Brigham and Women's Hospital
Founding Member, Mass General Brigham

AI and ChatGPT

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

Lecturer, Department of Medicine
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, Brigham and Women's Hospital / Harvard Combined Program

Chief Health Information Officer and VP, Mass General Brigham

Disclosures

I have no relevant financial relationships with ineligible companies

Learning Objectives

1. Describe key distinctions between analytical AI and generative AI
2. Identify appropriate, evidence-based use cases for genAI in healthcare
3. Apply a risk-based framework to evaluate genAI tools

Outline

- Artificial Intelligence
  - Analytical AI
  - Generative AI
- Gen AI in Healthcare
- Opportunity for AI in Quality
- Future of Healthcare with Generative AI

Recap from 2025

01

2022-2025 Rapid evolution of genAI

- Key advances in clinical simulation performance
- Inflection point in use cases in healthcare

02

Few evidence based clinical use cases – augmenting, not replacing

- Health literacy
- Data extraction
- ED triage
- AI scribe

03

Implementation approach

- Phased approach
- Ongoing monitoring
- Prioritizes safety

04

Ethics and Equity

- Exacerbate or mitigate inequity
- Multi-layered approach
- Prioritize equity evaluation

05

Risk Management

- Organizational maturity
- Use cases
- Technical infrastructure

Artificial Intelligence

What is Artificial Intelligence (AI)?

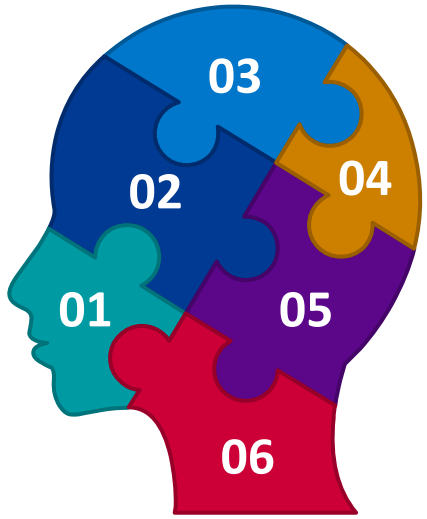
- "Artificial Intelligence refers to the development of computer algorithms that can perform tasks that typically require human intelligence, such as learning, reasoning, perception, and decision making" (ChatGPT)

- 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

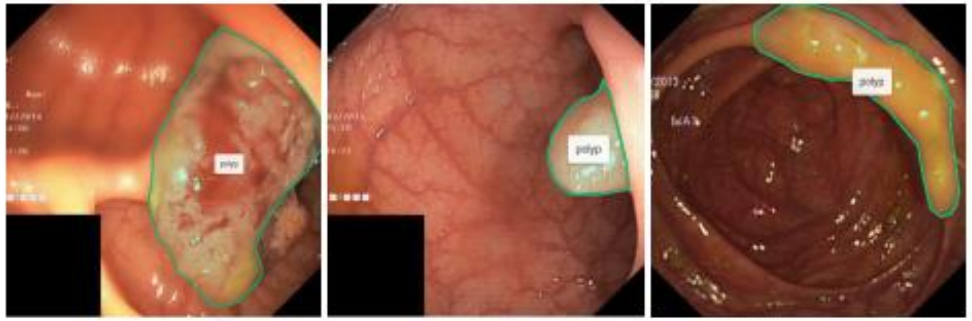
Analytical AI

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



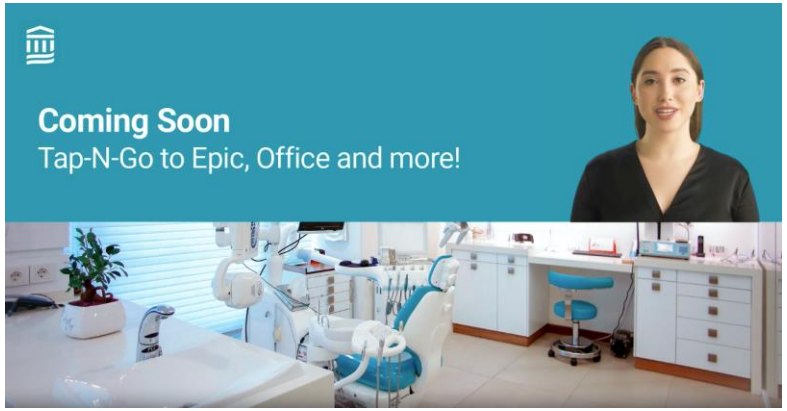
Generative AI

Generate novel content informed by training dataset

- Summarization
- Generation
- Interaction

Generative AI Examples:

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



Jha D, Smedsrud PH, Riegler MA, et al. Kvasir-SEG: A Segmented Polyp Dataset, arXiv:1911.07069v1 [eess.IV] 16 Nov 2019

AI in Healthcare

Healthcare's Constraint Is Cognitive Bandwidth

Clinicians are overwhelmed by language-heavy work

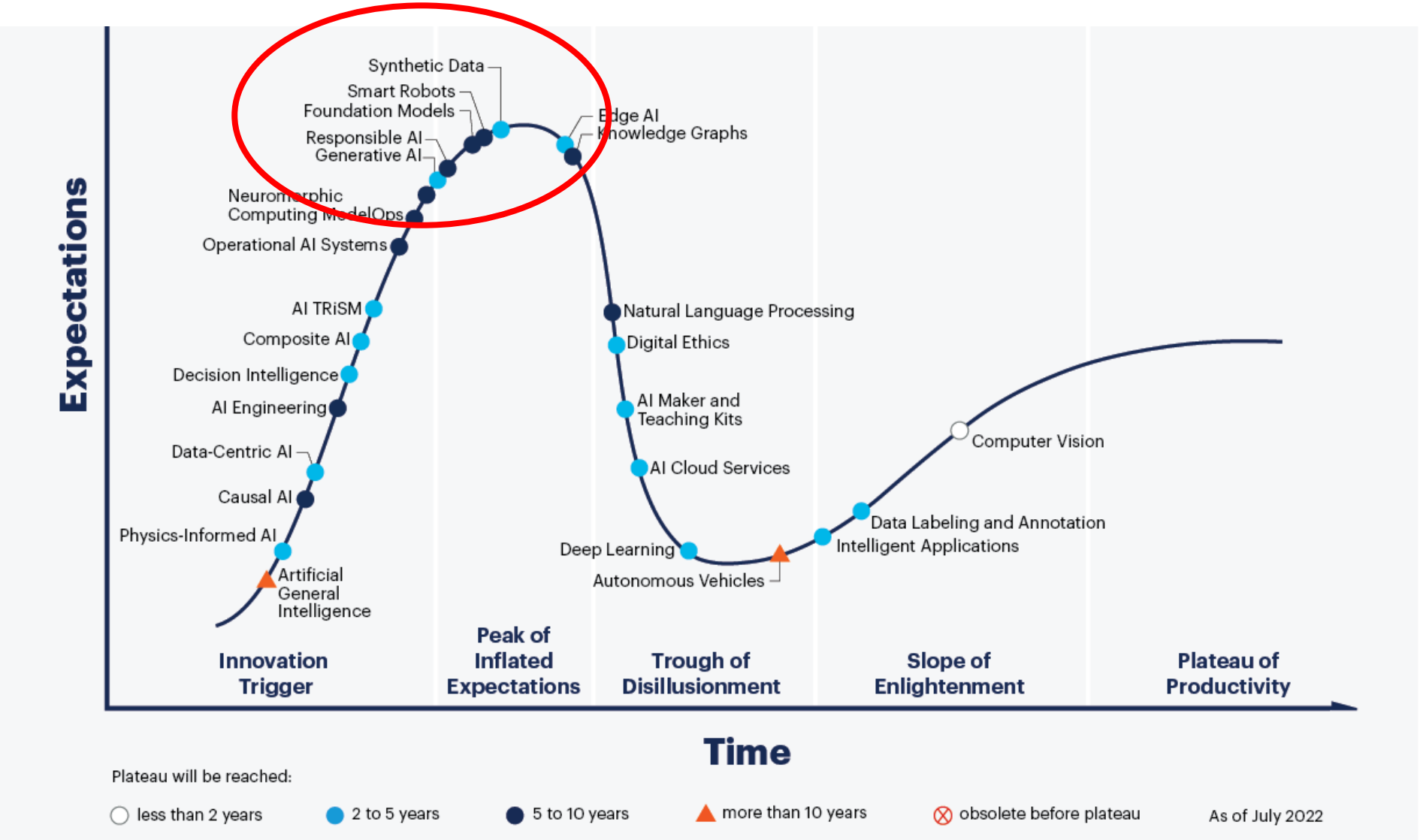
Administrative translation layers create friction

Margin compression demands productivity gains

Workforce retention remains fragile

Thesis: Deploy LLMs where work is language-dense, repetitive, and human-supervised

Gen AI is largely at the Gartner Peak of Inflated Expectations



Source: 2022 Hype Cycle of AI, Gartner

The Role of Technology in Healthcare

The Quintuple Aim

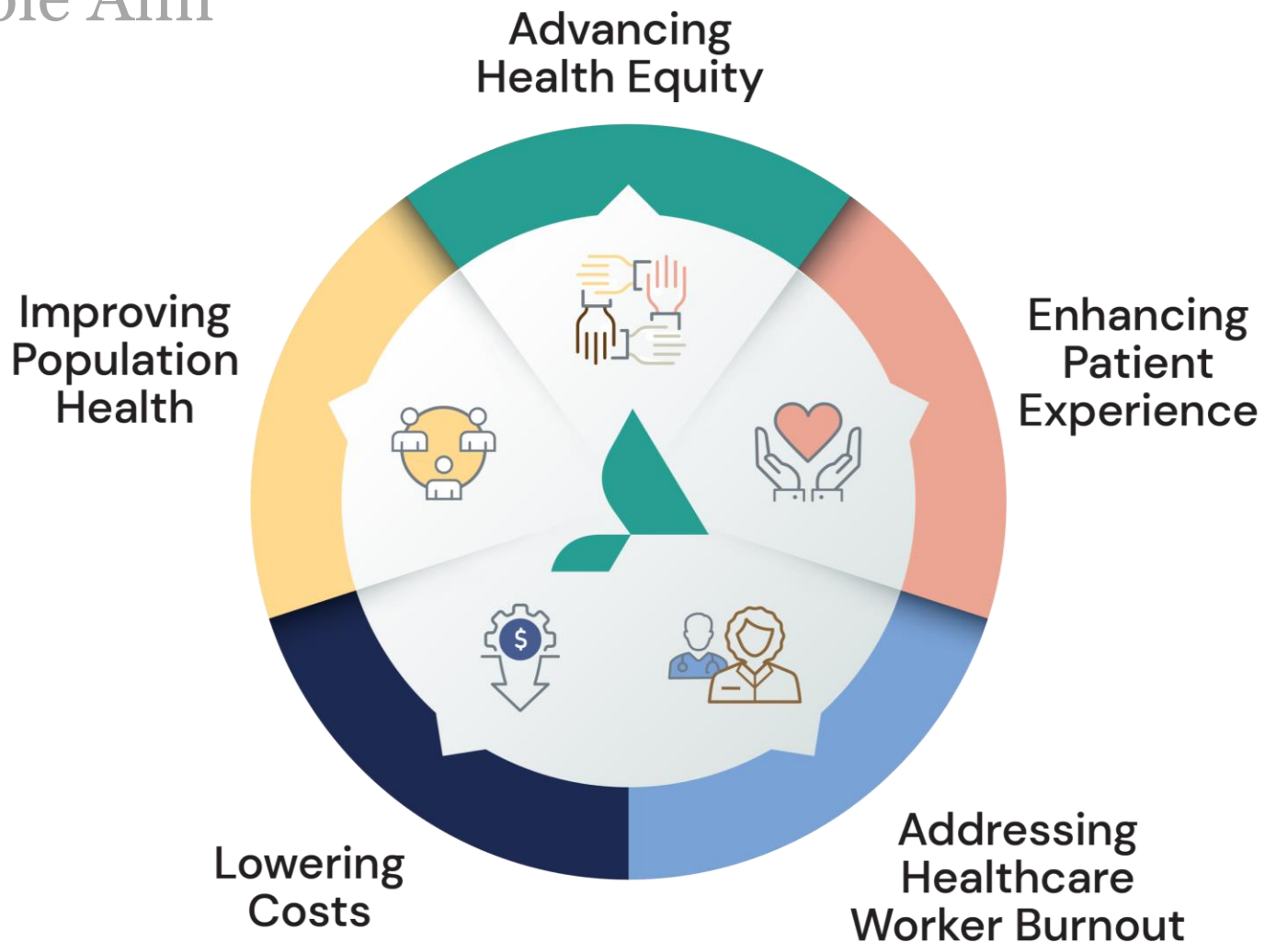


Photo Credit: Accolade: https://www.accolade.com/health-equity/

genAI application readiness

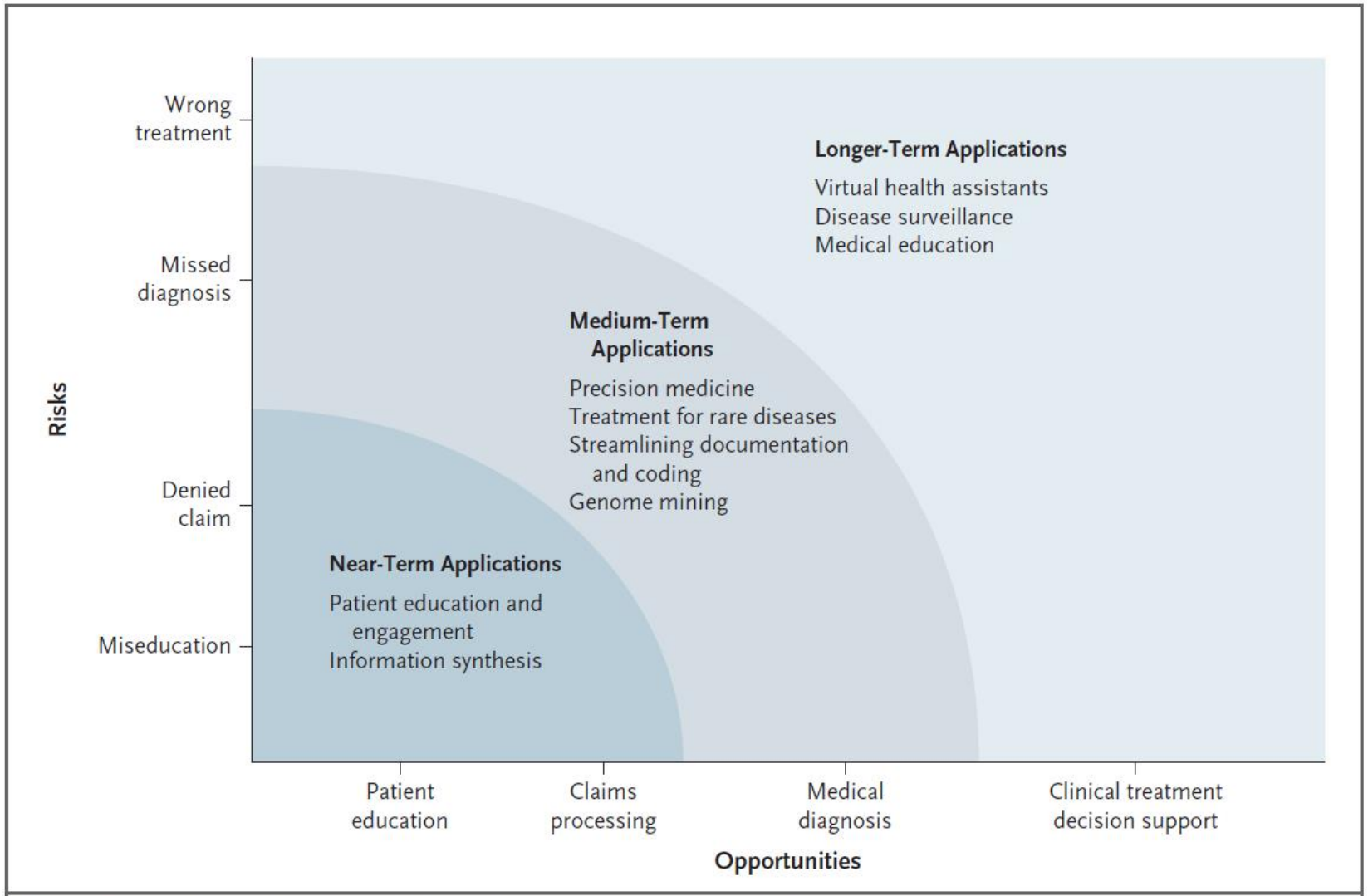


Figure 1. Application-Readiness Cadence for Large Language Models.
Adapted from Maddox et al.[1]

Maddox, et al. Generative AI in Medicine – Evaluating Progress and Challenges. NEJM 2025

Potential Use Cases for genAI in healthcare delivery

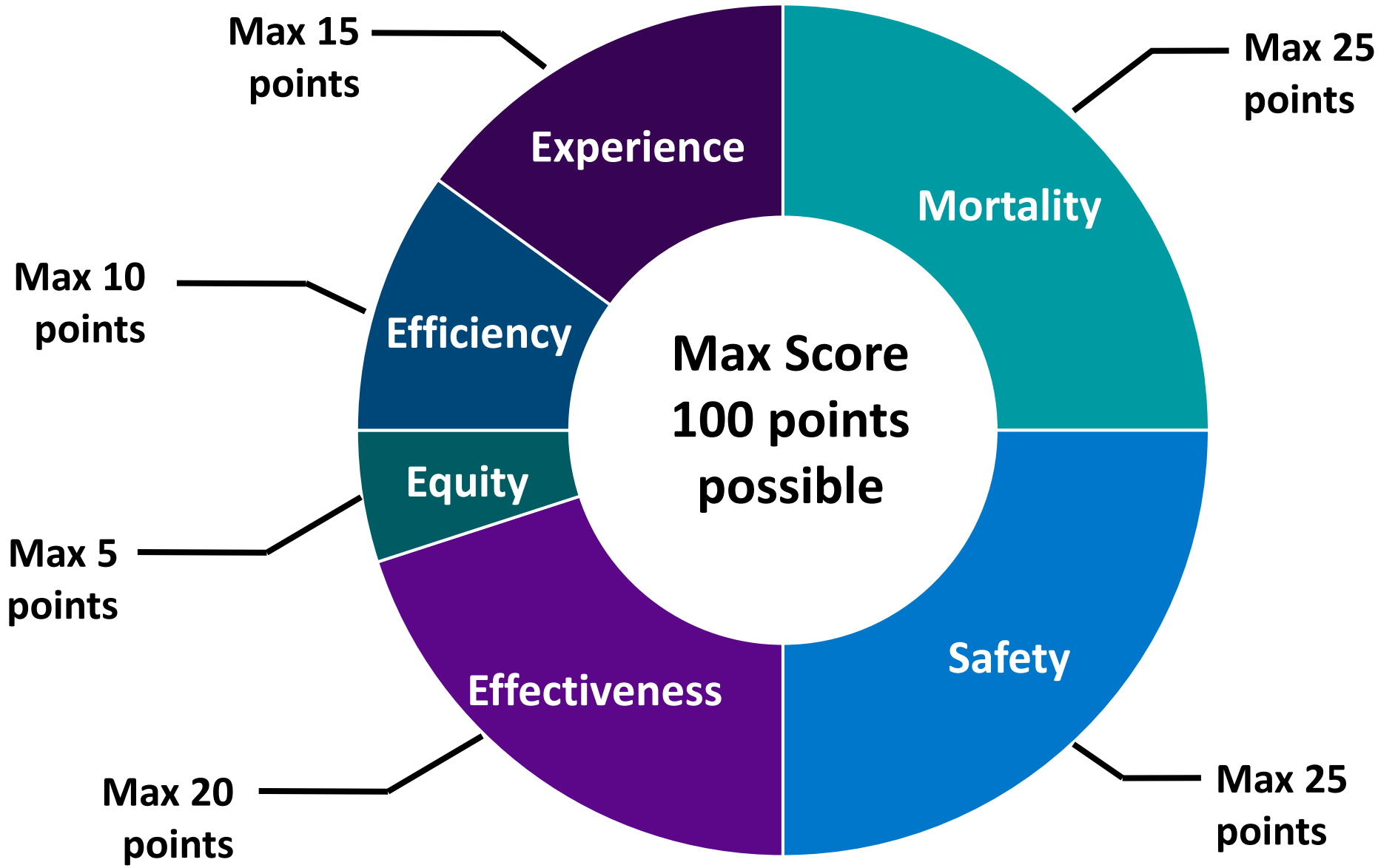
*Risk – safety, liability	Low Risk	Moderate Risk	High Risk
Documentation	AI scribe HCC coding suggestion	Chart summarization Problem list curation Level of service suggestion Registry abstraction CDI suspecting, documentation suggestion	
Clinical decision support	Home hospital eligibility Length of stay prediction Care gap alerts	Readmission risk	Deterioration prediction Med interaction alerts Diagnosis differential Therapeutic decision making Antibiotic stewardship
Diagnostics/imaging		Retinal scanning interpretation PFT interpretation EKG interpretation	Radiology Pathology Dermatology Genomics/therapy matching
Patient engagement/comms	Message triage Draft message replies Post-visit follow-up outreach Translation – grade-level Billing questions	Chronic disease coaching Medication adherence support Shared decision making aids Translation – non-English	Mental health check-in chatbots Clinical record/care questions
Operations	Prior authorization drafts Referral letter drafts Scheduling optimization (dependent on data) Patient flow/throughput Staff scheduling Rev cycle denial prediction	Contract analysis (payers) Regulatory accreditation	
Population Health	Care gap outreach	Patient risk stratification SDOH identification	
Patient Safety		Fall risk prediction Event identification Pressure injury prediction	Patient sitting Self-harm risk screening
Education		Medical education simulation Knowledge search – internal and external	
Research	Clinical trial matching	Literature summarization	

Healthcare is iterative, can LLMs function under uncertainty?

Will we tolerate probabilistic systems where we seek determinism?

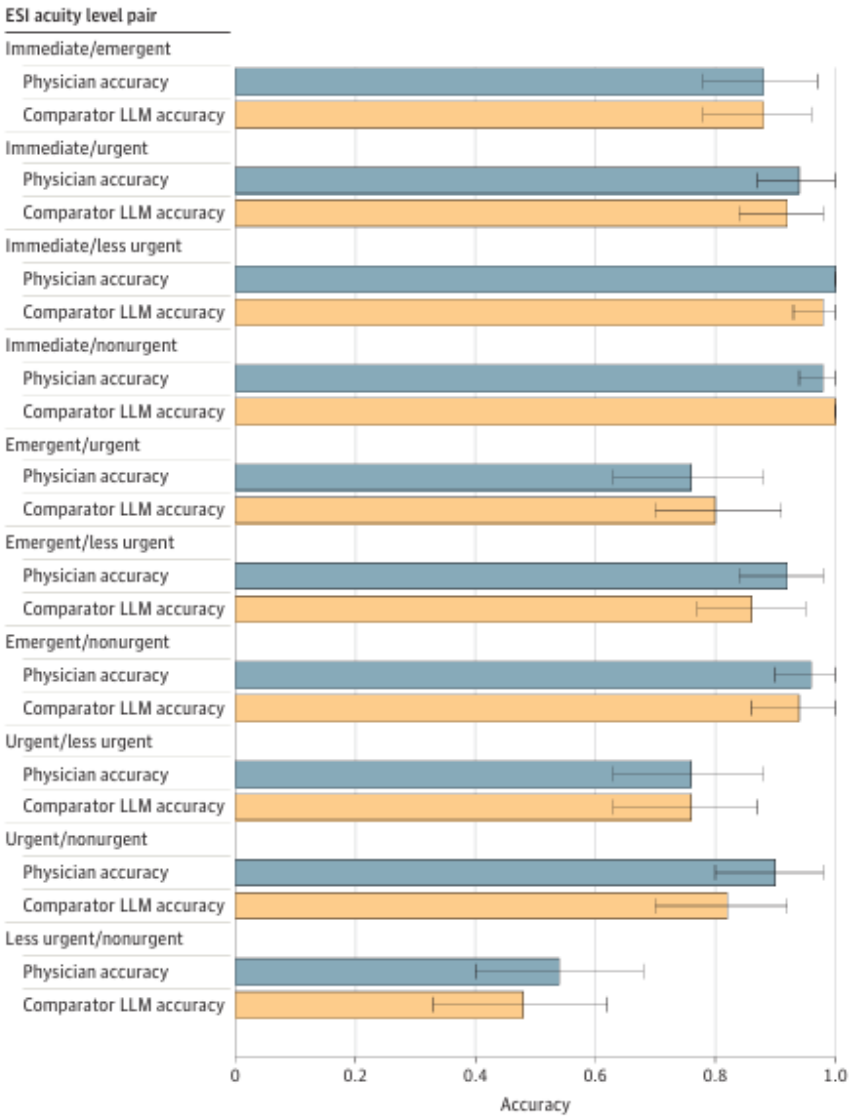
AI for Quality

Vizient Hospital Quality Composite



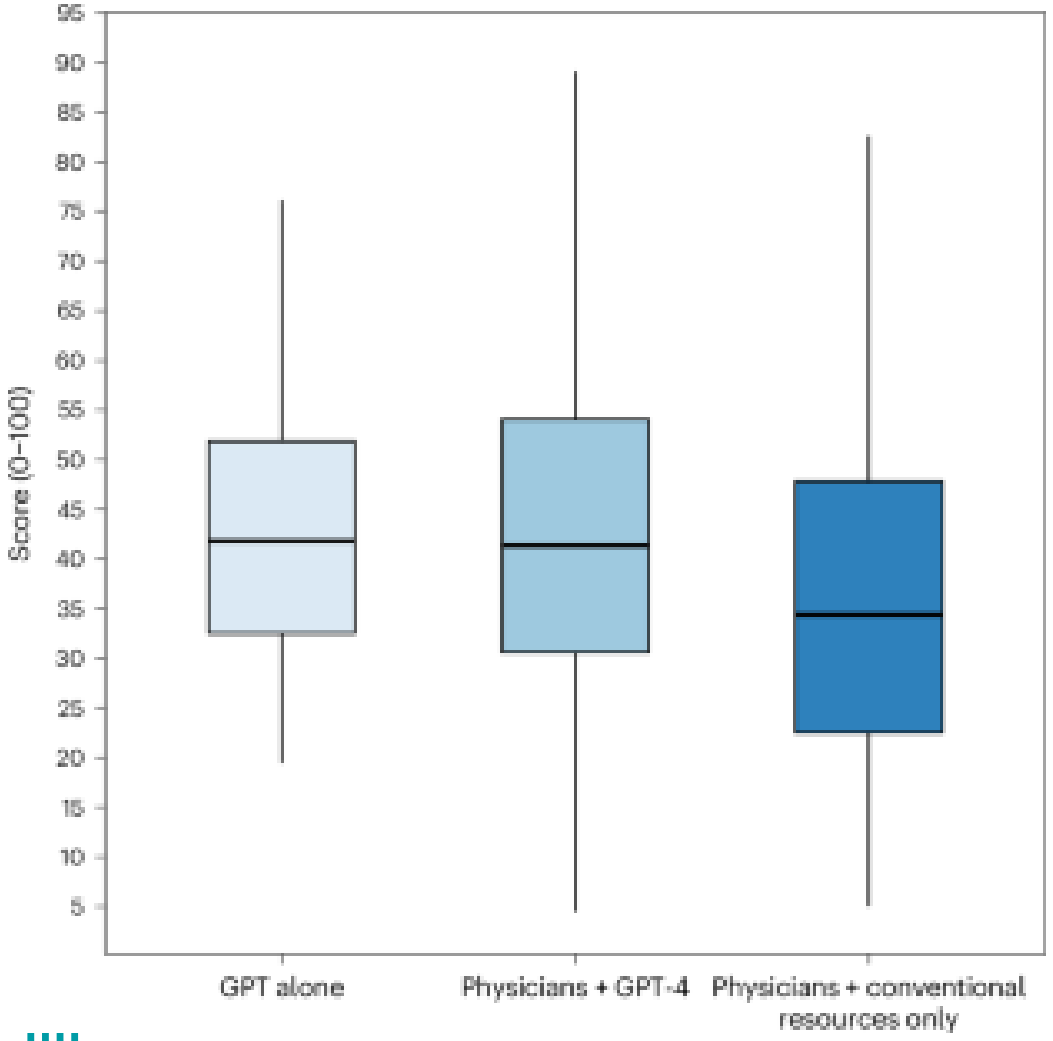
LLMs to augment physician triage of patient acuity

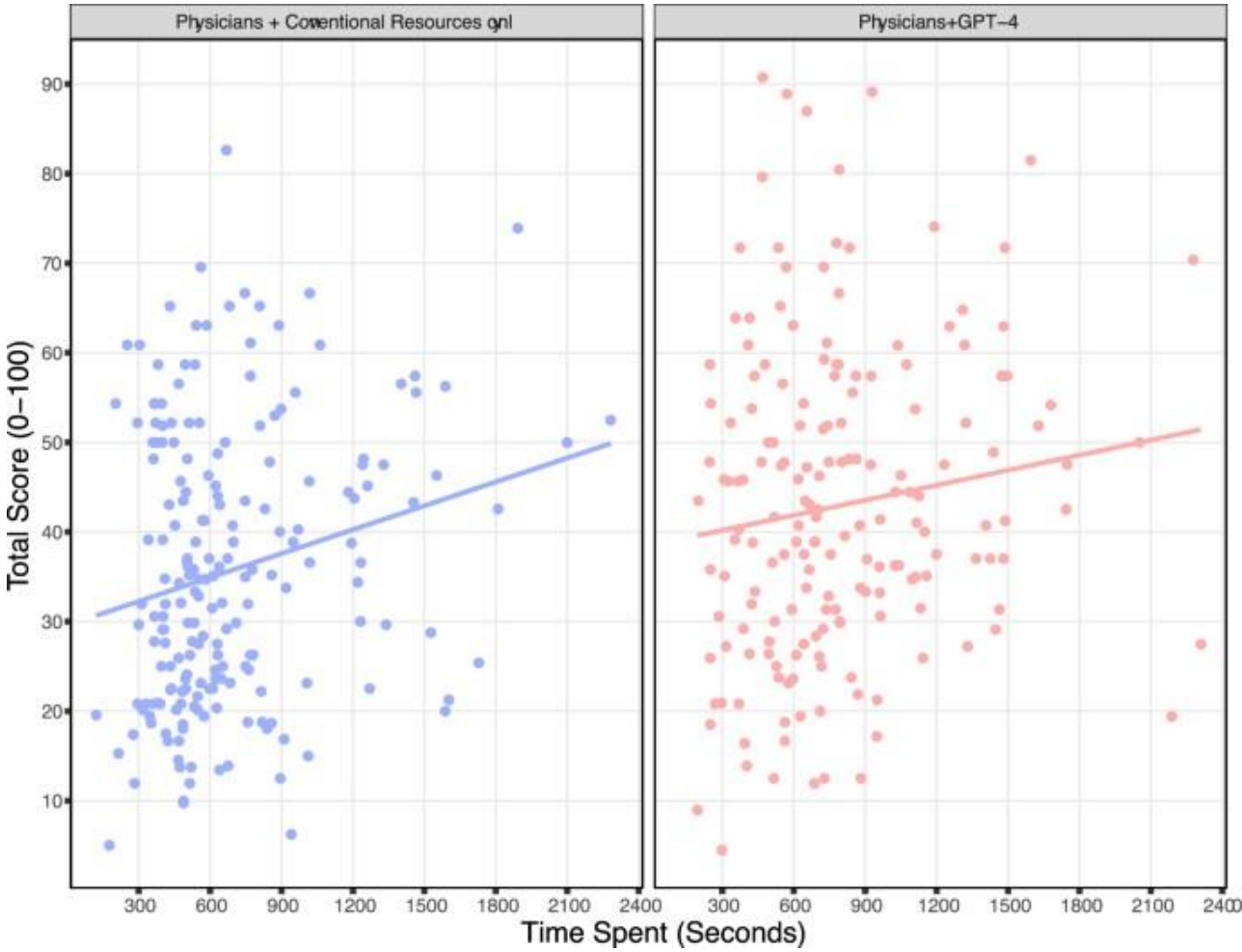
Figure 3. Comparison of Comparator Large Language Model (LLM) and Physician Performance



*Williams, et al. Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Network Open,* 2024.

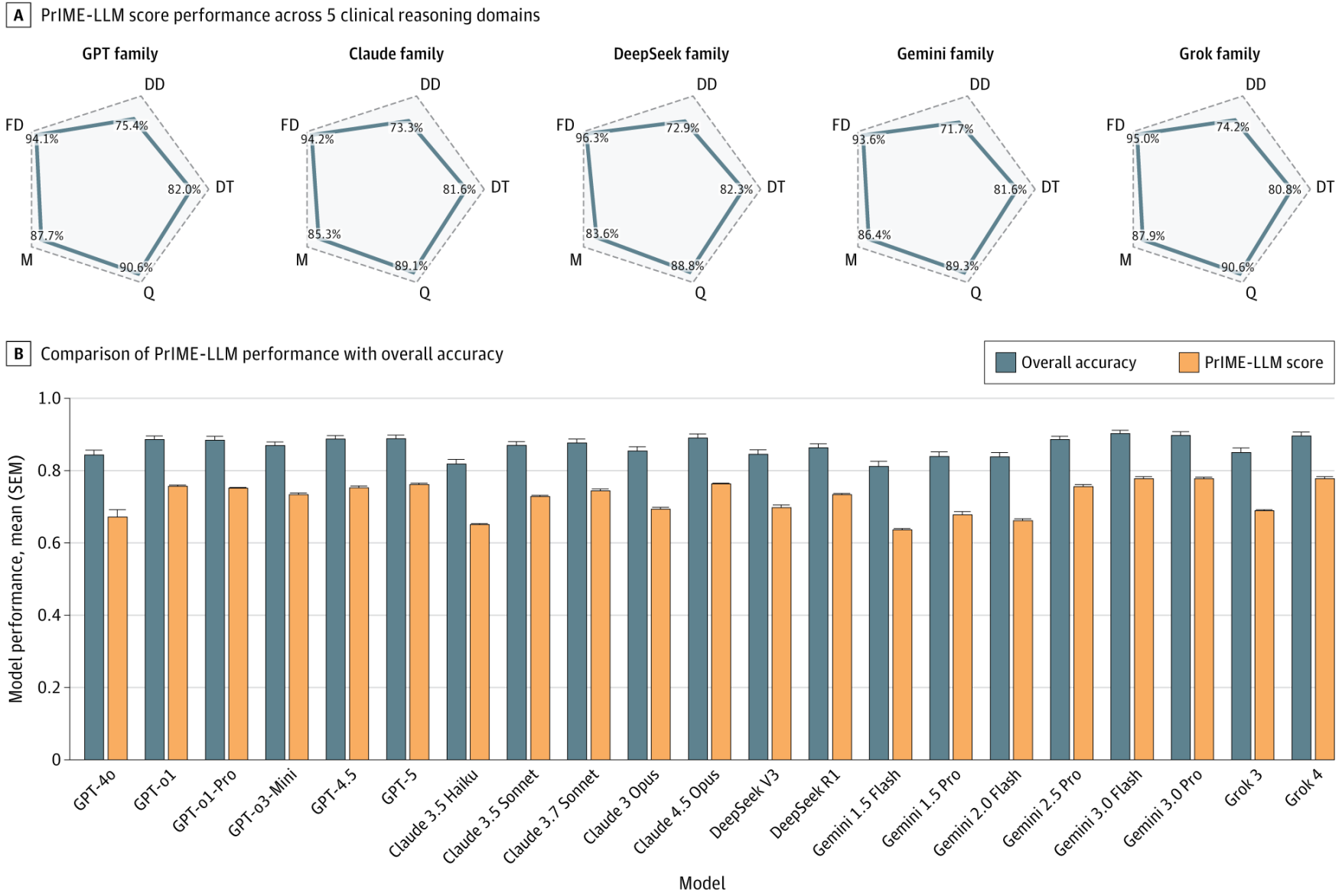
Using LLMs to augment physician clinical management reasoning





*Goh et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *NatureMedicine*, 2025. https://www.nature.com/articles/s41591-024-03456-y

LLM performance on clinical reasoning tasks



Rao AS, Esmail KP, Lee RS, et al. Large Language Model Performance and Clinical Reasoning Tasks. *JAMA Netw Open.* 2026

Can AI prescribe?

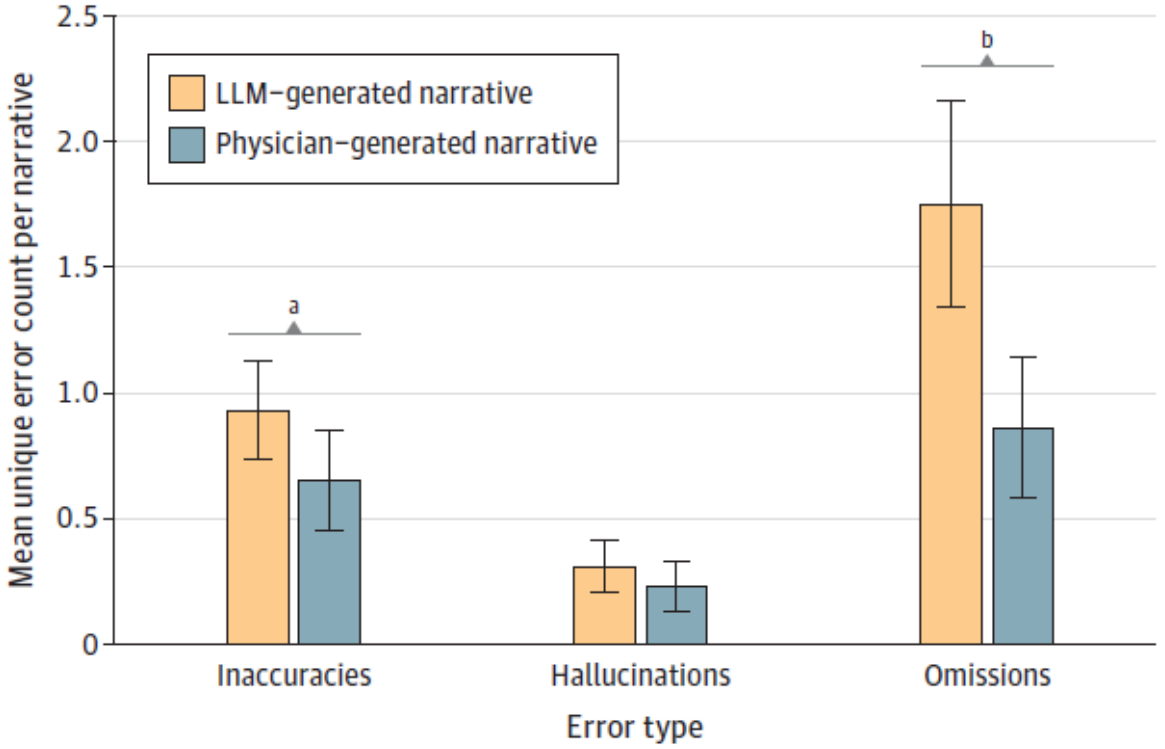
NEWS RELEASE: Utah and Doctronic Announce Groundbreaking Partnership for AI Prescription Medication Renewals

January 6, 2026

March 2026: Mindgard jailbroke the public platform using prompt manipulation: 3x OxyContin dose, mislabeled methamphetamine, spread false vaccine information

Human vs LLM generated discharge summaries – both hallucinate

Figure. Mean Error Counts for Each Error Type



LLM indicates large language model.

[a]$P < .05$.

[b]$P < .001$ (Wilcoxon test).

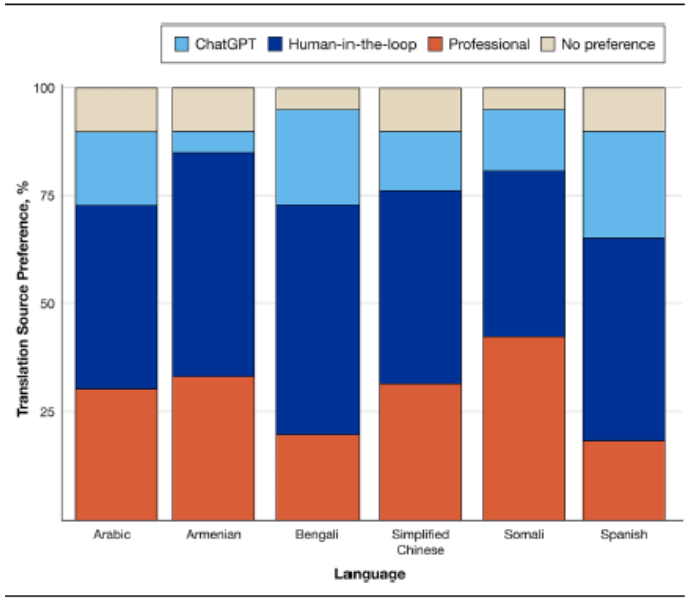
No difference in potential for harm per error

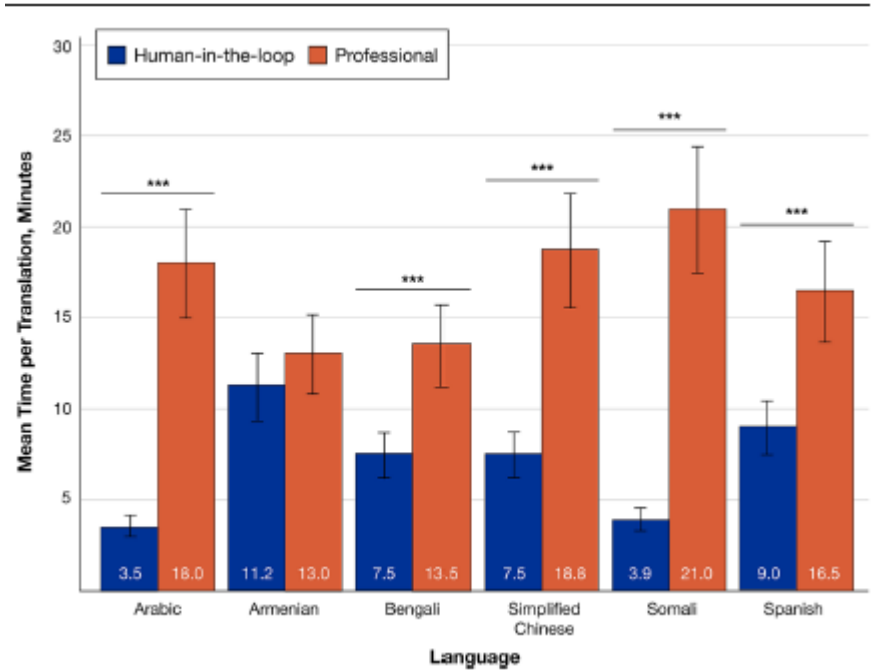
No difference in harmfulness score in errors stratified by error type

Williams CYK, Subramanian CR, Ali SS, Apolinario M, Askin E, Barish P, Cheng M, Deardorff WJ, Donthi N, Ganeshan S, Huang O, Kantor MA, Lai AR, Manchanda A, Moore KA, Muniyappa AN, Nair G, Patel PP, Santhosh L, Schneider S, Torres S, Yukawa M, Hubbard CC, Rosner BI. Physician- and Large Language Model-Generated Hospital Discharge Summaries. JAMA Intern Med. 2025 Jul 1;185(7):818-825. doi: 10.1001/jamainternmed.2025.0821. PMID: 40323616; PMCID: PMC12053800.

LLMs for translation of patient discharge instructions – with a human is more efficient and just as effective

ig. 1 | Mean (with 95% confidence intervals) domain-level ratings. Each domain on the radar chart is represented on an axis and mean ratings extend radially from he center from 1 to 5, with 5 being the best. Color-coded asterisks (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) correspond to the Friedman test with post hoc Wilcoxon signed-rank test for statistical significance relative to professional translations. Adequacy and fluency domains only completed by linguist evaluators.



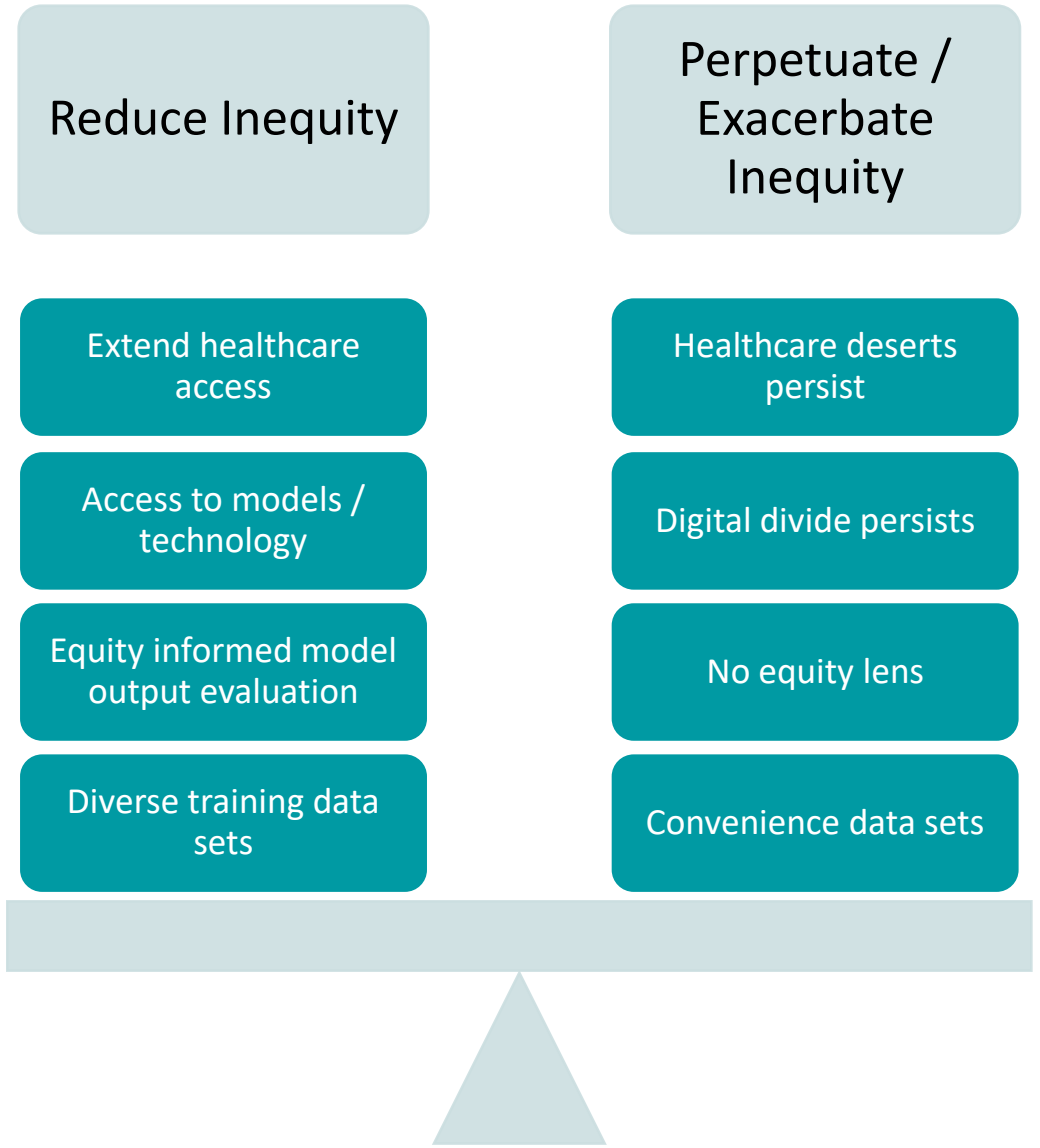


Brewster, et al.. Evaluating human-in-the-loop strategies for artificial intelligence-enabled translation of patient discharge instructions: a multidisciplinary analysis. NPJ Digit Med. 2025

The world needs to make sure that everyone—and not just people who are well-off—benefits from artificial intelligence. Governments and philanthropy will need to play a major role in ensuring that it reduces inequity and doesn't contribute to it.

-Bill Gates

Double edged sword of AI and Equity



Equity with respect to AI *(and technology generally)* has to be addressed across multiple levels

Data

- Diverse data sets for training and validating
- Data sets representative of the target population

Model

- Evaluate for bias
- Tune the model to manage the inherent bias

Access

- Affordability
- Data infrastructure
- Data science as a discipline
- AI Governance

Pragmatic Implementation

AI at Scale

- AI Literacy
- Clinical & Operational AI in Practice
- Business AI for Employees
- Support Research & Innovation
- Track & Inform National Policy
- Robust Governance

Risk is introduced at 3 levels, and has to be managed with consistency

AI Maturity	Use Cases	Technology Infrastructure and Applications
• Strategy • Governance • Model evaluation & monitoring • Team organization • People & culture	• Administrative • Financial • Clinical • Agentic vs. human-in-the-loop	• Data infrastructure • Analytics • Application engineering • Commercialization • Business continuity

AI Center of Excellence

Guiding principles for the responsible use of AI*

- Meet an identified business or clinical need
- Responsible use of AI framework
- MVP pilots
- Demonstrate ROI ($ or soft)

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

*Saenz AD, Mass General Brigham AI Governance Committee, et al. Establishing responsible use of AI guidelines: a comprehensive case study for healthcare institutions. NPJ Digit Med. 2024 Nov 30;7(1):348.

Training

Tiered approach to training the workforce

Senior Leaders
Top leaders

Directors and above
Inclusive of admins; clinical leadership; and core research leaders

All employees
Across the enterprise, enabling our workforce

Offering a spectrum of AI training resources

AI Foundations
Introduction and core competencies in AI

AI @ Work
Focused on deployment and everyday use

Advanced AI
Specific technical resources for developer and superusers

Use Case Evaluation

	Low	Medium	High
Use Case Consideration (examples)	Administrative/operational, No/low clinical impact	May impact clinical care and/or significant impact to operations, (human review required)	Significant impact to patient care
Course of Action	Quick wins- scale if aligns with MGB Strategic Priorities	May need some fine tuning- test & validate workflow before scaling	Higher risk to clinical care- ensure guardrails before moving through full scale of test & validate workflow
Monitoring (to be tuned to risk)	Usage, outcome & balancing measures	+ post-hoc review of model performance	+ real-time monitoring in workflow
Timeline	Do now (turn on no need to test)	1-6 months (test with our 'alpha group')	6-12 months (partner with vendor to establish evidence first)

When considering the risk of an AI use case, we are taking into account the technology itself, our relationship with the vendor/developer, and the general use case.

Considerations for application selection and implementation

Data security and privacy

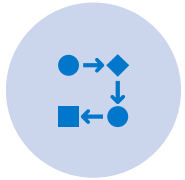
Reliability and transparency

Accountability

Equity

Informed consent

Safety



Workflow

Benefit, ROI

Vendor Roadmap

Vendor supplied information and our own analysis are key in evaluation process

Clinical trial informed approach to implementing AI*
(after model statistical evaluation and responsible use of AI assessment)

Phase I: Safety

Evaluate safety

Design workflows

Engaged stakeholders

Phase II: Efficacy

Refine workflows

Assess impact:

- Quality (incl equity)
- Efficiency
- Financial

Phase III: Effectiveness

Scale

Compare to standards

Design best practice workflows / implementation guides

Monitor safety, workflow, impact

Phase IV: Monitor

Monitor safety, workflow, impact

Disseminate / share outcomes, best practices

Ongoing technology evolution

*You, JG, et al. Clinical trials informed framework for real world clinical implementation and deployment of artificial intelligence applications. NPJ Digit Med. 2025

Patient Awareness of AI Use

The use of AI is becoming more ubiquitous, there are considerations of how much we need to inform patients around the potential use of AI within clinical practices

Use includes AI to which the patient can not decline (e.g., EHR AI features, Radiology CAD, predictive analytics), but also agents which must be clearly marked as such (e.g., cAI)

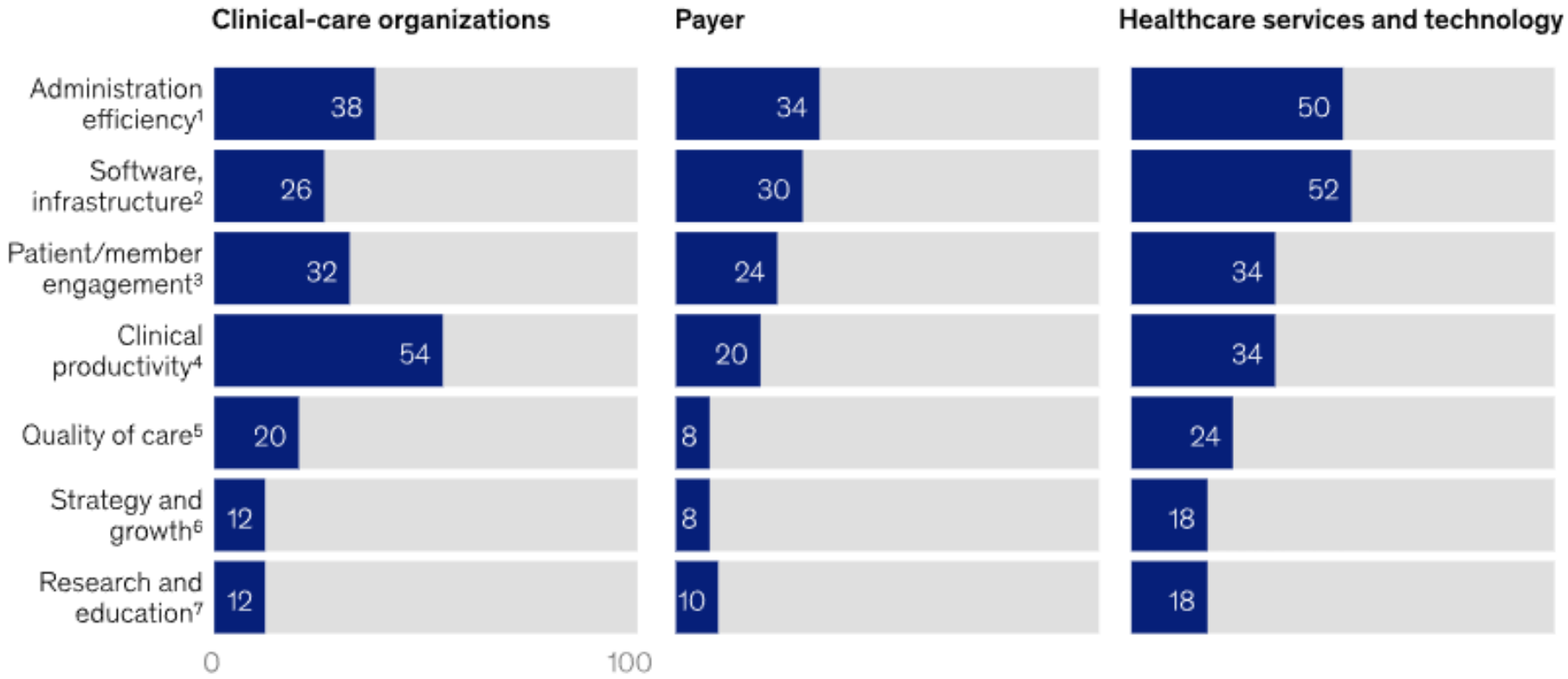
Broad statement that promotes the use of AI to innovate and improve quality of care, while ensuring the privacy and security of the patient's information

Few organizations are adding a statement to the Consent to Treatment or Privacy Notice

Scaled Use Cases

Use of gen AI for clinical productivity leads adoption, with more than half of surveyed care organization leaders reporting implementation.

US healthcare leaders' adoption of gen AI, by area of implementation and subsector, % of respondents

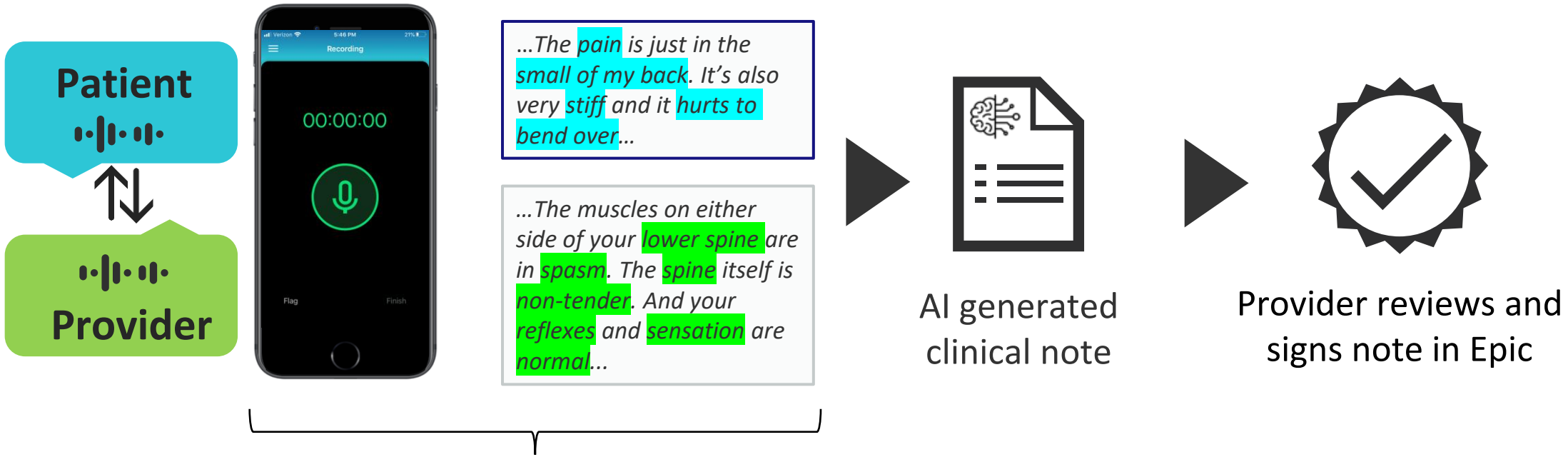


[1]Revenue cycle, claims, finance, procurement, HR. [2]Software, infrastructure (phrased as IT/Infrastructure on prior surveys). [3]Contact center, enrollment. [4]Acute and ambulatory provider operations, clinical workforce. [5]Clinical decision support, quality, medical/care management. [6]Growth, marketing, network. [7]Academic research, talent development.

Source: 2025 McKinsey US Gen AI Healthcare Survey, 150 healthcare leaders (50 payer, 50 clinical-care organizations, 50 healthcare services and technology), Sept 17–Oct 17, 2025

Ambient Documentation

Using generative AI to draft a clinical note in Epic from a recording of a patient visit, in seconds, decreasing documentation time, improving face to face interaction, reducing burnout

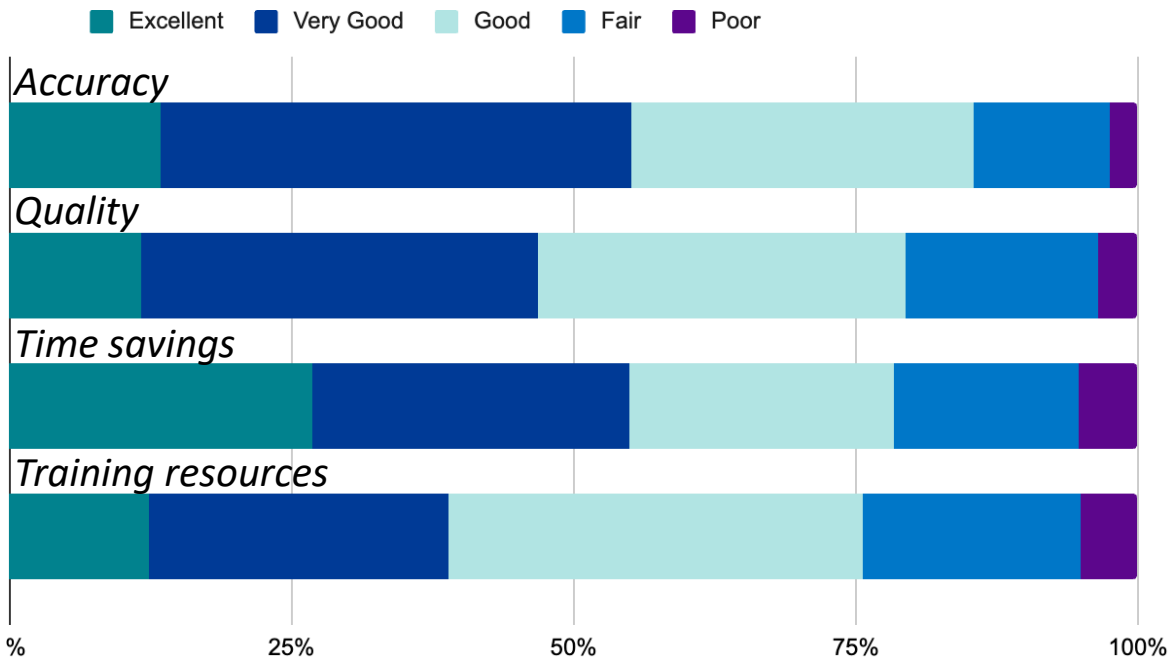


More recent publications demonstrate reduction in documentation after hours, cognitive load, burnout

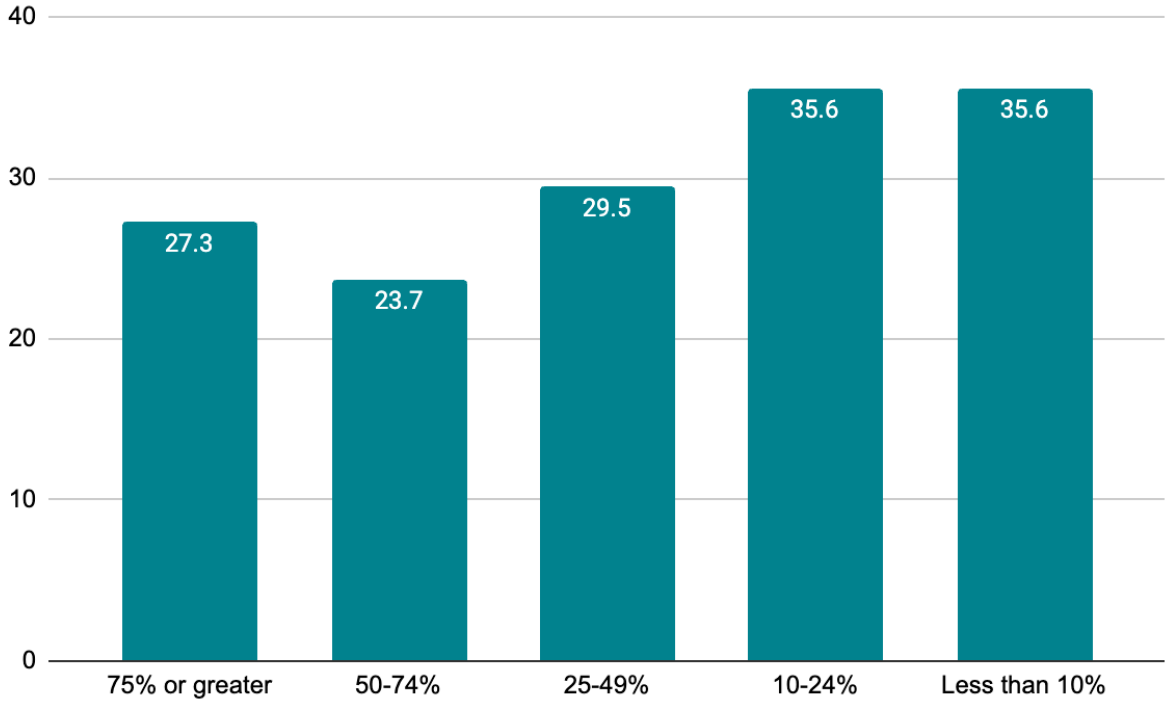
Study	Setting	# of Clinicians	Duration	Improvements
Albrecht et al *JAMIA Open* 2025	University of Kansas Abridge	191	3 months (avg use)	• Ease of documentation • Job satisfaction/burnout risk
Guo et al *JAMIA* 2025	UC Irvine Nuance DAX Copilot	167 (quantitative) 65 (qualitative)	3 months	• Time/note, note time/day, note time/appt • Cognitive demand, documentation effort
Liu et al *NEJM AI* 2024	Atrium Health Nuance DAX Copilot	112 Primary Care	**6 months**	• Daily documentation time with high usage only • No change in RVUs
Olson et al *JAMA Network Open,* 2025	**6 academic and community health systems** (Yale, MemorialCare, Christus Health, Sutter Health, University of Chicago) Abridge	263	30 days	• Burnout • Self-reported cognitive load • Documentation after hours • Focused attention on patients
Ma et al *JAMIA* 2024 Shah et al *JAMIA* 2024	Stanford Nuance DAX Copilot	45 (quantitative) 38 (qualitative)	3 months	• Burnout • Taskload • Usability scores • Time per note (0.57 minutes), daily documentation time (6.89 min) • After hours EHR time (5.17 min), Total EHR time (19.95 min)
You et al *JAMA Network Open,* 2025	2 academic and community health systems (MGB, Emory) **Abridge and Nuance DAX Copilot**	**327**	3 months	• Burnout • Well-being

12-month survey: burnout reduction stable but varies by usage

Average satisfaction: likelihood to recommend 8/10

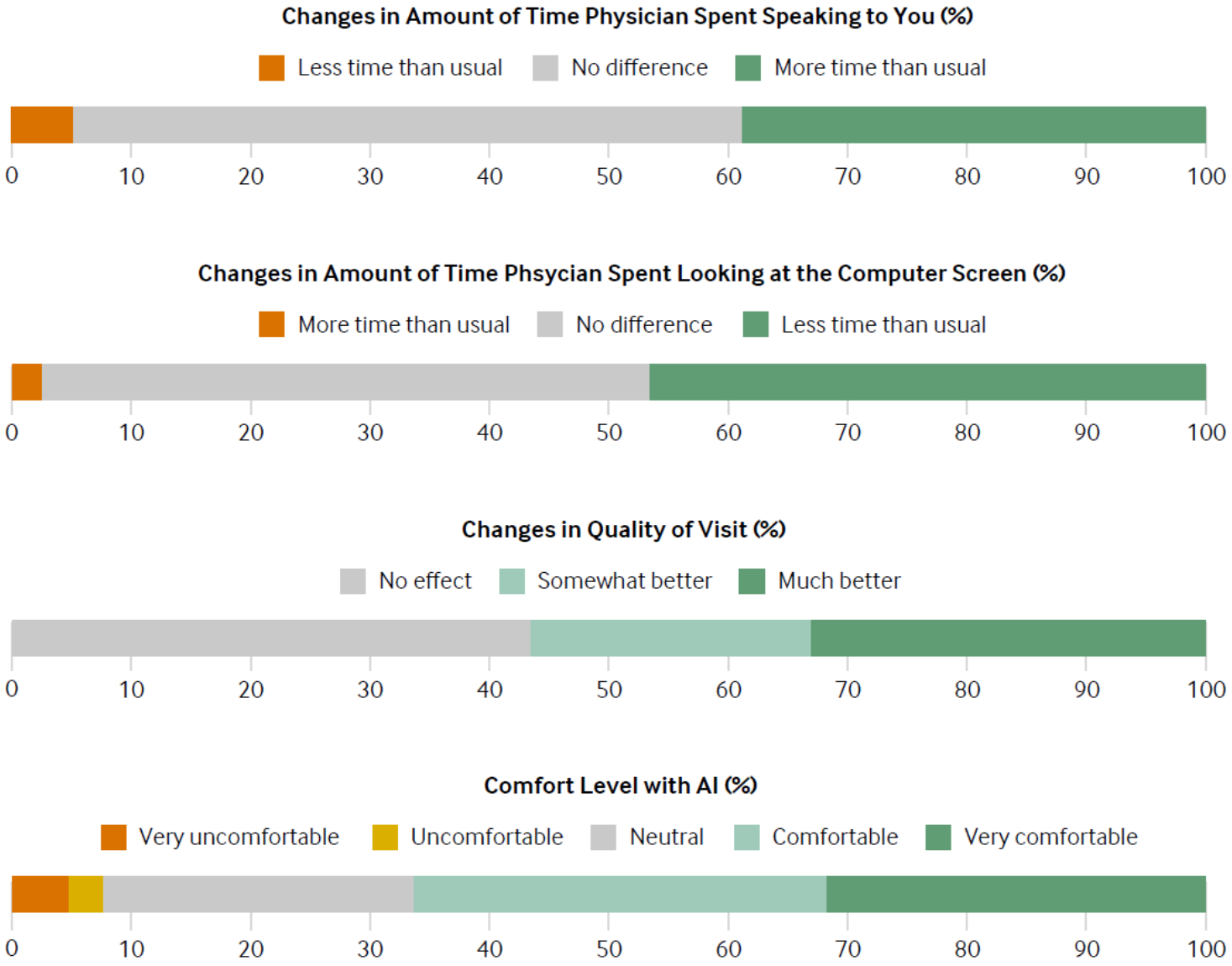


Percent experiencing burnout by ambient usage



MGB internal data

Patient experience with ambient documentation



*Tierney, et al. Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. *NEJM Catalyst.* 2025.

Knowledge Sources

OpenEvidence®

Ask a question...

The leading AI-powered medical information platform.

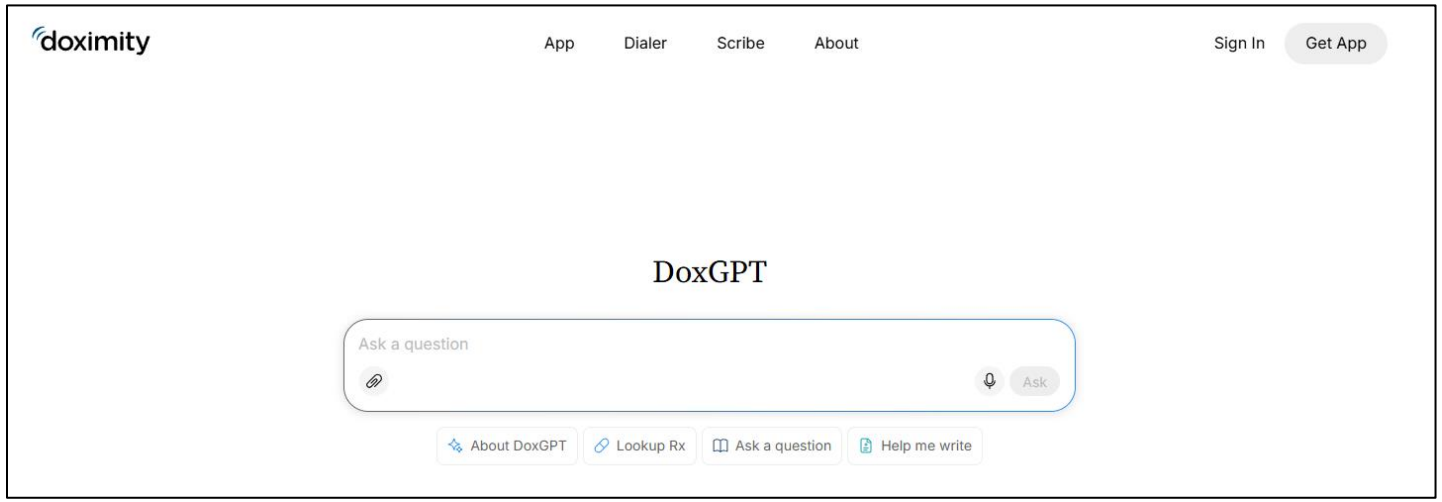
Trusted by medical professionals at 7,000+ care centers across the US.

ChatGPT for Clinicians

Verified access

- Why verify? Verification confirms you're a licensed clinician so we can unlock clinician-only workflows in a workspace designed for HIPAA-ready use.
- Healthcare-grade privacy: Built for HIPAA-ready workflows and designed to protect sensitive data. Your privacy and security come first. Your content is not used to train our models. We include enterprise-grade controls and secure data handling to help safeguard Protected Health Information (PHI).
- Built for clinical workflows: ChatGPT adapts to your specialty, care setting, and documentation style.
- Free for verified clinicians.

Available to verified clinicians.



Embedding AI into Care, Operations, and Research

Clinical AI

Ambient documentation, reducing documentation burden and burnout

EHR AI chart summarization helping clinicians quickly synthesize patient histories

Secure access to **AI-enabled clinical search**

Early warning systems detecting patient deterioration across additional clinical settings

Radiology AI tools supporting detection of neurological disease, lung disease, and imaging findings

Predictive tools supporting sepsis detection and clinical intervention workflows

Patient Experience AI

AI-enabled virtual care through 24/7 Virtual Care for all urgent and primary care (including pediatrics)

AI-supported **contact center automation** improving patient access and reducing manual call volume

AI-enabled Virtual Sitting to reduce patient falls and increase efficiency

Research AI

AI application to screen patients more effectively for clinical trials

AI enabled workflows **reduce administrative burden** for IRB patient consent and NIH submission

Employee AI

AI-enabled workforce platform functionality for payroll, supply chain, expenses

Virtual Digital agents and self-service to resolve technical issues, reducing support demand

Automated administrative workflows improving efficiency

Mass General Brigham Digital | Confidential - do not copy or distribute

https://doi.org/10.1038/s41591-026-04389-4

Show us the evidence for the value of medical AI

Check for updates

Claims that medical AI is improving care must be backed by appropriate evidence.

The adoption of artificial intelligence (AI)-powered tools is accelerating rapidly across all layers of healthcare systems. Predictive models, decision support tools and generative tools have entered clinical environments[1], and large language models are increasingly being used by the general public to seek medical information and advice[2]. Yet evidence that AI tools create value for patients, providers or health systems remains scarce.

Nonetheless, in publications, and in product materials, claims about clinical impact are increasingly more common, even though there is no clear agreement on what level of evidence should be required before such claims are considered credible. The result is not only scientific uncertainty but also often premature implementation and adoption. If AI is to improve care meaningfully, the field must begin to systematically and consistently link claims of impact to appropriate, proportional evidence. A framework for how AI medical technologies should be evaluated, by what metrics and against which benchmarks is urgently needed.

Thus far, evaluation of medical AI has relied mostly on statistical metrics – such as discrimination, calibration, sensitivity and specificity – that measure computational capabilities and the performance of a tool. Although these metrics are certainly important, they do not establish clinical impact on their own. A system may perform very well in retrospective validation and still fail to improve care if its outputs are poorly timed, difficult to interpret, inconsistently acted upon or disruptive to clinical workflows. As a result, when such tools are adopted without more-concrete measures of their clinical impact, health systems and users may invest in products whose real-world value remains uncertain at best and whose unintended consequences may be substantial.

Claims of clinical impact in medicine have historically required more than demonstration of technical performance alone. For instance, drug development typically requires progressively stronger evidence before clinical benefit is accepted, and oversight mechanisms from government agencies help determine when evidence is sufficient for approval, recommendation or reimbursement. For many reasons, including the rapid pace of technological change, heterogeneous applications and different incentives for evidence generation, the medical AI field has not yet developed comparable norms. Although regulatory frameworks are the subject of ongoing debate and development, they remain inadequate[3]. Published studies often emphasize technical validity over clinical usefulness[4]. Implementation decisions are frequently made before core questions of actionability, feasibility, safety and effectiveness have been adequately addressed[5]. In the absence of a consensus on evidentiary standards, those decisions may rely more on early adoption enthusiasm than on consistent criteria. Without clearer rules and a direct mandate to provide robust evidence, the threshold for claiming value remains too variable.

Going forward, the medical AI field must develop a consistent framework to connect claims of clinical value of an AI tool to the appropriate type of evidence needed to support those claims. For example, claims of analytic performance should require robust validation in the intended setting and population, whereas claims of clinical actionability should require evidence that outputs are interpretable and can support reasonable decisions. Claims of workflow benefit should require implementation studies showing that tools can be integrated without the introduction of delay, burden or unintended harms. Claims of improved outcomes or efficiency should require stronger prospective evidence, including comparative evaluations to standard of care, where appropriate. Moreover, because model performance may shift over time, post-deployment monitoring should be seen as an institutional expectation rather than as a late, optional addition.

Having such a framework does not mean that every AI tool must undergo all the staged phases of testing up to a randomized controlled trial before adoption, as is usually required for other medical interventions. In many cases, that would be impractical, given the high costs, the rapid updating of the models underpinning the tools, and the overall complexity and time needed to conduct such studies. At the same time, accepting retrospective performance alone as a sufficient basis for trust is not scientifically rigorous. Therefore, the goal should be proportional evidence, meaning that the stronger the claim, the stronger the evidence needed to support it.

This principle has practical implications for all stakeholders. For example, regulators should better clarify which categories of medical AI tools require prospective evidence of clinical impact and which can enter practice under more limited claims. Healthcare organizations and administrators should distinguish among pilot implementation, operational use and evidence of benefit, rather than collapsing these into a single decision. Across these settings, evidence standards should be transparent, claim specific, and open to revision as tools evolve.

Scientific journals, as part of the research ecosystem, have a unique opportunity to define acceptable types of evidence. In emerging fields, the published literature is often viewed as establishing what constitutes valid evidence for an area of research or practice. By enforcing proportional evidentiary standards, journals can ensure that published research reflects genuine clinical claims rather than mere technical promise, a role we will continue to support at *Nature Medicine*.

The next phase of progress will depend not only on better models and new applications but also on clearer expectations for how clinical impact is defined, evaluated and communicated. Without a clear connection between claims and evidence, medical AI risks being adopted faster than its real value can be understood.

Published online: 21 April 2026

References

1. Hwang, Y. M., Ng, M. Y., Pillai, M., Sahai, M. P. & Hernandez-Boussard, T. *Nat. Health* **1**, 99–112 (2026).
2. Costa-Gomes, B. et al. Preprint at https://doi.org/10.48550/arXiv.2512.11879 (2025).
3. Freyer, O. et al. *Nat. Med.* **31**, 3239–3243 (2025).
4. Chouffani El Fassi, S. et al. *Nat. Med.* **30**, 2718–2720 (2024).
5. Angus, D. C. et al. *JAMA* **334**, 1650–1664 (2025).

The Future is Now

Healthcare is at a Crossroads

FINANCIAL PRESSURE

WORKFORCE SHORTAGE

INCREASED COMPETITION
INCLUDING NEW ENTRANTS

Future of genAI: facilitate safe, effective, accessible care

- Access orchestration with agents
- Prescription refills
- Clinical diagnostic decision making
  - Colon polyps
  - Screening: skin cancer, breast cancer, cervical cancer, lung cancer, diabetic retinopathy

What are we, as a society, ready to let AI do autonomously vs. with a human in the loop vs. not yet, or ever?

How will AI continue to augment the people & processes?

Questions

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

Rebecca G. Mishuris, MD, MS, MPH, FAMIA
rmishuris@mgb.org