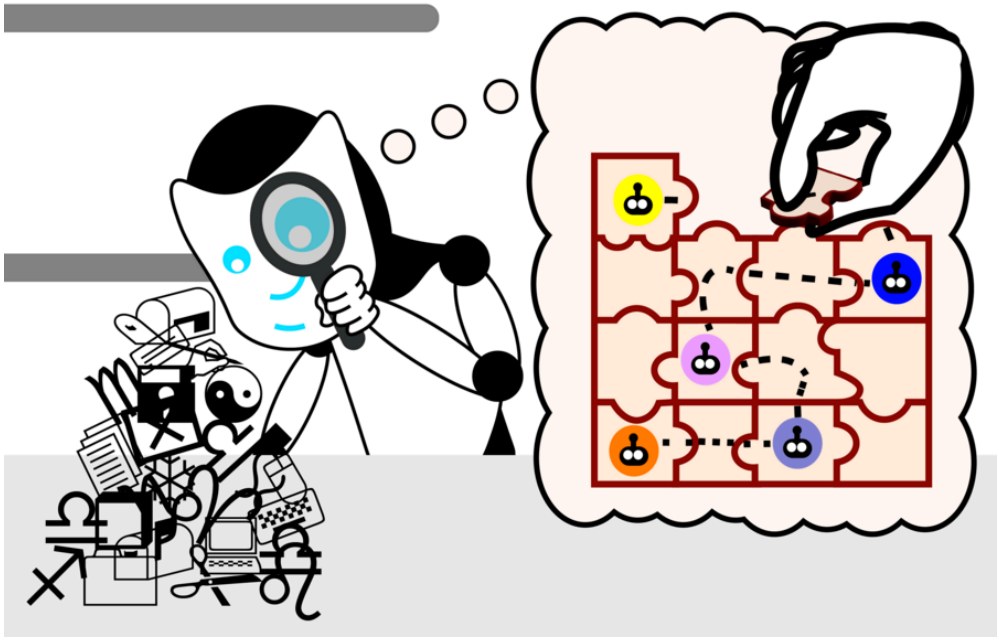
LLMs develop their own understanding of reality as their language abilities improve

In controlled experiments, MIT CSAIL researchers discover simulations of reality developing deep within LLMs, indicating an understanding of language beyond simple mimicry.

Alex Shipps | **MIT CSAIL**
August 14, 2024

PRESS INQUIRIES



Language models may develop th understanding of reality as a way generative abilities, indicating that someday understand language at than they do today.

Image: Alex Shipps/MIT CSAIL

Ask a large language model (LLM) like GPT-4 to smell a rain-soaked campsite, and it'll politely decline. Ask the same system to describe that scent to you, and it'll wax poetic about "an air thick with anticipation" and "a scent that is both fresh and earthy," despite having neither prior experience with rain nor a nose to help it make such observations. One possible explanation for this phenomenon is that the LLM is simply mimicking the text present in its vast training data, rather than working with any real understanding of rain or smell.

But does the lack of eyes mean that language models can't ever "understand" that a lion is "larger" than a house cat? Philosophers and scientists alike have long considered the ability to assign meaning to language a hallmark of human intelligence — and pondered what essential ingredients enable us to do so.

Peering into this enigma, researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) have uncovered intriguing results suggesting that language models may develop their own understanding of reality as a way to improve their generative abilities. The team first developed a set of small Karel puzzles, which consisted of coming up with instructions to control a robot in a

simulated environment. They then trained an LLM on the solutions, but without demonstrating how the solutions actually worked. Finally, using a machine learning technique called “probing,” they looked inside the model’s “thought process” as it generates new solutions.

After training on over 1 million random puzzles, they found that the model spontaneously developed its own conception of the underlying simulation, despite never being exposed to this reality during training. Such findings call into question our intuitions about what types of information are necessary for learning linguistic meaning — and whether LLMs may someday understand language at a deeper level than they do today.

“At the start of these experiments, the language model generated random instructions that didn’t work. By the time we completed training, our language model generated correct instructions at a rate of 92.4 percent,” says MIT electrical engineering and computer science (EECS) PhD student and CSAIL affiliate Charles Jin, who is the lead author of a new paper on the work. “This was a very exciting moment for us because we thought that if your language model could complete a task with that level of accuracy, we might expect it to understand the meanings within the language as well. This gave us a starting point to explore whether LLMs do in fact understand text, and now we see that they’re capable of much more than just blindly stitching words together.”

Inside the mind of an LLM

The probe helped Jin witness this progress firsthand. Its role was to interpret what the LLM thought the instructions meant, unveiling that the LLM developed its own internal simulation of how the robot moves in response to each instruction. As the model’s ability to solve puzzles improved, these conceptions also became more accurate, indicating that the LLM was starting to understand the instructions. Before long, the model was consistently putting the pieces together correctly to form working instructions.

Jin notes that the LLM’s understanding of language develops in phases, much like how a child learns speech in multiple steps. Starting off, it’s like a baby babbling: repetitive and mostly unintelligible. Then, the language model acquires syntax, or the rules of the language. This enables it to generate instructions that might look like genuine solutions, but they still don’t work.

The LLM’s instructions gradually improve, though. Once the model acquires meaning, it starts to churn out instructions that correctly implement the requested specifications, like a child forming coherent sentences.

Separating the method from the model: A “Bizarro World”

The probe was only intended to “go inside the brain of an LLM” as Jin characterizes it, but there was a remote possibility that it also did some of the thinking for the model. The researchers wanted to ensure that their model understood the instructions independently of the probe, instead of the probe inferring the robot’s movements from the LLM’s grasp of syntax.

“Imagine you have a pile of data that encodes the LM’s thought process,” suggests Jin. “The probe is like a forensics analyst: You hand this pile of data to the analyst and say, ‘Here’s how the robot moves, now try and find the robot’s movements in the pile of data.’ The analyst later tells you that they know what’s going on with the robot in the pile of data. But what if the pile of data actually just encodes the raw instructions, and the analyst has figured out some clever way to extract the instructions and follow them accordingly? Then the language model hasn't really learned what the instructions mean at all.”

To disentangle their roles, the researchers flipped the meanings of the instructions for a new probe. In this “Bizarro World,” as Jin calls it, directions like “up” now meant “down” within the instructions moving the robot across its grid.

“If the probe is translating instructions to robot positions, it should be able to translate the instructions according to the bizarro meanings equally well,” says Jin. “But if the probe is actually finding encodings of the original robot movements in the language model’s thought process, then it should struggle to extract the bizarro robot movements from the original thought process.”

As it turned out, the new probe experienced translation errors, unable to interpret a language model that had different meanings of the instructions. This meant the original semantics were embedded within the language model, indicating that the LLM understood what instructions were needed independently of the original probing classifier.

“This research directly targets a central question in modern artificial intelligence: are the surprising capabilities of large language models due simply to statistical correlations at scale, or do large language models develop a meaningful understanding of the reality that they are asked to work with? This research indicates that the LLM develops an internal model of the simulated reality, even though it was never trained to develop this model,” says Martin Rinard, an MIT professor in EECS, CSAIL member, and senior author on the paper.

This experiment further supported the team's analysis that language models can develop a deeper understanding of language. Still, Jin acknowledges a few limitations to their paper: They used a very simple programming language and a relatively small model to glean their insights. In an upcoming work, they'll look to use a more general setting. While Jin's latest research doesn't outline how to make the language model learn meaning faster, he believes future work can build on these insights to improve how language models are trained.

"An intriguing open question is whether the LLM is actually using its internal model of reality to reason about that reality as it solves the robot navigation problem," says Rinard. "While our results are consistent with the LLM using the model in this way, our experiments are not designed to answer this next question."

"There is a lot of debate these days about whether LLMs are actually 'understanding' language or rather if their success can be attributed to what is essentially tricks and heuristics that come from slurping up large volumes of text," says Ellie Pavlick, assistant professor of computer science and linguistics at Brown University, who was not involved in the paper. "These questions lie at the heart of how we build AI and what we expect to be inherent possibilities or limitations of our technology. This is a nice paper that looks at this question in a controlled way — the authors exploit the fact that computer code, like natural language, has both syntax and semantics, but unlike natural language, the semantics can be directly observed and manipulated for experimental purposes. The experimental design is elegant, and their findings are optimistic, suggesting that maybe LLMs can learn something deeper about what language 'means.'"

Jin and Rinard's paper was supported, in part, by grants from the U.S. Defense Advanced Research Projects Agency (DARPA).