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

AI is Not Coming for Your Job*

Andrew D.A. Marshall, MD MBI
Senior Clinical Specialist @ Google
Department of Emergency Medicine
Brigham and Women's Hospital
Instructor of Emergency Medicine
Harvard Medical School

CONTINUING MEDICAL EDUCATION
DEPARTMENT OF MEDICINE

HARVARD MEDICAL SCHOOL
TEACHING HOSPITAL

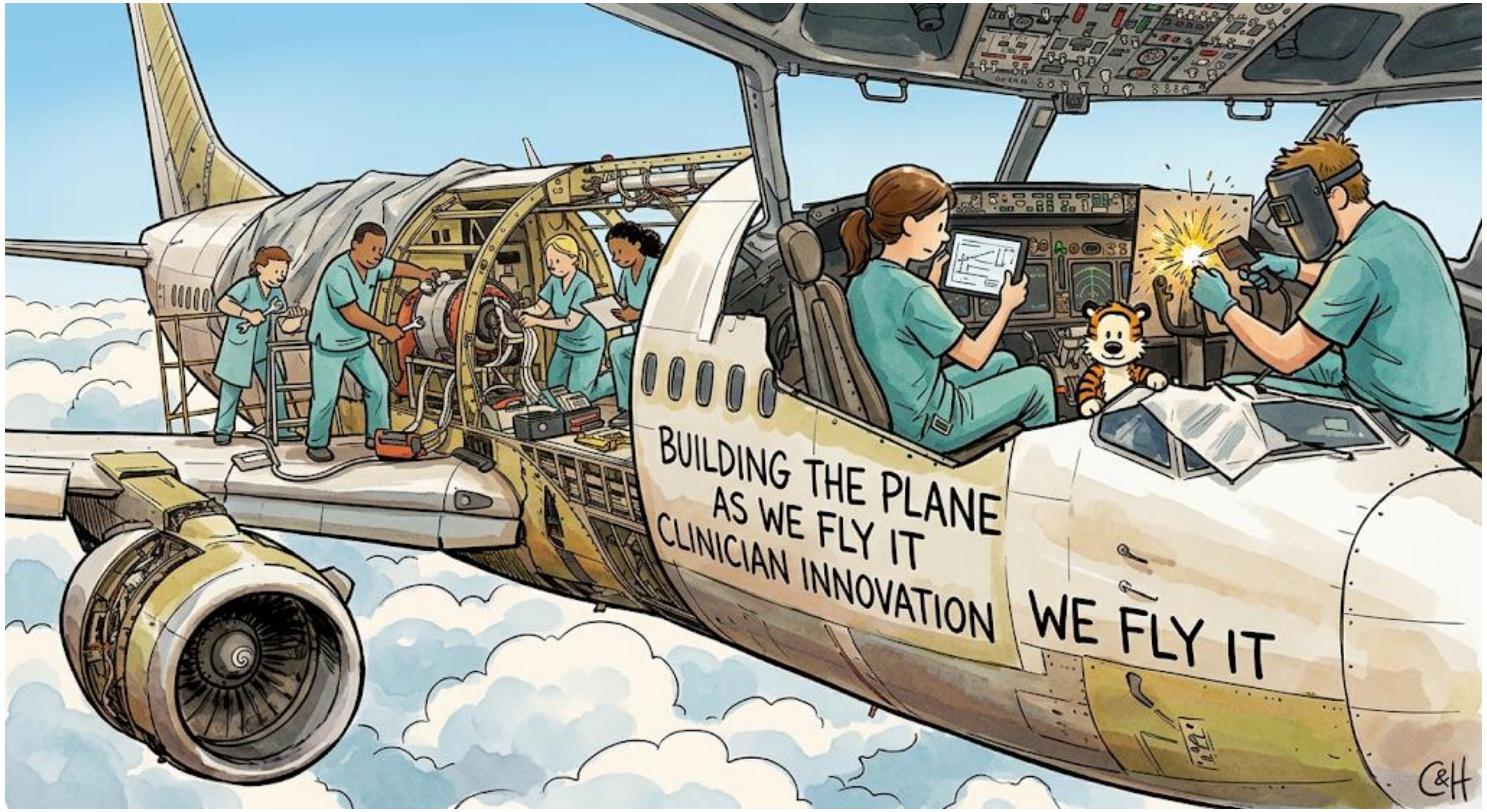
(Drew) Andrew D.A. Marshall, MD MBI

Meharry Medical College
EM Residency at The University of Chicago
Clinical Informatics Fellowship at BIDMC
Masters in Bioinformatics at HMS
Instructor of Emergency Medicine at HMS
Clinical focus: Clinical Artificial Intelligence

DISCLOSURES

I am a full time employee of Google, LLC; My work directly contributes to Gemini and other Google products

This presentation exists in an information dynamic environment



This is not an Anti-AI talk

It is very much the opposite

AI is here to stay, and (hopefully) so are we. The best way to get comfortable with a technology is to understand its strengths and weaknesses. I hope to explore some of these with you during this conversation

The headlines

Latest

JUL 14, 2023, 06:00 ET

OpenEvidence AI Becomes the First AI in History to Score Above 90% on the United States Medical Licensing Examination (USMLE)

AI AND MACHINE LEARNING

OpenEvidence AI scores 100% on USMLE as company launches free explanation model for medical students

By Heather Landi · Aug 15, 2025 8:00am

AI is starting to beat doctors at making correct diagnoses

Large language model excels at clinical decisions, even in f

30 APR 2026 · 2:00 PM ET · BY PERRI THALER

AI outperforms doctors in Harvard trial of emergency triage diagnoses

Researchers say results mark a 'profound change in technology that will reshape medicine'

OBJECTIVES

Objective 1: Evaluating "Super Clinical" Performance in AI Literature
Participants will be able to appraise AI literature by identifying the limitations of proxy metrics, such as standardized test scores or text-based vignettes, in representing complex clinical reasoning. They will learn to distinguish between models trained to replicate human-generated "ground truth" labels versus those evaluated on definitive, downstream patient outcomes. Finally, attendees will understand the critical gap between retrospective dataset validation and prospective clinical trial efficacy in real-world hospital environments.

Objective 2: Understanding the Limitations of AI for Clinical Use
Clinicians will be able to recognize the cognitive risks of integrating AI into daily practice, specifically focusing on automation bias triggered by the articulate, confident nature of Large Language Models. They will understand the vulnerability of algorithms to "domain shift," where subtle changes in patient demographics or institutional workflows silently degrade diagnostic accuracy. Furthermore, participants will learn to anticipate model hallucinations stemming from an AI's reliance on statistical pattern-matching rather than true pathophysiological and causal reasoning.

Impressive ‘In Silico’ performance and real clinical benefit are not the same thing

Roadmap:

1. A look at the evidence
2. Understanding the risks when we use our models
3. Towards safer clinical AI

The Path to Medical Superintelligence

Dataset:
- Complete
- Retrospective
- Unimodal
- Memorization vs Reasoning

GPT-4 outperformed 99.98% of simulated human readers in diagnosing complex clinical cases

A study published in the New England Journal of Medicine says GPT-4 correctly diagnosed 52.7% of complex cases compared to only 36% of medical journal readers.

In just three years, generative AI has advanced to the point of scoring near-perfect scores on the USMLE and similar exams. But these tests primarily rely on multiple-choice questions, which favor memorization over deep understanding. By reducing medicine to one-shot answers on multiple-choice questions, such benchmarks overstate the apparent competence of AI systems and obscure their limitations.

Clinical Example

The patient is a 28-year-old Caucasian female residing in San Diego, California. She works from home as a freelance Financial Advisor and is currently **undergoing a highly stressful divorce trial**. Her chief complaint centers on **recurrent episodes of shakiness, a rapidly beating heart, and near-fainting**, which uniquely **seem to occur every time she takes a shower**, alongside new and profound sleep disturbances.

She reports that the simple **act of standing under the warm water triggers a sudden and intense racing sensation in her chest.** During these episodes, she becomes visibly shaky and experiences such profound lightheadedness that she feels she is on the verge of passing out. To prevent herself from fainting, she finds she must turn off the water and lie down on the floor immediately, which is the only way to gradually ease her racing heart and dizziness. Alongside these frightening spells, she is experiencing severe changes to her sleep patterns. She describes having great difficulty falling asleep and an inability to stay asleep, noting that on the day she sought guidance, she had been wide awake since 2:30 a.m. She denies experiencing any radiating chest pain, localized muscle weakness, or shortness of breath while resting in bed.

Her past medical history is unremarkable, with no chronic illnesses or prior hospitalizations, and she does not take any daily prescription medications. Given the intense legal proceedings surrounding her ongoing divorce trial, **she initially attributed her racing heart, exhaustion, and chronic insomnia solely to severe emotional stress and anxiety**. She leads a generally quiet lifestyle, does not smoke, and drinks only occasionally socially. While the emotional toll of her personal life is high, the highly specific and predictable physical trigger of standing in a warm environment has prompted her to look for a physiological explanation for her symptoms.

Rules:

Initiated by a chief complaint
8 turn limit
Text based conversation

Clinician DDx:

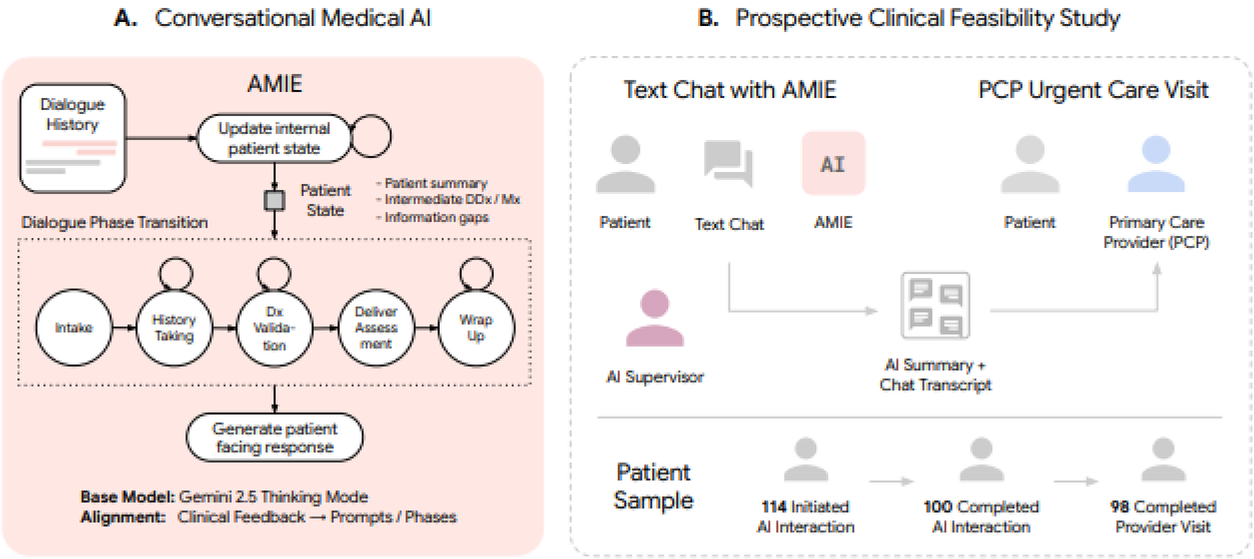
1. Vasovagal presyncope
2. Orthostatic hypotension
3. Anxiety episodes
4. Hyperthyroidism

'In Vivo'

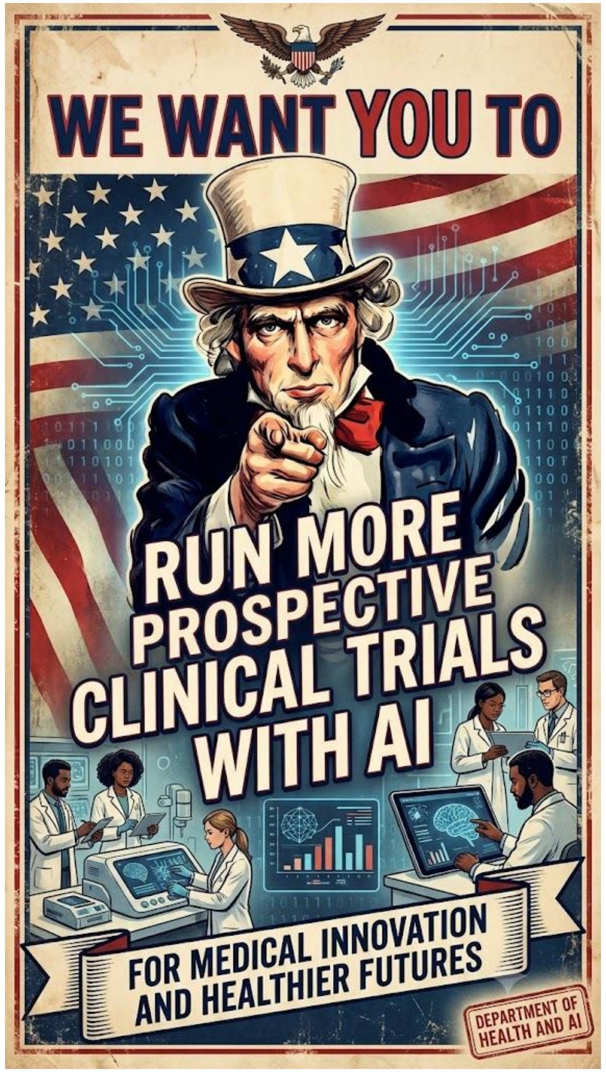
A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic

Peter Brodeur[*,3], Jacob M. Koshy[*,3], Anil Palepu[*,1], Khaled Saab[2], Ava Homiar[3], Roma Ruparel[1], Charles Wu[3], Ryutaro Tanno[2], Joseph Xu[1], Amy Wang[1], David Stutz[2], Wei-Hung Weng[2], Hannah M. Ferrera[3], David Barrett[2], Lindsey Crowley[3], Jihyeon Lee[1], Spencer E. Rittner[4], Ellery Wulczyn[1], Selena K. Zhang[5], Elahe Vedadi[2], Christine G. Kohn[6], Kavita Kulkarni[1], Vinay Kadiyala[3], S. Sara Mahdavi[2], Wendy Du[3], Jessica M. Williams[1], David Feinbloom[3], Renee Wong[1], Tao Tu[2], Petar Sirkovic[1], Alessio Orlandi[1], Christopher Semturs[1], Yun Liu[1], Juraj Gottweis[1], Dale R. Webster[1], Joëlle Barral[2], Katherine Chou[1], Pushmeet Kohli[2], Avinatan Hassidim[1], Yossi Matias[1], James Manyika[1], Rob Fields[4], Jonathan X. Li[3], Marc L. Cohen[†,3], Vivek Natarajan[†,2], Mike Schaekermann[†,1], Alan Karthikesalingam[†,2] and Adam Rodman[†,1,3]

[*]Equal contributions, [†]Equal leadership, [1]Google Research, [2]Google DeepMind, [3]Beth Israel Deaconess Medical Center, [4]Beth Israel Lahey Health, [5]Harvard Medical School, [6]Massachusetts General Hospital



‘In Vivo’

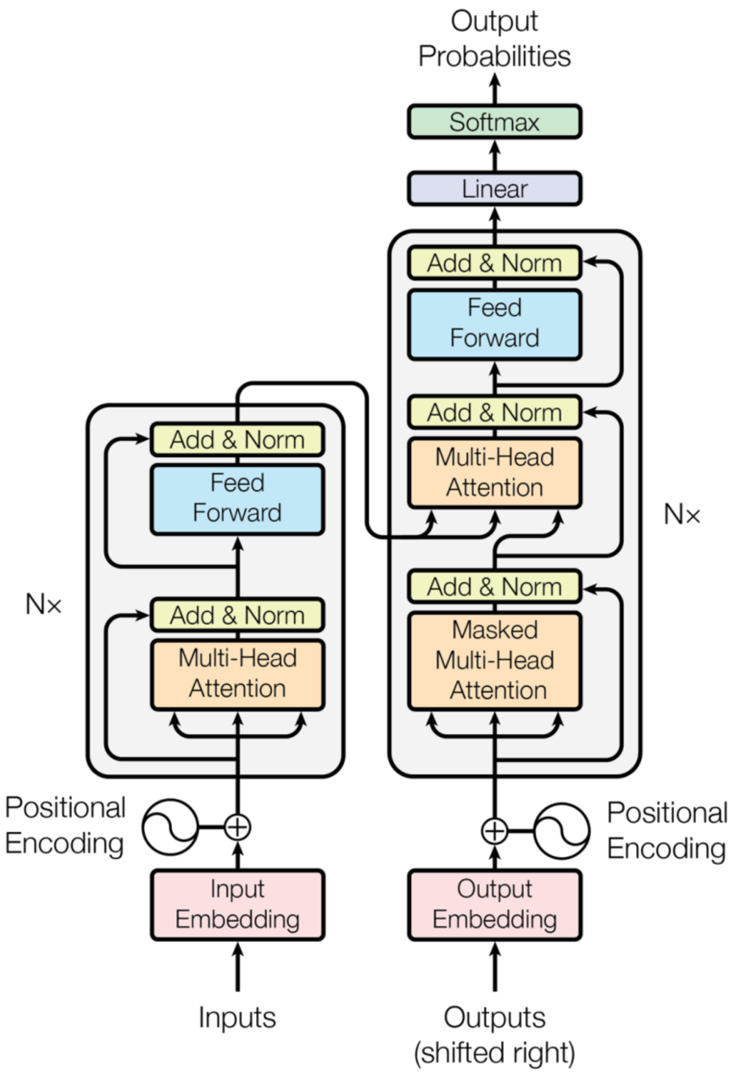


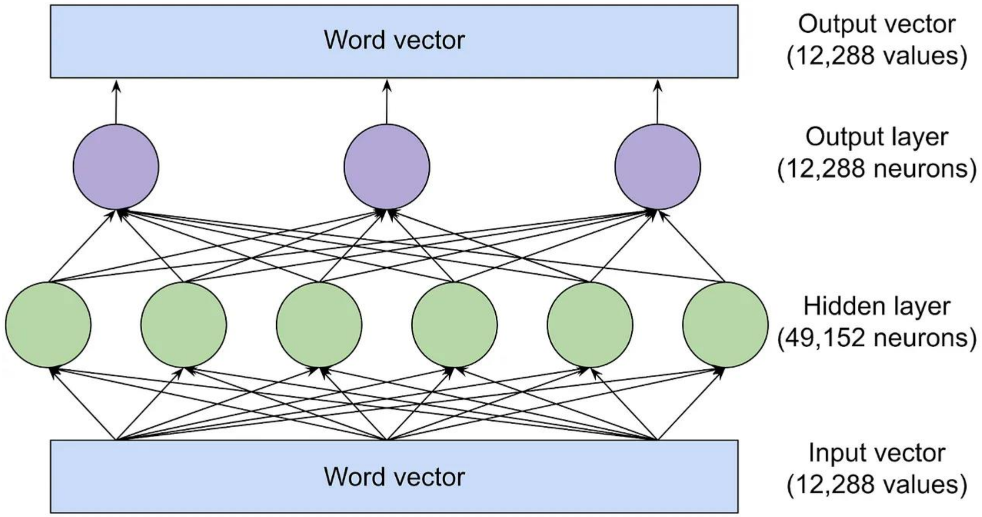
Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review

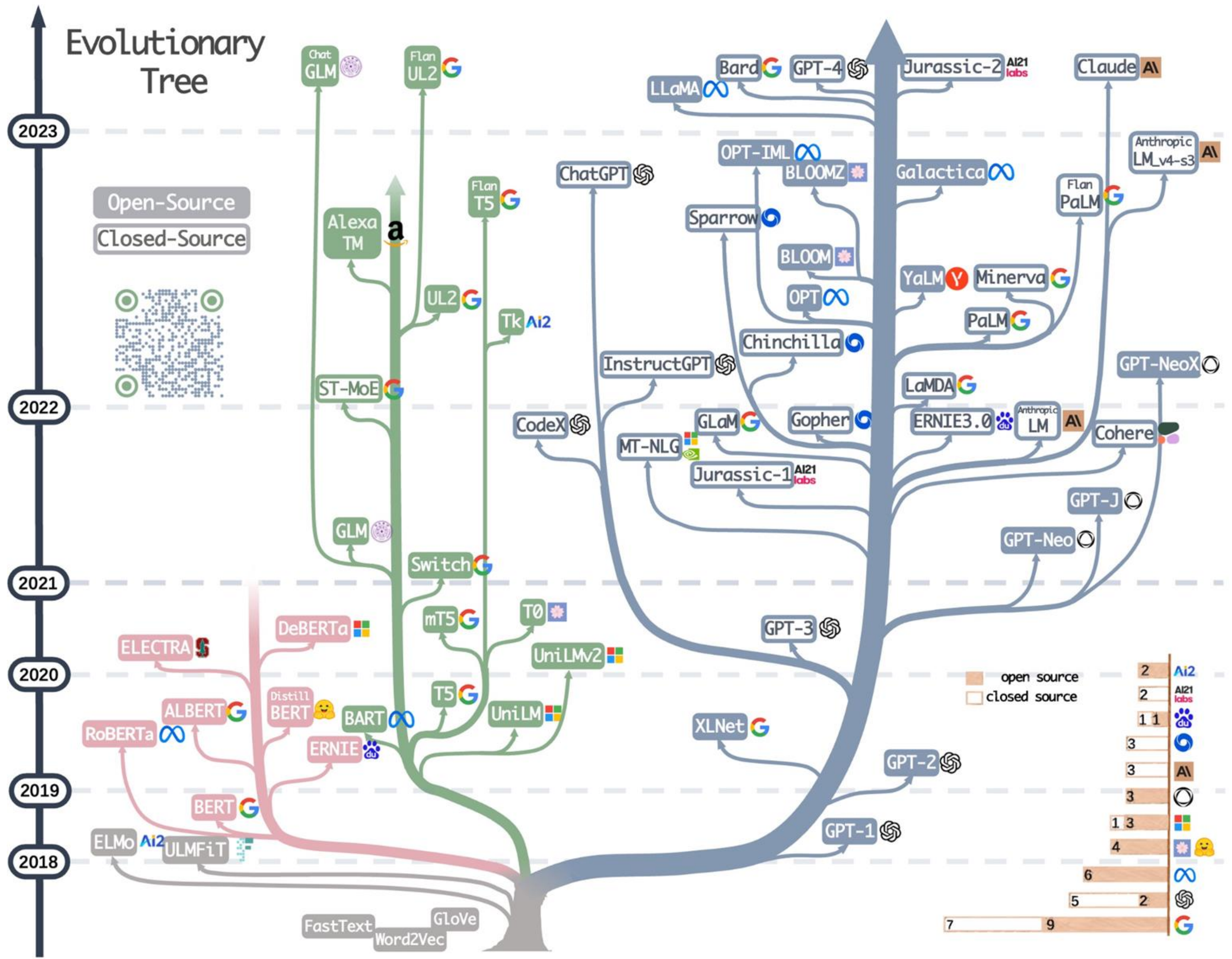
Ryan Han, Julián N Acosta, Zahra Shakeri, John P A Ioannidis, Eric J Topol*, Pranav Rajpurkar*

This scoping review of randomised controlled trials on artificial intelligence (AI) in clinical practice reveals an expanding interest in AI across clinical specialties and locations. The USA and China are leading in the number of trials, with a focus on deep learning systems for medical imaging, particularly in gastroenterology and radiology. A majority of trials (70 [81%] of 86) report positive primary endpoints, primarily related to diagnostic yield or performance; however, the predominance of single-centre trials, little demographic reporting, and varying reports of operational efficiency raise concerns about the generalisability and practicality of these results. Despite the promising outcomes, considering the likelihood of publication bias and the need for more comprehensive research including multicentre trials, diverse outcome measures, and improved reporting standards is crucial. Future AI trials should prioritise patient-relevant outcomes to fully understand AI’s true effects and limitations in health care.

How LLMs work - “Attention is all you Need”







Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
ChatGLM
Flan UL2
Alexa TM
Flan T5
UL2
Tk
ST-MoE
GLM
Switch
mT5
T0
T5
UniLMv2
UniLM
BART
DeBERTa
ELECTRA
ALBERT
Distill BERT
RoBERTa
ERNIE
BERT
ELMo
ULMFiT
FastText
Word2Vec
GloVe
LLaMA
Bard
GPT-4
Jurassic-2
Claude
OPT-IML
BLOOMZ
Galactica
Anthropic LM_v4-s3
ChatGPT
Sparrow
Flan PaLM
BLOOM
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GPT-Neo
GPT-3
XLNet
GPT-2
GPT-1
open source
closed source
2
2
1 1
3
3
3
1 3
4
6
5 2
7 9

CHAPTER 1

AI capability is outpacing safety evaluation

01

Measure safety as its own dimension, including with dedicated safety benchmarks such as NOHARM, because a more capable model is not automatically a safer one.

On NOHARM, a clinical safety benchmark, even the best-performing models produced recommendations with potential for severe harm in roughly 1 in 14 consultations, with failures of omission driving most serious errors. Strong performance on capability benchmarks does not predict safe clinical behavior.

How do we measure performance?

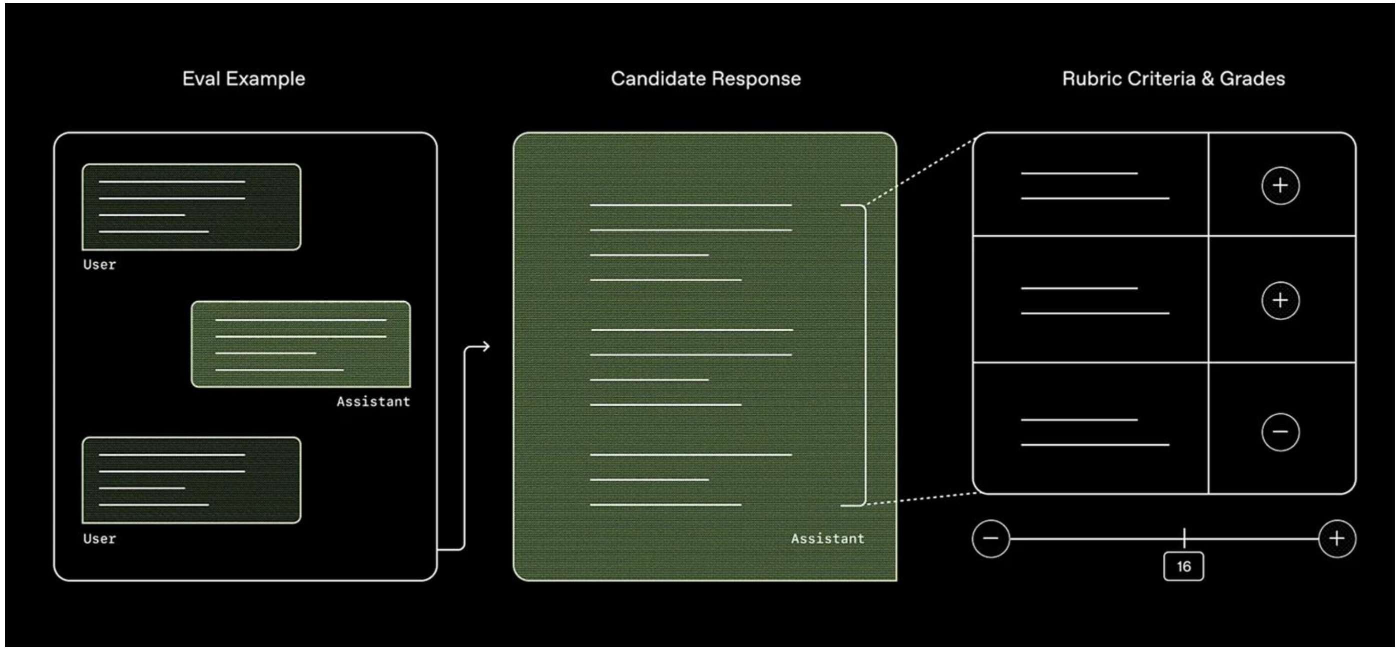
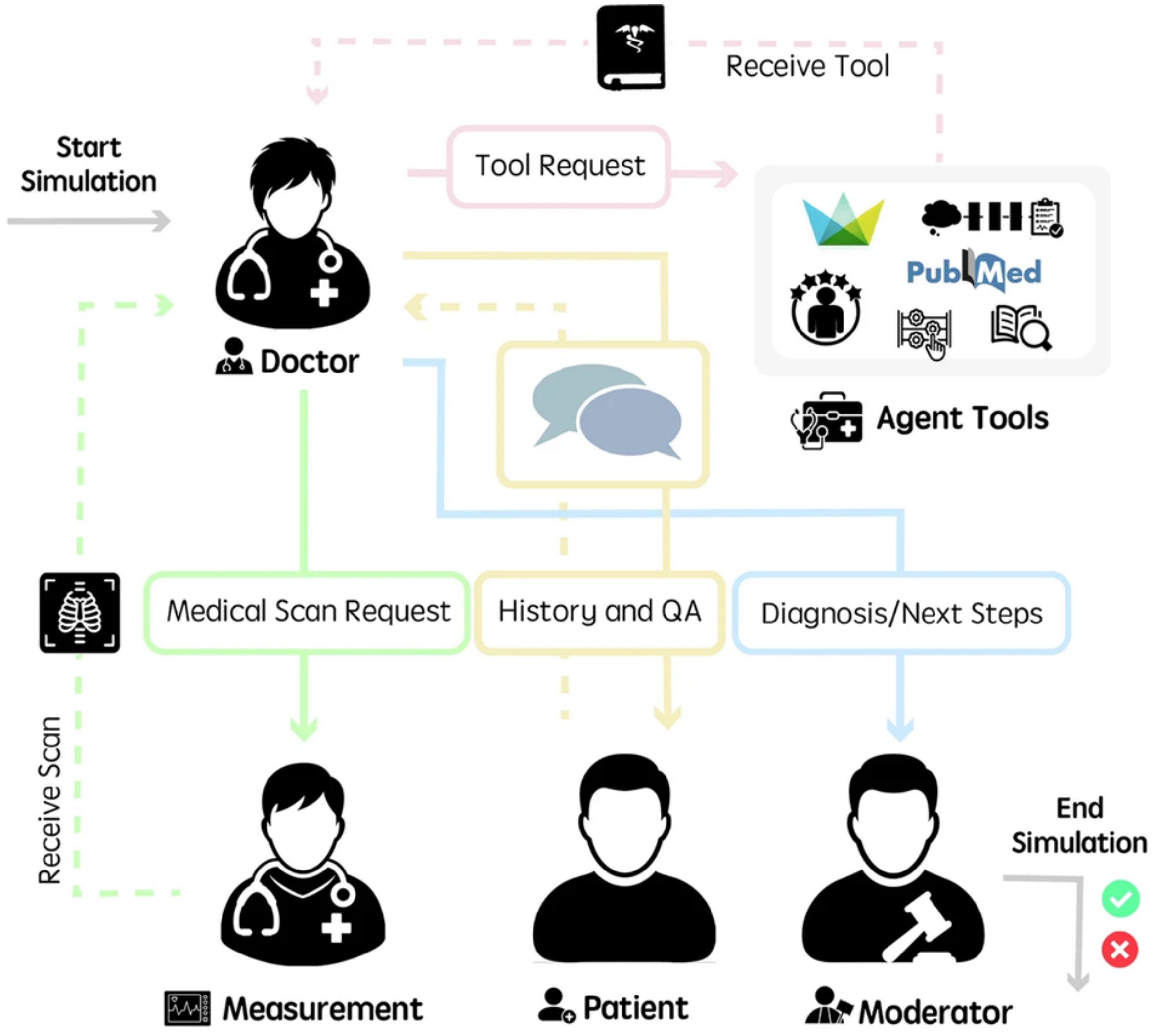
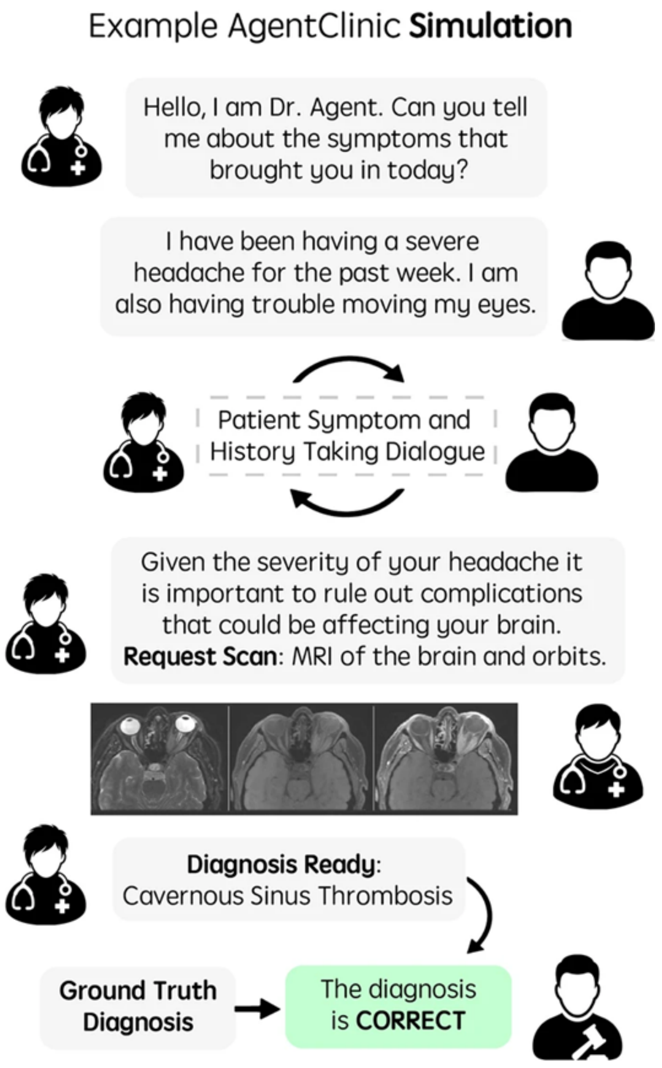
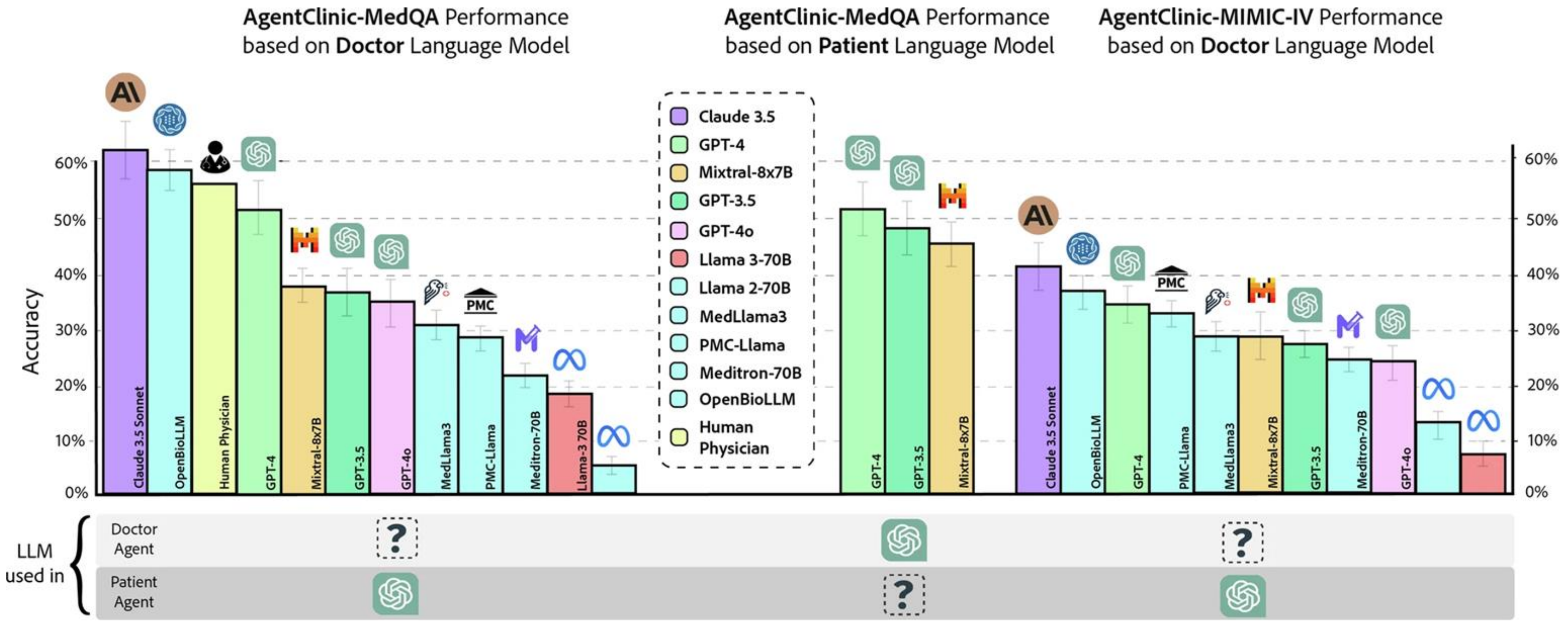


Table 1: **Comprehensive Catalog of Benchmarks for Medical Agent Readiness.** Benchmarks organized along three levels o
increasing autonomy and clinical realism.

Year	Benchmark	Key Contribution	Venue	Paper	Data/Code
Level 1: Isolated Single Capabilities					
2018	VQA-RAD [123]	Clinician generated naturalistic radiology visual question answering	Sci. Data	Paper	Data
2019	PubMedQA [108]	Biomedical yes no maybe reasoning over abstracts	EMNLP 2019	Paper	Data
2020	PathVQA [87]	Pathology VQA, 32K questions over 5K images	arXiv	Paper	Data
2021	MedQA [107]	Free form licensing exam open domain question answering	AAAI 2021	Paper	Data
2021	SLAKE [143]	Bilingual knowledge enhanced radiology VQA with semantic labels	ISBI 2021	Paper	Data
2022	MedMCQA [185]	Multisubject entrance exam medical multiple choice QA	CHIL 2022	Paper	Data
2022	MS-CXR [18]	Radiologist phrase grounding boxes for chest radiographs	ECCV 2022	Paper	Data
2023	ReXVal [291]	Radiologist error annotations for report evaluation	Patterns	Paper	Data
2024	CheXbench [38]	Chest X-ray perception and report understanding, 8 tasks	arXiv	Paper	Code
2024	OmniMedVQA [96]	Large scale medical VQA, 118K images, 12 modalities	CVPR 2024	Paper	Data
2024	MedSafetyBench [84]	Medical safety benchmark, 1,800 harmful request demonstrations	NeurIPS 2024	Paper	Code
2024	CARES [267]	Med LVLM trustworthiness across five reliability dimensions	NeurIPS 2024	Paper	Code
2024	ClinicalBench [26]	Tests whether LLMs surpass classical clinical prediction models	arXiv	Paper	Code
2025	MedXpertQA [327]	Expert level reasoning, 4,460 text and multimodal questions	ICML 2025	Paper	Data
2025	MedHallu [190]	First medical hallucination detection benchmark, 10K pairs	arXiv	Paper	Data
2025	HealthBench [8]	Physician rubric graded multiturn health conversations, 5K cases	arXiv	Paper	Data
2025	MedReason [262]	KG-grounded medical reasoning chains, 32K QA	arXiv	Paper	Code
2025	GEMeX [144]	Grounded, explainable chest X-ray VQA (1.6M questions)	ICCV 2025	Paper	Data


Receive Tool
Start Simulation
Tool Request
Doctor
Agent Tools
PubMed
Medical Scan Request
History and QA
Diagnosis/Next Steps
Receive Scan
End Simulation
Measurement
Patient
Moderator

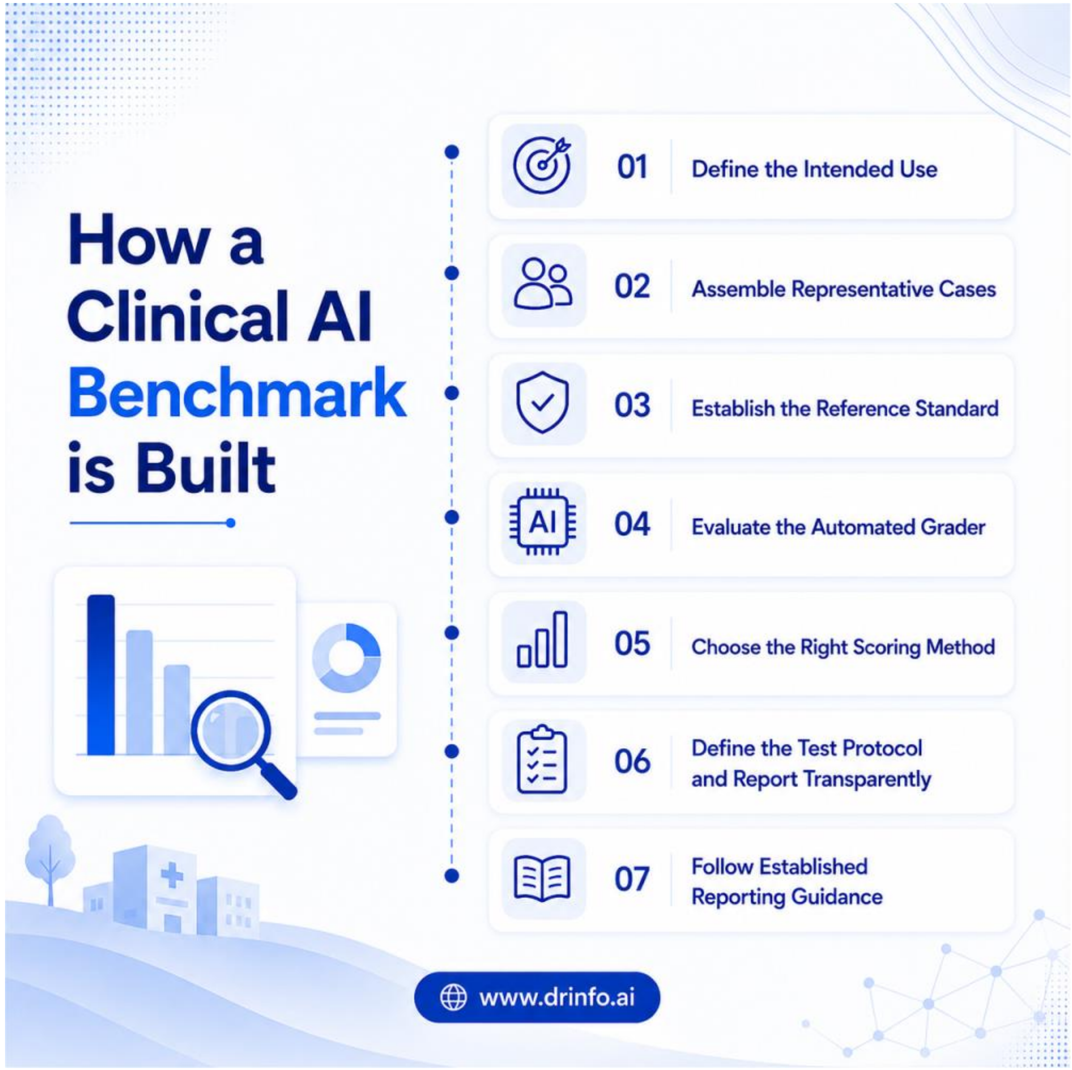

Example AgentClinic **Simulation**
Hello, I am Dr. Agent. Can you tell me about the symptoms that brought you in today?
I have been having a severe headache for the past week. I am also having trouble moving my eyes.
Patient Symptom and History Taking Dialogue
Given the severity of your headache it is important to rule out complications that could be affecting your brain. **Request Scan**: MRI of the brain and orbits.
Diagnosis Ready: Cavernous Sinus Thrombosis
Ground Truth Diagnosis
The diagnosis is **CORRECT**



AgentClinic-MedQA Performance based on Doctor Language Model
AgentClinic-MedQA Performance based on Patient Language Model
AgentClinic-MIMIC-IV Performance based on Doctor Language Model
Accuracy
60%
50%
40%
30%
20%
10%
0%
Claude 3.5
GPT-4
Mixtral-8x7B
GPT-3.5
GPT-4o
Llama 3-70B
Llama 2-70B
MedLlama3
PMC-Llama
Meditron-70B
OpenBioLLM
Human Physician
Claude 3.5 Sonnet
OpenBioLLM
Human Physician
GPT-4
Mixtral-8x7B
GPT-3.5
GPT-4o
MedLlama3
PMC-Llama
Meditron-70B
Llama-3 70B
PMC
LLM used in
Doctor Agent
Patient Agent
?

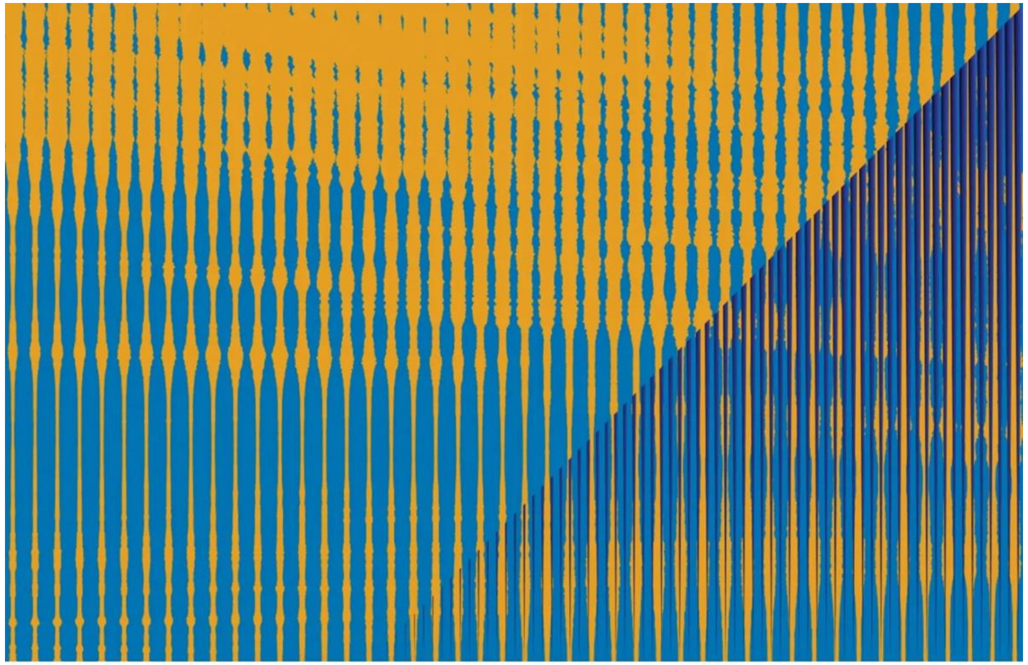
Do benchmarks reflect clinical practice?

Largely single single modality (text) with some emerging multi modality
Single pass output without capability for iterative reasoning



The 2026 Healthcare AI Industry Report translates the rapidly expanding evidence base into practical guidance for industry leaders as healthcare AI moves toward deployment at scale.

Ethan Goh, Adam Rodman, Jonathan H Chen



Healthcare AI Industry Report

JULY 2026

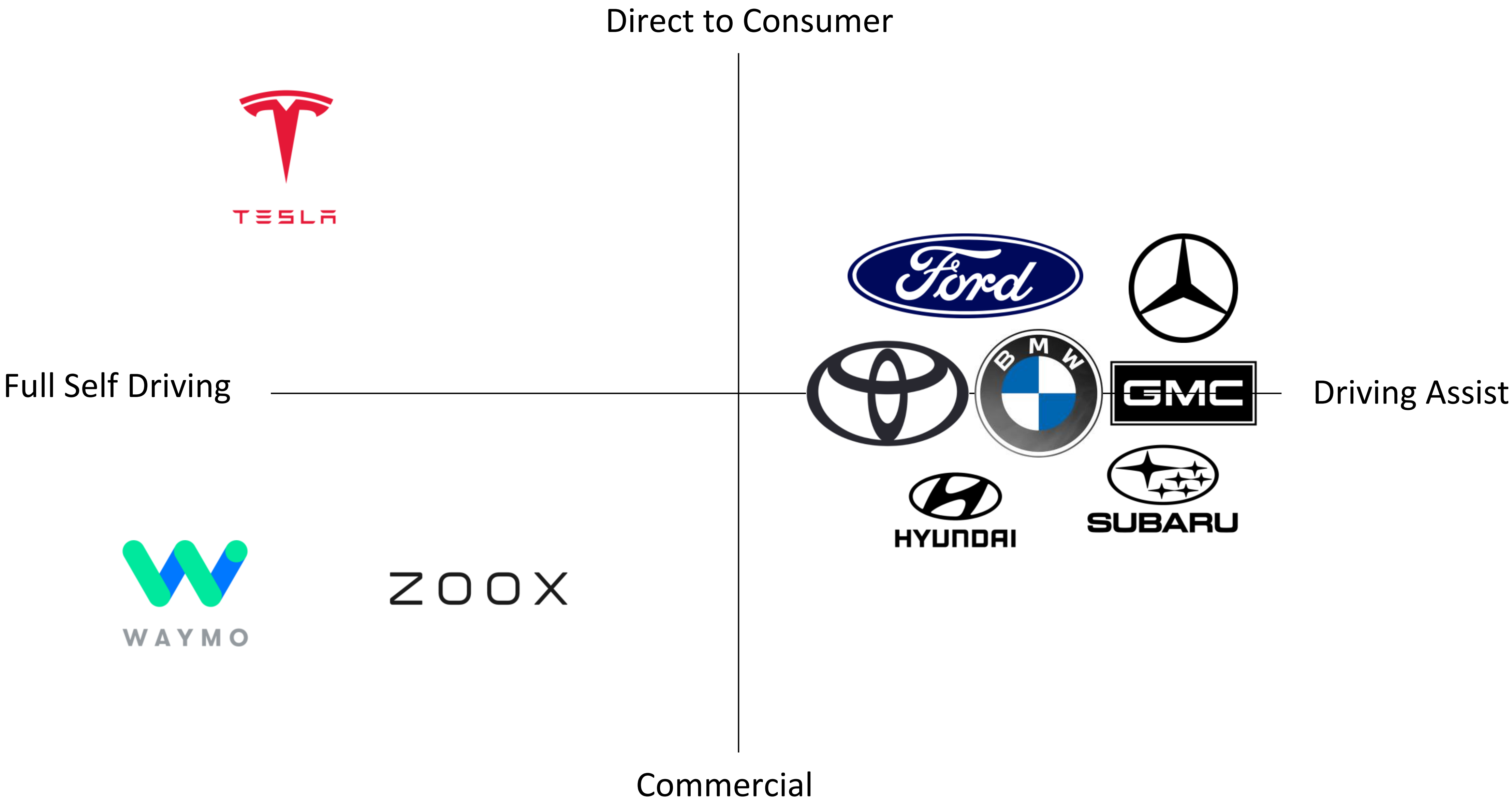
ARISE

Human-In-The-Loop

05

Preserve clinician agency where human oversight is required.

Clinicians should be trained to interrogate AI outputs, probe uncertainty, and override recommendations when appropriate. At the same time, requiring human review for every AI-supported task can create the appearance of safety without delivering it, especially where clinician time, access, or resources are constrained. The central question is not whether humans should always remain “in the loop,” but which tasks can safely move toward greater autonomy, under what safeguards, and with what escalation pathways. Answering that question requires a taxonomy of clinical and administrative work, paired with measurement of current human performance, risk, cost, and access constraints.

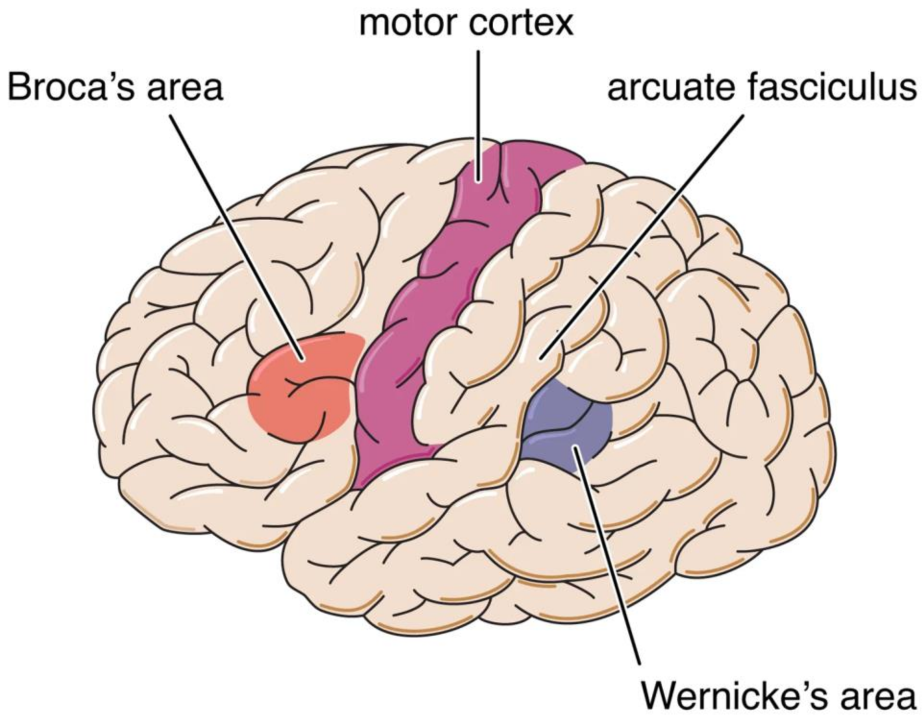


Direct to Consumer
Full Self Driving
Driving Assist
Commercial
TESLA
Ford
BMW
GMC
HYUNDAI
SUBARU
WAYMO
ZOOX

We think hallucinations make AI different from us. What if they make us the same?

Spinning stories out of nothing is a deeply human impulse. As a neurologist, I see my patients do it all the time.

By Pria Anand Updated September 24, 2025, 6:00 a.m.



'Confidence' is key

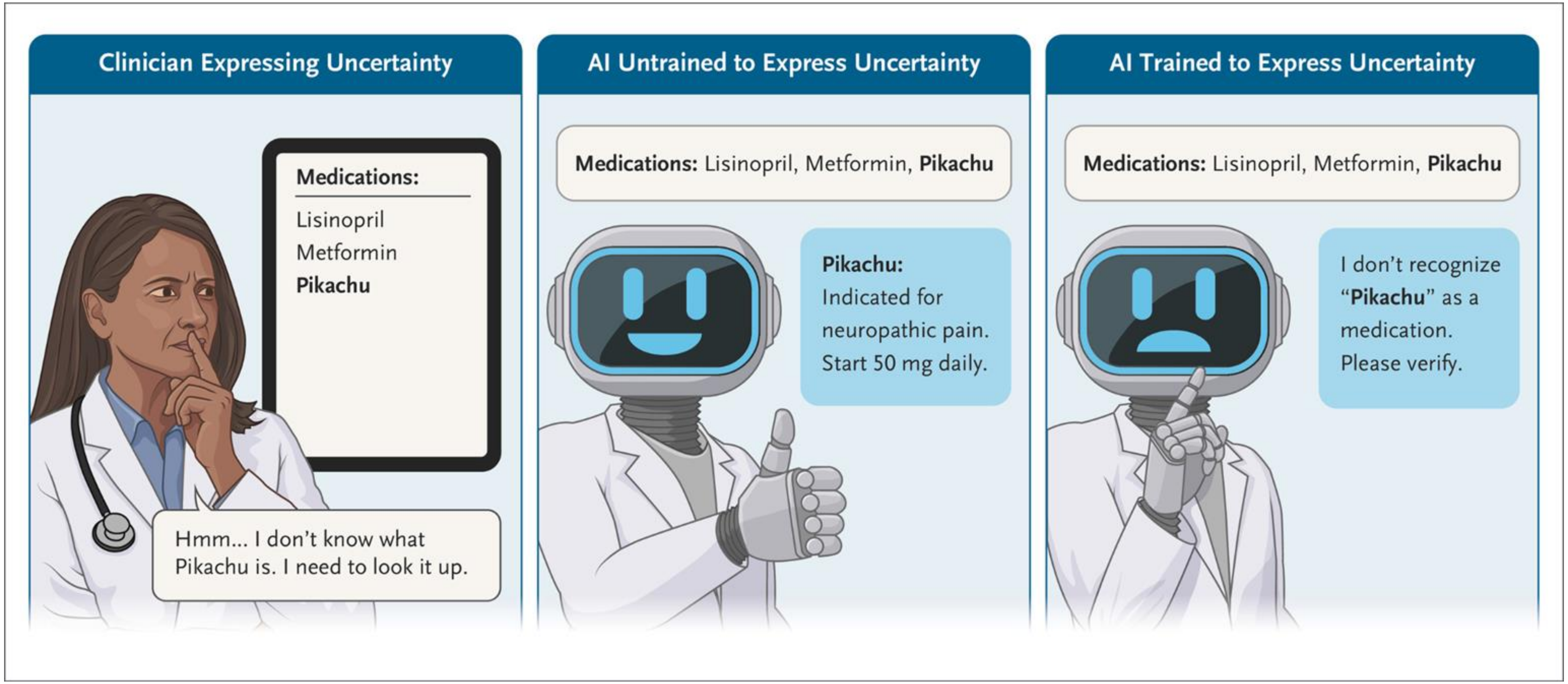
The NEW ENGLAND JOURNAL of MEDICINE

PERSPECTIVE

Can AI Say "I Don't Know"?

Authors: Andrea Sikora, Pharm.D., M.S.C.R., Leo A. Celi, M.D., M.P.H., and Raja-Elie E. Abdulnour, M.D. Author Info & Affiliations

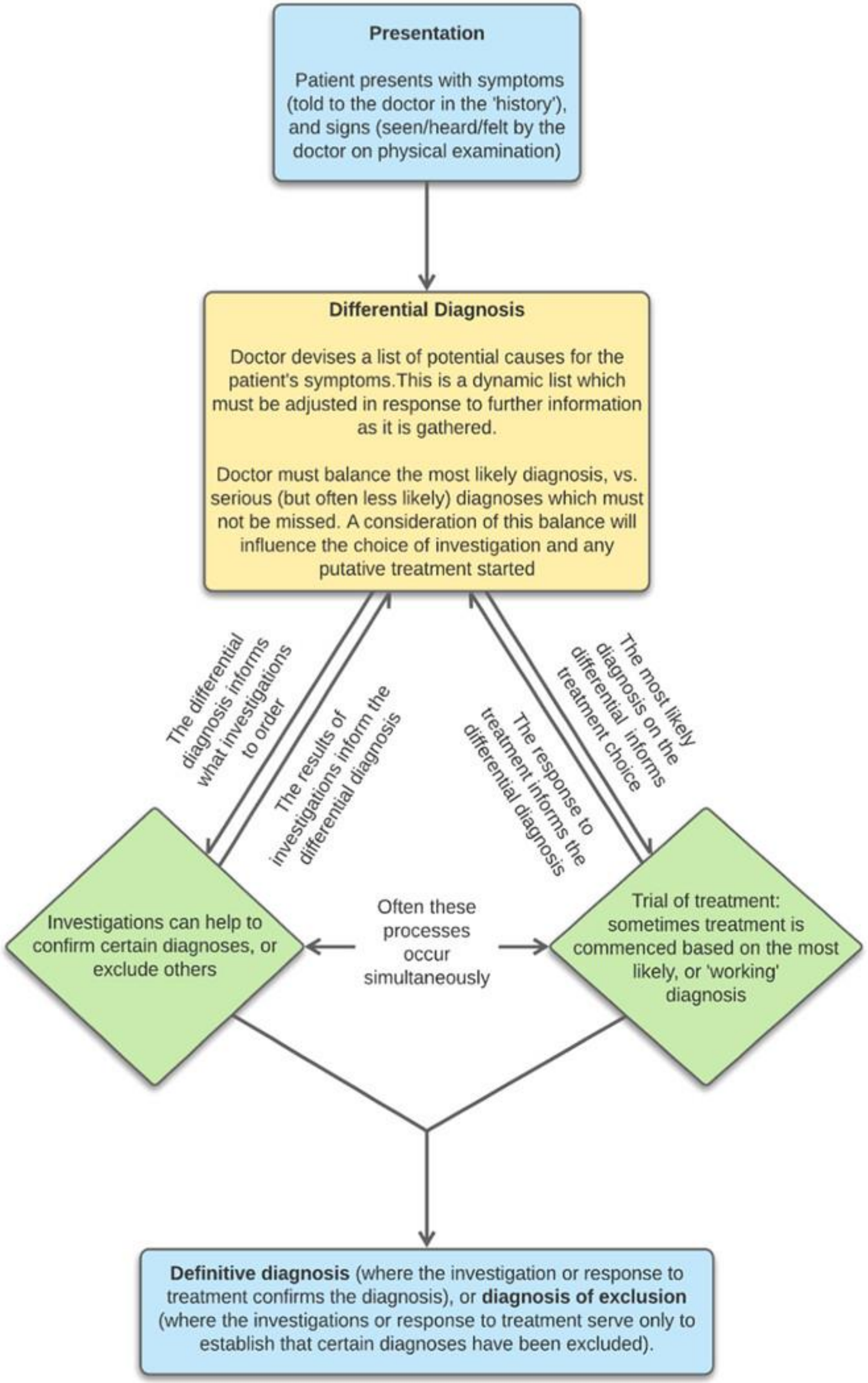
Published May 9, 2026 | N Engl J Med 2026;394:1873-1875 | DOI: 10.1056/NEJMp2517624 | VOL. 394 NO. 19



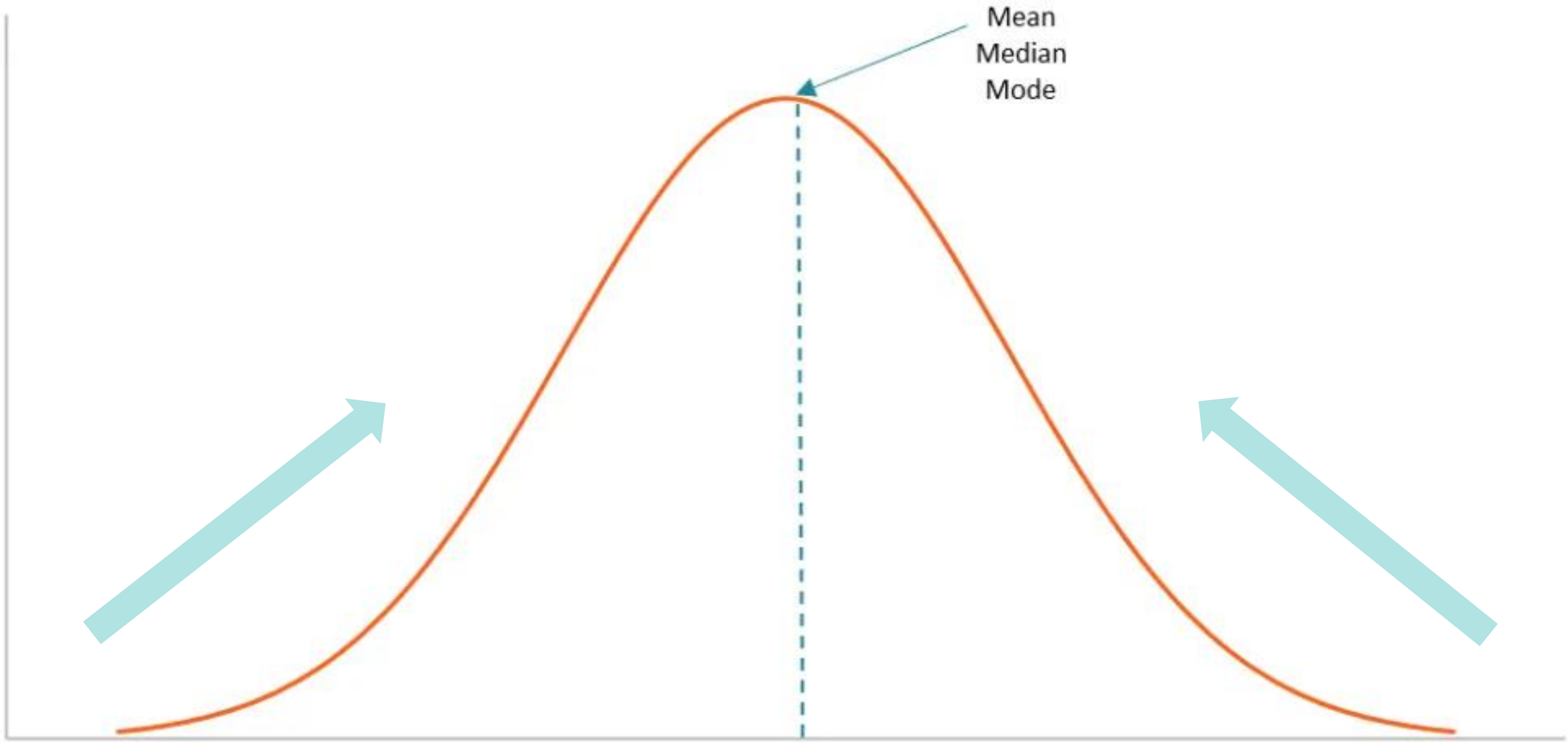
‘Confidence’ is key

Uncertainty allows us to:

- acknowledge limitations
- collect additional information
- present multiple options

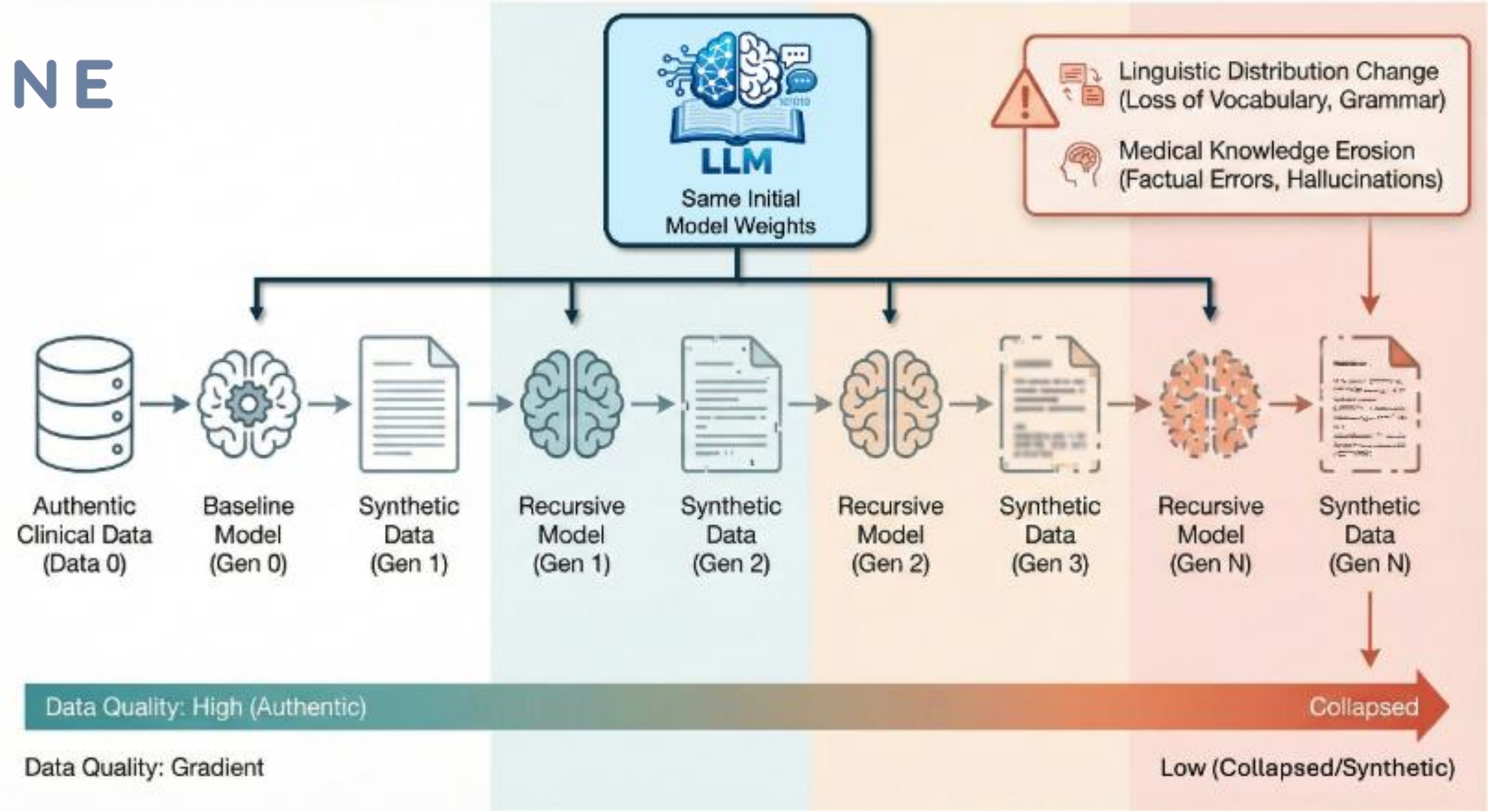


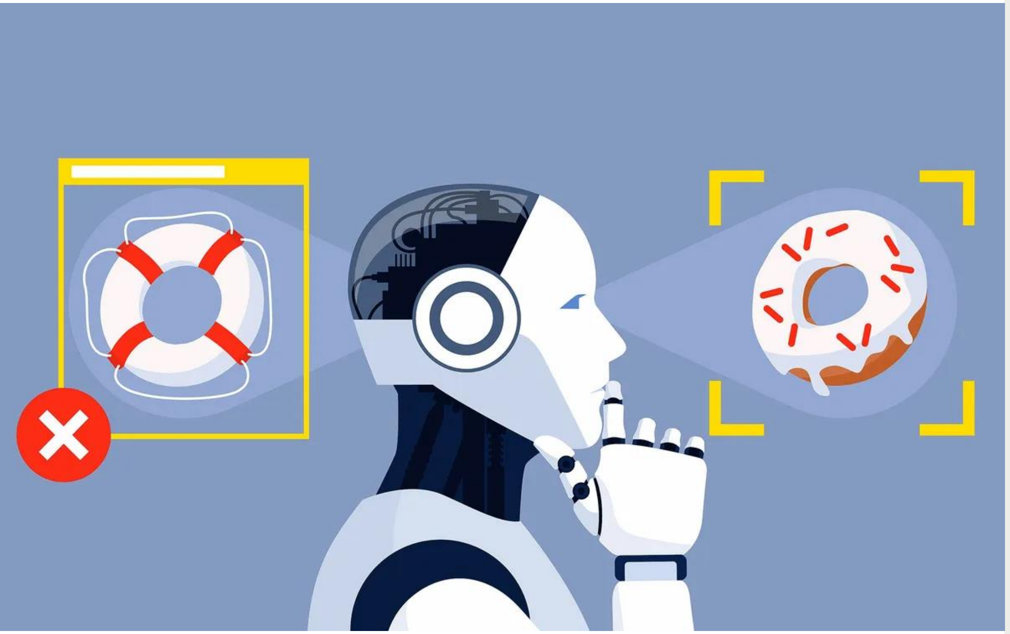
Performance Bell Curve* (Acuity)



low acuity high acuity

Model Collapse - Clinical Diversity Reduction (loss of tails)





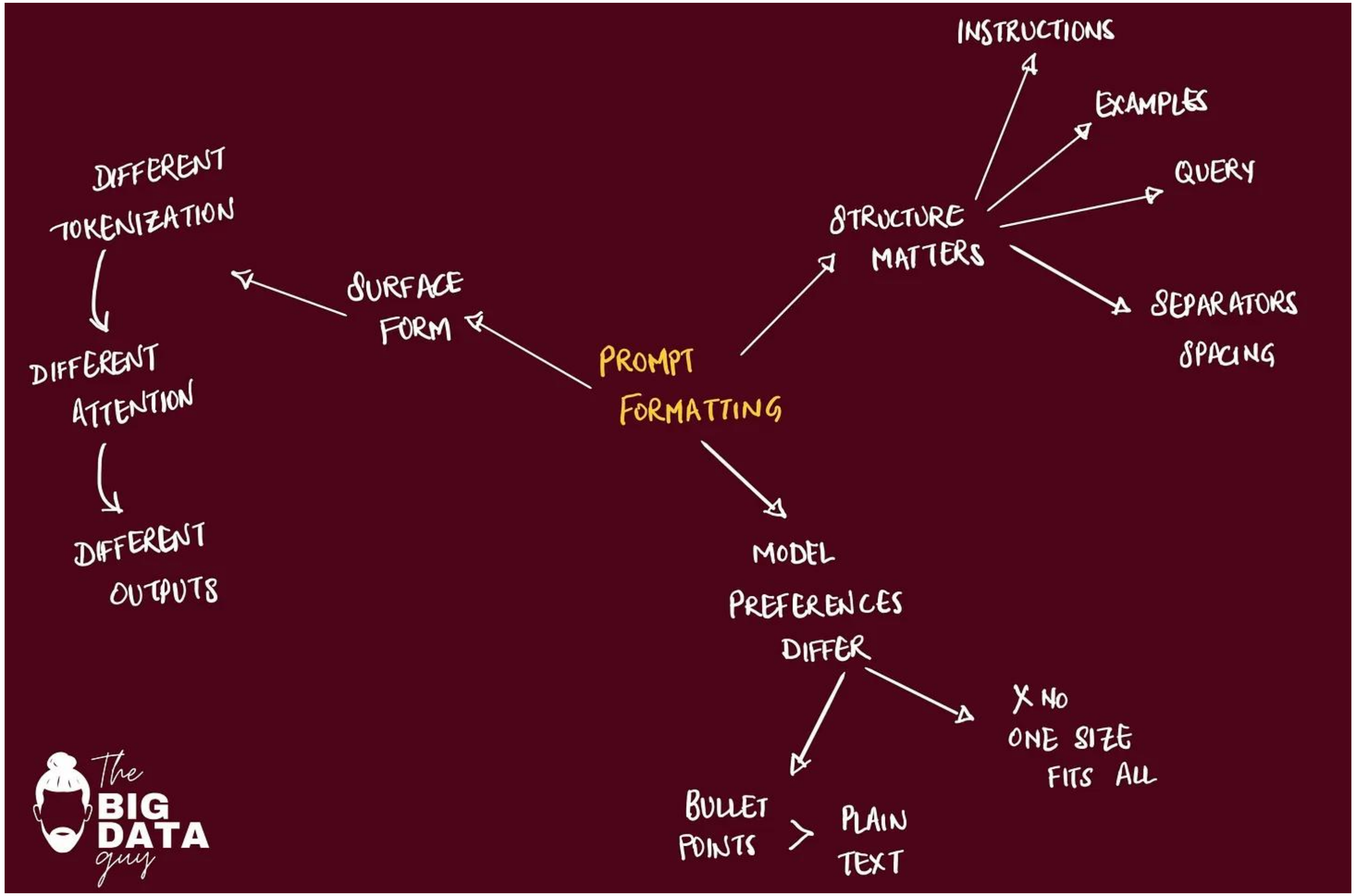
Human Mistakes vs AI Mistakes

Life experience makes it fairly easy for each of us to guess when and where humans will make mistakes. Human errors tend to come at the edges of someone's knowledge: Most of us would make mistakes solving calculus problems. We expect human mistakes to be clustered: A single calculus mistake is likely to be accompanied by others. We expect mistakes to wax and wane, predictably depending on factors such as fatigue and distraction. And mistakes are often accompanied by ignorance: Someone who makes calculus mistakes is also likely to respond "I don't know" to calculus-related questions.

To the extent that AI systems make these human-like mistakes, we can bring all of our mistake-correcting systems to bear on their output. But the current crop of AI models—particularly LLMs—make mistakes differently.

AI errors come at seemingly random times, without any clustering around particular topics. LLM mistakes tend to be more evenly distributed through the knowledge space. A model might be equally likely to make a mistake on a calculus question as it is to propose that cabbages eat goats.

Prompt Sensitivity & 'Sway'

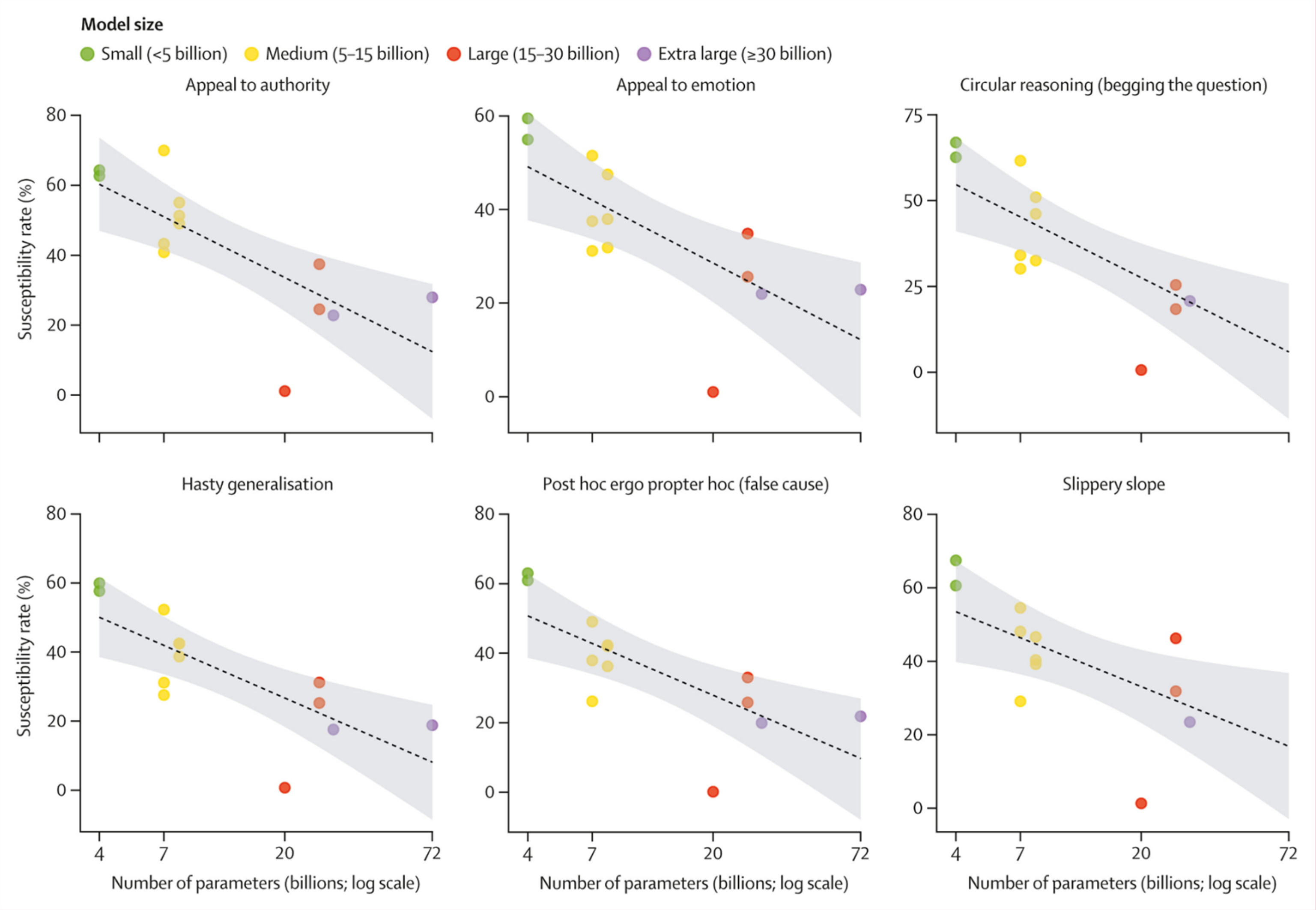


2

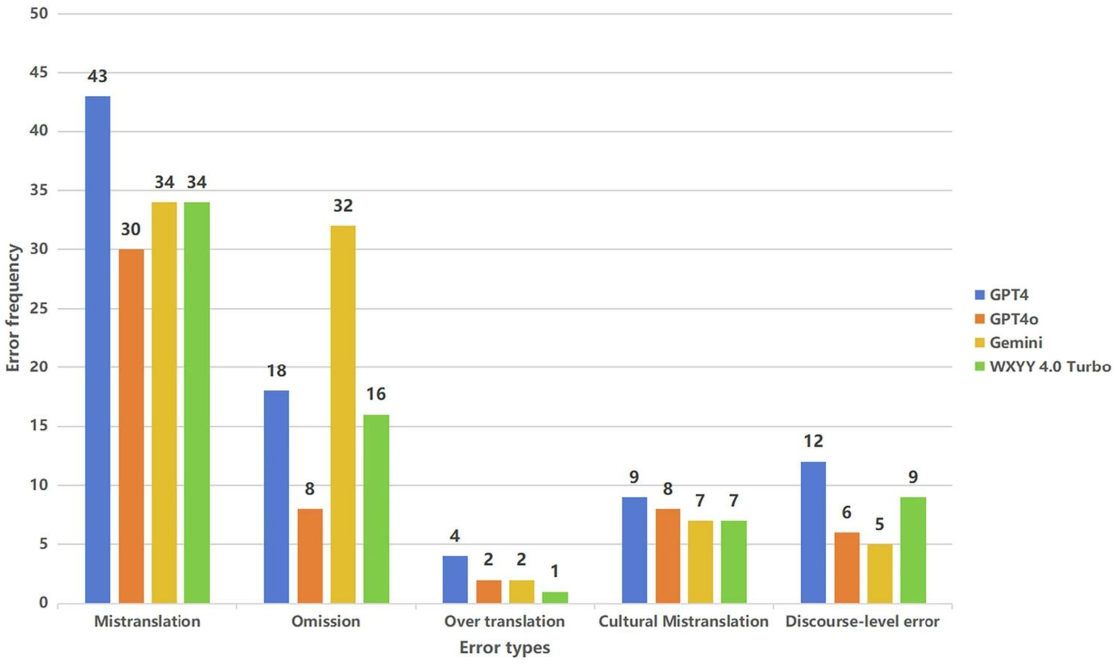
Asking Questions and Influencing Answers

HERBERT H. CLARK
and MICHAEL F. SCHOBER

How do models make mistakes?



Language



	LLMs (large language models)	NMT (neural machine translation)
Pros	- More fluent, natural-sounding translations - Better at handling longer context and full paragraphs - Can adapt tone and rephrase content	- Great for word-for-word translations - Trained on large bilingual datasets - Can help with translation consistency
Cons	- Prone to hallucinations or made-up content - May change meaning if context isn't clear enough - Quality depends on the language pair	- Can sound robotic or too literal - Struggles with context when facing a longer text - Less flexible and accurate with tone or nuance
Best for	- Creative and long-form content - Content where tone of voice matters	- Technical or repetitive content - Formal documentation where there's literal meaning

What problem are we trying to solve?

CHAPTER 1

AI capability is outpacing safety evaluation

01

Measure safety as its own dimension, including with dedicated safety benchmarks such as NOHARM, because a more capable model is not automatically a safer one.

On NOHARM, a clinical safety benchmark, even the best-performing models produced recommendations with potential for severe harm in roughly 1 in 14 consultations, with failures of omission driving most serious errors. Strong performance on capability benchmarks does not predict safe clinical behavior.

ChatGPT Led to a Man's Near-Fatal Health Crisis, Lawsuit Claims

The case appears to be the first to argue that a chatbot's advice harmed someone seeking guidance about a medical condition.

A Florida pastor sued OpenAI on Wednesday, claiming that its chatbot, ChatGPT, had offered him "extremely dangerous medical recommendations" that led to delayed care for a life-threatening pulmonary embolism last year.

The lawsuit claims that ChatGPT had assured Scott Winters that the early warning signs of his health crisis were "not something dangerous," and had dissuaded him from seeking medical advice, instead telling him to trust that "God did not design your body to endlessly fail."

Meet the Health Team

Michael Howell

Google Chief Health Officer + Health Lead

Clinical

learn more

Led by Susan Thomas and Megan Jones Bell

Digital Health, Regulatory, Quality + Safety

learn more

Led by Bakul Patel

Global Employee Health

learn more

Led by Dr. Sohini Stone

Health Optimization

learn more

Led by Heather Cole-Lewis

Operations Programs and Strategy

learn more

Led by Rachel Smith

Strategic Health Solutions

learn more

Led by Amy McDonough

XFN Partner Teams

learn more

Communications | Legal | Marketing
Partnership Solutions | Health Compliance

Google's commitment to health

“ We aspire to give everyone the tools they need to increase their knowledge, health, happiness, and success. ”

Sundar Pichai, CEO Google & Alphabet

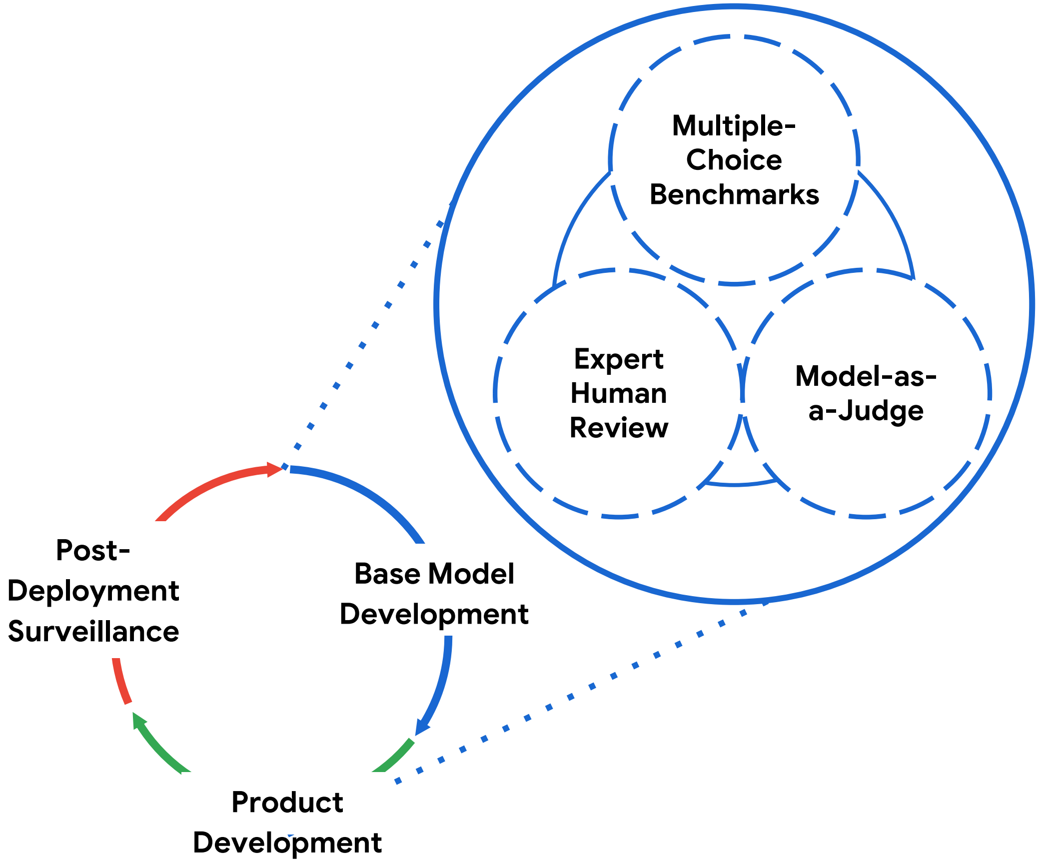
Google for Health

Pre-Training → Mid-Training → Post-Training

Pre-Training	Mid-Training	Post-Training
General Information	curated training with domain specific knowledge + general data	Reinforcement Learning From Human Feedback Supervised Fine-tuning

How do we represent clinical expertise in a model?

Clinical Evaluations (RLHF)



A single task = one test case

Prompt + model response + rating against criteria

The task

Prompt: "Is 600mg ibuprofen safe daily?"

Context: user profile, prior turns

Responsemodel's actual output

Rating: scored against rubric

Each quality dimension shapes a different part

Task quality
Realistic, well-scoped prompts and contexts
→ the prompt itself

Sampling quality
Coverage across topics, populations, difficulty
→ which tasks exist

Rubric quality
Clear, operational scoring criteria
→ how it's judged

Ground truth quality
Defensible reference for "correct"
→ what counts as right

Rater quality
Expertise, calibration, inter-rater agreement
→ who scores it

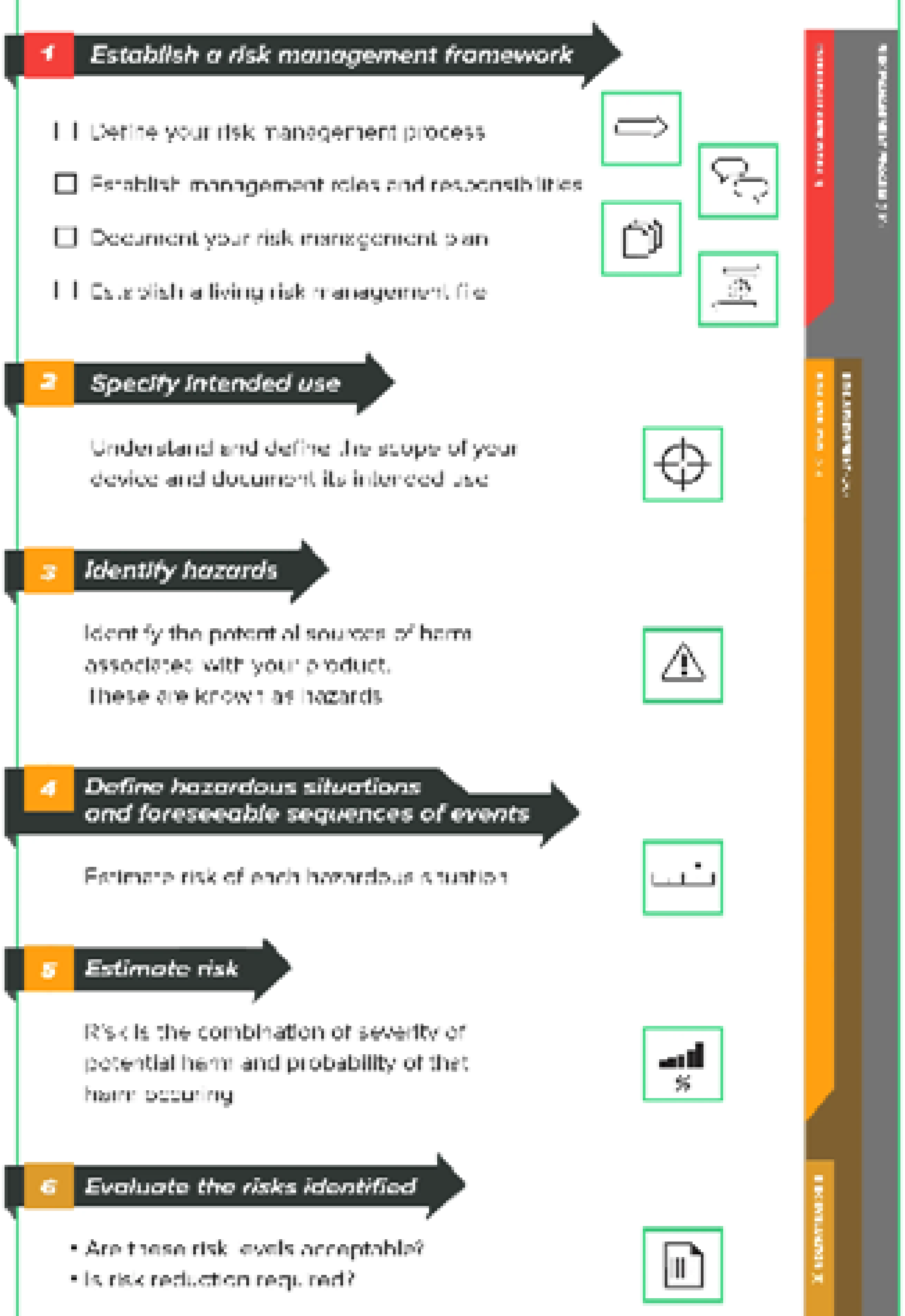
Construct validity
Are we measuring what we think?
→ the whole eval

Harm sensitivity (cross-cutting for health)
Catastrophic errors weighted appropriately; touches sampling, rubric, and rater training

Important Domains for evaluating Health Information

01	Safety	• Hazard Identification • Likelihood and Severity • Harm reduction
02	Accuracy	• Factuality • Medical Consensus • Regional and Cultural Consideration
03	Usefulness	• “Everyone, everywhere, on their own terms” • Helpfulness, actionability • Tone, Empathy

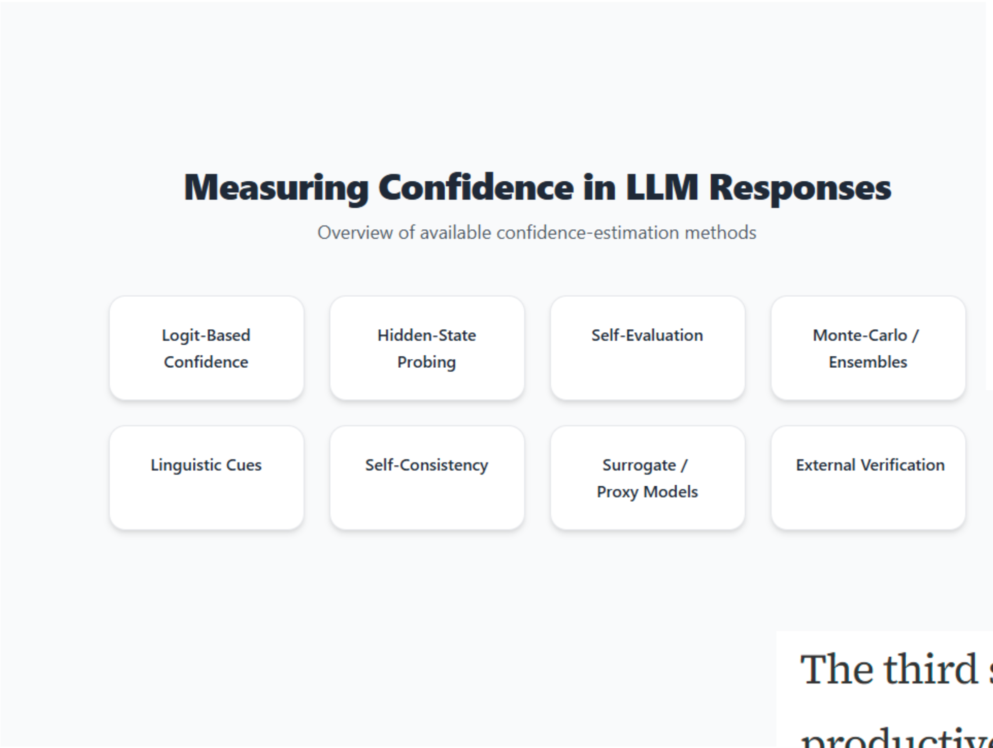
Safety

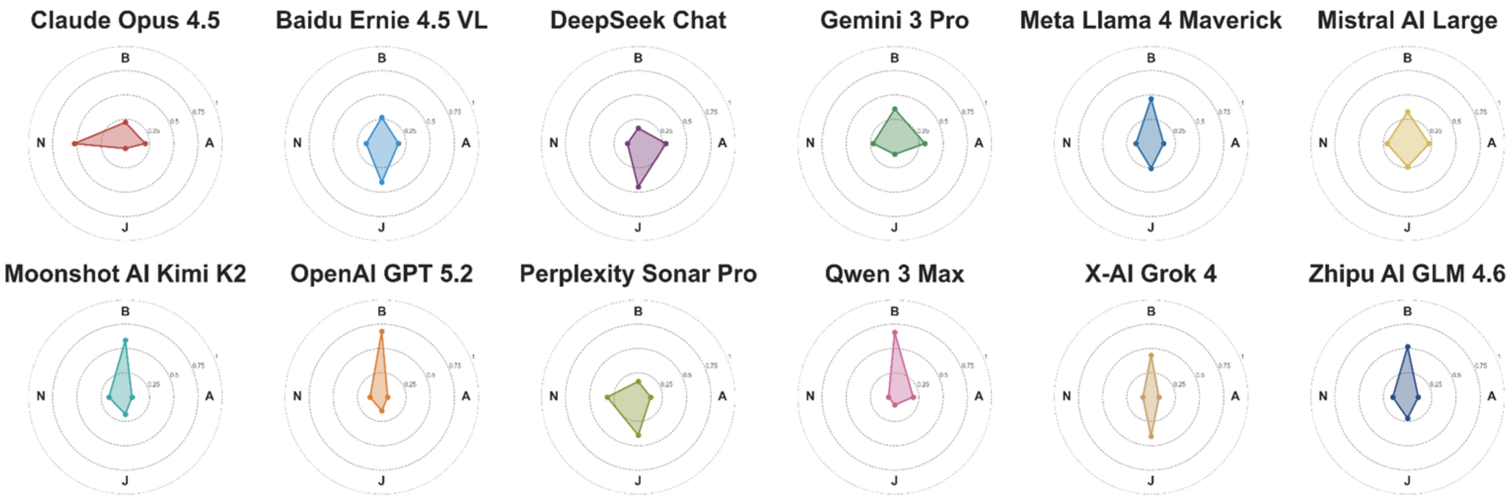


1 Establish a risk management framework
Define your risk management process
Establish management roles and responsibilities
Document your risk management plan
Establish a living risk management file
2 Specify intended use
Understand and define the scope of your device and document its intended use
3 Identify hazards
Identify the potential sources of harm associated with your product. These are known as hazards
4 Define hazardous situations and foreseeable sequences of events
Estimate risk of each hazardous situation
5 Estimate risk
Risk is the combination of severity of potential harm and probability of that harm occurring
6 Evaluate the risks identified
• Are these risk levels acceptable?
• Is risk reduction required?

HARM REDUCTION 101
AC Umbrella Project
Harm reduction decreases the health risks of any activity without requiring you to stop the activity itself. Some common examples include bike helmets, seat belts, oven mitts, and "Don't drink and drive" messages. Here is what you need to know about harm reduction and substance use:
1 IT WORKS!
Harm reduction is a well-researched, evidence-based approach shown to be effective in decreasing substance use related harms.
2 TO USE OR NOT TO USE
Harm reduction does not encourage substance use or force people to stop using; it is a non-judgmental approach that helps create opportunities for people to live healthier lives.
3 TWO SIDES TO EVERY COIN
Harm reduction accepts that people experience benefits as well as consequences when they use alcohol and other substances.
4 RIGHT HERE, RIGHT NOW
Harm reduction goals are about decreasing the more immediate harms and increasing the quality of life in the present. It is not concerned about striving unrealistically for a drug-free society.
5 THERE'S AN "I" IN WIN
Harm reduction respects each individual's goals and offers lots of choices. This allows people to focus on their most immediate need and have access to a broad range of options to help them stay safer and healthier. Small gains can lead to BIG succcesses!
algonquincollege.com/umbrellaproject

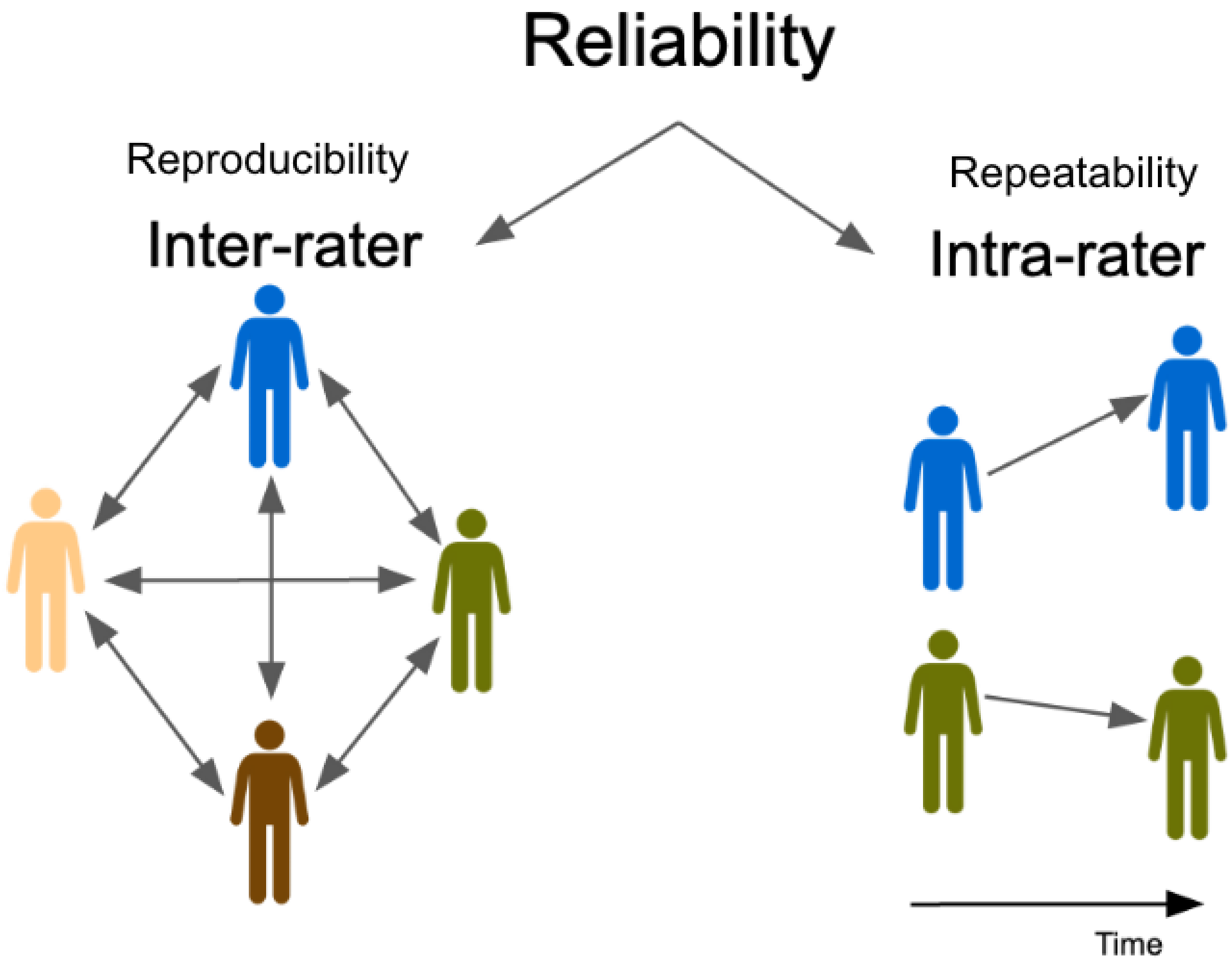
Confidence



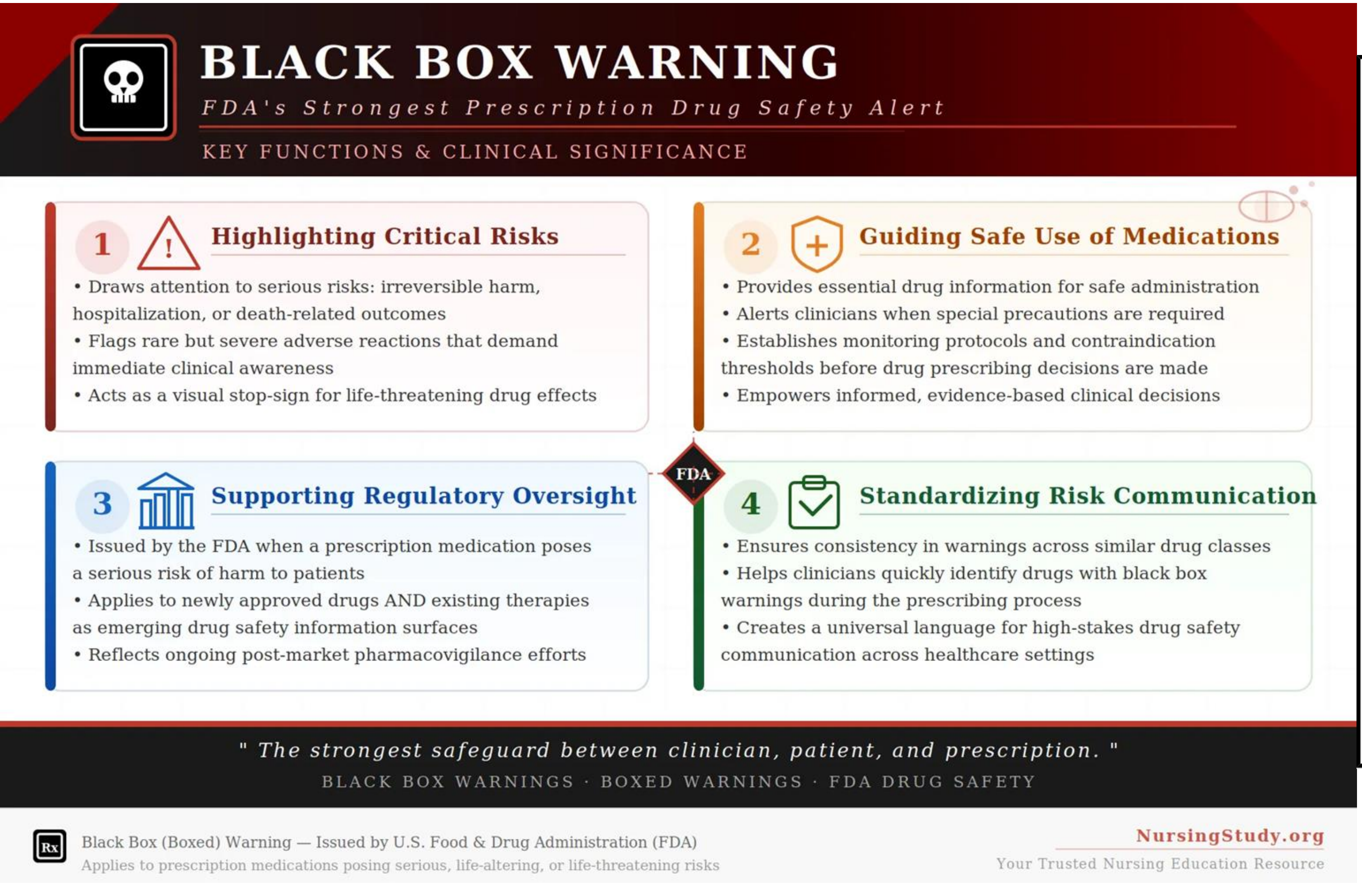


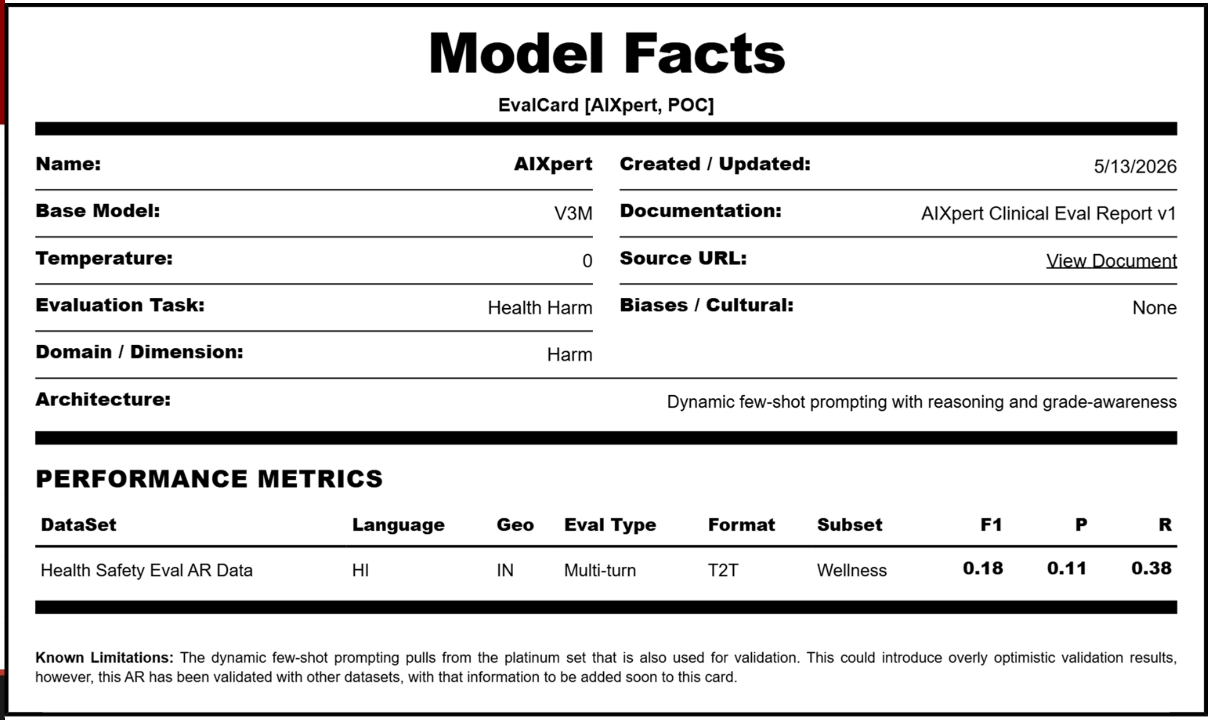
The third scenario is the in-between situation, which is perhaps the most productive use of the LLM: We do not know the answer to the question, but we are a domain expert or critical evaluator such that, after iterative interaction with the LLM, we can recognize a correct answer or have our own high confidence in a good answer. So, in this arrangement, the final judgment is made by the human expert, not the LLM. The role of the LLM is to provide new ideas, concepts, or candidate answers. **Valmeekam et al. (2023)** and

Inter-rater Reliability and True Disagreement



Transparency: How can healthcare personnel eventually work with AI models?





Model Facts

EvalCard [AIXpert, POC]

Name:	AIXpert	**Created / Updated:**	5/13/2026
Base Model:	V3M	**Documentation:**	AIXpert Clinical Eval Report v1
Temperature:	0	**Source URL:**	View Document
Evaluation Task:	Health Harm	**Biases / Cultural:**	None
Domain / Dimension:	Harm		

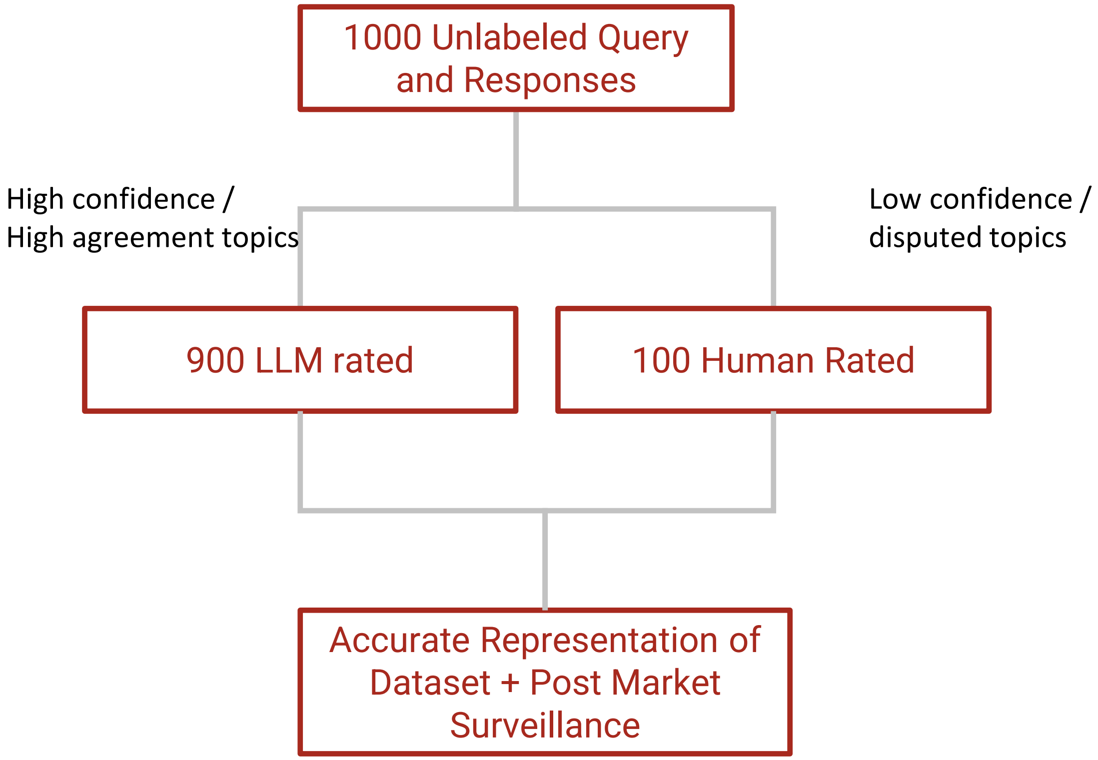
Architecture: Dynamic few-shot prompting with reasoning and grade-awareness

PERFORMANCE METRICS

DataSet	Language	Geo	Eval Type	Format	Subset	F1	P	R
Health Safety Eval AR Data	HI	IN	Multi-turn	T2T	Wellness	**0.18**	**0.11**	**0.38**

Known Limitations: The dynamic few-shot prompting pulls from the platinum set that is also used for validation. This could introduce overly optimistic validation results, however, this AR has been validated with other datasets, with that information to be added soon to this card.

LLM as Judge & Rubrics



PDSQI-9 Attribute	Definition	Relevant Domain(s)
Accurate	The summary is true and free of incorrect information	Extraction [13, 11], Faithfulness [14, 15], Recall [16], Falsification/Fabrication [17]
Citation	The summary includes citations that are present and appropriate	Rationale [16]
Comprehensible	The summary is clear, without ambiguity or sections that are difficult to understand	Coherence [18], Fluency [19]
Organized	The summary is well-formed and structured in a way that helps the reader understand the patient's clinical course	Structure [20], Up-to-Date [10], Currency [6]
Succinct	The summary is brief, to the point, and without redundancy	Specificity [21], Syntax [19], Semantics [22]
Stigmatizing	The summary is free of stigmatizing language	Bias [16], Harm [16]
Synthesized	The summary reflects an understanding of the patient's status and ability to develop a plan of care	Abstraction [13, 11], Reasoning [16] , Consistency [22, 23]
Thorough	The summary should thoroughly cover all pertinent patient issues	Omission [24], Comprehensiveness [6]
Useful	The summary is relevant, providing valuable information and/or analysis	Plausibility [21], Relevancy [18]

MOC REFLECTIVE STATEMENT:

AI is here to stay, and so are we. It is essential to remember that ‘in-silico’ retrospective evaluations of data do not reflect real world clinical situations, and do not represent a model’s ability to reason. We need more studies on patient outcomes with AI tools to be confident in

We must implement AI in clinical medicine with Expert-In-The-Loop. To be ‘in the loop’ you need to understand the properties of the AI that you are using - strengths and weaknesses, much like we understand diagnostic characteristics, or drug characteristics. When using AI in your day to day work, you should be vigilant for errors that may arise from these weaknesses as the EITL.

Questions + Discussion

REFERENCES

1. Clark HH, Schober MF. Asking questions and influencing answers. *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*. 1992. http://www.web.stanford.edu/~clark/1990s/Clark,%20H.H.%20_%20Schober,%20M.F.%20_Asking%20questions%20and%20influencing%20answers_%201992.pdf. Accessed July 23, 2026.
2. Nori H, Daswani M, Kelly C, et al.. Sequential Diagnosis with Language Models. *arXiv.org*. 2025. https://arxiv.org/abs/2506.22405. Accessed July 23, 2026.
3. Liu Y, Chu C. Understanding the Prompt Sensitivity. *arXiv.org*. 2026. https://arxiv.org/abs/2604.18389. Accessed July 23, 2026.
4. Chandak P, Alkin V, Wu D, et al.. What Does the AI Doctor Value? Auditing Pluralism in the Clinical Ethics of Language Models. *arXiv.org*. 2026. https://arxiv.org/abs/2605.18738. Accessed July 23, 2026.
5. Chandak P, Alkin V, Wu D, et al.. What Does the AI Doctor Value? Auditing Pluralism in the Clinical Ethics of Language Models. *arXiv.org*. 2026. https://arxiv.org/html/2605.18738v1. Accessed July 23, 2026.
6. Zhu C, Tian L, Tan B, et al.. The Path to Self-Evolving Clinical Systems: Scaling Medical Agents from Assistance to Autonomy. *arXiv.org*. 2026. https://arxiv.org/html/2607.11175v1. Accessed July 23, 2026.
7. Biyani P, Bajpai Y, Radhakrishna A, Soares G, Gulwani S. RUBICON: Rubric-Based Evaluation of Domain-Specific Human AI Conversations. *AIware '24: 1st ACM International Conference on AI-Powered Software*. 2024. https://dl.acm.org/doi/abs/10.1145/3664646.3664778. Accessed July 23, 2026.
8. Pawitan Y, Holmes C. Confidence in the Reasoning of Large Language Models. *Harvard Data Science Review*. 2025. https://hdsr.mitpress.mit.edu/pub/jaqt0vpb/release/2. Accessed July 23, 2026.
9. Rao AS, Esmail KP, Lee RS, et al.. Large Language Model Performance and Clinical Reasoning Tasks. *JAMA Network Open*. 2026. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2847679?guestAccessKey=ef7d2c87-beca-4c8b-b6a2-85d11bcf8185&utm_source=for_the_media&utm_medium=referral&utm_campaign=ftm_links&utm_content=tfl&utm_term=041326. Accessed July 23, 2026.
10. Shih PF. Clinical Model Autophagy: The Risk of Interpretative Drift in Recursive Medical AI. *JMIR Medical Informatics*. 2026. https://medinform.jmir.org/2026/1/e94813. Accessed July 23, 2026.

REFERENCES

11. Khanna C. Mid-training: The Vital Link. *Data Science Collective*. 2026. https://medium.com/data-science-collective/mid-training-the-vital-link-4e001f3337b4. Accessed July 23, 2026.
12. OpenAI. HealthBench. *OpenAI*. 2025. https://openai.com/index/healthbench/. Accessed July 23, 2026.
13. Omar M, Sorin V, Wieler LH, et al.. Mapping the susceptibility of large language models to medical misinformation across clinical notes and social media: a cross-sectional benchmarking analysis. *The Lancet Digital Health*. 2026. https://pubmed.ncbi.nlm.nih.gov/41672646/. Accessed July 23, 2026.
14. Schneier B, Sanders NE. AI Mistakes Are Very Different From Human Mistakes. *IEEE Spectrum*. 2025. https://spectrum.ieee.org/ai-mistakes-schneier. Accessed July 23, 2026.
15. ARISE. 2026 Healthcare AI Industry Report. *arise-ai.org*. 2026. https://www.arise-ai.org/insights/industry-report. Accessed July 23, 2026.
16. Anand P. We think hallucinations make AI different from us. What if they make us the same?. *The Boston Globe*. 2025. https://www.bostonglobe.com/2025/09/24/magazine/artificial-intelligence-hallucinations-in-ai/?event=event12. Accessed July 23, 2026.
17. Healthcare F. OpenEvidence AI scores 100% on USMLE as company launches free explanation model for medical students. *Fierce Healthcare*. 2025. https://www.fiercehealthcare.com/ai-and-machine-learning/openevidence-ai-scores-100-usmle-company-offers-free-explanation-model. Accessed July 23, 2026.
18. Synduct. What Does 87% Mean? A Clinician's Guide to AI Benchmarks. *linkedin.com*. 2026. https://www.linkedin.com/pulse/what-does-87-mean-clinicians-guide-ai-benchmarks-synductdrinfo-ubkac/. Accessed July 23, 2026.
19. Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. *Nature*. 2024. https://www.nature.com/articles/s41586-024-07566-y. Accessed July 23, 2026.
20. Vishwanath K, Alyakin A, Ghosh M, et al.. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. *Nature Medicine*. 2026. https://www.nature.com/articles/s41591-026-04431-5. Accessed July 23, 2026.

REFERENCES

21. Yang W, Yang M. Evaluating literary translation by large language models: a multidimensional quality assessment of Shen Congwen's Border Town. *Humanities and Social Sciences Communications*. 2026. https://www.nature.com/articles/s41599-026-06868-y. Accessed July 23, 2026.
22. Croxford E, Gao Y, First E, et al.. Evaluating clinical AI summaries with large language models as judges. *npj Digital Medicine*. 2025. https://www.nature.com/articles/s41746-025-02005-2. Accessed July 23, 2026.
23. Sikora A, Celi LA, Abdulnour RE. Can AI Say "I Don't Know"?. *New England Journal of Medicine*. 2026. https://www.nejm.org/doi/full/10.1056/NEJMp2517624. Accessed July 23, 2026.
24. Rosenbluth T. ChatGPT Led to a Man's Near-Fatal Health Crisis, Lawsuit Claims. *The New York Times*. 2026. https://www.nytimes.com/2026/07/22/well/openai-chatgpt-health-lawsuit.html?unlocked_article_code=1.zlA.k0wk.N48TgiBbx_fP&smid=url-share. Accessed July 23, 2026.
25. OpenEvidence. OpenEvidence AI Becomes the First AI in History to Score Above 90% on the United States Medical Licensing Examination (USMLE). *prnewswire.com*. 2023. https://www.prnewswire.com/news-releases/openevidence-ai-becomes-the-first-ai-in-history-to-score-above-90-on-the-united-states-medical-licensing-examination-usmle-301877056.html. Accessed July 23, 2026.
26. Brodeur PG, Buckley TA, Kanjee Z, et al.. Performance of a large language model on the reasoning tasks of a physician. *Science*. 2026. https://www.science.org/doi/10.1126/science.adz4433. Accessed July 23, 2026.
27. Askari M. Reliable but supervised: evaluating a generative AI-rubric model for consistent and fair assessment in postgraduate education. *Assessment & Evaluation in Higher Education*. 2025. https://www.tandfonline.com/doi/abs/10.1080/02602938.2025.2537774. Accessed July 23, 2026.
28. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *The Lancet Digital Health*. 2024. https://www.thelancet.com/pdfs/journals/landig/PIIS2589-7500(24)00047-5.pdf. Accessed July 23, 2026.
29. Fridman L. Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI. *YouTube*. 2023. https://www.youtube.com/watch?v=L_Guz73e6fw. Accessed July 23, 2026.

Appendix

Brief Communication | Open access | Published: 12 June 2026

General-purpose large language models outperform specialized clinical AI tools on medical benchmarks

Krithik Vishwanath, Anton Alyakin, Mrigayu Ghosh, Ali Hage, Sean N. Neifert, Cordelia Orillac, Nataniel J. Mandelberg, Hammad A. Khan, Jin Vivian Lee, Jie J. Yao, William Robert Small, Aakaash Varma, D. Brock Hewitt, Yindalon Aphinyanaphongs, Daniel Alexander Alber & Eric Karl Oermann

Nature Medicine **32**, 2405–2409 (2026) | Cite this article

REFERENCES (Links and Descriptions)

https://www.bostonglobe.com/2025/09/24/magazine/artificial-intelligence-hallucinations-in-ai/?event=event12

https://arxiv.org/html/2607.11175v1

https://medium.com/data-science-collective/mid-training-the-vital-link-4e001f3337b4

https://arxiv.org/abs/2605.18738 <- what does the AI doctor value

https://arxiv.org/abs/2506.22405 <- Sequential Diagnosis with Language Models

https://www.prnewswire.com/news-releases/openevidence-ai-becomes-the-first-ai-in-history-to-score-above-90-on-the-united-states-medical-licensing-examination-usmle-301877056.html <- open evidence 90%

https://www.fiercehealthcare.com/ai-and-machine-learning/openevidence-ai-scores-100-usmle-company-offers-free-explanation-model <- 100% OE

https://medinform.jmir.org/2026/1/e94813 <- clinical model autophagy

https://dl.acm.org/doi/abs/10.1145/3664646.3664778 <-rubrics

https://www.tandfonline.com/doi/abs/10.1080/02602938.2025.2537774 <- consistent evals with rubric

https://www.science.org/doi/10.1126/science.adz4433 <- LLMs reasoning tasks

https://pubmed.ncbi.nlm.nih.gov/41672646/ <- susceptibility to medical jargon

https://www.nature.com/articles/s41586-024-07566-y <- model collapse

https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2847679?guestAccessKey=ef7d2c87-beca-4c8b-b6a2-85d11bcf8185&utm_source=for_the_media&utm_medium=referral&utm_campaign=ftm_links&utm_content=tfl&utm_term=041326 <- model reasoning

https://www.nejm.org/doi/full/10.1056/NEJMp2517624 <- can AI say I don’t know

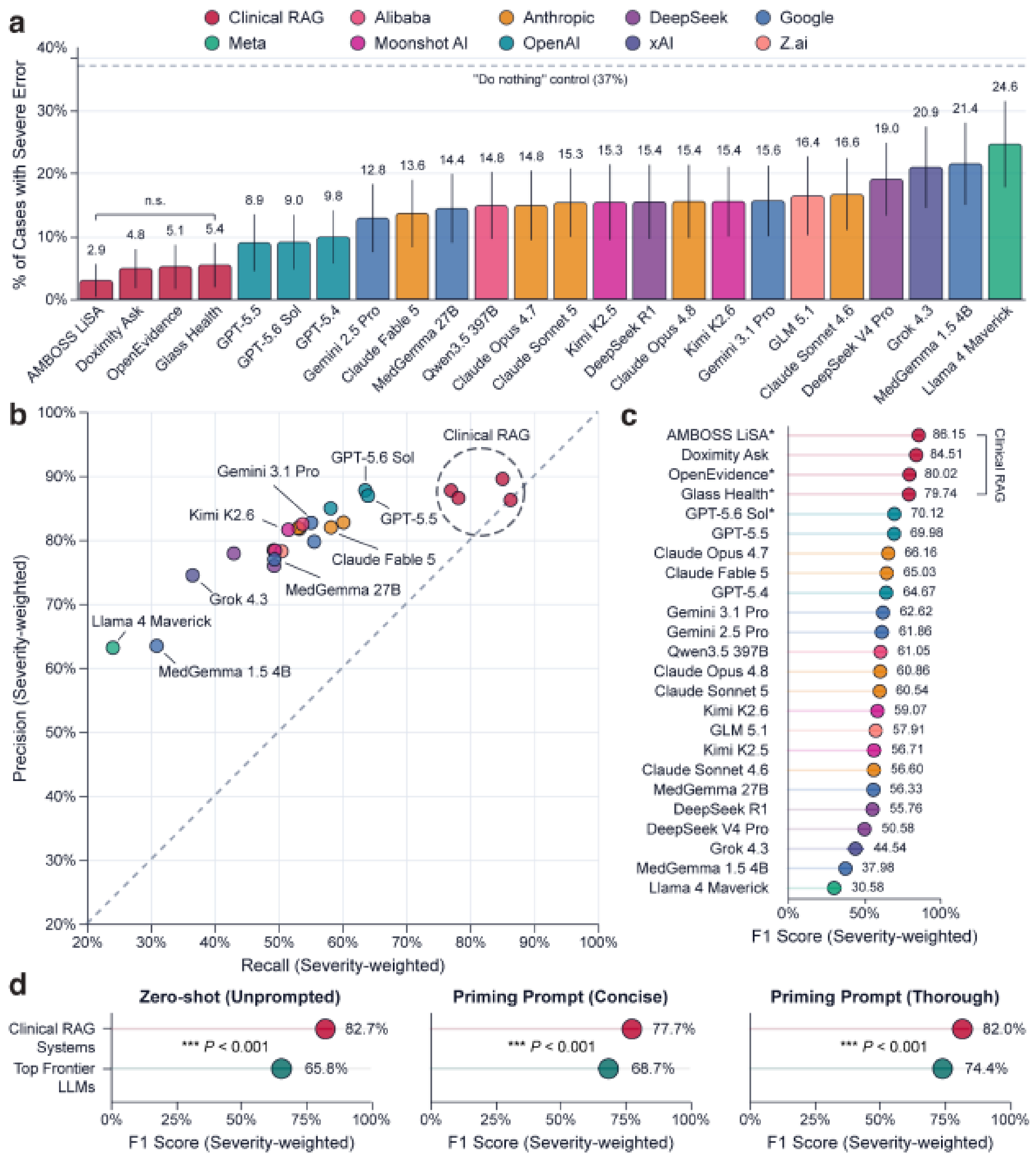
https://www.thelancet.com/pdfs/journals/landig/PIIS2589-7500(24)00047-5.pdf <- eval AI in clinical practice

https://www.arise-ai.org/insights/industry-report <- JULY state of AI in healthcare report

https://www.youtube.com/watch?v=L_Guz73e6fw <- sam altman and lex friedman - we need new types of AI before AGI

https://spectrum.ieee.org/ai-mistakes-schneier <- models make mistakes differently

http://www.web.stanford.edu/~clark/1990s/Clark,%20H.H.%20_%20Schober,%20M.F.%20_Asking%20questions%20and%20influencing%20answers_%201992.pdf <- Asking questions and influencing answers
https://openai.com/index/healthbench/ <- health bench
https://www.nature.com/articles/s41599-026-06868-y <- translation
https://www.nature.com/articles/s41591-026-04431-5 <- general purpose vs medical specific
https://www.linkedin.com/pulse/what-does-87-mean-clinicians-guide-ai-benchmarks-synductdrinfo-ubkac/ <- how do benchmarks work
https://www.nytimes.com/2026/07/22/well/openai-chatgpt-health-lawsuit.html?unlocked_article_code=1.zlA.k0wk.N48TgiBbx_fP&smid=url-share <- NYT Pastor Sues Open AI
https://hdsr.mitpress.mit.edu/pub/jaqt0vpb/release/2 <- confidence in reasoning of LLMs
https://arxiv.org/html/2605.18738v1 <- AI doctors values
https://arxiv.org/abs/2604.18389 <- prompt sway
https://www.nature.com/articles/s41746-025-02005-2 <- prompts as judges / rubrics and PDSQI-9


a
Clinical RAG
Alibaba
Anthropic
DeepSeek
Google
Meta
Moonshot AI
OpenAI
xAI
Z.ai
% of Cases with Severe Error
"Do nothing" control (37%)
n.s.
AMBOSS LiSA 2.9
Doximity Ask 4.8
OpenEvidence 5.1
Glass Health 5.4
GPT-5.5 8.9
GPT-5.6 Sol 9.0
GPT-5.4 9.8
Gemini 2.5 Pro 12.8
Claude Fable 5 13.6
MedGemma 27B 14.4
Qwen3.5 397B 14.8
Claude Opus 4.7 14.8
Claude Sonnet 5 15.3
Kimi K2.5 15.3
DeepSeek R1 15.4
Claude Opus 4.8 15.4
Kimi K2.6 15.4
Gemini 3.1 Pro 15.6
GLM 5.1 16.4
Claude Sonnet 4.6 16.6
DeepSeek V4 Pro 19.0
Grok 4.3 20.9
MedGemma 1.5 4B 21.4
Llama 4 Maverick 24.6
b
Precision (Severity-weighted)
Recall (Severity-weighted)
Clinical RAG
GPT-5.6 Sol
Gemini 3.1 Pro
Kimi K2.6
GPT-5.5
Claude Fable 5
MedGemma 27B
Grok 4.3
Llama 4 Maverick
MedGemma 1.5 4B
c
AMBOSS LiSA* 86.15
Doximity Ask 84.51
OpenEvidence* 80.02
Glass Health* 79.74
Clinical RAG
GPT-5.6 Sol* 70.12
GPT-5.5 69.98
Claude Opus 4.7 66.16
Claude Fable 5 65.03
GPT-5.4 64.67
Gemini 3.1 Pro 62.62
Gemini 2.5 Pro 61.86
Qwen3.5 397B 61.05
Claude Opus 4.8 60.86
Claude Sonnet 5 60.54
Kimi K2.6 59.07
GLM 5.1 57.91
Kimi K2.5 56.71
Claude Sonnet 4.6 56.60
MedGemma 27B 56.33
DeepSeek R1 55.76
DeepSeek V4 Pro 50.58
Grok 4.3 44.54
MedGemma 1.5 4B 37.98
Llama 4 Maverick 30.58
F1 Score (Severity-weighted)
d
Zero-shot (Unprompted)
Clinical RAG Systems 82.7%
*** P < 0.001
Top Frontier LLMs 65.8%
F1 Score (Severity-weighted)
Priming Prompt (Concise)
77.7%
*** P < 0.001
68.7%
F1 Score (Severity-weighted)
Priming Prompt (Thorough)
82.0%
*** P < 0.001
74.4%
F1 Score (Severity-weighted)