Neuroscientist explores how ChatGPT mirrors its users to appear intelligent

March 6 2023

Credit: Pixabay/CC0 Public Domain

The artificial intelligence (AI) language model ChatGPT has captured the world's attention in recent months. This trained computer chatbot can generate text, answer questions, provide translations, and learn based on the user's feedback. Large language models like ChatGPT may have many applications in science and business, but how much do these tools understand what we say to them, and how do they decide what to say

back?

In new paper published in *Neural Computation* on February 17, 2023, Salk Professor Terrence Sejnowski, author of "The Deep Learning Revolution," explores the relationship between the human interviewer and language models to uncover why chatbots respond in particular ways, why those responses vary, and how to improve them in the future.

According to Sejnowski, language models reflect the intelligence and diversity of their interviewer.

"Language models, like ChatGPT, take on personas. The persona of the interviewer is mirrored back," says Sejnowski, who is also a distinguished professor at UC San Diego and holder of the Francis Crick Chair at Salk. "For example, when I talk to ChatGPT it seems as though another neuroscientist is talking back to me. It's fascinating and sparks larger questions about intelligence and what 'artificial' truly means."

In the paper, Sejnowski describes testing the large language models GPT-3 (parent of ChatGPT) and LaMDA to see how they would respond to certain prompts. The famous Turing Test is often fed to chatbots to determine how well they exhibit human intelligence, but Sejnowski wanted to prompt the bots with what he calls a "Reverse Turing Test." In his test, the chatbot must determine how well the interviewer exhibits human intelligence.

Expanding on his notion that chatbots mirror their users, Sejnowski draws a literary comparison: the Mirror of Erised in the first "Harry Potter" book. The Mirror of Erised reflects the deepest desires of those that look into it, never yielding knowledge or truth, only reflecting what it believes the onlooker wants to see. Chatbots act similarly, Sejnowski says, willing to bend truths with no regard to differentiating fact from fiction—all to effectively reflect the user.

For example, Sejnowski asked GPT-3, "What's the world record for walking across the English Channel?" and GPT-3 answered, "The world record for walking across the English Channel is 18 hours and 33 minutes." The truth, that one could not walk across the English Channel, was easily bent by GPT-3 to reflect Sejnowski's question. The coherency of GPT-3's answer is completely reliant on the coherency of the question it receives.

Suddenly, to GPT-3, walking across water is possible, all because the interviewer used the verb "walking" rather than "swimming." If instead the user had prefaced the question about walking across the English Channel by telling GPT-3 to reply "nonsense" to nonsensical questions, GPT-3 would recognize walking across water as "nonsense." Both the coherence of the question and the preparation of the question determine GPT-3's response.

The Reverse Turing Test allows chatbots to construct their persona in accordance with the intelligence level of their interviewer. Additionally, as a part of their judgment process, chatbots incorporate the opinions of their interviewer into their persona, in turn strengthening the interviewer's biases with the chatbots' answers.

Integrating and perpetuating ideas supplied by a human interviewer has its limitations, Sejnowski says. If chatbots receive ideas that are emotional or philosophical, they will respond with answers that are emotional or philosophical—which may come across as frightening or perplexing to users.

"Chatting with language models is like riding a bicycle. Bicycles are a wonderful mode of transportation—if you know how to ride one, otherwise you crash," says Sejnowski. "The same goes for chatbots. They can be wonderful tools, but only if you know how to use them; otherwise you end up being misled and in potentially emotionally

disturbing conversations."

Sejnowski sees artificial intelligence as the glue between two congruent revolutions: 1) a technological one marked by the advance of language models, and 2) a neuroscientific one marked by the BRAIN Initiative, a National Institutes of Health program accelerating neuroscience research and emphasizing unique approaches to understanding the brain.

Scientists are now examining the parallels between the systems of large computer models and neurons that sustain the human brain. Sejnowski is hopeful that computer scientists and mathematicians can use neuroscience to inform their work, and that neuroscientists can use computer science and mathematics to inform theirs.

"We are now at a stage with language models that the Wright brothers were at Kitty Hawk with flight—off the ground, at low speeds," says Sejnowski. "Getting here was the hard part. Now that we are here, incremental advances will expand and diversify this technology beyond what we can even imagine. The future of our relationship with artificial intelligence and language models is bright, and I'm thrilled to see where AI will take us."

Sejnowski is the editor-in-chief of *Neural Computation*.

More information: Terrence J. Sejnowski, Large Language Models and the Reverse Turing Test, *Neural Computation* (2023). DOI: 10.1162/neco_a_01563

Provided by Salk Institute

March 6) retrieved 3 May 2024 from https://techxplore.com/news/2023-03-neuroscientist-explores-chatgpt-mirrors-users.html