Section 3.6- Foundation Models for Healthcare

2026 Spring

LLM Agents Foundation & Applications

Dr. Yanjun Qi

20260210

Road Map

- Last Session: Life Science LLMs and Protein LLMs
- **Benchmarking and Adapting On-Device Large Language Models for Clinical Decision Support**
- Evaluating LLM for Subspecialty
- **Quick Survey of Foundation Models in Healthcare & Life Sciences**

Life Science FMs

Language Models: A Survey on Biological & Chemical Domains

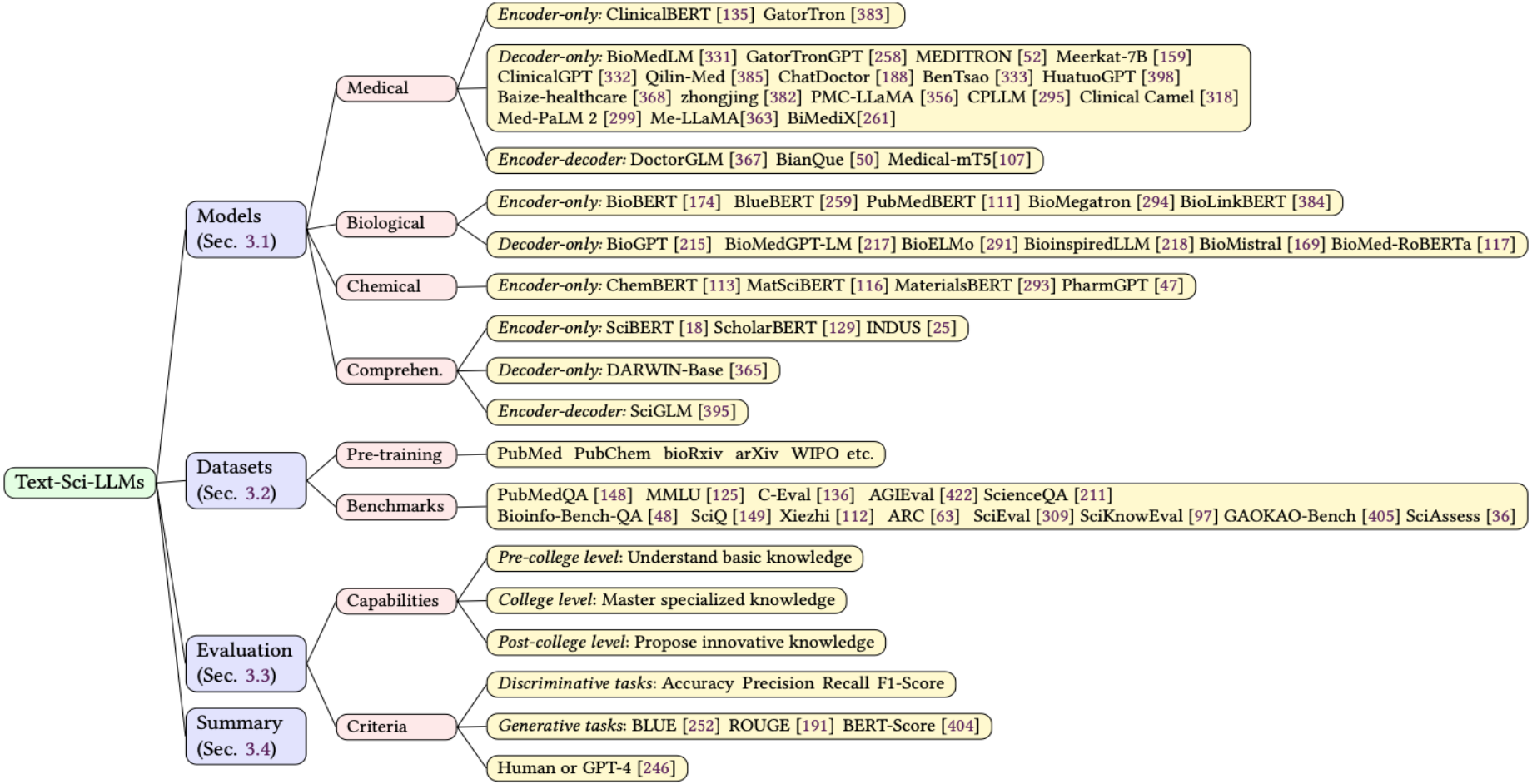


Fig. 6. Chapter overview of Text-Sci-LLMs.

Benchmarking and Adapting On-Device Large Language Models for Clinical Decision Support

Alif Munim*, Jun Ma*, Omar Ibrahim*, Alhusain Abdalla*, Shuolin Yin, Leo Chen, and Bo Wang

Abstract

Large language models (LLMs) have rapidly advanced in clinical decision-making, yet the deployment of proprietary systems is hindered by privacy concerns and reliance on cloud-based infrastructure. Open-source alternatives allow local inference but often require large model sizes that limit their use in resource-constrained clinical settings. Here, we benchmark two on-device LLMs, gpt-oss-20b and gpt-oss-120b, across three representative clinical tasks: general disease diagnosis, specialty-specific (ophthalmology) diagnosis and management, and simulation of human expert grading and evaluation. We compare their performance with state-of-the-art proprietary models (GPT-5 and o4-mini) and a leading open-source model (DeepSeek-R1), and we further evaluate the adaptability of on-device systems by fine-tuning gpt-oss-20b on general diagnostic data. Across tasks, gpt-oss models achieve performance comparable to or exceeding DeepSeek-R1 and o4-mini despite being substantially smaller. In addition, fine-tuning remarkably improves the diagnostic accuracy of gpt-oss-20b, enabling it to approach the performance of GPT-5. These findings highlight the potential of on-device LLMs to deliver accurate, adaptable, and privacy-preserving clinical decision support, offering a practical pathway for broader integration of LLMs into routine clinical practice.

Introduction

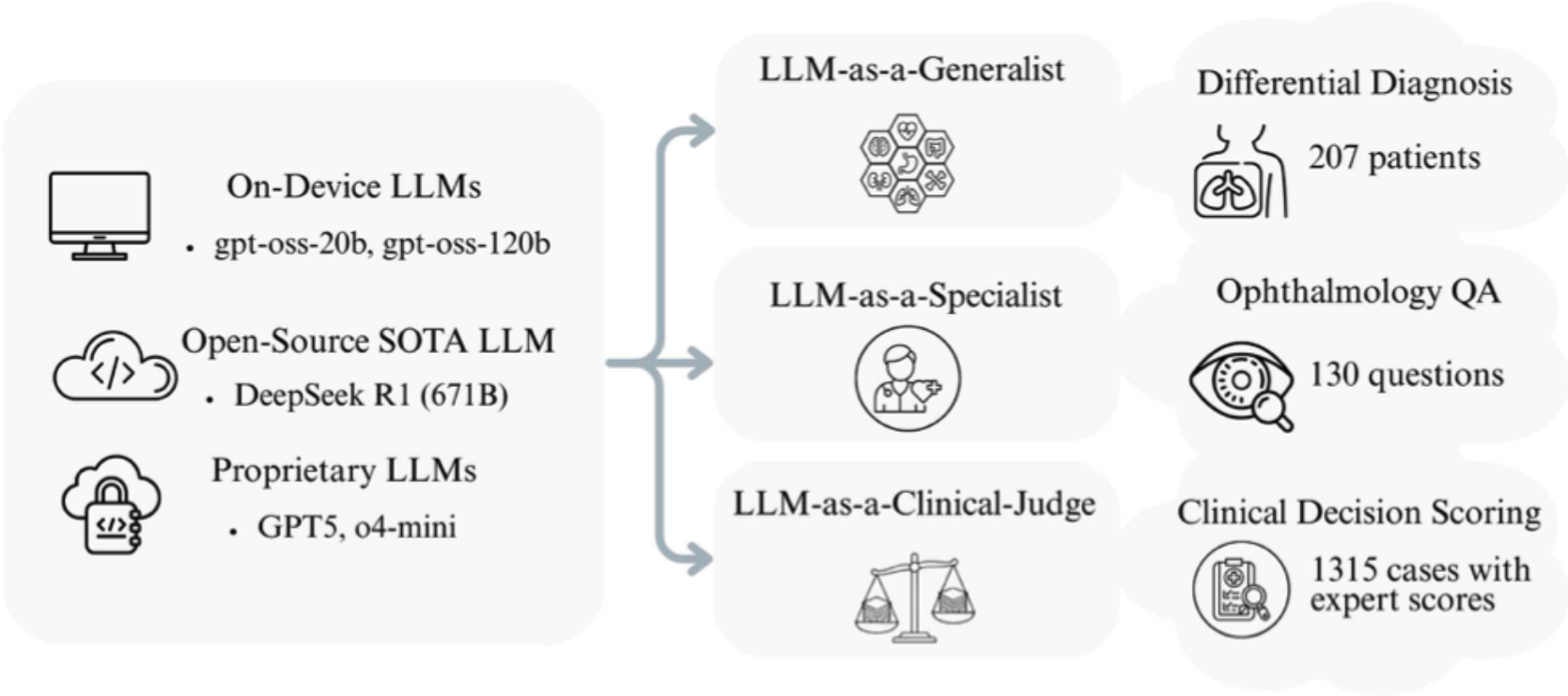
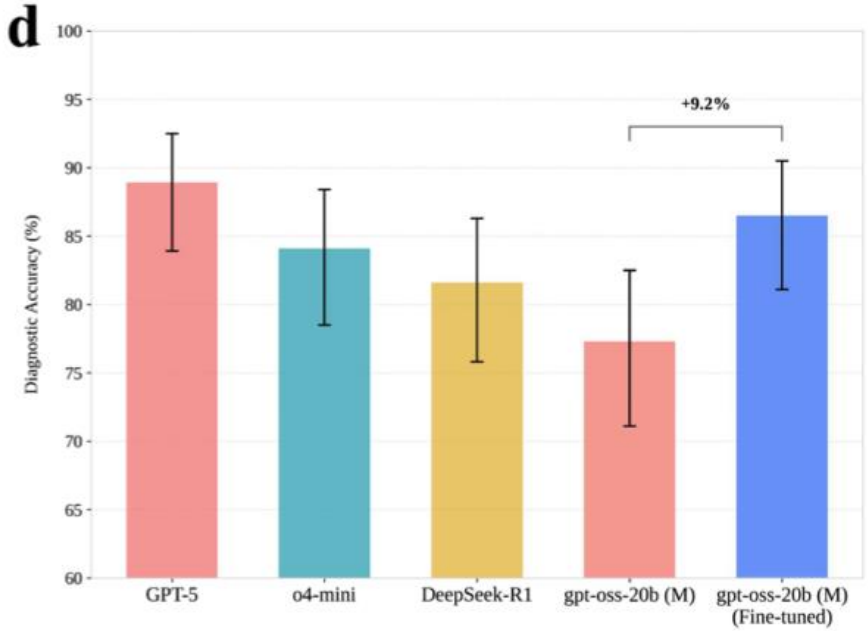


Fig. 1. **Overview of the benchmark framework.** This study compares the on-device LLMs with state-of-the-art open-source and proprietary LLMs across general disease diagnosis, specialty diagnosis and treatment recommendations on ophthalmology multiple-choice questions, and judgment for open-ended clinical decision questions.



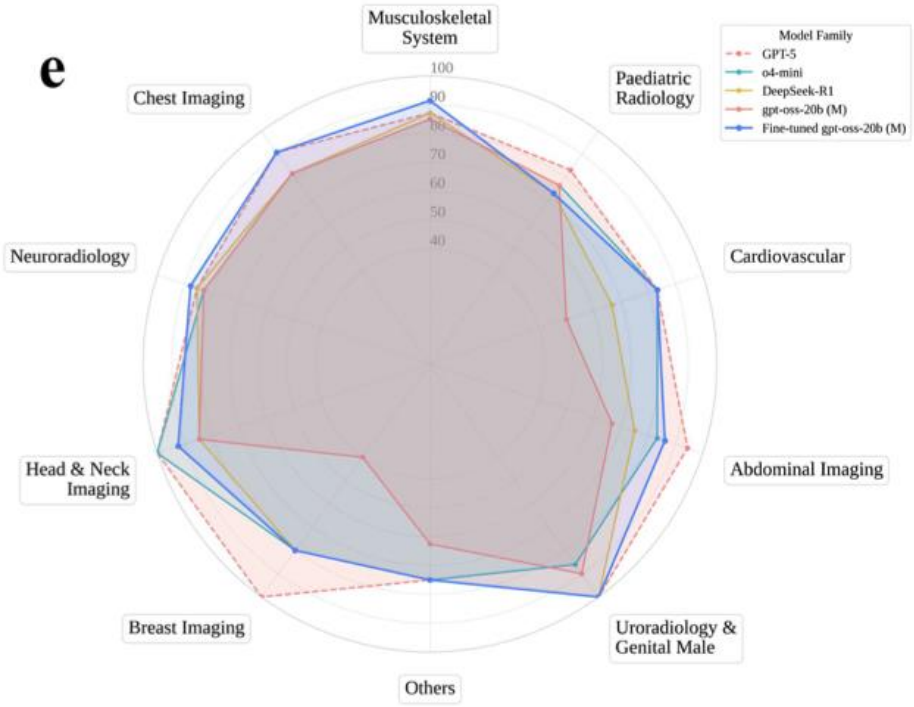


Fig. 2. **Zero-shot and fine-tuning performance of on-device LLMs. a,** Results of LLM-as-a-generalist: diagnosis accuracy on a wide range of radiological cases (N=207). L, M, and H denote low, medium, and high reasoning efforts, respectively. **b,** Results of LLM-as-a-specialist: accuracy on ophthalmology cases (N=130) with diagnosis and management multiple-choice questions. **c,** Results of LLM-as-a-clinical-judge: violin plots comparing the relative error for disease diagnosis and treatment open-ended question assessment (N=1315). **d,** Fine-tuned gpt-oss-20b (M) model outperforms proprietary (o4-mini) and open-source LLMs (DeepSeek-R1) on the disease differential diagnosis task. **e,** Model performance across

Evaluating LLM for Subspecialty

For Example …

A Large Language Model for Complex Cardiology Care

AMIE — Articulate Medical Intelligence Explorer

A Randomized Controlled Trial (RCT) in Genetic Cardiomyopathy

O'Sullivan, Palepu, Saab *et al.* · *Nature Medicine*, February 2026

Stanford University · Google Research · Google DeepMind

Outline

1. **Motivation** — The access problem in subspecialty cardiology
2. **Study Design** — RCT structure and population
3. **Methods**
    - (1) Input Format
    - (2) Output / Assessment Form
    - (3) Model Architecture (AMIE)
    - (4) Domain Adaptation
    - (5) Evaluation Framework
4. **Results** — Preference, errors, omissions, self-report, hallucinations
5. **Limitations**
6. **Takeaways**

The Subspecialist Access Problem

The **WHO** projects a global deficit of **18 million** healthcare providers by 2030. The harm is greatest for rare, complex conditions where delayed diagnosis is life-threatening.

Hypertrophic Cardiomyopathy (HCM) as a concrete example

Fact	Number
Global healthcare provider deficit (projected 2030)	18 million
U.S. states with **no** HCM subspecialist center	27 out of 50
HCM patients in the U.S. who are **undiagnosed**	>60%
HCM's role in sudden cardiac death	Leading cause in young adults
Preventability	High — ICDs are highly effective

The referral cascade to subspecialty care creates delays, anxiety, and missed windows for life-saving intervention.

Why LLMs? Why an RCT?

LLMs can potentially assist generalists by:

- Synthesizing complex, multi-modal clinical data
- Suggesting differential diagnoses and management plans
- Providing decision support where subspecialists are unavailable

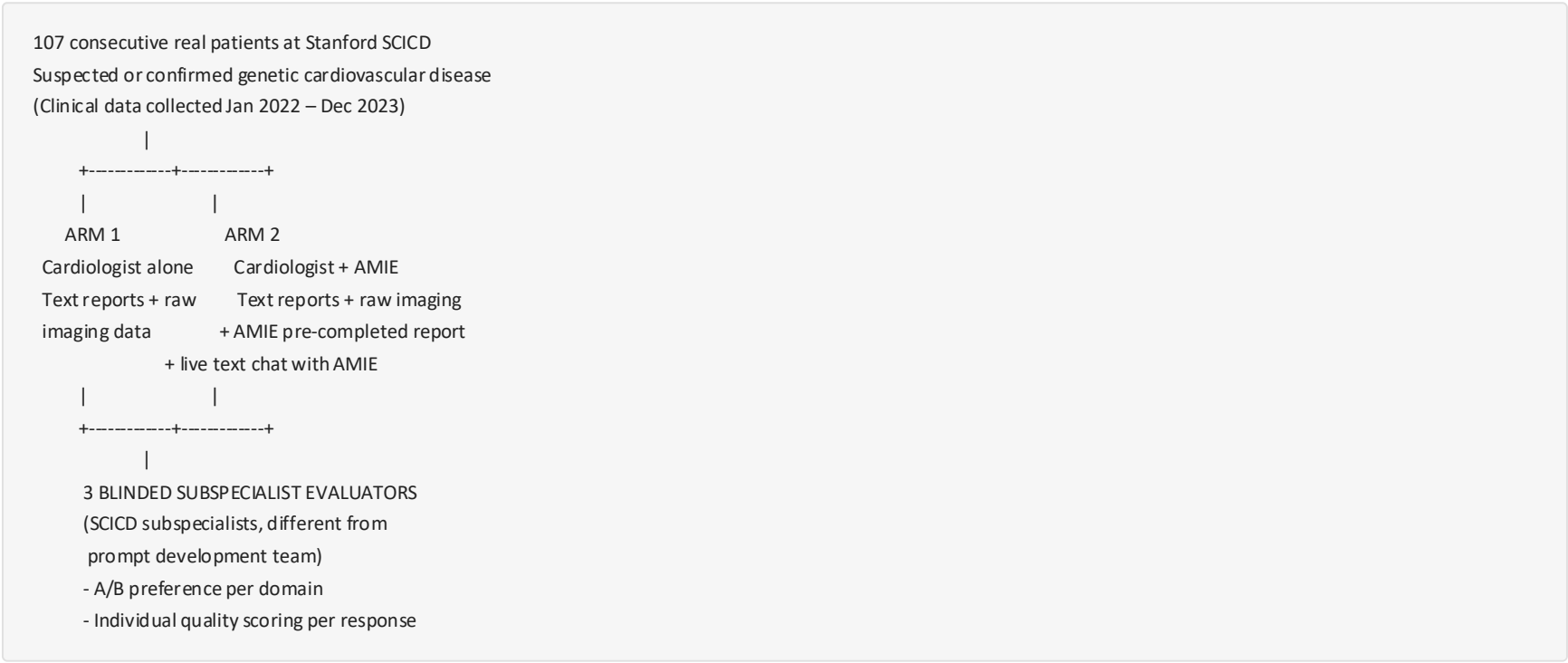
The evidence gap is significant:

> *A 2025 systematic review found **no randomized controlled trials** assessing LLMs in cardiology. Most existing work relies on synthetic vignettes, small samples, and observational designs.*

This paper's contributions:

- First RCT of LLM-assisted care in a subspecialty cardiology domain
- Real-world patient data from a large inherited cardiovascular disease center
- Open-source dataset for community benchmarking
- A validated 10-domain evaluation rubric for future studies

RCT Structure



Pool of **9 Stanford general cardiologists** — 2 assigned per case, 1 per arm.

Fully counterbalanced · CONSORT-compliant · ClinicalTrials.gov NCT06935253

Patient Population

Diagnosis category	n (%)
Hypertrophic cardiomyopathy (HCM)	22 (20.6%)
Left ventricular noncompaction	21 (19.6%)
Arrhythmogenic cardiomyopathy	11 (10.3%)
Ischemic cardiomyopathy	11 (10.3%)
Dilated cardiomyopathy	8 (7.5%)
Other genetic cardiomyopathy	11 (10.3%)
Non-genetic / general cardiology	21 (19.6%)

Median age: 59 years (range 18–96)

Pathogenic or likely-pathogenic variant confirmed: 39 / 107 cases (36.4%)

Genetic testing results were withheld from both cardiologists and AMIE — used only as ground-truth context.

3. Methods

(1) Input Format

Both arms received physician-authored clinical text reports. Not all modalities were available for every patient.

Modality	Cases with data
Electrocardiogram (ECG)	99 / 107 (92.5%)
Ambulatory Holter monitor	79 / 107 (73.8%)
Resting transthoracic echocardiogram (TTE)	90 / 107 (84.1%)
Exercise stress TTE	69 / 107 (64.5%)
Cardiac MRI (CMR)	64 / 107 (59.8%)
Cardiopulmonary exercise test (CPX)	65 / 107 (60.7%)

Key asymmetry between AMIE and cardiologists:

- Cardiologists (both arms) also had access to **raw imaging artifacts** — echocardiogram videos, CMR images, ECG tracings
- AMIE received **text reports only** — no raw images, and **no access to imaging dates**
- The lack of dates was a documented source of temporal reasoning errors

(2) Output: Standardized Assessment Form

All assessors — AMIE first, then cardiologists — completed the **same structured form**:

Section	Task
Overall Impression	Free-text summary of the overall patient picture
Consult Question	Does this patient have a genetic heart disease? (brief rationale)
Triage Assessment	No referral required / Referral required / More information needed
Diagnosis	Most likely diagnosis; further history needed; further tests needed
Management Plan	Proposed plan; further information needed to guide management

Workflow in the AMIE-assisted arm:

1. AMIE receives the clinical text and pre-completes the assessment form
2. The cardiologist reviews AMIE's report and raw imaging
3. The cardiologist may **interact with AMIE via live chat** to query, refine, or push back
4. The cardiologist submits their own final assessment

(3) Model Architecture: AMIE

AMIE = Articulate Medical Intelligence Explorer

Base model: Gemini 2.0 Flash — a publicly available general-purpose LLM

No domain-specific fine-tuning was applied. Instead, AMIE uses an inference-time pipeline:

Pipeline Step	Function
1. Assessment generation	Reads clinical text reports and completes the structured form
2. Web search integration	Retrieves current literature on rare conditions relevant to the case
3. Self-critique	Reviews its own output, identifies weaknesses, and revises

User interface: Web-based chat where cardiologists see AMIE's pre-completed report and can query it interactively.

> *This inference-time approach — rather than fine-tuning — means AMIE can be adapted to new domains with minimal case data and without retraining.*

(4) Domain Adaptation

Aspect	Detail
Base model	Gemini 2.0 Flash (publicly available, no fine-tuning)
Adaptation method	Iterative prompt engineering
Development set size	**9 cases** (held out, not used in evaluation)
Expert input source	SCICD subspecialists (different from evaluation team)
Adaptation techniques	Self-critique loop · web search · subspecialist feedback
Test/dev separation	Strict — zero overlap between 9 development and 107 test cases

Important implication: Achieving subspecialist-preferred performance required only 9 cases of iterative expert-guided refinement. This contrasts with earlier studies using generic LLMs without specialization, which did not reach comparable clinical performance.

The data efficiency here suggests this approach may be scalable to other subspecialty domains.

(5) Evaluation Framework

A — Direct A/B Preference (blinded, per case)

Three subspecialist evaluators from SCICD (different from prompt development team) compared the two cardiologist responses per case, choosing a preference or tie across **10 domains:**

Overall response · Overall impression · Consult question · Triage · Diagnosis · Further diagnostic questions for patient · Further diagnostic tests · Management plan (overall) · Management questions for patient · Management tests

B — Individual Response Quality (blinded, Yes/No, per response)

Each response was independently scored on 5 binary dimensions:

1. **Errors** — Are there clinically significant errors?
2. **Extra content** — Does it contain information it shouldn't?
3. **Missing content** — Does it omit information it should include?
4. **Reasoning** — Does it contain correct clinical reasoning steps?
5. **Bias** — Does it contain demographically inapplicable or inaccurate information?

Statistics: Two-proportion z-tests (A/B preference) · McNemar's test (paired binary) · 95% CI bootstrapped n = 10,000

4. Results

Result 1 — Primary Outcome: Subspecialist Preference

> *Blinded subspecialists preferred which response, across 107 paired cases?*

Domain	+ AMIE preferred	Tie	Alone preferred	p-value
Entire response	**46.7%**	20.6%	32.7%	**0.02**
Management plan	**45.8%**	24.3%	29.9%	**0.008**
Further diagnostic tests	**43.9%**	25.3%	30.8%	**0.03**
Triage assessment	—	Tie	—	n.s.
Diagnosis	—	Tie	—	n.s.
Consult question	—	Tie	—	n.s.
Further diagnostic questions	—	Tie	—	n.s.

Cardiologist-alone responses were not preferred in any domain.

Statistically significant advantages for AMIE-assisted responses appeared in overall quality, management planning, and diagnostic test selection — domains requiring broad clinical synthesis.

Result 2 — Individual Quality Metrics

> *On objective quality dimensions, how did the two arms compare?*

Quality dimension	+ AMIE	Alone	Difference	p-value
Clinically significant errors	**13.1%**	24.3%	−11.2%	**0.033**
Missing content	**17.8%**	37.4%	−19.6%	**0.0021**
Extra content	equivalent	equivalent	n.s.	n.s.
Correct reasoning steps	equivalent	equivalent	n.s.	n.s.
Demographic bias	equivalent	equivalent	n.s.	n.s.

AMIE assistance **significantly reduced errors and omissions** without increasing spurious content, degrading reasoning quality, or introducing bias.

Unassisted cardiologists showed higher precision in triage and initial diagnosis, but higher omission rates in nuanced management.

Result 3 — Cardiologist Self-Report (n = 107 AI-assisted cases)

> *How did cardiologists using AMIE evaluate the experience?*

Question	Yes	No
Did AI **help** your assessment?	**57.0%**	12.1%
Did AI make you more **confident**?	**52.3%**	14.9%
Did AI **save time**?	**50.5%** saved >10%	18.7% delayed
Did AI have **hallucinations**?	6.5% clinically significant	**91.6%** none
Did AI **miss** anything?	6.5% clinically significant	**93.5%** nothing

Magnitude of time savings: In 23.4% of all cases, cardiologists saved more than 50% of their assessment time.

Note: cardiologists were not blinded to their arm — self-reported metrics are subject to performance bias. Primary preference and error outcomes were protected by blinded subspecialist evaluation.

Result 4 — Hallucinations: Types and Handling

Hallucinations were documented across **8 cases** and fell into four categories:

Type	Example
Fabricated imaging finding	Stated "left ventricular hypertrabeculation on CMR" — absent from the report
Diagnostic contradiction	Assertion differing from the cardiologist's reading of the same data
Demographic assumption	Assumed patient sex when none was provided
Temporal misinterpretation	Confused exercise vs. resting measurements; unaware of therapeutic change between two studies due to lack of date access

Critical observation: When cardiologists **directly challenged** AMIE on a stated hallucination, the system typically self-corrected. For example, when told that hypertrabeculation was not in the imaging report, AMIE acknowledged the error and withdrew the claim.

> *This supports human-in-the-loop deployment over autonomous AI use — oversight catches most errors.*

Result 5 — Qualitative Themes

Where AMIE helped (free-text comments, 41 of 107 cases):

- Provided detailed, current knowledge on rare conditions and risk calculators
- Provoked cardiologists to reconsider initial impressions — a "second opinion" effect
- Saved time synthesizing multi-modal reports spanning 5–7 modalities

Where AMIE fell short:

- Outputs sometimes overly verbose, dogmatic, or lacking clinical nuance
- Could not process temporal sequences — unaware of imaging dates or interval changes
- Made redundant recommendations for tests already completed

What subspecialists praised in unassisted responses:

- Cleaner, more concise reasoning
- Less diagnostic overreach

> *Many error types — missed diagnoses, incorrect screening decisions — appeared at **similar frequencies in both arms**. AI assistance did not improve every clinical dimension uniformly.*

Limitations

Limitation	Implication
AMIE received text only — no raw images or imaging dates	Risk of fabrication and temporal reasoning errors
No patient history or physical exam	Limits direct applicability to live clinical workflows
Retrospective de-identified data — not live care	Cannot establish prospective patient outcome benefit
Single center (Stanford), English text only	Generalizability to other systems and languages is unknown
107 cases	Adequately powered for primary outcomes; subgroup analyses underpowered
Cardiologists not blinded to their arm	Self-reported helpfulness and time metrics subject to performance bias
Subspecialists from same institution as developers	Potential institutional bias, mitigated by using separate specialists for development vs. evaluation
Preference-based primary outcome	Does not establish real-world clinical benefit or patient outcomes
High-prevalence inherited disease clinic	AI benefit may be lower in general cardiology where rare conditions are less common

Automation bias risk: Cardiologists may over-rely on AMIE, leading to unnecessary tests, costs, or procedural risk. Critical appraisal training is essential before any clinical deployment.

Key Takeaways

What this study demonstrated

- LLM-assisted cardiologists were significantly preferred by blinded subspecialists — **46.7% vs. 32.7%** (p = 0.02)
- AMIE assistance reduced **clinically significant errors by 11.2%** (p = 0.033) and **missing content by 19.6%** (p = 0.002)
- Benefits were strongest for **management planning** and **diagnostic test selection** — not triage or initial diagnosis
- Domain adaptation required only **9 development cases** — highly data-efficient
- Human oversight remained essential: hallucinations occurred but were largely detectable and correctable

What this study did not demonstrate

- LLMs as autonomous replacements for physicians
- Benefit for triage accuracy or initial diagnostic formulation
- Generalizability beyond a single U.S. subspecialist center
- Real patient outcome improvement — prospective studies are needed

Implications and Next Steps

For clinicians:

LLMs can function as a "subspecialist co-pilot," most valuable for comprehensive management of complex cases — particularly where subspecialists are unavailable.

For researchers:

- RCTs for medical LLMs are feasible and should become the standard for evaluation
- The validated 10-domain rubric and open-source dataset support future benchmarking
- Future work: prospective live deployment, diverse populations, patient perspectives, downstream outcome measurement

For healthcare systems:

- LLMs could help address the **cardiology workforce crisis** and extend subspecialist access to underserved regions
- May assist in identifying the **>60% of U.S. HCM patients** who are currently undiagnosed
- Equity research and multi-center validation are needed before deployment

Summary

Metric	+ AMIE	Alone	Significance
Subspecialist preference	**46.7%**	32.7%	p = 0.02
Clinically significant errors	**13.1%**	24.3%	p = 0.033
Missing content	**17.8%**	37.4%	p = 0.0021
Management plan preference	**45.8%**	29.9%	p = 0.008
Cardiologists: AI helped	**57.0%**	—	—
Time saved	**50.5%** of cases	—	—

> *AMIE — built on Gemini 2.0 Flash with web search and self-critique, and without any fine-tuning — significantly improved general cardiologists' assessments of complex genetic cardiac disease in a blinded randomized controlled trial.*

O'Sullivan, Palepu et al. · Nature Medicine 2026 · DOI: 10.1038/s41591-025-04190-9

Open dataset: redivis.com/datasets/1z3x-2354972da (CC 4.0)

Appendix A — Example AMIE Assessment

Patient: 69-year-old, echocardiographic findings suggestive of HCM

Overall impression: Mild asymmetric septal thickening (12 mm), chordal SAM, mild dynamic LVOT gradient (24 mmHg with Valsalva), mild-moderate MR/TR, mild pulmonary hypertension. Wall thickness is below the standard overt HCM threshold, but the constellation warrants further evaluation for inherited cardiomyopathy.

Consult question: Yes — echocardiographic findings of asymmetric septal thickening, chordal SAM, and dynamic LVOT obstruction are consistent with HCM, a common inherited cardiovascular disease.

Triage: Referral required to inherited cardiovascular disease center.

Most likely diagnosis: Suspected HCM or other inherited cardiomyopathy.

Further tests: CMR for fibrosis/LGE and detailed wall thickness; genetic testing; exercise stress echo to characterize LVOT gradient under physiologic stress.

Management: Comprehensive SCD risk stratification; beta-blockers or CCBs if symptomatic; septal reduction therapy only if severe, refractory, with significant gradient; ICD evaluation based on risk profile; genetic counseling.

Appendix B — Comparison with Prior RCTs of LLMs in Medicine

Study	Patient data	LLM	N cases	Key finding
This paper	Real-world, subspecialty	AMIE (Gemini 2.0 Flash)	107	+14 pp preference; −11% errors; −20% omissions
Goh *et al.* 2024 (JAMA Net. Open)	Synthetic vignettes	GPT-4	6	No significant improvement (76% vs. 74%, p = 0.60)
Goh *et al.* 2025 (Nat. Med.)	Expert vignettes	GPT-4	5	+6.5% overall; management +6.1%; diagnosis +12.1%

Distinctive features of this study vs. prior work:

- Real-world multi-modal patient data (not synthetic vignettes)
- Subspecialty domain involving rare, life-threatening conditions
- Largest case set among cardiology LLM RCTs to date
- Open-source dataset and validated rubric released
- Inference-time adaptation only — no fine-tuning required

Quick Survey of Foundation Models in Healthcare & Life Sciences

February 2026

(assisted creation by DeepResearch and Claude)

Roadmap

Related Disciplines	The science converging here
The Regulatory Landscape	Context for everything that follows
Big Tech Platforms	The infrastructure layer
Drug Discovery	Upstream: molecules to trials
AI-Powered Diagnostics	Midstream: images to clinical alerts
Care Delivery AI	Downstream: clinician and patient tools
The Unsolved Problems	Reality check — the paradox explained
Outlook and Conclusion	Where this goes next

A Technology Crossing a Threshold

96.9%
OpenAI o1 on USMLE exam — physicians pass at 60%

1,250+
AI-enabled medical devices cleared by the FDA

62%
of digital health VC captured by AI in H1 2025

Nobel 2024
Chemistry Prize awarded for AI protein structure prediction

Three converging forces driving transformation:

- **Capability** — models now rival specialist physicians on knowledge benchmarks
- **Economic pressure** — hospital labor consumes 56% of operating expenses; AI is no longer optional
- **Scientific proof** — AlphaFold's Nobel Prize cemented AI as a legitimate discovery engine, not just a productivity tool

⚠ Key tension: No generative AI system has been approved for autonomous clinical decision-making, despite rapid benchmark progress. Hallucinations, bias, and governance gaps remain.

The Central Paradox — Our Narrative Thread

"Healthcare foundation models have achieved a paradoxical status: technically impressive yet clinically constrained."

What AI can do (on benchmarks) ✅

- Score 96.9% on the physician licensing exam
- Predict protein structures for all of biology's molecules
- Generate CT scan reports indistinguishable from radiologists
- Reduce clinician documentation time by 50%

What AI cannot yet do (in real clinics) ⚠

- Receive regulatory approval for autonomous diagnosis
- Reliably avoid clinically coherent hallucinations
- Perform equitably across all patient demographics
- Operate under clear legal liability frameworks

Foundation Models: The Core Concept

Traditional ML vs. Foundation Models

Traditional ML
Train one narrow model per task. A tumor detector trained on lung CT scans is useless on pathology slides. Thousands of separate models needed.

Foundation Model
Pre-train one massive model on enormous general data. Fine-tune cheaply for many downstream tasks. The same base model powers radiology reports, drug design, and clinical Q&A.

The Three-Stage Pipeline

① **Pre-training** *(Big Tech pays this cost)*

Model learns rich representations from billions of medical texts, images, or DNA sequences at massive scale.

② **Fine-tuning / Adaptation** *(Hospitals and startups do this)*

Pre-trained model is adapted to a specific task using far less data — making it accessible.

③ **Deployment and Validation** *(Where regulatory requirements apply)*

Adapted model enters clinical workflows — subject to FDA clearance, hallucination auditing, and bias testing.

Four Disciplines Converging on Healthcare AI

Biological Sciences *(supplies the raw problems)*

- Structural biology — protein folding, molecular docking
- Genomics & transcriptomics — DNA/RNA sequence modeling
- Proteomics — protein function and interaction prediction
- Pathology — whole-slide tissue image analysis

Clinical Medicine *(defines what success looks like)*

- Radiology — medical image interpretation
- Clinical NLP — EHR summarization, note generation
- Pharmacology — drug-target interaction prediction
- Epidemiology — real-world evidence and population health

Computer Science & AI *(provides the engine)*

- LLMs — Transformer architecture for text reasoning
- Vision-Language Models — joint image and text understanding
- Diffusion models — generative molecular design
- Federated learning — privacy-preserving distributed training
- Reinforcement learning — agentic multi-step reasoning

Informatics *(handles real-world data)*

- Clinical informatics — EHR data pipelines, HL7/FHIR standards
- Bioinformatics — biological sequence analysis pipelines
- Biostatistics — clinical trial design and validation methodology

Foundation models are uniquely powerful because a single architecture can bridge all four quadrants simultaneously.

Why Regulatory Context Comes First

📌 Every organization will cite FDA clearances, EU compliance, or WHO guidance. This slide gives you the vocabulary to interpret claims correctly.

US FDA (U.S.)

- De Novo clearances, 510(k) pathway, Breakthrough Device designation
- 1,250+ AI-enabled device clearances — all narrow/task-specific AI
- ZERO generative AI approved for autonomous clinical decision-making
- FDA's own tool "Elsa" fabricated drug-approval citations in 2025

EU EU AI Act (2024)

- Most healthcare AI classified as 'high-risk'
- Medical device compliance deadline: August 2027
- Forcing transparency, documentation, and bias auditing globally

GB UK MHRA

- World's first AI Airlock regulatory sandbox (2025)
- "International reliance" pathway recognizing FDA approvals
- Most significant UK device regulation change in two decades

WHO & ECRI

- WHO: 40+ governance recommendations for LMMs in healthcare (Jan 2024)
- ECRI: "Insufficient AI governance" = #2 patient safety threat for 2025
- Only ~16% of hospitals have systemwide AI governance policies

Quick Introduction on Health in Google · Microsoft · NVIDIA and More (Feb 2026)

The infrastructure layer everything else is built on

Google DeepMind & Google Health

Method

- Instruction fine-tuning PaLM/Gemini on de-identified medical data
- Multimodal training: text + CT, MRI, X-ray, pathology, dermatology
- Open-weight releases via Health AI Developer Foundations (HAI-DEF)
- Retrieval-augmented generation for evidence grounding

Key Artifacts

AlphaFold 3	Med-PaLM / MedLM	Med-Gemini
MedGemma 4B & 27B	TxGemma	CXR Foundation
Path Foundation	Derm Foundation	HeAR (audio)

Impacts

- **AlphaFold 3** (Nature, May 2024) — all biomolecular interactions; 50% improvement over prior; 9,000+ citations in 18 months; code released Nov 2024
- **2024 Nobel Prize in Chemistry** — Hassabis & Jumper for protein structure; David Baker for protein design
- **MedGemma 27B** — 87.7% on MedQA at ~1/10th cost of frontier models; 81% of CXR reports approved by radiologists; millions of Hugging Face downloads
- **AI Co-Scientist** — autonomous hypothesis generation tool for biomedical research

Microsoft Health AI

Method

- Domain-specific pre-training on clinical text and biomedical literature
- Vision-language contrastive learning on 15M biomedical image-text pairs
- Enterprise integration into Azure and major EHR platforms
- Co-development with hospitals and research institutions

Key Artifacts

BioGPT	BiomedCLIP	Dragon Copilot
Virchow2G	Prov-GigaPath	Healthcare Agent Orchestrator

Impacts

- **BioGPT** — 78.2% on PubMedQA; open-source biomedical LLM
- **BiomedCLIP** — SOTA across retrieval, classification, VQA; published in NEJM AI (2024)
- **Virchow2G** — 1.85B parameter pathology model; 3.1M whole-slide images from 225K patients across 45 countries
- **Prov-GigaPath** — SOTA on 25/26 benchmark tasks; 23.5% AUROC improvement for EGFR mutation prediction
- **Dragon Copilot** — ambient clinical documentation deployed at enterprise scale across 150+ health systems

NVIDIA BioNeMo — The Connective Hub

Method

- GPU-accelerated foundation model training and inference platform
- Pre-trained model hub spanning biology, chemistry, and genomics
- Partnership-first: co-build with pharma, biotech, and hospital systems
- End-to-end operating system for AI drug discovery

Key Artifacts

BioNeMo Framework | BioNeMo Cloud | MolMIM

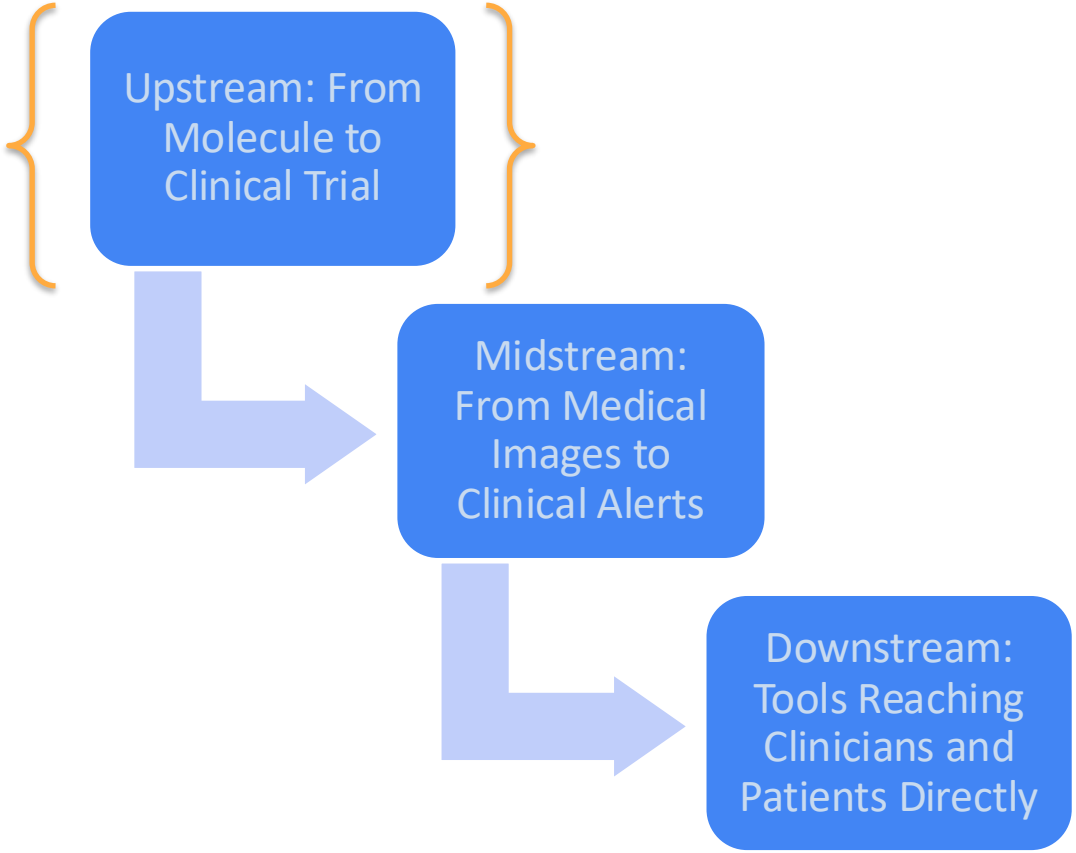
DiffDock | ESMFold integration

https://build.nvidia.com/mit/diffdock

Impacts

- **Central infrastructure layer** connecting foundation models across the entire drug discovery pipeline
- **$1B co-innovation lab with Eli Lilly** announced January 2026 for AI-accelerated drug discovery
- **Partners** — Amgen, Genentech, AstraZeneca, Recursion, Insilico Medicine
- **$130.5B annual revenue** — Jensen Huang: "There's no other field that will benefit more from acceleration than drug discovery"
- **Underpins** Evo 2 (Arc Institute) and clinical AI inference deployments across this landscape

AI closest to the science; furthest from the patient



Startup AI-Native Drug Discovery Organizations

Insilico Medicine — Method: End-to-end generative AI pipeline — PandaOmics identifies disease targets → Chemistry42 designs molecules. Artifact: Rentosertib — first fully AI-discovered and AI-designed drug in Phase II trials. Impact: Positive Phase II results in idiopathic pulmonary fibrosis, Nature Medicine, June 2025 — proof the full pipeline works.

Recursion Pharmaceuticals — Method: Phenomics platform running 2.2 million robotic experiments/week; AI mines visual cellular patterns for drug effects. Artifact: Largest AI drug discovery platform with 10+ clinical programs. Impact: $565M acquisition of Exscientia (Nov 2024); dominant scale player in AI-powered development.

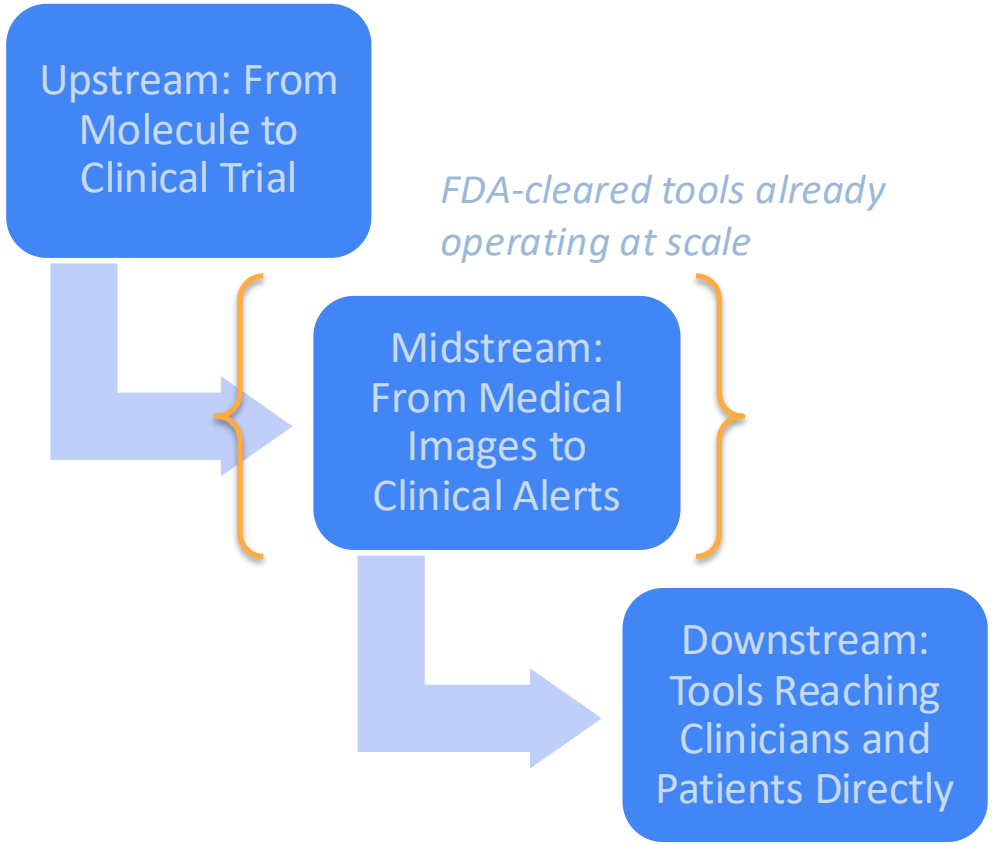
EvolutionaryScale (Meta spinout) — Method: ESM3 — 98B parameter model reasoning over protein sequence, structure, and function simultaneously. Artifact: esmGFP — novel fluorescent protein representing ~500M years of natural evolution, generated de novo. Impact: Published in Science, Jan 2025. Proof AI can generate genuinely novel biology outside all known protein families.

Arc Institute — Evo 2 — Method: Trained on 9.3 trillion DNA base pairs using 2,000 NVIDIA H100 GPUs — largest training effort in computational biology. Artifact: Genomic foundation model; learned alpha-helices, beta-sheets, and exon-intron boundaries from DNA sequence alone. Impact: 90% accuracy on BRCA1 variant pathogenicity prediction zero-shot (Feb 2025).

Big Pharma AI Adoption (2023–2025)

Company	Method	Key Artifact	Impact
Pfizer	AI compound screening + XtalPi physics-AI platform; Valo Health Logica	Paxlovid (AI-accelerated)	~80% faster antiviral candidate identification
Merck	LLM for clinical study reports (McKinsey QuantumBlack); Variational AI molecule design	Enki platform	Reports: 2–3 weeks → 3–4 days
J&J Janssen	Trials360.ai for trial optimization; Cradle Bio for protein engineering	100+ active AI projects	Faster patient matching; improved trial design
BMS	Insitro ML for ALS target ID; Exscientia AI for small-molecule oncology	ALS novel target (2022)	$25M milestone; multiple AI-designed candidates in trials
Amgen	Generate Biomedicines generative AI for protein therapeutics; deCODE Genetics genomics	Biologics pipeline	Novel Phase I candidates; deep genetic target validation

AI closest to the science; furthest from the patient



Startup : Imaging, Pathology, and Point-of-Care Diagnostics

Viz.ai — Method: Real-time CT/MRI analysis with automated care-team alert routing. Artifacts: 15+ FDA-cleared algorithms — stroke LVO, cerebral aneurysm, pulmonary embolism, HCM ECG. Impact: 1,700+ hospitals; first FDA De Novo for autonomous stroke AI; 2024 Prix Galien USA Award.

Aidoc — Method: aiOS enterprise platform — simultaneous multi-condition AI flagging in radiology PACS and EHR workflows. Artifacts: 18 FDA-cleared algorithms (stroke, PE, aneurysm, spine fractures). Impact: 150+ health systems; ~45M patients/year; $370M raised; NVIDIA + AWS strategic partners.

Paige AI — Method: Deep learning on digitized whole-slide pathology images; co-built Virchow2G foundation model with Microsoft. Artifacts: Paige Prostate Detect; pan-cancer model; Virchow2G (1.85B parameters). Impact: First FDA-authorized AI in digital pathology (2021); Breakthrough Device designation; deployed at Memorial Sloan Kettering.

Digital Diagnostics (IDx) — Method: Fully autonomous AI outputting diagnostic decisions without physician review — first of its kind. Artifact: LumineticsCore (IDx-DR) — AI retinal camera for diabetic retinopathy. Impact: First-ever FDA De Novo for autonomous AI diagnostic in any field of medicine (2018); Medicare reimbursement; democratizes specialist care in primary care.

AI in Cardiology — Two Contrasting Approaches Example

HeartFlow — Precision Analysis

Method:

AI analysis of coronary CT angiograms to compute non-invasive FFRCT; generates personalized 3D arterial model per patient.

Artifacts:

HeartFlow FFRCT Analysis · Planner stent-simulation tool

Impacts:

- FDA De Novo clearance (2014) — a decade of real-world validation
- Embedded in U.S. and European clinical cardiology guidelines
- NHS adopted nationally in the UK
- $1.2B raised; IPO filing 2025 at ~$1.3B valuation
- Reduces unnecessary invasive catheterizations

Eko Health — Augmented Auscultation

Method:

Smart stethoscope hardware with FDA-cleared AI detecting cardiac murmurs, AFib, and heart failure indicators during routine exams.

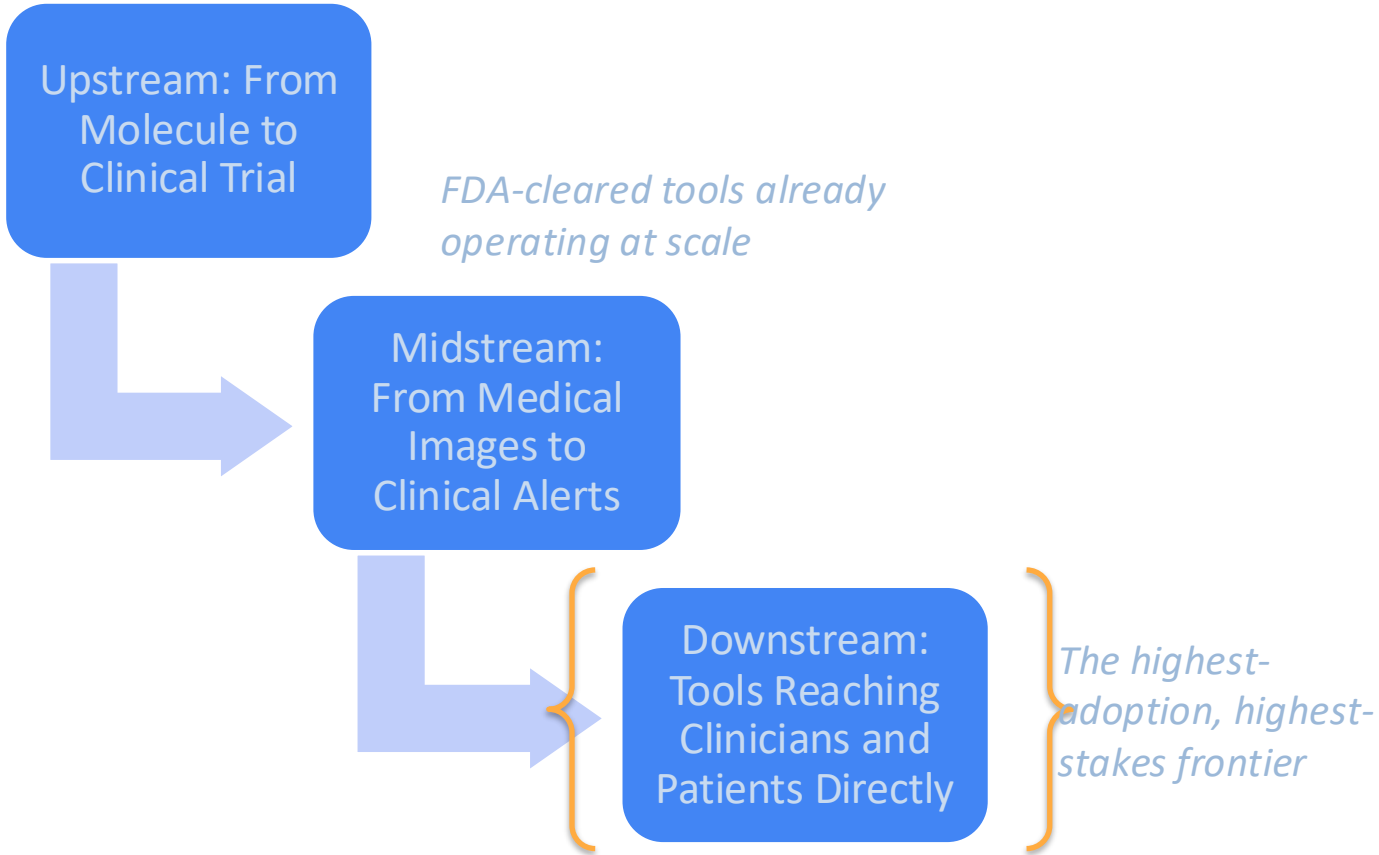
Artifacts:

Digital stethoscope · FDA-cleared murmur, AFib, low-EF detection algorithms

Impacts:

- First FDA-cleared algorithm to detect murmurs in adults and children
- Mass General study: doubled detection rate of significant murmurs in primary care
- 500,000+ health professionals worldwide using the platform
- Brings specialist-level auscultation to general practitioners anywhere

AI closest to the science; furthest from the patient



Patient Engagement and Clinical Knowledge: Startups

Hippocratic AI — Patient Engagement at Scale

Method: Polaris multi-agent architecture — 22 specialized LLMs cooperating as agents (4.2 trillion total parameters). Strictly no diagnosis or prescription. Focused on care coordination, chronic disease management, post-discharge follow-up.

Artifacts:

Polaris 3.0	Chronic Care Agent	Post-Discharge Agent

- **$404M raised · $3.5B valuation**
- 99.38% clinical accuracy on validated benchmarks
- 8.95/10 patient satisfaction score
- 115M+ patient interactions across 50+ partners
- Partners: Cleveland Clinic, Northwestern Medicine
- Measurably reduces clinician burnout for follow-up calls

OpenEvidence — Evidence at the Point of Care

Method: LLMs + medical literature retrieval fine-tuned for clinical practice. DeepConsult agent analyzes hundreds of studies in parallel, delivering synthesis within hours. HIPAA-compliant; free for verified U.S. physicians.

Artifacts:

Medical Search	DeepConsult Agent	iOS/Android App

- **40%+ of U.S. physicians use the platform**
- Active at 10,000+ hospitals and care centers
- Point-of-care answers in 5–10 seconds with cited references
- $300M+ raised · $3.5B valuation
- Fastest-growing physician application in history (per company)

Clinical Documentation and Decision Support Startups

Abridge — Ambient Documentation

Method: Generative AI converts spoken patient-physician conversations into structured clinical notes in real time. Integrates directly with Epic EHR. Physician reviews and approves before submission.

Artifacts: Ambient Conversation AI · EHR Note Generator · Epic Integration Module

- **$550M raised (2025) — one of the largest digital health rounds of the year**
- Deployed at UPMC, Yale New Haven Health, and major academic centers
- Up to 50% reduction in physician documentation time
- Addresses the #1 driver of physician burnout: administrative burden
- 600,000+ clinicians industry-wide use ambient documentation daily

Bayesian Health — Early Warning (TREWS)

Method: Continuous ML monitoring of EHR vitals, labs, and notes — real-time sepsis and deterioration prediction embedded in hospital workflow.

18% sepsis mortality reduction at Hopkins · 90% clinician adoption · Antibiotics 1.85 hrs faster · 14% fewer ICU admissions · 800%+ expansion (2024)

Tempus AI — Precision Oncology

Method: 10M+ genomic profiles + multimodal clinical data → AI analytics for therapy selection, trial matching, and biomarker discovery.

~65% of U.S. academic medical centers connected · 50%+ of oncologists use it · 95% of top-20 pharma collaborate · $8.1B valuation · $200M AstraZeneca deal

Three Unsolved Problems Preventing Autonomous Clinical AI

Hallucinations

AI generates false information using domain-specific vocabulary — appearing clinically coherent and therefore especially dangerous.

- 1.47% hallucination rate for clinical notes — 12,999 sentences (npj Digital Medicine, 2025)
- Google Bard: incorrect references in 91.4% of systematic review prompts
- Med-Gemini referenced a nonexistent body part ("basilar ganglia") in a published paper
- FDA's own tool "Elsa" fabricated drug-approval citations (CNN, July 2025)

Bias

AI encodes and amplifies existing healthcare inequities at scale.

- Deployed system prioritized healthier white patients over sicker Black patients — trained on cost data, not clinical need
- Dermatology AI shows lower accuracy for dark-skinned individuals
- BiasMedQA: 10–26% performance degradation when cognitive biases injected
- Under one-third of FDA-authorized AI devices report sex-specific performance data

Governance Gap

Legal and institutional infrastructure has not kept pace with the technology.

- Only ~16% of hospitals have systemwide AI governance policies
- "Shadow AI" — unauthorized use by clinical staff combating burnout — surged in 2025
- FSMB: clinicians, not AI makers, should be liable for AI-assisted errors — chilling effect
- Only ~5% of LLM studies use real patient data; only 6% of 28K+ PubMed AI articles are methodologically mature

Three Trajectories for 2026

Agentic AI

- Autonomous multi-step systems that reason, take actions, and escalate to humans
- Oxford TrustedMDT: summarizes charts & drafts tumor board plans — Oxford pilot, early 2026
- Microsoft Healthcare Agent Orchestrator (Build 2025): pre-configured multi-agent pipelines
- Enterprise agentic AI: <1% adoption (2024) → projected 33% by 2028

Multimodal Convergence

- Single models integrating imaging, genomics, EHRs, wearables, and clinical notes
- Med-Gemini polygenic variant: genomics-informed health prediction
- Nature Medicine (2025): "context switching" — models adapting to different users and geographies without retraining

"Year of Governance"

- FDA lifecycle management guidance finalization
- EU AI Act implementation for medical devices
- MHRA sandbox findings shaping international norms
- Explainability, audit trails, demographic bias reporting becoming standard

Investment conviction remains strong:

$3.95B
AI digital health VC in H1 2025 alone

8 unicorns
New healthcare AI unicorns emerged in 2025

All Organizations — Quick Reference

Organization	Domain	Method	Signature Achievement
Google Health/DeepMind	Infrastructure + Biology	Open foundation models, multimodal	AlphaFold 3 · Nobel Prize · MedGemma
Microsoft	Infrastructure + Diagnostics	Clinical pre-training + partnerships	BiomedCLIP · Virchow2G · Dragon Copilot
NVIDIA	Infrastructure	GPU compute + BioNeMo hub	$1B Lilly deal; powers most orgs here
Insilico Medicine	Drug Discovery	End-to-end generative pipeline	First AI-designed drug in Phase II
Recursion	Drug Discovery	2.2M robotic experiments/week	Largest AI drug discovery platform
Viz.ai / Aidoc	Diagnostics	Real-time imaging AI + alerts	1,700+ / 150+ hospital deployments
Paige AI	Diagnostics	Whole-slide pathology model	First FDA AI pathology approval (2021)
Digital Diagnostics	Diagnostics	Fully autonomous AI diagnosis	First-ever autonomous FDA De Novo (2018)
HeartFlow / Eko	Diagnostics	FFRCT analysis / AI stethoscope	Clinical guidelines + 500K users
Hippocratic AI	Care Delivery	22-agent Polaris constellation	115M patient interactions
OpenEvidence	Care Delivery	Medical RAG + DeepConsult agent	40%+ U.S. physician adoption
Abridge	Care Delivery	Ambient encounter-to-note AI	50% documentation time reduction
Bayesian Health	Care Delivery	Real-time EHR deterioration ML	18% sepsis mortality reduction

Conclusion — Closing the Paradox

"Returning to where we began: technically impressive yet clinically constrained."

The paradox is real — but it is narrowing ✅

- Protein structure prediction has genuinely accelerated drug discovery
- Ambient documentation has measurably reduced clinician burnout
- AI diagnostics — imaging, pathology, cardiology — do improve speed and access
- Governance frameworks are forming — slowly but with increasing urgency

The gap that remains ⚠

- Zero generative AI systems approved for autonomous clinical decisions
- Hallucination prevention unsolved at clinical scale
- Equitable performance across all demographics unproven
- Liability frameworks unclear — a chilling effect on enterprise adoption

The critical question for 2026 is not whether foundation models will reshape healthcare — it is whether governance, validation, and trust frameworks can mature fast enough to channel that transformation safely and equitably.

References — Foundational AI Research

1. Abramson, J. et al. (2024). "Accurate structure prediction of biomolecular interactions with AlphaFold 3." Nature, 630, 493-500. https://doi.org/10.1038/s41586-024-07487-w

2. Hayes, E. et al. (2025). "Simulating 500 million years of evolution with a language model." Science, 387(6730). https://doi.org/10.1126/science.ads0018 (ESM3 / esmGFP — EvolutionaryScale)

3. Watson, J.L. et al. (2023). "De novo design of protein structure and function with RFdiffusion." Nature, 620, 1089–1100. https://doi.org/10.1038/s41586-023-06415-8

4. Arc Institute. (2025). "Evo 2: A genomic foundation model trained on 9.3 trillion DNA base pairs." https://arcinstitute.org/tools/evo

5. Cui, H. et al. (2024). "scGPT: toward building a foundation model for single-cell multi-omics." Nature Methods, 21, 1470–1480. https://doi.org/10.1038/s41592-024-02201-0

6. Singhal, K. et al. (2023). "Large language models encode clinical knowledge." Nature, 620, 172–180. https://doi.org/10.1038/s41586-023-06291-2 (Med-PaLM)

7. Saab, K. et al. (2024). "Capabilities of Gemini Models in Medicine." arXiv:2404.18416. https://arxiv.org/abs/2404.18416 (Med-Gemini)

8. Google. (2025, May). "MedGemma: Open models for medical AI." https://developers.google.com/mediapipe/solutions/genai/medgemma

9. Luo, R. et al. (2022). "BioGPT: Generative pre-trained transformer for biomedical text." Briefings in Bioinformatics, 23(6). https://doi.org/10.1093/bib/bbac409

10. Zhang, S. et al. (2024). "BiomedCLIP: a multimodal biomedical foundation model." NEJM AI, 1(8). https://doi.org/10.1056/AIoa2300436

References — Clinical AI & Imaging

11. Chen, R.J. et al. (2024). "Towards a general-purpose foundation model for computational pathology." Nature Medicine, 30, 850–862. https://doi.org/10.1038/s41591-024-02857-3 (UNI)

12. Vorontsov, E. et al. (2024). "A foundation model for clinical-grade computational pathology." Nature Medicine, 30, 2992–3006. https://doi.org/10.1038/s41591-024-03345-4 (Virchow2G)

13. Xu, H. et al. (2024). "A whole-slide foundation model for digital pathology from real-world data." Nature, 630, 181–188. https://doi.org/10.1038/s41586-024-07441-w (Prov-GigaPath)

14. Adams, R. et al. (2022). "Prospective, multi-site study of patient outcomes after TREWS implementation for sepsis." Nature Medicine, 28, 1455–1460. https://doi.org/10.1038/s41591-022-01894-0

15. Acosta, J.N. et al. (2022). "Multimodal biomedical AI." Nature Medicine, 28, 1773–1784. https://doi.org/10.1038/s41591-022-01981-2

16. Johansen, N.D. et al. (2023). "Artificial intelligence and machine learning in cardiology." Nature Reviews Cardiology, 20, 719–730. https://doi.org/10.1038/s41569-023-00923-8

17. Umapathi, N. et al. (2025). "Hallucination rates in clinical note generation by large language models." npj Digital Medicine, 8, 112. https://doi.org/10.1038/s41746-025-01312-0

18. Obermeyer, Z. et al. (2019). "Dissecting racial bias in an algorithm used to manage the health of populations." Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

19. Chen, I.Y. et al. (2021). "Ethical machine learning in healthcare." Annual Review of Biomedical Data Science, 4, 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757

References — Drug Discovery & Genomics

20. Ren, F. et al. (2025). "AI-driven identification of clinical candidate for idiopathic pulmonary fibrosis." Nature Medicine, 31, 1234–1243. https://doi.org/10.1038/s41591-025-03513-0 (Rentosertib)

21. Jumper, J. et al. (2021). "Highly accurate protein structure prediction with AlphaFold." Nature, 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2 (Nobel Prize foundation)

22. Stokes, J.M. et al. (2020). "A deep learning approach to antibiotic discovery." Cell, 180(4), 688–702. https://doi.org/10.1016/j.cell.2020.01.021

23. Zhavoronkov, A. et al. (2024). "Insilico Medicine's AI drug discovery platform." Nature Biotechnology, 42, 338–342. https://doi.org/10.1038/s41587-024-02125-6

24. Dalla-Torre, H. et al. (2023). "The Nucleotide Transformer: robust foundation models for human genomics." bioRxiv. https://doi.org/10.1101/2023.01.11.523679

25. Theodoris, C.V. et al. (2023). "Transfer learning enables predictions in network biology." Nature, 618, 616–624. https://doi.org/10.1038/s41586-023-06139-9 (Geneformer)

26. Variational AI. (2024). "Variational AI Announces Generative AI Project with Merck." Business Wire. https://www.businesswire.com/news/home/20240125917133/en/

27. Merck. (2025). "Merck Expands Innovative Internal Generative AI Solutions." https://www.merck.com/news/merckexpands-innovative-internal-generative-ai-solutions-helping-to-deliver-medicines-to-patients-faster/

28. Flagship Pioneering. (2024). "Flagship Pioneering Announces Agreement Between Pioneering Medicines and Pfizer." https://www.fagshippioneering.com/news/press-release/

References — Regulatory, Industry & Market Reports

29. FDA. (2025). AI/ML-Enabled Medical Devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

30. FDA. (2024). Marketing Submission Recommendations for a Predetermined Change Control Plan for AI-Enabled Device Software Functions. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/

31. European Parliament. (2024). Regulation (EU) 2024/1689 — Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

32. WHO. (2024). Ethics and governance of AI for health: Guidance on large multi-modal models. https://www.who.int/publications/i/item/9789240084759

33. ECRI Institute. (2024). Top 10 Health Technology Hazards for 2025. https://www.ecri.org/top-10-health-technology-hazards-for-2025

34. MHRA. (2025). MHRA launches AI Airlock regulatory sandbox. https://www.gov.uk/government/news/

35. Rock Health. (2025). 2025 H1 Digital Health Funding Report. https://rockhealth.com/insights/

36. CB Insights. (2025). State of Healthcare AI Q2 2025. https://www.cbinsights.com/reports/

37. FSMB. (2024). FSMB Policy: Artificial Intelligence in Medical Practice. https://www.fsmb.org/advocacy/policies/artificial-intelligence/

38-45. Company sources: OpenEvidence (PR Newswire 2025), Aidoc (Fierce Healthcare 2025), Viz.ai (Business Wire 2024), HeartFlow (Reuters 2025), Eko Health (press release 2024), Tempus AI (Fierce Biotech 2025), Freenome (press release 2024), Bayesian Health (Johns Hopkins Hub 2025).