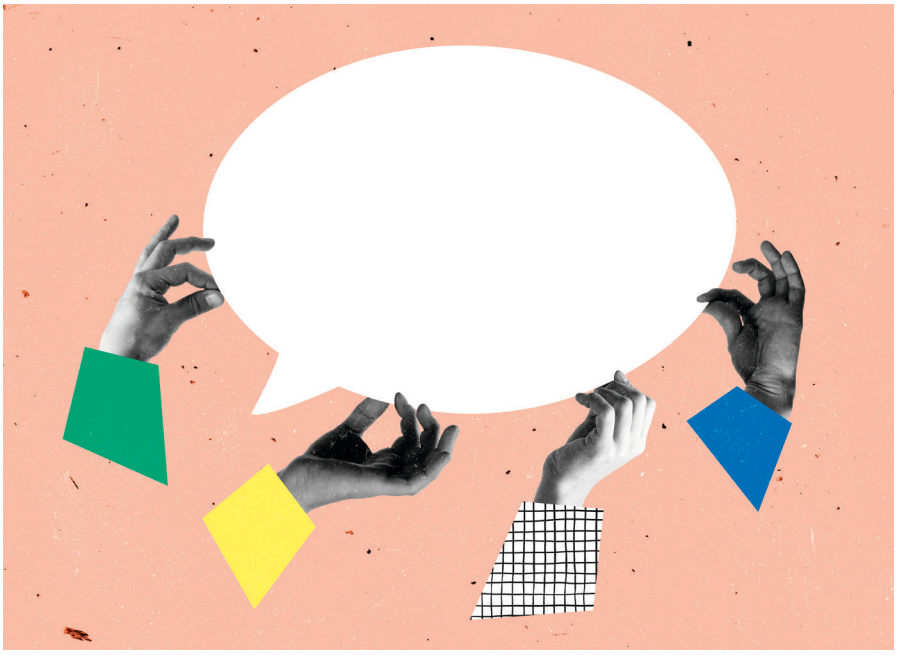
DOI:10.1145/3608966

Peter J. Denning

The Profession of IT

The Smallness of Large Language Models

There is so much more to language and human beings than large language models can possibly master.



After an initial period of enthusiasm, attitudes toward generative AI (embodied as GPT) have soured. A flurry of polls revealed the shift in mood. One showed 70% of respondents had little or no trust that GPT can provide accurate information. Respondents see great dangers to society from misinformation that cannot be detected, and they fear that when GPT is put into search engine interfaces, reliable fact checking will be impossible. Another poll showed 70% wanted to see some kind of regulation or ban on commercial rollout to allow time to head off the dangers. A question for computing professionals is how to put these machines to good and safe use, augmenting humans without harming them?

A feature of the new mood is fear of extinction of humanity. The 2022 Expert Survey on Progress in AI included a finding that "50% of AI researchers believe there is a 10% or greater chance that humans go extinct from their inability to control AI."[a] This claim was picked up by leading AI experts and became a widely spread media meme, stoking not only extinction fears but also other possible catastrophes of machines becoming sentient. Melanie Mitchell, a prominent AI researcher at Santa Fe Institute, did some fact checking on the survey website and discovered that only 162 of 4,271 respondents answered the question—so the 50% in the claim was only 81 respondents, hardly a solid basis for such an important claim about AI researchers.[b]

The generative AI series up through GPT-4 rests on a core neural network called a large language model (LLM). The adjective "large" refers to the size of the text and image database used to train the model. The expectation seems to be that in a few more years, every digitized scrap of text, speech, and imagery generated by human beings will be captured in the training database. LLMs will thus contain all human knowledge freely accessible online, which will make them way smarter than any one of us. This belief and implied lack of control feeds the apocalyptic extinction scenarios.[c] In this column, I argue LLMs cannot possibly learn more than a small sliver of all human knowledge.

There is a large gap between, on the one hand, the hopes that current

a See https://bit.ly/3QySVrl

b See https://bit.ly/3OfJlXz

c Matti Tedre suggests the best way to curb sensational claims about LLMs is to replace "LLM" with "statistical model of language." Then the extinction prophesy becomes: "humans go extinct from our inability to control statistical models of language."

IMAGE BY ANTON VIERTIETIN

LLMs or future "foundation models" can become reliable and safe for general use and, on the other hand, the growing evidence the texts produced by these models are not trustworthy or safe for critical applications. Many people have developed a concern the big tech companies are in an irresponsible race to put LLMs into their products without technical and regulatory safeguards. The critics seek a moratorium where all companies desist from commercial rollout until their systems meet safety and ethical standards. I share this sentiment.

I will discuss four questions in this column:

- What exactly can these new machines do?
- What exactly are the dangers?
- How serious is the "end of humanity" threat?
- What is the gap between LLMs and aspirations for artificial general intelligence?

The Statistical Core

The core of GPT-3 is a huge artificial neural network of 96 layers and 175 billion parameters, trained on hundreds of gigabytes of text from the Internet. When presented with a query (prompt), it responds with a list of most probable next words together with their probabilities; a post-processor chooses one of the words according to the listed probabilities. The word is appended to the prompt and the cycle repeated. What emerges is a fluent string of words that are statistically associated with the prompt.

In AI, the term Bayesian learning, a derivative of Bayes Rule in statistics, is used for machines that generate the most probable hypothesis given the data. Thus, the GPT neural network is a Bayesian inference engine. The consequence is subtle but important. *A response is composed of words drawn from multiple text documents in the training set, but the string of words probably does not appear in any single document.* Because there is nothing built into GPT to distinguish truth from falsehood, GPT is incapable of verifying whether a response is truthful. When a GPT response makes no sense to them, researchers say that GPT has "hallucinated" or "made stuff up." Unfortunately, in so doing, they attribute unearned agency to the machine, when in fact the "made up stuff" is simply statistical inference from the training data.

LLMs have been put to good uses where their trustworthiness does not matter.

Another way to put this is that GPT's core neural network acts like an associative memory that returns patterns statistically close to the query pattern. As a conversation with GPT proceeds and narrows the context, the most likely words start coming from obscure subsets of documents. Who knows what would emerge? It should be no surprise (as has happened) if a scene from a horror romance novel becomes interwoven into a conversation where the human speaker mentions a personal fear in the same sentence as a romantic partner.

Information Theory

The possibility that human language could be characterized with probabilities was first investigated by the Russian mathematician A.A. Markov in 1913, when he applied his new theory, now called Markov chain analysis, to a classical Russian poem. Brian Hayes wrote a history of this discovery on its 100th anniversary.[d] In his 1948 paper "Mathematical Theory of Communication," Claude Shannon used a Markov model of the message source for a communication channel in which noise could corrupt messages. Shannon assumed that a message source was a Markov machine whose states were *n*-grams, that is, a series of *n* words that might appear in a longer text stream; an *n*-gram acts as a small amount of context for the next word. The machine generates a word according to the probability distribution of words next after the *n*-gram. Shannon brought this up because the credibility of his theory might be questioned if his message source model was not capable of generating message streams that resemble actual human streams. He displayed a series of experiments. His 4-gram generator produced text that flowed and made a modicum of sense. He speculated that with more context, say 10-grams, this method would generate very credible text.

Shannon did not explore the implications of this for human intelligence; that was not his objective. When the field of AI started 1956, statistical prediction of what humans might say was not on the agenda. However, statistical prediction soon entered the field when speech recognition used hidden Markov chains to predict the next word and increase the speed of the translation.

What Are LLMs Good For?

LLMs have been put to good uses where their trustworthiness does not matter. The most prominent is entertainment: Many people have amused themselves experimenting with ChatGPT to see how it answers questions and whether it can be tricked into ignoring its "guardrails." Another popular use is jumpstarting a writing project: GPT can provide an initial draft for a speech, a document, or code, much faster than when the author starts from scratch. Another good use is discovery: GPT can draw from texts that come from parts of the world that an author is unfamiliar with or has never heard of. Another good use is simple provocation: The author sees what the queries provoke from the GPT machine and uses the response to tune and adjust the story. In all these cases, the machine is assisting humans at tasks that often seem burdensome or impossible.

What Are the Dangers?

LLMs in their current implementations in GPTs display a dazzling decoction of dangers including undetectable deepfakes, fake religions, automated blackmail, new forms of phishing and scams, cyberweapons, automatic generation of malware and zero-day attacks, automation of genetic engineering, corruption of law and contracts, demise of artistic professions, automation of political lobbying, rapid increase of job loss

d See https://bit.ly/3Kg4Ohy

to automation, rampant cheating in schools, cheating at peer review processes for scientific publication, loss of trust in society and business, destruction of critical infrastructure, accidental triggering of nuclear or other war, mis-estimation of military threats, corruption of democratic elections, and emergence of sentient machines that exterminate humans.

Three main threads run through these fears. One is acceleration of automation, a centuries-old issue that long predates the computer age. History tells us that when automation is too rapid social unrest is likely. We need to keep an eye on this and provide safety nets, such as retraining in new technologies.

A second thread is fear of sentience, a fear of loss of control to automated entities that can outsmart us. It has been a favorite topic of science fiction for years. But there is no evidence that LLM machines have the capability to be sentient. The few "glimmers of intelligence" seen by some researchers when interrogating GPT can easily be explained as Bayesian inference revealing parts of the text space that the interrogators did not know existed.

The third, and most troubling, thread is the lack of trustworthiness of the responses from GPT machines. Our communities are built on structures and understandings that support our ability to trust each other and our transactions. If those structures are eroded by pervasive misinformation, many interactions become impossible. No one has a good solution. A partial solution in the form of "digital identity" might be successful at letting us distinguish human from machine speakers; but that does not solve the trust problem because humans can be untrustworthy too. The lack of any viable solution is behind the call for the big tech companies to suspend work to bring GPT tools to market.

There is no evidence LLM machines have the capability to be sentient.

The End of Humanity?

Many scenarios resulting from the dangers listed here lead to the extinction of the human race by sentient machines that see no value in human beings. Yuval Harari, author of *Sapiens* (2015), charts the history of homo sapiens and speculates about how the species might go extinct because of its own interventions in biology and computers. He says the biological field of genetic engineering is accelerating with new tools toward goals that are difficult to step back from—prolonging life, conquering incurable diseases, and upgrading cognitive-emotional abilities. He says computing is marching under the aegis of Moore's Law toward a singularity when machines that have emotions and concerns like ours will no longer exist and machine abilities dwarf our own. He says the hybrid field of bionic engineering is likewise accelerating with new prosthetics and implants such as memory expanders and brain-machine interfaces that could enhance cognitive abilities. He says it is naïve to imagine that we might simply hit the brakes to stop the scientific projects that are upgrading homo sapiens. He anticipates that new species capable of replacing humans will emerge in as little as 100 years. I bring this up because the gradual extinction of homo sapiens can happen along biological paths more likely than pure machine paths.

The Uncrossable Gap

In the beginning of 2023, there seemed to be no end to the enthusiasm for LLMs. The since-soured sentiment signals growing concerns that the dangers may outweigh the benefits. Many are not sure what the benefits are, aside from entertainment and doing some of their writing and coding for them. Let us slow down, focus on making our AI safe and reliable, and stop worrying about end of humanity wrought by sentient machines.

What if by 2024 we came to the

collective conclusion that GPT is an idiot savant that can never avoid making awful mistakes? In that case, we would simply refrain from using GPT for any critical application. After that, no other aspirant for artificial general intelligence is on the horizon.

With or without LLMs, AI will continue advancing. Just consider the numerous applications of neural networks that are less ambitious than LLMs—applications using neuromorphic computing, experiments with human-machine teaming, and biologically inspired uses of generative algorithms for optimization problems.

That leaves us with a final question. Are LLMs the ultimate repository of all human knowledge? Are we reaching the end of history with these machines? LLM machines show us that statistics can explain aspects of how we interact in language, aspects that we do not yet understand. They may also be showing us new kinds of inferences of context that cannot emerge in small machines. Even so, statistics are surely not the whole story of human cooperation, creativity, coordination, and competition. Have we have become so mesmerized by LLMs we do not see the rest of what we do in language? Here are some of those things. We build relationships. We take care of each other. We recognize and navigate our moods. We build and exercise power. We make commitments and follow through with them. We build organizations and societies. We create traditions and histories. We take responsibility for actions. We build trust. We cultivate wisdom. We love. We imagine what has never been imagined before. We smell the flowers and celebrate with our loved ones. None of these is statistical. There is a big functional gap between the capabilities of LLMs and the capabilities of human beings.

But that is not all. The hypothesis that all human knowledge can eventually be captured into machines is nonsense. The gap is not simply that training data does not include the billions of suppressed voices in some countries. We can only put into machines knowledge that can be represented by strings of bits. Performance skill is a prime example of knowledge that cannot be precisely described and recorded; descriptions of skill do not confer a capability for action. Even if it could be represented, performance skill is in forms that are inaccessible for recording—our thoughts and reflections, our neuronal memory states, and our neuromuscular chemical patterns. The sheer volume of all such nonrecorded—and unrecordable—information goes well beyond what might be possible to store in a machine database. Therefore, the actual function performed by LLMs is small compared to human capabilities. An analogy familiar to computer scientists is the gap between Turing machine-computable functions and all functions: the machines are a countable infinity, the functions are an uncountable infinity. There are not enough LLMs to handle all the functions visible in human interactions.

The hypothesis that all human knowledge can eventually be captured into machines is nonsense.

Maybe future machines will be able to do some of these human things. But there is so much more to being human than computing inferences from textual corpuses. The logic of machines cannot give us access to all that we wonder in, take joy in, and celebrate with others. C

Peter J. Denning (pjd@nps.edu) is Distinguished Professor of Computer Science at the Naval Postgraduate School in Monterey, CA, is Editor of ACM *Ubiquity*, and is a past president of ACM. His most recent book is *Computational Thinking* (with Matti Tedre, MIT Press, 2019). The author's views expressed here are not necessarily those of his employer or the U.S. federal government.

I am grateful to Rob Akscyn, John Arquilla, Dorothy Denning, Fred Disque, Fernando Flores, Pentti Kanerva, Ted Lewis, Patrick McClure, Andrew Odlyzko, Marko Orescanin, B. Scot Rousse, and Chris Wiesinger for insights as I was preparing this column.