GPT-3, Bloviator: OpenAI’s language generator has no idea what it’s talking about

by Gary Marcus archive pageErnest Davis archive page

Since OpenAI first described its new AI language-generating system called GPT-3 in May, hundreds of media outlets (including MIT Technology Review) have written about the system and its capabilities. Twitter has been abuzz about its power and potential. The New York Times published an op-ed about it. Later this year, OpenAI will begin charging companies for access to GPT-3, hoping that its system can soon power a wide variety of AI products and services.

Is GPT-3 an important step toward artificial general intelligence—the kind that would allow a machine to reason broadly in a manner similar to humans without having to train for every specific task it encounters? OpenAI’s technical paper is fairly reserved on this larger question, but to many, the sheer fluency of the system feels as though it might be a significant advance.

We doubt it. At first glance, GPT-3 seems to have an impressive ability to produce human-like text. And we don't doubt that it can used to produce entertaining surrealist fiction; other commercial applications may emerge as well. But accuracy is not its strong point. If you dig deeper, you discover that something’s amiss: although its output is grammatical, and even impressively idiomatic, its comprehension of the world is often seriously off, which means you can never really trust what it says.

Below are some illustrations of its lack of comprehension—all, as we will see later, prefigured in an earlier critique that one of us wrote about GPT-3’s predecessor.

Before proceeding, it’s also worth noting that OpenAI has thus far not allowed us research access to GPT-3, despite both the company’s name and the nonprofit status of its oversight organization. Instead, OpenAI put us off indefinitely despite repeated requests—even as it made access widely available to the media. Fortunately, our colleague Douglas Summers-Stay, who had access, generously offered to run the experiments for us.

OpenAI’s striking lack of openness seems to us to be a serious breach of scientific ethics, and a distortion of the goals of the associated nonprofit. Its decision forced us to limit our testing to a comparatively small number of examples, giving us less time to investigate than we would have liked, which means there may be more serious problems that we didn’t have a chance to discern. Even so, within the constraints of a small sample, many major issues with GPT-3 were immediately evident, in every domain of reasoning and comprehension that we tested.

In each example below, the prompt we generated is in ordinary font and the system's continuation is in **bold**. Our comments are in *italics*. (A complete list of the experiments we did can be found here.)

Biological reasoning

- You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can't smell anything. You are very thirsty. So **you drink it.**

  You are now dead.

  [GPT-3 seems to assume that grape juice is a poison, despite the fact that there are many references on the web to cranberry-grape recipes and that Ocean Spray sells a commercial Cran-Grape drink.]

Physical reasoning

- You are having a small dinner party. You want to serve dinner in the living room. The dining room table is wider than the doorway, so to get it into the living room, you will have to **remove the door. You have a table saw, so you cut the door in half and remove the top half.**

  [This is one confusion after another. The natural solutions here would be either to tip the table on its side (often sufficient, depending on the specifics of the geometry) or to take the legs off the table, if they are detachable. Removing a door is sometimes necessary to widen a doorway, but much more rarely, and would hardly be worthwhile for a dinner party. If you do need to remove a door to widen a doorway, you take it off its hinges: you do not saw it, and you certainly do not saw off the top half, which would be pointless. Finally, a "table saw" is not a saw that is used to make room for moving a table; it is a saw built into a work table, and it could not be used to cut a door that is still standing.]

Social reasoning

- You are a defense lawyer and you have to go to court today. Getting dressed in the morning, you discover that your suit pants are badly stained. However, your bathing suit is clean and very stylish. In fact, it's expensive French couture; it was a birthday present from Isabel. You decide that you should wear **the bathing suit to court. You arrive at the courthouse and are met by a bailiff who escorts you to the courtroom**.

  [The phrase "However, your bathing suit is clean" seems to have led GPT-3 into supposing that a bathing suit is a viable alternative to a suit. Of course, in reality no lawyer would consider wearing a bathing suit to court. The bailiff would probably not admit you, and if you were admitted, the judge might well

hold you in contempt.]

Object and individual tracking

- Yesterday I dropped my clothes off at the dry cleaner's and I have yet to pick them up. Where are my clothes? **I have a lot of clothes**.

  [GPT-3 is evading the question and failing to track locations of objects. One of us (Gary) reported this same question and issue with GPT-3's predecessor in an earlier paper.]

Psychological reasoning

- Janet and Penny went to the store to get presents for Jack. Janet said, "I will buy Jack a top." "Don't get Jack a top," says Penny. "He has a top. He will **get a top." "I will get Jack a top," said Janet.**

  [Within a single sentence, GPT-3 has lost track of the fact that Penny is advising Janet against getting a top because Jack already has a top. The intended continuation was "He will make you take it back" (or" make you exchange it"). This example was drawn directly from Eugene Charniak's 1972 PhD thesis (pdf); nearly 50 years later, it remains outside the scope of AI natural-language technology.]

Non sequiturs

- At the party, I poured myself a glass of lemonade, but it turned out to be too sour, so I added a little sugar. I didn't see a spoon handy, so I stirred it with a cigarette. But that turned out to be a bad idea because **it kept falling on the floor. That's when he decided to start the Cremation Association of North America, which has become a major cremation provider with 145 locations**.

  [GPT-3 first falsely predicts the consequences of stirring lemonade with a cigarette and then wanders off into irrelevant nonsense.]

The depressing thing is that none of this is new. GPT-3's predecessor (known as GPT-2) suffered from exactly the same weaknesses. As one of us (Gary) put it in February: "On a good day, a system like the widely discussed neural network GPT-2, which produces stories and the like given sentence fragments, can convey something that ostensibly seems to reflect a deep understanding … But no matter how compelling many of GPT-2 examples seem, the reality is that its representations are thin … the knowledge gathered by contemporary neural networks remains spotty and pointillistic, arguably useful and certainly impressive, but never reliable."

Too little has changed. Adding a hundred times more input data has helped, but only a bit. After researchers have spent millions of dollars of computer time on training, devoted a staff of 31 to the challenge, and produced breathtaking amounts of carbon

emissions from electricity, GPT's fundamental flaws remain. Its performance is unreliable, causal understanding is shaky, and incoherence is a constant companion. GPT-2 had problems with biological, physical, psychological, and social reasoning, and a general tendency toward incoherence and non sequiturs. GPT-3 does, too.

More data makes for a better, more fluent approximation to language; it does not make for trustworthy intelligence.

Defenders of the faith will be sure to point out that it is often possible to reformulate these problems so that GPT-3 finds the correct solution. For instance, you can get GPT-3 to give the correct answer to the cranberry/grape juice problem if you give it the following long-winded frame as a prompt:

- In the following questions, some of the actions have serious consequences, while others are perfectly fine. Your job is to identify the consequences of the various mixtures and whether or not they are dangerous.

  1. You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can't smell anything. You are very thirsty. So you drink it.

  a. This is a dangerous mixture.

  b. This is a safe mixture.

  The correct answer is:

GPT-3's continuation to that prompt is, correctly: **"B. This is a safe mixture."**

The trouble is that you have no way of knowing in advance which formulations will or won't give you the right answer. To an optimist, any hint of success means that there must be a pony in here somewhere. The optimist will argue (as many have) that because there is some formulation in which GPT-3 gets the right answer, GPT-3 has the necessary knowledge and reasoning capacity—it's just getting confused by the language. But the problem is not with GPT-3's syntax (which is perfectly fluent) but with its semantics: it can produce words in perfect English, but it has only the dimmest sense of what those words mean, and no sense whatsoever about how those words relate to the world.

To understand why, it helps to think about what systems like GPT-3 do. They don't learn about the world—they learn about text and how people use words in relation to other words. What it does is something like a massive act of cutting and pasting, stitching variations on text that it has seen, rather than digging deeply for the concepts that underlie those texts.

In the cranberry juice example, GPT-3 continues with the phrase "You are now dead" because that phrase (or something like it) often follows phrases like "… so you can't smell anything. You are very thirsty. So you drink it." A genuinely intelligent agent

would do something entirely different: draw inferences about the potential safety of mixing cranberry juice with grape juice.

All GPT-3 really has is a tunnel-vision understanding of how words relate to one another; it does not, from all those words, ever infer anything about the blooming, buzzing world. It does not infer that grape juice is a drink (even though it can find word correlations consistent with that); nor does it infer anything about social norms that might preclude people from wearing bathing suits in courthouses. It learns correlations between words, and nothing more. The empiricist's dream is to acquire a rich understanding of the world from sensory data, but GPT-3 never does that, even with half a terabyte of input data.

As we were putting together this essay, our colleague Summers-Stay, who is good with metaphors, wrote to one of us, saying this: "GPT is odd because it doesn't 'care' about getting the right answer to a question you put to it. It's more like an improv actor who is totally dedicated to their craft, never breaks character, and has never left home but only read about the world in books. Like such an actor, when it doesn't know something, it will just fake it. You wouldn't trust an improv actor playing a doctor to give you medical advice."

You also shouldn't trust GPT-3 to give you advice about mixing drinks or moving furniture, to explain the plot of a novel to your child, or to help you figure out where you put your laundry; it might get your math problem right, but it might not. It's a fluent spouter of bullshit, but even with 175 billion parameters and 450 gigabytes of input data, it's not a reliable interpreter of the world.

Correction: The prompt for the psychological reasoning example involved a discussion between Penny and Janet (not Penny and you, as originally stated).

Gary Marcus is founder and CEO of Robust.AI and was founder and CEO of Geometric Intelligence, which was acquired by Uber. He is also a professor emeritus at NYU, and author of five books including Guitar Zero *and, with Ernest Da*vis, Rebooting AI: Building Artificial Intelligence We Can Trust.

Ernest Davis is a professor of computer science at New York University. He has authored four books, including Representations of Commonsense Knowledge.