CLOUDFLARE

電子書

AI 安全指南

CISO 指南：如何制定可擴展的 AI 策略

目錄

報告摘要

CISO，歡迎您！

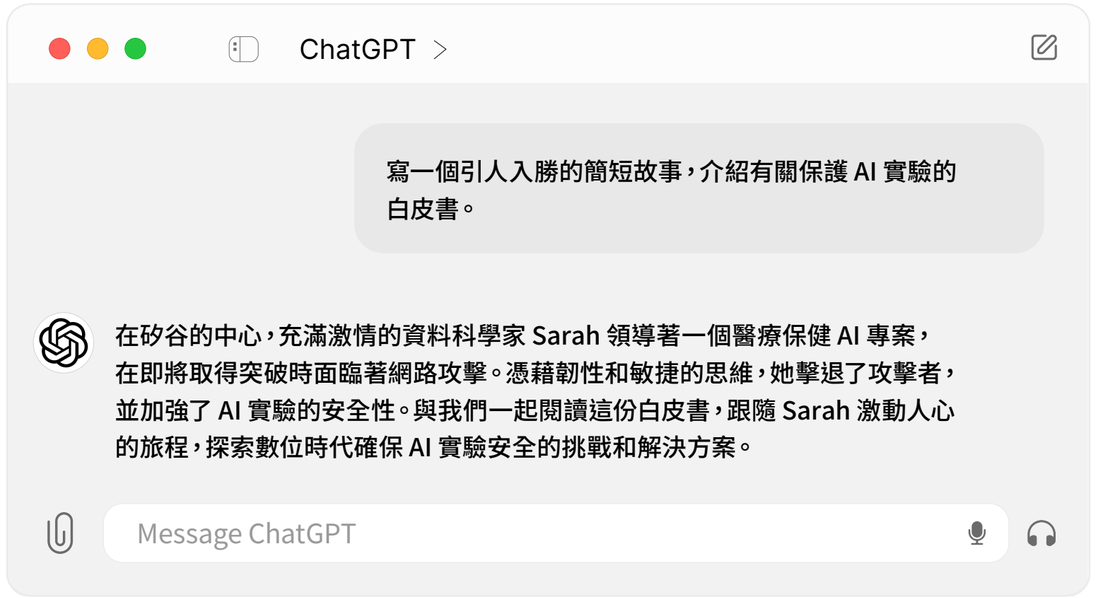
AI 可能是當今最熱門的詞彙，也是網路安全社群面臨的最緊迫問題之一。我們必須關注它所帶來的影響，因此，Cloudflare 撰寫了本指南，旨在幫助您思考如何在組織中安全地進行生成式人工智慧 (GenAI) 實驗。

AI 工具迅速發展，不僅功能越來越強大，也更易於獲取，為各個產業的創新提供了諸多機會。然而，與其他典範轉移一樣，GenAI 也面臨著獨特的安全性、隱私權與合規性挑戰。GenAI 的廣泛採用可能會引發不可預見的用量激增、使用者濫用、惡意行為以及危險的影子 IT 做法，所有這些情況都會增加資料外洩和敏感性資訊洩露的風險。

隨著工作場所中的 AI 採用範圍不斷擴大，您需要準備一份 GenAI 藍圖，提供有關如何大規模使用、構建和保護 AI 的指南。讓我們來討論一下風險，並研究相關技巧，以便您的團隊根據成熟度和使用情況來保護 GenAI。您的組織可以根據本指南，制定一個滿足業務需求的 GenAI 策略，同時保護資料並確保合規性。

——Cloudflare 產品行銷總監 Dawn Parzych

確保 GenAI 實驗安全



很抱歉告訴您，Sarah 的故事到此為止。不過，雖然我們告別了這個虛構角色，但隨著預測式和生成式 AI 的擴展，現實生活中將有無數的「Sarah」——每個人都是 IT 和開發人員團隊的英雄、業務技術人員以及員工個人。

AI 吸引了技術人員和日常使用者，激發了好奇心和探索精神。在我們努力釋放 AI 的全部潛力的過程中，這種實驗是必要的。但是，如果不加倍小心和防範，也可能導致安全性受損或不合規。

為了實現這種平衡，並更有效地理解和管理 AI 計畫，組織必須考慮三個關鍵領域：

1 使用 AI

使用第三方廠商提供的 AI 技術（例如 ChatGPT、Bard 和 GitHub Copilot），同時保護資產（例如敏感性資料、智慧財產權、原始碼等）並根據使用案例降低潛在風險

2 構建 AI

開發適合組織特定需求的自訂 AI 解決方案（例如用於預測分析的專有演算法、面向客戶的助理或聊天機器人以及 AI 驅動的威脅偵測系統）

3 保護 AI

保護 AI 應用程式和 AI 系統免受不良行為者的操縱，使其出現不可預測的行為

確保 GenAI 實驗安全

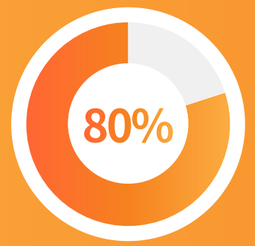
GenAI 轉型：現在和未來

GenAI 對消費者和組織的吸引力使其走上了前所未有的採用軌道。一小部分超級使用者快速成長，部分原因是活躍的開放原始碼社群以及消費者驅動的應用程式實驗，如 ChatGPT 和 Stable Diffusion。

在此過程中，使用者發現，機器人實際上不會「取代我們」。

GenAI 讓人類處於完善和增強的位置，而不是從頭開始創造一切，並且可以幫助企業提高員工效率。預測式 AI 提供了類似的好處，可以更輕鬆地利用資料來改善決策、建立更智慧的產品和個人化客戶體驗等一系列措施。

如今，**59% 的開發人員**正在其開發工作流程中使用 AI[1]



到 2026 年，超過 **80% 的企業**將使用在生產環境中部署的支援 GenAI 的 API、模型和/或應用程式（目前這一比例為 5%）[2]

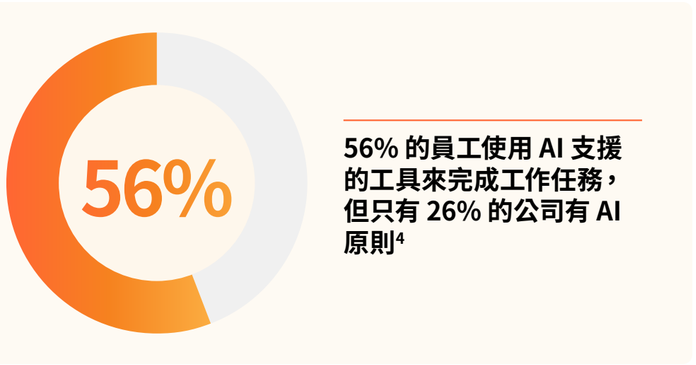
到 2030 年，GenAI 將增加 **50% 的知識工作者**任務，以提高生產力或提高平均工作品質（目前不到 1%）[3]

1. SlashData，「How developers interact with AI technologies」（開發人員如何與 AI 技術互動），2024 年 5 月
2. Gartner，「A CTO' s Guide to the Generative AI Technology Landscape」（生成式 AI 技術情勢 CTO 指南），2023 年 9 月
3. Gartner，「Emerging Tech: The Key Technology Approaches That Define Generative AI」（新興技術：定義生成式 AI 的關鍵技術方法），2023 年 9 月

安全使用 GenAI

AI 實驗涵蓋了從使用預先建置的 AI 工具和服務到從頭開始構建自訂 AI 解決方案的各個領域。雖然一些組織可能會進一步建立自己的 AI 模型和應用程式，但許多組織會堅持使用第三方 AI 工具。

在這些情況下，第三方 AI 工具會帶來新的風險，因為組織對其安全和隱私設定的直接控制能力有限。



56% 的員工使用 AI 支援的工具來完成工作任務，但只有 26% 的公司有 AI 原則[4]

員工現在可能透過 Microsoft 365 之類的 SaaS 套件、搜尋引擎或公用應用程式內建的聊天機器人，甚至 API，使用現成的 AI 工具來工作。

組織必須進行盡職調查以最大程度降低風險，包括：

- **評估**第三方工具的安全風險
- **解決**資料隱私問題
- **管理**對外部 API 的依賴（或過度依賴）
- **監控**潛在漏洞

例如，員工使用 ChatGPT 等公用 Web 應用程式時。輸入到提示詞的所有內容都會成為脫離組織控制的資料。使用者可能會過度分享敏感、機密或受監管的訊息，例如個人識別資訊 (PII)、財務資料、智慧財產權和原始碼。即使他們不分享明確的敏感性訊息，也可以透過將輸入中的上下文拼湊起來以推斷敏感性資料。

為了安全，員工可以切換設定以防止他們的輸入進一步訓練模型，但他們必須手動執行此操作。為了確保安全，組織需要採取措施阻止人們輸入私人資料。

為 AI 的安全隱患做好準備

資料暴露

使用者在多大程度上不恰當地與外部 AI 服務分享敏感性資料？匿名化/擬匿名化技術是否足夠？

API 風險

您將如何解決第三方 API 中可能成為攻擊者潛在入口的漏洞？

黑箱系統

外部 AI 模型的哪些決策過程可能會帶來意外風險？

廠商風險管理

您對第三方 AI 提供者的網路安全做法瞭解多少？以及更重要的，有哪些是您不知道的？

4. 世界大型企業聯合會，2023 年 9 月

保護 AI 使用的措施

1 管理治理和風險

- 制定有關如何以及何時使用 AI 的原則，包括組織允許使用者與 GenAI 分享哪些資訊、存取控制指南、合規性要求以及如何報告違規行為
- 執行影響評估以收集資訊，識別並量化使用 AI 的好處和風險

2 增加對安全性和隱私權的可見度和控制

- 記錄所有連線（包括與 AI 應用程式的連線），以持續監控使用者活動、AI 工具使用和資料存取模式，從而偵測異常
- 探索存在哪些影子 IT（包括 AI 工具），並就核准、封鎖或疊加其他控制措施做出決策
- 掃描 SaaS 應用程式設定是否存在潛在的安全風險（例如，從已核准的應用程式向未經授權的 AI 應用程式授予 OAuth 權限，導致產生資料外洩的風險）

3 檢查哪些資料進出 AI 工具，並篩選掉任何可能破壞IP、影響機密性或違反版權限制的內容

- 對使用者與 AI 工具互動的方式套用安全控制（例如停止上傳、阻止複製/貼上以及掃描和封鎖敏感/專有資料的輸入）
- 採取防護措施以阻止 AI 機器人剽竊您的網站
- 僅在無法進行其他控制的情況下才徹底封鎖 AI 工具。眾所周知，使用者會找到一些變通方法，使安全性脫離您的控制

4 控制對 AI 應用程式和基礎架構的存取

- 確保存取 AI 工具的每個使用者和裝置都經過嚴格的身分驗證，以限制誰可以使用 AI 工具
- 實作基於身分的 Zero Trust 存取控制。套用最低權限，以限制來自遭入侵帳戶或內部人員威脅的潛在損害

5 簡化成本並提高營運效率

- 透過分析和記錄瞭解人們使用 AI 應用程式的情況，以便您可以根據使用規模控制限速、快取以及請求重試和模型回退

保護您構建的內容

訓練 AI 模型

AI 管道正在擴大漏洞範圍。但是，憑藉在開始和整個開發過程中獲得的經驗，我們對取得成功的因素有了深入的瞭解。對於 AI 安全性，自然應當從您的模型著手。

作為 AI 應用程式的基礎，用於訓練 AI 模型的所有內容都將流向其輸出。在一開始就考慮如何保護這些資料，以避免日後產生負面影響。如果不採取保護措施，您可能會面臨攻擊面擴大並在未來產生應用程式問題的風險。

確保資料完整性的安全性對於緩解蓄意和意外資料外洩至關重要。AI 管道中的安全性風險可能包括：

- **資料中毒：**惡意資料集會影響結果並造成偏見
- **幻覺濫用：**威脅執行者將 AI 幻覺合法化（為了產生回應而編造資訊），導致惡意和非法資料集影響輸出

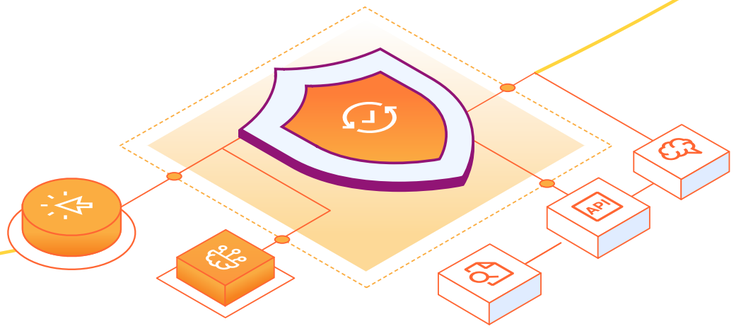
或者，如果您不訓練模型，您的內部 AI 一開始會選擇一個模型來執行工作。在這些情況下，您可能需要探索建立者是如何建立和保護模型的，因為它在推斷中發揮作用。

推斷是 AI 訓練之後的過程。模型訓練得越好，調整得越精細，推斷就越好，但它們永遠不能保證完美。即使是訓練有素的模型也會產生幻覺。

部署後安全性

構建並部署內部 AI 後，您需要保護其私人資料並保護對其的存取。除了我們在本文中提出的建議（包括對每個使用者強制執行權杖和限速）外，您還應該考慮：

- **管理配額：**使用限制來防止使用者的 API 金鑰遭到洩露和共用
- **封鎖某些自治號碼 (ASN)：**防止攻擊者向應用程式傳送大量流量
- **啟用 Waiting Room 或質詢使用者：**使請求變得更加困難或耗時，從而破壞攻擊者的經濟效益
- **構建和驗證 API 結構描述：**透過識別和編目所有 API 端點來概述預期用途，然後列出所有特定參數和類型限制
- **分析查詢的深度和複雜性：**協助防範直接的 DoS 攻擊和開發人員錯誤，保持來源健康並按預期向使用者提供請求
- **圍繞基於權杖的存取建立規則：**當權杖在中介軟體層或 API Gateway 中驗證時，防止存取遭受入侵



在 GenAI 實驗中提供強大的威脅防護

從採用到實作，GenAI 實驗範圍的每個階段都應在風險最小或可容忍的情況下進行。有了從本文中獲得的知識，無論您的組織正在使用、建置 AI，還是計劃未來以某種形式使用 AI，您都能夠控制您的數位環境。

在採用新功能時感到猶豫是很自然的，但總有一些資源可以讓您有信心安全地試驗 AI。在這些資源中，組織如今最需要的是連線所有 IT 和安全性的連接橋梁。它充當公共執行緒，透過處理環境中的所有內容來降低複雜性，隨處可用，並執行必要的安全性、網路和開發功能。

有了這種連接橋梁，您就可以對各種使用案例充滿信心，包括：

- 能夠偵測和控制受監管資料的移動，從而確保遵守法規
- 重新取得對 SaaS 應用程式、影子 IT 和新興 AI 工具中敏感性資料的可見度和控制
- 透過偵測和封鎖上傳和下載中的原始程式碼，保護開發人員程式碼。此外，還可以預防、查找和修復 SaaS 應用程式和雲端服務 (包括程式碼存放庫) 中的錯誤設定

隨著 AI 的不斷發展，不確定性是必然存在的。這就是為什麼，擁有像 Cloudflare 這樣的穩定力量將會大有助益。

防範三種類型的 LLM 中的 AI 風險

根據使用情況，AI 為組織帶來的風險暴露程度會有所不同。務必瞭解與使用和開發大型語言模型 (LLM) 相關的各種風險，然後積極參與各種 LLM 部署。

LLM 類型	主要風險
內部	存取敏感性資料和智慧財產權
產品	聲譽風險
公用	敏感性資料外洩

可擴展、易於使用且無縫整合

Cloudflare 的全球連通雲將控制權交給您，並提高了可見度和安全性，從而使 AI 實驗變得安全且可擴展。更棒的是，我們的服務能夠強化一切，確保無需在使用者體驗與安全性之間進行取捨。

鑑於大多數組織要么只使用 AI，要么會使用並構建 AI，利用 Cloudflare 意味著永遠不會在 AI 專案上停滯不前。

- 我們的**全球網路**讓您能夠在任何需要的地方快速擴展和實施控制
- 憑藉我們的**易用性**，您可以輕鬆部署和管理有關使用者如何使用 AI 的原則
- 一個**可程式設計的架構**，可讓您在正在建置的應用程式上疊加一層安全性，而不會中斷使用者對 AI 的使用

Cloudflare 的全球連通雲可保護 AI 實驗的方方面面，具體包括：

- 我們的 **Zero Trust 和安全存取服務邊緣 (SASE)** 服務有助於緩解員工**使用**第三方 AI 工具的風險
- 我們的**開發人員平台**可幫助您的組織安全高效地**構建**自己的 AI 工具和模型
- 為了**利用 AI 確保安全**，我們的平台利用 AI 和機器學習技術來建構威脅情報，然後用於在整個 AI 實驗中保護組織

	您如何使用 AI	您如何構建 AI
我們的全球網路規模	在世界各地一致地擴展和實施控制	加速推斷、查詢和快取
管理的簡便性	具有簡單部署和原則的單一控制平面	可快速入門的範本
我們統一且可程式設計的網路架構	在不中斷 AI 使用方式的情況下疊加新的安全性	內建隱私權與合規性

後續步驟

從保護您的組織使用 AI 的方式到保護您構建的 AI 應用程式，Cloudflare for AI 都能滿足您的需求。使用我們的服務，您可以按任何順序採用新功能，並擁有無限的互通性和靈活的整合。

與專家討論

如需更多資訊，請造訪 cloudflare.com/zh-tw/

+ 886 8 0185 7030 | enterprise@cloudflare.com | cloudflare.com/zh-tw

REV:BDES-6307.2024AUG19