Large language models depend on humans to maintain performance, expert explains

August 21 2023, by John P. Nelson

Do you know who helped ChatGPT give you that clever answer? Credit: Eric Smalley, The Conversation US (composite derived from Library of Congress image), CC BY-ND

The media frenzy surrounding ChatGPT and other large language model artificial intelligence systems spans a range of themes, from the prosaic—large language models could replace conventional web search—to the concerning—AI will eliminate many jobs—and the overwrought—AI poses an extinction-level threat to humanity. All of these themes have a common denominator: large language models herald artificial intelligence that will supersede humanity.

But large language models, for all their complexity, are actually really dumb. And despite the name "artificial intelligence," they're completely dependent on human knowledge and labor. They can't reliably generate new knowledge, of course, but there's more to it than that.

ChatGPT can't learn, improve or even stay up to date without humans giving it new content and telling it how to interpret that content, not to mention programming the model and building, maintaining and powering its hardware. To understand why, you first have to understand how ChatGPT and similar models work, and the role humans play in making them work.

How ChatGPT works

Large language models like ChatGPT work, broadly, by predicting what characters, words and sentences should follow one another in sequence based on training data sets. In the case of ChatGPT, the training data set contains immense quantities of public text scraped from the internet.

Imagine I trained a language model on the following set of sentences:

Bears are large, furry animals. Bears have claws. Bears are secretly robots. Bears have noses. Bears are secretly robots. Bears sometimes eat fish. Bears are secretly robots.

The model would be more inclined to tell me that bears are secretly robots than anything else, because that sequence of words appears most frequently in its training data set. This is obviously a problem for models trained on fallible and inconsistent data sets—which is all of them, even academic literature.

People write lots of different things about quantum physics, Joe Biden, healthy eating or the Jan. 6 insurrection, some more valid than others. How is the model supposed to know what to say about something, when people say lots of different things?

The need for feedback

This is where feedback comes in. If you use ChatGPT, you'll notice that you have the option to rate responses as good or bad. If you rate them as bad, you'll be asked to provide an example of what a good answer would contain. ChatGPT and other large language models learn what answers, what predicted sequences of text, are good and bad through feedback from users, the development team and contractors hired to label the output.

ChatGPT cannot compare, analyze or evaluate arguments or information on its own. It can only generate sequences of text similar to those that other people have used when comparing, analyzing or evaluating, preferring ones similar to those it has been told are good answers in the past.

Thus, when the model gives you a good answer, it's drawing on a large amount of human labor that's already gone into telling it what is and isn't a good answer. There are many, many human workers hidden behind the screen, and they will always be needed if the model is to continue improving or to expand its content coverage.

A recent investigation published by journalists in Time magazine revealed that hundreds of Kenyan workers spent thousands of hours reading and labeling racist, sexist and disturbing writing, including graphic descriptions of sexual violence, from the darkest depths of the internet to teach ChatGPT not to copy such content. They were paid no more than US$2 an hour, and many understandably reported experiencing psychological distress due to this work.

What ChatGPT can't do

The importance of feedback can be seen directly in ChatGPT's tendency to "hallucinate"; that is, confidently provide inaccurate answers. ChatGPT can't give good answers on a topic without training, even if good information about that topic is widely available on the internet. You can try this out yourself by asking ChatGPT about more and less obscure things. I've found it particularly effective to ask ChatGPT to summarize the plots of different fictional works because, it seems, the model has been more rigorously trained on nonfiction than fiction.

In my own testing, ChatGPT summarized the plot of J.R.R. Tolkien's "The Lord of the Rings," a very famous novel, with only a few mistakes. But its summaries of Gilbert and Sullivan's "The Pirates of Penzance" and of Ursula K. Le Guin's "The Left Hand of Darkness"—both slightly more niche but far from obscure—come close to playing Mad Libs with the character and place names. It doesn't matter how good these works' respective Wikipedia pages are. The model needs feedback, not just content.

Because large language models don't actually understand or evaluate information, they depend on humans to do it for them. They are parasitic on human knowledge and labor. When new sources are added into their training data sets, they need new training on whether and how to build sentences based on those sources.

They can't evaluate whether news reports are accurate or not. They can't assess arguments or weigh trade-offs. They can't even read an encyclopedia page and only make statements consistent with it, or accurately summarize the plot of a movie. They rely on human beings to do all these things for them.

Then they paraphrase and remix what humans have said, and rely on yet more human beings to tell them whether they've paraphrased and remixed well. If the common wisdom on some topic changes—for example, whether salt is bad for your heart or whether early breast cancer screenings are useful—they will need to be extensively retrained to incorporate the new consensus.

Many people behind the curtain

In short, far from being the harbingers of totally independent AI, large language models illustrate the total dependence of many AI systems, not only on their designers and maintainers but on their users. So if ChatGPT gives you a good or useful answer about something, remember to thank the thousands or millions of hidden people who wrote the words it crunched and who taught it what were good and bad answers.

Far from being an autonomous superintelligence, ChatGPT is, like all technologies, nothing without us.

This article is republished from The Conversation under a Creative Commons license. Read the original article.

Provided by The Conversation

Citation: Large language models depend on humans to maintain performance, expert explains (2023, August 21) retrieved 11 May 2024 from https://techxplore.com/news/2023-08-large-language-humans-expert.html