The international journal of science / 27 July 2023

nature

Understanding ChatGPT is a bold new challenge for science

Despite their wide use, large language models are still mysterious. Revealing their true nature is urgent and important.

“I propose to consider the question, ‘Can machines think?” So began a seminal 1950 paper by British computing and mathematics luminary Alan Turing (A. M. Turing *Mind* **LIX**, 433–460; 1950).

But as an alternative to the thorny task of defining what it means to think, Turing proposed a scenario that he called the “imitation game”. A person, called the interrogator, has text-based conversations with other people and a computer. Turing wondered whether the interrogator could reliably detect the computer – and implied that if they could not, then the computer could be presumed to be thinking. The game captured the public’s imagination and became known as the Turing test.

Although an enduring idea, the test has largely been considered too vague – and too focused on deception, rather than genuinely intelligent behaviour – to be a serious research tool or goal for artificial intelligence (AI). But the question of what part language can play in evaluating and creating intelligence is more relevant today than ever. That’s thanks to the explosion in the capabilities of AI systems known as large language models (LLMs), which are behind the ChatGPT chatbot, made by the firm OpenAI in San Francisco, California, and other advanced bots, such as Microsoft’s Bing Chat and Google’s Bard. As the name ‘large language model’ suggests, these tools are based purely on language.

With an eerily human, sometimes delightful knack for conversation – as well as a litany of other capabilities, including essay and poem writing, coding, passing tough exams and text summarization – these bots have triggered both excitement and fear about AI and what its rise means for humanity. But underlying these impressive achievements is a burning question: how do LLMs work? As with other neural networks, many of the behaviours of LLMs emerge from a training process, rather than being specified by programmers. As a result, in many cases the precise reasons why LLMs behave the way they do, as well as the mechanisms that underpin their behaviour, are not known – even to their own creators.

As *Nature* reports in a Feature, scientists are piecing together both LLMs’ true capabilities and the underlying mechanisms that drive them. Michael Frank, a cognitive scientist at Stanford University in California, describes the task as similar to investigating an “alien intelligence”.

Revealing this is both urgent and important, as researchers have pointed out (S. Bubeck *et al.* Preprint at https://arxiv.org/abs/2303.12712; 2023). For LLMs to solve problems and increase productivity in fields such as medicine and law, people need to better understand both the successes and failures of these tools. This will require new tests that offer a more systematic assessment than those that exist today.

> **“For LLMs to solve problems, people need to better understand the successes and failures of these tools.”**

Breezing through exams

LLMs ingest enormous reams of text, which they use to learn to predict the next word in a sentence or conversation. The models adjust their outputs through trial and error, and these can be further refined by feedback from human trainers. This seemingly simple process can have powerful results. Unlike previous AI systems, which were specialized to perform one task or have one capability, LLMs breeze through exams and questions with a breadth that would have seemed unthinkable for a single system just a few years ago.

But as researchers are increasingly documenting, LLMs’ capabilities can be brittle. Although GPT-4, the most advanced version of the LLM behind ChatGPT, has aced some academic and professional exam questions, even small perturbations to the way a question is phrased can throw the models off. This lack of robustness signals a lack of reliability in the real world.

Scientists are now debating what is going on under the hood of LLMs, given this mixed performance. On one side are researchers who see glimmers of reasoning and understanding when the models succeed at some tests. On the other are those who see their unreliability as a sign that the model is not as smart as it seems.

AI approvals

More systematic tests of LLMs’ capabilities would help to settle the debate. These would provide a more robust understanding of the models’ strengths and weaknesses. Similar to the processes that medicines go through to attain approval as treatments and to uncover possible side effects, assessments of AI systems could allow them to be deemed safe for certain applications and could enable the ways they might fail to be declared to users.

In May, a team of researchers led by Melanie Mitchell, a computer scientist at the Santa Fe Institute in New Mexico, reported the creation of ConceptARC (A. Moskvichev *et al.* Preprint at https://arxiv.org/abs/2305.07141; 2023): a series of visual puzzles to test AI systems’ ability to reason about abstract concepts. Crucially, the puzzles systematically test whether a system has truly grasped 16 underlying concepts by testing each one in 10 ways (spoiler alert: GPT-4 performs poorly). But ConceptARC addresses just one facet of reasoning and generalization; more tests are needed.

Confidence in a medicine doesn’t just come from observed safety and efficacy in clinical trials, however. Understanding the mechanism that causes its behaviour is also important, allowing researchers to predict how it will function in different contexts. For similar reasons, unravelling the mechanisms that give rise to LLMs' behaviours – which can be thought of as the underlying 'neuroscience' of the models – is also necessary.

Researchers want to understand the inner workings of LLMs, but they have a long road to travel. Another hurdle is a lack of transparency – for example, in revealing what data models were trained on – from the firms that build LLMs. However, scrutiny of AI companies from regulatory bodies is increasing, and could force more such data to be disclosed in future.

Seventy-three years after Turing first proposed the imitation game, it's hard to imagine a more important challenge for the field of AI than understanding the strengths and weaknesses of LLMs, and the mechanisms that drive them.

The UK is turning its back on researcher mobility

Scientific strength does not come from severing long-standing relationships or turning away international talent.

Science has long transcended geographical borders. Collaborative relationships formed when scientists from different countries are free to travel and work together are at the heart of today's research teams. As the innovation specialist Charles Leadbeater wrote 15 years ago in a monograph called 'The Difference Dividend', innovation thrives when people are allowed to work freely across borders and cultures (go.nature.com/3oqq6xx).

Yet some research-intensive nations are now erecting barriers. China and the United States are dismantling two decades of research cooperation that has benefited both nations. Opportunities for collaboration are also diminishing elsewhere. The necessity of obtaining a visa means that it has never been easy for researchers in low- and middle-income countries (LMICs) to collaborate with colleagues in high-income countries. Earlier this month, seismologist Sujania Talavera-Soza plotted the likelihood of a passport holder needing a visa to study or work abroad against their country's gross domestic product, and concluded that people in the poorest countries almost always need visas (S. Talavera-Soza *Nature Geosci.* **16**, 550–551; 2023). These are getting increasingly difficult to come by.

For some time, UK universities have been reporting that researchers whom they invite to attend conferences, give talks or receive awards struggle to get visas. In one instance, an event on science in LMICs had few speakers from these countries because most of those invited had their visa applications refused.

> "The UK government urgently needs to change its closed-door attitude towards scientists from low- and middle-income countries."

The Royal Society, the country's national science academy, has been investigating researchers' experiences of UK visa applications, and last week published its findings (see go.nature.com/3pv6kpw). The report reveals that, out of ten countries classed as leading science nations by Nature Index 2022 for which data are available, the United Kingdom is third, after Sweden and France, for rejecting business visa applications. In 2022, the ten countries from which the greatest proportion of people were refused a standard visitor visa – which allows individuals to come to the United Kingdom for up to six months for business or study – were all in Africa. More than 50% of applicants from these countries were turned down. This is shocking and unacceptable.

The Royal Society is recommending that the UK government create a special short-term researcher-mobility visa; it also says communication with visa applicants should be clearer. That should apply to all applicants, not just scientists. Everyone deserves to be treated fairly.

Going backwards

The society is also among many concerned that there is still no agreement on whether UK scientists will be allowed to participate fully in the European Union's Horizon Europe research programme. A decision on whether to join the €95.5-billion (US$106-billion) project is understood to have been postponed until after the summer. Scientists are aghast. The government's former chief scientific adviser, Patrick Vallance, said last week that joining needed to have happened "yesterday". There is a risk it won't happen at all.

The delay is down to disagreement over how much the United Kingdom should pay to join the project, and what should happen if its researchers end up winning more or less in grants than the nation pays in. The problem with a prolonged delay is that the Horizon project is time-limited – it started in 2021 and will run until the end of 2027. The United Kingdom has already lost two years, and researchers fear that if two becomes three, there might not be enough time left to run some projects. In case the Horizon bid fails, UK policymakers have a plan B that they are calling Pioneer: this would involve a new set of funding schemes for the United Kingdom and other nations. But Pioneer is untested and has few fans in the research world.

UK membership of EU science agreements goes back decades, and the disruption caused by Brexit has been damaging. UK negotiators need to ask themselves whether a prolonged argument is worth derailing an entire scheme from which the country, its researchers and the course of discovery, innovation and invention will all benefit.

Whether the UK government decides that the nation's future lies with Pioneer or Horizon, it urgently needs to change its closed-door attitude towards scientists from LMICs who apply to visit for academic purposes. UK leaders are fond of describing the country as a science superpower. Stopping researchers from African countries visiting, and allowing continued delays to prevent the nation from joining the world's largest international collaboration scheme, undermine that ambition.