THE EASY INTELLIGENCE TESTS THAT AI CHATBOTS FAIL

Scientists are racing to find new ways to clarify the differences between the capabilities of humans and of large language models. **By Celeste Biever**

The world's best artificial intelligence (AI) systems can pass tough exams, write convincingly human essays and chat so fluently that many find their output indistinguishable from people's. What can't they do? Solve simple visual logic puzzles.

In a test consisting of a series of brightly coloured blocks arranged on a screen, most people can spot the connecting patterns. But GPT-4, the most advanced version of the AI system behind the chatbot ChatGPT and the search engine Bing, gets barely one-third of the puzzles right in one category of patterns and as little as 3% correct in another, according to a report by researchers this May[1].

The team behind the logic puzzles aims to provide a better benchmark for testing the capabilities of AI systems — and to help address a conundrum about large language models (LLMs) such as GPT-4. Tested in one way, they breeze through what once were considered landmark feats of machine intelligence. Tested another way, they seem less impressive, exhibiting glaring blind spots and an inability to reason about abstract concepts.

"People in the field of AI are struggling with how to assess these systems," says Melanie Mitchell, a computer scientist at the Santa Fe Institute in New Mexico whose team created the logic puzzles.

In the past two to three years, LLMs have blown previous AI systems out of the water in terms of their ability across multiple tasks. They work simply by generating plausible next words when given an input text, based on the statistical correlations between words in billions of online sentences they are trained on. For chatbots built on LLMs, there is an extra element: human trainers have provided extensive feedback to tune how the bots respond.

What's striking is the breadth of capabilities that emerges from this autocomplete-like algorithm trained on vast stores of human language. Other AI systems might beat the LLMs at any one task, but they have to be trained on data relevant to a specific problem, and cannot generalize from one task to another.

Broadly speaking, two camps of researchers have opposing views about what is going on under the hood of LLMs, says Tomer Ullman, a cognitive scientist at Harvard University in Cambridge, Massachusetts. Some attribute the algorithms' achievements to glimmers of reasoning, or understanding, he says. Others (including himself and researchers such as Mitchell) are much more cautious.

"There's very good smart people on all sides of this debate," says Ullman. The reason for the split, he says, is a lack of conclusive evidence supporting either opinion. "There's no Geiger counter we can point at something and say 'beep beep beep — yes, intelligent'," Ullman adds.

Tests such as the logic puzzles that reveal differences between the capabilities of people and AI systems are a step in the right direction, say researchers from both sides of the discussion. Such benchmarks could also help to show what is missing in today's machine-learning systems, and untangle the ingredients of human intelligence, says Brenden Lake, a cognitive computational scientist at New York University.

Research on how best to test LLMs and what those tests show also has a practical point. If LLMs are going to be applied in real-world domains — from medicine to law — it's important to understand the limits of their capabilities, Mitchell says. "We have to understand what they can do and where they fail, so that we can know how to use them in a safe manner."

Is the Turing test dead?

The most famous test of machine intelligence has long been the Turing test, proposed by the British mathematician and computing luminary Alan Turing in 1950, when computers were still in their infancy. Turing suggested an assessment that he called the imitation game[2]. This was a scenario in which human judges hold short, text-based conversations with a hidden computer and an unseen person. Could the judge reliably detect which was the computer? That was a question equivalent to 'Can machines think?', Turing suggested.

Turing did not specify many details about the scenario, notes Mitchell, so there is no exact rubric to follow. "It was not meant as a literal test that you would actually run on the machine — it was more like a thought experiment," says François Chollet, a software engineer at Google who is based in Seattle, Washington.

But the idea of leveraging language to detect whether a machine is capable of thought endured. For several decades, the businessman and philanthropist Hugh Loebner funded an annual Turing test event known as the Loebner Prize. Human judges engaged in text-based dialogues with both machines and people, and tried to guess which was which. But these annual gatherings stopped after 2019, because Loebner had died and the money to do it ran out, says computer scientist Rob Wortham. He is co-director of the UK Society for the Study of Artificial Intelligence and Simulation of Behaviour, which hosted the competition on Loebner's behalf, starting
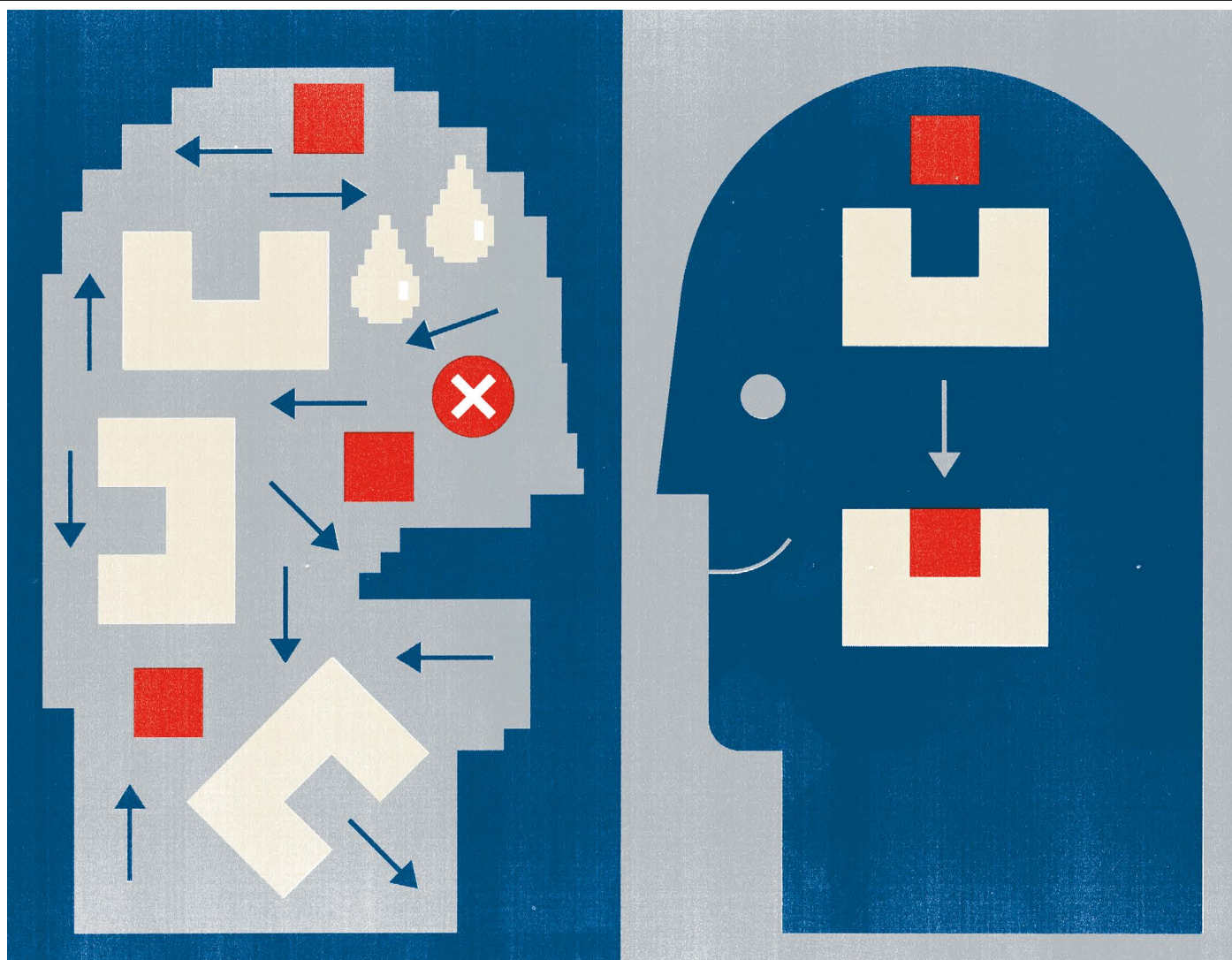
"WE HAVE TO UNDERSTAND WHAT THEY CAN DO AND WHERE THEY FAIL."

in 2014. He says that LLMs would now stand a good chance of fooling humans in such a contest; it's a coincidence that the events ended shortly before LLMs really took off.

Other researchers agree that GPT-4 and other LLMs would probably now pass the popular conception of the Turing test, in that they can fool a lot of people, at least for short conversations. In May, researchers at the company AI21 Labs in Tel Aviv, Israel, reported that more than 1.5 million people had played their online game based on the Turing test. Players were assigned to chat

for two minutes, either to another player or to an LLM-powered bot that the researchers had prompted to behave like a person. The players correctly identified bots just 60% of the time, which the researchers note is not much better than chance[3].

It's the kind of game that researchers familiar with LLMs could probably still win, however. Chollet says he'd find it easy to detect an LLM — by taking advantage of known weaknesses of the systems. "If you put me in a situation where you asked me, 'Am I chatting to an LLM right now?' I would definitely be able to tell you," says Chollet.

The key, he says, is to take the LLM outside of its comfort zone. He suggests presenting it with scenarios that are variations on ones the LLM will have seen a lot in its training data. In many cases, the LLM answers by spitting out words that are most likely to be associated with the original question in its training data, rather than by giving the correct answer to the new scenario.

Chollet and others, however, are sceptical about using a test centred around deceit as a goal for computer science. "It's all about trying to deceive the jury," says Chollet. The test

incentivizes chatbot developers to get an AI to perform tricks, instead of developing useful or interesting capabilities.

The perils of benchmarking

Rather than the Turing test, researchers instead typically assess AI systems using benchmarks intended to evaluate performance on specific capabilities, such as language ability, common-sense reasoning and mathematical capacity. Increasingly, teams are also turning to academic and professional examinations designed for people.

When GPT-4 was released in March this year, the firm behind it — OpenAI in San Francisco, California — tested its performance on a series of benchmarks designed for machines, including reading comprehension, mathematics and coding. GPT-4 aced most of them, OpenAI reported[4]. The company also set GPT-4 around 30 exams, including: various subject-specific tests designed for US high-school students, known as Advanced Placement; an exam to assess the current state of US physicians' clinical knowledge; and a standard test used in the selection process for US graduate studies, called the GRE. In the Uniform Bar Exam,

which forms part of the qualification process for lawyers in many US states, GPT-4 attained a score that would place it in the top 10% of people, OpenAI reported.

"A lot of these language models can do really well on these benchmarks," says Mitchell. "But often, the conclusion is not that they have surpassed humans in these general capacities, but that the benchmarks are limited." One challenge that researchers mention is that the models are trained on so much text that they could already have seen similar questions in their training data, and so might, in effect, be looking up the answer. This issue is known as contamination.

OpenAI says it checked for this by looking for similar strings of words in the questions and training data. When it tested the LLMs before and after removing the similar strings, there was little difference in performance, suggesting that successes couldn't be attributed largely to contamination. However, some researchers have questioned whether this test is stringent enough.

Sam Bowman, a language-technology scientist at New York University, who also works for the AI firm Anthropic in San Francisco, cautions against writing off GPT-4's abilities by dismissing its exam scores as merely the result of memorization. Contamination "complicates the claims a little bit, but I don't think it really changes the big picture that much", he says.

Researchers also note that LLMs' success on exam questions can be brittle and might not translate into the robust capability needed to get examples right in the real world. It's possible to change the exam questions slightly and get them to fail, says Mitchell. She took a question from an exam given to master's students in business administration that ChatGPT had passed, for instance, and rephrased it slightly. A person who could answer this question would be able to answer the rephrased version. But ChatGPT flunked it.

And there is a deeper problem in interpreting what the benchmarks mean. For a person, high scores across these exams would reliably indicate general intelligence — a fuzzy concept, but, according to one definition, one that refers to the ability to perform well across a range of tasks and adapt to different contexts. That is, someone who could do well at the exams can generally be assumed to do well at other cognitive tests and to have grasped certain abstract concepts. But that is not at all the case for LLMs, says Mitchell; these work in a very different way from people. "Extrapolating in the way that we extrapolate for humans won't always work for AI systems," she says.

This might be because LLMs learn only from language; without being embodied in the physical world, they do not experience language's connection to objects, properties and feelings, as a person does. "It's clear that they're not understanding words in the same way that people do," Lake says. In his opinion, LLMs currently demonstrate "that you can have very fluent language without genuine understanding".

On the flip side, LLMs also have capabilities that people don't — such as the ability to know the connections between almost every word humans have ever written. This might allow the models to solve problems by relying on quirks of language or other indicators, without necessarily generalizing to wider performance, says Mitchell.

Nick Ryder, a researcher at OpenAI, agrees that performance on one test might not generalize in the way it does for a person who gets the

GPT-4 CERTAINLY DOES NOT THINK LIKE A PERSON."

same score. "I don't think that one should look at an evaluation of a human and a large language model and derive any amount of equivalence," he says. The OpenAI scores are "not meant to be a statement of human-like capability or human-like reasoning. It's meant to be a statement of how the model performs on that task."

Researchers have also probed LLMs more broadly than through conventional machine benchmarks and human exams. In March, Sébastien Bubeck at Microsoft Research in Redmond, Washington, and his colleagues created waves with a preprint[5] entitled 'Sparks of Artificial General Intelligence: Early experiments with GPT-4'. Using an early version of GPT-4, they documented a range of surprising capabilities — many of which were not directly or obviously connected to language. A notable feat was that it could pass tests used by psychologists to assess theory of mind, a core human ability that allows people to predict and reason about the mental states of others. "Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system," they wrote.

Still, as Bubeck clarifies to *Nature*, "GPT-4 certainly does not think like a person, and for any capability that it displays, it achieves it in its own way".

Although provocative, the report does not probe the LLM's capabilities in a systematic way, says Mitchell. "It's more like anthropology," she says. Ullman says that to be convinced that a machine has theory of mind, he would need to see evidence of an underlying cognitive process corresponding to human-like theory of mind, and not just that the machine can output the same answers as a person.

To work out LLMs' strengths and weaknesses, more extensive and stringent audits are needed, say AI researchers. The colourful logic puzzles might be one candidate.

Fresh puzzles

In 2019, before LLMs exploded onto the scene, Chollet posted online a new kind of logic test for AI systems that he had created, called the Abstraction and Reasoning Corpus (ARC)[6]. Solvers look at several visual demonstrations of a grid of squares changing to another pattern, and show they have grasped the underlying rule for the change by indicating how the next grid would transform. "It is supposed to test for your ability to adapt to things you have not seen before," says Chollet, who argues that this is the essence of intelligence.

ARC captures a "hallmark of human intelligence", says Lake: the ability to make abstractions from everyday knowledge, and apply those to previously unseen problems.

Chollet organized an ARC competition for bots in 2020, before LLMs had gained much traction. The winning bot was an AI system that was specifically trained to solve ARC-like tasks but, unlike LLMs, had no general capabilities; it got only 21% of the problems right. People, by contrast, solve ARC problems correctly 80% of the time[7]. Several teams of researchers have now used the ARC to test the capabilities of LLMs; none has come close to human performance.

Mitchell and her colleagues made a set of fresh puzzles — known as ConceptARC — that were inspired by ARC, but differed in two key ways[1]. The ConceptARC tests are easier: Mitchell's team wanted to ensure the benchmark would not miss progress in machines' capabilities, even if small. The other difference was that the team chose specific concepts to test and then created a series of puzzles for each concept that are variations on a theme.

For example, to test the concept of sameness, one puzzle requires the solver to keep objects in the pattern that have the same shapes; another to keep objects that are aligned along the same axis. The goal of this was to reduce the chances that an AI system could pass the test without grasping the concepts (see 'An abstract-thinking test that defeats machines').

What poor performance means

The researchers fed the ConceptARC tasks to GPT-4 and to 400 people enlisted online. The humans scored, on average, 91% on all concept groups (and 97% on one); GPT-4 got 33% on one group and less than 30% on all the rest.

"We showed that the machines are still not

# AN ABSTRACT-THINKING TEST THAT DEFEATS MACHINES

Artificial-intelligence systems have so far been unable to achieve human-level performance on the ConceptARC test. This logic puzzle asks solvers to show how grid patterns will change after the solver has seen multiple demonstrations of an underlying abstract concept. Here are two sample tasks based on the concept of 'sameness' — between shapes in Task A and between orientations in Task B. See go.nature.com/43v6fzk for the answers.
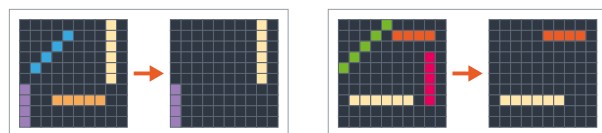
TASK A

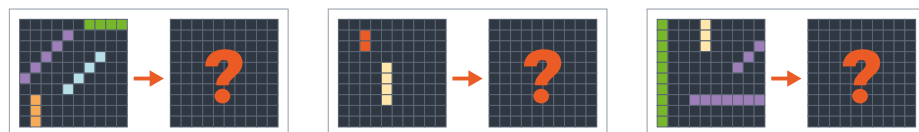Demonstrations:
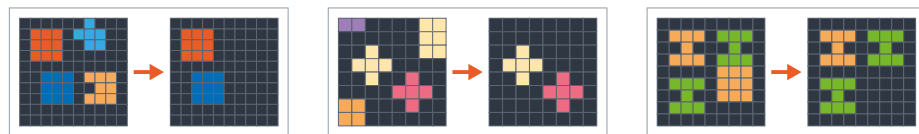
Test:
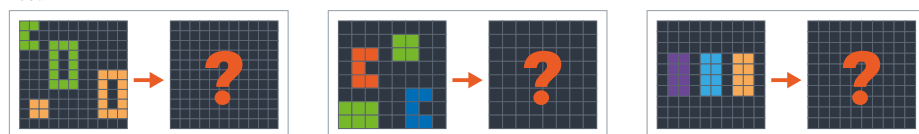
TASK B

Demonstrations:

Test:

able to get anywhere near the level of humans," says Mitchell. "It was surprising that it could solve some of the problems, because it had never been trained on them," she adds.

The team also tested the leading bots from Chollet's contest, which were not general-ability systems like LLMs, but were designed to solve visual puzzles such as ARC. Overall, they did better than GPT-4, but performed worse than people, with the best scoring 77% in one category but less than 60% in most[1].

Bowman, however, says GPT-4's struggles with ConceptARC don't prove that it lacks underlying capabilities in abstract reasoning. He says ConceptARC is skewed against GPT-4 — among other things, because it is a visual test. "Even if you suppose that these models are very good at this kind of reasoning, I don't think you'd really expect this experiment to have worked," he says.

Limitations to the way the test is done probably made it harder for GPT-4. The publicly available version of the LLM can accept only text as an input, so the researchers gave GPT-4 arrays of numbers that represented the images. (A blank pixel might be 0, and a colourful square a number, for instance.) By contrast, the human participants simply saw the images. "We are comparing a language-only system with humans, who have a highly developed visual system," says Mitchell. "So it might not be a totally fair comparison."

OpenAI has created a 'multimodal' version of GPT-4 that can accept images as input. Mitchell and her team are waiting for that to become publicly available so they can test ConceptARC on it, although she doesn't think the multimodal GPT-4 will do much better. "I don't think these systems have the same kind of abstract concepts and reasoning abilities that people have," she says.

Sam Acquaviva, a computational cognitive scientist at the Massachusetts Institute of Technology in Cambridge, agrees. "I would be shocked," he says. He notes that another team of researchers has tested GPT-4 on a benchmark called 1D-ARC, in which patterns are confined to a single row rather than being in a grid[8]. That should erase some of the unfairness, he says. Even though GPT-4's performance improved, it was not enough of a gain to suggest that the LLM was reliably grasping the underlying rule and reasoning about it, says Acquaviva.

Argument for reasoning

Bowman points to other experiments that, taken together, suggest to him that LLMs have acquired at least a rudimentary ability to reason about abstract concepts. In one example, computer scientist Kenneth Li at Harvard University and his colleagues used a digital version of the board game Othello, in which two players compete by placing black and white discs on a 8 × 8 grid. Their aim was to examine whether LLMs rely on the memorized surface statistics of language to generate text, or if they might be building internal representations of the world, as people do.

When they trained an LLM by feeding it lists of moves made by players, it became very good at spitting out accurate suggestions for next legal moves. The researchers argued that they had evidence that the LLM was keeping track of the state of the board — and that it was using this representation to suggest moves, rather than just coming up with textual suggestions[9].

Bowman acknowledges that the reasoning capabilities of LLMs in general are "spotty" and more limited than in people — but he says that they are there, and seem to improve with model size, which indicates to him that future LLMs will be even better. "These systems are definitely not anywhere near as reliable or as general as we want, and there probably are some particular abstract reasoning skills that they're still entirely failing at," he says. "But I think the basic capacity is there."

One thing Bowman, Mitchell and others agree on is that the best way to test LLMs for abstract reasoning abilities and other signs of intelligence remains an open, unsolved problem. Michael Frank, a cognitive scientist at Stanford University in Palo Alto, California, does not expect a single, catch-all test to emerge as a successor to the Turing test. "There's no Rubicon, no one line," he says. Rather, he thinks that researchers need lots of tests to quantify the strengths and weaknesses of various systems. "These agents are great, but they break in many, many ways and probing them systematically is absolutely critical," he says.

Wortham offers advice to anyone trying to understand AI systems — avoid what he calls the curse of anthropomorphization. "We anthropomorphize anything which appears to demonstrate intelligence," he says.

"It is a curse, because we can't think of things which display goal-oriented behaviour in any way other than using human models," he says. "And we're imagining that the reason it's doing that is because it's thinking like us, under the covers."

Celeste Biever is *Nature*'s chief news and features editor, based in London.

1. Moskvichev, A., Odouard, V. V. & Mitchell, M. Preprint at https://arxiv.org/abs/2305.07141 (2023).
2. Turing, A. M. *Mind* **LIX**, 433–460 (1950).
3. Jannai, D., Meron, A., Lenz, B., Levine, Y. & Shoham, Y. Preprint at https://arxiv.org/abs/2305.20010 (2023).
4. OpenAI. Preprint at https://arxiv.org/abs/2303.08774 (2023).
5. Bubeck, S. *et al.* Preprint at https://arxiv.org/abs/2303.12712 (2023).
6. Chollet, F. Preprint at https://arxiv.org/abs/1911.01547 (2019).
7. Johnson, A., Vong, W. K., Lake, B. M. & Gureckis, T. M. Preprint at https://arxiv.org/abs/2103.05823 (2021).
8. Xu, Y., Li, W., Vaezipoor, P., Sanner. S. & Khalil, E. B. Preprint at https://arxiv.org/abs/2305.18354 (2023).
9. Li, K. *et al.* Proc. Eleventh Int. Conf. Learn. Represent. https://openreview.net/forum?id=DeG07_TcZvT (2023).