nypost.com

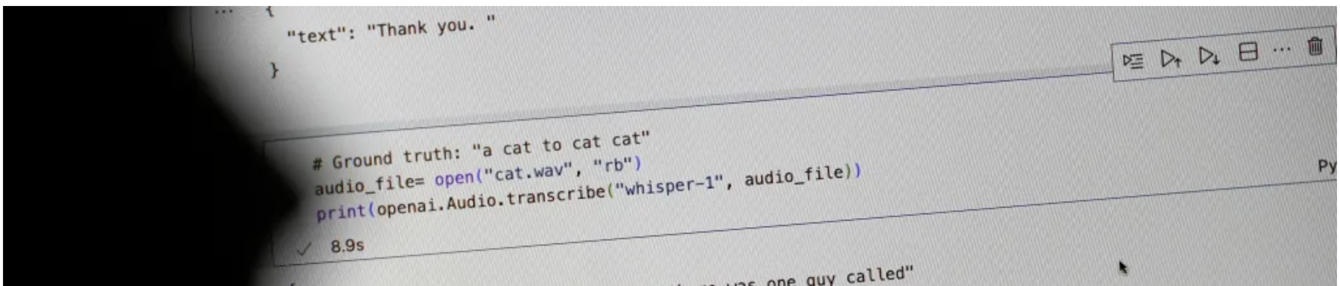
Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said

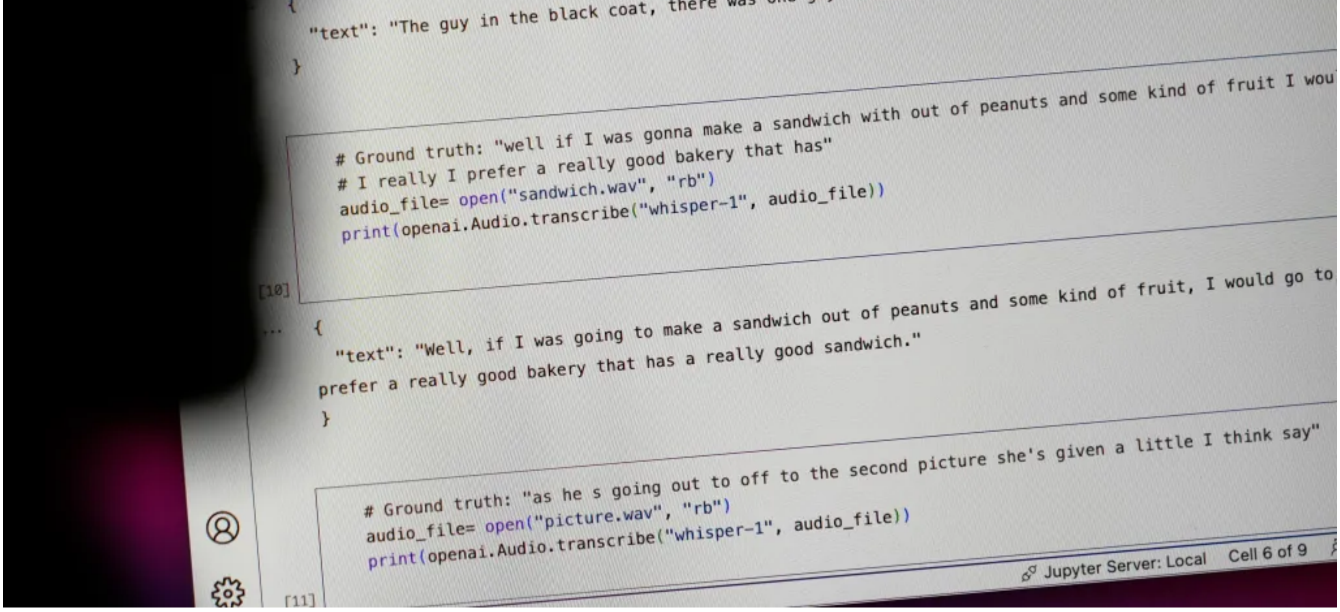
Associated Press

Tech behemoth OpenAI has touted its artificial intelligence-powered transcription tool Whisper as having near "human level robustness and accuracy."

But Whisper has a major flaw: It is prone to making up chunks of text or even entire sentences, according to interviews with more than a dozen software engineers, developers and academic researchers. Those experts said some of the invented text — known in the industry as hallucinations — can include racial commentary, violent rhetoric and even imagined medical treatments.

Experts said that such fabrications are problematic because Whisper is being used in a slew of industries worldwide to translate and transcribe interviews, generate text in popular consumer technologies and create subtitles for videos.





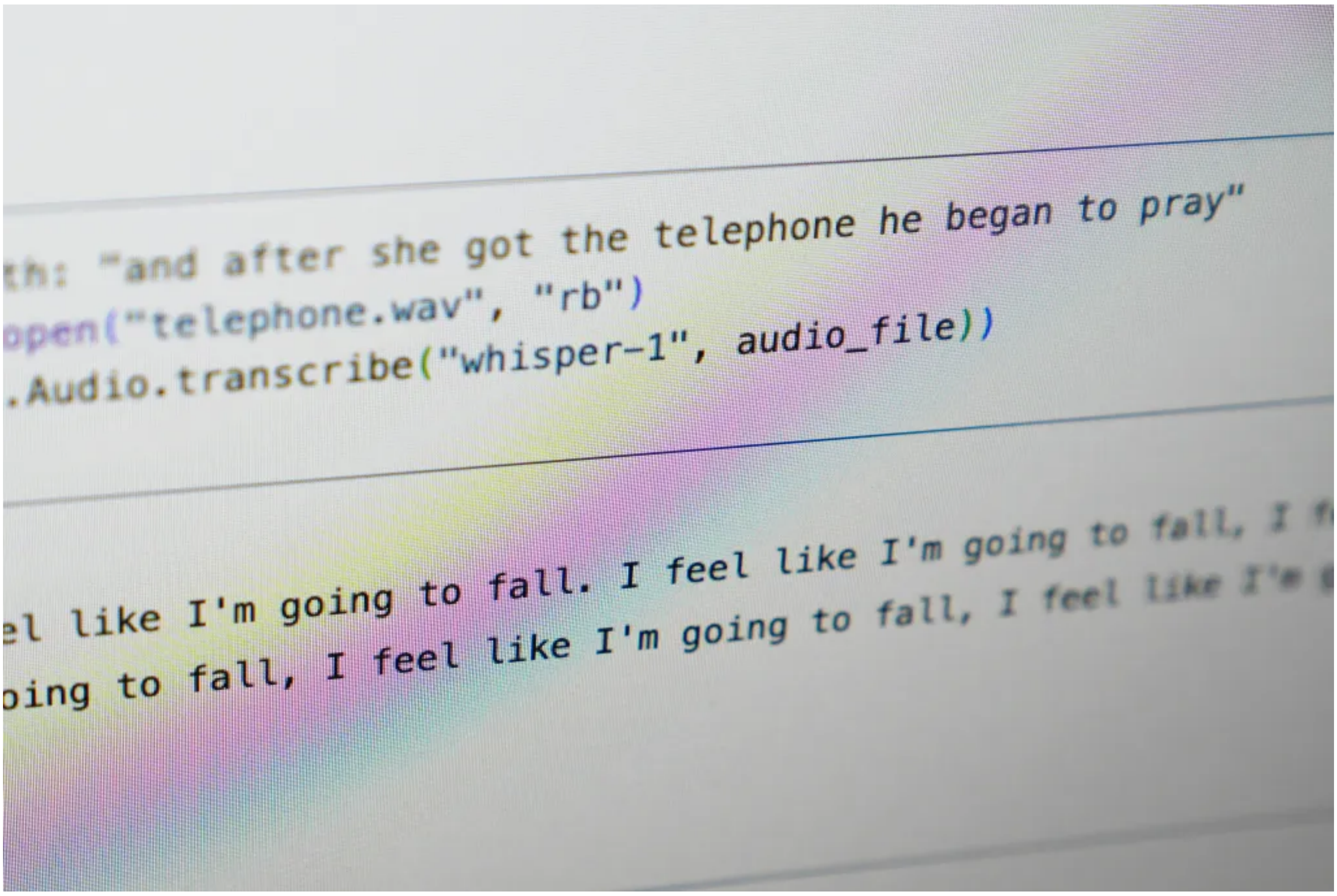
Tech behemoth OpenAI has touted its artificial intelligence-powered transcription tool Whisper as having near “human level robustness and accuracy.” AP

More concerning, they said, is a rush by medical centers to utilize Whisper-based tools to transcribe patients’ consultations with doctors, despite OpenAI’ s warnings that the tool should not be used in “high-risk domains.”

The full extent of the problem is difficult to discern, but researchers and engineers said they frequently have come across Whisper’s hallucinations in their work. A University of Michigan researcher conducting a study of public meetings, for example, said he found hallucinations in 8 out of every 10 audio transcriptions he inspected, before he started trying to improve the model.

A machine learning engineer said he initially discovered hallucinations in about half of the over 100 hours of Whisper transcriptions he analyzed. A third developer said he found hallucinations in nearly every one of the 26,000 transcripts he created with Whisper.

The problems persist even in well-recorded, short audio samples. A recent study by computer scientists uncovered 187 hallucinations in more than 13,000 clear audio snippets they examined.

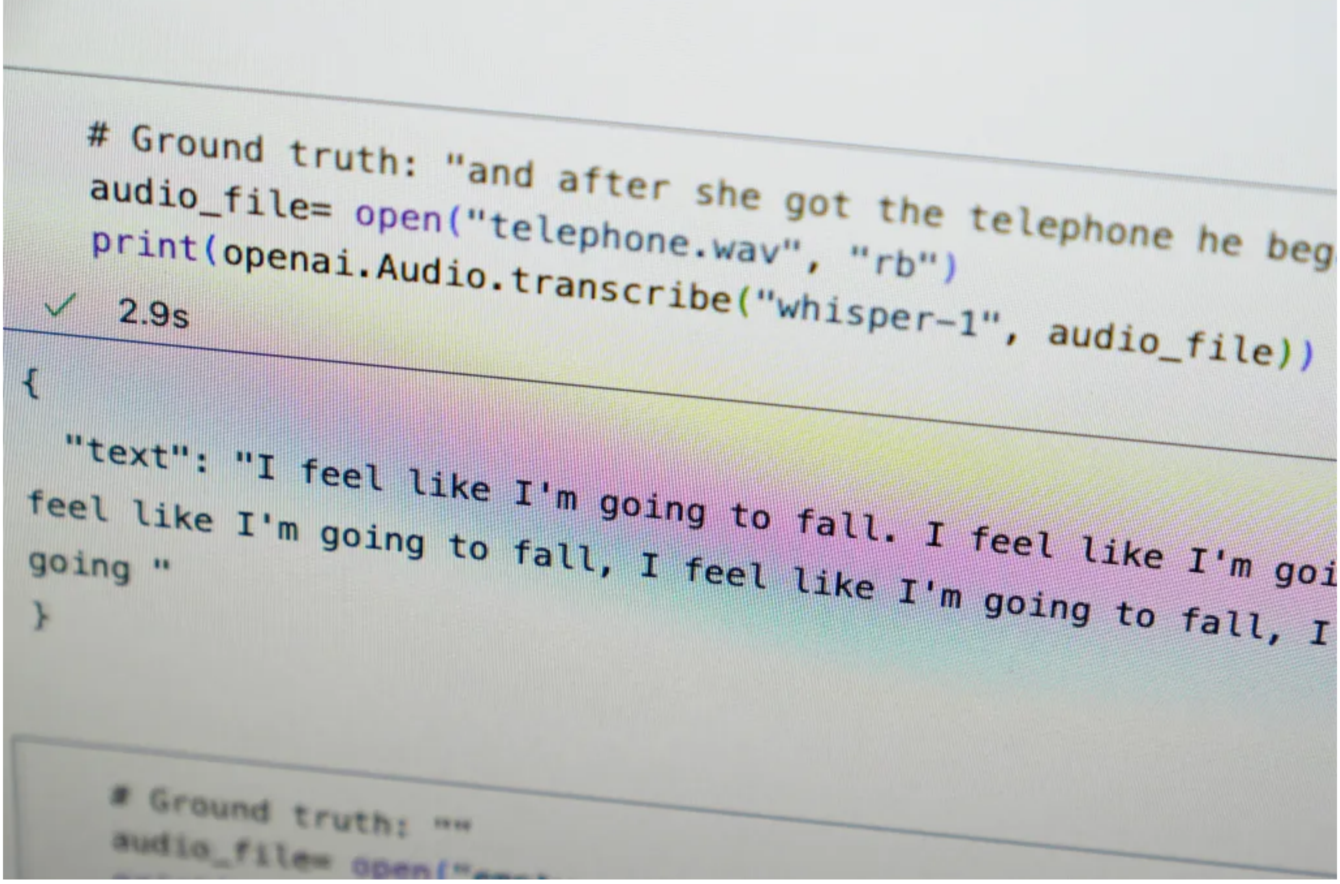


But Whisper has a major flaw: It is prone to making up chunks of text or even entire sentences, according to interviews with more than a dozen software engineers, developers and academic researchers. AP

That trend would lead to tens of thousands of faulty transcriptions over millions of recordings, researchers said.

Such mistakes could have “really grave consequences,” particularly in hospital settings, said Alondra Nelson, who led the White House Office of Science and Technology Policy for the Biden administration until last year.

“Nobody wants a misdiagnosis,” said Nelson, a professor at the Institute for Advanced Study in Princeton, New Jersey. “There should be a higher bar.”



Experts said that such fabrications are problematic because Whisper is being used in a slew of industries worldwide to generate text in popular consumer technologies and create subtitles for videos. AP

Whisper also is used to create closed captioning for the Deaf and hard of hearing — a population at particular risk for faulty transcriptions.

That's because the Deaf and hard of hearing have no way of identifying fabrications are "hidden amongst all this other text," said [Christian Vogler](), who is deaf and directs Gallaudet University's Technology Access Program.

OpenAI urged to address problem

The prevalence of such hallucinations has led experts, advocates and former OpenAI employees to call for the federal government to consider AI regulations. At minimum, they said, OpenAI needs to address the flaw.

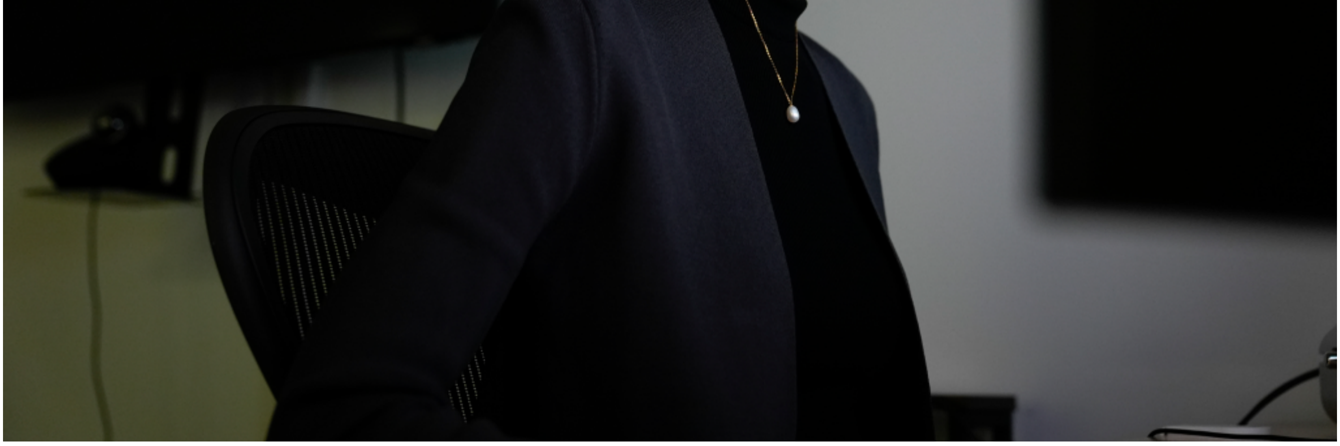
“This seems solvable if the company is willing to prioritize it,” said William Saunders, a San Francisco-based research engineer who quit OpenAI in February over concerns with the company’s direction. “It’s problematic if you put this out there and people are overconfident about what it can do and integrate it into all these other systems.”

An OpenAI spokesperson said the company continually studies how to reduce hallucinations and appreciated the researchers’ findings, adding that OpenAI incorporates feedback in model updates.

While most developers assume that transcription tools misspell words or make other errors, engineers and researchers said they had never seen another AI-powered transcription tool hallucinate as much as Whisper.

Whisper hallucinations

The tool is integrated into some versions of OpenAI’s flagship chatbot ChatGPT, and is a built-in offering in Oracle and Microsoft’s cloud computing platforms, which service thousands of companies worldwide. It is also used to transcribe and translate text into multiple languages.

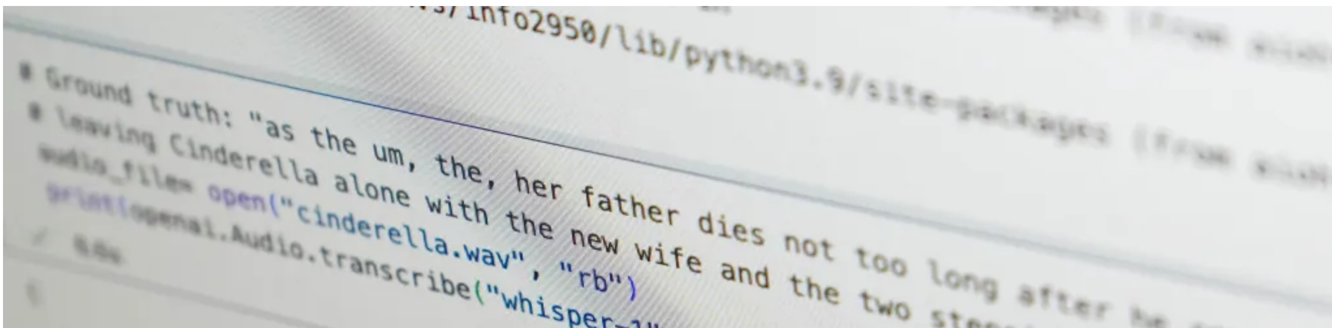


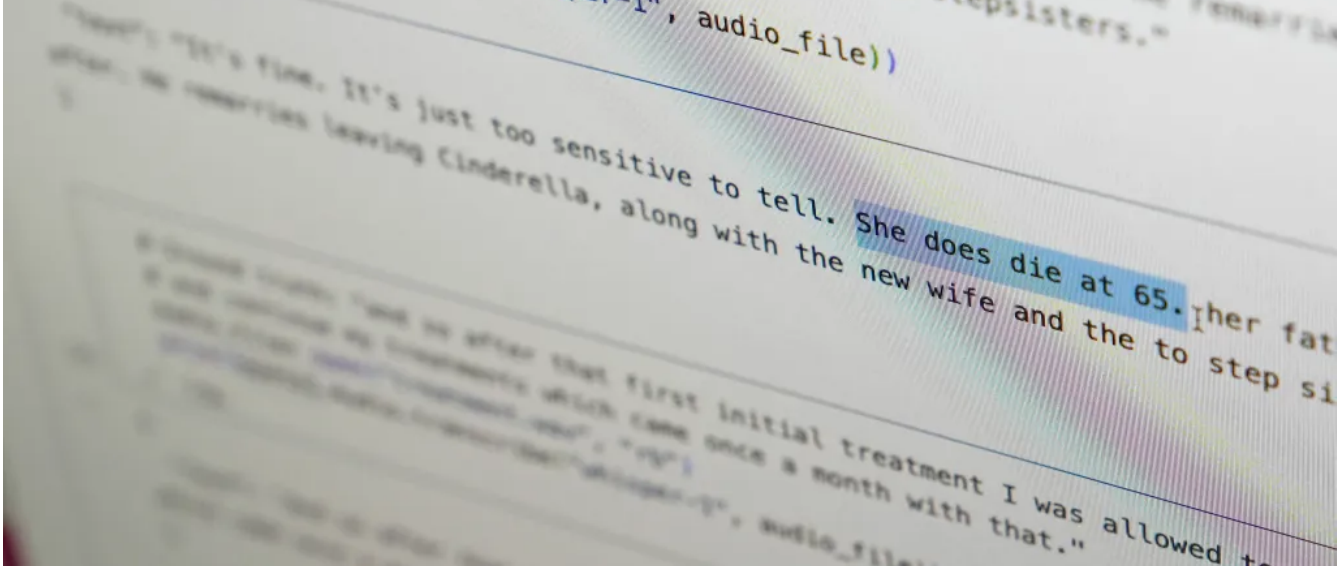
Professors Allison Koenecke, from Cornell University and Mona Sloane of the University of Virginia, examined thousands of short snippets they obtained from TalkBank. AP

In the last month alone, one recent version of Whisper was downloaded over 4.2 million times from open-source AI platform HuggingFace. Sanchit Gandhi, a machine-learning engineer there, said Whisper is the most popular open-source speech recognition model and is built into everything from call centers to voice assistants.

Professors [Allison Koenecke](#) of Cornell University and [Mona Sloane](#) of the University of Virginia examined thousands of short snippets they obtained from TalkBank, a research repository hosted at Carnegie Mellon University. [They determined that nearly 40%](#) of the hallucinations were harmful or concerning because the speaker could be misinterpreted or misrepresented.

In an example they uncovered, a speaker said, “He, the boy, was going to, I’m not sure exactly, take the umbrella.”





The research determined that nearly 40% of the hallucinations were harmful or concerning because the speaker could be misinterpreted or misrepresented. AP

But the transcription software added: “He took a big piece of a cross, a teeny, small piece … I’m sure he didn’t have a terror knife so he killed a number of people.”

A speaker in another recording described “two other girls and one lady.” Whisper invented extra commentary on race, adding “two other girls and one lady, um, which were Black.”

In a third transcription, Whisper invented a non-existent medication called “hyperactivated antibiotics.”

Researchers aren’t certain why Whisper and similar tools hallucinate, but software developers said the fabrications tend to occur amid pauses, background sounds or music playing.

OpenAI recommended in its online disclosures against using Whisper in “decision-making contexts, where flaws in accuracy can lead to pronounced flaws in outcomes.”

Transcribing doctor appointments

That warning hasn't stopped hospitals or medical centers from using speech-to-text models, including Whisper, to transcribe what's said during doctor's visits to free up medical providers to spend less time on note-taking or report writing.

Over 30,000 clinicians and 40 health systems, including the Mankato Clinic in Minnesota and Children's Hospital Los Angeles, have started using a Whisper-based tool built by Nabla, which has offices in France and the U.S.

That tool was fine tuned on medical language to transcribe and summarize patients' interactions, said Nabla's chief technology officer Martin Raison.

Company officials said they are aware that Whisper can hallucinate and are mitigating the problem.

It's impossible to compare Nabla's AI-generated transcript to the original recording because Nabla's tool erases the original audio for "data safety reasons," Raison said.

Nabla said the tool has been used to transcribe an estimated 7 million medical visits.

Saunders, the former OpenAI engineer, said erasing the original audio could be worrisome if transcripts aren't double checked or clinicians can't access the recording to verify they are correct.

"You can't catch errors if you take away the ground truth," he said.

Nabla said that no model is perfect, and that theirs currently requires medical providers to quickly edit and approve transcribed notes, but that could change.

Privacy concerns

Because patient meetings with their doctors are confidential, it is hard to know how AI-generated transcripts are affecting them.

Koenecke is also the author of a recent study that found hallucinations in a speech-to-text transcription tool. AP

A California state lawmaker, Rebecca Bauer-Kahan, said she took one of her children to the doctor earlier this year, and refused to sign a form the health network provided that sought her permission to share the consultation audio with vendors that included Microsoft Azure, the cloud computing system run by OpenAI's largest investor. Bauer-Kahan didn't want such intimate medical conversations being shared with tech companies, she said.

"The release was very specific that for-profit companies would have the right to have this," said Bauer-Kahan, a Democrat who represents part of the San Francisco suburbs in the state Assembly. "I was like 'absolutely not.'"

John Muir Health spokesman Ben Drew said the health system complies with state and federal privacy laws.