



**Brigham and Women's Hospital**

Founding Member, Mass General Brigham

## **Internal Medicine in the time of AI: Harnessing the Hype**

Rebecca G. Mishuris, MD, MS, MPH  
Chief Medical Information Officer and VP, Digital  
Mass General Brigham

Member of the Faculty  
Harvard Medical School



# Rebecca G. Mishuris, MD, MS, MPH, FAMIA



Medicine Residency @ Boston Medical Center  
Chief Medical Resident @ Boston Medical Center  
General Internal Medicine Fellowship @ BWH

Chief Medical Information Officer and VP @ Mass  
General Brigham

# Disclosures

Advisor for Elsevier, Inc.



# Learning Objectives

1. Understand generative AI: what it is and the basics of how it works
2. Identify opportunities and applications of generative AI in internal medicine
3. Evaluate ethical and practical considerations for using generative AI in healthcare



# Outline



Artificial Intelligence

Analytical AI  
Generative AI



Gen AI in Healthcare and IM

Clinician Experience  
Research and Education



Scaling Gen AI



Future of IM with Generative AI



# Artificial Intelligence



# What is Artificial Intelligence (AI)?

- “Artificial Intelligence refers to the development of computer algorithms that can perform tasks that typically require human intelligence, such as learning, reasoning, perception, and decision making” (ChatGPT)
- AI is based on machine learning algorithms and other computational techniques:

## ✓ Probabilistic Reasoning

Machine Learning  
Predictive Modeling  
Deep Learning  
Decision Trees

## ✓ Computational Logic

Rule-Based Systems  
Logic Programming  
Heuristic Techniques  
Case Based Reasoning

## ✓ Optimization Techniques

Constraint Satisfaction  
Constraint-Based Reasoning  
Linear Programming  
Genetic Algorithms



**Agent Based  
Orchestration**

## Perception Systems/Ambient Intelligence

Computer Vision  
Olfactive & Haptic  
Auditory/Speech

## Knowledge Representation, Learning/Search

Knowledge Graphs  
Semantic Networks

## Natural Language Processing

Text Analytics  
Natural Language Understanding  
Chatbots  
Natural Language Generation  
Dialog Management



# Two Types of Artificial Intelligence

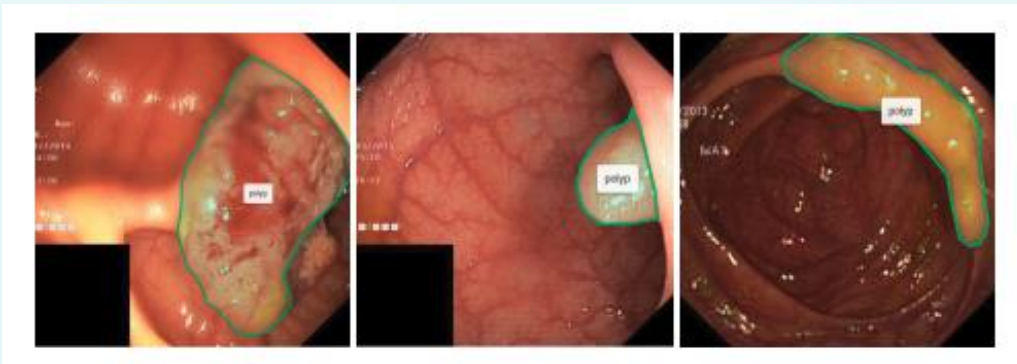
## Analytical AI

Analyze datasets to reveal novel insights

- Classification
- Prediction
- Recognition
- Other models

### *Segmentation of GI Polyps*

Use deep learning convolutional neural networks (CNNs) to identify gastrointestinal polyps



## Generative AI

Generate novel content informed by training dataset

- Summarization
- Generation
- Interaction

### Generative AI examples:

- Text (Bard, ChatGPT, MedPalm)
- Images (DALL-E2)
- Code (Github Copilot)
- Video/speech (Synthesia)







# Generative AI: ChatGPT

## What is GPT?

- Large language model (LLM) chatbot developed by OpenAI based on generative pre-trained transformer (GPT)-3.5
- LLMs predict the next word in a series of words in a sentence and the next sentences
- GPT-3 has 175 billion parameters and was trained on ~45 terabytes of text and information from the internet
- Also trained using human feedback (a technique called Reinforcement Learning with Human Feedback) so that the AI learned what humans expected when they asked a question
- Interacts in a text-based conversational way

## What is ChatGPT? (As answered by ChatGPT)

 What is ChatGPT?

 I am ChatGPT, a large language model developed by OpenAI. I have been trained on a massive dataset of text using deep learning techniques, which allows me to understand and generate human-like responses to a wide range of questions and prompts. As a language model, my primary function is to generate natural language responses to user inputs, whether they are in the form of text or speech. I can assist with a variety of tasks, from answering questions and providing information to engaging in conversations and even generating creative writing prompts or stories.



# ChatGPT

---

## Early Success

[Performed at or near the passing threshold \(60%\) on the US Medical Licensing Examination \(USMLE\)](#)

---

[Performed better than other Symptom Checkers on standard clinical vignettes](#)

---

## Limitations

ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers

---

Can exhibit undesirable behavior, including known racial, gender, and religious biases

---

Sensitive to tweaks in the input phrasing or attempting the same prompt multiple times

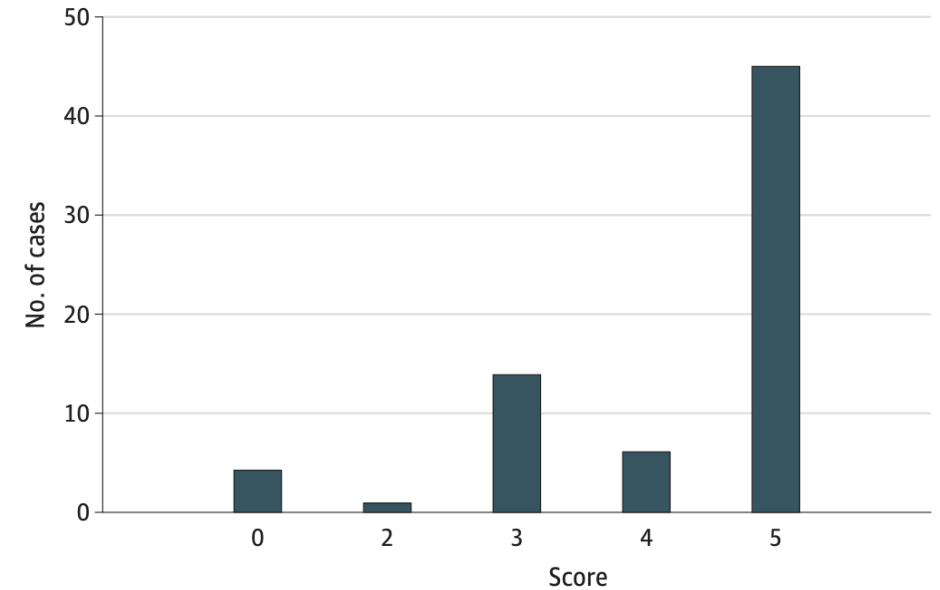
---



# Accuracy of GPT4 for Complex Diagnostic Challenge

- *New England Journal of Medicine* clinico-pathologic conferences from Jan 2021 – Dec 2022 (only diagnostic dilemmas)
- Outcome: model's top diagnosis matched the final case diagnosis (5-point scale)
- Results:
  - GPT4 top diagnosis agreed with the final diagnosis in 39% (27/70) of cases.
  - In 64% of cases (45/70), the model included the final diagnosis in its differential
- Compares favorably with existing differential diagnosis generators
- Generative AI is a promising adjunct to human cognition in diagnosis

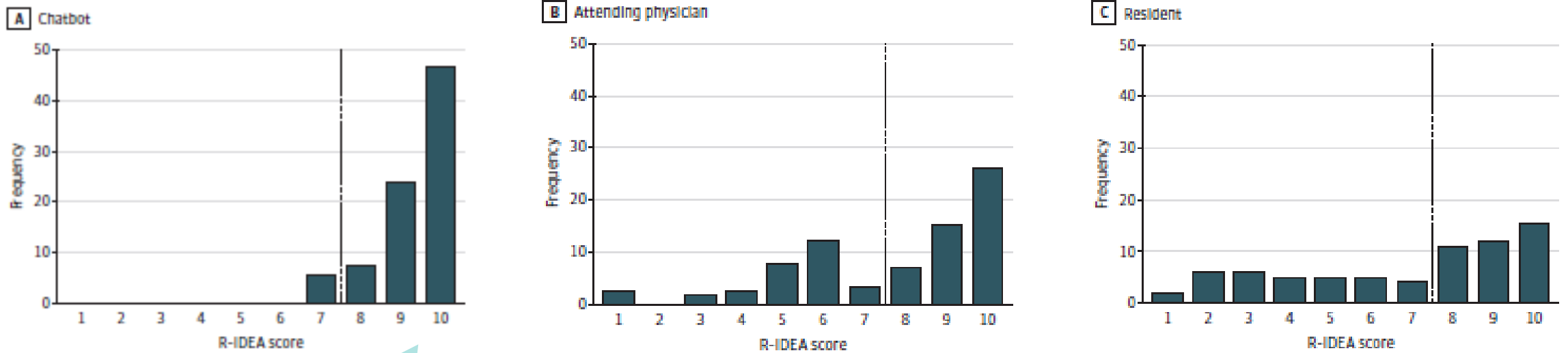
Figure. Performance of Generative Pre-trained Transformer 4 (GPT-4)



Histogram of GPT-4's performance. Performance scale scores (Bond et al<sup>2</sup>): 5 = the actual diagnosis was suggested in the differential; 4 = the suggestions included something very close, but not exact; 3 = the suggestions included something closely related that might have been helpful; 2 = the suggestions included something related, but unlikely to be helpful; 0 = no suggestions close to the target diagnosis. (The scale does not contain a score of 1.)

# LLM can perform clinical reasoning – with some limitations

Figure. Distribution of Revised-IDEA (R-IDEA) Scores



Distribution of R-IDEA scores is displayed across 232 sections, stratified by respondent type: chatbot (n = 80), attending physicians (n = 80), and residents (n = 72). The dashed vertical line delineates the low score (0-7) and high score (8-10) categories used in the regression analysis. For visual clarity, scores of 0 were aggregated with scores of 1.

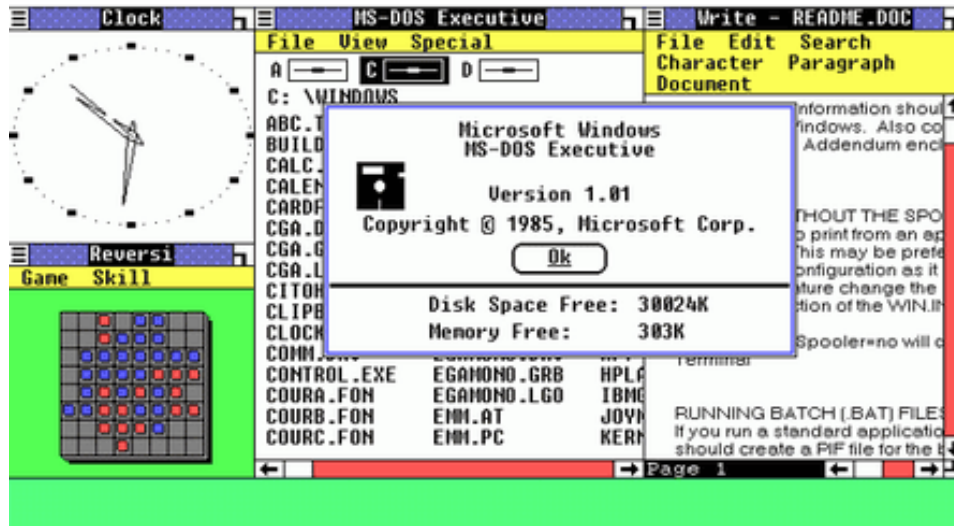
71 yo DDX: ectopic pregnancy



# The Age of AI has Begun

“In my lifetime, I’ve seen two demonstrations of technology that struck me as revolutionary.” Bill Gates

## Graphical User Interface



## ChatGPT

AD What is ChatGPT?

I am ChatGPT, a large language model developed by OpenAI. I have been trained on a massive dataset of text using deep learning techniques, which allows me to understand and generate human-like responses to a wide range of questions and prompts. As a language model, my primary function is to generate natural language responses to user inputs, whether they are in the form of text or speech. I can assist with a variety of tasks, from answering questions and providing information to engaging in conversations and even generating creative writing prompts or stories.

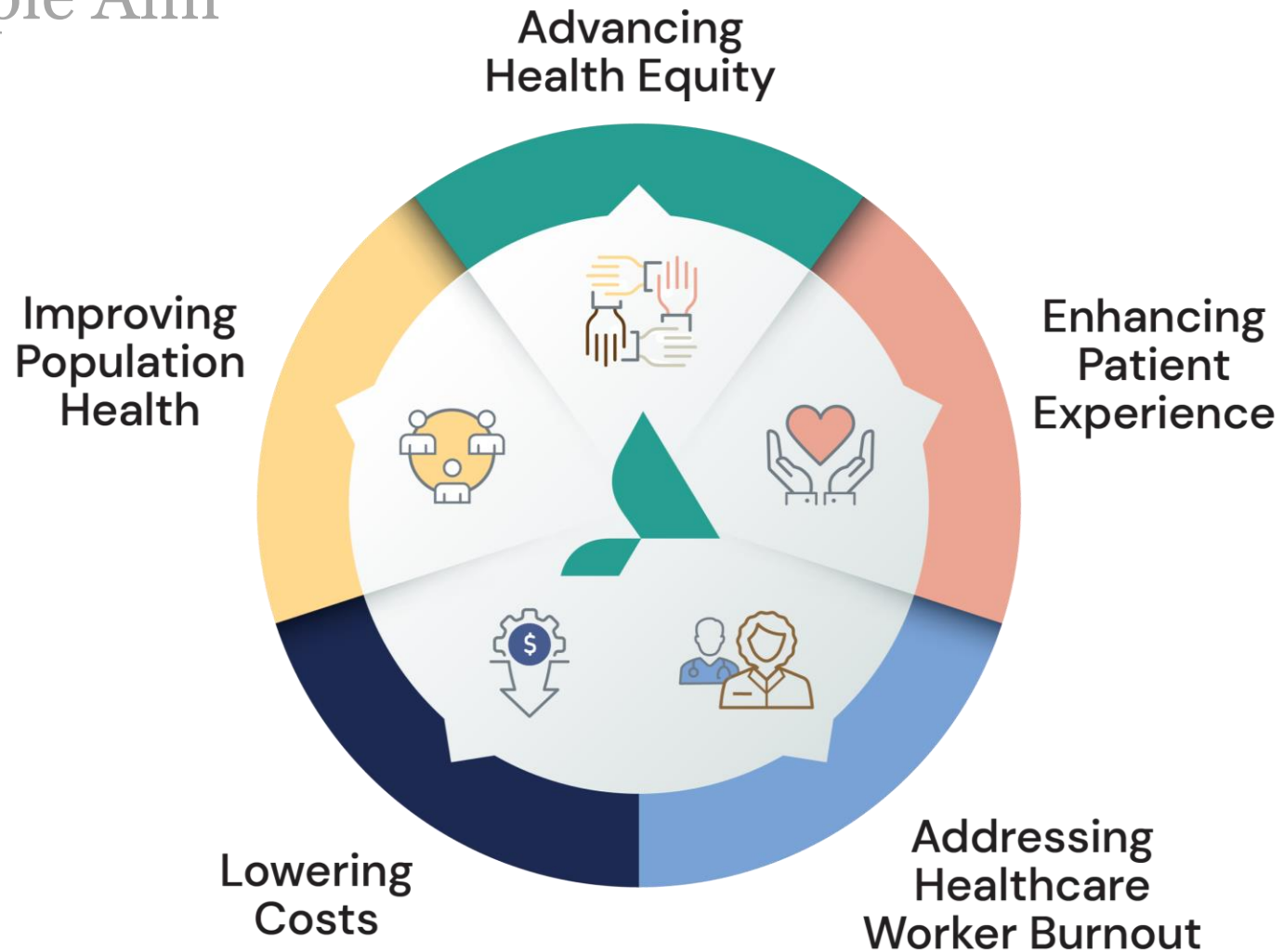


# Gen AI in Healthcare and Internal Medicine

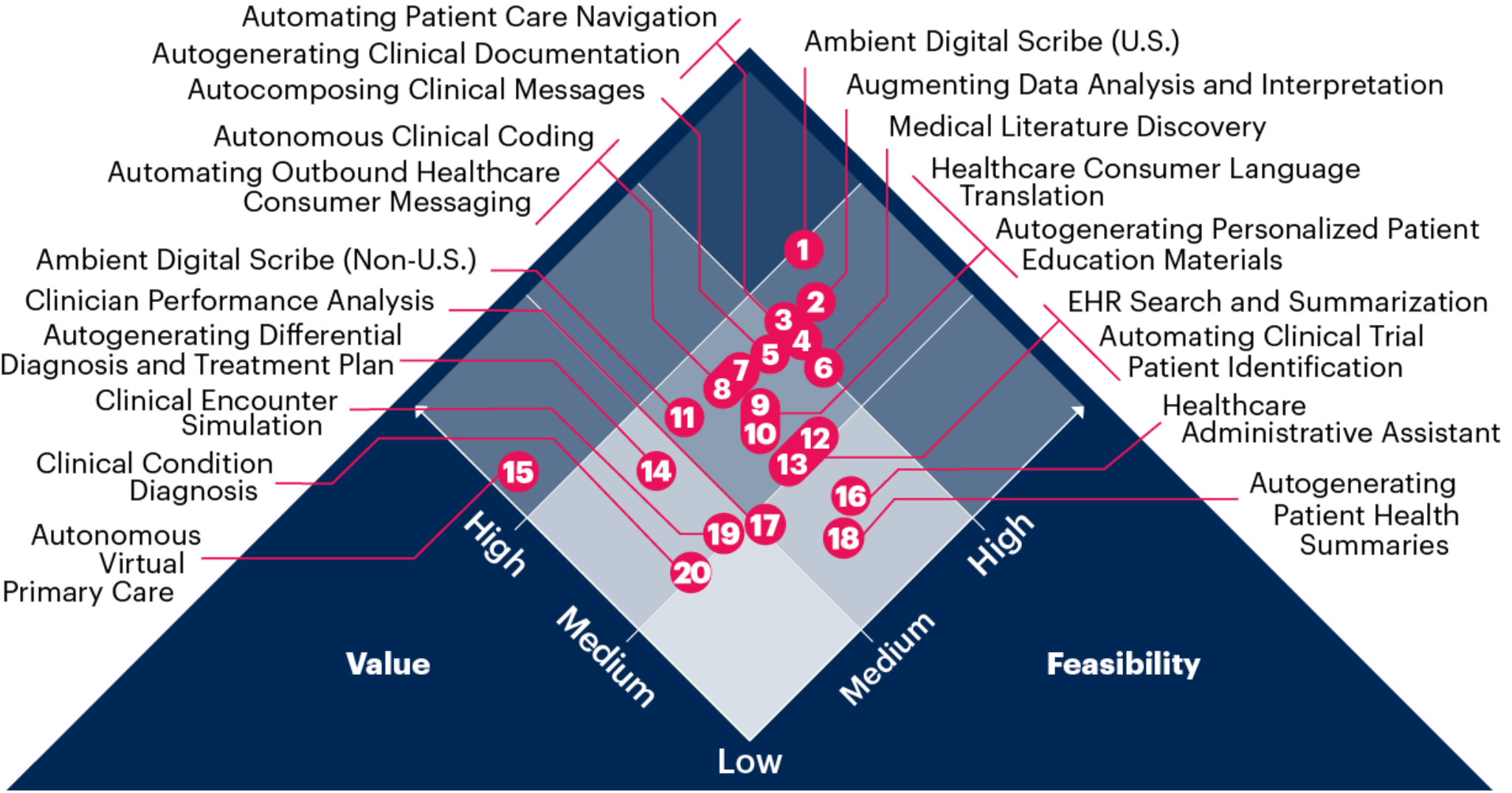


# The Role of Technology in Healthcare

## The Quintuple Aim



# GenAI Use-Case Prism for Healthcare Providers



Source: Gartner  
797404\_C

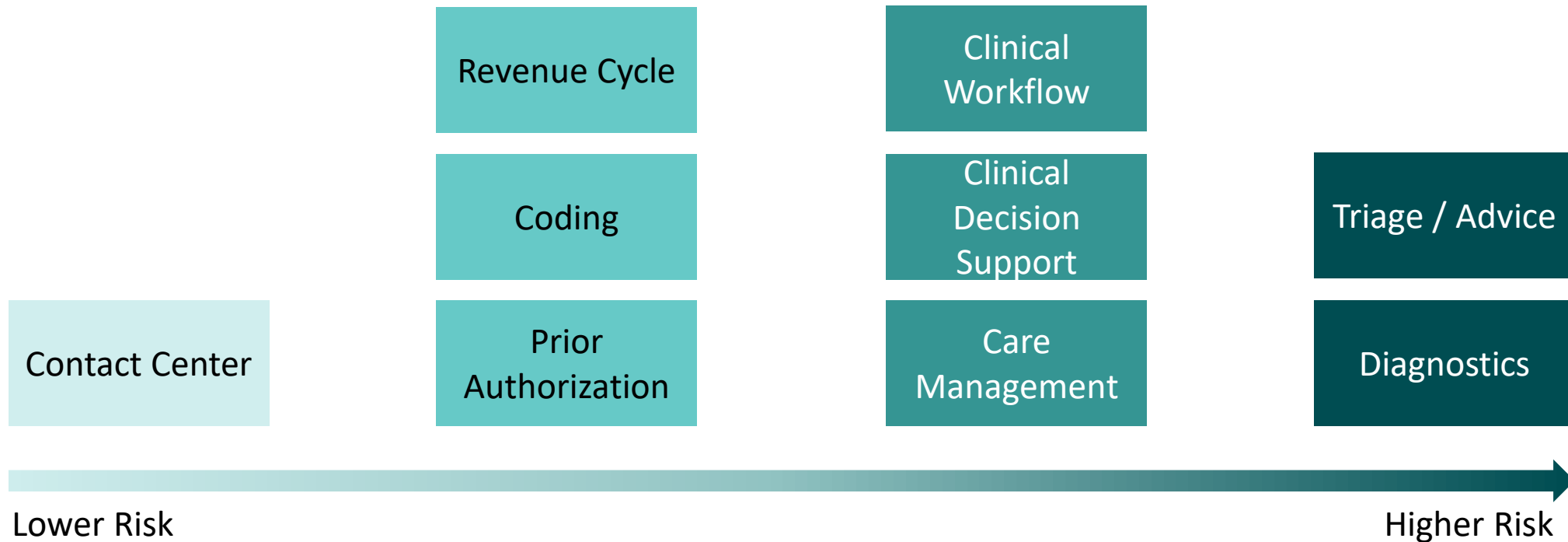




# The Opportunity for AI in Healthcare

We are scratching the surface of applying AI to address the Quintuple Aim and challenges in healthcare

## Sample AI Use Cases by Risk



# AI use cases in healthcare

## Care Team Experience

- Reduce administrative burden / optimize use of staff time
- Decrease burnout



Ambient note documentation



Draft replies to portal messages



Chart summarization



## Patient Experience

- Improve access to care
- Streamline interactions with health system



Call handling



Portal message categorization



Text translation

## Employee Experience

- Facilitate use of AI as business efficiency tool
- Simplify administrative functions



Microsoft Office CoPilot



AI video generator for training

## Researcher/Educator Exp.

- Increase research efficiency
- Support safe and secure investigation



Lit Review, Manuscript Drafting



Secure GPT instance

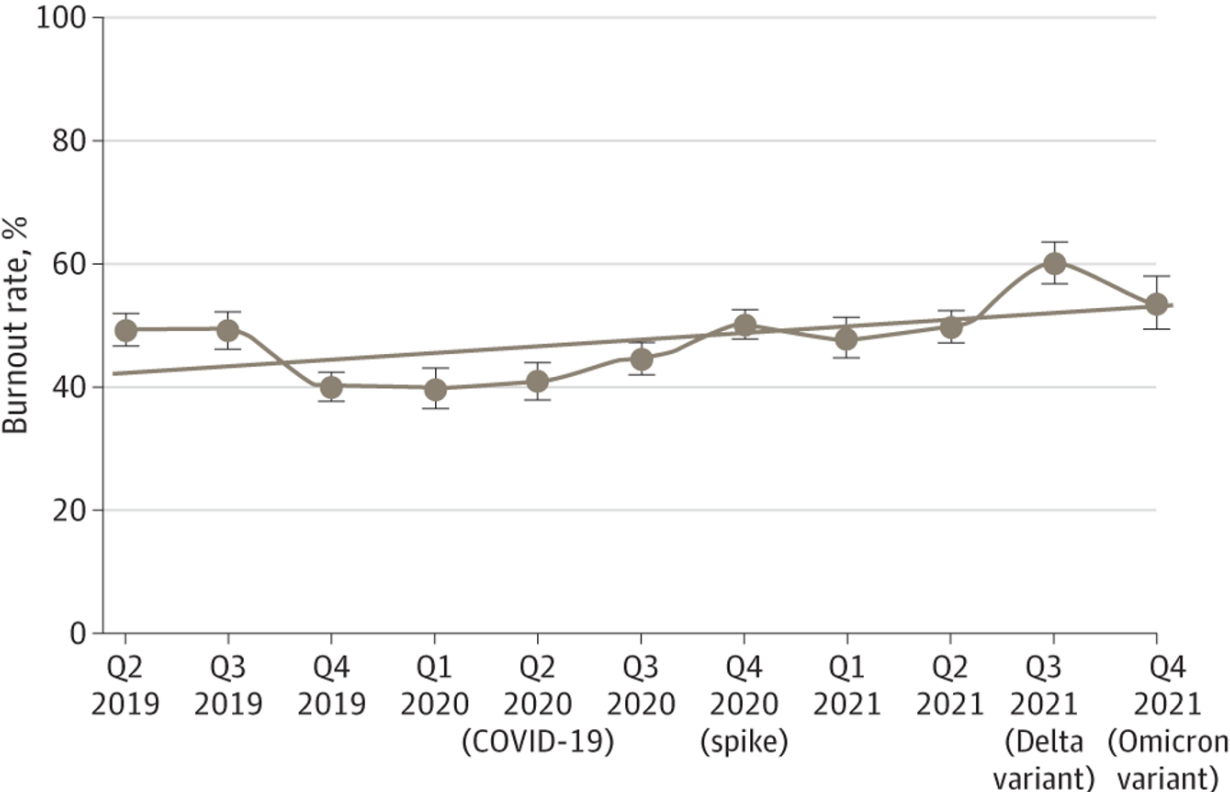


Teaching, Board Exam Questions

# Generative AI to Improve Clinician Experience



# Burnout is highly prevalent and multi-factorial



Linzer et al. JAMA Health Forum. 2022;3(11):e224163. doi:10.1001/jamahealthforum.2022.4163  
 Graphic (AHRQ): <https://www.ahrq.gov/prevention/clinician/ahrq-works/burnout/index.html>



# Consequences of Physician Burnout

## Reduces Access

- Burned out physicians more likely to leave practice
  - 67% reduction in productivity
  - 28% reduction in professional effort

## Threatens Patient Safety

- Depersonalization leads to poor interactions
- Burned out physicians also suffer from impaired attention, memory and executive physician
  - 11% increase in medical error rate

## Individual Risks

- Alcohol/substance abuse
- Broken relationships
- Poor physical and emotional QOL
- Anxiety/panic/depression /suicidal ideation
- Suicide



Physician Burnout. Content last reviewed July 2017. Agency for Healthcare Research and Quality, Rockville, MD. <http://www.ahrq.gov/professionals/clinicians-providers/ahrq-works/burnout/index.htm>  
Shanafelt TD, Balch CM, Beschamps G. et al. Burnout and medical errors among American surgeons, *Ann Surg* 2010;251(6):995-1000.  
Shanafelt TD, Goh J, Sinsky CA. The business case for investing in physician well-being. *JAMA Intern Med* 2017.



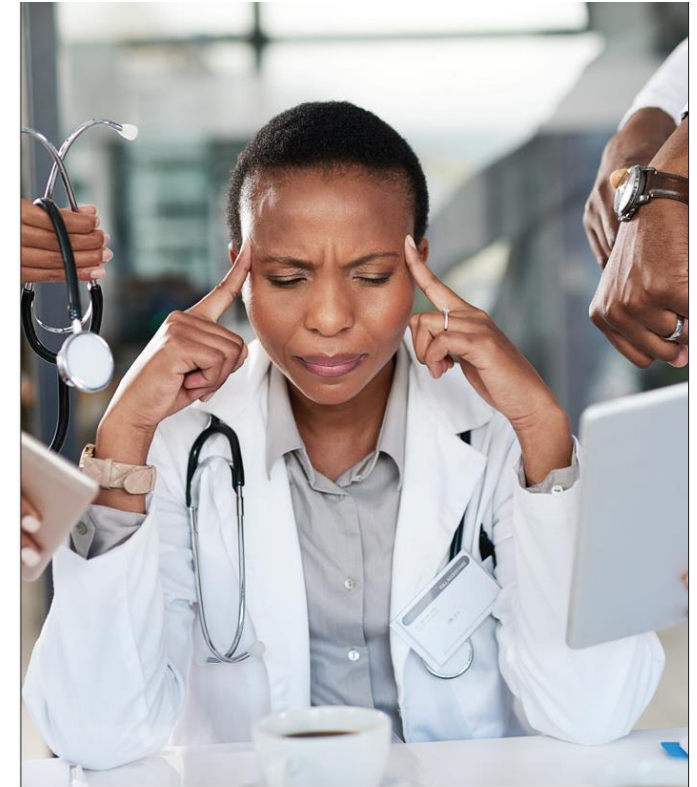
# Electronic Health Record time is Associated with Burnout

Better EHR usability = less workload → lower odds of burnout<sup>1</sup>

For every 8 hours patient scheduled time, ambulatory physicians spend ≥5 hours working in the EHR<sup>2</sup>

Physicians with ≤5 after-hours charting/week are 2x more likely to report less burnout than those with ≥6 hours<sup>3</sup>

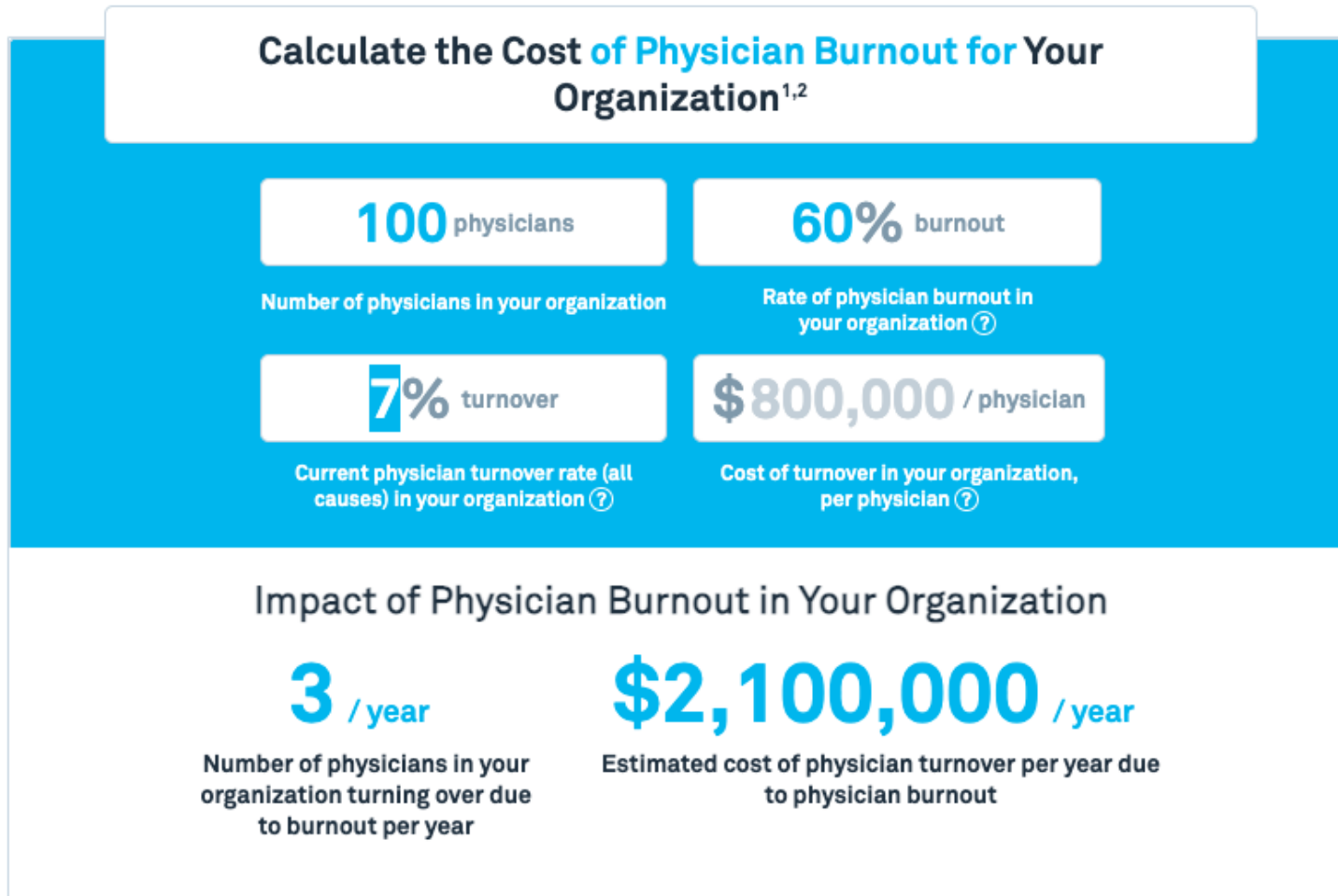
Female physicians = higher WOW<sub>8</sub>, EHR-Time<sub>8</sub> and Note-Time<sub>8</sub><sup>4</sup>



1. Melnick ER, et al. Perceived Electronic Health Record Usability as a Predictor of Task Load and Burnout Among US Physicians: Mediation Analysis. *J Med Internet Res*. 2020 Dec 22;22(12):e23382.
2. Melnick ER, et al. Characterizing physician EHR use with vendor derived data: a feasibility study and cross-sectional analysis, *Journal of the American Medical Informatics Association*, Volume 28, Issue 7, July 2021, Pages 1383–1392
3. Eschenroeder HC, et al. Associations of physician burnout with organizational electronic health record support and after-hours charting. *J Am Med Inform Assoc*. 2021 Apr 23;28(5):960-966.
4. Rotenstein LS, Fong AS, Jeffery MM, et al. Gender Differences in Time Spent on Documentation and the Electronic Health Record in a Large Ambulatory Network. *JAMA Netw Open*. 2022;5(3):e223935.

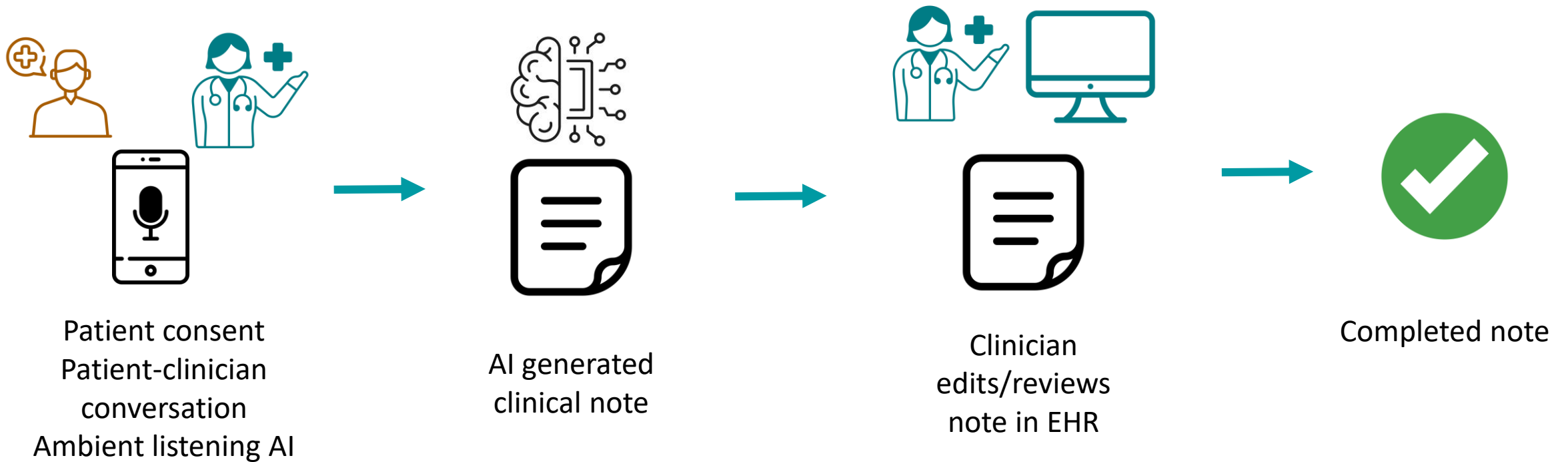


# Organizational Cost of Physician Burnout



# Ambient Documentation

- Autonomous documentation is rapidly improving due to advanced machine learning, including new generative AI/large language models (LLMs).
- Ambient clinical documentation platforms record patient visits and summarize into a draft note within minutes, reducing clinician time completing notes.





# Using AI to Summarize Patient Visits into Clinical Documentation: Initial Outcomes

## Proof of Concept with 18 Mass General Brigham physicians

- Recorded 400+ sessions
- Average time to draft note ready = 1.5 minutes
- 91% of drafted text is retained by the user
- More promising in primary care, EM
  - Subjectively less pajama time, more patient face time

## Pilot of 2 platforms

- Epic integration
- 800 clinicians from across all areas of MGB
- Assess Usability & Satisfaction as well as measures of time savings/efficiency, and ROI

Feedback: Epic integration crucial for workflow efficiency

*"When using ambient documentation during a visit, I have noticed that I am able to have more face-to-face time with my patients, rather than staring at my computer screen"*

*- Physician POC participant*



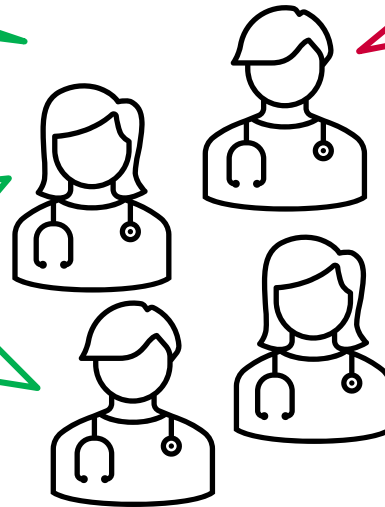
(Ambient) is **truly life changing** for me. I conduct over **50% of my visits in Spanish** and documentation has always been a struggle for me. There is a **huge weight off** my shoulders at the end of the day. - *Pediatrician*

I am in love with (ambient). Already I am **much less likely to change careers** in the next year. - *Primary Care (after 2 days of use)*

I'm realizing how much I'm doing for my patients." - *Neurologist*

I look at the patient's eyes and that is magical. I'm addicted. I want it to be a part of every encounter I have. - *Neurologist*

In short this is the **greatest change to my practice of medicine in the 26 years** I have been working. I can **never go back** and again **am enjoying my job** rather than dreading being behind on notes. - *Urologist*



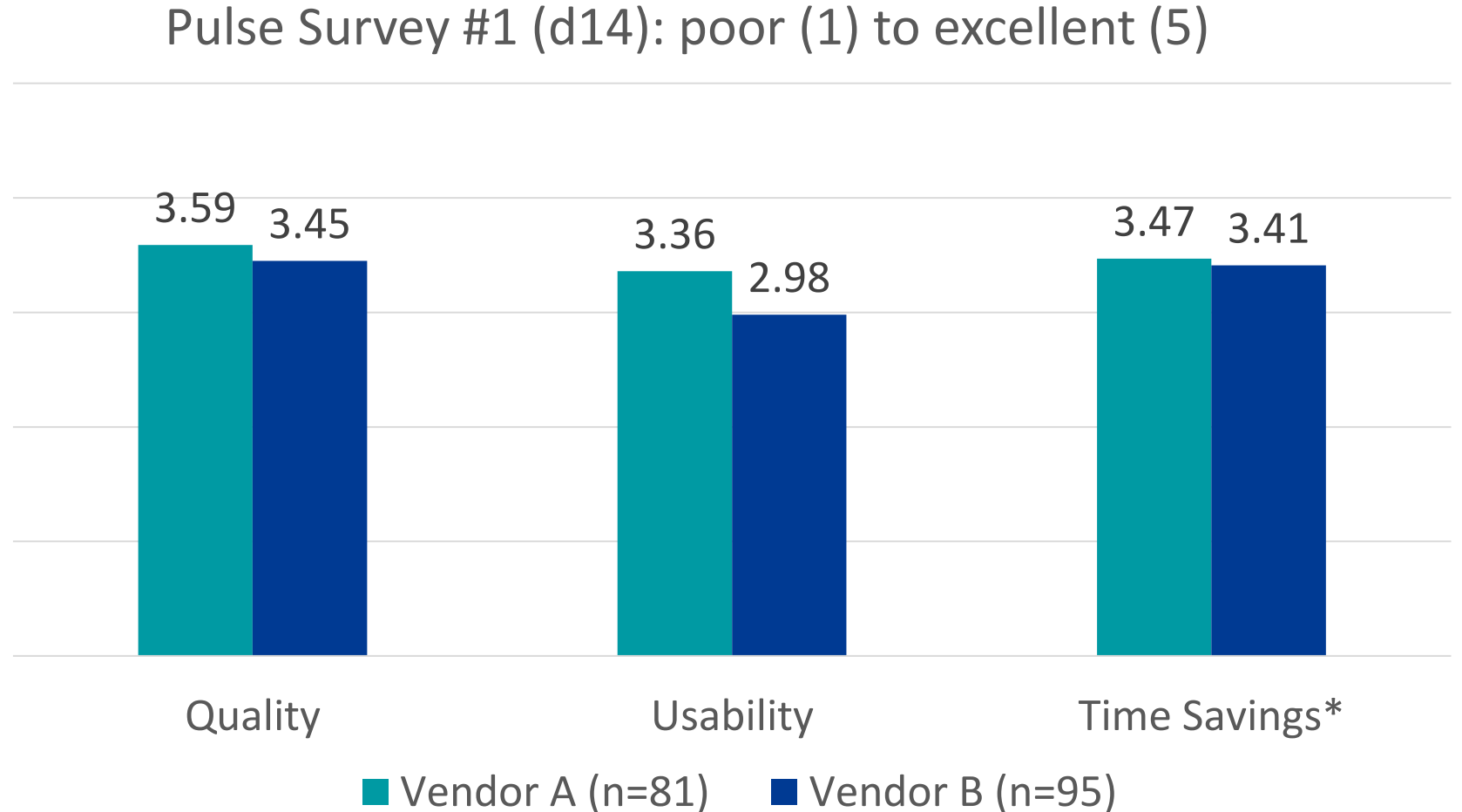
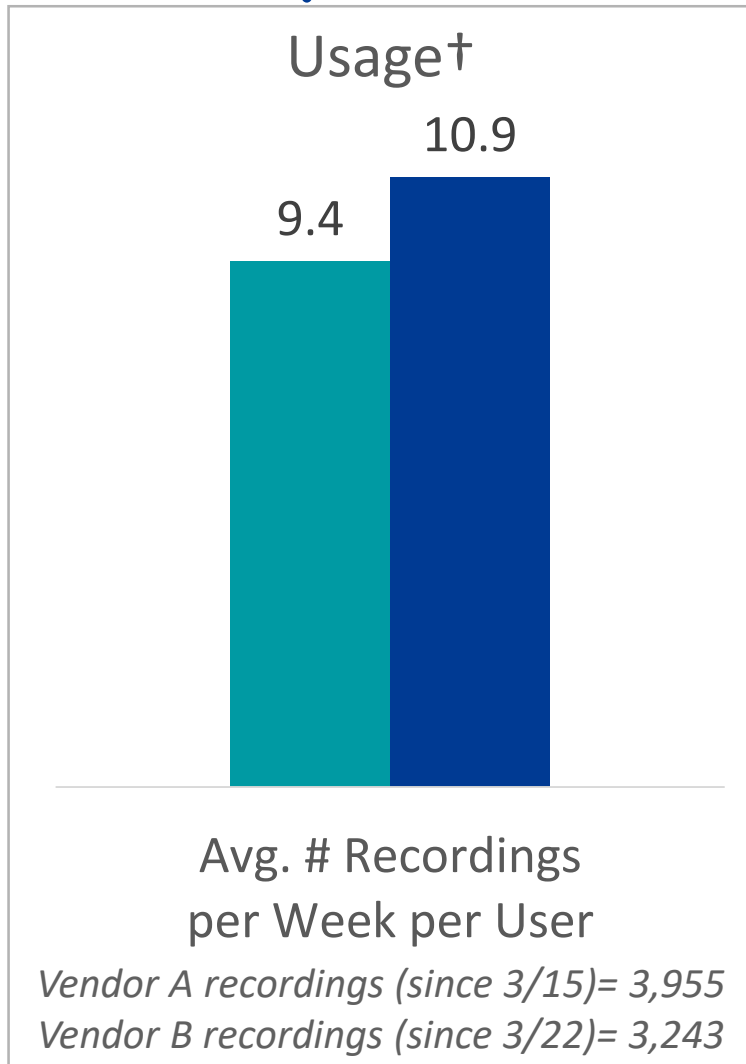
I'm still typing because what if there's a bug that it doesn't work. Then I've looked at the (ambient) note and it's perfect, so I delete everything I typed. - *Primary Care*

I spent a lot of time explaining to my patients why they weren't having an emergency...and very little of that made it into my note. I would like to see more of that in my note. - *Emergency Medicine*

(Ambient) is a **complete game changer**. The first time in 5 years can I **actually finish my notes on the day of clinic** and it allows me to focus on the patient conversation rather than on documentation. - *Orthopedic Surgeon*



# Initial feedback shows increasing usage and largely positive



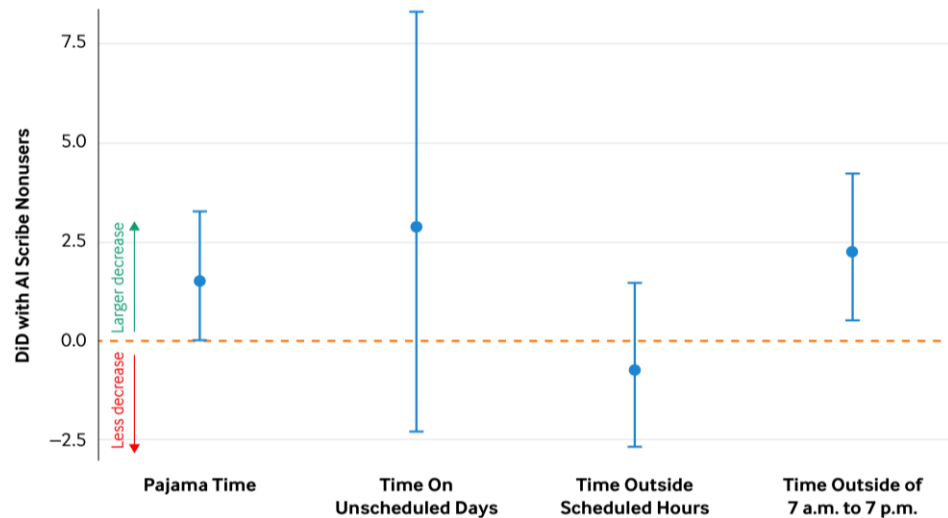
†Calculations based on vendor usage data (Vendor A n=141, Vendor B n=109)

\*Quantitative calculations based on EHR usage data will be available in June

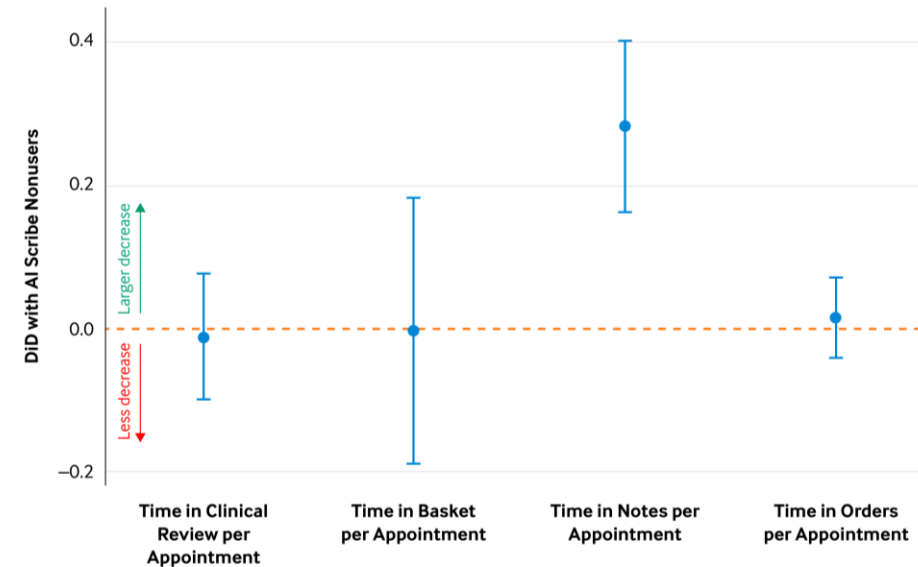
# Ambient Documentation Experience at Kaiser Permanente-Northern California

- Provided access to un-integrated ambient scribes for 10,000 users
- Evaluated after 10 weeks of pilot
- 3,442 unique users used the tool for 303,266 patient-physician encounters (10 weeks)

Panel A. Primary Care Physician Time Spent in the EHR-Related Activities



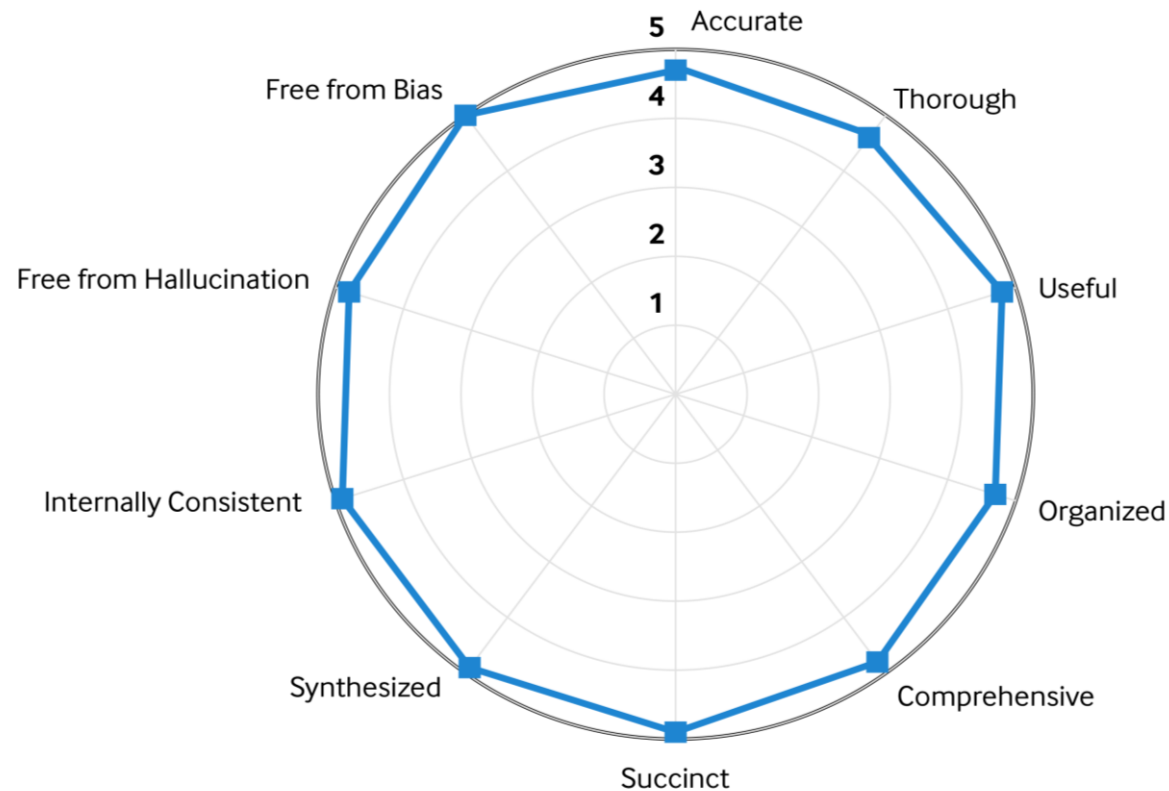
Panel B. Primary Care Physician Time Spent in Appointment-Related Activities



Tierney AA, Gayre G, Hoberman B, et al. "Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation." NEJM Catalyst, January 31, 2024. <https://catalyst.nejm.org/doi/full/10.1056/CAT.23.0404>

# Ambient Documentation Experience at Kaiser Permanente-Northern California

High levels of performance across 10 modified PDQI-9 domains on 35 randomly assessed notes across multiple clinical specialties



# In Basket Message Categorization & Routing

- Proof of concept testing ChatGPT 3+4 to categorize inbasket messages on de-identified historical messages
- Incoming messages are processed and scanned to categorize and route to best person/group for messages
  - This can lead to reduction of messages received by clinicians that are better triaged by other staff members
- Working with EHR vendor to develop this within EHR workflows to facilitate appropriate routing of incoming messages



The screenshot shows an email client window titled "Details Info, [1]". It includes action buttons for Archive, X-Parts: 1, Reply, Forward, Recipients: 1, and Call. The message details are as follows:

- Subject: Walter Smith
- Sender: Matheus Turra, Date: 2023-05-10 15:32 GMT
- Message body (JSON):

```
{
  "Id": "93d493f9-f80b-4a80-9407-babe7f4b7172",
  "Message": "Yesterday afternoon I tripped down the last step and grabbed the",
  "Recipient_Queue": "Patient Advice Questions",
  "Labels": [
    {
      "Label": "Patient Advice Questions",
      "NLP_Label": "Medical Condition Worsening",
      "Score": 0.9,
      "NLP": "OpenAI",
      "IsFinal": true,
    }
  ]
}
```


## Categories Of Responses

- Refills
- Results
- Paperwork
- General Questions




# In Basket Auto-Drafted Response

- GPT-4 driven In Basket **draft replies** to patients based on chart data.
- Drafts require approval by a care team member before the message is sent
- Considerations
  - Response prompts are customized by the organization – provide clinical recommendations?
  - Identifying appropriate user role
  - Barriers to change in workflow



 **Automatically Drafted Response** Pilot feature


Hello Barbara,

You have 5 refills remaining for your SYNTHROID TABS 0.1 MG OR. Here is how you can request a refill - In MyChart, go to Health, then Medications. Then **hover over** the medication you want to refill and **click** the Request Refill button.

 [Learn more](#) Reply options:


Did you find the drafted response helpful?

 This was helpful  This was not helpful

 **Med refill** (Newest Message First)

[View All Conversations on this Encounter](#)

**Barbara Cutler** → P Triage Nurses (supporting Walker, Drew, MD) 1 minute ago (11:45 AM)

 Dr Walker, I am running out of my thyroid meds, and would like a refill



# Addressing in basket burden: draft replies to patient portal messages

MGB, Stanford, UCSD

Low utilization: 20% of drafts, at best

Variable likelihood to recommend by role & specialty



Improvements in cognitive task load, work exhaustion



No change in time in messages, quality, workflow experience



Increase in length of messages



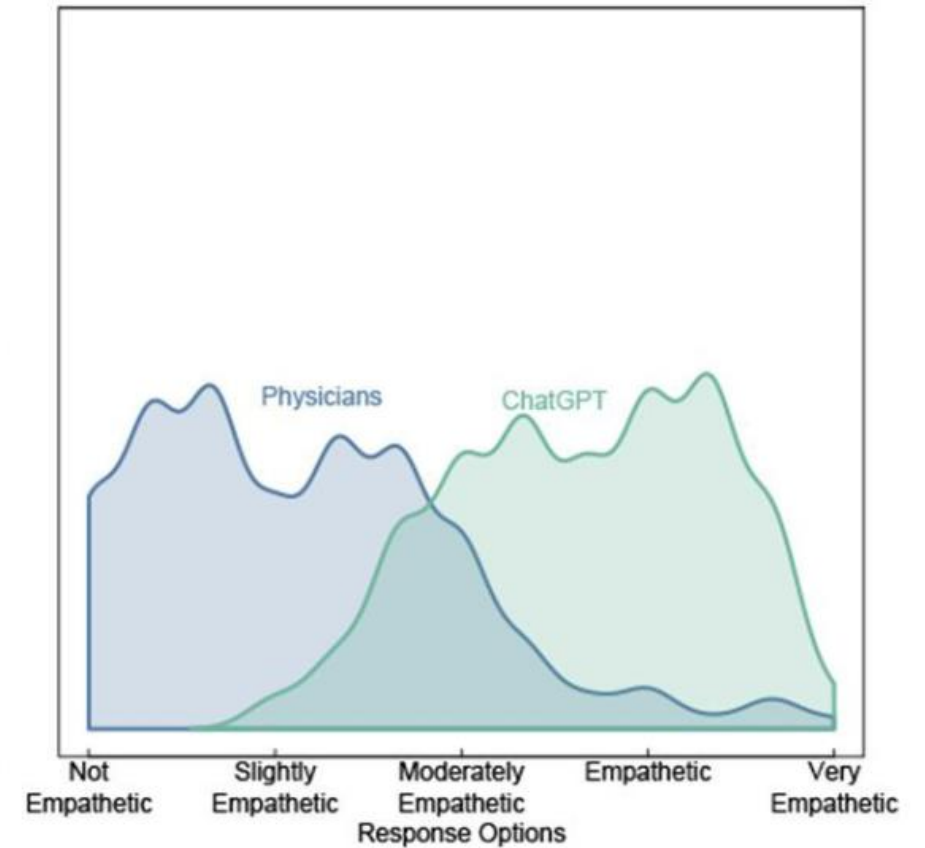
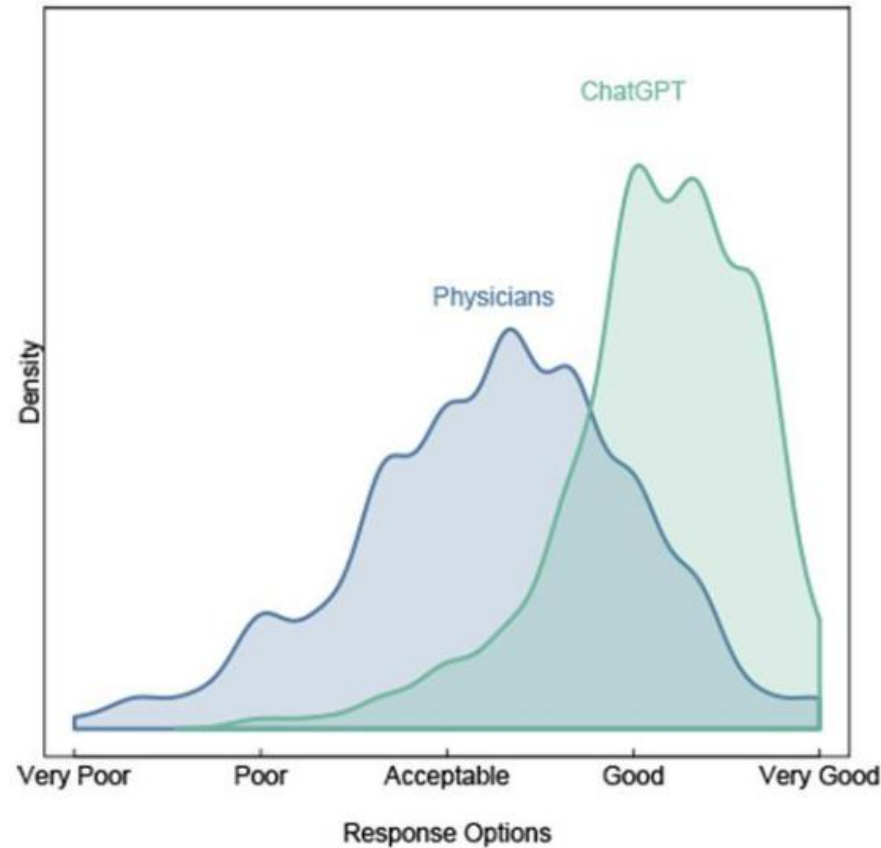


# ChatGPT is more empathetic than physicians?

ChatGPT responses to patient\* messages were more likely to be:

- Higher quality
- More empathetic

\*posted publicly on Reddit



# Patient Chart Summarization

Notes This Visit Summarize Notes

### Generate Summary From Notes

Lookback Time (Days)  
0 Since your last visit 2 days 15 months 3 years

Length  
Very Short Short Medium Long Very Long  Write bullet points

Written For  
You Patient Nurse Surgeon

Focus On  
Topic to focus the summary on

Generate Summary Leave Feedback

#### Auto-Generated Summary

Generated at: 6/29/2023 3:17 PM.  
The patient presents with a history of chronic kidney disease (CKD) and has undergone multiple interventions, including left and right renal artery angioplasty and stent placement, to improve blood flow to the kidneys [4], [8]. Despite these interventions, the patient continues to experience occasional fatigue and mild edema [9]. Throughout the treatment process, the patient has been notably preoccupied with the upcoming release of a new video game, which may impact his adherence to treatment and self-care recommendations [6], [9], [10]. The importance of medication compliance, dietary modifications, and adequate rest has been reinforced, and the patient has been encouraged to find a balance between his interests and prioritizing his health [9], [10].

References

Input Information

- Allows clinician to quickly review medical developments for a patient they haven't seen recently or have never seen before
- LLM lets clinician select timeframe, summary length, audience, and focus to create a customized summary based on the 20 most recent notes in the specified time
- Use Cases:
  - “Since your last visit”
  - Discharge summary
  - Patient summary



# Trial informed approach to implementing AI



## Phase I: Safety

Evaluate safety

Design workflows



## Phase II: Efficacy

Refine workflows

Assess impact:

- Quality
- Efficiency
- Financial



## Phase III: Effectiveness

Scale

Compare to standards



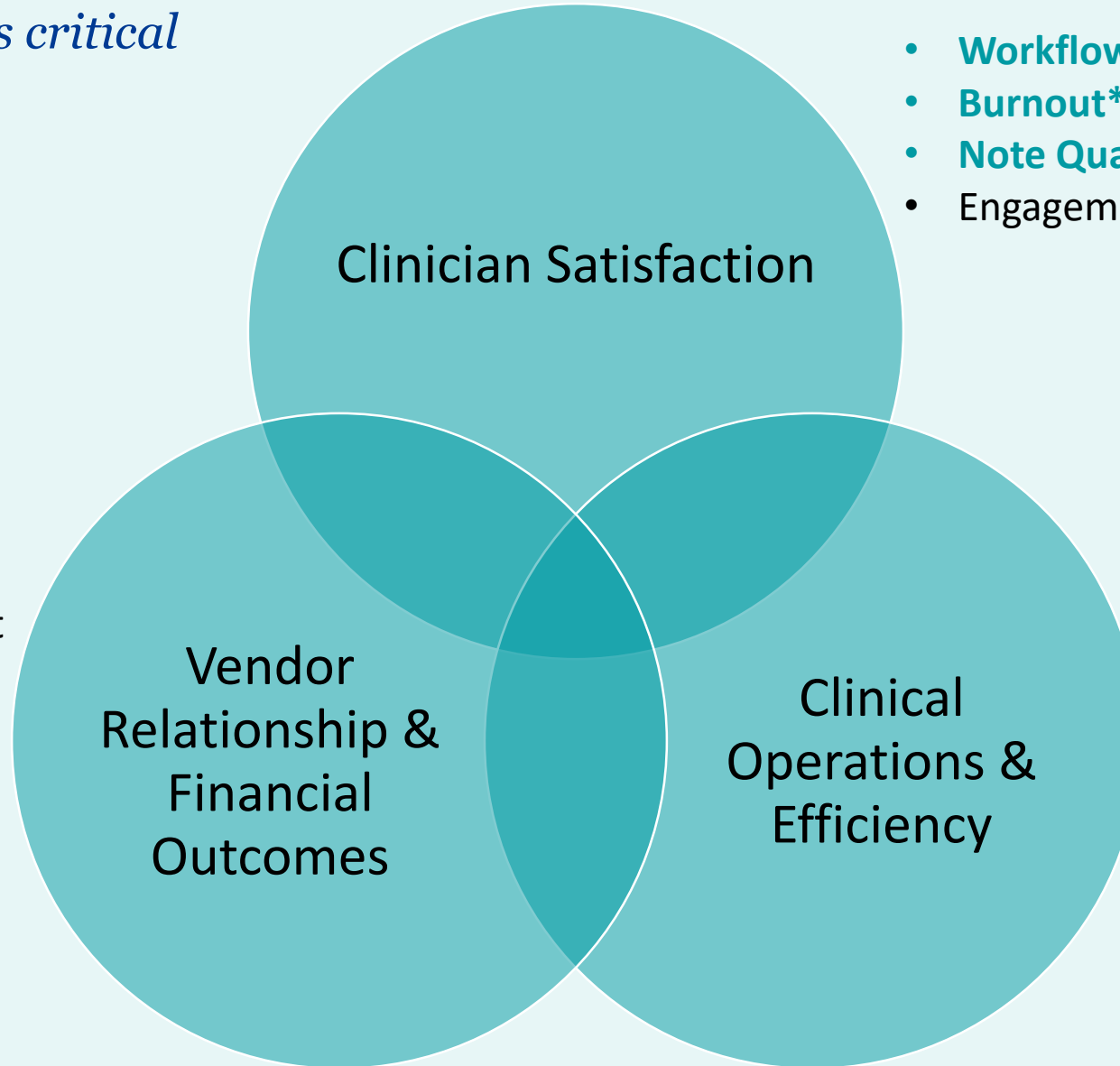
## Phase IV: Monitor

Monitor safety,  
workflow, impact

Ongoing technology  
evolution

# Outcome Assessment Plan

*Participant feedback is critical*



- **Workflow Satisfaction\***
- **Burnout\***
- **Note Quality, Accuracy\***
- Engagement/Dropoff rate

- Responsiveness
- Training/Onboarding cost
- Feature Assessment
- Product Cost
- Coding/Reimbursement Enhancement

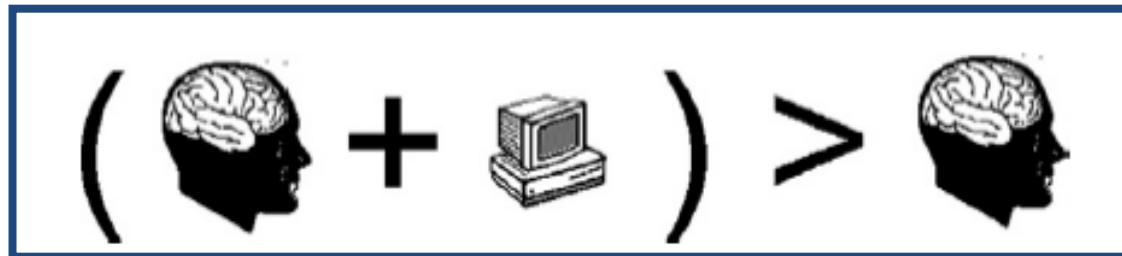
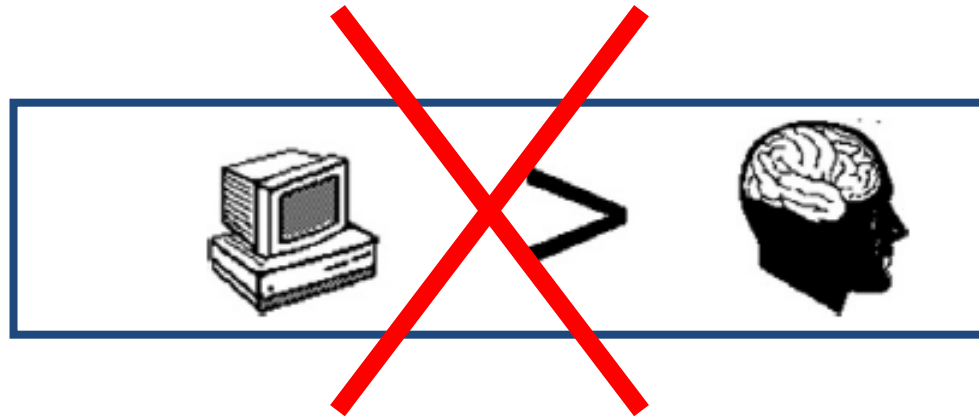
- **Time Savings\***
- Note Completion Time
- In Basket Time
- Time in Epic
- **Usability\***
- **Safety Events\***



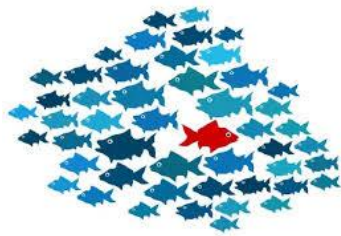
# Scaling Gen AI



# Fundamental Theorem of Informatics



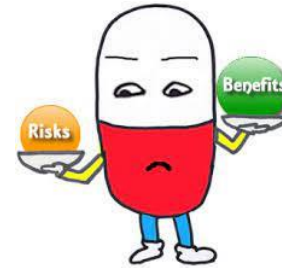
# What is needed to Advance Health AI?



**Reliability**



**Safety**



**Benefit/ROI**



**FDA Approval**



**Bias**



**Privacy/Security**



**Transparency**



**Informed Consent**

# Guiding principles

- Meet an identified business or clinical need
- Responsible use of AI framework
- MVP pilots
- Demonstrate ROI (tangible and intangible)

Characteristics of Responsible Use of AI	Sub-areas
<b>Fairness</b>	Patient-centered Equitable
<b>Transparent and Explainable</b>	Documentation of data and development Performance metrics / confidence intervals Patient education
<b>Responsible and Accountable</b>	Responsibility across model lifecycle AI governance structure ROI
<b>Robust and Reliable</b>	Model performance across shifts in data Performance monitoring and thresholds
<b>Privacy</b>	De-identified data used for model training Access to output Role of Informed consent and IRB
<b>Safety and Security</b>	User interaction Education Feedback loops / AE reporting Cybersecurity
<b>Benefit</b>	Patient outcomes and satisfaction Clinician and staff wellness Financial ROI



# Assistive vs. Autonomous AI Algorithms



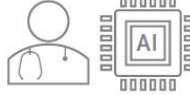


	Assistive AI algorithms		Autonomous AI algorithms		
	Level 1	Level 2	Level 3	Level 4	Level 5
	 Data presentation	 Clinical decision-support	 Conditional automation	 High automation	 Full automation
Event monitoring	AI	AI	AI	AI	AI
Response execution	Clinician	Clinician and AI	AI	AI	AI
Fallback	Not applicable	Clinician	AI, with a backup clinician available at AI request	AI	AI
Domain, system, and population specificity	Low	Low	Low	Low	High
Liability	Clinician	Clinician	Case dependent	AI developer	AI developer
Example	AI analyses mammogram and highlights high-risk regions	AI analyses mammogram and provides risk score that is interpreted by clinician	AI analyses mammogram and makes recommendation for biopsy, with a clinician always available as backup	AI analyses mammogram and makes biopsy recommendation, without a clinician available as backup	Same as level 4, but intended for use in all populations and systems

Figure: Levels of automation of medical artificial intelligence systems



Bitterman, Danielle S., Hugo JWL Aerts, and Raymond H. Mak. "Approaching autonomy in medical artificial intelligence." *The Lancet Digital Health* 2.9 (2020): e447-e449.

Credit: Agustina Saenz

# Potential Liability for Physicians using AI

Scenario	AI recommendation	AI accuracy	Physician action	Patient outcome	Legal outcome (probable)
1	Standard of care	Correct	Follows	Good	No injury and no liability
2			Rejects	Bad	Injury and liability
3		Incorrect (standard of care is incorrect)	Follows	Bad	Injury but no liability
4			Rejects	Good	No injury and no liability
5	Nonstandard care	Correct (standard of care is incorrect)	Follows	Good	No injury and no liability
6			Rejects	Bad	<i>Injury but no liability</i>
7		Incorrect	Follows	Bad	<i>Injury and liability</i>
8			Rejects	Good	No injury and no liability

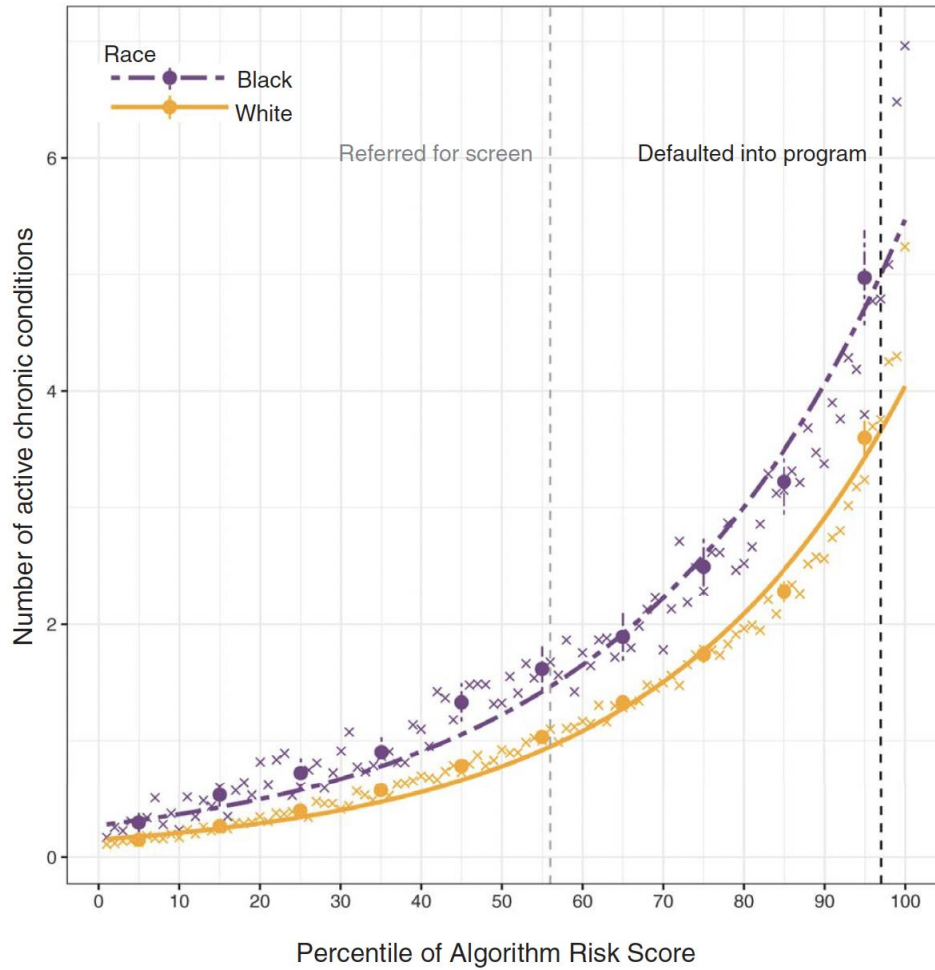
1. If there is no injury, there is no liability
2. If the standard of care is followed (even with bad outcome), there is generally no liability
- 3. If standard of care not followed and there is injury, then there is liability**

Note: Standard of care may change over time (and may even include AI in the future)

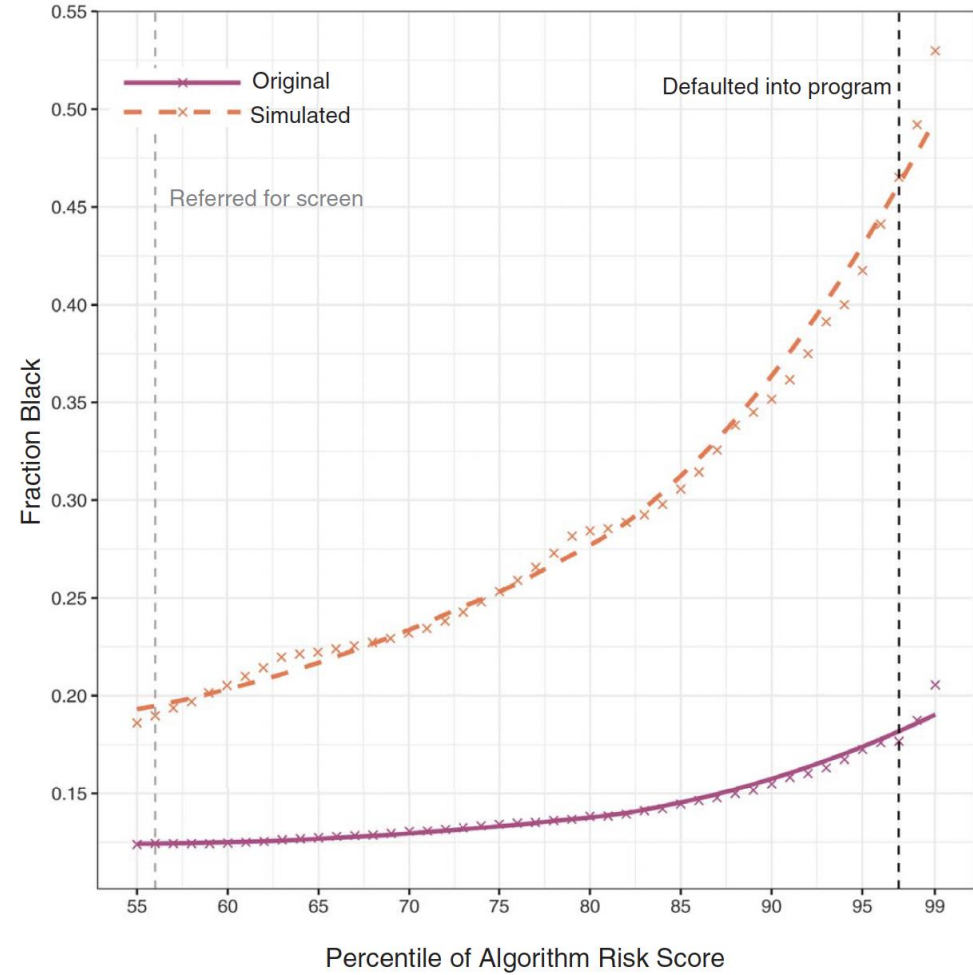


# Bias in AI

A



B



# Intellectual Property

The New York Times

---

## ***The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work***

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

- ChatGPT generates answers based on NYT articles when asked questions about current events or noteworthy topics
- Lawsuit cites examples where ChatGPT provided near-verbatim excerpts from NYT articles (behind paywall)



# White House Commitment to the Safe, Secure, and Trustworthy Use and Purchase of AI in Healthcare

## Our Commitments

1. We commit to vigorously developing AI solutions to optimize healthcare delivery and payment by advancing health equity, expanding access, making healthcare more affordable, improving outcomes through more coordinated care, improving patient experience, and reducing clinician burnout.
2. We will work with our peers and partners to ensure outcomes are aligned with fair, appropriate, valid, effective, and safe (FAVES) AI principles.
3. We will deploy trust mechanisms that inform users if content is largely AI-generated and not reviewed or edited by a human.
4. We will adhere to a risk management framework that includes comprehensive tracking of applications powered by frontier models and an accounting for potential harms and steps to mitigate them.
5. We will research, investigate, and develop swiftly but will do so responsibly.



# Future of Medicine with Generative AI





# Healthcare is at a Crossroads



**FINANCIAL PRESSURE**



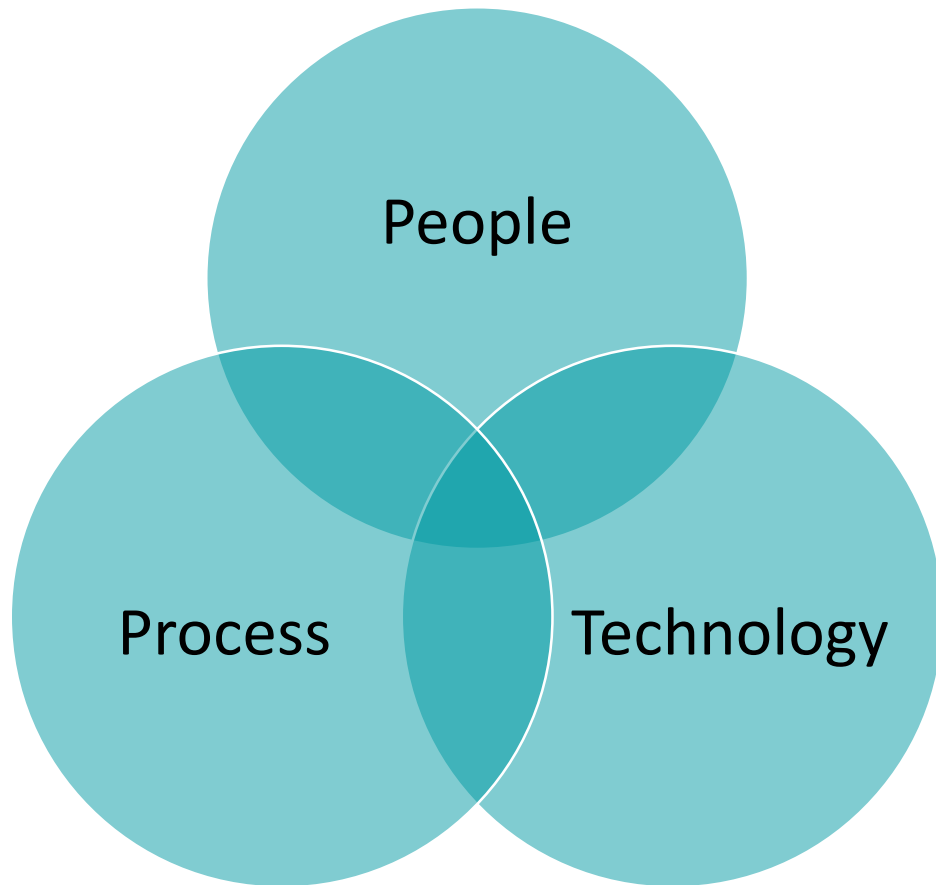
**WORKFORCE SHORTAGE**



**INCREASED COMPETITION INCLUDING NEW ENTRANTS**



# AI is Increasingly Important for Healthcare Delivery



In addition to AI (technology), we need to consider people and workflow

Automate repetitive, routine, and low risk tasks (e.g., registration, scheduling)

Use AI to augment humans to improve the quality or speed of manual tasks (e.g., prior authorization, documentation)

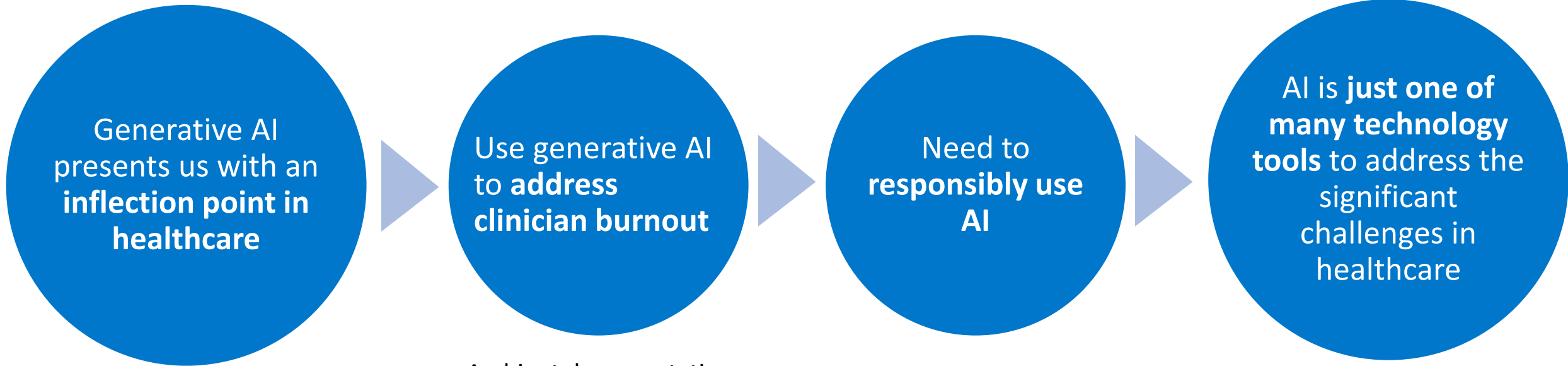
Upskill staff – especially those whose work has been automated or augmented by AI

Continue care delivery innovations (virtual care, chat bots for low acuity conditions, home hospital)





# Summary



Potential for dramatic change: interactions with existing technology/data, patient experience, high quality equitable care delivery, workforce challenges

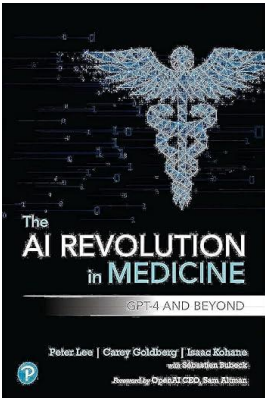
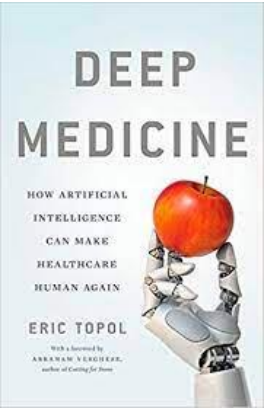
- Ambient documentation: promising early results, need to understand who benefits & how
- Other LLM use cases are in pilot and planned for future to address administrative burden, patient experience/engagement, clinical decision making, clinical deserts

Ensure see the benefits from this technology in a safe, secure, effective way

Technology, people, and process holds great potential to automate repetitive tasks, improve the quality and efficiency of manual tasks, upskill staff, and continue care delivery innovation



# Learn More



# Special Thanks to Team Members



**Matthew Butler**



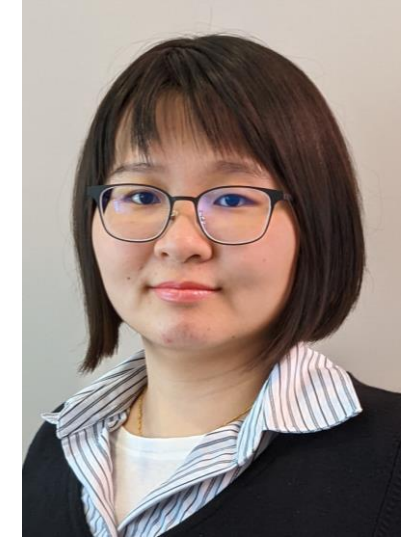
**Amanda Centi, PhD**



**Adam Landman, MD**



**David Y. Ting, MD**



**Jacqueline You, MD**

# Questions

*“Using technology to deliver better care and deliver care better”*

**Rebecca G. Mishuris, MD, MS, MPH, FAMIA**

[rmishuris@mgb.org](mailto:rmishuris@mgb.org)

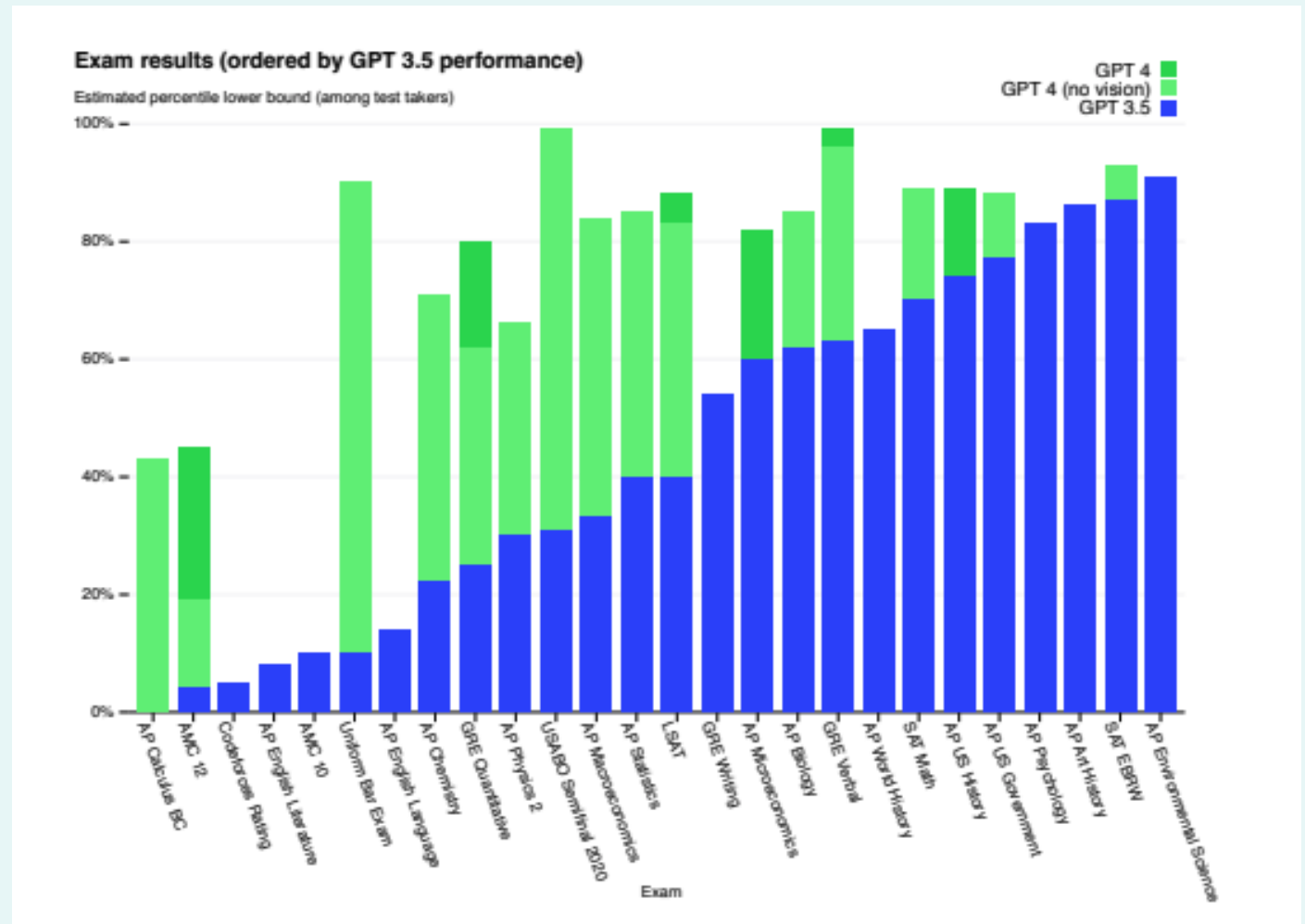




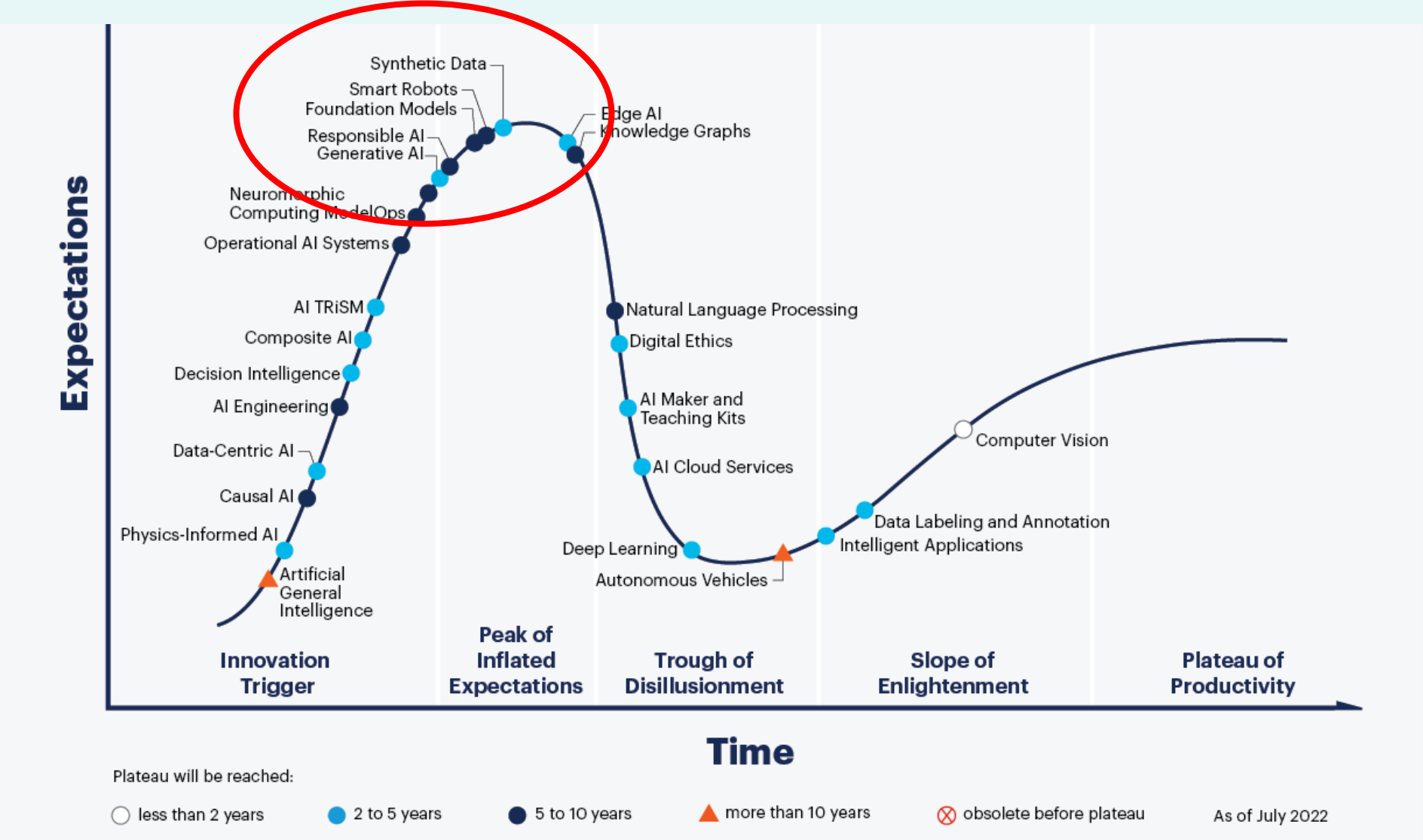
**Mass General Brigham**

# GPT-4 Multimodal Model with Improved Performance

- GPT4 is a large multimodal model that can process image and text inputs and produce text outputs
- Improved ability to understand and generate natural language text, in more complex and nuanced scenarios
- Improved performance on simulated bar exam (top 10% vs GPT-3.5 in bottom 10%)
- Similar limitations to prior GPT models



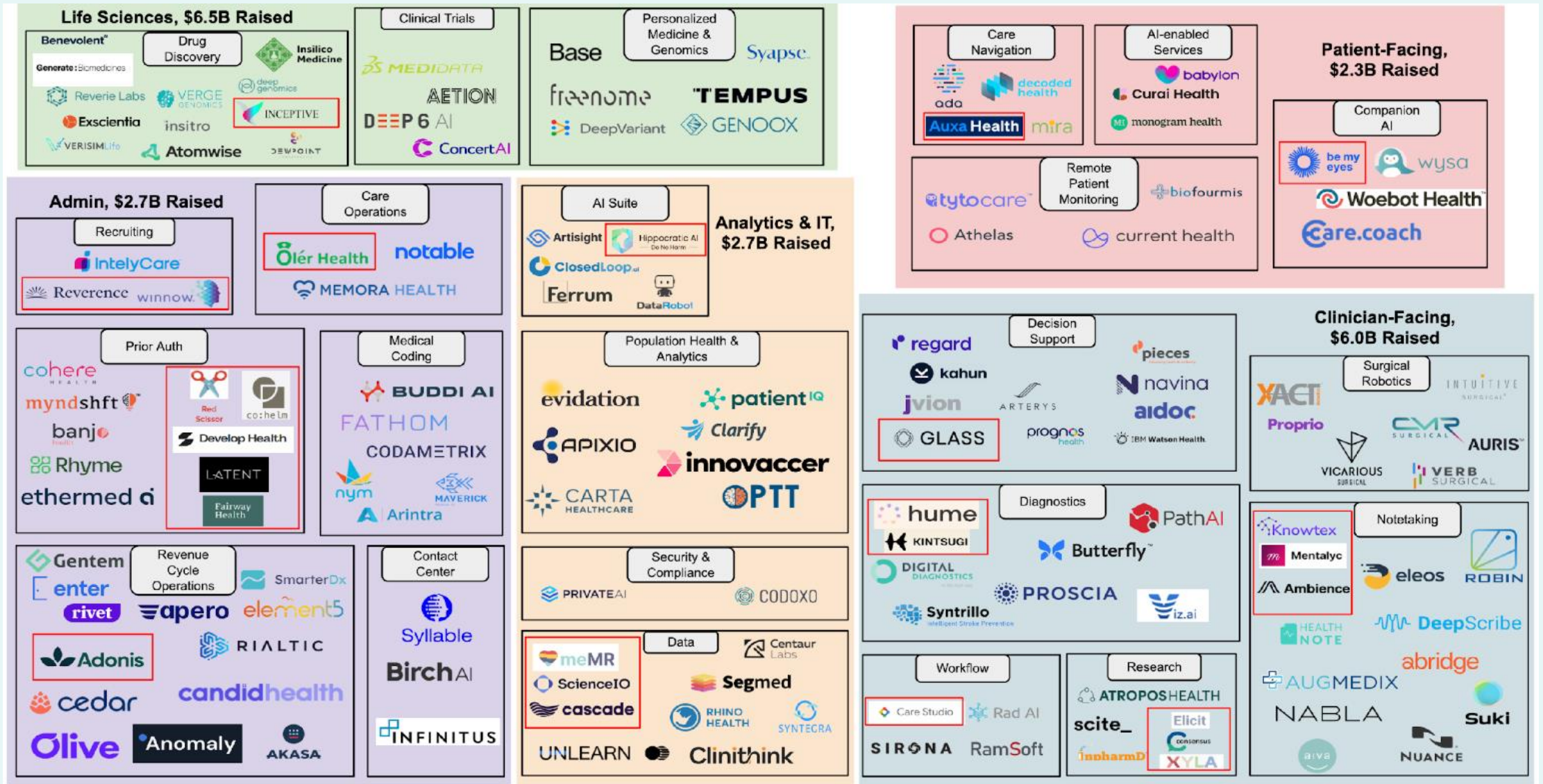
# Gen AI is largely at the Gartner Peak of Inflated Expectations



Source: 2022 Hype Cycle of AI, Gartner



# The Healthcare AI Landscape



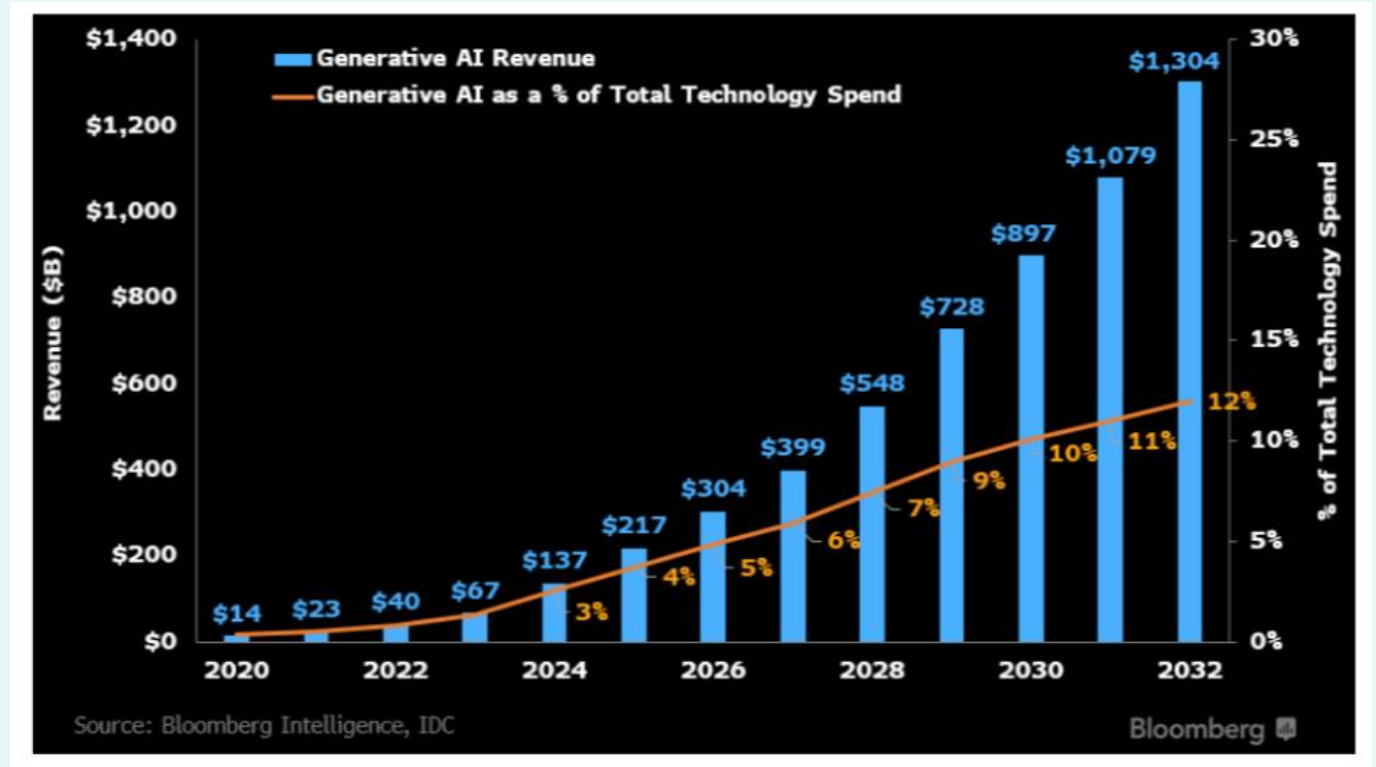


# The Opportunity for Generative AI in Healthcare

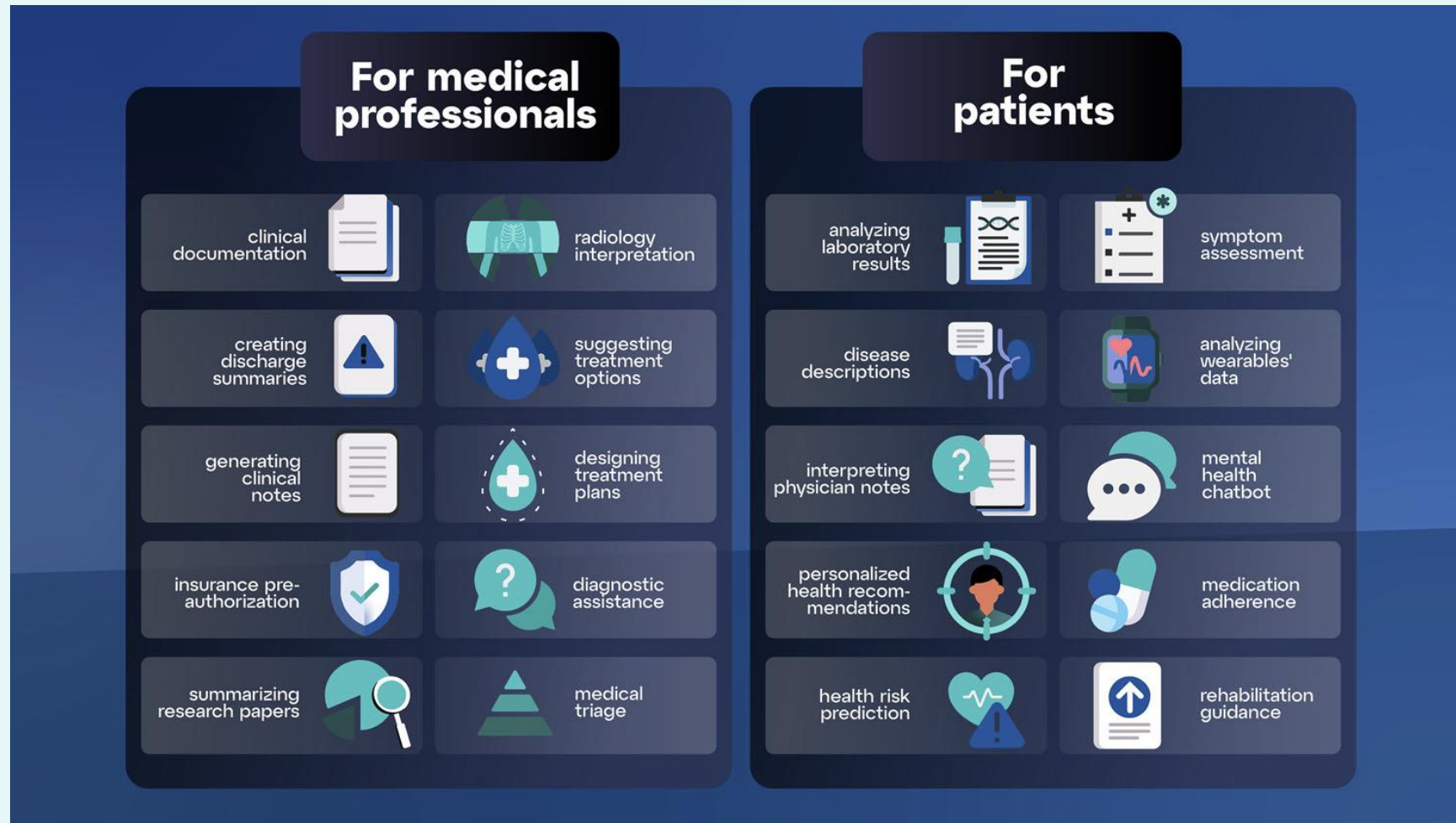
- Market **may** hit \$800-900 billion by 2030
- Healthcare use-cases are expected to make up 25% of the generative AI market by 2023

Top projected medical cases:

- healthcare analytics
- documentation
- customer interaction
- medical simulation/education
- decision support
- clinical trials
- drug development



# Opportunities for LLMs in Healthcare



# Generative AI to Improve Research and Education



# Common Research and Education Workflows

Literature Reviews

Writing Manuscripts

Promotions

- Preparing CVs
- Writing Letters of Recommendations

Gen AI  
Research/Evaluation

Teaching

Board Examinations



# Using AI to Search and Synthesize Scientific Literature

- Challenge: Researchers spend hours searching and digging through primary literature
- System platform ingests articles from PubMed every day and leverages AI to synthesize and contextualize scientific literature
  - Stored in graph database
  - 14M findings, 4.5M variables on 5M studies

STUDY	JOURNAL	AUTHORS	EXTRACTED FINDINGS	STUDY POPULATION	CITED BY	PUBLISH DATE
Cardiovascular Outcomes with Ertugliflozin in Type 2 Diabetes	The New England Journal of Medicine	Christopher P. Cannon +14 more	Ertugliflozin is associated with no change in the hazard of death from cardiovascular causes or hospitalization for heart failure	Adults with type 2 diabetes and atherosclerotic cardiovascular disease	752	2020
FINDING		STATISTICS		PLOT		<a href="#">Suggest revision</a>
Ertugliflozin is associated with no change in the hazard of a major adverse cardiovascular event		Hazard Ratio <b>0.970</b> p-value < 0.001 95% CI (0.850 to 1.110)				
FINDING		STATISTICS		PLOT		<a href="#">Suggest revision</a>
Ertugliflozin is associated with no change in the hazard of death from cardiovascular causes or hospitalization for heart failure		Hazard Ratio <b>0.880</b> p-value = 0.110 95% CI (0.750 to 1.030)				
FINDING		STATISTICS		PLOT		<a href="#">Suggest revision</a>
Ertugliflozin is associated with no change in the hazard of death from cardiovascular causes		Hazard Ratio <b>0.920</b> p-value: N/A 95% CI (0.770 to 1.110)				
FINDING		STATISTICS		PLOT		<a href="#">Suggest revision</a>
Ertugliflozin is associated with no change in the hazard of death from renal causes, renal replacement therapy, or doubling of the serum creatinine level		Hazard Ratio <b>0.810</b> p-value: N/A 95% CI (0.630 to 1.040)				

Improved Search

Export Structured Data

Get Synthesis of Research



# Assistance with Manuscript Writing

Controversial due to concerns about authorship, originality, factual inaccuracies, and “hallucinations”

*Science* prohibits use of AI generated text or figures

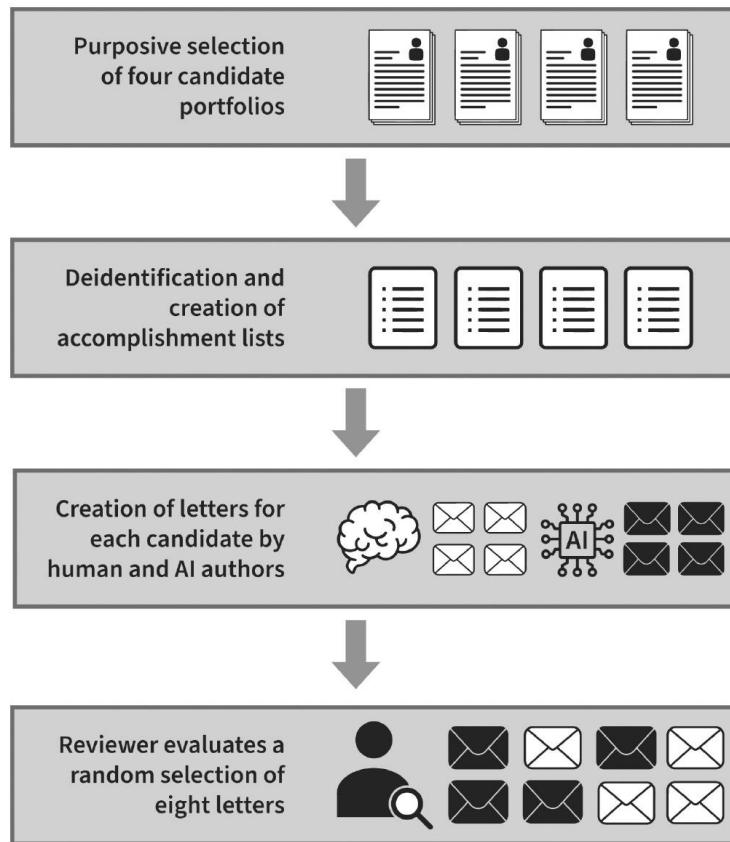
## *NEJM AI* Supports and Encourages Use of LLMs

- Potential benefits: produce better output, help non-English native speakers
- Ban is likely unenforceable
- Use must be acknowledged (like other tools used)
- Authors accountable for accuracy of work



# Assistance with Writing Letters of Recommendation

“AI may be able to partially or completely generate LORs for promotion and relieve the burden of letter writing for academicians”



Author	Candidate	Percent gender-biased language	GPTZero		OpenAI classification	Reviewer classification	
			Classification	"Perplexity"			"Burstiness"
AI 1	Dr. Chat GPT1	2% male	Human	195.529	255.381	Very unlikely AI	46.7% AI
AI 1	Dr. Chat GPT2	20% male	Human	148.611	130.822	Very unlikely AI	43.8% AI
AI 1	Dr. Chat GPT3	5% male	Human	144.647	149.835	Unlikely AI	52.6% AI
AI 1	Dr. Chat GPT4	9% male	Human	180.529	260.337	Unlikely AI	43.8% AI
AI 2	Dr. Chat GPT1	38% female	Human	295.56	530.609	Unlikely AI	60% AI
AI 2	Dr. Chat GPT2	16% female	Human	319.68	534.872	Unlikely AI	64.3% AI
AI 2	Dr. Chat GPT3	10% female	Human	249.7	462.411	Unlikely AI	62.5% AI
AI 2	Dr. Chat GPT4	4% male	Human	244.708	511.034	Unlikely AI	50.0% AI
Human 1	Dr. Chat GPT1	33% female	Human	283.688	467.376	Very unlikely AI	44.4% AI
Human 1	Dr. Chat GPT2	3% female	Human	297.343	456.176	Very unlikely AI	40.0% AI
Human 1	Dr. Chat GPT3	19% female	Human	314.303	485.842	Very unlikely AI	26.7% AI
Human 1	Dr. Chat GPT4	14% female	Human	319.294	485.003	Very unlikely AI	31.3% AI
Human 2	Dr. Chat GPT1	33% female	Human	649.5	2480.154	Very unlikely AI	29.4% AI
Human 2	Dr. Chat GPT2	27% female	Human	750.651	2458.046	Very unlikely AI	38.9% AI
Human 2	Dr. Chat GPT3	32% female	Human	779	2515.031	Very unlikely AI	44.4% AI
Human 2	Dr. Chat GPT4	38% female	Human	743.214	2426.486	Very unlikely AI	37.5% AI

Abbreviations: AI, artificial intelligence; LOR, letter of recommendation; P&T, promotion and tenure.





# LLMs have strong performance on Board Exams

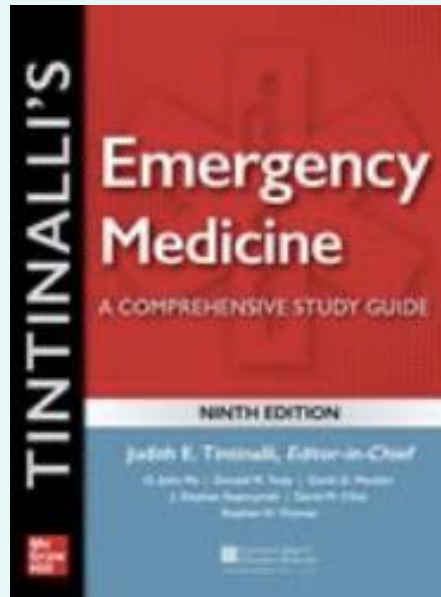
Exam	GPT 3.5	GPT 4	Notes
Neurology Board question bank	66.8%	85%	Human (mean score): 73.8%
Ophthalmology Basic and Clinical Science Course Self-Assessment	58.8%	75.8%	
Clinical Informatics Board Review questions	74%		
USMLE Step 3	53.41%	83.52%	
Emergency Medicine	???	???	

- Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. March 20, 2023. [https://www.microsoft.com/en-us/research/uploads/prod/2023/03/GPT-4\\_medical\\_benchmarks.pdf](https://www.microsoft.com/en-us/research/uploads/prod/2023/03/GPT-4_medical_benchmarks.pdf)
- Antaki F, Milad D, Chia MA, Giguère CÉ, Touma S, El-Khoury J, Keane PA, Duval R. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol. 2023 Nov 3:bjo-2023-324438. doi: 10.1136/bjo-2023-324438. Epub ahead of print. PMID: 37923374.
- Schubert MC, Wick W, Venkataramani V. Performance of Large Language Models on a Neurology Board-Style Examination. JAMA Netw Open. 2023 Dec 1;6(12):e2346721. doi: 10.1001/jamanetworkopen.2023.46721. PMID: 38060223; PMCID: PMC10704278.
- Kumah-Crystal Y, Mankowitz S, Embi P, Lehmann CU. ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? J Am Med Inform Assoc. 2023 Aug 18;30(9):1558-1560. doi: 10.1093/jamia/ocad104. PMID: 37335851; PMCID: PMC10436139.





# Can LLMs generate board questions?



Chapter 71A: Acute Abdominal Pain



Clinical Informatics  
Obesity Medicine  
Occupational & Environmental

Question: Write a multiple choice question testing readers knowledge of a single topic in this document? The output should include the question, 4 mutually exclusive answer options, and the explained answer to the question.

Multiple Choice Question: What is the most common surgical cause of abdominal pain in elderly patients?

- A. Appendicitis
- B. Cholecystitis
- C. Bowel obstruction
- D. Perforated viscus

Answer: B. Cholecystitis Explanation: According to the document, cholecystitis is the most common surgical cause of abdominal pain in elderly patients. ①

# Enabling Research with AI Capabilities with Data Infrastructure

Provisioned Azure Cloud APIs for ChatGPT 4 and GPT4-V

ChatGPT 4 Research Initiatives:

- Chart summarization to find patients with statin intolerance
- Paper identification for variant pathogenicity in support of genomics research
- LLMs for clinician decision support



# Microsoft CoPilot - Changing How We Work



- Combines the power of large language models (ChatGPT) with data from calendar, emails, chats, documents, meetings with the Microsoft 365 apps

- Integrated directly into Microsoft 365 apps (Word, PowerPoint, Excel, Outlook, Teams)

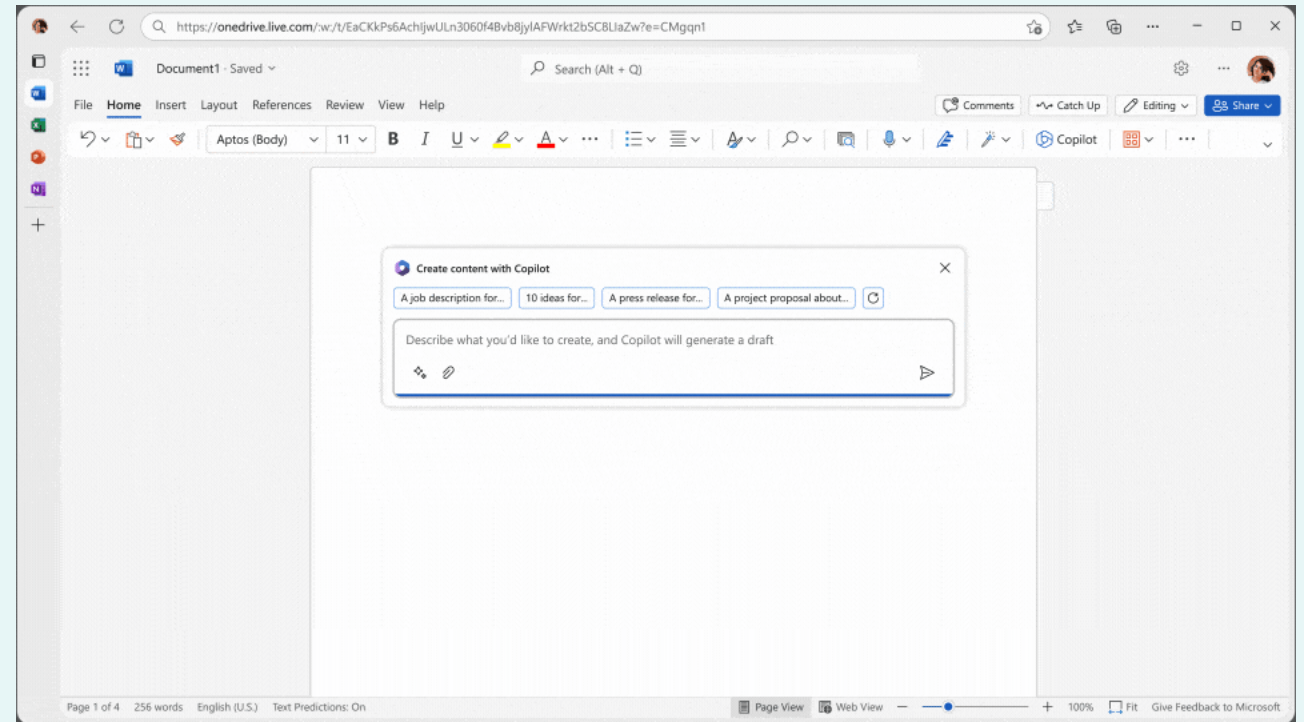
## Some examples may include:

**Word:** Draft a two-page project proposal based on data from [a document] and [a spreadsheet]

**Excel:** Give a breakdown of the sales by type and channel. Insert a table.

**PowerPoint:** Consolidate this presentation to 3 slides

**Outlook:** Summarize the emails I missed while out last week. Flag any important items.



# Clinical Informaticists Lead Health Care Digital Transformation

Informaticists analyze, design, implement and evaluate information and communication systems that enhance individual and population health outcomes, improve patient care, and strengthen the clinician-patient relationship

Clinical Informatics is an American Board of Medical Specialties approved subspecialty Practice Pathway (open through 2025)

- 3 years clinical informatics practice (25%+ FTE)

Fellowship Pathway

- ACGME-accredited fellowship of at least 24 months



# Key Considerations for Digital Health Tool Selection

Key Consideration	Description
Product Selection	Was the tool developed internally, by a third-party, by the health system's existing EHR vendor, or through a private sector partnership?
Financial Value	What framework does the tool leverage to generate financial value, and does it outweigh the costs associated with deployment and maintenance?
Clinical Value	Is there a clear, meaningful outcome metric consistent with the Quadruple Aim that would be improved by adopting this tool?
Data Assets	Does the health system have access to the data necessary for tool functionality? Is the tool interoperable with the health system's IT infrastructure? Is training data needed to tune the tool to the local environment? Has data governance been established?
Internal Champion	Is there an advocate within the health system with the leverage and motivation to facilitate adoption and implementation?
Executive Sponsors	Is there a senior executive-level advocate who will help support and pay for the tool's adoption and deployment?
Institutional Priorities	Does the tool align with institutional goals or aid in regulatory compliance?
Implementation	What IT, training, and workflow modification resources are needed for implementation?
Long-term Operational Home	Who will continue to provide technical support and quality assessment of the tool into the future?

# General LLM use to promote work of the future

Provision secure LLM access for all areas of healthcare organization:

- Operations
- Clinical
- Research
- Education

Guideline considerations:

- Bias & accuracy
- Suitability
- Disclosure
- IP/Copyright
- Minimum necessary data
- Escalation/control

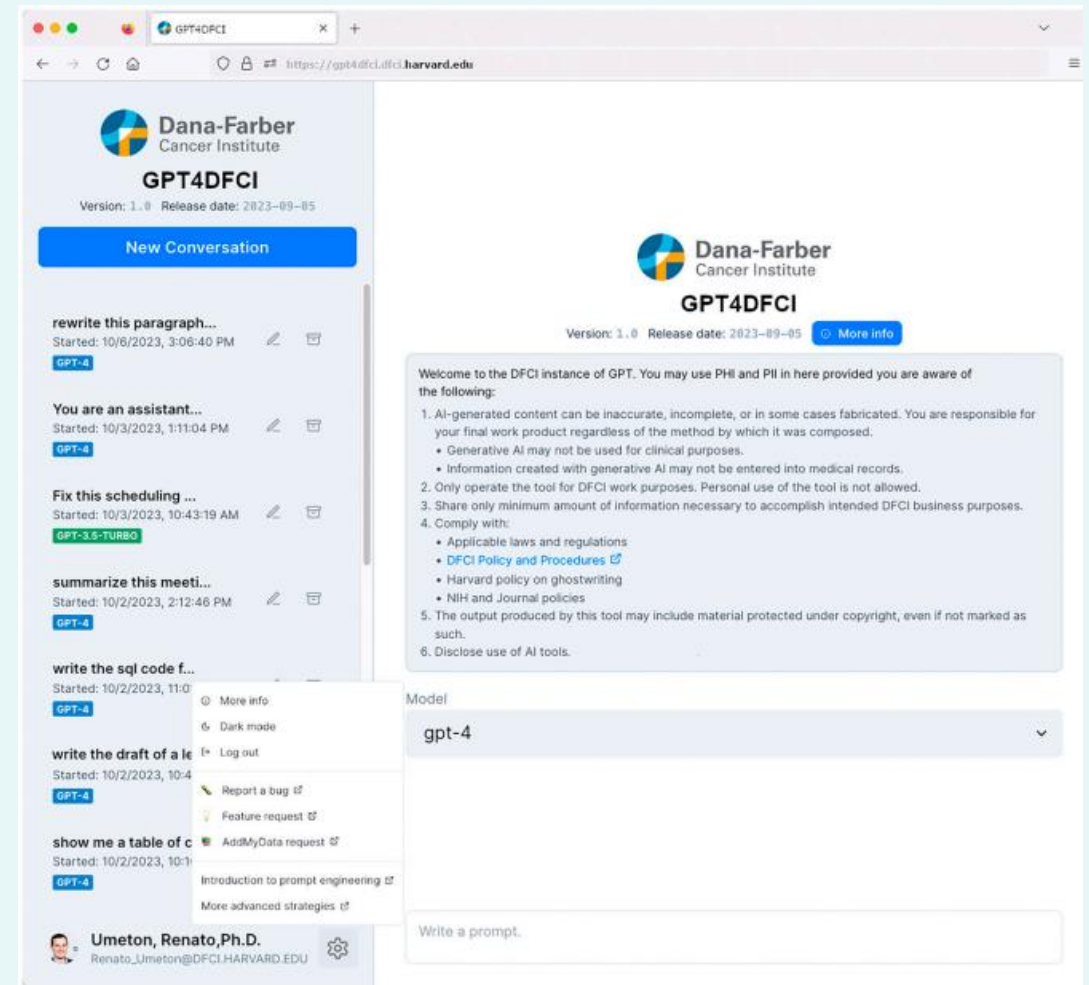
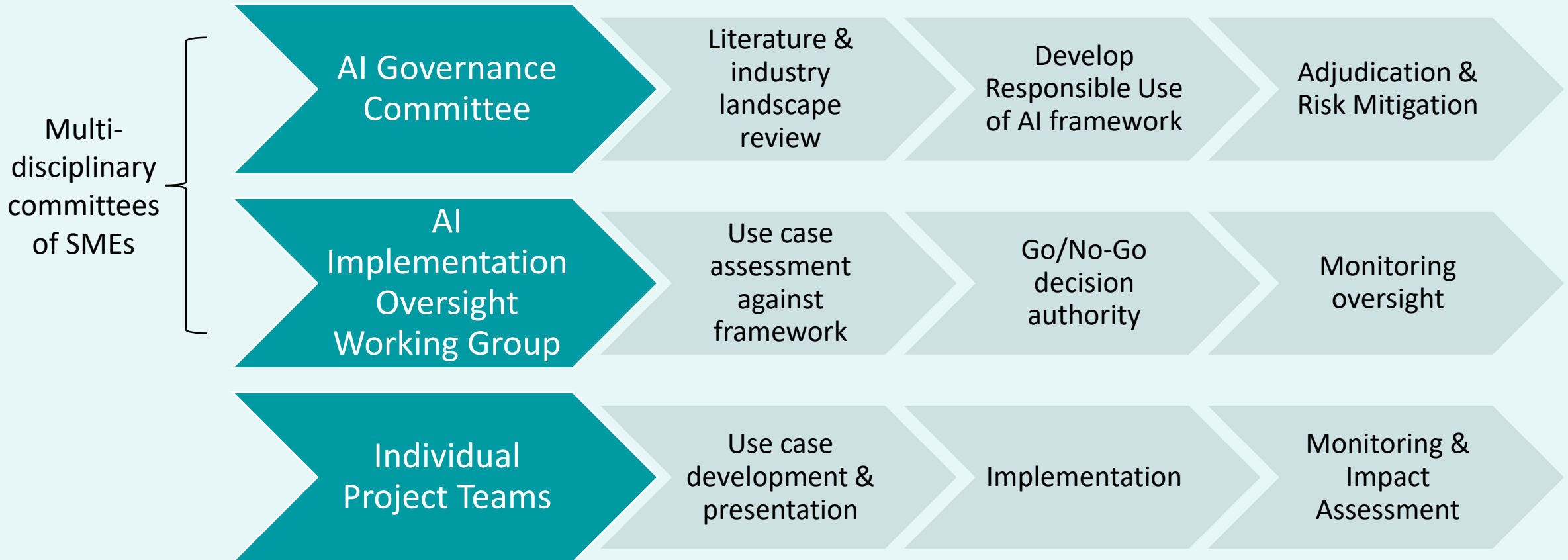


Figure 1. User Interface of GPT4DFCI v1.0 as Released on September 5, 2023. Previous conversations appear on the left, and a summary of use appears on the right, along with model selection. Also shown is a mechanism for users to report bugs, request features, consult the user guide on prompt engineering and “More advanced strategies,” and initiate the process for the creation of a retrieval-augmented generation project. GPT4DFCI denotes Generative Pretrained Transformer models for Dana-Farber Cancer Institute.

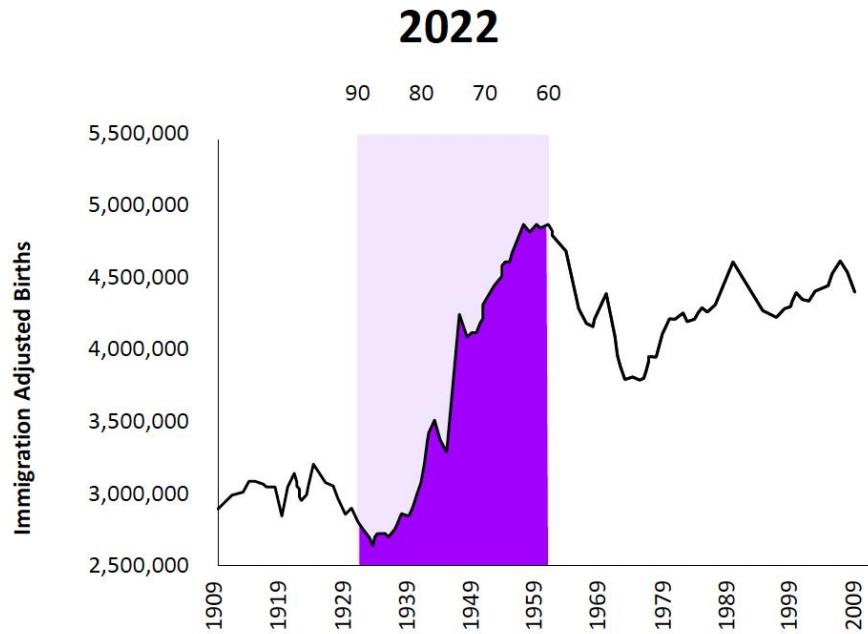




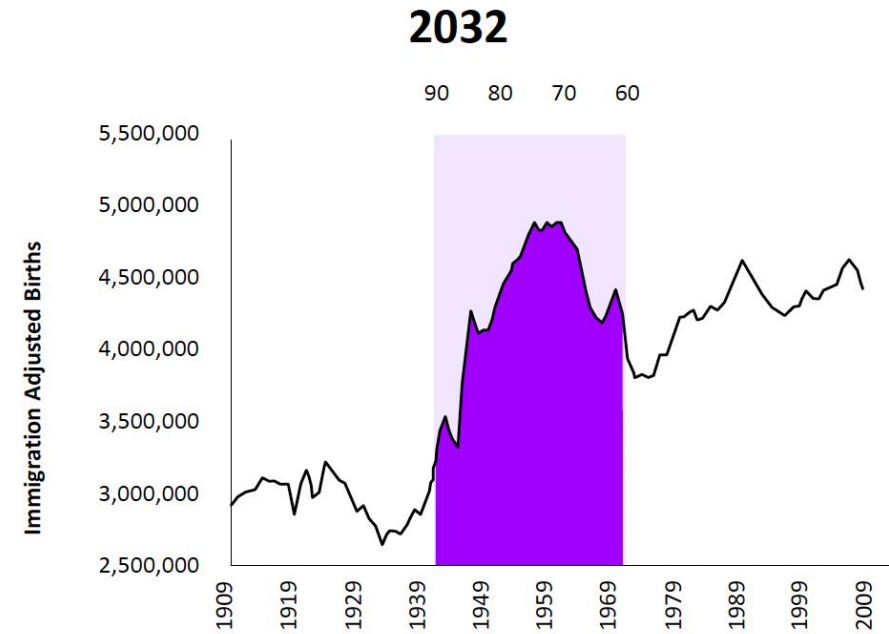
# AI Governance Approach



# The Supply/Demand Problem Continues to Get Worse



~54 million



~78 million

46% more 60-90 year-olds in next decade  
17% fewer working age person per senior

