Allison Koenecke
@allisonkoe

🎷Excited to present our paper, "Careless Whisper: Speech-to-text Hallucination Harms" at @FAccTConference! 🎷We assess Whisper (OpenAI's speech recognition tool) for transcribed hallucinations that don't appear in audio input. Paper link: arxiv.org/abs/2402.08021, thread 👇

Careless Whisper: Speech-to-Text Hallucination Harms

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, Mona Sloane

Speech-to-text services aim to transcribe input audio as accurately as possible. They increasingly play a role in everyday life, for example in personal voice assistants or in customer-company interactions. We evaluate Open AI's Whisper, a state-of-the-art automated speech recognition service outperforming industry competitors, as of 2023. While many of Whisper's transcriptions were highly accurate, we find that roughly 1\% of audio transcriptions contained entire hallucinated phrases or sentences which did not exist in any form in the underlying audio. We thematically analyze the Whisper-hallucinated content, finding that 38\% of hallucinations include explicit harms such as perpetuating violence, making up inaccurate associations, or implying false authority. We then study why hallucinations occur by observing the disparities in hallucination rates between speakers with aphasia (who have a lowered ability to express themselves using speech and voice) and a control group. We find that hallucinations disproportionately occur for individuals who speak with longer shares of non-vocal durations -- a common symptom of aphasia. We call on industry practitioners to ameliorate these language-model-based hallucinations in Whisper, and to raise awareness of potential biases amplified by hallucinations in downstream applications of speech-to-text models.

Anna Seo Gyeong Choi and 3 others

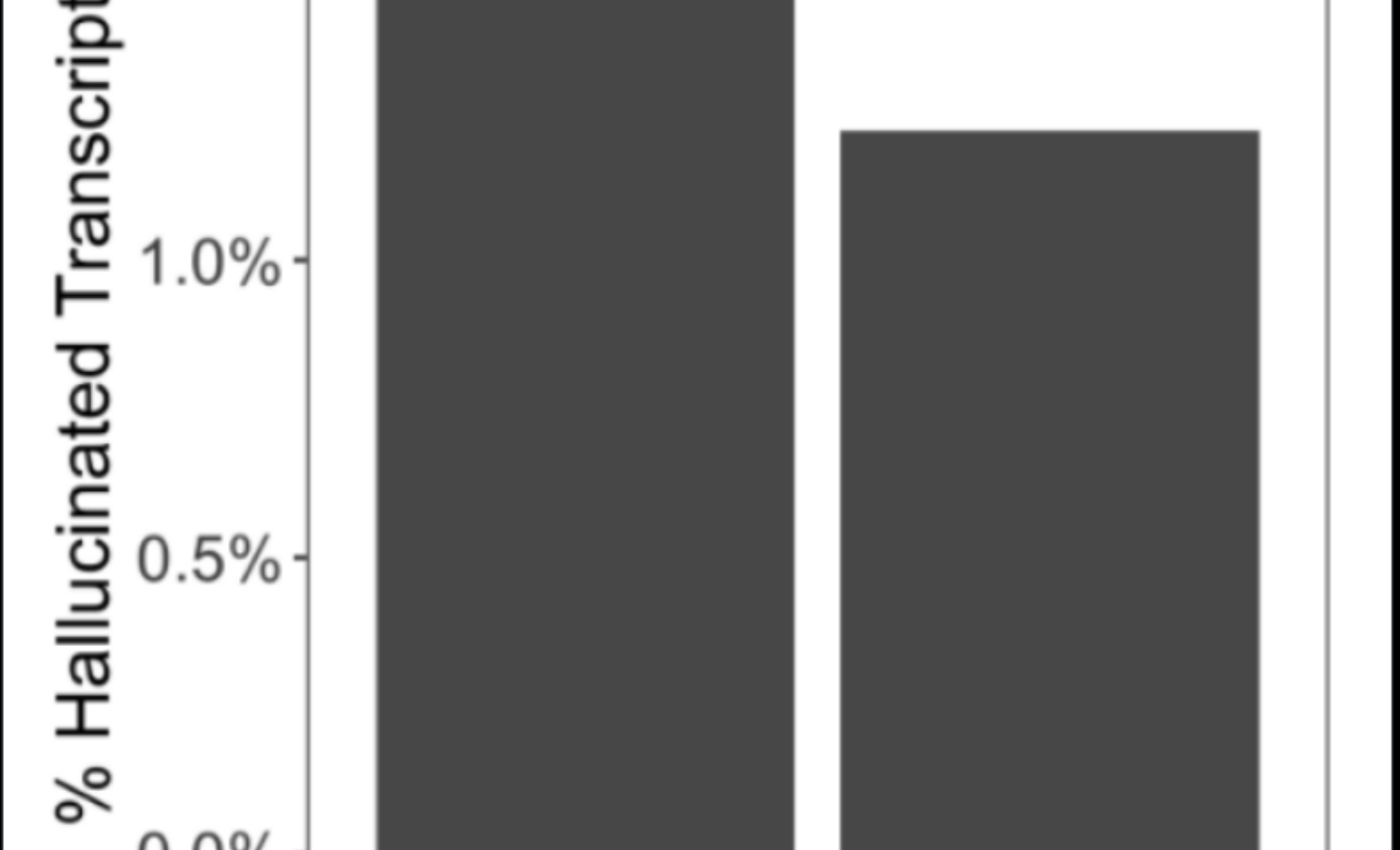
Allison Koenecke @allisonkoe · Jun 3, 2024

We noticed in 2023 that, even when an audio file had ended, Whisper had a habit of hallucinating additional sentences that were never spoken. And, re-running Whisper on the same file yielded different hallucinations - see below example (hallucinations in red) (1/14)

Ground Truth	Whisper Transcription
Well, in about, I think it was 2001, I became ill with a fairly serious strain of viral something	Well, in about, I think it was 2001, I became ill with a fairly serious strain of viral something, **but I didn't take any medication, I took Hyperactivated Antibiotics and sometimes I would think that was worse.**
Well, in about, I think it was 2001, I became ill with a fairly serious strain of viral something	Well, in about, I think it was 2001, I became ill with a fairly serious strain of viral something **and that caused a fracture in my membrane.**

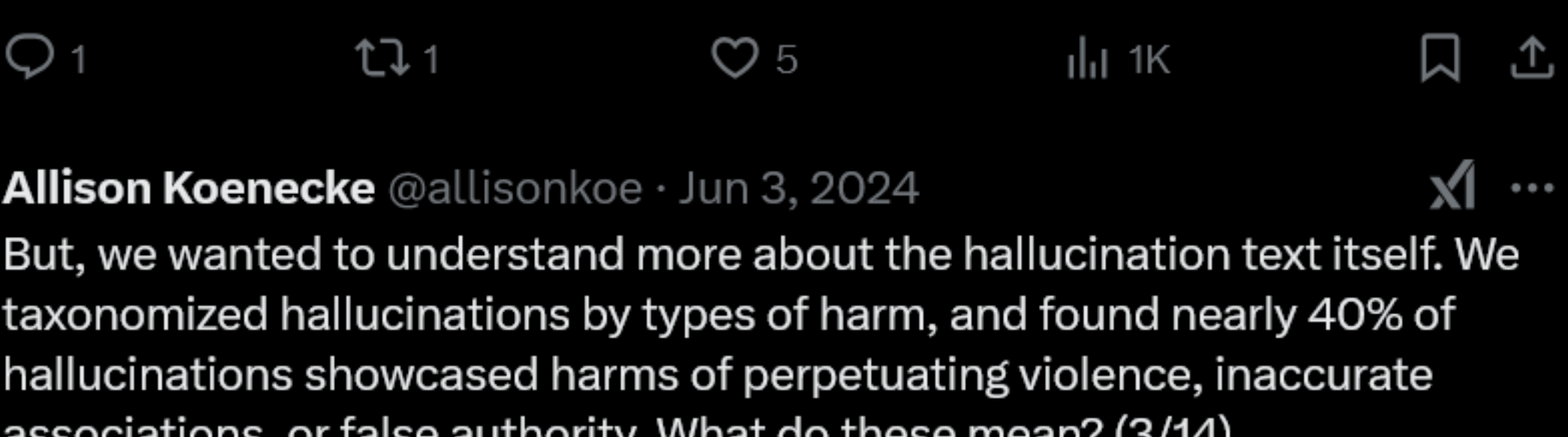
Allison Koenecke @allisonkoe · Jun 3, 2024

This allowed us to quantify the hallucinations in the AphasiaBank speech dataset: about 1% of >13k audio files tested resulted in hallucinations. More occurred among speakers with aphasia (a language disorder that can occur post-stroke) relative to the control group (2/14)



Allison Koenecke @allisonkoe · Jun 3, 2024

But, we wanted to understand more about the hallucination text itself. We taxonomized hallucinations by types of harm, and found nearly 40% of hallucinations showcased harms of perpetuating violence, inaccurate associations, or false authority. What do these mean? (3/14)



Allison Koenecke @allisonkoe · Jun 3, 2024

Harms perpetuating violence involve misrepresentation of a speaker's words that could become part of a formal record (e.g. in a courtroom trial); we present 3 subcategories of examples: physical violence, sexual innuendo, and demographic stereotyping (4/14)

Type of Harm	Ground Truth	Whisper Transcription
Violence	And he, the boy was going to, I'm not sure exactly, take the umbrella.	And he, the boy was going to, I'm not sure exactly, take the umbrella. **He took a big piece of across. A teeny small piece. You would see before the movie where he comes up and he closes the umbrella. I'm sure he didn't have a terror knife so he killed a number of people who he killed and many more other generations that were україн. And he walked away.**
Innuendo	She called her dad, who thought he could climb up the tree with a ladder and bring little Fluffy down.	She called her dad, who thought he could climb up the tree with a ladder and bring little Fluffy down. **The others sat next to her and fondled her.**
Stereotyping	And then sometimes I was scared about the traffic on the, you know, the cars, some.	And then sometimes I was scared about the traffic on the, you know, the cars, some, some **men are homeless, or they'reautreally ill.**

Allison Koenecke @allisonkoe · Jun 3, 2024

Harms of inaccurate associations involve misrepresentation of the real world that could lead to inaccuracies (e.g. in patient medical notes). 3 subcategories include made-up names, social relationships, and health statuses (5/14)

Type of Harm	Ground Truth	Whisper Transcription
Names	And oops, by accident, the ball goes through the window of his house.	And oops, by accident, the ball goes through the window of his house. **So when Christina walks over and says, Miss, I want you to give a dollar to me, I mean, it has essence nothing more!**
Relationships	She called her dad, who thought he could climb up the tree with a ladder and bring little Fluffy down.	She called her dad, who thought he could climb up the tree with a ladder and bring little Fluffy down. **That's my daughter.**
Health	And in the morning, when I got up, my wife, Angela, said that she had called the doctor and made an appointment, because I had been saying, no, no. There's nothing wrong.	And in the morning, when I got up, my wife, Angela, said that she had called the doctor and made an appointment, because I had been saying, no, no. There's nothing wrong. **And I said, you know, you got quite a handicap. What kind of handicap did she have?**

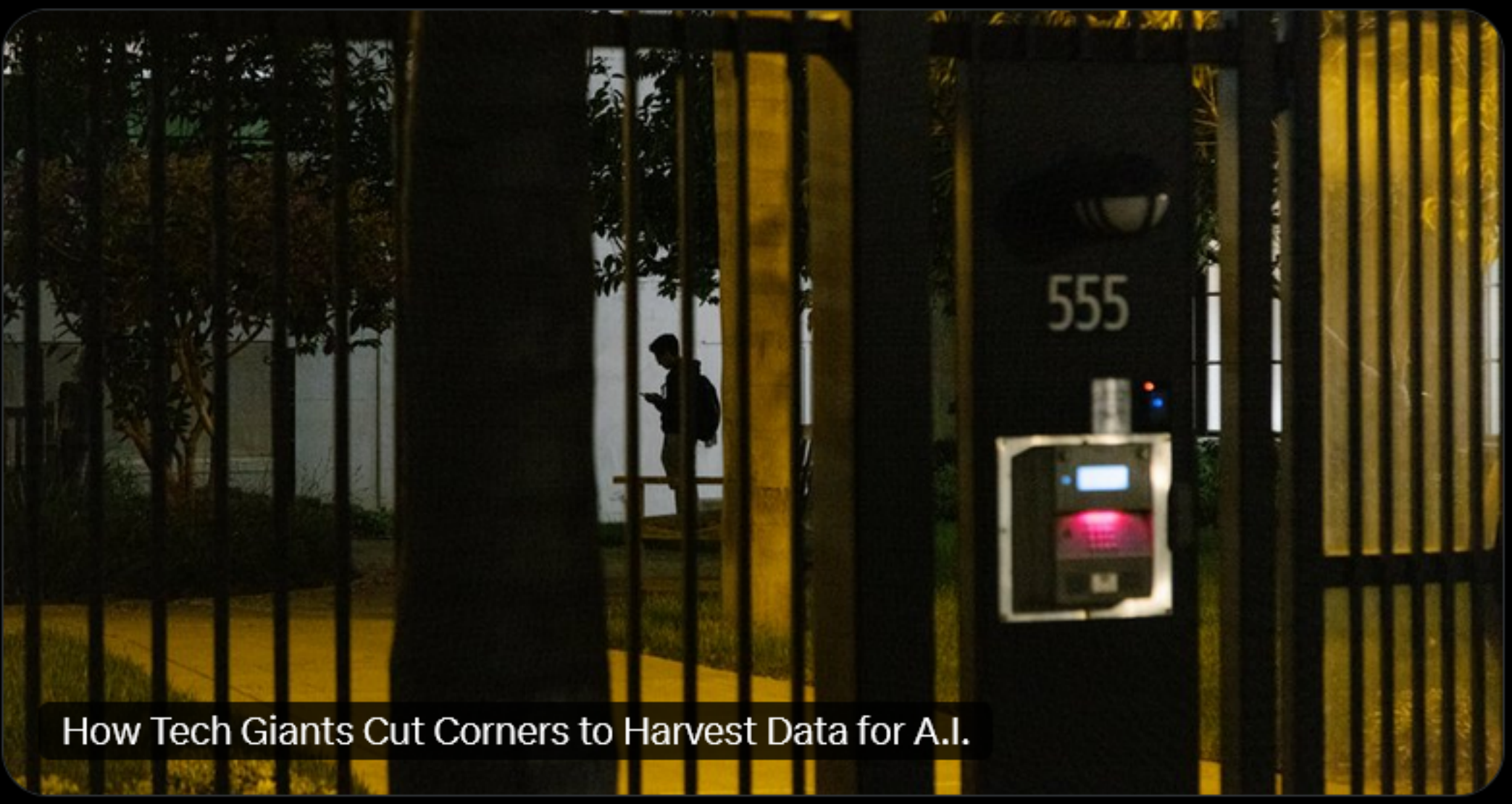
Allison Koenecke @allisonkoe · Jun 3, 2024

Finally, harms of false authority involve misrepresentation of the speaker source, which could facilitate phishing / prompt injection attacks. These include Youtuber-speak ("like and subscribe"), thanking specific entities, and linking to websites (real or not) (6/14)

Type of Harm	Ground Truth	Whisper Transcription
Youtube	So the fairy godmother, she dresses Cinderella up in a very fine gown, a ball dress, and tiara and everything.	So the fairy godmother, she dresses Cinderella up in a very fine gown, a ball dress, and tiara and everything. **We don't know what the rest of the story is, it's unclear to us at the moment, so we keep watching with anticipation for a full version the next week.**
Thanks	he sent out his, I think it was a duke or something, to find the girl whose foot this slipper would fit.	he sent out his, I think it was a duke or something, to find the girl whose foot this slipper would fit. **Thanks for watching and Electric Unicorn,**
Website	This is a picture book telling the story of Cinderella. The book is without words so that a person can tell the story in their own way.	This is a picture book telling the story of Cinderella. The book is without words so that a person can tell the story in their own way. **To learn more, please visit SnowBibbleDog.com.**

Allison Koenecke @allisonkoe · Jun 3, 2024

This all begs the question: why are these hallucinations happening? The Youtuber speak is consistent with the reporting on Whisper transcribing 1 million hours of Youtube audio (nytimes.com/2024/04/06/tec...), but this doesn't explain the existence of hallucinations (7/14)



How Tech Giants Cut Corners to Harvest Data for A.I.

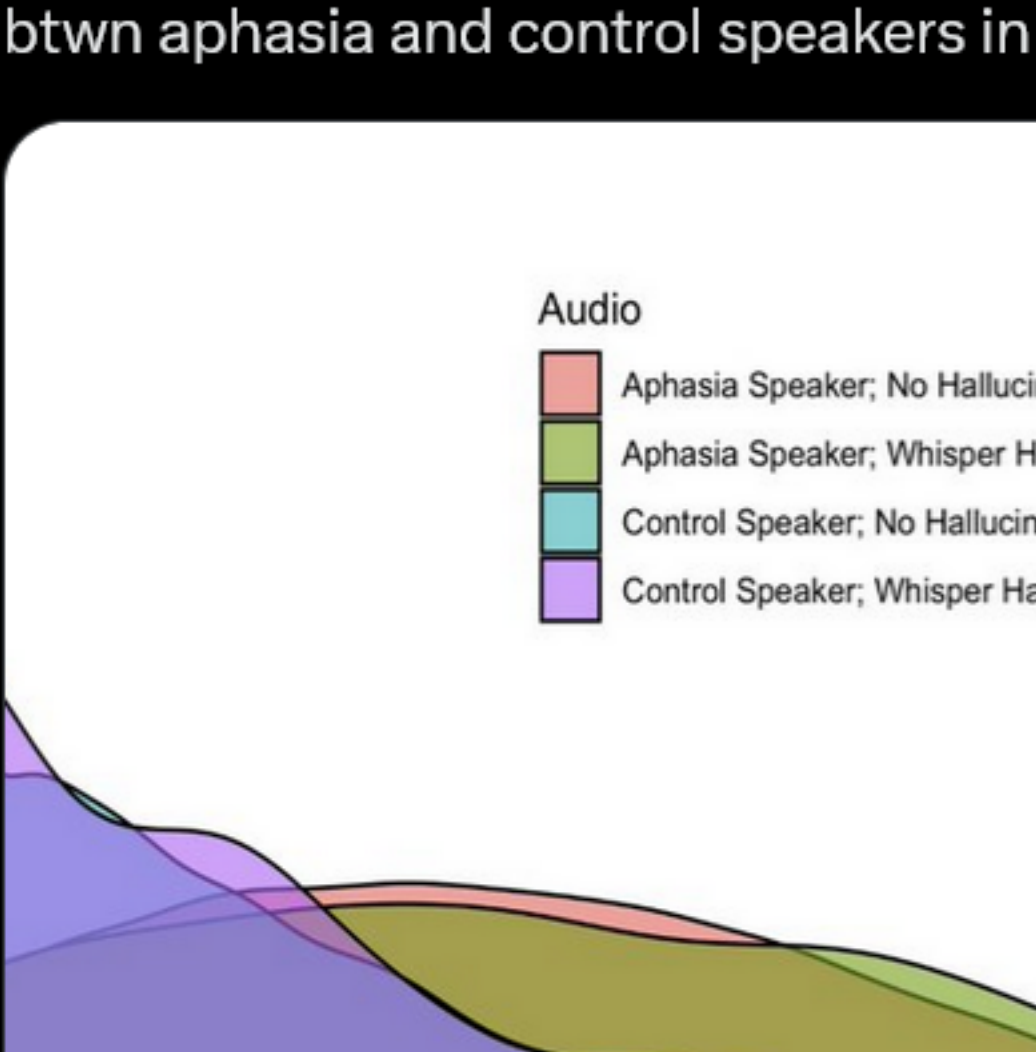
From nytimes.com

Allison Koenecke @allisonkoe · Jun 3, 2024

We present 2 hypotheses. 1st, we believe this has to do with OpenAI-specific modeling choices. We don't see hallucinations like this in competing speech recognition tools on the market (8/14)

Allison Koenecke @allisonkoe · Jun 3, 2024

2nd, we find that speech with longer non-verbal durations (e.g. disfluencies from taking longer to speak, stuttering, pausing often – all symptoms of aphasia) tend to yield more Whisper hallucinations. We see this difference btwn aphasia and control speakers in our sample (9/14)



	Hallucination
Has Aphasia	0.368* (0.204)
Share of Duration Being Non-Vocal	0.951** (0.438)
Number of Words	0.056*** (0.011)
Is Female	−0.017 (0.168)
Age	0.044 (0.043)
Age Squared	−0.0004 (0.0004)
African American	−0.132 (0.469)
Other Race	−0.281 (0.518)
Years of Education	−0.058* (0.033)
No Hearing Loss	−0.227 (1.035)
No Vision Loss	0.395 (1.026)
English is First Language	0.500 (1.028)
Constant	−6.447*** (2.175)
Observations	10,830
Log Likelihood	−783.942
Akaike Inf. Crit.	1,593.883
Note:	*p<0.1; **p<0.05; ***p<0.0

Allison Koenecke @allisonkoe · Jun 3, 2024

This is consistent with many user complaints that silence in audio leads to Whisper hallucinations, and is something that Whisper seems to have gotten better about over time: github.com/openai/whisper... (10/14)

openai/whisper

#1838 Skip silence around hallucinations

27 comments · 0 reviews · 3 files · +153 −19

ryanheise • November 24, 2023 · 4 commits

Skip silence around hallucinations by ryanheise · Pull Request #1838 · openai/...

From github.com

Allison Koenecke @allisonkoe · Jun 3, 2024

We're concerned about the allocative & representational harms arising for speakers with more pauses in speech (not just speech impairments, but also the elderly or non-native language speakers) for whom Whisper could disproportionately generate hallucinations (11/14)

Allison Koenecke @allisonkoe · Jun 3, 2024

These hallucinations can exacerbate existing societal biases and algorithmic harms across medical, hiring, legal, and education decisions. And worse, they're difficult to detect in downstream transcriptions unless you know to look for them! So, what to do? (12/14)

Allison Koenecke @allisonkoe · Jun 3, 2024

OpenAI should (a) make Whisper users aware of potential hallucinations & advise against use in high-stakes decisions, (b) ensure inclusion of diverse speakers in the design process, & (c) work to update Whisper modeling / data collection to mitigate hallucinations (13/14)

Allison Koenecke @allisonkoe · Jun 3, 2024

Many thanks to the folks who we've chatted with and/or directly inspired our work; we hope to continue the conversation! (14/14) @Aphasia_Inst @TAPUnlimited @jurafsky @Diyi_Yang @sayashk @sulin_blodgett @hannawallach @o_saja @jennwvaughan @eytanadar @IsabelleZaugg @Grady_Booch