

Meet GPT-3. It Has Learned to Code (and Blog and Argue).

The latest natural-language system generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs.



By **Cade Metz**

Nov. 24, 2020 Updated 9:48 a.m. ET

This summer, an artificial intelligence lab in San Francisco called OpenAI unveiled a technology several months in the making. This new system, GPT-3, had spent those months learning the ins and outs of natural language by analyzing thousands of digital books, the length and breadth of Wikipedia, and nearly a trillion words posted to blogs, social media and the rest of the internet.

Mckay Wrigley, a 23-year-old computer programmer from Salt Lake City, was one of the few invited to tinker with the system, which uses everything it has learned from that vast sea of digital text to generate new language on its own. Mr. Wrigley wondered if it could imitate public figures — write like them, perhaps even chat like them.

One of his experiments involved a pop psychologist, Scott Barry Kaufman. The system took in Mr. Kaufman's name and a topic for discussion: creativity. Then, when asked "How do we become more creative?" GPT-3 responded instantly:

I think creative expression is a natural byproduct of growing up in a diverse world. The more diverse the world is, the more you get exposed to different people, to different opportunities, to different places and to different challenges. And the more diverse that is, the more likely you'll be to be able to put the dots together to form something new. And in many ways, I think if you want to be creative, you have to go for it. If you want to be a writer, you have to write, if you want to be a musician, you have to create music, if you want to be a comedian, you have to create comedy, and the more you create, the more likely it is that you'll stumble onto some interesting stuff.

Later, when Mr. Wrigley posted the paragraph on Twitter, somebody looped in the real Scott Barry Kaufman. He was stunned. "It definitely sounds like something I would say," the real Mr. Kaufman tweeted, later adding, "Crazy accurate A.I."

In the weeks since its arrival, GPT-3 has spawned dozens of other experiments that raise the eyebrows in much the same way. It generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs, all with very little prompting. Some of these skills caught even the experts off guard.

For many artificial intelligence researchers, it is an unexpected step toward machines that can understand the vagaries of human language — and perhaps even tackle other human skills.

“It is surprising to me, and to a lot of people,” said Melanie Mitchell, an A.I. researcher at the Santa Fe Institute, an independent lab in New Mexico, who is among those experimenting with the system. “It is hard to figure out exactly what this thing is doing.”

GPT-3 is far from flawless. It often spews biased and toxic language. And if you ask for 10 paragraphs in the style of Scott Barry Kaufman, it might give you five that are convincing — and five others that are not. When Mr. Kaufman actually chatted with his A.I. doppelgänger, he was not nearly as impressed as he was at first:

Scott Barry Kaufman: What is imagination?

“Scott Barry Kaufman”: What is imagination? You should tell me!

Scott Barry Kaufman: You should know!

“Scott Barry Kaufman”: Nope, you should know. You are the one who knows what imagination is. What do you think it is?

But even this looked like real conversation. And notably, it showed what the system could do entirely on its own, without extra coding or fine-tuning.

GPT-3 is the culmination of several years of work inside the world’s leading artificial intelligence labs, including OpenAI, an independent organization backed by \$1 billion dollars in funding from Microsoft, as well as labs at Google and Facebook. At Google, a similar system helps answer queries on the company’s search engine.

These systems — known as universal language models — can help power a wide range of tools, like services that automatically summarize news articles and “chatbots” designed for online conversation. So far, their impact on real-world technology has been small. But GPT-3 — which learned from a far larger collection of online text than previous systems — opens the door to a wide range of new possibilities, such as software that can speed the development of new smartphone apps, or chatbots that can converse in far more human ways than past technologies.

As software designers, entrepreneurs, pundits and artists explore this system, each new experiment stokes an already heated debate over how powerful this breed of technology will ultimately be. While some say it may be a path toward truly intelligent machines, others argue that these experiments, while endlessly fascinating, are also misleading.

“It is very fluent,” said Mark Riedl, a professor and researcher at the Georgia Institute of Technology. “It is very articulate. It is very good at producing reasonable-sounding text. What it does not do, however, is think in advance. It does not plan out what it is going to say. It does not really have a goal.”

An ‘emergent quality’

Jordan Singer is a product designer at Square, the Silicon Valley mobile-payments company. He helps design the company’s smartphone apps, building the graphics, menus, buttons and other widgets that define an app’s look and feel. When he heard about GPT-3, he wondered if this automated system could do his job.

He fed the system a simple description of a smartphone app, and the computer code needed to create the app. The description was in plain English. The code was built inside Figma, a specialized design tool used by professionals like Mr. Singer.

He did this a few more times, feeding the system several more English-language descriptions alongside the matching Figma code. And when he was done, GPT-3 could write such code on its own.

If he described a simple app for posting and viewing photos as a user would on Instagram, the system generated the code needed to build it. This code was sometimes flawed. But typically, if Mr. Singer made just a tweak or two, it worked as he wanted. “It’s not absolutely perfect,” he said. “But it is very, very close.”

This behavior was entirely new, and it surprised even the designers of GPT-3. They had not built GPT-3 to generate computer code, just as they had not built it to write like Mr. Kaufman or generate tweets or translate languages. They had built it to do just one thing: predict the next word in a sequence of words.

GPT-3 is what artificial intelligence researchers call a neural network, a mathematical system loosely modeled on the web of neurons in the brain. This is the same technology that identifies faces in the photos you post to Facebook and recognizes the commands you bark into your iPhone.

A neural network learns such skills by pinpointing patterns in vast amounts of digital data. By analyzing thousands of cat photos, for instance, it can learn to recognize a cat.

About three years ago, researchers at Google and top labs like OpenAI started designing neural networks that learned from enormous amounts of prose, including unpublished books and Wikipedia articles by the thousands. These universal language models could be applied not just to one task, like translation, but to many.

GPT-3 analyzed digital prose on an unprecedented scale, spending months looking for patterns in huge amounts of text posted to the internet. In this way, it learned to predict the next word in a sequence. If you type a few words into GPT-3, it will keep going, completing your thought with entire paragraphs of text.

But in acquiring this specific skill, it learned much more. During its months of training, GPT-3 identified more than 175 billion parameters — mathematical representations of patterns — in that sea of books, Wikipedia articles and other online texts. These patterns amount to a map of human language: a mathematical description of the way we piece characters together, whether we are writing blogs or coding software programs. Using this map, GPT-3 can perform all sorts of tasks it was not built to do.

Before asking GPT-3 to generate new text, you can focus it on particular patterns it may have learned during its training, priming the system for certain tasks. You can feed it descriptions of smartphone apps and the matching Figma code. Or you can show it reams of human dialogue. Then, when you start typing, it will complete the sequence in a more specific way. If you prime it with dialogue, for instance, it will start chatting with you.

“It has this emergent quality,” said Dario Amodei, vice president for research at OpenAI. “It has some ability to recognize the pattern that you gave it and complete the story, give another example.”

Previous language models worked in similar ways. But GPT-3 can do things that previous models could not, like write its own computer code. And, perhaps more important, you can prime it for specific tasks using just a few examples, as opposed to the thousands of examples and several hours of additional training required by its predecessors. Researchers call this “few-shot learning,” and they believe GPT-3 is the first real example of what could be a powerful phenomenon.

“It exhibits a capability that no one thought possible,” said Ilya Sutskever, OpenAI’s chief scientist and a key figure in the rise of artificial intelligence technologies over the past decade. “Any layperson can take this model and provide these examples in about five minutes and get useful behavior out of it.”

This is both a blessing and a curse.

Unsafe for work?

OpenAI plans to sell access to GPT-3 via the internet, turning it into a widely used commercial product, and this year it made the system available to a limited number of beta testers through their web browsers. Not long after, Jerome Pesenti, who leads the Facebook A.I. lab, called GPT-3 “unsafe,” pointing to sexist, racist and otherwise toxic language the system generated when asked to discuss women, Black people, Jews and the Holocaust.

With systems like GPT-3, the problem is endemic. Everyday language is inherently biased and often hateful, particularly on the internet. Because GPT-3 learns from such language, it, too, can show bias and hate. And because it learns from internet text that associates atheism with the words “cool” and “correct” and that pairs Islam with “terrorism,” GPT-3 does the same thing.

This may be one reason that OpenAI has shared GPT-3 with only a small number of testers. The lab has built filters that warn that toxic language might be coming, but they are merely Band-Aids placed over a problem that no one quite knows how to solve.

“They are doing the right thing by not just publicly releasing GPT-3,” said Allison Koenecke, a Stanford researcher who explores unwanted bias in A.I. systems. “A lot is still up in the air.”

The onus is ultimately on OpenAI to ensure that this behavior remains in check, said Liz O’Sullivan, a vice president with Arthur, a company that helps businesses manage the behavior of artificial intelligence technologies. As it stands, she said, OpenAI is “passing along legal and reputation risk to anyone who might want to use the model in consumer-facing applications.”

Other experts worry that these language models could help spread disinformation across the internet, amping up the kind of online campaigns that may have helped sway the 2016 presidential election. GPT-3 points to a future in which we are even less sure if what we are reading is real or fake. That goes for tweets, online conversations, even long-form prose.

At the end of July, Liam Porr, a student at the University of California, Berkeley, generated several blog posts with GPT-3 and posted them on the internet, where they were read by 26,000 people. Sixty viewers were inspired to subscribe to the blog, and only a few suspected that the posts were written by a machine.

They were not necessarily gullible people. One of the blog posts — which argued that you can increase your productivity if you avoid thinking too much about everything you do — rose to the top of the leader board on Hacker News, a site where seasoned Silicon Valley programmers, engineers and entrepreneurs rate news articles and other online content. (“In

order to get something done, maybe we need to think less,” the post begins. “Seems counterintuitive, but I believe sometimes our thoughts can get in the way of the creative process.”)

But as with most experiments involving GPT-3, Mr. Porr’s is not as powerful as it might seem.

The flaws nobody notices

In the mid-1960s, Joseph Weizenbaum, a researcher at the Massachusetts Institute of Technology, built an automated psychotherapist he called ELIZA. Judged from our vantage point in 2020, this chatbot was exceedingly simple.

Unlike GPT-3, ELIZA did not learn from prose. It operated according to a few basic rules defined by its designer. It pretty much repeated whatever you said to it, only in the form of a question. But much to Dr. Weizenbaum’s surprise, many people treated the bot as if it were human, unloading their problems without reservation and taking comfort in the responses.

When dogs and other animals exhibit even small amounts of humanlike behavior, we tend to assume they are more like us than they really are. The same goes for machines, said Colin Allen, a professor at the University of Pittsburgh who explores cognitive skills in both animals and machines. “People get sucked in,” he said, “even if they know they are being sucked in.”

That is part of what is happening with GPT-3. Because it can generate convincing tweets, blog posts and computer code, we read humanity into this digital system — and pay less attention to its limits.

In practice, the system fails about as often as it succeeds. We overlook that the computer code it writes requires some fine-tuning from human programmers — a line removed here or added there. We do not notice that its talent for conversation breaks down after a few exchanges, when it cannot “remember” what it said just a few seconds before. We do not quite realize that although the system generated a convincing blog post for Mr. Porr, he provided the headline and the photo and the first few sentences, and he removed some sentences that were less convincing.

Mr. Porr does not believe GPT-3 is an enormous threat to the battle against disinformation in the short term, because it still requires so much help from humans. A tool like this becomes truly dangerous only if it can generate enormous amounts of convincing disinformation entirely on its own, exceeding what a team of hired hands can do with relative ease today.

Similarly, when app designers ask Mr. Singer of Square if GPT-3 is a threat to their careers, he assures them it is not — at least not yet. He sees it as a way of making their jobs easier. “If it can get 70 percent of the way there, that is a lot of tedious work taken out of the equation,” he said.

What we do not know is how much this technology will continue to improve in the months and years to come.

Smarter, faster, even more expensive

While the researchers at OpenAI were training GPT-3 on more than a trillion words posted to the internet, they ran a second experiment, training a similar system on tens of thousands of digital photos. That system could analyze all of those photos and learn to build images in much the same way that GPT-3 builds paragraphs. Given half of a cat photo, it could generate the rest of the cat.

For some researchers, the experiment indicates that such a system could ultimately handle tasks across multiple dimensions — language, sight, sound — much like humans do. Even when trained solely on language, they say, the system could already reach into other areas, whether computer programming, playing chess or generating guitar tabs.

But continuing to improve this technology is far from trivial. Processing all of that internet data requires a specialized supercomputer running for months on end, an undertaking that is enormously expensive. When asked if such a project ran into the millions of dollars, Sam Altman, OpenAI's chief executive, said the costs were actually "higher," running into the tens of millions.

Mr. Amodei, OpenAI's vice president for research, said there was still room to improve the technique, using more processing power to analyze more data. But he also said the approach might be close to running out of "juice."

At the very least, GPT-3 is a new tool for a world of A.I. researchers and entrepreneurs, a way of building all sorts of new technologies and new products. Mr. Wrigley, the computer programmer, recently quit his day job to start a company called LearnFromAnyone, which aims to build a kind of automated tutor using GPT-3 that can assume the guise of everyone from scientist Douglas Hofstadter to venture capitalist Peter Thiel. Others are building companies that aim to automatically generate code for computer programmers and automatically write promotional emails and tweets for marketing professionals.

But it is unclear how effective these services will ultimately be. If GPT-3 generates the right text only half of the time, can it satisfy professionals? And it is unclear whether this technique is a path to truly conversational machines, let alone truly intelligent systems. Additional progress on the long road to machines that can mimic the human brain, Mr. Amodei said, will require entirely new ideas.

"It is kind of like a chemistry reaction," he said. "We have this one ingredient. But other ingredients are required as well."

[Like the Science Times page on Facebook. | Sign up for the Science Times newsletter.]

Cade Metz is a technology correspondent, covering artificial intelligence, driverless cars, robotics, virtual reality, and other emerging areas. He previously wrote for Wired magazine. @cademetz

A version of this article appears in print on , Section D, Page 6 of the New York edition with the headline: Meet GPT-3. It Has Learned To Code (and Blog and Argue).