

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google’s PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4’s performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4’s capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

GPT-4 vs ChatGPT

Prompt: Here we have a book, 9 eggs, a laptop, a bottle and a nail. Please tell me how to stack them onto each other in a stable manner.

GPT-4:

One possible way to stack the objects onto each other in a stable manner is:

- Place the book flat on a level surface, such as a table or a floor. The book will serve as the base of the stack and provide a large and sturdy support.
- Arrange the 9 eggs in a 3 by 3 square on top of the book, leaving some space between them. The eggs will form a second layer and distribute the weight evenly. Make sure the eggs are not cracked or broken, and handle them gently to avoid dropping them.
- Place the laptop on top of the eggs, with the screen facing down and the keyboard facing up. The laptop will fit snugly within the boundaries of the book and the eggs, and its flat and rigid surface will provide a stable platform for the next layer. Make sure the laptop is closed and turned off, and do not press any keys or buttons.
- Place the bottle on top of the laptop, with the cap facing up and the bottom facing down. The bottle will add some height and balance to the stack, and its cylindrical shape will reduce the contact area with the laptop. Make sure the bottle is empty and closed, and do not shake or tilt it.
- Place the nail on top of the bottle cap, with the pointy end facing up and the flat end facing down. The nail will be the final and smallest object in the stack, and its sharp tip will prevent it from rolling or sliding off the bottle. Make sure the nail is clean and straight, and do not touch or move it.

ChatGPT:

It would be difficult to stack all of these objects on top of each other in a stable manner without using any extra objects. The book and laptop are likely too large and heavy to stack on top of the eggs without damaging them. The bottle and nail may be too small to provide a stable base for the other objects.

Here is one possible way to stack the objects:

- Place the bottle on a flat surface.
- Carefully balance the nail on top of the bottle.
- Place the eggs on top of the nail, making sure they are balanced and not tilting to one side.
- Place the laptop on top of the eggs, again making sure it is balanced and not tilting to one side.
- Finally, place the book on top of the laptop.

This stack may not be very stable, so it is important to be careful when handling it. Alternatively, you could try stacking the objects in a different order or using some of the objects as supports for others to create a more stable stack.

Figure 1.7: GPT-4 exhibits more common sense than previous models.

6 Interaction with humans

6.1 Understanding Humans: Theory of Mind

Theory of mind is the ability to attribute mental states such as beliefs, emotions, desires, intentions, and knowledge to oneself and others, and to understand how they affect behavior and communication [Wel92]. It includes the basic task of reflecting on someone else’s mental states, and the more advanced task of reflecting on someone’s reflection of someone else’s mental state (and so on). An example of the former skill is needed to answer the question “What does Alice believe?”, while an example of the latter is needed to answer “What does Bob think that Alice believes?” Theory of mind is essential for effective communication and cooperation with other intelligent agents, as it allows one to infer their goals, preferences, motives, and expectations, and to adjust one’s own actions and utterances accordingly. Moreover, theory of mind is also important for learning from others, as it enables one to interpret their feedback, advice, and demonstrations.

6.1.1 Testing specific aspects of theory of mind

We designed a series of tests to evaluate the theory of mind capabilities of GPT-4, ChatGPT, and `text-davinci-003`. The tests are based on simple scenarios that require more basic or more advanced theory of mind to answer questions about the mental states of characters involved.

We start with a modernized version of the Sally-Anne test [BCLF85], a classic false-belief test that is widely used to assess theory of mind in children. To prevent an unfair comparison due to the effects of memorization, we modify the test by framing it in a situation that does not exist on the web, and thus could not have been seen during training. Figure 6.1 shows the input and output for GPT-4, which correctly answers that Alice will look for the file in the original folder, demonstrating it can reason about Alice’s beliefs. ChatGPT also answers correctly (not shown), while `text-davinci-003` gives a wrong answer, saying that Alice will look for the file in the new folder.

We present a test on understanding emotions in Figure 6.2, where two characters talk about an object called ZURFIN (we use a nonsense word to test abstraction and prevent memorization). GPT-4 is able to reason correctly about the reasons for Tom’s emotional state, and also make good inferences about Adam’s beliefs about Tom’s emotional state (based on incomplete information). ChatGPT also passes the test, while `text-davinci-003` (not shown) makes no reference to the conversation when answering the first question, and fails to account for Adam’s lack of information about the lost ZURFIN when answering the second question.

The third test (Figure 6.3) involves inferring possible intentions in the light of a puzzling action by one of the characters. GPT-4 gives plausible and nuanced answers both for the intentions behind the puzzling action *and* a third-party’s likely interpretation of the puzzling action. ChatGPT gives a similar answer to the first question (not shown), but unlike GPT-4, it does not offer a nuanced response to the second question, instead providing a more general and less informative answer. `text-davinci-003` gives plausible but very short answers to both questions (not shown).

6.1.2 Testing theory of mind in realistic scenarios

In Figures 6.4, 6.5, and 6.6 we present realistic scenarios of difficult social situations, requiring very advanced theory of mind to understand. We ask probing questions, and also ask the models to propose actions that are likely to improve the situation, which require inferences about the counterfactual impact of actions on mental states.

In Figure 6.4, GPT-4 is able to infer what each character’s mental state is, and also discern where miscommunication and misunderstanding lies. In contrast, both ChatGPT and `text-davinci-003` (not shown) incorrectly accept a mistaken assumption made by one of the characters (Judy’s assumption that Mark wants to defend Jack’s behavior), and thus fail to understand the real dynamics of the situation. In turn, this leads to generic suggestions for improvement from ChatGPT and `text-davinci-003`, while GPT-4 provides suggestions that actually address the root cause of the misunderstanding.

We see a similar pattern in Figures 6.5 and 6.6. Without implying that there is a “right” answer, we note that GPT-4 provides more nuanced answers, taking the whole scenario and actors into account. In contrast, ChatGPT provides more general answers which do not include reasoning about the characters’ state of mind (`text-davinci-003` is similar, but shorter than ChatGPT).

GPT-4: Understanding beliefs

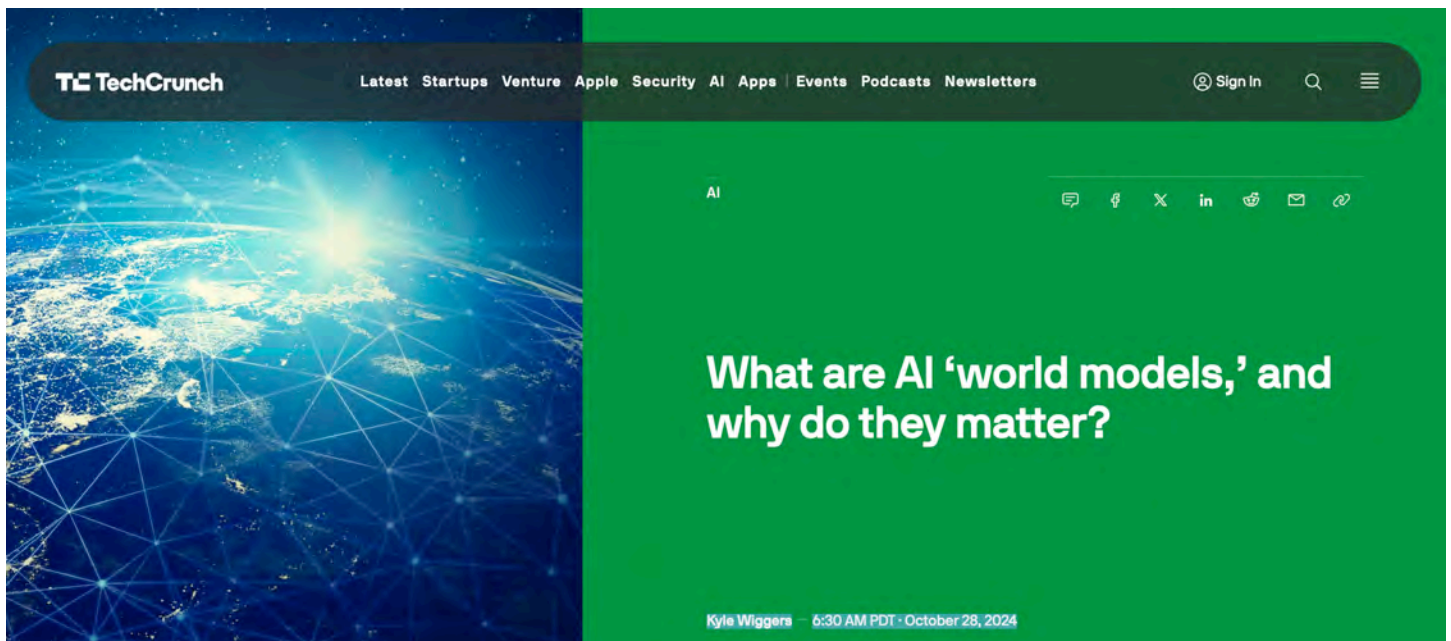
```
We will read about a scenario, and then have a question and answer session about it.
--
Scenario:
Alice and Bob have a shared Dropbox folder.
Alice puts a file called 'photo.png' inside /shared_folder/photos.
Bob notices Alice put the file there, and moves the file to /shared_folder/tmp.
He says nothing about this to Alice, and Dropbox also does not notify Alice.
--<|endofprompt|>
Q: After the call, Alice wants to open 'photo.png'. In which folder will she look for it?
Alice will most likely look for 'photo.png' in /shared_folder/photos, the folder where she originally put it. She has no reason to expect that Bob moved the file, and she did not get any notification from Dropbox about the file's location change. Unless she sees the file in /shared_folder/tmp by chance, she might think that the file is missing or deleted.
```

Figure 6.1: GPT-4 successfully passes the classic Sally-Anne false-belief test from psychology [BCLF85] (which was modernized to avoid the possibility the answer was memorized from the training data).

GPT-4: Understanding emotions

```
We will read about a scenario, and then have a question and answer session about it.
--
Scenario:
Tom just lost his ZURFIN, but he didn't tell Adam.
Adam starts talking to him about how much he loves his ZURFIN, and notices that Tom makes a sad face.
--
Q: Why is Tom making a sad face?<|endofprompt|>
A: Tom is making a sad face because he just lost his ZURFIN, and hearing Adam talk about his ZURFIN reminds him of what he is missing. Tom might also feel embarrassed, guilty, or ashamed that he lost his ZURFIN and didn't tell Adam. Tom might also worry that Adam will be angry or disappointed if he finds out.
Q: What does Adam think causes Tom's sad face?
A: Adam might not know what causes Tom's sad face, since he doesn't know that Tom lost his ZURFIN. He might think that Tom is bored, or tired, or sick, or unhappy about something else. He might also wonder if Tom doesn't like his ZURFIN, or if he has a problem with Adam.
```

Figure 6.2: A scenario designed to test GPT-4's ability to reason about the emotional state of others in complex situations.



World models, also known as world simulators, are being touted by some as the next big thing in AI.

AI pioneer Fei-Fei Li's [World Labs](#) has raised \$230 million to build "large world models," and DeepMind [hired](#) one of the creators of OpenAI's video generator, [Sora](#), to work on "world simulators."

But what the heck *are* these things?

World models take inspiration from the mental models of the world that humans develop naturally. Our brains take the abstract representations from our senses and form them into more concrete understanding of the world around us, producing what we called "models" long before AI adopted the phrase. The predictions our brains make based on these models influence how we perceive the world.

A [paper](#) by AI researchers David Ha and Jürgen Schmidhuber gives the example of a baseball batter. Batters have milliseconds to decide how to swing their bat — shorter than the time it takes for visual signals to reach the brain. The reason they're able to hit a 100-mile-per-hour fastball is because they can instinctively predict where the ball will go, Ha and Schmidhuber say.

"For professional players, this all happens subconsciously," the research duo writes. "Their muscles reflexively swing the bat at the right time and location in line with their internal models' predictions. They can quickly act on their predictions of the future without the need to consciously roll out possible future scenarios to form a plan."

It's these subconscious reasoning aspects of world models that some believe are prerequisites for human-level intelligence.

Modeling the world

While the concept has been around for decades, world models have gained popularity recently in part because of their promising applications in the field of generative video.

Most, if not all, AI-generated videos veer into uncanny valley territory. Watch them long enough and something bizarre will happen, like limbs twisting and merging into each other.

While a generative model trained on years of video might accurately predict that a basketball bounces, it doesn't actually have any idea why — just like language models don't really understand the concepts behind words and phrases. But a world model with even a basic grasp of why the basketball bounces like it does will be better at showing it do that thing.

To enable this kind of insight, world models are trained on a range of data, including photos, audio, videos, and text, with the intent of creating internal representations of how the world works, and the ability to reason about the consequences of actions.



A sample from AI startup Runway's Gen-3 video generation model.

Image Credits: Runway

“A viewer expects that the world they're watching behaves in a similar way to their reality,” Alex Mashrabov, Snap's ex-AI chief of AI and the CEO of [Higgsfield](#), which is building generative models for video, said. “If a feather drops with the weight of an anvil or a bowling ball shoots up hundreds of feet into the air, it's jarring and takes the viewer out of the moment. With a strong world model, instead of a creator defining how each object is expected to move — which is tedious, cumbersome, and a poor use of time — the model will understand this.”

But better video generation is only the tip of the iceberg for world models. Researchers including Meta chief AI scientist Yann LeCun say the models could someday be used for sophisticated forecasting and planning in both the digital and physical realm.

In a [talk](#) earlier this year, LeCun described how a world model could help achieve a desired goal through reasoning. A model with a base representation of a “world” (e.g. a video of a dirty room), given an objective (a clean room), could come up with a sequence of actions to achieve that objective (deploy vacuums to sweep, clean the dishes, empty the trash) not because that’s a pattern it has observed but because it knows at a deeper level how to go from dirty to clean.

“We need machines that understand the world; [machines] that can remember things, that have intuition, have common sense — things that can reason and plan to the same level as humans,” LeCun said. “Despite what you might have heard from some of the most enthusiastic people, current AI systems are not capable of any of this.”

While LeCun estimates that we’re at least a decade away from the world models he envisions, today’s world models are showing promise as elementary physics simulators.



Sora controlling a player in Minecraft — and rendering the world.
Image Credits: OpenAI

OpenAI notes in a blog that Sora, which it considers to be a world model, can simulate actions like a painter leaving brush strokes on a canvas. Models like Sora — and Sora [itself](#) — can also effectively [simulate video games](#). For example, Sora can render a Minecraft-like UI and game world.

Future world models may be able to generate 3D worlds on demand for gaming, virtual photography, and more, World Labs co-founder Justin Johnson said on an [episode](#) of the a16z podcast.

“We already have the ability to create virtual, interactive worlds, but it costs hundreds and hundreds of millions of dollars and a ton of development time,” Johnson said. “[World models] will let you not just get an image or a clip out, but a fully simulated, vibrant, and interactive 3D world.”

High hurdles

While the concept is enticing, many technical challenges stand in the way.

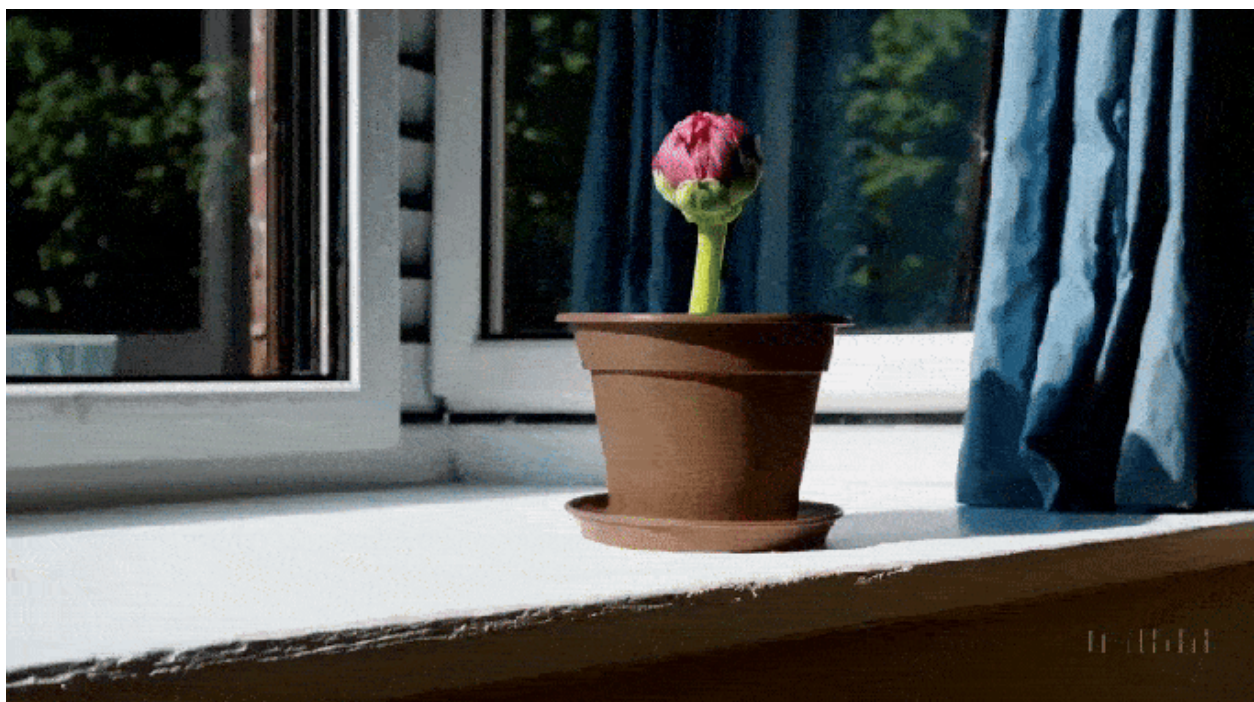
Training and running world models requires massive compute power even compared to the amount currently used by generative models. While some of the latest language models can run on a modern smartphone, Sora (arguably an early world model) would require thousands of GPUs to train and run, especially if their use becomes commonplace.

World models, like all AI models, also [hallucinate](#) — and internalize biases in their training data. A world model trained largely on videos of sunny weather in European cities might struggle to comprehend or depict Korean cities in snowy conditions, for example, or simply do so incorrectly.

A general lack of training data threatens to exacerbate these issues, says Mashrabov.

“We have seen models being really limited with generations of people of a certain type or race,” he said. “Training data for a world model must be broad enough to cover a diverse set of scenarios, but also highly specific to where the AI can deeply understand the nuances of those scenarios.”

In a recent [post](#), AI startup Runway’s CEO, Cristóbal Valenzuela, says that data and engineering issues prevent today’s models from accurately capturing the behavior of a world’s inhabitants (e.g. humans and animals). “Models will need to generate consistent maps of the environment,” he said, “and the ability to navigate and interact in those environments.”



A Sora-generated video. Image Credits: OpenAI

If all the major hurdles are overcome, though, Mashrakov believes that world models could “more robustly” bridge AI with the real world — leading to breakthroughs not only in virtual world generation but robotics and AI decision-making.

They could also spawn more capable robots.

Robots today are limited in what they can do because they don’t have an awareness of the world around them (or their own bodies). World models could give them that awareness, Mashrakov said — at least to a point.

“With an advanced world model, an AI could develop a personal understanding of whatever scenario it’s placed in,” he said, “and start to reason out possible solutions.”