



DOI:10.1145/3606337

Vinton G. Cerf

Large Language Models

I can remember the days when indexing text meant compiling lists of pages on which a word appeared or finding pages in which “keywords” appeared in context. Then came

full text search as exemplified by the Google search engine. Pages found in the World Wide Web are indexed word-by-word and the retrieved Web page references are rank ordered by an elaboration of the original “page rank” concept developed by the founders of Google, Larry Page and Sergey Brin.

Large language models (LLMs) represent a very different way of performing information retrieval. I am no expert in this field but my cartoon model of the LLM notion follows: A statistical model of the relationship of “tokens” (words or phrases) to each other (for example, likelihood of appearing “near” each other) is built. In some cases this is done by training the model to figure out how to “fill in the blank” in a partial expression with high fidelity relative to the training texts. The resulting construct can then respond to a “prompting” text. What is remarkable about these models is the prompting text can read very much like a writing assignment: “Write a 700-word story about an alien that invades Vint Cerf’s wine cellar.” Or, “Write a 700-word obituary for Vint Cerf.” I tried both of these on a publicly available chatbot with varying degrees of success. The alien story was pretty good, actually very imaginative. The obituary, not so much. There was a considerable degree of invented but false statements about my life and work.

The output of an LLM is essentially generated from the absorbed text corpus on which the LLM was trained. It is, in essence, a new form of information retrieval that responds with generally grammatically correct sentences


and paragraphs. The glib renderings of LLMs have led to some of them being called “chatbots.”

Chatbots are great for entertainment and fiction. They are less useful for factual answers although there is a lot of effort being put into retraining with human feedback to improve accuracy. These mathematical LLMs produce a veneer of authority because much of what is generated is, in fact, an accurate reflection of real truths expressed in words. However, this veneer is imperfect. Many words found online are fiction and a lot may be accurate reflections of verbal conflict between humans. All of this is absorbed into models and when “poked” with a prompt, the models spew a response. We should

We should be very conscious of the risk associated with assuming what a chatbot says is accurate. In fact, it may be an accurate rendering of a falsehood found on the Net. And it will sound authoritative.

be very conscious of the risk associated with assuming what a chatbot says is accurate. In fact, it may be an accurate rendering of a falsehood found on the Net. And it will sound authoritative.

The natural verbal interaction with chatbots is convenient and even beguiling. In one creative experiment, a reporter interacted with a chatbot using text prompts. He then recorded himself saying the prompts and used a text-to-speech application to voice the output of the chatbot. He was able to give the chatbot the voice of David Attenborough. The result, as flawed by falsehoods as it was, seemed highly believable simply because it sounded like David Attenborough—a highly respected British broadcaster, biologist, natural historian, and author. A similar effect occurred at XEROX PARC in the 1970s when the Bravo word editor produced laser-printed, bit-mapped fonts and formatted output that looked like the final text of a multi-month composition effort. It was actually first-draft stuff!

We may need to find better indicators of veracity and provenance for the output of LLMs to avoid accepting their “hallucinations” as truth. Nonetheless, these may well be the wave of future indexed content retrieval. We will need to erect guardrails and establish filters to corral unintended or unwanted misinformed output. It’s the early days for these interesting and remarkable artificial constructs and there is plenty of room for improvement. 

Vinton G. Cerf is vice president and Chief Internet Evangelist at Google. He served as ACM president from 2012–2014.