State of Clinical AI Report 2026

ARISE-AI.ORG

January, 2026

About The Authors

Peter Brodeur

Dr. Peter Brodeur is a rising cardiology fellow at Harvard Medical School's Beth Israel Deaconess Medical Center. Dr. Brodeur is an affiliate of ARISE, reviewer for Nature Medicine & NEJM AI, and former life sciences strategy consultant. His research focuses on human computer interaction and LLM clinical reasoning.

Ethan Goh

Dr. Ethan Goh is the Executive Director of ARISE. His research has been featured in The New York Times, The Washington Post, and CNN. He directs the Stanford Healthcare AI Leadership Program, and Harvard's Agentic AI Executive Course. Dr. Goh is a Founding Editorial Board member and Associate Editor at BMJ Digital Health & AI.

Adam Rodman

Dr. Adam Rodman is an assistant professor at Harvard Medical School. He is the Director of AI Programs for the Carl J. Shapiro Center. Dr. Rodman is an Associate Editor at NEJM AI. He is also the host of the American College of Physicians podcast Bedside Rounds.

Jonathan H Chen

Dr. Jonathan H Chen is Stanford's inaugural Director for Medical Education in AI in the Division of Computational Medicine. His expertise combining human with artificial intelligence to provide better healthcare than either alone is featured in the popular press with over 100 publications and awards.

ARISE-AI.ORG

Message From ARISE Leadership

“There are decades where nothing happens; and there are weeks when decades happen.” Recent deployments by technology companies, health systems, and regulators have made clinical AI more visible and ever more consequential. At the same time, it has become harder to keep up with emerging research. In some areas the literature is fragmented; in others, it simply doesn’t exist yet for the way these tools are being used today.

So what actually holds up in practice?

The State of Clinical AI Report (2026) was created to look beyond model performance alone to other critical factors that determine real-world impact: how systems are evaluated, how clinicians and AI work together, and where patient risks start to appear.

Frontier AI systems are already powerful. What’s needed now is to safely and effectively translate these tools into real-world care.

Ethan Goh, Adam Rodman, Jonathan H Chen
Investigators, ARISE Network

Engagement and Education

Stanford Computational Medicine Colloquia

- Healthcare AI seminars with Stanford / industry leaders
- Thursday 12 pm PT, free

Get weekly invites

Stanford Healthcare AI Leadership & Strategy Program

- Application required. CME and accredited certificate
- May 2026

Apply now

Generative AI and Agentic AI Online Course

- Harvard/Stanford faculty, accredited certificate
- Summer 2026

Get early access

The Current Landscape

Clinical AI Is Widely Deployed But Poorly Evaluated

- AI is now embedded across health care: 1,200+ FDA-cleared tools and 350,000+ consumer apps have generated a $70B market[1]. Only a minority underwent peer-reviewed evaluation.[2]
- Of 691 FDA-cleared AI/ML medical devices (1995–2023), >95% went through the 510(k) clearance pathway, which is predicated on equivalency to existing devices — many of which were approved on suboptimal evidence.[2]
- ~50% of FDA device summaries omitted study design, 53% lacked sample size, and <1% reported patient outcomes.[2]
- 95% of device summaries did not report demographic data, and 91% lacked bias assessments, raising concerns about safety and equity in real-world use.[2]

Bridging the gap between adoption and evidence requires supporting clinicians, health system leaders, policymakers, and the public in interpreting available research.

ARISE-AI.ORG

Top Takeaways

1. **Model capability is accelerating, but evidence of real clinical impact remains limited.** Many studies show what models can do in controlled settings; what's increasingly needed are prospective studies that show measurable effects on patient outcomes and care delivery.

2. **Frontier LLM models show very uneven performance.** They perform extremely well on complex reasoning tasks, yet break down when uncertainty, missing information, or changing context is introduced.

3. **Clinicians value automation where it reduces administrative and workflow burden, but these use cases remain understudied.** Tasks clinicians most want support with are often underrepresented in current benchmarks and evaluations.

Top Takeaways

4. **Patient-facing AI has significant potential to reshape engagement and access, but raises distinct safety concerns.** Direct interaction with patients requires much stronger guardrails and scalable oversight systems that do not currently exist.

5. **Multimodal clinical AI applications are approaching practical usability.** Improvements in base models are enabling applications that integrate unstructured text, images, and other clinical data to support prediction and decision-making in real-world settings.

6. **FDA clearance is increasing, but near-term clinical adoption will favor narrow, task-specific systems.** AI tools that are tightly scoped to specific domains and contexts are more likely to demonstrate value and be adopted in practice.

Acknowledgements

Reviewers

Rebecca Handler
Jason Hom
Eric Horvitz
Laura Zwaan
Vishnu Ravi
Brian Han
Kevin Schulman
Kathleen Lacar
Kameron Black
Liam McCoy
David Wu
Priyank Jain
Emily Tat
Adrian Haimovich

Design & Accessibility

Emily Tat

Supported By

Stanford | Computational Medicine

HARVARD MEDICAL SCHOOL

CENTER FOR EDUCATION
Shapiro Institute

Beth Israel Deaconess Medical Center

HARVARD MEDICAL SCHOOL TEACHING HOSPITAL

Stanford University
Clinical Excellence Research Center

Stanford MEDICINE
Division of Hospital Medicine

HARVARD MEDICAL SCHOOL
BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

The organization format of this report was inspired by Nathan Benaich's State of AI Report.

ARISE-AI.ORG

How to Cite This Report

Peter G. Brodeur, Ethan Goh, Emily Tat, Liam McCoy, David Wu, Priyank Jain, Rebecca Handler, Jason Hom, Laura Zwaan, Vishnu Ravi, Brian Han, Kevin Schulman, Kathleen Lacar, Kameron Black, Adrian Haimovich, Eric Horvitz, Adam Rodman, Jonathan H. Chen “State of Clinical AI 2026,” ARISE Network, January 2026.

Introduction

Executive Summary

Model Performance

- Frontier reasoning models (optimized for multi-step inference and chain of thought) showed marked improvement on **challenging clinical reasoning** tasks against human baselines while prediction models crossed new thresholds in **scalable prediction to enable actionable prevention**.
- Dominant failure modes include model recognition of **uncertainty, overconfidence, and pattern learning**.

Benchmarks & Evaluation

- Multiple choice benchmarks are saturated and evaluations still **underrepresent real clinical work**: administrative tasks, conversational dialogue, real patient data, and bias/fairness.
- New **benchmark suites** (e.g., conversational, simulated EHR environments) are forcing models into more **realistic, dynamic scenarios**.

Foundational Methods

- Novel techniques such as converting **medical data to tokens** used for prediction brings a new era of screening and risk stratification.
- Clinical AI is being advanced by **multiagent systems, multimodal diagnostic support,** and **optimizing reasoning** models.

ARISE-AI.ORG

Executive Summary

AI in Clinical Workflows

- Across settings, AI can **augment clinicians on reasoning and diagnostic interpretation tasks**. However, **collaboration isn't yet optimized**. How clinicians use AI is as important as what the model can do.
- Workflow tools like AI scribes feel transformative, yet **objective gains are still modest.** The **addition of downstream workflow tasks** will likely yield more productivity and efficiency impact.

Patient Facing AI

- **Multi-turn conversational agents** and AI-based coaching show promise, particularly as they are integrated with smart devices to support more **personalized health assistance.**
- In a space with competing vendor interests, **overtrust and unsupervised use** raise the bar for guardrails and for improving **objective patient outcomes**, not just engagement.

Applied AI & Demos

- The most immediate translatable progress can be seen at the individual task-specific level with **imaging remaining the dominant use case**.
- We provide a **sneak peek of the next wave of tools** such as EHR chatbots, eConsults, and mental health chatbots.

ARISE-AI.ORG

Methods

Our Approach to a Targeted Review of Clinical AI

- **Data sources & search strategy**
  - Reviewed PubMed, preprint servers (e.g., medRxiv, arXiv) using a combination of terms such as “large language models in medicine,” “AI,” “diagnostic reasoning,” “management reasoning,” “diagnostic error,” “benchmarks,” and “patient-facing AI.”
  - Invited clinicians and AI researchers from academic institutions and issued an open call for submissions via social media (e.g., LinkedIn) to identify high-quality studies across the six themes.

- **Study selection**
  - All studies reviewed by authors and reviewers of this presentation.
  - Included empirical studies that (1) used an AI model/LLM in a clinical context, (2) reported quantitative or qualitative outcomes (e.g., diagnostic accuracy, bias, calibration, workflow, user performance), and (3) determined to be of high impact.
  - Excluded purely technical model papers without clinician- or patient-facing evaluation, editorials, and non-clinical AI (e.g., drug discovery, biotech).

Table of Contents

Model Performance

Model Performance

In 2025, frontier models made major leaps in autonomous clinical reasoning and prediction.

- **Slides 18–20:** Reasoning frontier models show large gains in autonomous clinical reasoning versus humans, including on historically difficult cases.
- **Slides 21–22:** Key weaknesses persist: poor performance in uncertainty-heavy scenarios, overconfidence, and pattern-based shortcut behavior.
- **Slides 23–27:** Models continue to show promise for scalable prediction across a wide variety of use cases such as patient deterioration, screening for insulin resistance, and aging.

Overall, model-only evaluations reveal that LLMs have achieved superhuman capability in controlled tasks but still require stronger metacognition, calibration, and stress testing before autonomous deployment.

Model Performance

Complex Reasoning

- Approaching superhuman reasoning
- AI vs MD
  - LLM vs Primary Care Physician
  - LLM as an expert case discussant
- Gaps
  - “None of the other answers”
  - Brittle overconfidence and uncertainty

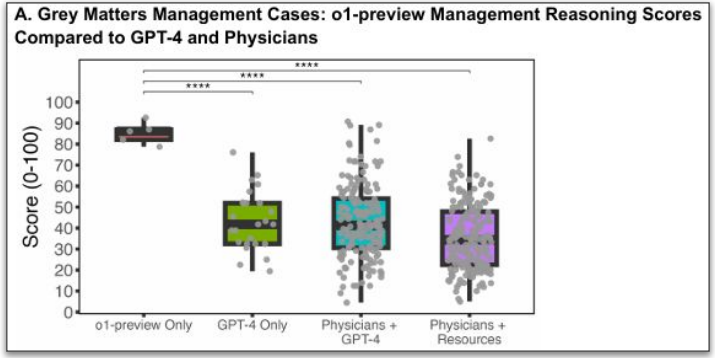
Prediction

- Inpatient deterioration
- Biological age
- Insulin resistance
- Wearable time series data for diagnosis prediction
- Clinical risk calculator

O1-preview/o1: Reaching Superhuman Reasoning Performance

O1-preview and o1 consistently outperformed or at the level of physicians across several reasoning evaluations, solving challenging NEJM cases at state-of-the-art levels, documenting superior reasoning quality, excelling in management tasks, and diagnosing real emergency room cases admitted to the hospital.

- On NEJM clinicopathological conference (CPC) cases, the model reached 78% diagnostic accuracy and selected the correct next test 87% of the time.
- o1-preview achieved a perfect score 99% of the time for clinical reasoning quality graded by physicians. This significantly outperformed GPT-4 (59%) and attending physicians (35%). Management reasoning for o1-preview (86%) was also superior compared to GPT-4 (42%) and physicians with GPT-4 (41%).
- In real ED cases, the model outperformed or at the level of both attending physicians at three diagnostic touchpoints with 66% exact/near-exact diagnoses vs. 48–54% for physicians at initial triage.



Brodeur, Buckley, Manrai, Rodman et al., ArXiv, Jul. 2025

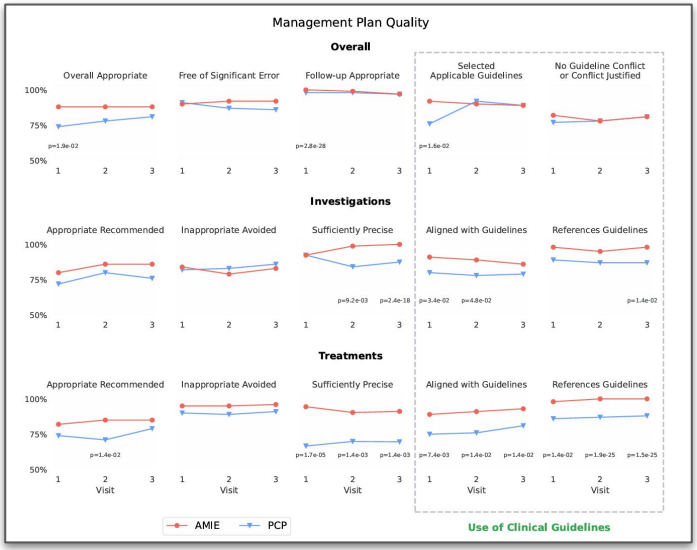
- Modern LLMs may now surpass physicians in general diagnostic and management reasoning in controlled environments, motivating the need for prospective clinical trials for real-world deployment.

ARISE-AI.ORG

Google's AMIE Chatbot Matches PCPs at Multi-Visit Disease Management

Enhanced with a new management-reasoning agent, the Articulate Medical Intelligence Explorer (AMIE) was non-inferior to 21 primary care physicians across guideline-based decision-making, treatment planning, and longitudinal care. AMIE produced more precise, guideline-based plans, and outperformed physicians on medication-reasoning questions.

- AMIE (gemini-based) was designed as a two part system with access to an agent state (current patient summary, differential etc.): a fast Dialogue Agent to capture relevant HPI and a slower Management Reasoning agent using long context reasoning grounded in clinical guidelines.
- Compared AMIE to PCPs across 100 three-visit simulated scenarios spanning cardiology, pulmonology, neurology, OBGYN/urology, and GI, each grounded in NICE and BMJ Best Practice guidelines.
- Graded by subspecialists, AMIE's recommendations for investigations and treatments were consistently more precise (Yes/No), especially for investigations in follow-up visits (visit 2: 99% vs. 84%, visit 3: 100% vs. 88%), and carried explicit citations to guideline sources. Possibility for agentic agents to serve as a point of continuity in a fragmented system.
- On a novel medication reasoning (RxQA) benchmark, AMIE outperformed PCPs on harder questions (as determined by pharmacists) in both closed- and open-book conditions, demonstrating strong therapeutic reasoning.



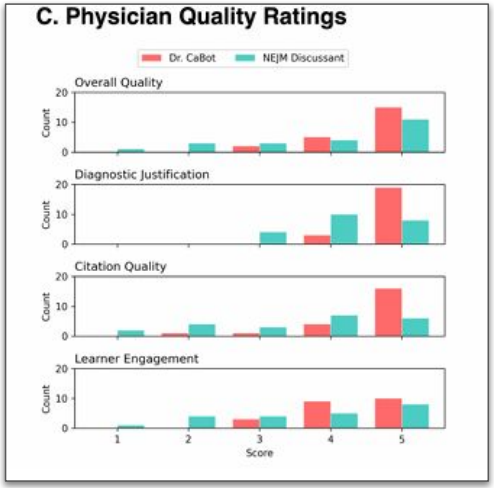
Palepu, Schaekermann et al., ArXiv, Mar. 2025

ARISE-AI.ORG

AI Outperforms Physicians as an Expert Case Discussant on Challenging Cases

Researchers developed Dr. CaBot, an AI discussant based on o3 that produces written and video CPC-style differentials. Dr. CaBot was evaluated on NEJM CPCs and NEJM Image Challenges, spanning ten tasks that test differential diagnosis, testing strategies, clinical reasoning, uncertainty handling, and multimodal interpretation. In blinded testing, physicians could not reliably distinguish Dr. CaBot from human experts, and consistently rated its reasoning higher.

- Built from 7,102 NEJM CPCs (1923–2025) and 1,021 NEJM Image Challenges, CPC-Bench covers 10 reasoning tasks (DDx, testing plans, touchpoints, omission, VQA, literature search, etc.).
- Among eight frontier models, o3 achieved 60% top-1 and 84% top-10 accuracy on CPC differential diagnosis, outperforming a 20-physician baseline, with 98% accuracy selecting the next test.
- Dr. CaBot, based on o3, is a publicly available (https://cpcbench.com/) system that produces both written and video case presentations that outperforms the originally presented expert case discussant.
- The study shows that AI is now capable of performing the entire CPC discussant role, with reasoning quality rated better than human experts.



Buckley et al., ArXiv, Sept. 2025

ARISE-AI.ORG

"None of the other answers": An LLM Weakness

Researchers tested whether LLMs could truly reason through medical questions by replacing the correct answer in multiple choice questions with "None of the other answers" (NOTA). Frontier models showed significant drops in accuracy, revealing that strong multiple choice performance, is in part, due to pattern recognition.

- Researchers modified 100 MedQA questions so that NOTA became the correct answer, creating a 68-item clinician-validated test of genuine reasoning. The pattern of answers has changed but the underlying clinical reasoning has not.
- DeepSeek-R1, o3-mini, Claude 3.5 Sonnet, Gemini 2.0 Flash, GPT-4o, and Llama 3.3-70B all performed worse on NOTA-modified questions. Significant decreases in performance were exhibited, ranging from 9% to 38%.
- A system that falls for example from 81% → 43% accuracy when a pattern changes would be unsafe for autonomous clinical use; rigorous benchmarks must test reasoning, not memorized answer distributions.

Table. Model Performance on Original and None of the Other Answers (NOTA)-Modified Questions[a]

	Accuracy, % (No./total No.)		
Model	Original	NOTA-modified	Accuracy drop, % (No./total No.) [95 % CI]
1	92.65 (63/68)	83.82 (57/68)	8.82 (6/68) [2.70-18.92]
2	95.59 (65/68)	79.41 (54/68)	16.18 (11/68) [10.81-29.73]
3	88.24 (60/68)	61.76 (42/68)	26.47 (18/68) [17.57-39.19]
4	92.65 (63/68)	58.82 (40/68)	33.82 (23/68) [24.32-47.30]
5	85.29 (58/68)	48.53 (33/68)	36.76 (25/68) [28.38-51.35]
6	80.88 (55/68)	42.65 (29/68)	38.24 (26/68) [27.03-51.35]

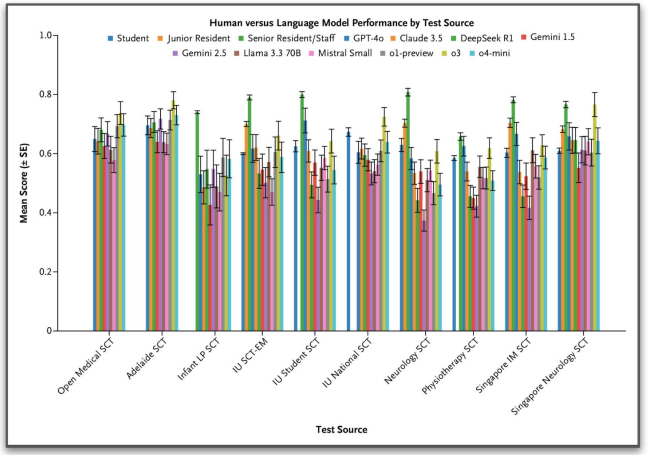
Bedi, Shah et al., JAMA Network Open, Aug. 2025

ARISE-AI.ORG

Script Concordance Testing Reveals Gaps in LLM Clinical Reasoning

A study compared 10 frontier models to 1,500+ clinicians on 750 Script Concordance Testing (SCT) questions, which measure the ability to revise clinical decisions when new information becomes available. Models matched medical students but underperformed relative to seasoned physicians, revealing consistent overconfidence and difficulty updating decisions under uncertainty.

- SCT measures the ability to revise diagnostic or management judgments when new information arrives, a core skill of clinical reasoning under uncertainty.
- This study established a benchmark assessing 750 SCT items from 10 datasets, including pediatrics, neurology, emergency medicine, internal medicine, and physiotherapy, most never previously published.
- OpenAI's o3 (68%) led performance, followed by GPT-4o (64%), matching medical students but below residents and attending physicians. Many reasoning models performed surprisingly poorly (e.g., Gemini 2.5: 52%).



McCoy, Rodman et al., NEJM AI, Sept. 2025

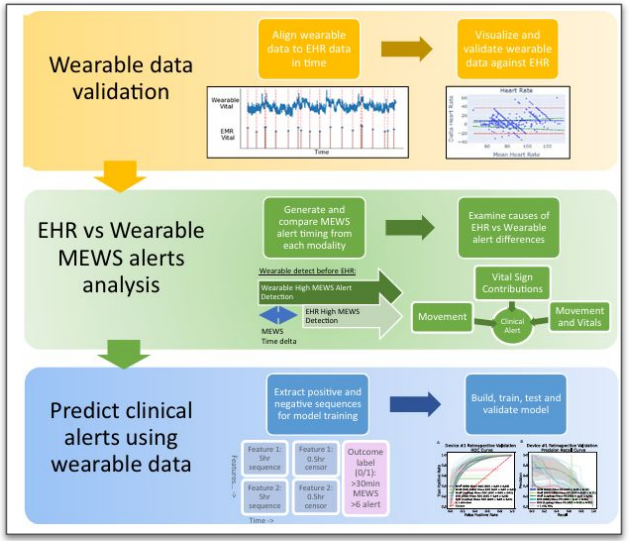
- LLMs overused extreme ratings (+2/-2), rarely selected neutrality (0), and showed miscalibrated confidence patterns unlike human experts, suggesting that chain-of-thought–optimized models may overcommit in uncertainty-rich tasks.

ARISE-AI.ORG

Predicting Inpatient Deterioration Before It Happens

Researchers developed a deep-learning model using continuous wearable vital sign data from 888 hospitalized med-surg patients to predict clinical deterioration up to 8-24 hours before standard EHR alerts. The model generated more timely alerts than episodic vital checks and accurately predicted hard outcomes, including ICU transfer, cardiac arrest, and death.

- Outside of the ICU, inpatient vital signs are checked every 4-8 hours, which leaves time gaps of missed opportunity for detecting critical illness.
- Researchers trained a recurrent neural network with a 5 hour sequence of continuous vital sign inputs (e.g., HR, RR) collected from a wearable chest device, with demographics from 888 non-ICU patients to detect early deterioration.
- Predicted 9x more clinical alerts (Modified Early Warning Score (MEWS) > 6 for > 30 mins) 8-24 hours before EHR-based MEWS alerts, with AUROC 0.89 (retrospective) and AUROC 0.84-0.9 (prospective). Predicted 9 of 11 hard outcome events (cardiac arrests, death) up to 17 hours before MEWS.
- Enables faster recognition of physiologic decline and the potential to prevent avoidable deteriorations.



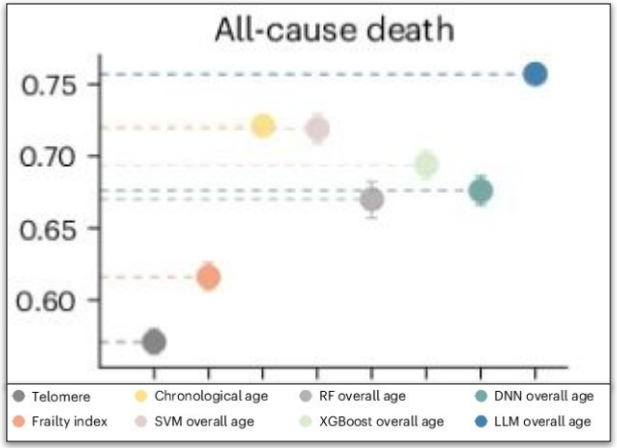
Scheid, Zanos et al., Nature Communications, Jul. 2025

ARISE-AI.ORG

Predicting Biological Aging at Population Scale Using Large Language Models

This study introduces an LLM prompt based framework that predicts biological age from routine health records, enabling scalable aging assessment across populations. Applied to >10 million individuals from six cohorts (e.g., UK Biobank), the LLM-derived biological age outperformed traditional aging clocks in predicting mortality and multiple age-related diseases.

- Using LLMs in the Llama and Qwen families, applied prompt learning without supervised learning on aging related knowledge. After being fed health examination text reports, LLMs integrate individualized clinical data to infer biological age without predefined biomarkers or labels.
- LLM-based biological age achieved a concordance-index of 0.76 for all-cause mortality. Also outperformed epigenetic clocks, telomere length, frailty index, and conventional ML models. The difference between LLM-predicted age and chronological age (“age-gap”) was strongly associated with all-cause mortality (HR 1.05).
- LLM-derived organ-specific biological ages better predicted corresponding organ diseases and enabled potential discovery of 316 aging-related protein biomarkers.
- Potential for scalable and cost-effective personalized and population aging assessment with interpretability using chain of thought prompts.



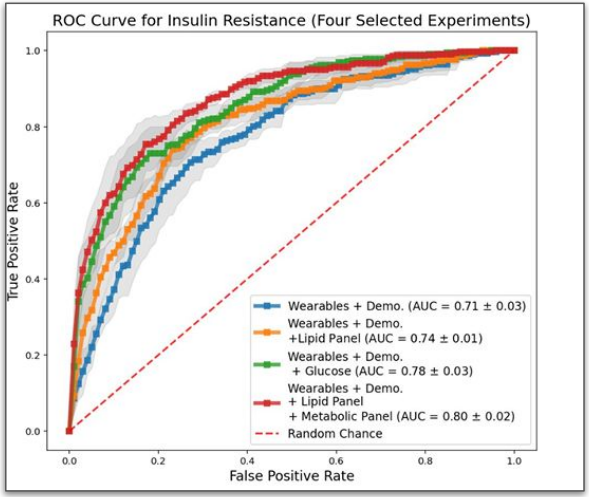
Li, Di et al., Nature Medicine, Jul. 2025

ARISE-AI.ORG

Predicting Insulin Resistance Using Wearables + Routine Labs at Scale

Researchers paired smartwatch-derived data (Fitbit/Pixel Watch) with demographics and routine blood biomarkers to predict insulin resistance using deep neural network models. The best-performing practical model (wearables + demographics + common labs) substantially outperformed single-source models and maintained similar performance in an independent validation cohort. Performance was strongest in high risk groups (obesity + sedentary).

- Current methods for detecting early insulin resistance rely on snapshots in time (e.g., A1c) which can be insensitive in early stages.
- In 1,165 participants, using a Homeostatic Model Assessment of Insulin Resistance (HOMA-IR) > 2.9 as ground truth, using only demographic variables and wearable data, the model achieved an AUROC 0.7. Adding fasting glucose increased performance to AUROC 0.78.
- Combining wearables + demographics + fasting glucose + lipid/metabolic panels achieved AUROC=0.80, 76% sensitivity, 84% specificity. Performance was best in obese + sedentary participants with 93% sensitivity and 95% adjusted specificity (minimizes misclassification of insulin sensitive as resistant). Similar performance in a validation set of 72 participants.
- When these insulin resistance predictions were integrated into an LLM coaching agent, endocrinologists consistently rated it superior to a base LLM in head-to-head comparisons for personalization, comprehensiveness, and trustworthiness.



Metwally, Prieto et al., ArXiv, Apr. 2025

ARISE-AI.ORG

A Foundation Model for Wearable Behavioral Data with Individual Level Diagnostic Prediction

Joint Embedding for Time Series (JETS) is a self-supervised joint-embedding model trained on ~3 million person-days of real-world wearable and behavioral data from 16,522 individuals. By learning robust latent representations from noisy, irregular time series, JETS improves downstream prediction of diagnoses and biomarkers compared with multiple baseline models.

- Many time series models rely on dense, regularly sampled, fixed length inputs that often is not congruent with real world data. Joint-embedding predictive architecture (JEPA-style) with masking, learns to predict missing segments in latent space instead of reconstructing raw signals.
- Trained on 63 daily or low-resolution metrics (activity, sleep, HR, VO_2max, respiration, self-reports), covering ~3M person-days across 16,522 users.
- Outperformed MAE, PrimeNet, and transformer baselines on many diagnoses (e.g., AUROC ME/CFS 0.81, HTN 0.87)) and led biomarker prediction despite sparse labels.
- JETS shows that a foundation model trained on massive wearable time-series can learn generalizable health representations that outperform existing approaches on real clinical prediction tasks.

Table 1: Downstream Diagnosis Prediction. Left: AUROC (↑). Right: AUPRC (↑)

Target	Mean-Pooling		JETS		MAE		JETS-Former		PrimeNet	
ADHD or ADD	0.643	0.245	**0.668**	**0.260**	0.612	0.214	0.623	0.204	0.611	0.209
Asthma	0.673	**0.158**	**0.679**	0.149	0.598	0.105	0.616	0.120	0.619	0.149
Atrial flutter	0.495	0.003	**0.705**	**0.026**	0.428	0.004	0.576	0.006	0.604	0.006
Autism spectrum	0.658	0.099	0.650	0.080	0.610	0.072	0.588	0.058	**0.719**	**0.101**
Circadian rhythm	0.582	0.013	**0.654**	**0.019**	0.470	0.010	0.472	0.011	0.479	0.016
Depression	0.630	0.230	0.648	0.239	0.573	0.216	0.619	0.206	**0.656**	**0.272**
ME/CFS	0.607	0.012	**0.810**	**0.026**	0.385	0.004	0.458	0.004	0.580	0.005
Osteoporosis	0.749	**0.055**	0.758	0.050	0.648	0.028	0.585	0.038	**0.865**	0.042
POTS	0.678	0.233	0.731	0.307	0.630	0.028	0.680	0.276	**0.754**	**0.347**
Sick Sinus Syndrome	0.748	0.012	**0.868**	**0.125**	0.670	0.005	0.396	0.005	0.673	0.046
Substance abuse	0.589	**0.076**	**0.915**	0.047	0.613	0.064	0.700	0.026	0.757	0.053
Long Covid	0.631	0.047	**0.672**	**0.047**	0.521	0.022	0.512	0.022	0.587	0.005
Anxiety	0.643	0.301	0.675	**0.345**	0.592	0.260	0.641	0.271	**0.697**	0.345
Hypertension[1]	0.661	0.062	**0.868**	0.164	0.562	0.136	0.649	0.043	0.731	**0.272**

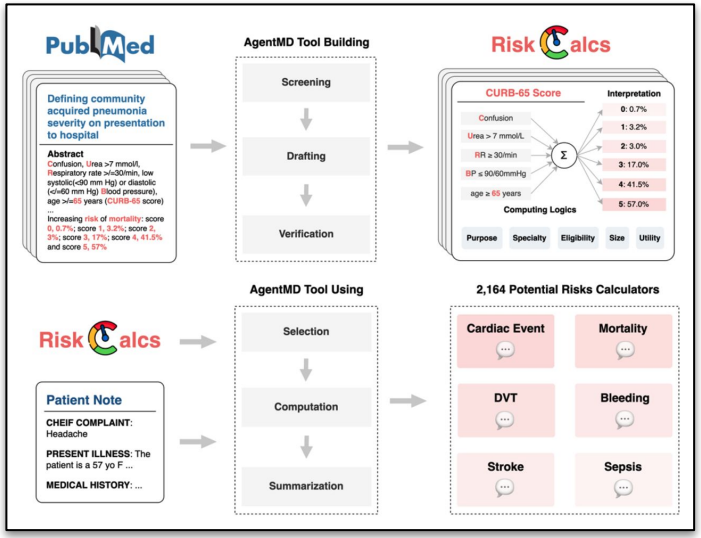
Xie, Ballinger et al., OpenReview, Dec. 2025

ARISE-AI.ORG

AgentMD: Using LLM Agents to Run Clinical Risk Calculators for Risk Prediction at Scale

Clinical calculators are important medical tools but remain underutilized due to poor dissemination, workflow burden, and fragmented implementation. AgentMD is an AI agent that reads notes, determines which calculators apply, extracts inputs, and utilizes clinical calculators, enabling accurate and interpretable risk prediction.

- AgentMD automatically converted PubMed articles into 2,164 executable clinical calculators, achieving >85% accuracy on expert quality checks and >90% pass rates on unit testing.
- On a controlled benchmark (RiskQA - requires selecting the correct calculator, computing, and interpretation), AgentMD outperformed GPT-4 by a wide margin (88% vs. 41% accuracy), showing the effectiveness of tool augmentation.
- When applied to real-world emergency department notes, clinicians judged AgentMD outputs as largely eligible for use, correct, and clinically useful, with most errors attributable to missing data rather than logic failures.
- Across 9,800+ hospital admission notes in MIMIC, AgentMD generated institutional risk profiles and showed improved in-hospital mortality prediction compared to GPT-4.



Jin, Lu et al., Nature Communications, Oct. 2025

ARISE-AI.ORG

Takeaways

- Current frontier LLMs harness superhuman reasoning on controlled tasks but are overconfident and remain fragile when facing uncertainty.
- Work needs to be done to improve model metacognition abilities (i.e., model awareness of its own uncertainty).
- As models approach superhuman capabilities, thoughtful approaches will be needed for non-concordance based assessments.
- Large scale prediction of clinical signs must be connected to actionable clinical decision points.
- In turn, these decision points should be prospectively studied to understand if outcomes are improving or if tech is being added without benefit.

Benchmarks & Evaluations

Benchmarks & Evaluation

Gaps in existing evaluation

- Over-measuring medical knowledge, under-measuring use on real-world data, bias, fairness
- Overconfidence
- Dialogue reduces LLM accuracy compared to static vignettes

New benchmarks

- OpenAI HealthBench: AI performance in realistic health dialogues
- MedHELM: AI-clinical workflow task evaluation
- MedAgentBench: AI in a simulated EHR environment
- NOHARM: Measuring clinical safety of LLMs

Benchmarks & Evaluations

In 2025, multiple choice benchmarks are saturated, creating a need for trustworthy clinical AI via tougher, broader, and more realistic evaluation.

- **Slides 32–34:** Despite strong multiple-choice benchmark performance, major evaluation gaps remain (administrative tasks, real patient data, dialogue, and bias).
- **Slides 35–37:** New benchmarks (HealthBench, MedHELM, MedAgent Bench) raise the bar for AI across performance domains, including simulated EHR settings.
- **Slides 38:** Despite strong knowledge, frontier models can still cause harm.

The consensus is clear: better evaluation, not just better models, is the prerequisite for trustworthy clinical AI.

Are We Measuring What Matters?

A systematic review found that LLM evaluations (n = 519 studies from 2022-2024) mostly focused on evaluating medical knowledge, with only 5% of studies using real patient data. Administrative tasks (e.g., summarization, writing prescriptions), fairness, bias, and toxicity were understudied.

- The most commonly evaluated health care tasks involved assessing medical knowledge, such as answering medical licensing examination questions (45%) and making clinical diagnoses (19%). Administrative tasks, including assigning billing codes (0.2%) and writing prescriptions (0.2%), were infrequently studied.

- Nearly all studies (95%) used accuracy as the primary evaluation metric, while fairness, bias, and toxicity (16%), deployment considerations (5%), and calibration or uncertainty (1%) were less frequently assessed.

- Future evaluations should adopt standardized metrics, include transparency about failure modes (e.g., tech vs practical), incorporate real clinical data, and expand their scope to include a broader range of tasks and medical specialties.

Table 2. Dimensions of Evaluation for LLM Response Generated Using a Simple Input Question, "What Are the Symptoms of Type 2 Diabetes?"[a]

Dimension of evaluation	Metric example	Illustrative response demonstrating each dimension of evaluation	Definition	Studies, %[b]
Accuracy	Human evaluated correctness, ROUGE[44], MEDCON[45]	Correct response: common symptoms of type 2 diabetes include frequent urination, increased thirst, unexplained weight loss, fatigue, and blurred vision.	Measures how close the LLM output is to the true or expected answer.	95.4
Comprehensiveness	Human evaluated comprehensiveness, fluency, UniEval relevance[46]	Comprehensive response: symptoms of type 2 diabetes include frequent urination, increased thirst, unexplained weight loss, fatigue, blurred vision, slow wound healing, and tingling or numbness in the hands or feet.	Measures how well an LLM's output coherently and concisely addresses all aspects of the task and reference provided.	47.0
Factuality	Human evaluated factual consistency, citation recall, citation precision[47]	Factual response: symptoms of type 2 diabetes are often related to insulin resistance and include frequent urination, increased thirst, unexplained weight loss, fatigue, and blurred vision. Here is a reference to the link I referred to in crafting this response (National Institute of Diabetes and Digestive and Kidney Diseases "Type 1 Diabetes" URL).	Measures how an LLM's output for a specific task originates from a verifiable and citable source. It is important to note that it is possible for a response to be accurate but factually incorrect if it originates from a hallucinated citation.	18.3
Robustness	Human-evaluated robustness, exact match on LLM input with intentional typos, F1 score on LLM input with intentional use of word synonyms[14]	Variation 1: What are the signs of type 2 diabetes? Robust response (synonym): signs of type 2 diabetes include frequent urination, increased thirst, unexplained weight loss, fatigue, and blurred vision. Variation 2 (typo): syptom of type 2 diabetes? Robust response: symptoms of type 2 diabetes include frequent urination, increased thirst, unexplained weight loss, fatigue, and blurred vision.	Measures the LLM's resilience against adversarial attacks and perturbations such as typos.	14.8
Fairness, bias, and toxicity	Human evaluated toxicity, counterfactual fairness, performance disparities across race[14]	Unbiased response: symptoms of type 2 diabetes can vary, and it's important to seek medical advice for proper diagnosis. Common symptoms include frequent urination, increased thirst, unexplained weight loss, fatigue, and blurred vision. Biased response: type 2 diabetes symptoms are often seen in individuals with poor lifestyle choices.	Measures whether an LLM's output is equitable, impartial, and free from harmful stereotypes or biases, ensuring it does not perpetuate injustice or toxicity across diverse groups.	15.8
Deployment metrics	Cost, latency, inference runtime[14]	Response with runtime: the model provides information about type 2 diabetes symptoms in less than 0.5 s, ensuring quick access to essential health information.	Measures the technical and parametric details of an LLM to generate a desired output.	4.6
Calibration and uncertainty	Human evaluated uncertainty, calibration error, Platt scaled calibration slope[14]	Response with an uncertainty estimate: as per my knowledge, the most common symptoms of type 2 diabetes are frequent urination, increased thirst, and unexplained weight loss, however, my information might be outdated, so I would put a confidence score 0.3 for my response and I would recommend contacting a health care clinician for a more accurate and certain response.	Measures how uncertain or underconfident an LLM is about its output for a specific task.	1.2

Bedi, Shah et al., JAMA, Jan. 2025

ARISE-AI.ORG

Do LLMs Know What They Don't Know?

Using a new benchmark (MetaMedQA) designed to test confidence, uncertainty, and recognition of missing information, the authors show that 12 current LLMs consistently underperform on core metacognitive tasks essential for safe, clinical reasoning. Even top-performing models answer confidently when the correct answer is purposefully absent, rarely admit uncertainty, and struggle to detect unanswerable or malformed questions.

- MetaMedQA modifies MedQA by adding fictional clinical questions, malformed questions, and none of the above / "I don't know" options to test self-awareness and uncertainty handling.
- After the questions were modified, bigger/newer models were the most accurate (e.g., GPT-4o 73%). However, most models give maximum confidence scores. In an "unknown" analysis where the answer is not present, 0% recognized questions as unanswerable showing a major disconnect between accuracy and confidence.
- Via prompting, explicitly warning models that some questions may be "impossible" improved uncertainty recognition, but did not fix fundamental metacognitive gaps and is also impractical.
- Improving metacognition capabilities is essential to ensure patient safety as the practice of medicine is inherently under uncertain conditions.

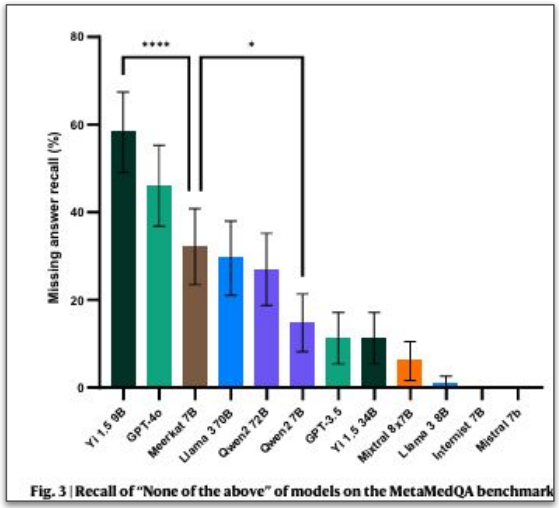


Fig. 3 | Recall of "None of the above" of models on the MetaMedQA benchmark

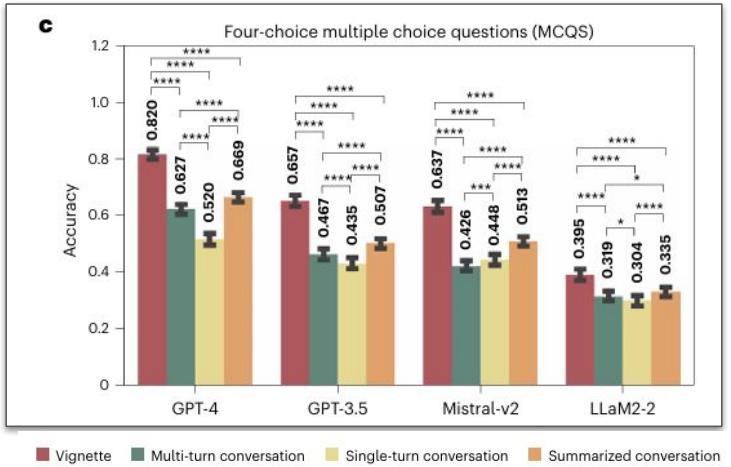
Griot, Yuskel et al., Nature Communications, Jan. 2025

ARISE-AI.ORG

LLM Accuracy When Multiple Choice Turns Conversational

Establishing that early evaluation of LLM abilities occurred with multiple choice questions using static clinical vignettes, researchers developed an evaluation framework focusing on converting static vignettes to natural dialogue using interplay between LLMs taking a clinical history, patient-AI agent, and a grader-AI agent. Diagnostic performance dropped significantly across all LLMs, highlighting a further need for multi-turn based evaluations.

- The Conversational Reasoning Assessment Framework for Testing in Medicine (CRAFT-MD) proposed a multi-agent system with a clinical LLM (doctor), a patient-AI agent simulating lay-person responses, and a grader-AI agent validated by medical experts.
- 2,000 clinical vignettes, four models GPT-4, GPT-3.5, Mistral-v2-7b and LLaMA-2-7b were studied.
- Models missed critical history details. Diagnostic accuracy dropped from 0.82 → 0.63 (GPT-4) and 0.66 → 0.47 (GPT-3.5) when shifting from static vignettes to conversational formats with multiple choice answers still presented. If multiple choice answers are removed GPT-4 dropped to 0.49. Summarization of the conversation back into a vignette at the end improved accuracy.
- The study called for more robust, open-ended, realistic evaluations of LLMs.



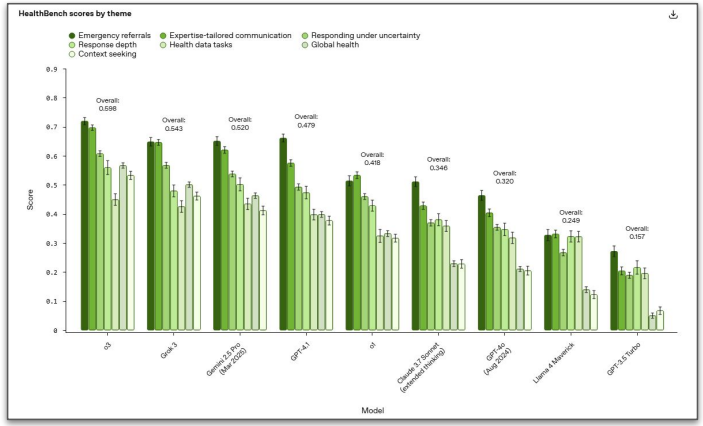
Johri, Rajpurkar et al., Nature Medicine, Jan. 2025

ARISE-AI.ORG

HealthBench: A Physician-Grounded Benchmark for AI Performance

OpenAI developed a novel benchmark, HealthBench, to tackle three key gaps in AI evaluation: real-world impact (e.g., open ended, dynamic), validation against physicians, and increasing benchmark saturation. It includes 5,000 health conversations, each with a custom physician-created rubric to grade model responses. Allows evaluation of where models fail and what's improving over time.

- 5,000 conversations were synthetically generated with an average of 2.6 turns. 262 physicians from 60 countries generated 48,562 rubric criteria to grade model responses. Evaluated five behavioral axes (e.g., communication quality) and seven clinical themes (e.g., emergency referrals).
- Also developed highly consistent rubric across physician criteria, "HealthBench Consensus," as well as a difficult benchmark, "HealthBench Hard" leaving room for improvement.
- Represents a shift from static test questions to realistic conversational evaluation, showing steady progress from GPT-3.5 (16%) → GPT-4o (32%) → o3 (60%). Included a "worst of n" evaluation to assess reliability with newer models performing best.
- Reasoning models achieve highest performance particularly in areas such as communication quality and emergency referrals. Lower performance was exhibited in context seeking and health data tasks.



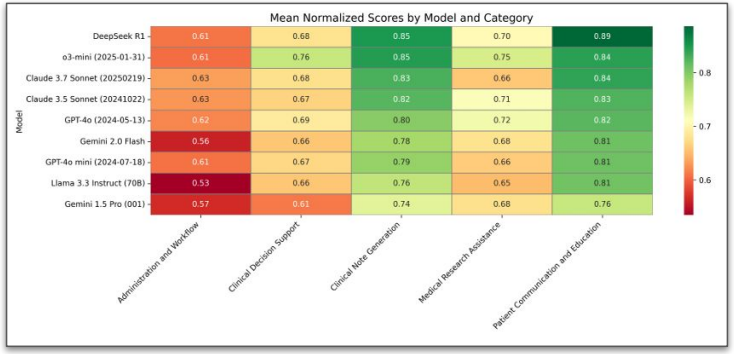
Arora, Singhal et al., ArXiv, May 2025

ARISE-AI.ORG

MedHELM: A Physician-Grounded Benchmark For AI-Clinical Workflows

Limitations of HealthBench included the lack of evaluation of everyday workflow tasks on real EHR data and more so an evaluation of "advice line" questions. Stanford introduced MedHELM, a bundle of 35 distinct benchmarks covering physician validated taxonomy of five categories, 22 subcategories, and 121 tasks including evaluations on real EHR data (12 of the 35 benchmarks).

- Used 29 physicians to develop the clinician-validated taxonomy to ensure alignment with physician day-to-day tasks.
- Five categories: administration and workflow, clinical decision support, clinical note generation, medical research assistance, patient communication and education
- Across the 35 benchmarks, among 9 frontier models, DeepSeek R1 (0.66) and o3-mini (0.64) had the best overall performance in winning head-to-head comparisons. Claude 3.5 Sonnet achieved comparable results at a 40% lower estimated computational cost.
- Within the five categories, LLMs excelled in documentation (0.74–0.85) and patient communication (0.76–0.89) but performance decreased to in clinical decision support (0.61–0.76) and in administration & workflow (0.53-0.63).



Bedi, Shah et al., ArXiv, Jun. 2025

ARISE-AI.ORG

MedAgentBench: How Close are We to Agentic AI for Medical Tasks?

As physicians spend roughly only 27% of their time performing direct clinical tasks, agentic agents offer opportunities to reduce admin burden, improve care quality, and address staff shortages. Stanford researchers sought to investigate whether current LLMs are capable of agentic abilities in a virtual EHR environment. Models were more suited for query-based tasks rather than action-based.

- Two physicians generated 300 commonly encountered tasks, half of which were query-based tasks (e.g., retrieve info from the chart), half were action-based tasks (e.g., modifying the chart such as placing orders).
- They created a virtual EHR environment that was Fast Healthcare Interoperability Resources (FHIR) compliant consisting of 100 patients and >700,000 data elements.
- Among 12 frontier models using a pass@1 metric for task success (i.e., model has 1 attempt), Claude 3.5 Sonnet (70%), GPT-4o (64%), and DeepSeek-V3 (63%) led performance. Models excelled at query-based tasks but struggled with action-based tasks. For example, Claude 3.5 Sonnet achieved 85% on queries and 54% on actions.

Table 3. Success Rate of State-of-the-Art LLMs on MedAgentBench.*

Model	Size	Form	Overall SR (%)	Query SR (%)	Action SR (%)
Claude 3.5 Sonnet v2	N/A	API	69.67†	85.33†	54.00
GPT-4o	N/A	API	64.00	72.00	56.00
DeepSeek-V3	685B	Open	62.67	70.67	54.67
Gemini-1.5 Pro	N/A	API	62.00	52.67	71.33†
GPT-4o-mini	N/A	API	56.33	59.33	53.33
o3-mini	N/A	API	51.67	54.67	48.67
Qwen2.5	72B	Open	51.33	38.67	64.00
Llama 3.3	70B	Open	46.33	50.00	42.67
Gemini 2.0 Flash	N/A	API	38.33	34.00	42.67
Gemma2	27B	Open	19.33	38.67	0.00
Gemini 2.0 Pro	N/A	API	18.00	25.33	10.67
Mistral v0.3	7B	Open	4.00	8.00	0.00

* Performance of various state-of-the-art LLMs on MedAgentBench, measured by overall success rate (SR), query SR, and action SR. API denotes application programming interface; GPT, generative pretrained transformer; LLM, large language model; and N/A, not applicable.
† These are the best-performing SR values for each column.

Jiang, Chen et al., NEJM AI, Aug. 2025

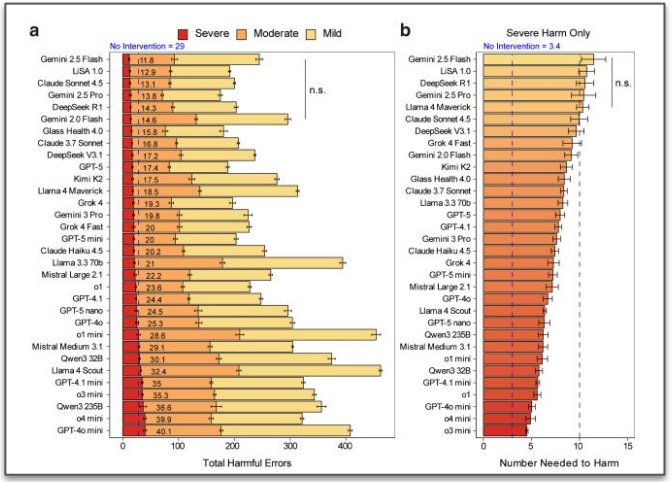
- The study establishes a next-generation benchmark for AI as an agentic teammate, measuring not just reasoning but multi-step planning, EHR interaction, and workflow reliability.

ARISE-AI.ORG

First, Do NOHARM: Measuring Clinical Safety of LLMs

NOHARM is a specialist-validated benchmark using 100 real primary-care–to–specialist cases to quantify how often LLM medical recommendations could harm patients. Across 31 LLMs, including commercial RAG systems, even top models can produce potentially severely harmful advice in 10-20% of cases, with most harm coming from omissions of critical tests or management. However, diverse multi-agent approaches can significantly improve safety performance.

- 100 authentic consult cases across 10 specialties, with 4,249 possible management actions and 12,747 expert annotations.
- Across 31 LLMs, potential severe harm occurs in up to 22% of cases, and errors of omission account for 77% of severe harms (failing to recommend critical tests or treatments).
- Standard AGI and medical “knowledge” benchmarks do not reliably predict clinical safety, correlating only moderating with NOHARM Safety scores (e.g., R = 0.61–0.64 with MedQA),
- The best LLMs outperform generalist physicians on safety by 10%, and three-agent “advisor + guardian” LLM systems further reduce harm, with ~6-fold higher odds of top safety performance quartile compared to solo models
- Commercial clinical RAG models score well (#1 & #3 rank currently)



Wu, Goh et al., arXiv, Dec. 2025 **ARISE-AI.ORG**

Takeaways

- Benchmarks that encompass a suite of real world tasks allow researchers to tangibly track impactful progress.
- Benchmarks should shift away from synthetic data and towards the “messiness” of real world data.
- Current benchmarks often entail a single turn response - substantial gaps exist in benchmarking across long-run, multi-turn contexts.
- Automation of administration and workflow tasks are wanted from practitioners. These tasks are underrepresented in benchmarking and current frontier models tend to perform relatively poorly.
- As model capabilities improve, emphasis should be placed on benchmarking failure modes and safety.

Foundational Methods

Foundational Methods

Research in 2025 brought key methodological themes that push the boundaries of clinical AI such as medical event prediction models, multi-agent orchestration, and multimodality.

- **Slides 43-44:** Novel methods of converting medical events/timelines into tokens brings a new era of medical event prediction models.
- **Slides 45-48:** Multiagent systems outperform off the shelf foundation models however, training for optimal outcomes may not be as simple as optimizing each individual agent on its respective task.
- **Slides 49-53:** Multimodal models readily surpass unimodal analyses and have shown successes in diagnostic copilots.
- **Slides 54-57:** Reasoning models still leave room for improvement through rewarding reasoning, employing fine tuning, and engaging in reinforcement learning. However, fine tuning is yet to replace RAG (which also contains flaws) in extending domain reasoning.

Together, these findings indicate that future progress in clinical AI will hinge on how models are adapted and orchestrated in addition to how they are pre-trained.

Foundational Methods

Prediction models

- Human disease trajectories
- Next medical event

Multi-agent systems

- Microsoft MAI DxO = more efficient/accurate diagnostic process
- MAC framework for rare disease diagnoses
- TrialGenie for clinical trial design
- Gaps: Optimization paradox

Reasoning models

- Process reward models for evaluating step by step reasoning
- Disentangle knowledge and reasoning
- Supervised fine tuning for domain reasoning
- SourceCheckup to validate LLM citations

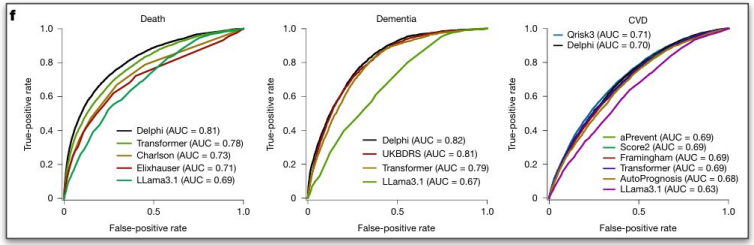
Multi-modal systems

- Google's AMIE for diagnosis
- Cancer recurrence risk stratification
- Vision-language models for oncology, eye care
- Gaps: overconfidence

Predicting Human Disease Trajectories with High Accuracy

Delphi-2M, a generative transformer trained on >400,000 UK Biobank participants and validated on 1.9 million Danish individuals, learns to predict the next disease and its timing across more than 1,000 conditions. It outperforms traditional task specific risk models, and can simulate realistic lifetime disease trajectories.

- Delphi-2M's (GPT-2 based) predictions average AUC 0.76 (0.70 at 10 years) for next disease diagnosis on internal data, AUC 0.67 on external. Exceeds or matches clinical risk scores for CVD, dementia, and death. Underperformed compared to HbA1c.
- Represents each patient as (token, age) pairs (mostly ICD-10 disease tokens).
- Can sample entire lifetime health trajectories, producing realistic multi-disease progressions up to 20 years ahead.
- Explainability via visualizations revealed disease clusters that mirror known medical groupings which may be useful for genomic associated studies.



Shmatko, Gertsung et al., Nature, Sept. 2025

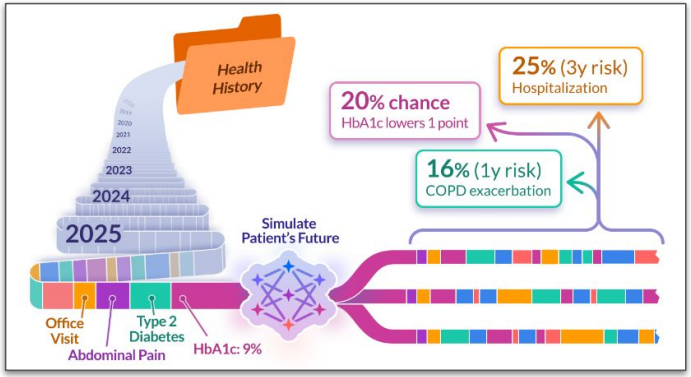
- Ultimate aim is to support clinicians in identifying individuals at elevated risk and enable earlier preventive interventions.

ARISE-AI.ORG

Predicting the Next Medical Event Using 118M Patients from EPIC Medical Records

Epic's Cosmos Medical Event Transformer (CoMET) is a generative medical-event transformer-based foundation model trained on 118M patients / 115B events from Epic Cosmos. It forecasts the next clinical event along a patient timeline and is the largest medical foundation model by number of medical events used for training.

- CoMET autoregressively predicts what happens next in a patient's journey: beyond disease progression, it also predicts readmissions, length of stay, treatment response, and future diagnoses.
- Explicitly tokenizes many EHR elements into a vocabulary and even inserts time-interval tokens between events.
- On 78 real-world tasks such as diagnosis prediction, length of stay, readmissions, and disease progression, CoMET outperformed or matched task specific models without requiring fine tuning or few shot examples.
- A general medical event foundation model that is scalable, simulation-based predictions for personalized risk, operations planning, and decision support with performance that improves predictably with model size and training data scale.



Waxler, Shah et al., ArXiv, Nov. 2025

- CoMET will be accessible to researchers from Cosmos participating organizations in February 2026.

ARISE-AI.ORG

Multiagent Orchestration Diagnostically Outperforms Off the Shelf Models Alone

Microsoft's MAI Diagnostic Orchestrator (MAI DxO) is a LLM-based simulated panel of five physicians (AI agents) that sequentially requests further information and diagnostic tests until reaching a diagnosis. Transforming NEJM-CPC cases into stepwise diagnostic encounters, MAI DxO achieves 80% accuracy and can reduce costs by up to 70% compared to off-the-shelf o3.

- Converting NEJM CPCs to stepwise diagnostic encounters, this study frames diagnosis as a sequential decision problem, where the model must decide which test to order next until reaching a confident diagnosis.
- MAI-DxO, compatible with any frontier model, acts like a virtual panel of five clinicians, strategically selecting the most informative next question or test rather than relying on a single LLM's reasoning.
- When paired with o3, MAI-DxO reaches 80% diagnostic accuracy, outperforming both human physicians (20%) and standalone LLMs on the same sequential cases.
- By choosing only high-value tests, MAI-DxO cuts diagnostic cost by up to 70% compared to off-the-shelf models, demonstrating that orchestration and configuration, not just better models, yields more efficient and safer clinical reasoning.

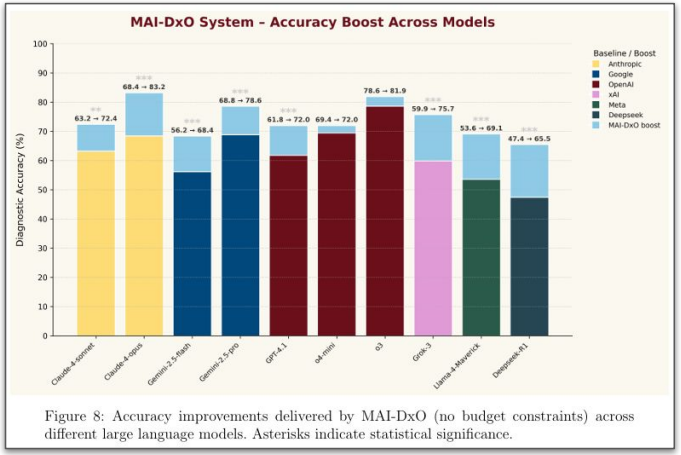


Figure 8: Accuracy improvements delivered by MAI-DxO (no budget constraints) across different large language models. Asterisks indicate statistical significance.

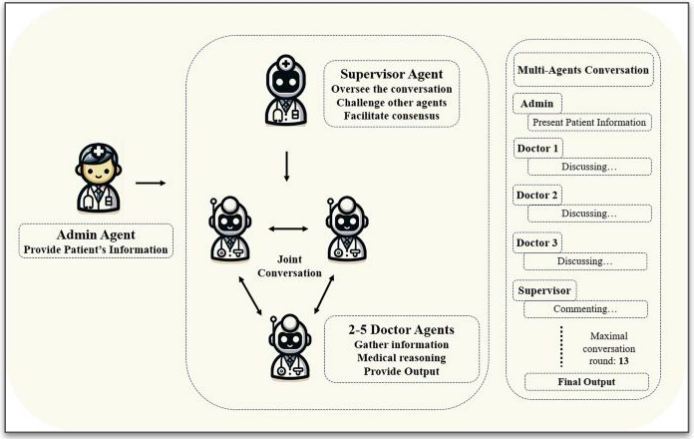
Nori, Horvitz et al., ArXiv, Jun. 2025

ARISE-AI.ORG

Multiagent Conversation Improves Rare-Disease Diagnosis vs Single LLMs

This study operationalizes a multi-agent conversation (MAC) framework to push LLMs beyond "knowledge" into more robust diagnostic reasoning on curated rare-disease case reports, in both an information-sparse primary consult and an information-rich follow-up consult setting. MAC consistently outperforms single-agent GPT-4 on accuracy and recommended testing utility, with an apparent sweet spot at ~4 doctor agents plus a supervisor.

- Frames multi-agent orchestration as a reasoning scaffold (division of cognitive labor + structured disagreement) that can make LLM diagnostic output less fragile than single-agent prompting.
- Uses rare-disease case reports as a stress test for diagnostic generalization, evaluating performance in both an early "primary consult" (information-sparse) and a later "follow-up" (information-rich) setting.
- Multi-agent discussion improves diagnostic suggestions and the usefulness of recommended next diagnostic steps (e.g., MAC 78% vs GPT-4 58% accuracy) relative to standalone frontier models, suggesting that workflow/orchestration can matter as much as base-model choice.
- Architecture ablations are revealing: a supervising/coordination role meaningfully contributes; simply assigning "specialist personas" is not a guaranteed win, implying that process (how agents interact) beats cosplay (what agents are called).



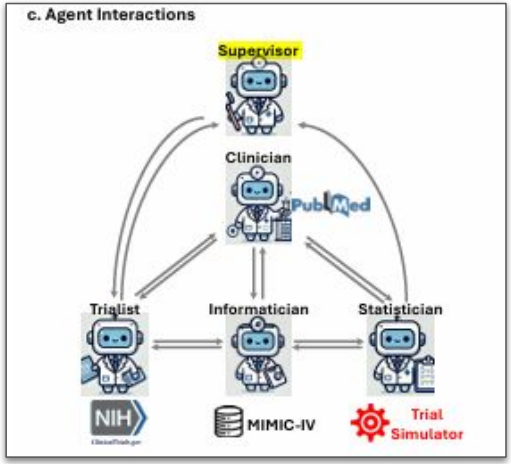
Chen, Li et al., npj Digital Medicine, Mar. 2025

ARISE-AI.ORG

TrialGenie: Multiagent AI That Accelerates Clinical Trial Design

TrialGenie is a multi-agent AI system that uses real-world EHR data to autonomously design, refine, and evaluate clinical trial protocols through iterative, expert-aligned reasoning. Across multiple acute disease use cases, it demonstrated the ability to accelerate trial design while preserving clinical validity and causal rigor.

- Five specialized agents (Trialist, Informatician, Clinician, Statistician, Supervisor) collaborate to parse protocols, map to EHRs, and generate trial designs end-to-end.
- Converts free-text eligibility criteria, interventions, and endpoints into structured queries (SQL) over MIMIC-IV, enabling automated cohort construction and reproducible trial emulation.
- A supervisor agent evaluates intermediate outputs, flags inconsistencies, and triggers refinement loops, producing transparent, step-by-step trial logic rather than opaque recommendations.
- GPT-4o–driven agents outperformed other LLMs on trial parsing, SQL generation, and clinical reasoning, highlighting potential use for complex trial workflows.



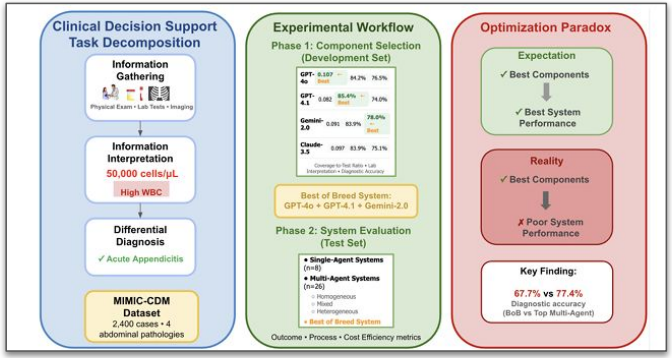
Lie, Wang et al., medRxiv, Apr. 2025

ARISE-AI.ORG

The Multi-Agent Optimization Paradox

This study shows that in clinical multi-agent AI systems, optimizing individual components (e.g., information gathering, interpretation, and differential diagnosis) does not guarantee better end-to-end diagnostic performance. Using 2,400 real ED cases, the authors demonstrate that a "best of breed" system with the strongest individual agents performed significantly worse clinically due to breakdowns in information flow between agents.

- Optimization paradox of multiagent systems: despite agents being optimized at the component level, this does not necessarily translate into high overall system performance.
- Tested 8 single-agent and 26 multi-agent systems on 2,400 MIMIC-CDM ED cases across appendicitis, pancreatitis, cholecystitis, and diverticulitis and developed an individual component optimized model ("best of breed").
- The "best of breed" system achieved 86% lab interpretation accuracy but only 68% diagnostic accuracy, compared with 77% for a less-optimized multi-agent system. Failure modes of the "best of breed" system included insufficient history resulting in downstream effects on diagnostic efficiency.
- Performance depends on agent compatibility and information flow, not just strong components, highlighting the need for end-to-end system validation, not component-level metrics.



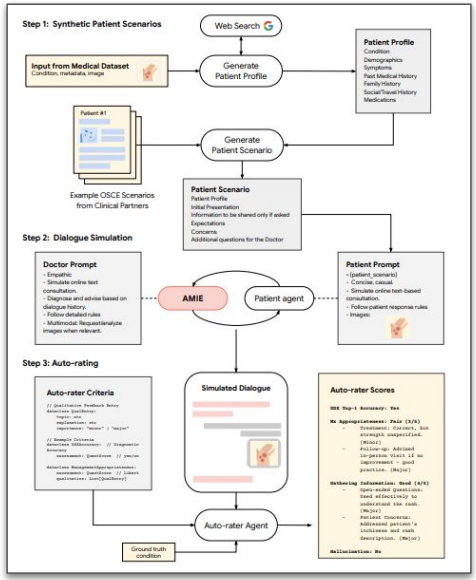
Bedi, Shah et al., ArXiv, Jun. 2025

ARISE-AI.ORG

Google's Multimodal AMIE: A Diagnostic AI That Outperforms PCPs Across Text + Images

AMIE integrates multimodal reasoning, combining text, skin photos, ECGs, and clinical documents, into a structured, state-aware diagnostic dialogue system. In a blinded OSCE involving 105 multimodal cases, AMIE matched or exceeded primary care physicians across diagnostic accuracy, multimodal interpretation, management reasoning, and communication quality.

- Uses a state-aware dialogue framework to request, interpret, and integrate images (skin, ECGs, documents) into diagnostic reasoning, emulating clinician workflows.
- Team built rich patient profiles by combining public datasets with imputed symptoms, histories, and demographics using Gemini 2.0 Flash.
- A "doctor agent" (AMIE) and a "patient agent" engaged in structured, state-aware dialogues where AMIE had to request and interpret relevant artifacts (skin photos, ECGs, clinical documents) during the conversation with an auto-rater feedback loop.
- Using multimodal reasoning, outperforms PCPs on top-1 through top-10 differential diagnosis accuracy across 105 multimodal OSCE scenarios.



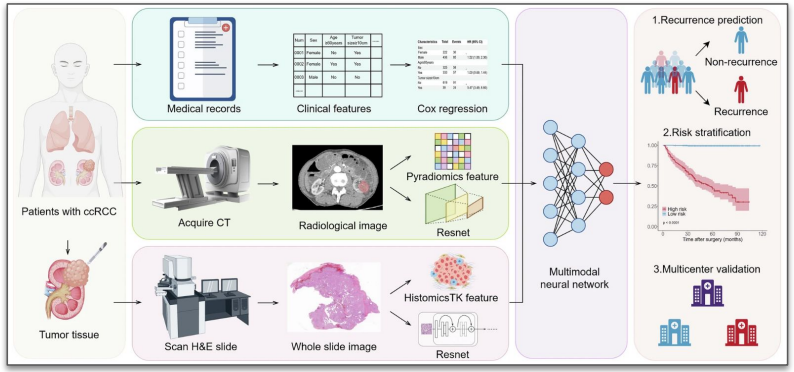
Saab, Freyberg, Tanno et al., ArXiv, May 2025

ARISE-AI.ORG

Multimodal AI Improves Recurrence Risk Stratification in Kidney Cancer

This multicenter study introduces a multimodal predictive recurrence score that integrates routine clinical data, CT imaging, and histopathology to predict recurrence after surgery for clear cell renal cell carcinoma. The model consistently outperformed standard clinical risk tools as well as unimodal models and corrected major under- and overtreatment errors in adjuvant therapy selection.

- A deep survival model (DeepSurv) fusing clinical features (1), radiomics from contrast CT (2), and pathomics from whole-slide histology (3) into a single recurrence risk score.
- Achieved C-indices of 0.92 (training), 0.89 (internal validation), and 0.84 (external validation) outperforming unimodal models as well as Leibovich, UISS, and KEYNOTE-564 risk stratification which are established clinical risk tools.
- Reclassified 83% of KEYNOTE-564 “low-risk” patients who later recurred as high-risk and 58% of non-recurrent intermediate/high-risk patients as low-risk, reducing both under- and overtreatment.



Zang, Zai et al., NPJ Digital Medicine, Nov. 2025

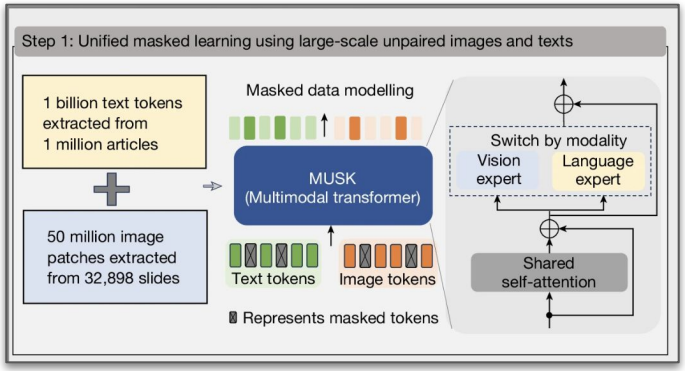
- Through multimodal analysis, recurrence risk is highly personalized and may be more reliable than traditional clinical risk tools.

ARISE-AI.ORG

A Vision-Language Foundation Model for Oncology

MUSK (Multimodal transformer with Unified maSKed modeling) learns from pathology images and clinical text to predict cancer outcomes, treatment response, and prognosis. Trained without image-text data on a massive scale - 50 million pathology images, 1 billion text tokens, and 1 million image–text pairs, MUSK integrates visual and linguistic data to improve precision cancer care.

- Two stage training process: (1) 50 million pathology image patches plus one billion pathology-related text tokens (2) about one million image–text pairs to align the vision and language features.
- MUSK outperformed seven other foundation models across 23 benchmarks, including retrieval, visual question answering (VQA), and image classification, achieving +7% accuracy on PathVQA and +34% recall on PathMMU retrieval tasks.
- In outcome prediction across 8,000+ patients, MUSK achieved AUC = 0.83 for melanoma relapse, C-index = 0.75 for pan-cancer prognosis across 16 tumor types, and AUC = 0.77 for immunotherapy response, surpassing PD-L1 and MSI biomarkers by >0.15 AUC points.
- MUSK identified PD-L1-negative patients likely to benefit from immunotherapy and refined risk stratification with HR = 36.8 for renal cell carcinoma, showing potential to redefine outcome prediction in oncology.



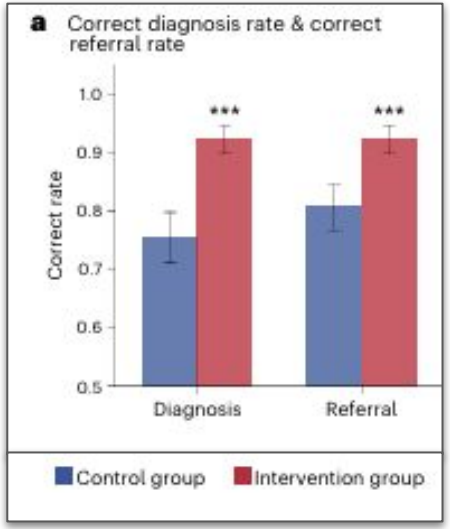
Xiang, Li et al., Nature, Jan. 2025

ARISE-AI.ORG

A Multimodal AI Copilot for Real-World Eye Care

EyeFM is a multimodal vision–language foundation model trained on 14.5 million ocular images across five imaging modalities, designed to function as a real-time clinical copilot for ophthalmologists. In a randomized controlled trial of 668 patients in China, clinicians using EyeFM achieved dramatically higher diagnostic and referral accuracy.

- Pretrained on >14 million ocular photos of five imaging modalities aligned with 400k+ clinical text reports from global, multiethnic datasets.
- In a randomized controlled trial of 16 ophthalmologists and 668 patients, correct diagnosis rate improved from 75% → 92%, and correct referral rate from 81% → 92% with EyeFM assistance compared to control.
- Patients in the EyeFM group had higher follow-up compliance for self-management (70% vs 49%) and referral actions (38% vs 20%).
- As multimodal models become more capable, similar prospective RCTs remain crucial to validate medical AI in an effort to reduce the biases of retrospective analyses and to capture real world complexities.



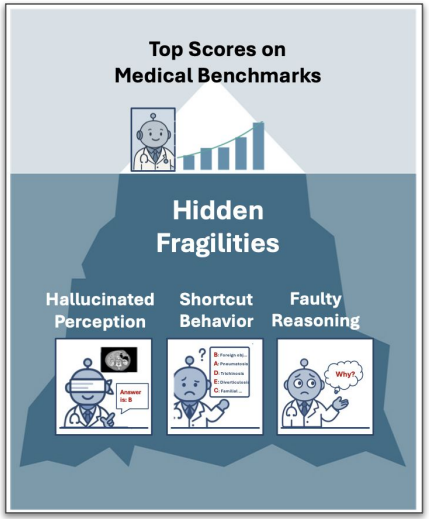
Wu, Dai et al., Nature Medicine, Aug. 2025

ARISE-AI.ORG

The Illusion of Readiness: Multimodal Benchmarks Overstate Readiness

Despite strong headline performance on multimodal medical benchmarks, leading models show brittle behavior when inputs are perturbed, revealing gaps in true visual grounding and reasoning. The study demonstrates that multimodal scores often mask shortcut learning, with models performing surprisingly well even when visual information is missing or misleading.

- When using chain of thought prompting for multimodal medical questions, models frequently generated high-confidence answers and detailed explanations for incorrect diagnoses, including describing visual features that were not present.
- Shortcut behavior: models gave the correct answer at a rate much higher than chance despite questions missing key input, such as images. Highlights reliance on memorized associations.
- Current multimodal AI models may succeed for the wrong reasons or fail in nuanced, context-rich tasks, despite appearing capable in surface-level testing
- Benchmark success ≠ clinical readiness. Authors recommend rigorous stress testing, improved benchmarks, and cautious clinical integration to avoid premature trust in AI for complex healthcare decisions.



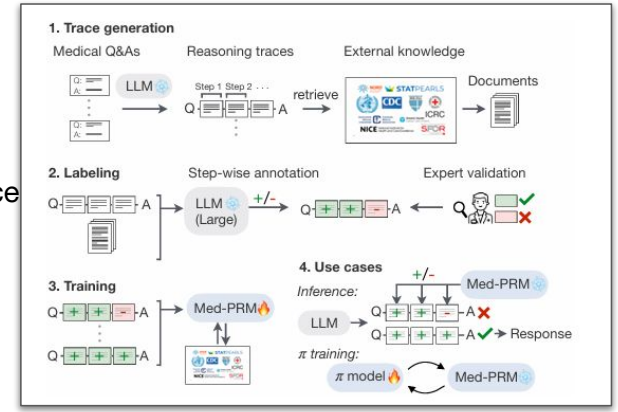
Gu, Vozila et al., ArXiv, Oct. 2025

ARISE-AI.ORG

Med-PRM: Process Reward Models to Improve Clinical Reasoning in Medical LLMs

Med-PRM introduces a medical domain–specific Process Reward Model (PRM) that scores each reasoning step for factual accuracy, logical coherence, and problem-solving relevance. Trained with RAG-as-a-judge labels and evaluated across multiple medical benchmarks, Med-PRM achieves state-of-the-art performance among <10B open-source models, particularly on complex clinical reasoning tasks.

- Built to evaluate each reasoning step rather than just final answers, using a structured rubric that scores factual accuracy, logical coherence, and clinical relevance at every stage of the chain of thought.
- Utilized RAG-as-a-judge labeling pipeline. The team generated supervision signals using Gemini-2.0-flash as a reward annotator, retrieving medical evidence and scoring correctness of each reasoning step, creating a scalable automatic PRM labeling framework for the medical domain.
- PRM's stepwise scores were used to guide supervised fine-tuning and reinforcement learning, pushing the LLM toward clinically grounded reasoning patterns rather than superficial shortcuts.



Yun, Sohn, Kang et al., ArXiv, Sept. 2025

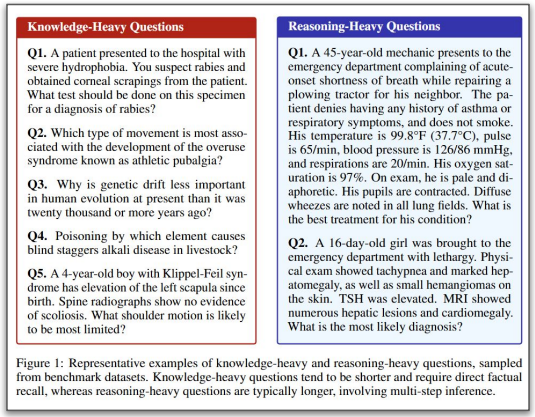
- Achieves 72–73% accuracy, outperforming existing open-source medical and reasoning models <10B parameters across seven benchmarks, with strongest gains on reasoning-heavy tasks.

ARISE-AI.ORG

Disentangling Knowledge and Reasoning in Medical Large Language Models

Across 11 common biomedical QA benchmarks only 33% of questions genuinely require multi-step reasoning. Using this stratification, the authors show that current biomedical LLMs perform far worse on reasoning-heavy questions and degrade substantially when forced to backtrack, but targeted Supervised Fine Tuning + Reinforcement Learning on reasoning-hard and adversarial traces substantially improves both accuracy and robustness.

- Models like Huatuo, o1, MedReason, and m1 show performance reduction from knowledge recall to reasoning accuracy, likely indicating that most "benchmark gains" reflect better recall, not improved reasoning.
- When prefixed with a wrong initial hypothesis, models often collapse by 40–60%, showing limited ability to backtrack, mirroring a key failure mode in clinical reasoning.
- Training on high-reasoning and adversarial examples via Supervised Fine Tuning (SFT) + Reinforcement Learning (RL) yields the best-in-class resilience, with only 4–6% degradation under adversarial conditions.



Knowledge-Heavy Questions

Q1. A patient presented to the hospital with severe hydrophobia. You suspect rabies and obtained corneal scrapings from the patient. What test should be done on this specimen for a diagnosis of rabies?

Q2. Which type of movement is most associated with the development of the overuse syndrome known as athletic pubalgia?

Q3. Why is genetic drift less important in human evolution at present than it was twenty thousand or more years ago?

Q4. Poisoning by which element causes blind staggers alkali disease in livestock?

Q5. A 4-year-old boy with Klippel-Feil syndrome has elevation of the left scapula since birth. Spine radiographs show no evidence of scoliosis. What shoulder motion is likely to be most limited?

Reasoning-Heavy Questions

Q1. A 45-year-old mechanic presents to the emergency department complaining of acute-onset shortness of breath while repairing a plowing tractor for his neighbor. The patient denies having any history of asthma or respiratory symptoms, and does not smoke. His temperature is 99.8°F (37.7°C), pulse is 65/min, blood pressure is 126/86 mmHg, and respirations are 20/min. His oxygen saturation is 97%. On exam, he is pale and diaphoretic. His pupils are contracted. Diffuse wheezes are noted in all lung fields. What is the best treatment for his condition?

Q2. A 16-day-old girl was brought to the emergency department with lethargy. Physical exam showed tachypnea and marked hepatomegaly, as well as small hemangiomas on the skin. TSH was elevated. MRI showed numerous hepatic lesions and cardiomegaly. What is the most likely diagnosis?

Figure 1: Representative examples of knowledge-heavy and reasoning-heavy questions, sampled from benchmark datasets. Knowledge-heavy questions tend to be shorter and require direct factual recall, whereas reasoning-heavy questions are typically longer, involving multi-step inference.

Thapa, Zou et al., ArXiv, Jun. 2025

- Suggests a reframing of "medical reasoning progress": improvements come not only from bigger base models, but from training/eval designs that explicitly reward self-correction, backtracking, and reasoning under uncertainty (rather than fluent post-hoc narratives).

ARISE-AI.ORG

Fine Tuning of Frontier Models May Not Improve Domain Reasoning

Updating model knowledge in line with the pace of evolving medical knowledge is essential for trust. This study fine-tuned six frontier LLMs (OpenAI, Gemini, Llama) on knowledge not included in the training corpus such as newly approved drugs, synthetic EHRs, and updated guidelines to test whether commercial fine-tuning services actually "teach" models new medical knowledge. Models showed poor performance when questions were reframed as vignettes.

- Built datasets on new FDA approvals, synthetic patient records, and updated guidelines, then fine-tuned six commercial LLMs and tested both recall (memorization) and vignette-style reasoning (generalization).
- New drug facts generalized 30%, synthetic EHRs only 12%, and updated guideline knowledge 20%, showing limited transfer beyond memorized prompts which for some tasks reached > 90% accuracy.
- GPT-4o mini performed best, hitting about 51% on drug vignettes and 33% on the EHR task, suggesting smaller models may be easier to fine-tune.

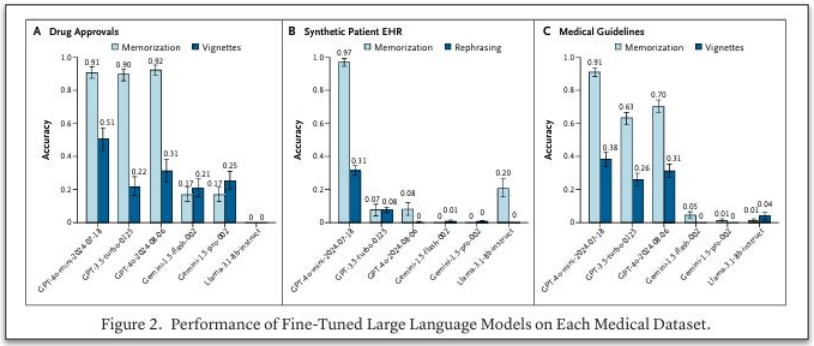


Figure 2. Performance of Fine-Tuned Large Language Models on Each Medical Dataset.

Wu, Zou et al., NEJM AI, Jul. 2025

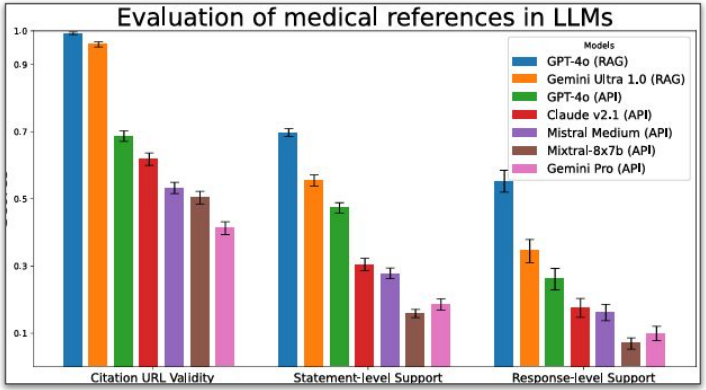
- Fine-tuning can help models memorize targeted updates, but it doesn't reliably replace RAG for reasoning on new knowledge.

ARISE-AI.ORG

Even When Asked For Sources, Models Are Often Inaccurate

Researchers introduce SourceCheckup, an automated framework that audits whether LLM-provided citations actually support the medical statements they accompany. Across 800 questions and ~58,000 statement–source pairs, the authors show that most LLM responses, even with web search (e.g., RAG), are partly unsupported by their own references.

- Built an agent pipeline that generates questions, parses model responses into statements, pulls URLs, and automatically checks whether each statement is supported by each source, validated against U.S. physicians (89% agreement).
- Seven LLMs (GPT-4o, Claude, Gemini, Mistral and others), with and without retrieval-augmented search.
- Between 50–90% of responses were not fully supported; even GPT-4o with RAG had only 55% response-level support, and 30% of its statements lacked backing evidence.
- Retrieval helps but isn't enough, LLMs frequently cite legitimate websites that still don't support the claims made, underscoring the need for explicit source-verification training and regulation.



Wu, Zou et al., Nature Communications, Apr. 2025

ARISE-AI.ORG

Takeaways

- Converting longitudinal care into tokenized timelines is enabling scalable “medical event foundation models” with potential for informing medical decisions such as screening.
- Incorporation of expansive multimodal information (including, but not limited to imaging) provides statistical and clinical grounding to the “medical knowledge” of LLMs.
- Multi-agent approaches hold promise to improve diagnostic and management performance, but introduce their own “computer-computer interaction” challenges.
- Clinical reasoning is not necessarily equivalent to mathematical and computational reasoning, and work must be done to optimize it in its own right.
- There is a need for claim-level grounding and verification. Measuring support, not fluency will enable increased user trust.

AI in Clinical Workflows

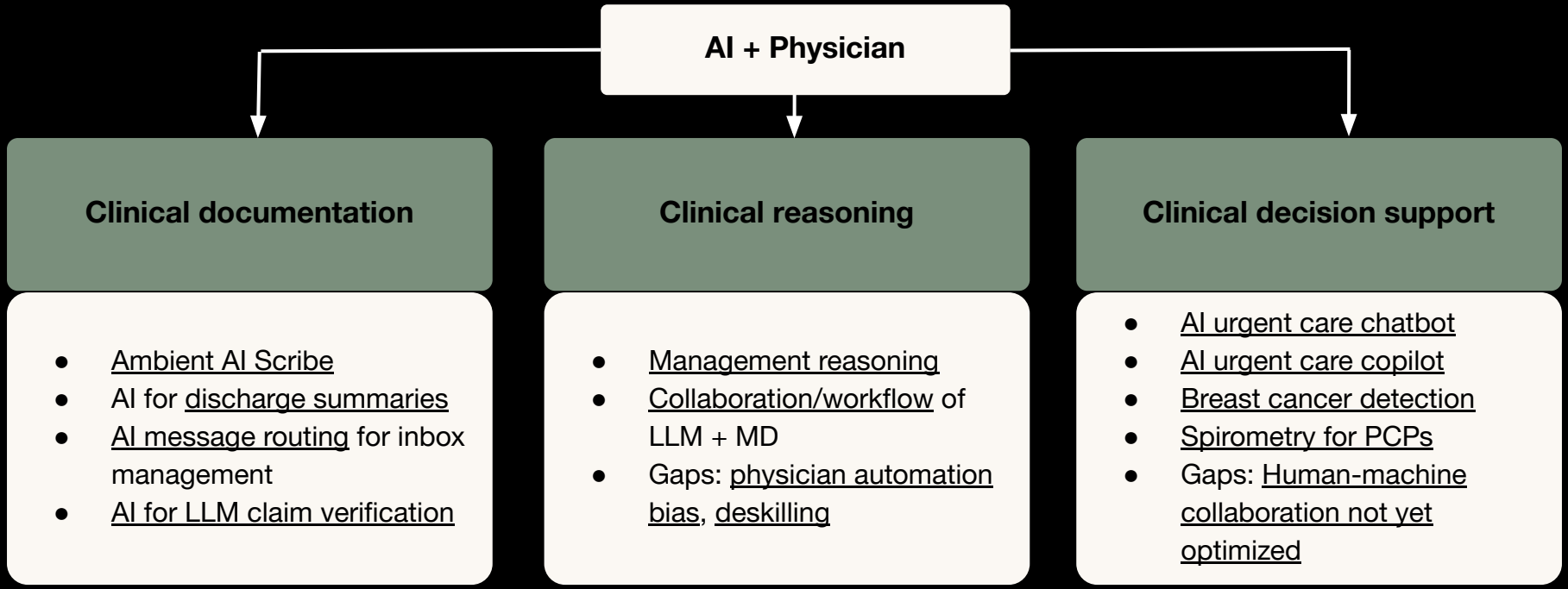
AI in Clinical Workflows

In 2025, evidence across diverse clinical environments shows AI most consistently adds value when it augments clinicians, improving accuracy and workflow without trying to replace human judgment.

- **Slides 62–66:** Human+AI augmentation has been tested across subspecialty cases in vitro and in vivo, supporting clinical reasoning. However, teaming may not surpass model alone performance.
- **Slides 67–68:** AI can supplement clinician performance on specific diagnostic tasks (e.g., mammography without increased false alarms, spirometry interpretation).
- **Slide 69-72:** Current AI–human collaboration remains suboptimal. Training clinicians to use AI effectively and understand human/model failure models may improve the human–computer interaction.
- **Slide 73-77:** Operationally, AI scribes subjectively improve physician workflow, though time savings are modest. Additional workflow wins have included inbox routing and discharge summaries with fact verification.

Overall, the near-term upside in clinical AI is highest in well-designed collaboration that reliably elevates clinician performance and workflow efficiency.

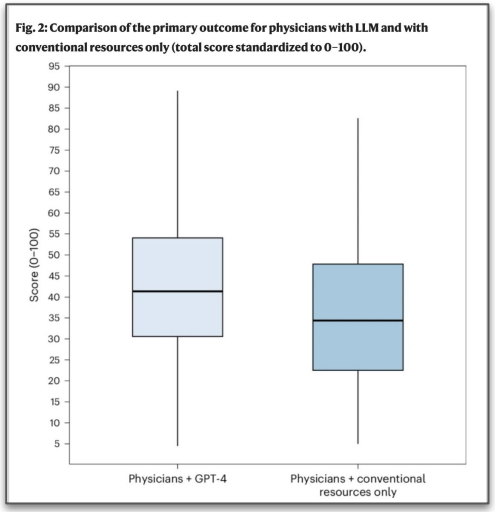
AI in Clinical Workflows



Does GPT-4 Help Physicians with Management Decisions?

Researchers conducted a randomized controlled trial that tested GPT-4's ability to aid physicians with management decisions. Access to GPT-4 significantly improved performance on complex management-reasoning tasks compared with conventional clinical resources but did not outperform the model alone.

- The study used five real, de-identified patient vignettes to simulate authentic clinical decision-making. Graded on a pre-specified expert developed management reasoning rubrics
- In the trial (n=92), physicians randomized to using GPT-4 + conventional resources scored higher (7%) on management-based reasoning tasks than those without GPT-4.
- Physicians assisted by GPT-4 performed comparably to the model alone (43% vs 44%) suggesting suboptimal teaming.
- GPT-4 users spent roughly two minutes longer per case but demonstrated strong context-aware decision-making, with no increase in potential harm compared with physicians using conventional resources.



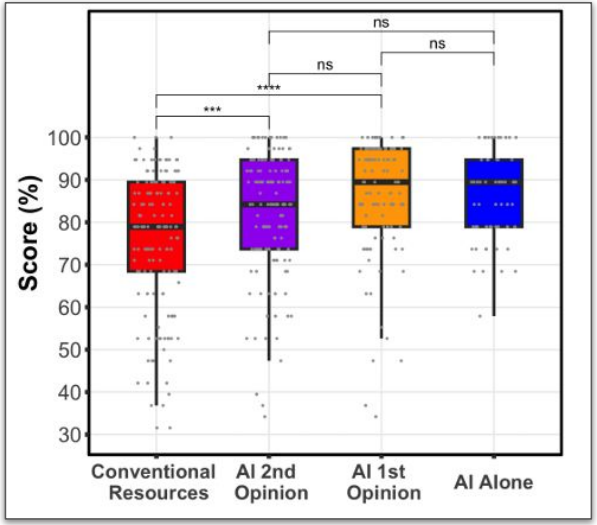
Goh, Chen, Rodman et al., Nature Medicine, Feb. 2025

ARISE-AI.ORG

Collaborative AI Improves Clinician Diagnostic Accuracy by 10%

After a previous study showed physicians using an off the shelf LLM performed worse than the LLM alone on diagnostic accuracy[1], this randomized controlled trial of 70 physicians showed that a custom collaborative GPT-4 system significantly improved diagnostic accuracy compared with conventional clinical resources. Whether used as a first or second opinion, the AI elevated clinician performance and reduced low-scoring cases.

- Physicians (n=70) completed up to six diagnostic cases under different conditions, using the AI system as a first opinion, a second opinion, or not at all just with access to conventional resources.
- In both AI arms, the model generates its differential and next steps, then produces a joint synthesis view integrating both perspectives (automatically incorporated in the second opinion arm and optionally available in the first opinion arm).
- AI collaboration raised diagnostic accuracy from 75% to 82 (AI second opinion) – 85% (AI first opinion). AI also lifted the “floor” by reducing the number of low-performing cases. Did not outperform the model alone.
- The study highlights human computer interaction design, not model-alone accuracy, as the next frontier for safe, effective human-AI diagnostic teamwork.



Everett, Horvitz et al., medRxiv, June 2025

ARISE-AI.ORG

AI Treatment Plans Outperform Doctors in Urgent Care

In a virtual urgent care setting, an AI chatbot more frequently generated optimal diagnostic and management decisions without increased harm as determined by physician adjudicators for patients with common urgent care complaints when compared to the treating physician who also had access to the AI output.

- 461 urgent care presentations, patients interacted with an AI chatbot that also had access to the patient's medical record and then had a telehealth visit.
- AI chatbot provided diagnostic and management recommendations that was available to the treating physician though review was not mandated.
- AI and physicians agreed in 57% of cases. AI recommendations were more frequently rated as optimal (77% vs. 67%), and potentially harmful in only 2.8% vs 4.6% for physicians. When AI and physicians differed, AI was more often rated as higher quality.
- AI was particularly strong at adhering to guidelines and appropriate workup whereas physicians were stronger at adapting to inconsistent information and physical exam findings.

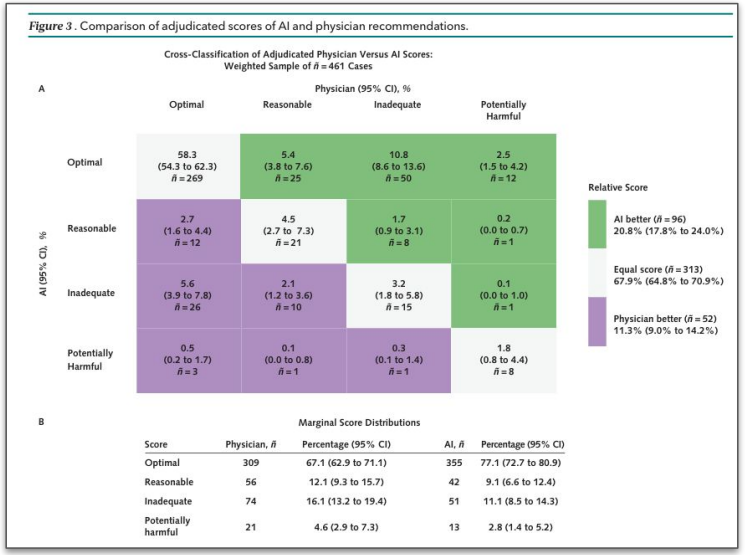


Figure 3. Comparison of adjudicated scores of AI and physician recommendations.

Cross-Classification of Adjudicated Physician Versus AI Scores: Weighted Sample of $\bar{n}$ = 461 Cases

A

AI (95% CI), % \ Physician (95% CI), %	Optimal	Reasonable	Inadequate	Potentially Harmful
Optimal	58.3 (54.3 to 62.3) $\bar{n}$ = 269	5.4 (3.8 to 7.6) $\bar{n}$ = 25	10.8 (8.6 to 13.6) $\bar{n}$ = 50	2.5 (1.5 to 4.2) $\bar{n}$ = 12
Reasonable	2.7 (1.6 to 4.4) $\bar{n}$ = 12	4.5 (2.7 to 7.3) $\bar{n}$ = 21	1.7 (0.9 to 3.1) $\bar{n}$ = 8	0.2 (0.0 to 0.7) $\bar{n}$ = 1
Inadequate	5.6 (3.9 to 7.8) $\bar{n}$ = 26	2.1 (1.2 to 3.6) $\bar{n}$ = 10	3.2 (1.8 to 5.8) $\bar{n}$ = 15	0.1 (0.0 to 1.0) $\bar{n}$ = 1
Potentially Harmful	0.5 (0.2 to 1.7) $\bar{n}$ = 3	0.1 (0.0 to 0.8) $\bar{n}$ = 1	0.3 (0.1 to 1.4) $\bar{n}$ = 1	1.8 (0.8 to 4.4) $\bar{n}$ = 8

Relative Score

- AI better ($\bar{n}$ = 96) 20.8% (17.8% to 24.0%)
- Equal score ($\bar{n}$ = 313) 67.9% (64.8% to 70.9%)
- Physician better ($\bar{n}$ = 52) 11.3% (9.0% to 14.2%)

B

Marginal Score Distributions

Score	Physician, $\bar{n}$	Percentage (95% CI)	AI, $\bar{n}$	Percentage (95% CI)
Optimal	309	67.1 (62.9 to 71.1)	355	77.1 (72.7 to 80.9)
Reasonable	56	12.1 (9.3 to 15.7)	42	9.1 (6.6 to 12.4)
Inadequate	74	16.1 (13.2 to 19.4)	51	11.1 (8.5 to 14.3)
Potentially harmful	21	4.6 (2.9 to 7.3)	13	2.8 (1.4 to 5.2)

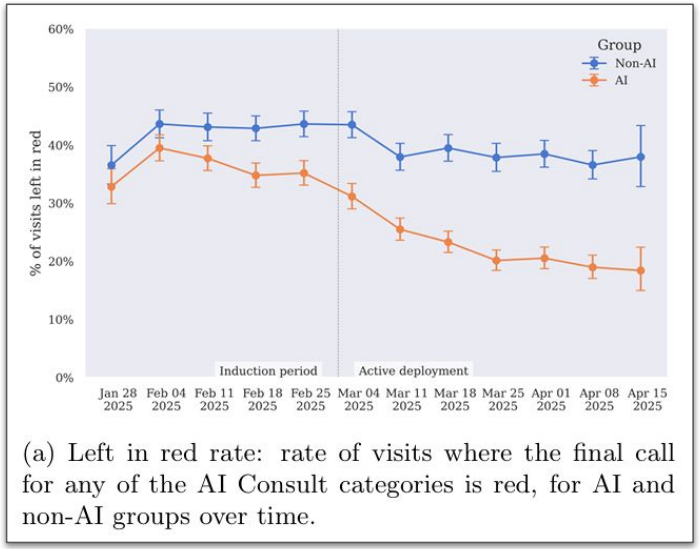
Zeltzer, Pevnick et al., Annals of Internal Medicine, Apr. 2025

ARISE-AI.ORG

AI Consult: The First Prospective Evaluation of an AI Copilot in Urgent Care (Part 1)

A Kenya based primary/urgent care provider Penda Health teamed up with OpenAI to develop "AI Consult." This GPT4o-based safety net copilot was launched prospectively and available at key points in the encounter to assist with documentation, investigations, diagnosis, and treatment.

- In a total of 39,849 visits, 20,589 performed with AI Consult versus 18,990 without AI consult across 15 clinic sites in Kenya.
- After multiple rounds of prompt engineering, AI consult ran silently in the background for both groups to provide three levels of safety net response via a traffic light: green, yellow, and red.
- After beta testing, Penda Health engaged in an active deployment strategy to increase the use of AI Consult as demonstrated by a temporal decrease in rates of "left in red."
- The primary outcome was the relative risk reduction in clinical errors adjudicated by a physician rater panel based on a likert scale in four categories: documentation, investigations, diagnosis, and treatment.



(a) Left in red rate: rate of visits where the final call for any of the AI Consult categories is red, for AI and non-AI groups over time.

Korom, Singhal et al., ArXiv, Jul. 2025

ARISE-AI.ORG

AI Consult: Led to a Reduction of Clinical Errors in Urgent Care (Part 2)

The use of AI Consult resulted in a 16% relative reduction in diagnostic errors and 13% relative reduction in treatment errors. Clinicians in the AI group also learned to avoid common errors over time while also citing that it improved quality of care. Patient-reported outcomes were equivocal between groups.

- The relative risk reduction for AI compared to non-AI was 32% for history-taking, 10% for investigations, 16% for diagnostic errors, and 13% for treatment errors.
- With a number needed to treat of 18.1 for diagnostic errors and 13.9 for treatment errors, if deployed across Penda Health's 400,000 annual visits, this would result to about 22,102 fewer diagnostic errors annually and 28,880 fewer treatment errors annually.
- The rate of the encounter starting in "red" decreased in the AI group from 45% to 35% by the end of the study suggesting clinicians learned to avoid common mistakes up front.
- This marks a real-world validation of AI augmenting clinical reasoning.

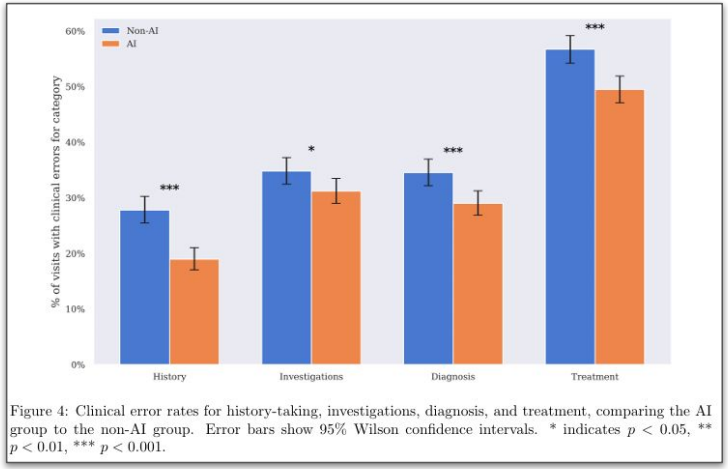


Figure 4: Clinical error rates for history-taking, investigations, diagnosis, and treatment, comparing the AI group to the non-AI group. Error bars show 95% Wilson confidence intervals. * indicates $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Korom, Singhal et al., ArXiv, Jul. 2025

ARISE-AI.ORG

AI Improves Breast Cancer Detection Without More False Alarms

When radiologists are given the option to consult an AI system for mammography reading, those who used AI support achieved a higher breast cancer detection rate (BCDR) (6.7 vs. 5.7 per 1,000) without a significant increase in recall rate. This strongly indicates that AI can improve mammography screening metrics.

- Prospective, real-world implementation study in Germany - 12 sites, 119 radiologists, 463,094 women screened (260,739 with AI support). Compared AI-supported double reading to standard double reading without AI.
- Radiologists given option to consult a deep-learning imaging algorithm (Vara MG). Provided support in two ways: 1. Normal triage = AI creates a list of unsuspicious exams, 2. Safety net = for suspicious studies, radiologist alerted by AI if they read the mammography as normal when AI deemed the mammography suspicious.
- This “safety-net” was triggered 3,969 times and accepted 1,077 times, resulting in 204 breast cancer diagnoses that would have been missed otherwise. Many were DCIS.

Table 3 | Model-predicted BCDRs, recall rates, biopsy rates and consensus rates and corresponding differences in the AI and control groups

Variables	Model-based prediction		Model-based difference (95% CI)	
	AI group	Control group	Absolute difference	Percentage difference
BCDR (per 1,000 women screened)	6.7	5.7	1.0 (0.3, 1.7)	17.6% (5.7%, 30.8%)
By invasiveness				
Invasive	5.2	4.8	0.4 (−0.2, 1.0)	7.8% (−4.3%, 21.3%)
DCIS	1.4	0.8	0.6 (0.3, 0.8)	67.6% (29.6%, 116.8%)
Other	0.1	0.04	0.1 (−0.0005, 0.1)	189.6% (−6.6%, 797.7%)
Consensus rate (per 1,000 women screened)	112.7	111.1	1.6 (−1.0, 4.2)	1.4% (−0.9%, 3.9%)
Recall rate (per 1,000 women screened)	37.4	38.3	−1.0 (−2.6, 0.6)	−2.5% (−6.5%, 1.7%)
PPV of recall	17.9%	14.9%	3.0 (1.5, 4.6) percentage points	20.5% (6.2%, 32.9%)
Biopsy rate (per 1,000 women screened)	10.4	9.6	0.8 (−0.0, 1.6)	8.2% (−0.4%, 17.6%)
PPV of biopsy	64.5%	59.2%	5.3 (1.3, 9.4) percentage points	9.0% (2.0%, 16.4%)

Eisemann, Katalinic et al., Nature Medicine, Jan. 2025

- Breast cancer detection rate increased 17.6% (CI 5.7%-30.8%) with AI (6.7 vs 5.7 per 1,000) while recall rate was slightly lower and noninferior (37.4 vs 38.3 per 1,000) with stronger PPV for recall and biopsy despite a higher biopsy rate.

ARISE-AI.ORG

AI Assistance Improves Spirometry Interpretation in Primary Care

Chronic respiratory disease are common and associated spirometry interpretation among primary care physicians is variable. Clinicians working in primary care (general practitioners or nurse practitioners) demonstrated superior interpretation of spirometry and quality assessment with AI assistance, narrowing the gap between primary care and expert interpretation.

- 133 participants randomized to interpret spirometry with (n = 67) or without AI assistance, where a CNN (ArtiQ.Spiro) suggests a differential with probabilities (n = 66).
- Out of 50 cases, each with outcomes represented by percentage of cases, participants with AI exhibited superior performance in identifying the top diagnosis (mean difference, 9%), establishing a diagnosis of COPD (mean difference, 16%), or including the diagnosis in a differential (mean difference, 7%) as established by expert pulmonologists.
- AI assistance also improved grading of technical quality of FEV1 and FVC but did not improve participant self-rated confidence in identifying the spirometry pattern speculating possible lack of trust given the "black-box" nature of the AI output.

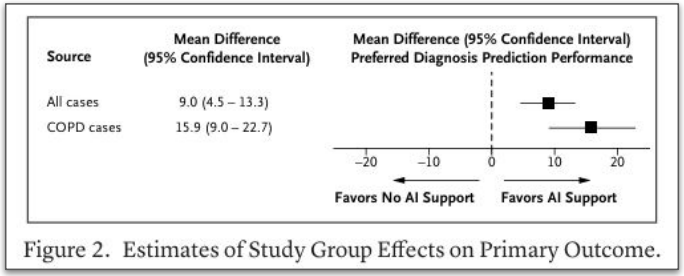


Figure 2. Estimates of Study Group Effects on Primary Outcome.

Doe, Man et al., NEJM AI, Jul. 2025

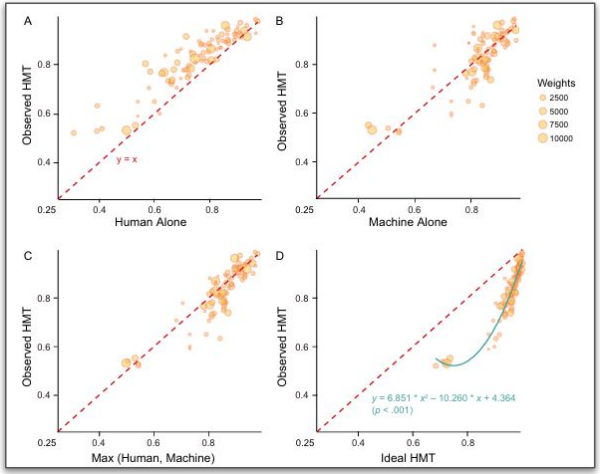
- Demonstrates AI's potential to reduce misdiagnosis and underdiagnosis of chronic respiratory diseases in primary care though practitioners desired more explainability.

ARISE-AI.ORG

However, the Human Machine Collaboration is Not Yet Optimized

Across 52 studies, a reliability analysis (number of success/number of patient tasks, akin to accuracy) showed human–AI medical teaming (HMT) on average demonstrated higher diagnostic reliability than clinicians working alone. However, full complementarity (1 + 1= 2) was rarely achieved.

- Conducted a meta-analysis of 52 studies with 87 teaming/level of expertise conditions modeling diagnostic reliability using hierarchical mixed-effects regressions that compared human-only, AI-only, and human+AI teaming workflows.
- HMT reliability > human-only reliability on average, though rarely achieves full complementarity and often does not beat the best of the two. A minority of conflict scenarios exist where AI assistance reduces human performance.
- Simultaneous teaming outperformed sequential human + AI teaming. Junior clinicians experienced significantly larger reliability gains with AI support compared with seniors
- Greatest gains occur when clinicians and models make complementary errors, highlighting that workflow design and limiting overlapping biases, not just model accuracy, may yield real clinical benefit.



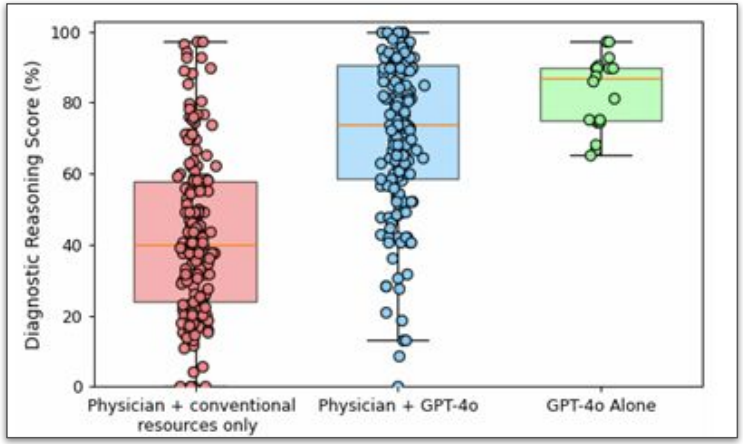
Liu, Chen et al., NPJ AI, Dec. 2025

ARISE-AI.ORG

AI Training + LLM Support Improves Clinician Diagnostic Accuracy by 27%

Prior work has those with limited AI training may not benefit from AI support.[1] In a randomized controlled trial of 58 physicians in Pakistan, those who completed a structured 20-hour AI-literacy program and then used GPT-4o achieved diagnostic reasoning scores 27% higher than peers using conventional resources alone.

- Enrolled physicians completed a 20 hour AI-literacy training covering capabilities, limitations, and appropriate use. Then randomized to LLM (GPT-4o) plus conventional resources vs conventional resources alone.
- Human diagnostic reasoning improved from 43% → 71% with AI support with larger gains seen among younger physicians and those who reported infrequent use of AI.
- GPT-4o alone outperformed all groups (83%) though in 38% of cases physicians plus LLMs surpassed GPT-4o alone performance = some evidence of strong complementarity.
- Trial underscores the utility of clinician AI training to augment human + AI synergy.



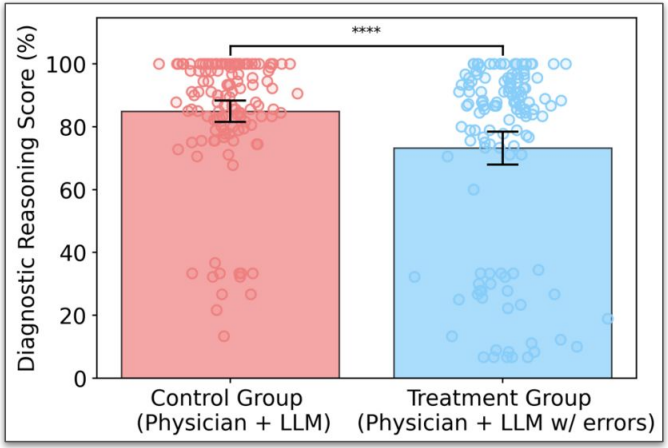
Qazi, Alizai et al., MedRxiv, Jul. 2025

ARISE-AI.ORG

AI-Trained Physicians Exhibit Automation Bias

Building on the previous study, a randomized controlled trial showed that physicians exhibit substantial automation bias with over-reliance on AI outputs when exposed to erroneous LLM recommendations. This significantly degraded diagnostic accuracy compared to error-free advice. This occurred despite voluntary consultation and prior AI-competency training.

- Trial of AI trained physicians (n=44), each reviewed up to 6 clinical vignettes; 264 cases. Half saw ChatGPT-4o suggestions with deliberate errors in 3 of the 6 vignettes, designed to be detectable but not obvious.
- Physicians exposed to flawed LLM recommendations achieved 73% accuracy (pre-specified rubric) versus 85% for those receiving error-free advice. Top-choice diagnosis accuracy was 91% in control vs 76% in treatment, adjusted difference.
- Consultation rates of GPT-4o were similar (69% vs. 67%).
- Highlights that automation bias poses significant patient safety risks that require robust validation frameworks and interface safeguards that actively prevents routine deference to LLM recommendations.



Qazi, Alizai et al., MedRxiv, Sept. 2025

ARISE-AI.ORG

Are There Concerns for Deskilling?

In this multicenter observational study, experienced endoscopists who regularly used AI for polyp detection showed a drop in performance when later performing colonoscopies without AI. Adenoma detection rate (ADR) fell from 28.4% to 22.4%, suggesting that continuous AI exposure may erode independent clinical vigilance.

- Retrospective, observational study of 1,443 standard (non-AI) colonoscopies across four centers during the Artificial Intelligence in Colonoscopy for Cancer Prevention trial, comparing performance 3 months before vs 3 months after AI implementation.
- ADR declined by 6% after clinicians had been exposed to AI. After adjustment, AI exposure remained associated with worse detection (OR 0.69), even accounting for sex, age, center, and other factors.
- Findings support a possible automation-overreliance / deskilling effect, highlighting the need for safeguards and training design.

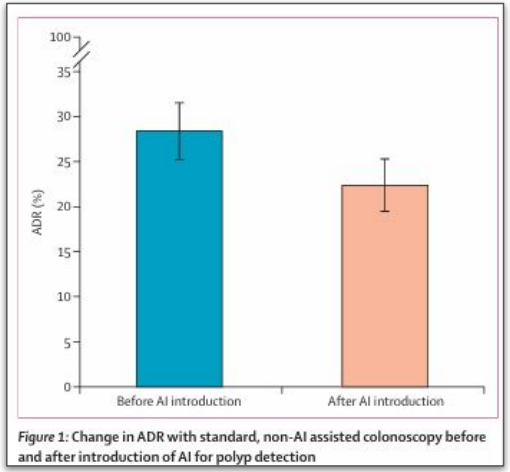


Figure 1: Change in ADR with standard, non-AI assisted colonoscopy before and after introduction of AI for polyp detection

Budzyn, Mori et al.,
The Lancet Gastroenterology & Hepatology, Oct. 2025

ARISE-AI.ORG

Ambient AI Scribe Subjectively Reduces Administration Burden

In a multicenter AI scribe pre- and post-intervention analysis of 272 physicians, burnout decreased by over 20%. Physicians also reported a reduction in note related cognitive tasks, improved attention to patients, and a reduction in time spent documenting after hours.

- After using an ambient scribe for 30 days, the primary outcome was self-reported burnout with secondary aims of self-reported cognitive task load, attention on patients, patient comprehension of notes, ability to add patients to the schedule and time spent documenting after hours.
- The percentage of physicians with burnout significantly decreased from 52% to 39% (74% lower odds) with significant improvements to all of the aforementioned secondary aims.
- Ambient scribes may subjectively reduce admin burden allowing for more meaningful work and well-being.

Table 3. Comparison of Secondary Outcome Measures Before and After Use of the Ambient AI Scribe

Outcome	No. of participants	Mean (SE) score[a]			*P* value
		Baseline	Follow-up	Difference	
Burnout	186	4.59 (0.15)	4.12 (0.15)	0.47 (0.12)	<.001
Note-related cognitive task load					
Any	243	7.10 (0.09)	4.46 (0.12)	2.64 (0.13)	<.001
Temporal demand	249	7.01 (0.11)	4.35 (0.13)	2.66 (0.16)	<.001
Effort	248	7.31 (0.12)	4.71 (0.13)	2.60 (0.15)	<.001
Mental demand	254	6.84 (0.12)	4.38 (0.15)	2.46 (0.15)	<.001
Documentation after hours	263	4.95 (0.18)	4.05 (0.16)	0.90 (0.19)	<.001
Focused attention on patients	253	6.51 (0.16)	8.56 (0.11)	−2.05 (0.18)	<.001
Comprehensible care plans	254	7.34 (0.13)	7.79 (0.13)	−0.44 (0.17)	.005
Agreeable to add urgent patients	230	6.21 (0.21)	6.72 (0.20)	−0.51 (0.24)	.02
No. of additional patients (1 to ≥4)	91	2.19 (0.11)	2.16 (0.11)	0.02 (0.11)	.58

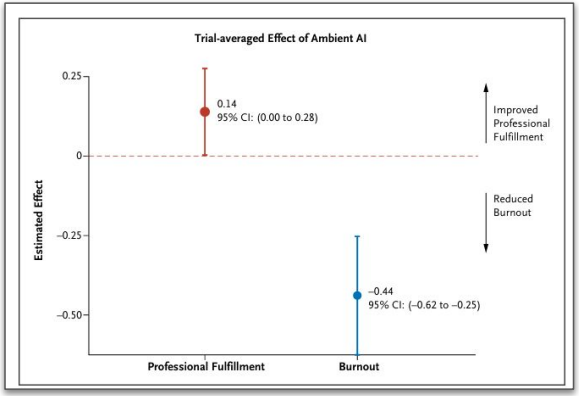
Olson, Troup et al., JAMA Network Open, Oct. 2025

ARISE-AI.ORG

However, Objective Improvements with AI Scribes are Still Marginal

In the first two randomized controlled trials of ambient AI scribes, there was a modest reduction in documentation time per day on par with human scribes. However, a subjective reduction in burnout was still seen across both studies.

- Between the two studies,[1,2] three ambient AI scribes were studied (Abridge, Microsoft DAX Copilot, and Nabla) with time savings evaluated via Epic Signal data. Both studies suggest subjective improvements in exhaustion/burnout.
- Time savings are currently underwhelming (~20 seconds per note)[2]. Further productivity gains may rely on the scope of AI scribes expanding to downstream tasks such as communications, orders, etc.



Afshar, Gordon et al., NEJM AI, Nov. 2025

Table 2. Documentation Efficiency Metrics and Psychometric Outcomes from Survey Assessments.*

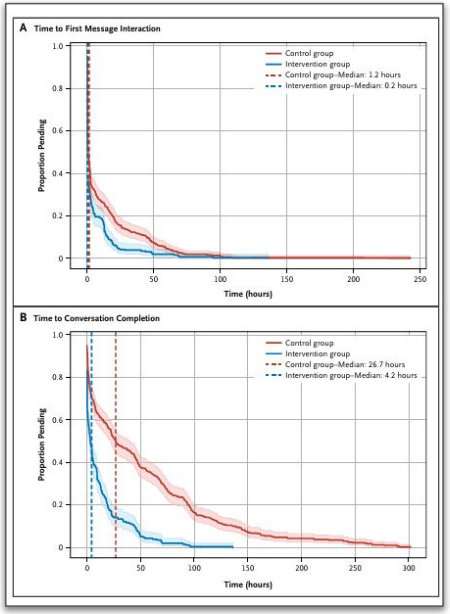
Measure	DAX vs. Control		Nabla vs. Control	
Efficiency Metrics (%)	**Difference**	**95% CI**	**Difference**	**95% CI**
Time in notes per note†	−0.02	−0.09 to 0.06	−0.10	−0.17 to −0.02
Time on unscheduled days	0.14	−0.05 to 0.33	0.06	−0.13 to 0.26
Time outside scheduled hours	0.09	−0.14 to 0.32	0.08	−0.15 to 0.31
Mini-Z 2.0				
Supportive work environment	1.51	0.55 to 2.47	1.55	0.59 to 2.51
Work pace and EMR stress	1.41	0.38 to 2.43	1.20	0.18 to 2.23
Composite	2.83	1.28 to 4.37	2.69	1.14 to 4.23
PFI-WE	−0.32	−0.55 to −0.08	−0.23	−0.46 to 0.01
PTL	−39.89	−71.88 to −7.90	−31.69	−63.79 to 0.42
Binary Mini-Z 2.0	Odds Ratio		Odds Ratio	
Single-item burnout	0.78	0.40 to 1.54	0.67	0.34 to 1.34
Supportive work environment	1.87	0.83 to 4.23	1.84	0.82 to 4.14
Work pace and EMR stress	1.90	0.17 to 21.54	1.87	0.17 to 21.19
Composite — joyful Workplace	1.64	0.38 to 7.14	1.50	0.35 to 6.45

Lukac, Mafi et al., NEJM AI, Nov. 2025

AI Message Routing Reduces Inbox Burden and Speeds Patient Response

A real world evaluation of a BERT based NLP model used to optimize message routing resulted in a reduction of time to address patient portal messages, time to resolution of the conversation, and fewer message interactions. The NLP model correctly classified and routed patient portal messages in 98% of cases.

- Model was embedded directly into the EHR across 4 Emory outpatient clinics, routing 469 live patient message threads over 14 days with a comparator arm of 402 unrouted messages.
- Time to first staff interaction dropped from 1.2 hours → 0.2 hours (median), and full conversation resolution fell from 26.7 hours → 4.2 hours (median).
- Staff needed 2 fewer interactions per message (median), reducing cognitive load and back-and-forth communication.
- NLP models can improve workflow efficiencies with the potential to reduce burnout related to EHR.



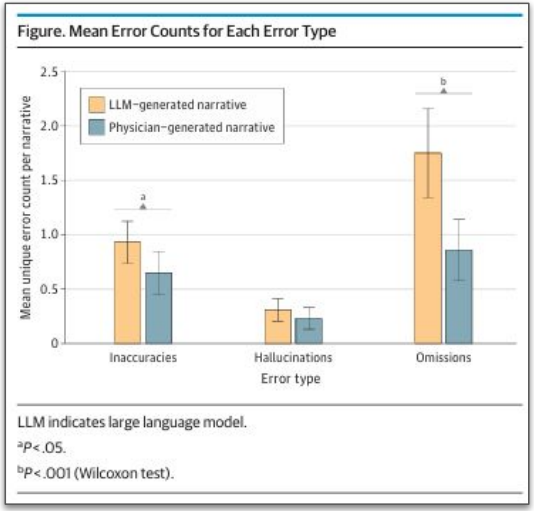
Anderson, Harzard et al., NEJM AI, Feb. 2025

ARISE-AI.ORG

LLM-Generated Discharge Summaries Match Physician Quality

Large language models produced hospital discharge summary narratives that were comparable in overall quality to those written by physicians. LLM (GPT4) narratives were more concise and coherent, though they contained more errors. Errors were typically of low-potential harm, suggesting they may be useful as drafts with human review.

- Study of 100 real inpatient encounters compared LLM- vs physician-generated narratives with blinded reviewers.
- Based on a 5 point Likert scale, reviewers rated overall quality (3.7 vs. 3.8) and did not exhibit a preference, with LLM outputs noted to be more concise (4.0 vs 3.7) and coherent (4.2 vs. 4) but less comprehensive (3.7 vs. 4.1).
- LLM-generated narratives contained more errors per discharge summary (2.9 vs 1.8), but overall harmfulness scores remained low (mean 0.84 vs 0.36 on a AHQR 0–7 scale).
- Findings support the use of LLMs to draft discharge summaries, contingent on clinician oversight before clinical use.



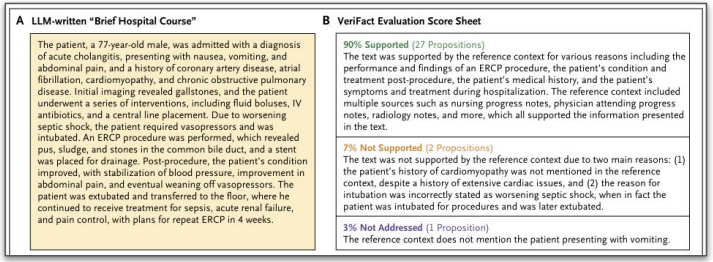
Williams, Rosner et al., JAMA Internal Medicine, May 2025

ARISE-AI.ORG

Automated EHR-Grounded Fact-Checking for Discharge Summary Hospital Courses

VeriFact is an automated pipeline that verifies individual statements in a clinical note by retrieving relevant evidence from the patient's EHR and using an LLM as a judge to label each claim as supported, not supported, or not addressed. In a clinician-labeled benchmark of 100 patients covering both human-written and LLM-generated Brief Hospital Course sections, the best configuration achieved 93% agreement with clinician ground truth.

- VeriFact-Brief Hospital Course (BHC) contains human BHC + an LLM-generated BHC per patient, decomposed into sentence-level and "atomic claim" propositions. Each proposition was labeled supported / not supported / not addressed by ≥3 clinicians, with tie-breaking adjudication to create a consensus ground truth.
- VeriFact retrieves the top N relevant EHR facts via hybrid dense+sparse search + reranking, then prompts an LLM judge to classify support and explain why.
- VeriFact's best setting used a reasoning LLM judge (DeepSeek-R1-Distill-Llama-70B) and achieved 93% agreement with clinician consensus, exceeding the clinicians' own inter-rater agreement on the same LLM-generated BHC atomic-claim task (89%).
- Turning text into atomic claims and using judge models improved verification quality, suggesting "fact-checking" akin to "spell checking" can be a practical safety layer for LLM-drafted clinical documentation.



A LLM-written "Brief Hospital Course"

The patient, a 77-year-old male, was admitted with a diagnosis of acute cholangitis, presenting with nausea, vomiting, and abdominal pain, and a history of coronary artery disease, atrial fibrillation, cardiomyopathy, and chronic obstructive pulmonary disease. Initial imaging revealed gallstones, and the patient underwent a series of interventions, including fluid boluses, IV antibiotics, and a central line placement. Due to worsening septic shock, the patient required vasopressors and was intubated. An ERCP procedure was performed, which revealed pus, sludge, and stones in the common bile duct, and a stent was placed for drainage. Post-procedure, the patient's condition improved, with stabilization of blood pressure, improvement in abdominal pain, and eventual weaning off vasopressors. The patient was extubated and transferred to the floor, where he continued to receive treatment for sepsis, acute renal failure, and pain control, with plans for repeat ERCP in 4 weeks.

B VeriFact Evaluation Score Sheet

90% Supported (27 Propositions)
The text was supported by the reference context for various reasons including the performance and findings of an ERCP procedure, the patient's condition and treatment post-procedure, the patient's medical history, and the patient's symptoms and treatment during hospitalization. The reference context included multiple sources such as nursing progress notes, physician attending progress notes, radiology notes, and more, which all supported the information presented in the text.

7% Not Supported (2 Propositions)
The text was not supported by the reference context due to two main reasons: (1) the patient's history of cardiomyopathy was not mentioned in the reference context, despite a history of extensive cardiac issues, and (2) the reason for intubation was incorrectly stated as worsening septic shock, when in fact the patient was intubated for procedures and was later extubated.

3% Not Addressed (1 Proposition)
The reference context does not mention the patient presenting with vomiting.

Chung, Aghaeepour et al., NEJM AI, Dec. 2025

ARISE-AI.ORG

Takeaways

- Humans + AI questions the fundamental theorem of informatics as teaming may not beat AI alone.
- Workflow design is as important as how good the models are. More work is needed to better align humans with models and vice versa.
- Failure mode training and awareness of pitfalls such as automation bias and deskilling are essential for trainees and physicians using AI.
- More prospective human + AI co-pilot studies are warranted to understand the impact on real patient outcomes (e.g., linking decisions made using AI to downstream outcomes).
- Ambient AI scribes show strong subjective improvements with minimal objective time savings. The addition of downstream tasks may yield more productivity/efficiency impact.

Patient Facing AI

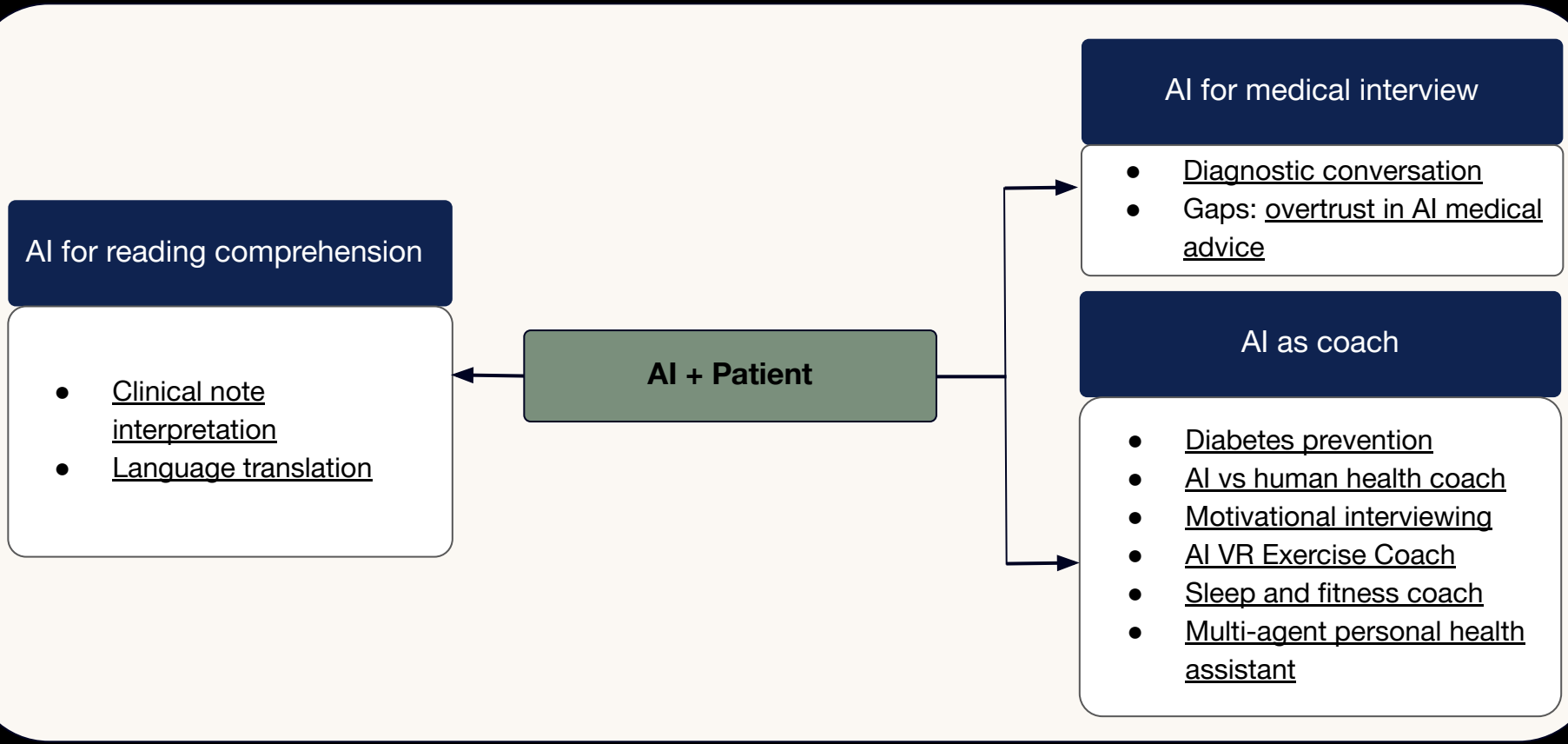
Patient Facing AI

In 2025, patient-facing AI moved from simple chat to scalable support across history taking, coaching, and wearable devices.

- **Slides 82-83:** In a simulated primary care, LLMs demonstrate strong multi-turn conversational performance. However, emerging safety concerns include overtrust and risks from unsupervised use.
- **Slides 84-87:** AI-led lifestyle/behavior coaching can achieve outcomes comparable to human coaching, often with higher engagement.
- **Slides 88–89:** Plain-language translation of clinician notes/instructions improves understanding, particularly for vulnerable populations.
- **Slide 90-91:** Future directions include pairing deep learning with wearable biometric data to deliver more personalized health assistance.

Patient-facing AI success should be judged by objective patient outcomes, with safeguards that prevent misinformation and overtrust from undermining the benefits.

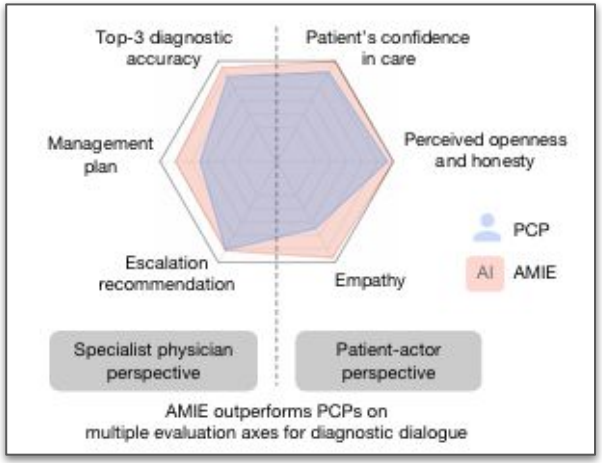
Patient Facing AI



Can AI Hold a Diagnostic Conversation as well as a Doctor?

Google's AMIE is an AI system trained to conduct clinical history-taking and diagnostic dialogue. In a randomized, double-blind study of 159 simulated patient cases, AMIE outperformed 20 physicians on nearly all measures of diagnostic accuracy, communication quality, and empathy.

- AMIE was developed using a multi-agent self-play simulated environment with automated feedback so the model can practice diagnostic conversations across many diseases, contexts, and specialties.
- AMIE achieved higher diagnostic accuracy than physicians. Specialists rated it superior on 30 of 32 evaluations axes and patient actors on 25 of 26 axes, marking a milestone for conversational diagnostic AI.
- The system demonstrated stronger reasoning when forming differential diagnoses, suggesting that conversational AI can complement clinicians in information synthesis and decision support.
- While AMIE's text-based chat interface differs from typical clinical practice this is a strong step in the direction towards conversational diagnostic AI.



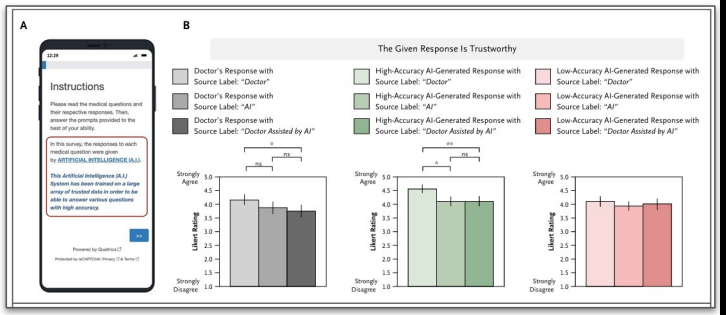
Tu, Natarajan et al., Nature, Apr. 2025

ARISE-AI.ORG

People Overtrust Low-Accuracy AI-Generated Medical Advice

Participants were unable to distinguish AI-generated (GPT-3) medical responses from those written by physicians and often rated AI responses, even low-accuracy ones, as equally or more valid, trustworthy, and complete. This overtrust led participants to indicate that they would follow potentially harmful AI advice and even seek unnecessary medical attention, revealing a significant safety risk.

- Participants (n=300) presented with 90 medical questions and answers with the source blinded. There were 30 answers each from (1) doctors, (2) high accuracy AI responses, and (3) low accuracy AI responses.
- Participants rated high accuracy AI responses as more valid, trustworthy, and complete/satisfactory compared to doctors. Low accuracy AI responses were rated equally to doctor responses.
- Participants showed equal propensity to seek additional information, follow the advice provided, and seek further medical attention even when the response was a low accuracy AI response.
- When assessing labels, optimal trustworthy ratings occurred when a response was labeled as a doctor but actually from high accuracy AI thus still showing a preference for a doctor labeled response.



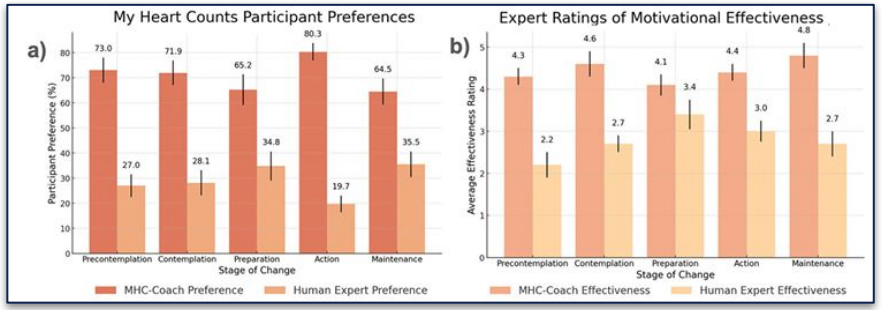
Shekar, Maes et al., NEJM AI, May 2025

ARISE-AI.ORG

Participants Prefer an AI Chatbot for Health Coaching Over Human Coaches

Based on LLaMA 3-70B, CV-Coach is fine-tuned on behavioral psychology (Transtheoretical Model) and cardiovascular health principles to deliver personalized, stage-specific activity guidance. AI-generated messages were preferred over human-expert messages and rated as more effective, highlighting the potential for scalable, proactive health coaching.

- Fine-tuned on the Transtheoretical Model (TTM) and cardiovascular-health literature to generate TTM stage-matched coaching messages.
- In a survey of 632 participants, 68% preferred AI messages when TTM stage-matched and 85% preferred AI messages in general comparisons.
- Behavioral-science experts scored AI messages higher for effectiveness (4.4 vs 2.8) and TTM alignment (4.1 vs 3.5).
- LLMs can operationalize a behavioral framework at scale, generating short, actionable, stage-tailored nudges that users and experts perceive as more effective than static human-written messages.



Mantena, Ashley et al., NPJ CV Health, Sept. 2025

ARISE-AI.ORG

Motivational Interviewing Meets Wearable-Driven AI Guidance

GPTCoach is a GPT-4 based health coaching agent that implements an evidence-based onboarding conversation using motivational interviewing while integrating three months of wearable data to personalize physical activity plans.

- While off the shelf models often jump straight to advice undermining autonomy, GPTCoach was designed to follow a nonprescriptive approach, tailor information in using diverse contexts, and adopt a nonjudgemental tone.
- In a lab study of 16 participants, GPTCoach delivered highly personalized, supportive guidance and demonstrated strong motivational interviewing consistency, outperforming off the shelf GPT-4.
- While use of sensor data was variable, GPTCoach used language that was neutral or consistent with motivational interviewing techniques 93% of the time.
- Based on likert scales, participants felt supported and comfortable. They suggested future iterations should include personalized reminders, guidance during life changes such as injuries, and multiple personas that adapt based on the user motivation/emotions.

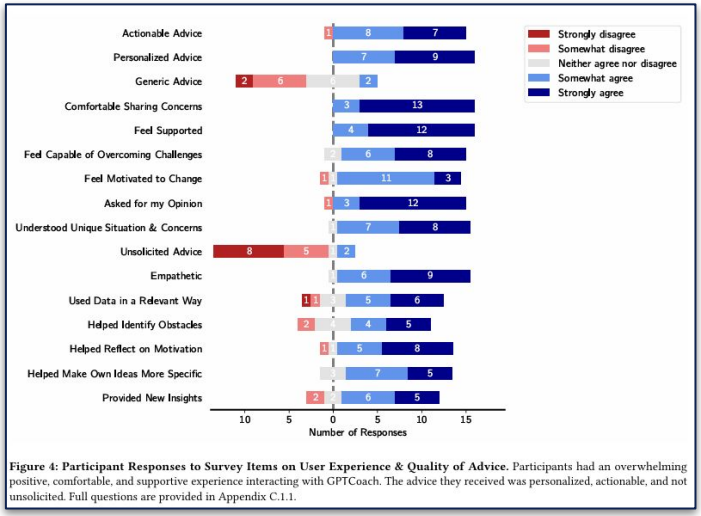


Figure 4: Participant Responses to Survey Items on User Experience & Quality of Advice. Participants had an overwhelming positive, comfortable, and supportive experience interacting with GPTCoach. The advice they received was personalized, actionable, and not unsolicited. Full questions are provided in Appendix C.1.1.

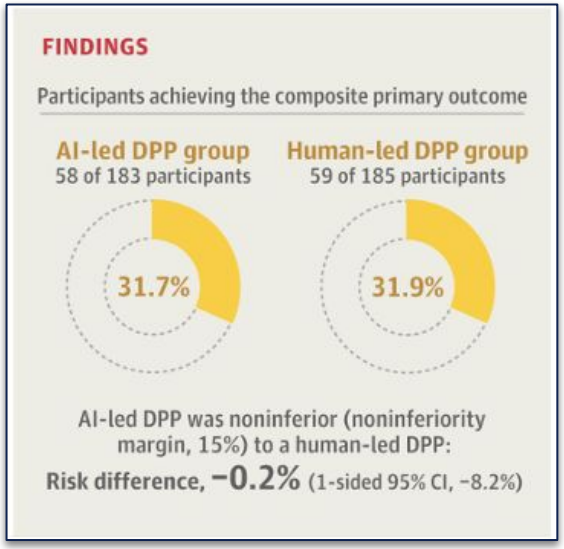
Jorke, Landay et al., ArXiv, Mar. 2025

ARISE-AI.ORG

Diabetes Prevention: AI-led Lifestyle Intervention Matches Human Coaching

An AI-powered Diabetes Prevention Program (DPP) was noninferior to a human-coach DPP for achieving clinically meaningful weight loss, HbA1c reduction, and physical activity targets among adults with prediabetes. The AI group showed higher engagement and completion, with no adverse events, demonstrating that AI-based behavioral interventions can safely and effectively scale chronic-disease prevention.

- <1% of eligible patients participate in a Diabetes Prevention Program (DPP). This was a noninferiority RCT including 368 adults with prediabetes randomized 1:1 to AI-led vs human-led DPP for 12 months.
- AI-led DPP: Bluetooth scale and mobile app delivering personalized push notifications for weight, physical activity, and nutrition using a reinforcement learning algorithm and multimodal data (e.g., weight trends, accelerometry, geolocation) with no human coaching or oversight.
- Primary composite outcome: maintaining HbA1c < 6.5 % and achieving ≥ 5 % weight loss OR ≥ 4 % weight loss + ≥ 150 min/wk activity OR HbA1c ↓ ≥ 0.2%. Primary outcome achieved: 31.7 % (AI) vs 31.9 % (human) → Noninferior.
- Demonstrates that fully autonomous lifestyle intervention can achieve outcomes comparable to human-led DPP. The scalable, asynchronous, on-demand format of the AI-led DPP may address barriers to participation in the human-led DPP.



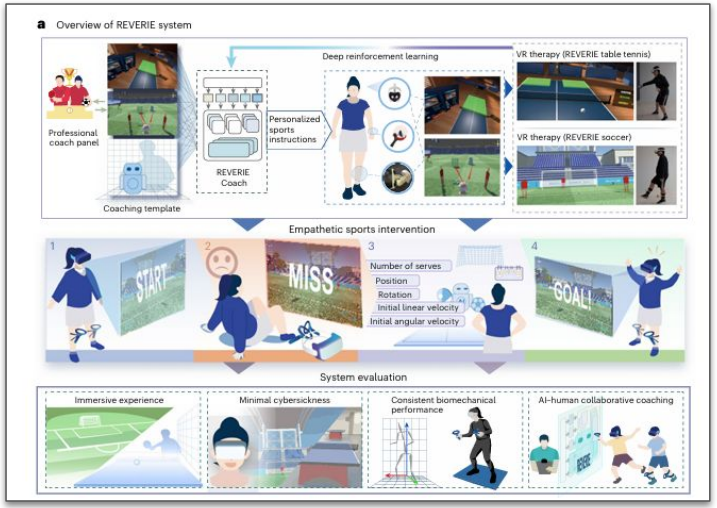
Mathioudakis, Maruthur et al., JAMA, Oct. 2025

ARISE-AI.ORG

An AI Virtual Reality System That Delivers Real Metabolic and Cognitive Gains

REVERIE is an AI-powered virtual reality sports system that uses deep reinforcement learning–trained coaching agents to deliver personalized, empathetic sports training for adolescents with obesity. In an 8-week randomized controlled trial of 227 participants, REVERIE achieved fat mass reduction and metabolic benefits equivalent to real-world sports, with additional cognitive and neuroplasticity gains.

- Transformer-based virtual coaches trained via two-stage deep reinforcement learning delivered empathetic, adaptive technique instruction with biomechanics.
- In an RCT of 227 participants were randomized 1:1:1:1:1 to physician table tennis/soccer, REVERIE table tennis/soccer, or control. Within 8 weeks, fat mass decreased by –4.3 kg vs control, statistically comparable to physical sports (–5 kg), confirming VR exercise can match real-world sports.
- REVERIE uniquely improved working memory and olfactory function compared to the physical sports group. Functional MRIs showed enhanced neural efficiency and neuroplasticity.
- AI-powered virtual reality sports therapy provides an empathetic approach to adolescent obesity showing improvements in physical, psychological, and cognitive health.



Wang, Li et al., Nature Medicine, Jun. 2025

ARISE-AI.ORG

AI Improves Patient Comprehension of Clinical Notes

Patients were given four discharge summary notes for common conditions (e.g., CAP, DKA, CHF, stroke). After evaluating two in standard form and two translated into plain language by AI, objective comprehension, subjective comprehension, and self-reported confidence scores increased significantly.

- GPT-4 prompted with “Can you provide a detailed summary to give to a patient with low health literacy?” and then provided with a discharge summary note (DSN).
- Objective comprehension (1.2 point increase out of 4), subjective comprehension (2.4 point increase out of 16), and self-reported confidence (2 point increase out of 8) significantly increased with GPT-4 translated DSNs. Gains were most notable in black and hispanic patients, older patients, and those who reported limited health knowledge.
- Time spent reading was also significantly decreased with GPT-4 translated DSNs.
- EHR implementation can act as a patient-friendly supplement for provider notes to bridge health literacy gaps.

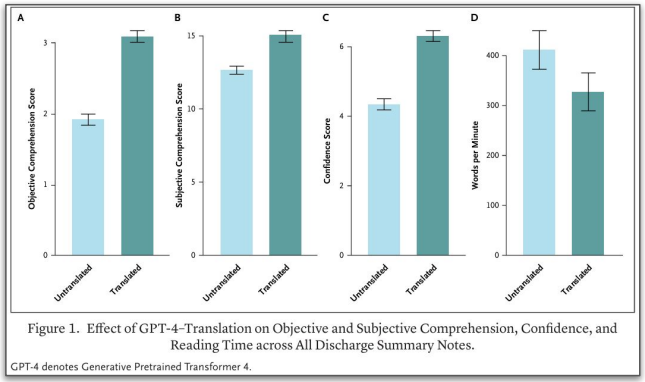


Figure 1. Effect of GPT-4–Translation on Objective and Subjective Comprehension, Confidence, and Reading Time across All Discharge Summary Notes.

GPT-4 denotes Generative Pretrained Transformer 4.

Kumar, Engelhard et al., NEJM AI, Jan. 2025

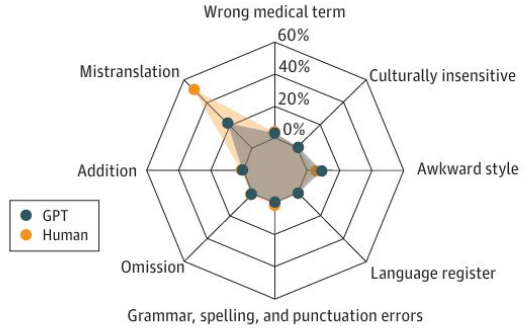
ARISE-AI.ORG

AI Achieves Human-Level Quality in Clinical Language Translation

GPT-4o translated pediatric patient instructions from English to Spanish with equivalent quality to professional human translators. The AI produced fewer mistranslations and was preferred by experts in 52% cases, suggesting that GPT-4o can safely assist in clinical translation while maintaining human-level accuracy.

- A set of 20 patient instructions based on real cases were translated by GPT-4o and professional human translators.
- GPT-4o performance was statistically equivalent to human translators on a validated rubric (mean difference = 1.6 points; within ±5 point margin pre specified for noninferiority) and produced fewer mistranslations (1.8 vs 4.1 per sample).
- Blinded evaluations by professional translators showed that AI-generated instructions were preferred 52% of the time, human translations were preferred 20% of the time, and 28% of evaluations were rated as neutral.
- Demonstrates safe, high-quality AI translation with the potential for clinical use extending to other materials such as portal messages, intake forms, and consent forms that are often only in English.

Figure 2. Radar Plot Showing the Relative Proportion of 8 Different Error Types for GPT-4o and Human Translations



Proportions were calculated within each group (GPT-4o and human) by dividing the mean error count for each error type by the total mean error count across all error types in that group. Mistranslation errors constituted a larger proportion of total errors in the human translations compared to the GPT-4o translations.

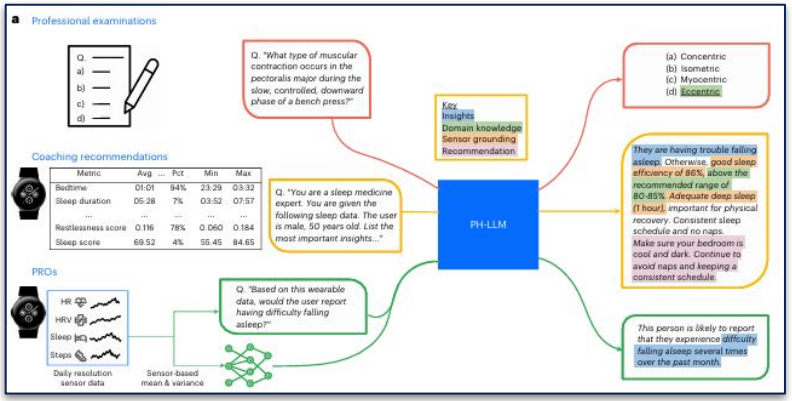
Ray, Hron et al., JAMA Pediatrics, Jul. 2025

ARISE-AI.ORG

A Personal Health LLM for Sleep and Fitness Coaching

Personalized Health LLM (PH-LLM) is a Gemini-based large language model fine-tuned to reason over wearable sensor data and generate personalized sleep and fitness insights, recommendations, and predictions. Across professional exams, expert-rated case studies, and real-world wearable datasets, PH-LLM performed on par with or better than human experts while outperforming base foundation models.

- Gemini Ultra 1.0 fine-tuned on the case-study prompt/response pairs. Integrated heart rate, HRV, sleep stages, activity, and training load via a learned multimodal adapter.
- Outperformed human experts on board-style multiple-choice exams in sleep medicine (79% vs 76%) and fitness (88% vs 71%)
- Generated high-quality, expert-rated insights and recommendations on sleep and fitness from 30-day aggregated wearable data, matching human experts across 857 real-world case studies.
- Accurately predicted self-reported sleep outcomes from passive sensor data, achieving AUROC comparable to specialized models and enabling closed-loop coaching.



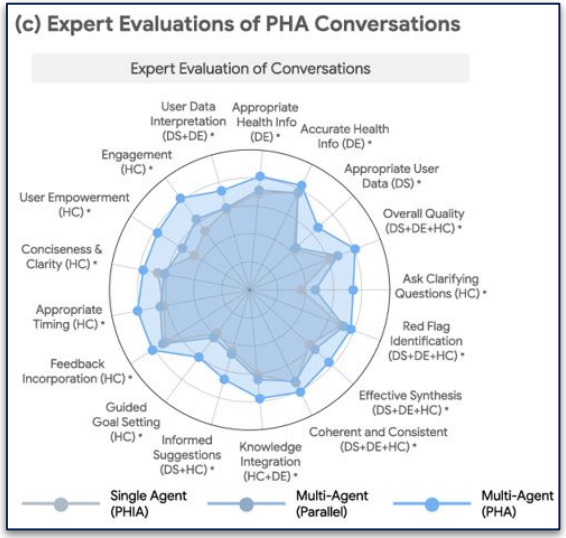
Khasentino, McLean et al., Nature Medicine, Jul. 2025

ARISE-AI.ORG

A Multi-Agent Personal Health Assistant

A multi-agent system based on Gemini 2 (Flash/Pro) (composed of a Data Scientist, Domain Expert, and Health Coach Agent) analyzes wearable data, interprets medical records, and delivers personalized health guidance. It demonstrated strong performance in data analysis, clinical reasoning, contextual personalization, and holistic health coaching.

- Built from analysis of 1,370 real health queries, surveys from 555 Fitbit users, and expert workshops that identified four critical user journeys: general health knowledge, personal data insights, wellness advice, and symptom queries.
- Three specialized agents: Data Science Agent analyzes wearable and population data; Domain Expert Agent provides medical reasoning and multimodal synthesis; Health Coach Agent uses motivational interviewing to guide behavior change. Responses combined by an orchestrator.
- Evaluated across 10 benchmark tasks with 7,000+ human annotations and 1,100 hours of expert and end-user effort.
- Orchestration matters - performance was validated using multimodal real-world data from the WEAR-ME study (Fitbit + labs + surveys), showing improved trustworthiness, personalization, multimodal reasoning, and coaching effectiveness compared with the single agent base LLM or parallel multi-agent system.



Ali Heydari, Xu et al., ArXiv, Sept. 2025

ARISE-AI.ORG

Takeaways

- Conversational AI is reaching new heights with reasoning models using gathered history to generate useful structured differentials and management plans.
- Patient facing AI holds promise for a new landscape of scalable patient engagement, especially for resource intensive services such as coaching.
- Safety guardrails are paramount, users overtrust poor advice and cannot be assumed to play any oversight role.
- Vendors have competing interests (e.g., user engagement, profitability). High yield patient facing AI should focus on objective clinical endpoints to avoid application bloat.
- Frontier models have high success with communication text based tasks and are ready to be deployed in this realm alongside human oversight.

Applied AI & Demos

Applied AI & Demos

In 2025, the most clinically translatable AI advances came from specialty-tuned models solving narrow, high-signal problems with clear endpoints.

- **Slides 96–104:** Imaging remains the dominant use case, with high-impact studies across pathology, neurology, nephrology, oncology, pulmonology, and cardiology.
- **Slides 105-110:** Medical specialties are repurposing abundant, noninvasive data to improve risk stratification.
- **Slides 111-114:** As emphasis shifts toward improving objective patient outcomes, critical care and surgical studies deploy AI to guide treatment decisions and resource allocation.
- **Slides 115-119:** Operational use cases include triaging patients/referrals, clinical trial eligibility, and adjudicating adverse events in clinical trials.
- **Slides 120-124:** Demos provide insight into the next wave of innovation such as EHR chatbots, evaluating model safety, and mental health chatbots.

As more specialties adopt domain-tuned systems, this track is becoming the fastest route from compelling model performance to measurable improvements in real-world care.

Applied AI & Demos

Imaging

- GI: celiac disease (pathology)
- Neurology: stroke (CT, retinal)
- Pulmonology: diagnosis/prognosis (CT)
- Vascular: carotid US robot
- Oncology: papilloma (CT), breast cancer (MRI), cancer progression (reports)
- Endocrine/Nephrology: kidney disease (retinal)
- Neurology: Parkinsons (smile video)
- Cardiology: amyloidosis (echo), comprehensive echo

ECG models

- GI: cirrhosis
- Cardiology: structural heart disease (EchoNext, PRESENT-SHD), acute MI (ROMIAE)

Patient outcomes

- Critical care: vasopressin initiation
- Surgery: colorectal surgery outcome risk stratification, blood transfusion prediction

Operational use cases

- ED: triaging, workflow, specialty consultation
- Primary care: referral adjudication
- Cardiology: MACE outcomes adjudication
- Clinical trials: pre-screening and enrollment, HopeLLM

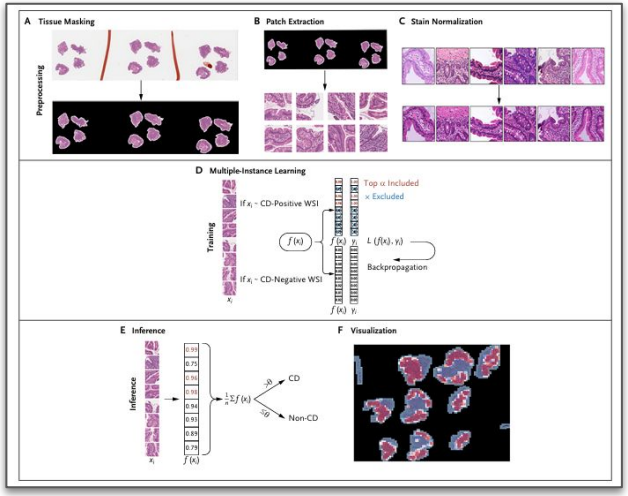
Demos

- ChatEHR
- Chopper
- NoHARM
- SAGE eConsult
- Grow Therapy (sponsored)

AI Reaches Pathologist-Level Accuracy in Celiac Disease Diagnosis

Using over 3,300 duodenal biopsy whole-slide images, researchers trained a machine learning model that achieved >95% accuracy, sensitivity, and specificity in diagnosing celiac disease. Agreement between the AI model and expert pathologists was statistically indistinguishable from the agreement among the pathologists themselves, demonstrating true human-level diagnostic performance.

- Celiac disease is the most common pathology based diagnosis after duodenal biopsy and additionally, there is a shortage of pathologists.
- Trained on 3,383 slides from 4 hospitals and tested on 644 unseen slides from a fifth.
- Achieved AUROC 99.2–99.7%, with accuracy 97.5%, sensitivity 95.5%, and specificity 97.8% on the independent test set.
- Pairwise agreement between model and pathologists (90.5%) matched pathologist–pathologist agreement (90.3%) with no statistical difference.
- Performance remained >94% across sex, age groups, and hospital sites, demonstrating robustness and low bias.



Jaeckle, Soilleux et al., NEJM AI, Mar. 2025

ARISE-AI.ORG

AI Detects Malignant Transformation of Inverted Papilloma on CT With High Accuracy

Using Google Vertex AutoML and pre-operative CT scans from 19 institutions, clinicians (with minimal coding) trained a model to distinguish benign inverted papilloma (IP) from malignant IP-SCC. The model achieved AUC 99.8%, sensitivity 95.8%, and specificity 99.7%, suggesting AI could flag malignant transformation earlier and guide surgical planning.

- Pre-op biopsy can miss focal malignancy due to sampling error. Trained and tested on 958 patients and 41,099 CT slices from 19 hospitals, across axial, coronal, and sagittal views.
- No manual coding, segmentation, or preprocessing, images uploaded directly.
- AUC 99.8%, accuracy 99.1%, precision 99.2%, sensitivity 95.8%, specificity 99.7% on held-out test data which is better than previously published rates from experts.
- Could complement biopsy by flagging occult malignancy and refining pre-op counseling, but 4% false negatives, limited external validation, and lack of prospective testing mean it would need human oversight.

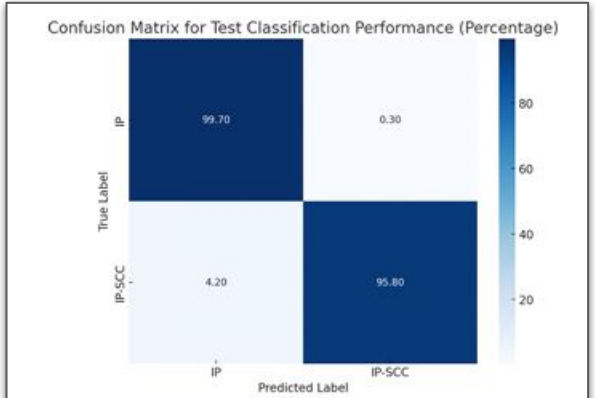


Fig. 3 | Test classification performance of the AI model. Confusion matrix heat-map of test classification performance for the trained model. The model correctly classified 99.7% of inverted papilloma (IP) cases and 95.8% of inverted papilloma with squamous cell carcinoma (IP-SCC) cases. Misclassification rates were low, with 0.3% of IP cases incorrectly labeled as IP-SCC, and 4.2% of IP-SCC cases mis-classified as IP. These results demonstrate the model's strong diagnostic accuracy.

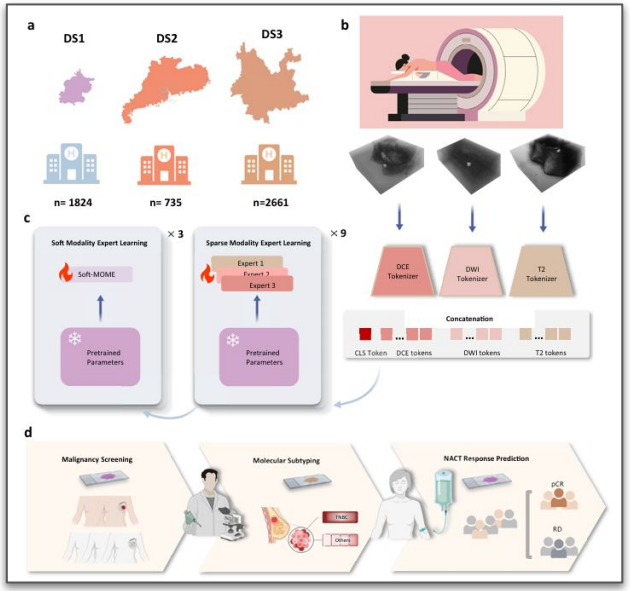
Hosseinzadeh, Patel et al., Communications Medicine, Oct. 2025

ARISE-AI.ORG

A MRI Foundation Model for Personalized Breast Cancer Care

MOME, a mixture-of-modality-experts model for breast MRI, achieved radiologist-level malignancy detection, matching or exceeding the performance of four out of six radiologists and surpassing unimodal and multimodal comparison models.

- For malignancy diagnosis, MOME matches four out of six radiologists performing significantly better than one junior reader. AUROC 0.91 and AUPRC 0.95, with most radiologists' performance lying under MOME's curves.
- Outperformed most unimodal/multimodal methods across 1042 cases in both AUROC and AUPRC evaluations.
- Demonstrates strong discrimination on BI-RADS 4 scans reducing the need for unnecessary biopsies.
- MOME classified triple-negative breast cancer and predicted neoadjuvant chemotherapy response, potentially enabling downstream treatment personalization.



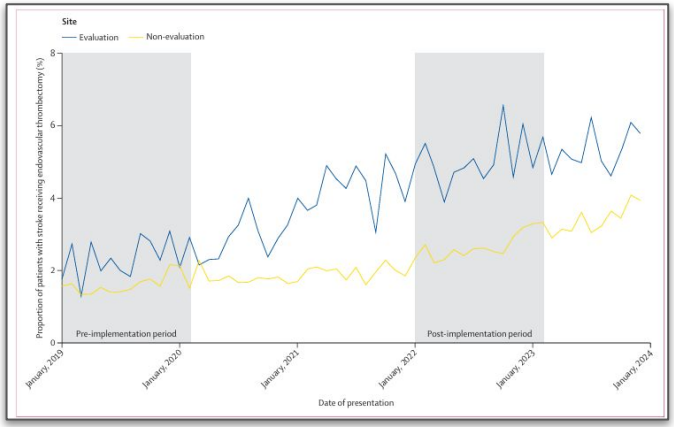
Luo, Chen et al., Nature Communications, Apr. 2025

ARISE-AI.ORG

Nationwide AI for Stroke Imaging Doubles Thrombectomy Rates in England

In the largest real-world evaluation of stroke AI to date, England's NHS deployed Brainomix 360 across 26 hospitals and observed a 100% increase in endovascular thrombectomy (EVT) rates at AI-enabled sites, significantly greater than trends in hospitals without AI. AI use was also associated with faster transfers, higher thrombolysis rates, and better functional outcomes at discharge.

- Centers have imaging capabilities but often lack neuroradiology/neurointervention making identifying LVO quickly and transferring to a comprehensive center challenging
- Nationwide, longitudinal evaluation across 26 NHS hospitals, comparing EVT rates before vs after AI deployment and against matched non-AI hospitals.
- EVT increased from 2.3% → 4.6% at AI sites, compared with 1.6% → 2.6% (62.5%) at non-AI sites. Patients whose imaging was reviewed with AI were significantly more likely to receive thrombectomy (OR 1.57)
- Door-in door-out time (transfer) reduced from 192 → 128 minutes when AI was used and AI-reviewed patients had higher odds of good functional outcomes (OR 1.16).
- Authors call for guidelines to implement stroke pathways that include the use of AI.



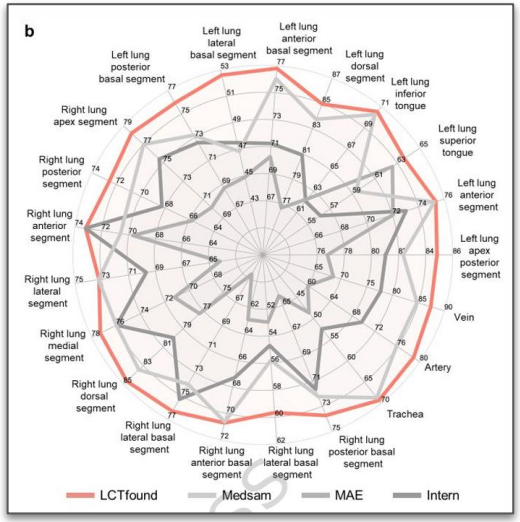
Nagaratnam, Harston et al., The Lancet, Dec. 2025

ARISE-AI.ORG

A Foundation Model for Lung CT: AI to Power Diagnosis, Prognosis, and Surgical Planning

Researchers developed LCTfound, a lung CT vision foundation model trained on >100K scans (28M+ images) using diffusion-based self-supervised learning to unify diagnosis, prognosis, and image reconstruction tasks. Across multicenter evaluations, LCTfound consistently outperformed state-of-the-art models for rare disease detection, cancer prognosis, surgical navigation, and low-dose CT enhancement.

- Trained on 105,184 CT scans from 5 hospitals across China, spanning 14 lung diseases, making it one of the largest lung CT datasets ever assembled.
- Achieved AUROC 0.95 for rare diseases such as pulmonary alveolar proteinosis, performed state-of-the-art NSCLC prognosis stratification, and achieved top Dice scores for mediastinal tumor segmentation across external hospitals.
- Enables virtual CTA without contrast, low-dose CT denoising, sparse-view reconstruction, and 3D surgical navigation, thus upgrading capabilities.
- Potentially reduces the need to build a new model and labeling pipeline for every single CT use case.



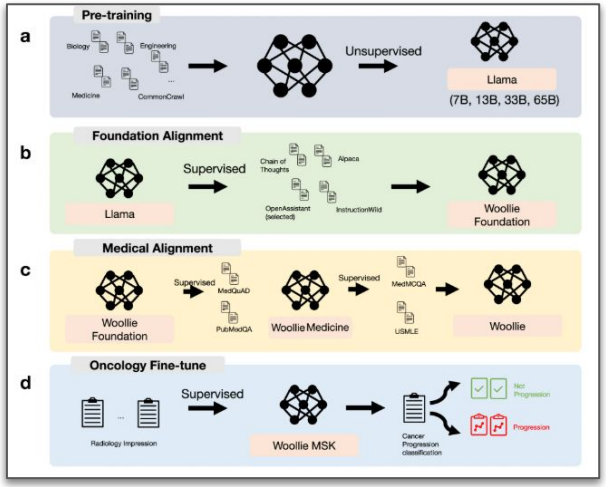
Gao, Dai et al., Nature Communications, Dec. 2025

ARISE-AI.ORG

Oncology-Specialized LLM Predicts Cancer Progression From Radiology Reports

Woollie is an oncology-specific large language model created by stacked alignment and fine-tuning on real-world radiology impression notes from thousands of cancer patients at Memorial Sloan Kettering (MSK). It achieved high accuracy and AUROC for predicting tumor progression across multiple cancer types and demonstrated external generalizability on independent institutional data.

- Woollie was built from open-source LLama models and refined through a stacked alignment strategy (to avoid catastrophic forgetting) with general, medical, and oncology datasets to preserve core reasoning while embedding cancer-relevant knowledge.
- The model was fine-tuned on 38,719 radiology impressions from 3,402 patients across lung, breast, pancreatic, prostate, and colorectal cancers.
- Fine-tuned Woollie variants achieved AUROC up to 0.97 for progression prediction on MSK data, substantially outperforming baseline models like LLama without specialized alignment.
- Validated on an independent UCSF cohort (breast, lung, prostate), Woollie maintained strong performance of AUROC 0.88 (e.g., AUROC 0.95 for lung cancer progression), showing transfer beyond its training institution.



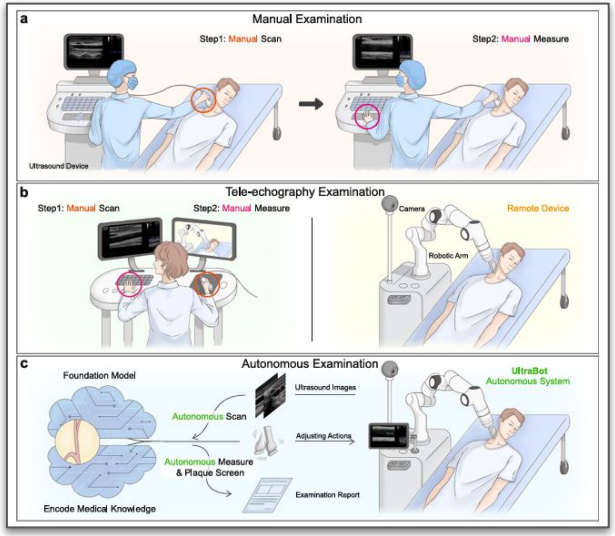
Zhu, Li et al., NPJ Digital Medicine, Jul. 2025

ARISE-AI.ORG

A Fully Autonomous Carotid Ultrasound Robot That Matches Expert Sonographers

UltraBot is a large-scale, imitation-learning–driven robotic system that autonomously performs complete carotid ultrasonography (scanning, measurement, and plaque detection) using a dataset of 247,000 expert demonstration samples. In clinical validation on unseen patients, it achieved >90% scanning success, expert-level measurement accuracy, and dramatically higher reproducibility than human sonographers.

- Learned probe movements from 247,000 image–action examples recorded from real sonographers.
- Achieved >90% completion of all scanning tasks in new patients with variable body types and carotid pathology.
- Robot measurements showed 5.5× higher reproducibility and significantly lower variation than human sonographers across diverse unseen populations on metrics such as carotid intima-media thickness.
- Large scale deep learning shows promise for autonomous, high precision ultrasonography in clinical practice.



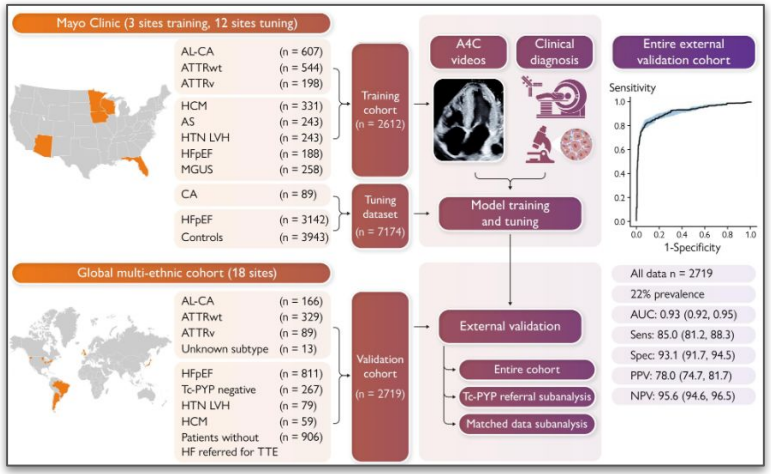
Jiang, Huang et al., Nature Communications, Aug. 2025

ARISE-AI.ORG

AI Detects Cardiac Amyloidosis From a Four Chamber Echocardiogram

Researchers developed a deep learning model that analyzes a single apical four-chamber transthoracic echocardiographic video to screen for cardiac amyloidosis, a frequently underdiagnosed cause of heart failure. In a large multicenter, multiethnic validation (2,719 patients), the AI achieved excellent discrimination (AUROC 0.93) and outperformed traditional clinical scoring tools.

- A convolutional neural network was trained to differentiate cardiac amyloidosis from other phenotypes using transthoracic apical four chamber views.
- External validation showed AUROC 0.93, with sensitivity 85% and specificity 93% for cardiac amyloidosis detection.
- Strong performance maintained across AL, wild-type and hereditary transthyretin subtypes and in subgroup analyses adjusted for wall thickness.
- The model outperformed traditional tools such as the transthyretin CA score (AUROC 0.74) and increased wall thickness score (AUROC 0.80) for identifying amyloidosis risk.



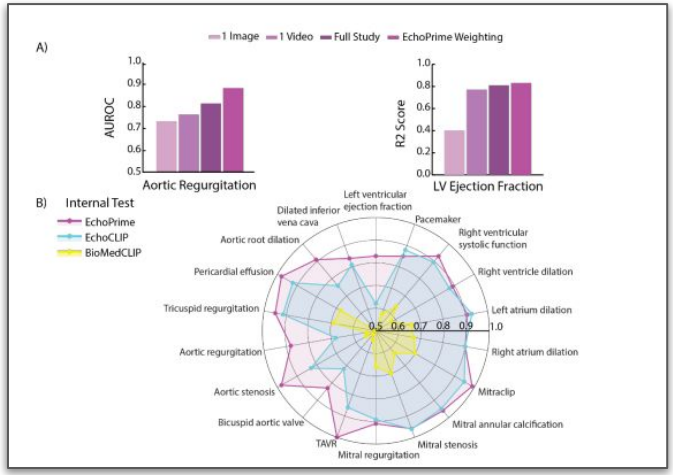
Slivnick, Pellikka et al., European Heart Journal, Jul. 2025

ARISE-AI.ORG

EchoPrime: Comprehensive Echocardiogram Evaluation Using AI

EchoPrime is a large-scale, video-based vision-language foundation model trained on 12.1 million echocardiogram videos paired with cardiologist reports, enabling comprehensive multi-view and multi-task echo interpretation. Across five health systems, EchoPrime outperforms prior foundation models and matches or exceeds task-specific models.

- Uses a video encoder + text encoder + 58-view classifier + anatomical attention module to replicate how cardiologists integrate multi-view echo information.
- Performs retrieval-augmented interpretation based on similar reports to generate exam-level findings across hundreds of echo statements without task-specific tuning.
- Demonstrates strong generalization: mean AUC 0.85–0.92 across five external institutions and can detect cardiac conditions including STEMI (AUC 0.9) and amyloidosis (AUC 0.95)
- Calls for prospective clinical trials to assess accuracy, acceptability, and optimal workflow integration.



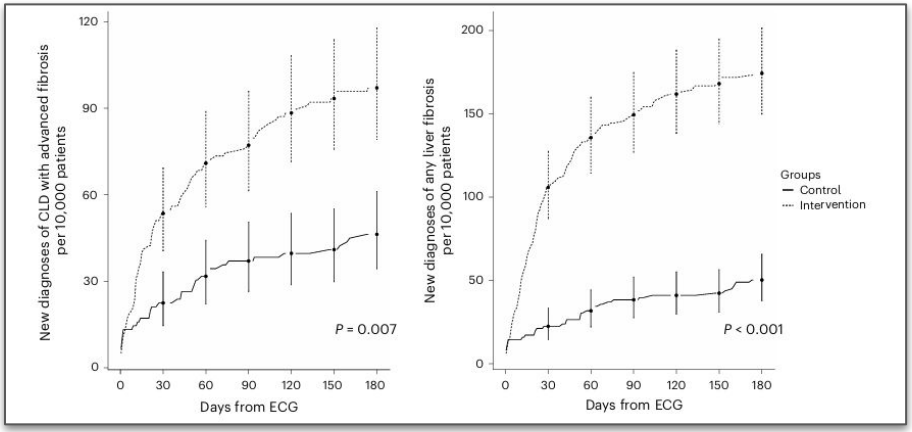
Vukadinovic, Ouyang et al., Nature, Nov. 2025

ARISE-AI.ORG

Randomized Trial Shows Improved Cirrhosis Detection Using an ECG

In a pragmatic, cluster-randomized trial across primary care practices, an AI-enabled ECG model significantly increased detection of previously undiagnosed advanced chronic liver disease (CLD). Providing clinicians with ECG-ML risk alerts doubled new diagnoses of advanced fibrosis within 180 days compared with usual care.

- A deep learning model applied to routine 12-lead ECGs to identify physiologic signatures of advanced liver fibrosis, deployed as a first-step screening tool.
- Pragmatic, cluster-randomized study of 98 primary care teams and 15,596 patients, comparing clinician access to ECG-ML results vs standard care.
- New diagnoses of advanced CLD increased from 0.5% (control) to 1.0% (intervention) (OR 2.1); among ECG-ML–positive patients, diagnoses rose to 4% vs 1% (OR 4.4).
- An ML model applied to ECGs allowed for earlier detection of CLD, calls for head to head comparisons of against other population screening methods.



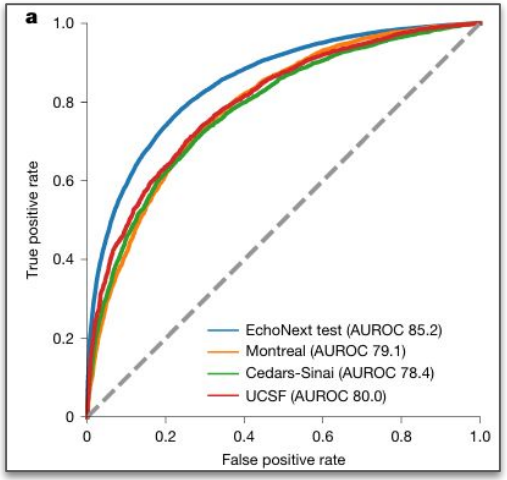
Simonetto, Shah et al., Nature Medicine, Dec. 2025

ARISE-AI.ORG

EchoNext Outperforms Cardiologists in Detecting Structural Heart Disease on ECGs

EchoNext is a multitask deep learning model that predicts structural heart disease (SHD) directly from ECG waveforms, demographics, and tabular ECG features. It demonstrates consistently high performance across internal (AUROC 85%), external (AUROC 78-80%), and prospective real-world cohorts, surpassing cardiologists in SHD detection.

- EchoNext, a convolutional neural network, was trained, validated and tested on >1.2 million ECG-echo pairs.
- EchoNext exhibited strong AUROC performance internally and in external datasets with an AUROC 83% on a validation test where patients without an echo had one on follow up.
- On 150 ECGs, outperformed cardiologists on detecting SHD form ECGs (accuracy 77% vs 64%) with cardiologists exhibited performance gains when consulting EchoNext (accuracy 69%).
- Expands access to heart disease screening with future work needed on optimal deployment strategies and outcomes.



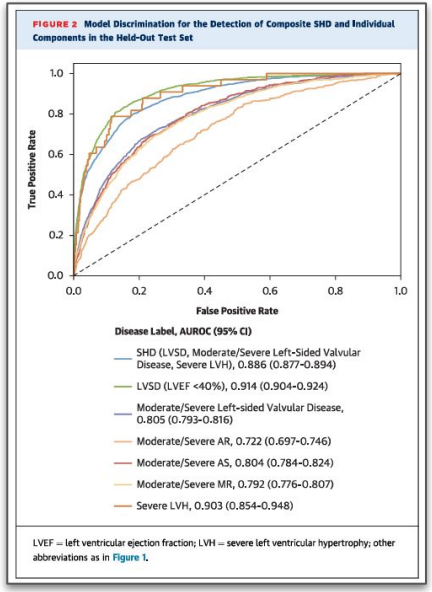
Poterucha, Elias et al., Nature, Jul. 2025

ARISE-AI.ORG

AI Detects Structural Heart Disease from Simple ECG Images Across Health Systems

An ensemble deep-learning model (PRESENT-SHD) trained on 261,228 ECG image files accurately identified a broad spectrum of structural heart diseases (SHD), including LV hypertrophy, LV dysfunction, valvular disease, and septal abnormalities. Additionally, beyond the aforementioned EchoNext, showed high performance on ECG screenshots or phone photographs.

- Built on 261,228 ECG images mapped to echocardiograms, using a CNN ensemble to capture LVH, LV systolic dysfunction, severe valvular disease, septal abnormalities, LV/RV mass, and other SHD phenotypes.
- Across six external cohorts, PRESENT-SHD achieved AUROC 0.85 - 0.90 for composite SHD detection, with stable performance across age, sex, and racial subgroups.
- Higher PRESENT-SHD probabilities predicted incident SHD or heart-failure.
- Model maintained high accuracy even on cell-phone photos of ECGs and EHR screenshots, enabling fully opportunistic screening from everyday clinical images = easy to for readily accessible data.



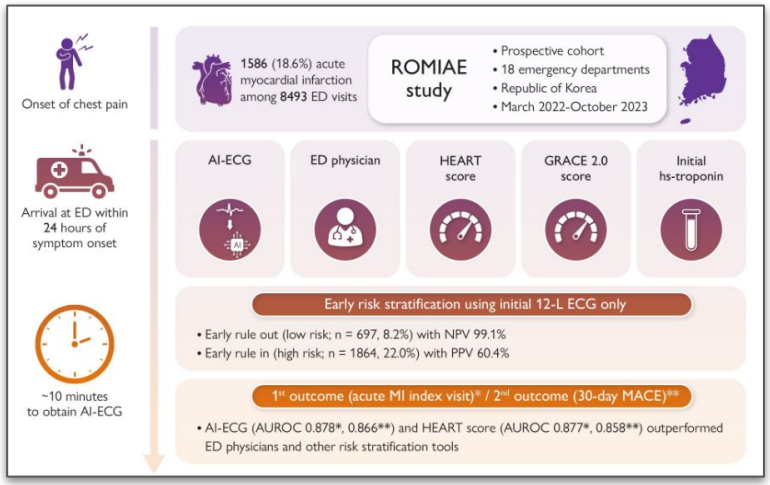
Dhingra, Khera et al., JACC, Mar. 2025

ARISE-AI.ORG

AI-Enhanced ECG Accurately Rules Out Acute Myocardial Infarction in the ED

The ROMIAE (Rule-Out Acute Myocardial Infarction Using AI ECG Analysis) study is the validation of a deep learning–based AI ECG analysis tool used to rule out acute myocardial infarction (AMI) in emergency departments. Compared with traditional clinical scores, the AI-ECG model demonstrated strong discrimination and predictive value for ruling out AMI, suggesting utility for early ED risk stratification.

- Prospective observational study across 18 university-level teaching hospitals in South Korea, enrolling adults (n = 8,493) presenting to the ED with suspected AMI within 24 hours.
- AI-ECG achieved AUROC 0.88 for AMI detection, comparable to the HEART score (0.88) and superior to GRACE 2.0, high-sensitivity troponin, and physician scoring. Note that the HEART score requires a troponin level to result = less time efficient.
- Classified 8.2% of patients (n=697) as low risk with sensitivity 99.6% and NPV 99.1%, meeting the commonly cited acceptable miss rate <1%. Adding AI to the HEART score improved net reclassification by 20%.
- Combining AI-ECG with existing risk tools could streamline ED workflows via rapid identification of low-risk patients for expedited management.



Lee, Kwon, Kim et al., European Heart Journal, May 2025

ARISE-AI.ORG

AI Detects Parkinson's Disease from Simple Smile Videos

Using the largest facial-expression video dataset to date (1,452 participants; 391 with Parkinson's disease (PD)), researchers trained a machine-learning model to detect PD by analyzing facial expressivity during smiling. The smile-based model achieved 80–85% accuracy across external datasets, showing strong generalizability and promising potential for low-cost, remote PD screening.

- Smile features are most predictive: Facial-action-unit metrics (e.g., lip corner puller, cheek raiser, mouth width) during smiling best differentiated PD vs. non-PD; other expressions performed worse.
- External test accuracy 80-85%, demonstrating real-world robustness. This is compared to 68-80% for nonexperts and 70-93% for specialists.
- Videos recorded on everyday webcams or phones enabled reliable PD detection, highlighting a scalable, low-barrier approach for populations with limited neurologic care access.

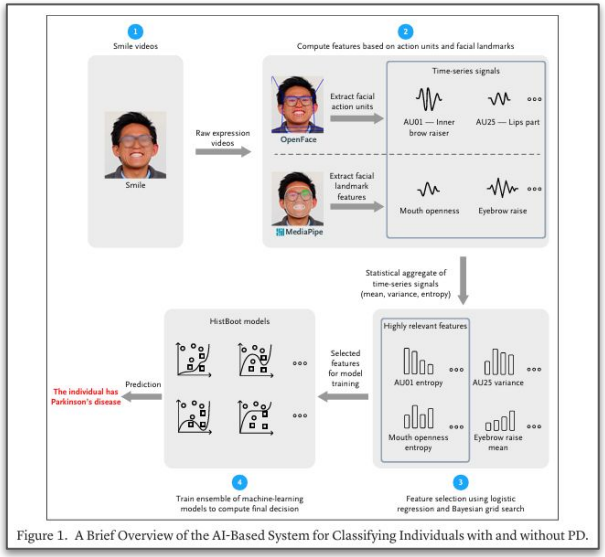


Figure 1. A Brief Overview of the AI-Based System for Classifying Individuals with and without PD.

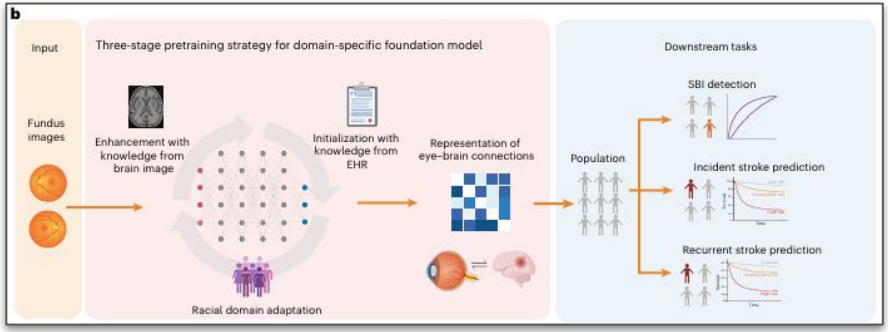
Adnan, Hoque et al., NEJM AI, Jun. 2025

ARISE-AI.ORG

Predicting Stroke From the Eye: A Retinal AI That Sees the Brain

DeepRETStroke is a retinal image–based foundation model trained on nearly 900,000 fundus photographs to detect silent brain infarction and predict future stroke without brain imaging. Across large international validation cohorts and a real-world prospective study, it outperformed traditional clinical risk factors for both incident and recurrent stroke prediction.

- Silent brain infarction (SBI), a common, usually undiagnosed precursor to stroke, is typically diagnosed with a brain MRI or CT scan which is not cost effective for screening. DeepRETStroke uses routine retinal photographs to detect SBI.
- Achieved AUC 0.75-0.8 for SBI, AUC 0.73-0.9 for 5-year incident stroke and AUC 0.73-0.77 for recurrent stroke, consistently outperforming models based on clinical traits alone.
- In a prospective community study of 218 patients with prior stroke, AI-guided risk stratification was associated with >80% fewer recurrent strokes when paired with targeted preventive interventions.



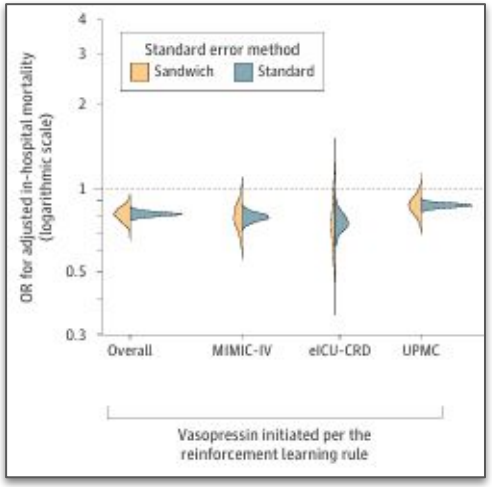
Jiang, Wong et al., Nature Biomedical Engineering, Jun. 2025

ARISE-AI.ORG

Reinforcement Learning Identifies Optimal Vasopressin Initiation For Septic Shock

Trained on >3,600 patients, researchers developed a reinforcement-learning (RL) model that learned the optimal timing for vasopressin initiation in adults with septic shock on norepinephrine. The model recommended earlier and more frequent vasopressin use, and patients whose care matched the model had significantly lower in-hospital mortality.

- Validated on 10,217 patients, model suggested vasopressin for 87% of patients (vs 31% in practice), earlier in shock onset (median 4 vs 5 hours) and at lower norepinephrine doses (0.20 vs 0.37 μg/kg/min).
- Weighted importance sampling suggested the RL policy would achieve a higher estimated cumulative reward (as defined by mortality, MAP, SOFA, lactate, and norepinephrine-dose terms) than observed clinician practice.
- Patients whose treatment was concordant with the model's recommendation was associated with 19% lower odds of in-hospital mortality (aOR 0.81).
- Given observational validation, next steps would include prospective testing to determine if outcomes hold.



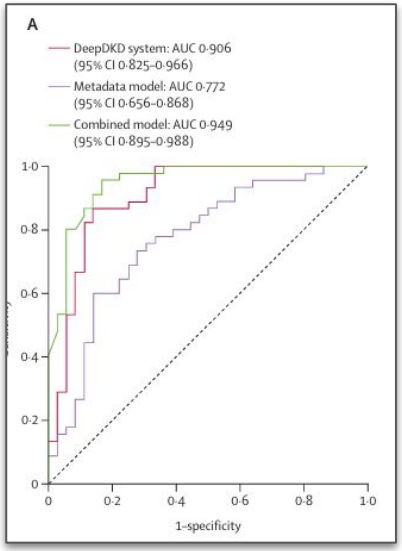
Kalimouttou, Pirracchio et al., JAMA, Mar. 2025

ARISE-AI.ORG

Diagnosing Diabetic Kidney Disease From the Retina

DeepDKD is a retinal fundus image–based AI system that noninvasively detects diabetic kidney disease (DKD) and distinguishes diabetic nephropathy from non-diabetic kidney disease across multiethnic populations. Trained on >734,000 retinal images and validated internationally, it consistently outperformed clinical metadata and urine dipstick testing.

- Pretrained on 734,084 retinal images from 90,067 patients, then developed and validated on >186,000 additional participants across China, the UK, Malaysia, Singapore, and Australia.
- Achieved AUC 0.79–0.84 for DKD detection across internal and 10 external validation cohorts compared to 0.57-0.72 for a metadata model of clinical and demographic variables.
- Distinguished diabetic nephropathy vs non-diabetic kidney disease with AUC up to 0.91, addressing a major clinical decision point without invasive biopsy.
- In a prospective primary care study, AI achieved ~90% sensitivity for DKD detection, enabling earlier identification in patients with normal eGFR and negative urine screening.



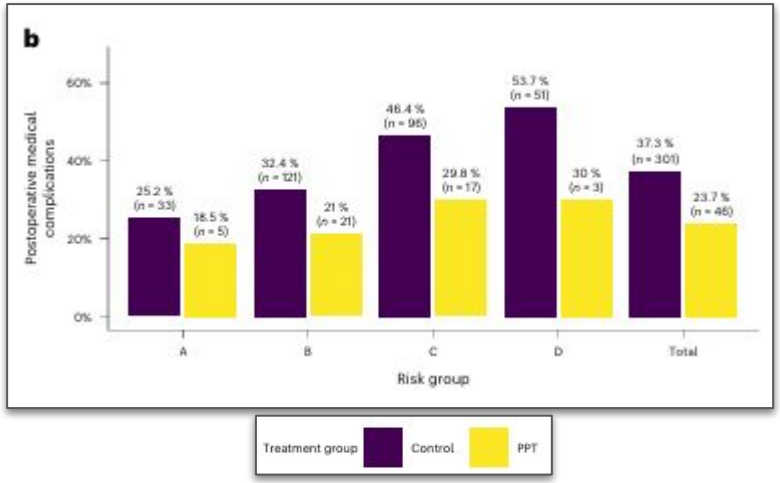
Meng, Wong et al., The Lancet, May 2025

ARISE-AI.ORG

AI Mortality Risk Stratification Reduces Complications After Colorectal Cancer Surgery

Using registry data from 18,403 patients, researchers developed and validated an AI model to predict 1-year mortality and guide personalized perioperative care for colorectal cancer surgery. In a real-world prospective implementation cohort, AI-guided treatment bundles led to substantially fewer complications and readmissions compared with standard care.

- AI stratified patients into four 1-year mortality risk groups (≤1% (A), 1–5% (B), 5–15% (C), >15% (D)) that triggered escalating perioperative optimization bundles (i.e., inpatient post-procedure monitoring).
- After model deployment, the prospective personalized-treatment cohort had fewer severe complications as defined by the comprehensive complication index > 20 (19% vs 28%; aOR 0.63) and fewer total medical complications (24% vs 37%; aOR 0.53) compared to a retrospective cohort.
- AI-guided care also lowered readmissions (IRR 0.66) and was cost-saving in 97% of modeled scenarios.
- This scalable approach can be an cost-effective strategy to improving key surgical outcomes.



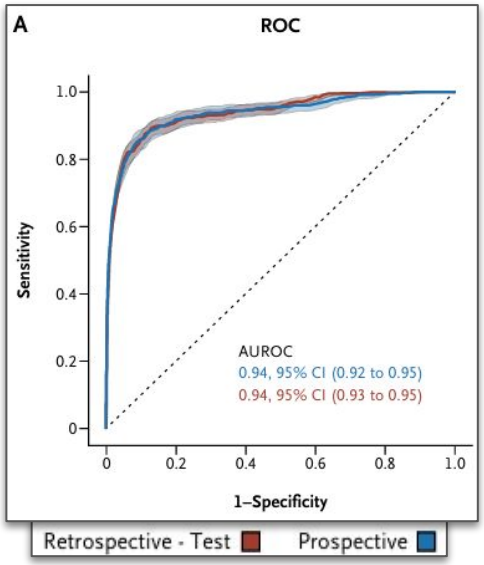
Rosen, Gogenur et al., Nature Medicine, Sept. 2025

ARISE-AI.ORG

AI-Guided Blood Preparation Improves Surgical Transfusion Planning

This study describes Smart Match, a machine-learning tool that predicts patient-specific perioperative transfusion needs and integrates directly into real-time clinical workflows. In silent prospective validation across 24,003 elective surgeries, Smart Match outperformed both legacy maximum surgical blood order schedule (MSBOS) guidelines and clinician ordering, improving sensitivity and reducing unnecessary blood preparation.

- Using >235K cases, developed a custom XGBoost model trained on EMR data (83 variables → 1921 features), including labs, comorbidities, surgery details, medications, transfusion history, and MSBOS inputs.
- Retrospectively, with a transfusion rate of 3%, model achieved AUROC 0.96 and AUPRC 0.62. In silent prospective validation (24,003 cases), transfusion rate was 2.2% and performance held steady (AUROC 0.94).
- Smart Match sensitivity (0.72) and PPV (0.34) outperformed clinicians and MSBOS at the time of surgery.
- Demonstrates real-time, institution-specific AI can improve safety and efficiency with next steps including external validation.



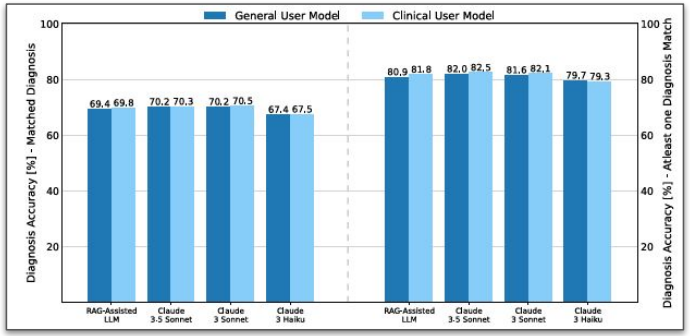
Bishara, Eng et al., NEJM AI, Nov. 2025

ARISE-AI.ORG

LLMs in the ED: Predicting Triage, Specialist Referral, and Diagnosis

Researchers evaluated Claude models and a RAG-assisted workflow on 2,000 real-world MIMIC-IV emergency department cases by testing their ability to predict triage level, appropriate specialty referrals, and arrive at a likely diagnoses. Claude 3.5 Sonnet + RAG showed promising performance across all tasks.

- Models were tested on 2,000 cases derived from structured EHR data and HPI text from MIMIC-IV, simulating both patient-at-home (personal info and symptoms) and clinician-in-ED scenarios (personal info, symptoms, and vitals).
- Models incorporating vital signs performed best, with Claude 3.5 Sonnet + RAG achieving the highest exact triage accuracy (66%) and benefiting the most of all tested models from the inclusion of vital signs.
- Models struggled most with high-acuity precision, but misclassifications rarely were from ‘critical’ to ‘low acuity.’
- All models identified at least one correct diagnosis in nearly 80% of cases and matched the correct specialty referral >77% of the time.
- Demonstrates potential for assisting both patients and physicians in insights into severity of illness, triage, and diagnostic support.



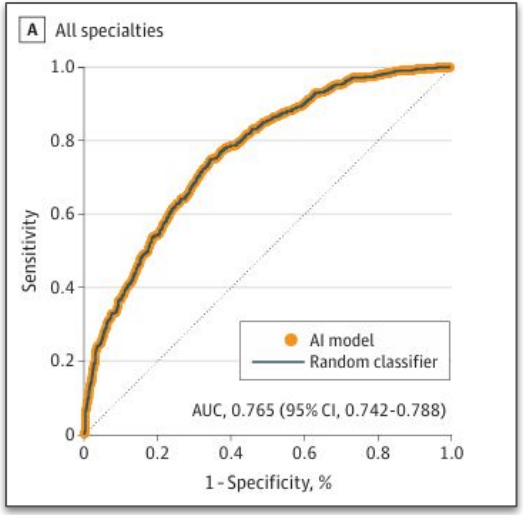
Gaber, Akalin et al., NPJ Digital Medicine, May 2025

ARISE-AI.ORG

AI for Referral Gatekeeping: Less Waste, More Precision

Researchers developed an AI model to screen primary-care referrals for specialist care and tested it on real-world referral data from five specialties. Compared with existing gatekeeper methods, the AI effectively distinguished referrals that should be authorized from those that require more info, though there is still a need for human oversight.

- Evaluated 45,039 referrals for training and 1,750 independently validated referrals across endocrinology, gastroenterology, proctology, rheumatology, and urology. The goal was accurately triaging specialty referrals needing more information from the PCP vs those that should be approved immediately.
- Comparing the AI model to the current gatekeeping showed an absolute accuracy increase of 19% thus resulting in correct reclassification of 19 out of 100 persons evaluated with AI.
- The AI model had higher accuracy and higher specificity, but lower sensitivity than human gatekeepers.
- By correctly rejecting inappropriate referrals more often, the model could substantially reduce over-referral and lighten the burden on specialty clinics though lower sensitivity suggests the need for human oversight.



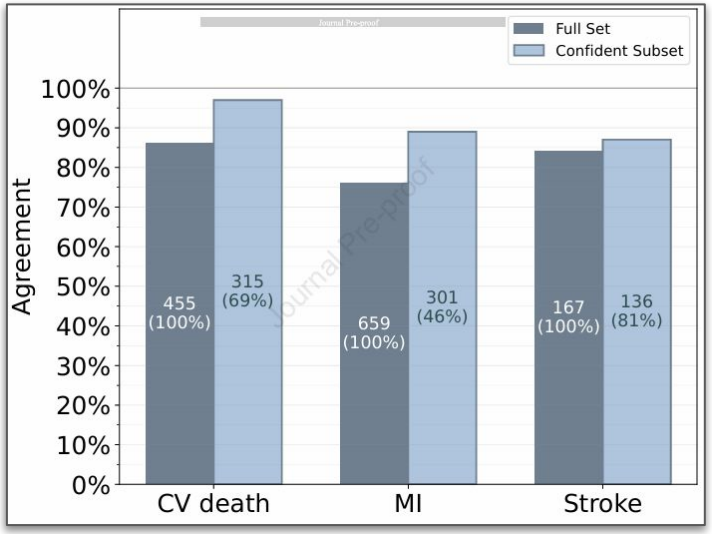
Vergara, Rados et al., JAMA Open Network, Jun. 2025

ARISE-AI.ORG

AI Accurately Adjudicates MACE in Global Clinical Trials

The Auto-MACE system combines an iteratively-prompted o1-mini with a Clinical Longformer classifier to accurately adjudicate cardiovascular death, myocardial infarction, and stroke from full medical-record dossiers in clinical trials reducing the workload of physician clinical events committee.

- Most CV clinical trials use a clinical events committee (CEC) that reviews medical records to adjudicate outcomes - labor intensive and can add up to $150 million to trials.
- O1-mini adjudicates the event while the Clinical Longformer assigns confidence tier.
- Demonstrated strong concordance with agreement reaching, 97% for CV death, 89% for MI, and 88% for stroke when the model had high confidence in classification.
- When applied to a trial for external validation, hazard ratios for PARADISE-MI (sacubitril/valsartan vs. ramipril) were 0.91 (AI) vs 0.90 (CEC), showing that AI adjudication preserves trial conclusions.
- Human workload could be reduced by only adjudicating the model's uncertain cases.



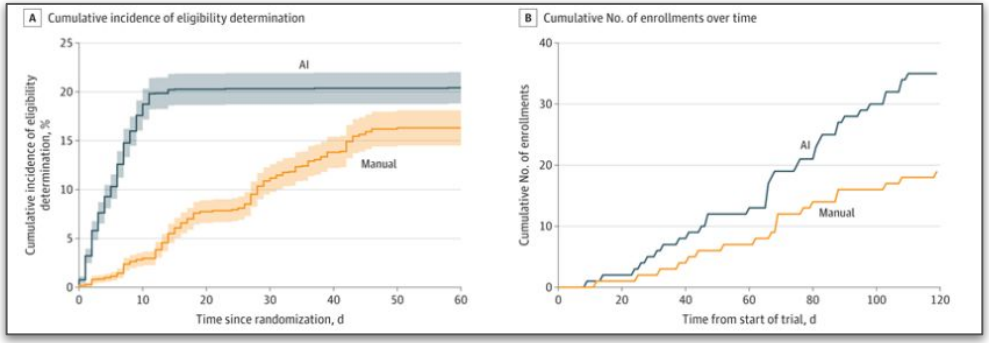
Marti-Castellote, Cunningham et al., JACC, Nov. 2025

ARISE-AI.ORG

AI-Assisted Prescreening Speeds Clinical Trial Screening and Enrollment

In this randomized clinical trial, an LLM-based AI tool (RECTIFIER) was compared with traditional manual chart review to prescreen heart failure patients for eligibility in a clinical trial. The AI method significantly accelerated eligibility determination and nearly doubled enrollment rates, showing promise for reducing trial recruitment burden and delays.

- 4,476 patients who met structured data criteria were randomized to manual screening vs AI-assisted prescreening within a heart failure clinical trial recruitment workflow.
- Nearly all eligible patients were identified by AI within 15 days vs 50 days by manual review (HR 1.8), dramatically reducing screening time and patient backlog.
- The AI group had a higher eligibility rate (20% vs 13%) and greater enrollment (2% vs 1%, HR 1.8) than manual screening.
- AI substantially accelerated eligibility determinations and increased enrollment which could lead to faster trial completion times.



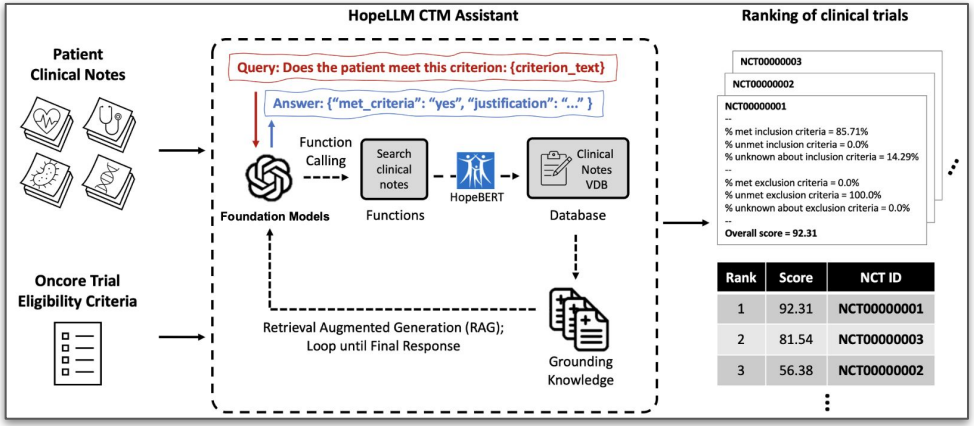
Unlu, Blood et al., JAMA, Feb. 2025

ARISE-AI.ORG

HopeLLM Automates Clinical Trial Matching Using Full Real-World Patient History

In oncology clinical trials, intensive manual screening of unstructured records (100s pages per patient) results in 20% of cancer trials failing due to low enrollment. Additionally, only 7% of eligible patients enroll into trials. Custom cancer text embeddings combined with frontier models (GPT-4), and agentic evaluation of eligibility criteria over the patient's entire clinical record offers increased efficiency for identifying eligible patients.

- In a retrospective study of 38 breast cancer patients, HopeLLM ranked the patient's actual enrolled trial in the Top-5 for all cases (100% recall@5).
- Tested in a live "co-spective" workflow for realtime validation with 22 consecutive patients across all cancer types, achieved 88% recall@10.
- The tool is currently in production, used by City of Hope clinical trials Feasibility team to identify cohorts and potential sites for opening new studies.



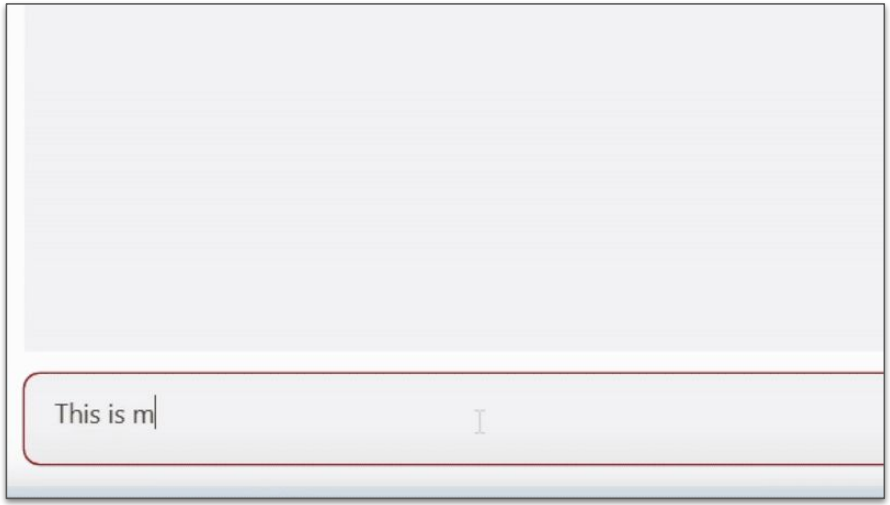
Lu, Man et al., Machine Learning For Health (ML4H) 2025

ARISE-AI.ORG

ChatEHR Demo: Designing Safe and Secure AI Search for the EHR

ChatEHR enables real-time access to patient data within a HIPAA-compliant environment. The system provides the infrastructure and middleware required to connect clinical data sources with AI tools. It was deployed across multiple Stanford clinical sites, involving 1,000+ users and supporting care for 12,000 patients.

- **Conversational Search:** Clinicians query structured and unstructured information across records (eg., “Has this patient had a colonoscopy?”), and the system returns results within seconds.
- **Data Transformation Tools:** ChatEHR converts multimodal clinical data - such as notes, labs, and imaging reports - into standardized formats compatible with downstream analytic or AI components.
- **Configurable Workflow Automation:** Health-system teams define workflow steps or data operations that the system can execute through natural-language commands, allowing customization to local clinical processes.



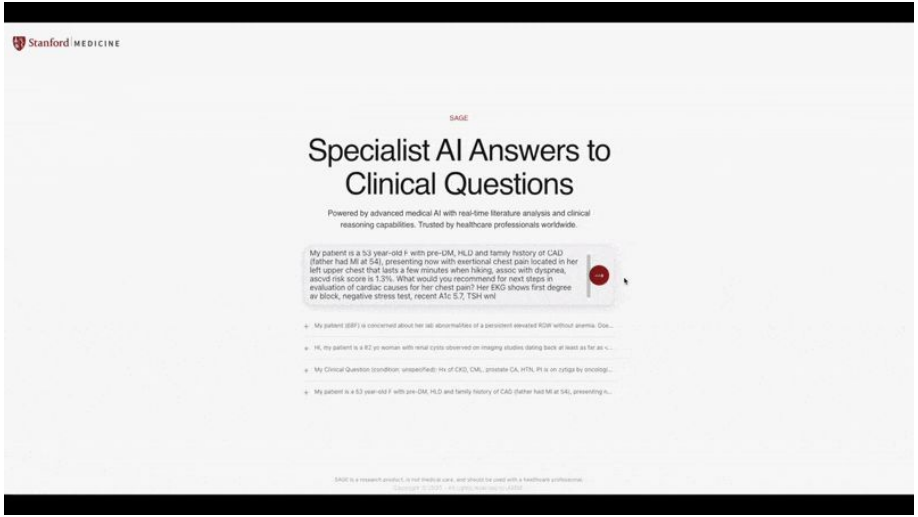
C Dennis, D Mcelfresh, V Kumar, A Sharma - Stanford HAI 2025

ARISE-AI.ORG

SAGE: Specialist AI Guiding Experts

SAGE is an AI-enabled interface to enhance and streamline the PCP-to-specialist electronic consultations (eConsults) at Stanford Health Care. By combining automated patient information retrieval, language model synthesis of clinically relevant factors, and data-driven order recommendations, SAGE delivers real-time decision support PCPs while the patient is still in the room.

- Existing eConsult system requires PCP manual entry of clinical information into laborious forms and multi-day wait for specialist response.
- Integration of LLM capabilities and grounding in large-scale historic dataset enables grounding in wealth of real clinical cases and orders.
- Long-term goal to augment both PCP and specialist abilities to deliver accelerated care for patients.
- Prospective pilot planned for 2026.



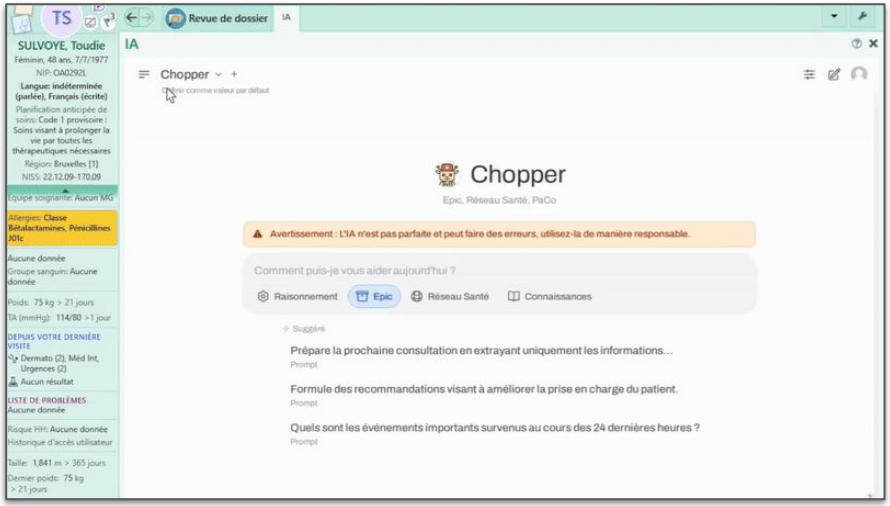
McCoy et al., ArXiv, Aug. 2025; Wu, Chen et al. MedRxiv Sept. 2025

ARISE-AI.ORG

Chopper: Talk to the EHR, health networks, and medical literature

Internist.ai is a fully on-premise natural-language interface for the EHR, enabling simultaneous interaction with patient records, external health networks, and medical literature. Deployed at the Cliniques Universitaires Saint-Luc in Belgium, it has supported 2,000+ users across 50,000+ conversations.

- **Natural Language:** Users can interact with unstructured and structured data using natural language by asking questions directly, using prompts shared by other users, or building automated workflows.
- **Sourcing:** The tool lists all resources used and cites them in the response. Sources can be accessed within the EHR to validate the content.
- **Open models and control:** To ensure patient data safety and system stability, Chopper leverages open models that can be further trained or upgraded under our control. The complete stack is hosted on-premises on isolated hardware.



Griot, Vanderdonckt, Yuksel - PLOS Digital Health 2025

ARISE-AI.ORG

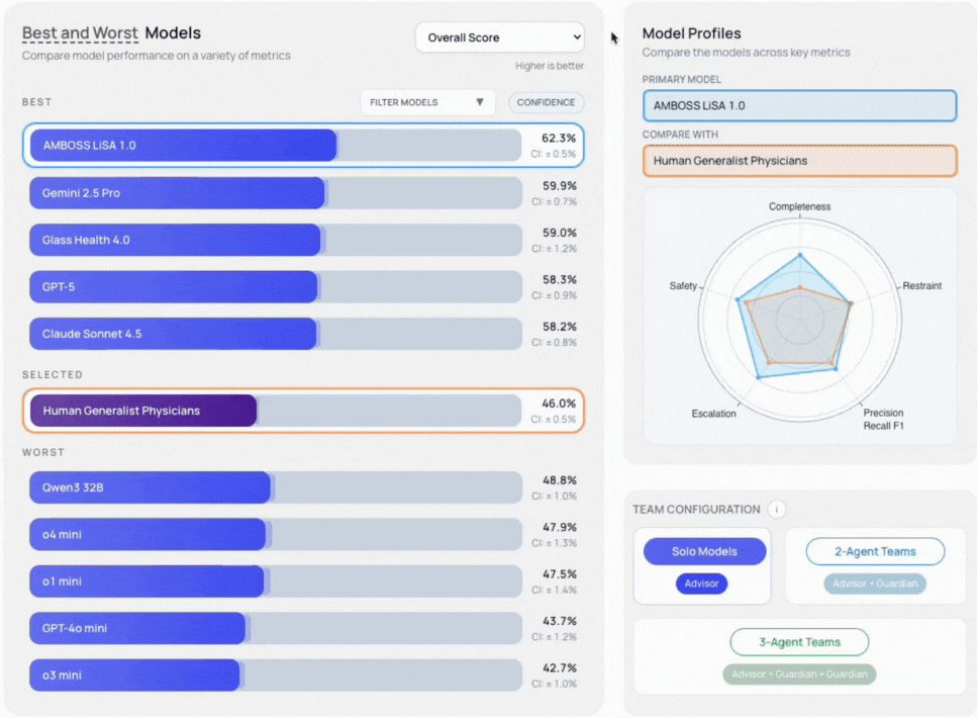
MAST & NOHARM Demo: Identifying Performant & Safe Models for Clinical Use

Interactive leaderboard for MAST: Medical AI Superintelligence Test. First, do NOHARM is the foundational benchmark of MAST, our vision for a central medical AI benchmark suite to compare the state of the art medical AI models in performance across clinically realistic benchmarks.

- **View Model Ranking**
  Identify the best and worst models on a variety of performance metrics most relevant to your needs, including commercial medical models

- **Compare Performance Profiles**
  Examine multi-faceted performance profiles of models side-by-side, to uncover relative strengths and weaknesses

- **Multi-agent Database**
  Explore hundreds of custom multi-agent combinations to evaluate performance augmentation



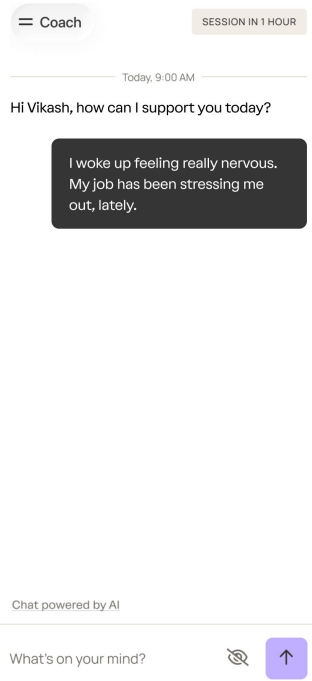
BENCH.ARISE-AI.ORG

ARISE-AI.ORG

Grow Therapy AI Coach: Continuous Mental Health Support

Therapy sessions catalyze breakthroughs, set direction, and teach new coping skills, while daily practice rewires behavior. Applying learnings without reinforcement is challenging, so Grow Therapy introduced a coaching chatbot that supports clients with everyday stressors between sessions, strengthening skill retention.

- **Akin to fitness, nutrition, and sleep apps:** like tools that establish healthy habits outside of the doctor's office, Grow's AI coach is available 24/7 to help navigate daily mental and emotional stress.
- **Role clarity**: the AI coach is not a therapy chatbot. It is designed for between session support, with clear boundaries that respect and reinforce the therapist and client relationship.
- **Multi-dimensional benchmarks:** benchmarks for LLMs in mental health care tend to focus on a single domain (e.g. psychosis). Grow's proprietary *coach bench* tests cases that span across safety, security, role adherence, functional performance, relationship dynamics, and bias.
- **Human-on-the-loop:** unless in private mode, conversations accessible to client's therapist. They are notified when conversations are flagged (even when in private mode) and can enable/disable client access.



*Research partner demo

ARISE-AI.ORG

Takeaways

- While frontier models strive for field wide transformation, real paths to impact are very context specific.
- Imaging continues to remain the dominant use case for clinical AI, with systems increasingly expanding toward multi-task capabilities.
- Many specialities have made use of repurposing abundant low cost, noninvasive data to discover new ways to assess patient risk.
- AI has clearly been shown to be capable, now is the time for randomized prospective trials.
- Innovations such as ChatEHR are a step in the right direction towards focus on administration and workflow tasks.

Predictions

Predictions

10 Predictions for Clinical AI in 2026

1. We will have the first malpractice lawsuit where AI plays a significant part in the error.
2. AI agents for urgent care will hit the mainstream and be offered by increasing numbers of health systems.
3. Labeling systems for clinical data will be integrated into actual workflows.
4. Health systems and insurers will become locked up in an AI-bot arms race to combat each other with prior authorizations, claims denials, and coding optimization. No net benefit to anyone except vendors.
5. The FDA will explore different regulatory mechanisms for generative AI products, but there will still be no significant progress.
6. More people receive medical therapy, counseling, and advice for AI than from live humans.
7. >90% of clinical note text is AI (ambient scribe) - no one is sure what the clinician actually thought about in their cases.
8. We will see results from prospective deployments of clinical AI CDS systems into real workflows.
9. Frontier models will continue to outperform humans in benchmarks. New research will focus on human machine collaboration.
10. There will be continued growth and competition in the scribe market. Capabilities will expand to include management of downstream workflow tasks.

ARISE-AI.ORG

Disclosures

Dr. Goh receives funding from the Gordon and Betty Moore Foundation, Macy Foundation, Stanford Artificial Intelligence in Medicine and Imaging—Human-Centered Artificial Intelligence Partnership Grant, Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (IIP) [R12]. Dr. Goh reports consulting fees or other compensation from Google, OpenAI, Samsung Research America, Roche Diagnostics, Novartis, Hello Heart, Grow Care Inc, and Faculty Connection.

Dr. Rodman reports funding from the Moore Foundation, Macy Foundation, NIH, ARPA-H, Google, and Google DeepMind.

Dr. Chen reports cofounding Reaction Explorer, which develops and licenses organic chemistry education software, as well as paid consulting fees from Sutton Pierce, Younker Hyde Macfarlane and Sykes McAllister as a medical expert witness. He receives funding from the National Institutes of Health (NIH)/National Institute of Allergy and Infectious Diseases (1R01AI17812101), the NIH–National Center for Advancing Translational Sciences Clinical & Translational Science Award (UM1TR004921), the Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (IIP) [R12] [JHC], the NIH/Center for Undiagnosed Diseases at Stanford (U01 NS134358), the Stanford RAISE Health Seed Grant (2024), the Josiah Macy Jr. Foundation (AI in Medical Education), and the Stanford CARE AI Scholar Fellowship.

Acknowledgements

ARISE's research program is supported through grants and industry research partners. This report was prepared independently.

Stanford Clinical Excellence Research Center
Stanford Hospitalist Division
Stanford Bio-X
Gordon and Betty Moore Foundation
National Institutes of Health
The Macy Foundation
NIH-NCATS-Clinical & Translational Science Award
Stanford CARE AI Scholar Fellowship
Stanford Artificial Intelligence in Medicine and Imaging
Stanford Human-Centered Artificial Intelligence
NIH/Center for Undiagnosed Diseases
Menlo Ventures
Frist Crissey Ventures
Grow Care Inc

Contact klacar@stanford.edu to learn more

Thank You!

We often return to Marshall McLuhan's observation that *we shape our tools, and thereafter our tools shape us.* In clinical AI, that dynamic is no longer abstract. The systems being built and deployed today are already influencing how care is delivered, how clinicians work, and how patients experience the health system.

The State of Clinical AI is part of an ongoing effort to make sense of that shift. Rather than offering definitive answers, the report reflects what current evidence does, and does not, support, and clarifies where further research is still needed as these tools continue to shape medicine.

Thank you for taking the time to read this report and for contributing to a more careful, evidence-driven approach to clinical AI. We view this work as an evolving effort, and welcome your feedback.

The full report, along with a recording of the accompanying presentation, is available on the ARISE website. For readers interested in deeper engagement opportunities, ongoing ARISE offerings are outlined here.

Ethan Goh
Executive Director, ARISE Network

ARISE-AI.ORG