AI Security

從風險到防禦的多層次設計

From Risk to Defense-by-Design

多租戶地端 LLM 輔助 SOC 營運的脈絡隔離治理

PART I — OPENING & FRAMING

第一部分　開場與定位：AI Security 的典範轉移

為什麼不是傳統資安的延伸

傳統資安建立在「程式碼與資料分離」之上。
LLM 的顛覆：指令與資料被壓進同一個權杖序列（Token Sequence）——系統提示詞、使用者輸入、檢索文件、工具回傳值，全在同一脈絡視窗被同等對待。
因此提示詞注入（Prompt Injection）不是可修補的漏洞，而是架構層級的固有特性。

對應 OWASP Top 10 for LLM：LLM01 Prompt Injection、LLM06 Sensitive Information Disclosure。

開場案例：教育場域跨租戶脈絡洩漏

同一所大學的共享 LLM 平台：
行政處擬公文、教師批改作業（含學生個資）、研究團隊處理未發表資料。
三方共用同一推論服務，若隔離不當，學生成績可能透過 KV-cache 殘留或語意鄰近，洩漏給不相干人員。
此即個人資料保護法（PDPA）下的特定目的外利用風險。

PART II — THE RISK LANDSCAPE

第二部分　風險全景：六大風險群

座標系：OWASP Top 10 for LLM × MITRE ATLAS × NIST AI RMF　|　R1–R26 投射於多租戶地端 LLM 輔助 SOC

1 群一　脈絡洩漏

Context Leakage

R1 租戶脈絡洩漏、R3 KV-cache 殘留、R4 系統提示詞逆向萃取

2 群二　注入與繞過

Injection & Bypass

R2 提示詞繞過隔離、R7 經由 RAG 的間接注入

3 群三　知識庫污染

Knowledge Base Contamination

R5 惡意注入、R6 非蓄意污染

4 群四　生命週期與殘留

Lifecycle & Residue

R19 GPU VRAM 硬體記憶殘留、R21 脈絡銷毀不完整

5 群五　代理人架構

Agentic Architecture

R13 Harness 單點故障、R18 湧現行為失控

6 群六　法遵落差

Compliance Gap

R23 AI 基本法操作缺口、R25 委外規範缺口、R26 國際標準銜接空白

PART III — FIVE LINES OF DEFENSE

第三部分　五道防線：風險與設計原則的橋樑

五個正交（Orthogonal）的治理問題——彼此獨立、合起來完整覆蓋。

WHO	**身份與信任**	DP1 身份確定性	回應 群一、群二
WHAT	**外部介面信任與隔離**	DP3 結構安全優先性	回應 群三
HOW	**代理人架構治理**	DP14 代理人通訊矩陣最小化	回應 群五
WHEN	**生命週期與銷毀**	DP8 推論基礎設施微隔離與安全強化	回應 群四
PROVE	**稽核與合規**	DP9 可稽核性、DP12 法遵與技術協調	回應 群六

橫向決策準則

Lateral Criteria

DP7
結構性隔離優於語意過濾

DP11
受控情資跨租戶共享

貫穿全部五道防線，不綁定任一道。

第四部分　技術深潛

DP7　結構性隔離優於語意過濾

模型「遵守指令」是機率性、非強制性的；提示詞裡的禁令都可能被精心構造的輸入繞過。唯有結構性隔離（不同租戶用不同模型實例／記憶體空間／知識庫索引）才提供可證明的安全保證。

記憶殘留的硬體真相

KV-cache 跨租戶殘留（R3）與 GPU VRAM 硬體記憶殘留（R19）共通點：發生在抽象層之下。快取層選擇性隔離與記憶體歸零是解方，但須在效能與安全間取捨。

DP8　推論基礎設施微隔離

容器執行期隔離梯度：runC（最快、隔離最弱）→ gVisor（使用者空間核心攔截）→ Kata（輕量虛擬機、最強隔離、開銷最大）。Mavridis 等人（2020）提供量化依據；搭配 TEE 可信執行環境形成硬體信任根。

PART V — GOVERNANCE & COMPLIANCE

第五部分　治理與法遵

在地法規遵循

人工智慧基本法：提供原則性框架。
資通安全管理法第 10 條：規範委外資安義務。
（第 21 條另處理中央主管機關稽核，兩者勿混。）

數位發展部「公部門 AI 應用參考手冊」：提供操作指引。

可證明的稽核（PROVE）

DP9 隔離狀態可稽核性：
每一次隔離決策都留下可驗證的證據鏈。

DP12 法遵與技術協調性：
技術控制能對映到法規條文，讓稽核員看得懂、查得到、證得明。

保留「PROVE」而非僅用 Compliance，是為了不丟失證據與稽核維度。

第六部分　收尾與落地

落地優先順序

理由：沒有身份就談不上隔離；沒有隔離就無從稽核。

反方觀點與侷限（誠實面對）

- 成本爭議：結構性隔離（DP7）在資源有限的教育機構可能成本過高。
- 張力無銀彈：完全隔離與情資共享（DP11）之間的取捨沒有單一解。
- 覆蓋缺口：開源 LLM 的模型供應鏈安全（後門、模型投毒）尚未被本框架完整涵蓋，列為未來研究。