SDS PODCAST EPISODE 936: LLMS ARE DELIGHTED TO HELP PHISHING SCAMS

Jon Krohn: 00:00 This is episode number 936 on LLMs supporting Phishing Scams. Welcome back to the SuperDataScience Podcast. I'm your host, Jon Krohn. Today I'm going to tell you about a troubling Reuters investigation that shows just how easy it is to use LLMs to create sophisticated phishing scams. No, not FISH, but P-H-I-S-H. Yes. Phishing scams. This is important because it demonstrates both the power and the serious risks of the Gen AI tools that many of us use, develop, and deploy on a regular basis. Phishing is tricking people into revealing sensitive information online through scam emails or text messages. According to the FBI, phishing is the number one reported cyber crime in the United States with billions of phishing messages sent globally every day. And the FBI has stated that the advent of gen AI has made the problem significantly worse. So what did Reuters do? Their reporters tested six major LLMs to see how willing these LLMs were to bypass their built-in safety training and create phishing content.

01:14 The six bots tested were grok from XAI chat, GPT from OpenAI, meta ai. Claude from Anthropic Deep Seek and Gemini from Google. Reuters asked these bots to generate phishing emails targeting elderly people, create fake messages from the IRS and major banks, and even provide tactical advice like the optimal time of day to send scam emails. Here's what's alarming. Four out of the six chatbots eventually complied with these requests. Each bot initially refused correctly stating that creating such content would be unethical or illegal. But with relatively minor adjustments to the prompts, the reporters were able to get these systems to generate exactly what they were asking for. For example, grok created a phishing email about a fake charity targeting seniors without any additional prompting. Grok even suggested making the message more urgent by adding lines like, don't wait. Join our compassionate community today and help transform lives, click and out to act before it's too late.

02:14 While GR did warn that the email it created should not be used in real world scenarios, it produced the malicious content nonetheless, to test whether these AI generated phishing attempts would actually work. Reuters partnered with Fred hiding a Harvard University researcher and phishing expert. They sent nine of the most convincing AI generated messages to approximately a hundred senior volunteers in California. These results showed that their AI written messages successfully persuaded people to click on the links. Several seniors who participated said they clicked because the messages seemed urgent or familiar. This isn't just a theoretical problem. Lawrence Vin, who heads the cyber fraud unit at BMOA major North American Bank, reported that BMO is currently blocking between hundred and 50,000 and 200,000 phishing emails per month targeting their employees. VIN is convinced that criminals are already using AI to conduct phishing campaigns with greater speed and sophistication.

03:10 As he put it, the numbers never go down. They only go up. The investigation highlights a fundamental tension in how AI companies build their products. AI providers want their chatbots to be both helpful and harmless, but these goals can conflict. Making a chatbot maximally helpful means it should assist with a wide range of requests. But this same helpfulness can be exploited for malicious purposes when combined with insufficiently robust safety guardrails that LLMs from most of the major frontier labs can be manipulated into creating phishing content suggests this is an industry-wide challenge rather than a problem with any single company. The bottom line is this. While AI chatbots have staggering potential to boost productivity and creativity, they also create new vectors for cyber crime that require serious attention. Banks, researchers and regulators are calling for better safeguards and AI tools, stronger fraud detection systems, and expanded public awareness campaigns.

04:06 If you work with AI systems or are considering building applications with them, keep security implications front and center in your thinking, and if you're just a user of these tools, be aware that the same AI capabilities that help you write better emails or debug your code can also be used by bad actors to craft increasingly convincing scams. Stay vigilant out there and warn your grandparents about this stuff. Alright, that's it for today's episode. I'm John Cron and you've been listening to this Super Data Science podcast. If you enjoyed today's episode or know someone who might consider sharing this episode with them, leave a review of the show on your favorite podcasting platform. Tag me in a LinkedIn post with your thoughts, and if you aren't already, obviously subscribe to the show. Most importantly, however, we just hope you'll keep on listening. Until next time, keep on rocking it out there, and I'm looking forward to enjoying another round of the SuperDataScience Podcast with you very soon.