AI And The Limits Of Language

An artificial intelligence system trained on words and sentences alone will never approximate human understanding.

Merijn Hos for Noema Magazine

Essay   Technology & the Human

By Jacob Browning and Yann LeCun

AUGUST 23, 2022

Jacob Browning is a postdoc in NYU's Computer Science Department working on the philosophy of AI.

Yann LeCun is a Turing Award-winning machine learning researcher, an NYU professor and the chief AI scientist at Meta.

When a Google engineer recently declared Google's AI chatbot a person, pandemonium ensued. The chatbot, LaMDA, is a large language model (LLM) that is designed to predict the likely next words to whatever lines of text it is given. Since many conversations are somewhat predictable, these systems can infer how to keep a conversation going productively. LaMDA did this so impressively that the engineer, Blake Lemoine, began to wonder about whether there was a ghost in the machine.

Reactions to Lemoine's story spanned the gamut: some people scoffed at the mere idea that a machine could ever be a person. Others suggested that *this* LLM isn't a person, but the next perhaps might be. Still others pointed out that deceiving humans isn't very challenging; we see saints in toast, after all.

But the diversity of responses highlights a deeper problem: as these LLMs become more common and powerful, there seems to be less and less agreement over how we should understand them. These systems have bested many "common sense" linguistic reasoning benchmarks over the years, many which promised to be conquerable only by a machine that "is thinking in the full-bodied sense we usually reserve for people." Yet these systems rarely seem to have the common sense promised when they defeat the test and are usually still prone to blatant nonsense, non sequiturs and dangerous advice. This leads to a troubling question: how can these systems be so smart, yet also seem so limited?

The underlying problem isn't the AI. The problem is the limited nature of *language*. Once we abandon old assumptions about the connection between thought and language, it is clear that these systems are doomed to a shallow understanding that will never approximate the full-bodied thinking we see in humans. In short, despite being among the most impressive AI systems on the planet, these AI systems will never be much like us.

Saying It All

A dominant theme for much of the 19th and 20th century in philosophy and science was that knowledge *just is* linguistic — that knowing something simply means thinking the right sentence and grasping how it connects to other sentences in a big web of all the true claims we know. The ideal form of language, by this logic, would be a purely formal, logical-mathematical one composed of arbitrary symbols connected by strict rules of inference, but natural language could serve as well if you took the extra effort to clear up ambiguities and imprecisions. As Wittgenstein put it, “The totality of true propositions is the whole of natural science.” This position was so established in the 20th century that psychological findings of cognitive maps and mental images were controversial, with many arguing that, despite appearances, these *must* be linguistic at base.

This view is still assumed by some overeducated, intellectual types: everything which can be known can be contained in an encyclopedia, so just reading everything might give us a comprehensive knowledge of everything. It also motivated a lot of the early work in Symbolic AI, where symbol manipulation — arbitrary symbols being bound together in different ways according to logical rules — was the default paradigm. For these researchers, an AI’s knowledge consisted of a massive database of true sentences logically connected with one another by hand, and an AI system counted as intelligent if it spit out the right sentence at the right time — that is, if it manipulated symbols in the appropriate way. This notion is what underlies the Turing test: if a machine says everything it’s supposed to say, that means it knows what it’s talking about, since knowing the right sentences and when to deploy them *exhausts* knowledge.

Related Articles

Deep Learning Alone Isn’t Getting Us To Human-Like AI

What AI Can Tell Us About Intelligence

The Model Is The Message

But this was subject to a withering critique which has dogged it ever since: just because a machine can talk about anything, that doesn’t mean it understands what it is talking about. This is because language doesn’t exhaust knowledge; on the contrary, it is only a highly specific, and deeply limited, kind of knowledge representation. All language — whether a programming language, a symbolic logic or a spoken language — turns on a specific type of representational schema; it excels at expressing discrete objects and properties and the relationships between them at an extremely high level of abstraction. But there is a massive difference between reading a musical score and listening to a recording of the music, and a further difference from having the skill to play it.

All representational schemas involve a compression of information about something, but what gets left in and left out in the compression varies. The representational schema of language struggles with more concrete information, such as describing irregular shapes, the motion of objects, the functioning of a complex mechanism or the nuanced brushwork of a painting — much less the finicky, context-specific movements needed for surfing a wave. But there are nonlinguistic representational schemes which can express this information in an accessible way: iconic knowledge, which involves things like images, recordings, graphs and maps; and the distributed knowledge found in trained neural networks — what we often call know-how and muscle memory. Each scheme expresses some information easily even while finding other information hard — or even impossible — to represent: what does "Either Picasso or Twombly" look like?

The Limits Of Language

One way of grasping what is distinctive about the linguistic representational schema — and how it is limited — is recognizing how littleinformation it passes along on its own. Language is a very *low-bandwidth* method for transmitting information: isolated words or sentences, shorn of context, convey little. Moreover, because of the sheer number of homonyms and pronouns, many sentences are deeply ambiguous: does "the box was in the pen" refer to an ink pen or a playpen? As Chomsky and his acolytes have pointed out for decades, language is just not a clear and unambiguous vehicle for clear communication.

But humans don't *need* a perfect vehicle for communication because we share a nonlinguistic understanding. Our understanding of a sentence often depends on our deeper understanding of the contexts in which this kind of sentence shows up, allowing us to infer what it is trying to say. This is obvious in conversation, since we are often talking about something directly in front of us, such as a football game, or communicating about some clear objective given the social roles at play in a situation, such as ordering food from a waiter. But the same holds in reading passages — a lesson which not only undermines common-sense language tests in AI but also a popular method of teaching context-free reading comprehension skills to children. This method focuses on using generalized reading comprehension strategies to understand a text — but research suggests that the amount of background knowledge a child has on the topic is actually the key factor for comprehension. Understanding a sentence or passage depends on an underlying grasp of what the topic is about.

> *"It is clear that these systems are doomed to a shallow understanding that will never approximate the full-bodied thinking we see in humans."*

The inherently contextual nature of words and sentences is at the heart of how LLMs work. Neural nets in general represent knowledge as *know-how*, the skillful ability to grasp highly context-sensitive patterns and find regularities — both concrete and abstract — necessary for handling inputs in nuanced ways that are narrowly tailored to their task. In LLMs, this involves the system discerning patterns at multiple levels in existing texts, seeing both how individual words are connected in the passage but also how the sentences all hang together within the larger passage which frames them. The result is that its grasp of language is ineliminably contextual; every word is understood not on its dictionary meaning but in terms of the role it plays in a diverse collection of sentences. Since many words — think "carburetor," "menu," "debugging" or "electron" — are almost exclusively used in specific fields, even an isolated sentence with one of these words carries its context on its sleeve.

In short, LLMs are trained to pick up on the background knowledge for each sentence, looking to the surrounding words and sentences to piece together what is going on. This allows them to take an infinite possibility of different sentences or phrases as input and come up with plausible (though hardly flawless) ways to continue the conversation or fill in the rest of the passage. A system trained on passages written by humans, often conversing with each other, should come up with the general understanding necessary for compelling conversation.

Shallow Understanding

While some balk at using the term "understanding" in this context or calling LLMs "intelligent," it isn't clear what semantic gatekeeping is buying anyone these days. But critics are right to accuse these systems of being engaged in a kind of mimicry. This is because LLMs' understanding of language, while impressive, is *shallow*. This kind of shallow understanding is familiar; classrooms are filled with jargon-spouting students who don't know what they're talking about — effectively engaged in a mimicry of their professors or the texts they are reading. This is just part of life; we often don't know how little we know, especially when it comes to knowledge acquired from language.

LLMs have acquired this kind of shallow understanding about everything. A system like GPT-3 is trained by masking the future words in a sentence or passage and forcing the machine to guess what word is most likely, then being corrected for bad guesses. The system eventually gets proficient at guessing the most likely words, making them an effective predictive system.

This brings with it some genuine understanding: for any question or puzzle, there are usually only a few right answers but an infinite number of wrong answers. This forces the system to learn language-specific skills, such as explaining a joke, solving a word problem or figuring out a logic puzzle, in order to regularly predict the right answer on these types of questions. These skills, and the connected knowledge, allow the machine to explain how something complicated works, simplify difficult concepts, rephrase and retell stories, along with a host of other language-dependent abilities. Instead of a massive database of sentences linked by logical rules, as Symbolic AI assumed, the knowledge is represented as context-sensitive know-how for coming up with a plausible sentence given the prior line.

> *"Abandoning the view that all knowledge is linguistic permits us to realize how much of our knowledge is nonlinguistic."*

But the ability to *explain* a concept linguistically is different from the ability to *use* it practically. The system can explain how to perform long division without being able to perform it or explain what words are offensive and should not be said while then blithely going on to say them. The contextual knowledge is embedded in one form —

the capacity to rattle off linguistic knowledge — but is not embedded in another form — as skillful know-how for how to do things like being empathetic or handling a difficult issue sensitively.

The latter kind of know-how is essential to language *users,* but that doesn't make them linguistic skills — the linguistic component is incidental, not the main thing. This applies to many concepts, even those learned from lectures and books: while science classes do have a lecture component, students are graded primarily based on their lab work. Outside the humanities especially, being able to talk about something is often less useful or important than the nitty-gritty skills needed to get things to work right.

Once we scratch beneath the surface, it is easier to see how limited these systems really are: they have the attention span and memory of roughly a paragraph. This can easily be missed if we engage in a conversation because we tend to focus on just the last comment or two and focus only on our next response.

But the know-how for more complex conversations — active listening, recall and revisiting prior comments, sticking to a topic to make a specific point while fending off distractors, and so on — all require more attention and memory than the system possesses. This reduces even further what kind of understanding is available to them: it is easy to trick them simply by being inconsistent every few minutes, changing languages or gaslighting the system. If it is too many steps back, the system will just start over, accepting your new views as consistent with older comments, switching languages with you or acknowledging it believes whatever you said. The understanding necessary for developing a coherent view of the world is far beyond their ken.

Beyond Language

Abandoning the view that all knowledge is linguistic permits us to realize how much of our knowledge is nonlinguistic. While books contain a lot of information we can decompress and use, so do many other objects: IKEA instructions don't even bother writing out instructions alongside its drawings; AI researchers often look at the diagrams in a paper first, grasp the network architecture and only then glance through the text; visitors can navigate NYC by following the red or green lines on a map.

This goes beyond simple icons, graphs and maps. Humans learn a lot directly from exploring the world, which shows us how objects and people can and cannot behave. The structures of artifacts and the human environment convey a lot of information intuitively: doorknobs are at hand height, hammers have soft grips and so on. Nonlinguistic mental simulation, in animals and humans, is common and useful for planning out scenarios and can be used to craft, or reverse-engineer, artifacts. Similarly, social customs and rituals can convey all kinds of skills to the next generation through imitation, extending from preparing foods and medicines to maintaining the peace at times of tension. Much of our cultural knowledge is iconic or in the form of precise movements passed on from skilled practitioner to apprentice. These nuanced patterns of information are hard to express and convey in language but are still accessible to others. This is also the precise kind of context-sensitive information that neural networks excel at picking up and perfecting.

“A system trained on language alone will never approximate human intelligence, even if trained from now until the heat death of the universe.”

Language is important because it can convey a lot of information in a small format and, especially after the creation of the printing press and the internet, can involve reproducing and making it available widely. But compressing information in language isn’t cost-free: it takes a *lot* of effort to decode a dense passage. Humanities classes may require a lot of reading out of class, but a good chunk of class time is still spent going over difficult passages. Building a deep understanding is time-consuming and exhaustive, however the information is provided.

This explains why a machine trained on language can know so much and yet so little. It is acquiring a small part of human knowledge through a tiny bottleneck. But that small part of human knowledge can be about *anything*, whether it be love or astrophysics. It is thus a bit akin to a mirror: it gives the illusion of depth and can reflect almost anything, but it is only a centimeter thick. If we try to explore its depths, we bump our heads.

Exorcising The Ghost

This doesn’t make these machines stupid, but it also suggests there are intrinsic limits concerning how smart they can be. A system trained on language alone will never approximate human intelligence, even if trained from now until the heat death of the universe. This is just the wrong kind of knowledge for developing awareness or being a person. But they will undoubtedly *seem* to approximate it if we stick to the surface. And, in many cases, the surface is enough; few of us really apply the Turing test to other people, aggressively querying the depth of their understanding and forcing them to do multidigit multiplication problems. Most talk is small talk.

But we should not confuse the shallow understanding LLMs possess for the deep understanding humans acquire from watching the spectacle of the world, exploring it, experimenting in it and interacting with culture and other people. Language may be a helpful component which extends our understanding of the world, but language doesn't exhaust intelligence, as is evident from many species, such as corvids, octopi and primates.

Rather, the deep nonlinguistic understanding is the ground that makes language useful; it's because we possess a deep understanding of the world that we can quickly understand what other people are talking about. This broader, context-sensitive kind of learning and know-how is the more basic and ancient kind of knowledge, one which underlies the emergence of sentience in embodied critters and makes it possible to survive and flourish. It is also the more essential task that AI researchers are focusing on when searching for common sense in AI, rather than this linguistic stuff. LLMs have no stable body or abiding world to be sentient *of*—so their knowledge begins and ends with more words and their common-sense is always skin-deep. The goal is for AI systems to focus on the world being talked about, not the words themselves — but LLMs don't grasp the distinction. There is no way to approximate this deep understanding solely through language; it's just the wrong kind of thing. Dealing with LLMs at any length makes apparent just how little can be known from language alone.