ChatGPT is everywhere. Here's where it came from

OpenAI's breakout hit was an overnight sensation—but it is built on decades of research.

Will Douglas Heaven

Tech Review Explains: Let our writers untangle the complex, messy world of technology to help you understand what's coming next. You can read more here.

We've reached peak ChatGPT. Released at the end of November as a web app by the San Francisco–based firm OpenAI, the chatbot exploded into the mainstream almost overnight. According to some estimates, it is the fastest-growing internet service ever, reaching 100 million users in January, just two months after launch.

Through OpenAI's $10 billion deal with Microsoft, the tech is now being built into Office software and the Bing search engine. Stung into action by its newly awakened onetime rival in the battle for search, Google is fast-tracking the rollout of its own chatbot, based on its large language model PaLM. Even my family WhatsApp is filled with ChatGPT chat.

But OpenAI's breakout hit did not come out of nowhere. The chatbot is the most polished iteration to date in a line of large language models going back years. This is how we got here.

1980s–'90s: Recurrent Neural Networks

ChatGPT is a version of GPT-3, a large language model also developed by OpenAI. A large language model (or LLM) is a type of neural network that has been trained on lots and lots of text. (Neural networks are software inspired by the way neurons in animal brains signal one another.) Because text is made up of sequences of letters and words of varying lengths, language models require a type of neural network that can make sense of that kind of data. Recurrent neural networks, invented in the 1980s, can handle sequences of words, but they are slow to train and can forget previous words in a sequence.

In 1997, computer scientists Sepp Hochreiter and Jürgen Schmidhuber fixed this by inventing **LSTM (Long Short-Term Memory)** networks, recurrent neural networks with special components that allowed past data in an input sequence to be retained for longer. LSTMs could handle strings of text several hundred words long, but their language skills were limited.

2017: Transformers

The breakthrough behind today's generation of large language models came when a team of Google researchers invented transformers, a kind of neural network that can track where each word or phrase appears in a sequence. The meaning of words often depends on the meaning of other words that come before or after. By tracking this contextual information, transformers can handle longer strings of text and capture the meanings of words more accurately. For example, "hot dog" means very different things in the sentences "A hot dog should be given lots of water" and "A hot dog should be eaten with mustard."

2018–2019: GPT and GPT-2

OpenAI's first two large language models came just a few months apart. The company wants to develop multi-skilled, general-purpose AI and believes that large language models are a key step toward that goal. GPT (short for Generative Pre-trained Transformer) planted a flag, beating state-of-the-art benchmarks for natural-language processing at the time.

GPT combined transformers with unsupervised learning, a way to train machine-learning models on data (in this case, lots and lots of text) that hasn't been annotated beforehand. This lets the software figure out patterns in the data by itself, without having to be told what it's looking at. Many previous successes in machine-learning had relied on supervised learning and annotated data, but labeling data by hand is slow work and thus limits the size of the data sets available for training.

But it was GPT-2 that created the bigger buzz. OpenAI claimed to be so concerned people would use GPT-2 "to generate deceptive, biased, or abusive language" that it would not be releasing the full model. How times change.

2020: GPT-3

GPT-2 was impressive, but OpenAI's follow-up, GPT-3, made jaws drop. Its ability to generate human-like text was a big leap forward. GPT-3 can answer questions, summarize documents, generate stories in different styles, translate between English, French, Spanish, and Japanese, and more. Its mimicry is uncanny.

One of the most remarkable takeaways is that GPT-3's gains came from supersizing existing techniques rather than inventing new ones. GPT-3 has 175 billion parameters (the values in a network that get adjusted during training), compared with GPT-2's 1.5 billion. It was also trained on a lot more data.

But training on text taken from the internet brings new problems. GPT-3 soaked up much of the disinformation and prejudice it found online and reproduced it on demand. As OpenAI acknowledged: “Internet-trained models have internet-scale biases.”

December 2020: Toxic text and other problems

While OpenAI was wrestling with GPT-3’s biases, the rest of the tech world was facing a high-profile reckoning over the failure to curb toxic tendencies in AI. It’s no secret that large language models can spew out false—even hateful—text, but researchers have found that fixing the problem is not on the to-do list of most Big Tech firms. When Timnit Gebru, co-director of Google’s AI ethics team, coauthored a paper that highlighted the potential harms associated with large language models (including high computing costs), it was not welcomed by senior managers inside the company. In December 2020, Gebru was pushed out of her job.

January 2022: InstructGPT

OpenAI tried to reduce the amount of misinformation and offensive text that GPT-3 produced by using reinforcement learning to train a version of the model on the preferences of human testers (a technique called reinforcement learning from human feedback, or RLHF). The result, InstructGPT, was better at following the instructions of people using it—known as “alignment” in AI jargon—and produced less offensive language, less misinformation, and fewer mistakes overall. In short, InstructGPT is less of an asshole—unless it’s asked to be one.

May–July 2022: OPT, BLOOM

A common criticism of large language models is that the cost of training them makes it hard for all but the richest labs to build one. This raises concerns that such powerful AI is being built by small corporate teams behind closed doors, without proper scrutiny and without the input of a wider research community. In response, a handful of collaborative projects have developed large language models and released them for free to any researcher who wants to study—and improve—the technology. Meta built and gave away OPT, a reconstruction of GPT-3. And Hugging Face led a consortium of around 1,000 volunteer researchers to build and release BLOOM.

December 2022: ChatGPT

Even OpenAI is blown away by how ChatGPT has been received. In the company's first demo, which it gave me the day before ChatGPT was launched online, it was pitched as an incremental update to InstructGPT. Like that model, ChatGPT was trained using reinforcement learning on feedback from human testers who scored its performance as a fluid, accurate, and inoffensive interlocutor. In effect, OpenAI trained GPT-3 to master the game of conversation and invited everyone to come and play. Millions of us have been playing ever since.