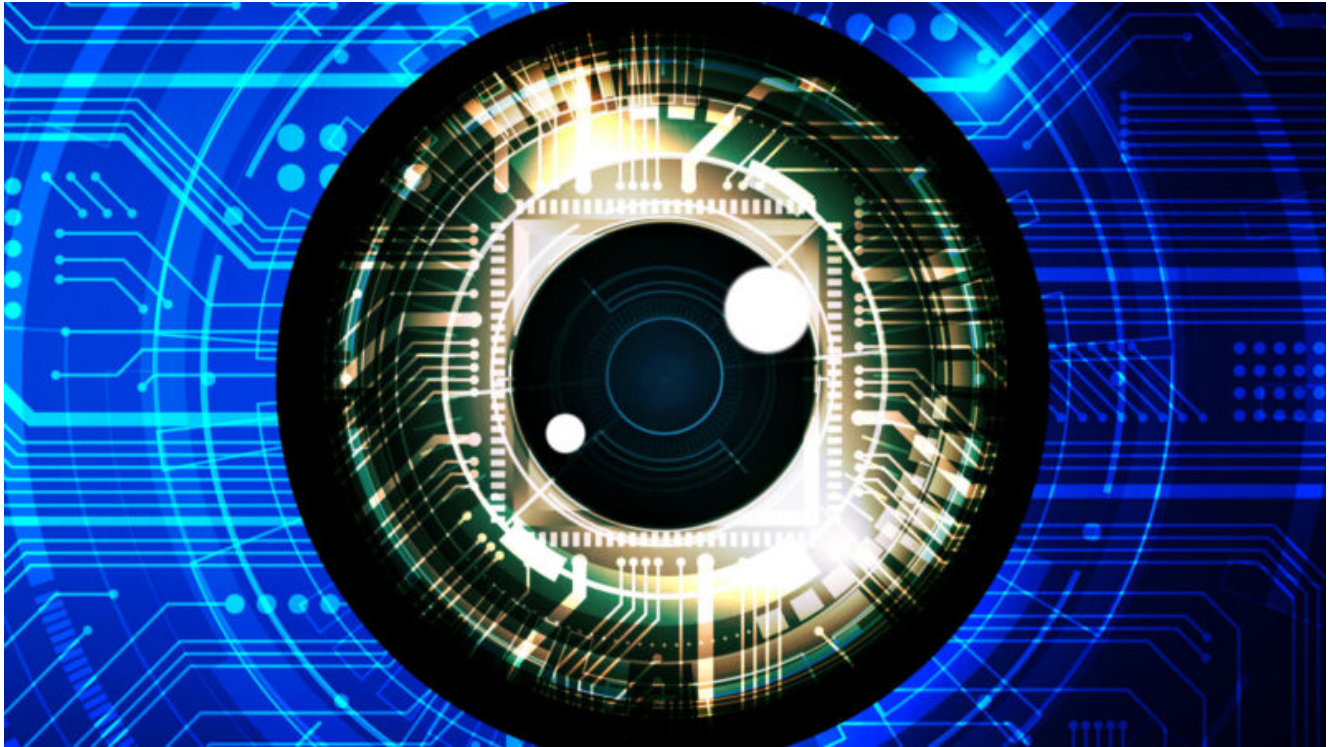


ChatGPT update enables its AI to “see, hear, and speak,” according to OpenAI



reader comments

61 with

On Monday, OpenAI announced a significant update to ChatGPT that enables its GPT-3.5 and GPT-4 AI models to analyze images and react to them as part of a text conversation. Also, the ChatGPT mobile app will add speech synthesis options that, when paired with its existing speech recognition features, will enable fully verbal conversations with the AI assistant, OpenAI says.

OpenAI is planning to roll out these features in ChatGPT to Plus and Enterprise subscribers “over the next two weeks.” It also notes that speech synthesis is coming to iOS and Android only, and image recognition will be available on both the web

interface and the mobile apps.

OpenAI says the new image recognition feature in ChatGPT lets users upload one or more images for conversation, using either the GPT-3.5 or GPT-4 models. In its promotional blog post, the company claims the feature can be used for a variety of everyday applications: from figuring out what's for dinner by taking pictures of the fridge and pantry, to troubleshooting why your grill won't start. It also says that users can use their device's touch screen to circle parts of the image that they would like ChatGPT to concentrate on.

- A shot taken from an OpenAI promotional video where ChatGPT analyzes user photos to help adjust a bike seat.
- A shot taken from an OpenAI promotional video where ChatGPT analyzes user photos to help adjust a bike seat.
- A shot taken from an OpenAI promotional video where ChatGPT analyzes user photos to help adjust a bike seat.

On its site, OpenAI provides a promotional video that illustrates a hypothetical exchange with ChatGPT where a user asks how to raise a bicycle seat, providing photos as well as an instruction manual and an image of the user's toolbox. ChatGPT reacts and advises the user how to complete the process. We have not tested this feature ourselves, so its real-world effectiveness is unknown.

Advertisement

So how does it work? OpenAI has not released technical details of how GPT-4 or its multimodal version, GPT-4V, operate under the hood, but based on known AI research from others (including OpenAI partner Microsoft), multimodal AI models typically transform text and images into a shared encoding space, which enables them to process various types of data through the same neural network. OpenAI may use CLIP to bridge the gap between visual and text data in a way that aligns image and text representations in the same latent space, a kind of vectorized web of data relationships. That technique

could allow ChatGPT to make contextual deductions across text and images, though this is speculative on our part.

Meanwhile in audio land, ChatGPT's new voice synthesis feature reportedly allows for back-and-forth spoken conversation with ChatGPT, driven by what OpenAI calls a "new text-to-speech model," although text-to-speech has been solved for a long time. Once the feature rolls out, the company says that users can engage the feature by opting in to voice conversations in the app's settings and then selecting from five different synthetic voices with names like "Juniper," "Sky," "Cove," "Ember," and "Breeze." OpenAI says these voices have been crafted in collaboration with professional voice actors.

OpenAI's Whisper, an open source speech recognition system we covered in September of last year, will continue to handle the transcription of user speech input. Whisper has been integrated with the ChatGPT iOS app since it launched in May. OpenAI released the similarly capable ChatGPT Android app in July.

<https://arstechnica.com/?p=1970737>