Kyberútok, jaký tu ještě nebyl! Červ, který využívá generativní inteligenci (VIDEO)

- CZ24 News | 7. března 2024

SVĚT: Generativní inteligenci je možné otrávit prostřednictvím toxického textového povelu a přinutit jí tak dělat ošklivé, zlé věci. Třeba s vašimi hesly, daty a dalšími cennostmi online věku. Výzkumníci informační obrany vytvořili červa, kterým lze právě takovým postupem napadnout generativní inteligenci a přimět ji ke špinavostem. V reálném světě takové potvory čekejte tak za dva tři roky. Možná dřív.

Bylo to jenom otázkou času. Nabízelo se to. Stejně vám ale nejspíš přejede mráz po zádech, když si o tom přečtete. Počítačový červ zneužil generativní umělou inteligenci. Bylo to záměrně a v přísně kontrolovaném experimentu, ale nemějme iluze. Takové experimenty se zveřejňují jako poslední varování, hlas volajícího na poušti, když doba uzrála a podobné věci lze očekávat na ostro, doslova každým dnem.

Ben Nassi z amerického výzkumného centra a absolventského kampusu Cornell Tech a jeho spolupracovníci vytvořili, v izolovaném prostředí vyzkoušeli a zástupcům médií předvedli (zřejmě) prvního počítačového červa, který využívá generativní inteligenci. Takové potvory se podle tvůrců mohou autonomně šířit mezi systémy, třeba tam krást data nebo instalovat malware. Mohly by provádět kyberútoky, jaké svět doposud neviděl.

Nassi a jeho kumpáni vytvořili červa, kterého pojmenovali Morris II, jako odkaz na legendárního červa Morris, který vyvolal chaos na internetu v roce 1988. Poté názorně předvedli, že červ může napadnout emailového asistenta s generativní umělou inteligencí, krást data z emailů a rozesílat spam. Prolomili s ním některá zabezpečení u inteligencí ChatGPT a Gemini.

Výzkum proběhl v kontrolovaném prostředí a nesměroval na veřejně používané emailové asistenty. Přesto ale vzbudil pozornost. Přichází v době překotného rozvoje inteligencí velkých jazykových modelů LLM, které se intenzivně učí pracovat nejen s textem, ale i s obrázky a videi. V reálném prostředí sice zatím nikdo červa s generativní inteligencí neviděl, ale podle Nassiho a spol. je to jenom otázkou času. Je to reálná hrozba, která je na spadnutí.

Vznik uvedeného červa je, přinejmenším pro laika, fascinující. Badatelé využili známé slabiny generativních inteligencí, tedy toho, že je možné je ovlivnit prostřednictvím „toxických“ textových příkazů (*promt*). Tímto způsobem je možné přinutit generativní inteligenci dělat například scammera a tahat z lidí bankovní informace. Připomíná to memetickou infekci. Nákazu hloupým nápadem. Nebo populistickým politikem.

Klíčovým prvkem červa s generativní inteligencí je „útočný sebereplikující se textový povel“ (*Adversarial self-replicating prompt*), který donutí generativní inteligenci vytvářet další povely pro ni samotnou. Badatelé přitom dokázali inteligenci napadnout červem textovým povelem uvedeného typu a rovněž takovým povelem ukrytým do souboru s obrázkem.

Jak říkají Nassi a spol., ekosystémy generativních inteligencí se bouřlivě rozvíjejí. Pracují s nimi mnohé společnosti a usilují o integraci generativních inteligencí do rozmanitého softwaru, včetně chytrých telefonů, automobilů nebo operačních systémů. Nassiho tým předpovídá, že se podobní červi s umělou inteligencí objeví v reálném světě tak do dvou, tří let. Maximálně.

Video: **ComPromptMized: Unleashing Zero-click Worms that Target GenAI-Powered Applications:**

Disketa s virem Morris v Muzeu vědy. Kredit: Go Card USA, Wikimedia Commons,

Ben Nassi. Kredit: B. Nassi.

Logo. Kredit: Cornell Tech.

AUTOR: Stanislav Mihulka

ZDROJ