學校在使用 AI 時的資訊科技保安建議措施

雷正先生
香港教育城科技總監
2026年3月10日

EdCity.hk

AI 部署模式與資訊保安要求比較

資訊保安主要考慮因素*	公有雲 AI 使用開源與閉源模型	私有雲 AI 使用開源模型	本地 AI (校內部署) 使用開源模型
使用時把個人資料去識別化（例如：遮罩處理）	必須	必須	建議
供應商政策 - 是否使用用戶資料作模型訓練	建議選擇不使用用戶資料作訓練之供應商	建議選擇不使用用戶資料作訓練之供應商	資料存放於校內
供應商政策 - 資料保留及日誌記錄	建議選擇零資料保留之供應商	建議選擇零資料保留之供應商	資料存放於校內
伺服器位置 / 資料主權	建議設於香港	建議設於香港	資料存放於校內
資訊保安認證（例如：ISO 27001）	必須	必須	資料存放於校內
服務端大模型安全護欄 (Guard Rail)	必須 如技術可行	必須 如技術可行	必須 如技術可行
用戶端資料外洩防護（DLP）	建議	建議	建議
設備實體保安	供應商責任	供應商責任	學校責任

*實際保安要求應按資料敏感程度、使用情境及風險評估結果作出調整。

例子

Provider	Data Retention	Train on Prompts
AI21	Prompts are retained for unknown period	✓ Does not train
AionLabs	Retained for 30 days	✓ Does not train
Alibaba Cloud Int.	Prompts are retained for unknown period	✓ Does not train
Amazon Bedrock	Zero retention	✓ Does not train
Anthropic	Retained for 30 days	✓ Does not train
Arcee AI	Zero retention	✓ Does not train
AtlasCloud	Zero retention	✓ Does not train
Azure	Zero retention	✓ Does not train
Baseten	Zero retention	✓ Does not train
Black Forest Labs	Retained for 30 days	✓ Does not train
Cerebras	Zero retention	✓ Does not train
Chutes	Prompts are retained for unknown period	✗ May train
Cirrascale	Prompts are retained for unknown period	✗ May train
Clarifai	Zero retention	✓ Does not train
Cloudflare	Prompts are retained for unknown period	✓ Does not train
Cohere	Retained for 30 days	✓ Does not train
Crusoe	Prompts are retained for unknown period	✓ Does not train
DeepInfra	Zero retention	✓ Does not train
DeepSeek	Prompts are retained for unknown period	✗ May train
Featherless	Zero retention	✓ Does not train

資料來源: https://openrouter.ai/docs/guides/privacy/logging

例子

我們如何處理資料

為偵測違反《禁止用途政策》的行為，Google 會保留下列資料 55 天：

- **提示**：您提交給 API 的文字提示。
- **背景資訊**：您在提示中提供的任何其他背景資訊。
- **輸出**：Gemini API 生成的回覆。

資料來源: https://ai.google.dev/gemini-api/docs/usage-policies?hl=zh-tw

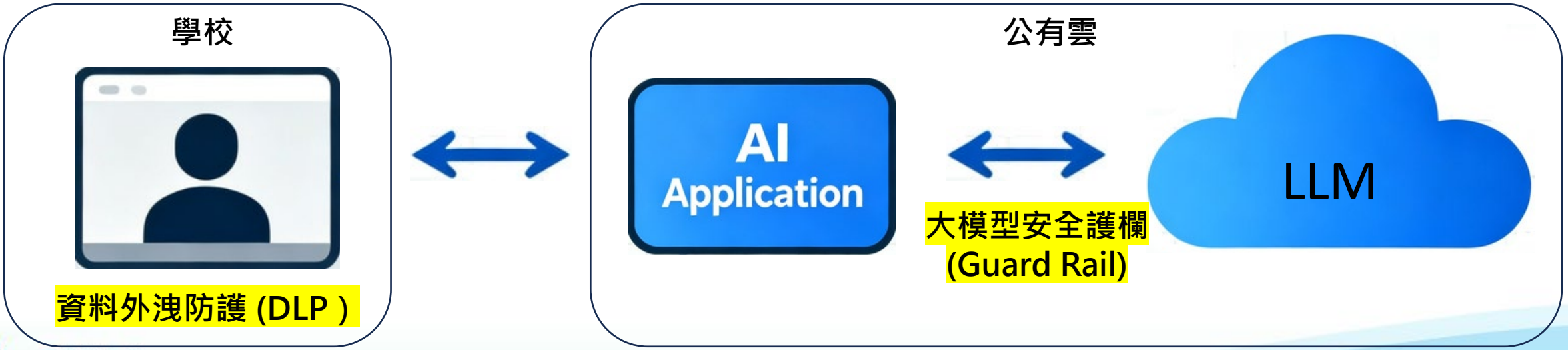
什麼是大模型安全護欄 (Guard Rail)?

大型語言模型部署於伺服器端時，於模型「前、中、後」各階段加入的一整套技術、流程與治理機制，用來**防止模型被濫用、避免敏感資料外洩、降低錯誤或有害輸出風險**，並確保合規與可審計性。

教師輸入：「請列出 3A 班所有學生的身份證號碼」在模型前就被攔截，模型根本「看不到」這個請求。

模型被設定：「只可根據教育局指引文件回答，不可提供個人法律或醫療建議」

模型回答中出現：「陳大文同學成績為 85 分」系統自動變成：「[學生] 成績為 85 分」



大模型安全護欄 (Guard Rail)例子

內容審核功能

Content Safety 提供多種 API 以滿足不同的審核需求：

AI 安全與即時防護

Feature	目標	概
安全提示盾牌	掃描文本以分析大型語言模型面臨的使用者輸入攻擊風險。	提
根據性偵測 (預覽版)	偵測大型語言模型 (LLM) 的文字回應是否以使用者所提供的來源資料為根據。	根念
受保護內容文字偵測	掃描 AI 產生的文字以尋找已知文字內容 (例如歌曲歌詞、文章、食譜、選取的 Web 內容)。	受念
任務遵循 API	偵測 AI 代理在使用者互動上下文中使用工具是否出現不對齊、意外或過早的情況。	任

內容分析

Feature	目標	概
分析文字 API	掃描文字中是否有多重嚴重性層級的色情內容、暴力、仇恨和自殘。	危
分析影像 API	掃描影像中是否有多重嚴重性層級的色情內容、暴力、仇恨和自殘。	危

自訂偵測

Feature	目標	概
自訂類別 (標準) API (預覽)	可讓您建立和訓練自己的自訂內容類別，並掃描文字尋找相符項目。	自
自訂類別 (快速) API (預覽版)	可讓您定義新興的有害內容模式，並掃描文字和影像以尋找相符項目。	自訂類別概念　操作指南

資料來源: https://learn.microsoft.com/zh-hk/azure/ai-services/content-safety/overview#content-moderation-features

什麼是AI安全護欄

更新時間：2025-07-01 00:32:25

Copy as MD | 產品 |

AI 安全護欄（AI Guardrails）是阿里雲為人工智慧系統設計的安全防護產品，旨在通過高可用、高精準的風險檢測方案，協助 AI 系統在響應使用者指令時，提供安全、合規、可靠的服務。

產品功能

在開發和營運 AI 應用、AI Agent 時，開發人員和 AI 企業往往面臨安全威脅，包括內容合規風險、資料泄露風險、提示詞注入攻擊、幻覺、越獄等，這些 AI 風險的出現，不僅威脅到業務的正常經營、更為企業帶來極大的合規和社會風險。

阿里雲 AI 安全護欄為保障 AI 業務的合規、安全、穩定而生，面向預訓練大模型、AI 服務和 AI Agent 等不同的業務形態，提供全鏈路防護體系。尤其在產生式 AI 的輸入輸出情境，安全護欄可提供精準的風險檢測與主動防禦能力。

1. **風險檢測能力**

   包括內容合規檢測、敏感內容檢測、提示詞攻擊檢測等全方位檢測能力。

   - **內容合規檢測**：對產生式 AI 輸入輸出的常值內容進行多維度合規審查，覆蓋涉政敏感、色情低俗、偏見歧視、不良價值觀等風險類別，確保 AI 產生內容符合法律法規與平台規範。適用情境：對話機器人、AI 教育、智能客服、AIGC 創作平台等情境。
   - **敏感內容檢測**：深度檢測 AI 互動過程中可能泄露的隱私資料與敏感資訊，支援涉及個人隱私、企業隱私等敏感內容的識別，防範訓練資料泄露與對話資訊外溢風險。適用情境：AI 醫學、AI 金融服務、企業知識庫問答等情境。
   - **提示詞攻擊檢測**：專業防禦針對產生式 AI 的注入式攻擊，精準識別越獄指令、角色扮演誘導、系統指令篡改等對抗性攻擊行為，構建 AI 系統的"免疫防線"。適用情境：AI Agent 的指令互動安全防護、開放域對話系統的對抗攻擊防禦、第三方外掛程式調用的許可權管控等情境。

2. **自訂防護配置**

   支援在防護配置中更改精細化的風險檢測項。您可通過點擊登入 AI 安全護欄產品控制台，隨時開啟或關閉相關的風險檢測內容，以適合適的風險檢測模板。

   - **自訂檢測項**：對內容合規檢測中的精細化標籤進行配置。
   - **自訂風險閾值**：對精細化標籤的命中閾值進行配置，在模型輸出的 0-100 置信分中，支援最小配置步長 1。
   - **自訂過濾詞**：對需要檢測和攔截的敏感詞（如競爭者名字等）進行配置，支援增、刪、改等詞庫管理操作。

資料來源: https://www.alibabacloud.com/help/tc/content-moderation/latest/what-is-ai-security-barrier

什麼是資料外洩防護 (Data Loss Prevention)?

資料外洩防護 (DLP) 是一套用來防止敏感資料被未經授權存取、傳送或洩漏的資安技術，常見於企業與政府機構，用來保護機密與合規資料。

核心原理: 辨識資料 → 監控流動 → 採取動作 (允許 / 警告 / 阻擋 / 記錄)

常見的 DLP 產品例子:

- Microsoft Purview DLP（Microsoft 365）
- Google Data Loss Prevention (DLP)
- Symantec / Broadcom DLP
- Forcepoint DLP
- McAfee DLP

（學校或需聯絡相關供應商安排設定，並可能須繳付額外授權費用。）

規劃與實施資料外洩防護 (DLP)

1. 管治與問責
2. 資料分類與盤點
3. 敏感度標籤（DLP基礎控制)
4. AI工具專屬資料外洩防護
5. 提示層面(Prompt)保護（人為錯誤控制）
6. 存取權限與過度共享
7. 監察與持續風險管理
8. 審計、事故應變及證據保留
9. 非核准 AI 工具
10. 教職員培訓與意識提升

DLP 能管哪些行為？

例子:

- 寄 Email（內外部）
- 上傳到 OneDrive / SharePoint / Copilot
- Teams / Chat 傳送
- Copy 到 USB
- 列印、截圖（Endpoint DLP）

1. AI 使用批准及責任分工

AI 審核及批准
學校在引入 AI 工具時，應先作審慎評估；涉及大型或全校性 AI 平台的部署， 建議按既定機制提交管理層審核及批准。

AI 支援教學及行政
AI 僅用於輔助教學與行政工作， 非取代專業判斷的決策。

指定負責人
指定 AI 與資料保護負責人，確保事故時有明確處理流程。

使用範圍文件記錄
明確記錄 AI 使用範圍，界定允許及禁止用途，保障資料安全與私隱。

2. 識別及標示敏感資料

全面識別敏感資料
學校需識別所有敏感資料例如學生個人及評核資料 ，確保資料完整掌握。

標示高風險存放位置
應標示資料存放雲端及電腦等高風險位置。

利用 DLP 保護數據
清楚分佈有助防止AI 工具未授權存取及資料濫用。

支持風險評估與合規
識別和標示工作是DLP技術控制的重要基礎。

3. 建立及應用敏感度標籤 (Sensitivity Labels)

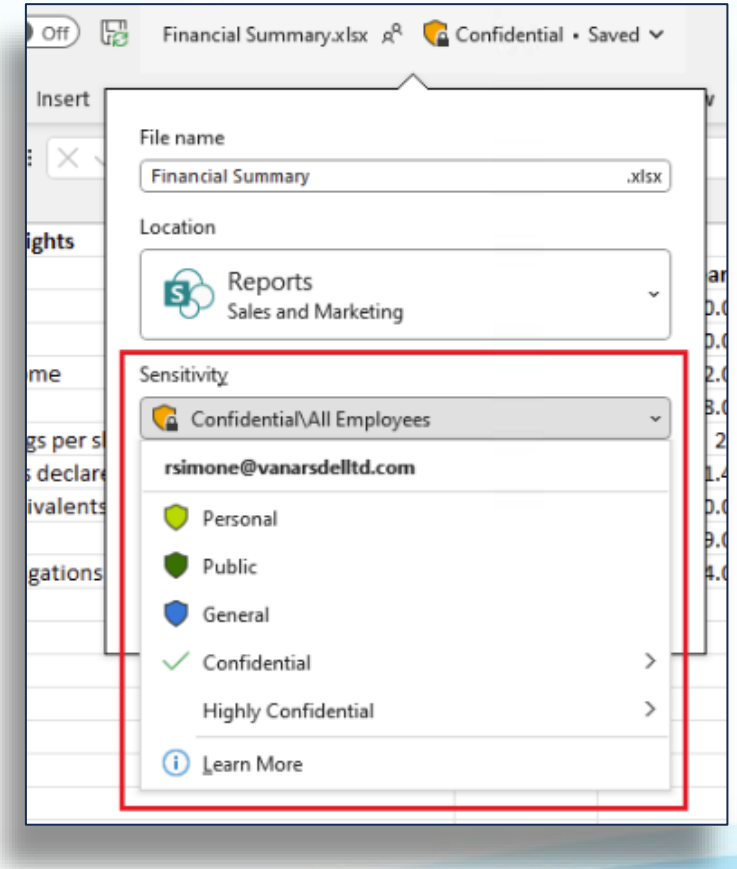
統一敏感度標籤政策
學校制定公開、內部、機密等標籤，並通知所有教職員執行。統一管理提升資料安全。

自動標籤功能啟用
DLP系統自動識別學生資料及評核關鍵字，自動標記敏感資訊，提升效率。

高風險資料自動加密
對機密及以上標籤資料自動加密，限制未授權存取，強化資料保護。

AI 工具強制執行標籤
AI繼承標籤政策，持續監控並強制保護敏感資料安全。

例子



資料來源: https://learn.microsoft.com/en-us/purview/sensitivity-labels

4. 限制 AI 處理敏感內容

設定敏感資料政策 (DLP Policy)
學校需制定政策阻止 AI 處理學生個人資料和敏感文件; 如必須使用，須先做遮罩/去識別化/代碼化。

限制電郵敏感內容
限制 AI 摘要含有敏感標籤的電郵，防止資料外洩。

保護本機及離線檔案
確保本地及離線檔案納入保護，防止人為疏忽洩漏資料。

5. 防止敏感資料輸入 AI 提示 (Prompt)

技術措施防護
學校需設置DLP技術手段，防止輸入敏感學生資料。系統應能即時偵測並阻擋此類內容。

清晰阻擋訊息
系統偵測敏感資料時，應顯示清晰訊息，解釋原因並引導正確行為。

控制人為錯誤
多數資料外洩為意外，強調減少因操作錯誤導致的私隱風險。

6. 審查及限制資料存取

定期檢查存取權限
學校需定期審查雲端文件儲存及分享空間的資料存取權限，防止不當開放。

嚴格限制敏感資料夾
考試及學生資料夾必須嚴格限制存取，保障資料安全。

清理遺留與繼承權限
清理過時或繼承權限，避免權限錯誤與資料洩漏風險。

嚴格管理對外分享
對外分享需嚴格把關，尤其涉及學生資料的資料夾。

7. AI 使用監察及風險評估

監察 AI 使用情況
集中監察 AI 工具使用及異常存取行為，保障系統安全。

定期風險檢討
每學期進行風險評估，及早發現及修正潛在問題。

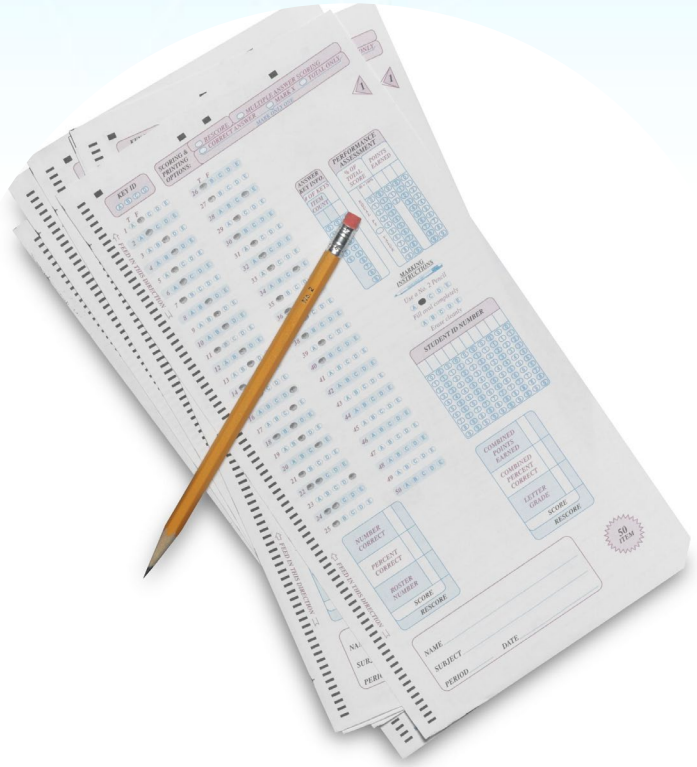
持續風險管理
持續採用技術和管理措施，確保合規及資料安全。

8. 建立審計及事故處理機制

AI 互動審計
AI 互動可被審計並保留調查用的日誌資料。

明確事故應變
制定事故應變程序，處理資料外洩與意外披露情境。

證據提供能力
遇查詢時能即時提供合理保護措施的證據。

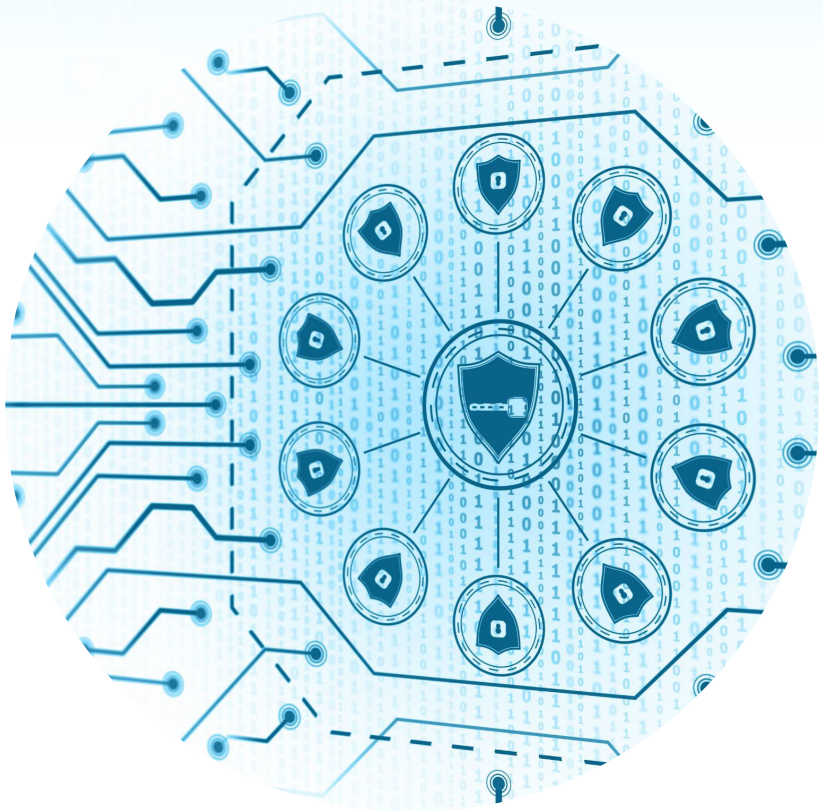


9. 限制使用不安全的 AI 工具

限制未授權 AI 工具
學校禁止使用可疑的 AI 工具，防止學生資料外洩。

阻止資料貼上行為
阻止教職員在瀏覽器貼上學生個人資料至 AI 網站。如必須使用，須先做遮罩/去識別化/代碼化。

明確 AI 工具政策
清晰溝通哪些 AI 工具可用，哪些屬違規，保障資料安全。



10. AI與資料保護培訓及指引

專門教職員培訓
學校為教職員提供專門的 AI 工具與資料保護培訓，強調合規行為。

簡明行為指引
發放清晰『可做 / 不可做』指引，明確區分允許與禁止行為。

持續文化建立
持續培訓幫助提升資料保護意識，降低資料外洩風險。

謝謝!

特此聲明，對於本簡報或演示過程中所提及的任何第三方品牌、產品或服務，並無作出任何形式的認可、批准、推薦，亦不構成任何合作、隸屬、代理或其他關聯關係。相關提述僅作說明或示例用途，不得詮釋為 香港教育城對該等第三方之任何承諾、保證或責任承擔。