Confidare i propri segreti – personali o professionali – a **ChatGPT** non è mai una buona idea. E non solo perché tali informazioni possono essere potenzialmente accessibili all'azienda che controlla lo strumento, **OpenAI**, ma anche perché continuano a emergere criticità che dimostrano come i dati condivisi con il chatbot possano, in determinate circostanze, fuoriuscire dai contesti privati delle conversazioni. A suggerirlo è il fatto che, solamente negli ultimi giorni, la comunità della sicurezza informatica ha puntato i riflettori su ben sette diverse vulnerabilità e tecniche d'attacco che **mettono a rischio la riservatezza dei dati personali**, garantiscono la persistenza di istruzioni malevole e mettono in dubbio la sicurezza delle interazioni tra utenti e modelli linguistici.

A individuare queste falle è stata la società di cybersecurity **Tenable**, la quale ha pubblicato un resoconto dei suoi rinvenimenti sotto il pittoresco titolo di "HackedGPT". Gran parte dei problemi emersi riconduce a un fenomeno già ben noto nel mondo delle intelligenze artificiali generative: la ***prompt injection***, ossia la capacità di un attore malevolo di inserire nell'interazione con il modello delle istruzioni non previste, né desiderate, dall'utente. Ciò può avvenire non solo tramite prompt diretti, ma anche attraverso metodi più subdoli. Nel caso analizzato, i tecnici di Tenable hanno nascosto dei comandi all'interno della **sezione commenti** di un blog creato ad hoc, osservando come ChatGPT, interagendo con la pagina, finisse per eseguire gli ordini celati in calce.

Ancora più preoccupante è che forme di *prompt injection* siano state riscontrate anche durante l'uso della **funzione di ricerca su internet**. Lo staff di Tenable è infatti riuscito a indurre ChatGPT a seguire istruzioni occulte semplicemente creando una pagina web appositamente ottimizzata perchè venisse privilegiata tra le fonti di SearchGPT. In questo modo, gli attaccanti sono riusciti ad aggirare i meccanismi di difesa del sistema con una vulnerabilità che non richiede l'intervento attivo dell'utente, un cosiddetto "**zero-click**": l'utente basta porre una domanda al chatbot affinché quest'ultimo finisca con l'assorbire comandi nascosti all'interno di una fonte apparentemente legittima. Considerando che un numero crescente di persone utilizza oggi i modelli di intelligenza artificiale come **sostituti dei motori di ricerca** tradizionali, questo scenario apre la strada a un vettore d'attacco potenzialmente ampio e mirato, capace di colpire gruppi di utenti in base a interessi specifici — dalle preferenze di consumo alle convinzioni politiche.

È significativo notare che le vulnerabilità riscontrate non riguardano soltanto versioni più obsolete dello strumento, con la ricerca di Tenable che ha evidenziato come alcune falle siano ancora attive anche nell'ultimo modello di OpenAI, **GPT-5**. L'intervento dell'azienda di sicurezza ha effettivamente portato alla correzione di parte dei problemi, tuttavia alcune vulnerabilità restano tuttora sfruttabili da eventuali malintenzionati. La combinazione di

ricerca web, memoria conversazionale e capacità di navigazione amplifica la portata dei *prompt injection*, esponendo le **debolezze strutturali** dei grandi modelli linguistici. Un problema che, prevedibilmente, non è affatto limitato a OpenAI. Lo scorso ottobre è emerso che anche il **Gemini di Google** potrebbe essere suscettibile a vulnerabilità simili: la profonda integrazione con servizi come Gmail e Google Calendar consentirebbe di nascondere istruzioni malevole direttamente nelle email e negli appuntamenti segnati in agenda.

“HackedGPT mette in luce una debolezza fondamentale nel modo in cui i modelli linguistici di grandi dimensioni giudicano **di quali informazioni possono fidarsi**“, ha dichiarato **Moshe Bernstein**, Senior Research Engineer di Tenable. “Individualmente, questi difetti sembrano piccoli, ma insieme formano una catena di attacco completa, dall’iniezione e dall’evasione al furto di dati e alla persistenza. Il report evidenzia che i sistemi di intelligenza artificiale non sono solo potenziali bersagli; possono essere trasformati in **strumenti di attacco** che raccolgono silenziosamente informazioni dalle chat o dalla navigazione quotidiana”.

Walter Ferri

Giornalista milanese, per *L’Indipendente* si occupa della stesura di articoli di analisi nel campo della tecnologia, dei diritti informatici, della privacy e dei nuovi media, indagando le implicazioni sociali ed etiche delle nuove tecnologie. È coautore e curatore del libro *Sopravvivere nell'era dell'Intelligenza Artificiale*.

Sette vulnerabilità hanno messo a rischio gli utenti di ChatGPT