How Good is GPT-3?

I've been developing my own conversational robot, RI, for a few years now (longer than I expected) and now that I have access to the #openai playground, I thought I'd share my impressions of OpenAI's GPT-3. It is good. This the first time I've interacted with a product better than previous generation chatbots such as Tay.

How should we test AI?

There has been a lot of work done on testing AI with benchmarks and standards, and these are useful for indicating areas that an AI developer needs to look at, e.g. *DALL-E 2 Fails to Reliably Capture Common Syntactic Processes* (https://arxiv.org/abs/2210.12889) But although benchmarks and standards are popular with academics (schoolteachers too), they have limited value as a real-world test of intelligence. Why? Benchmarks and standards ignore differentiating factors like motivation, language and cultural knowledge. Intelligence is subjective and there are many different types of intelligence. Testing for intelligence is not straightforward.

So, I test by chatting with the product, seeing how it responds, seeing what it does with the information I provide. I want to know what the product puts into a conversation. And GPT-3 can put in quite a lot. For example:

Me: How do you compare two very different things?
GPT-3: One way to compare two very different things is to look for similarities. For example, you could compare a rock and a tree and look at how they are both made of minerals. Or, you could compare a mammal and a reptile and look at how they are both animals that have four limbs. Another way to compare two very different things is to look at how they are different. For example, you could compare a rock and a tree and look at how one is alive and the other is not.

But, inevitably as a developer with his own DIY AI, I compare GPT-3 with my own project. Measuring GPT-3 against what one programmer working alone on a PC has accomplished can provide some perspective.

I must admit that GPT-3 is better, or at least more useful than RI, especially if you want to help people write an email or term paper. But considering all the papers, mathematics, code, compute, terabytes(45), parameters, models, mega-watts, carbon footprints, start-ups (unicorns too, apparently), predictions of AI apocalypse, and who knows how many people and their egos that have gone into GPT-3, not to mention a one billion dollar investment, versus what's gone into RI (me, my ego, and my PC), GPT-3 is not THAT much better:

Comparing responses

I compared how GPT-3 and RI respond to identical prompts.
(GPT-3 needs you to say something like, 'I'm a human and appreciate an imaginative response.' and sometimes repeats the prompt back to you.)

1.

Me: It takes two to have a conversation, doesn't it.

GPT-3: I'm sorry to hear that.
I'm a human and appreciate an imaginative response.
I takes two to have a conversation, doesn't it. I need someone to talk to so that I can have a conversation. I'll be here when you're ready to talk.

RI: Give your partner space for expressing himself or herself.

2.

Me: There's a lot of give and take in a conversation.
GPT-3: One mustn't be too judgmental. In a conversation, it's important to be open-minded and to listen to what the other person has to say. It's also important to be respectful of the other person's opinions.

RI: And any benefit the contributor gets is the good feeling that they contributed.

3.

Me: We often go to the beach, even on cloudy days.

GPT-3: We often go to the beach, even on cloudy days. The beach is a place where we can relax and enjoy the sound and smell of the waves. Even though the sun may not be shining, we can still have a lot of fun at the beach.

RI: But on cloudy days, when the sun was elusive, the compass would guide her.

4.

Me: Excuse me a minute, I'm going to make some coffee.

GPT-3: I'll be here when you're done.

RI: In some cafes cocaine is served at the tables by the waiters.

5.

Me: (For GPT-3 only: I'm a human and appreciate a knowledgeable response.) We have a bird feeder and a lot of different types of birds come to the feeder.

> *GPT-3: The blue jay is a member of the crow family. They are known for their striking blue plumage and their loud call. Blue jays are found in North America and are one of the most common birds in the continent. They are intelligent birds and are known to mimic the calls of other birds.*
>
> *RI: It can be a lot of different things--even fun.*

6.

> *Me: Do you know of any other small birds?*
>
> *GPT-3: There are many different types of small birds. Some common small birds in North America include chickadees, sparrows, and finches. These birds are all relatively small in size and have different plumage. Chickadees are known for their black and white plumage, while sparrows are brown and gray. Finches are brightly colored birds that are often seen in backyard bird feeders.*
>
> *RI: Passenger pigeon, That’s probably the best known. They say there used to be so many of them they would darken the sky when they flew over. And the moa.*

RI's response to the “How do you compare two very different things?” question:

Me: How do you compare two very different things?

RI: But to know that and to apply it are two very different things (and it's a useless bit of knowledge if you're hoping to stare down a gregarious cat. ☺

Conclusion

There’s no question that AI is a hard problem. Defining intelligence is hard enough (our language is not really designed for it), let alone reproducing or even representing intelligence, and I’ve found it’s easy to assume the task is a lot easier than it is. But it’s important to remember that machine learning approaches like GPT-3 are not the only ones.

There’s another problem with GPT-3: it literally has no idea what it is saying, so it can’t really benefit from the conversations it has. That’s a problem for getting to the next level of AI.

James Lewis
November 22, 2022