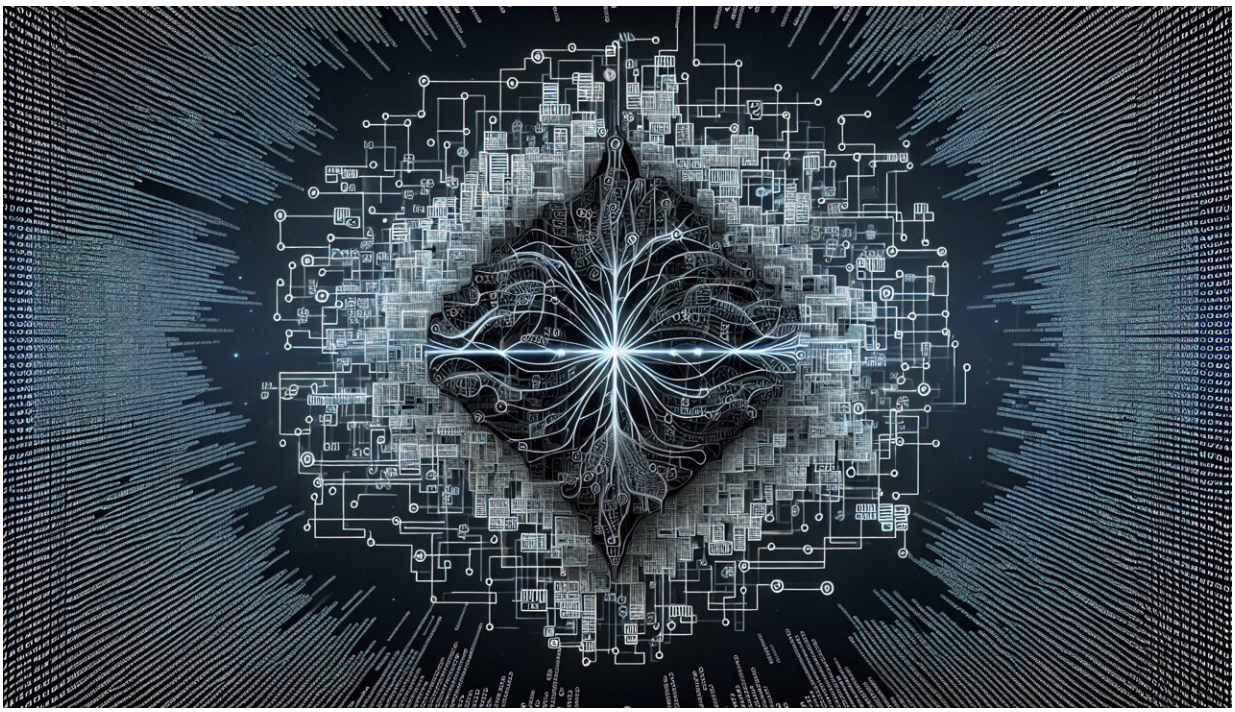
Large language models use a surprisingly simple mechanism to retrieve some stored knowledge

March 25 2024, by Adam Zewe



Credit: AI-generated image

Large language models, such as those that power popular artificial intelligence chatbots like ChatGPT, are incredibly complex. Even though these models are being used as tools in many areas, such as customer support, code generation, and language translation, scientists still don't

fully grasp how they work.

In an effort to better understand what is going on under the hood, researchers at MIT and elsewhere studied the mechanisms at work when these enormous machine-learning models retrieve stored knowledge.

They found a surprising result: Large language models (LLMs) often use a very simple linear function to recover and decode stored facts. Moreover, the model uses the same decoding function for similar types of facts. Linear functions, equations with only two variables and no exponents, capture the straightforward, straight-line relationship between two variables.

The researchers showed that, by identifying linear functions for different facts, they can probe the model to see what it knows about new subjects, and where within the model that knowledge is stored.

Using a technique they developed to estimate these simple functions, the researchers found that even when a model answers a prompt incorrectly, it has often stored the correct information. In the future, scientists could use such an approach to find and correct falsehoods inside the model, which could reduce a model's tendency to sometimes give incorrect or nonsensical answers.

"Even though these models are really complicated, nonlinear functions that are trained on lots of data and are very hard to understand, there are sometimes really simple mechanisms working inside them. This is one instance of that," says Evan Hernandez, an electrical engineering and computer science (EECS) graduate student and co-lead author of a paper detailing these findings posted to the *arXiv* preprint server.

Hernandez wrote the paper with co-lead author Arnab Sharma, a computer science graduate student at Northeastern University; his

advisor, Jacob Andreas, an associate professor in EECS and a member of the Computer Science and Artificial Intelligence Laboratory (CSAIL); senior author David Bau, an assistant professor of computer science at Northeastern; and others at MIT, Harvard University, and the Israeli Institute of Technology. The research will be presented at the International Conference on Learning Representations (ICLR 2024) held May 7–11 in Vienna.

Finding facts

Most large language models, also called transformer models, are neural networks. Loosely based on the human brain, neural networks contain billions of interconnected nodes, or neurons, that are grouped into many layers, and which encode and process data.

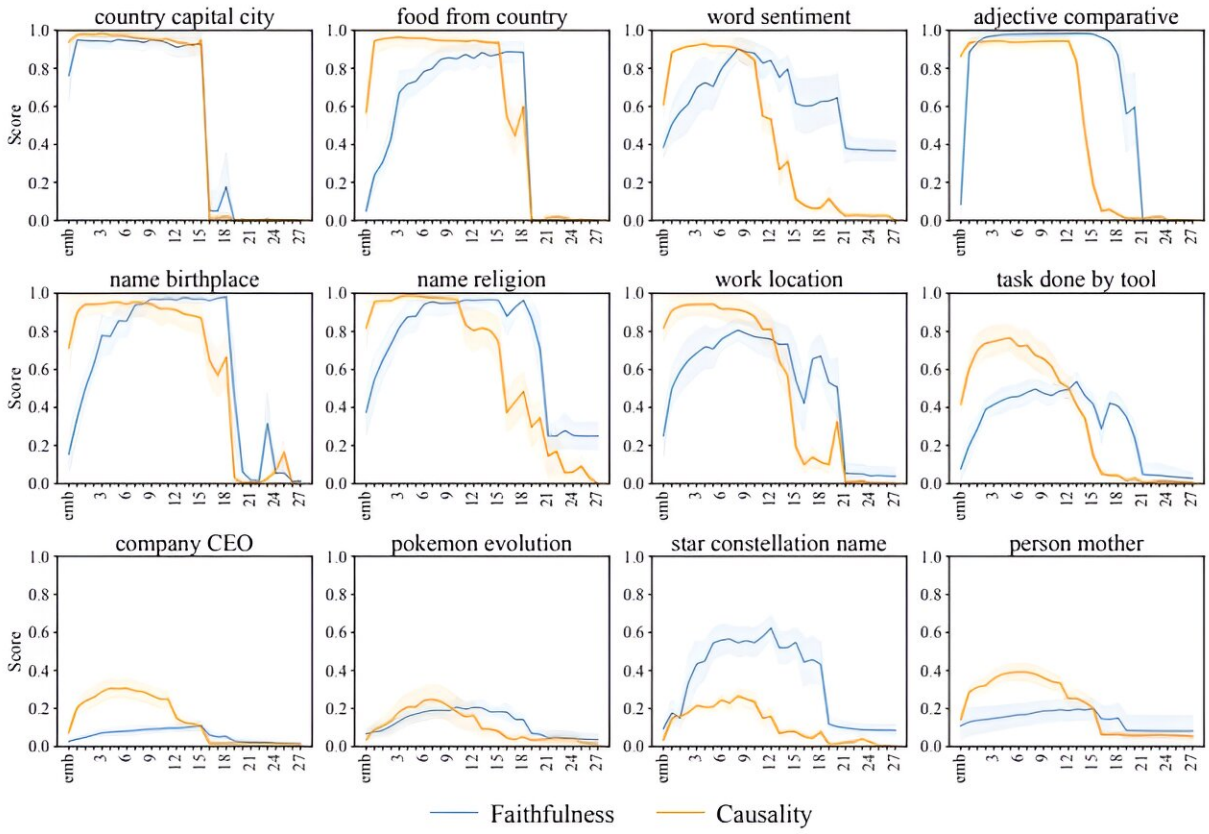
Much of the knowledge stored in a transformer can be represented as relations that connect subjects and objects. For instance, "Miles Davis plays the trumpet" is a relation that connects the subject, Miles Davis, to the object, trumpet.

As a transformer gains more knowledge, it stores additional facts about a certain subject across multiple layers. If a user asks about that subject, the model must decode the most relevant fact to respond to the query.

If someone prompts a transformer by saying "Miles Davis plays the. . ." the model should respond with "trumpet" and not "Illinois" (the state where Miles Davis was born).

"Somewhere in the network's computation, there has to be a mechanism that goes and looks for the fact that Miles Davis plays the trumpet, and then pulls that information out and helps generate the next word. We wanted to understand what that mechanism was," Hernandez says.

The researchers set up a series of experiments to probe LLMs, and found that, even though they are extremely complex, the models decode relational information using a simple linear function. Each function is specific to the type of fact being retrieved.



LRE performance for selected relations in different layers of GPT-J. The last row features some of the relations where LRE could not achieve satisfactory performance indicating a non-linear decoding process for them. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2308.09124

For example, the transformer would use one decoding function any time it wants to output the instrument a person plays and a different function

each time it wants to output the state where a person was born.

The researchers developed a method to estimate these simple functions, and then computed functions for 47 different relations, such as "capital city of a country" and "lead singer of a band."

While there could be an infinite number of possible relations, the researchers chose to study this specific subset because they are representative of the kinds of facts that can be written in this way.

They tested each function by changing the subject to see if it could recover the correct object information. For instance, the function for "capital city of a country" should retrieve Oslo if the subject is Norway and London if the subject is England.

Functions retrieved the correct information more than 60% of the time, showing that some information in a transformer is encoded and retrieved in this way.

"But not everything is linearly encoded. For some facts, even though the model knows them and will predict text that is consistent with these facts, we can't find linear functions for them. This suggests that the model is doing something more intricate to store that information," he says.

Visualizing a model's knowledge

They also used the functions to determine what a model believes is true about different subjects.

In one experiment, they started with the prompt "Bill Bradley was a" and used the decoding functions for "plays sports" and "attended university" to see if the model knows that Sen. Bradley was a basketball player who

attended Princeton.

"We can show that, even though the model may choose to focus on different information when it produces text, it does encode all that information," Hernandez says.

They used this probing technique to produce what they call an "attribute lens," a grid that visualizes where specific information about a particular relation is stored within the transformer's many layers.

Attribute lenses can be generated automatically, providing a streamlined method to help researchers understand more about a model. This visualization tool could enable scientists and engineers to correct stored knowledge and help prevent an AI chatbot from giving false information.

In the future, Hernandez and his collaborators want to better understand what happens in cases where facts are not stored linearly. They would also like to run experiments with larger models, as well as study the precision of linear decoding functions.

"This is an exciting work that reveals a missing piece in our understanding of how large language models recall factual knowledge during inference. Previous work showed that LLMs build information-rich representations of given subjects, from which specific attributes are being extracted during inference.

"This work shows that the complex nonlinear computation of LLMs for attribute extraction can be well-approximated with a simple linear function," says Mor Geva Pipek, an assistant professor in the School of Computer Science at Tel Aviv University, who was not involved with this work.

More information: Evan Hernandez et al, Linearity of Relation Decoding in Transformer Language Models, *arXiv* (2023). DOI: 10.48550/arxiv.2308.09124

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology