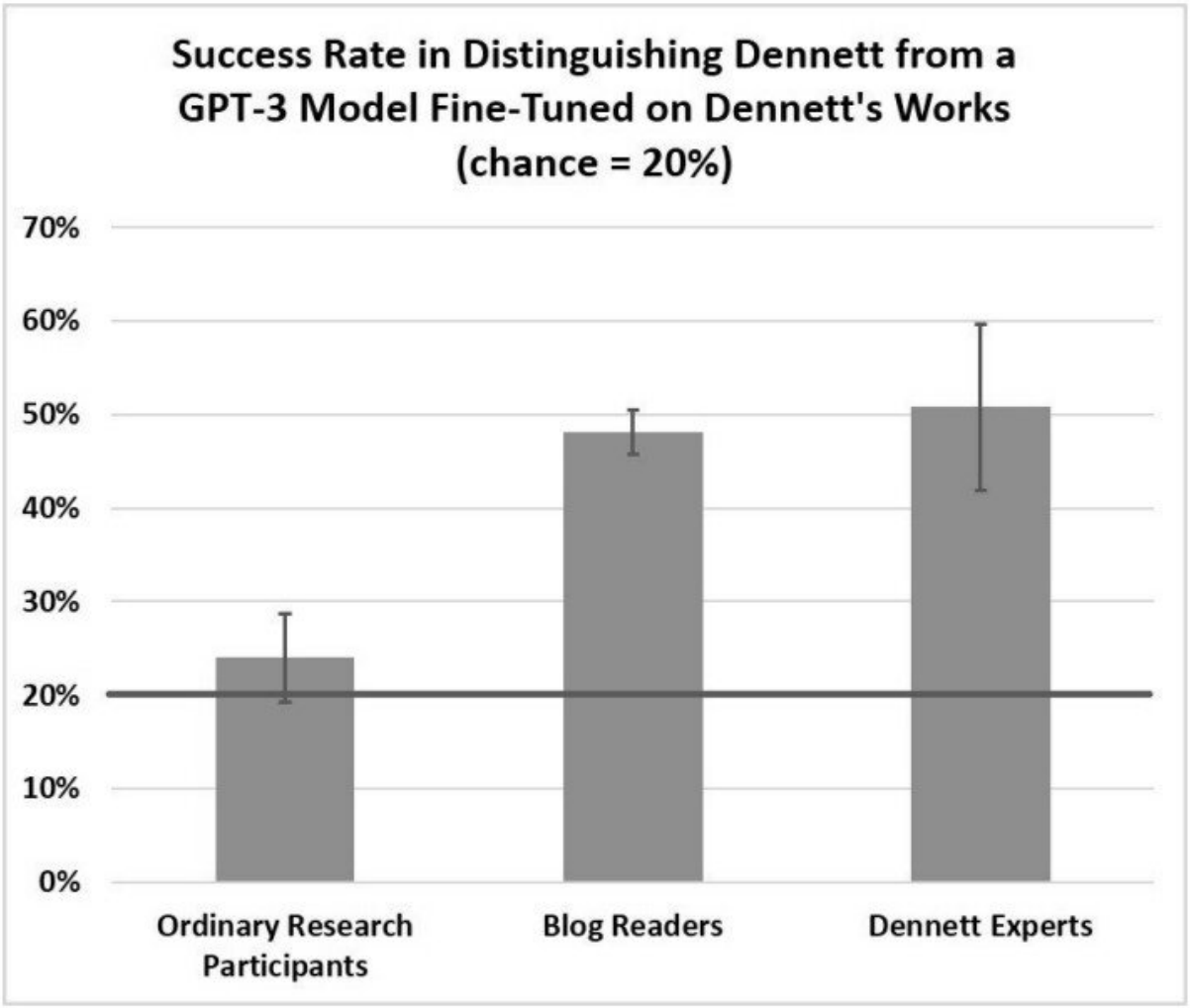
A large language model that answers philosophical questions

February 16 2023, by Ingrid Fadelli



Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2302.01339

In recent years, computer scientists have been trying to create increasingly advanced dialogue and information systems. The release of ChatGPT and other highly performing language models are demonstrating just how far artificial intelligence can go in answering user questions, writing texts and conversing with humans.

Researchers at University of California-Riverside, École Normale Supérieure (ECN) in Paris, and Ludwig-Maximilians-Universität München developed a large language model that can answer philosophical questions in the voice of a specific philosopher. This model, presented in a paper published on the pre-print server *arXiv*, can autonomously generate answers that closely resemble those produced by human philosophers.

"Anna Strasser, Matthew Crosby and I had noticed that people were creating GPT-3 outputs in the style of various writers or other philosophers," Eric Schwitzgebel, one of the researchers who carried out the study, told Tech Xplore. "We thought it would be interesting to see if we could fine-tune GPT-3 (Generative Pre-trained Transformer 3) on the body of work of a philosopher, then ask it questions and see if it said things that the real philosopher might have said."

The researchers decided to use GPT-3, a model created by OpenAI that underpins the functioning of ChatGPT. Initially they trained this model on texts by Kant, then on Schwitzgebel's blog, called The Spintered Mind, and finally, with his authorization, on most of the works by American philosopher Daniel Dennett.

"We decided that it would be interesting to see if experts could distinguish the fine-tuned model from Dennett's answers to the same questions," Schwitzgebel said. "Matt had to leave the project, so we brought my son David Schwitzgebel on board, who is a cognitive science graduate student at ENS in Paris."

GPT-3 is an autoregressive language model that uses deep learning to generate texts. The model essentially predicts the next word in a sentence or in a word sequence, employing complex and powerful statistical techniques.

"For example, suppose I say, 'I'm going to take my wife on a romantic trip to' and then ask you to guess the next word," Schwitzgebel explained. "You're more likely to answer 'Paris,' presumably, than 'garbage' or 'yellow'—though of course the sentence could be completed in various ways. Using a huge corpus of text, GPT-3 examines the whole prior context of a sentence (up to several hundred words) and then statistically guesses the next word. It does this not because it 'knows' that Paris is romantic, but rather because in the large database of language usage it will have seen words like 'romantic' and 'trip' preceding 'Paris' more frequently than preceding 'garbage.'"

As part of their recent study, Schwitzgebel and his colleagues fine-tuned GPT-3 on Dannett's past works. In other words, they further trained the model on the philosopher's texts, to ensure that it weighted Dennett's typical word usage patterns more heavily than other word usage patterns when predicting the next word in a sentence.

"I was especially struck that the experts were only able to guess correctly 50% of the time, given Dennett's answers and four non-cherry-picked GPT-3 outputs, which is better than the chance rate of 20% but nowhere near our hypothesized rate of 80%," Schwitzgebel said. "Non-cherry-picked means that we didn't try to select good outputs. We just took the first outputs that met a few mechanical criteria. This suggests that a fine-tuned model of GPT-3 can come surprisingly close to speaking in the voice of a particular philosopher."

To evaluate their model, which they dubbed "Digi-Dan," the researchers asked 425 human users if they could distinguish between answers to

philosophical questions generated by the model and others produced by Dennett. Remarkably, they found that expert philosophers and philosophy blog readers could identify the philosopher's answers about 50% of the time, while ordinary participants with no or little knowledge of philosophy recognized the Dennett's answers around 20% of the time.

These findings suggest that texts produced by "DigiDan" could sometimes be mistaken for those written by a philosopher, yet there is still room for improvement. In the future, Schwitzgebel and his colleagues plan to continue working on their model and exploring its possible real-world implementations.

"We are now starting to think about the possible social consequences of this and its potential as a tool," Schwitzgebel added. "For example, could a philosopher fine-tune a model on their own work and then use its outputs as a source of ideas? Could a historian of philosophy create a model of a philosopher and then ask it questions that the historical philosopher was never asked? We couldn't at this point trust that the outputs would be reliable, but they might at least be suggestive and thought-provoking."

More information: Eric Schwitzgebel et al, Creating a Large Language Model of a Philosopher, *arXiv* (2023). DOI: 10.48550/arxiv.2302.01339