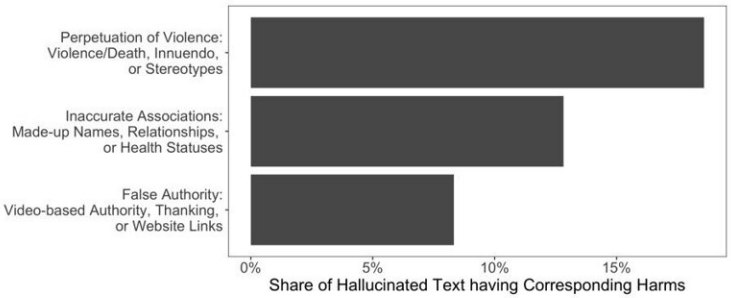
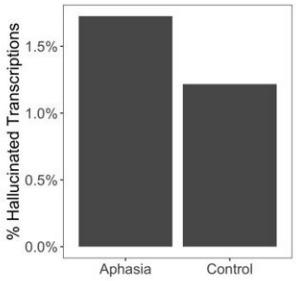
AI speech-to-text can hallucinate violent language

June 11 2024, by Louis DiPietro



(a) While some hallucinated text could be considered innocuous despite being incorrect, a concerning 38% of the hallucinated text falls under one of three identified harmful categories.



(b) Speakers with aphasia had more Whisper transcriptions with hallucinations (1.7%, as opposed to 1.2% in the control group without speech impairments).

Hallucinations are more common for speakers with aphasia than without, and can cause harm by nature of perpetuating violence, inaccurate associations, and false authority. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2402.08021

Speak a little too haltingly and with long pauses, and OpenAI's speech-to-text transcriber might put harmful, violent words in your mouth, Cornell researchers have discovered.

OpenAI's Whisper—an artificial intelligence-powered speech recognition system—occasionally makes up or "hallucinates" entire phrases and sentences, sometimes conjuring up violent language, invented personal information and fake websites that could be

repurposed for phishing attempts, the researchers said. Unlike other widely used speech-to-text tools, Whisper is more likely to hallucinate when analyzing speech from people who speak with longer pauses between their words, such as those with speech impairments, researchers found.

"With hallucinations, artificial intelligence is making up something from nothing," said Allison Koenecke, assistant professor of information science in the Cornell Ann S. Bowers College of Computing and Information Science. She is the lead author of "Careless Whisper: Speech-to-Text Hallucination Harms", presented at the ACM Conference on Fairness, Accountability, and Transparency (FAccT), beginning June 3. The findings are published on the *arXiv* preprint server.

She said, "That can lead to huge downstream consequences if these transcriptions are used in the context of AI-based hiring, in courtroom trials or patient notes in medical settings."

Released in 2022, OpenAI's Whisper is trained on 680,000 hours of audio data and, according to OpenAI, can transcribe audio data with near human-level accuracy. OpenAI has improved its model behind Whisper since researchers carried out this work last year, and the hallucination rate has decreased, Koenecke said.

In their analysis, researchers found that roughly 1% of Whisper's audio transcriptions contained entire hallucinated phrases—including references to websites, real and fake, that could be reappropriated for cyberattacks. For instance, in one sound bite, Whisper correctly transcribed a single, simple sentence, but then hallucinated five additional sentences that contained the words "terror," "knife" and "killed," none of which were in the original audio.

In other examples of hallucinated transcriptions, Whisper conjured random names, fragments of addresses and irrelevant—and sometimes completely fake—websites. Hallucinated traces of YouTuber lingo, like "Thanks for watching and Electric Unicorn," also wormed into transcriptions.

Researchers ran more than 13,000 speech clips into Whisper. The audio data came from AphasiaBank, a research-specific repository of audio recordings of people with aphasia, a condition that limits or completely impairs a person's ability to speak. The repository also includes clips from people with no speech impairments. From their analysis, researchers hypothesize that longer pauses and silences between words are more likely to trigger harmful hallucinations.

"It seems that the underlying large language model-type technology is seeding silence as some sort of word," Koenecke said, adding that Whisper hallucinated "Thank you" after analyzing a silent audio file.

Koenecke cautions that even a small percentage of harmful hallucinations could do real damage.

"On average, your transcriptions are going to be great, but on the margin, for a very small share of these, you might have a massive potential harm," she said.

Training the large language models (LLMs) underlying Whisper with audio data from diverse speakers is a good start, Koenecke said, but a better solution is for OpenAI to tweak its model to account for the different ways people speak, particularly those with speech impairments.

"The problem is less about the quantity of data, and more about the actual modeling choices that are being used here," Koenecke said. "We hope that more can be done to ameliorate these hallucinations, especially

for speakers who tend to speak with longer nonvocal durations."

Along with Koenecke, paper co-authors are doctoral student Anna Seo Gyeong Choi, Katelyn Mei of the University of Washington, Hilke Schellmann of New York University, and Mona Sloane of the University of Virginia.

More information: Allison Koenecke et al, Careless Whisper: Speech-to-Text Hallucination Harms, *arXiv* (2024). DOI: 10.48550/arxiv.2402.08021

Provided by Cornell University