GPT-3 transforms chemical research

February 6 2024

Credit: CC0 Public Domain

Artificial intelligence is growing into a pivotal tool in chemical research, offering novel methods to tackle complex challenges that traditional approaches struggle with. One subtype of artificial intelligence that has seen increasing use in chemistry is machine learning, which uses algorithms and statistical models to make decisions based on data and

perform tasks that it has not been explicitly programmed for.

However, to make reliable predictions, machine learning also demands large amounts of data, which isn't always available in chemical research. Small chemical datasets simply do not provide enough information for these algorithms to train on, which limits their effectiveness.

Scientists, in the team of Berend Smit at EPFL, have found a solution in large language models such as GPT-3. Those models are pre-trained on massive amounts of texts, and are known for their broad capabilities in understanding and generating human-like text. GPT-3 forms the basis of the more popular artificial intelligence ChatGPT.

The study, published in *Nature Machine Intelligence*, unveils a novel approach that significantly simplifies chemical analysis using artificial intelligence. Contrary to initial skepticism, the method doesn't directly ask GPT-3 chemical questions.

"GPT-3 has not seen most of the chemical literature, so if we ask ChatGPT a chemical question, the answers are typically limited to what one can find on Wikipedia," says Kevin Jablonka, the study's lead researcher.

"Instead, we fine-tune GPT-3 with a small data set converted into questions and answers, creating a new model capable of providing accurate chemical insights."

This process involves feeding GPT-3 a curated list of Q&As. "For example, for high-entropy alloys, it is important to know whether an alloy occurs in a single phase or has multiple phases," says Smit. "The curated list of Q&As are of the type: Q= 'Is the (name of the high entropy alloy) single phase?' A= 'Yes/No.'"

He continues, "In the literature, we have found many alloys of which the answer is known, and we used this data to fine-tune GPT-3. What we get back is a refined AI model that is trained to only answer this question with a yes or no."

In tests, the model, trained with relatively few Q&As, correctly answered over 95% of very diverse chemical problems, often surpassing the accuracy of state-of-the-art machine-learning models. "The point is that this is as easy as doing a literature search, which works for many chemical problems," says Smit.

One of the most striking aspects of this study is its simplicity and speed. Traditional machine learning models require months to develop and demand extensive knowledge. In contrast, the approach developed by Jablonka takes five minutes and requires zero knowledge.

The implications of the study are profound. It introduces a method as easy as conducting a literature search, applicable to various chemical problems. The ability to formulate questions like "Is the yield of a [chemical] made with this (recipe) high?" and receive accurate answers can revolutionize how chemical research is planned and carried out.

In the paper, the authors say, "Next to a literature search, querying a foundational model (e.g., GPT-3,4) might become a routine way to bootstrap a project by leveraging the collective knowledge encoded in these foundational models." Or, as Smit succinctly puts it, "This is going to change the way we do chemistry."

More information: Kevin Maik Jablonka, Is GPT all you need for low-data discovery in chemistry?, *Nature Machine Intelligence* (2024). DOI: 10.1038/s42256-023-00788-1

Provided by Ecole Polytechnique Federale de Lausanne