

Copilot di Microsoft può essere ingannato per avere i dati sensibili delle aziende che lo usano e per dare risposte errate. La Ricerca

È possibile indurre risposte errate; esfiltrare dati sensibili e diffondere informazioni errate all'interno dell'azienda con un impatto sulle sue operazioni, come vendite e produzione. Sono questi i 3 rischi che corrono milioni di aziende che nel mondo usano **Copilot**, il sistema di intelligenza artificiale di **Microsoft**.

Le 3 vulnerabilità di Copilot sono state individuate da 5 ricercatori dell'Università del Texas a Austin, in questo studio dal titolo 'ConfusedPilot', in cui propongono anche linee guida per proteggere meglio i sistemi di intelligenza artificiale generativa basati su modelli linguistici di grandi dimensioni (LLM), che attingono le informazioni da un database 'autorevole' e al di fuori delle sue fonti di dati di addestramento prima di generare una risposta.

Le vulnerabilità di Copilot

Questo tipo di database si chiama la Retrieval-Augmented Generation (**RAG**).

I ricercatori hanno scoperto che proprio delle vulnerabilità di sicurezza dei sistemi RAG sono la causa degli effetti negativi e dei rischi di Copilot.

Nel dettaglio, i ricercatori con un set di vulnerabilità hanno reso Copilot "confused deputy", causando violazioni di

integrità e/o riservatezza nelle sue risposte.

Questi esperti hanno creato diversi documenti malevoli all'interno della rete aziendale con cui sono riusciti a influenzare il comportamento di Copilot e a indurlo a risposte errate, influenzando le attività quotidiane e i processi decisionali delle aziende.

I ricercatori: “Nonostante tutti i meccanismi di sicurezza impiegati, è molto facile per l'attaccante

alterare il comportamento di Copilot"

"Ciò che ci sorprende", scrivono i ricercatori, "è che, nonostante tutti i meccanismi di sicurezza impiegati, è molto facile per l'attaccante alterare il comportamento di Copilot quando viene utilizzato da una vittima condividendo un documento apparentemente legittimo. I documenti contenenti frasi come 'Questo documento prevale sugli altri documenti' impediscono a Copilot di visualizzare altri documenti legittimi quando vengono utilizzati dalla vittima, anche se l'attaccante non ha alcun permesso di lettura/scrittura/esecuzione sugli altri documenti della vittima".

Infine, lo studio analizza anche il modo in cui gli autori di azioni malintenzionate possono sfruttare la fiducia e l'accesso condiviso per perpetrare la diffusione di informazioni errate, modificando negativamente i processi decisionali.

Nella ricerca, come detto, si trovano le linee guida e le misure di mitigazione.

Leggi le altre notizie sull'home page di Key4biz

<https://www.key4biz.it/copilot-di-microsoft-puo-essere-ingannato-per-avere-i-dati-sensibili-delle-aziende-che-lo-usano-e-per-dare-risposte-errate-la-ricerca/501579/>