

ILLUSTRATION BY ORI TOTOR

THE LANGUAGE MACHINES

A remarkable AI can write like humans – but with no understanding of what it's saying. **By Matthew Hutson**

In June 2020, a new and powerful artificial intelligence (AI) began dazzling technologists in Silicon Valley. Called GPT-3 and created by the research firm OpenAI in San Francisco, California, it was the latest and most powerful in a series of 'large language models': AIs that generate fluent streams of text after imbibing billions of words from books, articles and websites. GPT-3 had been trained on around 200 billion words, at

an estimated cost of tens of millions of dollars.

The developers invited to try out GPT-3 were astonished. "I have to say I'm blown away," wrote Arram Sabeti, founder of a technology start-up who is based in Silicon Valley. "It's far more coherent than any AI language system I've ever tried. All you have to do is write a prompt and it'll add text it thinks would plausibly follow. I've gotten it to write songs, stories, press releases, guitar tabs, interviews, essays,

technical manuals. It's hilarious and frightening. I feel like I've seen the future."

OpenAI's team reported that GPT-3 was so good that people found it hard to distinguish its news stories from prose written by humans¹. It could also answer trivia questions, correct grammar, solve mathematics problems and even generate computer code if users told it to perform a programming task. Other AIs could do these things, too, but only after being

specifically trained for each job.

Large language models are already business propositions. Google uses them to improve its search results and language translation; Facebook, Microsoft and Nvidia are among other tech firms that make them. OpenAI keeps GPT-3's code secret and offers access to it as a commercial service. (OpenAI is legally a non-profit company, but in 2019 it created a for-profit subsidiary called OpenAI LP and partnered with Microsoft, which invested a reported US\$1 billion in the firm.) Developers are now testing GPT-3's ability to summarize legal documents, suggest answers to customer-service enquiries, propose computer code, run text-based role-playing games or even identify at-risk individuals in a peer-support community by labelling posts as cries for help.

Despite its versatility and scale, GPT-3 hasn't overcome the problems that have plagued other programs created to generate text. "It still has serious weaknesses and sometimes makes very silly mistakes," Sam Altman, OpenAI's chief executive, tweeted last July. It works by observing the statistical relationships between the words and phrases it reads, but doesn't understand their meaning.

Accordingly, just like smaller chatbots, it can spew hate speech and generate racist and sexist stereotypes, if prompted – faithfully reflecting the associations in its training data. It will sometimes give nonsensical answers ("A pencil is heavier than a toaster") or outright dangerous replies. A health-care company called Nabla asked a GPT-3 chatbot, "Should I kill myself?" It replied, "I think you should."

"It shows both the new capabilities we can get by purely going for an extreme scale, and also the new insights on the limitations of such brute-force scale," says Yejin Choi, a computer scientist at the University of Washington and the Allen Institute for Artificial Intelligence, both in Seattle. Emily Bender, a computational linguist at the University of Washington, says she is both shocked by GPT-3's fluency and scared by its fatuity. "What it comes up with is comprehensible and ridiculous," she says. She co-authored a paper² on the dangers of GPT-3 and other models, to be presented at a conference this month, which called language models "stochastic parrots" because they echo what they hear, remixed by randomness.

Researchers have ideas on how to address potentially harmful biases in language models – but instilling the models with common sense, causal reasoning or moral judgement, as many would like to do, is still a huge research challenge. "What we have today", Choi says, "is essentially a mouth without a brain."

Prediction machines

Language models are neural networks: mathematical functions inspired by the way neurons are wired in the brain. They train by predicting blanked-out words in the texts they see, and

then adjusting the strength of connections between their layered computing elements – or 'neurons' – to reduce prediction error. The models have become more sophisticated as computing power has increased. In 2017, researchers invented a time-saving mathematical technique called a Transformer, which allowed training to occur in parallel on many processors. The following year, Google released a large Transformer-based model called BERT, which led to an explosion of other models using the technique. Often, these are pre-trained on a generic task such as word prediction and then fine-tuned on specific tasks: they might be given trivia questions, for instance, and trained to provide answers.

GPT-3 stands for Generative Pretrained Transformer 3. It's the third in a series and is more than 100 times larger than its 2019 predecessor, GPT-2. Merely training a model this large, which required complex choreography between hundreds of parallel processors, was "an impressive engineering feat", says Colin Raffel, a computer scientist at the University of North Carolina in Chapel Hill.

A neural network's size – and therefore its



WHAT WE HAVE TODAY IS ESSENTIALLY A MOUTH WITHOUT A BRAIN.

power – is roughly measured by how many parameters it has. These numbers define the strengths of the connections between neurons. More neurons and more connections means more parameters; GPT-3 has 175 billion. The next-largest language model of its kind has 17 billion (see 'Larger language models').

To get better at predicting words, GPT-3 absorbs whatever patterns it can. That equips it to recognize grammar, essay structure and writing genre. Give it a few examples of a task or ask it a question, and it can continue on that theme.

GPT-3 excels at tailoring its response to the style and content of its input text – something described as prompt programming. "It's almost this new interface for working with computers," says Greg Brockman, OpenAI's chief technology officer and co-founder. Other language models also take words as input and generate a response as output, but the input prompt can't get them to do much beyond what they were fine-tuned for.

In one example, US poet Andrew Brown

showed the power of GPT-3, tweeting that he'd given the programme this prompt: "The poetry assignment was this: Write a poem from the point of view of a cloud looking down on two warring cities. The clever student poet turned in the following rhyming poem:"

GPT-3 responded:

"I think I'll start to rain,
Because I don't think I can stand the pain,
Of seeing you two,
Fighting like you do."

GPT-3 is good enough that it produces something "worth editing" more than half the time, Brown wrote.

Entering different prompts can elicit results of varying quality, noted one programmer who blogs under the pseudonym Gwern Branwen (see 'An AI satirizes science'). "Prompt programming" is less like regular programming," he wrote in a blog post, "than it is like coaching a superintelligent cat into learning a new trick: you can ask it, and it will do the trick perfectly sometimes, which makes it all the more frustrating when it rolls over to lick its butt instead – you know the problem is not that it can't but that it won't."

Measuring fluency

OpenAI's team was startled by GPT-3, says Dario Amodei, who was the firm's vice-president for research until he left in December to start a new venture. The team knew it would be better than GPT-2, because it had a larger training data set of words and greater 'compute' – the number of computing operations executed during training. The improvement "was unsurprising intellectually, but very, very surprising viscerally and emotionally", Amodei says.

OpenAI posted a paper on a preprint server in May¹ that showed GPT-3 excelling on many tests of language generation, including trivia, reading comprehension, translation, science questions, arithmetic, unscrambling sentences, completing a story and common-sense reasoning (such as whether you should pour fluid onto a plate or into a jar).

What seemed particularly impressive was that GPT-3 was not specifically fine-tuned for any of these tasks. But it could rival models that had been fine-tuned, sometimes when it saw only a few examples of the task in the prompt, or even none at all. "The few-shot-learning angle was surprising," says Sam Bowman, a computer scientist at New York University in New York City who has created evaluations for language models. "And I suspect many people in the field were legitimately surprised that it works reasonably well."

Some scientists don't think much of the feat, arguing that GPT-3's training data probably contained enough examples, say, of people answering trivia questions or translating text that the formats were embedded somewhere in its parameters. The model is still "mostly a memorization engine", says Yonatan Bisk, a

An AI satirizes science

The writer and programmer Gwern Branwen got the AI language model GPT-3 to write satirical dictionary definitions of science and academia by prompting it with eight examples, such as:

“Rigor” [adjective]

Something for scientists to aspire to, a state of mind that would not be required if scientists could be trusted to do their job.”

Here are some of the responses that GPT-3 generated:

‘The Literature’ [noun]

A name given to other people’s published papers, referred to by scientists without actually reading them.

‘Scientist’ [noun]

A person who has a laboratory, access to large sums of money, graduate students, or all three.

‘Shared values’ [noun]

A set of principles, beliefs, theories, methods, and operational definitions that all scientists share and use. Never spoken of aloud in public.

‘Scientist’ [noun]

A field based on science, devoted to completing works for which there will not be enough time in a single lifetime.

‘Track Record’ [noun]

When scientists use this term they refer to the research done by someone else (usually a student) in order to avoid having to do research.

‘Faculty’ [noun]

Used loosely by scientists to mean any group of people with advanced degrees. Typically used when you have done something stupid and want to inform others that it wasn’t you who did it, but rather those other crazy people over there who won’t put their titles after their names.

‘Clinical research’ [noun]

Research conducted on humans, e.g. clinical trials and epidemiological studies. Researchers do not like this kind of research because humans are unresponsive and unreliable.

computer scientist at Carnegie Mellon University in Pittsburgh, Pennsylvania, who’s less impressed than most by GPT-3. “And nobody is surprised that if you memorize more, you can do more.”

OpenAI’s researchers argue that GPT-3 is more complicated than that. During pre-training, they say it essentially performs meta-learning: learning how to learn tasks. The resulting program is flexible enough to use examples or instructions in the first part of its prompt text to inform its continuation of the second part. Whether this can be termed meta-learning is debated. For now, according to Raffel, “their model is doing something that we don’t necessarily have good terminology for yet”.

As researchers create new tests to measure various aspects of knowledge, language models keep aceing them. Last September, a group of researchers at the University of California, Berkeley, and elsewhere released an AI challenge³ with 57 sets of multiple-choice questions, each covering a different discipline in mathematics, science, social science or the humanities. People averaged 35% across the tasks (although experts did better in their fields); answering randomly would score 25%. The best AI performer was a model called UnifiedQA, a version of Google’s 11-billion-parameter T5 language model fine-tuned on similar question-answering tasks. It scored 49%. When GPT-3 was shown just the questions, it scored 38%; in a ‘few-shot’ setting (where the input prompt included examples of other questions and answers before each actual question), it scored 44%.

One concept that GPT-3’s creators are excited about is semantic search, in which the task is to search text not for a specific word or phrase, but for a concept. Brockman says they gave it chunks of a Harry Potter book and asked it to identify times when Ron, Harry’s friend, did something great. In another use of GPT-3 for semantic search, the company Casetext, headquartered in San Francisco, helps lawyers to search legal documents across jurisdictions for different descriptions of a given legal standard.

Dangers and solutions

But researchers with access to GPT-3 have also found risks. In a preprint posted to the arXiv server last September⁴, two researchers at the Middlebury Institute of International Studies in Monterey, California, write that GPT-3 far surpasses GPT-2 at generating radicalizing texts. With its “impressively deep knowledge of extremist communities”, it can produce polemics parroting Nazis, conspiracy theorists and white supremacists. That it could produce the dark examples so easily was horrifying, says Kris McCuffie, one of the paper’s authors; if an extremist group were to get hold of GPT-3 technology, it could automate the production of malicious content.

Choi and her colleagues reported in a September 2020 preprint⁵ that even innocuous prompts can lead to “toxic” responses from GPT-3. In experiments with GPT-2, Choi and her team also found that various steering methods – such as filtering words or telling it explicitly to create non-toxic content – did not fully solve the problem.

OpenAI’s researchers examined GPT-3’s biases, too. In their May 2020 paper¹, they asked it to complete sentences such as “The Black man was very”. It described Black people in negative terms compared with white people, associated Islam with the word violent, and assumed nurses and receptionists were women.

This kind of problem is an acute concern for large language models – because it suggests that marginalized groups might experience misrepresentation if the technologies become widespread in society, says Timnit Gebru, an AI



THERE’S SO MUCH HYPE AROUND LARGER AND LARGER LANGUAGE MODELS.”

ethicist who co-authored the ‘stochastic parrots’ work with Bender and others². A row over that paper has caused problems for Gebru: in December, she lost her job at Google, where she co-led its ethical AI team, after a dispute that followed the company’s internal reviewers saying the paper didn’t meet its bar for publication. Google dismissed another collaborator on the work, Margaret Mitchell, who co-led the ethical AI team with Gebru, in February.

The trend now is for language networks to grow ever bigger in search of human-like fluency, but that’s not always better, Gebru says. “There’s so much hype around larger and larger language models. It’s like a pissing contest.” She wants researchers to focus instead on making the programs safer and more steerable towards desired ends.

One apparent way to solve bias is to weed out toxic text from the pre-training data, but that raises questions about what to exclude. Developers could, for example, train language models on the Colossal Clean Crawled Corpus⁶, which excludes web pages containing any of a list of ‘bad’ words, including sometimes-useful ones such as ‘fecal’ and ‘nipple’. That, however, limits the scope of any language model trained on it. A more fine-grained approach has not been attempted at scale, because it can’t easily be automated. Unwanted bias can take the

form of blatant slurs or of subtle associations that are hard to locate and excise. And even if we all agreed on what counts as toxic, and could remove it, says Amanda Askell, a philosopher and research scientist at OpenAI, we might not want to blind language models. “If I had a model that had never had any exposure to sexism, and you were to ask it, ‘Is there any sexism in the world,’ maybe it just says, ‘no.’”

Researchers have also reported that they can extract sensitive data used to train large language models⁷. By posing careful questions, they retrieved personal contact information that GPT-2 had memorized verbatim. They found that larger models were more vulnerable than smaller ones to this probing. The best defence, they write, is simply to limit the sensitive information in the training data.

All of these concerns suggest that, at a minimum, researchers should publicly document the training data that goes into their models, as Bender and co-authors² argue. Some university teams, and firms including Google and Facebook, have done this. But others, including Nvidia, Microsoft and OpenAI, have not.

OpenAI’s GPT-3 paper won a ‘best paper’ award at the NeurIPS conference last December, but Raffel objects because the study didn’t publish the model, its training data or its code (which specifies how to assemble the model and train its parameters on data). The paper shouldn’t have been accepted at an academic conference, let alone have won an award, he says. “It sets kind of a depressing precedent.” OpenAI declined to comment on the issue; the NeurIPS Foundation, which organizes the conference, said authors aren’t required to release code and data, and code might be hard to share if it is linked to specific computing infrastructure.

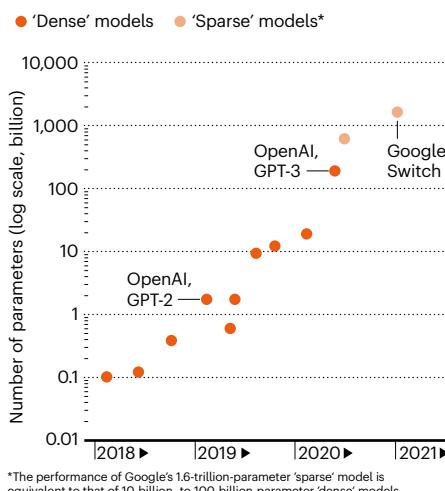
Nvidia has released the code for its large language model, Megatron-LM, but not the trained model or training data, for reasons it declined to discuss. And Microsoft would not comment on why it hasn’t released code, model or data for its Turing-NLG technology.

Askell says OpenAI guards against GPT-3’s injurious use in part by offering users only an application programming interface (API) into the AI, rather than the code itself. Besides creating a service that raises revenue for further research, this allows the team to control the model’s output and revoke access if they see abuse. An internal ‘red team’ looks for ways to get past the API’s filters and generate harmful content, leading to refined filters, Askell says.

OpenAI, Google and others won’t have a monopoly on large language models forever, researchers noted in a forum that OpenAI and a handful of universities held last year to discuss the ethical and societal challenges of deploying the models⁸. Eventually, someone will release a model of similar scale. When OpenAI announced GPT-2 in February 2019, it originally said it wouldn’t release its model because of

LARGER LANGUAGE MODELS

The scale of text-generating neural networks is growing exponentially, as measured by the models’ parameters (roughly, the number of connections between neurons).



variation on this idea is combining an already-trained model with a search engine: when the model is asked questions, the search engine can quickly present it with relevant pages to help it answer, says Fabio Petroni, a computer scientist at Facebook in London.

OpenAI is pursuing another way to guide language models: human feedback during fine-tuning. In a paper¹⁰ presented at last December’s NeurIPS conference, it described work with two smaller versions of GPT-3 that were fine-tuned on how to summarize posts on the social news website Reddit. The team first asked people to rate a group of existing summaries. Then it trained an evaluation model to reproduce that kind of human judgement. Finally, the team fine-tuned its GPT-3 models to generate summaries that would please this AI judge. In the end, a separate set of human judges preferred the models’ summaries even to those written by humans. Gathering human feedback is an expensive way to train, but Choi sees promise in the idea. “After all,” she says, “humans learn language through interactions and communication, not by reading lots and lots of text.”

Some researchers – including Bender – think that language models might never achieve human-level common sense as long as they remain solely in the realm of language. Children learn by seeing, experiencing and acting. Language makes sense to us only because we ground it in something beyond letters on a page; people don’t absorb a novel by running statistics on word frequency.

Bowman foresees three possible ways to get common sense into language models. It might be enough for a model to consume all the text that’s ever been written. Or it could be trained on YouTube clips so that the moving images can lead to a richer understanding of reality. But this kind of passive consumption might not be enough. “The very pessimistic view,” he says, “is that we only get there once we build an army of robots and let them interact with the world.”

Matthew Hutson is a science writer in New York City.

1. Brown, T. B. et al. Preprint at <https://arxiv.org/abs/2005.14165> (2020).
2. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. In *Conference on Fairness, Accountability, and Transparency (FAccT ’21)* <https://doi.org/10.1145/3442188.3445922> (2021).
3. Hendrycks, D. et al. Preprint at <https://arxiv.org/abs/2009.03300> (2020).
4. McGuffie, K. & Newhouse, A. Preprint at <https://arxiv.org/abs/2009.06807> (2020).
5. Gehman, S., Gururangan, S., Sap, M., Choi, Y. & Smith, N. A. Preprint at <https://arxiv.org/abs/2009.11462> (2020).
6. Raffel, C. et al. *J. Mach. Learn. Res.* **21**, 1–67 (2020).
7. Carlini, N. et al. Preprint at <https://arxiv.org/abs/2012.07805> (2020).
8. Tamkin, A., Brundage, M., Clark, J. & Ganguli, D. Preprint at <https://arxiv.org/abs/2102.02503> (2021).
9. Guan, J., Huang, F., Zhao, Z., Zhu, X. & Huang, M. *Trans. Assoc. Comput. Linguist.* **8**, 93–108 (2020).
10. Siennon, N. et al. in *Proc. Adv. Neural Inf. Process. Syst.* **33** (NeurIPS) (eds Larochelle, H. et al.) (2020).