Netskope One AI Guardrails

資料表

防範 AI 威脅和濫用

AI 帶來傳統的安全性工具無法察覺的全新威脅向量。攻擊者現在利用操縱性提示繞過安全控制措施，而使用者可能在無意中產生有害、歧視性或受版權保護的內容，對企業造成重大的法律和聲譽風險。

為何 Netskope 是最佳選擇？

Netskope One AI Guardrails 專為現代 AI 企業而設計，為 AI 環境提供專用的執行階段防禦層。它對所有流量進行深入的即時分析以緩解複雜的攻擊，包括提示注入和越獄嘗試。它也可作為人類和代理式互動的自動化審核器，確保持續的原則合規性和資料完整性。

執行階段威脅防護和內容審核

- **阻止威脅並確保模型完整性**
  封鎖試圖規避系統規則或竊取資料的對抗性攻擊。檢查每個要求和回應，以找出複雜的語言漏洞利用程式背後的多階段意圖。
- **落實負責任 AI 使用和品牌安全**
  自動過濾有害或歧視性內容，包括仇恨言論、暴力、武器和犯罪。這可確保 AI 使用保持在組織的風險容許範圍內，並維護企業聲譽。
- **降低生成內容的法律和 IP 風險**
  識別並禁止傳遞 AI 回應中受專利或版權保護的資料。這可主動預防與生成式模型輸出相關的新興法律責任。
- **將偵測與 DLP 和進階威脅防護相關聯**
  AI Guardrails 與 Netskope One DLP 和威脅防護無縫整合。
  透過 Netskope One 平台的 AI (SkopeAI) 將相關的原則違反偵測統一呈現於單一總覽中，提供更豐富的脈絡並加快調查速度。

主要效益和能力

提高 AI 投資報酬

建立並實施明確的安全邊界，讓您針對高風險、高價值的商業使用案例部署 AI。

提升 SecOps 與合規性效率

將偵測對映至 MITRE ATLAS 和 OWASP LLM Top 10。這種統一檢視可縮短調查時間，並讓團隊與最新 TTP 同步。

可追溯，為稽核做好準備

保留與原則觸發條件相符的可搜尋對話紀錄，並採用以角色為基礎的存取控制，確保只有獲授權的調查人員才可檢視提示紀錄。

深度使用者意圖分析

利用行為訊號區分合法使用與惡意活動，在資料外洩實際發生前加以預防。

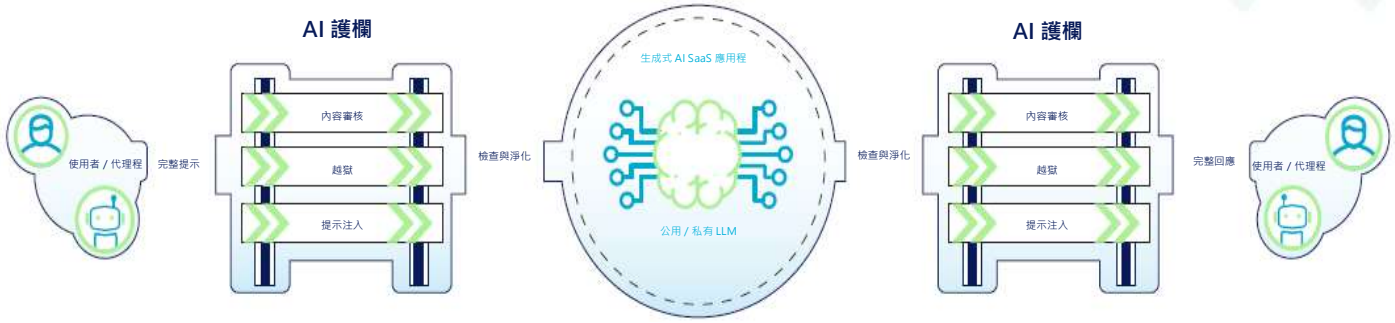
不間斷的使用者體驗

低延遲護欄能以現代企業 AI 的規模運作，助您維持創新速度。

「由於影子 AI 和工具暴增，42% 的 GenAI 原則違反涉及原始碼，其次是受管制資料 (32%) 和智慧財產 (16%)。」

–Netskope Threat Labs，雲端和威脅報告：2026 年雲端和威脅報告

Netskope One Agentic Broker



netskope

Netskope 的獨特之處

Netskope One AI Guardrails 重新定義資料治理，從零散的工具轉為主動式整合防禦系統。此解決方案在關鍵的投入時刻套用智慧控制，以獨特方式將內容審核的專用 AI 護欄和 AI 特有威脅防護與 Netskope 領先業界的 DLP 和威脅防護結合。此整合提供重要的風險脈絡，建立所有 LLM 威脅的單一事件視圖，以降低警報疲勞並加快調查速度。

這種統一方法可讓平台檢視脈絡中的提示和回應以理解使用者和代理程式的真實意圖，區分正常工作與惡意操縱。資安團隊受益於原則觸發條件的可搜尋對話紀錄，直接對映至 MITRE ATLAS、OWASP LLM Top 10 等框架，確保他們始終領先於不斷演變的對抗戰術。無論是管理公用生成式 AI SaaS、企業計畫、Amazon Bedrock 上的私有部署或自主代理式工作流程，Netskope 都能提供一致的內容審核與威脅防護。將這些能力整合至 Netskope One 平台，組織最終可從實驗階段邁入下一階段，安全地充分利用 AI。

效益	說明
阻止惡意 AI 攻擊	即時檢查提示和回應以阻止複雜技術，例如提示注入和惡意越獄嘗試。
落實負責任 AI 使用	內容審核引擎會自動偵測不當內容類別，例如暴力、歧視、武器、色情內容、盜版和受版權保護的資料。
為完整提示和回應提供多語支援	AI Guardrails 為完整提示和回應提供多語支援，包括中文、英文、法文、西班牙文、葡萄牙文、德文、義大利文、俄文、日文、韓文、越南文、泰文、阿拉伯文等。
防止敏感資料外洩	AI Guardrails 與 Netskope One DLP 整合，可識別並阻止 PII、原始碼或專有機密進入 AI 模型。
統一檢視事件	將 AI 原則違反（包括內容審核、DLP 和威脅）對映至 MITRE ATLAS 和 OWASP Top 10 框架，統一檢視事件。
保護智慧財產權	Netskope 可偵測並阻止從 AI 產生的回應中擷取或分享受版權和專利保護的資料。
稽核與治理	保留可搜尋紀錄並採用以角色為基礎的存取控制，確保只有獲授權的調查人員才可檢視敏感的聊天紀錄。