# Team creates software to block AI phishing scams

June 21 2024, by Brian Lopez

Credit: Pixabay/CC0 Public Domain

A team of researchers at the University of Texas at Arlington has developed software that prevents artificial intelligence (AI) chatbots such as ChatGPT from creating phishing websites—a growing concern

as cybercriminals have been utilizing the technology for designing scams.

Created by Shirin Nilizadeh, assistant professor in the Department of Computer Science and Engineering, and her doctoral students Sayak Saha Roy and Poojitha Thota, the software allows AI chatbots to better detect and reject instruction prompts entered by users that could be used to create phishing websites.

Currently, AI chatbots have some inbuilt detection capabilities, but Dr. Nilizadeh said her team has found loopholes that could easily bypass them and exploit the chatbots to create these attacks. With the emergence of AI chatbots, launching online scams has become highly accessible, even for attackers with minimal technical skills. Now, one does not need coding expertise to create a website, as AI can build one almost instantly.

"These tools are very powerful, and we are showing how they can be misused by attackers," Nilizadeh said.

To develop their tool, the group initially identified various instruction prompts that could be used to create phishing websites, Saha Roy said. Leveraging this knowledge, they successfully trained their software to recognize and react to those specific keywords and patterns, enhancing its ability to detect and block such malicious prompts from being executed by the chatbots.

The team's work has captured significant attention within the cybersecurity industry, highlighted by their recent publication at the IEEE Symposium on Security and Privacy (IEEE S&P 2024). In May, the researchers not only shared their findings but also received the Distinguished Paper Award, further underscoring the impact of their research.

"I want people to be receptive to our work and see the risk," Saha Roy said. "It starts with the security community and trickles down from there."

The researchers have reached out to the major tech companies that drive these chatbots, including Google and OpenAI, aiming to integrate their findings into broader AI security strategies. Both Saha Roy and Thota expressed a strong commitment to their research's implications for cybersecurity.

"I'm really happy that I was able to work on this important research," Thota added. "I'm also looking forward to sharing this work with our colleagues in the cybersecurity space and finding ways to further our work."

**More information:** Sayak Saha Roy et al, From Chatbots to Phishbots?: Phishing Scam Generation in Commercial Large Language Models, *2024 IEEE Symposium on Security and Privacy (SP)* (204). DOI: 10.1109/SP54263.2024.00182. www.computer.org/csdl/proceedi … 3000a221/1WPcYLpYFHy

Provided by University of Texas at Arlington