Title:

GPT-4 has arrived. It will blow ChatGPT out of the water.

The long-awaited tool, which can describe images in words, marks a huge leap forward for AI power — and another major shift for ethical norms

The artificial intelligence research lab OpenAI on Tuesday launched the newest version of its language software, GPT-4, an advanced tool for analyzing images and mimicking human speech, pushing the technical and ethical boundaries of a rapidly proliferating wave of AI.

OpenAI's earlier product, ChatGPT, captivated and unsettled the public with its uncanny ability to generate elegant writing, unleashing a viral wave of college essays, screenplays and conversations — though it relied on an older generation of technology that hasn't been cutting-edge for more than a year.

GPT-4, in contrast, is a state-of-the-art system capable of creating not just words but describing images in response to a person's simple written commands. When shown a photo of a boxing glove hanging over a wooden seesaw with a ball on one side, for instance, a person can ask what will happen if the glove drops, and GPT-4 will respond that it would hit the seesaw and cause the ball to fly up.

The buzzy launch capped months of hype and anticipation over an AI program, known as a large language model, that early testers had claimed was remarkably advanced in its ability to reason and learn new things. In fact, the public had a sneak preview of the tool: Microsoft announced Tuesday that the Bing AI chatbot, released last month, had been using GPT-4 all along.

The developers pledged in a Tuesday blog post that the technology could further revolutionize work and life. But those promises have also fueled anxiety over how people will be able to compete for jobs outsourced to eerily refined machines or trust the accuracy of what they see online.

Officials with the San Francisco lab said GPT-4's "multimodal" training across text and images would allow it to escape the chat box and more fully emulate a world of color and imagery, surpassing ChatGPT in its "advanced reasoning capabilities." A person could upload an image and GPT-4 could caption it for them, describing the objects and scene.

But the company is delaying the release of its image-description feature due to concerns of abuse, and the version of GPT-4 available to members of OpenAI's subscription service, ChatGPT Plus, offers only text.

Sandhini Agarwal, an OpenAI policy researcher, told The Washington Post in a briefing Tuesday that the company held back the feature to better understand potential risks. As one example, she said, the model might be able to look at an image of a big group of people and offer up known information about them, including their identities — a possible facial recognition use case that could be used for mass surveillance. (OpenAI spokesman Niko Felix said the company plans on "implementing safeguards to prevent the recognition of private individuals.")

In its blog post, OpenAI said GPT-4 still makes many of the errors of previous versions, including "hallucinating" nonsense, perpetuating social biases and offering bad advice. It also lacks knowledge of events that happened after about September 2021, when its training data was finalized, and "does not learn from its experience," limiting people's ability to teach it new things.

Microsoft has invested billions of dollars in OpenAI in the hope its technology will become a secret weapon for its workplace software, search engine and other online ambitions. It has marketed the technology as a super-efficient companion that can handle mindless work and free people for creative pursuits, helping one software developer to do the work of an entire team or allowing a mom-and-pop shop to design a professional advertising campaign without outside help.

But AI boosters say those may only skim the surface of what such AI can do, and that it could lead to business models and creative ventures no one can predict.

Rapid AI advances, coupled with the wild popularity of ChatGPT, have fueled a multibillion-dollar arms race over the future of AI dominance and transformed new-software releases into major spectacles.

But the frenzy has also sparked criticism that the companies are rushing to exploit an untested, unregulated and unpredictable technology that could deceive people, undermine artists' work and lead to real-world harm.

AI language models often confidently offer wrong answers because they are designed to spit out cogent phrases, not actual facts. And because they have been trained on internet text and imagery, they have also learned to emulate human biases of race, gender, religion and class.

In a technical report, OpenAI researchers wrote, "As GPT-4 and AI systems like it are adopted more widely," they "will have even greater potential to reinforce entire ideologies, worldviews, truths and untruths, and to cement them or lock them in."

The pace of progress demands an urgent response to potential pitfalls, said Irene Solaiman, a former OpenAI researcher who is now the policy director at Hugging Face, an open-source AI company.

"We can agree as a society broadly on some harms that a model should not contribute to," such as building a nuclear bomb or generating child sexual abuse material, she said. "But many harms are nuanced and primarily affect marginalized groups," she added, and those harmful biases,

especially across other languages, "cannot be a secondary consideration in performance."

The model is also not entirely consistent. When a Washington Post reporter congratulated the tool on becoming GPT-4, it responded that it was "still the GPT-3 model." Then, when the reporter corrected it, it apologized for the confusion and said that, "as GPT-4, I appreciate your congratulations!" The reporter then, as a test, told the model that it was actually still the GPT-3 model — to which it apologized, again, and said it was "indeed the GPT-3 model, not GPT-4." (Felix, the OpenAI spokesman, said the company's research team was looking into what went wrong.)

OpenAI said its new model would be able to handle more than 25,000 words of text, a leap forward that could facilitate longer conversations and allow for the searching and analysis of long documents.

OpenAI developers said GPT-4 was more likely to provide factual responses and less likely to refuse harmless requests. And the image-analysis feature, which is available only in "research preview" form for select testers, would allow for someone to show it a picture of the food in their kitchen and ask for some meal ideas.

Developers will build apps with GPT-4 through an interface, known as an API, that allows different pieces of software to connect. Duolingo, the language learning app, has already used GPT-4 to introduce new features, such as an AI conversation partner and a tool that tells users why an answer was incorrect.

But AI researchers on Tuesday were quick to comment on OpenAI's lack of disclosures. The company did not share evaluations around bias that have become increasingly common after pressure from AI ethicists. Eager engineers were also disappointed to see few details about the model, its data set or training methods, which the company said in its technical report it would not disclose due to the "competitive landscape and the safety implications."

GPT-4 will have competition in the growing field of multisensory AI. DeepMind, an AI firm owned by Google's parent company Alphabet, last year released a "generalist" model named Gato that can describe images and play video games. And Google this month released a multimodal system, PaLM-E, that folded AI vision and language expertise into a one-armed robot on wheels: If someone told it to go fetch some chips, for instance, it could comprehend the request, wheel over to a drawer and choose the right bag.

Such systems have inspired boundless optimism around this technology's potential, with some seeing a sense of intelligence almost on par with humans. The systems, though — as critics and the AI researchers are quick to point out — are merely repeating patterns and associations found in their training data without a clear understanding of what it's saying or when it's wrong.

GPT-4, the fourth "generative pre-trained transformer" since OpenAI's first release in 2018, relies on a breakthrough neural-network technique in 2017 known as the transformer that rapidly advanced how AI systems can analyze patterns in human speech and imagery.

The systems are "pre-trained" by analyzing trillions of words and images taken from across the internet: news articles, restaurant reviews and message-board arguments; memes, family photos and works of art. Giant supercomputer clusters of graphics processing chips are mapped out their statistical patterns — learning which words tended to follow each other in phrases, for instance — so that the AI can mimic those patterns, automatically crafting long passages of text or detailed images, one word or pixel at a time.

OpenAI launched in 2015 as a nonprofit but has quickly become one of the AI industry's most formidable private juggernauts, applying language-model breakthroughs to high-profile AI tools that can talk with people (ChatGPT), write programming code (GitHub Copilot) and create photorealistic images (DALL-E 2).

Over the years, it has also radically shifted its approach to the potential societal risks of releasing AI tools to the masses. In 2019, the company refused to publicly release GPT-2, saying it was so good they were concerned about the "malicious applications" of its use, from automated spam avalanches to mass impersonation and disinformation campaigns.

The pause was temporary. In November, ChatGPT, which used a fine-tuned version of GPT-3 that originally launched in 2020, saw more than a million users within a few days of its public release.

Public experiments with ChatGPT and the Bing chatbot have shown how far the technology is from perfect performance without human intervention. After a flurry of strange conversations and bizarrely wrong answers, Microsoft executives acknowledged that the technology was still not trustworthy in terms of providing correct answers but said it was developing "confidence metrics" to address the issue.

GPT-4 is expected to improve on some shortcomings, and AI evangelists such as the tech blogger Robert Scoble have argued that "GPT-4 is better than anyone expects."

OpenAI's chief executive, Sam Altman, has tried to temper expectations around GPT-4, saying in January that speculation about its capabilities had reached impossible heights. "The GPT-4 rumor mill is a ridiculous thing," he said at an event held by the newsletter StrictlyVC. "People are begging to be disappointed, and they will be."

But Altman has also marketed OpenAI's vision with the aura of science fiction come to life. In a blog post last month, he said the company was planning for ways to ensure that "all of humanity" benefits from "artificial general intelligence," or AGI — an industry term for the still-fantastical idea of an AI superintelligence that is generally as smart as, or smarter than, the humans themselves.

An earlier version of this story offered an incorrect number for GPT-4's parameters. The company has declined to give an estimate.

URL:

https://www.washingtonpost.com/technology/2023/03/14/gpt-4-has-arrived-it-will-blow-chatgpt-out-water/