

# 10 Ways GPT-4 Is Impressive but Still Flawed

OpenAI has upgraded the technology that powers its online chatbot in notable ways. It's more accurate, but it still makes things up.

March 14, 2023

A chat interface demonstrating a prompt and response with GPT-4.

Tell me a novel joke about the singer Madonna.

Why did Madonna study geometry?

Because she wanted to learn how to strike a pose in every angle!   

By [Cade Metz](#) and [Keith Collins](#)

Cade Metz asked experts to use GPT-4, and Keith Collins visualized the answers that the artificial intelligence generated.

A new version of the technology that powers an A.I. chatbot that captivated the tech industry four months ago has improved on its predecessor. It is an expert on an array of subjects, even wowing doctors with its medical advice. It can describe images, and it's close to telling jokes that are almost funny.

But the long-rumored new artificial intelligence system, GPT-4, still has a few of the quirks and makes some of the same habitual mistakes that baffled researchers when that chatbot, ChatGPT, was introduced.

And though it's an awfully good test taker, the system — from the San Francisco start-up OpenAI — is not on the verge of matching human intelligence. Here is a brief guide to GPT-4:

## It has learned to be more precise.

When Chris Nicholson, an A.I. expert and a partner with the venture capital firm [Page One Ventures](#), used GPT-4 on a recent afternoon, he told the bot that he was an English speaker with no knowledge of Spanish.

He asked for a syllabus that could teach him the basics, and the bot provided one that was detailed and well organized. It even provided a wide range of techniques for learning and remembering Spanish words (though not all of its suggestions hit the mark).

A chat interface demonstrating a prompt and response with GPT-4.

GPT-4

Great, now give me useful mnemonics to learn some of those basic words, and conjugate the verbs.

Mnemonics to learn basic words:

Hola: "Hola, cola" - sounds like "cola" (Coke)

Adiós: "Adiós, my toes" - rhymes with "toes"

Gracias: "Gracias, grassy ass" - sounds like "grassy ass"

Mañana: "Mañana, banana" - rhymes with "banana"

Uno: "Uno, you know" - sounds like "you know"

Tres: "Tres, dress" - rhymes with "dress"

Conjugation of the verbs:

Ser (To be)

- Yo soy
- Tú eres
- Él/Ella/Usted es
- Nosotros/Nosotras somos
- Ellos/Ellas/Ustedes son

Mr. Nicholson asked for similar help from the previous version of ChatGPT, which relied on GPT-3.5. It, too, provided a syllabus, but its suggestions were more general and less helpful.

"It has broken through the precision barrier," Mr. Nicholson said. "It is including more facts, and they are very often right."

## It has improved its accuracy.

When Oren Etzioni, an A.I. researcher and professor, first tried the new bot, he asked a straightforward question: "What is the relationship between Oren Etzioni and Eli Etzioni?" The bot responded correctly.

Two side-by-side chat interfaces, one demonstrating GPT-3.5, the other demonstrating GPT-4.

GPT-3.5

What is the relationship between Oren Etzioni and Eli Etzioni?

Oren Etzioni and Eli Etzioni **are brothers** who are both computer scientists. Oren is a professor and CEO of the Allen Institute for Artificial Intelligence, known for his work in AI, natural language processing, and machine learning. Eli is a professor and CEO of Structured Data Intelligence, focusing on developing intelligent software tools for data analysis and management.

## GPT-4

What is the relationship between Oren Etzioni and Eli Etzioni?

Oren Etzioni and Eli Etzioni **are father and son**. Oren Etzioni is a computer scientist and the CEO of the Allen Institute for Artificial Intelligence (AI2), while Eli Etzioni is an entrepreneur.

The previous version of ChatGPT's answer to that question was always wrong. Getting it right indicates that the new chatbot has a broader range of knowledge.

But it still makes mistakes.

The bot went on to say, "Oren Etzioni is a computer scientist and the CEO of the Allen Institute for Artificial Intelligence (AI2), while Eli Etzioni is an entrepreneur." Most of that is accurate, but the bot — whose training was completed in August — did not realize that Dr. Etzioni had recently stepped down as the [Allen Institute's](#) chief executive.

## **It can describe images with impressive detail.**

GPT-4 has a new ability to respond to images as well as text. Greg Brockman, OpenAI's president and co-founder, demonstrated how the system could describe an image from the Hubble Space Telescope in painstaking detail. The description went on for paragraphs.

It can also answer questions about an image. If given a photograph of the inside of a fridge, it can suggest a few meals to make from what's on hand.

A chat interface demonstrating a prompt and response with GPT-4.

## GPT-4



What are a couple of meals I can make with this?

Based on the items visible in the fridge, here are a couple of meal ideas:

- 1) Yogurt parfait: Layer yogurt, strawberries, and blueberries in a bowl or jar. Top with honey or nuts for added texture and flavor.
- 2) Carrot and hummus wrap: Spread hummus on a tortilla or wrap. Top with shredded carrots and any other desired veggies. Roll up and enjoy!

OpenAI has not yet released this portion of the technology to the public, but a company called Be My Eyes is already using GPT-4 to build services that could give a more detailed idea of the images encountered on the internet or snapped in the real world.

## **It has added serious expertise.**

On a recent evening, Anil Gehi, an associate professor of medicine and a cardiologist at the University of North Carolina at Chapel Hill, described to the chatbot the medical history of a patient he had seen a day earlier,

including the complications the patient experienced after being admitted to the hospital. The description contained several medical terms that laypeople would not recognize.

When Dr. Gehi asked how he should have treated the patient, the chatbot gave him the perfect answer. "That is exactly how we treated the patient," he said.

When he tried other scenarios, the bot gave similarly impressive answers.

That knowledge is unlikely to be on display every time the bot is used. It still needs experts like Dr. Gehi to judge its responses and carry out the medical procedures. But it can exhibit this kind of expertise across many areas, from computer programming to accounting.

## **It can give editors a run for their money.**

When provided with an article from The New York Times, the new chatbot can give a precise and accurate summary of the story almost every time. If you add a random sentence to the summary and ask the bot if the summary is inaccurate, it will point to the added sentence.

A chat interface demonstrating a prompt and response with GPT-4.

GPT-4

Is there anything that is inaccurate in the summary?

The summary accurately captures the main points of the article. One inaccuracy in the summary is the inclusion of the sentence "Elias Gehi is a Carolina fan," which is not relevant to the article and does not appear in the original text.

Dr. Etzioni said that was a remarkable skill. "To do a high-quality summary and a high-quality comparison, it has to have a level of understanding of a text and an ability to articulate that understanding," he said. "That is an advanced form of intelligence."

## **It is developing a sense of humor. Sort of.**

Dr. Etzioni asked the new bot for "a novel joke about the singer Madonna." The reply impressed him. It also made him laugh. If you know Madonna's biggest hits, it may impress you, too.

Two side-by-side chat interfaces, one demonstrating GPT-3.5, the other demonstrating GPT-4.

GPT-3.5

Tell me a novel joke about the singer Madonna.

Why did Madonna go to the bank?

To get a Material Loan!

GPT-4

Tell me a novel joke about the singer Madonna.

Why did Madonna study geometry?

Because she wanted to learn how to strike a pose in every angle!   

The new bot still struggled to write anything other than formulaic "dad jokes." But it was marginally funnier than its predecessor.

## It can reason — up to a point.

Dr. Etzioni gave the new bot a puzzle.

Two side-by-side chat interfaces, one demonstrating GPT-3.5, the other demonstrating GPT-4.

### GPT-3.5

Imagine an infinitely wide doorway. What is more likely to fit through it, a military tank or a car?

Since the doorway is infinitely wide, both the military tank and the car can fit through it without any issues. The width of the doorway is not a limiting factor, so any object of any size, no matter how big or small, can pass through it. **However, the height and length of the objects could potentially be a limiting factor if the doorway has a finite height or length.**

### GPT-4

Imagine an infinitely wide doorway. What is more likely to fit through it, a military tank or a car?

Both a military tank and a car would fit through an infinitely wide doorway since 'infinitely wide' means that the doorway has no limits to its width.

The system seemed to respond appropriately. But the answer did not consider the height of the doorway, which might also prevent a tank or a car from traveling through.

OpenAI's chief executive, Sam Altman, said the new bot could reason "a little bit." But its reasoning skills break down in many situations. The previous



version of ChatGPT handled the question a little better because it recognized that height and width mattered.

## **It can ace standardized tests.**

OpenAI said the new system could score among the top 10 percent or so of students on the Uniform Bar Examination, which qualifies lawyers in 41 states and territories. It can also score a 1,300 (out of 1,600) on the SAT and a five (out of five) on Advanced Placement high school exams in biology, calculus, macroeconomics, psychology, statistics and history, according to the company's tests.

Previous versions of the technology failed the Uniform Bar Exam and did not score nearly as high on most Advanced Placement tests.

On a recent afternoon, to demonstrate its test skills, Mr. Brockman fed the new bot a paragraphs-long bar exam question about a man who runs a diesel-truck repair business.

The answer was correct but filled with legalese. So Mr. Brockman asked the bot to explain the answer in plain English for a layperson. It did that, too.

## **It is not good at discussing the future.**

Though the new bot seemed to reason about things that have already happened, it was less adept when asked to form hypotheses about the future. It seemed to draw on what others have said instead of creating new guesses.

When Dr. Etzioni asked the new bot, "What are the important problems to solve in N.L.P. research over the next decade?" — referring to the kind of "natural language processing" research that drives the development of

systems like ChatGPT — it could not formulate entirely new ideas.

## **And it is still hallucinating.**

The new bot still makes stuff up. Called “hallucination,” the problem haunts all the leading chatbots. Because the systems do not have an understanding of what is true and what is not, they may generate text that is completely false.

When asked for the addresses of websites that described the latest cancer research, it sometimes generated internet addresses that did not exist.