AETERNA NON CADUCA

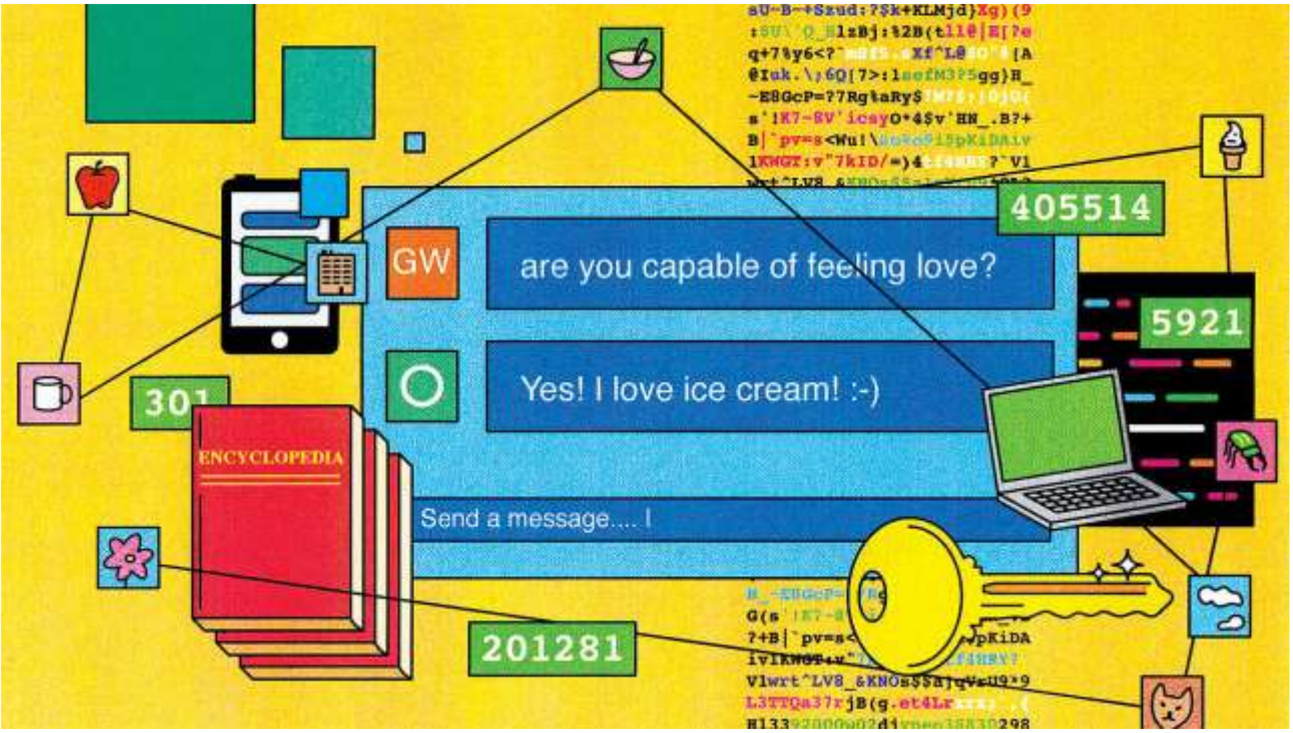
SAINTS PERSPECTIVES

*Things Scientific and Technical*

# Large, Creative AI Models will Transform Lives and Labour Markets

## They bring enormous promise and peril. But how do they work?

[The Economist, Science & Technology, April 22nd 2023](#)



are you capable of feeling love?

Yes! I love ice cream! :-)

Since November 2022, when **OpenAI**, the company which makes **ChatGPT**, first opened the chatbot to the public, there has been little else that the tech elite has wanted to talk about. As this article was being written, the founder of a London technology company messaged your correspondent unprompted to say that this kind of AI is “essentially all I’m thinking about these days”. He says he is in the process of redesigning his company, valued at many billions of dollars, around it. He is not alone.

**ChatGPT** embodies more knowledge than any human has ever known. It can converse cogently about mineral extraction in Papua New Guinea, or about TSMC, a Taiwanese semiconductor firm that finds itself in the geopolitical crosshairs. **GPT-4**, the artificial neural network which powers ChatGPT, has aced exams that serve as gateways for people to enter careers in law and medicine in America. It can generate songs, poems and essays. Other "**generative AI**" models can churn out digital photos, drawings and animations.

Running alongside this excitement is deep concern, inside the tech industry and beyond, that generative AI models are being developed too quickly. GPT-4 is a type of generative AI called a **large language model (LLM**). Tech giants like Alphabet, Amazon and Nvidia have all trained their own LLMS, and given them names like palm, Megatron, Titan and Chinchilla.

## The lure grows greater

The London tech boss says he is "incredibly nervous about the existential threat" posed by AI, even as he pursues it, and is "speaking with [other] founders about it daily". Governments in America, Europe and China have all started mulling new regulations. Prominent voices are calling for the development of artificial intelligence to be paused, lest the software somehow run out of control and damage, or even destroy, human society. To calibrate how worried or excited you should be about this technology, it helps first to understand where it came from, how it works and what the limits are to its growth.

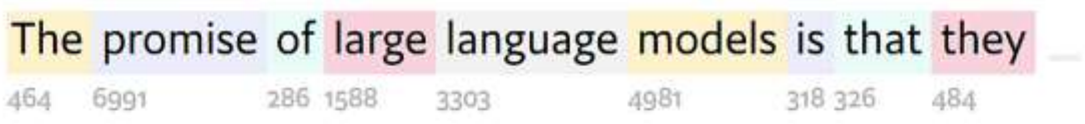
The contemporary explosion of the capabilities of AI software began in the early 2010s, when a software technique called "**deep learning**" became popular. Using the magic mix of vast datasets and powerful computers running neural networks on Graphics Processing Units (GPUS), deep learning dramatically improved computers' abilities to recognise images, process audio and play games. By the late 2010s computers could do many of these tasks better than any human.

But neural networks tended to be embedded in software with broader functionality, like email clients, and non-coders rarely interacted with these AIS directly. Those that did often described their experience in near-spiritual terms. Lee Sedol, one of the world's best players of **Go**, an ancient Chinese board game, retired from the game after Alphabet's neural-net-based **AlphaGo** software crushed him in 2016. "Even if I become the number one," he said, "there is an entity that cannot be defeated."

By working in the most human of mediums, conversation, ChatGPT is now allowing the internet-using public to experience something similar, a kind of intellectual vertigo caused by software which has improved suddenly to the point where it can perform tasks that had been exclusively in the domain of human intelligence.

Despite that feeling of magic, an LLM is, in reality, a giant exercise in statistics. Prompt ChatGPT to finish the sentence: “The promise of large language models is that they…” and you will get an immediate response. How does it work?
First, the language of the query is **converted from words, which neural networks cannot handle, into a representative set of numbers** (see graphic). GPT-3, which powered an earlier version of ChatGPT, does this by splitting text into chunks of characters, called **tokens**, which commonly occur together. These tokens can be words, like “love” or “are”, affixes, like “dis” or “ised”, and punctuation, like “?”. GPT-3’s dictionary contains details of 50,257 tokens.

## Tokenisation

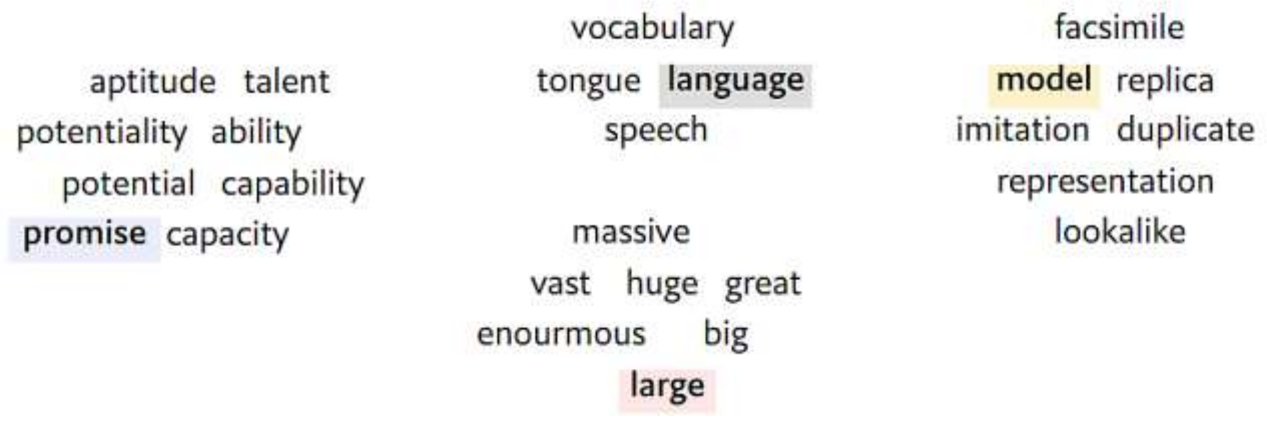


| The | promise | of | large | language | models | is | that | they |
|---|---|---|---|---|---|---|---|---|
| 464 | 6991 | 286 | 1588 | 3303 | 4981 | 318 | 326 | 484 |

GPT-3 is able to process a maximum of 2,048 tokens at a time, which is around the length of a long article in The Economist. GPT-4, by contrast, can handle inputs up to 32,000 tokens long - a novella. The more text the model can take in, the more context it can see, and the better its answers will be. There is a catch—the required computation rises non-linearly with the length of the input, meaning slightly longer inputs need much more computing power.

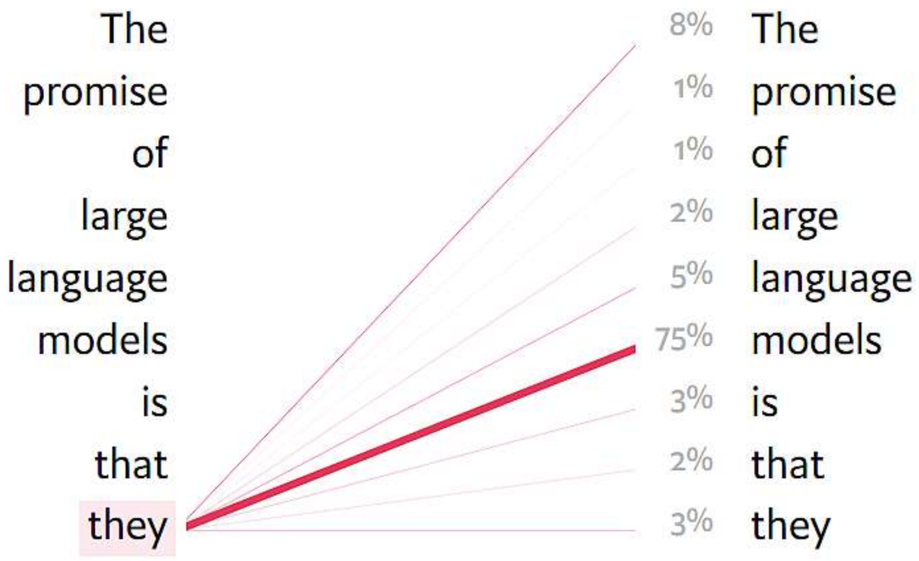
The tokens are then **assigned the equivalent of definitions** by placing them into a “meaning space” where words that have similar meanings are located in nearby areas.

**Embedding**



The LLM then deploys its "**attention network**" to make connections between different parts of the prompt. Someone reading our prompt, "the promise of large language models is that they…", would know how English grammar works and understand the concepts behind the words in the sentence. It would be obvious to them which words relate to each other—it is the model that is large, for example. An LLM, however, must learn these associations from scratch during its training phase—over billions of training runs**, its attention network slowly encodes the structure of the language it sees as numbers (called "weights") within its neural network.** If it understands language at all, an LLM only does so in a statistical, rather than a grammatical, way. It is much more like an abacus than it is like a mind.

**Attention**

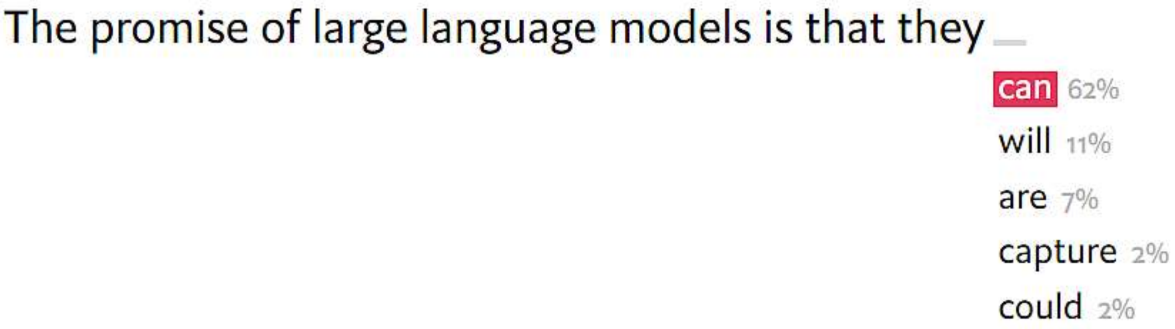


Once the prompt has been processed, the LLM initiates a response. At this point, for each of the tokens in the model's vocabulary, the attention network has produced a probability of that token being the most appropriate one to use next in the sentence it is generating. The token with the highest probability score is not always the one chosen for the response—how the LLM makes this choice depends on how creative the model has been told to be by its operators.

The LLM **generates a word** and then feeds the result back into itself. The first word is generated based on the prompt alone. The second word is generated by including the first word in the response, then the third word by including the first two generated words, and so on. This process—called autoregression—repeats until the LLM has finished

**Completion**

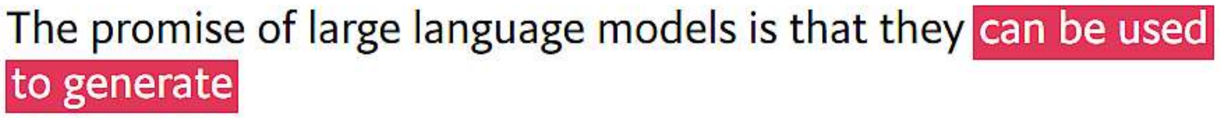
**a)**



The promise of large language models is that they

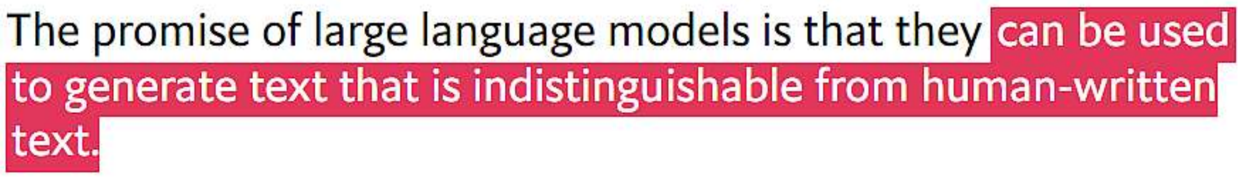
|  |  |
|---|---|
| can | 62% |
| will | 11% |
| are | 7% |
| capture | 2% |
| could | 2% |

b)



The promise of large language models is that they can be used to generate

c)



The promise of large language models is that they can be used to generate text that is indistinguishable from human-written text.

Although it is possible to write down the rules for how they work, LLMS' outputs are not entirely predictable; it turns out that these extremely big abacuses can do things which smaller ones cannot, in ways which surprise even the people who make them. Jason Wei, a researcher at OpenAI, has counted 137 so-called “emergent” abilities across a variety of different LLMS.

The abilities that emerge are not magic—they are all represented in some form within the LLMS’ training data (or the prompts they are given) but they do not become apparent until the LLMS cross a certain, very large, threshold in their size. At one size, an LLM does not know how to write gender-inclusive sentences in German any better than if it was doing so at random. Make the model just a little bigger, however, and all of a sudden a new ability pops out. GPT-4 passed the American Uniform Bar Examination, designed to test the skills of lawyers before they become licensed, in the 90th percentile. The slightly smaller GPT-3.5 flunked it.

Emergent abilities are exciting, because they hint at the untapped potential of LLMS. Jonas Degrave, an engineer at DeepMind, an AI research company owned by Alphabet, has shown that ChatGPT can be convinced to act like the command line terminal of a computer, appearing to compile and run programs accurately. Just a little bigger, goes the thinking, and the models may suddenly be able to do all manner of useful new things. But experts worry for the same reason. One analysis shows that certain social biases emerge when models become large. It is not easy to tell what harmful behaviours might be lying dormant, waiting for just a little more scale in order to be unleashed.

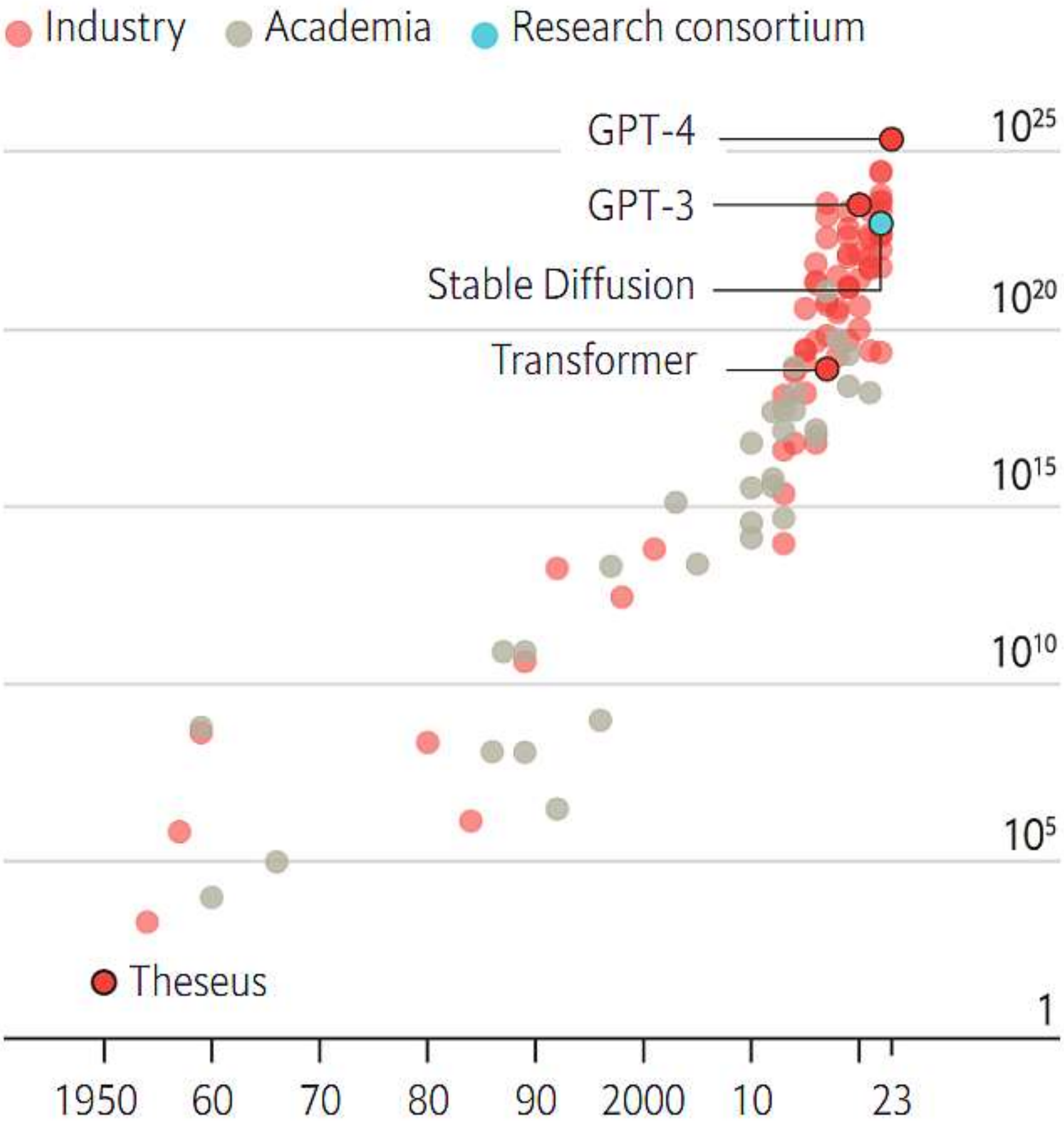
### Process the data

The recent success of LLMS in generating convincing text, as well as their startling emergent abilities, is due to the coalescence of three things: gobsmacking quantities of data, algorithms capable of learning from them and the computational power to do so (see chart on next page). The details of GPT-4’s construction and function are not yet public, but those of GPT-3 are, in a paper called “Language Models are Few-Shot Learners”, published in 2020 by OpenAI.

## Faster, higher, more calculations

Computing power used in training AI systems

Selected systems, floating-point operations, log scale



Sources: Sevilla et al., 2023; Our World in Data

Before it sees any training data, the weights in GPT-3's neural network are mostly random. As a result, any text it generates will be gibberish. Pushing its output towards something which makes sense, and eventually something that is fluent, requires training. GPT-3 was trained on several sources of data, but the bulk of it comes from snapshots of the entire internet between 2016 and 2019 taken from a database called Common Crawl. There's a lot of junk text on the internet, so the initial 45 terabytes were filtered using a different machine-learning model to select just the high-quality text: 570 gigabytes of it, a dataset that could fit on a modern laptop. In addition, GPT-4 was trained on an unknown quantity of images, probably several terabytes. By comparison AlexNet, a neural network that reignited image-processing excitement in the 2010s, was trained on a dataset of 1.2m labelled images, a total of 126 gigabytes—less than a tenth of the size of GPT-4's likely dataset.

To train, the LLM quizzes itself on the text it is given. It takes a chunk, covers up some words at the end, and tries to guess what might go there. Then the LLM uncovers the answer and compares it to its guess. Because the answers are in the data itself, these models can be trained in a "self-supervised" manner on massive datasets without requiring human labellers.

The model's goal is to make its guesses as good as possible by making as few errors as possible. Not all errors are equal, though. If the original text is "I love ice cream", guessing "I love ice hockey" is better than "I love ice are". How bad a guess is, is turned into a number called the loss. After a few guesses, the loss is sent back into the neural network and used to nudge the weights in a direction that will produce better answers.

### Trailblazing a daze

The LLM'S attention network is key to learning from such vast amounts of data. It builds into the model a way to learn and use associations between words and concepts even when they appear at a distance from each other within a text, and it allows it to process reams of data in a reasonable amount of time. Many different attention networks operate in parallel within a typical LLM and this parallelisation allows the process to be run across multiple GPUS. Older, non-attention-based versions of language models would not have been able to process such a quantity of data in a reasonable amount of time. "Without attention, the scaling would not be computationally tractable," says Yoshua Bengio, scientific director of Mila, a prominent AI research institute in Quebec. The sheer scale at which LLMS can process data has been driving their recent growth. GPT-3 has hundreds of layers, billions of weights, and was trained on hundreds of billions of words. By contrast, the first version of GPT, created five years ago, was just one ten-thousandth of the size.

But there are good reasons, says Dr Bengio, to think that this growth cannot continue indefinitely. The inputs of LLMS—data, computing power, electricity, skilled labour—cost money. Training GPT-3, for example, used 1.3 gigawatt-hours of electricity (enough to power 121 homes in America for a year), and cost OpenAI an estimated $4.6m. GPT-4, which is a much larger model, will have cost disproportionately more (in the realm of $100m) to train. Since computing-power requirements scale up dramatically faster than the input data, training LLMS gets expensive faster than it gets better. Indeed, Sam Altman, the boss of OpenAI, seems to think an inflection point has already arrived. On April 13th he told an audience at the Massachusetts Institute of Technology: "I think we're at the end of the era where it's going to be these, like, giant, giant models. We'll make them better in other ways."

But the most important limit to the continued improvement of LLMS is the amount of training data available. GPT-3 has already been trained on what amounts to all of the high-quality text that is available to download from the internet. A paper published in October 2022 concluded that "the stock of high-quality language data will be exhausted soon; likely before 2026." There is certainly more text available, but it is locked away in small amounts in corporate databases or on personal devices, inaccessible at the scale and low cost that Common Crawl allows.

Computers will get more powerful over time, but there is no new hardware forthcoming which offers a leap in performance as large as that which came from using GPUS in the early 2010s, so training larger models will probably be increasingly expensive—perhaps why Mr Altman is not enthused by the idea. Improvements are possible, including new kinds of chips such as Google's Tensor Processing Unit, but the manufacturing of chips is no longer improving exponentially through Moore's law and shrinking circuits.

There will also be legal issues. Stability AI, a company which produces an image-generation model called Stable Diffusion, has been sued by Getty Images, a photography agency. Stable Diffusion's training data comes from the same place as GPT-3 and GPT-4, Common Crawl, and it processes it in very similar ways, using attention networks. Some of the most striking examples of AI's generative prowess have been **images**. People on the internet are now regularly getting caught up in excitement about apparent photos of scenes that never took place: the pope in a Balenciaga jacket; Donald Trump being arrested.

Getty points to images produced by Stable Diffusion which contain its copyright watermark, suggesting that Stable Diffusion has ingested and is reproducing copyrighted material without permission (Stability AI has not yet commented publicly on the lawsuit). The same level of evidence is harder to come by when examining ChatGPT's text output, but there is no doubt that it has been trained on copyrighted material. OpenAI will be hoping that its text generation is covered by "fair use", a provision in copyright law that allows limited use of copyrighted material for "transformative" purposes. That idea will probably one day be tested in court.

### A major appliance

But even in a scenario where LLMS stopped improving this year, and a blockbuster lawsuit drove OpenAI to bankruptcy, the power of large language models would remain. The data and the tools to process it are widely available, even if the sheer scale achieved by OpenAI remains expensive.

Open-source implementations, when trained carefully and selectively, are already aping the performance of GPT-4. This is a good thing: having the power of LLMS in many hands means that many minds can come up with innovative new applications, improving everything from medicine to the law.

But it also means that the catastrophic risk which keeps the tech elite up at night has become more imaginable. LLMS are already incredibly powerful and have improved so quickly that many of those working on them have taken fright. The capabilities of the biggest models have outrun their creators' understanding and control. That creates risks, of all kinds. . ■

## Related Articles

**Artificial intelligence - Wikipedia**
https://en.wikipedia.org/wiki/Artificial_intelligence

**Artificial intelligence Encyclopaedia Britanica**
https://www.britannica.com/technology/artificial-intelligence/The-Turing-test