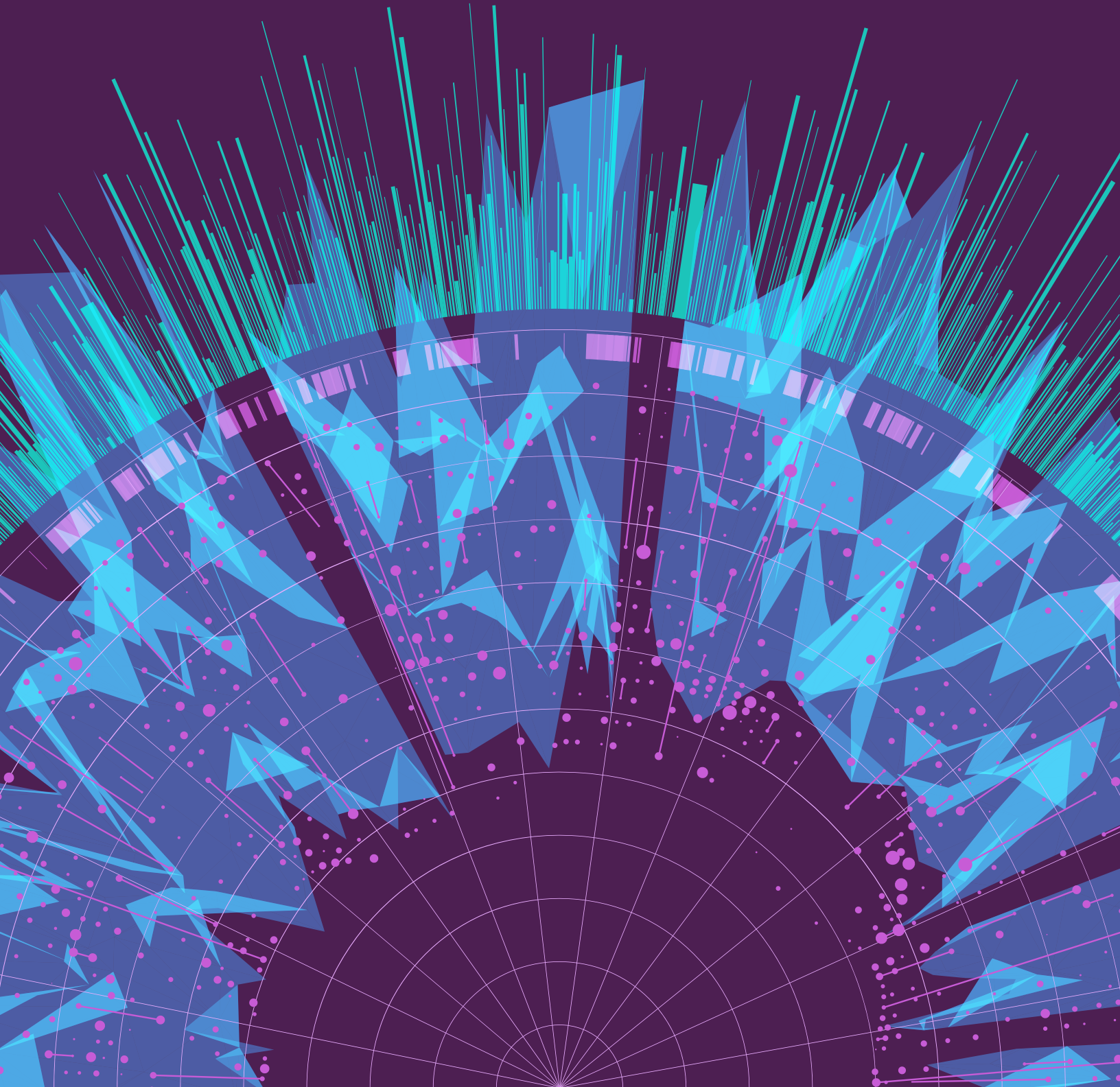




Artificial Intelligence  
Index Report 2024

# CHAPTER 5: Science and Medicine



# Preview

---

Overview	3
Chapter Highlights	4
<b>5.1 Notable Scientific Milestones</b>	<b>5</b>
AlphaDev	5
FlexiCubes	6
Synbot	8
GraphCast	9
GNoME	10
Flood Forecasting	11
<b>5.2 AI in Medicine</b>	<b>12</b>
Notable Medical Systems	12
SynthSR	12
Coupled Plasmonic Infrared Sensors	14
EVEscape	15
AlphaMissence	17
Human Pangenome Reference	18
Clinical Knowledge	19
MedQA	19
<b>Highlighted Research:</b> GPT-4 Medprompt	20
<b>Highlighted Research:</b> MediTron-70B	22
Diagnosis	23
<b>Highlighted Research:</b> CoDoC	23
<b>Highlighted Research:</b> CT Panda	24
Other Diagnostic Uses	25
FDA-Approved AI-Related Medical Devices	26
Administration and Care	28
<b>Highlighted Research:</b> MedAlign	28
<b>Appendix</b>	<b>30</b>

ACCESS THE PUBLIC DATA

# Overview

---

This year's AI Index introduces a new chapter on AI in science and medicine in recognition of AI's growing role in scientific and medical discovery. It explores 2023's standout AI-facilitated scientific achievements, including advanced weather forecasting systems like GraphCast and improved material discovery algorithms like GNoME. The chapter also examines medical AI system performance, important 2023 AI-driven medical innovations like SynthSR and ImmunoSEIRA, and trends in the approval of FDA AI-related medical devices.

# Chapter Highlights

---

**1. Scientific progress accelerates even further, thanks to AI.** In 2022, AI began to advance scientific discovery. 2023, however, saw the launch of even more significant science-related AI applications—from AlphaDev, which makes algorithmic sorting more efficient, to GNoME, which facilitates the process of materials discovery.

---

**2. AI helps medicine take significant strides forward.** In 2023, several significant medical systems were launched, including EVEscape, which enhances pandemic prediction, and AlphaMissence, which assists in AI-driven mutation classification. AI is increasingly being utilized to propel medical advancements.

---

**3. Highly knowledgeable medical AI has arrived.** Over the past few years, AI systems have shown remarkable improvement on the MedQA benchmark, a key test for assessing AI's clinical knowledge. The standout model of 2023, GPT-4 Medprompt, reached an accuracy rate of 90.2%, marking a 22.6 percentage point increase from the highest score in 2022. Since the benchmark's introduction in 2019, AI performance on MedQA has nearly tripled.

---

**4. The FDA approves more and more AI-related medical devices.** In 2022, the FDA approved 139 AI-related medical devices, a 12.1% increase from 2021. Since 2012, the number of FDA-approved AI-related medical devices has increased by more than 45-fold. AI is increasingly being used for real-world medical purposes.

---

This section highlights significant AI-related scientific breakthroughs of 2023 as chosen by the AI Index Steering Committee.

# 5.1 Notable Scientific Milestones

## AlphaDev

### AlphaDev discovers faster sorting algorithms

AlphaDev is a new AI reinforcement learning system that has improved on decades of work by scientists and engineers in the field of computational algorithmic enhancement. AlphaDev developed algorithms with fewer instructions than existing human benchmarks for

fundamental sorting algorithms on short sequences such as Sort 3, Sort 4, and Sort 5 (Figure 5.1.1). Some of the new algorithms discovered by AlphaDev have been incorporated into the LLVM standard C++ sort library. This marks the first update to this part of the library in over 10 years and is the first addition designed using reinforcement learning.

### AlphaDev vs. human benchmarks when optimizing for algorithm length

Source: Mankowitz et al., 2023 | Chart: 2024 AI Index report

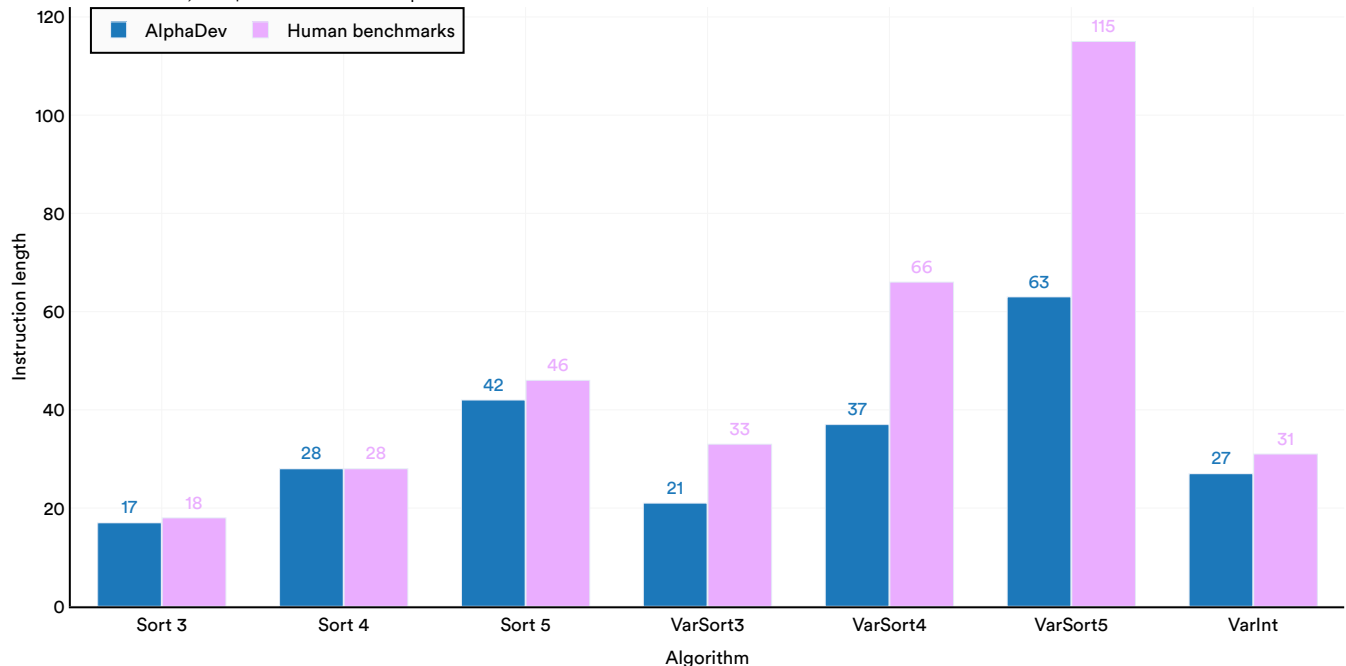


Figure 5.1.1

## FlexiCubes

### 3D mesh optimization with FlexiCubes

3D mesh generation, crucial in computer graphics, involves creating a mesh of vertices, edges, and faces to define 3D objects. It is key to video games, animation, medical imaging, and scientific visualization. Traditional isosurface extraction algorithms often struggle with limited resolution, structural rigidity, and numerical instabilities, which subsequently impacts

quality. FlexiCubes addresses some of these limitations by employing AI for gradient-based optimization and adaptable parameters (Figure 5.1.2). This method allows for precise, localized mesh adjustments. Compared to other leading methods that utilize differentiable isosurfacing for mesh reconstruction, FlexiCubes achieves mesh extractions that align much more closely with the underlying ground truth (Figure 5.1.3).

### Sample FlexiCubes surface reconstructions

Source: [Nvidia, 2023](#)

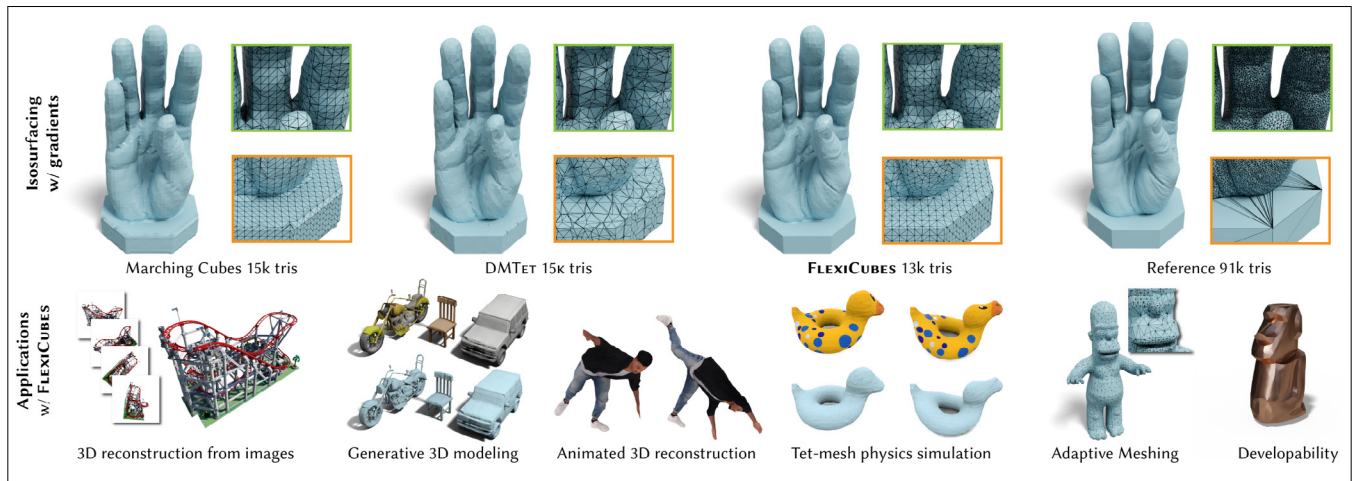


Figure 5.1.2

### Select quantitative results on 3D mesh reconstruction

Source: Shen et al., 2023 | Chart: 2024 AI Index report

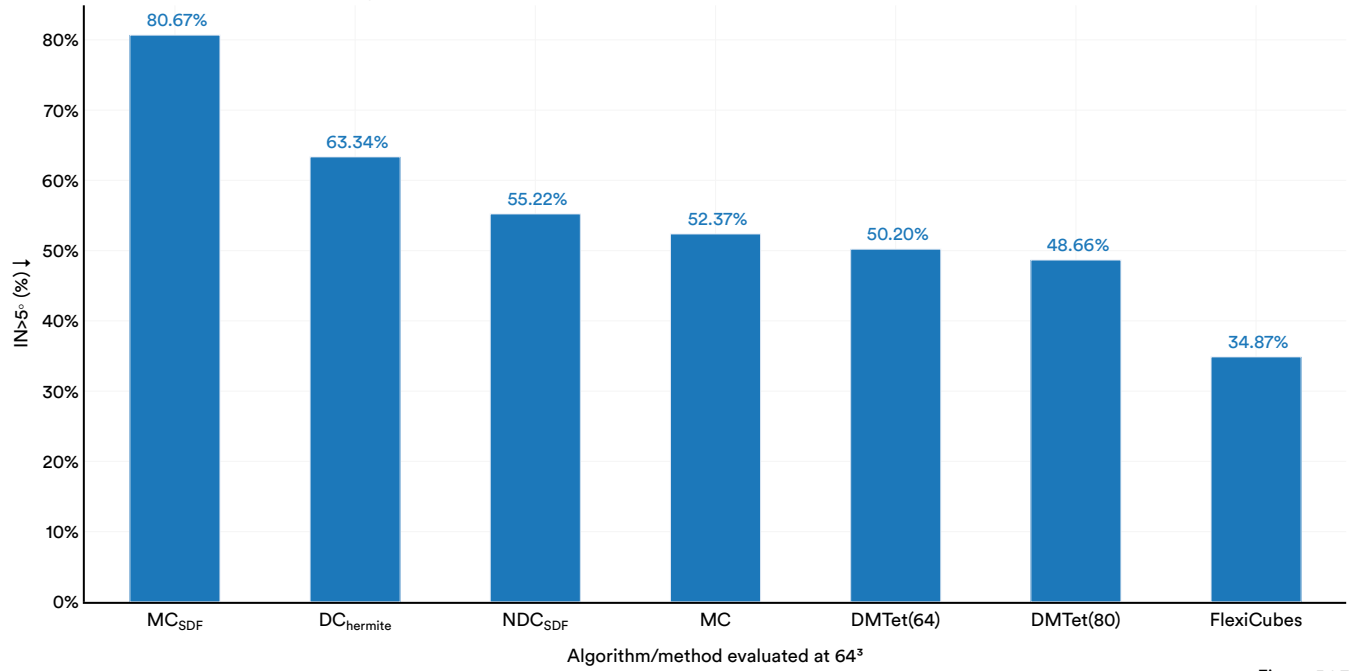


Figure 5.1.3

## Synbot

### AI-driven robotic chemist for synthesizing organic molecules

Synbot employs a multilayered system, comprising an AI software layer for chemical synthesis planning, a robot software layer for translating commands, and a physical robot layer for conducting experiments. The closed-loop feedback mechanism between the AI and the robotic system enables Synbot to develop synthetic recipes with yields equal to or exceeding established references (Figure 5.1.4). In an experiment aimed at synthesizing M1 [4-(2,3-dimethoxyphenyl)-1H-pyrrolo[2,3-b]pyridine], Synbot developed multiple synthetic formulas that achieved conversion yields surpassing

### Synbot design

Source: [Ha et al., 2023](#)

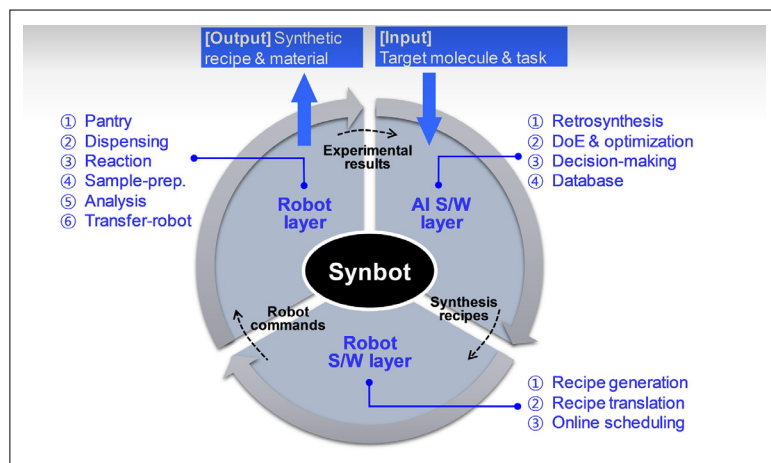


Figure 5.1.4

the mid-80% reference range and completed the synthesis in significantly less time (Figure 5.1.5). Synbot's automation of organic synthesis highlights AI's potential in fields such as pharmaceuticals and materials science.

### Reaction kinetics of M1 autonomous optimization experiment, Synbot vs. reference

Source: [Ha et al., 2023](#) | Chart: 2024 AI Index report

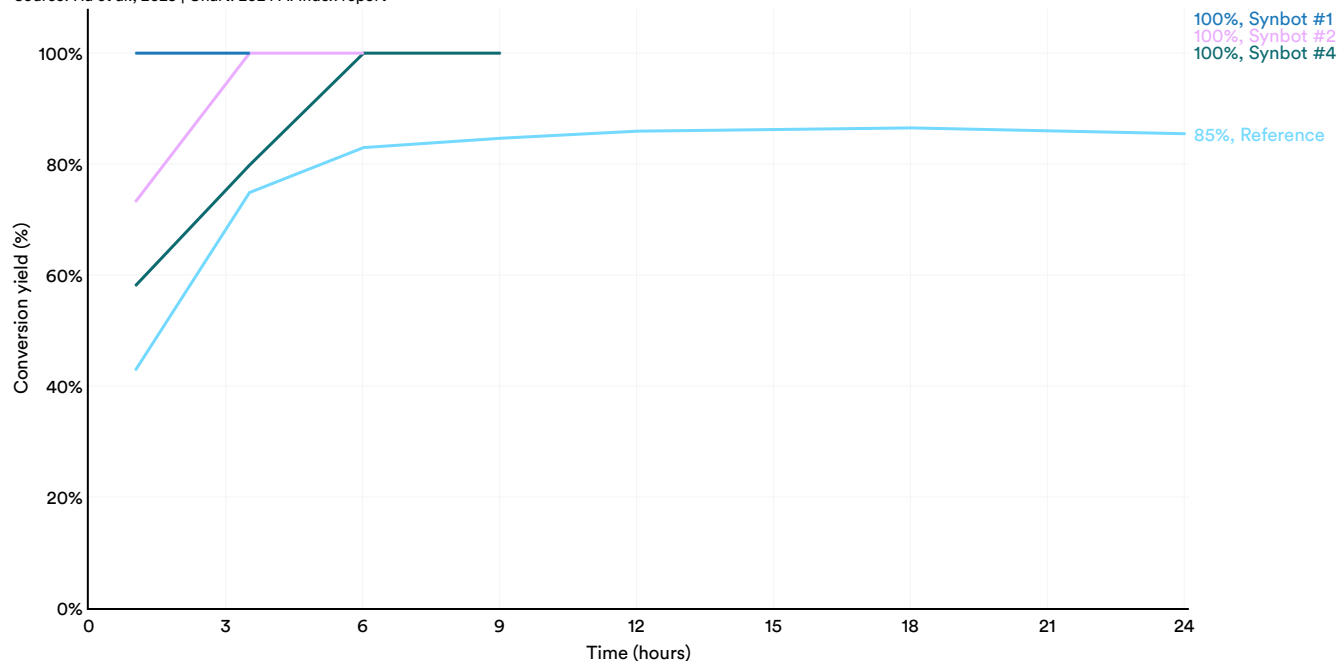


Figure 5.1.5



## GraphCast

### More accurate global weather forecasting with GraphCast

GraphCast is a new weather forecasting system that delivers highly accurate 10-day weather predictions in under a minute (Figure 5.1.6). Utilizing graph neural networks and machine learning, GraphCast processes vast datasets to forecast temperature, wind speed, atmospheric conditions,

and more. Figure 5.1.7 compares the performance of GraphCast with the current industry state-of-the-art weather simulation system: the High Resolution Forecast (HRES). GraphCast posts a lower root mean squared error, meaning its forecasts more closely correspond to observed weather patterns. GraphCast can be a valuable tool in deciphering weather patterns, enhancing preparedness for extreme weather events, and contributing to global climate research.

### GraphCast weather prediction

Source: DeepMind, 2023

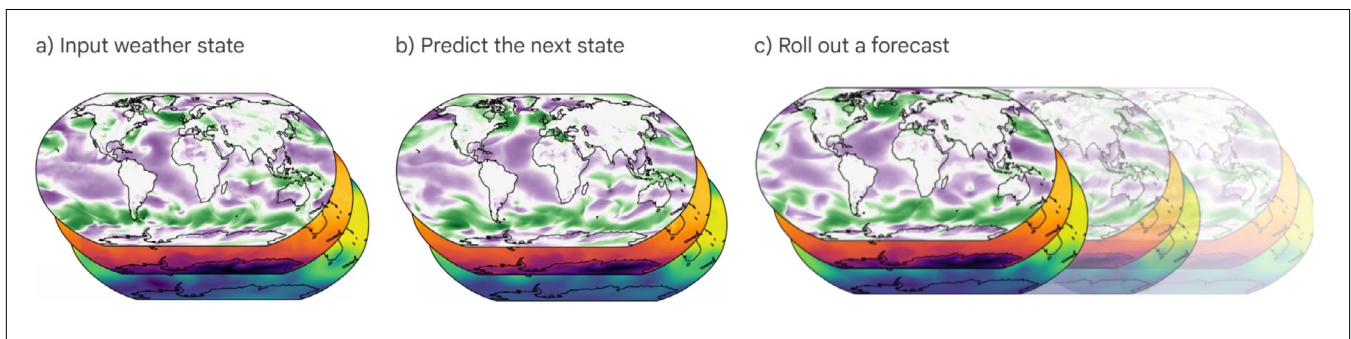


Figure 5.1.6

### Ten-day z500 forecast skill: GraphCast vs. HRES

Source: Lam et al., 2023 | Chart: 2024 AI Index report

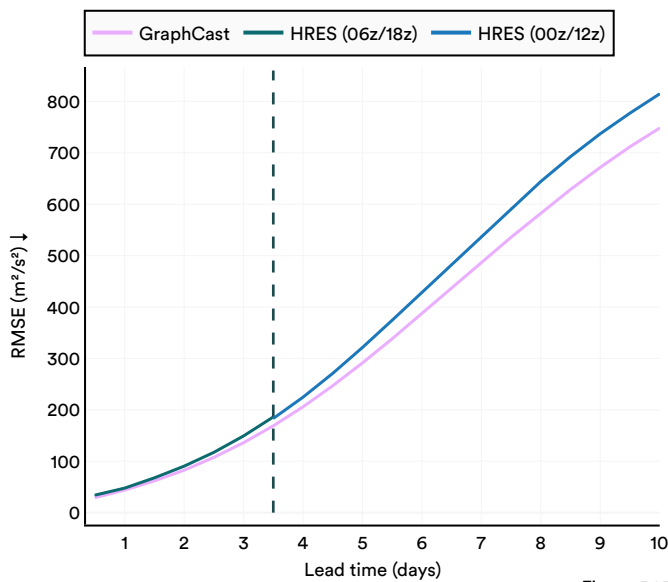


Figure 5.1.7

## GNoME

### Discovering new materials with GNoME

The search for new functional materials is key to advancements in various scientific fields, including robotics and semiconductor manufacturing. Yet this discovery process is typically expensive and slow. Recent advancements by Google researchers have demonstrated that graph networks, a type of AI model, can expedite this process when trained on large datasets. Their model, GNoME, outperformed the Materials Project, a leading method in materials discovery, by identifying a significantly larger number of stable crystals (Figure 5.1.8). GNoME has unveiled 2.2 million new crystal structures, many overlooked by human researchers (Figure 5.1.9 and Figure 5.1.10). The success of AI-driven projects like GNoME highlights the power of data and scaling in speeding up scientific breakthroughs.

### Sample material structures

Source: Merchant et al., 2023

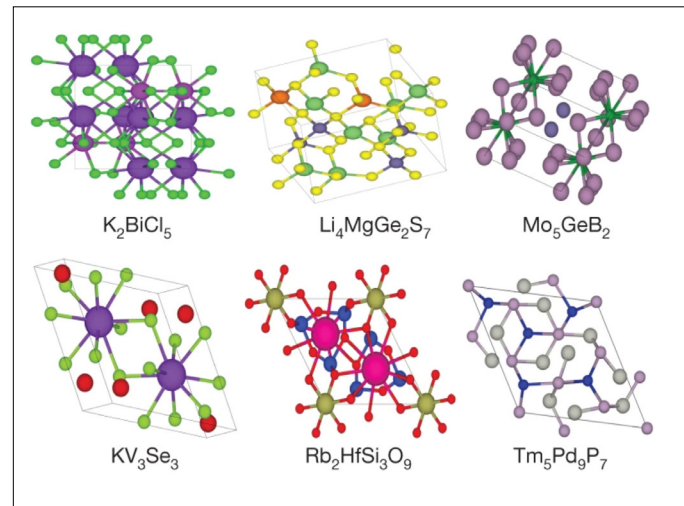
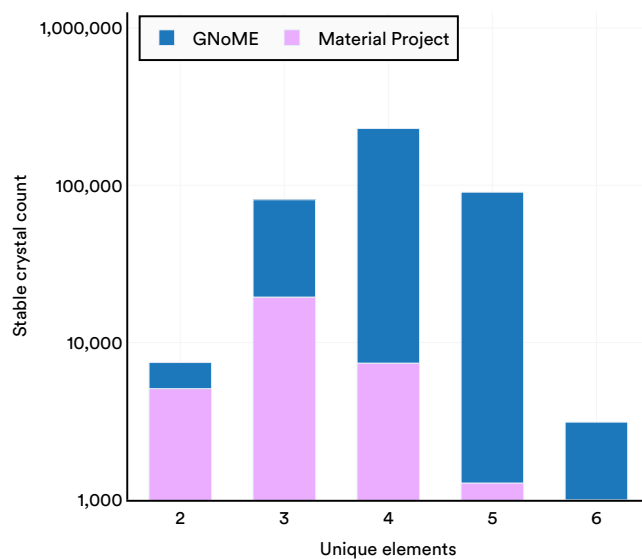


Figure 5.1.8

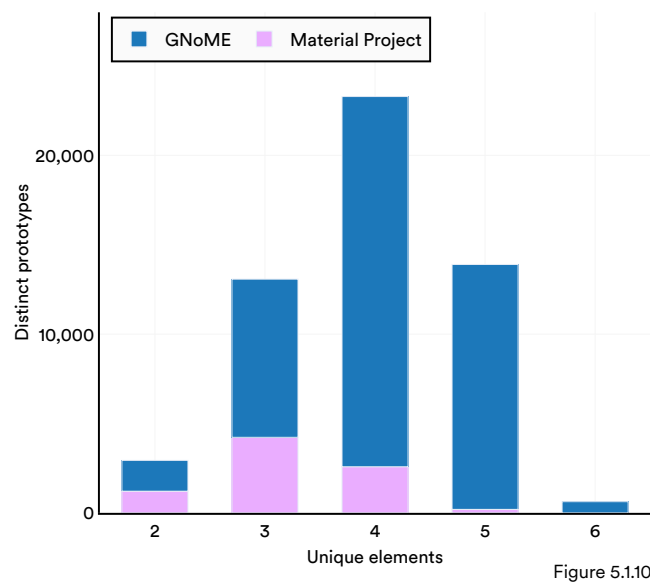
### GNoME vs. Materials Project: stable crystal count

Source: Merchant et al., 2023 | Chart: 2024 AI Index report



### GNoME vs. Materials Project: distinct prototypes

Source: Merchant et al., 2023 | Chart: 2024 AI Index report



## Flood Forecasting

### AI for more accurate and reliable flood forecasts

New research introduced in 2023 has made significant progress in predicting large-scale flood events. Floods, among the most common natural disasters, have particularly devastating effects in less developed countries where infrastructure for prevention and mitigation is lacking. Consequently, developing more accurate prediction methods that can forecast these events further in advance could yield substantial positive impacts.

A team of Google researchers has used AI to develop highly accurate hydrological simulation models that are also applicable to ungauged basins.<sup>1</sup> These innovative methods can predict certain extreme flood events up to five days in advance, with accuracy that matches or surpasses current state-of-the-art models, such as GloFAS. The AI model demonstrates superior precision (accuracy of positive predictions) and recall (ability to correctly identify all relevant instances) across a range of return period events, outperforming the leading contemporary method (Figure 5.1.11).<sup>2</sup> The model is open-source and is already being used to predict flood events in over 80 countries.

### Predictions of AI model vs. GloFAS across return periods

Source: Nearing et al., 2023 | Chart: 2024 AI Index report

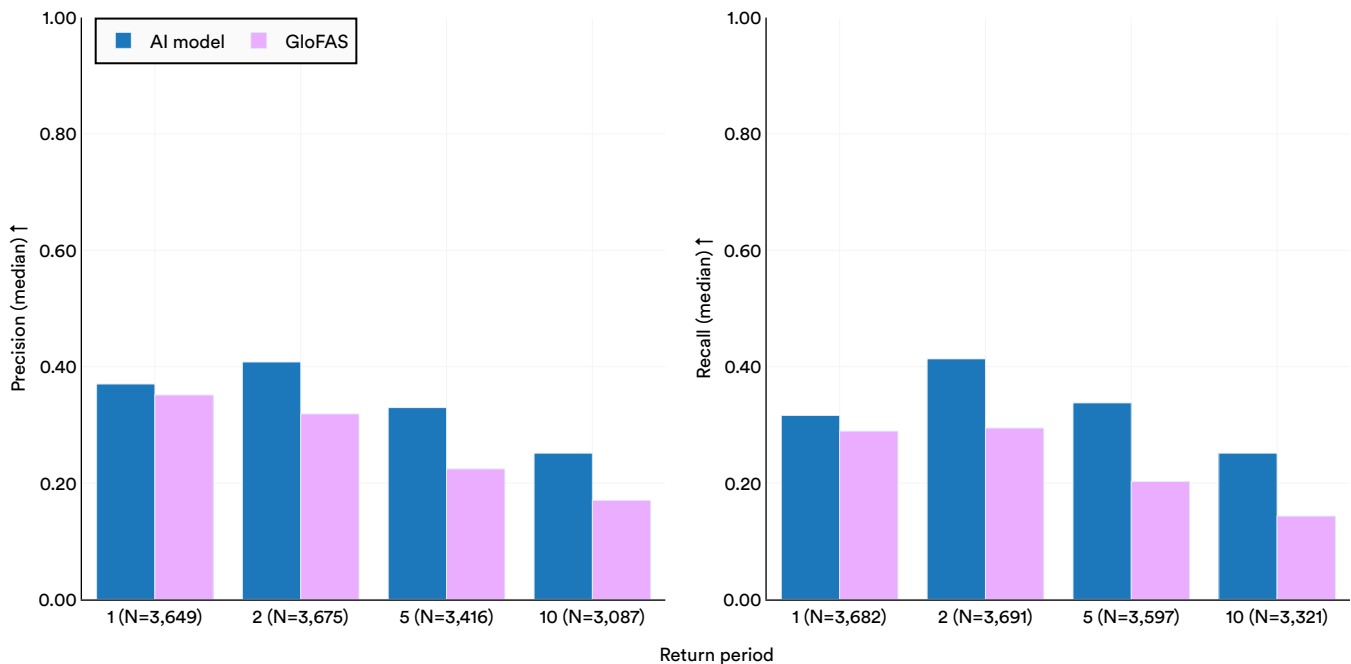


Figure 5.1.11

<sup>1</sup> An ungauged basin is a watershed for which there is insufficient streamflow data to model hydrological flows.

<sup>2</sup> A return period (recurrence interval) measures the likelihood of a particular hydrological event recurring within a specific period. For example, a 100-year flood means there is a 1% chance of the event being equaled or exceeded in any given year.

AI models are becoming increasingly valuable in healthcare, with applications for detecting polyps to aiding clinicians in making diagnoses. As AI performance continues to improve, monitoring its impact on medical practice becomes increasingly important. This section highlights significant AI-related medical systems introduced in 2023, the current state of clinical AI knowledge, and the development of new AI diagnostic tools and models aimed at enhancing hospital administration.

## 5.2 AI in Medicine

### Notable Medical Systems

This section identifies significant AI-related medical breakthroughs of 2023 as chosen by the AI Index Steering Committee.

#### SynthSR

##### Transforming brain scans for advanced analysis

SynthSR is an AI tool that converts clinical brain scans into high-resolution T-1 weighted images (Figure 5.2.1). This advancement addresses the issue of scan quality variability, which previously limited the use of many scans in advanced research. By transforming these scans into T1-weighted images, known for their high contrast and clear brain structure depiction, SynthSR facilitates the creation of detailed 3D brain renderings. Experiments using SynthSR demonstrate robust correlations between observed volumes at both scan and subject levels, suggesting that SynthSR generates images closely resembling those produced by high-resolution T1 scans. Figure 5.2.2 illustrates the extent to which SynthSR scans correspond with ground-truth observations across selected brain regions. SynthID significantly improves the visualization and analysis of brain structures, facilitating neuroscientific research and clinical diagnostics.

#### SynthSR generations

Source: [Iglesias et al., 2023](#)

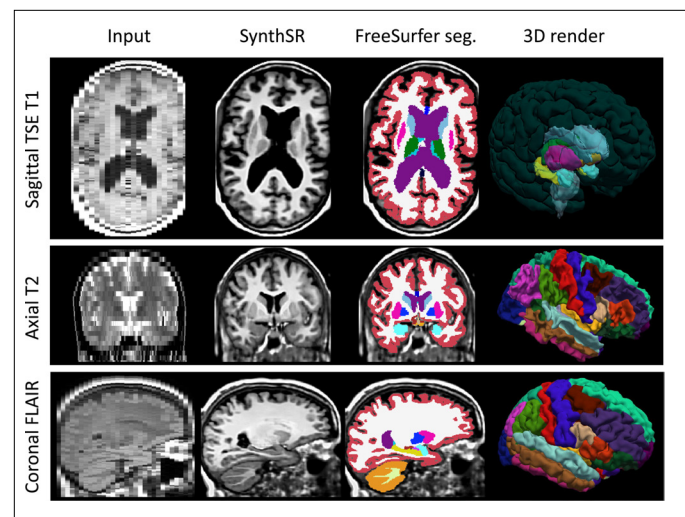


Figure 5.2.1

### SynthSR correlation with ground-truth volumes on select brain regions

Source: Iglesias et al., 2023 | Chart: 2024 AI Index report

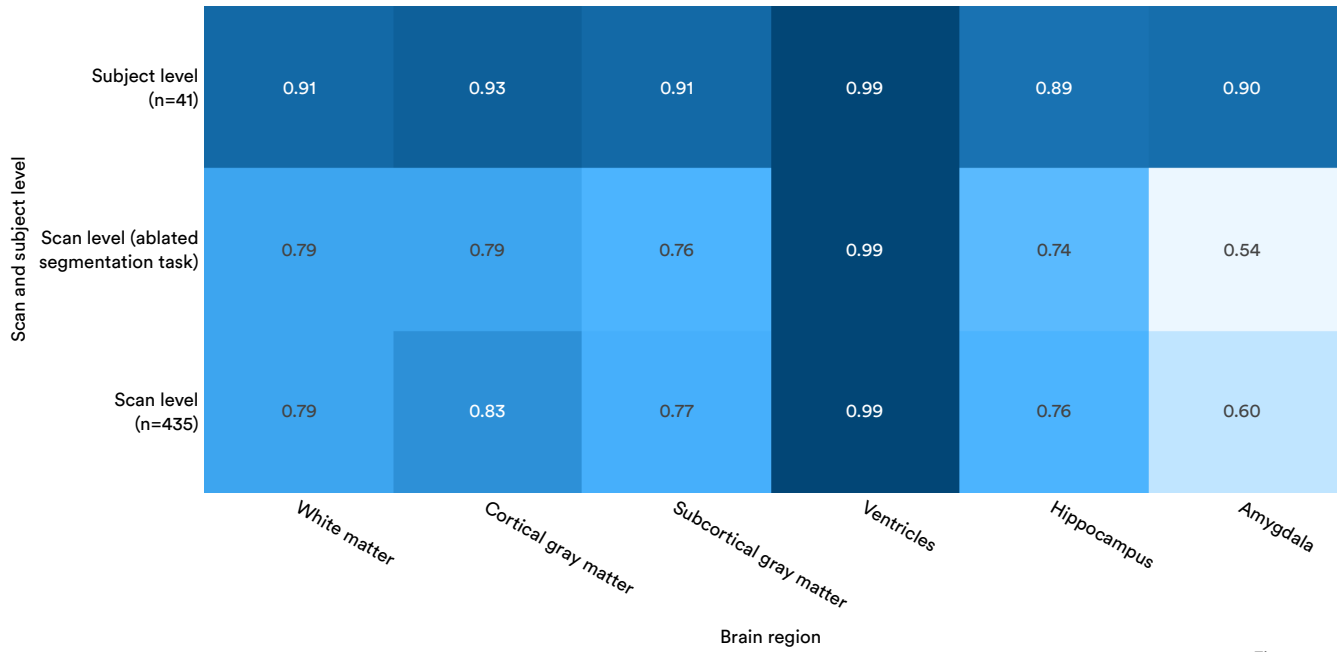


Figure 5.2.2

### Coupled Plasmonic Infrared Sensors

#### Coupled plasmonic infrared sensors for the detection of neurodegenerative diseases

Diagnosis of neurodegenerative diseases such as Parkinson’s and Alzheimer’s depends on fast and precise identification of biomarkers. Traditional methods, such as mass spectrometry and ELISA, are useful in that they can focus on quantifying protein levels; however, they cannot discern changes in structural states. This year, researchers uncovered a new method for neurodegenerative disease diagnosis that combined AI-coupled plasmonic infrared sensors that use Surface-Enhanced Infrared Absorption (SEIRA) spectroscopy with an immunoassay technique (ImmunoSEIRA; Figure 5.2.3). In tests that compared actual fibril percentages with predictions made by AI systems, the accuracy of the predictions was found to very closely match the actual reported percentages (Figure 5.2.4).

#### ImmunoSEIRA detection principle and the setup

Source: Kavungal et al., 2023

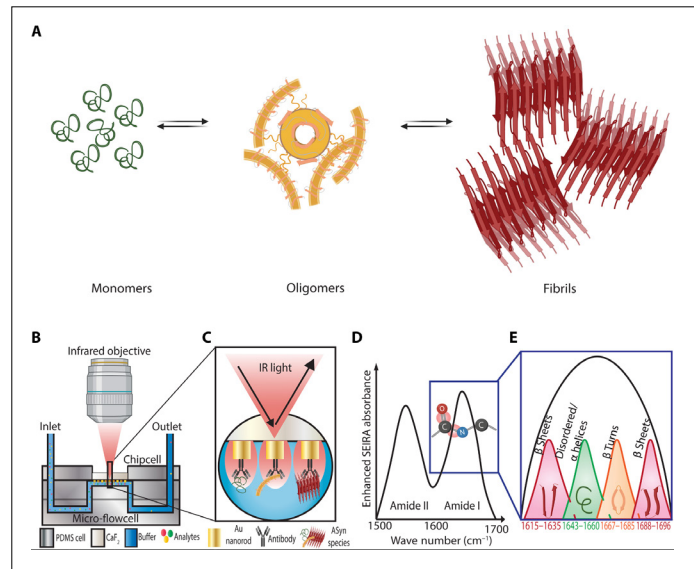


Figure 5.2.3

#### Deep neural network predicted vs. actual fibrils percentages in test samples

Source: Kavungal et al., 2023 | Chart: 2024 AI Index report

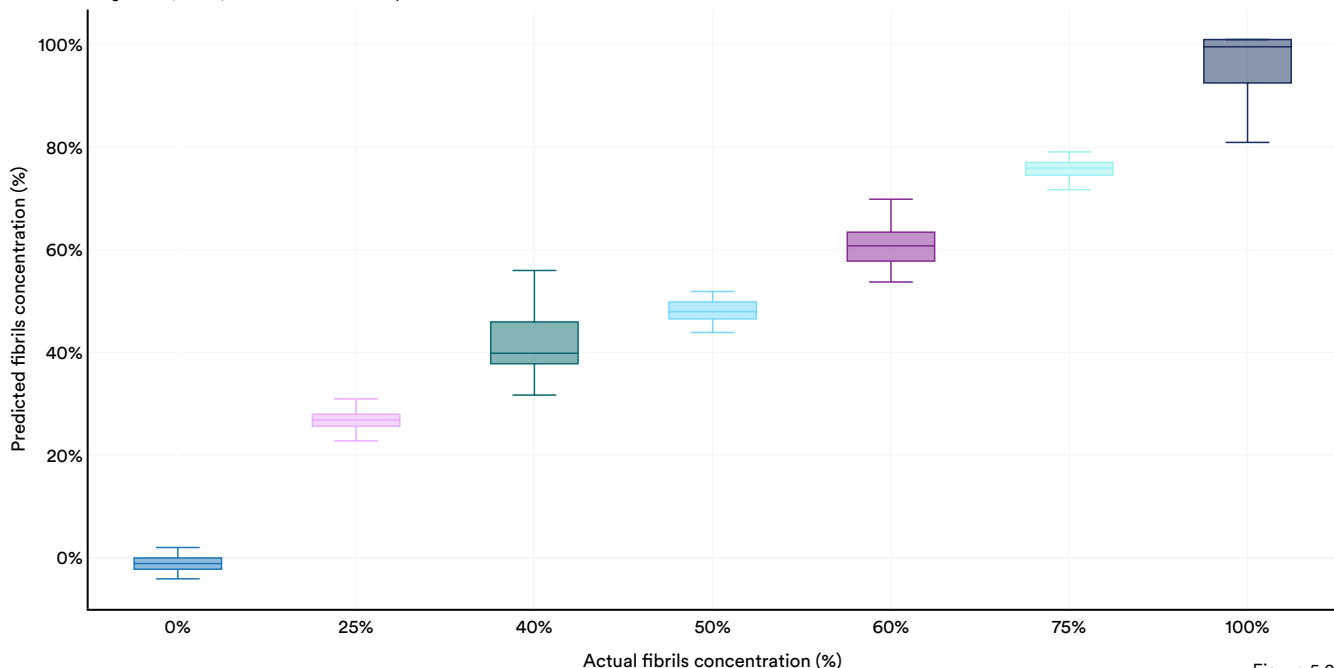


Figure 5.2.4

## EVEscape

### Forecasting viral evolution for pandemic preparedness

Predicting viral mutations is vital for vaccine design and pandemic minimization. Traditional methods, which rely on real-time virus strain and antibody data, face challenges during early pandemic stages due to data scarcity. EVEscape is a new AI deep learning model trained on historical sequences and biophysical and structural information that predicts the evolution

of viruses (Figure 5.2.5). EVEscape evaluates viral escape independently of current strain data predicting 50.0% of observed SARS-CoV-2 mutations, outperforming traditional lab studies which predicted 46.2% and 32.3%, as well as a previous model, which predicted only 24% of mutations (Figure 5.2.6). This performance highlights EVEscape’s potential as a valuable asset for enhancing future pandemic preparedness and response efforts.

### EVEscape design

Source: Thadani et al., 2023

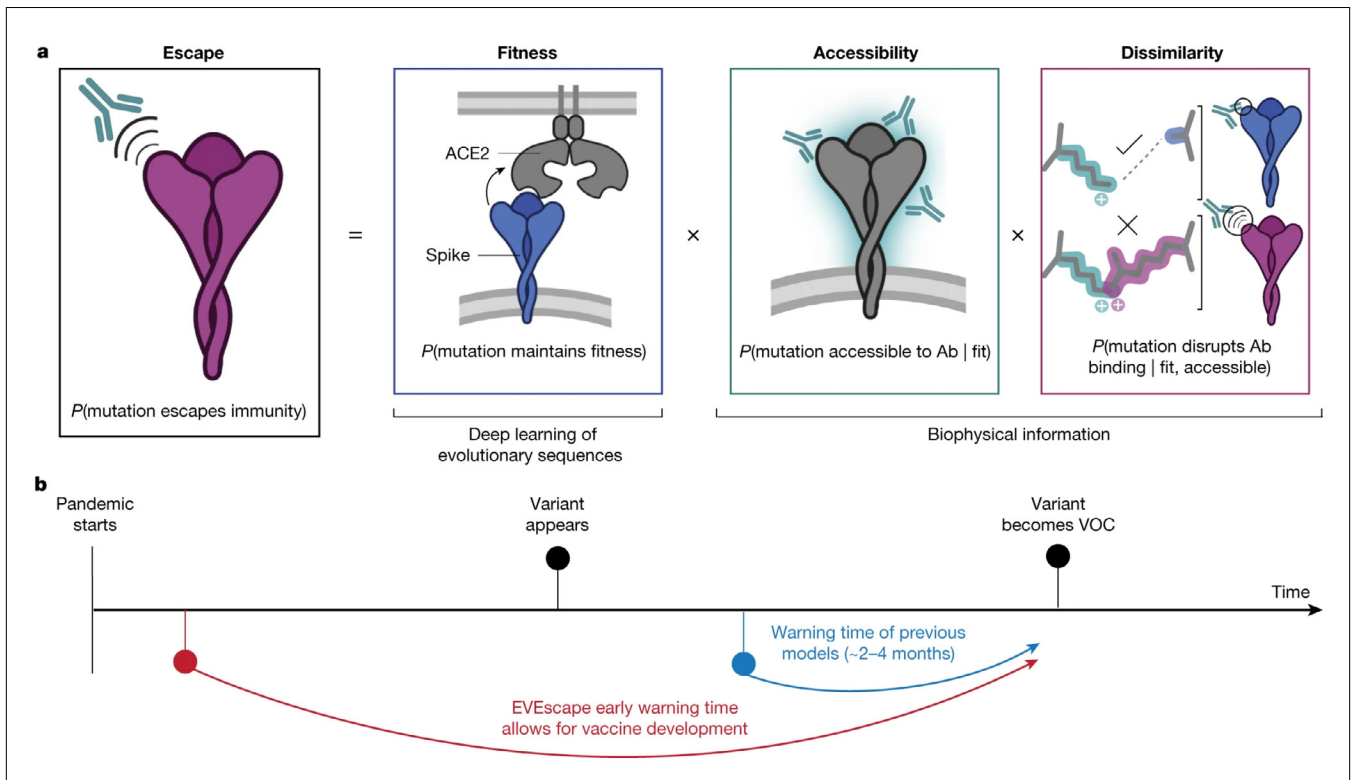


Figure 5.2.5

### EVEscape vs. other models on SARS-CoV-2 RBD mutation prediction

Source: Thadani et al., 2023 | Chart: 2024 AI Index report

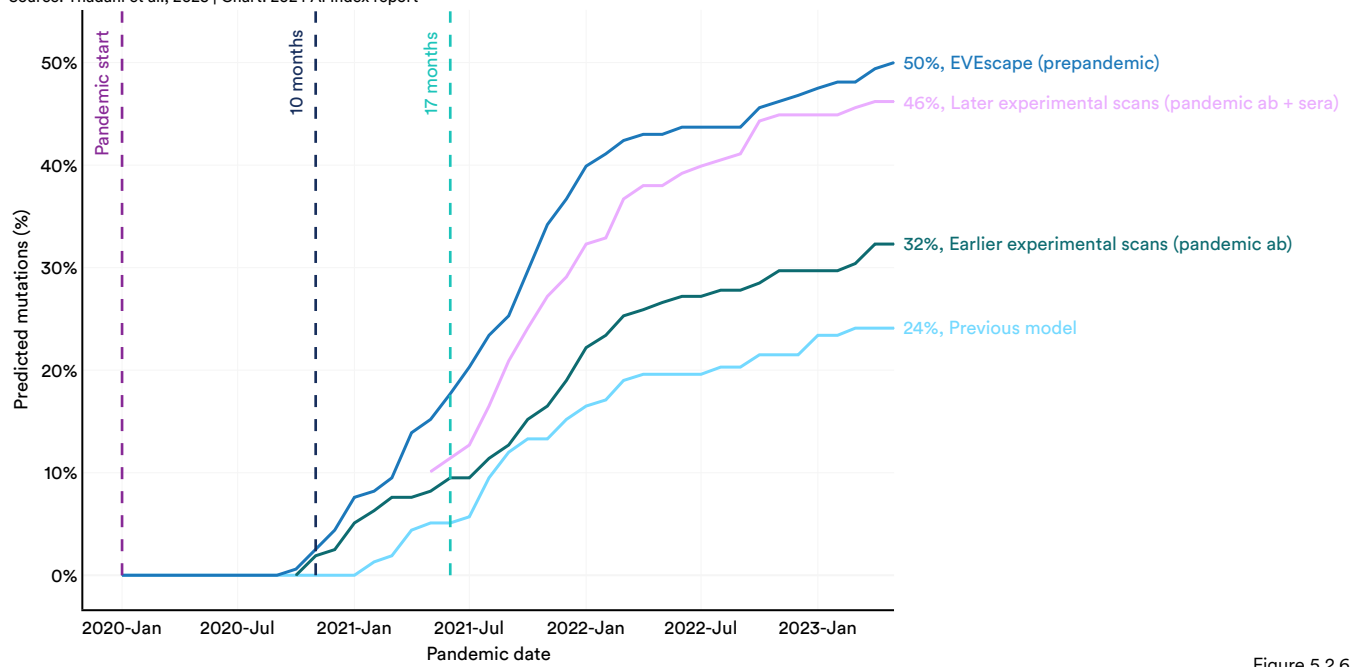


Figure 5.2.6



## AlphaMissense

### Better classification of AI mutations

Scientists still do not fully understand which genetic mutations lead to diseases. With millions of possible genetic mutations, determining whether a mutation is benign or pathogenic requires labor-intensive experiments.

In 2023, researchers from Google DeepMind unveiled AlphaMissense, a new AI model that predicted the pathogenicity of 71 million missense variants. Missense mutations are genetic alterations that impact the functionality of human proteins (Figure 5.2.7) and can lead to various diseases, including cancer. Of the 71 million possible missense variants, AlphaMissense classified 89%, identifying 57% as likely benign and 32% as likely pathogenic, while the remainder were categorized as uncertain (Figure 5.2.8). In contrast, human annotators have only been able to confirm the nature of 0.1% of all missense mutations.

### Hemaglobin subunit beta (HBB)

Source: Google DeepMind, 2023

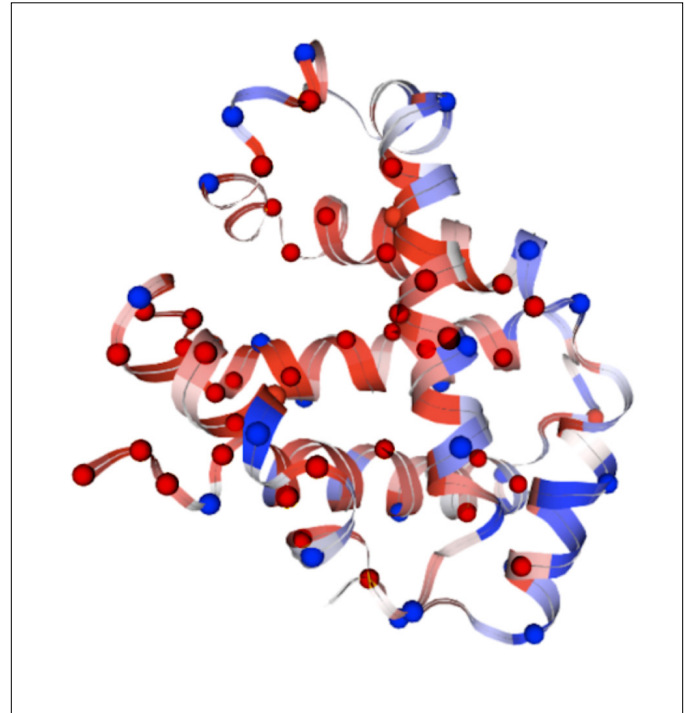


Figure 5.2.7

### AlphaMissense predictions

Source: Google DeepMind, 2023 | Chart: 2024 AI Index report

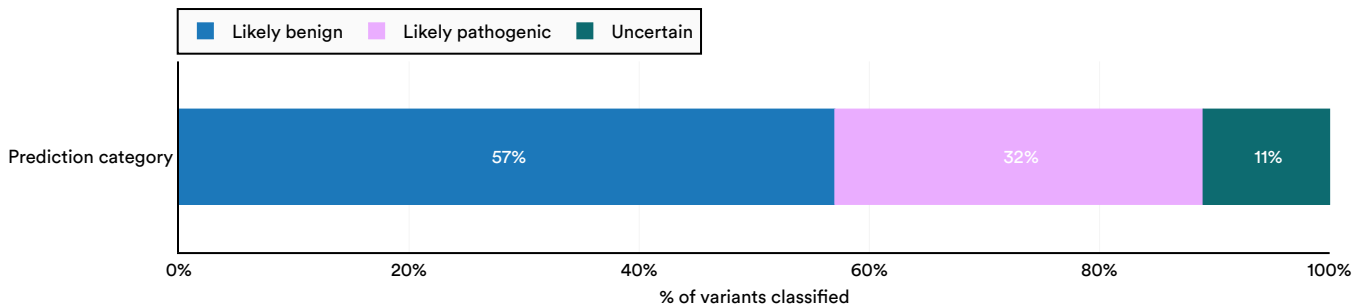


Figure 5.2.8

### Human Pangenome Reference

#### Using AI to map the human genome

The human genome is a set of molecular instructions for a human. The first human genome draft was released in 2000 and updated in 2022. However, the update was somewhat incomplete. It did not incorporate various genetic mutations, like blood type, and did not as completely map diverse ancestry groups. Therefore, under the existing genome reference, it would be difficult to detect diseases or find cures in certain groups of people.

In 2023, the Human Pangenome Research Consortium, comprising 119 scientists from 60 institutions, used AI to develop an updated and more representative human genome map (Figure 5.2.9). The researchers achieved remarkable accuracy, annotating a median of 99.07% of protein-coding genes, 99.42% of protein-coding transcripts, 98.16% of noncoding genes, and 98.96% of noncoding transcripts, as detailed in Figure 5.2.10.

#### Graph genome for the MHC region of the genome

Source: Google Research, 2023

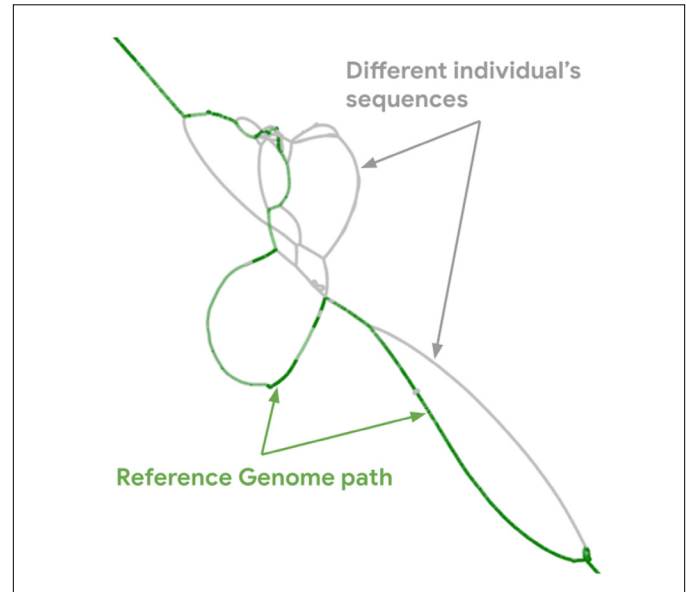


Figure 5.2.9

This latest version of the genome represents the most comprehensive and genetically diverse mapping of the human genome to date.

#### Ensembl mapping pipeline results

Source: Liao et al., 2023 | Chart: 2024 AI Index report

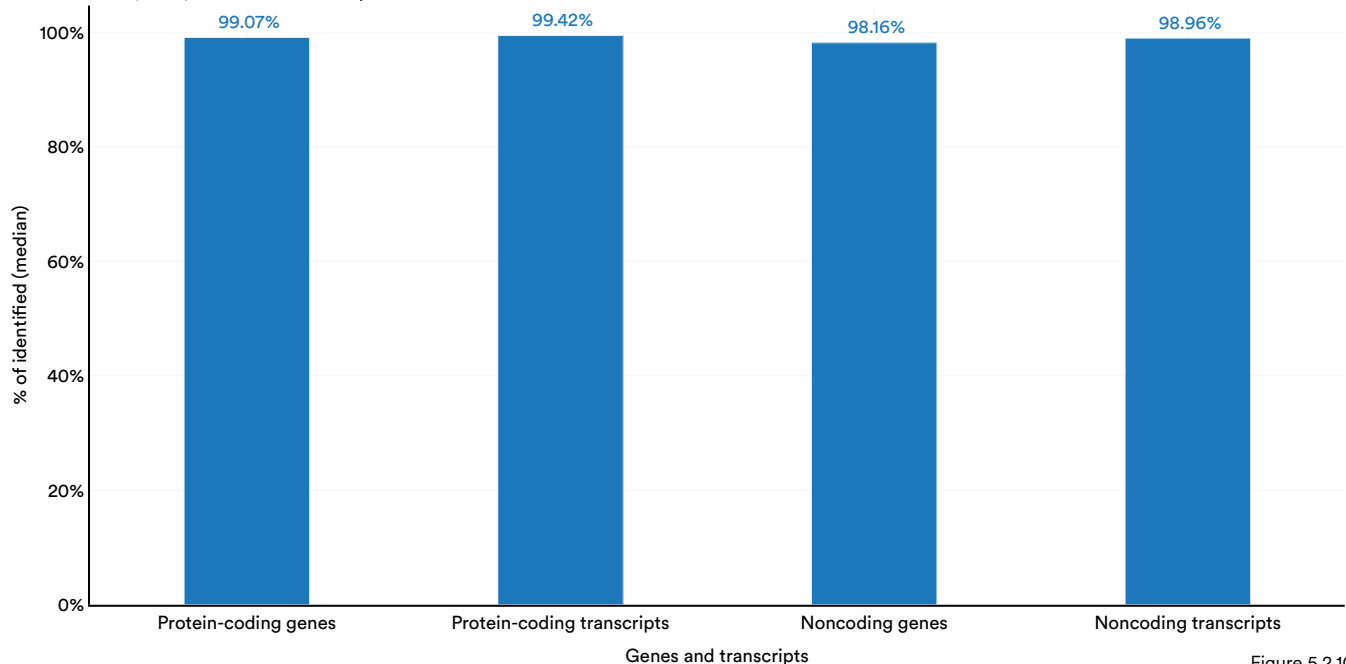


Figure 5.2.10

## Clinical Knowledge

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

### MedQA

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 seen remarkable improvement, with the leading system, GPT-4 Medprompt, achieving an accuracy rate of 90.2%—an increase of 22.6 percentage points from the top score in 2022 (Figure 5.2.11). Since MedQA's inception, AI capabilities on this benchmark have nearly tripled, showcasing the rapid improvements of clinically knowledgeable AI systems.

### MedQA: accuracy

Source: Papers With Code, 2023 | Chart: 2024 AI Index report

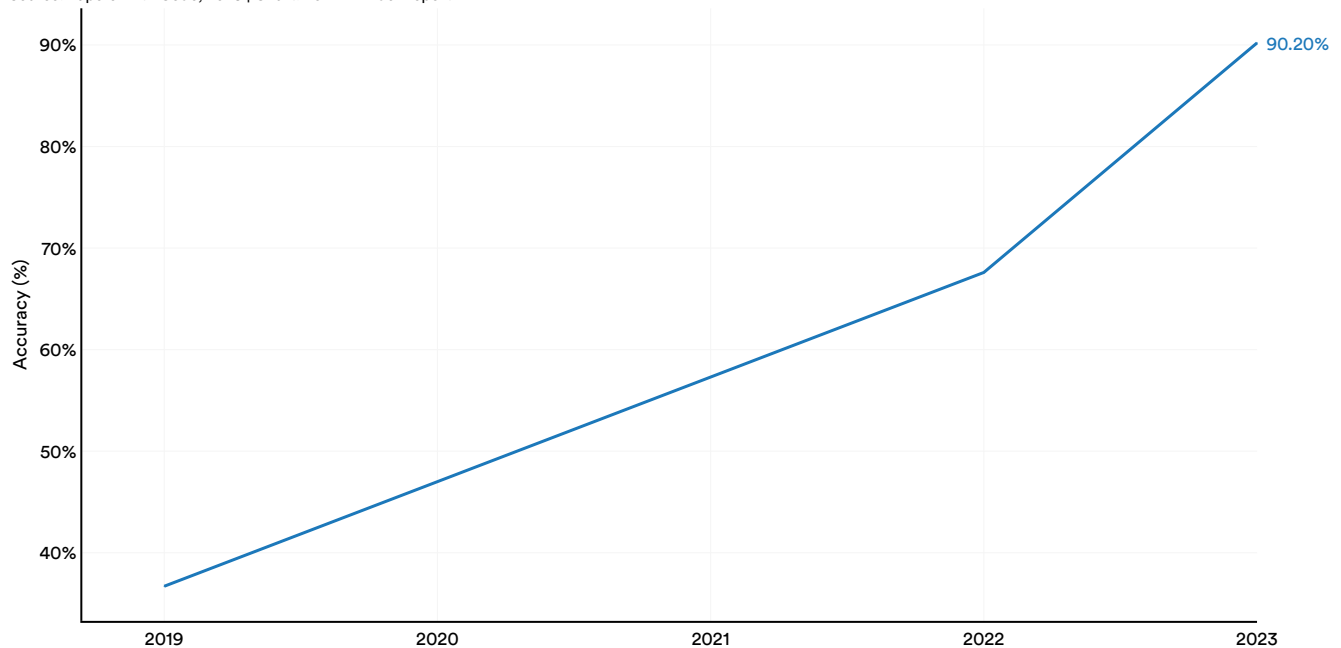


Figure 5.2.11

Highlighted Research:

## GPT-4 Medprompt

Although LLMs exhibit impressive general knowledge, it is commonly assumed that significant fine-tuning is required for them to excel at specialized knowledge, such as answering medical questions. Fine-tuning entails training an LLM on domain-specific data.

Research from Microsoft in late 2023 has overturned this assumption. This study employed prompt engineering to direct GPT-4 toward achieving remarkable performance on the MultiMedQA benchmark suite, a group of four challenging medical benchmarks (Figure 5.2.12). GPT-4 Medprompt exceeded the performance of the top 2022 model, Flan-PaLM 540B, in the multiple-choice sections of several renowned medical benchmarks, including PubMedQA, MedMCQA, and MMLU, by 3.0, 21.5, and 16.2 percentage points, respectively. It also exceeded the performance of the then state-of-the-art Med-PaLM 2 (Figure 5.2.13).


Moreover, as noted earlier, GPT-4 Medprompt was the first to surpass the 90% accuracy mark on the MedQA benchmark. This breakthrough not only underscores GPT-4 Medprompt’s exceptional and potentially clinically useful medical capabilities but also demonstrates that fine-tuning may not always be necessary for adapting models to specialized domains. Prompt engineering has shown to be a promising alternative strategy.

### GPT-4 vs. Med-PaLM 2 answering a medical question

Source: [Nori et al., 2023](#)


**Question:** A 22-year-old male marathon runner presents to the office with the complaint of right-sided rib pain when he runs long distances. Physical examination reveals normal heart and lung findings and an exhalation dysfunction at ribs 4-5 on the right. Which of the following muscles or muscle groups will be most useful in correcting this dysfunction utilizing a direct method?

(A) anterior scalene (B) latissimus dorsi (C) pectoralis minor (D) quadratus lumborum

  
Expert

Hand-crafted CoT from Med PaLM 2

Let's solve this step-bystep, referring to authoritative sources as needed. Among the options, only pectoralis minor muscle origins from the outer surfaces of the 3rd to 5th ribs.

  
GPT-4

GPT-4 generated CoT

The primary muscle involved in rib exhalation is the internal intercostal muscle. However, this option is not listed among the answer choices. Among the provided options, the pectoralis minor can contribute to rib movement, as it originates from the 3rd-5th ribs and can act to depress the ribs during exhalation when the scapula is fixed. Therefore, it could potentially be useful in correcting an exhalation dysfunction at ribs 4-5.

Figure 5.2.12

## Highlighted Research: GPT-4 Medprompt (cont'd)

### Model performance on MultiMedQA sub-benchmarks

Source: Nori et al., 2023 | Chart: 2024 AI Index report

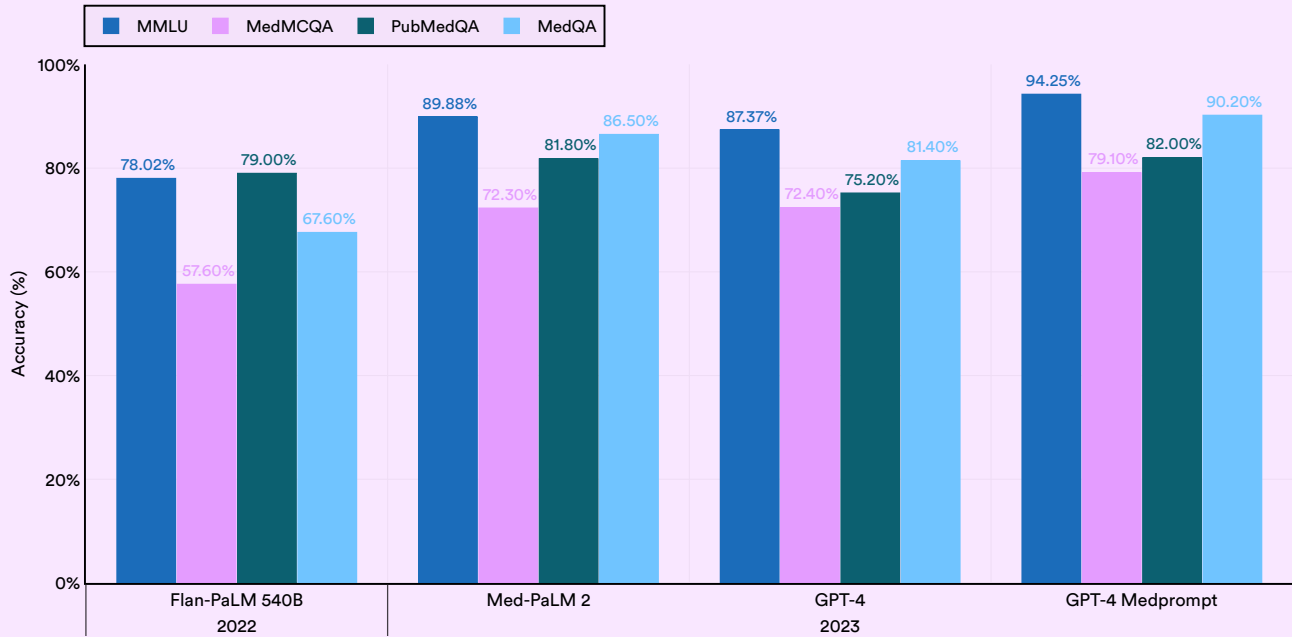


Figure 5.2.13

Highlighted Research:

## MediTron-70B

GPT-4 Medprompt is an impressive system; however, it is closed-source, meaning its weights are not freely available to the broader public for use. New research in 2023 has also sought to advance the capabilities of open-source medical LLMs. Among this new research, MediTron-70B stands out as particularly promising. This model achieves a respectable 70.2% accuracy on the MedQA benchmark. Although this is below the performance of GPT-4 Medprompt and Med-

PaLM 2 (both closed models), it represents a significant improvement over the state-of-the-art results from 2023 and surpasses other open-source models like Llama 2 (Figure 5.2.14). MediTron-70B’s score on MedQA is the highest yet achieved by an open-source model. If medical AI is to reach its fullest potential, it is important that its capabilities are widely accessible. In this context, MediTron represents an encouraging step forward.

### Performance of select models on MedQA

Source: Chen et al., 2023 | Table: 2024 AI Index report

Model	Release date	Access type	Score on MedQA
GPT-4 Medprompt	November 2023	Closed	90.20%
Med-PaLM 2	April 2023	Closed	86.20%
MediTron-70B	November 2023	Open	70.20%
Med-PaLM	December 2022	Closed	67.20%
Llama 2	July 2023	Open	63.80%

Figure 5.2.14

## Diagnosis

AI tools can also be used for diagnostic purposes including, for example, in radiology or cancer detection.

### Highlighted Research:

## CoDoC

AI medical imaging systems demonstrate robust diagnostic capabilities, yet there are instances where they overlook diagnoses that clinicians catch, and vice versa. This observation suggests a logical integration of AI systems and clinicians' diagnostic abilities. In 2023, researchers unveiled CoDoC (Complementarity-Driven Deferral to Clinical Workflow), a system designed to discern when to rely on AI for diagnosis and when to defer to traditional clinical methods. CoDoC notably enhances both sensitivity (the ability to correctly identify individuals with a disease) and specificity

(the ability to accurately identify those without it). Specifically, across four medical datasets, CoDoC's sensitivity surpasses clinicians' by an average of 4.5 percentage points and a standalone AI model's by 6.5 percentage points (Figure 5.2.15). In terms of specificity, CoDoC outperforms clinicians by an average of 2.7 percentage points across tested datasets and a standalone predictive model by 5.7 percentage points. Moreover, CoDoC has been shown to reduce clinical workflow by 66%. These findings suggest that AI medical systems can be integrated into clinical workflows, thereby enhancing diagnostic accuracy and efficiency.

### CoDoC vs. standalone predictive AI system and clinical readers: sensitivity

Source: Dvijotham et al., 2023 | Chart: 2024 AI Index report

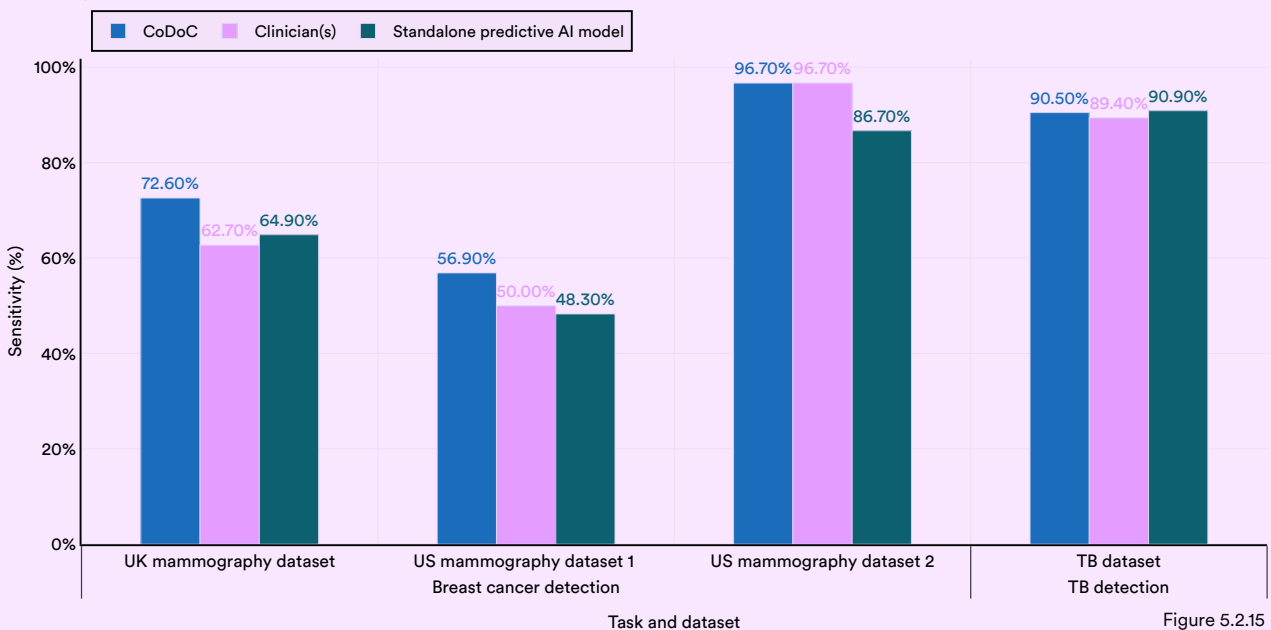


Figure 5.2.15

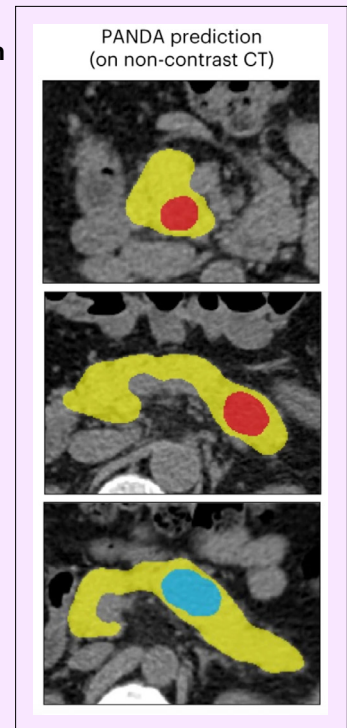
**Highlighted Research:**

## CT Panda

Pancreatic ductal adenocarcinoma (PDAC) is a particularly lethal cancer, often detected too late for surgical intervention. Screening for PDAC in asymptomatic individuals is challenging due to its low prevalence and the risk of false positives. This year, a Chinese research team developed PANDA (pancreatic cancer detection with artificial intelligence), an AI model capable of efficiently detecting and classifying pancreatic lesions in X-rays (Figure 5.2.16). In validation tests, PANDA surpassed the average radiologist in sensitivity by 34.1% and in specificity by 6.3% (Figure 5.2.17). In a large-scale, real-world test involving approximately 20,000 patients, PANDA achieved a sensitivity of 92.9% and a specificity of 99.9% (Figure 5.2.18). AI medical tools like PANDA represent significant advancements in diagnosing challenging conditions, offering cost-effective and accurate detection previously considered difficult or prohibitive.

**PANDA detection**

Source: Cao et al., 2023  
Figure 5.2.16



**PANDA vs. mean radiologist on multicenter validation (6,239 patients)**

Source: Cao et al., 2023 | Chart: 2024 AI Index report

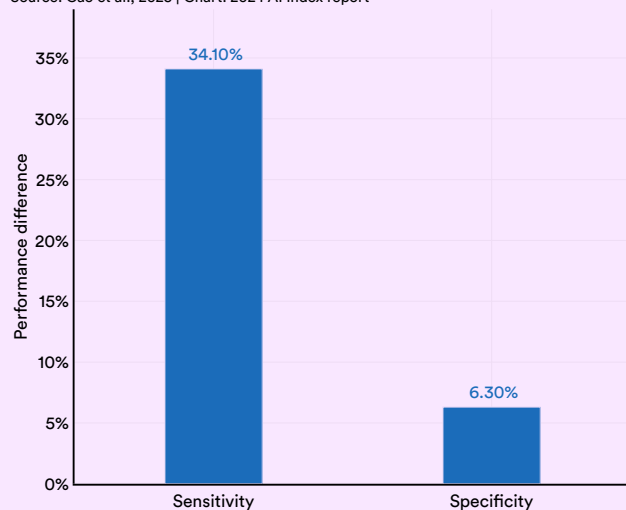


Figure 5.2.17

**PANDA performance on real-world multi-scenario validation (20,530 patients)**

Source: Cao et al., 2023 | Chart: 2024 AI Index report

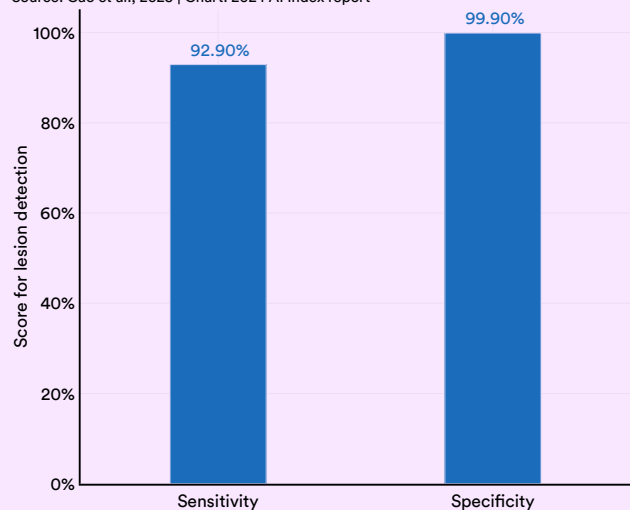


Figure 5.2.18



### Other Diagnostic Uses

New research published in 2023 highlights how AI can be used in other diagnostic contexts. Figure 5.2.19 summarizes some of the findings.

#### Additional research on diagnostic AI use cases

Source: AI Index, 2024

Research	Use case	Findings
Schopf et al., 2023	Breast cancer	The authors conducted a meta-review of the literature exploring mammography-image-based AI algorithms. They discovered that predicting future breast cancer risk using only mammography images achieves accuracy that is comparable to or better than traditional risk assessment tools.
Dicente Cid et al., 2023	X-ray interpretation	The researchers developed two open-source neural networks, X-Raydar and X-Raydar-NLP, for classifying chest X-rays using images and free-text reports. They found that these automated classification methods perform at levels comparable to human experts and demonstrate robustness when applied to external data sets.

Figure 5.2.19

### FDA-Approved AI-Related Medical Devices

The U.S. Food and Drug Administration (FDA) maintains a [list](#) of AI/ML-enabled medical devices that have received approval. The devices featured on this list meet the FDA’s premarket standards, which include a detailed review of their effectiveness and safety. As of October 2023, the FDA has not approved any devices that utilize generative AI or are powered by LLMs.

Figure 5.2.20 illustrates the number of AI medical devices approved by the FDA over the past decade. In 2022, a total of 139 AI-related medical devices received FDA approval, marking a 12.1% increase from the total approved in 2021. Since 2012, the number of these devices has increased by more than 45-fold.

#### Number of AI medical devices approved by the FDA, 2012–22

Source: FDA, 2023 | Chart: 2024 AI Index report

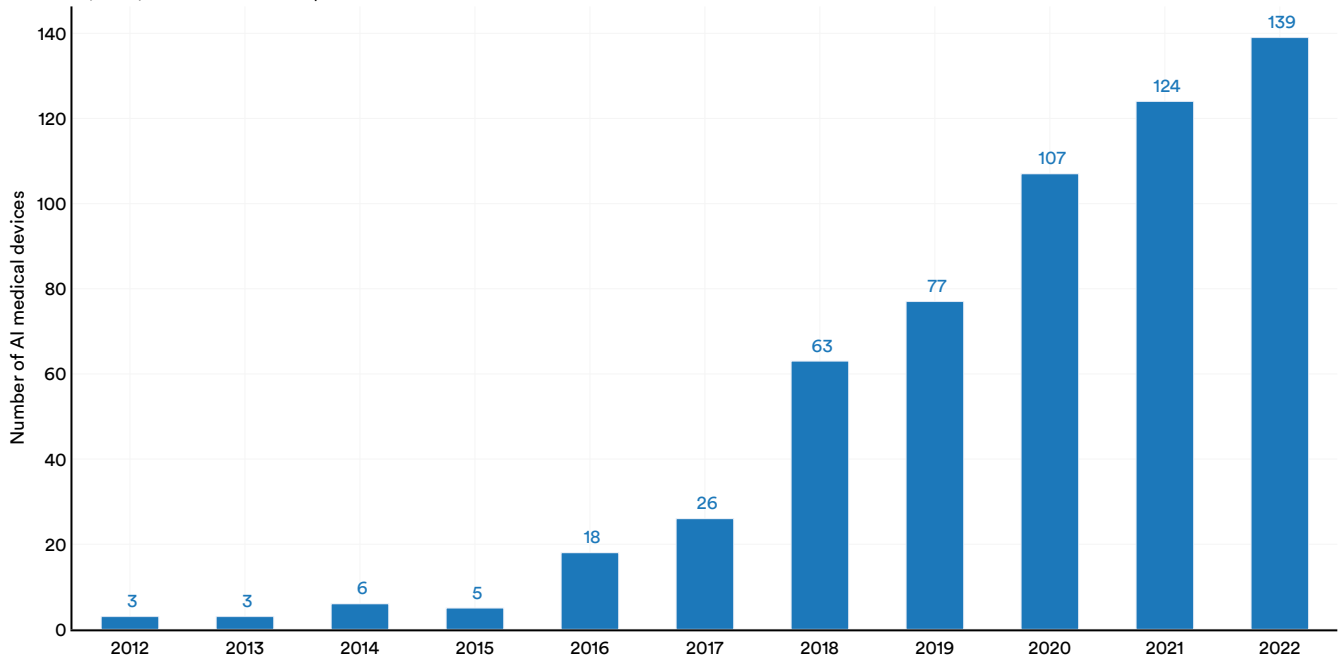


Figure 5.2.20

<sup>3</sup> The FDA last updated the list in October 2023, meaning that the totals for 2023 were incomplete. Consequently, the AI Index limited its data presentation to include only information up to 2022.

Figure 5.2.21 illustrates the specialties associated with FDA-approved medical devices. Of the 139 devices approved in 2022, a significant majority, 87.1%, were related to radiology. The next most common specialty was cardiovascular, accounting for 7.2% of the approvals.

### Number of AI medical devices approved by the FDA by specialty, 2012–22

Source: FDA, 2023 | Chart: 2024 AI Index report

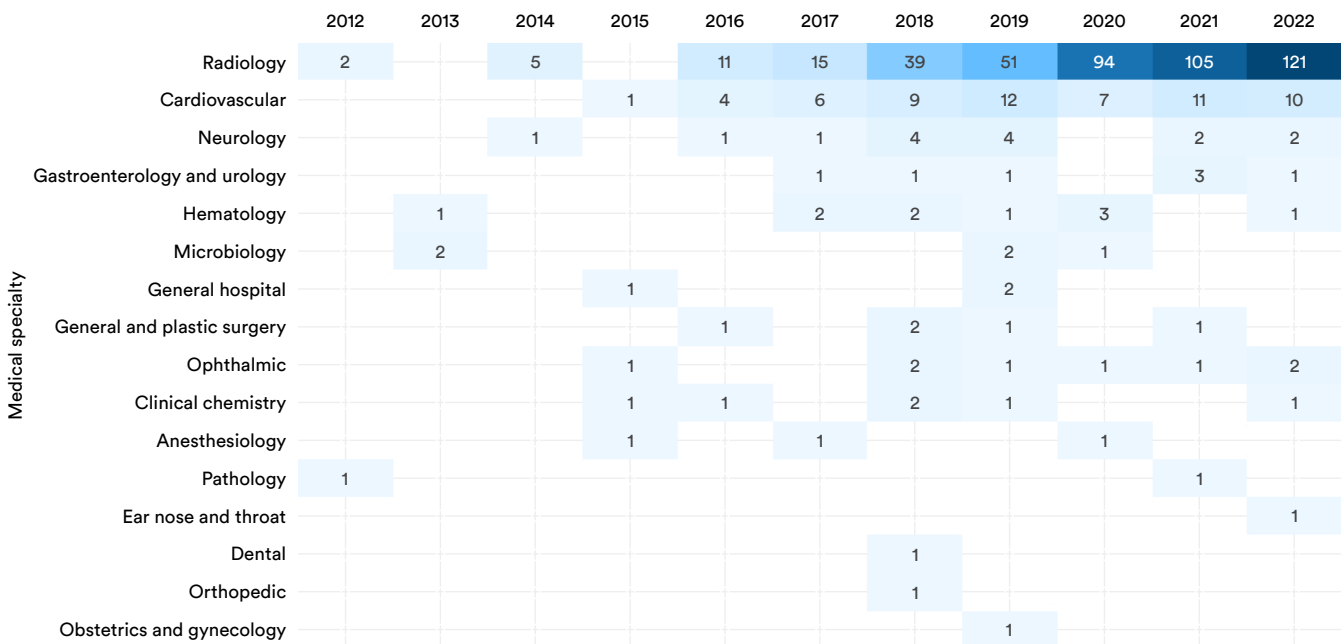


Figure 5.2.21

## Administration and Care

AI tools also hold the potential to enhance medical administration efficiency and elevate the standard of patient care.

### Highlighted Research:

## MedAlign

Despite significant advances in AI for healthcare, existing benchmarks like MedQA and USMLE, focused on knowledge-based questions, do not fully capture the diverse tasks clinicians perform in patient care. Clinicians often engage in information-intensive tasks, such as creating tailored diagnostic plans, and spend a significant proportion of their working hours on administrative tasks. Although AI has the potential to streamline these processes, there is a lack of suitable electronic health records (EHR) datasets for benchmarking and fine-tuning medically administrative LLMs. This year researchers have made strides to address this gap by introducing MedAlign: a comprehensive EHR-based

benchmark with 983 questions and instructions and 303 clinician responses, drawn from seven different medical specialties (Figure 5.2.22). MedAlign is the first extensive EHR-focused benchmark.

The researchers then tested various existing LLMs on MedAlign. Of all LLMs, a GPT-4 variant using multistep refinement achieved the highest correctness rate (65.0%) and was routinely preferred over other LLMs (Figure 5.2.23). MedAlign is a valuable milestone toward using AI to alleviate administrative burdens in healthcare.

### MedAlign workflow

Source: Fleming et al., 2023

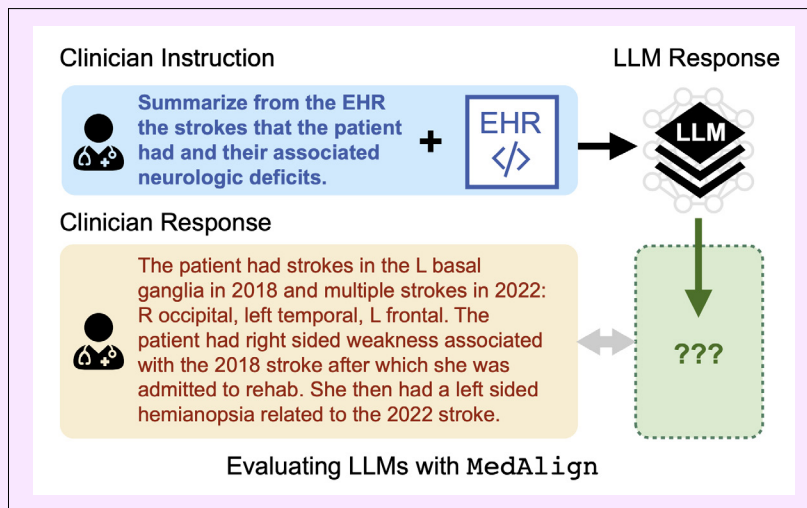


Figure 5.2.22

## Highlighted Research: MedAlign (cont'd)

### Evaluation of model performance: human vs. COMET ranks

Source: Fleming et al., 2023 | Chart: 2024 AI Index report

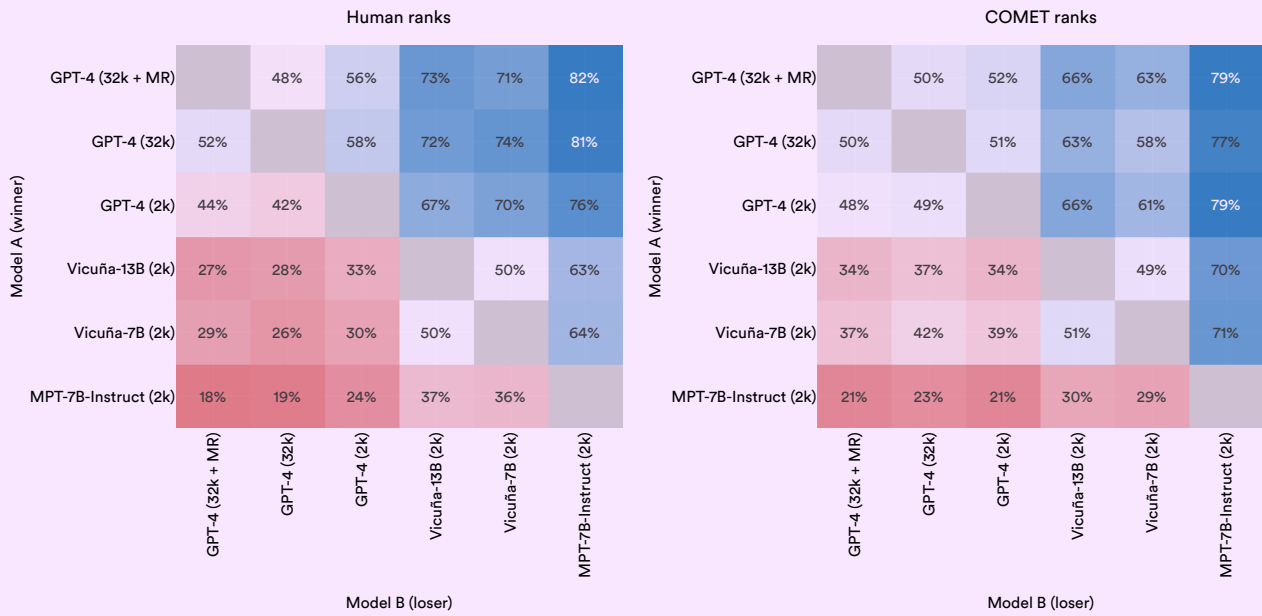


Figure 5.2.23

# Appendix

## Acknowledgments

The AI Index would like to acknowledge Emma Williamson for her work surveying the literature on significant AI-related science and medicine trends.

## Benchmarks

1. **MedQA:** Data on MedQA was taken from the [MedQA Papers With Code leaderboard](#) in January 2024. To learn more about MedQA, please read the [original paper](#).

## FDA-Approved AI-Medical Devices

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

## Works Cited

Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., Nogues, I., Li, X., Guo, W., Wang, Y., Fang, W., Qiu, M., Hou, Y., Kovarnik, T., Vocka, M., Lu, J. (2023). “Large-Scale Pancreatic Cancer Detection via Non-contrast CT and Deep Learning.” *Nature Medicine* 29, no. 12: 3033–3043. <https://doi.org/10.1038/s41591-023-02640-w>.

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M.-A., Jaggi, M. & Bosselut, A. (2023). *MEDITRON-70B: Scaling Medical Pretraining for Large Language Models* (arXiv:2311.16079). arXiv. <http://arxiv.org/abs/2311.16079>.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P. & Avsec, Ž. (2023). “Accurate Proteome-Wide Missense Variant Effect Prediction With AlphaMissense.” *Science* 381. <https://doi.org/10.1126/science.adg7492>.

Cid, Y. D., Macpherson, M., Gervais-Andre, L., Zhu, Y., Franco, G., Santeramo, R., Lim, C., Selby, I., Muthuswamy, K., Amlani, A., Hopewell, H., Indrajeet, D., Liakata, M., Hutchinson, C. E., Goh, V. & Montana, G. (2024). “Development and Validation of Open-Source Deep Neural Networks for Comprehensive Chest X-Ray Reading: A Retrospective, Multicentre Study.” *The Lancet Digital Health* 6, no. 1: e44–e57. [https://doi.org/10.1016/S2589-7500\(23\)00218-2](https://doi.org/10.1016/S2589-7500(23)00218-2).

Fleming, S. L., Lozano, A., Haberkorn, W. J., Jindal, J. A., Reis, E. P., Thapa, R., Blankemeier, L., Genkins, J. Z., Steinberg, E., Nayak, A., Patel, B. S., Chiang, C.-C., Callahan, A., Huo, Z., Gatidis, S., Adams, S. J., Fayanju, O., Shah, S. J., Savage, T., ... Shah, N. H. (2023). *MedAlign: A Clinician-Generated Dataset for Instruction Following With Electronic Medical Records* (arXiv:2308.14089). arXiv. <http://arxiv.org/abs/2308.14089>.

Ha, T., Lee, D., Kwon, Y., Park, M. S., Lee, S., Jang, J., Choi, B., Jeon, H., Kim, J., Choi, H., Seo, H.-T., Choi, W., Hong, W., Park, Y. J., Jang, J., Cho, J., Kim, B., Kwon, H., Kim, G., ... Choi, Y.-S. (2023). “AI-Driven Robotic Chemist for Autonomous Synthesis of Organic Molecules.” *Science Advances* 9, no. 44. <https://doi.org/10.1126/sciadv.adj0461>.

Iglesias, J. E., Billot, B., Balbastre, Y., Magdamo, C., Arnold, S. E., Das, S., Edlow, B. L., Alexander, D. C., Golland, P. & Fischl, B. (2023). “SynthSR: A Public AI Tool to Turn Heterogeneous Clinical Brain Scans into High-Resolution T1-Weighted Images for 3D Morphometry.” *Science Advances* 9, no. 5. <https://doi.org/10.1126/sciadv.add3607>.

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; Version 1). arXiv. <http://arxiv.org/abs/2009.13081>.

Kavungal, D., Magalhães, P., Kumar, S. T., Kolla, R., Lashuel, H. A. & Altug, H. (2023). “Artificial Intelligence–Coupled Plasmonic Infrared Sensor for Detection of Structural Protein Biomarkers in Neurodegenerative Diseases.” *Science Advances* 9, no. 28. <https://doi.org/10.1126/sciadv.adg9644>.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S. & Battaglia, P. (2023). “Learning Skillful Medium-Range Global Weather Forecasting.” *Science* 382. <https://doi.org/10.1126/science.adi2336>.

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., ... Paten, B. (2023). “A Draft Human Pangenome Reference.” *Nature* 617: 312–24. <https://doi.org/10.1038/s41586-023-05896-x>.

Mankowitz, D. J., Michi, A., Zhernov, A., Gelmi, M., Selvi, M., Paduraru, C., Leurent, E., Iqbal, S., Lespiau, J.-B., Ahern, A., Köppe, T., Millikin, K., Gaffney, S., Elster, S., Broshear, J., Gamble, C., Milan, K., Tung, R., Hwang, M., ... Silver, D. (2023). “Faster Sorting Algorithms Discovered Using Deep Reinforcement Learning.” *Nature* 618: 257–63. <https://doi.org/10.1038/s41586-023-06004-9>.

- Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G. & Cubuk, E. D. (2023). “Scaling Deep Learning for Materials Discovery.” *Nature* 624: 80–85. <https://doi.org/10.1038/s41586-023-06735-9>.
- Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T., Weitzner, D. & Matias, Y. (2023). *AI Increases Global Access to Reliable Flood Forecasts* (arXiv:2307.16104). arXiv. <http://arxiv.org/abs/2307.16104>.
- 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. (2023a). *Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine* (arXiv:2311.16452; Version 1). arXiv. <http://arxiv.org/abs/2311.16452>.
- Schopf, C. M., Ramwala, O. A., Lowry, K. P., Hofvind, S., Marinovich, M. L., Houssami, N., Elmore, J. G., Dontchos, B. N., Lee, J. M. & Lee, C. I. (2024). “Artificial Intelligence-Driven Mammography-Based Future Breast Cancer Risk Prediction: A Systematic Review.” *Journal of the American College of Radiology* 21, no. 2: 319–28. <https://doi.org/10.1016/j.jacr.2023.10.018>.
- Shen, T., Munkberg, J., Hasselgren, J., Yin, K., Wang, Z., Chen, W., Gojcic, Z., Fidler, S., Sharp, N. & Gao, J. (2023). “Flexible Isosurface Extraction for Gradient-Based Mesh Optimization.” *ACM Transactions on Graphics* 42, no. 4: 1–16. <https://doi.org/10.1145/3592430>.
- Thadani, N. N., Gurev, S., Notin, P., Youssef, N., Rollins, N. J., Ritter, D., Sander, C., Gal, Y. & Marks, D. S. (2023). “Learning From Prepandemic Data to Forecast Viral Escape.” *Nature* 622: 818–25. <https://doi.org/10.1038/s41586-023-06617-0>.