

CHAPTER 5:

Science and Medicine



Chapter 5: Science and Medicine

Overview	3	Diagnostic Reasoning With LLMs	25
Chapter Highlights	4	Highlight: LLMs Influence Diagnostic Reasoning	25
5.1 Notable Medical and Biological AI Milestones	6	Management Reasoning and Patient Care Decisions	25
Protein Sequence Optimization	6	Highlight: GPT-4 Assistance on Patient Care Tasks	26
Aviary	7	Ambient AI Scribes	27
AlphaProteo	8	Deployment, Implementation, Deimplementation	29
Human Brain Mapping	8	FDA Authorization of AI-Enabled Medical Devices	29
Virtual AI Lab	9	Successful Use Cases: Stanford Health Care	29
GluFormer	10	Screening for Peripheral Arterial Disease	30
Evolutionary Scale Modeling v3 (ESM3)	10	Social Determinants of Health	31
AlphaFold 3	11	Extracting SDoH From EHR and Clinical Notes	31
5.2 The Central Dogma	12	AI Adoption Across Medical Fields and the Integration of SDoH	32
Protein Sequence Analysis	12	Synthetic Data	32
AI-Driven Protein Sequence Models	12	Clinical Risk Prediction	32
Public Databases for Protein Science	14	Drug Discovery	33
Research and Publication Trends	15	Data Generation Platforms	33
AI-Driven Protein Science Publications	15	Electronic Health Record System	34
Image and Multimodal AI for Scientific Discovery	16	Clinical Decision Support	36
5.3 Clinical Care, Imaging	17	5.5 Ethical Considerations	38
Data: Sources, Types, and Needs	17	Meta Review	38
Advanced Modeling Approaches	19	5.6 AI in Physics, Chemistry, and Other Scientific Domains	41
5.4 Clinical Care, Non-Imaging	21	Highlight: Notable Model Releases	41
Clinical Knowledge	21	Appendix	44
MedQA	21		
Highlight: AI Doctors and Cost-Efficiency Considerations	22		
Evaluation of LLMs for Healthcare Performance	23		
Overview	23		

ACCESS THE PUBLIC DATA

CHAPTER 5: Science and Medicine

Overview

This chapter explores key trends in AI-driven science and medicine, reflecting the technology's growing impact in these fields. It begins with notable AI milestones from 2024, followed by an analysis of AI in protein folding, an important area of scientific advancement. The chapter then examines AI's role in clinical care, spanning both imaging and non-imaging applications. This includes a review of clinical knowledge capabilities in new language models, diagnostic and clinical management capabilities of AI systems, real-world AI deployments in medicine, synthetic data applications, and social determinants of health. Finally, the chapter concludes with an exploration of ethical trends in AI medical research.

This chapter was prepared by [RAISE Health](#) (Responsible AI for Safe and Equitable Health), a collaboration between Stanford Medicine and the Stanford Institute for Human-Centered Artificial Intelligence (HAI). Since its launch in 2023, RAISE Health has worked to advance responsible AI innovation in biomedical research, education, and patient care, with a focus on ensuring that these technologies benefit everyone.

Fostering collaborative research and knowledge sharing are central to RAISE Health's mission. As part of that commitment, RAISE Health partnered with the AI Index Steering Committee to expand the group's focus to include key developments in science and medicine. In 2024, this collaboration produced the inaugural chapter on science and medicine, highlighting major AI advancements at Stanford and beyond. The 2025 chapter builds on that foundation with contributions from members of the RAISE Health [faculty research council](#), Stanford School of Medicine faculty, postdoctoral fellows, and undergraduate students from the schools of Medicine and Engineering.

CHAPTER 5:

Science and Medicine

Chapter Highlights

1. Bigger and better protein sequencing models emerge. In 2024, several large-scale, high-performance protein sequencing models, including ESM3 and AlphaFold 3, were launched. Over time, these models have grown significantly in size, leading to continuous improvements in protein prediction accuracy.

2. AI continues to drive rapid advances in scientific discovery. AI's role in scientific progress continues to expand. While 2022 and 2023 marked the early stages of AI-driven breakthroughs, 2024 brought even greater advancements, including Aviary, which trains LLM agents for biological tasks, and FireSat, which significantly enhances wildfire prediction.

3. The clinical knowledge of leading LLMs continues to improve. OpenAI's recently released o1 set a new state-of-the-art 96.0% on the MedQA benchmark—a 5.8 percentage point gain over the best score posted in 2023. Since late 2022, performance has improved 28.4 percentage points. MedQA, a key benchmark for assessing clinical knowledge, may be approaching saturation, signaling the need for more challenging evaluations.

4. AI outperforms doctors on key clinical tasks. A new study found that GPT-4 alone outperformed doctors—both with and without AI—in diagnosing complex clinical cases. Other recent studies show AI surpassing doctors in cancer detection and identifying high-mortality-risk patients. However, some early research suggests that AI-doctor collaboration yields the best results, making it a fruitful area of further research.

5. The number of FDA-approved, AI-enabled medical devices skyrockets. The FDA authorized its first AI-enabled medical device in 1995. By 2015, only six such devices had been approved, but the number spiked to 223 by 2023.

6. Synthetic data shows significant promise in medicine. Studies released in 2024 suggest that AI-generated synthetic data can help models better identify social determinants of health, enhance privacy-preserving clinical risk prediction, and facilitate the discovery of new drug compounds.

CHAPTER 5:

Science and Medicine

Chapter Highlights (cont'd)

7. Medical AI ethics publications are increasing year over year. The number of publications on ethics in medical AI quadrupled from 2020 to 2024, rising from 288 in 2020 to 1,031 in 2024.

8. Foundation models come to medicine. In 2024, a wave of large-scale medical foundation models were released, ranging from general-purpose multimodal models like Med-Gemini to specialized models such as EchoCLIP for echocardiology and ChexAgent for radiology.

9. Publicly available protein databases grow in size. Since 2021, the number of entries in major public protein science databases has grown significantly, including UniProt (31%), PDB (23%), and AlphaFold (585%). This expansion has important implications for scientific discovery.

10. AI research wins two Nobel Prizes. In 2024, AI-driven research received top honors, with two Nobel Prizes awarded for AI-related breakthroughs. Google DeepMind's Demis Hassabis and John Jumper won the Nobel Prize in Chemistry for their pioneering work on protein folding with AlphaFold. Meanwhile, John Hopfield and Geoffrey Hinton received the Nobel Prize in Physics for their foundational contributions to neural networks.

Chapter 5: Science and Medicine

5.1 Notable Medical and Biological AI Milestones

This section highlights significant AI-related medical and biological breakthroughs in 2024 as chosen by the RAISE Health AI Index Workgroup and AI Index Steering Committee.

5.1 Notable Medical and Biological AI Milestones

Protein Sequence Optimization

LLMs optimize protein sequence optimization

LLMs have recently, albeit unintentionally, gained a new biological capability: optimizing protein sequences. Traditionally, protein engineering requires extensive lab studies to refine sequences for improved functionality. However, a [recent study](#) found that LLMs—without fine-tuning—are becoming remarkably effective at this task. In other words, this is a hidden strength of existing LLMs, exemplified in this case by an adapted version of Llama-3.1-8B-Instruct. Using a directed evolutionary approach, researchers demonstrated that LLMs can generate protein sequences that outperform conventional algorithms across both synthetic and experimental fitness landscapes.

Figure 5.1.1 illustrates the researchers' findings. The objective in this case is to maximize the fitness value, with higher scores indicating better performance. The researchers compared their proposed method's fitness score against that of the default evolutionary algorithm (EA) approach.¹ The study revealed that this optimization extends beyond single-objective tasks to include constrained, budget-limited, and multiobjective scenarios. This compelling finding highlights the emergent properties of state-of-the-art LLMs, suggesting that as these general-purpose models continue to improve, their impact on scientific fields will only grow.

Single-objective optimization results for fitness optimization

Source: [Wang et al., 2024](#)

Dataset	Method	Population × iteration	Fitness score		
			Top 1	Top 10	Top 50
GB1	EA	32×4	5.38±1.77	3.81±1.10	2.31±0.71
		48×4	4.88±0.33	3.72±0.38	2.17±0.27
		96×4	5.72±0.56	4.32±0.53	2.84±0.60
	Ours	32×4	4.34±0.53	3.22±0.23	1.94±0.28
		48×4	4.31±0.82	3.76±0.82	2.45±0.61
		96×4	4.80±0.52	4.09±0.19	3.04±0.19
TrpB	EA	32×4	0.20±0.18	0.14±0.12	0.07±0.05
		48×4	0.67±0.14	0.52±0.11	0.19±0.04
		96×4	0.74±0.01	0.59±0.03	0.35±0.10
	Ours	32×4	0.60±0.10	0.50±0.07	0.35±0.07
		48×4	0.68±0.04	0.58±0.01	0.36±0.01
		96×4	0.78±0.20	0.60±0.16	0.39±0.16
Syn-3bfo	EA	32×8	0.57±0.21	-0.44±0.11	-1.35±0.17
		48×8	1.29±0.36	0.42±0.24	-0.63±0.07
		96×8	1.85±0.47	1.10±0.28	0.07±0.28
	Ours	32×8	2.51±0.23	1.33±0.14	0.28±0.20
		48×8	2.35±0.26	1.36±0.11	0.04±0.09
		96×8	2.83±0.20	2.02±0.36	0.96±0.36
AAV	EA	32×8	0.42±0.03	0.36±0.01	0.32±0.00
		48×8	0.44±0.00	0.38±0.01	0.33±0.00
		96×8	0.44±0.00	0.40±0.01	0.36±0.00
	Ours	32×8	0.74±0.00	0.69±0.02	0.62±0.03
		48×8	0.75±0.01	0.71±0.01	0.64±0.02
		96×8	0.76±0.03	0.73±0.03	0.68±0.03
GFP	EA	32×8	0.43±0.13	0.21±0.02	0.12±0.01
		48×8	0.43±0.14	0.26±0.05	0.12±0.01
		96×8	0.50±0.11	0.34±0.05	0.18±0.01
	Ours	32×8	0.96±0.02	0.94±0.01	0.88±0.03
		48×8	0.96±0.02	0.93±0.01	0.84±0.02
		96×8	0.97±0.01	0.95±0.01	0.92±0.01

Figure 5.1.1

¹ Evolutionary algorithms (EA) simulate key aspects of biological evolution within a computer program to tackle complex problems—especially those without precise or fully satisfactory solutions—by finding approximate answers.

Chapter 5: Science and Medicine

5.1 Notable Medical and Biological AI Milestones

Aviary

Training LLM agents for biological tasks

As AI systems become increasingly useful, particularly for scientific use cases, one challenge has been designing language models that can interact with tools as they reason through complex tasks. *Aviary* introduces a structured framework for training language agents for three particularly challenging scientific tasks: DNA manipulation (for molecular cloning), answering research questions (through accessing scientific papers), and engineering protein stability. Figure 5.1.2 compares the performance of different models across various *Aviary* environments. It contrasts a baseline Claude 3.5

Sonnet model, which attempts tasks without environmental access, with models integrated into agent frameworks within the *Aviary* environment. Across nearly all tasks, the agentic models outperform the baseline. This research demonstrates that (1) although general-purpose LLMs perform well at many scientific tasks, fine-tuning models alongside domain experts often helps models yield superior results, and (2) AI-driven scientific research can be accelerated not only by model size but also through interaction with external tools, capabilities now commonly referred to as “agentic AI.”

Performance of LLMs and language agents to solve tasks using *Aviary* environments

Source: Narayanan et al., 2024 | Chart: 2025 AI Index report

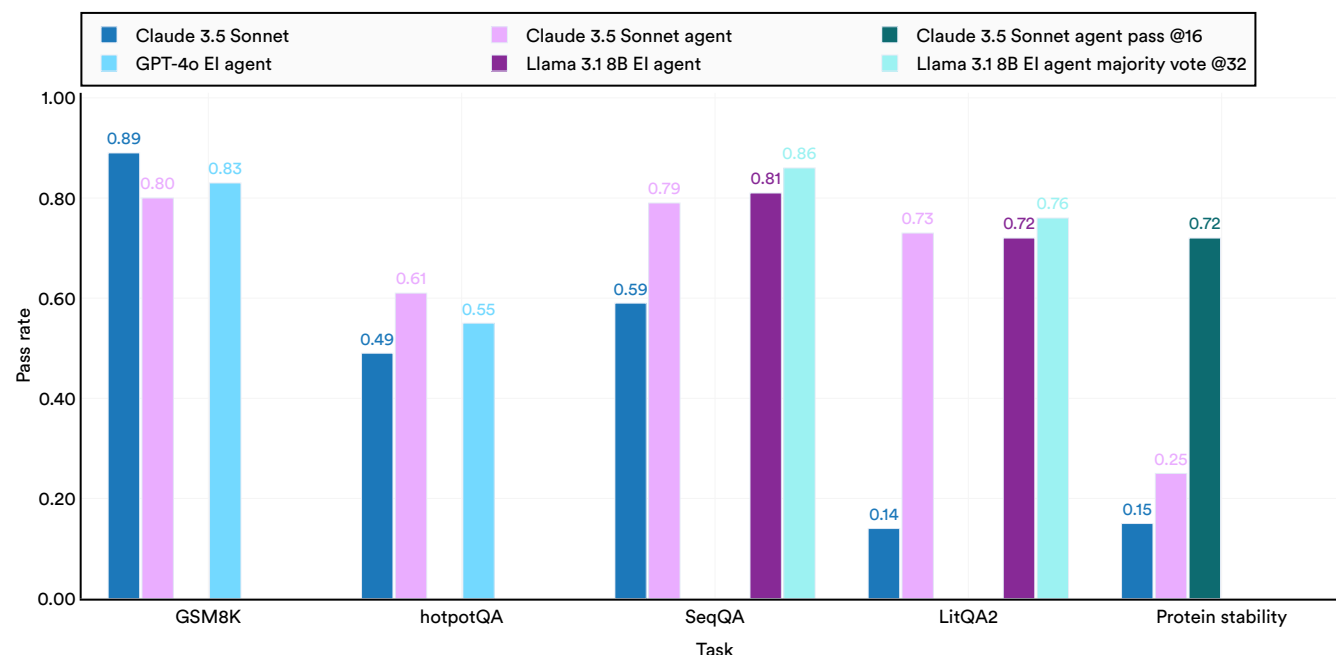


Figure 5.1.2

Chapter 5: Science and Medicine

5.1 Notable Medical and Biological AI Milestones

AlphaProteo

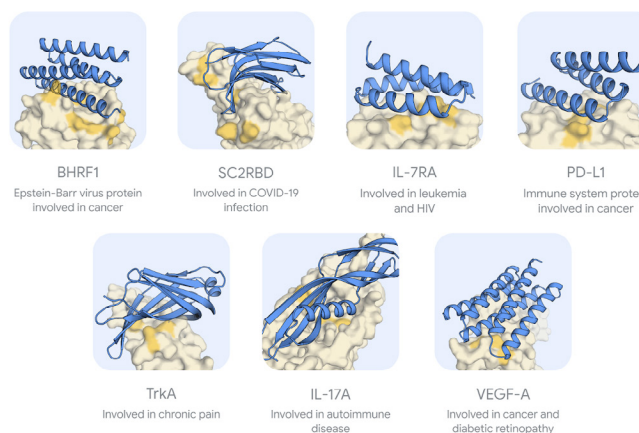
AI for novel, high-affinity protein binders

AlphaProteo is Google DeepMind's model focused on creating novel, high-affinity protein binders that attach to specific target molecules. Figure 5.1.3 illustrates the predicted structures of seven target proteins for which AlphaProteo created successful binders. AlphaProteo has designed the first protein binders for many targets, including VEGF-A, a protein linked to cancer and diabetes. Many of the tool's binding strengths are significantly better than current state-of-the-art solutions; in fact, the team estimates that some of their binders are up to 300 times more effective than anything currently available on the seven target proteins they tested. For the viral protein BHRF1, 88% of their designed binders successfully bound when tested in DeepMind's wet lab. Based on the tested targets, AlphaProteo binders hold together roughly 10 times more strongly than those created using existing state-of-the-art design methods, making it a true bioengineering breakthrough. The model is being used for drug development, diagnostics, and biotech applications.

AlphaProteo generating successful binders

Source: [Google DeepMind, 2024](#)

Figure 5.1.3



Human Brain Mapping

Synaptically reconstructing a small piece of the human brain

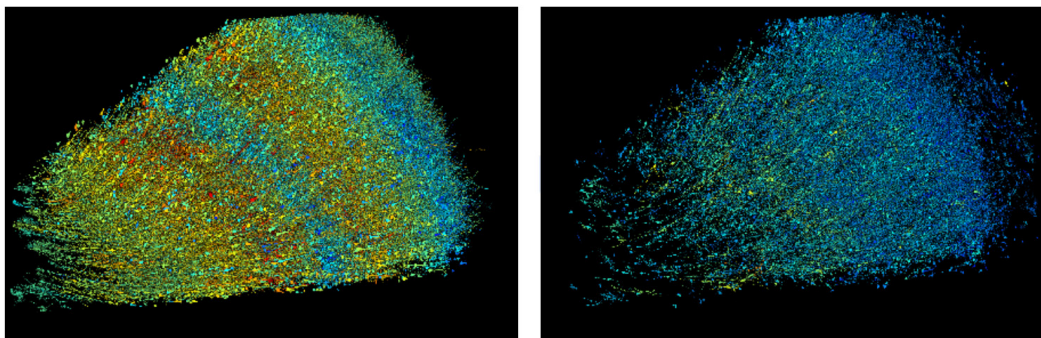
A team at Google's Connectomics project has reconstructed a one-cubic-millimeter section of the human brain at the synaptic level—hailed by Wired as “the most detailed map of brain connections ever made.” The sample, taken from an epileptic patient's left anterior temporal lobe during surgery, was imaged with a multibeam scanning electron microscope. Over 5,000 ultra-thin slices (30 nanometers each) captured around 57,000 cells—including neurons, glial cells, and blood vessels—along with 150 million synapses. Figure 5.1.4

visualizes the results: excitatory neurons on the left, inhibitory neurons on the right. To process this massive dataset, the team developed machine learning tools like flood-filling networks (for neuron reconstruction without manual tracing), SegCLR (for cell type identification), and TensorStore (for managing the multidimensional dataset). The dataset is publicly available via Neuroglancer, a web-based exploration tool; and CAVE, a Neuroglancer extension for annotation refinement. This project marks a major step in understanding neural circuitry and could inform future neurological treatments.

3D brain map images

Source: [Google Research, 2024](#)

Figure 5.1.4



Chapter 5: Science and Medicine

5.1 Notable Medical and Biological AI Milestones

Virtual AI Lab

Virtual AI lab supercharges biomedical research

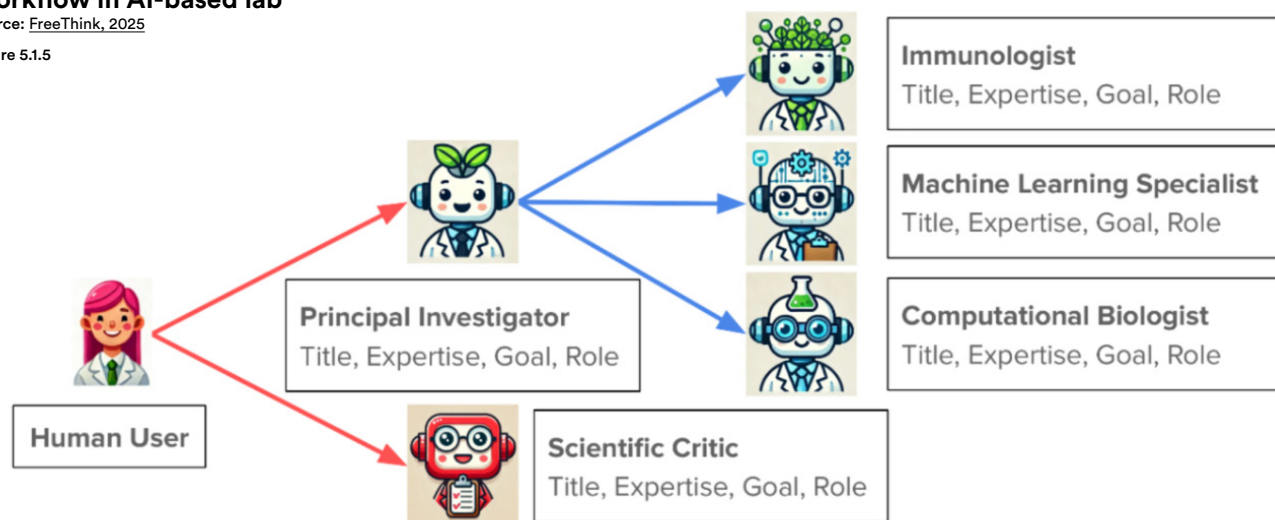
AI's role in science is shifting from a passive tool to an active collaborator. A [recent Stanford study](#) introduced a virtual AI laboratory, where multiple AI-powered scientists (technically LLMs) specialize in different disciplines and autonomously collaborate as agents. In one experiment, human researchers tasked this AI-driven lab with designing nanobodies—antibody fragments—capable of binding to SARS-CoV-2, the virus that causes COVID-19. The lab generated 92 nanobodies, with over 90% successfully binding to the virus

in validation studies. The virtual lab was structured similar to a computational biology lab, comprising a principal investigator (PI), a scientific critic AI, and three discipline-specific scientists specializing in immunology, computational biology, and machine learning (Figure 5.1.5). The PI model created these expert scientists and guided their research. Tools like AlphaFold and Rosetta were used for protein design, but the real significance of this study lies not in its specific findings, but in demonstrating that an entirely autonomous, LLM-powered lab can generate meaningful scientific discoveries.

Workflow in AI-based lab

Source: [FreeThink, 2025](#)

Figure 5.1.5



Chapter 5: Science and Medicine

5.1 Notable Medical and Biological AI Milestones

GluFormer

Continuous glucose monitoring with AI

GluFormer, a foundation model developed by Nvidia Tel Aviv, the Weizmann Institute, and others, analyzes continuous glucose monitoring (CGM) data to predict long-term health outcomes. Trained on over 10 million glucose measurements from nearly 11,000 individuals—most without diabetes—it forecasts health trajectories up to four years in advance. For instance, GluFormer can identify individuals at risk of developing diabetes or worsening glycemic control long before symptoms appear. In a 12-year study of 580 adults, it accurately flagged 66% of new-onset diabetes cases and 69% of cardiovascular-related deaths within their respective top-risk quartiles. The model's results have also generalized across 19 external cohorts (n=6,044) in five countries and diverse health conditions. GluFormer often outperforms standard CGM-based metrics like the glucose management indicator (GMI) (Figure 5.1.6). In the near and long term, models like GluFormer will shift diabetes care from reactive treatment to proactive prevention, enabling earlier clinical intervention.

Evolutionary Scale Modeling v3 (ESM3)

Simulating evolutionary processes to generate novel proteins

EvolutionaryScale's ESM3 is a groundbreaking model designed to generate novel proteins by simulating evolutionary processes. The model was trained on 2.78 billion protein sequences, and hosts 98 billion parameters. Like many other AI models, it is available in three sizes (small, medium, and large) and is available both via API and their partners' platforms. Perhaps ESM3's most notable achievement is designing esmGFP, a new artificial green fluorescent protein which the company estimates would take nature 500 million years to develop. This was done through human-led chain-of-thought prompting. Figure 5.1.7 illustrates the performance of various ESM3 models in generating proteins that satisfy atomic coordination prompts. The results show that larger ESM3 models solve twice as many tasks. ESM3 is also open-sourced, promoting collaboration in synthetic biology and protein engineering projects which hope to use code and data from the project—with applications in drug discovery, materials science, and environmental engineering.

GluFormer versus glucose management indicator

Source: Lutsker et al., 2024

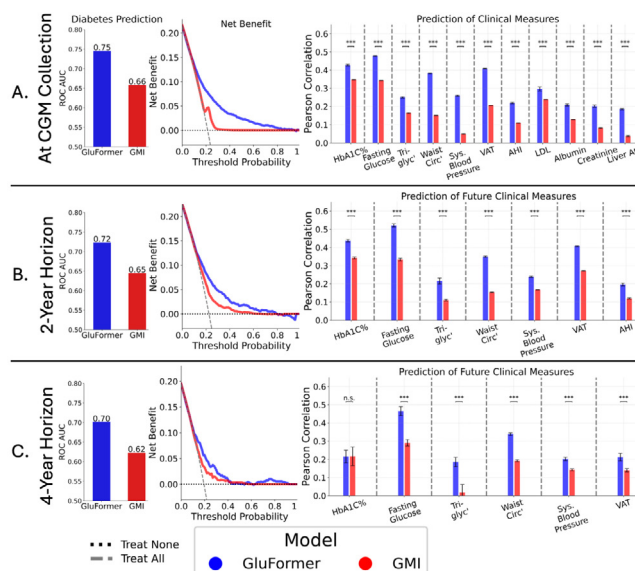


Figure 5.1.6

ESM3 models evaluated on protein generation from atomic coordination prompts

Source: ESM3, 2024 | Chart: 2025 AI Index report

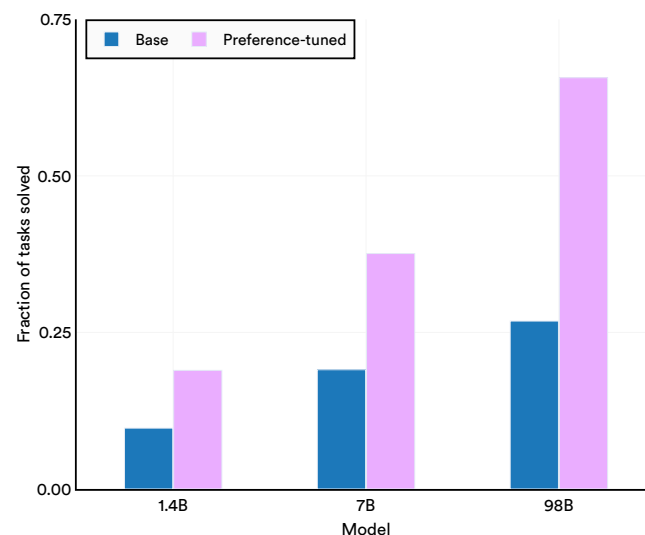


Figure 5.1.7

Chapter 5: Science and Medicine

5.1 Notable Medical and Biological AI Milestones

AlphaFold 3

Predicting the structure and interactions of all of life's molecules

Google and Isomorphic Lab's latest in the AlphaFold series, [AlphaFold 3](#), goes beyond predicting protein structures to more accurately modeling their interactions with key biomolecules (DNA, RNA, ligands, antibodies). Figure 5.1.8 [compares](#) AlphaFold 3's accuracy in predicting protein-ligand interactions against other top docking tools (e.g., [Vina](#) and [Gnina](#)) based on the percentage of predictions with a root mean square deviation (RMSD) below 2 Å, an

important measure of docking accuracy.^{2 3} AlphaFold 3 is competitive with previous state-of-the-art methods and particularly effective when the binding pocket is predefined, meaning that the docking algorithm is given prior knowledge about the specific region on the protein where the small molecule (ligand) is expected to bind. AlphaFold 3 can accelerate drug development by modeling small molecule-protein interactions, which is important for disease research. Moreover, AlphaFold 3's open-source access empowers scientists globally.

AlphaFold 3 vs. baselines for protein-ligand docking

Source: ESM3, 2024 | Chart: 2025 AI Index report

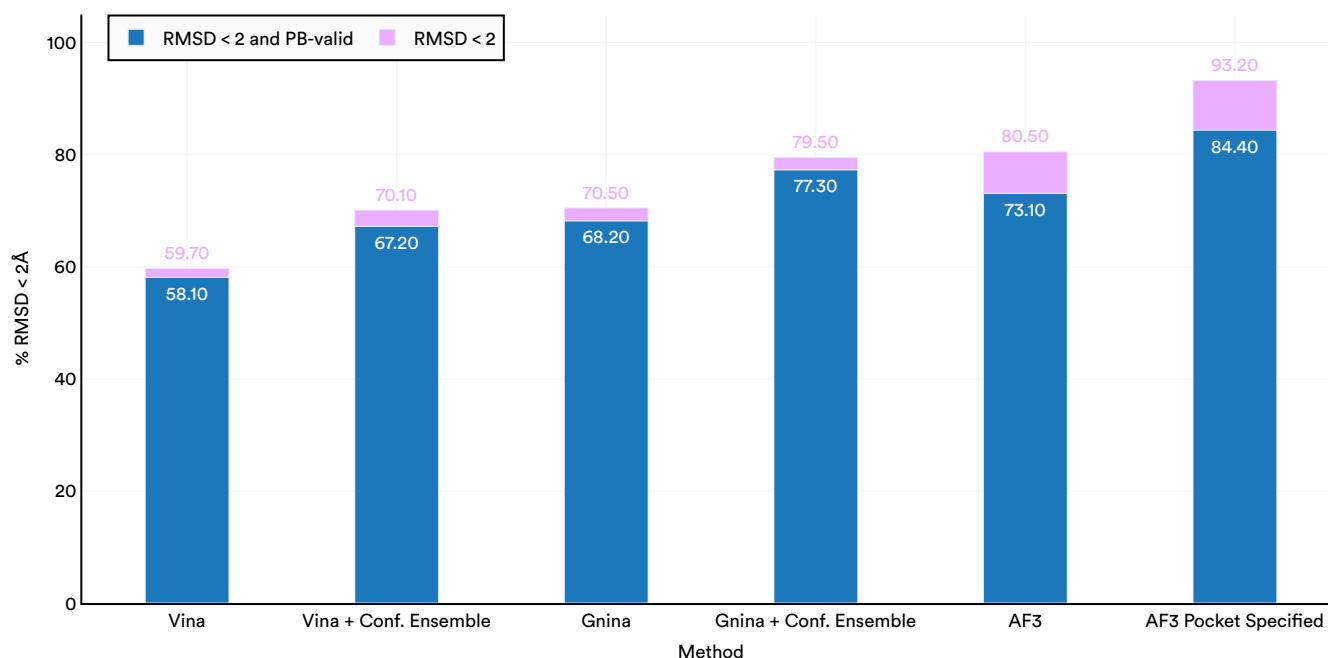


Figure 5.1.8

² A docking tool, like Vina, is a computational program used in molecular docking—a process that predicts how small molecules (such as drugs) interact with target proteins. These tools help scientists model and visualize how a molecule might bind to a protein's active site, which is crucial in drug discovery.

³ The chart uses two shades of bars to represent different accuracy criteria in molecular docking predictions. The lighter bars indicate the percentage of docking results with a root mean square deviation (RMSD) below 2 Å, meaning the predicted pose is structurally accurate. The darker bars apply a stricter criterion, showing the proportion of predictions that are not only within 2 Å RMSD but also correctly positioned within the binding pocket (PB-valid). This distinction highlights the difference between general docking accuracy and more precise, biologically relevant binding predictions.

Chapter 5: Science and Medicine

5.2 The Central Dogma

AI has transformed numerous scientific fields, with protein science being one of the most impacted areas. Understanding protein sequences is fundamental to biology, influencing drug discovery, synthetic biology, and disease research. Recent AI advancements have enabled scientists to analyze and predict protein functions, structures, and interactions with unprecedented accuracy. As the field evolves, these developments will affect healthcare, biotechnology, and regulatory frameworks. This section highlights key advancements in AI-driven protein analysis over the past year, focusing on public databases, research trends, and emerging policy considerations.

5.2 The Central Dogma

Protein Sequence Analysis

AI-Driven Protein Sequence Models

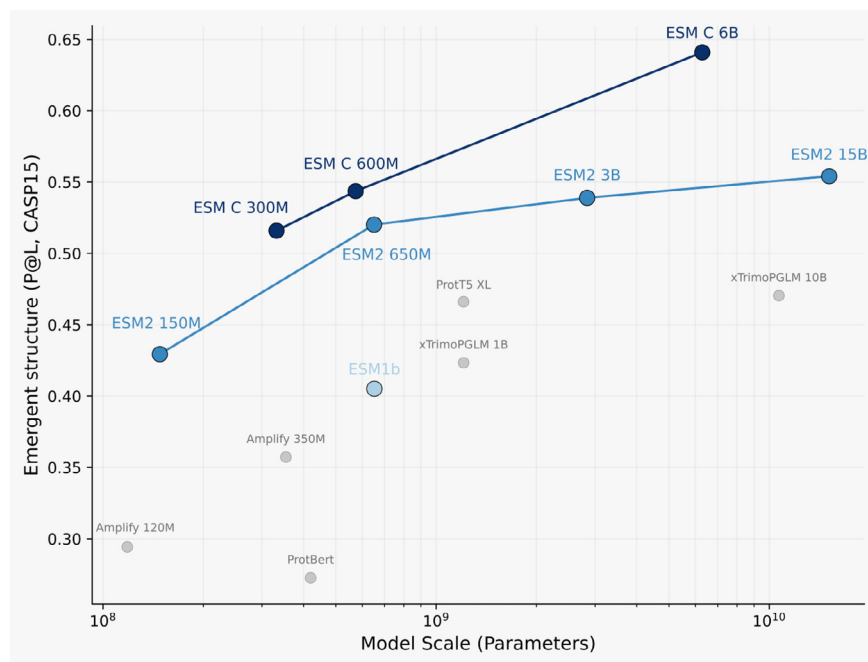
The past year has witnessed remarkable progress in AI models applied to protein sequences. Large-scale machine learning models have improved our ability to predict protein properties, accelerating research in structural biology and molecular engineering. As noted above, several notable protein sequencing models, like AlphaFold, ESM2, and ESM3, have recently been released.

ESM3 integrates multimodal inputs—sequence, structure, and interaction data—while its larger parameter size improves representativeness and predictive accuracy. As the ESM family has expanded in scale, protein prediction performance has improved. Newer models, such as ESM C, released in 2024, have achieved greater accuracy in predicting protein structures in the Critical Assessment of Structure Prediction (CASP15) challenge (Figure 5.2.1).

Emergent structure prediction success, CASP15

Source: [EvolutionaryScale, 2024](#)

Figure 5.2.1



Chapter 5: Science and Medicine

5.2 The Central Dogma

Other significant advancements include ProGen, a generative AI model that, in demonstrating the ability to design functional protein sequences, has highlighted the potential of AI-assisted protein engineering. Similarly, transformer-based models such as ProtT5 leverage deep learning to predict protein function and interactions directly from sequence data, advancing the field of computational biology. Figure

5.2.2 showcases key protein sequencing models and their parameter sizes, arranged by release date. As noted earlier, there is a clear trend toward increasingly larger models trained on ever-expanding datasets. These AI-driven approaches have transformed protein science by minimizing reliance on costly, time-intensive experimental methods, enabling rapid exploration of protein function and design.

Size of protein sequencing models, 2020–24

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

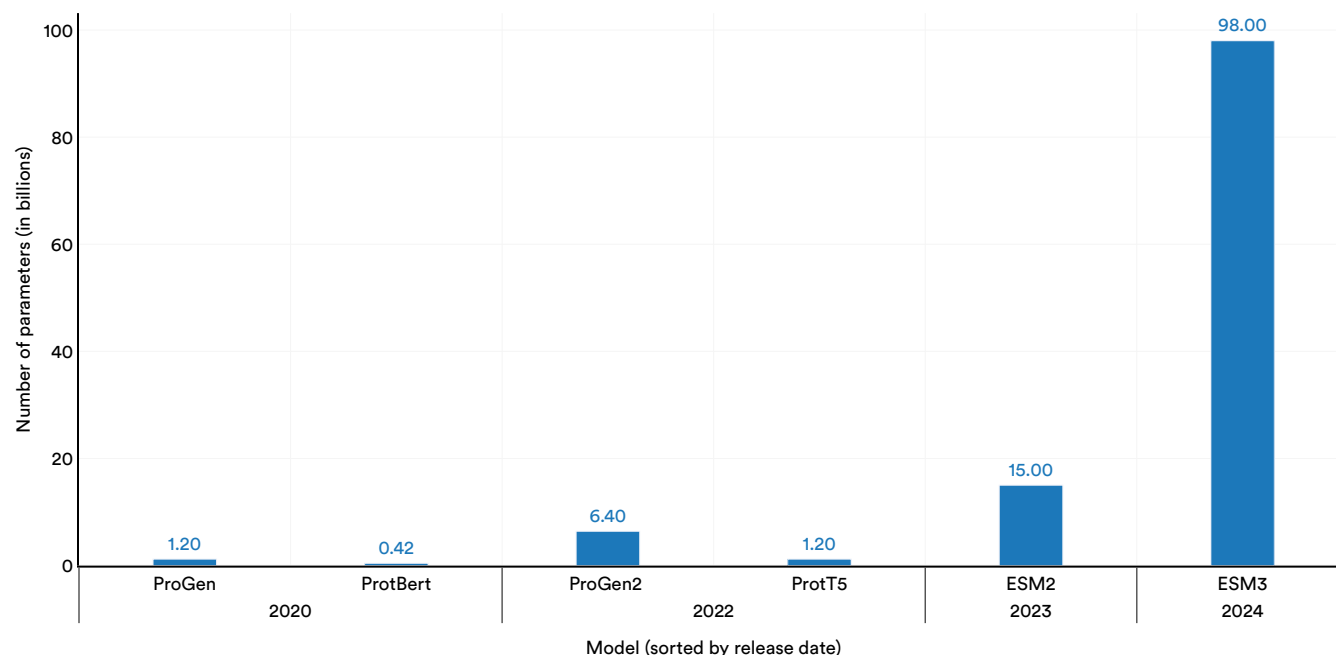


Figure 5.2.2

Chapter 5: Science and Medicine

5.2 The Central Dogma

Public Databases for Protein Science

The expansion of public databases has been crucial for AI applications in protein science. Well-curated, large-scale datasets enable AI models to train on diverse biological

sequences, enhancing their predictive power. Figure 5.2.3 provides information on several key protein science databases and their release date.

Key protein science databases

Source: AI Index, 2025

Dataset	Release date	Description
Protein Data Bank (PDB)	1971	A database of experimentally solved protein structures. When first released, it was the first open-access digital resource in the biological sciences.
Pfam	1995	A comprehensive database of protein families, providing annotations and multiple sequence alignments generated through hidden Markov models.
STRING	2000	Dataset offering valuable information on protein interactions and evolutionary relationships.
UniProt	2002	Still the gold standard for protein sequence and function annotation, with AI-assisted curation improving accuracy.
PDBbind	2004	A subset of the PDB that contains protein biomolecular complexes, including protein-ligand, protein-protein, and protein-nucleic acid complexes.
AlphaFold Database	2021	An essential resource for structural biology, now integrating AI-driven models to predict missing experimental data.

Figure 5.2.3

The number of entries in various public protein science databases has also steadily grown over time (Figure 5.2.4). The increasing availability of AI-generated protein insights has made these databases indispensable tools for researchers and industry professionals. However, maintaining data quality and preventing biases in AI models remain ongoing challenges.

Growth of public protein science databases, 2019–25

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

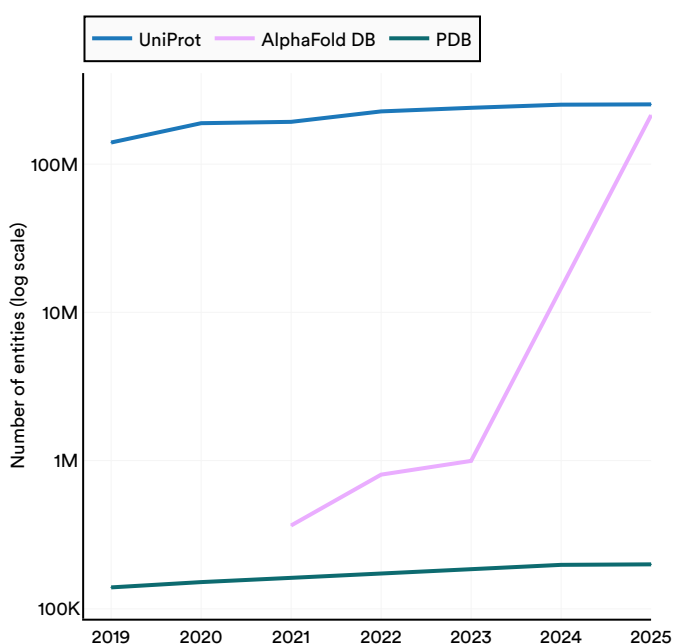


Figure 5.2.4

Research and Publication Trends

AI-Driven Protein Science Publications

AI applications in protein science have gained significant traction in academic research, as evidenced by an increase in AI-driven studies on PubMed and bioRxiv preprints over the past year. These studies focus on several key areas. Protein structure prediction has become more accessible due to advances in machine learning, providing deeper structural insights. AI models now infer biochemical functions from raw sequence data with greater accuracy, enhancing function prediction. In addition, AI models are being developed that can predict protein-drug interactions and even create

new drugs from scratch that can target specific proteins. Both of these tasks are crucial for drug discovery and drug development. Furthermore, AI-generated proteins with novel functions are emerging, particularly in enzyme engineering and therapeutic applications, marking a significant step forward in synthetic protein design. Figure 5.2.5 illustrates the proportion of protein AI-driven research within biological sciences in 2024. The most researched topic was function prediction (8.4%), followed by protein structure prediction (7.6%) and protein-drug interactions (3.0%)

Proportion of AI-driven protein research in the biological sciences, 2024

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

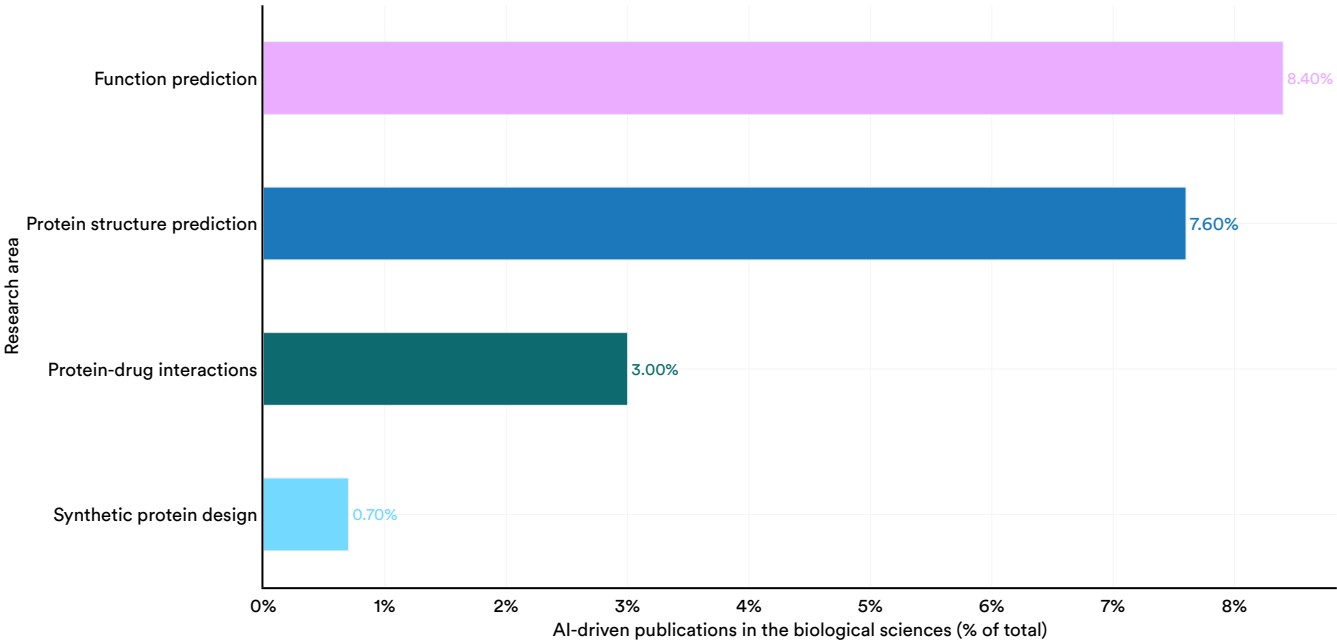


Figure 5.2.5

Image and Multimodal AI for Scientific Discovery

Advances in cryo-electron microscopy, high-throughput fluorescence microscopy, and whole-slide imaging allow scientists to examine and analyze atomic, subcellular context and tissue-level structures with high precision to reveal new insights into complex biological processes. To achieve this, researchers interpret and contextualize image findings with existing scientific knowledge to link observations to biological functions and disease relevance. Given the rise of high-throughput microscopy, active research has increasingly focused on the intersection of vision, vision-language, and, more recently, vision-omics foundation models. The number of microscopy foundation models has increased over time across various techniques (Figure 5.2.6). Light-based models doubled from four to eight in 2024, and, while no electron or fluorescence models were released in 2023, four models for each technique emerged in 2024. Overall, foundation models for microscopy are increasing as more data is collected and made publicly available.

Number of foundation models per microscopy techniques, 2023–24

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

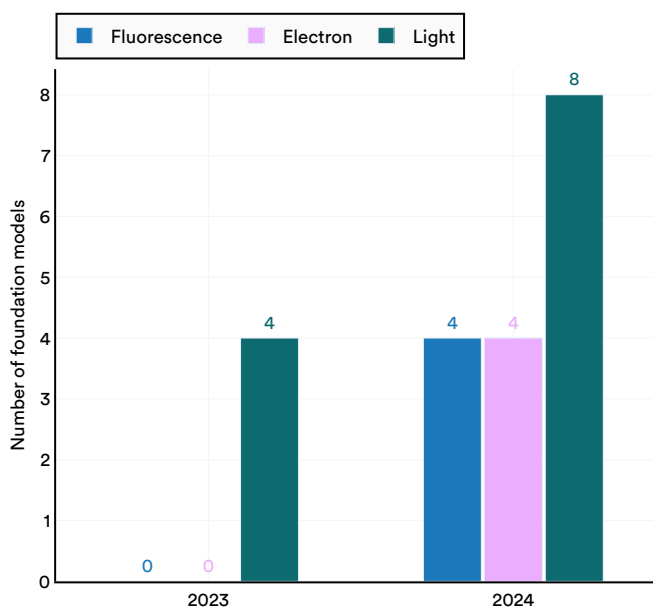


Figure 5.2.6

5.3 Clinical Care, Imaging

Data: Sources, Types, and Needs

AI in medical imaging is rapidly evolving, expanding into new data modalities, and addressing increasingly complex clinical questions. More than 80% of FDA-cleared machine learning software targets the analysis of medical images. Currently, AI is predominantly applied to two-dimensional (2D) data settings, where conventional image-processing architectures, such as convolutional neural networks (CNNs) and transformers, can be effectively utilized. However, despite a number of successes in this field, many AI applications in medical imaging rely on highly limited training datasets.

In histopathology, for example, while staining patient biopsies for histological analysis is routine, only a small fraction of these samples is digitized and made publicly available. Even fewer datasets contain the necessary matched annotations or omics data required for advanced classification tasks. Publicly

available histopathology cohorts rarely exceed 10,000 patient samples, with The Cancer Genome Atlas (TCGA) providing one of the most comprehensive collections—comprising 11,125 patient samples with matched clinical annotations, genomic sequencing, and protein expression data across 32 cancer types. As a result, histopathology AI models are often trained on fewer than 1,000 patient samples, particularly when genomic or proteomic data serve as labels. Limited training sets increase the risk of data overfitting and poor generalization.

Figure 5.3.1 illustrates the geographic distribution of U.S. cohorts used to train deep learning algorithms. Most cohorts originate from California, Massachusetts, and New York, raising concerns about the limited scope of the datasets used to train these algorithms.

US patient cohorts used to train clinical machine learning algorithms by state, 2015–19

Source: Kaushal et al., 2020 | Chart: 2025 AI Index report

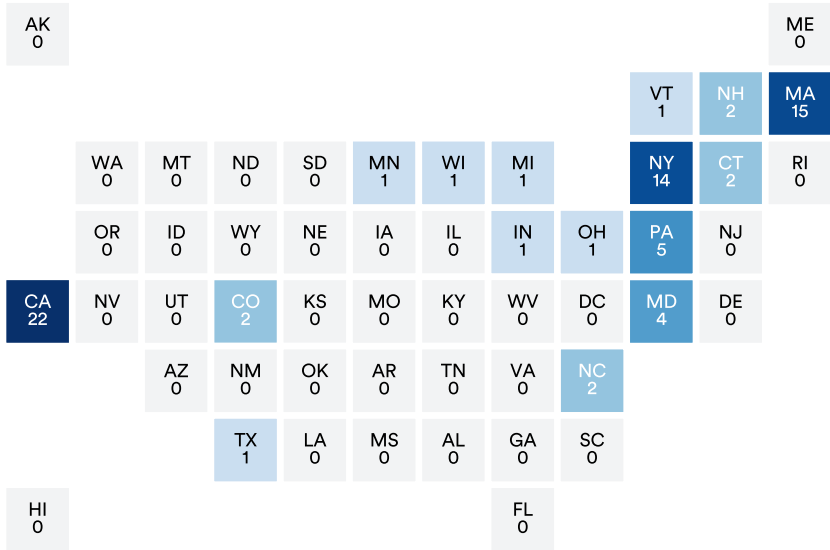


Figure 5.3.1

Chapter 5: Science and Medicine

5.3 Clinical Care, Imaging

These data limitations are more pronounced for three-dimensional (3D) medical imaging. While AI has traditionally focused on 2D modalities such as chest X-rays, histopathology slides, and fundus photography, recent advancements have expanded its application to 3D imaging modalities, including computed tomography (CT), magnetic resonance imaging (MRI), and 3D histopathology analysis. Three-dimensional analysis provides richer data, enabling AI models to learn patterns from volumetric structures and complex surfaces that may not be apparent in 2D slices. Although promising approaches have been developed for the use of AI to analyze 3D medical images, similar data limitations and needs persist. Publicly available 3D datasets remain limited, with [UK Biobank](#) (around 100,000 MRI scans) and [TCIA](#) (around 50,000 studies) among the largest. Although 3D samples are routinely collected in histopathology, 3D imaging is not standard practice, resulting in an absence of publicly available 3D histopathology datasets. Standardization challenges persist due to acquisition variability in pathology. Differences in instrument settings, staining techniques, and institutional practices introduce batch effects, which are further exacerbated by limited training datasets.

Training accurate AI models requires large datasets: CNNs have succeeded with around 10,000 labeled images, but transformers need orders of magnitude more data. [MIMIC-CXR](#) (377,000 images) and [CheXpert Plus](#) (around 226,000 frontal-view radiographs with aligned radiology reports and patient metadata) are important resources but remain smaller than ImageNet (around 14 million images). Data completeness and bias issues remain key challenges.

Figure 5.3.2 illustrates the token volume in text and image datasets used to train various leading medical language and imaging models, in comparison to various all-purpose text and image models. [GatorTron](#), a large clinical LLM designed to extract patient information from unstructured electronic health records, was trained on 82 billion tokens. In contrast, [Llama 3](#) was trained on 15 trillion tokens—nearly 182 times more. On the imaging side, [RadImageNet](#), an open radiologic deep learning research dataset, contains 16 million image-equivalent tokens, while [DALL-E](#), an early OpenAI image generator, was trained on approximately 6 billion—roughly 375 times more.

Training dataset token volumes: medical vs. nonmedical language and imaging models

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

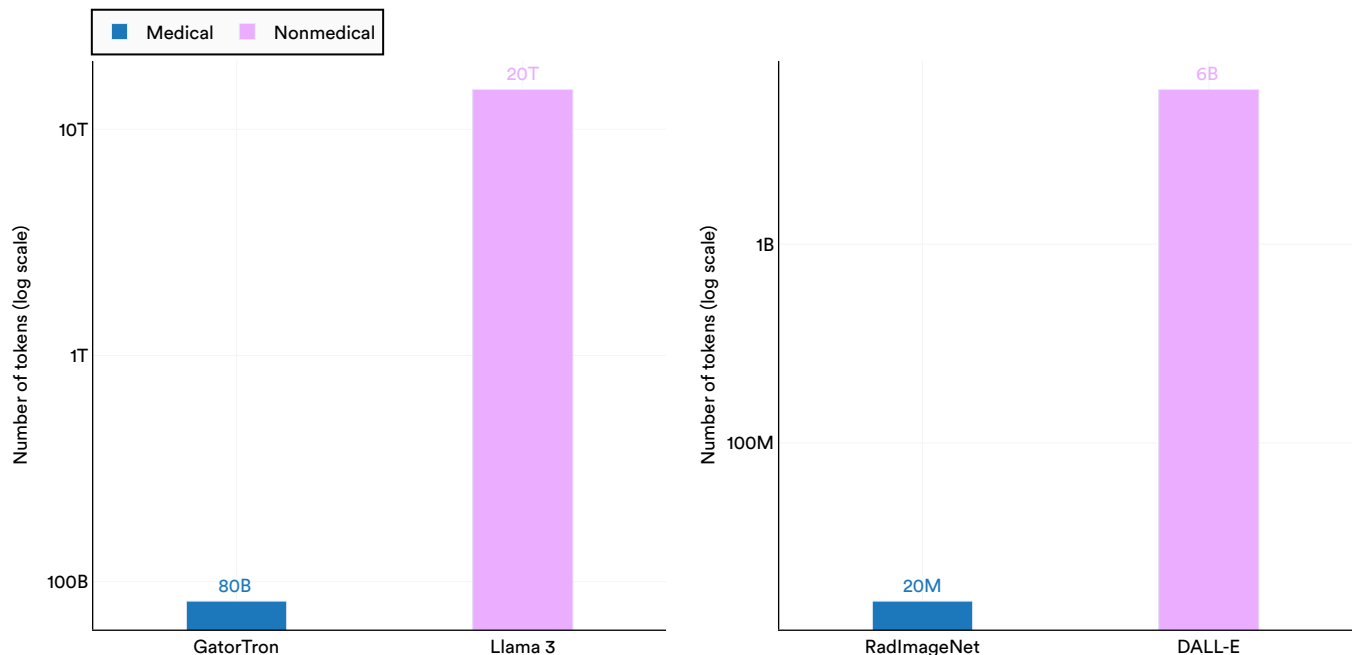


Figure 5.3.2

Chapter 5: Science and Medicine

5.3 Clinical Care, Imaging

Longitudinal imaging is important for modeling disease progression but remains underrepresented. [ADNI](#) (around 2,000 participants over 15-plus years) exemplifies such efforts, but scalable multimodal longitudinal datasets are rare. Addressing these gaps requires privacy-preserving data-sharing (e.g., [federated learning](#)), synthetic data generation, and improved annotation strategies.

To train and validate robust medical imaging AI models, larger, more comprehensive, and multicohort collections of training data are required. By increasing the availability of high-quality, labeled training data, models can be expected to achieve improved performance. Additionally, better validation practices will bolster confidence in these models, facilitating their transition into clinical practice.

Advanced Modeling Approaches

Figure 5.3.3 presents leading clinical imaging modeling approaches, notable releases per approach, and key challenges associated with each.

Imaging modeling approaches and notable AI models

Source: AI Index, 2025

Modeling approach	Notable releases	Advantages	Challenges
Diffusion models	<ol style="list-style-type: none"> RoentGen (2022) RNA-CDM (2023) XReal (2024) 	Generate synthetic medical images for enhanced training, privacy, and pathology-specific augmentation. Outperform GANs in stability and diversity.	Dataset biases, hallucinated artifacts, diagnostic uncertainty.
Large vision-language models (LVLMs)	<ol style="list-style-type: none"> CheXagent (2024) Merlin (2024) Med-Gemini (2024) PathChat (2024) TITAN (2024) PRISM (2025) BiomedParse (2025) 	Integrate medical images with text for improved diagnosis, segmentation, and report automation. LVLMs extend multimodal capabilities.	Data scarcity, generalization to low-resource settings, computational demands.
2D vision-only foundation models	<ol style="list-style-type: none"> CTransPath (2022) Virchow (2024) UNI (2024) MedSAM(2024) 	Pan-cancer detection, biomarker prediction, and image segmentation. Reduce annotation burdens.	Domain generalization, cross-modal adaptability.
Multiscale/slide-level models	<ol style="list-style-type: none"> HIPT (2022) MEGT (2023) MG-Trans (2023) HIGT (2023) Prov-GigaPath (2024) 	Enhance whole-slide imaging analysis using hierarchical transformers and graph-based models for spatial relationships. Improve diagnostic fidelity and interpretability.	Scalability, computational efficiency, dataset variability.

Figure 5.3.3

Chapter 5: Science and Medicine

5.3 Clinical Care, Imaging

In recent years, there has been a notable rise in foundation models being used for medical imaging purposes. Figure 5.3.4 categorizes notable models by medical discipline. In recent

years, the number of medical imaging foundation models has risen sharply, with a particularly high concentration of newly launched pathology models.

Medical disciplines and notable AI models

Source: AI Index, 2025

Discipline	Notable releases
Echocardiology	1. EchoCLIP (2024)
Oncology	1. MUSK (2025)
Ophthalmology	1. RETFound (2023) 2. VisionFM (2024)
Pathology	1. CTransPath (2022) 2. CHIEF (2024) 3. Prov-GigaPath (2024) 4. PathChat (2024) 5. TITAN (2024) 6. Virchow (2024) 7. UNI (2024)
Radiology	1. RoentGen (2022) 2. CheXagent (2024) 3. Merlin (2024) 4. PRISM (2025)

Figure 5.3.4

5.4 Clinical Care, Non-Imaging

Clinical Knowledge

The following section examines the performance of LLMs and recent AI models on key medical knowledge benchmarks.

MedQA

Evaluating the clinical knowledge of AI models involves determining the extent of their medical expertise, particularly knowledge applicable in a clinical setting.

Introduced in 2020, MedQA is a comprehensive dataset derived from professional medical board exams, featuring over 60,000 clinical questions designed to challenge

doctors. AI performance on the MedQA benchmark has advanced significantly. A team of Microsoft and OpenAI researchers recently tested o1, which achieved a new state-of-the-art score of 96.0%—a substantial 5.8 percentage point improvement over the record set in 2023 (Figure 5.4.1). Since late 2022, performance on the benchmark has increased by 28.4 percentage points. As with other general knowledge benchmarks discussed in Chapter 2, MedQA may be approaching a saturation point, indicating the need for more challenging evaluations.

MedQA: test accuracy

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

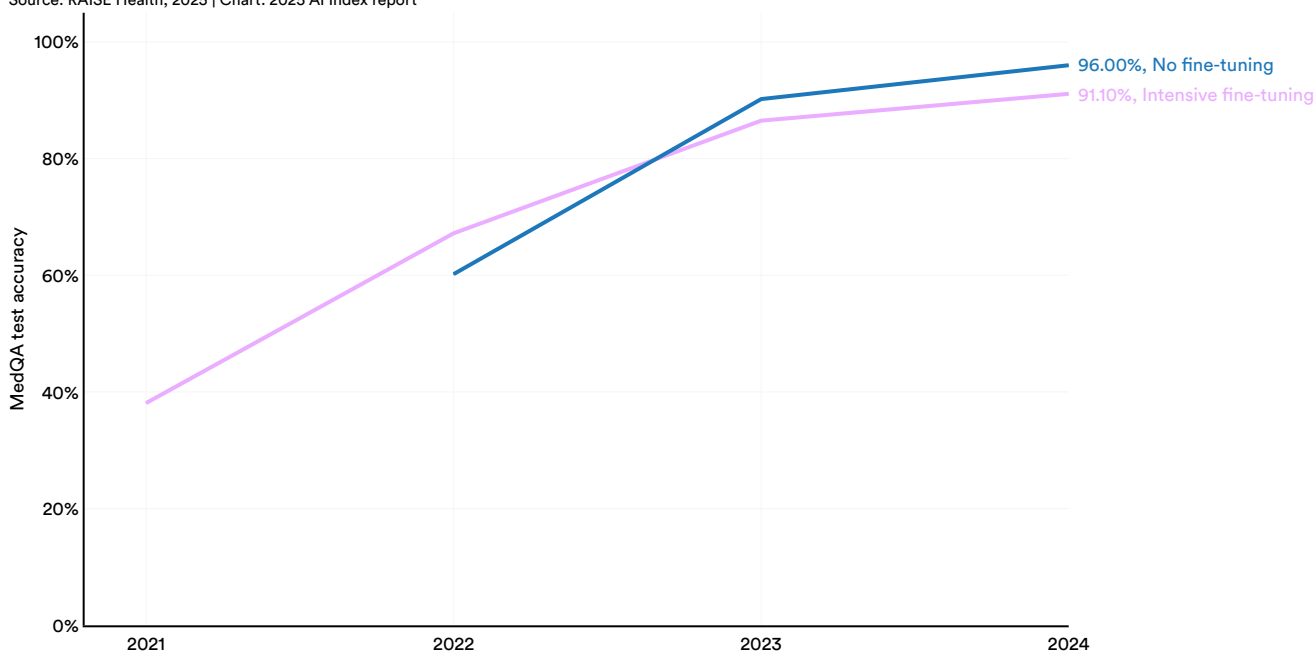


Figure 5.4.1

Highlight:

AI Doctors and Cost-Efficiency Considerations

Some researchers argue that evaluating medical LLMs requires more comprehensive benchmarks than MedQA, those that span a broader range of medical domains. Relying solely on standard medical QA benchmarks like MedQA—while valuable—may overlook the complexities of real-world clinical applications. Alternatively, using multiple benchmarks can offer greater clinical relevance and a more robust assessment of model performance.

This year, new research from UC Santa Cruz, the University of Edinburgh, and the National Institutes of Health has taken a more expansive approach to testing AI medical systems. The study evaluated five leading large language models, including the newly developed o1, which features chain-of-thought reasoning. The other models assessed were GPT-3.5, Llama 3-8B, GPT-4, and Meditron-70B—the last of which is a specialized medical model. These models were tested on a diverse set of medical benchmarks covering various tasks, including concept recognition, text summarization, knowledge-based QA, clinical decision support, and medical calculations. Figure 5.4.2 presents the average performance of these five LLMs across 19 medical datasets. The findings indicate that clinical knowledge performance in LLMs is improving, particularly for newer models like o1 equipped with real-time reasoning capabilities. However, persistent challenges remain, including issues with hallucinations and inconsistent multilingual performance.

Previous research, cited in last year's AI Index, demonstrated that prompting techniques like Medprompt can significantly enhance LLM performance on medical benchmarks without additional fine-tuning. OpenAI's recently released o1 model incorporates some of these insights by employing runtime reasoning before generating final responses. Researchers found that o1 outperforms the GPT-4 series with Medprompt, even without specialized prompting techniques. However, their analysis also highlights the accuracy-cost trade-off associated with o1. While it achieves a 5.8 percentage point higher score than GPT-4 Turbo with Medprompt, it is approximately 1.5 times more

Performance of select LLMs on medical datasets

Source: Xie et al., 2024 | Chart: 2025 AI Index report

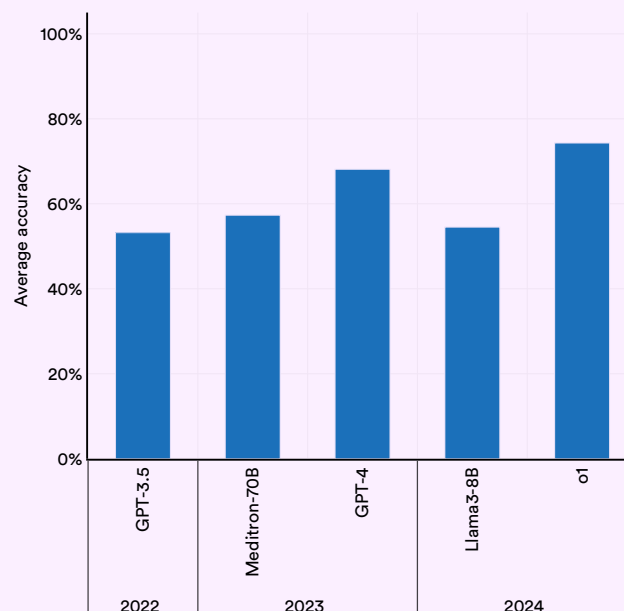


Figure 5.4.2

Enhanced pareto frontier: accuracy vs. cost

Source: Nori et al., 2024

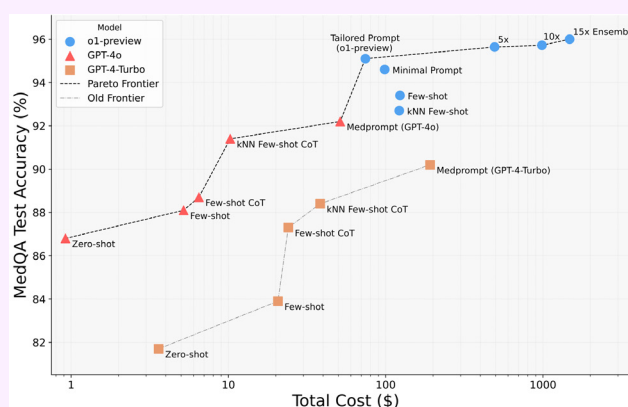


Figure 5.4.3

expensive. Figure 5.4.3 illustrates the cost versus accuracy trade-off on the MedQA benchmark. This trade-off highlights a key consideration for medical professionals deploying AI in clinical settings: the need to balance performance gains with computational costs.

Evaluation of LLMs for Healthcare Performance

Overview

There has been an explosion in interest in the evaluation of language model performance on healthcare tasks. A PubMed search for “large language model” returned 1,566 papers starting in 2019 with 1,210 published in 2024 alone (Figure 5.4.4).

Number of publications on large language models in PubMed, 2019–24

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

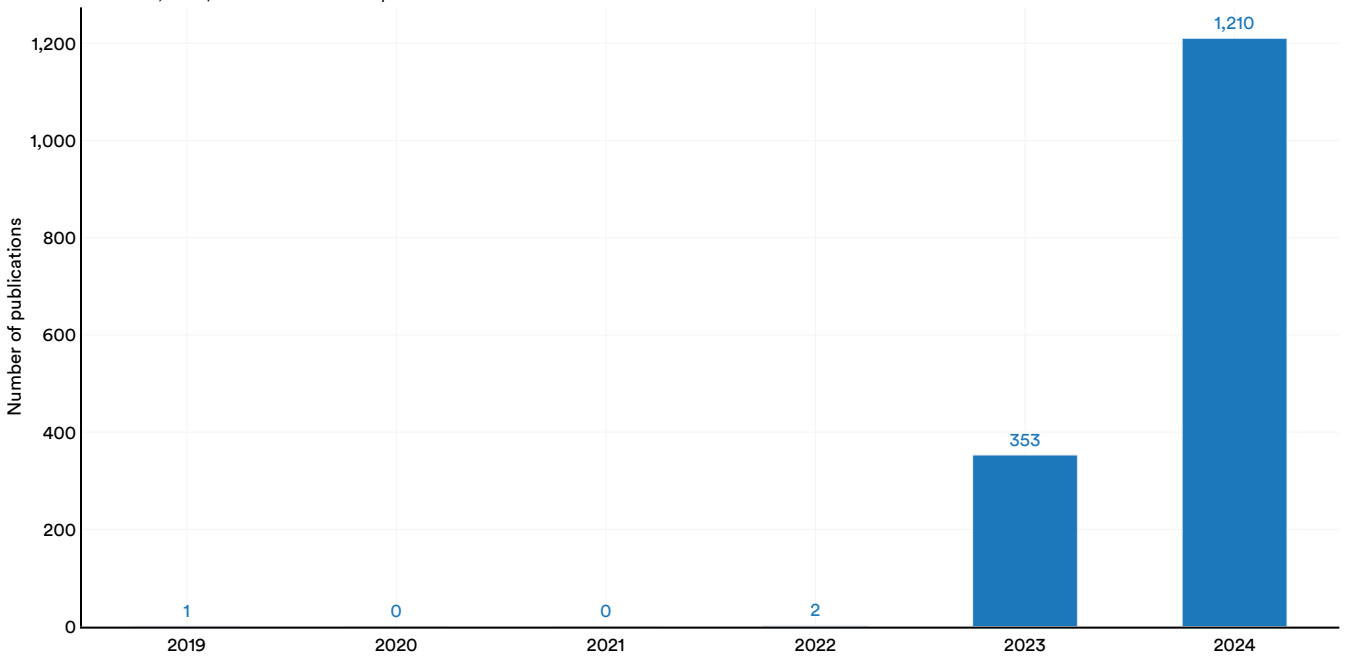


Figure 5.4.4

Chapter 5: Science and Medicine

5.4 Clinical Care, Non-Imaging

A systematic review in early 2024 identified over 500 papers evaluating the performance of NLP on healthcare tasks with a heavy emphasis on medical decision-making (Figure

5.4.5). Most of the healthcare studies that evaluated the performance of NLP systems focused on enhancing medical knowledge (419) and making diagnoses (178).

Healthcare tasks, NLP and NLU tasks, and dimensions of evaluation across 519 studies

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

Task							
	Accuracy	Comprehensiveness	Factuality	Robustness	Fairness, bias, and toxicity evaluation	Deployment metrics	Calibration and uncertainty
Enhancing medical knowledge	222	91	44	33	16	10	3
Making diagnoses	100	38	11	11	14	4	0
Educating patients	88	68	32	22	18	3	2
Making treatment recommendations	47	22	9	8	3	1	0
Communicating with patients	35	29	8	15	22	1	0
Care coordination and planning	36	24	5	5	7	1	0
Triaging patients	24	7	5	2	8	8	0
Carrying out a literature review	18	7	3	2	2	2	0
Synthesizing data for research	16	7	2	3	2	2	0
Generating medical reports	8	8	2	0	3	0	0
Conducting medical research	8	7	3	3	3	0	0
Providing asynchronous care	8	5	3	3	1	1	0
Managing clinical knowledge	5	5	1	1	0	0	0
Clinical note-taking	6	2	1	1	0	0	1
Generating clinical referrals	3	0	0	0	0	0	0
Enhancing surgical operations	3	3	1	1	0	0	0
Biomedical data mining	2	0	0	0	0	0	0
Generating billing codes	1	0	0	0	0	0	0
Writing prescriptions	1	0	0	0	0	0	0
Question answering*	398	194	71	61	54	14	5
Text classification*	29	10	6	5	10	2	0
Information extraction*	29	12	8	5	4	6	0
Summarization*	29	21	7	3	8	0	1
Conversational dialogue*	6	6	1	1	5	1	0
Translation*	5	1	2	2	1	2	0

4 The asterisks represent tasks in NLP and NLU.

Figure 5.4.5⁴

Diagnostic Reasoning With LLMs

Diagnostic errors account for substantial patient harm, and many organizations are exploring AI as a tool to improve the diagnostic process.

Highlight:

LLMs Influence Diagnostic Reasoning

A 2024 single-blind, randomized trial tested GPT-4 assistance against conventional resources in tackling complex clinical vignettes. The study involved 50 U.S.-licensed physicians and evaluated whether AI-enhanced decision-making could improve diagnostic accuracy and efficiency. The results revealed no significant improvement when physicians used GPT-4 alongside traditional resources. In fact, physicians with AI assistance performed only slightly better (76%) than those who relied solely on conventional tools (74%). However, in a secondary analysis, GPT-4 alone outperformed both groups, achieving a 92% diagnostic reasoning score, a 16-percentage-point increase over physicians working without AI (Figure 5.4.6). Despite AI's superior standalone performance, integrating it into clinical workflows proved challenging. There was no clear advantage in time efficiency, as case completion times remained statistically unchanged across conditions.

While purely autonomous AI outperformed physician-only efforts, simply giving doctors access to an LLM did not enhance their performance. This underscores a phenomenon seen in other AI-human collaborations: Bridging the gap between excellent model performance in isolation and effective synergy with clinicians requires rethinking workflows, user training, and interface design.

LLM performance in clinical diagnosis

Source: Goh et al., 2024 | Chart: 2025 AI Index report

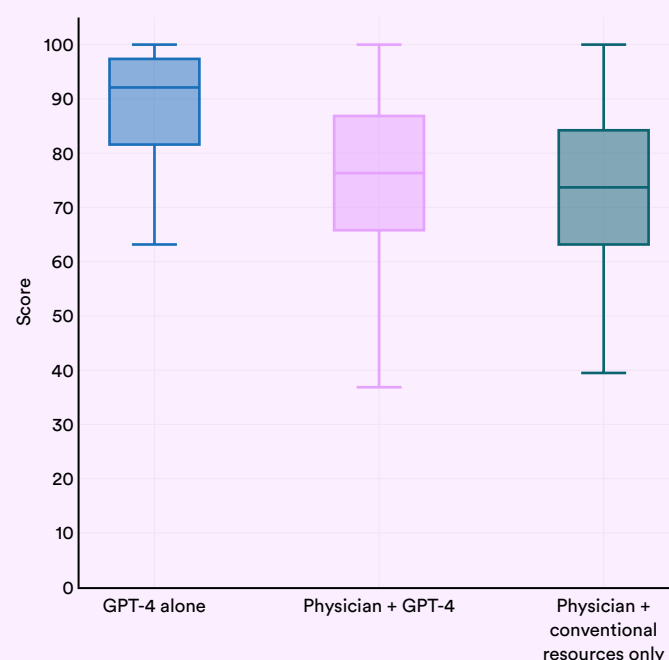


Figure 5.4.6

Management Reasoning and Patient Care Decisions

Beyond diagnosis, physicians must juggle treatment decisions, risk-benefit trade-offs, and patient preferences—collectively

referred to as “management reasoning.” Researchers tested whether LLMs could improve these complex, context-dependent skills.

Highlight:

GPT-4 Assistance on Patient Care Tasks

A 2024–25 prospective, randomized, controlled trial evaluated the impact of GPT-4 assistance on complex clinical management decisions. The study involved 92 physicians, with half using GPT-4 alongside standard resources and the other half relying solely on conventional references. Physicians assisted by GPT-4 outperformed the control group by approximately 6.5 percentage points (Figure 5.4.7). Interestingly, GPT-4 alone performed on par with GPT-4-assisted physicians, suggesting that in certain well-defined scenarios, near-autonomous AI-driven management support may be feasible. However, AI assistance came with a trade-off, as physicians using GPT-4 spent slightly longer on each scenario—a delay researchers attributed to deeper reflection and analysis. Generative AI can meaningfully improve clinical decision-making, but its impact may be qualitative rather than purely efficiency-driven.

Impact of LLM assistance on clinical management

Source: Goh et al., 2025 | Chart: 2025 AI Index report

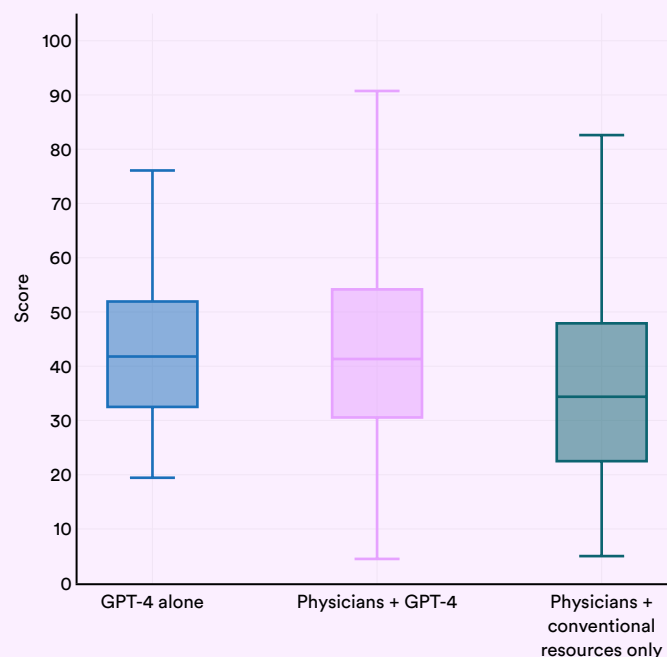


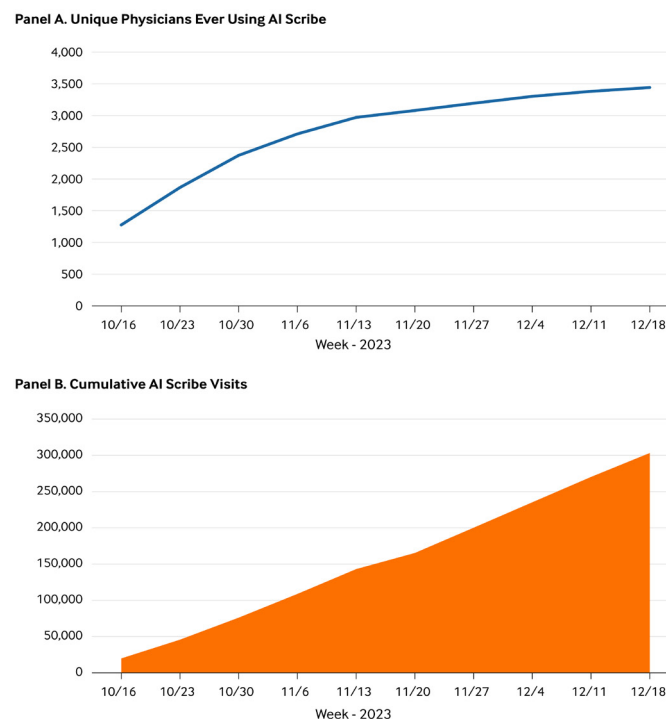
Figure 5.4.7

Ambient AI Scribes

Clinical documentation has long been a source of clinician burden and burnout. Ambient scribe technology has rapidly evolved to integrate LLMs into the processing pipeline for physician-patient recordings. The first study, published in *NEJM Catalyst*, describes the launch of ambient AI scribe technology at Kaiser Permanente Northern California in late 2023. The technology was eventually adopted by thousands of clinicians before the end of the pilot (Figure 5.4.8). This was followed by a second study, published in *JAMIA*, that describes the pilot experience at Intermountain Health. Both studies were based on earlier versions of the technology that were not fully automated or integrated into the electronic health record (EHR).

Cumulative Use of the Ambient Artificial Intelligence (AI) Scribe Tool, October 16–December 24, 2023

Between go-live on October 16, 2023, and December 24, 2023, there were 3,442 unique physician and staff users (Panel A) and a total of 303,266 patient–physician encounters in which the AI scribe was enabled and in which the encounter lasted at least 2 minutes (Panel B).



AI = artificial intelligence.
Source: The authors
NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

Source: Tierney et al., 2024
Figure 5.4.8

Chapter 5: Science and Medicine

5.4 Clinical Care, Non-Imaging

Researchers at [Stanford](#) conducted a two-part study on the use of ambient AI scribe technology, building on prior work by testing a fully integrated, automated AI scribe system. The study demonstrated improvements in both objective measures, such as documentation time, and subjective measures of physician experience. Adoption was strong, with an average uptake of 55% among physicians. The AI scribe provided notable efficiency gains, saving physicians

approximately 30 seconds per note and reducing overall EHR time by about 20 minutes per day (Figure 5.4.9). Additionally, physicians reported significant reductions in burden and burnout, with average decreases of 35% and 26%, respectively. These findings suggest that AI-powered scribe technology can meaningfully improve physician workflow and well-being, offering both time savings and relief from administrative strain.

Impact of AI Scribe on physician EHR usage

Source: Ma et al., 2024 | Chart: 2025 AI Index report

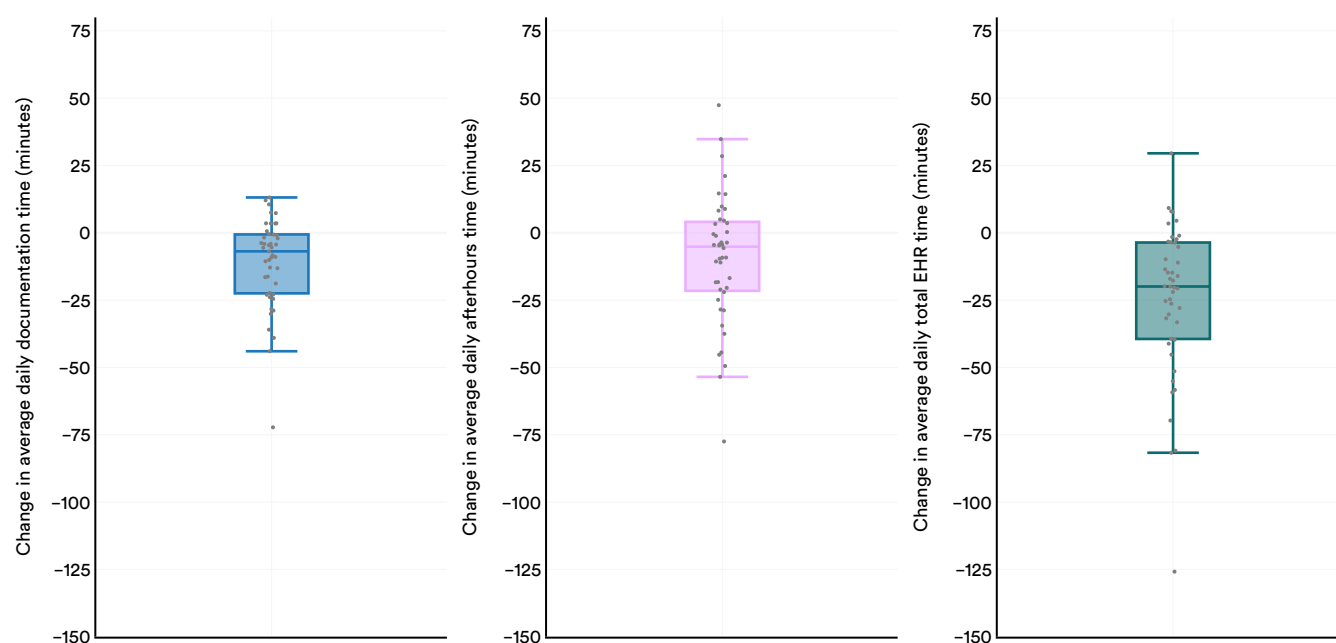


Figure 5.4.9

Investment in ambient scribe technology is [reported](#) to reach almost \$300 million in 2024. While clinical documentation has been the starting point for the technology and the evaluations performed to date, optimists envision ubiquitous

ambient listening technology in both outpatient and inpatient settings that will eventually support order placement, billing and coding, and real-time clinical decision support.

Deployment, Implementation, Deimplementation

FDA Authorization of AI-Enabled Medical Devices

The deployment of AI in clinical settings has grown exponentially over the past decade, highlighted by the dramatic increase in the number of AI-enabled medical devices authorized by the U.S. Food and Drug Administration (FDA).

The FDA authorized its first AI-enabled medical device in 1995. For the next two decades, annual approvals remained in the single digits. In 2015 alone, six AI medical devices were approved. Since then, the number of yearly approvals has surged, peaking at 223 in 2023 (Figure 5.4.10).

Number of AI medical devices approved by the FDA, 1995–2023

Source: FDA, 2024 | Chart: 2025 AI Index report

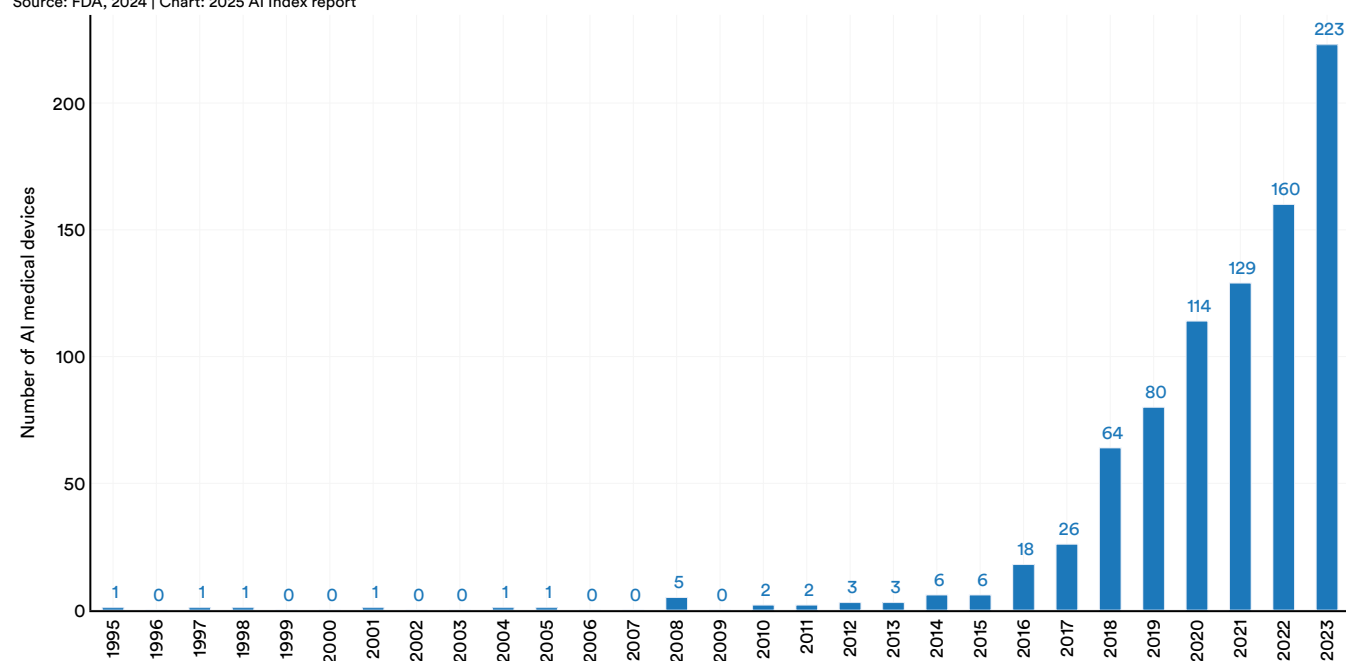


Figure 5.4.10

Successful Use Cases: Stanford Health Care

In practice, transitioning AI models into real-world use requires a framework that ensures fairness, utility, and reliability. Stanford Health Care has led the way by evaluating and implementing AI tools using its FURM (Fair, Useful,

Reliable, Measurable) framework. Among the six AI use cases assessed, two have been successfully implemented: (1) screening for peripheral arterial disease (PAD) and (2) improving documentation and coding for inpatient care. This section details screening for peripheral arterial disease.

Chapter 5: Science and Medicine

5.4 Clinical Care, Non-Imaging

Screening for Peripheral Arterial Disease

Peripheral arterial disease (PAD) is a chronic vascular condition that often goes undiagnosed in its early stages, leading to severe complications such as critical limb ischemia and amputation. To improve early detection and intervention, Stanford Health Care developed and implemented an AI-enabled PAD classification model designed to enhance screening and optimize patient care.

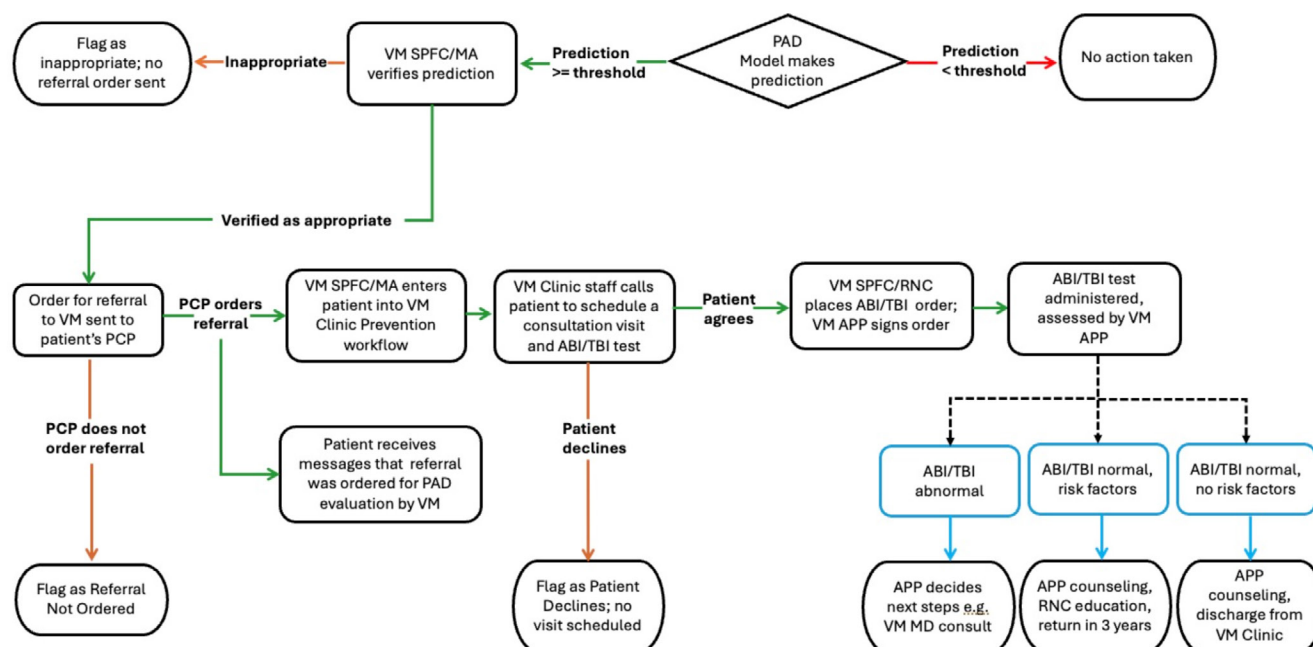
The primary goal of the PAD screening tool is to facilitate earlier diagnosis in primary care populations, allowing for medical or surgical intervention before the disease leads to

severe complications. By identifying high-risk patients, the model also helps optimize resource allocation, ensuring that those most in need receive immediate follow-up and care.

To integrate seamlessly into clinical workflows, the AI tool was designed to automatically assess PAD risk and flag high-risk individuals for further evaluation. If the condition is confirmed, the patient is referred for a vascular consultation. Figure 5.4.11 illustrates the proposed model and workflow details for integrating PAD screening into clinical workflows, including risk assessment, referrals, and patient follow-up.

Proposed model and workflow for integrating PAD screening into clinical practice

Source: Callahan et al., 2024



ABI/TBI – ankle/toe brachial index; APP – advanced practice provider; MA – medical assistant; RNC – registered nurse coordinator; SPFC – specialty patient flow coordinator; VM – vascular medicine

Figure 5.4.11

Following a successful pilot phase, the PAD screening tool advanced to Stage 2 and was fully implemented at Stanford Health Care. The model is expected to impact approximately 1,400 patients annually. Beyond its clinical benefits, the program has demonstrated financial sustainability, operating

independently without external funding. By increasing early PAD detection, reducing the likelihood of severe complications, and improving patient outcomes, this AI-driven approach is reshaping the standard of care for PAD management.

Social Determinants of Health

The integration of LLMs and AI-based clinical decision support (CDS) systems is transforming medicine, though adoption varies by specialty. While some embrace LLMs, others remain cautious. This review explores research and innovations, emphasizing the role of a strong evidence base. A key aspect is addressing social determinants of health (SDoH), such as socioeconomic status and environment. In 2024, AI advancements targeted SDoH, improving patient care and health equity.

Extracting SDoH From EHR and Clinical Notes

Fine-tuned multilabel classifiers (Flan-T5 XL) outperformed ChatGPT-family models in identifying SDoH in clinical notes and were less sensitive to demographic descriptors. They also exhibited lower bias, with reduced discrepancies when race, ethnicity, or gender was introduced. Figure 5.4.12 illustrates the performance of various models on SDoH identification tasks in a radiotherapy test set. Newer, larger models like Flan-T5-XXL, augmented with synthetic and gold data (SDoH-labeled sentences), showed superior performance. As models have scaled and incorporated more data over time, their ability to identify SDoH has improved.

Model performance on in-domain RT test dataset (any SDoH)

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

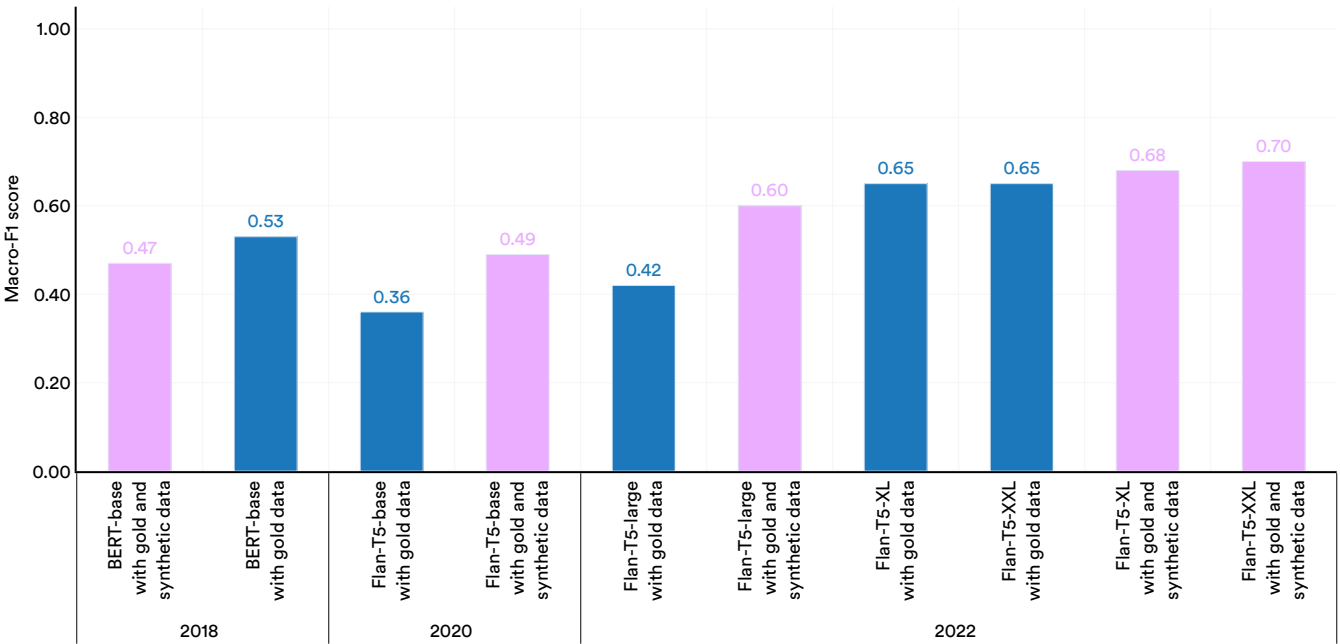


Figure 5.4.12

Extracting SDoH from EHRs helps healthcare providers address social needs like housing instability or food insecurity. These findings highlight LLMs' potential to enhance SDoH

documentation, resource allocation, and health equity while emphasizing the need for bias mitigation and robust synthetic data methods.

Chapter 5: Science and Medicine

5.4 Clinical Care, Non-Imaging

AI Adoption Across Medical Fields and the Integration of SDoH

Figure 5.4.13 highlights various medical fields and illustrates how AI integrates social determinants of health in each.

Field	Recent research	Description of integration
Oncology	Istasy et al., 2024	In cancer care, AI-driven tools have been developed to consider SDoH in treatment planning. By incorporating factors such as a patient's access to care and support systems, these tools assist oncologists in creating personalized treatment plans that are both effective and feasible for patients.
Cardiology	Snowdon et al., 2023 Quer et al., 2024	AI models in cardiology have been enhanced to include SDoH, improving the accuracy of risk assessments for conditions like hypertension and heart failure . This inclusion allows for more comprehensive patient evaluations and tailored management strategies.
Psychiatry	Stade et al., 2024	LLMs have been applied to analyze community-level SDoH data, aiding in the allocation of mental health resources. By identifying areas with high social risk factors, healthcare systems can prioritize interventions and support services in communities with the greatest need.

Figure 5.4.13

Synthetic Data

Synthetic data is revolutionizing healthcare by enhancing privacy-preserving analytics, clinical modeling, and AI training. It optimizes workflows, simulates rare cases, and supports AI-driven innovations. However, scalability concerns, as noted in the first chapter of this year's AI Index, call for cautious adoption.

Clinical Risk Prediction

A recent [study](#) validated synthetic data for privacy-preserving

clinical risk prediction. Using ADSGAN, PATEGAN, and DPGAN, researchers modeled lung cancer risk in ever-smokers from the UK Biobank.⁵ The figure below compares PCA eigenvalues, showing how ADSGAN and PATEGAN closely match real data distributions, enabling reliable clustering and feature selection (Figure 5.4.14). These findings demonstrate that synthetic datasets can preserve statistical fidelity, support exploratory analysis, and develop predictive models without real and identifiable patient data.

Principal component analysis

Source: [Qian et al., 2024](#)

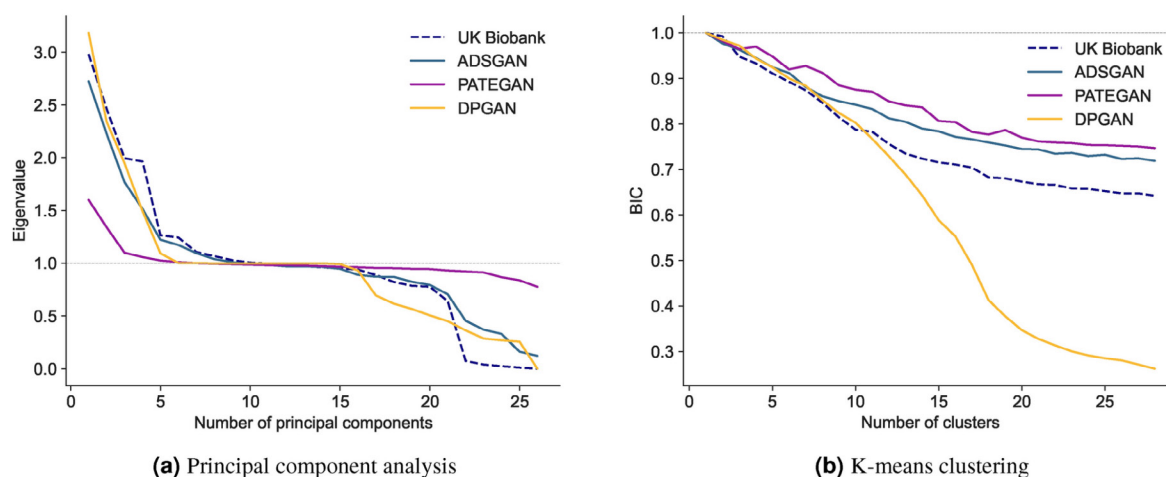


Figure 5.4.14

⁵ An ever-smoker is someone who has smoked at least 100 cigarettes in their lifetime.

Chapter 5: Science and Medicine

5.4 Clinical Care, Non-Imaging

Drug Discovery

A recent Nature [study](#) introduced a generative AI approach for in silico formulation optimization and particle engineering in drug development. Using an image generator guided by critical quality attributes, it creates digital formulations for analysis without extensive physical testing. The study validated this method by predicting the percolation threshold of microcrystalline cellulose (MCC) in oral tablets. Figure 5.4.15 compares the tortuosity calculations of real tablet volumes (green squares) with AI-synthesized volumes (red circles).⁶ Their close alignment suggests that synthetic data holds promise for modeling drug properties and improving AI-driven drug discovery.

Percolation threshold prediction and validation based on AI-generated synthetic structures

Source: [Hornick et al., 2024](#)

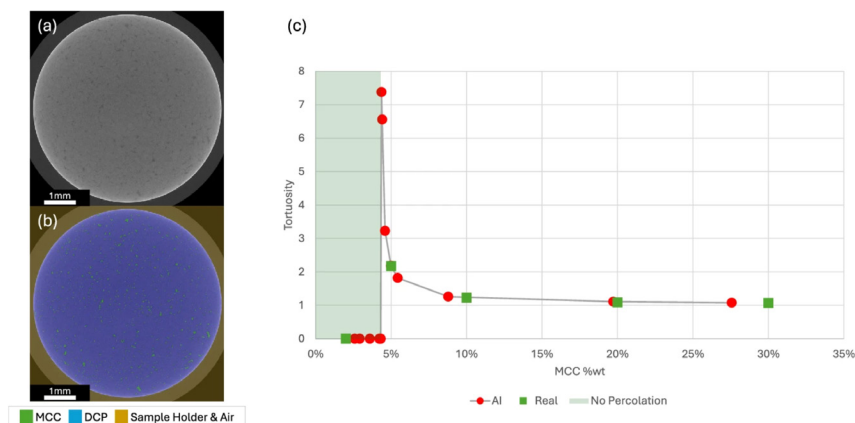


Figure 5.4.15

Data Generation Platforms

Platforms are necessary to demonstrate, standardize, and automate the creation of synthetic data. [Recently published research](#) has demonstrated that large-scale synthetic data generation and validation is not only feasible but also capable of significantly enhancing AI applications in medicine with their synthetic tabular neural generator (STNG) framework. Figure 5.4.16 compares the area-under-the-curve values for real and synthetic heart disease datasets to evaluate the effectiveness of different synthetic data generation methods. In many cases, there is a fairly close overlap between the real datasets and the synthetic datasets, showing the ability of synthetic data to model complex health conditions closely. Advancements in synthetic data generation methodologies can improve data fidelity while minimizing privacy risks.

Areas under the curve for evaluating synthetic heart disease datasets

Source: [Rashidi et al., 2024](#)

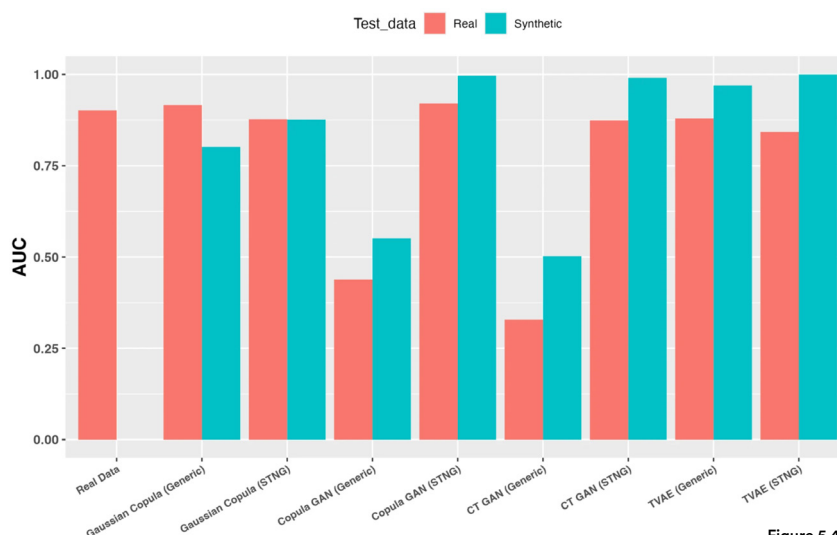


Figure 5.4.16

⁶ Tortuosity is a measure of how convoluted or twisted a path is compared to the shortest possible straight-line distance between two points.

Chapter 5: Science and Medicine

5.4 Clinical Care, Non-Imaging

Electronic Health Record System

AI integration in electronic health records (EHRs) can ease healthcare burdens by streamlining administration, enhancing clinical decision support, and improving patient care. With major vendors—Epic, Oracle Health (formerly Cerner), Meditech, and TruBridge (formerly CPSI)—dominating the market, their AI tools can be widely adopted within their networks. As of 2021, EHR adoption had approached 90% for any system and 80% for certified EHR systems.

A 2023 American Hospital Association IT survey found that most hospitals using ML or predictive models in their EHRs relied on a dominant vendor for inpatient care (Figure 5.4.17). Adoption was highest with Epic, Cerner, and Meditech. While Epic, Cerner, and CPSI hospitals primarily used vendor-developed models, Meditech and others more often adopted third-party or in-house solutions (Figure 5.4.18).

Predictive model use across primary inpatient EHR vendor

Source: AHA survey, 2024 | Chart: 2025 AI Index report

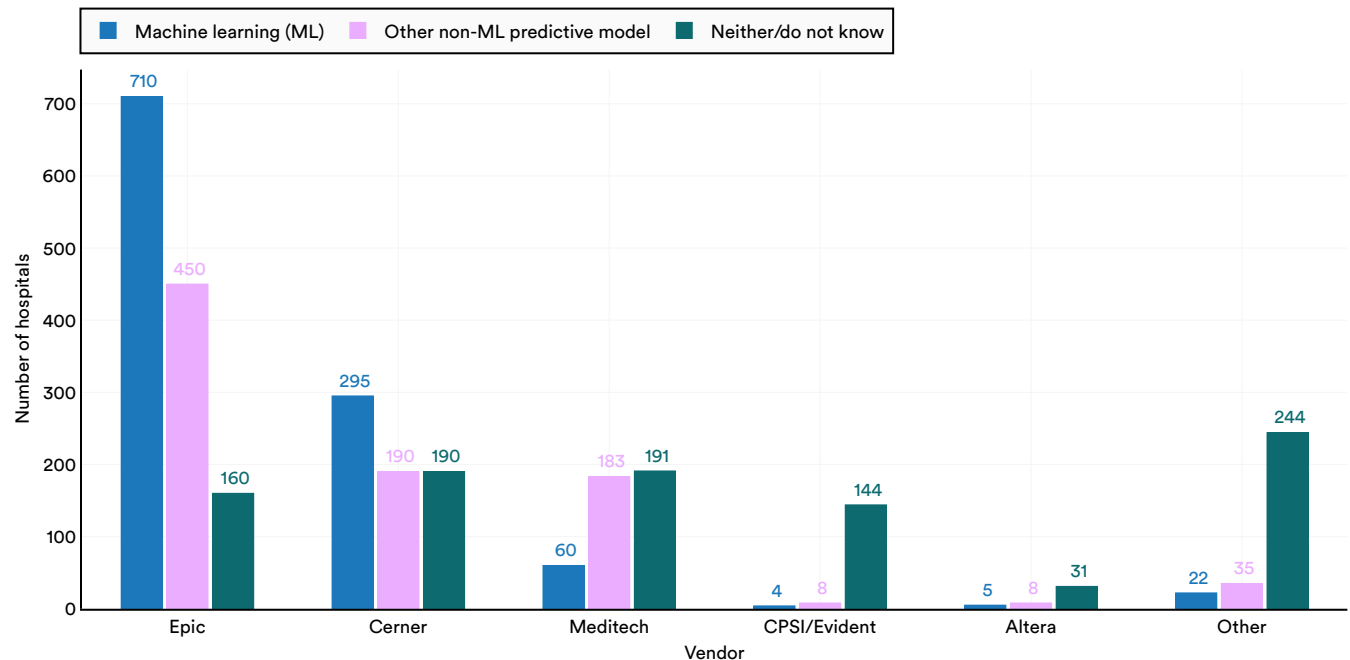


Figure 5.4.17

Developer of predictive models across EHR vendor

Source: AHA survey, 2024 | Chart: 2025 AI Index report

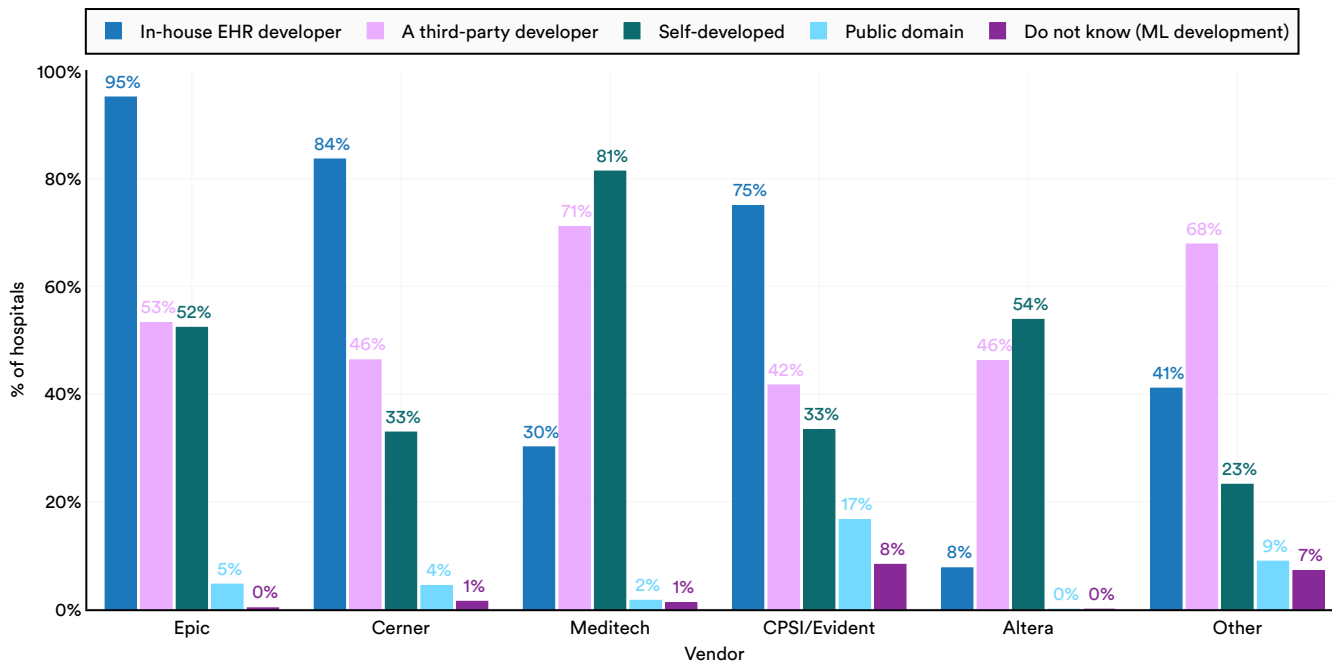


Figure 5.4.18

AI integration into EHRs could streamline clinical workflows and enhance provider and patient experiences. However, it remains unclear whether AI-enabled health IT will benefit underserved communities, which often struggle with technological adoption. Rural areas, for example, face barriers like limited broadband access, weak healthcare

IT infrastructure, and EHR functionality constraints—key enablers of AI-driven healthcare. Additionally, it is important to assess whether AI tools are equitably developed for both basic and comprehensive EHR systems, as many resource-limited settings still rely on the former.

Clinical Decision Support

AI has transformed how healthcare providers diagnose, predict, and manage diseases with an increasing focus on rigorous evaluation of AI-based systems in clinical trials. The evolution of AI in clinical decision support (CDS) reflects a shift from reactive interventions—e.g., during

the COVID-19 pandemic—to proactive, data-driven clinical decision-making with clinical trials increasing over the years. The number of clinical trials that have included mentions of artificial intelligence is steadily rising (Figure 5.4.19).

Number of clinical trials that have included mentions of AI, 2014–24

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

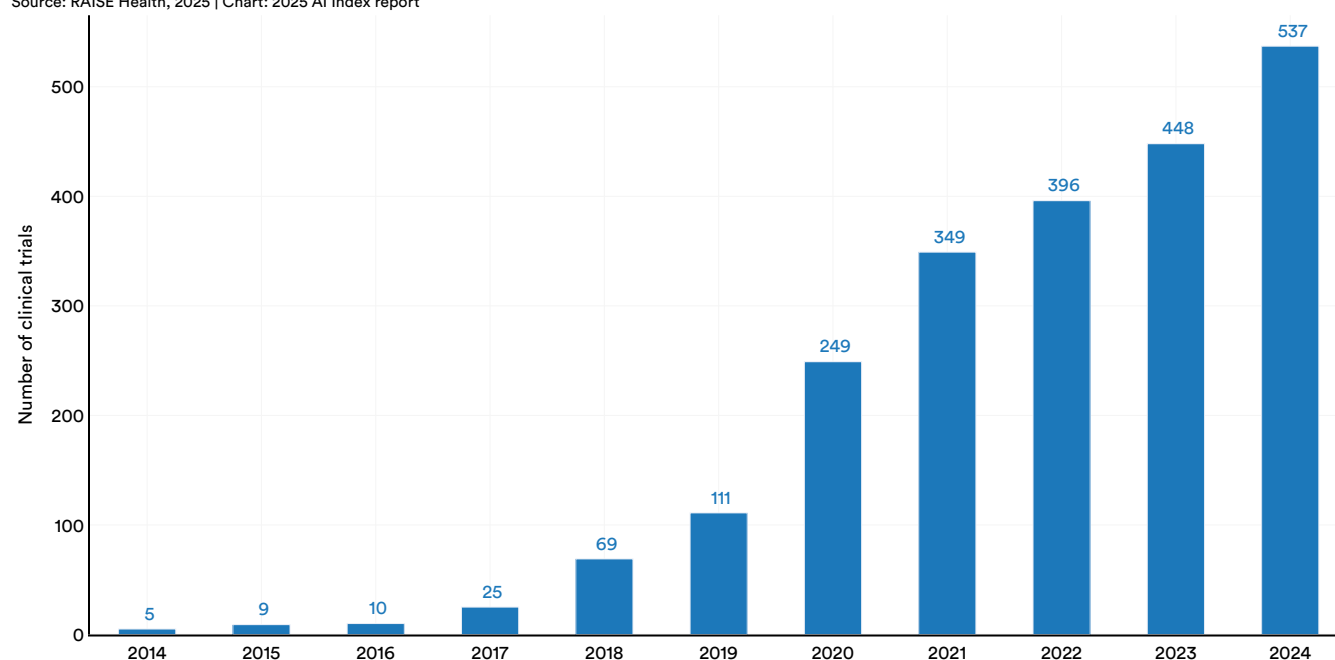


Figure 5.4.19

Chapter 5: Science and Medicine

5.4 Clinical Care, Non-Imaging

The COVID-19 pandemic accelerated AI adoption in triage, resource allocation, and outcome prediction, showcasing the technology's potential in real-time CDS. Post-pandemic, AI expanded beyond emergency response to managing chronic disease, optimizing procedures, and streamlining workflows. Trials like the [CERTAIN Study](#) demonstrated how AI-driven real-time procedural support could improve diagnostic

accuracy in gastrointestinal procedures. By 2023, AI in CDS extended to medication safety and workflow optimization, as seen in [Preventing Medication Dispensing Errors in Pharmacy Practice](#), which used AI to detect real-time medication errors. Globally, AI-driven clinical trials have sharply risen, with China (105 trials), the U.S. (97), and Italy (42) leading in 2024 (Figure 5.4.20).

Number of clinical trials that have included mentions of AI by select geographic areas, 2021–24

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

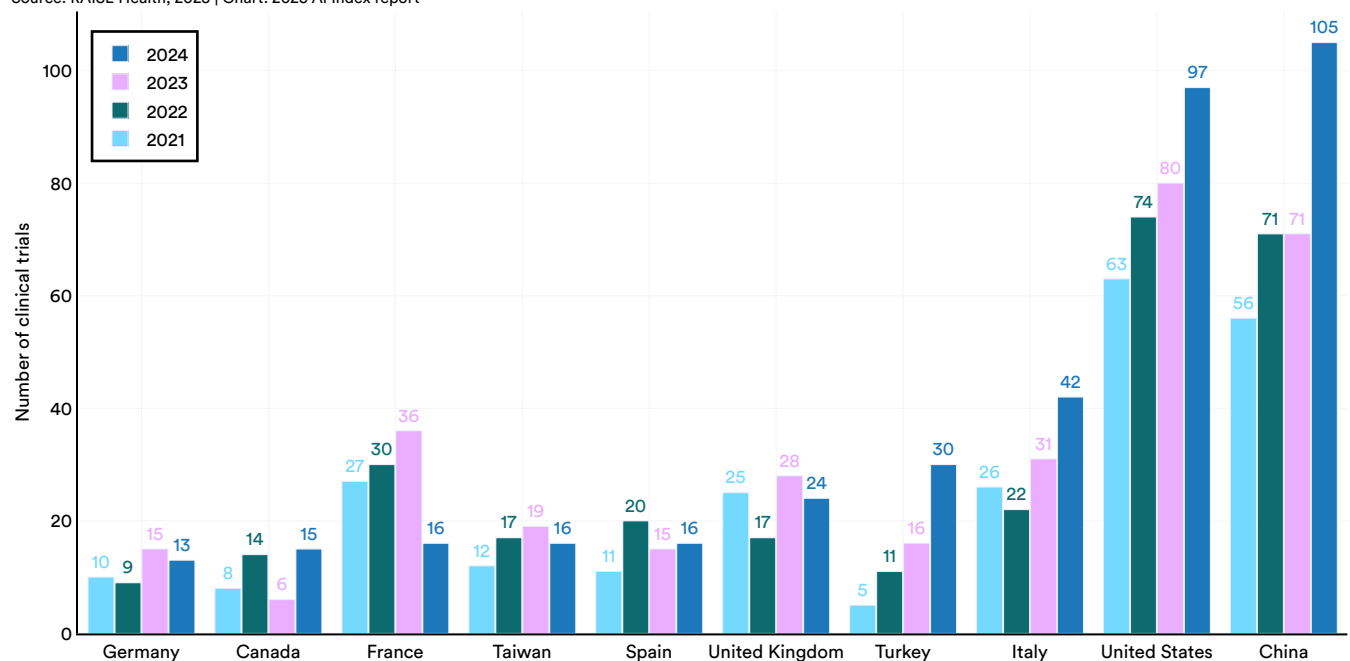


Figure 5.4.20

Chapter 5: Science and Medicine

5.5 Ethical Considerations

The increasing integration of AI in medical research and clinical care as discussed in previous sections brings both promises and challenges. AI systems lean heavily on large amounts of data for training. The collection, use, and sharing of this data—especially in high-stakes domains such as healthcare—can raise various ethical concerns.

5.5 Ethical Considerations

Meta Review

For this section, the AI Index conducted a meta review of thousands of medical ethics studies to glean insights on the state of the field. The team's methodology is highlighted in Figure 5.5.1.

Attention to the ethical issues in medical AI has increased in each of the past five years. The number of publications related to ethics and medical AI increased fourfold from 2020 to 2024 (Figure 5.5.2).

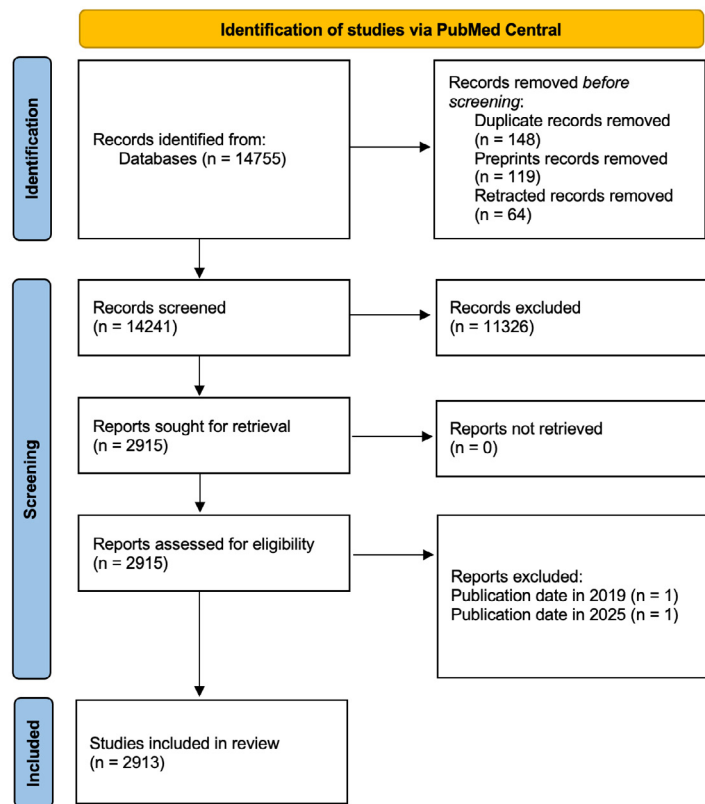


Figure 5.5.1

Number of medical AI ethics publications, 2020–24

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

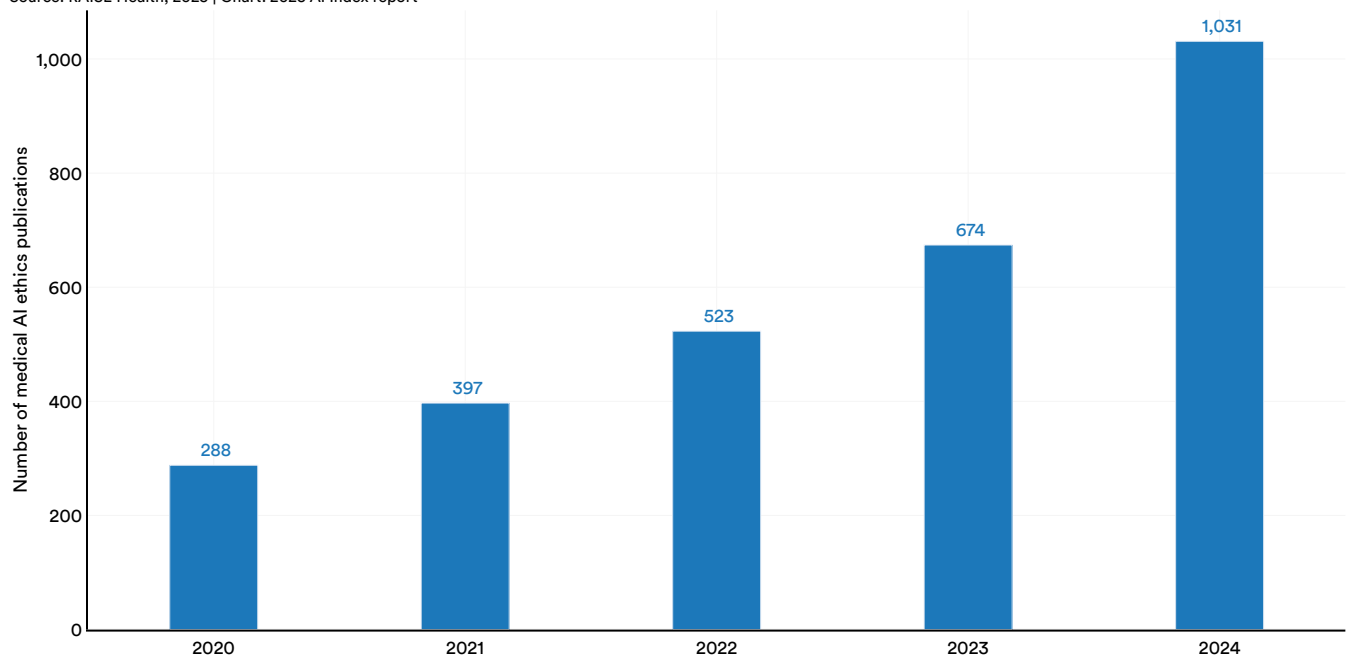


Figure 5.5.2

Chapter 5: Science and Medicine

5.5 Ethical Considerations

The focus of AI applications in medical ethics literature has evolved over time. Figure 5.5.3 illustrates the ethical issues discussed in AI medical papers from 2020 to 2024. In 2024, bias and privacy were the most frequently cited concerns, followed by equity. In contrast, privacy was a more prominent topic than bias in 2020, but this trend has since shifted.

Top 10 ethical concerns discussed in medical AI ethics publications, 2020–24

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

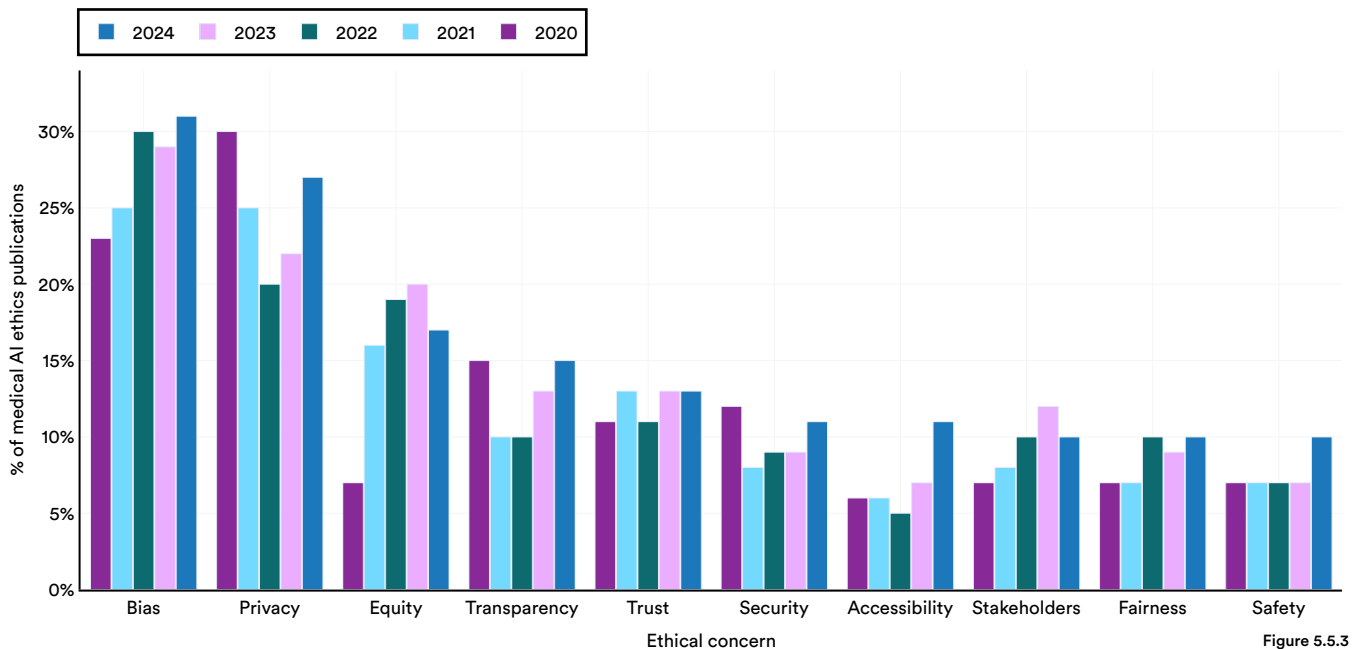


Figure 5.5.3

In terms of AI tools, much attention has been paid in medical ethics literature to OpenAI's GPT series (e.g., ChatGPT) (Figure 5.5.4). This reflects an expanding interest in large-language models over the past few years.

AI tools discussed in medical AI ethics publications, 2020–24

Source: RAISE Health, 2025 | Chart: 2025 AI Index report

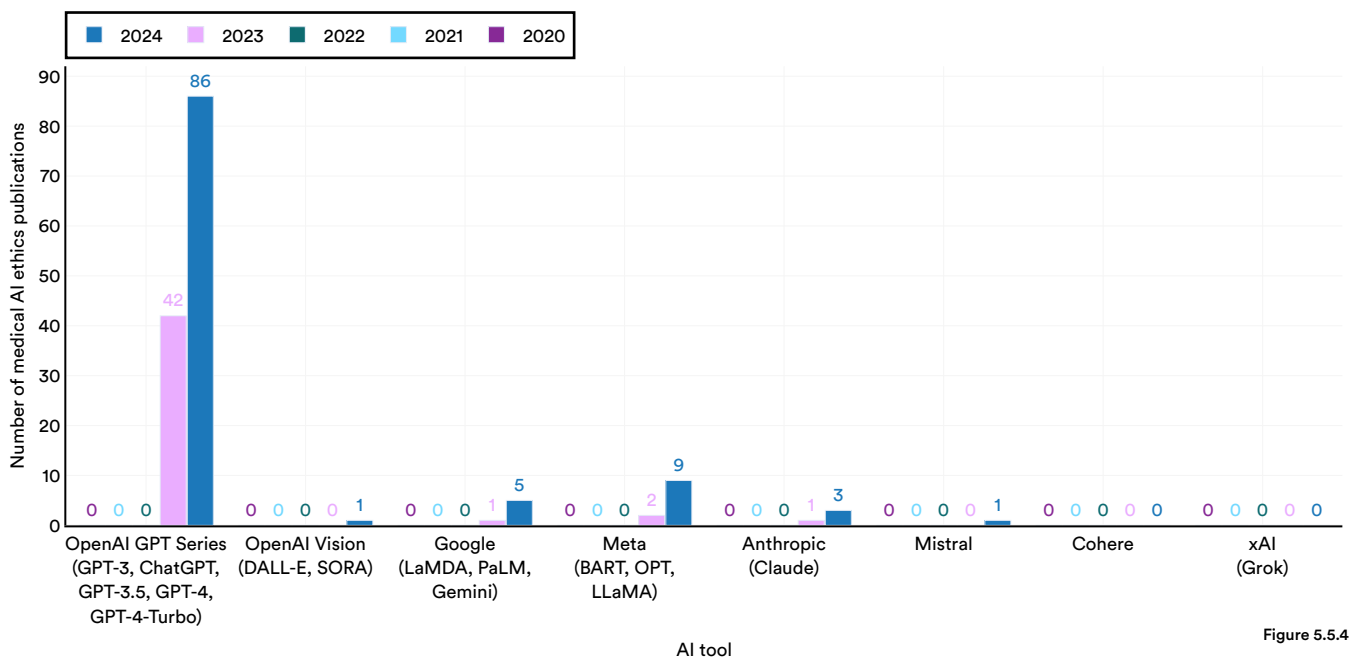
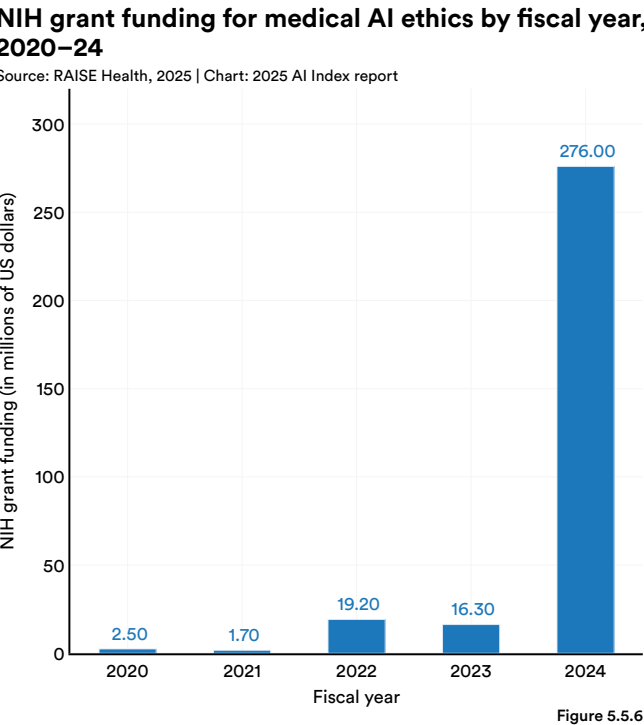
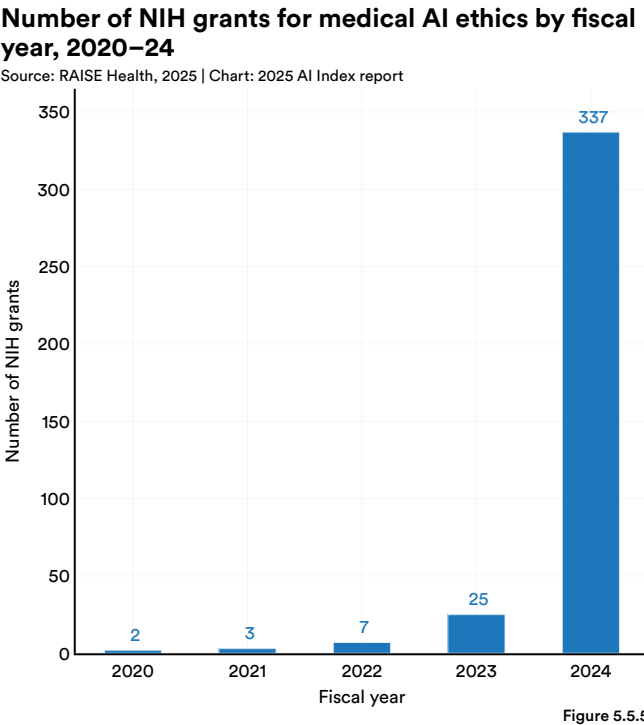


Figure 5.5.4

Chapter 5: Science and Medicine
5.5 Ethical Considerations

Figure 5.5.5 and Figure 5.5.6 show the number and total funding of NIH grants for medical AI ethics projects by fiscal year. The number of grants skyrocketed from 25 in 2023 to

337 in 2024 (Figure 5.5.5). Similarly, total funding soared from \$16 million in 2023 to \$276 million in 2024—an almost 17-fold increase in just one year.



Chapter 5: Science and Medicine

5.6 AI Foundation Models in Science

This year, dozens of foundation models have been developed across various scientific fields. Some are refined large language models, adapted for specific domains using relevant publications; others are trained from scratch with specialized data, such as time series or weather data. These foundation models are then fine-tuned for targeted scientific tasks or applications.

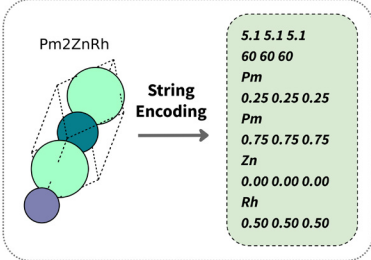
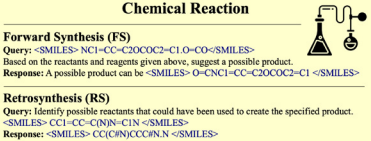
5.6 AI Foundation Models in Science

Highlight:

Notable Model Releases

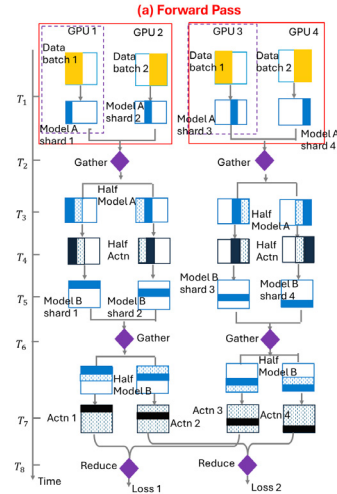
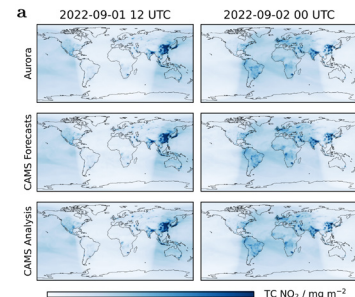
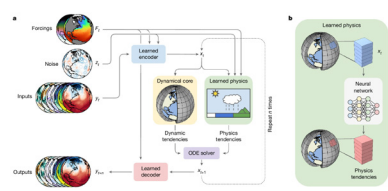
AI has driven significant progress in other scientific domains, including physics, chemistry, and geosciences. The table below highlights some of the most notable recent launches in these areas, alongside newly released resources that further track these developments. This

analysis represents an initial effort by the AI Index, which aims to expand and deepen its coverage of AI-driven scientific progress across a broader range of disciplines in the future.

Date	Name	Domain	Significance	Image
Feb 6, 2024	CrystalLLM	Materials science	Researchers fine-tuned LLaMA-2 70B on text-encoded atomistic data to generate stable materials, achieving nearly double the metastability rate of a leading diffusion model (49% vs. 28%) while maintaining physical plausibility. The approach enables flexible applications like unconditional generation, structure infilling, and text-guided design, with model scale enhancing symmetry awareness.	 <p>Figure 5.6.1 Source: Gruver et al., 2024</p>
Feb 14, 2024	LlaSMol	Chemistry	To address LLMs' poor performance on chemistry tasks, researchers introduce SMolInstruct, a high-quality dataset with over 3 million samples across 14 tasks; and LlaSMol, a set of models fine-tuned on it. Among them, the Mistral-based LlaSMol outperforms GPT-4 and Claude 3 Opus by a wide margin, approaching task-specific model performance while tuning just 0.58% of parameters, demonstrating the power of domain-specific instruction tuning.	 <p>Figure 5.6.2 Source: Yu et al., 2024</p>

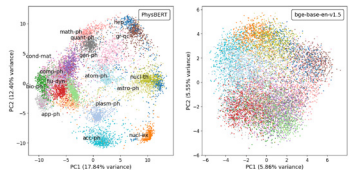

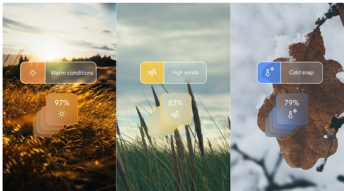
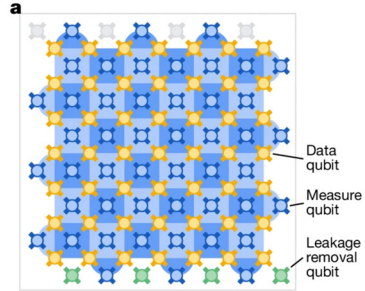
Highlight:

Notable Model Releases (cont'd)

Apr 23, 2024	ORBIT	Earth science	Oak Ridge National Lab introduced ORBIT, a 113B-parameter vision transformer and the largest AI model ever built for climate science—1,000 times larger than prior models. Trained using a novel parallelism technique and tested on the Frontier supercomputer, ORBIT achieved up to 1.6 exaFLOPS of sustained performance. This breakthrough sets a new bar for AI-driven Earth system prediction.	 <p>Figure 5.6.3 Source: Wang et al., 2024</p>
May 20, 2024	Aurora	Earth science	Aurora is a large-scale foundation model trained on over a million hours of Earth system data, delivering state-of-the-art forecasts for air quality, ocean waves, cyclone tracks, and high-resolution weather. It outperforms traditional systems while operating at a fraction of the computational cost, and can be fine-tuned across domains with minimal resources—marking a major step toward accessible, AI-driven Earth system forecasting.	 <p>Figure 5.6.4 Source: Bodnar et al., 2024</p>
Jul 22, 2024	NeuralGCM	Weather forecasting	This study introduces NeuralGCM, a hybrid model that combines a differentiable, physics-based solver with machine learning components to simulate both weather and climate. It matches or exceeds leading ML and physics-based models in short- and medium-term forecasts, accurately tracks climate metrics over decades, and captures complex phenomena like tropical cyclones—all while offering massive computational savings.	 <p>Figure 5.6.5 Source: Kochkov et al., 2024</p>

Highlight:

Notable Model Releases (cont'd)

Aug 18, 2024	PhysBERT	Physics	Physics texts are notoriously difficult for NLP due to their specialized language and complex concepts. PhysBERT, the first physics-specific, text-embedding model, addresses this by outperforming general-purpose models on physics-specific tasks. Trained on 1.2 million arXiv papers and fine-tuned with supervised data, it significantly boosts performance in information retrieval and subdomain fine-tuning.	 <p>Figure 5.6.6 Source: Hellert et al., 2024</p>
Sep 16, 2024	FireSat	Fire prediction	Google's FireSat is a satellite-based wildfire detection system that uses AI to identify fires as small as 5x5 meters within 20 minutes of ignition by analyzing real-time imagery and environmental data. Developed in partnership with Earth Fire Alliance and Muon Space, it not only enhances disaster response but also advances global wildfire research.	 <p>Figure 5.6.7 Source: Google, 2024</p>
Dec 4, 2024	GenCast	Weather prediction	Google DeepMind's GenCast is an AI-powered weather model that delivers highly accurate 15-day forecasts using a diffusion-based approach, outperforming traditional systems like the ENS on nearly all metrics. It generates forecasts in minutes instead of hours and has broad applications in disaster response, renewable energy, and agriculture.	 <p>Figure 5.6.8 Source: Google, 2024</p>
Dec 9, 2024	AlphaQubit	Quantum computing	In late 2024, Google DeepMind and Google Quantum AI released AlphaQubit, an AI-based decoder with state-of-the-art quantum error detection. Soon after, they introduced Willow, the first quantum chip to achieve exponential error suppression and correction below the surface code threshold—a major milestone in the field. Willow also completed a benchmark task in under five minutes that would take the fastest supercomputer over 10 septillion years, longer than the age of the known universe.	 <p>Figure 5.6.9 Source: Google, 2024</p>

Appendix

Acknowledgments

The AI Index would like to acknowledge Armin Hamrah for his work in surveying the literature on significant trends in AI-related science and medicine.

Benchmarks

1. **MedQA:** Data on MedQA was taken from the [MedQA Papers With Code leaderboard](#) in February 2025. To learn more about MedQA, please read the [original paper](#).

AI-Driven Protein Science Publications

The AI Index used [Dimensions](#)' AI document search function to measure the number of manuscripts published in a year. The searches were restricted to the 2024 publication year and the biological sciences category (987,717 publications). Then a search was conducted for each key term, which had to be present in both the title and the abstract. This requirement limited the number of manuscripts returned that might only have mentioned the key term in passing, rather than describing research about the key term. Once the number of manuscripts was identified, the percent of total biological sciences manuscripts about each key term was calculated.

Image and Multimodal AI for Scientific Discovery

The AI Index used Semantic Scholar and Google Scholar to measure the number of manuscripts published from 2023 to 2025. A search was then performed for each key term (e.g., “foundation models,” “microscopy,” “electron microscopy,” “fluorescence microscopy,” “light microscopy”) with the requirement that the terms be present in both the title and the abstract. Furthermore, the search was refined to strictly comply with the definition of a foundation model—specifically, a model trained on vast datasets that can be applied across a wide range of use cases. To this end, any

model alleged to be a foundation model that had been trained on fewer than 1 million data points or not evaluated on multiple tasks was discarded.

FDA-Approved AI Medical Devices

Data on FDA-approved AI medical devices was sourced from the [FDA website](#), which tracks artificial intelligence and machine learning (AI/ML)–enabled medical devices.

Ethical Considerations

The AI Index used PubMedCentral's API to query for English-language indexed articles published between Jan. 1, 2020, and Dec. 31, 2024, using search terms regarding artificial intelligence, medicine, and ethical issues. In order to obtain only articles at the intersection of those three topics, the AI Index further narrowed the articles to those with an abstract including a keyword related to: (a) artificial intelligence, (b) medicine, and (c) at least one ethical issue. After removing preprints, retracted articles, and articles that failed to satisfy the inclusion criteria, 2,916 articles remained. The AI Index used the frequency of ethical issues mentioned in abstracts across this pool of articles to conduct its analysis.

API query:

(“artificial intelligence”[MeSH] OR “machine learning”[MeSH] OR “deep learning”[All Fields] OR “AI”[All Fields] OR “ML”[All Fields] OR “predictive analytics”[All Fields]) AND ((“ethics”[MeSH] OR “ethical implications”[All Fields] OR “fair*”[All Fields] OR “unfair*”[All Fields] OR “bias”[All Fields] OR “accountability”[All Fields] OR “transparency”[All Fields] OR “explainability”[All Fields] OR “privacy”[All Fields] OR “trustworthy AI”[All Fields]) OR (“bioethics”[MeSH] OR “ELSI”[All Fields] OR “autonomy”[All Fields] OR “equity”[All Fields] OR “equitab*”[All Fields] OR “justice”[All Fields] OR “beneficence”[All Fields] OR “non-maleficence”[All Fields] OR “independent review”[All Fields] OR “oversight”[All

Chapter 5: Science and Medicine

Appendix

Fields] OR “racis*”[All Fields] OR “prejud*”[All Fields] OR “inequit*”[All Fields] OR “community engagement”[All Fields] OR “misuse”[All Fields] OR “dual use”[All Fields])) AND (“medicine”[MeSH] OR “medical AI”[All Fields] OR “clinical decision support”[All Fields] OR “health informatics”[All Fields]) AND (“2020/01/01”[PubDate] : “2024/12/31”[PubDate])

Date of search: 2/14/2025

Abstract inclusion criteria:

Therefore, includes only articles that discuss medicine, artificial intelligence, and at least one ethical issue within the abstract (N = 2,916).

- **AI keywords:** “artificial intelligence,” “AI,” “algorithm,” “ML,” “machine learning,” “deep learning,” predictive analytics.
- **Medicine keywords:** “medicine,” “medical,” “health,” “healthcare.”
- **Ethics keywords:** “ethic*,” “fairness,” “bias,” “accountability,” “transparency,” “explainability,” “privacy,” “trustworthy AI,” “bioethics,” “ELSI,” “autonomy,” “equit*,” “justice,” “beneficence,” “non-maleficence,” “independent review,” “oversight,” “racism,” “inequit*,” community engagement, misuse, dual use.

Works Cited

- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., ... Jumper, J. M. (2024). Accurate Structure Prediction of Biomolecular Interactions With AlphaFold 3. *Nature*, 630(8016), 493–500. <https://doi.org/10.1038/s41586-024-07487-w>
- Acharya, R., Abanin, D. A., Aghababaie-Beni, L., Aleiner, I., Andersen, T. I., Ansmann, M., Arute, F., Arya, K., Asfaw, A., Astrakhantsev, N., Atalaya, J., Babbush, R., Bacon, D., Ballard, B., Bardin, J. C., Bausch, J., Bengtsson, A., Bilmes, A., Blackwell, S., ... Google Quantum AI and Collaborators. (2025). Quantum Error Correction Below the Surface Code Threshold. *Nature*, 638(8052), 920–26. <https://doi.org/10.1038/s41586-024-08449-y>
- Blankemeier, L., Cohen, J. P., Kumar, A., Veen, D. V., Gardezi, S. J. S., Paschali, M., Chen, Z., Delbrouck, J.-B., Reis, E., Truys, C., Bluethgen, C., Jensen, M. E. K., Ostmeier, S., Varma, M., Valanarasu, J. M. J., Fang, Z., Huo, Z., Nabulsi, Z., Ardila, D., ... Chaudhari, A. S. (2024). *Merlin: A Vision Language Foundation Model for 3D Computed Tomography* (arXiv:2406.06512). arXiv. <https://doi.org/10.48550/arXiv.2406.06512>
- Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Vaughan, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., Gupta, J. K., Thambiratnam, K., Archibald, A. T., Wu, C.-C., Heider, E., Welling, M., Turner, R. E., & Perdikaris, P. (2024). *A Foundation Model for the Earth System* (arXiv:2405.13063). arXiv. <https://doi.org/10.48550/arXiv.2405.13063>
- Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods in Molecular Biology* (Clifton, N.J.), 1607, 627–41. https://doi.org/10.1007/978-1-4939-7000-1_26
- Callahan, A., McElfresh, D., Banda, J. M., Bunney, G., Char, D., Chen, J., Corbin, C. K., Dash, D., Downing, N. L., Jain, S. S., Kotecha, N., Masterson, J., Mello, M. M., Morse, K., Nallan, S., Pandya, A., Revri, A., Sharma, A., Sharp, C., ... Shah, N. H. (2024). Standing on FURM Ground: A Framework for Evaluating Fair, Useful, and Reliable AI Models in Health Care Systems. *NEJM Catalyst*, 5(10), CAT.24.0131. <https://doi.org/10.1056/CAT.24.0131>
- Campanella, G., Chen, S., Verma, R., Zeng, J., Stock, A., Croken, M., Veremis, B., Elmas, A., Huang, K., Kwan, R., Houldsworth, J., Schoenfeld, A. J., & Vanderbilt, C. (2024). *A Clinical Benchmark of Public Self-Supervised Pathology Foundation Models* (arXiv:2407.06508). arXiv. <https://doi.org/10.48550/arXiv.2407.06508>
- Carrillo-Perez, F., Pizurica, M., Zheng, Y., Nandi, T. N., Madduri, R., Shen, J., & Gevaert, O. (2023). RNA-to-Image Multi-cancer Synthesis Using Cascaded Diffusion Models. *bioRxiv: The Preprint Server for Biology*, 2023.01.13.523899. <https://doi.org/10.1101/2023.01.13.523899>
- Chambon, P., Bluethgen, C., Delbrouck, J.-B., Sluijs, R. V. der, Połacin, M., Chaves, J. M. Z., Abraham, T. M., Purohit, S., Langlotz, C. P., & Chaudhari, A. (2022). *RoentGen: Vision-Language Foundation Model for Chest X-ray Generation* (arXiv:2211.12737). arXiv. <https://doi.org/10.48550/arXiv.2211.12737>

Chambon, P., Delbrouck, J.-B., Sounack, T., Huang, S.-C., Chen, Z., Varma, M., Truong, S. Q., Chuong, C. T., & Langlotz, C. P. (2024). *CheXpert Plus: Augmenting a Large Chest X-ray Dataset With Text Radiology Reports, Patient Demographics and Additional Image Formats* (arXiv:2405.19538). arXiv. <https://doi.org/10.48550/arXiv.2405.19538>

Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., & Mahmood, F. (2022). *Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning* (arXiv:2206.02647). arXiv. <https://doi.org/10.48550/arXiv.2206.02647>

Chen, Z., Varma, M., Xu, J., Paschali, M., Veen, D. V., Johnston, A., Youssef, A., Blankemeier, L., Bluethgen, C., Altmayer, S., Valanarasu, J. M. J., Muneer, M. S. E., Reis, E. P., Cohen, J. P., Olsen, C., Abraham, T. M., Tsai, E. B., Beaulieu, C. F., Jitsev, J., ... Langlotz, C. P. (2024). *A Vision-Language Foundation Model to Enhance Efficiency of Chest X-ray Interpretation* (arXiv:2401.12208). arXiv. <https://doi.org/10.48550/arXiv.2401.12208>

Christensen, M., Vukadinovic, M., Yuan, N., & Ouyang, D. (2024). Vision–Language Foundation Model for Echocardiogram Interpretation. *Nature Medicine*, 30(5), 1481–88. <https://doi.org/10.1038/s41591-024-02959-y>

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., & Prior, F. (2013). The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6), 1045–57. <https://doi.org/10.1007/s10278-013-9622-7>

Ding, S., Li, J., Wang, J., Ying, S., & Shi, J. (2023). *Multi-scale Efficient Graph-Transformer for Whole Slide Image Classification* (arXiv:2305.15773). arXiv. <https://doi.org/10.48550/arXiv.2305.15773>

Ding, T., Wagner, S. J., Song, A. H., Chen, R. J., Lu, M. Y., Zhang, A., Vaidya, A. J., Jaume, G., Shaban, M., Kim, A., Williamson, D. F. K., Chen, B., Almagro-Perez, C., Doucet, P., Sahai, S., Chen, C., Komura, D., Kawabe, A., Ishikawa, S., ... Mahmood, F. (2024). *Multimodal Whole Slide Foundation Model for Pathology*(arXiv:2411.19666). arXiv. <https://doi.org/10.48550/arXiv.2411.19666>

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, 7(10), e2440969. <https://doi.org/10.1001/jamanetworkopen.2024.40969>

Goh, E., Gallo, R. J., Strong, E., Weng, Y., Kerman, H., Freed, J. A., Cool, J. A., Kanjee, Z., Lane, K. P., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Hom, J., Chen, J. H., & Rodman, A. (2025). GPT-4 Assistance for Improvement of Physician Performance on Patient Care Tasks: A Randomized Controlled Trial. *Nature Medicine*, 1–6. <https://doi.org/10.1038/s41591-024-03456-y>

Gruver, N., Sriram, A., Madotto, A., Wilson, A. G., Zitnick, C. L., & Ulissi, Z. (2024). *Fine-Tuned Language Models Generate Stable Inorganic Materials as Text* (arXiv:2402.04379). arXiv. <https://doi.org/10.48550/arXiv.2402.04379>

Guevara, M., Chen, S., Thomas, S., Chaunzwa, T. L., Franco, I., Kann, B. H., Moningi, S., Qian, J. M., Goldstein, M., Harper, S., Aerts, H. J. W. L., Catalano, P. J., Savova, G. K., Mak, R. H., & Bitterman, D. S. (2024). Large Language Models to Identify Social Determinants of Health in Electronic Health Records. *Npj Digital Medicine*, 7(1), 1–14. <https://doi.org/10.1038/s41746-023-00970-0>

- Guo, Z., Zhao, W., Wang, S., & Yu, L. (2023). *HIGT: Hierarchical Interaction Graph-Transformer for Whole Slide Image Analysis* (arXiv:2309.07400). arXiv. <https://doi.org/10.48550/arXiv.2309.07400>
- Haberle, T., Cleveland, C., Snow, G. L., Barber, C., Stookey, N., Thornock, C., Younger, L., Mullahkhel, B., & Ize-Ludlow, D. (2024). The Impact of Nuance DAX Ambient Listening AI Documentation: A Cohort Study. *Journal of the American Medical Informatics Association*, 31(4), 975–79. <https://doi.org/10.1093/jamia/ocae022>
- Hashmi, A. U. R., Almakky, I., Qazi, M. A., Sanjeev, S., Papineni, V. R., Jagdish, J., & Yaqub, M. (2024). *XReal: Realistic Anatomy and Pathology-Aware X-ray Generation via Controllable Diffusion Model* (arXiv:2403.09240). arXiv. <https://doi.org/10.48550/arXiv.2403.09240>
- Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y. A., Mishra, C., Kim, C., ... Rives, A. (2024). *Simulating 500 Million Years of Evolution With a Language Model* (p. 2024.07.01.600583). bioRxiv. <https://doi.org/10.1101/2024.07.01.600583>
- Hellert, T., Montenegro, J., & Pollastro, A. (2024). *PhysBERT: A Text Embedding Model for Physics Scientific Literature* (arXiv:2408.09574). arXiv. <https://doi.org/10.48550/arXiv.2408.09574>
- Hornick, T., Mao, C., Koynov, A., Yawman, P., Thool, P., Salish, K., Giles, M., Nagapudi, K., & Zhang, S. (2024). In Silico Formulation Optimization and Particle Engineering of Pharmaceutical Products Using a Generative Artificial Intelligence Structure Synthesis Method. *Nature Communications*, 15(1), 9622. <https://doi.org/10.1038/s41467-024-54011-9>
- Istasy, P., Lee, W. S., Iansavichene, A., Upshur, R., Gyawali, B., Burkell, J., Sadikovic, B., Lazo-Langner, A., & Chin-Yee, B. (2022). The Impact of Artificial Intelligence on Health Equity in Oncology: Scoping Review. *Journal of Medical Internet Research*, 24(11), e39748. <https://doi.org/10.2196/39748>
- Jiang, J. X., Qi, K., Bai, G., & Schulman, K. (2023). Pre-pandemic Assessment: A Decade of Progress in Electronic Health Record Adoption Among U.S. Hospitals. *Health Affairs Scholar*, 1(5), qxad056. <https://doi.org/10.1093/haschl/qxad056>
- Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2020). *What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset From Medical Exams* (arXiv:2009.13081). arXiv. <https://doi.org/10.48550/arXiv.2009.13081>
- Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., & Horng, S. (2019). MIMIC-CXR, a De-identified Publicly Available Database of Chest Radiographs With Free-Text Reports. *Scientific Data*, 6(1), 317. <https://doi.org/10.1038/s41597-019-0322-0>
- Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., & Hoyer, S. (2024). Neural General Circulation Models for Weather and Climate. *Nature*, 632(8027), 1060–66. <https://doi.org/10.1038/s41586-024-07744-y>

- Kudiabor, H. (2024). Virtual Lab Powered by ‘AI Scientists’ Super-Charges Biomedical Research. *Nature*, 636(8043), 532–33. <https://doi.org/10.1038/d41586-024-01684-3>
- Kumar, A., Kriz, A., Havaei, M., & Arbel, T. (2025). *PRISM: High-Resolution & Precise Counterfactual Medical Image Generation Using Language-Guided Stable Diffusion* (arXiv:2503.00196). arXiv. <https://doi.org/10.48550/arXiv.2503.00196>
- Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Zhao, M., Chow, A. K., Ikemura, K., Kim, A., Pouli, D., Patel, A., Soliman, A., Chen, C., Ding, T., Wang, J. J., Gerber, G., Liang, I., Le, L. P., Parwani, A. V., Weishaupt, L. L., & Mahmood, F. (2024). A Multimodal Generative AI Copilot for Human Pathology. *Nature*, 634(8033), 466–73. <https://doi.org/10.1038/s41586-024-07618-3>
- Lutsker, G., Sapir, G., Shilo, S., Merino, J., Godneva, A., Greenfield, J. R., Samocha-Bonet, D., Dhir, R., Gude, F., Mannor, S., Meirom, E., Chechik, G., Rossman, H., & Segal, E. (2025). *From Glucose Patterns to Health Outcomes: A Generalizable Foundation Model for Continuous Glucose Monitor Data Analysis* (arXiv:2408.11876). arXiv. <https://doi.org/10.48550/arXiv.2408.11876>
- Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment Anything in Medical Images. *Nature Communications*, 15(1), 654. <https://doi.org/10.1038/s41467-024-44824-z>
- Ma, S. P., Liang, A. S., Shah, S. J., Smith, M., Jeong, Y., Devon-Sand, A., Crowell, T., Delahaie, C., Hsia, C., Lin, S., Shanafelt, T., Pfeffer, M. A., Sharp, C., & Garcia, P. (2025). Ambient Artificial Intelligence Scribes: Utilization and Impact on Documentation Time. *Journal of the American Medical Informatics Association*, 32(2), 381–85. <https://doi.org/10.1093/jamia/ocae304>
- Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., & Naik, N. (2023). Large Language Models Generate Functional Protein Sequences Across Diverse Families. *Nature Biotechnology*, 41(8), 1099–1106. <https://doi.org/10.1038/s41587-022-01618-2>
- Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., ... Kopp-Schneider, A. (2018). Why Rankings of Biomedical Image Analysis Competitions Should Be Interpreted With Care. *Nature Communications*, 9(1), 5217. <https://doi.org/10.1038/s41467-018-07619-7>
- Mei, X., Liu, Z., Robson, P. M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K. E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z. A., & Yang, Y. (2022). RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. *Radiology: Artificial Intelligence*, 4(5), e210315. <https://doi.org/10.1148/ryai.210315>
- Narayanan, S., Braza, J. D., Griffiths, R.-R., Ponnampati, M., Bou, A., Laurent, J., Kabeli, O., Wellawatte, G., Cox, S., Rodrigues, S. G., & White, A. D. (2024). *Aviary: Training Language Agents on Challenging Scientific Tasks* (arXiv:2412.21154). arXiv. <https://doi.org/10.48550/arXiv.2412.21154>
- Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S. M., Ness, R. O., Poon, H., Qin, T., Usuyama, N., White, C., & Horvitz, E. (2023). *Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine* (arXiv:2311.16452). arXiv. <https://doi.org/10.48550/arXiv.2311.16452>

- Nori, H., Usuyama, N., King, N., McKinney, S. M., Fernandes, X., Zhang, S., & Horvitz, E. (2024). *From Medprompt to o1: Exploration of Run-Time Strategies for Medical Challenge Problems and Beyond* (arXiv:2411.03590). arXiv. <https://doi.org/10.48550/arXiv.2411.03590>
- Pokharel, S., Pratyush, P., Heinzinger, M., Newman, R. H., & Kc, D. B. (2022). Improving Protein Succinylation Sites Prediction Using Embeddings From Protein Language Model. *Scientific Reports*, 12(1), 16933. <https://doi.org/10.1038/s41598-022-21366-2>
- Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., & Willson, M. (2025). Probabilistic Weather Forecasting With Machine Learning. *Nature*, 637(8044), 84–90. <https://doi.org/10.1038/s41586-024-08252-9>
- Qian, Z., Callender, T., Cebere, B., Janes, S. M., Navani, N., & van der Schaar, M. (2024). Synthetic Data for Privacy-Preserving Clinical Risk Prediction. *Scientific Reports*, 14(1), 25676. <https://doi.org/10.1038/s41598-024-72894-y>
- Qiu, J., Wu, J., Wei, H., Shi, P., Zhang, M., Sun, Y., Li, L., Liu, H., Liu, H., Hou, S., Zhao, Y., Shi, X., Xian, J., Qu, X., Zhu, S., Pan, L., Chen, X., Zhang, X., Jiang, S., ... Yuan, W. (2024). Development and Validation of a Multimodal Multitask Vision Foundation Model for Generalist Ophthalmic Artificial Intelligence. *NEJM AI*, 1(12), Aloa2300221. <https://doi.org/10.1056/Aloa2300221>
- Quer, G., & Topol, E. J. (2024). The Potential for Large Language Models to Transform Cardiovascular Medicine. *The Lancet Digital Health*, 6(10), e767–71. [https://doi.org/10.1016/S2589-7500\(24\)00151-1](https://doi.org/10.1016/S2589-7500(24)00151-1)
- Rashidi, H. H., Albahra, S., Rubin, B. P., & Hu, B. (2024). A Novel and Fully Automated Platform for Synthetic Tabular Data Generation and Validation. *Scientific Reports*, 14(1), 23312. <https://doi.org/10.1038/s41598-024-73608-0>
- Shah, S. J., Devon-Sand, A., Ma, S. P., Jeong, Y., Crowell, T., Smith, M., Liang, A. S., Delahaie, C., Hsia, C., Shanafelt, T., Pfeffer, M. A., Sharp, C., Lin, S., & Garcia, P. (2025). Ambient Artificial Intelligence Scribes: Physician Burnout and Perspectives on Usability and Documentation Burden. *Journal of the American Medical Informatics Association*, 32(2), 375–80. <https://doi.org/10.1093/jamia/ocae295>
- Shapson-Coe, A., Januszewski, M., Berger, D. R., Pope, A., Wu, Y., Blakely, T., Schalek, R. L., Li, P. H., Wang, S., Maitin-Shepard, J., Karlupia, N., Dorkenwald, S., Sjostedt, E., Leavitt, L., Lee, D., Troidl, J., Collman, F., Bailey, L., Fitzmaurice, A., ... Lichtman, J. W. (2024). A Petavoxel Fragment of Human Cerebral Cortex Reconstructed at Nanoscale Resolution. *Science*, 384(6696), eadk4858. <https://doi.org/10.1126/science.adk4858>
- Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., & Bakas, S. (2020). Federated Learning in Medicine: Facilitating Multi-institutional Collaborations Without Sharing Patient Data. *Scientific Reports*, 10(1), 12598. <https://doi.org/10.1038/s41598-020-69250-1>
- Shi, J., Tang, L., Gao, Z., Li, Y., Wang, C., Gong, T., Li, C., & Fu, H. (2023). MG-Trans: Multi-scale Graph Transformer With Information Bottleneck for Whole Slide Image Classification. *IEEE Transactions on Medical Imaging*, 42(12), 3871–83. <https://doi.org/10.1109/TMI.2023.3313252>

- Snel, B., Lehmann, G., Bork, P., & Huynen, M. A. (2000). STRING: A Web-Server to Retrieve and Display the Repeatedly Occurring Neighbourhood of a Gene. *Nucleic Acids Research*, 28(18), 3442–44. <https://doi.org/10.1093/nar/28.18.3442>
- Snowdon, J. L., Scheufele, E. L., Pritts, J., Le, P.-T., Mensah, G. A., Zhang, X., & Dankwa-Mullan, I. (2023). Evaluating Social Determinants of Health Variables in Advanced Analytic and Artificial Intelligence Models for Cardiovascular Disease Risk and Outcomes: A Targeted Review. *Ethnicity & Disease*, 33(1), 33–43. <https://doi.org/10.18865/1704>
- Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation. *Npj Mental Health Research*, 3(1), 1–12. <https://doi.org/10.1038/s44184-024-00056-z>
- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3), e1001779. <https://doi.org/10.1371/journal.pmed.1001779>
- Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Kipnis, P., Liu, V., & Lee, K. (2024). Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. *NEJM Catalyst*, 5(3), CAT.23.0404. <https://doi.org/10.1056/CAT.23.0404>
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Židek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., ... Velankar, S. (2022). AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space With High-Accuracy Models. *Nucleic Acids Research*, 50(D1), D439–44. <https://doi.org/10.1093/nar/gkab1061>
- Veitch, D. P., Weiner, M. W., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Morris, J. C., Petersen, R. C., Saykin, A. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., & Alzheimer's Disease Neuroimaging Initiative. (2019). Understanding Disease Progression and Improving Alzheimer's Disease Clinical Trials: Recent Highlights From the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 15(1), 106–52. <https://doi.org/10.1016/j.jalz.2018.08.005>
- Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., Yang, E., Mathieu, P., van Eck, A., Lee, D., Viret, J., Robert, E., Wang, Y. K., Kunz, J. D., Lee, M. C. H., ... Fuchs, T. J. (2024). A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection. *Nature Medicine*, 30(10), 2924–35. <https://doi.org/10.1038/s41591-024-03141-0>
- Wang, R., Fang, X., Lu, Y., & Wang, S. (2004). The PDBbind Database: Collection of Binding Affinities for Protein–Ligand Complexes With Known Three-Dimensional Structures. *Journal of Medicinal Chemistry*, 47(12), 2977–80. <https://doi.org/10.1021/jm030580l>
- Wang, X., Liu, S., Tsaris, A., Choi, J.-Y., Aji, A., Fan, M., Zhang, W., Yin, J., Ashfaq, M., Lu, D., & Balaprakash, P. (2024). ORBIT: Oak Ridge Base Foundation Model for Earth System Predictability (arXiv:2404.14712). arXiv. <https://doi.org/10.48550/arXiv.2404.14712>

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., & Han, X. (2022a). Transformer-Based Unsupervised Contrastive Learning for Histopathological Image Classification. *Medical Image Analysis*, 81, 102559. <https://doi.org/10.1016/j.media.2022.102559>

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., & Han, X. (2022b). Transformer-Based Unsupervised Contrastive Learning for Histopathological Image Classification. *Medical Image Analysis*, 81, 102559. <https://doi.org/10.1016/j.media.2022.102559>

Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., Wang, F., Peng, Y., Zhu, J., Zhang, J., Jackson, C. R., Zhang, J., Dillon, D., Lin, N. U., Sholl, L., ... Yu, K.-H. (2024). A Pathology Foundation Model for Cancer Diagnosis and Prognosis Prediction. *Nature*, 634(8035), 970–78. <https://doi.org/10.1038/s41586-024-07894-z>

Wang, Y., He, J., Du, Y., Chen, X., Li, J. C., Liu, L.-P., Xu, X., & Hassoun, S. (2025). *Large Language Model Is Secretly a Protein Sequence Optimizer* (arXiv:2501.09274). arXiv. <https://doi.org/10.48550/arXiv.2501.09274>

Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., Yu, K.-H., Willens, S., Olguin, F. M., Nirschl, J. J., Neal, J., Diehn, M., Yang, S., & Li, R. (2025). A Vision–Language Foundation Model for Precision Oncology. *Nature*, 638(8051), 769–78. <https://doi.org/10.1038/s41586-024-08378-w>

Xie, Y., Wu, J., Tu, H., Yang, S., Zhao, B., Zong, Y., Jin, Q., Xie, C., & Zhou, Y. (2024). *A Preliminary Study of o1 in Medicine: Are We Closer to an AI Doctor?* (arXiv:2409.15277). arXiv. <https://doi.org/10.48550/arXiv.2409.15277>

Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., ... Poon, H. (2024). A Whole-Slide Foundation Model for Digital Pathology From Real-World Data. *Nature*, 630(8015), 181–88. <https://doi.org/10.1038/s41586-024-07441-w>

Yang, L., Xu, S., Sellergren, A., Kohlberger, T., Zhou, Y., Ktena, I., Kiraly, A., Ahmed, F., Hormozdiari, F., Jaroensri, T., Wang, E., Wulczyn, E., Jamil, F., Guidroz, T., Lau, C., Qiao, S., Liu, Y., Goel, A., Park, K., ... Golden, D. (2024). *Advancing Multimodal Medical Capabilities of Gemini* (arXiv:2405.03162). arXiv. <https://doi.org/10.48550/arXiv.2405.03162>

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). *GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records* (arXiv:2203.03540). arXiv. <https://doi.org/10.48550/arXiv.2203.03540>

Yu, B., Baker, F. N., Chen, Z., Ning, X., & Sun, H. (2024). *LlaSMol: Advancing Large Language Models for Chemistry With a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset* (arXiv:2402.09391). arXiv. <https://doi.org/10.48550/arXiv.2402.09391>

Zambaldi, V., La, D., Chu, A. E., Patani, H., Danson, A. E., Kwan, T. O. C., Frerix, T., Schneider, R. G., Saxton, D., Thillaisundaram, A., Wu, Z., Moraes, I., Lange, O., Papa, E., Stanton, G., Martin, V., Singh, S., Wong, L. H., Bates, R., ... Wang, J. (2024). *De Novo Design of High-Affinity Protein Binders with AlphaProteo* (arXiv:2409.08022). arXiv. <https://doi.org/10.48550/arXiv.2409.08022>

Chapter 5: Science and Medicine

Appendix

Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H. H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J., Moungh-Wen, C., Piening, B., Bifulco, C., Wei, M., Poon, H., & Wang, S. (2025). A Foundation Model for Joint Segmentation, Detection and Recognition of Biomedical Objects Across Nine Modalities. *Nature Methods*, 22(1), 166–76. <https://doi.org/10.1038/s41592-024-02499-w>

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., Liu, T., Xu, M., Lozano, M. G., Woodward-Court, P., Kihara, Y., Altmann, A., Lee, A. Y., Topol, E. J., Denniston, A. K., Alexander, D. C., & Keane, P. A. (2023). A Foundation Model for Generalizable Disease Detection From Retinal Images. *Nature*, 622(7981), 156–63. <https://doi.org/10.1038/s41586-023-06555-x>