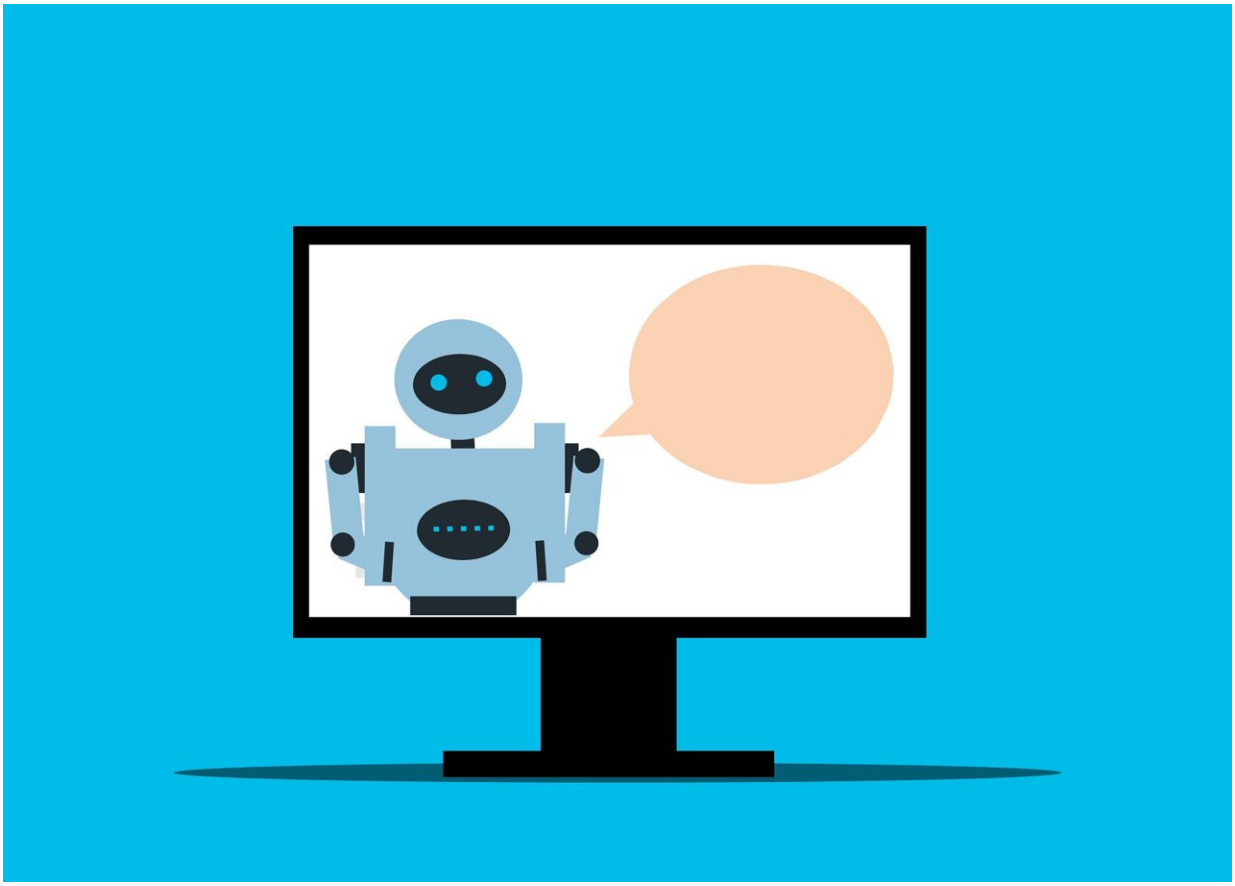
Q&A: How AI models teach themselves to learn new things

July 18 2023, by Jan Overney



Credit: Pixabay/CC0 Public Domain

Despite their huge success, the inner workings of large language models such as OpenAI's GPT model family and Google Bard remain a mystery,

even to their developers. Researchers at ETH and Google have uncovered a potential key mechanism behind their ability to learn on-the-fly and fine-tune their answers based on interactions with their users.

Johannes von Oswald is a doctoral student in the group headed by Angelika Steger, ETH Professor for Theoretical Computer Science, and researches learning algorithms for neural networks. His new paper will be presented at the International Conference on Machine Learning (ICML) in late July. It is currently available on the *arXiv* preprint server.

The T in GPT stands for transformers. What are transformers and why did they become so prevalent in modern AI?

Johannes von Oswald: Transformers are a particular artificial neural network architecture. It is for example used by large language models such as ChatGPT, but was put on the map in 2017 by researchers at Google, where it led to state-of-the-art performance in language translation. Intriguingly, a slightly modified version of this architecture was already developed by the AI-Pioneer Jürgen Schmidhuber back in 1991.

And what distinguishes this architecture?

Before the recent breakthrough of Transformers, different tasks, e.g., image classification and language translation, had used different model architectures that were each specialized on these specific domains. A crucial aspect that sets transformers apart from these previous AI models is that they seem to work extremely well on any kind of task. Because of their widespread use, it is important to understand how they work.

What did your research reveal?

While neural networks are generally regarded a black box that spit out output when provided with input, we showed that transformers can learn on their own to implement algorithms within their architecture. We were able to show that they can implement a classic and powerful machine learning algorithm that learns from the recent information it receives.

Can you give an example when this type of learning can occur?

You might, for instance, provide the language model with several texts and the sentiment—either positive or negative—associated with each of them. You can go on to present the model with a text it hasn't seen before, and it will predict whether it is positive or negative based on the examples you have provided.

So you're saying that the model teaches itself a technique to learn new things?

Yes, it's surprising but true. Driven simply by the pressure to improve on its training objective, namely to predict the immediate future, it develops a technique that enables it to learn from the conversations it has with its users, for example. This type of learning is what we refer to as in-context learning.

All these models get is text input. Can you describe how transformers use this minimal information to optimize their output?

One way to achieve this—and our paper shows that it's a likely possibility—is to learn what you might call a world model that allows you to make predictions. What is interesting is that this learning takes

place inside the transformer that has already been trained. Learning would normally involve changing the connections in the model's neural network. We showed that the transformer model is somehow able to simulate the same learning process within its fixed neural architecture instead.

How does this capability emerge in transformers?

We hypothesized in our paper that the transformer architecture has an inductive bias towards learning. This means that its ability to develop these learning mechanisms is implicitly built into its basic design, even before the model is trained.

GPT-3, the model behind ChatGPT, has 175 billion parameters. How do you study such a large system?

There are different ways to try to understand these systems. Some researchers take a psychological approach and analyze how the models respond when confronted with standardized tests or conflicting situations such as moral dilemmas. We studied this system mechanistically—as neuroscientists you could say. Taking this analogy further, because our model runs on a computer, we were able to record every neuron and every connection in its neural network—something that would be unthinkable when studying the biological brains of animals or humans. Investigating these systems at the level of individual neurons is currently only feasible when studying very specific phenomena on relatively small architectures.

Can you provide more information on the system you used in your paper?

The transformer we used in our study is roughly identical to the

commonly used transformer architecture. Rather than training our system on all the texts on the internet, we trained it on examples of a simple problem known as linear regression. Because this problem and its solution are so well understood, we were able to compare the known solution with what we observed inside the transformer. We confirmed that it implements a very well-known and powerful learning algorithm within itself called gradient descent.

Would you expect other behavior to emerge that is entirely new to computer science?

That is possible. In our case, we were able to show that the transformer was not simply performing plain gradient descent but an improved version of it. Two independent studies from MIT and UC Berkeley have now analyzed the algorithm that the transformer learned. A long-term goal of this line of research could be to determine whether transformers can discover algorithms or even prove theorems and develop mathematics that we are not yet familiar with. This would be truly remarkable and groundbreaking.

More information: Johannes von Oswald et al, Transformers learn in-context by gradient descent, *arXiv* (2022). DOI: 10.48550/arxiv.2212.07677

Provided by ETH Zurich