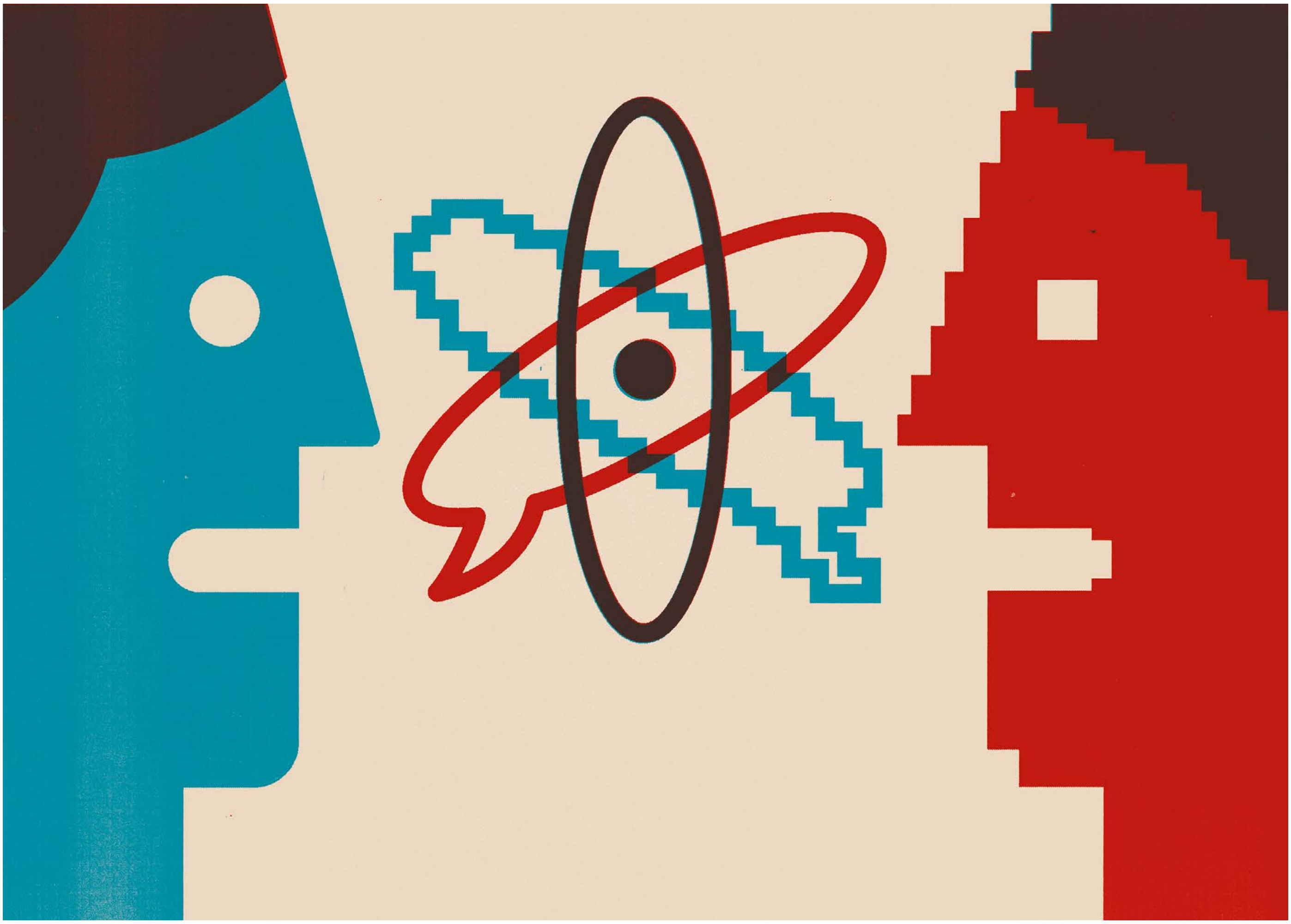

ILLUSTRATION BY THE PROJECT TWINS

CHATGPT FOR SCIENCE: HOW TO TALK TO YOUR DATA

Companies are using artificial-intelligence tools to help scientists to query their data without the need for programming skills. **By Julian Nowogrodzki**

"Computer, analyse." In science fiction, characters don't need programming skills to extract meaningful information from their data, they simply ask for it.

Now a growing number of companies are attempting to make that fiction a reality – sort of – using large language models (LLMs). These powerful but focused artificial intelligence (AI) tools let researchers ask natural-language questions of their data, such as "what's the difference between the control group and the experimental group?". But unlike their science-fiction counterparts, the answers these AIs spit out still need to be taken with a grain of salt and double-checked before they can be used safely. Think ChatGPT, for data.

The reason for the tools is simple: sifting through and prioritizing biological data is laborious and challenging, and requires specialized skills. "Biological data has become increasingly complex," says Alexandro Trevino, scientific manager at Enable Medicine in San Francisco, California, a company that is building an atlas of spatial gene-expression and protein-localization data for its drug-development clients. "The scale has increased vastly, the complexity of these data sets has increased, and with that I think we have scaled the challenges of mining and effectively understanding and interpreting those data."

In theory, dedicated LLMs allow researchers to extract insights from their data without knowing the data's intricacies, or how to program. And some of these tools can already answer remarkably complex questions. But they remain works in progress. And like other LLM-based tools, they can 'hallucinate' or make up answers. As a result, their developers say that they should be used only with some degree of oversight by humans.

Why talk to your data?

There is no shortage of online data, nor tools to query it. The CZ CELLxGENE data portal, for instance, provides pre-built tools that allow researchers to interrogate single-cell gene-expression data sets. Utilities such as ChatPDF allow researchers to upload PDFs, such as scientific papers, and ask questions of them. But more sophisticated analyses require knowing the structure of the underlying data and the names and types of their variables.

To make such interactions easier,

biotechnology company Genentech in San Francisco is building its LLM-based tool from scratch. Led by Stephen Ra, the company's director of frontier research in New York City, this LLM aims to address "a vast array of problems across the drug-discovery and development pipeline", he says, "from target identification, discovery, safety, assessment, prioritization, all the way to how do we make better decisions, or de-risk certain clinical trial phases, or understand patient trajectories and adverse outcomes better".

The resulting LLM could ease tasks that are currently manual and onerous, Ra says. For example, a scientist might put one of their data sets aside for a while, but then want to summarize those data later. They could ask, "give me all the results for this particular assay, at this particular time, for this strain", Ra says. The system should be able to understand the query, and the data, well enough to fulfil the request, and "many teams" across Genentech and its parent company Roche are beta testing it.

Similarly, Enable Medicine's LLM aims to allow the company to interrogate its biological atlas on behalf of its clients, mostly pharmaceutical companies in oncology and autoimmune disease, says chief executive Kamni Vijay.

Researchers can ask questions such as "does a patient respond to therapy, and what differentiates patients who respond to a therapy from those who do not?", or "what biomarkers would influence or predict disease progression?", Vijay says. Enable is building on several existing LLMs, she adds, and training with petabytes (1 petabyte is 1 million gigabytes) of molecular and cellular data from tens of thousands of samples. They are still experimenting, however. "Part of our research explores whether this type of interface can be scientifically valid and valuable."

What do they look like?

Some tools in this space emulate ChatGPT's popular question-and-answer format. For instance, PathChat, built by computational pathologist Faisal Mahmood, at Brigham and Women's Hospital in Boston, Massachusetts, allows users to input pathology images, such as tumour biopsy results, as well as descriptive data such as "this tumour stained positive for markers A, B and C". (M. Y. Lu *et al. Nature* https://doi.org/gtzht8 (2024). Users can then ask natural-language questions about these data, such as, "what is your assessment of the primary origin of the tumour?" The exchanges appear visually like the back-and-forth text bubbles of a WhatsApp conversation.

Enable's system, however, diverges from the question-and-answer format, says Vijay. It is a more complex automated system that allows for natural-language queries, she says.

Still other tools output code instead of words. Mergen is an LLM-based R programming language library built by Altuna Akalin, a bioinformatician at the Max Delbrück Center in Berlin. Akalin created the library (or 'package') because his team was getting more requests to analyse genomic data than it could handle. Intended for genomics researchers rather than computational scientists, Mergen analyses pre-processed genomics data sets to answer questions such as, "can you give me all the genes that are overexpressed in a certain set of individuals?" Instead of an answer, the tool returns executable code that can perform the analysis. As with all LLMs, however, that code should be double-checked by a person before it is used, Akalin warns, because even if the code is executable, it might contain logical errors.

"In this very effective approach, the model itself is concretely learning new stuff."

How are they made?

What does it take to build an LLM that allows researchers to converse with their data? As with all AI systems, the answer is lots of training data. But the balance of data types is equally important, and his team puts considerable effort into achieving the right balance, says Ra. "The value for us lies in being able to take something that's broadly useful to many groups [in Genentech] and allow those groups to also fine-tune their own model." Genentech trained its model on a combination of in-house and external information covering multiple projects and fields, including omics and clinical data, Ra says.

Trevino says that there are two main ways to transform a generalist LLM into a system that enables users to converse with their data. One is to fine-tune the generalist LLM using field-specific information, such as pathology data. In this "very effective" approach, he says, the model itself "is concretely learning new stuff". The other approach, called contextualization, doesn't change the underlying generalist LLM but gives it tailored context, such as a database of medical literature, as part of the query. Trevino declined to say which approach Enable uses.

To build PathChat, Mahmood and his team started with the generalist LLM Llama 2, developed by Facebook parent company Meta. They hooked the LLM up to two vision-language models that they had built for pathology, called UNI and CONCH, each of which was trained on millions of pathology images and captions, to make a multimodal LLM. The researchers then refined that multimodal LLM using half a million pathology conversations extracted from case reports and educational articles that follow the complete trajectories of cases, mostly from Brigham and Women's Hospital and Massachusetts General Hospital, to yield PathChat, Mahmood says. Some pathologists at Brigham and Women's are now using the system to interpret microscopy images and write morphological descriptions that a pathologist can then check, he says.

Are they trustworthy?

Confirmation is important: just because an LLM provides an answer doesn't mean that answer is correct. LLMs can make answers up or leave information out, and how best to ensure that a model's response is verifiable and replicable remains unsettled, Trevino says. "It's an active area of research, how to vet the results."

One crucial aspect, says Ra, is feedback from field-specific experts. There are different ways to incorporate such checks – users could provide a simple thumbs up or thumbs down, for instance, a more detailed response, or there could be iterative interaction between a person and an LLM. In any event, the hope is that over time the model will evolve to require less input, because such feedback isn't scalable as data sets expand.

Trevino and Ra say that understanding and trusting what's going on in the underlying model is especially important in research-specific LLMs. One challenge, says Trevino, is to "open up that black box a little bit" to understand better why it answers in the way it does. This could help to minimize hallucinations.

Indeed, one of Genentech's motivations for building its LLM from scratch, Ra says, is that it wants to know it can trust and understand every bit of data that goes into it. "That's incredibly important in an environment where we're often dealing with privileged information or very sensitive information", such as patient data, he says.

With off-the-shelf, 'black box' LLMs, it isn't always clear how they are trained, Ra explains. "I think this has been a common criticism of some of the commercial LLM solutions, that oftentimes there's not enough data transparency."

Another persistent challenge, as in the field of LLMs as a whole, is bias in the underlying data. Groups that are under-represented in the training data will be misrepresented by the resulting model, and current genomic data hugely over-represent people of European descent. The solution, say Trevino and Vijay, is to improve the diversity of the underlying data. But there's not really an endpoint for when the underlying data is sufficiently diverse, they say.

Should these challenges be overcome, however, "there are going to be very real benefits" to these types of model, Trevino says. The important thing is "to make sure that that benefit is realized and maximally democratized," and that the gain is worth all the work still left to do.

Julian Nowogrodzki is a science writer and editor in Boston, Massachusetts.