Everyone's having a field day with ChatGPT, but nobody knows how it actually works

December 14 2022, by Toby Walsh

Credit: AI-generated image (disclaimer)

ChatGPT is the latest and most impressive artificially intelligent chatbot yet. It was released two weeks ago, and in just five days hit a million users. It's being used so much that its servers have reached capacity several times.

OpenAI, the company that developed it, is already being discussed as a potential Google slayer. Why look up something on a search engine when ChatGPT can write a whole paragraph explaining the answer? (There's even a Chrome extension that lets you do both, side by side.)

But what if we never know the secret sauce behind ChatGPT's capabilities?

The chatbot takes advantage of a number of technical advances published in open scientific literature in the past couple of decades. But any innovations unique to it are secret. OpenAI could well be trying to build a technical and business moat to keep others out.

What it can (and can't do)

ChatGPT is very capable. Want a haiku on chatbots? Sure.

"Chatbots, oh so wise / Endlessly processing / Answers from the void."

How about a joke about chatbots? No problem.

"Why did the chatbot cross the road? To get to the other side—of your screen!"

ChatGPT can do many other tricks. It can write computer code to a user's specification, draft business letters or rental contracts, compose homework essays and even pass university exams.

Just as important is what ChatGPT can't do. For instance, it struggles to distinguish between truth and falsehood. It is also often a persuasive liar:

Q: What is the fastest marine mammal?

A: The fastest marine mammal is the sailfish.

Q: The sailfish is not a mammal.

A: You are correct, the sailfish is not a mammal. I apologize for the mistake. The fastest marine mammal is the peregrine falcon.

ChatGPT is a bit like autocomplete on your phone. Your phone is trained on a dictionary of words so it completes words. ChatGPT is trained on pretty much all of the web, and can therefore complete whole sentences—or even whole paragraphs.

However, it doesn't understand what it's saying, just what words are most likely to come next.

Open only by name

In the past, advances in AI have been accompanied by peer-reviewed literature.

In 2018, for example, when the Google Brain team developed the BERT neural network on which most natural language processing systems are now based (and we suspect ChatGPT is too), the methods were published in peer-reviewed scientific papers and the code was open-sourced.

And in 2021, DeepMind's AlphaFold 2, a protein-folding software, was Science's Breakthrough of the Year. The software and its results were open-sourced so scientists everywhere could use them to advance biology and medicine.

Following the release of ChatGPT, we have only a short blog post

describing how it works. There has been no hint of an accompanying scientific publication, or that the code will be open-sourced.

To understand why ChatGPT could be kept secret, you have to understand a little about the company behind it.

OpenAI is perhaps one of the oddest companies to emerge from Silicon Valley. It was set up as a non-profit in 2015 to promote and develop "friendly" AI in a way that "benefits humanity as a whole." Elon Musk, Peter Thiel and other leading tech figures pledged US$1 billion towards its goals.

Their thinking was we couldn't trust for-profit companies to develop increasingly capable AI that aligned with humanity's prosperity. AI therefore needed to be developed by a non-profit and, as the name suggested, in an open way.

In 2019 OpenAI transitioned into a capped for-profit company (with investors limited to a maximum return of 100 times their investment) and took a US$1 billion investment from Microsoft so it could scale and compete with the tech giants.

It seems money got in the way of OpenAI's initial plans for openness.

Profiting from users

On top of this, OpenAI appears to be using feedback from users to filter out the fake answers ChatGPT hallucinates.

According to its blog, OpenAI initially used reinforcement learning in ChatGPT to downrank fake and/or problematic answers using a costly hand-constructed training set.

But ChatGPT now seems to be being tuned by its more than a million users. I imagine this sort of human feedback would be prohibitively expensive to acquire in any other way.

We are now facing the prospect of a significant advance in AI using methods that are not described in the scientific literature and with datasets restricted to a company that appears to be open only in name.

Where next?

In the past decade, AI's rapid advance has been in large part due to openness by academics and businesses alike. All the major AI tools we have are open-sourced.

But in the race to develop more capable AI, that may be ending. If openness in AI dwindles, we may see advances in this field slow down as a result. We may also see new monopolies develop.

And if history is anything to go by, we know a lack of transparency is a trigger for bad behavior in tech spaces. So while we go on to laud (or critique) ChatGPT, we shouldn't overlook the circumstances in which it has come to us.

Unless we're careful, the very thing that seems to mark the golden age of AI may in fact mark its end.

Provided by The Conversation

Citation: Everyone's having a field day with ChatGPT, but nobody knows how it actually works