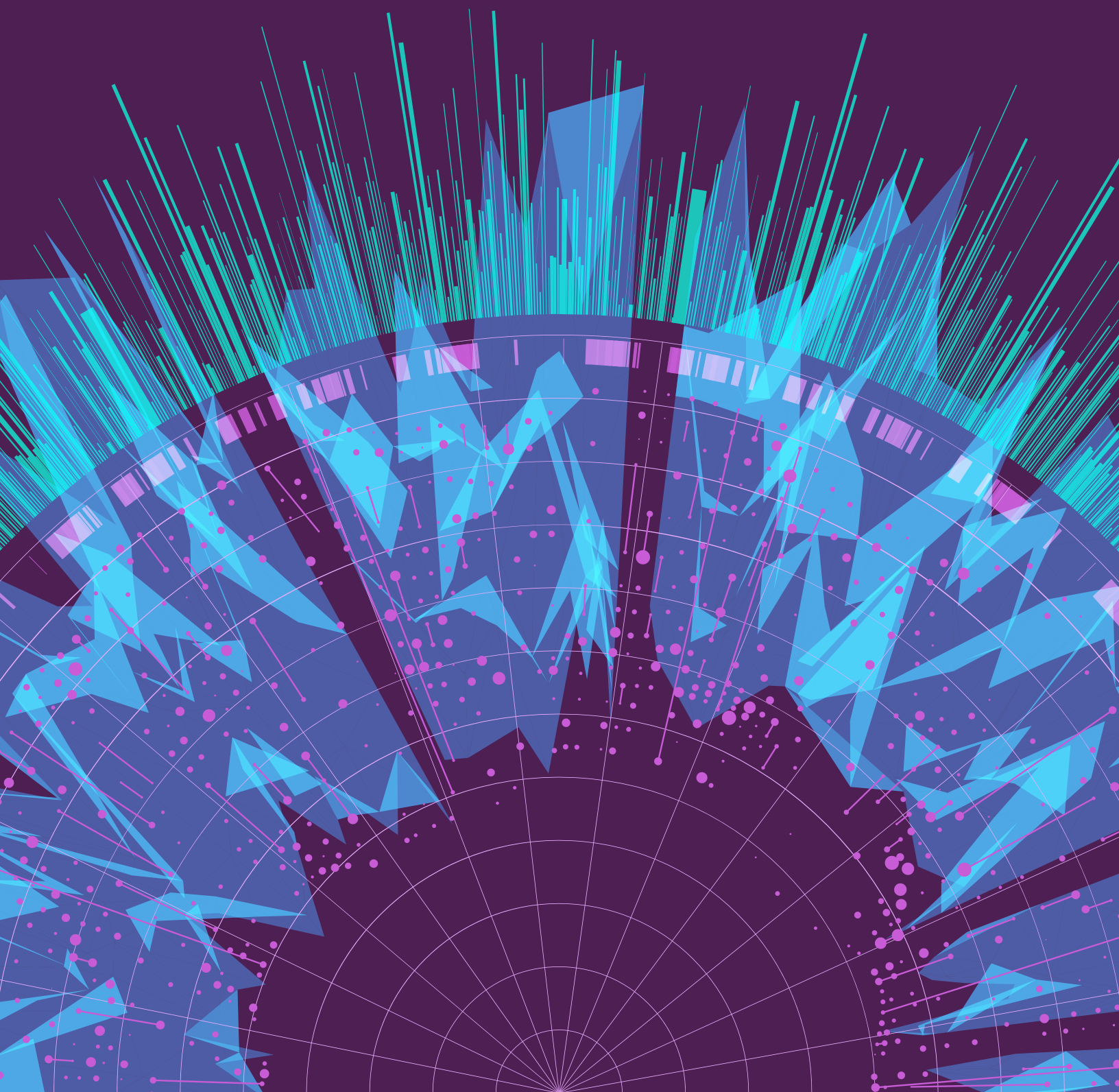




Artificial Intelligence
Index Report 2024

CHAPTER 5: Science and Medicine



Preview

Overview	298
Chapter Highlights	299

5.1 Notable Scientific Milestones	300
AlphaDev	300
FlexiCubes	301
Synbot	303
GraphCast	304
GNoME	305
Flood Forecasting	306

5.2 AI in Medicine	307
Notable Medical Systems	307
SynthSR	307
Coupled Plasmonic Infrared Sensors	309
EVEscape	310
AlphaMissence	312
Human Pangenome Reference	313
Clinical Knowledge	314
MedQA	314
Highlighted Research: GPT-4 Medprompt	315
Highlighted Research: MediTron-70B	317
Diagnosis	318
Highlighted Research: CoDoC	318
Highlighted Research: CT Panda	319
Other Diagnostic Uses	320
FDA-Approved AI-Related Medical Devices	321
Administration and Care	323
Highlighted Research: MedAlign	323

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

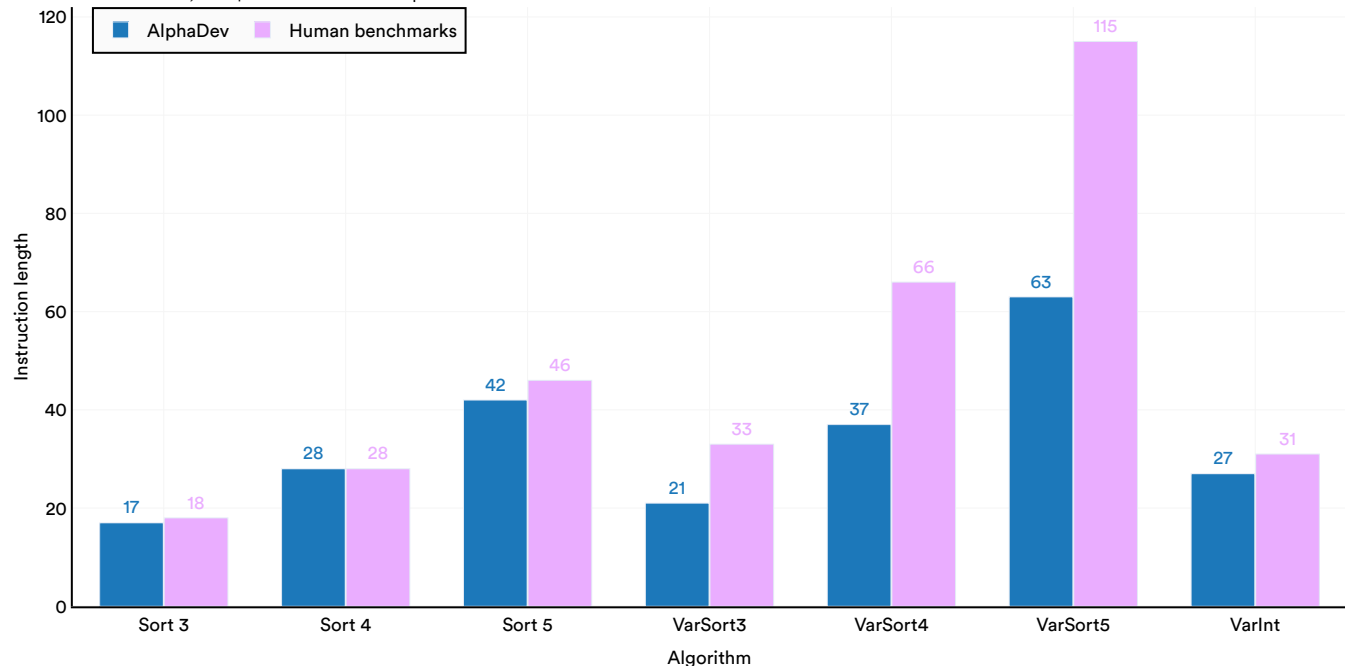


Figure 5.1.1

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](#)

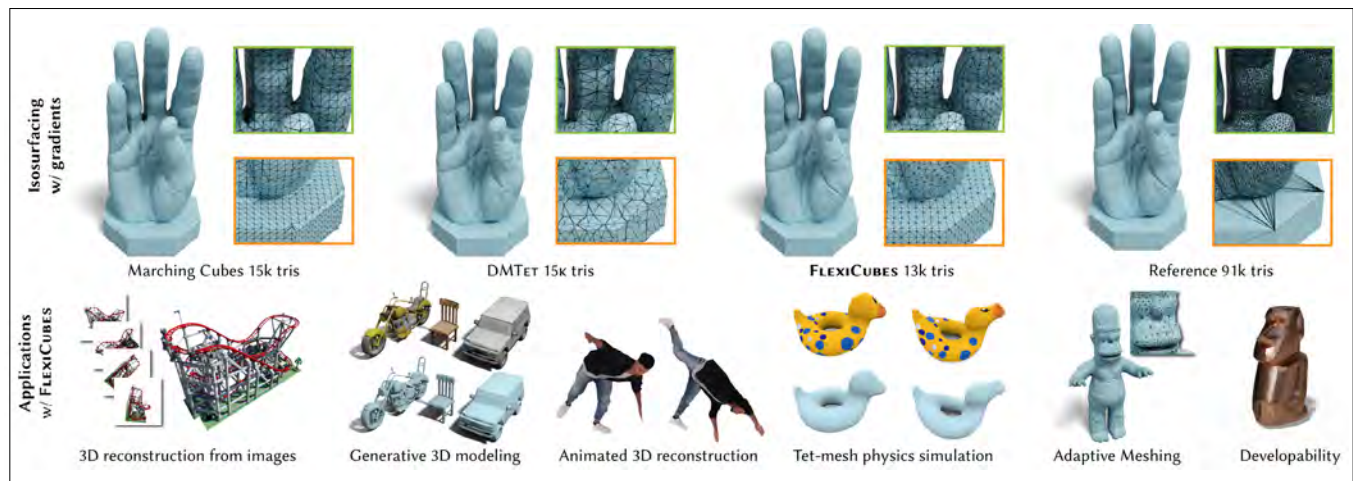


Figure 5.1.2

Select quantitative results on 3D mesh reconstruction

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

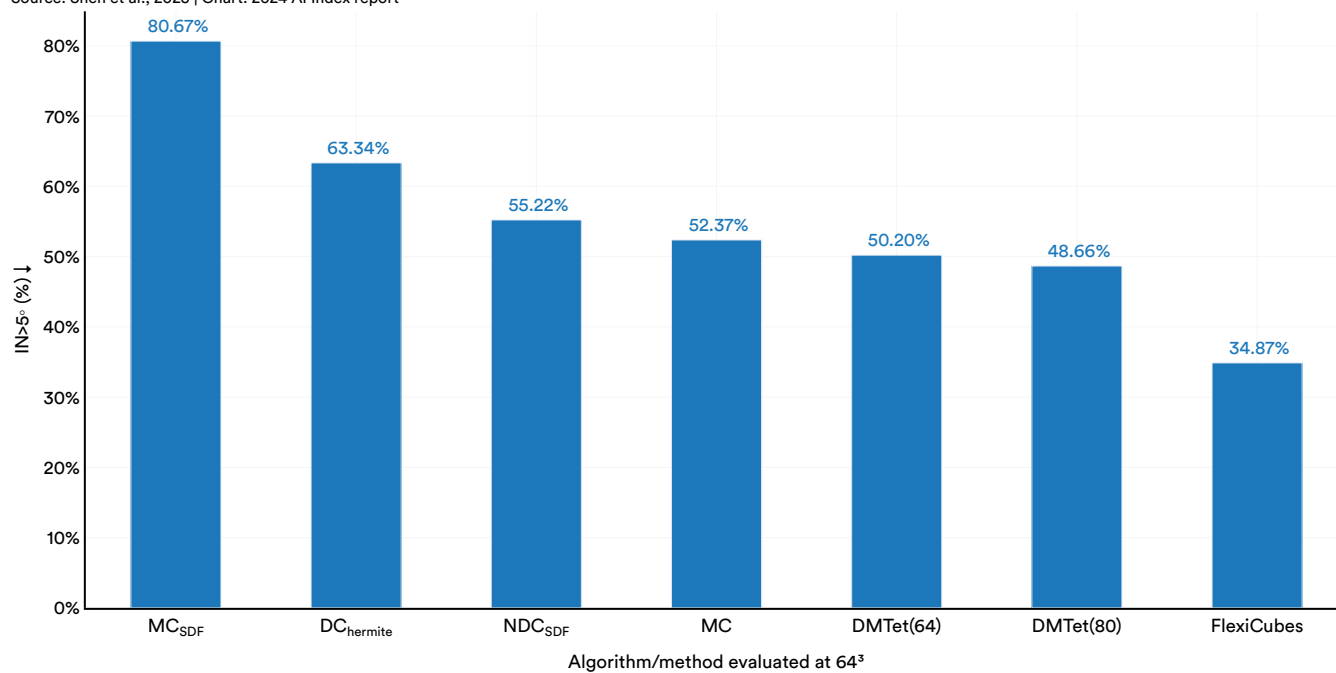


Figure 5.1.3

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](#)

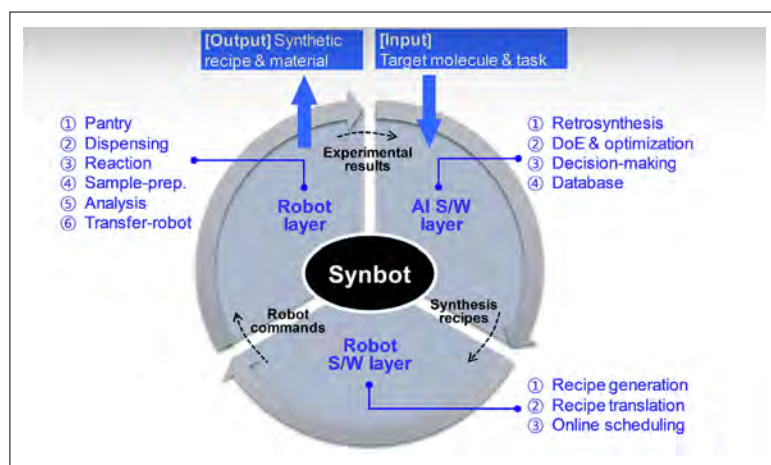


Figure 5.1.4

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

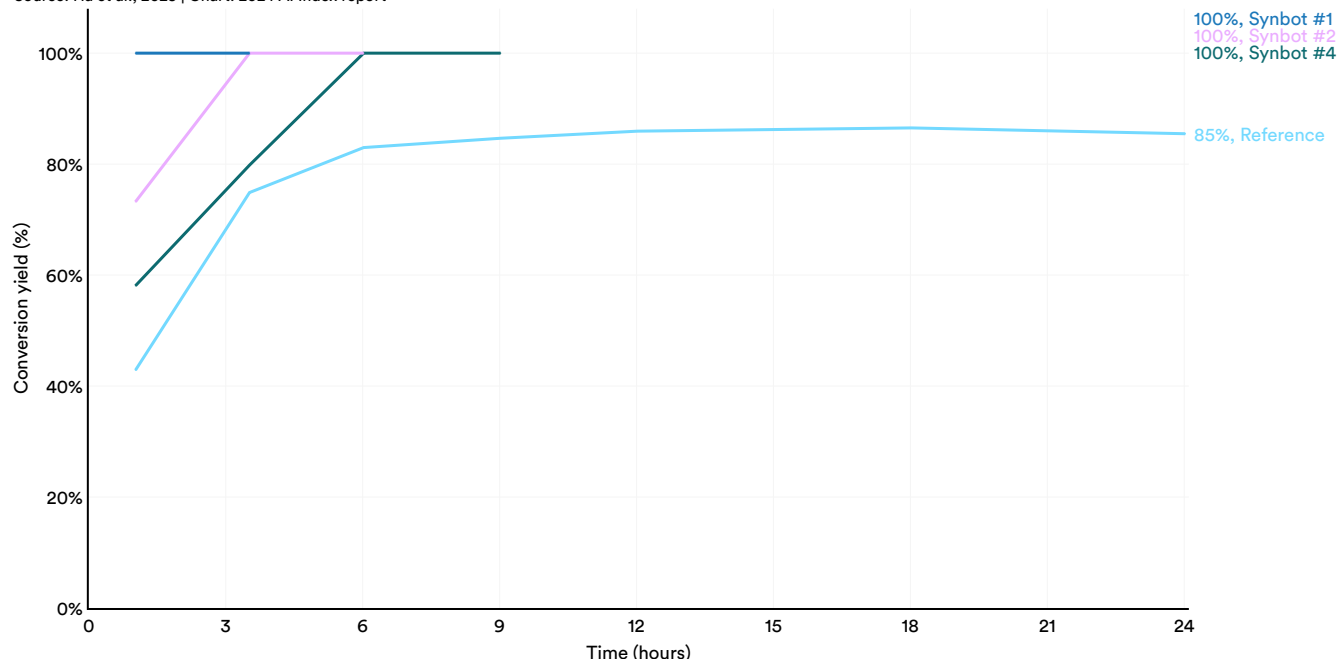


Figure 5.1.5

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](#)

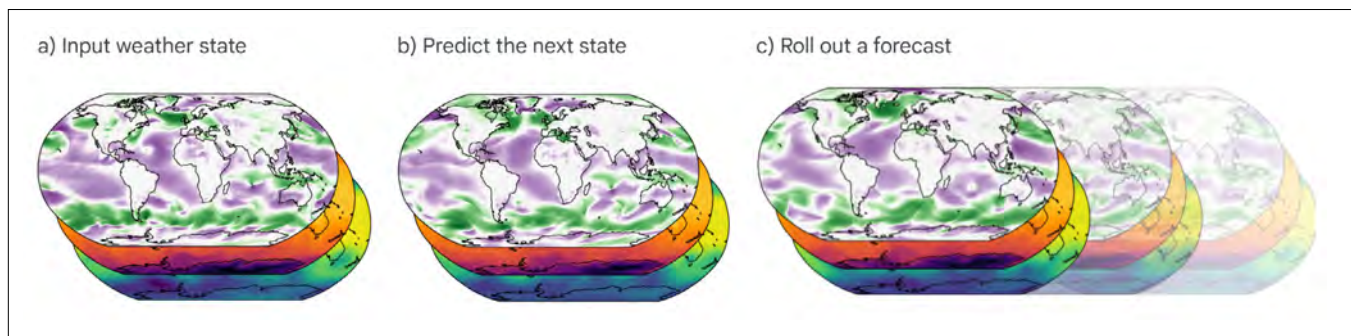


Figure 5.1.6

Ten-day z500 forecast skill: GraphCast vs. HRES

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

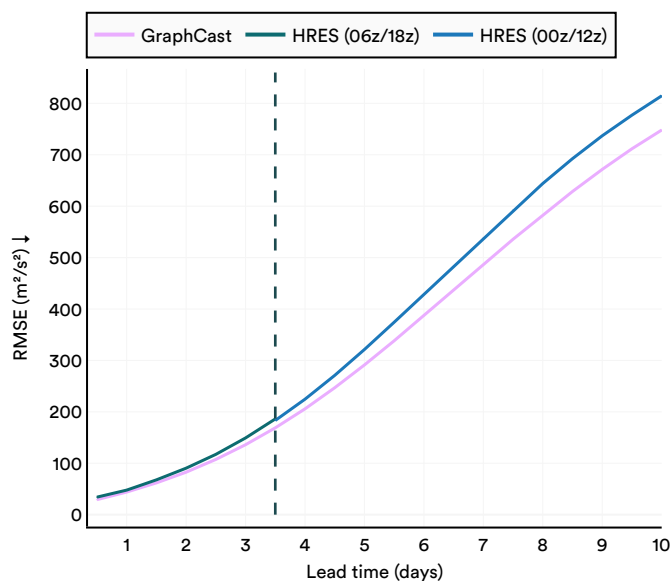


Figure 5.1.7

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

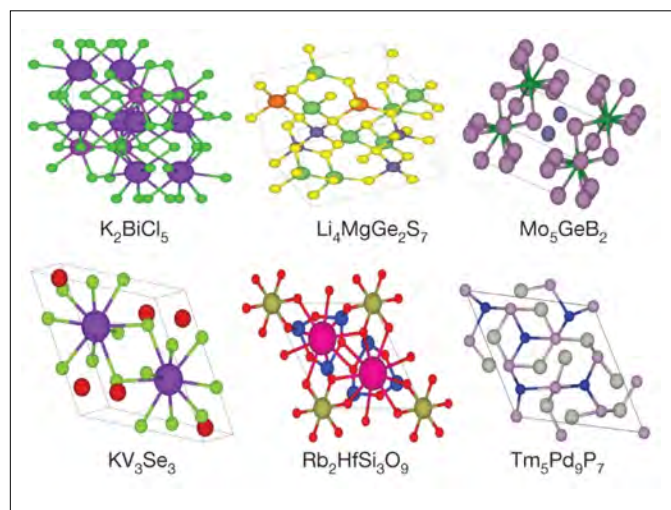


Figure 5.1.8

GNoME vs. Materials Project: stable crystal count

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

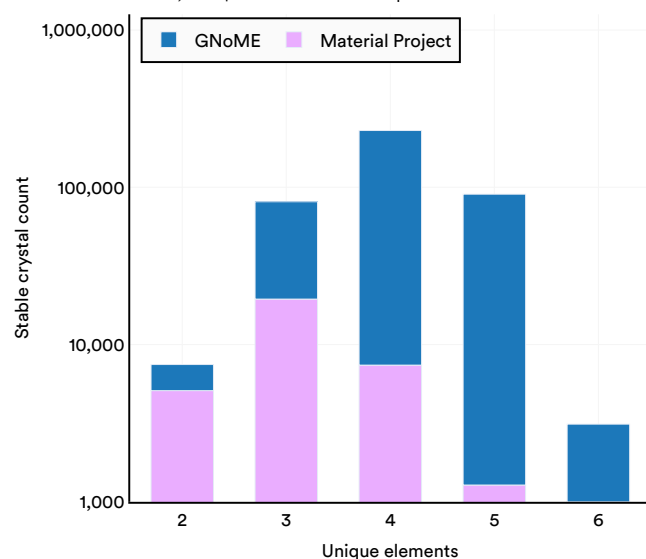


Figure 5.1.9

GNoME vs. Materials Project: distinct prototypes

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

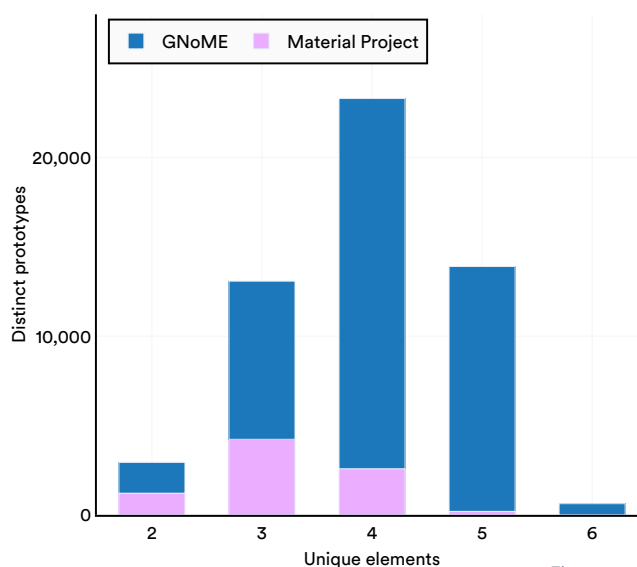


Figure 5.1.10

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.¹ 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).² 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

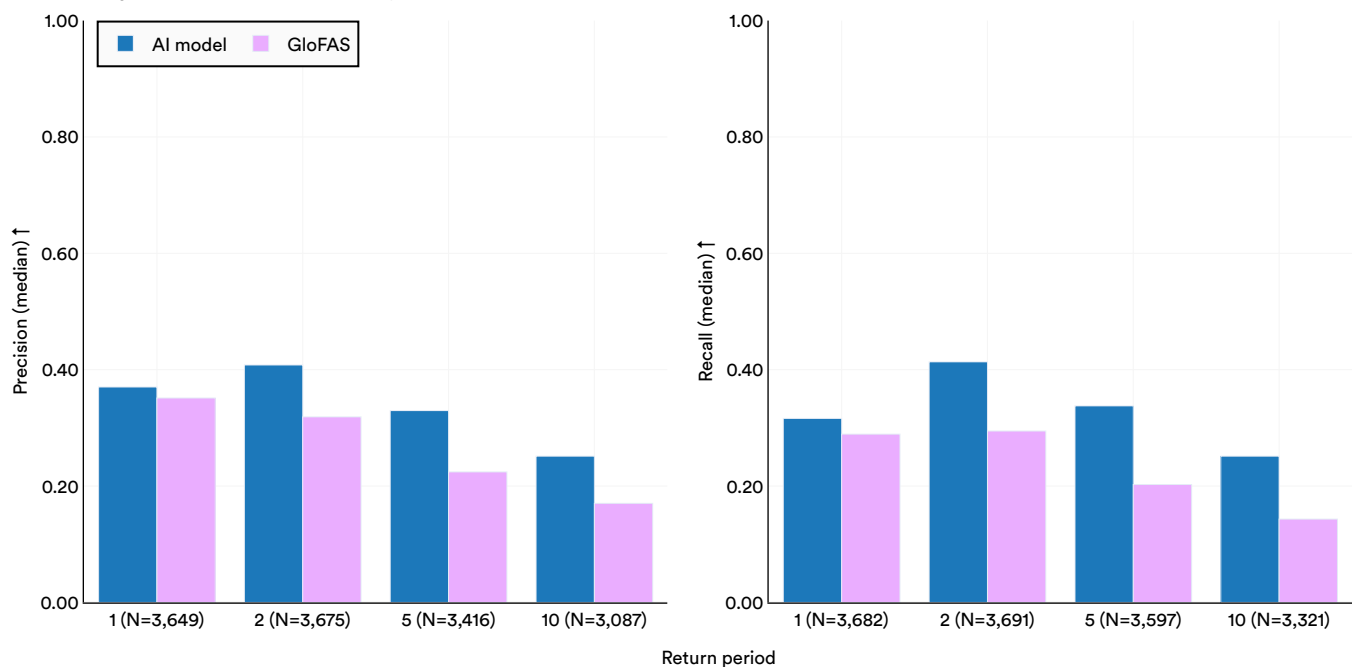


Figure 5.1.11

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

² 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](#)

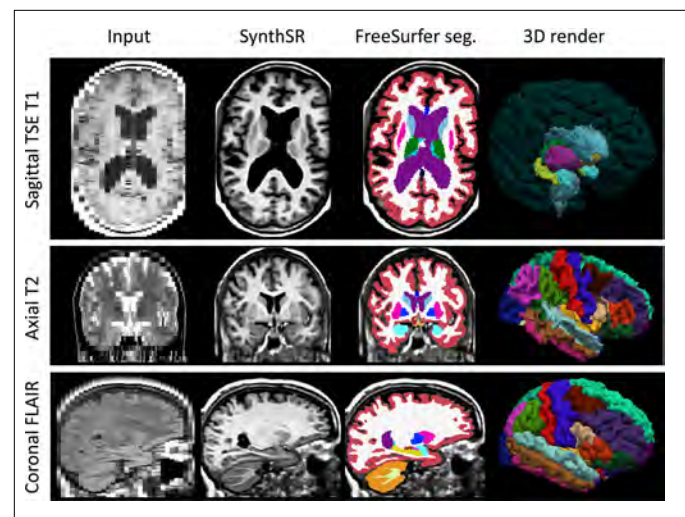


Figure 5.2.1

SynthSR correlation with ground-truth volumes on select brain regions

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

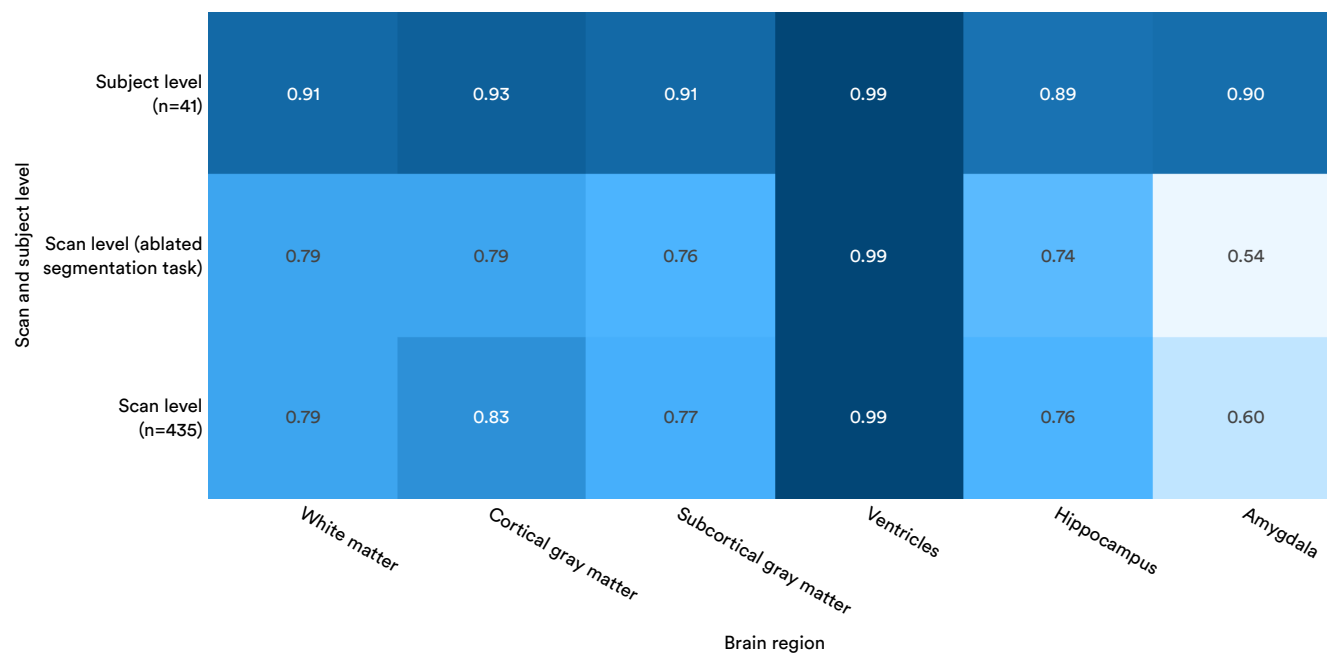


Figure 5.2.2

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

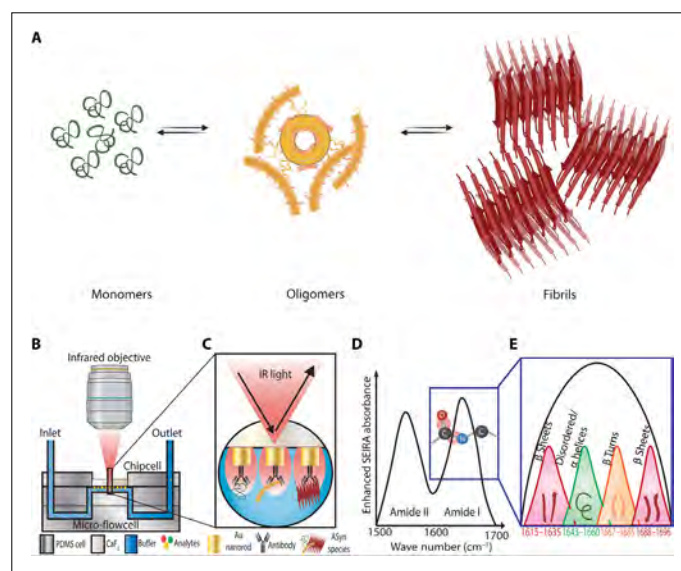


Figure 5.2.3

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

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

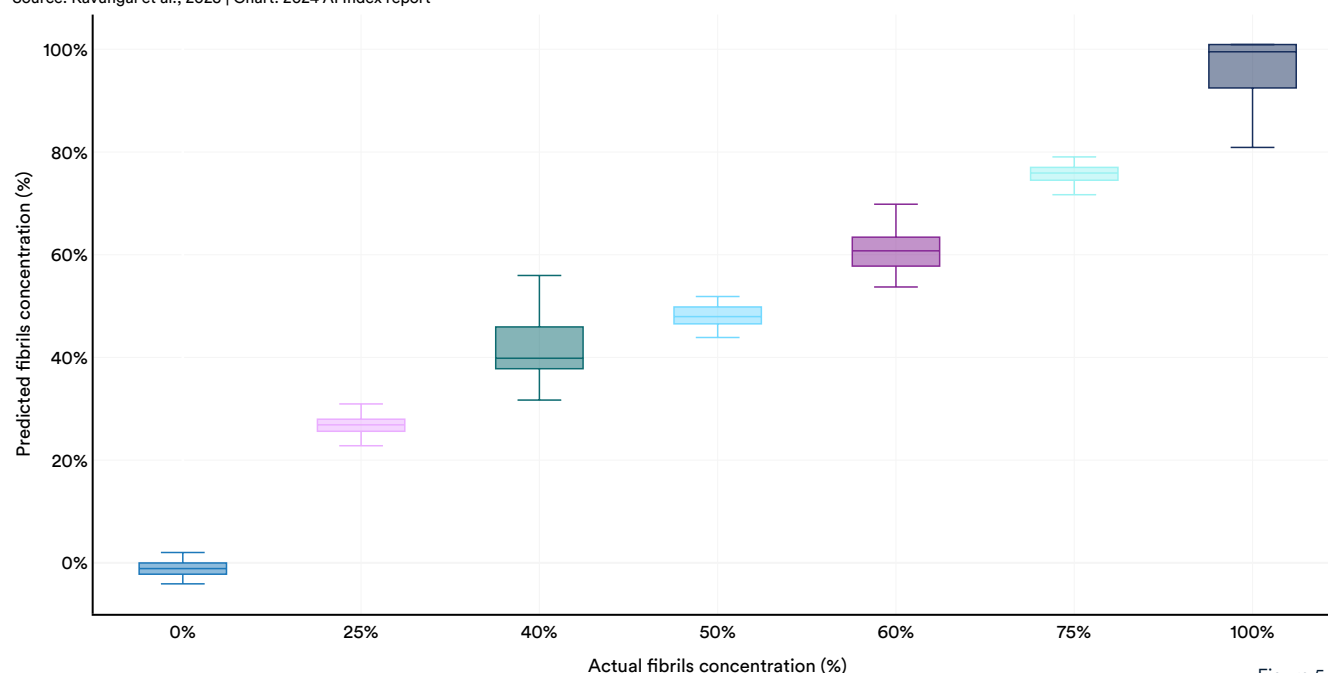


Figure 5.2.4

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

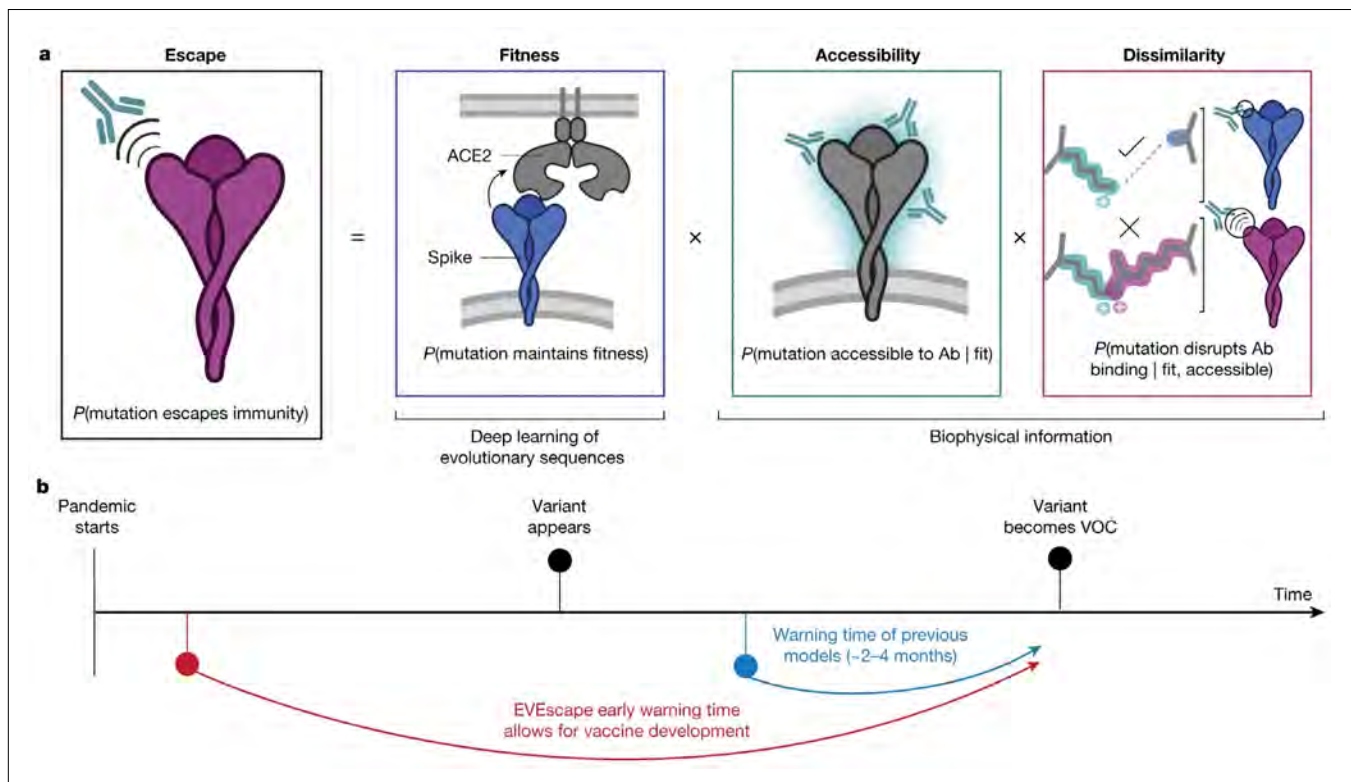


Figure 5.2.5

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

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

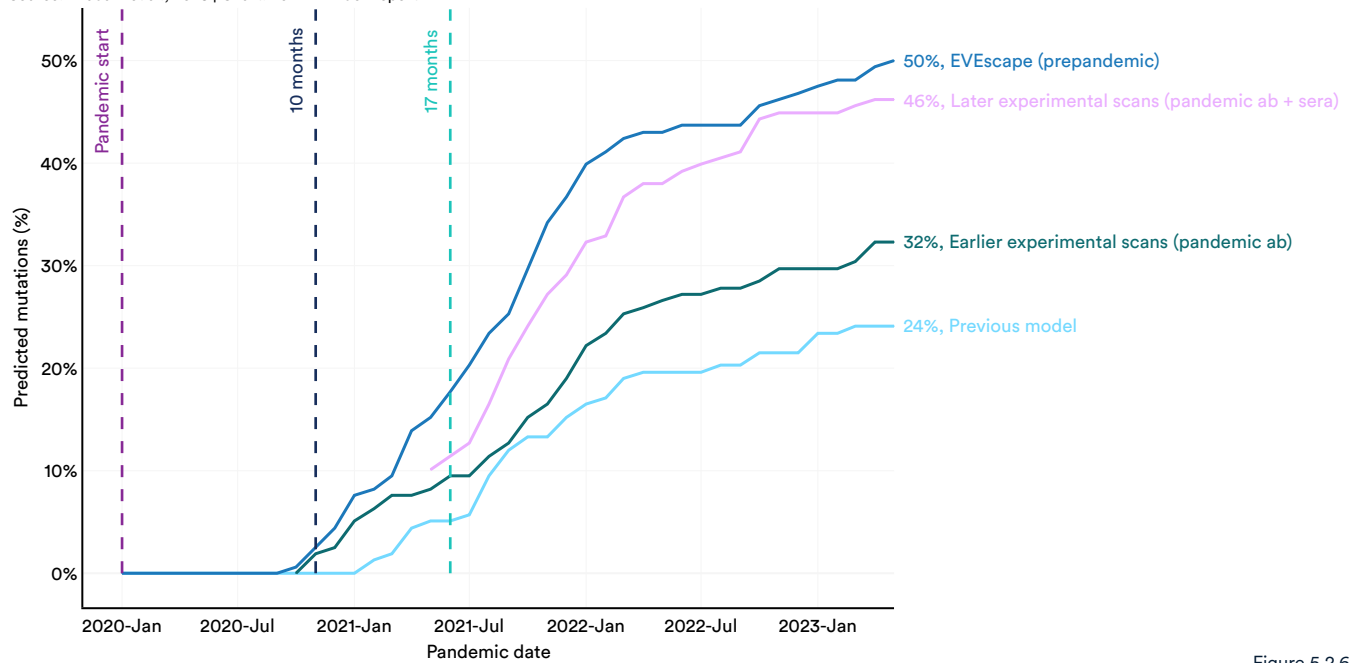


Figure 5.2.6

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

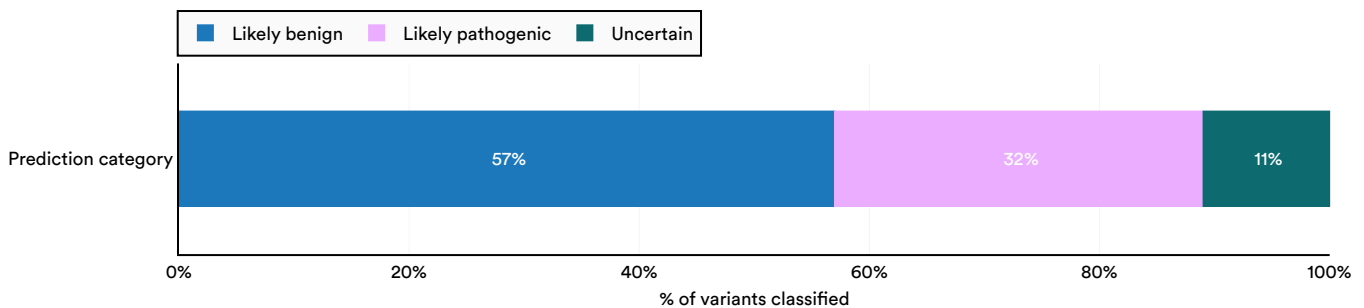


Figure 5.2.8

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

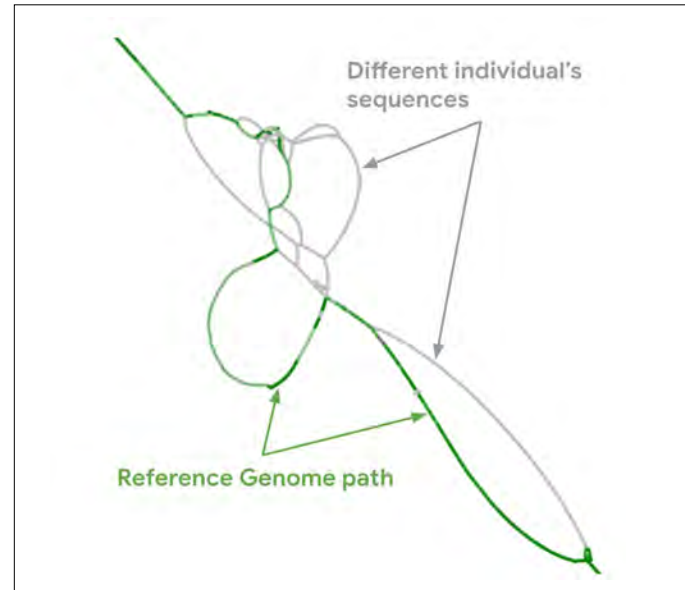


Figure 5.2.9

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

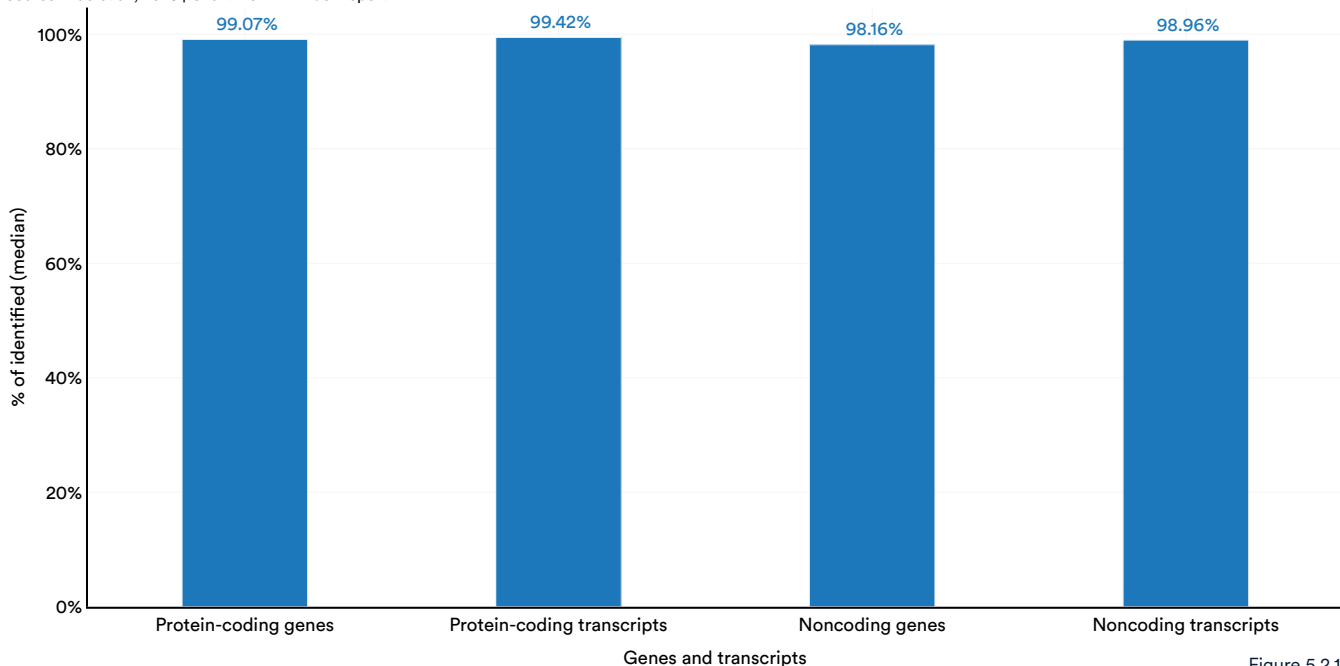


Figure 5.2.10

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.

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.

AI performance on the MedQA benchmark has seen

MedQA: accuracy

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

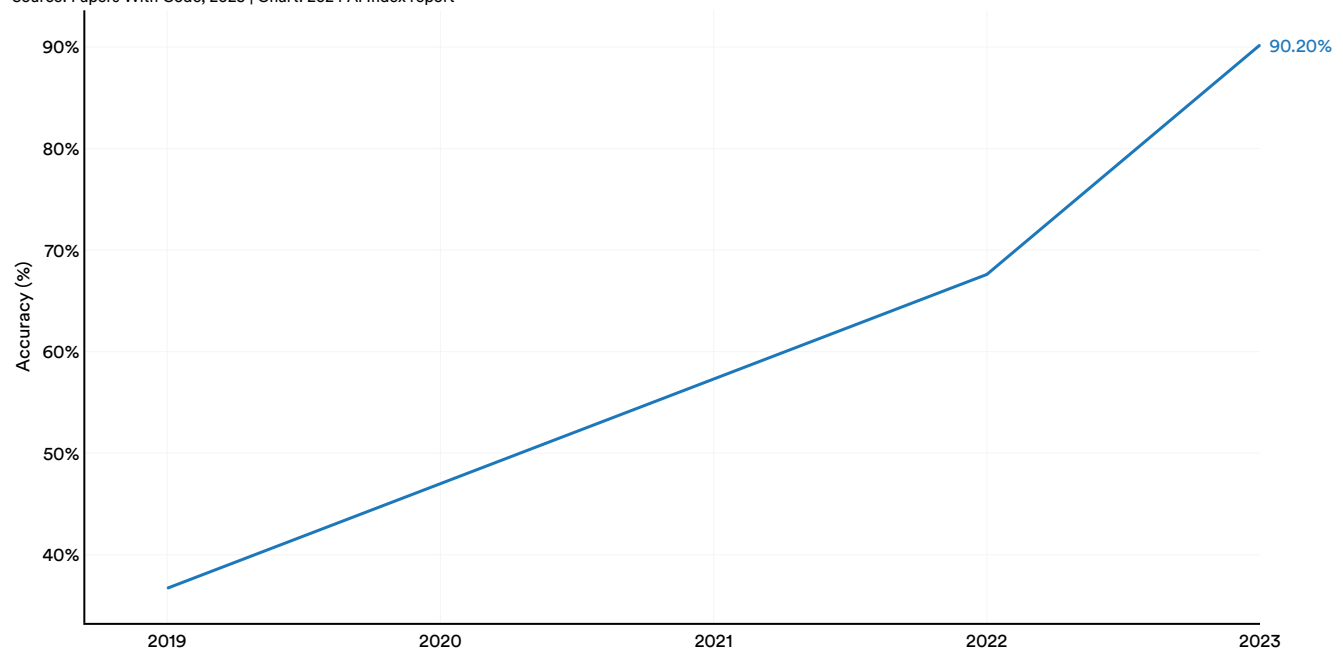


Figure 5.2.11

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

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 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

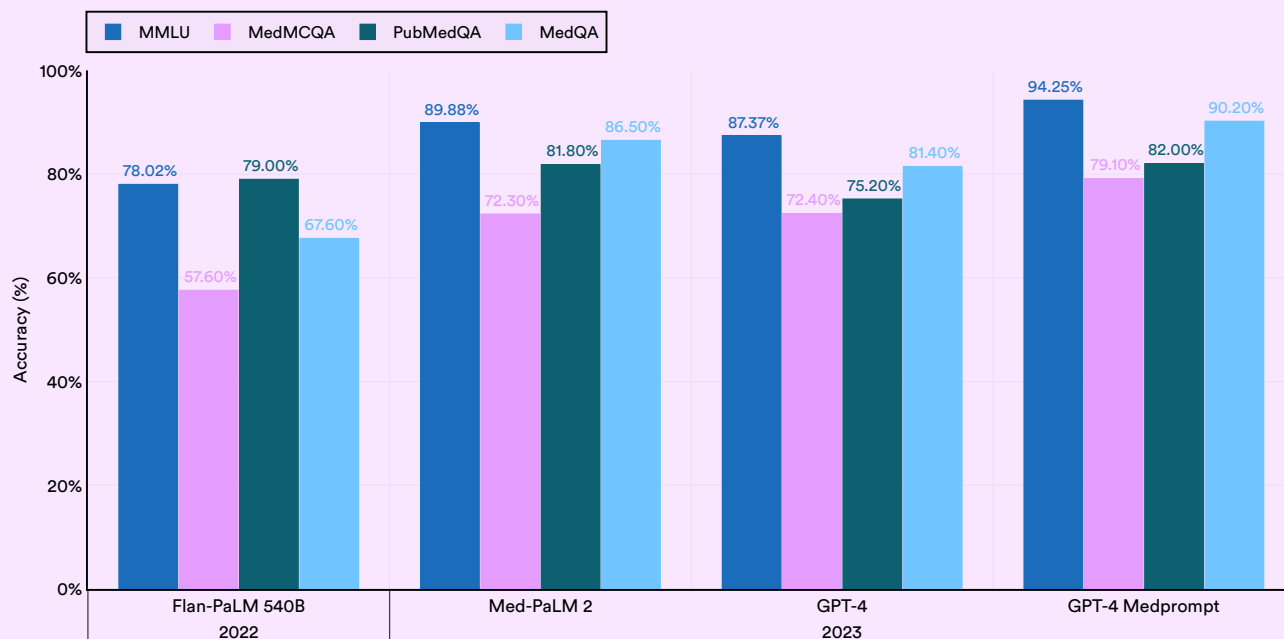


Figure 5.2.13

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:

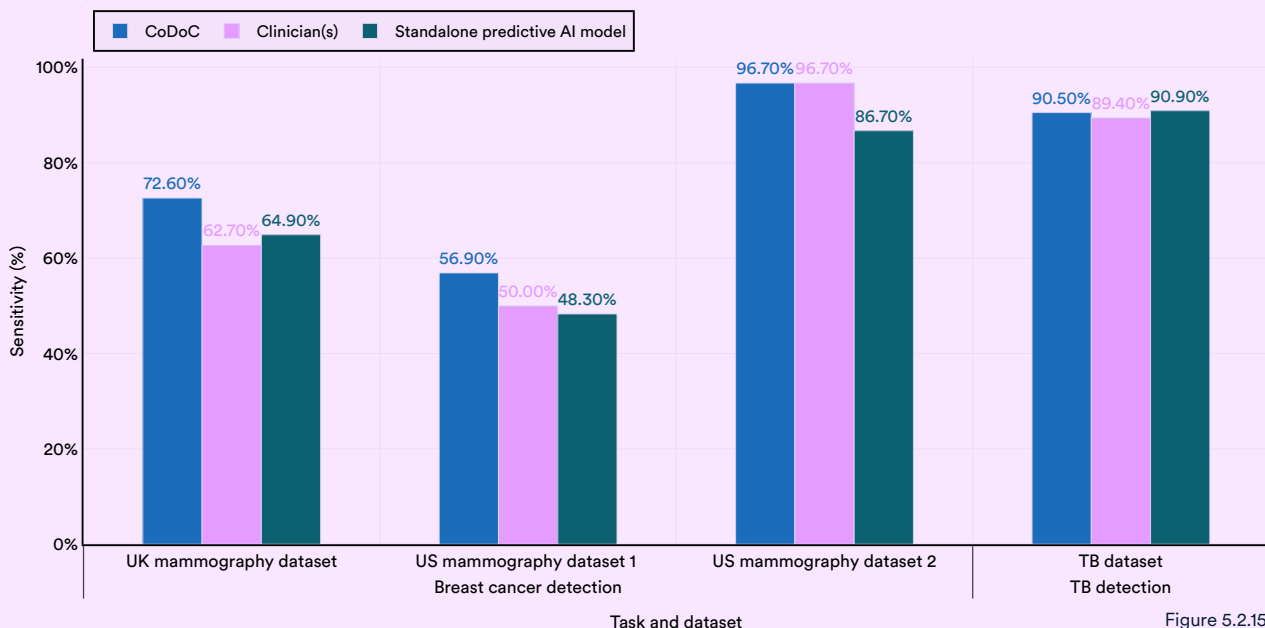
CoDoC

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



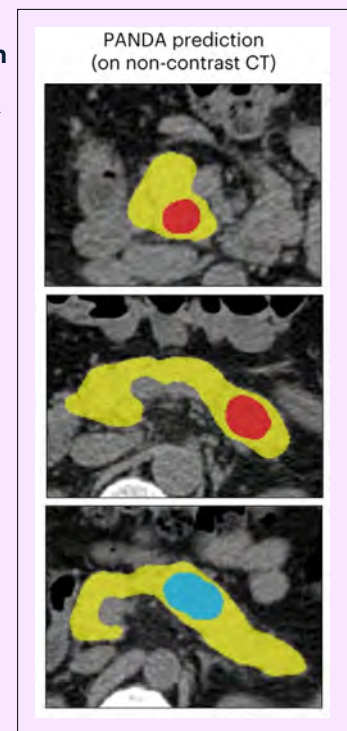
Highlighted Research:

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

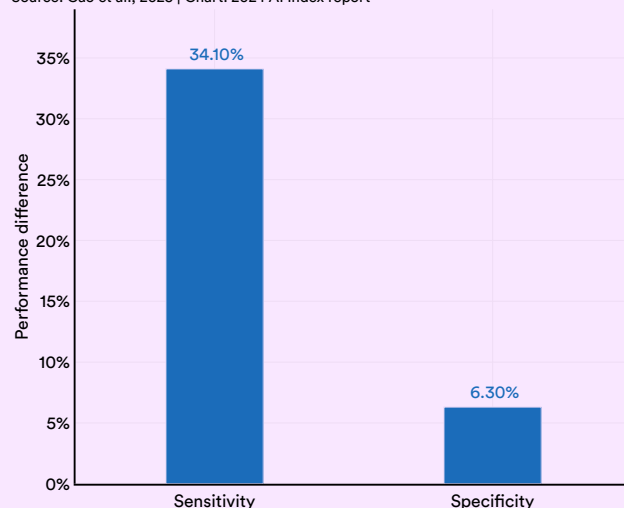


Figure 5.2.17

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

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

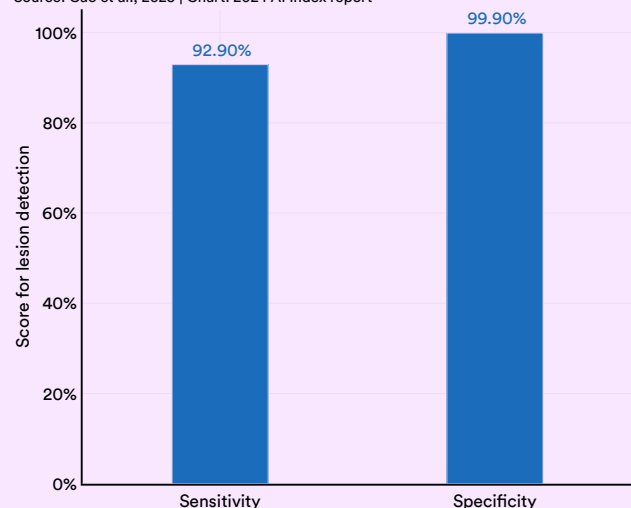


Figure 5.2.18

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

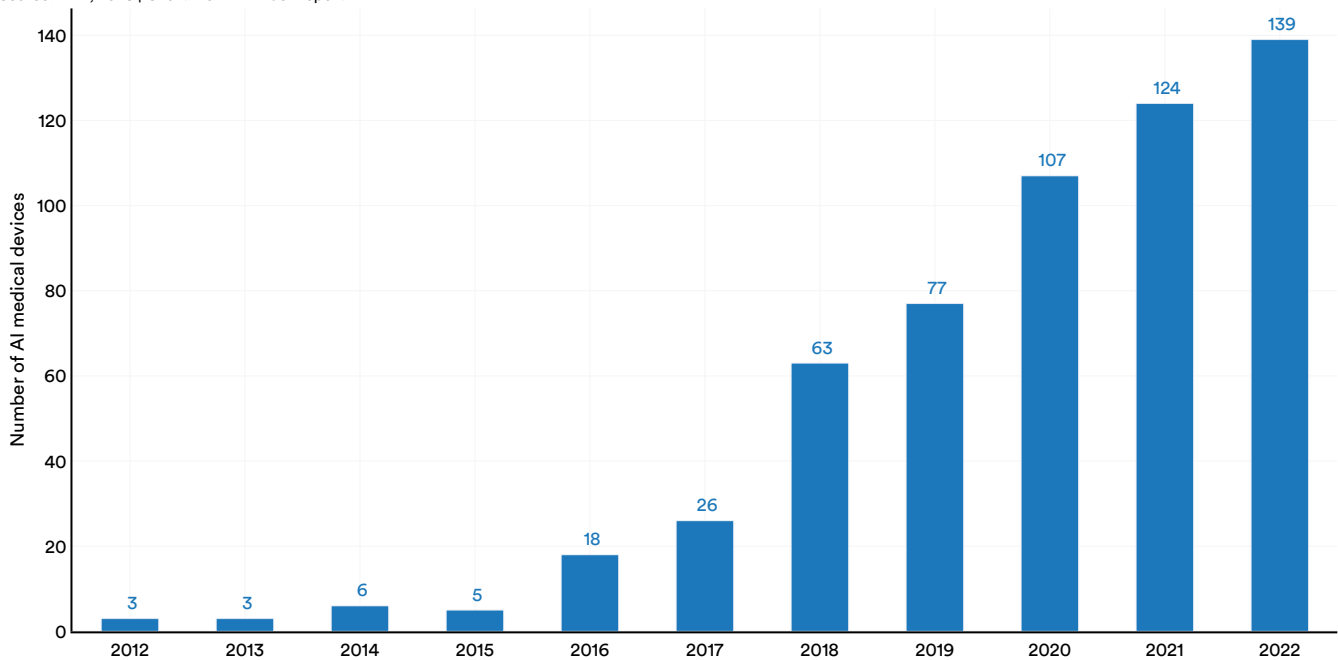


Figure 5.2.20

³ 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

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Radiology	2		5		11	15	39	51	94	105	121
Cardiovascular				1	4	6	9	12	7	11	10
Neurology			1		1	1	4	4		2	2
Gastroenterology and urology						1	1	1		3	1
Hematology		1				2	2	1	3		1
Microbiology		2						2	1		
General hospital				1				2			
General and plastic surgery					1		2	1		1	
Ophthalmic				1			2	1	1	1	2
Clinical chemistry				1	1		2	1			1
Anesthesiology				1		1			1		
Pathology	1									1	
Ear nose and throat											1
Dental							1				
Orthopedic							1				
Obstetrics and gynecology								1			

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](#)

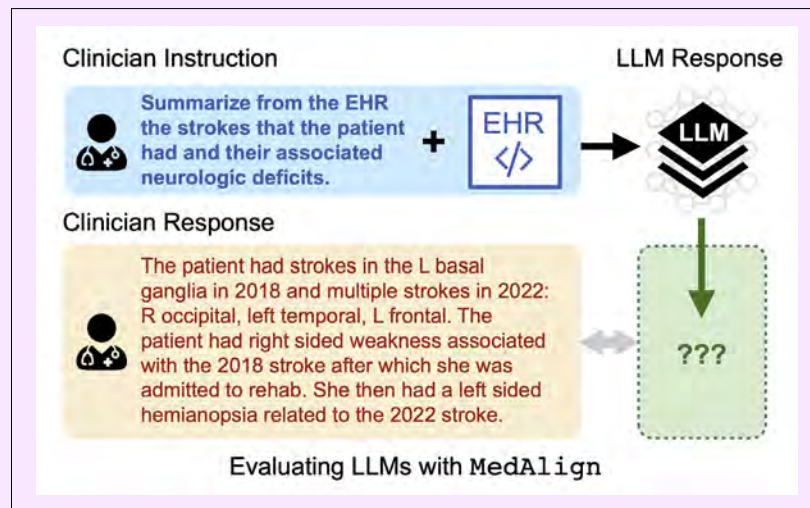


Figure 5.2.22

Highlighted Research:

MedAlign (cont'd)

Evaluation of model performance: human vs. COMET ranks

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

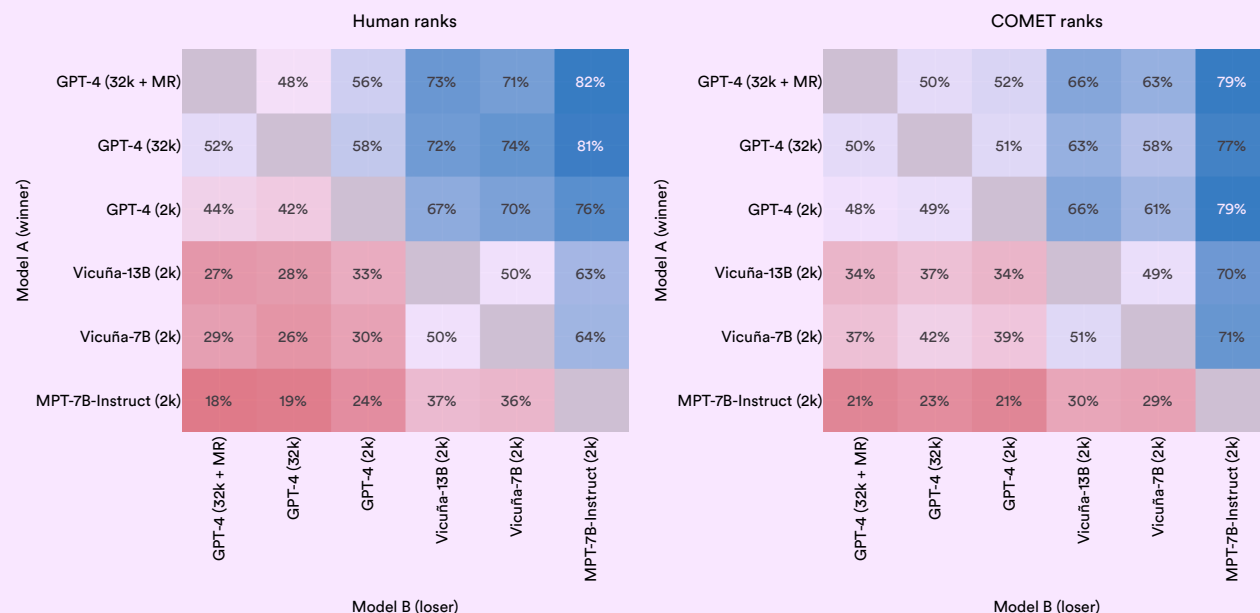
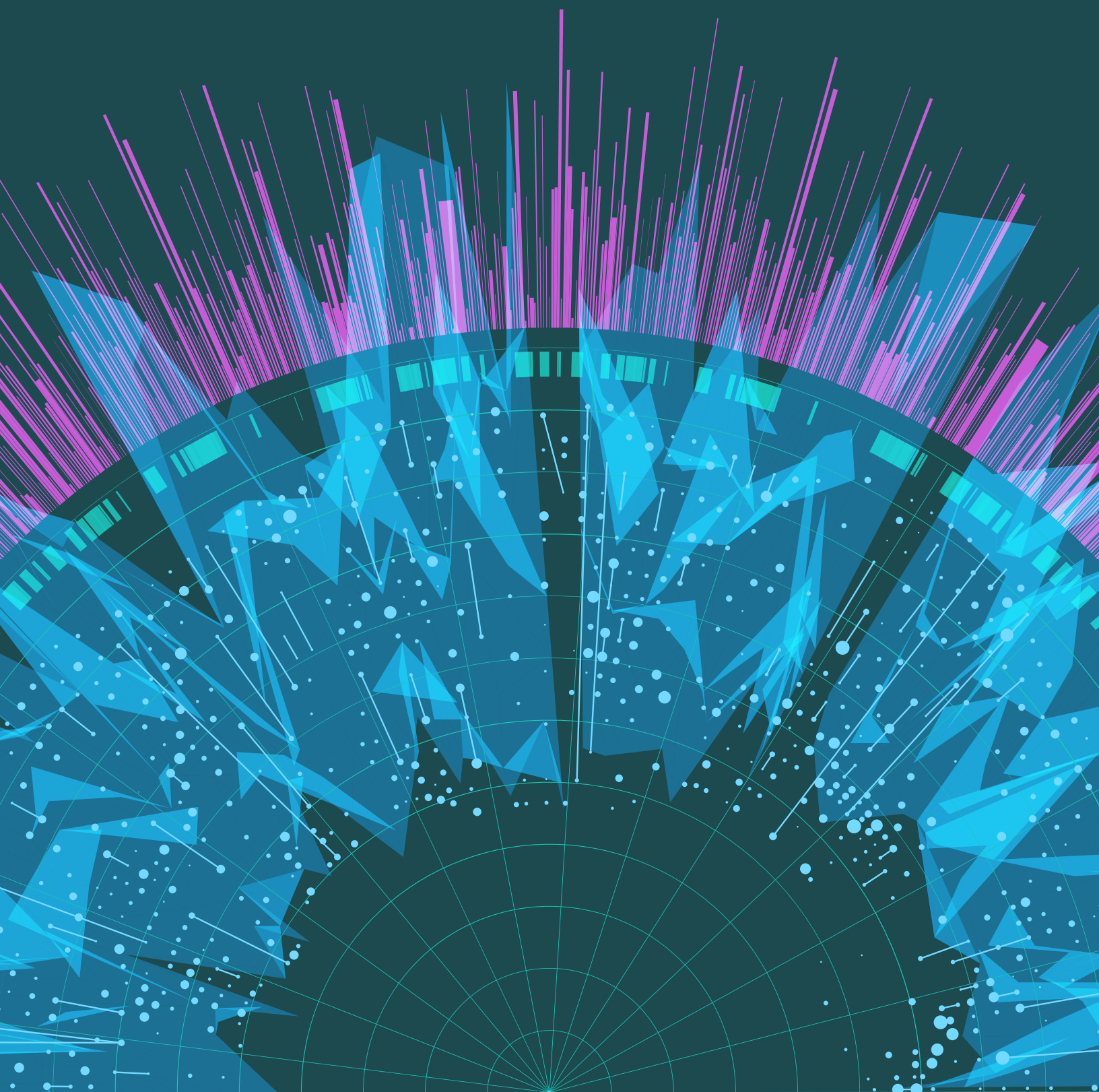


Figure 5.2.23



Artificial Intelligence
Index Report 2024

CHAPTER 6: Education



Preview

Overview	327
Chapter Highlights	328

6.1 Postsecondary CS and AI Education 329

United States and Canada	329
CS Bachelor's Graduates	329
CS Master's Graduates	331
CS PhD Graduates	333
CS, CE, and Information Faculty	336
Europe	344
Informatics, CS, CE, and IT Bachelor's Graduates	344
Informatics, CS, CE, and IT Master's Graduates	347
Informatics, CS, CE, and IT PhD Graduates	351
AI-Related Study Programs	355
Total Courses	355
Education Level	356
Geographic Distribution	357

6.2 K–12 CS and AI Education 359

United States	359
State-Level Trends	359
AP Computer Science	361
Highlight: Access Issues	363
Highlight: ChatGPT Usage Among Teachers and Students	364

ACCESS THE PUBLIC DATA

Overview

This chapter examines trends in AI and computer science (CS) education, focusing on who is learning, where they are learning, and how these trends have evolved over time. Amid growing concerns about AI's impact on education, it also investigates the use of new AI tools like ChatGPT by teachers and students.

The analysis begins with an overview of the state of postsecondary CS and AI education in the United States and Canada, based on the Computing Research Association's annual Taulbee Survey. It then reviews data from Informatics Europe regarding CS education in Europe. This year introduces a new section with data from Studyportals on the global count of AI-related English-language study programs.

The chapter wraps up with insights into K–12 CS education in the United States from Code.org and findings from the Walton Foundation survey on ChatGPT's use in schools.

Chapter Highlights

1. The number of American and Canadian CS bachelor's graduates continues to rise, new CS master's graduates stay relatively flat, and PhD graduates modestly grow. While the number of new American and Canadian bachelor's graduates has consistently risen for more than a decade, the number of students opting for graduate education in CS has flattened. Since 2018, the number of CS master's and PhD graduates has slightly declined.

2. The migration of AI PhDs to industry continues at an accelerating pace. In 2011, roughly equal percentages of new AI PhDs took jobs in industry (40.9%) and academia (41.6%). However, by 2022, a significantly larger proportion (70.7%) joined industry after graduation compared to those entering academia (20.0%). Over the past year alone, the share of industry-bound AI PhDs has risen by 5.3 percentage points, indicating an intensifying brain drain from universities into industry.

3. Less transition of academic talent from industry to academia. In 2019, 13% of new AI faculty in the United States and Canada were from industry. By 2021, this figure had declined to 11%, and in 2022, it further dropped to 7%. This trend indicates a progressively lower migration of high-level AI talent from industry into academia.

4. CS education in the United States and Canada becomes less international. Proportionally fewer international CS bachelor's, master's, and PhDs graduated in 2022 than in 2021. The drop in international students in the master's category was especially pronounced.

5. More American high school students take CS courses, but access problems remain. In 2022, 201,000 AP CS exams were administered. Since 2007, the number of students taking these exams has increased more than tenfold. However, recent evidence indicates that students in larger high schools and those in suburban areas are more likely to have access to CS courses.

6. AI-related degree programs are on the rise internationally. The number of English-language, AI-related postsecondary degree programs has tripled since 2017, showing a steady annual increase over the past five years. Universities worldwide are offering more AI-focused degree programs.

7. The United Kingdom and Germany lead in European informatics, CS, CE, and IT graduate production. The United Kingdom and Germany lead Europe in producing the highest number of new informatics, CS, CE, and information bachelor's, master's, and PhD graduates. On a per capita basis, Finland leads in the production of both bachelor's and PhD graduates, while Ireland leads in the production of master's graduates.

This section provides an overview of postsecondary education in CS and AI, highlighting graduation statistics across North America and Europe for various degrees including bachelor's, master's, and PhDs. It also covers information on AI-related courses offered in English.

6.1 Postsecondary CS and AI Education

United States and Canada

This subsection presents an analysis of data from the Computing Research Association's [Taulbee Survey](#), which evaluates the state of CS and AI postsecondary education in the United States and Canada. The survey covers 297 PhD-granting CS departments across the United States and Canada.¹

CS Bachelor's Graduates

Over the past decade, the total number of new CS bachelor's graduates in North America has steadily risen, increasing more than threefold, with a 7.9% year-over-year rise from 2021 to 2022 (Figure 6.1.1).

New CS bachelor's graduates in the United States and Canada, 2010–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

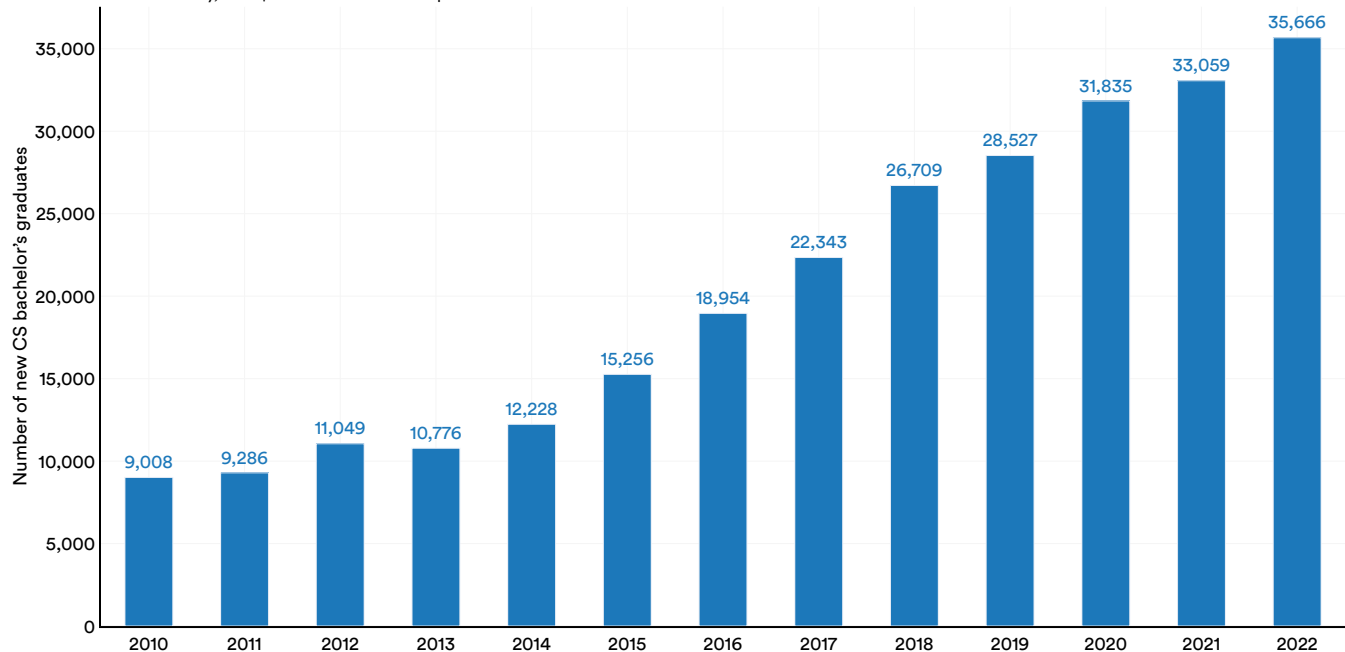


Figure 6.1.1

¹ It is important to note that not all PhD-granting departments targeted in the survey provided responses. Out of the 297 departments targeted, only 182 responded, yielding an overall response rate of 61%.

For the first time in almost eight years, the proportion of international students among CS bachelor's graduates in American and Canadian universities declined, falling from 16.3% in 2021 to 15.2% in 2022 (Figure 6.1.2). This decline likely reflects the increased difficulty of obtaining study visas during the early years of the Trump administration, an impact that is only now beginning to manifest in

the data. The decline is also partially attributable to international travel restrictions that were imposed during the COVID-19 pandemic, affecting the ability of international students to study in the United States and Canada. Despite this recent drop, the overall trend over the last decade shows a steady increase in the proportion of international students.

New international CS bachelor's graduates (% of total) in the United States and Canada, 2010–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

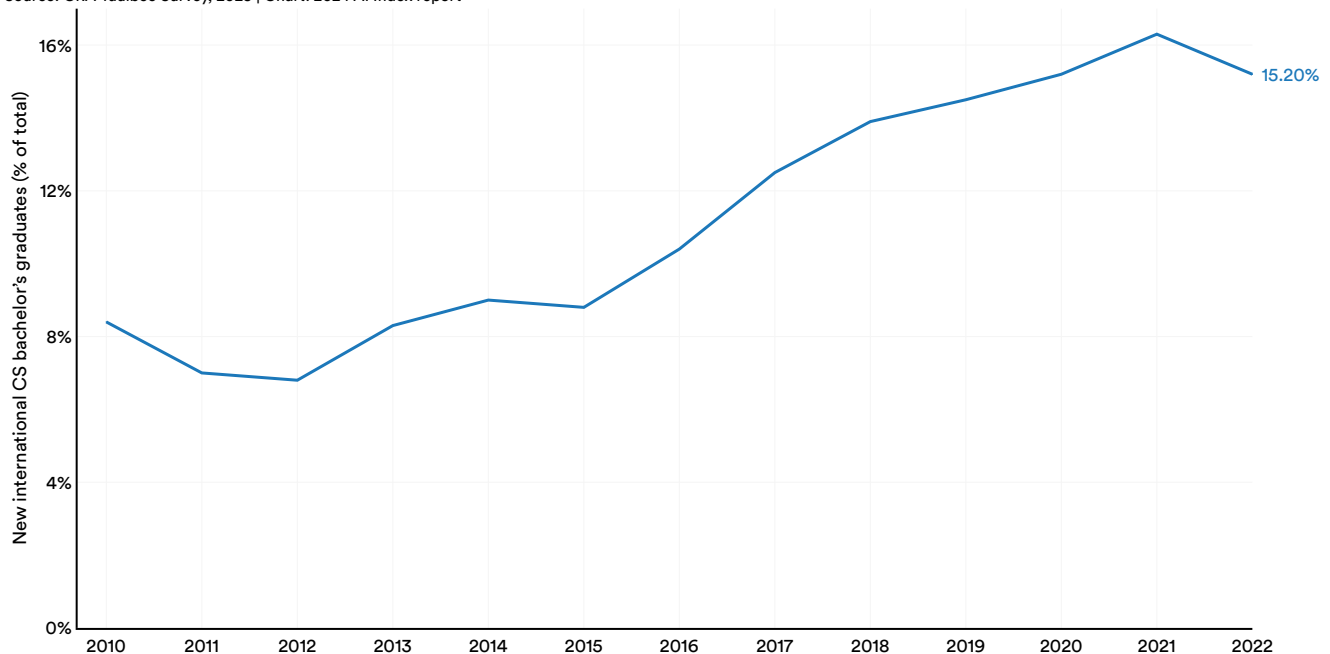


Figure 6.1.2



CS Master's Graduates

AI courses are commonly included in CS master's degree programs. While the total number of new CS master's graduates from American and Canadian universities more than doubled over the past decade,

the number appears to have leveled out since 2018 and slightly decreased, by 2.5%, last year (Figure 6.1.3). This leveling is a reflection of the decline in international master's students shown in the following graph.

New CS master's graduates in the United States and Canada, 2010–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

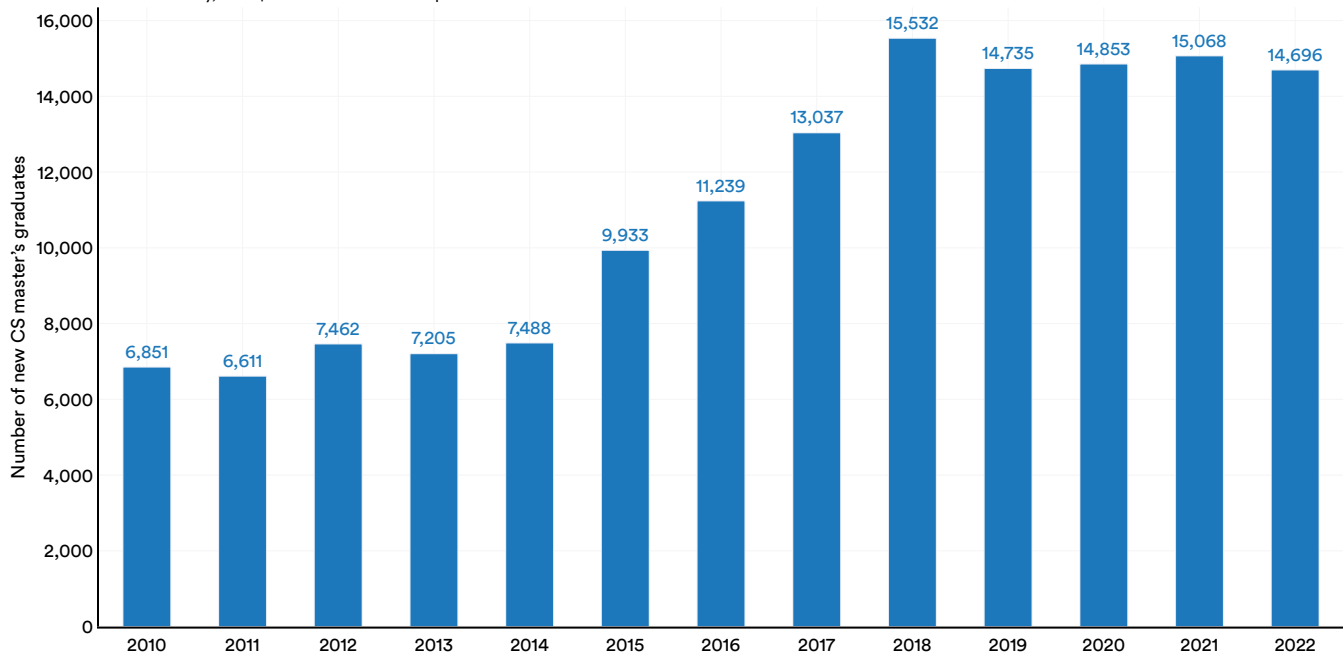


Figure 6.1.3

In 2022, American and Canadian universities experienced a notable decrease in international CS master's students. This downward trend began around 2017, but the decline was most pronounced last year, at 14.8 percentage points (Figure 6.1.4). Currently, the split between international and domestic CS master's graduates is roughly even.

New international CS master's graduates (% of total) in the United States and Canada, 2010–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

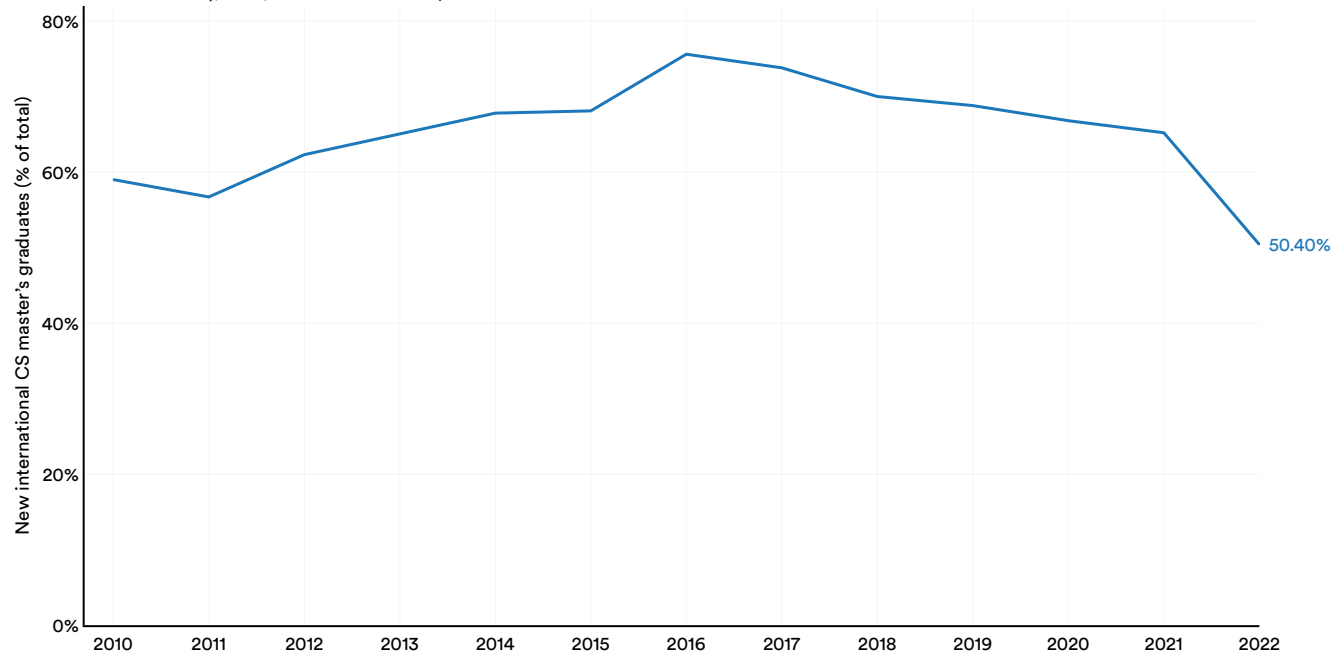


Figure 6.1.4



CS PhD Graduates

For the first time in a decade, there has been a significant increase in the number of new CS PhD graduates at American and Canadian universities. In 2022, the number of CS PhD graduates reached 2,105, the highest since 2010 (Figure 6.1.5).

New CS PhD graduates in the United States and Canada, 2010–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

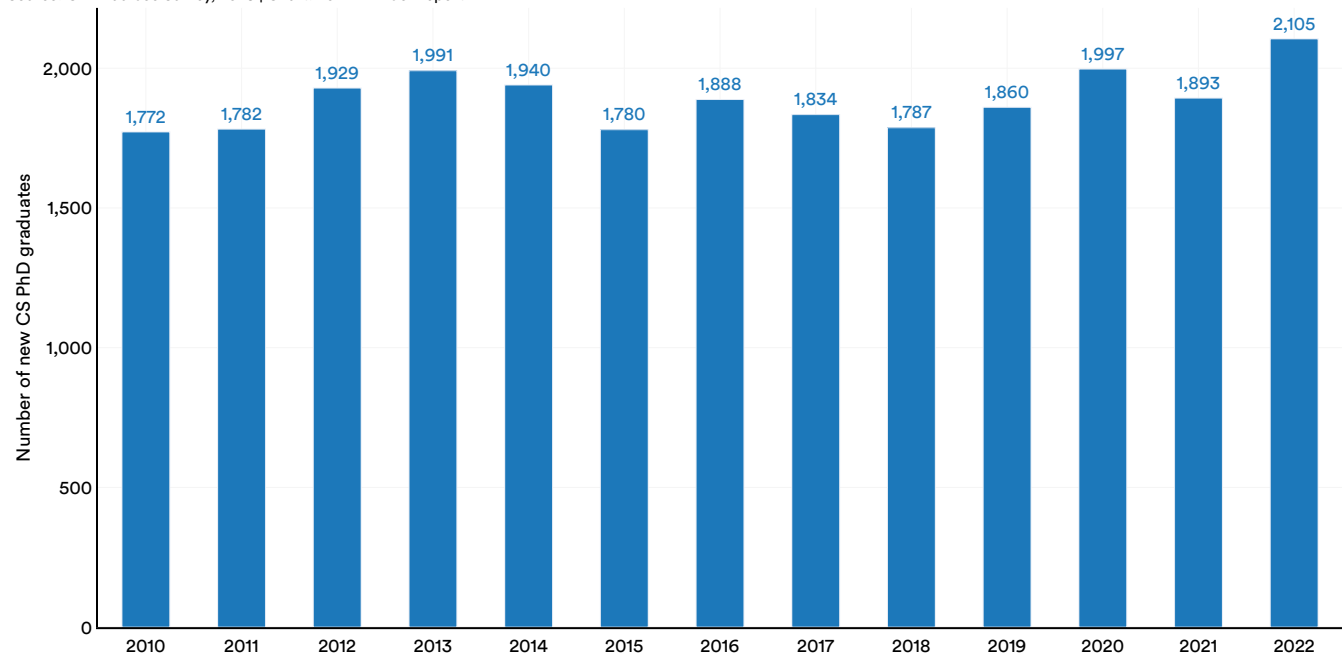


Figure 6.1.5

While the proportion of international students among CS PhD graduates has risen over the past decade, there was a slight decrease in this proportion in the last year, dropping from 68.6% in 2021 to 65.9% in 2022 (Figure 6.1.6).

New international CS PhD graduates (% of total) in the United States and Canada, 2010–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

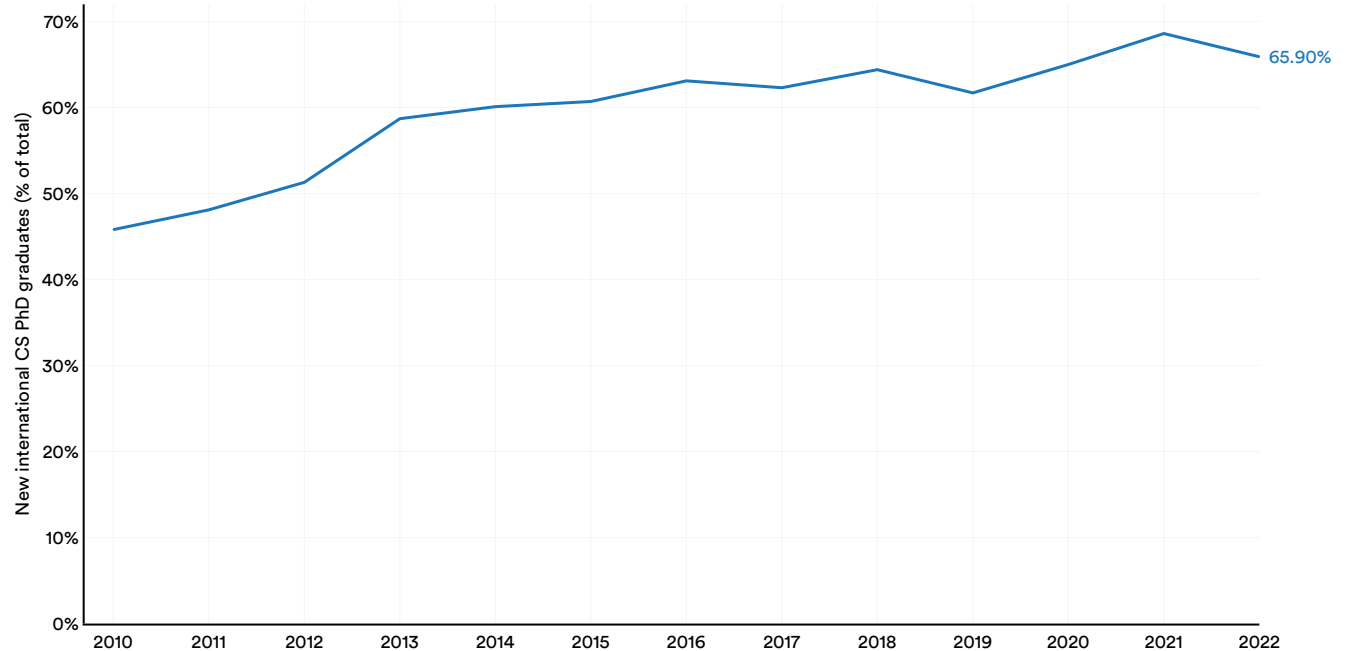


Figure 6.1.6

Where do newly minted AI PhDs choose to work after graduating? Following a trend highlighted in last year's AI Index report, a growing share of AI doctoral recipients are pursuing careers in industry (Figure 6.1.7 and Figure 6.1.8). In 2011, around the same percentage took jobs in industry (40.9%) as in academia (41.6%).

However, by 2022, a significantly larger proportion (70.7%) joined industry after graduation compared to those entering academia (20.0%). The percentage of new AI PhDs going into government roles has remained relatively low and steady at around 0.7% over the past five years.

Employment of new AI PhDs (% of total) in the United States and Canada by sector, 2010–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

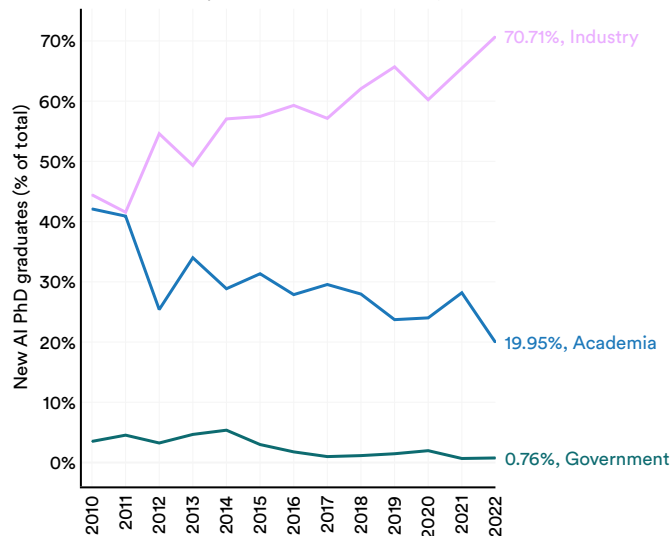


Figure 6.1.7²

Employment of new AI PhDs in the United States and Canada by sector, 2010–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

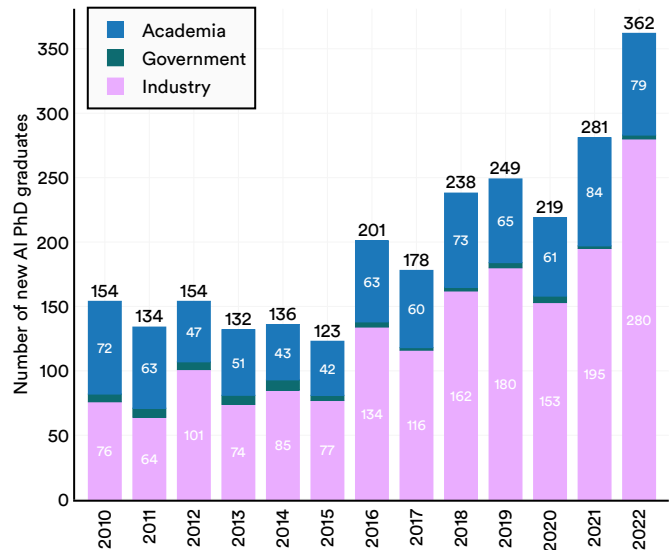


Figure 6.1.8

² The sums in Figure 6.1.7 do not add up to 100, as there is a subset of new AI PhDs each year who become self-employed, unemployed, or report an "other" employment status in the CRA survey. These students are not included in the chart.

CS, CE, and Information Faculty

To better understand trends in CS and AI education, it is helpful to examine data on CS faculty. Last year, the total number of CS, CE, and information faculty in American and Canadian universities increased 7.2% (Figure 6.1.9). Since 2011, the increase is 42.4%.

Number of CS, CE, and information faculty in the United States and Canada, 2011–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

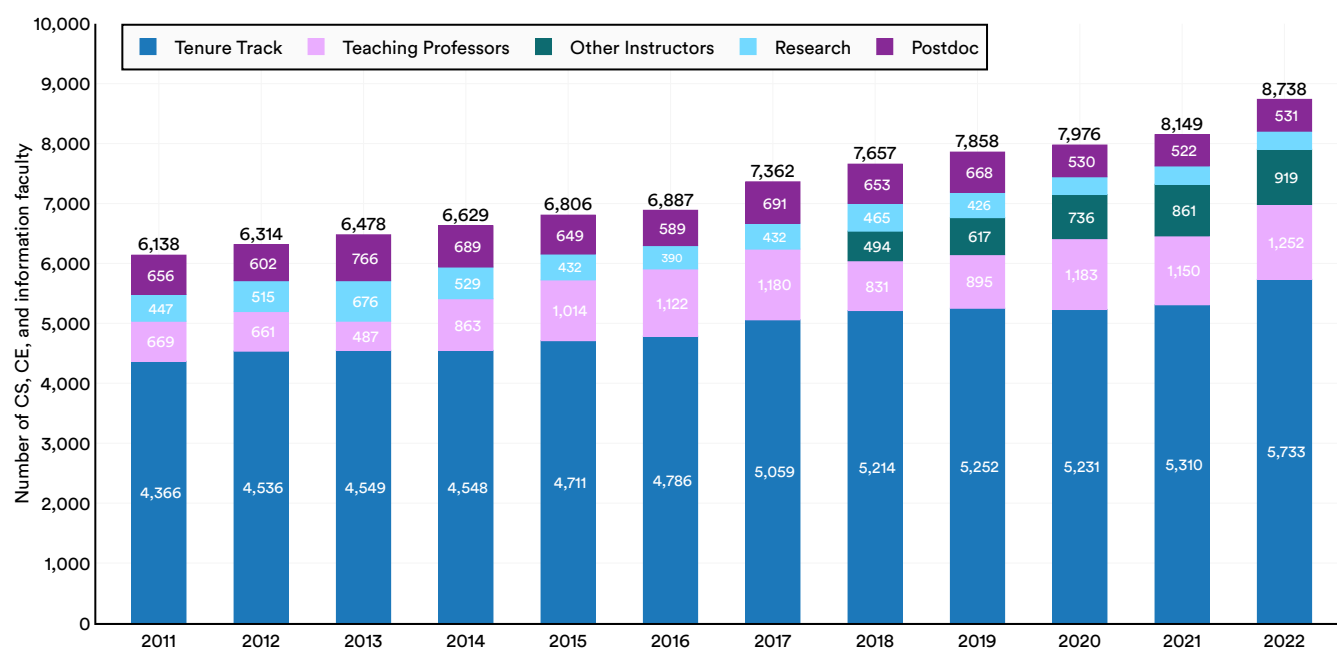


Figure 6.1.9

In 2022, the United States had 7,084 CS faculty members, with the majority (65.7%) on the tenure track (Figure 6.1.10). The total number of American CS faculty has risen 4.4% since 2021 and 45.0% since 2011.

Number of CS faculty in the United States, 2011–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

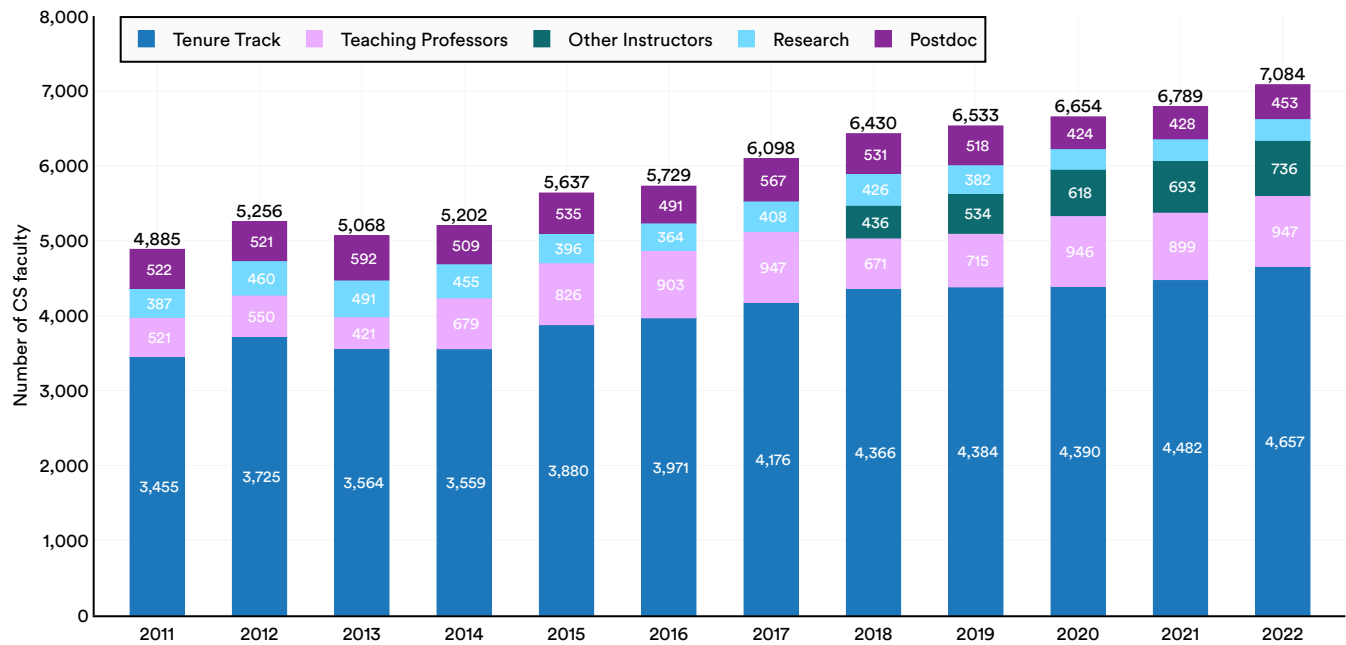


Figure 6.1.10

Last year, 915 new faculty were hired across CS, CE, and information disciplines in North America, a decade high. 455 of these positions were tenure track. (Figure 6.1.11).

New CS, CE, and information faculty hires in the United States and Canada, 2011–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

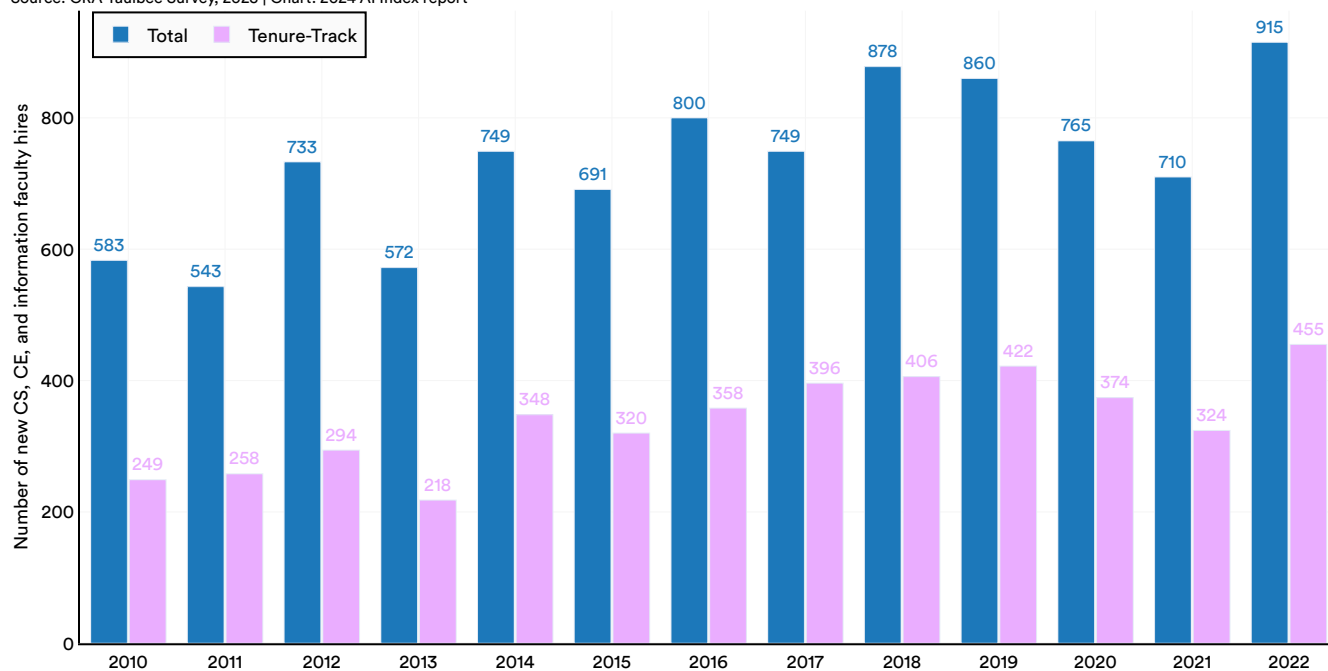


Figure 6.1.11

In 2022, 43% of new faculty appointments came from other academic positions, indicating a “churn” within the academic workforce (Figure 6.1.12). Since these “new” faculty members vacated positions elsewhere, their previous roles will eventually need to be filled. Additionally, the proportion of faculty transitioning from industry in 2022 fell to 7% from 11% in the previous year and 13% in 2019.

Source of new faculty in American and Canadian CS, CE, and information departments, 2018–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

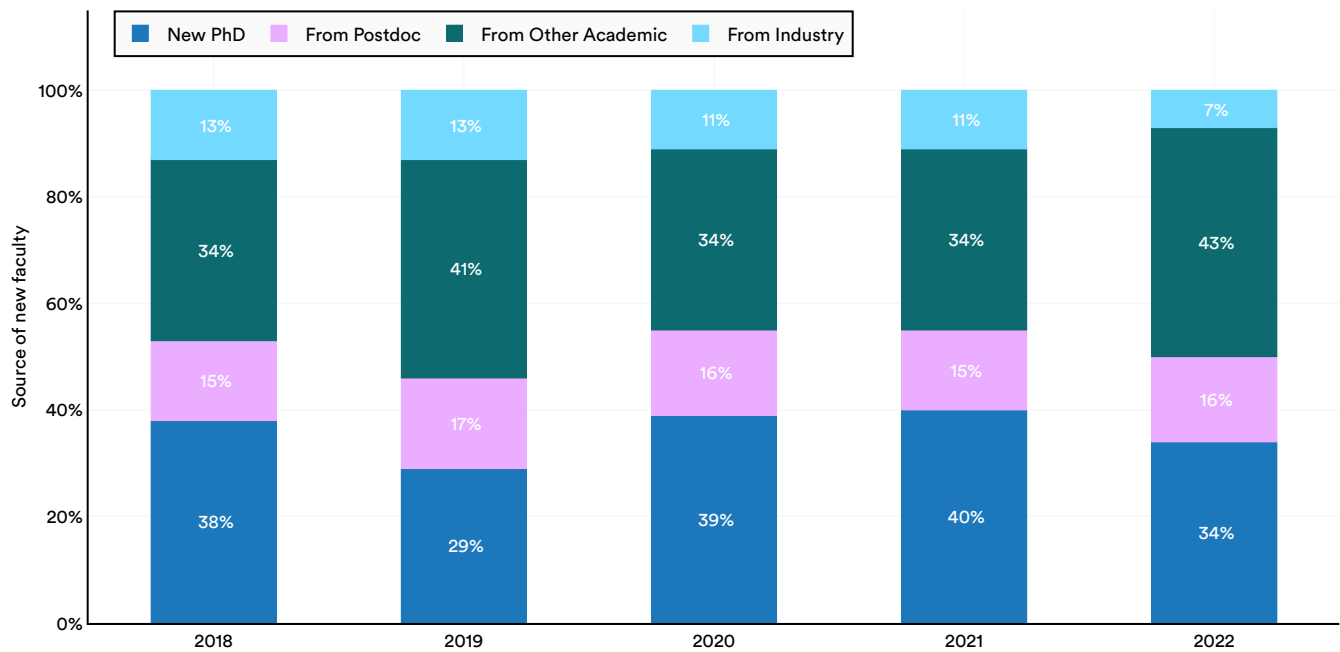


Figure 6.1.12

The reasons for faculty positions remaining unfilled have varied over the past decade. In 2011, 37% of failed searches were due to no offer being made, while 34% were because the offer made was declined (Figure 6.1.13). In contrast, in 2022, only 15% ended with no offer being made, while 55% involved offers that

were turned down. This trend appears to reflect an increasingly competitive market for new CS faculty. However, it remains unclear whether this indicates heightened competition with other academic positions or with industry positions.

Reason why new CS, CE, and information faculty positions remained unfilled (% of total), 2011–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

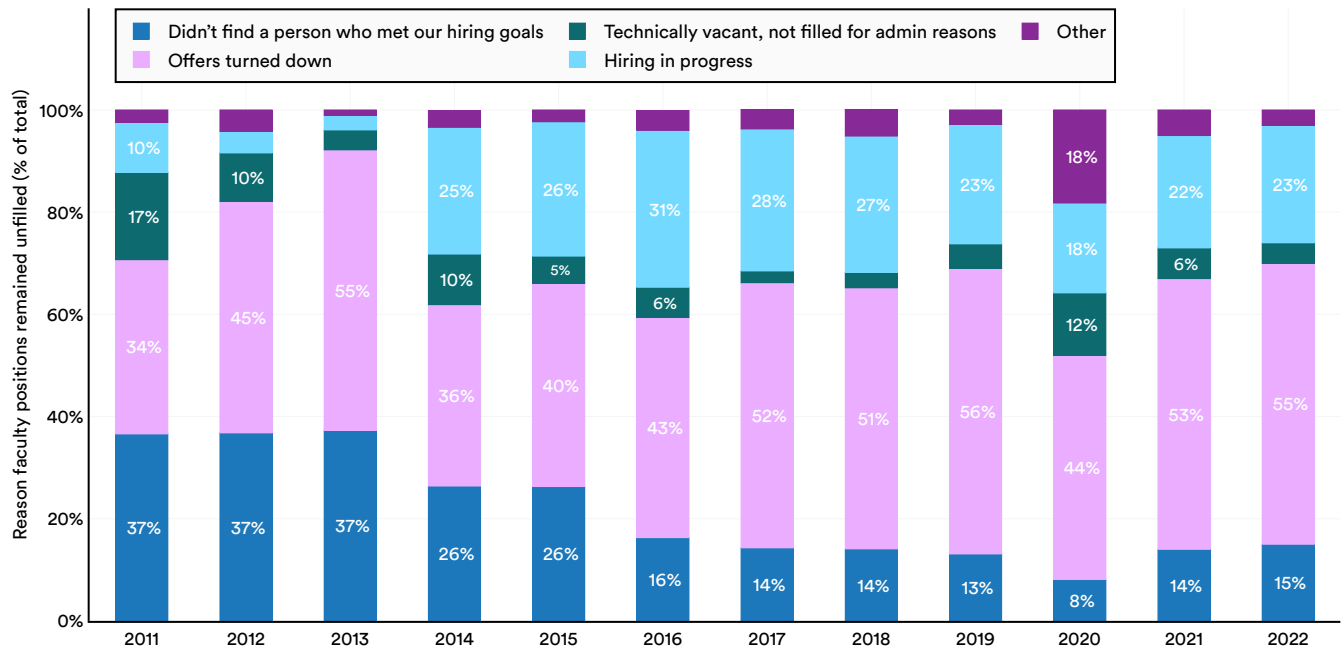


Figure 6.1.13

In 2022, North American departments in CS, CE, and information disciplines experienced a significant increase in faculty departures, totaling 405, compared to 303 in 2021 (Figure 6.1.14). Of these losses, 38.5% left for other academic positions, while 16.3% moved to nonacademic roles, maintaining a trend consistent with previous years.

Faculty losses in American and Canadian CS, CE, and information departments, 2011–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

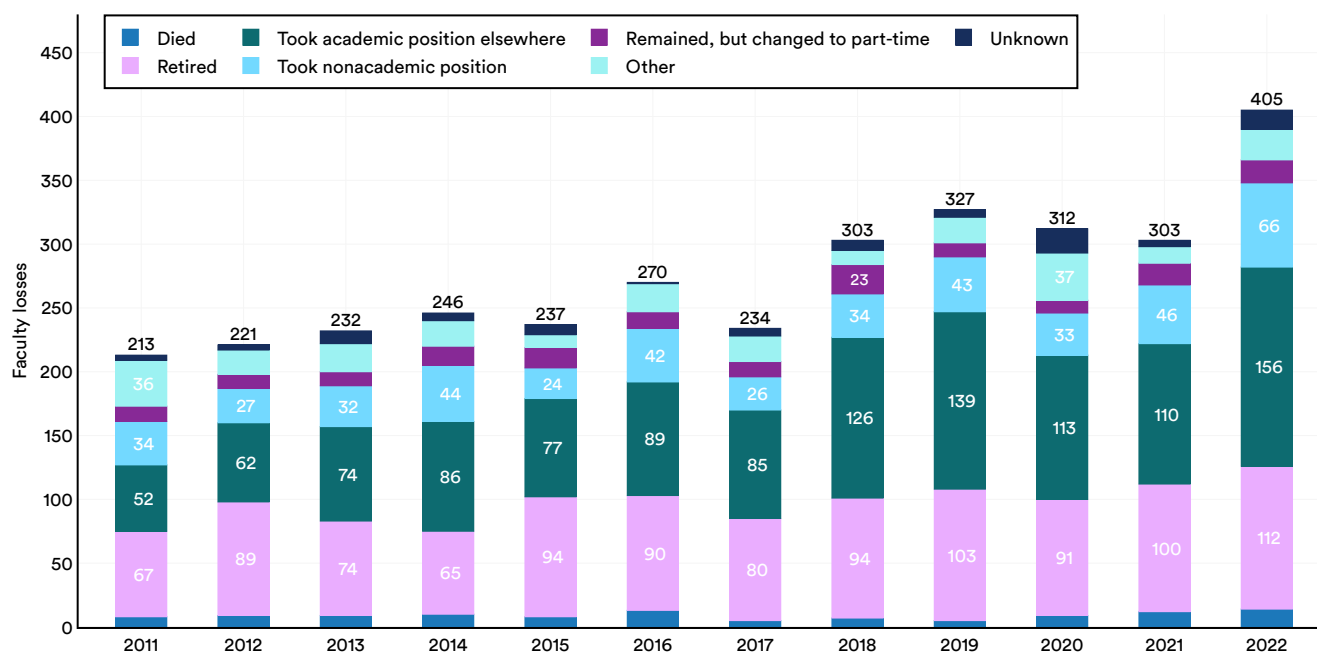


Figure 6.1.14

Since 2015, the increase in median nine-month salaries for full professors has slightly fallen below U.S. inflation rates, whereas median salaries for assistant and associate professors have seen slight increases above inflation. In 2022, a full professor's salary was 3.2% higher than in 2021, which did not keep pace with the 7% U.S. inflation rate, and 16.4% higher than in 2015, still below the 19% inflation increase over those years (Figure 6.1.15).

Median nine-month salary of CS faculty in the United States, 2015–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

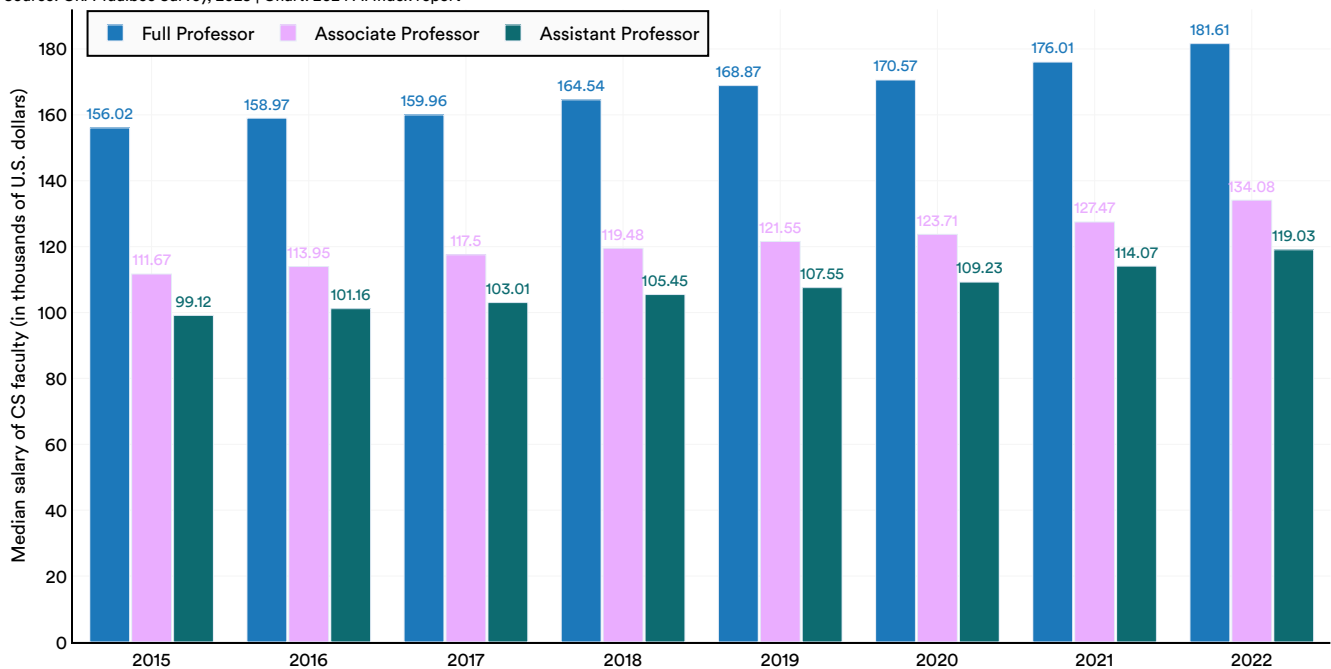


Figure 6.1.15

In 2022, the proportion of international hires among new tenure-track faculty in CS, CE, and information disciplines significantly increased to 19.3% from 13.2% the previous year (Figure 6.1.16). This marked the second-highest percentage recorded in the past decade, only surpassed by 2013.

New international CS, CE, and information tenure-track faculty hires (% of total) in the United States and Canada, 2010–22

Source: CRA Taulbee Survey, 2023 | Chart: 2024 AI Index report

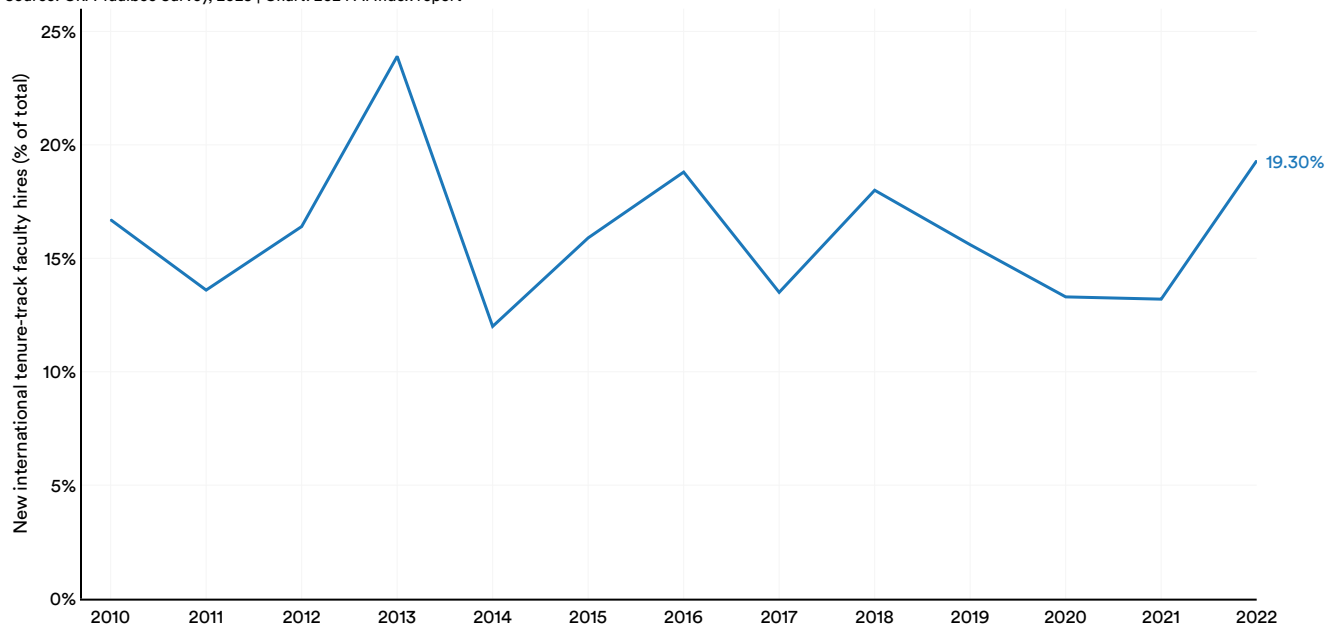


Figure 6.1.16

Europe

Data on European CS graduates comes from Informatics Europe, an academic and research community that, among other goals, monitors the state of informatics education in Europe.³ Informatics Europe gathers data on graduates in informatics, CS, CE, computing, and information technology (IT) disciplines from statistical offices of European governments.⁴

Informatics, CS, CE, and IT Bachelor's Graduates

In 2022, the United Kingdom led with the highest number of new graduates in informatics, CS, CE, and IT at the bachelor's level, totaling approximately 25,000 (Figure 6.117).⁵ Germany and Turkey followed closely. Most countries in the sample saw an increase in graduates in these fields compared to a decade ago, though there were exceptions like Poland, Spain, and the Czech Republic (Figure 6.118).

New informatics, CS, CE, and IT bachelor's graduates by country in Europe, 2022

Source: Informatics Europe, 2023 | Chart: 2024 AI Index report

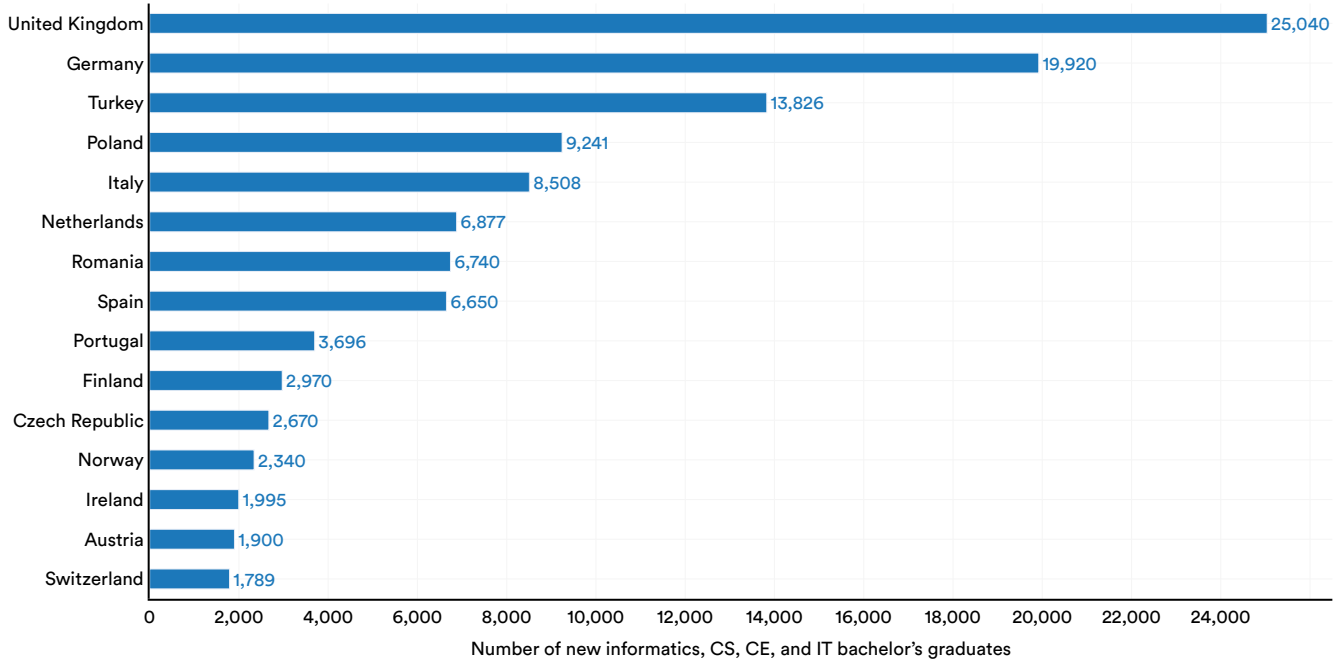


Figure 6.117

³ There is no singular term for CS education that is used uniformly across European countries. Across Europe, CS education can be reflected in terms such as informatics, computer science (CS), computer engineering (CE), computing, information technology (IT), information and communication technology (ICT), and information science and technology (IST). The full list of subject names (and English translations) that Informatics Europe uses to identify informatics studies programs can be found at the [following link](#).

⁴ Readers are cautioned against making per capita comparisons between the CRA North American data and the European CS graduate data detailed in subsequent sections, as the European data is collected from national statistical offices and boasts broader coverage.

⁵ Note that not all countries for which the AI Index has data are visualized in the figures in this section. To access the complete data, please view the public data associated with this chapter. Moreover, the year label refers to the year in which an academic year ends. For example, the figures visualizing new graduates for 2022 reflect the number of graduates reported for the 2021/2022 academic year. For the sake of visual simplicity, the Index opts to focus on the year in which students graduated.

Percentage change of new informatics, CS, CE, and IT bachelor's graduates by country in Europe, 2012 vs. 2022

Source: Informatics Europe, 2023 | Chart: 2024 AI Index report

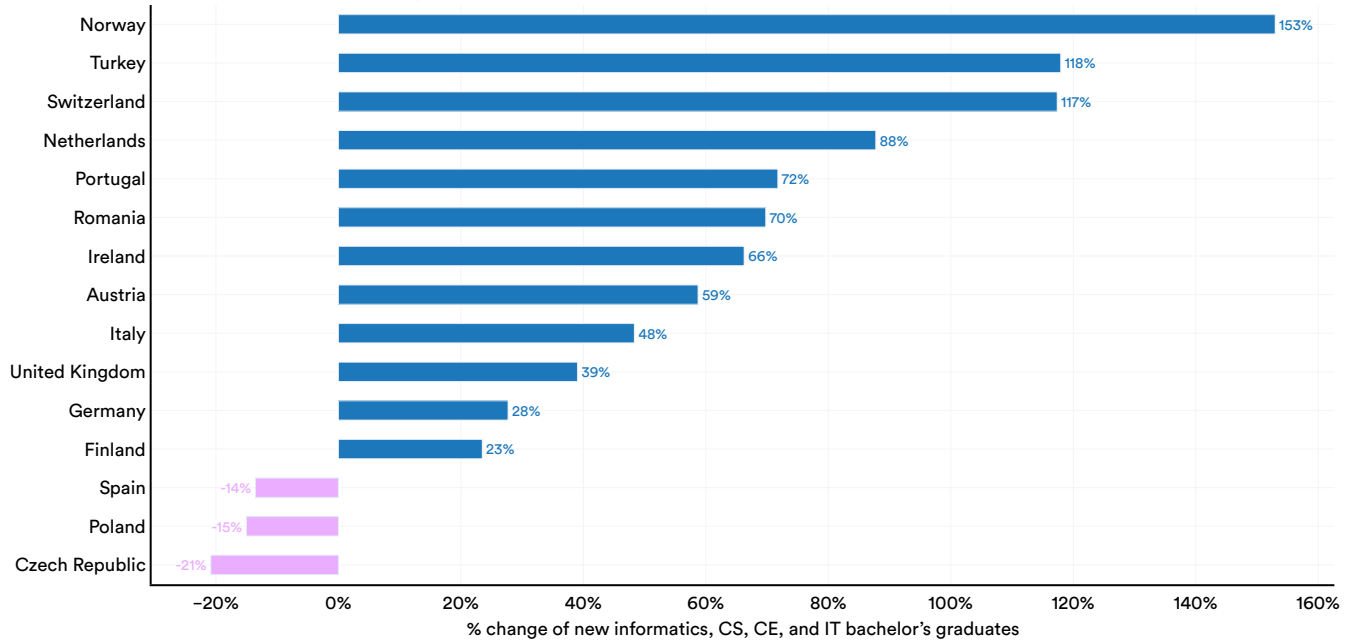


Figure 6.118

Finland (53.4), Norway (42.6), and the Netherlands (38.6) lead in the number of new bachelor's graduates in informatics CS, CE, and IT per 100,000 inhabitants (Figure 6.119). On a per capita basis, most sampled European countries have seen increases in the total number of informatics, CS, CE, and IT bachelor's graduates (Figure 6.120).