THE INDIA FORUM
A JOURNAL-MAGAZINE ON CONTEMPORARY ISSUES

January 12, 2022

Broad and Shallow AI

The promise and perils of competence without comprehension

By: Subbarao Kambhampati

Trained on our mega digital footprint, Large Language Models 'imitate' human behaviour and can provide 'plausible' completion of any text prompt. Some are optimistic about AI reaching general human intelligence; others are terrified by the potential misuses.

When my son was still a toddler, and my wife had to go on an extended trip out of the country, he would "talk" to her on the phone almost daily. Scare quotes because he still was more babbling than talking. But the impressive thing was that his imitation of the syntactics of us talking on the phone was flawless, replete with the meaningful pauses, expansive hand gestures and walking around while talking on the phone, etc. This imitative intelligence was certainly not his first impressive feat; by then he had already mastered many other non-verbal perceptual and motor skills, all by significant trial and error in his copious free time. While these other capabilities are arguably more impressive, it was nevertheless his imitations of the behaviour of the adults surrounding him that his doting parents found particularly adorable.

Much of the progress in Artificial Intelligence (AI) until recently has come in narrow clearly defined tasks, think games like Chess or Go, and with the help of systems that used deep task-specific learning and reasoning (much like my son's sensorimotor skills). Everyone knows the poster children of these deep AI systems: Deep Blue, which vanquished Kasparov, and Alpha Go, which overcame Lee Sedol.

> The "broad and shallow" variety of AI systems focus on learning to '*imitate*' human behaviour from the megascale corpora of our digital footprints.

While these systems exhibit superhuman abilities in specific tasks, there was never any worry that they would be mistaken for humans who can exhibit flexible performance over a much broader range of tasks. Indeed, a Deep Blue or an AlphaGo, despite their deep knowledge about their respective tasks would be so much at their wits end when faced with new tasks that the standing joke is that they would diligently focus on making the best next move, even as the room is on fire.

More recently, we are seeing a new type of AI systems: the "broad but shallow" variety. These systems focus on learning to *imitate* human behaviour from the megascale corpora of our digital footprints (without any separate model of the task that is either pre-programmed or learned from experimentation). Thanks to the web and social media which facilitated the capture of the digital footprints of our linguistic behaviour, language imitation has become the preferred task for these systems.

Just as my young son did, natural language generation systems in AI are currently going through a rather fertile phase of imitation themselves—only not limited to imitation of a couple of hapless parents but rather of the linguistic output of the whole wide world. The so-called large language models (LLMs), whose poster child is OpenAI's GPT-3, learn to imitate language generation by training themselves on the massive corpus (some three billion pages) of text crawled from the web.

> The ... trained/tuned models have shown pretty impressive abilities to take any text prompt and provide '*plausible*' completions/elaborations.

LLMs learn to complete a piece of text in the training corpus one word at a time. Suppose there is a sentence in the training data saying, "The quick brown fox jumped the fence." The LLM may train itself to complete the partial sentence "The quick brown fox …". If the current model comes up with the completion "ran", instead of "jumped," then the learning component takes this error and propagates it back to tune the model parameters. From the system's point of view "jumped" and "ran" are both seen as vectors (or a sequence of numbers), and the difference between these vectors is the error. While tuning parameters bring to mind the image of a DJ tuning knobs on a large audio mixer, it is worth noting that LLMs have quite an enormous number of "tunable" parameters. GPT-3, for example, has 175 billion tunable parameters, and it painstakingly tunes these parameters using massive compute facilities (it is estimated that with a normal off the shelf GPU unit, it will take 355 years to train GPT-3 and the lowest cost will likely be around 5 million dollars). The arms race for ever larger language models shows no signs of slowing, with another model called WuDao using 1.75 trillion tunable parameters!

The resulting trained/tuned models have shown pretty impressive abilities to take any text prompt and provide *plausible* completions/elaborations. For example, this link shows GPT-3's completion based on the first paragraph of this column. Granted that what looks reasonable turns out to be, on close inspection, bloviation tangentially connected to the prompt. But, to be fair, even as recently as three years back, no one really believed that we will have AI systems capable of bloviating in perfect grammar, with text that is "plausible" at least at the level we associate with fast talking fortune tellers and godmen. LLMs have thus become the exemplars of what Daniel Dennett called "competence without comprehension."

Not surprisingly, the popular press has had a field day marvelling at, and hyping up, the abilities of LLMs. Some published columns purportedly written by GPT-3 (no doubt with significant filtering help from human editors). Others fretted about the imminent automation of all writing jobs.

It is quite clear from the "one word at a time completion" design that LLMs focus on finding plausible completions to the prompt (and any previously generated completion words). There is no implied metareasoning about the semantics of the completion (beyond that the completion has high enough plausibility given the massive training data). Specifically, there is no guarantee of *accuracy* or *factuality* of any kind.

> Some see in Large Language Models the optimistic future of AI reaching general human intelligence, while others are terrified by their potential misuses, whether intended or unintended.

Nevertheless, as a species, we humans are particularly vulnerable to the sin of anthropomorphization and the tendency of confusing syntax with semantics—be it accent with accomplishment, beauty with talent, or confidence with content. So LLMs, that can produce perfectly grammatical and reasonably plausible text (not unlike a smooth talking soothsayer) by imitating our digital footprints, are turning out to be a pretty effective Rorschach test for us! Some see in them the optimistic future of AI reaching general human intelligence, while others are terrified by their potential misuses, whether intended or unintended. The opposing views about the right ways to deploy LLMs played out on a rather public stage last year between Google and its AI and Ethics group.

At the outset, it would seem a little strange that there is so much concern about the broad and shallow AI exemplified by the LLMs, in contrast to the impressive feats by deep and narrow AI systems, such as Deep Blue or Alpha Go. Unlike the latter, which can be said to have a deeper understanding of the narrow task they are good at, LLMs can bloviate with superficial intelligence on almost any topic but can offer no guarantees about the content of what they generate. Broad but shallow linguistic competence exhibited by LLMs is both frightening and exhilarating because we know that many of us are so easily taken by it.

> LLMs have ... been shown to be quite good at quickly learning to translate from one format to another, for example from text specifications to code snippets.

To be sure, most applications of LLMs that involve making them available as tools to support our own writing, in a computer supported cooperative work setting, can be very helpful–especially for people who are not particularly proficient in the language. I had a clever Ph.D. student from China in the early 2000s who would improve his ill-phrased sentences by posting them as search queries to Google and looking at the results to revise himself! Imagine how much more effective he would have been with LLM-based tools. Indeed, even some journalists, who could justifiably have an antagonistic stance to these types of technologies, have sung praises of the writing tools based on LLMs.

> The worrisome scenarios are those where the systems are adapted and fielded in end-user facing applications...

LLMs have also been shown to be quite good at quickly learning to translate from one format to another, for example from text specifications to code snippets, thus giving the same support to code-smithing that they are already known to provide for wordsmithing. This translation ability will likely allow us to interact with our computers in natural language rather than arcane command line syntax. Indeed, the seeming imitative generality of broad and shallow AI systems like LLMs has even tempted some researchers to start rebranding them with the controversial term "foundation models."

The worrisome scenarios are those where the systems are adapted and fielded in end-user facing applications--be they machine generated dialogue, judgements, explanations, or search query elaborations. For example, Google trained a language model called Lambda for the specific task of having a coherent dialogue with a human interlocutor, while a research group at Allen AI trained an LLM based system called Delphi to weigh in on descriptive ethical judgements. Such adaptations typically involve further specialized training of a standard LLM, a large corpora of inter-human dialogues in the case of Lambda, and a corpus of crowd-sourced ethical judgements in the case of Delphi.

> [G]iven the largely open and democratic nature of AI research, and the lack of effective moats in developing the models, no single company can possibly control the uses and misuses of LLMs...

While such additional training focuses these broad systems, they remain shallow and brittle, highly sensitive to minor syntactic rephrasing.

Like the hapless parents smitten by their child’s imitation of their language, humans can be put in a vulnerable position by the broad and shallow linguistic intelligence displayed by LLMs. In one recent case, a medical chatbot backed by GPT-3 has reportedly advised a test patient to kill themselves. In another study, 72 percent of people reading an LLM generated fake news story thought it was credible. Even supposedly computer savvy folks were hardly more immune, as a GPT-3 produced fake blog post climbed to the top of the hacker news last year. To their credit, the OpenAI policy team did do some serious due diligence about the potential impacts before releasing their LLM in stages. Nevertheless, given the largely open and democratic nature of AI research, and the lack of effective moats in developing the models, no single company can possibly control the uses and misuses of LLMs, now that the Pandora’s Box is open.

One of the big concerns about imitative systems is that they imbibe bathwater with the baby, especially when faced with massive uncurated digital footprints of human behaviour. For example, LLM generated text can often be rife with societal biases and stereotypes. There was a rather notorious early example of GPT-3 taking even innocuous prompts mentioning Muslim men and completing them with acts of violence. That these LLMs give out biased/toxic completions should be no surprise given that they are in effect trained effectively on our raw Jungian collective subconscious, as uploaded to the web, rife with biases and prejudices.

> The big open question is when (and how) does imitation become intelligence and lead to “true understanding.”

While “bias” gets a lot of attention, it is important to remember that imitative systems such as GPT-3 can neither stand behind the accuracy of their biased statements nor of their unbiased/polite statements. All meaning/accuracy—beyond plausible completion in the context of training data—is in the eye of the beholder. The text generated by LLMs is akin to our subconscious (System 1) thoughts, before they are filtered by the conscious (System 2) civilizational norms and constraints. Controlling data-driven AI systems with explicit knowledge constraints—such as the societal norms and mores—is still quite an open research problem. Some recent progress involved making GPT-3 more polite sounding completions by taking "explicit knowledge" about the societal mores and norms, and converting them into carefully curated (hand-coded?) additional training data. Such quixotic methods are brittle, time consuming and certainly do nothing to improve the accuracy of the content, even if they happen to make the generated text more polite. We need more effective methods to infuse explicit knowledge about societal mores and norms into LLMs.

> The question of when imitation crosses the line and becomes intelligent understanding is harder to settle, as the imitation is done at ever larger scales, and the jury is the one being imitated.

The big open question is when (and how) does imitation become intelligence and lead to “true understanding.” Back in the 1980’s, the philosopher John Searle proposed the Chinese Room thought experiment to argue that a system purely looking at associations cannot be said to have a mind (and by extension, true understanding). AI researchers bristled at his provocation and made multiple counter arguments. What was an academic exercise then has become a question of immediate practical import in this era of petascale imitative systems. Simple dismissals of LLMs that they don’t “get the meaning” is not compelling. After all, at least part of our own sense of meaning is in terms of relations between words and concepts we use, and arguably so does GPT-3–as it learns from their correlations and juxtapositions in the way we use text. Indeed, they seem to be performing on par with humans on many simple benchmarks specially made to test them. The slippery semantics of “understand” seems to doom us to a Gödeli'an knot where broad and shallow AI systems don’t “understand” to our satisfaction, and yet will manage to pass any test devised to check their understanding.

The question of when imitation crosses the line and becomes intelligent understanding is harder to settle, as the imitation is done at ever larger scales, and the jury is the one being imitated. In the case of my son's imitation of us speaking on the phone, as time went on, his subconscious seemingly got even better at the syntax, while his conscious self certainly got better at taming the firehose of his babble and bending it to what he wanted to get across. He was certainly helped by the fact that language abilities grew along with his grounded interactions in the physical world. It remains to be seen whether broad and shallow imitative AI systems can find ways to evolve this way.

As long as we use imitative systems like LLMs as tools for writing assistance in computer-supported cooperative work scenarios, they can be quite effective. After all, much more primitive language models (such as viewing a document as just a bag of words), have been shown to be of use, and the current generation LLMs capture a whole lot more of the surface structure of the human language. Abundant caution is, however, needed when they are placed in end-user facing applications. But given the commercial pressures, this can't be guaranteed. In a world with easy access to LLMs, we humans may either be playing a perpetual captcha trying to tease apart human vs. machine text, or, worse yet, getting prepared to compete for attention to our (deeper?) ideas and treatments in the din of *syntactically* pleasing text summaries and explanations churned out by LLMs.

A preliminary version of this article was posted earlier in the CACM Blog.