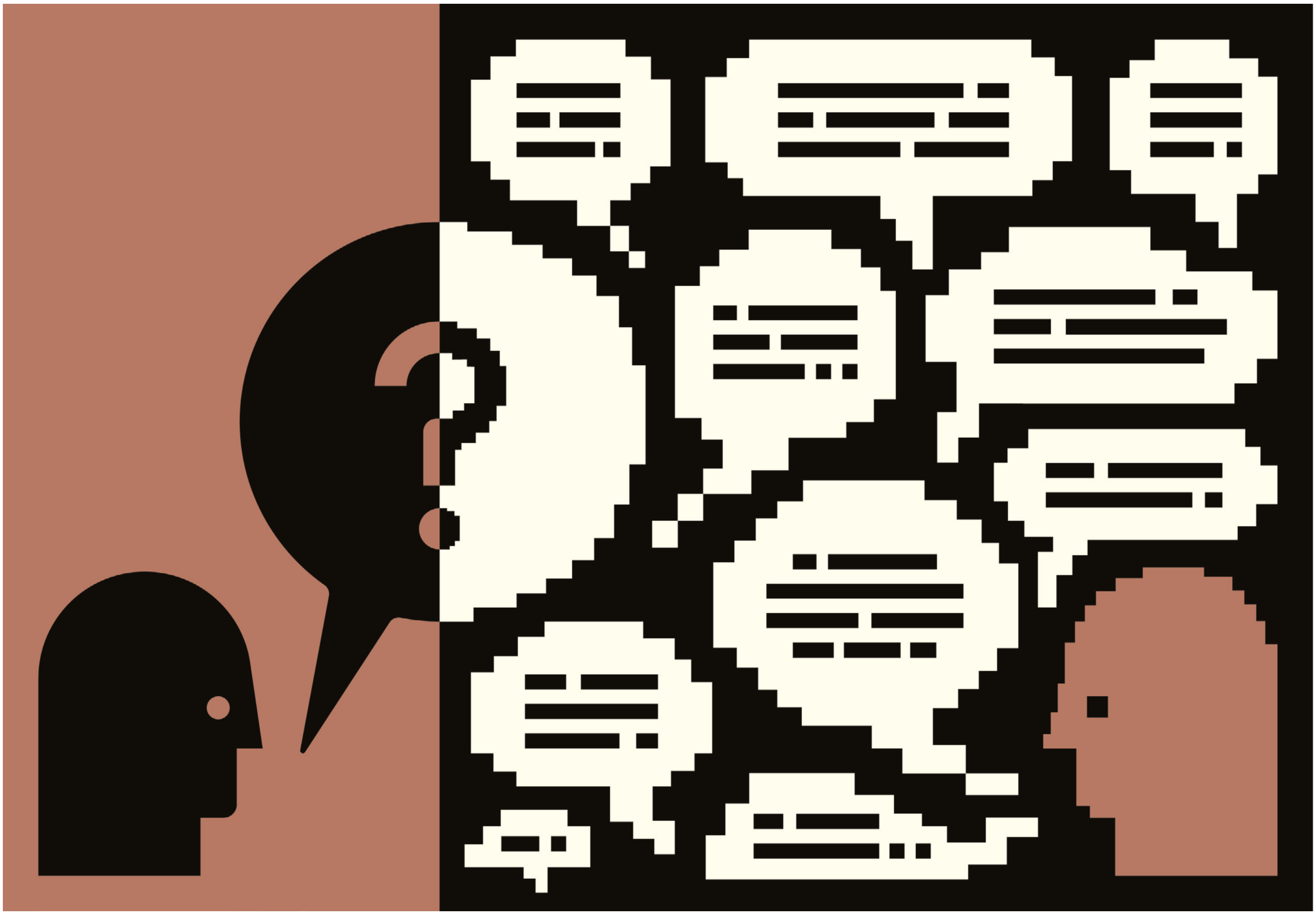

ILLUSTRATION BY THE PROJECT TWINS

COULD AI HELP YOU TO WRITE YOUR NEXT PAPER?

Large language models can draft abstracts or suggest research directions, but these artificial-intelligence tools are a work in progress. **By Matthew Hutson**

You know that text autocomplete function that makes your smartphone so convenient – and occasionally frustrating – to use? Well, now tools based on the same idea have progressed to the point that they are helping researchers to analyse and write scientific papers, generate code and brainstorm ideas.

The tools come from natural language processing (NLP), an area of artificial intelligence aimed at helping computers to 'understand' and even produce human-readable text. Called large language models (LLMs), these tools have evolved to become not only objects of study but also assistants in research.

LLMs are neural networks that have been trained on massive bodies of text to process and, in particular, generate language. OpenAI, a research laboratory in San Francisco, California, created the most well-known LLM, GPT-3, in 2020, by training a network to predict the next piece of text based on what came before. On Twitter and elsewhere, researchers have expressed amazement at its spookily human-like writing. And anyone can now use it, through the OpenAI programming interface, to generate text based on a prompt. (Prices start at about US$0.0004 per 750 words processed – a measure that combines reading the prompt and writing the response.)

"I think I use GPT-3 almost every day," says computer scientist Hafsteinn Einarsson at the University of Iceland, Reykjavik. He uses it to generate feedback on the abstracts of his papers. In one example that Einarsson shared at a conference in June, some of the algorithm's suggestions were useless, advising him to add information that was already included in his text. But others were more helpful, such as "make the research question more explicit at the beginning of the abstract". It can be hard to see the flaws in your own manuscript, Einarsson says. "Either you have to sleep on it for two weeks, or you can have somebody else look at it. And that 'somebody else' can be GPT-3."

Organized thinking

Some researchers use LLMs to generate paper titles or to make text more readable. Mina Lee, a doctoral student in computer science at Stanford University, California, gives GPT-3 prompts such as "using these keywords, generate the title of a paper". To rewrite troublesome sections, she uses an AI-powered writing assistant called Wordtune by AI21 Labs in Tel Aviv, Israel. "I write a paragraph, and it's basically like a doing brain dump," she says. "I just click

'Rewrite' until I find a cleaner version I like."

Computer scientist Domenic Rosati at the technology start-up Scite in Brooklyn, New York, uses an LLM called Generate to organize his thinking. Developed by Cohere, an NLP firm in Toronto, Canada, Generate behaves much like GPT-3. "I put in notes, or just scribbles and thoughts, and I say 'summarize this', or 'turn this into an abstract'," Rosati says. "It's really helpful for me as a synthesis tool."

Language models can even help with experimental design. For one project, Einarsson was using the game Pictionary as a way to collect language data from participants. Given a description of the game, GPT-3 suggested game variations he could try. Theoretically, researchers could also ask for fresh takes on experimental protocols. As for Lee, she asked GPT-3 to brainstorm things to do when introducing her boyfriend to her parents. It suggested going to a restaurant by the beach.

Encoding coding

OpenAI researchers trained GPT-3 on a vast assortment of text, including books, news stories, Wikipedia entries and software code. Later, the team noticed that GPT-3 could complete pieces of code, just like it can with other text. The researchers created a fine-tuned version of the algorithm called Codex, training it on more than 150 gigabytes of text from the code-sharing platform GitHub[1]. GitHub has now integrated Codex into a service called Copilot that suggests code as people type.

Computer scientist Luca Soldaini at the Allen Institute for AI (also called AI2) in Seattle, Washington, says at least half their office uses Copilot. It works best for repetitive programming, Soldaini says, citing a project that involves writing boilerplate code to process PDFs. "It just blurts out something, and it's like, 'I hope this is what you want'." Sometimes it's not. As a result, Soldaini says they are careful to use Copilot only for languages and libraries with which they are familiar, so they can spot problems.

Literature searches

Perhaps the most established application of language models involves searching and summarizing literature. AI2's Semantic Scholar search engine – which covers around 200 million papers, mostly from biomedicine and computer science – provides tweet-length descriptions of papers using a language model called TLDR (short for too long; didn't read). TLDR is derived from an earlier model called BART, by researchers at the social media platform Facebook, that's been fine-tuned on human-written summaries. (By today's standards, TLDR is not a large language model, because it contains only about 400 million parameters. The largest version of GPT-3 contains 175 billion.)

TLDR also appears in AI2's Semantic Reader, an application that augments scientific papers. When a user clicks on an in-text citation in Semantic Reader, a box pops up with information that includes a TLDR summary. "The idea is to take artificial intelligence and put it right into the reading experience," says Dan Weld, Semantic Scholar's chief scientist.

When language models generate text summaries, often "there's a problem with what people charitably call hallucination", Weld says, "but is really the language model just completely making stuff up or lying." TLDR does relatively well on tests of truthfulness[2] – authors of papers TLDR was asked to describe rated its accuracy as 2.5 out of 3. Weld says this is partly because the summaries are only about 20 words long, and partly because the algorithm rejects summaries that introduce uncommon words that don't appear in the full text.

In terms of search tools, Elicit debuted in 2021 from the machine-learning non-profit organization Ought in San Francisco, California. Ask Elicit a question, such as, "What are the effects of mindfulness on decision making?" and it outputs a table of ten papers. Users can ask the software to fill columns with content such as abstract summaries and metadata, as well as information about study participants, methodology and results. Elicit uses tools including GPT-3 to extract or generate this information from papers.

> **"It just blurts out something, and it's like, 'I hope this is what you want'."**

Joel Chan at the University of Maryland in College Park, who studies human–computer interactions, uses Elicit whenever he starts a project. "It works really well when I don't know the right language to use to search," he says. Neuroscientist Gustav Nilsonne at the Karolinska Institute, Stockholm, uses Elicit to find papers with data he can add to pooled analyses. The tool has suggested papers he hadn't found in other searches, he says.

Evolving models

Prototypes at AI2 give a sense of the future for LLMs. Sometimes researchers have questions after reading a scientific abstract but don't have the time to read the full paper. A team at AI2 developed a tool that can answer such questions, at least in the domain of NLP. It began by asking researchers to read the abstracts of NLP papers and then ask questions about them (such as "what five dialogue attributes were analysed?"). The team then asked other researchers to answer those questions after they had read the full papers[3]. AI2 trained a version of its Longformer language model – which can ingest a complete paper, not just the few hundred words that other models take in – on the resulting data set to generate answers to different questions about other papers[4].

A model called ACCoRD can generate definitions and analogies for 150 scientific concepts related to NLP, whereas MS^2, a data set of 470,000 medical documents and 20,000 multi-document summaries, was used to fine-tune BART to allow researchers to take a question and a set of documents and generate a brief meta-analytical summary.

And then there are applications beyond text generation. In 2019, AI2 fine-tuned BERT, a language model created by Google in 2018, on Semantic Scholar papers to create SciBERT, which has 110 million parameters. Scite, which has used AI to create a scientific search engine, further fine-tuned SciBERT so that when its search engine lists papers citing a target paper, it categorizes them as supporting, contrasting or otherwise mentioning that paper. Rosati says that that nuance helps people to identify limitations or gaps in the literature.

AI2's SPECTER model, also based on SciBERT, reduces papers to compact mathematical representations. Conference organizers use SPECTER to match submitted papers to peer reviewers, Weld says, and Semantic Scholar uses it to recommend papers based on a user's library.

Computer scientist Tom Hope, at the Hebrew University of Jerusalem and AI2, says that other research projects at AI2 have fine-tuned language models to identify effective drug combinations, connections between genes and disease, and scientific challenges and directions in COVID-19 research.

But can language models allow deeper insight or even discovery? In May, Hope and Weld co-authored a review[5] with Eric Horvitz, chief scientific officer at Microsoft, and others that lists challenges to achieving this, including teaching models to "[infer] the result of recombining two concepts". "It's one thing to generate a picture of a cat flying into space," Hope says, referring to OpenAI's DALL·E 2 image-generation model. But "how will we go from that to combining abstract, highly complicated scientific concepts?"

That's an open question. But LLMs are already making a tangible impact on research. "At some point," Einarsson says, "people will be missing out if they're not using these large language models."

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

1. Chen, M. *et al.* Preprint at https://arxiv.org/abs/2107.03374 (2021).
2. Cachola, I., Lo, K., Cohan, A. & Weld, D. S. In *Findings of the Association for Computational Linguistics* 4766–4777 (2020).
3. Dasigi, P. *et al.* In *Proc. 2021 Conference of the North American Chapter of the Association of Computational Linguistics* 4599–4610 (2021).
4. Beltagy, I., Peters, M. E. & Cohan, A. Preprint at https://arxiv.org/abs/2004.05150 (2020).
5. Hope, T. *et al.* Preprint at https://arxiv.org/abs/2205.02007 (2022).