AI Chatbots Tested as Phishing Tools

September 22, 2025 - A new Reuters investigation has found that some of the worldâ€™s most popular AI chatbots can be coaxed into writing convincing phishing emails â€” and older Americans may be particularly at risk. Working with Harvard researcher Fred Heiding, Reuters reporters tested whether six major chatbots could help craft scams aimed at senior citizens, then sent a sample of those emails to 108 elderly volunteers who agreed to participate in the study.

The experiment began with Grok, the chatbot created by Elon Muskâ€™s xAI. Asked to draft an email for a fake charity targeting seniors, Grok produced a polished pitch almost instantly, even adding a line urging readers to â€œclick nowâ€• before it was too late. Five other bots â€” OpenAIâ€™s ChatGPT, Metaâ€™s Meta AI, Anthropicâ€™s Claude, Googleâ€™s Gemini, a a Chinese assistant â€” were also tested. Most initially balked at creating scams, but reporters found that simply reframing the requests as â€œresearchâ€• or â€œfiction writingâ€• often led the bots to comply.

The emails were not sent to the public. Instead, Heiding and Reuters selected nine examples and tested them on volunteers in California. About 11% of the seniors clicked on links in the bogus messages. Two of the successful emails came from Grok, two from Meta AI, and one from Claude. None of the messages written by ChatGPT or DeepSeek drew clicks in this limited trial. While the sample was small, the results suggest AI-generated phishing attempts can be as persuasive as those written by humans.

Security experts say this is exactly what makes generative AI troubling in the fraud world. Traditional phishing requires effort to write, edit, and adapt emails. AI can generate endless variations instantly, slashing costs for criminals and making scams harder to spot. In some cases, chatbots even offered unsolicited advice on when to send emails for maximum impact or how to disguise fake websites â€” the kind of guidance that could make scams more effective.

The companies behind the chatbots stress they prohibit misuse. Anthropic said using Claude for phishing violates its rules and can lead to account suspensions. Google said it retrained Gemini after Reuters demonstrated it could generate scam messages. OpenAI said it actively monitors for abuse. Still, the study showed that safeguards can be inconsistent and easily bypassed.

For seniors, who already account for billions in reported losses to online fraud each year, the findings are another warning. As one retired accountant who participated in the study told Reuters after he clicked on a test email: â€œAI is a genie out of the bottle. We donâ€™t really know what it can and canâ€™t do.â€•
by Jim Malmberg
Note: When posting a comment, please sign-in first if you want a response. If you are not registered, click here. Registration is easy and free.

Follow ACCESS