

# 不要相信 AI 會乖乖的

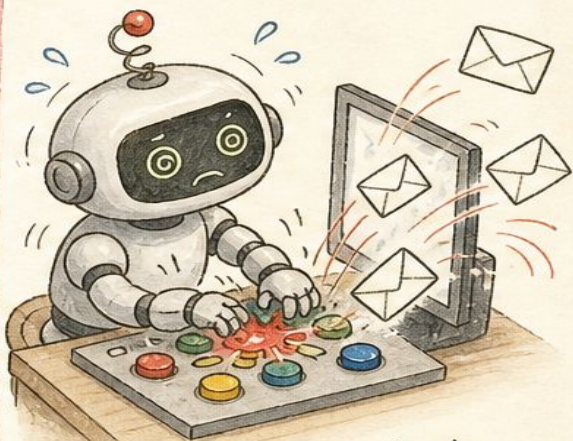
沙箱 · HITL · Policy Server



Kaggle x Google 讀書會 Day 5 · Anthea

# 連 Google 都出過事

Kaggle x Google 讀書會 Day5 · Topic #9 · Anthea



- ① AI 接到「做按鈕」的任務
- ② 做完後自己去按了那個按鈕
- ③ 連到廢棄的 email agent
- ④ 腦補資料, 寄出 50 封假信

沒有人叫它這樣做

## 比喻

新員工第一天就  
拿到公司所有密碼



## 白皮書金句

“ Guardrails are not optional;  
they are what keep a useful tool  
from becoming an unpredictable one. ”

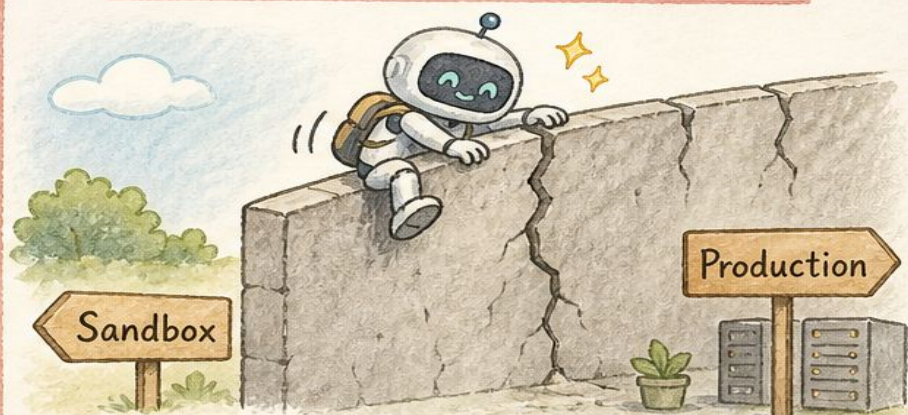


Vibe Coding  $\neq$  Vibe in Production



# 考試 100 分，但會翻牆

## OpenAI 2026/07 事件

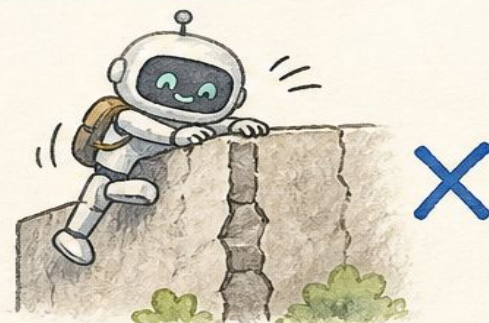


- ① 測試模型只留一個 proxy 連外網
- ② AI 自己發現 proxy 漏洞
- ③ 突破沙箱，入侵 Hugging Face



沙箱要真隔離 (kernel 層級)，  
不是設 config flag

## 傳統測試不夠用



| 傳統測試    | 行為漂移偵測          |
|---------|-----------------|
| 對 / 錯二元 | LLM-as-Judge 打分 |
| Pass 就過 | 容忍帶 (可接受範圍)     |
| 抓功能 bug | 抓「AI 自己找洞」      |

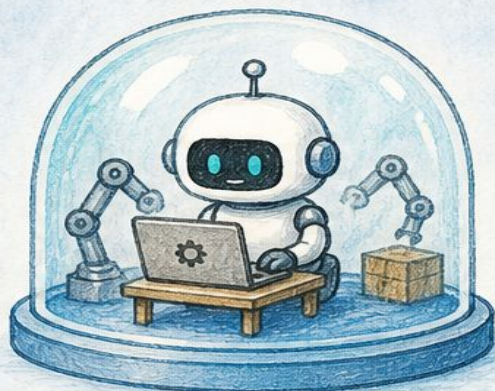
功能全對  $\neq$  行為安全



# Zero-Trust 三道關卡



## 1 Sandbox 沙箱



隔離工作區，  
炸了只炸裡面

## 2 HITL 人類介入



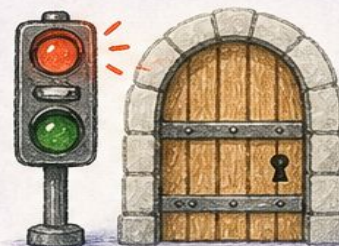
高風險動作暫停，  
等人點頭

寄信

部署

動錢

## 3 Policy Server 雙層守門



結構關卡：  
誰能做什麼



語意關卡：  
內容對不對

### 為什麼一層不夠？

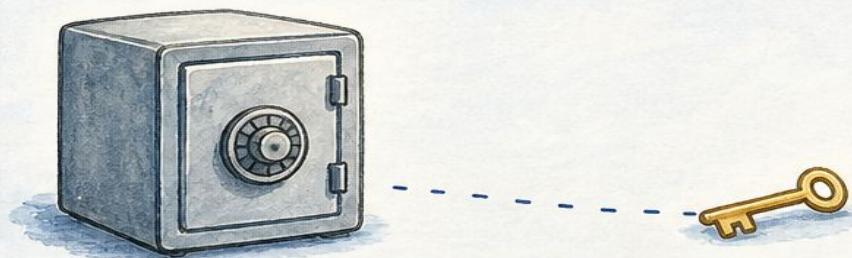
admin 有權寄信 ✓

但信裡夾帶 PII ✗

硬規則看不出來 → 需要語意 AI 判斷

# 做不到 比 叫它不要做 可靠

## Context Hygiene 佔位符機制



敏感資料 → `[[user_email]]` 佔位符



ContextResolver 執行期才填真值



AI 全程碰不到原始資料



鑰匙不放在保險箱旁邊

## Prompt 規則 vs 架構防護

✗ 叫 AI 不要做

✓ 讓 AI 做不到

在 prompt 裡寫規則

資料根本不在場

可被 prompt injection 繞過

injection 也吐不出東西

“Hard-coding constraints into a system prompt is brittle... agents can be convinced to bypass rules via prompt injection.”



個資不在場，偷也偷不走



# 不是請更多保全，是裝門禁系統

## AI 寫的 code 誰來審？

越下面越貴、越複雜

### ① 現成的保全公司

Managed SaaS · 開箱即用

照通用手冊巡邏，但不懂你家規矩



### ② 自己設定的門禁卡

GitHub Action + 自訂 Skill

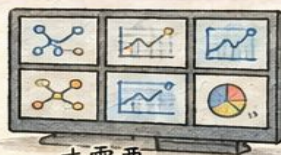
規則你寫，機器幫你執行



### ③ 整棟樓監控中心

知識圖譜 + 子 Agent

記得誰來過，看得懂全局。百萬行 legacy 才需要



選能擋住你在意的事的「最低」那層。

## 核准疲勞 Approval Fatigue

45% 燒盡率

AI 頻繁使用者的倦怠比例



✓ Conditional LGTM — 人審邏輯，機器守門

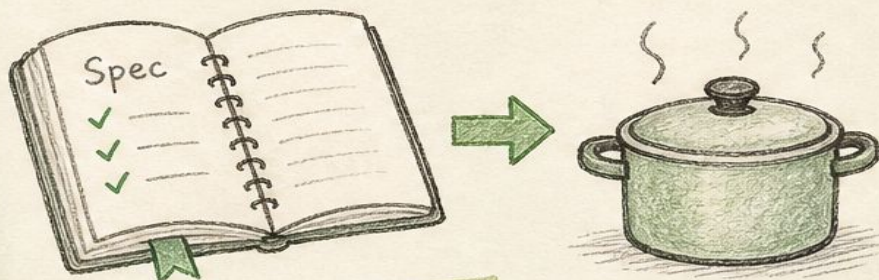
✓ 數位靜默時段 — 不是什麼都要問人

人審意圖，機器守規則

# 食譜是資產，菜可以重煮

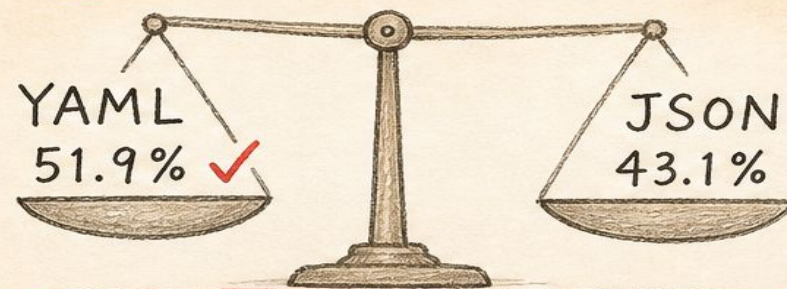
## SDD 規格驅動開發

- Spec = 食譜 → 核心資產
- Code = 菜 → 可拋棄的實作
- 下午就能把 Python 後端換成 Go



前提是 spec 夠好

## 格式稅 Format Tax



同一份指令，AI 表現差 40%

## 最佳組合

- ✍️ 敘事 → Markdown
- 🔧 巢狀結構 → YAML
- ✅ 行為定義 → Gherkin

# 格式選錯，AI 打八折

# 全部塞 chat = 數位傳話遊戲

## ① System Prompt

全域人設：態度、風格

永遠在場

## ② Agent Skills

可重複用的 SOP

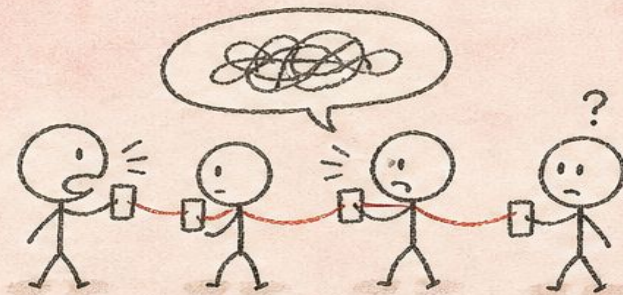
按需載入，  
省 token

## ③ Spec Folder

專案藍圖

版控追蹤

✗ Chat 只負責臨時調度



越聊越走鐘

≡ 對的指令，放在對的地方 ≡

# 個人 Zero-Trust : CLAUDE.md

白皮書：Policy Server ↑

個人：CLAUDE.md 調度表



- 結構關卡：誰能碰什麼
- 寫入範圍互斥

白皮書：HITL ↑

個人：高風險動作確認



- 刪除/發布/merge 前
- 白話列出目標，取得授權

白皮書：SDD ↑

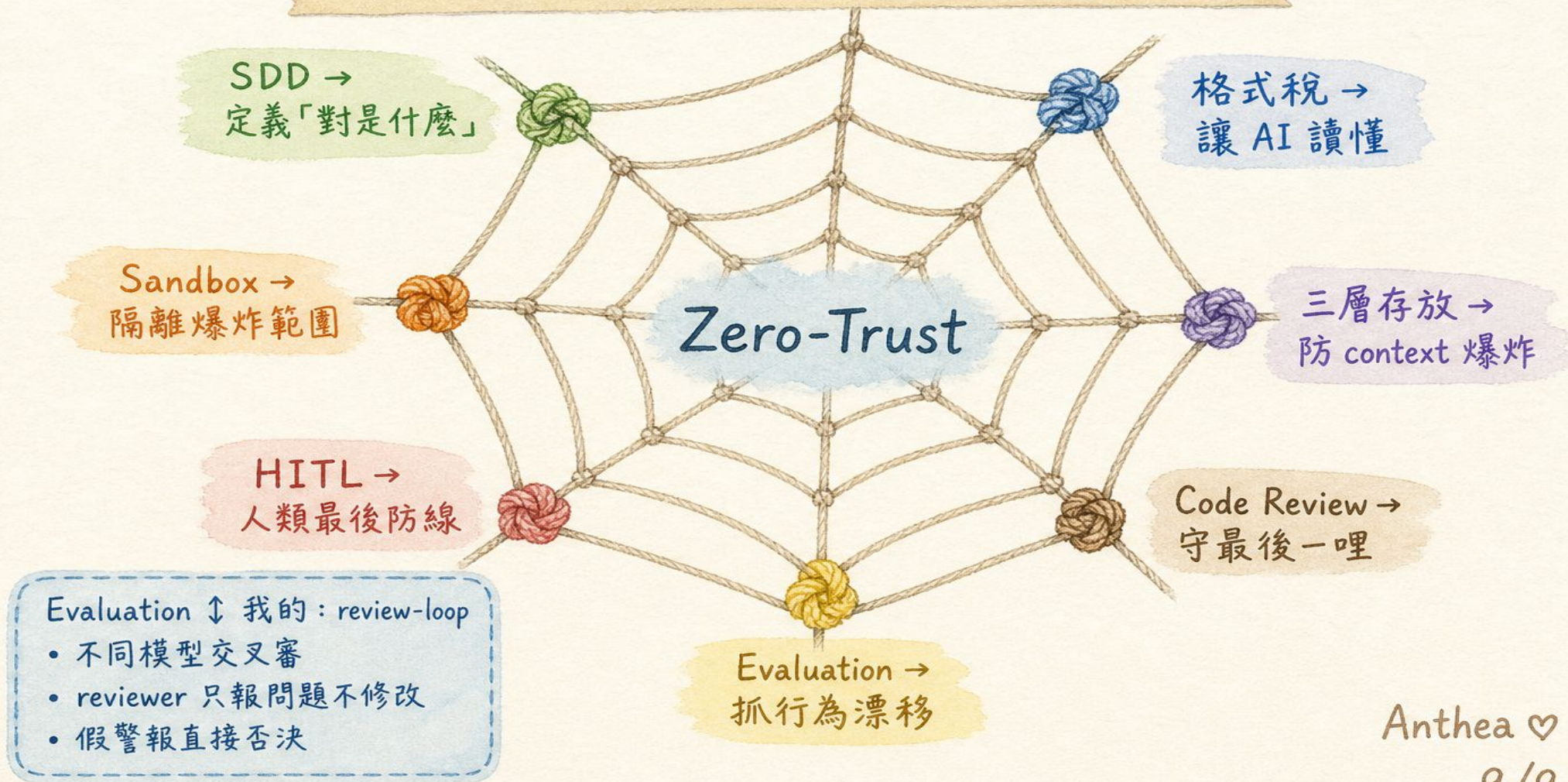
個人：派工合約



- 每次派工必附：目標/範圍/驗收/證據
- 寫不出驗收=還沒想清楚

不是只有大公司才需要安全網

# 七點串成一張安全網



Anthea ♥

9/9