Hacking Heathcare - Weekly Blog

Hacking Healthcare	TLP:WHITE	Alert Id: 3a1dfb11	2024-11-07 20:38:35

This week, Health-ISAC®'s Hacking Healthcare® examines research suggesting that a well-known artificial intelligence (AI) transcription model that is being used as the underlying tool in healthcare products may be worryingly prone to “hallucinating” words or even entire sentences. This week, we assess what the research has found and provide some general considerations for healthcare organizations eager to take advantage of AI capabilities.

Welcome back to Hacking Healthcare®.

Hallucinating AI Tool Prompts Healthcare Concern

AI developers and policymakers routinely tout the transformative capabilities of AI in sectors like healthcare. For example, within healthcare delivery organizations, AI has the potential to aid in enhancing medical imagery analysis, ease patient scheduling, or more effectively route IT help desk requests. However, recent news articles[i] [ii] have reiterated reasons why organizations should be careful when adopting emerging technologies that may not be as safe, secure, or reliable as they claim or are assumed to be.

Nabla’s Healthcare AI Assistant

The ability to reduce administrative burdens so that healthcare providers can spend more time focusing on the patient and caregiving is an understandably attractive quality for an AI tool. One such product advertised as providing just that is an “AI assistant” produced by Nabla. Nabla claims their AI assistant is capable of “pre-charting, medical codification, clinical decision prompting,” and more.[iii] It would appear that this suite of capabilities has been very well received, as Nabla’s website suggests that their product is already deployed in over 85 health organizations and is being used by more than 45,000 clinicians.[iv]

One of the highlights of Nabla’s AI assistant is its ability to transcribe clinician-patient interactions into appropriate clinical notes with a high degree of accuracy.[v] Accuracy is obviously critical in this context, given how inaccurate transcriptions may risk significant patient harm. For example, failing to accurately capture a patient’s allergen history or their current medicine regimen and dosage could lead to the wrong healthcare decisions down the road. It is this aspect of the tool that has come under scrutiny in recent weeks due to a study that has called into question the accuracy of the underlying AI model that Nabla’s tool is based on.

OpenAI and Tool Development

If you were ever curious how so many AI and AI-enabled products have been able to come to market so quickly despite the relative complexity and newness of AI, part of the answer is the use of existing tools as a basis upon which other companies can build something more specialized or complex. This is the case with Nabla’s AI assistant, which employs OpenAI’s *Whisper*[vi] as the underlying tool.[vii] For those unfamiliar, OpenAI’s *Whisper* is described as an automatic speech recognition (ASR) system that is described as “[approaching] human level robustness and accuracy on English speech recognition.”[viii]

So what’s the issue?

Whisper’s Accuracy Woes

According to recent research, OpenAI's *Whisper* is more error prone than might be appreciated.[ix] While you may be thinking, "it's only natural that something like a name might be misspelled or that a heavy accent might slightly skew a transcription," the errors reported were a bit more concerning. It has been reported that *Whisper* is "prone to making up chunks of text or even entire sentences" and that a University of Michigan researcher found "hallucinations in eight out of every 10 audio transcriptions" that they had reviewed.[x] Other users of *Whisper* supported this finding, with one claiming that they had found "hallucinations in nearly every one of the 26,000 transcripts he created with Whisper."[xi]

You can see how this may be concerning to users of Nabla's AI assistant. However, things are not quite as straightforward as intuiting that Nabla's product is inherently flawed or subject to the same concerning hallucination issues.

Adapting Whisper & Nabla's Response

Models like *Whisper* are designed to be built upon. Without getting too technical, they can be trained on new sources of data and adjustments can be made to numerous variables, such as how they weigh certain aspects or interpret instructions. In essence, they can be fine-tuned to better specialize in a particular task or subject matter.

According to Nabla, the limitations of *Whisper* were known to them and it is why they say they spent several years and millions of dollars to "[gather] and manually [annotate] a unique dataset of 7,000 hours of medical encounters audio" to better refine it.[xii] Furthermore, Nabla claims there are additional improvements and safeguards to "suppress" hallucinations and limit the potential for inaccuracies to make it onto a patient's record.[xiii]

In the Action & Analysis section below, we will provide some high-level takeaways for Health-ISAC members on how to think about employing AI tools, as well as some considerations specific to Nabla's case.

Action & Analysis

Included with Health-ISAC Membership

[i] https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14#

[ii] https://www.wired.com/story/hospitals-ai-transcription-tools-hallucination/

[iii] https://www.nabla.com/

[iv] https://www.nabla.com/

[v] Nabla's marketing refers to "95% note accuracy" in relation to "15 seconds note generation"

[vi] https://openai.com/index/whisper/

[vii] https://www.nabla.com/blog/how-nabla-uses-whisper/

[viii] https://openai.com/index/whisper/

[ix] https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14#

[x] https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14#

[xi] https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14#

[xii] https://www.nabla.com/blog/how-nabla-uses-whisper/

[xiii] https://www.nabla.com/blog/how-nabla-uses-whisper/

[xiv] https://www.nabla.com/blog/how-nabla-uses-whisper/

[xv] https://www.nabla.com/blog/assessing-reliability-nabla-speech-to-text/

[xvi] https://www.nabla.com/blog/how-nabla-uses-whisper/

Reference(s): nabla, openai, AP News, nabla, nabla, Wired
Report Source(s): Health-ISAC
Release Date: Nov 08, 2024 (UTC)

Tags: Emerging Technology, Regulation, Regulatory, Hacking Healthcare, United States (U.S.), Artifical Intelligence (AI), Artificial Intelligence
TLP:WHITE: Subject to standard copyright rules, TLP:WHITE information may be distributed without restriction.
Conferences, Webinars, and Summits:
https://h-isac.org/events/
Hacking Healthcare:
Hacking Healthcare is co-written by John Banghart and Tim McGiff.

John Banghart has served as a primary advisor on cybersecurity incidents and preparedness and led the National Security Councils efforts to address significant cybersecurity incidents, including those at OPM and the White House. John is currently the Senior Director of Cybersecurity Services at Venable. His background includes serving as the National Security Councils Director for Federal Cybersecurity, as Senior Cybersecurity Advisor for the Centers for Medicare and Medicaid Services, as a cybersecurity researcher and policy expert at the National Institute of Standards and Technology (NIST), and in the Office of the Undersecretary of Commerce for Standards and Technology.

Tim McGiff is currently a Cybersecurity Services Program Manager at Venable, where he coordinates the Health-ISACs annual Hobby Exercise and provides legal and regulatory updates for the Health-ISACs monthly Threat Briefing.

- John can be reached at jbanghart@h-isac.org and jfbanghart@venable.com.
- Tim can be reached at tmcgiff@venable.com.

Share Threat Intel:
For guidance on sharing indicators with Health-ISAC via HTIP, please visit the Knowledge Base article "HTIP - Share Threat Intel" here.

The "Share Threat Intel" feature allows for attributed or anonymous sharing across ISACs and other cybersecurity-related entities.
Turn off Categories:
For guidance on disabling alert categories, please visit the Knowledge Base article "HTIP Alert Categories" here.
Access the Health-ISAC Threat Intelligence Portal:
Enhance your personalized information-sharing community with improved threat visibility, alert notifications, and incident sharing in a trusted environment delivered to you via email and mobile apps. Contact membership@h-isac.org for access to Health-ISAC Threat Intelligence Portal (HTIP).
For Questions or Comments:
Please email us at toc@h-isac.org