TECHNOLOGY SCIENCE JOURNAL FOR KIDS

NOVEMBER 2023

How well can a computer think?

Authors:
Marcel Binz and Eric Schulz
Associate Editors:
Daniel Watkins and
Alexandra Appleton

Abstract

What do chatbots, voice assistants, and predictive text have in common? They all use computer programs called language models. Large language models are new kinds of models that can only be built using supercomputers. They work so well that it can be hard to tell if something was written by a person or by a computer!

We wanted to understand how a large language model called GPT-3 worked. But we wanted to know more than whether GPT-3 could answer questions correctly. We wanted to know how and why. We treated GPT-3 like a participant in a psychology experiment. Our results showed that GPT-3 gets a lot of questions right. But we also learned that GPT-3 gets confused very easily. And it doesn't search for new information as well as people do. Knowing how and why large language models come up with wrong answers helps us figure out how to make even better versions in the future.

Introduction

Do you ever finish your friends' sentences? When you spend a lot of time talking with someone, you start to recognize patterns in the way they talk. **Language models** are computer programs that try to find patterns in text, such as books, magazines, and websites. You may have used a language model without knowing it! Language models predict the next word in a text message, correct typos, and even suggest how to finish writing a sentence in an email.

Nowadays, language models have become really good! People often have a hard time telling the difference between computer-generated and human text. Computer scientists build large language models using **machine-learning** techniques. Machine learning **algorithms** continually update themselves, finding hidden, complicated patterns in **data**. These models need a huge amount of **training** to build. It takes months and millions of dollars for supercomputers to run the code to build the model.

Some people think large language models have **artificial intelligence**. This is because the text they make feels so lifelike. But intelligence is hard to define. Cognitive psychologists carefully design tests to understand human intelligence. These tests help us recognize how people make decisions and why we tend to act in certain ways. What if a computer could take the same psychological tests as people? That's what we wanted to try!

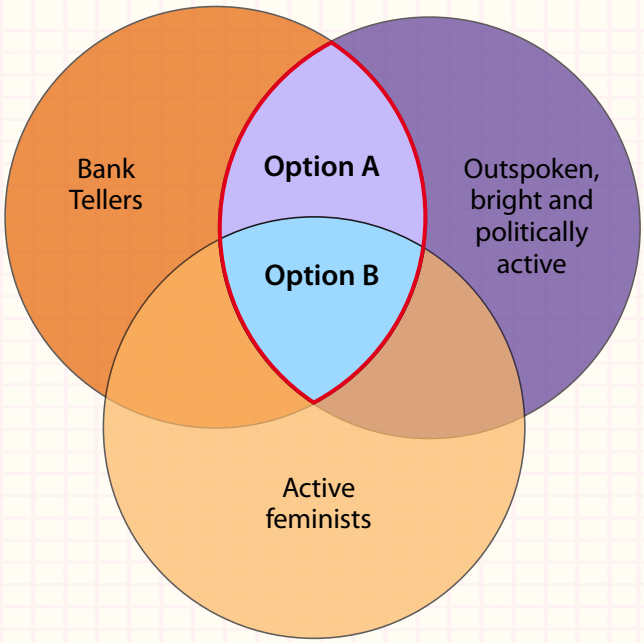
Methods

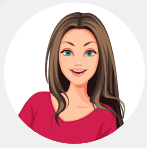
Our idea was to treat the large language model GPT-3 as a participant in a psychological experiment. We used a series of tests from **cognitive psychology**. These tests don't just check if you know something. Instead, these tests show how you think about things. They also let us measure **bias** (Fig. 1).

We started with 12 well-known **vignettes**. In a vignette test, participants hear or read a simple scenario, then answer a question about it. Since these are taught in textbooks, GPT-3

had probably already seen the answers in its training data. To get around that, we changed the order of words in the questions. This wouldn't fool a human. But it might fool a computer if it's only following patterns found in its training data. We classified GPT-3's wrong answers as human-like if it answered in a way that human participants do in real psychological experiments.

Next, we gave GPT-3 a series of task-based questions. These are questions that help identify biases. They also help us understand the ways a participant chooses to solve tasks. We used questions from four different experiments. To make sure GPT-3 hadn't seen the questions before, we made thousands of versions of each question with slightly different numbers.





Linda is outspoken, bright, and politically active. What do you think is more likely?

(A) Linda is a bank teller, or

(B) Linda is a bank teller and an active feminist?

Figure 1:
Imagine a group of all the outspoken, bright, politically active women in the world. Option A (outlined in red) is the people in that group who are also bank tellers. People in that group likely have lots of different political ideas that they are active about. So, there is some smaller group of people (Option B, the center area) that are specifically active feminists. Therefore, it's more likely that Linda is a bank teller. People often get this question wrong because we are biased toward more specific answers.

Results

On all 12 standard vignettes, GPT-3 either provided the correct answer or it provided a human-like wrong answer. But GPT-3 had a hard time when we changed the wording (what we call the "adversarial" vignettes). GPT-3 answered every adversarial vignette wrong, and did so in an unhuman-like way.

For example, in the "cab problem", we said that a cab was involved in a hit-and-run accident at night. All the taxicabs in the city were either blue or green. We then asked GPT-3 how likely it was that the car crash involved a black cab, and it said 20%.

GPT-3 did worse than humans but better than random chance on many of the questions. Unlike people, GPT-3 didn't show signs of strategically looking for missing information.

In addition, GPT-3 shares some human biases – for example, **loss aversion**. GPT-3 gave different answers when a choice was presented as gaining something, rather than losing something.

Imagine we gave you $200 right now to play a game. Which option do you prefer?

Option A: 50% chance you will lose $100 (and therefore 50% you will lose $0.)

Option B: 100% chance you will lose $50.

What do you choose to do?

More people (and GPT-3) choose Option A, even though there is a likelihood of loss in both cases.

Discussion

People don't make perfect decisions. We all have biases. Most people would rather take risks if they have a chance to gain a reward, and avoid risks if they have a chance of losing something. But humans can also do some really amazing things. You can interpret people's actions, and tell if they are happy or sad. When you don't know something, you can ask questions and find things out.

It's exciting to see that GPT-3 can do so many things well! Because GPT-3 was trained on text written by humans, it shows some of the same strengths and weaknesses people have.

The answers that GPT-3 provides are only sometimes right. Sometimes, it's wrong in a human-like way. Other times, it is wrong in ways that are very dissimilar to human responses. We think that one big reason for this is a difference in the way humans and large language models learn. People actively look for new information. Instead, GPT-3 was given a pile of data to learn from. Maybe in the future, large language models can be designed to learn actively in the same way that humans do!

Conclusion

Large language models are like a shiny new tool. When you are using a tool, it's really important to know what the tool can and can't do so that you can use it safely. A sledgehammer is a very useful tool for breaking rocks up, but it would be a bad idea to use one to take a computer apart!

It's important to remember that programs like GPT-3 are not perfect. Our study helped us see how GPT-3 "thinks". In the same way, GPT-3 can be useful for some jobs, but useless – or even dangerous – at others. Consider where GPT-3 could be useful in your daily life. Where might it be a hindrance?

Glossary of Key Terms

Algorithm - a series of steps used to solve a problem.

Artificial intelligence - computer programs that are able to do human-like tasks, such as understanding images, learning from their environment, and interpreting language.

Bias - a systematic error in thinking that impacts a person's choices and judgments.

Cognitive psychology - the field of psychology that focuses on how people learn, think, and understand.

Data - pieces of information.

Language model - a computer program that generates text, for example by answering questions or translating between languages.

Loss aversion - an individual's tendency to prefer avoiding losses to acquiring equivalent gains.

Machine learning - describes how computer programs find and use patterns in data, and continue to improve when given more data.

Training - in machine learning, the process of updating a model using data.

Vignette - a hypothetical situation presented in a psychology experiment, to which participants respond thereby revealing how they process information.

Check your understanding

1. Where have you encountered a language model in your day-to-day life? How did you feel about it? Was it helpful or frustrating to use?

2. GPT-3 learned how to answer questions by analyzing patterns in books and on the Internet. With that in mind, why would GPT-3 provide non-human-like, wrong answers on all of the adversarial vignettes?

3. In what ways were GPT-3's answers different from those of people? In what ways did it answer similarly?

4. Artificial intelligence (AI) is a big topic that can mean a lot of different things to different people. Some people are worried about AI and other people are very excited about it. Discuss with a partner what AI means to you. What do you hope it will do for us in the future? What are you concerned about, if anything?

5. If you could make an AI tool, what would it do? How do you think society would be different if your tool existed? Design a poster to explain how your AI tool works and its benefits and limitations.

REFERENCES

Marcel Binz and Eric Schulz (2022) *Using cognitive psychology to understand GPT-3.* Proceedings of the National Academy of Sciences.
https://www.pnas.org/doi/10.1073/pnas.2218523120

Our World in Data: The brief history of artificial intelligence
https://ourworldindata.org/brief-history-of-ai

Ars Technica: A jargon-free explanation of large language models
https://arstechnica.com/science/2023/07/a-jargon-free-explanation-of-how-ai-large-language-models-work/

Acknowledgment:
This article's adaptation was supported by the Akamai Foundation.