ILLUSTRATION BY FABIO BUONOCORE

IN AI, IS BIGGER BETTER?

As generative artificial-intelligence models get larger, some scientists advocate for leaner, more energy-efficient systems. **By Anil Ananthaswamy**

Artificial-intelligence systems that can churn out fluent text, such as OpenAI's ChatGPT, are the newest darlings of the technology industry. But when faced with mathematical queries that require reasoning to answer, these large language models (LLMs) often stumble. Take, for instance, this algebra problem:

> A line parallel to $y = 4x + 6$ passes through (5, 10).
> What is the y-coordinate of the point where this line crosses the y-axis?

Although LLMs can sometimes answer these types of question correctly, they more often get them wrong. In one early test of its reasoning abilities, ChatGPT scored just 26% when faced with a sample of questions from the 'MATH' data set of secondary-school-level mathematical problems[1].

This is to be expected: given input text, an LLM simply generates new text in accordance with statistical regularities in the words, symbols and sentences that make up the model's training data. It would be startling if just learning language patterns could allow LLMs to mimic mathematical reasoning reliably.

But back in June 2022, an LLM called Minerva, created by Google, had already defied these expectations — to some extent. Minerva scored 50% on questions in the MATH data set[2], a result that shocked some researchers in artificial intelligence (AI).

"The chatter in the community is that this is really kind of astounding," says Sébastien Bubeck, a machine-learning specialist at Microsoft Research in Redmond, Washington.

Minerva had the advantage that it was trained on mathematics-related texts. But Google's study suggested another important reason the model did so well — its sheer size. It was around three times the size of ChatGPT.

The Minerva results hint at something that some researchers have long suspected: that training larger LLMs, and feeding them more data, could give them the ability, through pattern-recognition alone, to solve tasks that are supposed to require reasoning. If so, some AI researchers say that this 'bigger is better'

strategy might provide a path to powerful AI.

But there are reasons to doubt this thesis. LLMs still make glaring errors, and some scientists suggest that larger models are merely getting better at answering queries that fall within the scope of the correlations in their training data, rather than acquiring an ability to answer entirely new questions.

The debate is now playing out on the frontiers of AI. Commercial firms have seen better results with bigger AI models, so they are rolling out ever-larger LLMs — each costing millions of dollars to train and run (see 'The drive to bigger AI models'). But these models have major downsides. Besides concerns that their output cannot be trusted, and that they might exacerbate the spread of misinformation, they are expensive and suck up huge amounts of energy.

Critics argue that, ultimately, big LLMs will never be able to mimic or acquire skills that allow them to answer reasoning problems consistently. Instead, some scientists say, smaller, more energy-efficient AI is the way to make progress — inspired, in part, by the way the brain seems to learn and make connections.

Big, bigger, better?

LLMs such as ChatGPT and Minerva are giant networks of computing units (also called artificial neurons), arranged in layers. An LLM's size is measured in how many parameters it has — the adjustable values that describe the strength of the connections between neurons. Training such a network involves asking it to predict masked portions of a known sentence and tweaking these parameters so that the algorithm does a little better next time.

Do this repeatedly over billions of human-written sentences, and the neural network learns internal representations that model how humans write language. At this stage, an LLM is said to be pre-trained: its parameters capture the statistical structure of written language that it saw during training, including all the facts, biases and errors in the texts. It can then be 'fine-tuned' on specialized data.

To make Minerva, for instance, researchers started with Google's Pathways Language Model (PaLM), which has 540 billion parameters and was pre-trained on a data set of 780 billion tokens[3]. A token can be a word, digit or some unit of information; in PaLM's case, the tokens were gleaned from English and multilingual web documents, books and code. Minerva was the result of fine-tuning PaLM on tens of billions of tokens from scientific papers and mathematics-related web pages.

Minerva can answer prompts such as: what is the largest multiple of 30 that is less than 520? The LLM appears to be thinking through the steps, and yet all it is doing is turning the questions into a sequence of tokens, generating a statistically plausible next token, appending it to the original sequence, generating another token, and so on: a process called inference.

Google researchers fine-tuned three sizes of Minerva using underlying pre-trained PaLM models of 8 billion, 62 billion and 540 billion parameters. Minerva's performance improved with scale. On the overall MATH data set, the smallest model had 25% accuracy, the medium-sized model reached 43% and the largest breached the 50% mark.

> "THE BIGGER MODELS KEEP DOING BETTER AND BETTER."

The biggest model also used the least amount of fine-tuning data — it was fine-tuned on only 26 billion tokens, whereas the smallest model looked at 164 billion tokens. But the biggest model took a month to fine-tune, on specialized hardware that had eight times as much computing capacity as used for the smallest model, which was fine-tuned for only two weeks. Ideally, the biggest model would have been fine-tuned on more tokens, says Ethan Dyer at Google Research in Mountain View, California, who is a member of the Minerva team; this could have led to an even better performance. But the team felt that the computational expense wasn't feasible.

Scaling laws

That the biggest Minerva model did best was in line with studies that have revealed scaling laws — rules that govern how performance improves with model size. A study in 2020 showed that models did better when given one of three things: more parameters, more training data or more 'compute' (the number of computing operations executed during training)[4]. Performance scaled according to a power law, meaning that it improved as some power of, for example, the number of parameters.

However, researchers don't exactly know why. "The laws are purely empirical," says Irina Rish, a computer scientist at the University of Montreal and Mila — Quebec Artificial Intelligence Institute in Montreal, Canada.

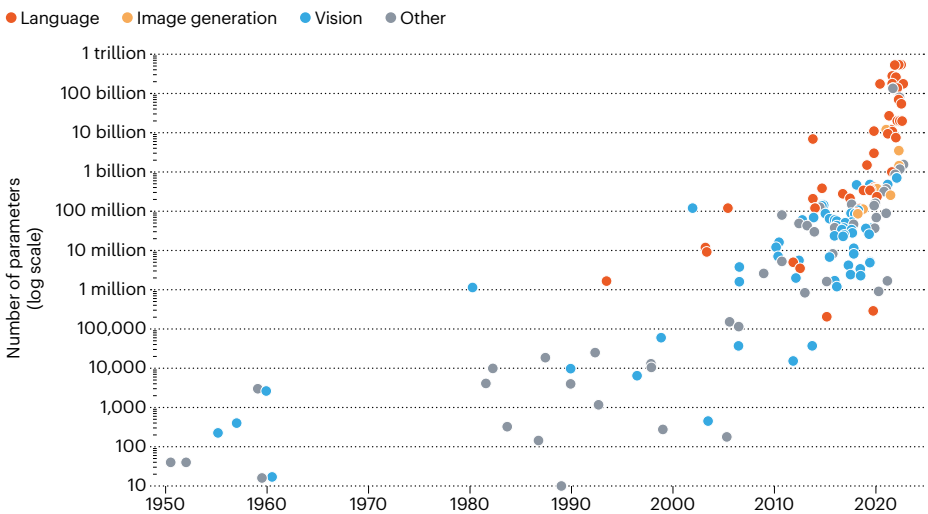
For the best results, the 2020 study suggested that as training data is doubled, model size should increase five times. Work last year slightly revised this. In March, the London-based AI firm DeepMind argued that it's best to scale up model size and training data together, and that smaller models trained on more data do better than bigger models trained on fewer data[5] (see 'Different routes to scale'). For example, DeepMind's Chinchilla model has 70 billion parameters, and was trained on 1.4 trillion tokens, whereas its 280-billion-parameter Gopher model, was trained on 300 billion tokens. Chinchilla outperforms Gopher on tasks designed to evaluate what the LLM has learnt.

Scientists at Meta Research built on this concept in February with their own small-parameter model called LLaMA, trained on up to 1.4 trillion tokens. The 13-billion-parameter version of LLaMA outperformed ChatGPT's forerunner GPT-3 (175 billion parameters), the researchers say, whereas the 65-billion-parameter version was competitive with Chinchilla and even PaLM (see go.nature.com/3kje2fj).

Last October, Ethan Caballero at McGill University in Montreal, together with Rish and others, reported finding more complex relationships between size and performance[6]. In some instances, multiple power laws can

THE DRIVE TO BIGGER AI MODELS

The scale of artificial-intelligence neural networks is growing exponentially, as measured by the models' parameters (roughly, the number of connections between their neurons)*.



*'Sparse' models, which have more than one trillion parameters but use only a fraction of them in each computation, are not shown.

ADAPTED FROM OUR WORLD IN DATA, AND FROM J. SEVILLA *ET AL.* PREPRINT AT ARXIV HTTPS://DOI.ORG/10.48550/ARXIV.2202.05924 (2022).

govern how performance scales with model size, the researchers say.

For instance, in one hypothetical scenario fitting a general equation that they found, performance improves first gradually and then more rapidly with a model's size, but then dips slightly as the number of parameters continues to rise, before increasing again. The characteristics of this complex relationship are dictated by the specifics of each model and how it is trained. Ultimately, the researchers hope to be able to predict this in advance as any particular LLM is scaled up.

A separate theoretical finding also supports the drive for bigger models — a 'law of robustness' for machine learning, introduced in 2021 by Bubeck and his colleagues[7]. A model is robust if its answers remain consistent despite small perturbations in its inputs. Some AIs are notoriously fragile. If trained to recognize images of dogs, for instance, they will misclassify a test image when it's modified by a small amount of noise that wouldn't fool a person.

The more robust the AI, the better it generalizes to unseen data. Bubeck and his colleagues have shown mathematically that increasing the number of parameters in a model increases robustness, and hence ability to generalize. The law proves that scaling up is necessary for generalization, but not that it's sufficient, says Bubeck. Nonetheless, it is being used to justify the move towards bigger models, he says. "Which I think is a reasonable thing."

Minerva also took advantage of a key innovation called chain-of-thought prompting. The user prefixes their question with text that includes a couple of examples of questions and solutions, including the reasoning — illustrating a typical chain of thought — that led to the answers. During inference, the LLM takes its cues from this context and provides a step-by-step answer that looks surprisingly like reasoning. This doesn't require updates to the model's parameters, and so doesn't involve the additional computing power that fine-tuning needs.

The ability to respond to chain-of-thought prompts shows up only in LLMs with more than about 100 billion parameters. Such discoveries have helped bigger models to improve in accordance with empirical scaling laws, says Blaise Agüera y Arcas at Google Research in Seattle, Washington. "The bigger models keep doing better and better."

Reasonable concerns

François Chollet, an AI researcher at Google in Mountain View, is among the sceptics who argue that no matter how big LLMs become, they will never get near to having the ability to reason (or mimic reasoning) well enough to solve new problems reliably. An LLM only appears to reason by using templates that it has encountered before, whether in the training data or in the prompt, he says. "It cannot, on the fly, make sense of something that it has not seen before."

The best that LLMs might be able to do is to slurp in so much training data that the statistical patterns of language alone allow them to respond to questions with answers that are very close to what they've already seen.

Agüera y Arcas, however, argues that LLMs do seem to have gained some surprising abilities that they weren't trained on specifically. In particular, he points to tests designed to show whether a person has what's called theory of mind — being able to theorize or gauge the mental states of others. Take this simple example. Alice puts her glasses away in a drawer. Then Bob, unbeknown to Alice, hides the glasses under a cushion. Where will Alice look for her glasses first? A child asked this question is being tested on whether they understand that Alice has her own beliefs that might not agree with what the child knows.

In his experiments with another of Google's LLMs called Language Model for Dialogue Applications (LaMDA), Agüera y Arcas found that LaMDA responded correctly in more extended conversations of this type. To him, this was suggestive of an LLM's ability to internally model the intentions of others. "These models that are doing nothing but predicting sequences develop an extraordinary range of capabilities, including theory of mind," says Agüera y Arcas. But he concedes that these models are error-prone, and he, too, is unsure whether scaling alone, although it seems necessary, is sufficient for reliable reasoning.

> "EVERY BIG TECH COMPANY WILL NOW ATTEMPT TO DEPLOY LLMS."

Even when LLMs get the answers right, however, there is no understanding involved, says Chollet. "When you try to probe it a little bit, it becomes immediately obvious that it's all empty. ChatGPT has no model of what it is talking about," he says. "You're watching a puppet show and believing the puppets are alive."

So far, LLMs still make absurd mistakes that humans never would, says Melanie Mitchell, who studies conceptual abstraction and analogy-making in AI systems at the Santa Fe Institute in New Mexico. This has contributed to the many concerns about the safety of unleashing LLMs without guardrails into society.

An issue with the debate over whether LLMs can ever tackle genuinely new, unseen problems is that we have no way of comprehensively testing for this ability, Mitchell adds. "The current benchmarks we have are not adequate," she says. "They're not probing things systematically. We don't really know yet how to do that."

The problems of scale

While the debate plays out, there are already pressing concerns over the trend towards larger language models. One is that the data sets, computing power and expense involved in training big LLMs restricts their development — and therefore their research direction — to companies with immense computing resources. OpenAI has not confirmed the costs of creating ChatGPT, but others have estimated on the basis of the compute involved that pre-training GPT-3 (a predecessor of ChatGPT) would have cost more than US$4 million. It's probably costing OpenAI millions of dollars each month to run ChatGPT, because of the number of queries that the free chatbot is now servicing. "We are already deep into this regime," says Bubeck. "There are just a few companies that have models bigger than 100 billion parameters."

Governments are starting to step in with support that might expand the playing field. In June last year, an international team of around 1,000 academic volunteers, with funding from the French government, a US AI company called Hugging Face and others, trained a model called BLOOM with about 175 billion parameters, using $7 million worth of computing time[8]. And in November, the US Department of Energy awarded supercomputing time to a project from Rish and her colleagues, to build large models to study their behaviour. "We hope to train a Chinchilla-like 70-billion-parameter model — not necessarily the largest, but rather the one whose performance scales more effectively," says Rish.

Regardless of who gets to build them, LLMs also raise concerns about electricity consumption. For example, Google reported that training PaLM took about 3.4 gigawatt-hours over about two months. That's the annual energy consumption of about 300 US households. Google trained PaLM at its Oklahoma data centre, which it said operated on 89% carbon-free energy, being powered mostly by wind and other renewable sources. But a survey of industry AI models has shown that the majority are trained using electricity grids that are still largely powered by fossil fuels[9].

Chollet's concern is that as multiple firms begin training and using bigger models, they could start to suck up more electricity. "Every big tech company will now attempt to deploy LLMs into their products, regardless of whether it's a good idea or not," he says.

Smarter and smaller?

For many scientists, then, there's a pressing need to reduce LLM's energy consumption — to make neural networks smaller and more efficient, as well as, perhaps, smarter. Besides

the energy costs of training LLMs (which, although substantial, are a one-off), the energy needed for inference – in which LLMs answer queries – can shoot up as the number of users increases. Big tech firms haven't commented on the usage costs of their models. Hugging Face, however, revealed that when its BLOOM model was deployed on the Google Cloud Platform for 18 days, in which time it answered 230,768 queries (many fewer than ChatGPT, which hit 100 million active users a month in February), it consumed, on average, 1,664 watts[10].

For comparison, our own brains are much more complicated and larger than any LLM, with 86 billion neurons and some 100 trillion synaptic connections. And yet, the human brain consumes somewhere between 20 and 50 watts of power, says Friedemann Zenke at the Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland.

So some researchers hope that mimicking aspects of the brain will help LLMs and other neural networks to become smaller, smarter and more efficient.

One source of the brain's overall intelligence and efficiency might be its recurrent, or feedback, connections. LLMs are, essentially, 'feedforward' networks. This means that information flows one way: from the input, through the layers of the LLM, to its output. The brain is wired differently. For example, in the human visual system, neurons connect regions of the brain that first receive visual information to areas further back. But there are also feedback connections that allow information transfer between neurons in the reverse direction. "There's maybe ten times as many feedback connections as there are feedforward connections in the [human] visual system," says Mitchell, but an LLM has no feedback connections.

Artificial neural networks that incorporate both feedforward and feedback connections are generically called recurrent neural networks (RNNs). Such networks (unlike feedforward LLMs) can discern patterns in data that change over time. That's "fundamental to how all natural intelligences experience the world and learn", says Kanaka Rajan, a computational neuroscientist at the Icahn School of Medicine at Mount Sinai in New York City. But RNNs come with challenges, says Rajan. For instance, they are hard and slow to train, making it difficult to scale them to the size of current LLMs.

Another reason brains are efficient is that biological neurons mostly remain quiet – they have only an occasional spike in activity. By contrast, the artificial neurons in most neural networks are modelled as being constantly on. Researchers are studying artificial neurons that spike (mimicking real ones), but it is difficult to adapt algorithms that train standard neural networks to networks that use spiking neurons. Still, research using small data sets (for example, 10,000 audio recordings used to train a network to recognize spoken digits) has shown that RNNs with spiking neurons outperform those with standard neurons, and, in theory, are three orders of magnitude more computationally efficient[11]. "Progress is rapid and impressive," says Sander Bohté, at Amsterdam's National Research Institute for Mathematics and Computer Science in the Netherlands (CWI), who works in this area.

As long as such spiking networks are only simulated in software, however, they cannot provide real efficiency gains (since the hardware simulating them still consumes power). Such computing elements will need to be built into hardware, on neuromorphic chips, to realize their benefits.

Energy-efficient LLMs

Meanwhile, researchers are experimenting with different ways to make existing LLMs more energy efficient, and smarter. In December 2021, DeepMind reported a system called RETRO, which combines an LLM with an external database. The LLM uses relevant text retrieved from this database during inference to help it make predictions. DeepMind's researchers showed that a 7.5-billion parameter LLM, coupled with a database of 2 trillion tokens, outperforms LLMs with 25 times more parameters[12]. The researchers wrote that this was a "more efficient approach than raw parameter scaling as we seek to build more powerful language models".

In the same month, scientists at Google Research reported another way to increase energy efficiency at scale. Their Generalist Language Model, or GLaM, has 1.2 trillion parameters[13]. But these parameters don't represent one giant neural network; internally, they are distributed between 64 smaller neural networks, alongside other layers. The LLM is trained such that during inference, it uses only two of its networks to complete a task; overall, the LLM uses only about 8% of its trillion-plus total parameters for inference, per token. According to Google, GLaM used the same amount of computing resources as were needed to train GPT-3, but consumed only about one-third of the power, because of improvements in training software and hardware. During inference, GLaM used half the computing resources that GPT-3 needed. And it outperformed GPT-3 when trained on the same amount of data.

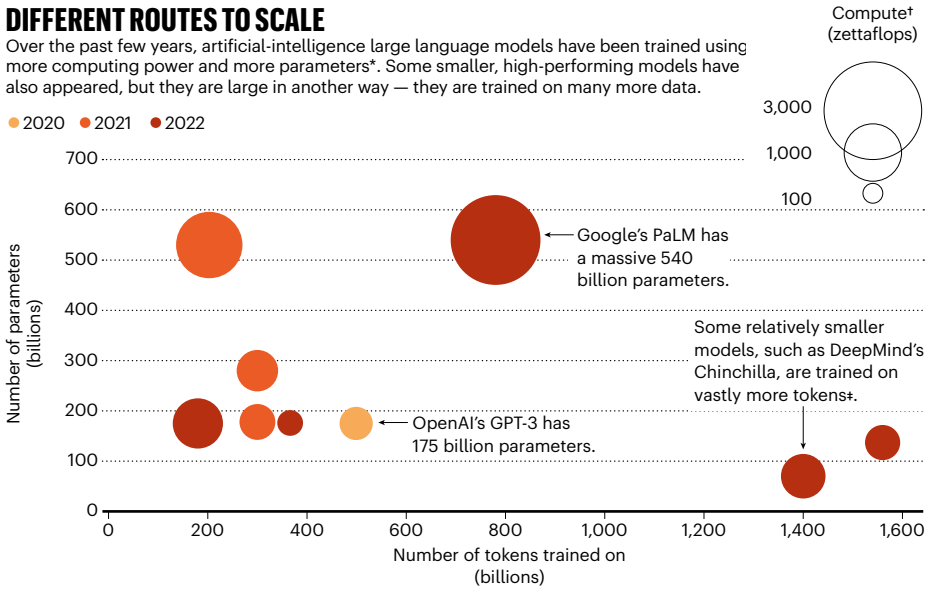
To improve further, however, even these more energy-efficient LLMs seem destined to become bigger, using up more data and compute. Researchers will be watching to see what new behaviours emerge with scale. "Whether it will fully unlock reasoning, I'm not sure," says Bubeck. "Nobody knows."

Anil Ananthaswamy is a freelance science writer based in California.

1. Frieder, S. *et al.* Preprint at https://arxiv.org/abs/2301.13867 (2023).
2. Lewkowycz, A. *et al.* Preprint at https://arxiv.org/abs/2206.14858 (2022).
3. Chowdhery, A. *et al.* Preprint at https://arxiv.org/abs/2204.02311 (2022).
4. Kaplan, J. *et al.* Preprint at https://arxiv.org/abs/2001.08361 (2020).
5. Hoffmann, J. *et al.* Preprint at https://arxiv.org/abs/2203.15556 (2022).
6. Caballero, E. *et al.* Preprint at https://arxiv.org/abs/2210.14891 (2022).
7. Bubeck, S. *et al.* Preprint at https://arxiv.org/abs/2105.12806 (2021).
8. Le Scao, T. *et al.* Preprint at https://arxiv.org/abs/2211.05100 (2022).
9. Luccioni, A. S. & Hernandez-Garcia, A. Preprint at https://arxiv.org/abs/2302.08476 (2023).
10. Luccioni, A. S., Viguier, S. & Ligozat, A.-L. Preprint at https://arxiv.org/abs/2211.02001 (2022).
11. Yin, B. *et al. Nature Mach. Intell.* **3**, 905–913 (2021).
12. Borgeaud, S. *et al.* Preprint at https://arxiv.org/abs/2112.04426 (2021).
13. Du, N. *et al.* Preprint at https://arxiv.org/abs/2112.06905 (2021).

DIFFERENT ROUTES TO SCALE

Over the past few years, artificial-intelligence large language models have been trained using more computing power and more parameters*. Some smaller, high-performing models have also appeared, but they are large in another way — they are trained on many more data.



*Parameters: roughly, the number of connections between neurons. †Compute: number of computing operations executed during training, measured as floating point operations (flops). ‡Tokens: words, digits or other units of information that models are trained on.

ADAPTED FROM OUR WORLD IN DATA, AND FROM J. SEVILLA *ET AL.* PREPRINT AT ARXV HTTPS://DOI.ORG/10.48550/ARXIV.2202.05924 (2022).