Netskope One AI 紅隊演練

資料表

為可上線 AI 提供自動化測試

從 SaaS 轉移到私有大型語言模型 (LLM) 會產生嚴重的安全性缺口。Netskope One AI 紅隊演練可填補此缺口，將對抗性模擬自動化以找出漏洞，確保私有模型具備安全性、合規性和韌性。

為何 Netskope 是最佳選擇？

透過 Netskope One AI 紅隊演練主動預防並消除私有 AI 部署的漏洞。我們將 LLM 暴露於數千個模擬提示中，以研究漂移如何產生漏洞。整合 Netskope One AI Guardrails 可支援原則建立和提示強化，保護從建置到執行階段的開發生命週期。

在生命週期中保護 AI 開發

- **自動化對抗性測試**
  利用超過 18,000 種對抗情境，以系統化方式對模型進行壓力測試。此自動化方法可跟上快速開發的步調，取代緩慢的手動測試。
- **在 AI 開發生命週期中持續進行安全性整合**
  使用 API 整合至 CI/CD 管道中，在每次正式發佈前自動過濾因程式碼變更而產生的漏洞或風險。
- **模擬複雜的多輪攻擊**
  確定複雜的萬能鑰匙攻擊和漸強式攻擊（誘導 LLM 繞過安全護欄）可能在上線前和上線後影響模型的位置。
- **追蹤不斷變化的風險評估**
  將模型測試從被動觀察轉變為主動防禦。定期執行紅隊演練模擬，呈現在對同一模型進行的所有測試中發現的風險變化。

主要效益和能力

彌補 AI 安全性缺口

透過自動化模擬確保私有模型具備韌性和安全性，放心從實驗階段邁向可上線 AI。

維持強健的對抗性防禦

維持一致的高強度對抗性防禦層，確保模型更新絕不會產生新的安全漏洞或增加風險。

確保嚴格的隱私合規性

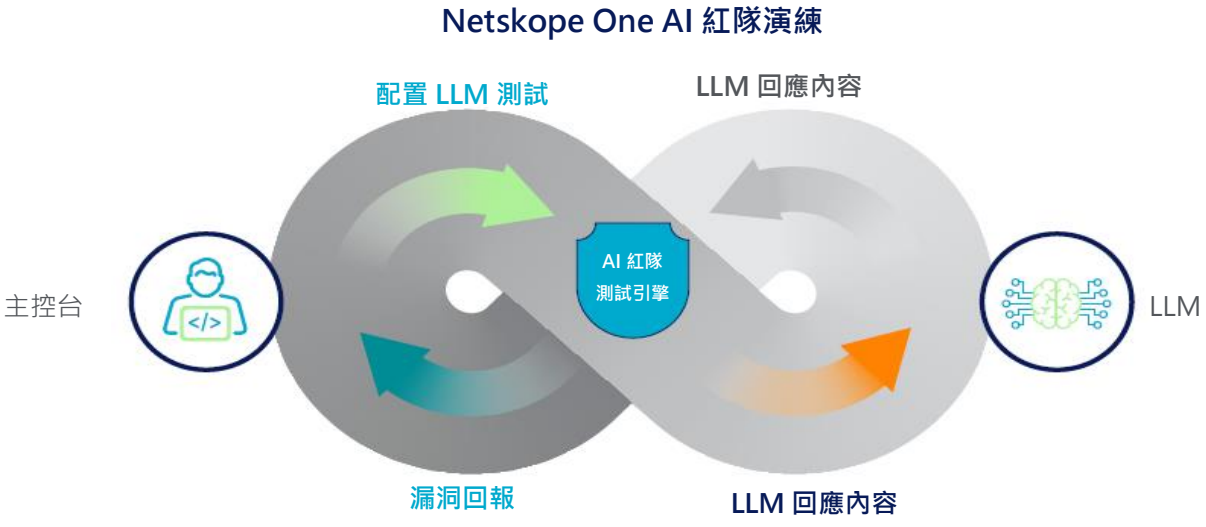
找出可能意外揭露內部系統提示、訓練紀錄或敏感智慧財產的模型漏洞，保護您的品牌。

加快安全的 AI 創新

以自動化測試取代手動安全性審核以加快開發週期，讓團隊更快速地部署 AI 功能而不影響安全。

防範不斷演變的威脅

透過持續性探測識別提示注入、越獄和以進階邏輯為基礎的對話式威脅，預防複雜攻擊。



netskope

Netskope 的獨特之處

Netskope One 為所有使用者和代理程式流量提供統一檢查點，提供現代 AI 安全性所需的即時可見性、威脅和資料保護。傳統工具難以應對新型 AI 計畫的規模，Netskope One AI 紅隊演練則可在上線前和上線後自動找出漏洞。我們提供超過 18,000 種對抗情境和測試案例，針對提示注入、越獄、資料外洩和惡意使用情境對模型進行壓力測試。除了簡單的過濾器之外，我們也模擬複雜的多輪攻擊技術，包括以繞過標準安全護欄為目標的萬能鑰匙攻擊和漸強式攻擊。此方法確保您的 AI 旅程設計具備安全性和可擴充性，完全整合於整體資料安全策略中。透過 Netskope，您可以放心從實驗進入上線階段，確保私有模型能在不斷演變的 AI 局勢中抵禦最複雜的對抗性威脅。

效益	說明
LLM 安全性測試	我們使用超過 18,000 種對抗性情境測試案例和種子提示，讓您能夠在將 LLM 部署至正式環境之前和之後，以系統化方式對其進行壓力測試以找出漏洞。
自動化 CI/CD 安全性整合	透過 API 將對抗性壓力測試直接整合至 CI/CD 管道中。在所有程式碼變更或模型更新進入正式環境之前自動過濾，找出新的安全風險和漏洞。
多輪攻擊模擬	模擬複雜的多輪攻擊，攻擊者會試圖誘導 LLM 或將多階段對話分層，以繞過缺乏完整對話脈絡的護欄。
資料外洩預防	偵測可能的資料外洩，例如模型漏洞可能揭露內部系統提示或喚回敏感資料（包括訓練紀錄和內部知識），確保符合嚴格的隱私標準。