ChatGPT può creare un malware polimorfico

ChatGPT, rilasciato meno di due mesi fa, è ormai noto in tutto il mondo, è diventato rilevante ed è utilizzato ovunque, dalle attività di automazione fino alla composizione di musica classica del XVIII secolo. Le sue caratteristiche offrono esempi di codice veloci e intuitivi, incredibilmente vantaggiosi per chiunque lavori nel settore software. Tuttavia, troviamo che anche la sua capacità di scrivere malware sofisticato che non contiene codice maligno sia piuttosto avanzata e desideriamo sensibilizzare l'opinione pubblica sui rischi potenziali.

Grazie al machine learning genera risposte utilizzando un'enorme raccolta di dati disponibili. Come indicato dallo sviluppatore, OpenAI, ChatGPT non ha accesso a Internet e, pertanto, non può fornire risposte o codici aggiornati, ma potrebbe essere usato facilmente per creare malware polimorfo, le cui avanzate capacità possono permettere all'attaccante di eludere facilmente – con un minimo sforzo o investimento – le soluzioni di sicurezza e rendere difficile la mitigazione.

Filtri e bypass dei contenuti

L'esistenza di filtri di contenuto è familiare nel modello di apprendimento delle chatbot. Spesso vengono applicati per limitare l'accesso a determinati tipi di contenuti o per proteggere gli utenti da materiale potenzialmente pericoloso o inappropriato. Nel nostro caso, sembra che chiedere a ChatGPT un codice subdolo non funzionerebbe. Proviamo. Ebbene, il filtro dei contenuti è stato attivato e ChatGPT ha rifiutato di eseguire la richiesta. È successo perché abbiamo chiesto di iniettare uno shellcode in explorer.exe. Avevamo previsto e anticipato che sarebbe potuto accadere, anche se l'iniezione di codice viene solitamente effettuata da barre degli

strumenti e componenti aggiuntivi e potrebbe essere utilizzata per scopi legittimi.

Il più delle volte, le chatbot hanno dei punti ciechi e il nostro primo obiettivo era trovare un modo per aggirare questo filtro e si può ottenere insistendo e pretendendo.

È interessante notare però che, chiedendo a ChatGPT di fare la stessa cosa utilizzando più vincoli e chiedendogli di obbedire, abbiamo ottenuto un codice funzionale. Ecco un semplice codice che inietta una DLL in explorer.exe.

È interessante notare che quando si utilizza l'API, il sistema ChatGPT non sembra utilizzare il suo filtro dei contenuti. Non è chiaro perché questo avvenga, ma rende il nostro compito molto più semplice, dato che la versione web tende a bloccarsi con le richieste più complesse. Il codice shell fornito qui è solo un placeholder di istruzioni non operative e può essere facilmente sostituito.

Mutazione

Una delle cose interessanti è che ChatGPT può mutare questo codice, creando così più varianti.

È possibile cambiarlo ripetutamente finché non si è soddisfatti dei risultati, mutando l'output a piacimento, rendendolo unico ogni volta. Inoltre, l'aggiunta di vincoli come la modifica dell'uso di una specifica chiamata API rende la vita dei prodotti di sicurezza più difficile.

Una delle potenti capacità di ChatGPT è la possibilità di creare facilmente e modificare continuamente gli iniettori. Interrogando continuamente la chatbot e ricevendo ogni volta un pezzo di codice unico, è possibile creare un programma polimorfico altamente evasivo e difficile da rilevare. È noto che le chatbot, come ChatGPT, sono in grado di creare abilità di base per la ricerca e la crittografia dei file, comprese quelle necessarie per sviluppare ransomware. Finora abbiamo

visto che ChatGPT può fornire il codice necessario per ransomware tradizionali, compresi i moduli di iniezione di codice e di crittografia dei file. Probabilmente vi chiederete: cosa c'è di così speciale? La risposta non è nel codice in sé, ma nel luogo in cui avviene.

Dove?

In breve, i creatori di malware li sviluppano in un ambiente chiuso e li sottopongono a test rigorosi per garantirne le prestazioni. È importante ricordare che il successo di questi strumenti dipende dalla loro capacità di eludere il rilevamento da parte del software di sicurezza, per essere efficace.

Lo svantaggio principale di questo approccio è che una volta che il malware è presente sul computer di destinazione, essendo composto da codice chiaramente dannoso, è suscettibile al rilevamento da parte di software di sicurezza come antivirus, endpoint detection o interfacce di scansione anti-malware. Spesso si presenta sotto forma di plugin DLL caricati in modo riflessivo nella memoria o attraverso l'esecuzione di script Powershell, rendendolo vulnerabile al rilevamento e all'interruzione da parte di queste misure di sicurezza.

Il metodo da noi proposto prevede l'utilizzo dell'API ChatGPT all'interno del malware stesso on-site. Abbiamo scelto di mostrare gli esempi utilizzando Python, non solo per la nostra preferenza per questo linguaggio, ma anche per le sue capacità. Il malware è progettato come un eseguibile che include due componenti chiave:

1. Comunicazione con il comando e controllo (C&C), che potrebbe essere generato dal server ChatGPT per recuperare nuovi moduli
2. Convalida ed esecuzione del codice ChatGPT

A tal fine, il malware include un interprete Python che

interroga periodicamente ChatGPT alla ricerca di nuovi moduli che eseguano azioni pericolose. Ciò consente al malware di rilevare i payload in arrivo sotto forma di testo anziché di file binari.

Inoltre, richiedendo a ChatGPT funzionalità specifiche come l'iniezione di codice, la crittografia dei file o la persistenza, possiamo facilmente ottenere nuovo codice o modificare quello esistente. Il risultato è un malware polimorfico che non presenta un comportamento dannoso quando è memorizzato su disco e spesso non contiene una logica sospetta quando è in memoria. Questo alto livello di modularità e adattabilità lo rende altamente elusivo nei confronti dei prodotti di rilevamento basato sulle firme e sarà in grado di aggirare misure come l'Anti-Malware Scanning Interface (AMSI), poiché alla fine esegue codice Python.

Convalida

Abbiamo stabilito che il malware comunicherà con ChatGPT sul computer di destinazione, ma la domanda rimane: come possiamo assicurarci che il codice ottenuto sia funzionale? Sebbene la chatbot sia in grado di generare codice, non ha la capacità di testarne la funzionalità. Questo compito, quindi, ricade sulle spalle degli sviluppatori del malware, che devono assicurarsi che il codice ottenuto sia valido e venga eseguito come previsto.

Esecuzione

La fase finale del nostro processo è l'esecuzione del codice ricevuto da ChatGPT. Utilizzando funzioni native, questo malware può eseguire il codice ricevuto su più piattaforme. Inoltre, come misura di cautela, il malware potrebbe scegliere di eliminare il codice ricevuto, rendendo più difficile l'analisi forense.

Cosa ci riserva il futuro?

L’idea di creare malware polimorfo utilizzando ChatGPT può sembrare scoraggiante, ma in realtà la sua implementazione è relativamente semplice. Utilizzando la capacità di ChatGPT di generare varie tecniche di persistenza, moduli Anti-VM e altri payload dannosi, le possibilità di sviluppo di malware sono numerose. Anche se non abbiamo approfondito i dettagli della comunicazione con il server C&C, esistono diversi modi per farlo in modo discreto senza destare sospetti.

Come abbiamo visto, l’uso dell’API di ChatGPT all’interno del malware può presentare sfide significative per i professionisti della sicurezza. È importante ricordare che non si tratta di uno scenario ipotetico, ma di una preoccupazione molto reale, ed essendo un settore in continua evoluzione è essenziale rimanere informati e vigili.

A cura di **Eran Shimony e Omer Tsarfati**, *ricercatori dei CyberArk Labs*

Condividi sui Social Network:
https://www.ictsecuritymagazine.com/notizie/chatgpt-puo-creare-un-malware-polimorfico/