

IT就業、技能替代與企業AI落地案例

臺大計資中心 106遠距教室

2026/8/20 14:00 – 17:00

👤 講者：黃國泰 | 想見科技總經理、程式驅動協會 理事長



自我介紹



台大計資中心
Computer and Information Networking
Center



黃國泰

一切都是最好的安排

總監

想見科技負責人
程式驅動理事長
經濟部工業局職能基準委員

程式驅動 想見科技

● 營利事業 – 想見科技負責人

- ✓ AI模型應用與產品整合
- ✓ 智慧長照應用開發 – 好管家、大腦活化、五十肩復健系統
- ✓ 智慧商圈 – MAZE人流分析系統
- ✓ 教育訓練 – 文大推廣、北護推廣、慈大、高科大、工研院
- ✓ 政府專案 – 勞動部、經濟部產發署

● 非營利事業 – 社團法人程式驅動理事長

- ✓ 2017-2026 AIoT營隊負責人&講師 (偏鄉教育) SDGs#4
- ✓ 數位賦能計畫
 1. 二手電腦回收與轉贈弱勢：【數位再生·偏鄉啟航】
 2. 提供雲端服務給非營利組織 (NPO)：【雲端共好·組織賦能】

● 個人興趣

- ✓ 寫PHP程式、玩 Kali-Linux、AIoT研究
- ✓ 網球、桌球、重機旅行

這 7 年來團隊累積的成果

817

位

AI大數據、Java、雲端資安及全端工程師

8

間

大專校院教育 + 政府專案合作

3

產業

產品研發專案合作

第一部分：全球大洗牌

第二部分：技術落地實務

第三部分：地端模型應用與比較評測

第四部分：AI 漏洞挖掘與系統提權

第五部分：政策紅利與人才引導

| 不是股市分析，是用企業需求來看AI(2022-2026)未來展望

Anthropic 剛剛(8/18)公佈年化營收(run rate) 650億 美金



 Yahoo股市

Anthropic年化營收突破650億美元 半年多暴增7倍、狠甩OpenAI

彭博引述消息人士報導，Anthropic按目前營運表現推算，年化營收已突破650 億美元，較去年底的水準成長逾7 倍。報到指出，截至7月底，Anthropic 的營...

1 天前

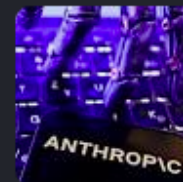


 UDN

Anthropic年化營收突破650億美元AI需求展望提振晶片股| 全球財經| 全球

AI新創公司Anthropic營收成長急遽加速，7月底年化營收突破650億美元，第2季調整後營業利益也呈正數，可見先進AI模型業者正把使用需求轉化為實際營收。

2 天前



 中央社 CNA

Anthropic增長勢頭猛年化營收衝破650億美元| 國際

知情人士今天透露，截至7月底，Claude聊天機器人開發商Anthropic年化營收已突破650 億美元，凸顯這家可能於今年稍晚掛牌上市的人工智慧（AI）新創公司...

1 天前

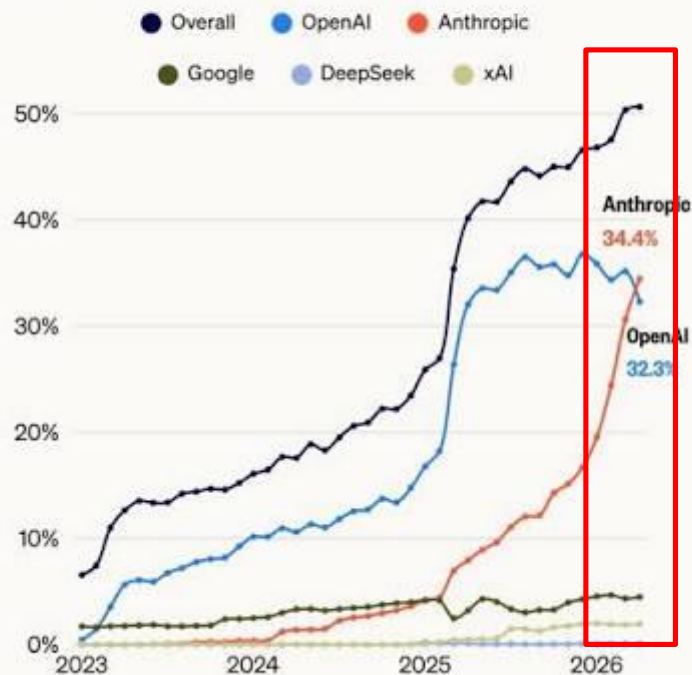


Anthropic 不久之前 (5/14) 企業用戶超車 OpenAI (ChatGPT)



MODEL ADOPTION RATE—
% OF U.S. BUSINESSES
W/PAID AI SUBSCRIPTIONS

SOURCE: RAMP AI INDEX



NOTE: OVERALL INCLUDES BUSINESSES SUBSCRIBED
TO ANY AI PRODUCT OR SERVICE

| Anthropic 個人20美元/每月 訂閱費大家覺得貴嗎？



20_{USD}

每個月訂閱費

| Anthropic 企業100美元/每月 訂閱費大家覺得貴嗎？



100

USD
每個月訂閱費

| Anthropic 外租機房 200 – 250 億 USD

 Claude



ASUS 窮人版 AI Server 準系統 – 當時的嗤之以鼻，現在的高不可攀

70-120_B

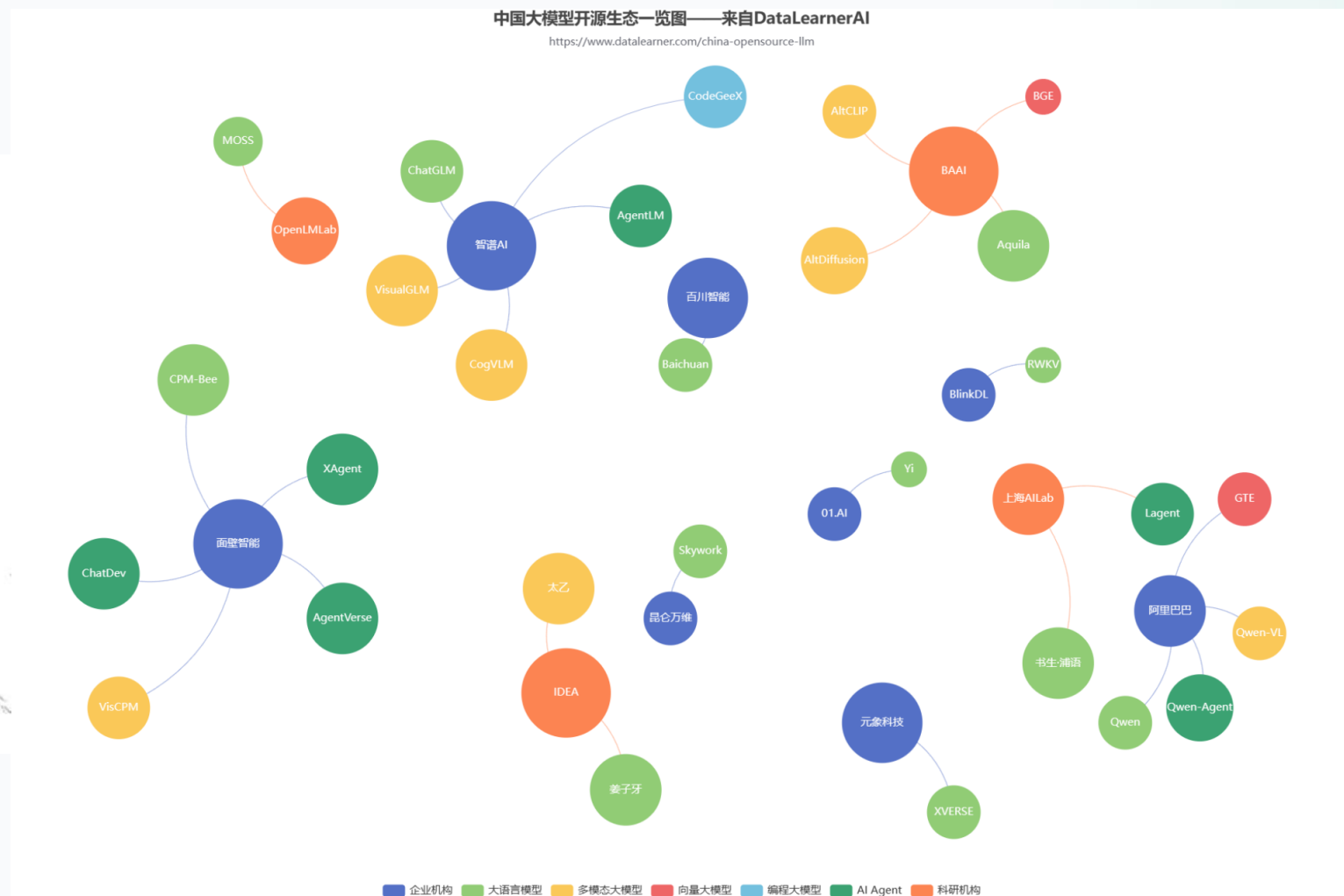
LLM



這款搭載 NVIDIA GB10 Grace Blackwell 超級晶片的迷你 AI 工作站，在市場上的定價與實際通路販售價格有著明顯的波動過程： 鉅亨網 +1

時間區間	市場事件與供需狀態	1TB 版本預估市場均價 (NT\$)	4TB 版本預估市場均價 (NT\$)
2025 年 10 月	華碩官方正式發表，宣布採取客訂接單，公告基礎建議售價。	\$97,900 (初始定價)	(尚未正式供貨)
2025 年 11~12 月	由於 NVIDIA Blackwell 晶片極度搶手，首批到貨前即被預約一空，市場出現供不應求的溢價現象。	\$123,900 ~ \$139,900	\$139,000 (初期預估)
2026 年上半年	評測解禁與電商通路陸續鋪貨，包含批發與拍賣通路價格因稀缺性拉抬。	\$125,900 ~ \$168,000	\$200,000 ~ \$228,000
2026 年 8 月 (目前)	大型主流電商平台（如 PChome、Momo、Autobuy）常態上架，價格趨於穩定但維持在高位。	\$175,900 (主流價)	\$245,900 (主流價)

還有那些來字神州的AI，使用龐大的經濟（不明）實力硬剛西方



AI讓預測變便宜了，為什麼就變貴了 – AI經濟的策略思維



26年前企業導入

LINUX

為了解決微軟授權金問題



全球職能大洗牌

從北美看 IT 工作現場、技能替代率，到不同年齡層就業數據深度解密

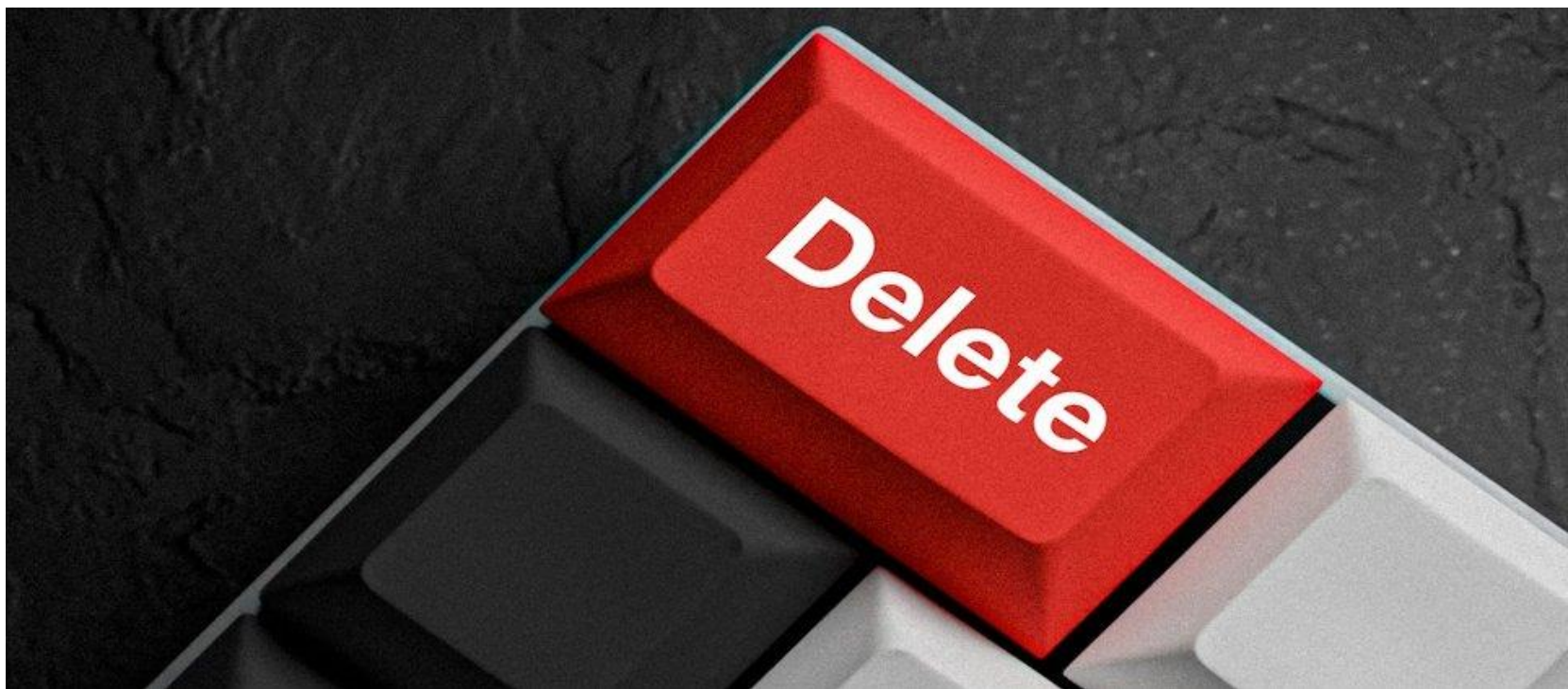
人會對**未知**充滿恐懼

9秒

能做什麼？

AI代理人9秒刪掉新創公司的資料庫及備份 – 2026/4/29

提供租車SaaS服務的新創公司PocketOS創辦人Jeremy Crane於上週發生慘劇後，[將此事在X上](#)公開警告大家。上週五，執行Anthropic旗艦模型Opus 4.6的Cursor刪掉其營運用的資料庫，之後透過一個API呼叫，將在其雲端基礎架構服務Railway上的備份也刪掉。整個過程花了9秒鐘。



Never Fucking Guess

Use API to FIX

Volume Delete

AI比六歲小孩還不會說謊

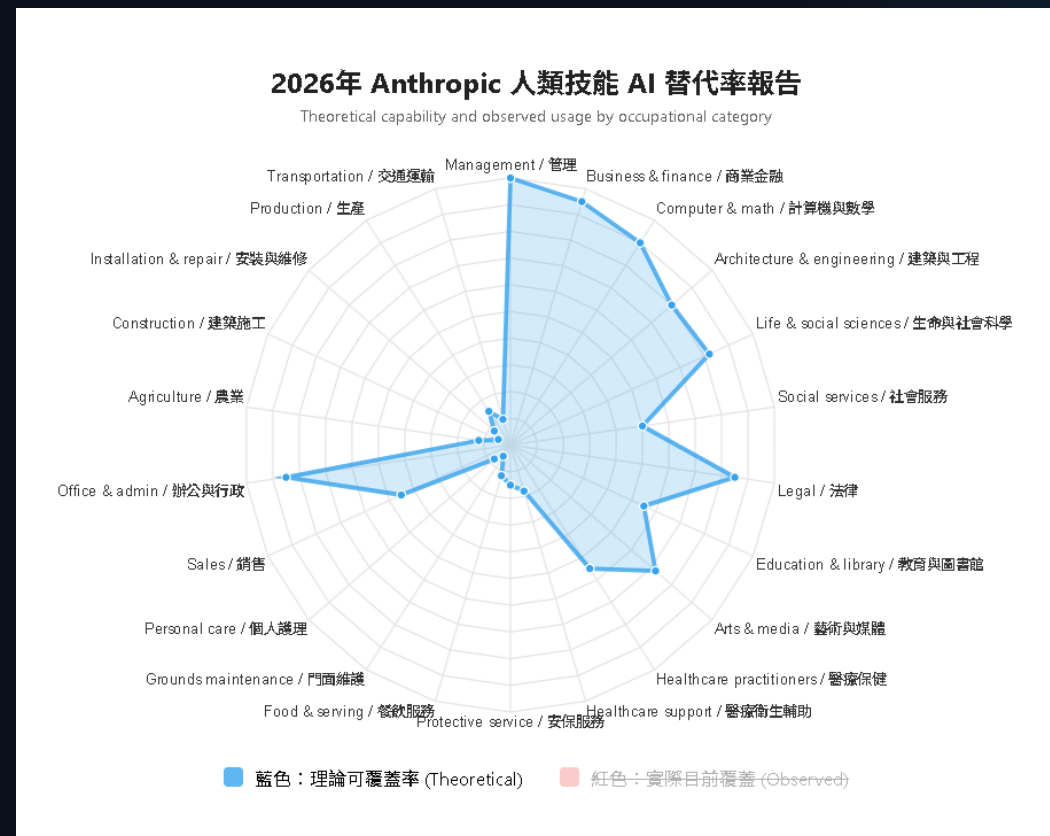
人類會**犯錯**

像人類的**AI**也會**犯錯**

但人會有道德與愧疚感，所以煞車需要**人**

北美 IT 工作現場：AI 帶來的範式轉移與職能重組

- 🔗 **Cursor / Copilot 標配化**：軟體開發從「Raw Coding」轉型為「AI 提示與系統架構組裝」。
- 👤 **初階工程師 (Junior SWE) 斷崖式縮水**：過去由新鮮人負責的常規模組、Debug 工作，高達 80% 已由 AI 代理人接管。
- ⚡ **10x Engineer 的實體化**：具備全棧與 AI 架構能力的資深工程師，其產出效率被放大 10 倍以上，成為企業獵才核心。
- ↻ **範式轉移**：軟體工程的核心價值，從「寫代碼的速度」變為「理解商業邏輯與精準系統解構的能力」。



Anthropic 人類技能 AI 替代率報告解析

☢高替代率風險區 (> 75%)

- 基礎代碼編寫與 Debug (82%)：規格明確的 API 撰寫、資料清理腳本與標準化單元測試。
- 系統常規維運與 Shell Scripting (78%)：標準化的備份、日誌監控與基礎 CI/CD 流程配置。
- 結構化文件初審 (75%)：標準合約比對、合規性查核與基礎 bug triage。

🛡低替代率黃金區 (< 25%)

- 模糊邊界問題定義 (12%)：從零開始釐清複雜的商業痛點，將混亂需求轉化為可執行的架構。
- 系統級架構設計 (15%)：複雜的混合雲、高併發 (High Concurrency) 和異質資料串接拓撲。
- 跨部門政治與談判 (18%)：涉及多方利益權衡、法規摩擦與高情感密度的組織溝通。

台灣企業不再任用「小白」當初階工程師？

多數企業（金融科技、智慧製造）定義新人任用標準，已經拉到
需有1-1.5年相關工作經驗（訓練 + 實習是被認可的）

傳統軟體委外專案的規模劃分維度

專案規模	人月與工期	系統架構與業務複雜度	團隊編制與交付重點
小型專案	1 – 6 MM 1 – 3 個月	單一功能模組、標準活動頁、小型形象官網或輕量內部工具；無複雜第三方系統整合。	1-2 位全端或前後端工程師，通常需兼任 PM 與測試工作。
中型專案	6 – 30 MM 3 – 9 個月	具備多角色權限控管、金物流串接、自訂業務流程（如中小企業 ERP/CRM、專屬電商平台）。	配置專職 PM、UI/UX 設計師、前後端分離開發團隊及 QA/測試人員。
大型專案	30 MM 以上 9 個月至數年	跨核心系統整合（如銀行核心、醫療 HIS）、高併發微服務架構、嚴苛合規資安稽核與跨組織協作。	跨部門架構師、專案總監、資安顧問及多組專責開發與維運團隊。

AI 時代：軟體工程師與委外開發的轉型差異



Focus Shift

產出瓶頸轉移

從過往耗時於「編寫基礎代碼與刻 UI」轉向「**問題精確定義與 Context 建構**」。開發瓶頸不再是語法實現，而是能否提供嚴謹的架構規格、業務規則與邊界條件。



Quality Assurance

品管與審查機制

工程師的核心職責從「親手撰寫」轉為「**Code Review 與驗證**」。重點在於及時識別 AI 幻覺 (Hallucination)、微小資安漏洞、效能未佳化的 SQL 查詢與隱性邏輯矛盾。



Competency Upgrade

架構與整合門檻提升

基礎體力活代碼被 AI 消化，拉高了初階工程師 (Junior) 入門門檻。工程師在發展早期即需具備「**系統全貌視角**」、資安合規、CI/CD 部署與跨系統串接整合能力。

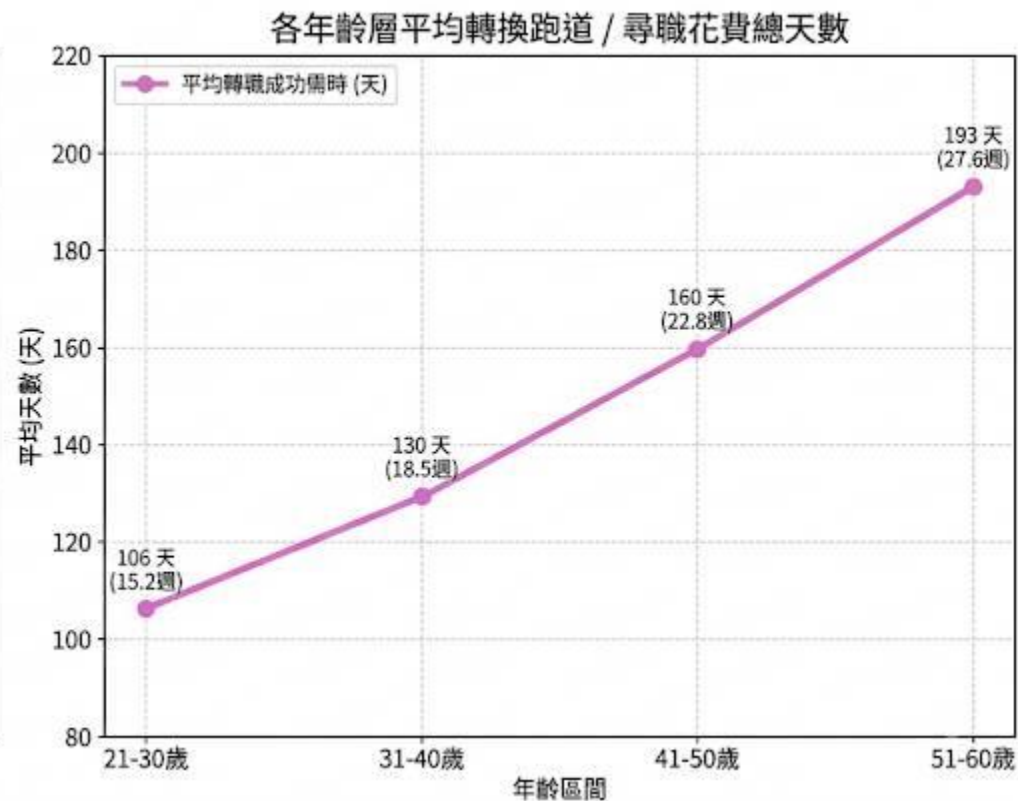
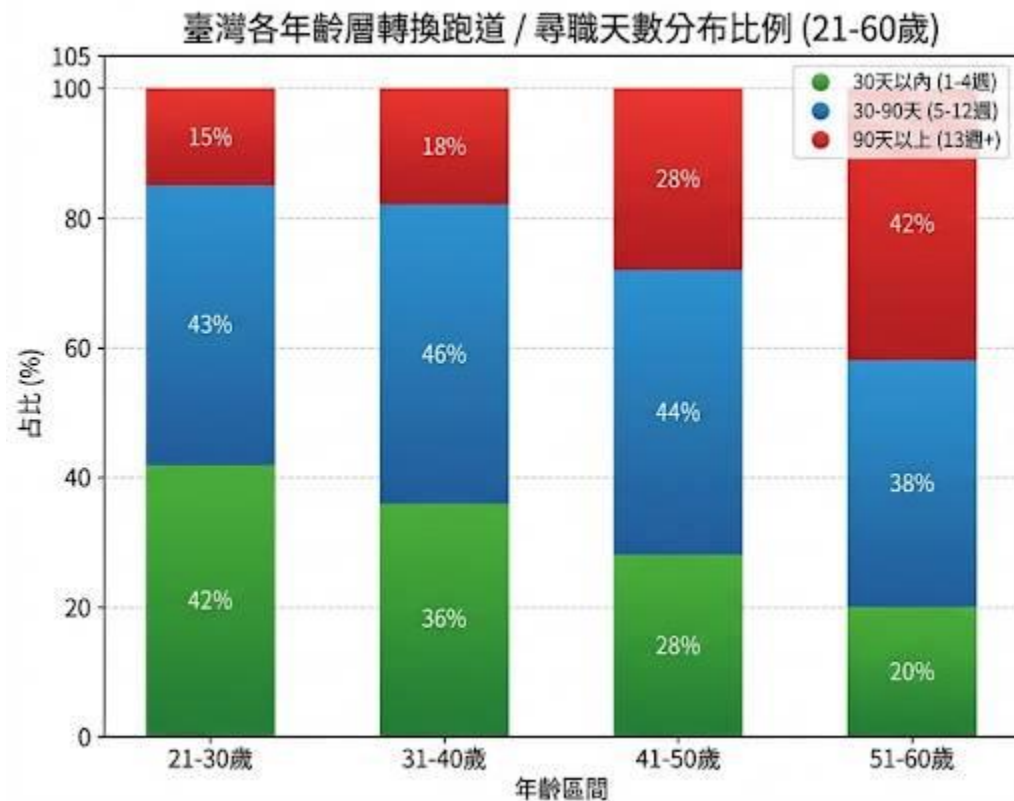


Business Model

委外商業模式衝擊

AI 大幅壓縮標準模組開發工時，使傳統「人月工時 (T&M)」計價受挑戰。委外廠商轉為競爭「**交付速度**」、「**產業領域知識 (Domain Knowledge)**」與「**價值/結果導向報價**」。

AI 時代：台灣年齡與求職時間分析



第一章：大浪襲來

從 ChatGPT 浪潮到 AI 時代的職場與企業衝擊

| 有時人(工作者)的焦慮是因為看到『數據』

22-55歲up 工作者的集體焦慮與衝擊

- ❗ 工作模式重塑：ChatGPT 爆發至今，全齡白領工作者日常被重整。
- ♻️ 管理者雙重焦慮：面臨員工「不會用 AI 落後」與「被 AI 取代恐慌」。
- ⚠️ 企業面臨三重考驗：數位轉型、ESG 減碳轉型、AI 智慧轉型並進。
- 🛡️ 強化營運韌性：因應國際情勢、供應鏈重組、美國關稅等外部衝擊。

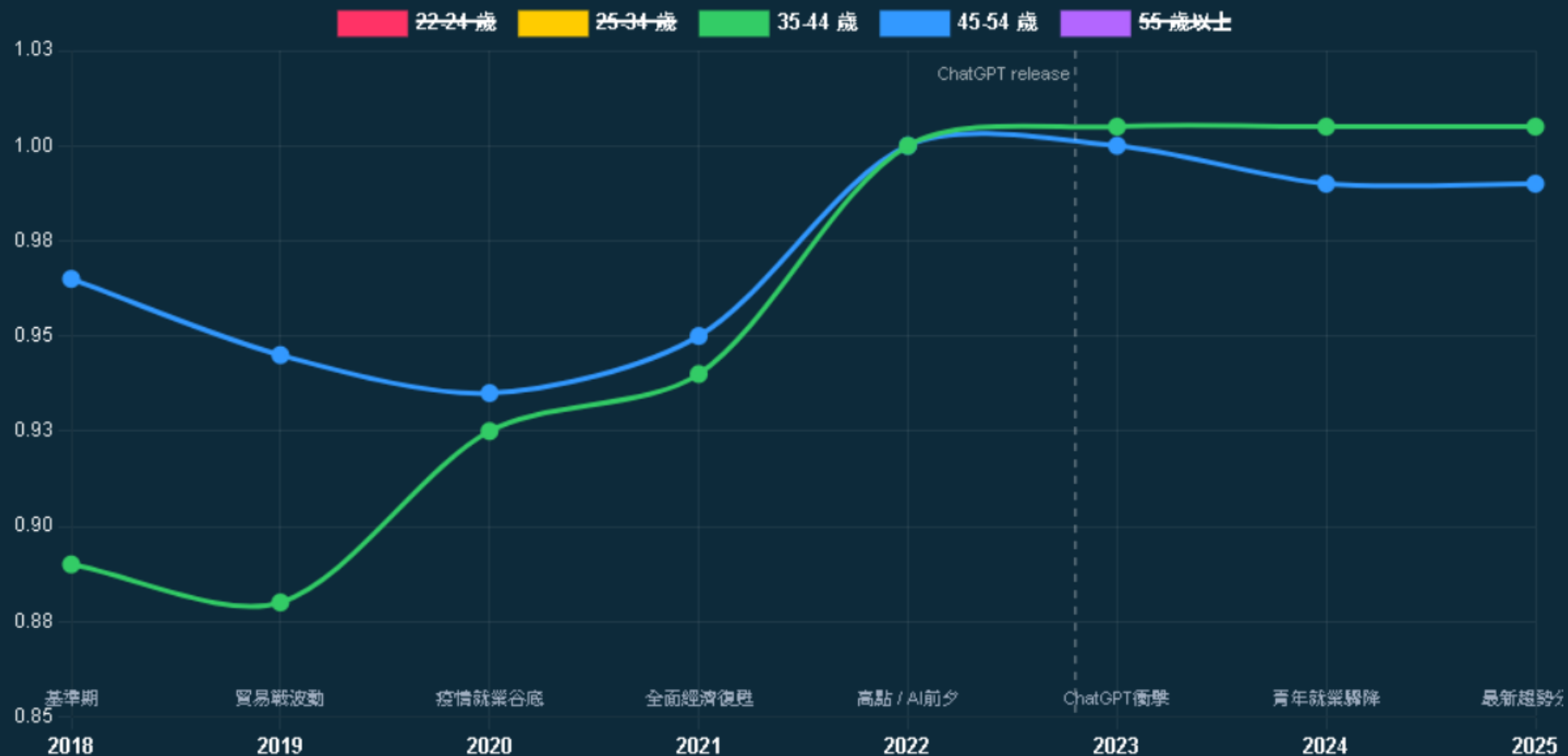


22-24
25-34
55up?

這兩個年齡區間？

Employment by age in AI-exposed industries

AI 暴露產業中不同年齡層的就業變化趨勢 (2018-2025)



資料來源：美國人口普查局經濟學家 Lee Tucker

Designed by FORESII

| 有時人(企業主)的焦慮是因為看到『數據』

300,749

1-4 月美企裁員總人數

-50%

較 2025 年同期下降
(去年聯邦員工大砍)

+33%

科技業裁員年增
共逾 8.5 萬人

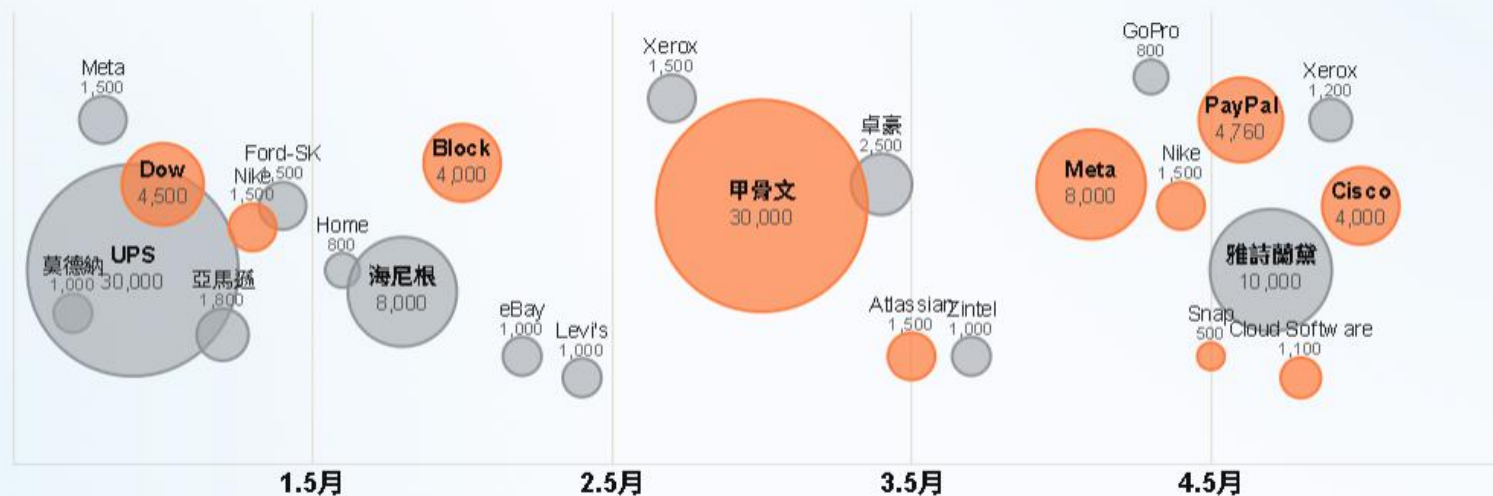
33

本表追蹤的
主要企業事件

FIGURE 01 · BUBBLE CHART

2026 主要裁員事件分布（按月份）

圖圖大小 = 裁員人數規模 (最大者為 UPS/甲骨文 各 3 萬人)



■ 動因明確指向 AI 重組 ■ 景氣/策略調整為主

註：Oracle 規模為 TD Cowen 估算；Walmart 含「裁員與內部調動」共約 1,000 人。資料來源：各公司聲明與 WSJ 報導。



第二部分：技術落地實務

企業數據架構重組、AI 訓練與推理應用趨勢深度剖析

2026/4/18 我接到一個 數據分析服務商 負責人的電話，希望我幫他
建立地端AI模型訓練/推理架構，要用在現有產品中服務客戶



Lenovo T470
2017年上市
Intel 第七代 CPU
16GB RAM DDR3

超輕量級模型 (SLM) 的企業核心價值：高效與隱私並行

在 16GB RAM 的標準設備上，超輕量模型 (Small Language Models) 的目標不是取代 ChatGPT，而是將 AI 能力嵌入企業既有的自動化流程 (Workflow) 中

維度	大型模型 (LLM/Cloud)	超輕量模型 (SLM/Local 16G)
部署成本	昂貴 (按 Token 計費/高階 GPU)	極低 (現有辦公電腦/零邊際成本)
處理速度	高延遲 (受網路影響)	極速 (本地運算，毫秒級響應)
數據隱私	資料需上雲，存在資安風險	100% 內網運行，資料不外流
主要任務	創意寫作、跨領域推理	結構化轉換、分類、自動化路由

各位學長姐，你們是否還不知道自建AI到底能幫企業做到什麼

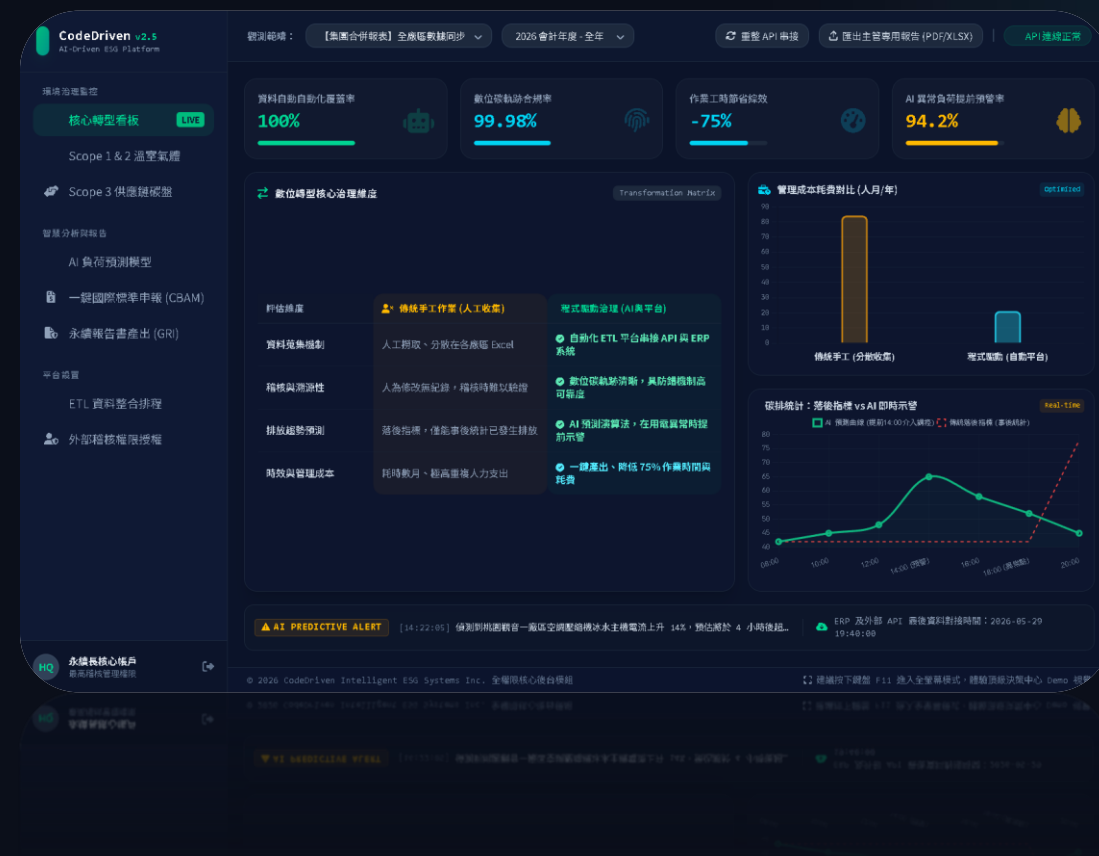
數據應用新架構：從傳統 ETL 走向 AI-Ready & RAG

 **非結構化資料崛起**：企業 80% 的智慧散落在合約、PDF與郵件中，傳統關聯式資料庫已不夠用。

 **從 ETL 到向量管道 (Vector Pipelines)**：重點在於如何將非結構化資料進行 Chunking、Embedding 並精準入庫。

 **RAG (檢索增強生成) 成為企業標配**：無需重新訓練模型，透過 RAG 就能讓 AI 準確讀取內部私有知識庫。

 **資料治理與權限控管**：企業落地最常遇到的瓶頸——必須在向量檢索階段，嚴格確保敏感資料不跨部門外洩。



2026 企業 AI 應用對比：訓練 vs 推理

評估指標	微調與訓練趨勢 (Fine-Tuning / Training)	落地推理應用趨勢 (Inference / Edge)
市場現況與支出	從「盲目煉丹」冷卻，回歸理性與特定垂直模型微調	爆發式成長！佔企業 AI 總預算支出的 80% 以上
部署路徑	SLM (領域小模型) 微調：如 Phi-4 等地端部署微調	RAG 與多步驟 Agent 協作：免訓練，直接呼叫 API
硬體與晶片部署	高階 AI 伺服器集群 (如 DGX H100)，多單次投資	邊緣運算 (Edge AI)：PC、Raspberry Pi 等端點部署
核心瓶頸	資料隱私與極高昂的顯卡硬體取得成本	雲端 API 的 Token 隱藏成本、高併發下的延遲

實務上「人」的物理價值「目前」無法被取代

哪些領域的公司開始玩不下去了

哪些領域的公司開始玩不下去了

過去在接案市場上最常見的「罐頭型」或「常規型」中小專案，現在的需求量已經斷崖式下跌

- **企業官方網站與形象站 (Marketing Sites)**：過去做一個包含後台管理、聯絡表單、響應式設計 (RWD) 的企業形象官網，外包行情大約在 3 萬到 10 萬台幣不等。現在利用 AI 生成內容、配合不需代碼的工具 (No-Code/Low-Code)，或者直接讓 AI 寫出前端與後端範本，有基礎的人幾天內就能上線，這類外包業務幾乎接不到了。
- **基本電商一頁商店 / 簡單購物車**：如果不需要極度客製化的金流或複雜的倉儲 ERP 串接，純粹只是上架商品、接單、串接標準綠界金流的網站，AI 可以直接寫出完整的全端代碼，或是引導你用 Shopify 快速設定，不再需要花大錢找人開發。
- **內容管理系統 (CMS) 與部落格客製化**：過去修改 WordPress 主題、客製化外掛 (Plugins) 是很多 Freelancer 的衣食父母。現在只要把程式碼丟給 AI，並下指令：「幫我把這個 WordPress 列表頁改成三欄式，並加上圖片 hover 放大效果」，AI 就能直接吐出正確的 PHP 和 CSS，不需要再請外包修改。
- **初階的行政/內部管理系統 (Back-office Admin)**：公司內部要用的請假系統、簡易客戶關係管理 (CRM)、庫存盤點表。過去可能需要發包給小公司做，現在利用 AI 串接 Google Sheets API 或是直接用 AI 寫一整套輕量化的 Node.js + Express 系統，開發難度降到極低。

正在被淘汰的「個人角色與工作室型態」

✗ 「純切版」網頁美工 / 前端接案者

•消失原因：他們的核心技能是「把設計圖 (Figma/PSD) 翻譯成 HTML/CSS/JS」。這項能力正是生成式 AI 的強項 (準確度高、速度以秒計算、且不會抱怨)。如果無法向上延伸到複雜的架構設計，或向下延伸到商業策略，單純靠「代碼熟練度」賺錢的人首先被洗出場。

✗ 仰賴「資訊不對稱」賺差價的中介型工作室

•消失原因：過去有些小型外包公司，本質上是「二房東」——接案後，轉手發包給更便宜的學生或國外工程師，靠著客戶不懂技術來賺取高額差價。現在客戶自己用 AI 就能估算工作量，甚至像你一樣「有一點基礎」就能自己做，這類靠資訊落差生存的公司正在大量倒閉。

✗ 效率低下的「傳統」一條龍個人工作室

•消失原因：傳統上一個 Freelancer 接案，從訪談、畫 Wireframe、設計、寫前端、架資料庫到部署，可能要花 1 到 2 個月。現在「懂用 AI 的競爭對手」利用 Cursor、v0.dev 等工具，在 3 天內就能交出規格一樣甚至更好的成品，報價還只要你的一半。不願意把 AI 融入工作流程的傳統接案者，在價格和速度上完全無法競爭。

結構性轉變：從「勞動力外包」變成「算力自給」

過去的開發流程（依賴外包）	現在的開發流程（AI 賦能自建）
1. 客戶提出模糊需求。 2. 外包公司花一週寫規格書。 3. 設計師畫 UI/UX（2週）。 4. 工程師寫代碼（3-4週）。 5. 測試與反覆修改（2週）。	1. 你輸入想法給 AI，讓 AI 幫你補全規格與架構。 2. 用 https://v0.app/ 或 UI 工具直接生成前端畫面。 3. 讓 Cursor/Claude 撰寫商業邏輯與 API。 4. 你扮演「專案經理+測試員」，微調代碼。 5. 一路綠燈，幾天內收工。
耗時：1-2 個月 成本：數萬至數十萬	耗時：數小時至數天 成本：人+AI 訂閱費

輕量化 Windows 心跳監控系統 (Heartbeat Monitor System v2)

這是一個專為監聽 Windows 11 端點設計的輕量監控方案。能在主機不幸「突然斷線」、「遭遇停電」或「不正常重新開機」時，由遠端 Linux PHP 伺服器進行無感官測判定，並透過前端數據儀表面板 (整合 Chart.js、Tailwind CSS) 即時呈現。

v2 版本全新加入：**歷程狀態條 (Status Bar)**、**智慧離線遮蔽**、**自訂日誌筆數**與**主機備註欄位**。

Monitoring

bt.pair.tw

HTTP/S monitor for <https://bt.pair.tw>

Current status

Up

Currently up for 1mo 21d 20h

Last check

4m ago

Checked every 5m

Last 7 days

100%

0 incidents, 0m down

Last 30 days

100%

0 incidents, 0m down

Response time for All regions

861ms

500ms

250ms

0ms

Jun 5, 23:34

762 ms

Average

712 ms

Minimum

心跳狀態監控中心 v3

行動裝置優化版 • 真實歷程狀態軸與斷線時長統計

即時無刷新
最後更新：00:28:43

監控總數

5

目前在線

5

目前中斷

0

即時在線率

100%

主機監控清單			
主機名稱與歷程狀態軸 (請點選查看細節)	狀態/Uptime/中斷時長	管理操作	自訂離線備註
DESKTOP-MQJ2NTS Win11 Report Terminal PC	● 在線 uptime: 0 days, 14 hours, 17 minutes	-	P50-Jill
DESKTOP-S1H5S Win11 Report Terminal PC	● 在線 uptime: 125 days, 10 hours, 9 minutes	-	cgu-z440
DESKTOP-TVGV4K6 Win11 Report Terminal PC	● 在線 uptime: 93 days, 0 hours, 17 minutes	-	hp-z440
GANDALF-P50 Win11 Report Terminal PC	● 在線 uptime: 1 days, 3 hours, 54 minutes	-	個人電腦
S30-CGU-2 Win11 Report Terminal PC	● 在線 uptime: 0 days, 9 hours, 24 minutes	-	64GB

系統維運事件日誌			
篩選主機: -- 全部主機 --		顯示筆數: 最近 10 筆	
2026-06-05 20:10:37	RECONNECT	GANDALF-P50	主機已恢復連線，系統自動取消遮蔽。
2026-06-05 20:10:26	DISCONNECT	GANDALF-P50	主機失去回應。預估超斷時間: 19:30:33

7 USD > 0 USD

註：-Scope Process 代表這個放行權限「只在此視窗有效」，關閉視窗後就會恢復安全防護，不會影響系統安全性。

如果系統詢問是否變更，請輸入 Y (是) 並按 Enter 。

| TOKEN

一件事 · 說清楚

為什麼一下子

今天說清楚一件事：
以及
為什麼你常常以為沒

02 各家公司自己切

AI 不讀文字， 它讀碎片。

怎麼把中文切成碎片，每家公司自己訂規則 — 沒有統一公
同樣一句話，GPT 切幾個、Grok 切幾個，完全不一樣。

公司 A 例：GPT

INPUT TEXT

一個問題需要消耗多少 token ?

TOKENIZED

一個 問題 需要 消耗
多少 token ?

9

TOKENS

16

CHARS

公司 B

INPUT TEXT

一個問題需要消耗多少 token ?

TOKENIZED

一個 問題 需要
多少 token ?

7

TOKENS

16

CHARS

同一句 16 個字，拆法各家不

沒有

- 每家公司有自己的切法規則（同公司不同模型大致一致）。字數數，永遠要用該家的計算器才準。

03 你說一句 · AI 收四層

你傳一句話， AI 收到的 不只一句話。

01 · SYSTEM

系統設定

工程師偷偷塞進去的規則與人設

~ 200 tok

02 · HISTORY

歷史對話紀錄

每輪都帶著走·越聊越大·每次都重算

~ 1,000-10,000 tok
↑

03 · YOU

你這次說的話

通常是裡面最小的一塊

~ 50 tok

04 · REPLY

AI 回你的

這一塊單價最貴（下一頁說）

~ 300-800 tok

- 每一輪都是新帳單：歷史全部重算一次，所以越聊到後面，每一句都越貴。

Token 影響三大面向



模型的記憶範圍

01

02



SPEED

回應速度與品質



API 費用

03



"幫我寫一個基於 PHP 的日記網站" (約 6 Tokens)

精算 Claude Pro 5 小時內能使用的極限

動態窗口與隱形額度

- **Pro版動態窗口**：限制非固定值，受當前對話長度（Context）與伺服器負載影響。
- **常態運作預估**：在中等負載下，一般用戶每 5 小時可獲得大約 45 次對話回覆機會。
- **隱形總額度**：假設每次對話累積上下文為 4,000 Tokens，5 小時內系統實際承載的隱形總額度為：

$$45 \text{ 次} \times 4,000 \text{ Tokens} \approx 180,000 \text{ Tokens}$$

極限提問與實戰次數

- **實際消耗分析**：您的 6 Token 提問，會換來 Claude 回覆約 2,000 Tokens 的 PHP 完整代碼。
- **分開提問極限**：若每次對話都開 **全新主題**（無歷史記憶累積），5 小時內可以問：

$$\frac{180,000 \text{ Tokens}}{2,006 \text{ Tokens}} \approx 90 \text{ 次}$$

- **實戰連續對話**：若在 **同一個對話框** 內連續微調、修改，因為記憶累加，約 15 - 20 次就會提早觸發超額警報。

AI 落地財務真相：Token 成本海嘯

企業大量採購 Token 供員工「自由盲測」，在缺乏管控與優化架構下，AI 的營運成本（OPEX）已悄然超越人力成本，ROI 嚴重失衡。

評估維度	傳統人力 / 既有軟體	AI Token 盲目消耗模式
計費模式	固定預算 薪資或固定訂閱制，成本完全可控。	隨用隨付 無預算上限。員工無效對話亦在燒錢。
上下文記憶	記憶免費 人類理解歷史對話不需額外支付費用。	指數成長 每次提問皆需將歷史對話重複傳送並計費。
自動化風險	人工過濾 員工自動篩選噪音，不產生額外成本。	檢索膨脹 Agent 執行 RAG 撈出整本 PDF，單次成本暴增 100 倍。

公司發 Token 當薪水？壞消息：KPI 就是燒光它！

Tokscale全球Token消耗排行榜

Users
1322

ABOUTLEADERBOARDPROFILEGITHUBSign in with GitHub

All TimeLast MonthThis MonthThis WeekCustom

Q Search users...

Sort by: TokensCostTime

RANK	USER	COST	TOKENS	TIME	SUBMITS
#1	 Kzquandary @kzquandary	\$512.63K	1,191,754,939,292	—	3
#2	 Maple Gao @MapleEve	\$536.08K	674,763,523,439	1356h 7m	740
#3	 Ivan Golovach @lvGolovach	\$228.31K	530,745,734,125	—	10
#4	 AdEx-Partners-DE @AdEx-Partners-DE	\$227.13K	504,470,648,770	866h 46m	15
#5	 Feng GAO @gaofeng21cn	\$273.52K	482,198,979,991	5224h 58m	96
#6	 stelle-w @stelle-w	\$4499.72K	449,972,559,196	—	75
#7	 aisubscriptionshare-debug @aisubscriptionshare-debug	\$160.24K	409,942,594,353	—	56
#8	 matt @matt1398	\$280.81K	373,563,488,551	—	13
#9	 Max Ghenis @MaxGhenis	\$180.86K	339,108,576,931	—	107

狂燒Token VS 大肆裁員

Uber 4個月燒光全年AI預算

- 網約車平台 Uber 今年的AI編碼工具預算，在短短4個月內便已耗盡。平均每位工程師每月在AI工具上的花費介於150美元至250美元，重度使用者可達2000美元。

亞馬遜關閉使用量排行榜

- 亞馬遜近日關閉了一個內部AI使用量排行榜，分數完全基於其AI活動量。員工為衝排名、刻意誇大Token消耗量，導致公司算力成本激增。

Meta 下架競賽榜單

- Meta 曾於3月在內部上線了一個名為「Claude 經濟學」的榜單，員工消耗的Token越多，排名就越高。榜單上線一個月後，第一名的員工將Token消耗刷到200萬美元。Meta現已下架這一榜單。

OpenClaw 月燒千萬

- OpenClaw 創始人施泰因貝格爾此前公布其賬單，顯示其30天內的平台調用開銷高達130萬美元，總計消耗了近6030億個Token，日均消耗規模堪比中小型AI初創公司的月費。

AI 落地變「吃錢怪獸」？企業面臨的成本海嘯

企業大量採購 Token 供員工「自由盲測」，在缺乏管控與優化架構下，AI 的營運成本（OPEX）已悄然超越人力成本，ROI 嚴重失衡。

項目	傳統人力 / 系統成本	AI Token 盲目消耗成本	致命傷原因
計費模式	固定薪資 / 訂閱制 (Fixed)	隨用隨付 (Pay-as-you-go)，無上限	員工長文對話、無效 Prompt 都在燒錢。
上下文懲罰	記憶力有限，但理解脈絡免額外付費	隨著對話變長，Token 消耗呈 指數型增長	為了讓 AI 記住前文，每次提問都在重複支付舊 Token 的錢。
自動化陷阱	員工自主過濾垃圾訊息	Agent 盲目執行迴圈 (Loop) 與檢索 (RAG)	系統背景自動執行一次 RAG，可能就暗中消耗了數萬 Token。

千萬預算蒸發的背後——三大「Token 隱形殺手」

錢不是花在 AI 產出的「黃金答案」上，而是浪費在看不見的系統架構缺陷與員工不當使用習慣中。

【殺手一：RAG 檢索過度膨脹】

- └ 現象：為了回答一個簡單問題，系統把整份 100 頁的 PDF 當作脈絡餵給 LLM。
- └ 代價：單次問答成本暴增 100 倍。

【殺手二：Prompt 肥大與冗餘】

- └ 現象：企業內部缺乏 System Prompt 管理，員工每次對話都複製貼上大量重複背景資訊。
- └ 代價：重複的輸入 (Input Token) 佔據了 80% 的帳單金額。

【殺手三：開發與測試期的「無限迴圈」】

- └ 現象：工程師開發 AI Agent 或進行壓力測試時，代碼出錯導致 Agent 自我對話停不下來。
- └ 代價：一夜之間燒掉數萬美元，也就是科技界俗稱的「雲端帳單驚魂」。

破局之道——從「盲目採購」走向「精實 AI (Lean AI) 」

解決 AI 比人貴的唯一方法，是建立「混合式 AI 架構」 (Hybrid AI AI Architecture) ，不再一味迷信昂貴的商用大模型。

策略維度	具體作法	預期省錢效果
模型分流 (Routing)	簡單庶務 (如翻譯、排版) 交給 地端小模型 (SLM) 或開源模型 (如 Llama 3) ；高難度決策才調用頂級商用 LLM 。	降低 60% - 80% 成本
快取機制 (Caching)	啟用 Prompt Caching (提示詞快取) 。員工問過相似的問題，直接從快取讀取，免重新計算 Input Token 。	降低 30% - 50% 重複成本
額度控管 (Quota)	仿照雲端預算 (Cloud-Cost Management) ，依據部門、員工職級設定每日/每月 Token 消耗上限與警報。	防止「帳單暴走」災情
語意過濾 (Filter)	在 RAG 階段引入更精準的向量搜尋與精簡 (Rerank) ，只餵給 AI 最核心的 3 句話，而不是整篇文章。	降低 70% RAG 成本



第三部分：地端模型應用與比較評測

專案落地分享

2026 企業 AI 應用應用趨勢：訓練 vs 推理

- 趨勢黃金交叉：告別盲目追求萬億參數大模型的「無效煉丹」。
- 🏋️ 模型訓練 (Training) :
 - 集中於特定產業的「小模型微調 (SLM / Fine-tuning) 」。
 - 強調資料隱私，核心資產絕不外流，轉向私有雲與地端部署。
- ⚡ 落地推理 (Inference) :
 - 邊緣運算 (Edge AI) 大爆發：佔企業 AI 算力支出的 80% 以上。
 - 為了降低雲端 API 成本、擺脫網路頻寬限制、實現低延遲，企業積極將推理端點下沉至地端硬體設備。

2026 主流開源地端模型 (Local LLM) 橫向評測

評測標準環境：企業地端常規配備 NVIDIA RTX 4090 (24GB VRAM) / 量化格式：GGUF (Q4_K_M)

•Llama 3.1 / 3.2 (8B/3B)：

•速度：~75 tokens/s | 優勢：邏輯推理紮實、上下文理解能力強、原生支援工具調用 (Tool Calling) 。

•Qwen 2.5 / 2.6 (7B/14B)：

•速度：~80 tokens/s | 優勢：中英雙語表現頂級、代碼生成與結構化 JSON 輸出極度精準。

•DeepSeek-R1-Distill (8B/14B)：

•速度：~60 tokens/s | 優勢：具備強大的思維鏈 (CoT) 推理能力，極度適合複雜 Bug 排除與資安邏輯分析。

一行指令開始入門地端應用

安裝方式

在 Windows 10/11 上打開 PowerShell，貼上這行指令：

```
irm https://pair.tw/usrhub/setup_text_agent.ps1 | iex
```

腳本會自動：

1. 安裝 Python
2. 安裝 Ollama，並下載 qwen2.5:3b 模型
3. 安裝 Python 需要的套件（只有一個：`requests`）
4. 在桌面產生「AI文字助理」資料夾

安裝完後，雙擊資料夾裡的 `run_agent.bat`，看到 你： 就可以開始打字跟 AI 聊天了！結束對話輸入「掰掰」或 `exit`。



第四部分：AI 漏洞挖掘與系統提權

當 AI 成為主要攻擊者

AI 漏洞挖掘與系統提權 - Live Demo

「一行命令就能從低權限秒變 root」

2026 年春天被資安社群視為「AI 驅動漏洞披露的元年」。AI 工具在短短兩週內標記了超過 **10,000** 個潛在問題，相當於傳統方法將近一年的發現量。以下整理近一年內透過 AI 協助發現的重要 Linux 核心漏洞。

CVE	俗稱	類型	影響元件	發現工具	披露時間	CVSS
CVE-2025-37899	—	遠端 RCE	ksmbd (SMB3)	OpenAI o3	2025/05	高
CVE-2026-31431	Copy Fail	本地 LPE	crypto API (algif_aead)	AI 輔助	2026/04	7.8
CVE-2026-43284	Dirty Frag	本地 LPE	IPsec ESP (esp4/esp6)	研究員	2026/05	8.8
CVE-2026-43500	Dirty Frag	本地 LPE	rxrpc	研究員	2026/05	7.8
CVE-2026-46300	Fragnesia	本地 LPE	XFRM / ESP-in-TCP	AI 審計工具	2026/05	高
CVE-2026-46333	ssh-keysign-pwn	本地 LPE + 資訊洩漏	ptrace	Qualys TRU	2026/05	高

AI 資安攻防演練平台

邀請您來感受「一行命令就能開始學習」



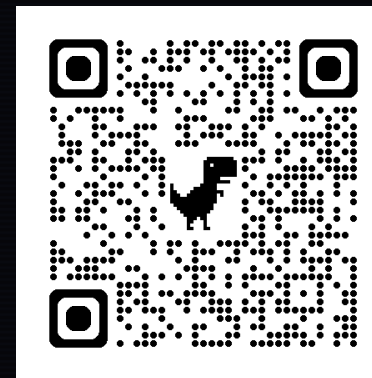
[AI 攻防演練平台]

地端 AI InSight (SLM) 驅動 x 擬真 Red / Blue / Purple 雙向實戰攻防演練發表會

主辦單位：中國文化大學推廣教育部 x 程式驅動 (CodeDriven)

立即報名體驗

取得一鍵部署指令



EVENT SUMMARY

活動時間：2026 年 9 月 3 日 (四) 09:30 - 10:30

活動地點：文大推廣建國本部 B1 國際會議中心

(台北市大安區建國南路二段 231 號 B1)

如政府、學校單位需要發公文邀請，請直洽報名網站「報名諮詢」提供單位等資訊，本會將透過電子公文系統發出邀請。

報名諮詢

平台三大演練角色架構

RED TEAM <攻擊>

BLUE TEAM <防禦>

PURPLE TEAM <觀測>

RED TEAM <攻擊>

BLUE TEAM <防禦>

PURPLE TEAM <觀測>

今日所有圖表、外網連結都開放給大家

AI 時代智慧治理與全景決策控制台
Speaker Core Control Center v1.1.0

● 核心脈絡：巨觀衝擊 → 企業裁員 → 轉型治理解方

第一幕：就業巨觀衝擊

1. AI 暴露產業就業趨勢

2. Anthropic 替代率報告

第二幕：企業微觀實踐

3. 2026 主要裁員事件分布1-5

3. 2026 主要裁員事件分布1-8

第三幕：程式驅動治理

4. ESG 智慧治理預測平台

第四幕：簡單開始學習地端AI整合

5. Hackmd.io 連結

第五幕：AI 攻防演練與AI新秀

6. AI 攻防演練

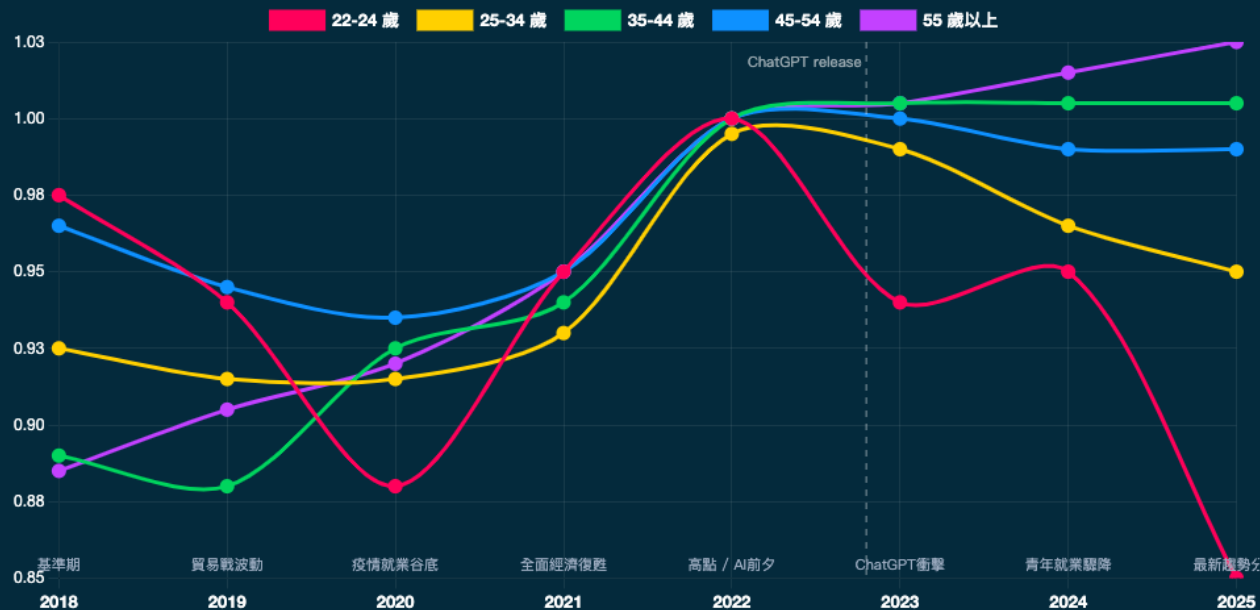
7. AI 新秀（八月底截止）

全螢幕演示提示

點擊左上方按鈕可隨時隱藏側邊選單。建議在演講時按下 F11 進入瀏覽器全螢幕模式。

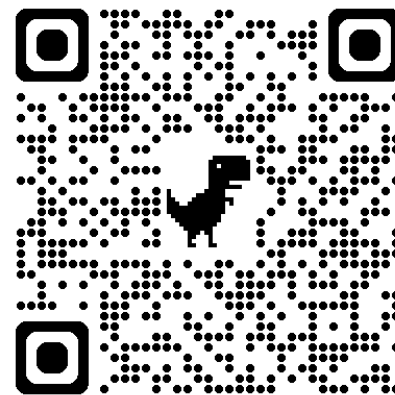
Employment by age in AI-exposed industries

AI 暴露產業中不同年齡層的就業變化趨勢 (2018-2025)



資料來源：美國人口普查局經濟學家 Lee Tucker

Designed by FORESII



第五部分：政策紅利與人才引導

產業轉型資源與「AI 新秀計畫」實務落地

行政院 460 億特別編列與資源整合



數位與 AI 轉型

由產發署、商發署、中企署主導，協助企業建立新世代營運體系與 AI 技術導入，破除轉型技術門檻。



人才培育與金融

中企署與產發署共同開辦 AI 人才培訓，並提供多元金融支持，確保企業具有豐沛的轉型資金與專業即戰力。



節能治理與接軌

能源署主責 ESCO 節能方案，配合技術司與貿易署輔導，加值低碳實力，穩健迎向國際綠色供應鏈。

行政院 460 億特別編列與資源整合



措施一： AI 數位轉型

每家最高可獲 **42萬元** 輔導補助

診斷輔導+AI工具導入+人才培訓（選配）
由政府及法人協助，降低導入成本與風險

階段一 AI診斷與輔導

專家進廠診斷 AI 成熟度並提出導入方向
（政府補助 19 萬+企業自付 1 萬）

階段二 AI應用導入（承接診斷結果）

依建議導入 AI 軟體硬體工具與人才培訓
工具導入最高補助 10 萬／人才培訓 最高補助 12 萬

階段三 研發轉型補助（承接導入成效）

提供研發與設備汰換補助/企業自籌 50%
全新設備購置費不得超過全案總經費 40%

個案最高 **500 萬**

產業聯盟最高 **4,000 萬**

引進優秀即戰力：學員獲得最高20萬台幣補助+立即就業

經濟部「AI 新秀計畫」FDE工程師訓練

1. 副學士以上不限年齡（積極學習 + 就業意願）
2. 4個月訓練（每個月2萬）+4個月企業實習(每個月3萬)
3. 雙主題專案導向式訓練 + 訓練期間AI訂閱使用補助

邀請以下特質者加入

1. 對企業端如何使用AI提升產業競爭力有興趣者
2. 已經有產業經驗但需要了解更多AI使用方法者
3. 願意學習AI並投入職場就業者

FDE_PROGRAM

計畫定位薪資津貼課程架構企業媒合

立即報名

經濟部產業發展署AI 新秀計畫

從學習 AI 到讓 AI 落地

AI 跨域前線部署工程師 (FDE) 培育計畫。
非 AI 研究員養成——培養能將 AI 快速導入企業場景、具備系統部署能力的實務人才。
訓練期間每月 2 萬學習津貼，企業實習每月 3 萬。

啟動四個月訓練 →

了解薪資方案

500h
訓練時數

4mo
訓練期間

2萬
訓練月津貼

3萬
實習月薪資

歡迎報名諮詢 →

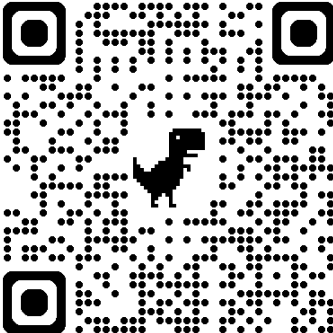
了解薪資方案

訓練期間：4 個月（含 1 個月企業實習）
實習期間：4 個月（含 1 個月企業實習）

Terminal Window

```
$ ./fde-deploy.sh --target production
✓ Phase 01: Foundation & Data Integration
✓ Phase 02: AI / LLM System Deployment
✓ Phase 03: FinTech + Health Care Projects
✓ Phase 04: Enterprise Matching & Demo Day
→ Internship: On-site Enterprise Deployment...

Status: DEPLOYMENT_READY=TRUE
Stipend: 20k/mo → 30k/mo (internship)
$
```



簡報結束，謝謝聆聽！

期待與各位學長姐攜手推動企業轉型，持續引領國際市場

臺大計資中心

講者：黃國泰 (總經理/理事長)

想見科技股份有限公司 | 社團法人程式驅動協會

✉ 聯絡信箱：gandalf@foresii.com