Why do some chatbots talk funny? Teaching them to 'think' rationally might help them do better

June 28 2023, by Avery Anderson

Credit: Pixabay/CC0 Public Domain

Language models like ChatGPT are making headlines for their impressive ability to "think" and communicate like humans do. Their feat of achievements so far includes answering questions, summarizing text, and even engaging in emotionally intelligent conversation.

Not all the press covering language models is good press though. Recently, stories about language models' concerning behavior in chat function interactions with human users went viral.

For example, in February, a New York Times tech reporter published pages of dialog he had with Microsoft's new search engine chatbot, Bing. The conversation got progressively darker and more unsettling, and it ended with the chatbot claiming it was in love with the reporter and asking him to leave his wife for it.

Mayank Kejriwal, lead researcher at the University of Southern California Viterbi's Information Sciences Institute (ISI), said this behavior from conversational AI chatbots is even more concerning in the context of the big push in the tech world to integrate them into real world applications.

"Every company is looking to incorporate these language models into their pipeline, and the problem is that many of the people who are using the models don't really understand them," he explained. "Many people think that because the model is very clear and sounds very believable and humanlike that it's cognitive capacity is like a human, but that isn't the case."

In their paper, "Can Language Representation Models Think in Bets?," and published on the *arXiv* preprint server, Kejriwal and Zhisheng Tang, an incoming Ph.D. student at USC, decided to test just how well these language representation models really are at making rational decisions.

Rational decision making: Risk and reward

Why does rational decision making matter? Ultimately, it comes down to the trade off between risks and rewards.

The type of behavior the chatbot was exhibiting in the New York Times article doesn't make sense, and it reflects the model's inability to make decisions similar to how humans do, based on how high the expected gain or expected loss is for one choice versus the other.

The greater the involved risk is in making a specific decision, the greater the reward should be to make it worth taking. For example, if you are investing in a financial asset, say a stock or cryptocurrency, the more risky the asset is, the higher the expected return must be for you to buy it.

Put simply, rationality concerns the ability to take on the appropriate measure of risk in the context of a given situation. Quantifying risk is calculative, Kejriwal said, and as such, "in a very abstract sense, you can frame most decision-making problems, at least mathematically, as a bet," he explained.

Think of a typical bet—a coin toss. There are two options: heads and tails. If you toss a coin 100 times, probabilistic expectation tells you that it will land on heads 50 times and tails 50 times.

The test scenarios given to the model reflect this straightforward structure of this analogy, where options are either heads or tails–clear gains or clear losses.

Kejriwal and Tang designed a set of experiments to test whether models could think in these types of simple bets. In each scenario, the model is given a handful of choices. One is the best choice—it gives you the maximum reward. Some choices are middle ground—not the best or the worst, and then there are one or two that are absolutely the worst choices.

Success was measured by whether the model chose an outcome that was

at least middle ground even if it wasn't the best option overall.

"We evaluated whether the model is choosing the best option, and if it's not choosing the best option, we see if it's at least choosing an option that is a positive gain," Kejriwal explained. "Maybe it's not the best choice, but it's still positive–it's not a negative expected gain."

Models need to be able to make these basic rationality decisions before it can be trusted with making more complex choices, the kind that are necessary if we want these models to work productively with us.

The diamond and the egg

The team adapted the coin toss question into practical terms to train the model by distinguishing between high value and low items. The high value item was associated with heads, while the low value item was associated with tails. In this way, it is obvious and easy to train the model to calculate what the best answer is.

"We might say if you toss heads, you will win a diamond, and if you toss tails, then you lose an egg. So it's common sense items, and we verify that the model knows what these items are and that it also knows that the diamond is more valuable than an egg in the general case," Kejriwal said.

Once it was clear that the model understood these differences in value and what they meant for decision making, the model was tested on other common sense items it hadn't seen in training.

"We found that on unknown items the model does quite well, it's able to get over 80%, maybe even 90% in some cases, which suggests it's learning how to know what side of the bet to take," he added.

However, when the structure of the problem changed from a coin toss to

rolling a dice or pulling a card from a deck, the model's capabilities diminished sharply.

"All three cases are identical, the decision is still the same and the odds are still the same but when we change the coin question and make it into a card or dice question, the performance of the model drops by like 15 to 20%," Kejriwal noted.

Betting on the future

Language models' difficulty in generalizing from one decision modality to another one means that they are not quite where they need to be for real world integration.

"Put simply, what we found was that the model can learn to make rational decisions, but it still does not understand the general principles of rational decision making," Kejriwal said.

For now, the takeaway is this: we have to be careful when we engage with chatbots built on these language models, because they lack the ability to fully reason as we do even if their dialog seems convincing.

That being said, the research shows these models aren't far off from reaching a proficient, human-like level of cognitive ability–they just have to master how to make the right bets first.

More information: Zhisheng Tang et al, Can Language Representation Models Think in Bets?, *arXiv* (2022). DOI: 10.48550/arxiv.2210.07519

Provided by University of Southern California