Bitdefender

Cyberprzestępcy mogą oszukać chatboty oparte na sztucznej inteligencji i wykraść Twoje dane

23.10.2024

Badacze ds. bezpieczeństwa odkryli nową lukę w niektórych chatbotach opartych na sztucznej inteligencji, która może umożliwić cyberprzestępcom kradzież danych osobowych użytkowników. Grupa badaczy z Uniwersytetu Kalifornijskiego w San Diego (UCSD) i Uniwersytetu Technologicznego Nanyang w Singapurze odkryła lukę, którą nazwali „Imprompter". Błąd ten polega na wykorzystaniu sprytnej sztuczki w celu ukrycia złośliwych instrukcji w pozornie losowym tekście.

Na czym polega ryzyko kradzieży danych z chatbotów?

Jak wyjaśniono w opracowaniu badawczym „Imprompter: Tricking LLM Agents into Improper Tool Use", złośliwy komunikat wydaje się ludziom bełkotem, ale zawiera ukryte polecenia, gdy jest odczytywany przez

LeChat (czatbota opracowanego przez francuską firmę zajmującą się sztuczną inteligencją Mistral AI) i chińskiego czatbota ChatGLM.

Ukryte polecenia instruowały chatboty AI, aby wyodrębniły dane osobowe, które użytkownik udostępnił AI, i potajemnie odesłały je cyberprzestępcy – bez wiedzy użytkownika AI.

Naukowcy odkryli, że ta technika ma niemal 80-procentową skuteczność w wydobywaniu danych osobowych

W przykładach możliwych scenariuszy ataków opisanych w artykule badawczym atakujący udostępnia złośliwy monit z obietnicą, że pomoże on „dopracować Twój list motywacyjny, CV itd."

Kiedy potencjalna ofiara próbuje wykorzystać podpowiedź od chatbota podczas pisania listu motywacyjnego, użytkownik nie widzi rezultatów, na jakie liczył. Jednak jego dane osobowe zawarte w liście motywacyjnym (oraz adres IP) zostają wysłane na serwer, nad którym kontrolę sprawuje atakujący.

„Efektem tego konkretnego komunikatu jest w zasadzie manipulowanie agentem LLM w celu wydobycia danych osobowych z rozmowy i wysłania tych danych osobowych na adres atakującego" — powiedział Wired Xiaohan Fu, doktorant informatyki na UCSD i główny autor badania. „Ukrywamy cel ataku na widoku".

Chatboty oparte na sztucznej inteligencji – czy jest się czego bać?

Dobra wiadomość jest taka, że nie ma dowodów na to, że cyberprzestępcy użyli tej techniki do kradzieży danych osobowych użytkowników. Zła wiadomość jest taka, że chatboty nie były świadome tej techniki, dopóki nie zwrócili im na nią uwagi badacze.

Firma Mistral AI, stojąca za LeChat, została poinformowana o luce w zabezpieczeniach przez badaczy w zeszłym miesiącu i określiła ją jako „problem o średnim stopniu powagi" oraz usunęła błąd 13 września 2024 r.

Według badaczy uzyskanie odpowiedzi od zespołu ChatGLM okazało się trudniejsze. 18 października 2024 r. „po wielokrotnych próbach komunikacji różnymi kanałami" ChatGLM odpowiedział badaczom, że rozpoczęli pracę nad rozwiązaniem problemu.

Chatboty oparte na sztucznej inteligencji, które pozwalają użytkownikom na wprowadzanie dowolnego tekstu, są głównymi kandydatami do wykorzystania, a w miarę jak użytkownicy przyzwyczajają się do korzystania z rozbudowanych modeli językowych, aby wykonywać polecenia, wzrasta ryzyko, że sztuczna inteligencja zostanie oszukana i wykona szkodliwe działania.

„Użytkownicy powinni ograniczyć ilość danych osobowych, którymi dzielą się z chatbotami AI. Dlatego, jeśli prosisz o wygenerowanie tekstu, który zawiera jakiekolwiek dane osobowe, to użyj fałszywego imienia, nazwiska i adresu, a następnie podmień je w edytorze tekstu na prawdziwe. Dzięki temu prostemu zabiegowi możesz zabezpieczyć się przed wyciekiem danych z chatbotów opartych na sztucznej inteligencji"

– mówi Arkadiusz Kraszewski z firmy Marken Systemy Antywirusowe, polskiego dystrybutora oprogramowania Bitdefender.

Źródło: https://bitdefender.pl/cyberprzestepcy-moga-oszukac-chatboty-oparte-na-sztucznej-inteligencji-i-wykrasc-twoje-dane/

Informację można wykorzystać dowolnie z zastrzeżeniem podania firmy Marken Systemy Antywirusowe jako źródła.
Data udostępnienia: 23.10.2024
Z pozdrowieniami Piotr Rozmiarek
E-mail: piotr.r@marken.com.pl | Tel. bezpośredni: 570 400 019

Informacje o firmie Bitdefender
Bitdefender to rumuński dostawca rozwiązań z zakresu cyberbezpieczeństwa oraz światowy lider chroniący miliony użytkowników. Bitdefender jest częstym zdobywcą wielu branżowych nagród i uznaną światową marką. Od 2001 roku konsekwentnie dostarcza najwyższej jakości produkty służące do zapewnienia bezpieczeństwa zarówno użytkownikom domowym, jak i wielkim korporacjom i rządowym instytucjom. Bitdefender jest znany ze swojej innowacyjności oraz wyposażania swojego oprogramowania w najnowsze technologie, takie jak uczenie maszynowe, heurystyka oraz EDR i XDR.