Comprendere il controllo dei contenuti dell'IA generativa e l'espansione della copertura degli strumenti dell'IA DLP

Sommario

Introduzione

Panoramica

- In che modo DLP può aiutare a controllare il contenuto generato da ChatGPT?
- Perché controllare i contenuti generati dall'IA?
- Come applicare la scansione DLP alle risposte ChatGPT?
- In che cosa consiste la categoria di applicazioni API generative nel DLP?
- È possibile applicare una regola di prevenzione della perdita dei dati all'intera categoria di applicazioni API generative?
- Dove è possibile trovare la documentazione correlata?
- Prevediamo di fare qualche annuncio nel prossimo Cisco Live Amsterdam riguardo questi interessanti casi di utilizzo della protezione Generative AI?

Introduzione

Questo documento descrive il nuovo controllo del contenuto AI generativo e l'espansione della copertura degli strumenti AI DLP per Umbrella.

Panoramica

Siamo lieti di annunciare la disponibilità generale di Generative AI Content Control.
Questa funzionalità consente di monitorare e, se necessario, bloccare il contenuto generato da ChatGPT.

Siamo inoltre entusiasti di poter condividere che abbiamo ampliato l'ambito della nostra copertura DLP in tempo reale per gli strumenti di intelligenza artificiale generativa. Inizialmente limitato a ChatGPT, ora supportiamo tutti i 70 strumenti di AI nella nostra categoria di applicazione Generative AI rilasciata di recente. Questa espansione significativa consente di ampliare il caso di utilizzo sicuro dell'IA, offrendo una soluzione più completa e solida per la protezione generativa dell'utilizzo dell'IA.

In che modo DLP può aiutare a controllare il contenuto generato da ChatGPT?

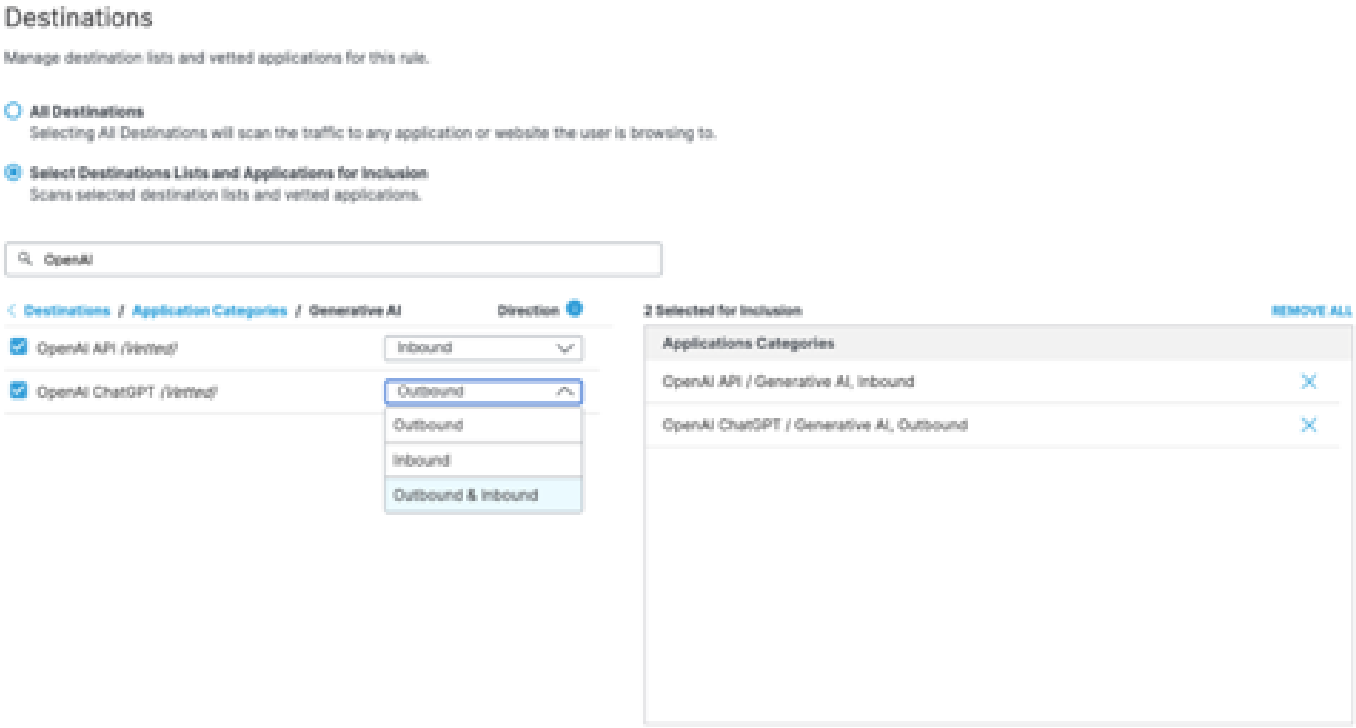
Il DLP può aiutare le organizzazioni a controllare i contenuti generati attraverso la scansione delle risposte ChatGPT utilizzando i criteri DLP in tempo reale. In questa versione è possibile scegliere di analizzare le risposte ChatGPT (ovvero, il traffico in entrata) per qualsiasi tipo di contenuto generato che si desidera monitorare o bloccare.

Perché controllare i contenuti generati dall'IA?

L'utilizzo di contenuti generati dall'IA comporta rischi per le organizzazioni per vari motivi, tra cui la violazione del copyright, informazioni imprecise, codice difettoso e così via.
Ad esempio, è possibile impedire agli utenti di utilizzare codice sorgente generato dall'IA per impedire l'utilizzo di codice protetto da copyright o non sicuro, mentre altri potrebbero voler impedire l'utilizzo di citazioni di tribunale generate dall'IA per paura di archiviare informazioni imprecise.

Come applicare la scansione DLP alle risposte ChatGPT?

In genere, il DLP in tempo reale analizza il traffico Web in uscita, come i prompt di ChatGPT, per evitare la perdita di dati sensibili. Con questa versione, introduciamo la possibilità di analizzare anche il traffico in entrata scegliendo la direzione del traffico che viene analizzato da Real-Time DLP, ovvero traffico in entrata, traffico in uscita o entrambi. Questa funzionalità è attualmente disponibile solo per ChatGPT (sia chatbot che API). Se si sceglie di analizzare il traffico in entrata, vengono analizzate le risposte di ChatGPT.



23281122679316

In che cosa consiste la categoria di applicazioni API generative nel DLP?

Prima di questa release, i criteri di destinazione nelle regole di Real-Time DLP includevano un elenco finito selezionabile di circa 20 applicazioni. Con questa release, Real-Time DLP consente ai clienti di scegliere una qualsiasi delle 38 categorie di applicazioni, inclusa Generative AI, o una qualsiasi delle ±4.600 applicazioni controllabili disponibili classificate al loro interno. La categoria Generative AI, che è stata lanciata solo pochi mesi fa con 20 applicazioni, ora ha 70 applicazioni, e siamo impegnati ad aggiornare continuamente questa categoria con strumenti di intelligenza

artificiale top-of-mental.

È possibile applicare una regola di prevenzione della perdita dei dati all'intera categoria di applicazioni API generative?

Sì, è possibile applicare una regola di prevenzione della perdita dei dati in tempo reale a un'intera categoria o a un sottoinsieme di applicazioni al suo interno.

Dove è possibile trovare la documentazione correlata?

- Per informazioni su come controllare la direzione di scansione per monitorare o bloccare le risposte ChatGPT, controllare:
  Aggiungere una regola in tempo reale ai criteri di prevenzione della perdita di dati
- Per informazioni su come verificare se è stata bloccata una richiesta chatGPT o una risposta chatGPT, selezionare la casella di controllo direzione di scansione nella casella seguente:
  Rapporto sulla prevenzione della perdita dei dati
- Per esaminare tutte le categorie di applicazioni ora disponibili nelle regole dei criteri di prevenzione della perdita dei dati in tempo reale, fare clic qui: Categorie di applicazioni

Prevediamo di fare qualche annuncio nel prossimo Cisco Live Amsterdam riguardo questi interessanti casi di utilizzo della protezione Generative AI?

Sì, abbiamo intenzione di tenere una sessione breakout intitolato Protecting Your Sensitive Data from Generative AI Usage in Cisco Live Amsterdam, martedì, 6 febbraio, 3:00 PM - 4:30 PM CET.

Per favore, si risparmi il posto!

Informazioni su questa traduzione

Cisco ha tradotto questo documento utilizzando una combinazione di tecnologie automatiche e umane per offrire ai nostri utenti in tutto il mondo contenuti di supporto nella propria lingua. Si noti che anche la migliore traduzione automatica non sarà mai accurata come quella fornita da un traduttore professionista. Cisco Systems, Inc. non si assume alcuna responsabilità per l’accuratezza di queste traduzioni e consiglia di consultare sempre il documento originale in inglese (disponibile al link fornito).