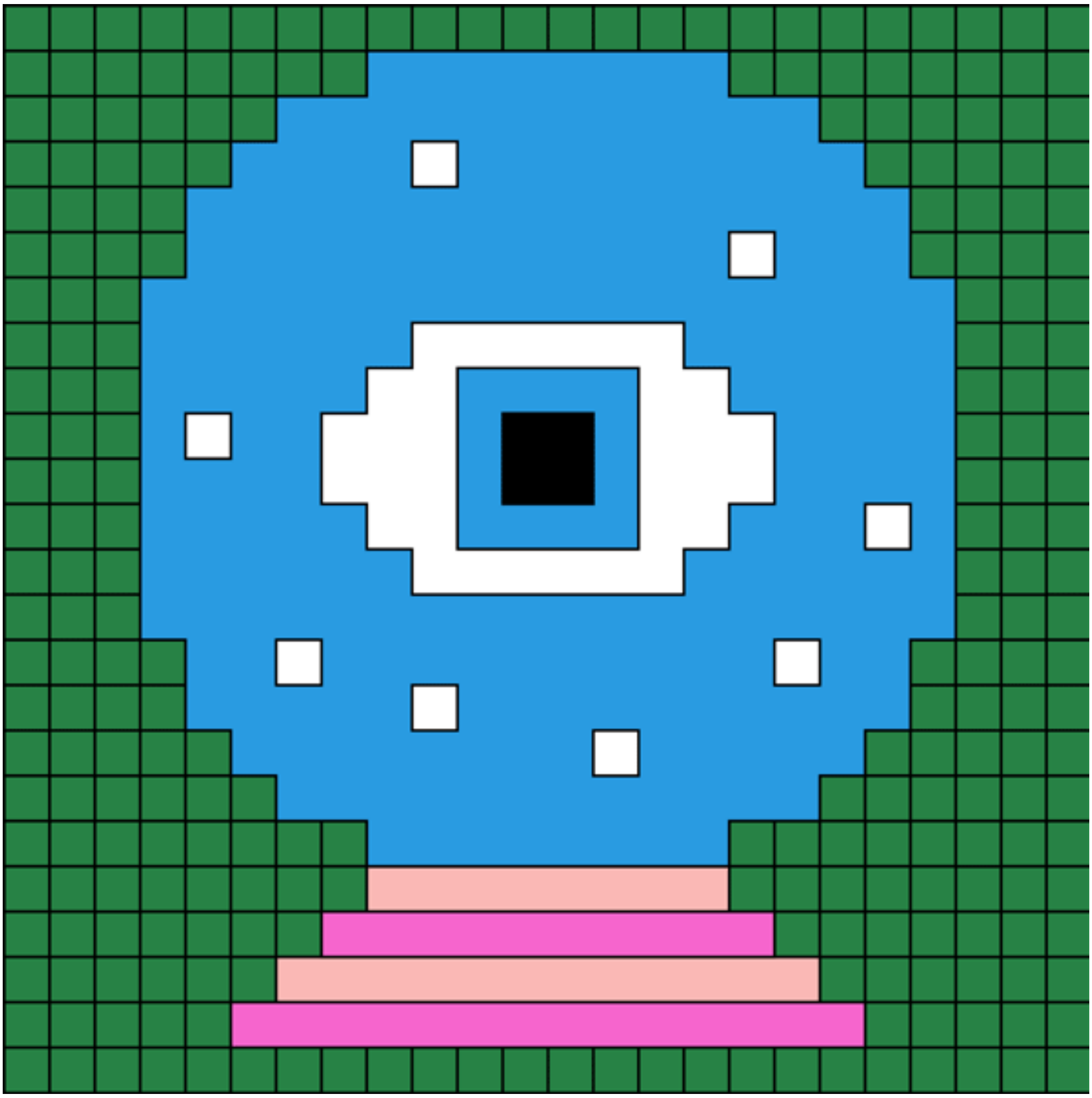
The New York Times

On Tech
Artificial Intelligence

FOR SUBSCRIBERS | MARCH 28, 2023

Continue reading the main story

PART 2 ● ● ○ ○ ○



Illustrations by Mathieu Labrecque

How does ChatGPT really work?

By Kevin Roose
Technology Columnist

In the second of our five-part series, I'm going to explain how the technology actually works.

The artificial intelligences that powers ChatGPT, Microsoft's Bing chatbot and Google's Bard can carry out humanlike conversations and write natural, fluid prose on an endless variety of topics. They can also perform complex tasks, from writing code to planning a kid's birthday party.

But how does it all work? To answer that, we need to peek under the hood of something called a large language model — the type of A.I. that drives these systems.

Large language models, or L.L.M.s, are relatively new on the A.I. scene. The first ones appeared only about five years ago, and they weren't very good. But today they can draft emails, presentations and memos and tutor you in a foreign language. Even more capabilities are sure to surface in the coming months and years, as the technology improves and Silicon Valley scrambles to cash in.

I'm going to walk you through setting one a large language model from scratch, simplifying things and leaving out a lot of hard math. Let's pretend that we're trying to build an L.L.M. to help you with replying to your emails. We'll call it MailBot.

Step 1: Set a goal

Every A.I. system needs a goal. Researchers call this an **objective function**. It can be simple — for example, "win as many chess games as possible" — or complicated, like "predict the three-dimensional shapes of proteins, using only their amino acid sequences."

Most large language models have the same basic objective function: Given a sequence of text, guess what comes next. We'll give MailBot more specific goals later on, but let's stick to that one for now.

Step 2: Collect lots of data

Next, we need to assemble the training data that will teach MailBot how to write. Ideally, we'll put together a colossally large repository of text, which usually means billions of pages scraped from the internet — like blog posts, tweets, Wikipedia articles and news stories.

To start, we'll use some free, publicly available data libraries, such as the Common Crawl repository of web data. But we'll also want to add our own secret sauce, in the form of proprietary or specialized data. Maybe we'll license some foreign-language text, so that MailBot learns to compose emails in French or Spanish as well as English. In general, the more data we have, and the more diverse the sources, the better our model will be.

Before we can feed the data into our model, we need to break it down into units called tokens, which can be words, phrases or even individual characters. Transforming text into bite-size chunks helps a model analyze it more easily.

Step 3: Build your neural network

Once our data is tokenized, we need to assemble the A.I.'s "brain" — a type of system known as a neural network. This is a complex web of interconnected nodes (or "neurons") that process and store information.

For MailBot, we're going to want to use a relatively new type of neural network known as a **transformer model**. They can analyze multiple pieces of text at the same time, making them faster and more efficient. (Transformer models are the key to systems like ChatGPT — whose full acronym stands for "Generative Pretrained Transformer.")

Step 4: Train your neural network

Next, the model will analyze the data, token by token, identifying patterns and relationships. It might notice "Dear" is often followed by a name, or that "Best regards" typically comes before your name. By identifying these patterns, the A.I. learns how to construct messages that make sense.

The system also develops a sense of context. For example, it might learn that "bank" can refer to a financial institution or the side of a river, depending on the surrounding words.

As it learns these patterns, the transformer model sketches a map: an enormously complex mathematical representation of human language. It keeps track of these relationships using numerical values known as **parameters**. Many of today’s best L.L.M.s have hundreds of billions of parameters or more.

Training could take days or even weeks, and will require immense amounts of computing power. But once it’s done, it will almost be ready to start writing your emails.

Weirdly, it may develop other skills, too. As L.L.M.s learn to predict the next word in a sequence, over and over and over again, they can pick up other, unexpected abilities, such as knowing how to code. A.I. researchers call these emergent behaviors, and they’re still sometimes mystified by them.

Step 5: Fine-tune your model

Once a large language model is trained, it needs to be calibrated for a specific job. A chatbot used by a hospital might need to understand medical terms, for example.

To fine-tune MailBot, we could ask it to generate a bunch of emails, hire people to rate them on accuracy and then feed the ratings back into the model until it improves.

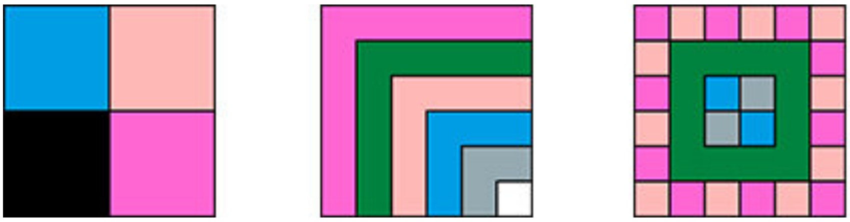
This is a rough approximation of the approach that was used with ChatGPT, which is known as **reinforcement learning with human feedback**.

Step 6: Launch, carefully

Congratulations! Once MailBot has been trained and fine-tuned, it’s ready to use. After you build some kind of user interface for it — like a Chrome extension that plugs into your email app — it can start cranking out emails.

But no matter how good it seems, you’re still going to want to keep tabs on your new assistant. As companies like Microsoft and Meta have learned the hard way, A.I. systems can be erratic and unpredictable, or even turn creepy and dangerous.

Tomorrow, we'll hear more about how things can go wrong in unexpected and sometimes disturbing ways.



Your homework

Let's explore one of the more creative abilities of L.L.M.s: the ability to combine disparate concepts and formats into something bizarre and new. For example, our colleagues at Well asked ChatGPT to "write a song in Taylor Swift's voice that uses themes from a Dr. Seuss book."

For today's homework, try to mix and match a format, a style and a topic — like, "Write a limerick in the style of Snoop Dogg about global warming."

Don't forget to share your creation as a comment here.

Continue reading the main story

Quiz

Question 1 of 3

What is the primary objective function of large language models like ChatGPT?

- Win as many chess games as possible
- Predict the 3-D shapes of proteins
- Guess what comes next in a sequence of text

Start the quiz by choosing your answer.

Glossary

- **Transformer model:** A neural network architecture useful for understanding language, which does not have to analyze words one at a time but can look at an entire sentence at once. A technique called self-attention allows the model to focus on the particular words that are important in understanding the meaning of the sentence.
- **Parameters:** Numerical values that define a large language model's structure and behavior, like clues that help it guess what words come next. Modern systems like GPT-4 are thought to have hundreds of billions of parameters.
- **Reinforcement learning:** A technique that teaches an A.I. model to find the best result by trial and error, receiving rewards or punishments from an algorithm based on its results. This system can be enhanced by humans giving feedback on its performance.