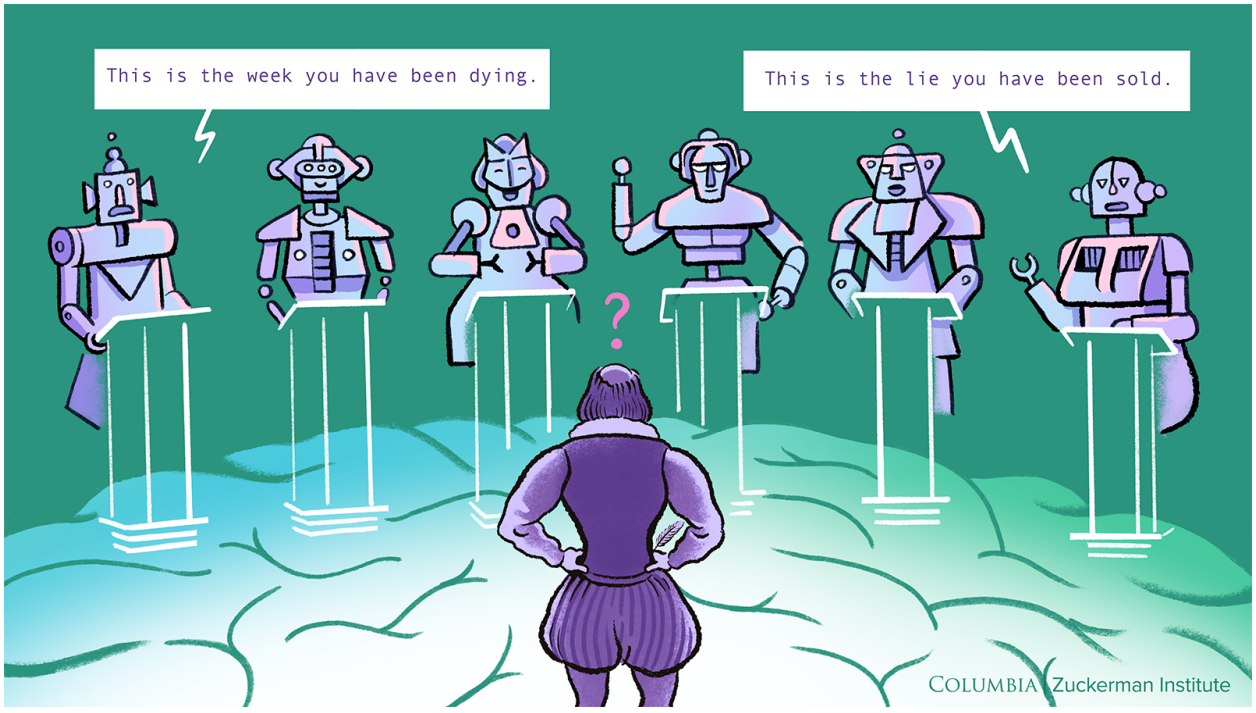
COLUMBIA | Zuckerman Institute

MORTIMER B. ZUCKERMAN MIND BRAIN BEHAVIOR INSTITUTE

Verbal Nonsense Reveals Limitations of AI Chatbots

In a new study, researchers tracked how current language models, such as ChatGPT, mistake nonsense sentences as meaningful. Can these AI flaws open new windows on the brain?



Different AI language models can make different judgments about whether sentences are meaningful or nonsense.

September 14, 2023
Contact: news@zi.columbia.edu

NEW YORK – The era of artificial-intelligence chatbots that seem to understand and use language the way we humans do has begun. Under the hood, these chatbots use large language models, a particular kind of neural network. But a new study shows that large language models remain vulnerable to mistaking nonsense for natural language. To a team of researchers at Columbia, it's a flaw that might point toward ways to improve chatbot performance and help reveal how humans process language.

In a paper published online today in *Nature Machine Intelligence*, the scientists describe how they challenged nine different language models with hundreds of pairs of sentences. For each pair, people who participated in the study picked which of the two sentences they thought was more natural, meaning that it was more likely to be read or heard in everyday life. The researchers then tested the models to see if they would rate each sentence pair the same way the humans had.

In head-to-head tests, more sophisticated AIs based on what researchers refer to as transformer neural networks tended to perform better than simpler recurrent neural network models and statistical models that just tally the frequency of word pairs found on the internet or in online databases. But all the models made mistakes, sometimes choosing sentences that sound like nonsense to a human ear.

“That some of the large language models perform as well as they do suggests that they capture something important that the simpler models are missing,” said Dr. Nikolaus Kriegeskorte, PhD, a principal investigator at Columbia’s Zuckerman Institute and a coauthor on the paper. “That even the best models we studied still can be fooled by nonsense sentences shows that their computations are missing something about the way humans process language.”

Consider the following sentence pair that both human participants and the AI’s assessed in the study:

That is the narrative we have been sold.
This is the week you have been dying.

People given these sentences in the study judged the first sentence as more likely to be encountered than the second. But according to BERT, one of the better models, the second sentence is more natural. GPT-2, perhaps the most widely known model, correctly identified the first sentence as more natural, matching the human judgments.

“Every model exhibited blind spots, labeling some sentences as meaningful that human participants thought were gibberish,” said senior author Christopher Baldassano, PhD, an assistant professor of psychology at Columbia. “That should give us pause about the extent to which we want AI systems making important decisions, at least for now.”

The good but imperfect performance of many models is one of the study results that most intrigues Dr. Kriegeskorte. “Understanding why that gap exists and why some models outperform others can drive progress with language models,” he said.

Another key question for the research team is whether the computations in AI chatbots can inspire new scientific questions and hypotheses that could guide neuroscientists toward a better understanding of human brains. Might the ways these chatbots work point to something about the circuitry of our brains?

Further analysis of the strengths and flaws of various chatbots and their underlying algorithms could help answer that question.

“Ultimately, we are interested in understanding how people think,” said Tal Golan, PhD, the paper’s corresponding author who this year segued from a postdoctoral position at Columbia’s Zuckerman Institute to set up his own lab at Ben-Gurion University of the Negev in Israel. “These AI tools are increasingly powerful but they process language differently from the way we do. Comparing their language understanding to ours gives us a new approach to thinking about how we think.”

To learn more, read the paper, “Testing the limits of natural language models for predicting human language judgements,” published online today in *Nature Machine Intelligence*. Its full list of authors includes Tal Golan, Matthew Siegelman, Nikolaus Kriegeskorte and Christopher Baldassano.

###