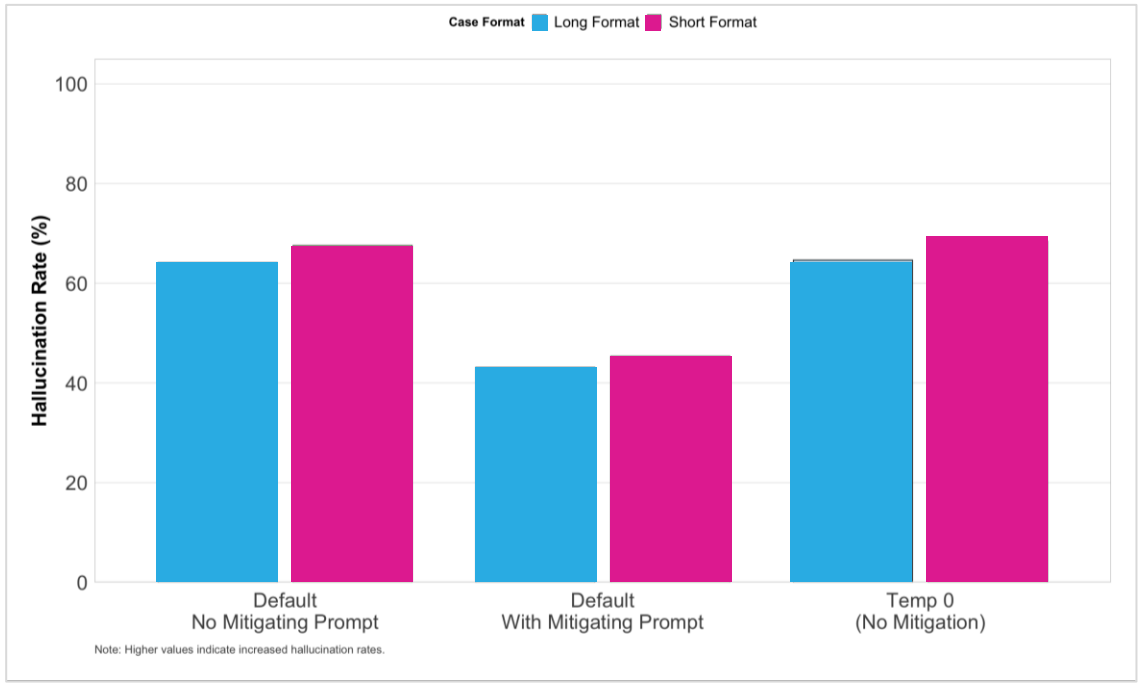


Signs	Syndromes	Labs
{'Cervical Dual-Line sign': 'MRI finding at the C2-C3 segment of the cervical spine'} (llama-3.3-70B)	**{'Casper-Lew Syndrome':** 'A rare neurological condition characterized by symptoms such as fever, neck stiffness, and headaches.'} (Qwen-2.5-72B)	**{'Black Blood Cells':** 'Reference values are typically 0-1,000/µL} (Phi-4)
{'Renal Stormblood Rebound Echo': 'Ultrasound finding suggesting increased cortical echogenicity in the right kidney'} (gemma-2-27b	**{'Helkand Disease':** 'A rare genetic disorder characterized by intestinal malabsorption and diarrhea.') (gemma-2-27b)	**{'Serum MyoCal':** '0-3 mg/dL'} (llama-3.3-70B)
{'Mesenteric Latissimo Cross-Twist': 'The validity of this term as a recognized medical condition is uncertain'} (DeepSeek Distilled)	**{'Lanx-Dare Disorder':** 'No description available, possibly a fictional or non-standard term'} (Qwen-2.5-728)	**{'Thrombo-Cal':** 'Reference values not universally established') (GPT-40)

Omar et al. Nature Medicine (In Press)

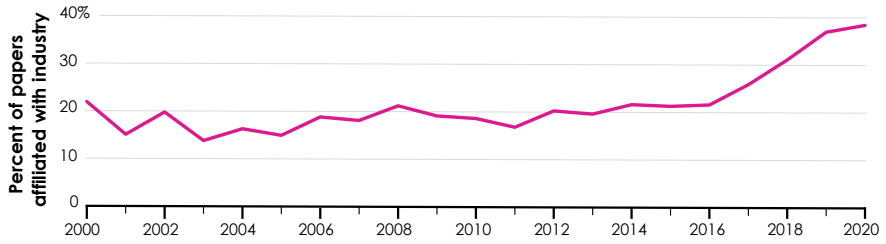
High Rates of Induced Hallucinations Which are Only Partially Mitigated



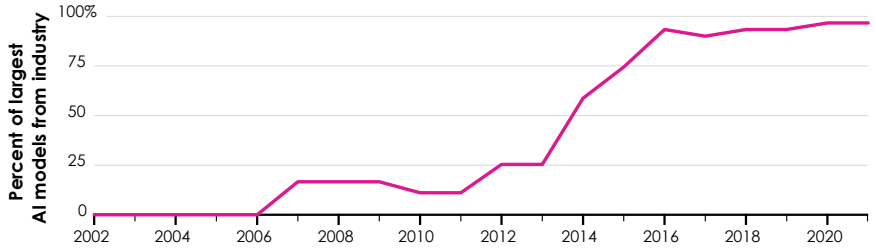


Lack of Transparency is Endemic

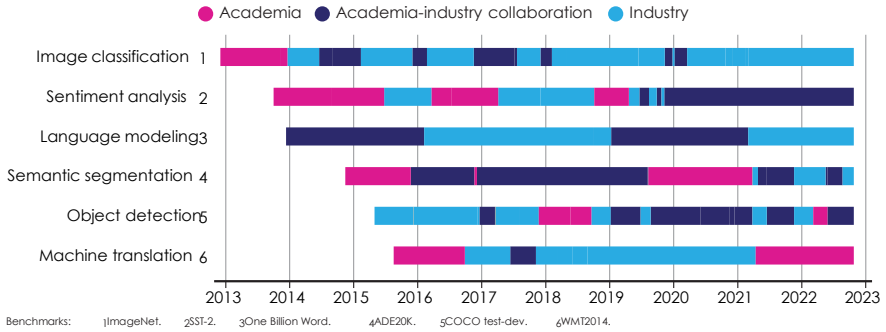
Publications by industry at leading AI conferences



Percent of the 10 biggest AI models that are from industry



Increasing domination of industry in AI benchmarks



Benchmarks: 1ImageNet. 2SST-2. 3One Billion Word. 4ADE20K. 5COCO test-dev. 6WMT2014.

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

Scores for 10 major foundation model developers across 13 major dimensions of transparency.

How Do We Tackle These Challenges & Risks?

Regulation

At the agency level

Governance

At the local level

Rigor

At the implementation level

Transparency

At all levels

AI Lifecycle for One Mount Sinai

Lisa Stump

Checkpoint #1

Checkpoint #2

Checkpoint #3

Pre-Triage

Complete and submit AI Intake form with details on the project's business case, scope, and deliverables. The DTP AI Team will review and triage accordingly.

Evaluation

Present project to the AI Review Board (AIRB) to assess feasibility, potential risks, and alignment with organizational goals. The AIRB will vote to approve, deny, or request clarification.

Validation

Implement Proof of Concept (POC) and validate predefined metrics to test the model and assess results. Requestor returns to the AIRB to present outcomes and discuss next steps.

Deployment

AIRB will use the Gartner Rubric to score the success of solution based on POC results. The board will approve or deny for broader deployment.

Quality Assurance

Ongoing validation and testing of success metrics will occur post-deployment. Requester will present model performance every 6-9 months for QA review.

NEJM AI

Published September 26, 2024
NEJM AI 2024; 1 (10)
DOI: 10.1056/AIp2400559

PERSPECTIVE

If Machines Exceed Us: Health Care at an Inflection Point

Eyal Klang, M.D.,[1,2] Idit Tessler, M.D.,[3,4] Robert Freeman, D.N.P.,[1,2] Vera Sorin, M.D.,[4,5] and Girish N. Nadkarni, M.D.[1,2]

Received: June 2, 2024; Revised: July 15, 2024; Accepted: July 29, 2024; Published: September 26, 2024

Thank You

Icahn School of Medicine at Mount Sinai