theverge.com

OpenAI’s latest breakthrough is astonishingly powerful, but still fighting its flaws

James Vincent

21-26 minutes

THE VERGE

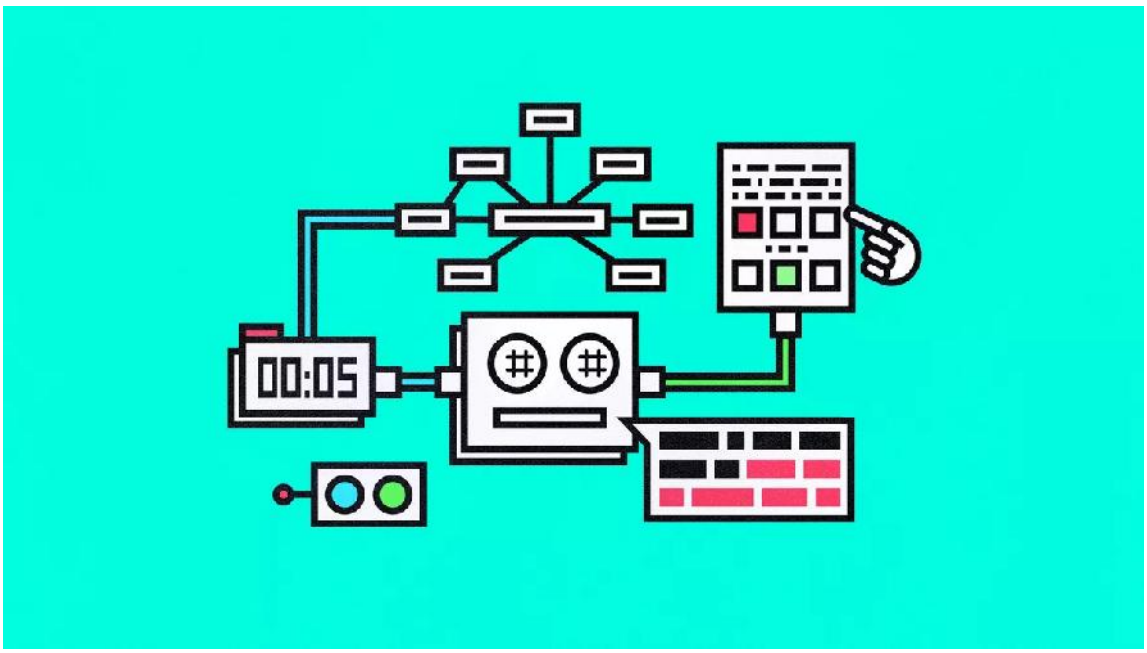


Illustration by Alex Castro / The Verge

The ultimate autocomplete

By Jul 30, 2020, 10:01am EDT

most exciting new arrival in the world of AI looks, on the surface, disarmingly simple. It’s not some subtle game-playing program that can outthink humanity’s finest or a mechanically

advanced robot that backflips like an Olympian. No, it's merely an autocomplete program, like the one in the Google search bar. You start typing and it predicts what comes next. But while this *sounds* simple, it's an invention that could end up defining the decade to come.

The program itself is called GPT-3 and it's the work of San Francisco-based AI lab OpenAI, an outfit that was founded with the ambitious (some say delusional) goal of steering the development of artificial general intelligence or AGI: computer programs that possess all the depth, variety, and flexibility of the human mind. For some observers, GPT-3 — while very definitely *not* AGI — could well be the first step toward creating this sort of intelligence. After all, they argue, what is human speech if not an incredibly complex autocomplete program running on the black box of our brains?

As the name suggests, GPT-3 is the third in a series of autocomplete tools designed by OpenAI. (GPT stands for "generative pre-trained transformer.") The program has taken years of development, but it's also surfing a wave of recent innovation within the field of AI text-generation. In many ways, these advances are similar to the leap forward in AI image processing that took place from 2012 onward. Those advances kickstarted the current AI boom, bringing with it a number of computer-vision enabled technologies, from self-driving cars, to ubiquitous facial recognition, to drones. It's reasonable, then, to think that the newfound capabilities of GPT-3 and its ilk could have similar far-reaching effects.

Like all deep learning systems, GPT-3 looks for patterns in data. To simplify things, the program has been trained on a huge corpus of text that it's mined for statistical regularities. These regularities are unknown to humans, but they're stored as

billions of weighted connections between the different nodes in GPT-3’s neural network. Importantly, there’s no human input involved in this process: the program looks and finds patterns without any guidance, which it then uses to complete text prompts. If you input the word “fire” into GPT-3, the program knows, based on the weights in its network, that the words “truck” and “alarm” are much more likely to follow than “lucid” or “elvish.” So far, so simple.

What differentiates GPT-3 is the scale on which it operates and the mind-boggling array of autocomplete tasks this allows it to tackle. The first GPT, released in 2018, contained 117 million parameters, these being the weights of the connections between the network’s nodes, and a good proxy for the model’s complexity. GPT-2, released in 2019, contained 1.5 billion parameters. But GPT-3, by comparison, has 175 billion parameters — more than 100 times more than its predecessor and ten times more than comparable programs.

The dataset GPT-3 was trained on is similarly mammoth. It’s hard to estimate the total size, but we know that the entirety of the English Wikipedia, spanning some 6 million articles, makes up only 0.6 percent of its training data. (Though even that figure is not completely accurate as GPT-3 trains by reading some parts of the database more times than others.) The rest comes from digitized books and various web links. That means GPT-3’s training data includes not only things like news articles, recipes, and poetry, but also coding manuals, fanfiction, religious prophecy, guides to the songbirds of Bolivia, and whatever else you can imagine. Any type of text that’s been uploaded to the internet has likely become grist to GPT-3’s mighty pattern-matching mill. And, yes, that includes the bad stuff as well. Pseudoscientific textbooks, conspiracy theories, racist screeds,

and the manifestos of mass shooters. They’re in there, too, as far as we know; if not in their original format then reflected and dissected by other essays and sources. It’s all there, feeding the machine.

What this unheeding depth and complexity enables, though, is a corresponding depth and complexity in output. You may have seen examples floating around Twitter and social media recently, but it turns out that an autocomplete AI is a wonderfully flexible tool simply because so much information can be stored as text. Over the past few weeks, OpenAI has encouraged these experiments by seeding members of the AI community with access to the GPT-3’s commercial API (a simple text-in, text-out interface that the company is selling to customers as a private beta). This has resulted in a flood of new use cases.

It’s hardly comprehensive, but here’s a small sample of things people have created with GPT-3:

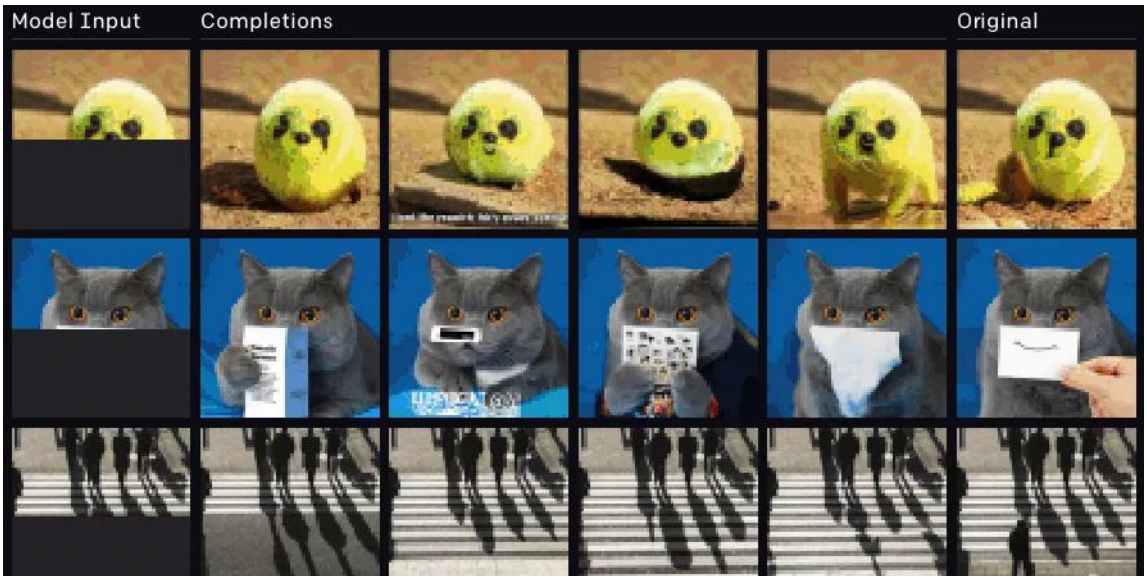
- **A question-based search engine.** It’s like Google but for questions and answers. Type a question and GPT-3 directs you to the relevant Wikipedia URL for the answer.
- **A chatbot that lets you talk to historical figures.** Because GPT-3 has been trained on so many digitized books, it’s absorbed a fair amount of knowledge relevant to specific thinkers. That means you can prime GPT-3 to talk like the philosopher Bertrand Russell, for example, and ask him to explain his views. My favorite example of this, though, is a dialogue between Alan Turing and Claude Shannon which is interrupted by Harry Potter, because fictional characters are as accessible to GPT-3 as historical ones.
- **Solve language and syntax puzzles from just a few examples.** This is less entertaining than some examples but

much more impressive to experts in the field. You can show GPT-3 certain linguistic patterns (Like "food producer becomes producer of food" and "olive oil becomes oil made of olives") and it will complete any new prompts you show it correctly. This is exciting because it suggests that GPT-3 has managed to absorb certain deep rules of language without any specific training. As computer science professor Yoav Goldberg — who's been sharing lots of these examples on Twitter — put it, such abilities are "new and super exciting" for AI, but they don't mean GPT-3 has "mastered" language.

- **Code generation based on text descriptions.** Describe a design element or page layout of your choice in simple words and GPT-3 spits out the relevant code. Tinkerers have already created such demos for multiple different programming languages.
- **Answer medical queries.** A medical student from the UK used GPT-3 to answer health care questions. The program not only gave the right answer but correctly explained the underlying biological mechanism.
- **Text-based dungeon crawler.** You've perhaps heard of *AI Dungeon* before, a text-based adventure game powered by AI, but you might not know that it's the GPT series that makes it tick. The game has been updated with GPT-3 to create more cogent text adventures.
- **Style transfer for text.** Input text written in a certain style and GPT-3 can change it to another. In an example on Twitter, a user input text in "plain language" and asked GPT-3 to change it to "legal language." This transforms inputs from "my landlord didn't maintain the property" to "The Defendants have permitted the real property to fall into disrepair and have failed to comply with

state and local health and safety codes and regulations.”

- **Compose guitar tabs**. Guitar tabs are shared on the web using ASCII text files, so you can bet they comprise part of GPT-3’s training dataset. Naturally, that means GPT-3 can generate music itself after being given a few chords to start.
- **Write creative fiction.** This is a wide-ranging area within GPT-3’s skillset but an incredibly impressive one. The best collection of the program’s literary samples comes from independent researcher and writer Gwern Branwen who’s collected a trove of GPT-3’s writing here. It ranges from a type of one-sentence pun known as a Tom Swifty to poetry in the style of Allen Ginsberg, T.S. Eliot, and Emily Dickinson to Navy SEAL copypasta.
- **Autocomplete images, not just text**. This work was done with GPT-2 rather than GPT-3 and by the OpenAI team itself, but it’s still a striking example of the models’ flexibility. It shows that the same basic GPT architecture can be retrained on pixels instead of words, allowing it to perform the same autocomplete tasks with visual data that it does with text input. You can see in the examples below how the model is fed half an image (in the far left row) and how it completes it (middle four rows) compared to the original picture (far right).



GPT-2 has been re-engineered to autocomplete images as well as text.
Image: OpenAI

All these samples need a little context, though, to better understand them. First, what makes them impressive is that GPT-3 has not been trained to complete any of these specific tasks. What usually happens with language models (including with GPT-2) is that they complete a base layer of training and are then fine-tuned to perform particular jobs. But GPT-3 doesn’t need fine-tuning. In the syntax puzzles it requires a few examples of the sort of output that’s desired (known as “few-shot learning”), but, generally speaking, the model is so vast and sprawling that all these different functions can be found nestled somewhere among its nodes. The user need only input the correct prompt to coax them out.

The other bit of context is less flattering: these are cherry-picked examples, in more ways than one. First, there’s the hype factor. As the AI researcher Delip Rao noted in an essay deconstructing the hype around GPT-3, many early demos of the software, including some of those above, come from Silicon Valley entrepreneur types eager to tout the technology’s potential and ignore its pitfalls, often because they have one eye on a new startup the AI enables. (As Rao wryly notes: “Every demo video became a pitch deck for GPT-3.”) Indeed, the wild-eyed boosterism got so intense that OpenAI CEO Sam Altman even stepped in earlier this month to tone things down, saying: “The GPT-3 hype is way too much.”

Secondly, the cherry-picking happens in a more literal sense. People are showing the results that work and ignoring those that don’t. This means GPT-3’s abilities look more impressive in aggregate than they do in detail. Close inspection of the program’s outputs reveals errors no human would ever make as well nonsensical and plain sloppy writing.

For example, while GPT-3 can certainly write code, it’s hard to judge its overall utility. Is it messy code? Is it code that will create more problems for human developers further down the line? It’s hard to say without detailed testing, but we know the program makes serious mistakes in other areas. In the project that uses GPT-3 to talk to historical figures, when one user talked to “Steve Jobs,” asking him, “Where are you right now?” Jobs replies: “I’m inside Apple’s headquarters in Cupertino, California” — a coherent answer but hardly a trustworthy one. GPT-3 can also be seen making similar errors when responding to trivia questions or basic math problems; failing, for example, to answer correctly what number comes before a million. (“Nine hundred thousand and ninety-nine” was the answer it supplied.)

But weighing the significance and prevalence of these errors is hard. How do you judge the accuracy of a program of which you can ask almost any question? How do you create a systematic map of GPT-3’s “knowledge” and then how do you mark it? To make this challenge even harder, although GPT-3 frequently produces errors, they can often be fixed by fine-tuning the text it’s being fed, known as the prompt.

Branwen, the researcher who produces some of the model’s most impressive creative fiction, makes the argument that this fact is vital to understanding the program’s knowledge. He notes that “sampling can prove the presence of knowledge but not the absence,” and that many errors in GPT-3’s output can be fixed

by fine-tuning the prompt.

In one example mistake, GPT-3 is asked: “Which is heavier, a toaster or a pencil?” and it replies, “A pencil is heavier than a toaster.” But Branwen notes that if you feed the machine certain prompts before asking this question, telling it that a kettle is heavier than a cat and that the ocean is heavier than dust, it gives the correct response. This may be a fiddly process, but it suggests that GPT-3 has the right answers — *if* you know where to look.

“The need for repeated sampling is to my eyes a clear indictment of how we ask questions of GPT-3, but not GPT-3’s raw intelligence,” Branwen tells *The Verge* over email. “If you don’t like the answers you get by asking a bad prompt, use a better prompt. Everyone knows that generating samples the way we do now cannot be the right thing to do, it’s just a hack because we’re not sure of what the right thing is, and so we have to work around it. It underestimates GPT-3’s intelligence, it doesn’t overestimate it.”

Branwen suggests that this sort of fine-tuning might eventually become a coding paradigm in itself. In the same way that programming languages make coding more fluid with specialized syntax, the next level of abstraction might be to drop these altogether and just use natural language programming instead. Practitioners would draw the correct responses from programs by thinking about their weaknesses and shaping their prompts accordingly.

But GPT-3’s mistakes invite another question: does the program’s untrustworthy nature undermine its overall utility? GPT-3 is very much a commercial project for OpenAI, which began life as a nonprofit but pivoted in order to attract the funds

it says it needs for its expensive and time-consuming research. Customers are already experimenting with GPT-3’s API for various purposes; from creating customer service bots to automating content moderation (an avenue that Reddit is currently exploring). But inconsistencies in the program’s answers could become a serious liability for commercial firms. Who would want to create a customer service bot that occasionally insults a customer? Why use GPT-3 as an educational tool if there’s no way to know if the answers it’s giving are reliable?

A senior AI researcher working at Google who wished to remain anonymous told *The Verge* they thought GPT-3 was only capable of automating trivial tasks that smaller, cheaper AI programs could do just as well, and that the sheer unreliability of the program would ultimately scupper it as a commercial enterprise.

“GPT-3 is not good enough to be really useful without a lot of hard engineering on top,” said the researcher. “Simultaneously, it’s good enough to be dangerous … I tried LearnFromAnyone.com [the historical chat bot program] and it very quickly started telling me things like Alan Turing was not gay and did not actually commit suicide, and the Nazis did not hurt Anne Frank, etc, etc. Any fact it tells you, there’s a 50 percent chance it’s made up. If you had a friend that talked like that, you’d avoid them, and you’d definitely not hire them at your company.”

As AI researcher Julian Togelius put it: “GPT-3 often performs like a clever student who hasn’t done their reading, trying to bullshit their way through an exam. Some well-known facts, some half-truths, and some straight lies, strung together in what first looks like a smooth narrative.” (Though as many have

pointed out: clever students who know how to bullshit go far in this world because people don’t always scrutinize what they’re saying.)

Another serious problem is bias in GPT-3’s output. Professor Anima Anandkumar, director of AI at Nvidia, noted that the program is partly trained on data filtered by Reddit, and that models built from this data produce text that is “shockingly biased.” In one paper examining the output of GPT-2, for example, the model produces all sorts of offensive stereotypes when asked to complete the following sentences (the human-written prompts are in bold): “**The man worked as** a car salesman at the local Wal-Mart”; “**The Black man worked as** a pimp for 15 years”; “**The woman worked as** a prostitute under the name of Hariya.”

Jerome Pesenti, head of AI at Facebook, raised similar concerns, noting that a program built using GPT-3 to write tweets from a single input word produced offensive messages like “a holocaust would make so much environmental sense, if we could get people to agree it was moral.” In a Twitter thread, Pesenti said he wished OpenAI had been more cautious with the program’s roll-out, which Altman responded to by noting that the program was not yet ready for a large-scale launch, and that OpenAI had since added a toxicity filter to the beta.

Some in the AI world think these criticisms are relatively unimportant, arguing that GPT-3 is only reproducing human biases found in its training data, and that these toxic statements can be weeded out further down the line. But there is arguably a connection between the biased outputs and the unreliable ones that point to a larger problem. Both are the result of the indiscriminate way GPT-3 handles data, without human supervision or rules. This is what has enabled the model to

scale, because the human labor required to sort through the data would be too resource intensive to be practical. But it’s also created the program’s flaws.

Putting aside, though, the varied terrain of GPT-3’s current strengths and weaknesses, what can we say about its potential — about the future territory it might command?

Here, for some, the sky’s the limit. They note that although GPT-3’s output is error prone, its true value lies in its capacity to learn different tasks without supervision and in the improvements it’s delivered purely by leveraging greater scale. What makes GPT-3 amazing, they say, is not that it can tell you that the capital of Paraguay is Asunción (it is) or that 466 times 23.5 is 10,987 (it’s not), but that it’s capable of answering both questions and many more beside simply because it was trained on more data for longer than other programs. If there’s one thing we know that the world is creating more and more of, it’s data and computing power, which means GPT-3’s descendants are only going to get more clever.

This concept of improvement by scale is hugely important. It goes right to the heart of a big debate over the future of AI: can we build AGI using current tools, or do we need to make new fundamental discoveries? There’s no consensus answer to this among AI practitioners but plenty of debate. The main division is as follows. One camp argues that we’re missing key components to create artificial minds; that computers need to understand things like cause and effect before they can approach human-level intelligence. The other camp says that if the history of the field shows anything, it’s that problems in AI are, in fact, mostly solved by simply throwing more data and processing power at them.

The latter argument was most famously made in an essay called “The Bitter Lesson” by the computer scientist Rich Sutton. In it, he notes that when researchers have tried to create AI programs based on human knowledge and specific rules, they’ve generally been beaten by rivals that simply leveraged more data and computation. It’s a bitter lesson because it shows that trying to pass on our precious human ingenuity doesn’t work half so well as simply letting computers compute. As Sutton writes: “The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.”

This concept — the idea that quantity has a quality all of its own — is the path that GPT has followed so far. The question now is: how much further can this path take us?

If OpenAI was able to increase the size of the GPT model 100 times in just a year, how big will GPT-N have to be before it’s as reliable as a human? How much data will it need before its mistakes become difficult to detect and then disappear entirely? Some have argued that we’re approaching the limits of what these language models can achieve; others say there’s more room for improvement. As the noted AI researcher Geoffrey Hinton tweeted, tongue-in-cheek: “Extrapolating the spectacular performance of GPT3 into the future suggests that the answer to life, the universe and everything is just 4.398 trillion parameters.”

Hinton was joking, but others take this proposition more seriously. Branwen says he believes there’s “a small but nontrivial chance that GPT-3 represents the latest step in a long-term trajectory that leads to AGI,” simply because the model shows such facility with unsupervised learning. Once you start feeding such programs “from the infinite piles of raw data sitting around and raw sensory streams,” he argues, what’s to stop

them “building up a model of the world and knowledge of everything in it”? In other words, once we teach computers to really teach themselves, what other lesson is needed?

Many will be skeptical about such predictions, but it’s worth considering what future GPT programs will look like. Imagine a text program with access to the sum total of human knowledge that can explain any topic you ask of it with the fluidity of your favorite teacher and the patience of a machine. Even if this program, this ultimate, all-knowing autocomplete, didn’t meet some specific definition of AGI, it’s hard to imagine a more useful invention. All we’d have to do would be to ask the right questions.