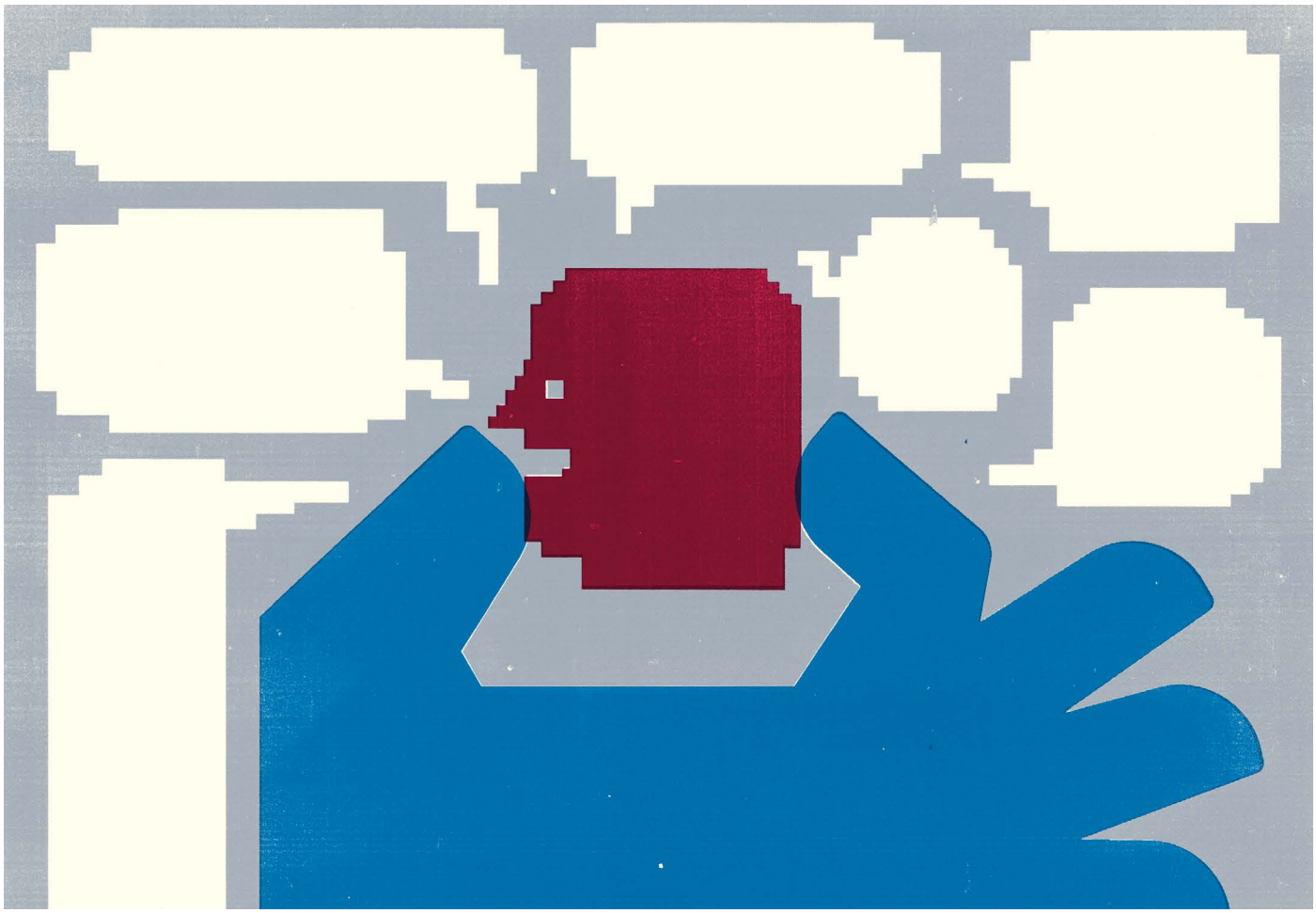

ILLUSTRATION: THE PROJECT TWINS

FORGET CHATGPT: HOW TO RUN AI LOCALLY ON A LAPTOP

Researchers typically use AIs online, but a host of openly available tools means they don't have to. **By Matthew Hutson**

The website histo.fyi is a database of structures of immune-system proteins called major histocompatibility complex (MHC) molecules. It includes images, data tables and amino-acid sequences, and is run by bioinformatician Chris Thorpe, who uses artificial intelligence (AI) tools called large language models (LLMs) to convert those assets into readable summaries. But he doesn't use ChatGPT, or any other web-based LLM. Instead, Thorpe runs the AI on his laptop.

Over the past couple of years, chatbots based on LLMs have won praise for their ability to write poetry or engage in conversations. Some LLMs have hundreds of billions of parameters – the more parameters, the greater the complexity – and can be accessed only online. But two more recent trends have blossomed. First, organizations are making 'open weights' versions of LLMs, in which the weights and biases used to train a model are publicly available, so that users can download and run them locally, if they have the computing power. Second, technology firms are making scaled-down versions that can be run on consumer hardware – and that rival the performance of older, larger models.

Researchers might use such tools to save money, protect the confidentiality of patients or corporations, or ensure reproducibility. Thorpe, who's based in Oxford, UK, and works at the European Molecular Biology Laboratory's European Bioinformatics Institute in Hinxton, UK, is just one of many researchers exploring what the tools can do. That trend is likely to grow, Thorpe says. As computers get faster and models become more efficient, people will increasingly have AIs running on their laptops or mobile devices for all but the most intensive needs. Scientists will finally have AI assistants at their fingertips – but the actual algorithms, not just remote access to them.

Big things in small packages

Several large tech firms and research institutes have released small and open-weights models over the past few years, including Google DeepMind in London; Meta in Menlo Park, California; and the Allen Institute for Artificial Intelligence in Seattle, Washington. ('Small' is relative – these models can contain some 30 billion parameters, which is large by comparison with earlier models.)

Although the California tech firm OpenAI hasn't open-weighted its current GPT models, its partner Microsoft in Redmond, Washington, has been on a spree, releasing the small

language models Phi-1, Phi-1.5 and Phi-2 in 2023, then four versions of Phi-3 and three versions of Phi-3.5 this year. The Phi-3 and Phi-3.5 models have between 3.8 billion and 14 billion active parameters, and two models (Phi-3-vision and Phi-3.5-vision) handle images[1]. By some benchmarks, even the smallest Phi model outperforms OpenAI's GPT-3.5 Turbo from 2023, rumoured to have 20 billion parameters.

Sébastien Bubeck, Microsoft's vice-president for generative AI, attributes Phi-3's performance to its training data set. LLMs initially train by predicting the next 'token' (iota of text) in long text strings. To predict the name of the killer at the end of a murder mystery, for instance, an AI needs to 'understand' everything that came before, but such consequential predictions are rare in most text. To get around this problem, Microsoft used LLMs to write millions of short stories and textbooks in which one thing builds on another. The result of training on this text, Bubeck says, is a model that fits on a mobile phone but has the power of the initial 2022 version of ChatGPT. "If you are able to craft a data set that is very rich in those reasoning tokens, then the signal will be much richer," he says.

Phi-3 can also help with routing — deciding whether a query should go to a larger model. "That's a place where Phi-3 is going to shine," Bubeck says. Small models can also help scientists in remote regions that have little cloud connectivity. "Here in the Pacific Northwest, we have amazing places to hike, and sometimes I just don't have network," he says. "And maybe I want to take a picture of some flower and ask my AI some information about it."

Researchers can build on these tools to create custom applications. The Chinese e-commerce site Alibaba, for instance, has built models called Qwen with 500 million to 72 billion parameters. A biomedical scientist in New Hampshire fine-tuned the largest Qwen model using scientific data to create Turbcat-72b, which is available on the model-sharing site Hugging Face. (The researcher goes only by the name Kal'tsit on the Discord messaging platform, because AI-assisted work in science is still controversial.) Kal'tsit says she created the model to help researchers to brainstorm, proof manuscripts, prototype code and summarize published papers; the model has been downloaded thousands of times.

Preserving privacy

Beyond the ability to fine-tune open models for focused applications, Kal'tsit says, another advantage of local models is privacy. Sending personally identifiable data to a commercial service could run foul of data-protection regulations. "If an audit were to happen and you show them you're using ChatGPT, the situation could become pretty nasty," she says.

Cyril Zakka, a physician who leads the health team at Hugging Face, uses local models to generate training data for other models (which are sometimes local, too). In one project, he uses them to extract diagnoses from medical reports so that another model can learn to predict those diagnoses on the basis of echocardiograms, which are used to monitor heart disease. In another, he uses the models to generate questions and answers from medical textbooks to test other models. "We are paving the way towards fully autonomous surgery," he explains. A robot trained to answer questions would be able to communicate better with doctors.

Zakka uses local models — he prefers Mistral 7B, released by the tech firm Mistral AI in Paris, or Meta's Llama-3 70B — because they're cheaper than subscription services such as ChatGPT Plus, and because he can fine-tune them. But privacy is also key, because he's not allowed to send patients' medical records to commercial AI services.

Johnson Thomas, an endocrinologist at the health system Mercy in Springfield, Missouri, is likewise motivated by patient privacy. Clinicians rarely have time to transcribe and summarize patient interviews, but most commercial services that use AI to do so are either too expensive or not approved to handle private medical data. So, Thomas is developing an alternative. Based on Whisper — an open-weight speech-recognition model from OpenAI — and on Gemma 2 from Google DeepMind, the system will allow physicians to transcribe conversations and convert them to medical notes, and also summarize data from medical-research participants.

Onur Karakaslar, a computational biologist at Leiden University Medical Center in the Netherlands, developed a pipeline named ceLLama to annotate cell types using local LLMs such as Llama 3.1. He highlights privacy as one advantage on his GitHub page, noting that ceLLama "operates locally, ensuring no data leaks". The similarly named CELLama, developed at the South Korean pharmaceutical company Portrai in Seoul, exploits LLMs to reduce information about a cell's gene expression and other characteristics to a summary sentence[2]. It then creates a numerical representation of this sentence, which can be used to cluster cells into types.

Putting models to good use

As the LLM landscape evolves, scientists face a fast-changing menu of options. "I'm still at the tinkering, playing stage of using LLMs locally," Thorpe says. He tried ChatGPT, but felt it was expensive, and the tone of its output wasn't right. Now he uses Llama locally, with either 8 billion or 70 billion parameters, both of which can run on his Mac laptop.

Another benefit, Thorpe says, is that local models don't change. Commercial developers, by contrast, can update their models at any moment, leading to different outputs and forcing Thorpe to alter his prompts or templates. "In most of science, you want things that are reproducible," he explains. "And it's always a worry if you're not in control of the reproducibility of what you're generating."

For another project, Thorpe is writing code that aligns MHC molecules on the basis of their 3D structure. To develop and test his algorithms, he needs lots of diverse proteins — more than exist naturally. To design plausible new proteins, he uses ProtGPT2, an open-weights model with 738 million parameters that was trained on about 50 million sequences[3].

Sometimes, however, a local app won't do. For coding, Thorpe uses the cloud-based GitHub Copilot as a partner. "It kind of feels like my arm's chopped off when for some reason I can't actually use Copilot," he says. Local LLM-based coding tools do exist (such as Google DeepMind's CodeGemma and one from California-based developers Continue), but in his experience they can't compete with Copilot.

Access points

So, how do you run a local LLM? Software called Ollama (available for Mac, Windows and Linux operating systems) lets users download open models, including Llama 3.1, Phi-3, Mistral and Gemma 2, and access them through a command line. Other options include the cross-platform app GPT4All and Llamafile, which can transform LLMs into a single file that runs on any of six operating systems, with or without a graphics processing unit.

Sharon Machlis, a former editor at the website InfoWorld, who lives in Framingham, Massachusetts, wrote a guide to using LLMs locally, covering a dozen options. "The first thing I would suggest," she says, "is to have the software you choose fit your level of how much you want to fiddle." Some people prefer the ease of apps, whereas others prefer the flexibility of the command line.

Whichever approach you choose, local LLMs should soon be good enough for most applications, says Stephen Hood, who heads open-source AI at the tech firm Mozilla in San Francisco. "The rate of progress on those over the past year has been astounding," he says.

As for what those applications might be, that's for users to decide. "Don't be afraid to get your hands dirty," Zakka says. "You might be pleasantly surprised by the results."

Matthew Hutson is a science writer based in New York City.

1. Abdin, M. *et al*. Preprint at arXiv https://doi.org/10.48550/arXiv.2404.14219 (2024).
2. Choi, H. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2024.05.08.593094 (2024).
3. Ferruz, N. *et al*. *Nature Commun.* **13**, 4348 (2022).

Correction

This Technology feature conflated two pieces of software. The quote "operates locally, ensuring no data leaks" actually came from the developer of ceLLama, not CELLama.

Corrected 24 September 2024