Understand New Generative AI Content Control and Expansion of DLP AI-tools Coverage

Contents

Introduction

Overview

- How can DLP help control ChatGPT generated content?
- Why to control AI-generated content?
- How can I apply DLP scanning to ChatGPT responses?
- What is the Generative AI application category in DLP?
- Can a DLP rule be applied to the entire Generative AI application category?
- Where can I find related documentation?
- Do we plan to make any announcement in the upcoming Cisco Live Amsterdam regarding these exciting Generative AI protection use cases?

Introduction

This document describes the new generative AI Content Control and the expansion of the DLP AI-tools coverage for Umbrella.

Overview

We are excited to announce the general availability of Generative AI Content Control.
This feature empowers you to monitor and, if necessary, block content generated by ChatGPT.

We are also thrilled to share that we have expanded the scope of our Real-Time DLP coverage for Generative AI tools. Initially limited to ChatGPT, we now support all 70 AI tools in our recently released Generative AI application category. This significant expansion empowers you to broaden the AI Safe Usage use case, offering a more comprehensive and robust solution for Generative AI usage protection.

How can DLP help control ChatGPT generated content?

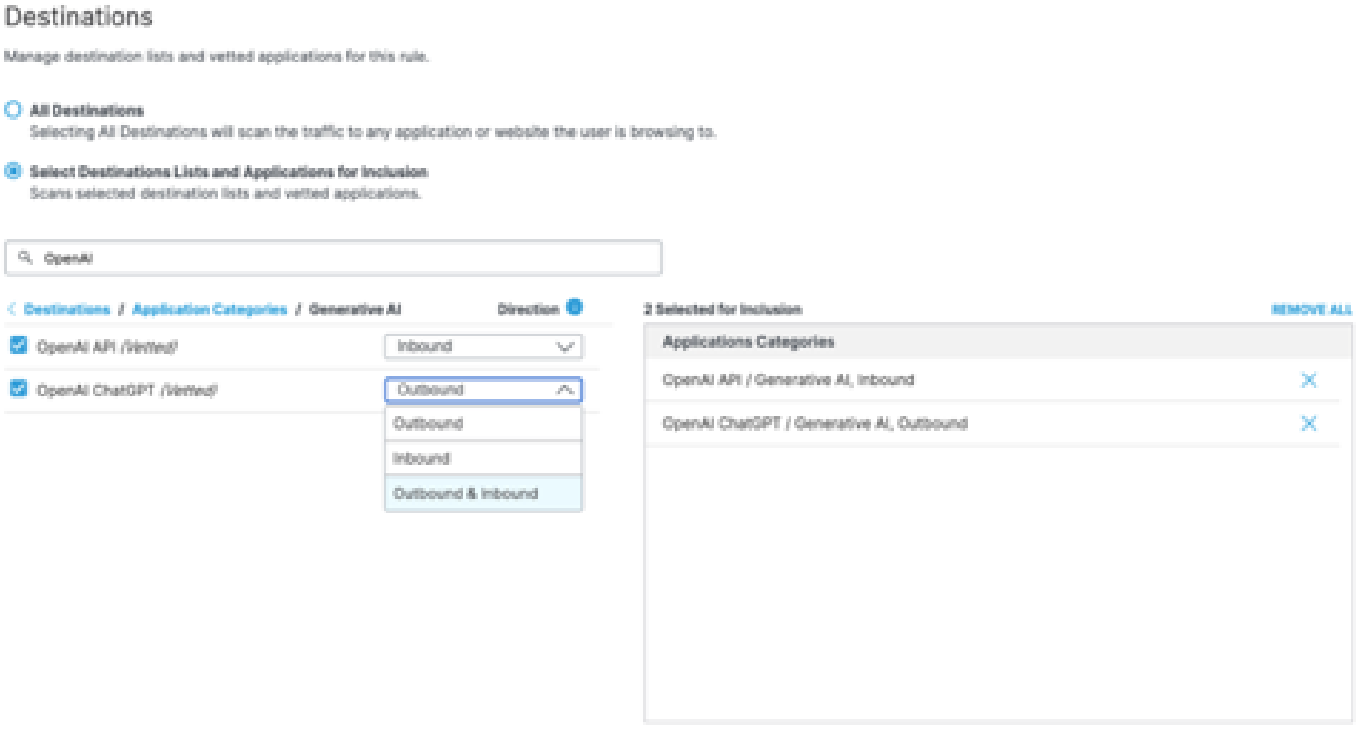
DLP can assist organizations in controlling generated content by scanning ChatGPT responses using Real-Time DLP policy. With this release, you can choose to scan ChatGPT responses (that is, inbound traffic) for any type of generated content you wish to monitor or block.

Why to control AI-generated content?

Using AI-generated content poses risks to organizations for various reasons, including copyright infringement, inaccurate information, faulty code, and so on.
For example, you might wish to stop users from using AI-generated source code to prevent the usage of copyrighted or unsafe code, while others might wish to prevent the usage of AI-generated court citations for fear of filing inaccurate information.

How can I apply DLP scanning to ChatGPT responses?

Generally, Real-Time DLP scans outbound web traffic, like ChatGPT prompts, to prevent sensitive data leakage. With this release, we are introducing the ability to also scan inbound traffic by choosing the direction of traffic that Real-Time DLP scans, that is, inbound traffic, outbound traffic, or both. This ability is currently only available for ChatGPT (both chatbot and API). Choosing to scan inbound traffic scans ChatGPT responses.



23281122679316

What is the Generative AI application category in DLP?

Before this release, the destination criteria in the Real-Time DLP rules included a finite selectable list of about 20 applications. With this release, Real-Time DLP allows customers to choose any one of our 38 application categories, including Generative AI, or any of the ±4,600 available controllable apps categorized within them. The Generative AI application category, which was launched only a few months ago with 20 apps, now has 70 apps, and we are committed to continuously updating this category with top-of-mind AI tools.

Can a DLP rule be applied to the entire Generative AI application category?

Yes, a Real-Time DLP rule can be applied to an entire category or to a subset of applications within it.

Where can I find related documentation?

- To learn how to control scanning direction to monitor or block ChatGPT responses check: Add a Real Time Rule to the Data Loss Prevention Policy
- To learn how to check whether it was a chatGPT prompt or chatGPT response that was blocked check scanning direction reporting here: Data Loss Prevention Report
- To review all application categories that are now available in Real-Time DLP policy rules check here: Application Categories

Do we plan to make any announcement in the upcoming Cisco Live Amsterdam regarding these exciting Generative AI protection use cases?

Yes, we are going to hold a breakout session titled Protecting Your Sensitive Data from Generative AI Usage in Cisco Live Amsterdam, on Tuesday, Feb 6, 3:00 PM - 4:30 PM CET.

Please save your seat!