of extra — often intrusive — tests needed to identify a tumour's origins, says Mahmood. The predictions were restricted to 12 common sources of cancer, including the lungs, ovaries, breasts and stomach. Some other forms of cancer, including those originating in the prostate and kidneys, could not be identified, because they don't typically spread to fluid deposits in the abdomen and lungs, says Li.

When tested on some 500 images, the model was better than human pathologists at predicting a tumour's origin. This improvement was statistically significant.

The researchers also retrospectively assessed a subset of 391 study participants some four years after they had had cancer treatment. They found that those who had received treatment for the type of cancer that the model predicted were more likely to have survived, and lived longer, than were participants for whom the prediction did not match. "This is a pretty convincing argument" for using the AI model in a clinical setting, says Mahmood.

Mahmood has previously used AI to predict the origin of cancers from tissue samples (M. Y. Lu *et al. Nature* **594**, 106–110; 2021), and other teams have used genomic data. Combining the three data sources — cells, tissue and genomics — could further improve outcomes for people with metastatic cancers of unknown origins, he says.

NEW BENCHMARKS NEEDED TO KEEP PACE WITH AI'S ADVANCE

Stanford University's 2024 AI Index charts the meteoric rise of artificial-intelligence tools.

By Nicola Jones

Artificial intelligence (AI) systems, such as the chatbot ChatGPT, have become so advanced that they now very nearly match or exceed human performance in tasks including reading comprehension, image classification and competition-level mathematics, according to a report (see 'Speedy advances'). Rapid progress in the development of these systems also means that many common benchmarks and tests for assessing them are quickly becoming obsolete.

These are just a few of the headline findings from the Artificial Intelligence Index Report 2024, which was published on 15 April by the Institute for Human-Centered Artificial Intelligence at Stanford University in California (see go.nature.com/44ihnhx). The report charts the meteoric progress in machine-learning systems over the past decade.

In particular, the report says, new ways of assessing AI — for example, evaluating their performance on complex tasks, such as reasoning — are becoming more and more necessary. "A decade ago, benchmarks would serve the community for five to ten years", whereas now they often become irrelevant in just a few years, says Nestor Maslej, a social scientist at Stanford and editor-in-chief of the AI Index. "The pace of gain has been startlingly rapid."

Stanford's annual AI Index, first published in 2017, is compiled by a group of academic and industry specialists to assess the field's technical capabilities, costs, ethics and more — with an eye to informing researchers, policymakers and the public. This year's report, which is more than 400 pages long and was copy-edited and tightened with the aid of AI tools, notes that AI-related regulation in the United States is sharply rising. But the lack of standardized assessments for responsible use of AI makes it difficult to compare systems in terms of the risks that they pose.

The rise in the use of AI in science is also highlighted in this year's edition: for the first time, it dedicates an entire chapter to scientific applications, highlighting projects including Graph Networks for Materials Exploration (GNoME), a project from Google DeepMind that aims to help chemists discover materials, and GraphCast, another DeepMind tool, which does rapid weather forecasting.

Growing up

The current AI boom — built on neural networks and machine-learning algorithms — dates back to the early 2010s. The field has since rapidly expanded. For example, the number of AI coding projects on GitHub, a common platform for sharing code, increased from about 800 in 2011 to 1.8 million last year. And journal publications about AI roughly tripled over this period, the report says.

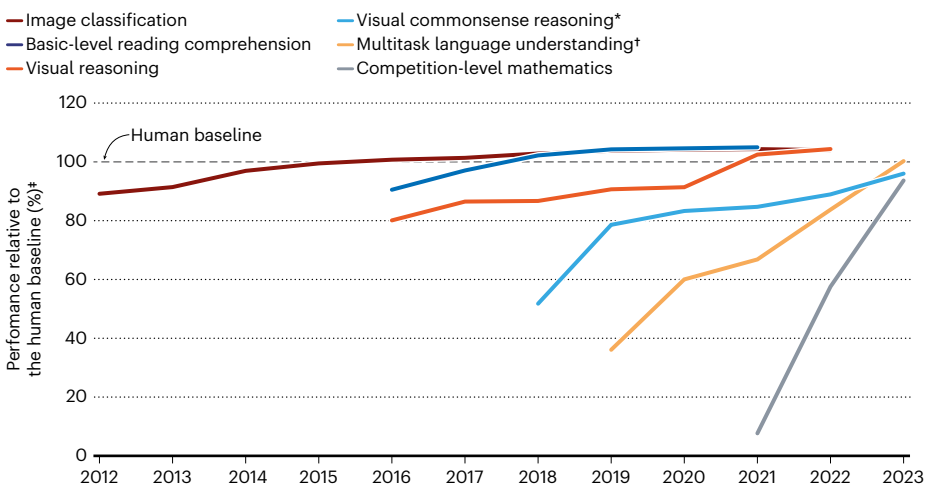
Much of the cutting-edge work on AI is being done in industry: that sector produced 51 notable machine-learning systems last year, whereas academic researchers contributed 15. "Academic work is shifting to analysing the models coming out of companies — doing a deeper dive into their weaknesses," says Raymond Mooney, director of the AI Lab at the University of Texas at Austin, who wasn't involved in the report.

That includes developing tougher tests to assess the visual, mathematical and even moral-reasoning capabilities of large language models (LLMs), which power chatbots. One of the latest tests is the Graduate-Level Google-Proof Q&A Benchmark (GPQA), developed last year by a team including machine-learning researcher David Rein at New York University (D. Rein *et al.* Preprint at arXiv https://doi.org/mr2k; 2023).

The GPQA, consisting of more than 400 multiple-choice questions, is tough: PhD-level scholars could correctly answer

SPEEDY ADVANCES

In the past several years, some AI systems have surpassed human performance on certain benchmark tests, and others have made rapid progress.



*Requires an AI system to answer questions about an image and provide a rationale for why its answers are true. †Tests an AI model's knowledge and problem-solving ability with regard to 57 subjects, including broader topics such as mathematics and history, and narrower areas such as law and ethics. ‡Data indicate the best performance of an AI model that year.

SOURCE: ARTIFICIAL INTELLIGENCE INDEX REPORT 2024.

questions in their field 65% of the time. The same scholars, when attempting to answer questions outside their field, scored only 34%, despite having access to the Internet during the test (randomly selecting answers would yield a score of 25%). As of last year, AI systems scored about 30–40%. This year, Rein says, Claude 3 — the latest chatbot released by AI company Anthropic, based in San Francisco, California — scored about 60%. "The rate of progress is pretty shocking to a lot of people, me included," Rein adds. "It's quite difficult to make a benchmark that survives for more than a few years."

Cost of business

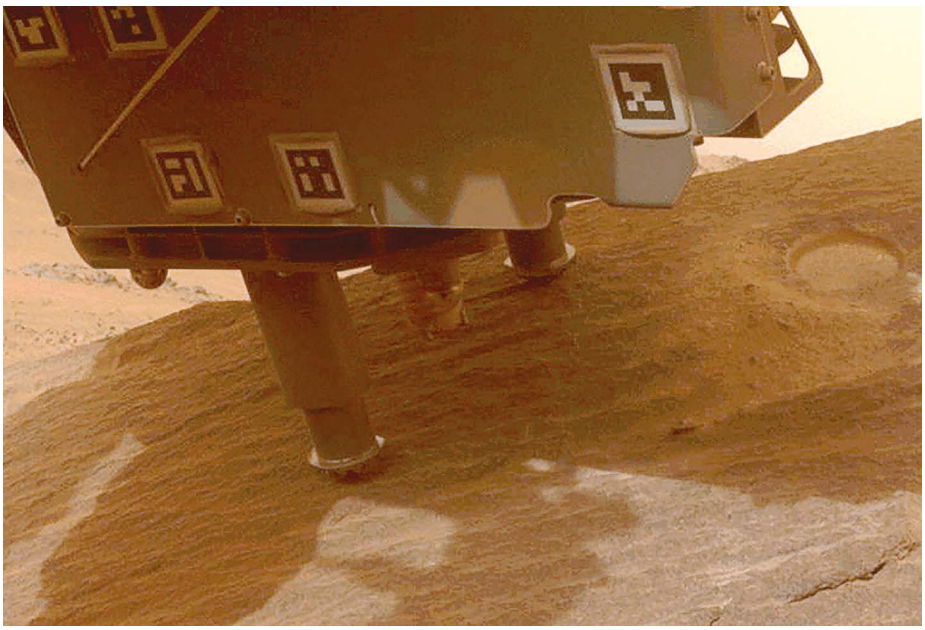
As performance is skyrocketing, so are costs. GPT-4 — the LLM that powers ChatGPT and that was released in March 2023 by San Francisco-based firm OpenAI — reportedly cost US$78 million to train. Google's chatbot Gemini Ultra, launched in December, cost $191 million. Many people are concerned about the energy use of these systems, as well as the amount of water needed to cool the data centres that help to run them (P. Li *et al*. Preprint at arXiv https://doi.org/mr2m; 2023). "These systems are impressive, but they're also very inefficient," Maslej says.

AI models' costs and energy use are high in large part because one of the main ways to make systems better is to make them bigger. This means training them on ever-larger stocks of text and images. The AI Index notes that some researchers now worry about running out of training data. Last year, according to the report, the non-profit research institute Epoch projected that supplies of high-quality language data could be exhausted as soon as this year. (However, the institute's most recent analysis suggests that 2028 is a better estimate.)

Ethical concerns about AI are also mounting. "People are way more nervous about AI than ever before, both in the United States and across the globe," says Maslej, who sees signs of a growing international divide. "There are now some countries very excited about AI, and others that are very pessimistic."

In the United States, the report notes a steep rise in regulatory interest. In 2016, there was just one US regulation that mentioned AI; last year, there were 25. "After 2022, there's a massive spike in the number of AI-related bills that have been proposed" by policymakers, Maslej says.

Regulatory action is increasingly focused on promoting responsible AI use. Although benchmarks are emerging that can score metrics such as an AI tool's truthfulness, bias and even likeability, not everyone is using the same models, Maslej says, which makes cross-comparisons hard. "This is a really important topic," he says. "We need to bring the community together on this."



NASA/JPL-CALTECH

NASA's Perseverance rover uses its robotic arm to drill into a Martian rock.

NASA SEEKS FRESH IDEAS FOR BRINGING MARS ROCKS TO EARTH

The agency's head says the current schedule for delivering samples to Earth is 'unacceptable'.

By Sumeet Kulkarni

NASA announced on 15 April that it is abandoning its long-standing plan for ferrying rock and soil samples from Mars to Earth. Instead, the agency will seek proposals for quicker and cheaper ways to deliver the samples to Earth.

An independent review board concluded last year that NASA's Mars sample return mission could cost as much as US$11 billion, more than the cost of launching the James Webb Space Telescope. In a report released on 15 April, a separate NASA review team concluded that even if the agency spent that much money, the samples would not reach Earth until 2040. NASA had originally sought to drop the samples on Earth in the early 2030s.

The $11-billion price tag is "too expensive", said NASA administrator Bill Nelson at a briefing, and "not returning the samples until 2040 is unacceptable". Nelson said the agency would bring "more than 30" of the 43 planned samples to Earth.

Scaling back

NASA's Perseverance rover has already collected more than 20 rock samples from Jezero Crater, where the rover landed in 2021. Scientists think that the crater was once filled with a lake of water, and samples from the crater and its surroundings could provide a window into the planet's history and, perhaps, evidence of past life on the red planet.

In the agency's original vision, a NASA spacecraft would have flown to Mars carrying a two-part retrieval system: a 2.3-tonne lander — which would have been the heaviest vehicle ever to land there — and a rocket to fly the lander and samples into Martian orbit. There, they were to meet a spacecraft launched by the European Space Agency that would fly the samples to Earth.

Now, NASA plans to solicit proposals from companies as well as NASA centres for a streamlined system, perhaps using a lighter lander, said Nicky Fox, associate administrator for NASA's Science Mission Directorate, at the briefing. The revised mission will be chosen later this year. Fox did not respond directly to questions about when the samples will reach Earth under the new scheme.

NASA recommends spending $200 million of its planetary-science budget for 2025 on assessing alternative architectures for Mars sample return, Fox said. Dedicating any more money to the mission threatened to "cannibalize" other planetary-science missions, Nelson said.