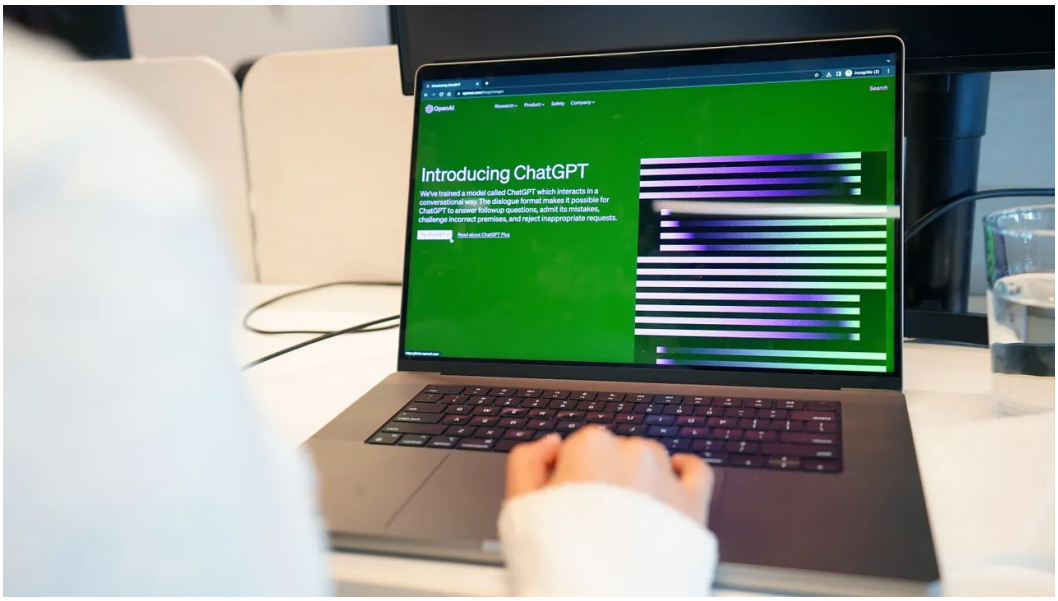
How does ChatGPT work?

We take a deep dive into the inner workings of the wildly popular AI chatbot, ChatGPT. If you want to know how its generative AI magic happens, read on.

Written by **David Gewirtz,** Senior Contributing Editor on March 10, 2023
Reviewed by **Alyson Windsor - ZDNet**



Google, Wolfram Alpha, and ChatGPT all interact with users via a single line text entry field and provide text results. Google returns search results, a list of web pages and articles that will (hopefully) provide information related to the search queries. Wolfram Alpha generally provides mathematically and data analysis-related answers.

Also: How to use ChatGPT: Everything you need to know

ChatGPT, by contrast, provides a response based on the context and intent behind a user's question. You can't, for example, ask Google to write a story or ask Wolfram Alpha to write a code module, but ChatGPT can do these sorts of things.

Fundamentally, Google's power is the ability to do enormous database lookups and provide a series of matches. Wolfram Alpha's power is the ability to parse data-related questions and perform calculations based on those questions. ChatGPT's power is the ability to parse queries and produce fully-fleshed out answers and results based on most of the world's digitally-accessible text-based information -- at least information that existed as of its time of training prior to 2021.

In this article, we'll look at how ChatGPT can produce those fully-fleshed out answers. We'll start by looking at the main phases of ChatGPT operation, then cover some of the core AI architecture components that make it all work.

In addition to the sources cited in this article (many of which are the original research papers behind each of the technologies), I used ChatGPT itself to help me create this

backgrounder. I asked it a lot of questions. Some answers are paraphrased within the overall context of this discussion.

The two main phases of ChatGPT operation

Let's use Google as an analogy again. When you ask Google to look up something, you probably know that it doesn't -- at the moment you ask -- go out and scour the entire web for answers. Instead, Google searches its database for pages that match that request. Google effectively has two main phases: the spidering and data gathering phase, and the user interaction/lookup phase.

Also: The best AI chatbots: ChatGPT and other fun alternatives to try

Roughly speaking, ChatGPT works the same way. The data gathering phase is called pre-training, while the user responsiveness phase is called inference. The magic behind generative AI and the reason it's suddenly exploded is that the way pre-training works has suddenly proven to be enormously scalable.

Pre-training the AI

Generally speaking (because to get into specifics would take volumes), AIs pre-train using two principle approaches: supervised and non-supervised. For most AI projects up until the current crop of generative AI systems like ChatGPT, the supervised approach was used.

Supervised pre-training is a process where a model is trained on a labeled dataset, where each input is associated with a corresponding output.

Also: 6 things ChatGPT can't do (and another 20 it refuses to do)

For example, an AI could be trained on a dataset of customer service conversations, where the user's questions and complaints are labeled with the appropriate responses from the customer service representative. To train the AI, questions like "How can I reset my password?" would be provided as user input, and answers like "You can reset your password by visiting the account settings page on our website and following the prompts." would be provided as output.

In a supervised training approach, the overall model is trained to learn a mapping function that can map inputs to

outputs accurately. This process is often used in supervised learning tasks, such as classification, regression, and sequence labeling.

As you might imagine, there are limits to how this can scale. Human trainers would have to go pretty far in anticipating all the inputs and outputs. Training could take a very long time and be limited in subject matter expertise.

But as we've come to know, ChatGPT has very few limits in subject matter expertise. You can ask it to write a resume for the character Chief Miles O'Brien from Star Trek, have it explain quantum physics, write a piece of code, write a short piece of fiction, and compare the governing styles of former presidents of the United States.

Also: I asked ChatGPT to write a short Star Trek episode. It actually succeeded

It would be impossible to anticipate all the questions that would ever be asked, so there really is no way that ChatGPT could have been trained with a supervised model. Instead, ChatGPT uses non-supervised pre-training -- and this is the game changer.

Non-supervised pre-training is the process by which a model is trained on data where no specific output is associated with each input. Instead, the model is trained to

learn the underlying structure and patterns in the input data without any specific task in mind. This process is often used in unsupervised learning tasks, such as clustering, anomaly detection, and dimensionality reduction. In the context of language modeling, non-supervised pre-training can be used to train a model to understand the syntax and semantics of natural language, so that it can generate coherent and meaningful text in a conversational context.

It's here where ChatGPT's apparently limitless knowledge becomes possible. Because the developers don't need to know the outputs that come from the inputs, all they have to do is dump more and more information into the ChatGPT pre-training mechanism, which is called transformer-base language modeling.

Transformer architecture

The transformer architecture is a type of neural network that is used for processing natural language data. A neural network simulates the way a human brain works by processing information through layers of interconnected nodes. Think of a neural network like a hockey team: each player has a role, but they pass the puck back and forth among players with specific roles, all working together to score the goal.

The transformer architecture processes sequences of words by using "self-attention" to weigh the importance of different words in a sequence when making predictions. Self-attention is similar to the way a reader might look back at a previous sentence or paragraph for the context needed to understand a new word in a book. The transformer looks at all the words in a sequence to understand the context and the relationships between the words.

Also: I asked ChatGPT to write a WordPress plugin I needed. It did it in less than 5 minutes

The transformer is made up of several layers, each with multiple sub-layers. The two main sub-layers are the self-attention layer and the feedforward layer. The self-attention layer computes the importance of each word in the sequence, while the feedforward layer applies non-linear transformations to the input data. These layers help the transformer learn and understand the relationships between the words in a sequence.

During training, the transformer is given input data, such as a sentence, and is asked to make a prediction based on that input. The model is updated based on how well its prediction matches the actual output. Through this process, the transformer learns to understand the context and relationships between words in a sequence, making it

a powerful tool for natural language processing tasks such as language translation and text generation.

Let's discuss the data that gets fed into ChatGPT first, and then take a look at the user-interaction phase of ChatGPT and natural language.

ChatGPT's training datasets

The dataset used to train ChatGPT is huge. ChatGPT is based on the GPT-3 (Generative Pre-trained Transformer 3) architecture. Now, the abbreviation GPT makes sense, doesn't it? It's generative, meaning it generates results, it's pre-trained, meaning it's based on all this data it ingests, and it uses the transformer architecture that weighs text inputs to understand context.

GPT-3 was trained on a dataset called WebText2, a library of over 45 terabytes of text data. When you can buy a 16 terabyte hard drive for under $300, a 45 terabyte corpus may not seem that large. But text takes up a lot less storage space than pictures or video.

Also: These experts are racing to protect AI from hackers. Time is running out

This massive amount of data allowed ChatGPT to learn patterns and relationships between words and phrases in natural language at an unprecedented scale, which is one of the reasons why it is so effective at generating coherent and contextually relevant responses to user queries.

While ChatGPT is based on the GPT-3 architecture, it has been fine-tuned on a different dataset and optimized for conversational use cases. This allows it to provide a more personalized and engaging experience for users who interact with it through a chat interface.

For example, OpenAI (developers of ChatGPT) has released a dataset called Persona-Chat that is specifically designed for training conversational AI models like ChatGPT. This dataset consists of over 160,000 dialogues between two human participants, with each participant assigned a unique persona that describes their background, interests, and personality. This allows ChatGPT to learn how to generate responses that are personalized and relevant to the specific context of the conversation.

Also: How to save a ChatGPT conversation to revisit later

In addition to Persona-Chat, there are many other conversational datasets that were used to fine-tune ChatGPT. Here are a few examples:

- **Cornell Movie Dialogs Corpus:** a dataset containing conversations between characters in movie scripts. It includes over 200,000 conversational exchanges between more than 10,000 movie character pairs, covering a diverse range of topics and genres.
- **Ubuntu Dialogue Corpus:** a collection of multi-turn dialogues between users seeking technical support and the Ubuntu community support team. It contains over 1 million dialogues, making it one of the largest publicly available datasets for research on dialog systems.
- **DailyDialog:** a collection of human-to-human dialogues in a variety of topics, ranging from daily life conversations to discussions about social issues. Each dialogue in the dataset consists of several turns, and is labeled with a set of emotion, sentiment, and topic information.

In addition to these datasets, ChatGPT was trained on a large amount of unstructured data found on the internet, including websites, books, and other text sources. This allowed ChatGPT to learn about the structure and patterns of language in a more general sense, which could then be fine-tuned for specific applications like dialogue management or sentiment analysis.

ChatGPT is a distinct model that was trained using a similar approach as the GPT series, but with some differences in architecture and training data. ChatGPT has

1.5 billion parameters, which is smaller than GPT-3's 175 billion parameters.

Also: The best AI art generators: DALL-E 2 and other fun alternatives to try

Overall, the training data used to fine-tune ChatGPT is typically conversational in nature and specifically curated to include dialogues between humans, which allows ChatGPT to learn how to generate natural and engaging responses in a conversational format.

Think of ChatGPT's unsupervised training this way: It was fed a lot of data and left to its own devices to find patterns and make sense of it all. This is the mechanism that allowed the new generative AI systems to scale up so quickly.

While the heavy lifting for ChatGPT's generative AI is being done by the pre-training, it also has to be able to understand questions and construct answers from all that data. That's done by the inference phase which consists of natural language processing and dialog management.

What about human involvement in pre-training?

Despite the inherent scalability of non-supervised pre-training, there is some evidence that human assistance may have been involved in the preparation of ChatGPT for public use.

The big reveal was in an article in TIME Magazine that discussed the use of human "data labelers" earning between $1.32 and $2/hour in Kenya. According to the TIME report, it was the responsibility of these workers to scan horrifying and sexually explicit Internet content to flag it for ChatGPT training.

Also: I'm using ChatGPT to help me fix code faster, but at what cost?

Another article, posted in Martechpost, an AI newsletter, states that the large language model was trained using a process called Reinforcement Learning from Human Feedback (RLHF). According to the article, "The training process involved a fine-tuned initial model using supervised learning, with human trainers playing both the role of the user and an AI assistant."

There are some subtleties here, when it comes to what the word "training" means. According to ChatGPT itself, "OpenAI did not use reinforcement learning with human feedback to train me. Instead, I was pre-trained using a combination of unsupervised and supervised learning techniques, such as language modeling, auto-encoding, and sequence prediction. My training involved processing massive amounts of text data from the internet, which allowed me to learn patterns and relationships between words and phrases."

The AI continued, "However, once I was pre-trained, researchers and developers can use reinforcement learning with human feedback to fine-tune me for specific tasks or domains, such as answering questions or generating text. In these cases, humans can provide feedback in the form of rewards or penalties, which can be used to update my parameters and improve my performance on similar tasks in the future."

Also: Generative AI could lower drug prices. Here's how

This seems to fit with the Marktechpost and TIME reports, in that the initial pre-training was non-supervised, allowing for a tremendous amount of data to be fed into the system. But in building the dialog responses that communicate with users (more on that below), the response engines were apparently trained both on the types of responses,

and trained to filter out inappropriate material -- and that training seems to have been human assisted.

I reached out to OpenAI (the maker of ChatGPT) for clarification, but haven't yet gotten a response. If the company gets back to me (outside of ChatGPT itself), I'll update the article with its answer.

Natural language processing

Natural language processing (NLP) focuses on enabling computers to understand, interpret, and generate human language. With the exponential growth of digital data and the increasing use of natural language interfaces, NLP has become a crucial technology for many businesses.

NLP technologies can be used for a wide range of applications, including sentiment analysis, chatbots, speech recognition, and translation. By leveraging NLP, businesses can automate tasks, improve customer service, and gain valuable insights from customer feedback and social media posts.

Also: What the New York Times and others are terribly getting wrong about ChatGPT

One of the key challenges in implementing NLP is dealing with the complexity and ambiguity of human language. NLP algorithms need to be trained on large amounts of data in order to recognize patterns and learn the nuances of language. They also need to be continually refined and updated to keep up with changes in language use and context.

The technology works by breaking down language inputs, such as sentences or paragraphs, into smaller components and analyzing their meanings and relationships to generate insights or responses. NLP technologies use a combination of techniques, including statistical modeling, machine learning, and deep learning, to recognize patterns and learn from large amounts of data in order to accurately interpret and generate language.

Dialogue management

You may have noticed that ChatGPT can ask follow-up questions to clarify your intent or better understand your needs, and provide personalized responses that take into account the entire conversation history.

This is how ChatGPT can have multi-turn conversations with users in a way that feels natural and engaging. It involves using algorithms and machine learning

techniques to understand the context of a conversation and maintain it over multiple exchanges with the user.

Also: Just how big is this generative AI? Think internet-level disruption

Dialogue management is an important aspect of natural language processing because it allows computer programs to interact with people in a way that feels more like a conversation than a series of one-off interactions. This can help to build trust and engagement with users, and ultimately lead to better outcomes for both the user and the organization using the program.

Marketers, of course, want to expand how trust is built up, but this is also an area that could prove scary because it's one way that an AI might be able to manipulate the people who use it.

And now you know

Even though we're pushing 2,500 words, this is still a very rudimentary overview of all that goes on inside of ChatGPT. That said, perhaps now you understand a bit more about why this technology has exploded over the past few months. The key to it all is that the data itself isn't "supervised," and the AI is able to take what it's been fed and make sense of it.

Also: ChatGPT and Bard: Are we looking for answers in all the wrong places?

Pretty awesome, really.

To wrap up, I fed a draft of this entire article to ChatGPT and asked the AI to describe the article in one sentence. Here you go:

ChatGPT is like Google and Wolfram Alpha's brainy cousin who can do things they can't, like write stories and code modules.

ChatGPT is supposed to be a technology without an ego, but if that answer doesn't just slightly give you the creeps, you haven't been paying attention.

What do you think? Are you using ChatGPT? What questions do you still have about how it works? Share with us in the comments below.

---

original article: https://www.zdnet.com/article/how-does-chatgpt-work/?ftag=TRE-03-10aaa6b&utm_email=14712f88bfda00eba532e946c7553b40ff3ee047fa5bd9c33127feb1458bc41a&utm_campaign_id=6364442&utm_email_id=90f9d59cc75582d7ecae68774dff1b734c1f34faae7b0270afa9d6e177d0eb72&utm_newsletter_id=92303&medium=email&source=iterable