May 9, 2024
By James A. Larson program co-chair, SpeechTEK 2021
Forward Thinking

Pitfalls Facing Conversational Assistants: Hallucinations and Deepfakes

Large language models are trained by using vast troves of text, audio, and video. The accuracy and consistency of data used to train LLMs are key to the results produced by conversational assistants, those AI-powered helpers, like Siri or Alexa, that speak to users via text, voice, and graphics. But LLMs can present two significant risks: hallucinations and deepfakes.

Hallucinations. Generative AI can create hallucinations, a seemingly authoritative but misleading output that does not seem to be justified by its training data. Some folks would call them exaggerations or lies. The spread of hallucinations could lead to negative consequences and damage people's mental and physical health.

Why does generative AI create hallucinations? Possible reasons include these:

- **Misinterpretation and inappropriate generalizations.** A conversational assistant using the LLM trained with phrases including "Red sky at night, sailors' delight" and "Blue skies smiling at me" might respond to the question "What color is the sky?" with "The sky is both red and blue." LLM models should be refined to avoid making these types of errors.
- **Bad data.** LLMs generated from a database containing gaps, errors, and oversights will pass these problems on to conversational assistants. LLMs trained on incomplete or inaccurate

- **Absence of an external fact-checking mechanism.** As human beings, our life experiences and common sense help us detect and discard hallucinations. Generative AI needs a way to check the factual accuracy of its responses. Ongoing research using cutting-edge algorithms like those investigated by researchers, including Microsoft's Weijia Xu et al., show promise in identifying misinformation. Until advanced detection tools are available, it is crucial to verify information from conversational assistants by critical thinking and cross-checking with reliable sources.

Deepfakes. Generative AI can be used to create deepfakes, which are audio and video files that appear to be real but are generated synthetically. Bad actors use generative AI to create fake audio and video of people saying or doing things they haven't in situations that never existed. These culprits use deepfakes to spread massive amounts of misinformation quickly and efficiently in a variety of forms, especially social media platforms. Fake audio and video could have profound consequences, such as damaging reputations, inciting violence, or upsetting or even terrorizing the targets' family members.

Detecting deepfakes made by generative LLMs is an ever-evolving game of cat and mouse. Techniques you can employ to improve your chances of spotting them include the following:

- **Verify the individual depicted.** It is becoming more difficult for users to recognize if a subject is a deepfake. However, if voice prints or facial characteristics for purported subjects are available, speech recognition or facial recognition algorithms might be able to validate the voice and face.
- **Look for unnatural features and behaviors.** Deepfakes may contain odd or impossible features like missing shadows, incorrect reflections in mirrors or glasses, unrealistic limb lengths, or overly smooth or robotic movements. Subjects may speak unusual phrases, repetitive word choices, or stilted sentence structures. If users detect odd or impossible features or behaviors, then the file is likely a fraud. Research is currently under way to train an AI to detect such abnormalities, but the results to date suffer from excessive false positives (declare a real file to be a fraud) or false negatives (fail to detect a fraudulent file).
- **Recognize inconsistencies with known facts.** If the content claims something that isn't true or contradicts the person's usual behavior or knowledge, that could be a red flag. Fact-checking software that detects false statements may indicate the file is a deepfake.
- **Determine if the source is credible.** If users can determine that the creator of the file is reliable, then the file is likely real. Watermarks (data that is inserted into a file that is not detectable by users but can be detected by software) can also show that a file is real.

While detecting deepfakes remains a challenge, advancements in technology and user awareness can help mitigate their harmful effects.

Regulation and laws. Hallucinations and deepfakes lead to periods of disbelief where the line between reality and hallucinations/deepfakes is blurred, affecting a society's normal functioning. Standards and regulatory bodies need to create laws and regulations to identify and prohibit the use of hallucinations and deepfakes in the commission of illegal activities. Penalties for crimes for using misinformation should be established and enforced. Misinformation is a societal problem that erodes users' confidence in what is read, heard, and seen. x

James A. Larson, an independent voice technology expert, can be reached at jim42@larson-tech.com.

Meeting the Rising Demand for Voice-Based Biometric Systems
Coming December 03, 2024

Avatar Platforms in Customer Service
Coming March 11, 2025

More Web Events

POPULAR ARTICLES

The Evolution and Importance of Voice Surveillance

Colossyan Introduces Instant Avatar Feature

Wondercraft Launches Director Mode

Agora Launches Conversational AI SDK