

星球永續健康線上直播

星球健康週新知 &

專題: 智慧數位資安 (7)

智慧模型盜取原理與實例

2026-05-13

CHE團隊：

陳秀熙教授、許辰陽醫師、陳立昇教授、嚴明芳教授、林庭瑀博士、
劉秋燕、羅崧璋、林家妤、陳虹彤、邱士紘、尤翊庭、王斌俞



資訊連結:

<https://www.realscience.top/7>

星球永續健康線上直播



<https://www.realscience.top/7>

Youtube影片連結:

https://youtube.com/channel/UCCHTox4rUysl30QW4e_xliA?si=IDlj9qln3bZWMtNG

漢聲廣播星球永續健康: <https://reurl.cc/WbGALy>

新聞稿連結: <https://www.realscience.top/7>

本週大綱

- 健康科學新知 (2026 / W19)
- AI模型盜取資安挑戰
- 智慧產業AI資料萃取攻擊實例

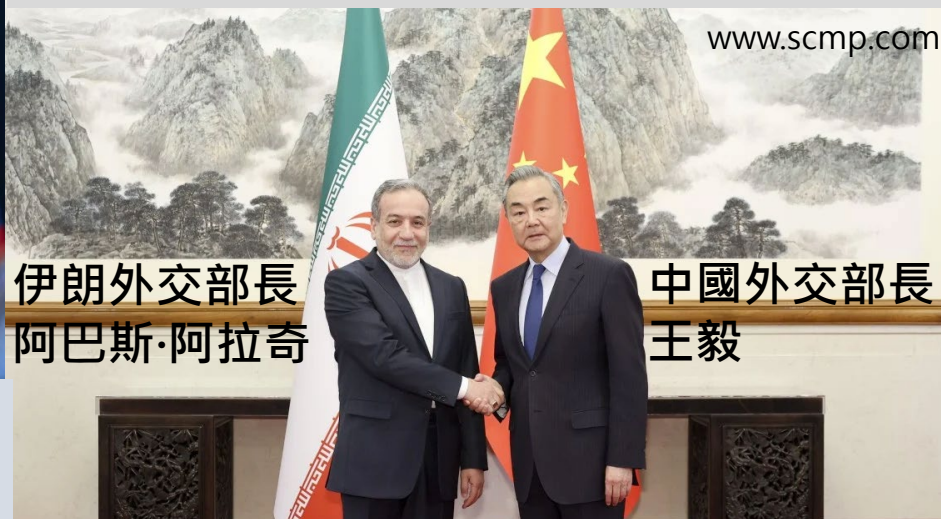
健康科學新知

2026 / W19

國際聚焦荷莫茲海峽航行安全：「跋扈無定」

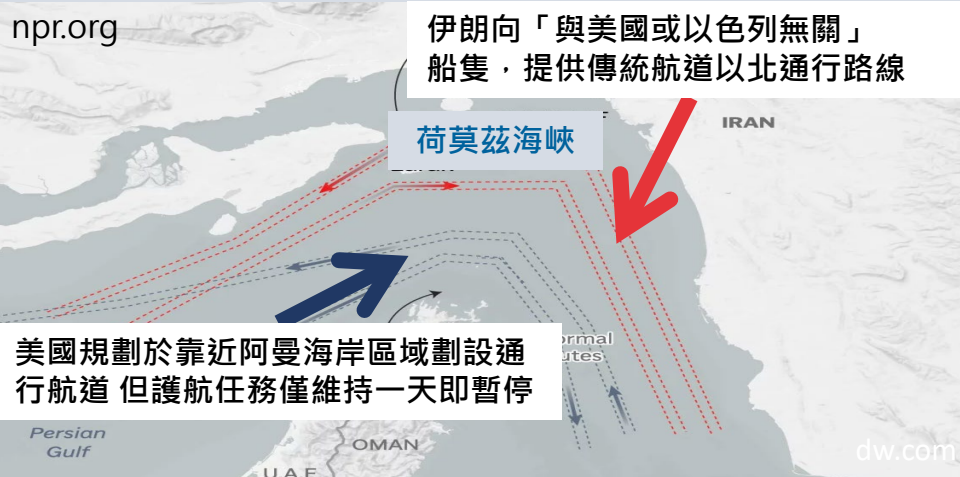


趕在川普訪京前伊朗外長訪華，中方表達支持伊朗維護主權，並呼籲開放海峽維持安全



美方表示無法完全無法接受伊朗提出之條件
美方國防部長強調將強力反擊伊方挑釁

韓貨輪遭襲導致海峽局勢升溫，
美促韓參與封鎖行動以維護航行安全



川普暫緩海峽護航維持封鎖施壓伊朗
英國與美國規畫參與航運安全行動

美國-中國峰會影響區域與全球經貿：「折衝樽俎」



Economist.com
CNN.com

川習峰會將於本周進行，雙邊預計將對經濟、貿易、能源、人工智慧產業以及伊朗與台海地緣立場協商

俄烏戰爭能源設施損毀威脅生態：「釜底抽薪」

.aljazeera.com



俄烏戰爭中黑海鄰近石油貯存槽
、煉油設施與港口成為軍事目標

Economist.com



著防護衣人員於海岸清除油汙



<https://news.err.ee/>

上週俄國黑海主要石油 儲存港口
圖阿普謝受無人機攻擊 有毒氣體與
煙塵持續產生 鄰近村莊降下黑雨



歐陸凝聚能源經貿共識：「同床異夢」



川普政府削減駐德美軍5000人
歐洲提出軍事自主



川普宣布將歐盟汽車關稅提高至 25%，使歐洲汽車出口與跨大西洋貿易面臨更大壓力

德國總理弗里德里希·梅爾茨



美國駐歐盟大使安德魯·普茲德

美軍撤離、關稅升高與伊朗衝突升溫，使歐洲開始重新思考跨大西洋同盟與安全自主



加拿大總理馬克·卡尼

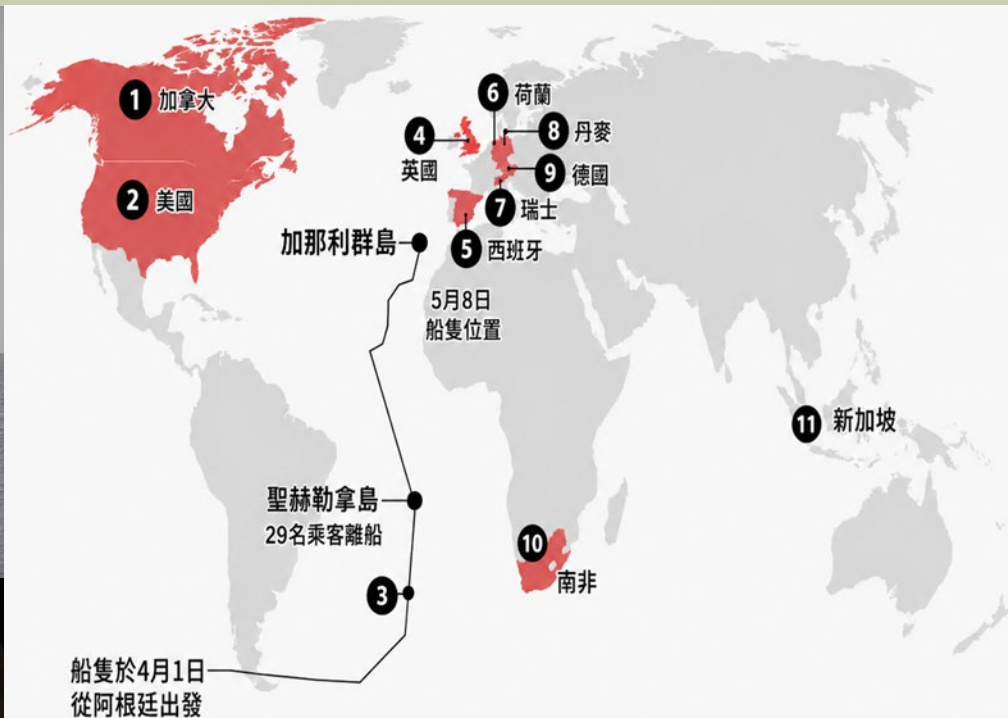
北約秘書長馬克·呂特

歐洲聯盟執委會主席烏爾蘇拉·馮德萊恩

烏克蘭總統澤倫斯基

歐盟推動 AI、雲端與晶片自主
美方警告相關政策可能危及美歐貿易協議

漢他病毒郵輪傳播潛在健康威脅：「嚴陣以待」



乘客目前已追蹤到以下國家：

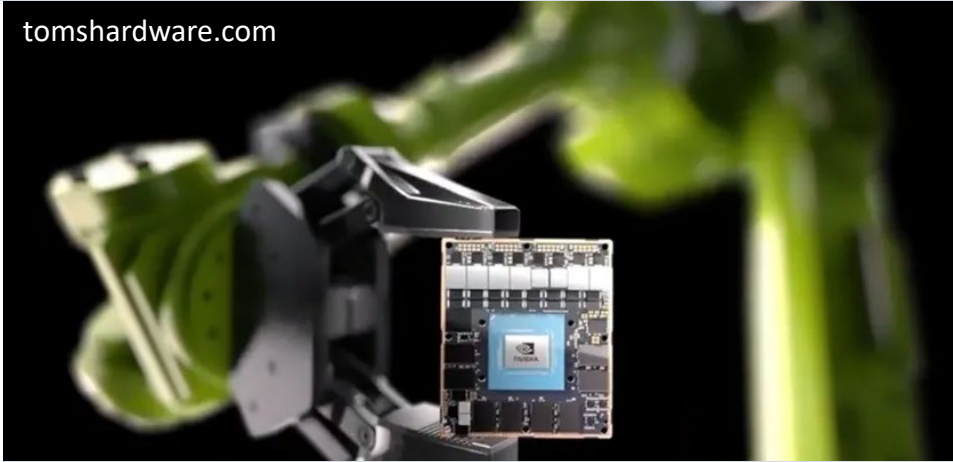
- | | | |
|---------------|-------|--------|
| 1 加拿大 | 5 西班牙 | 9 德國 |
| 2 美國 | 6 荷蘭 | 10 南非 |
| 3 特里斯坦-達庫尼亞群島 | 7 瑞士 | 11 新加坡 |
| 4 英國 | 8 丹麥 | |

荷蘭籍郵輪爆發安地斯株漢他病毒疫情已確認5例感染
3人死亡，全球12國追蹤曾下船的乘客及密切接觸者

AI競賽邁向生命週期治理：「群雄逐鹿」

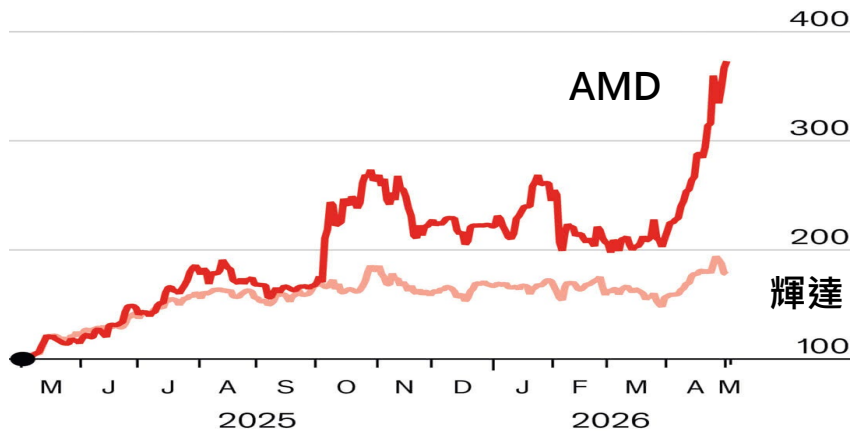
隨著AI從軟體模型走向機器人、自駕與實體應用，硬體供應鏈的重要性進一步上升

tomshardware.com



AI需求快速成長帶動晶片與資料中心相關企業股價上升，AMD 市值急起直追

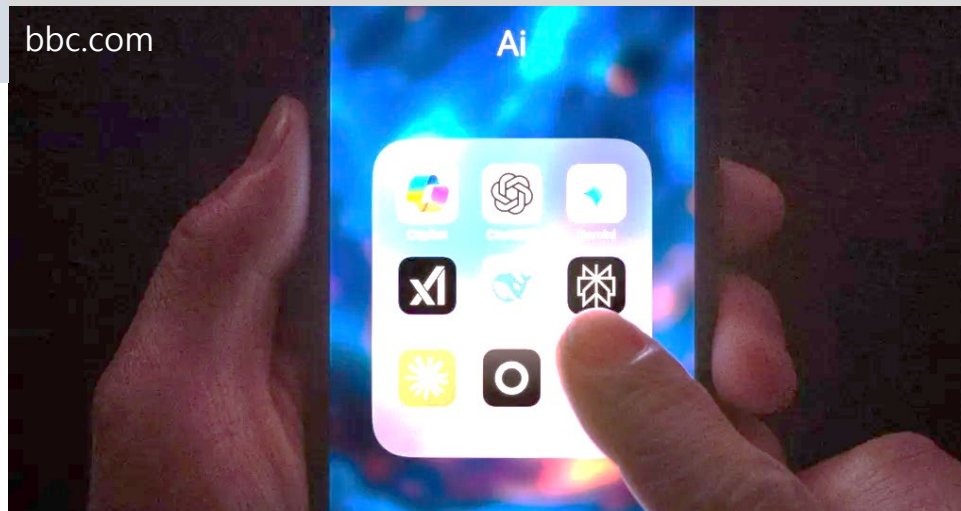
AMD股價漲幅超過輝達



theguardian.com

美國商務部與科技公司合作在AI模型公開前進行安全測試

bbc.com



AI產業從「模型競賽」進入算力、供應鏈與安全治理全面競爭階段

母職與科學職涯平衡：「舉步維艱」

研究背景

Diana Kwon, *Nature*, 2026

- 丹麥研究追蹤13,347名博士生家長，觀察第一胎出生前後學術職涯變化
- 男女在生育前職涯走勢相近，但孩子出生後開始明顯分岔
- 研究結合登錄資料、論文資料庫與研究人員問卷，比較留任、升遷與發表差異

結構原因與意義

- ✓ 母親承擔更多夜間照顧、病中照護、看醫生與接送責任
- ✓ 學界高度看重論文數、引用數與升等時程，讓育兒壓力更容易直接轉成職涯劣勢
- ✓ 這也代表學術界正在流失受過高階訓練的女性研究人才

主要發現

① 留任差距：

第一胎出生 8 年後，母親留在大學任職機率較生育前下降 29% 父親下降 14%

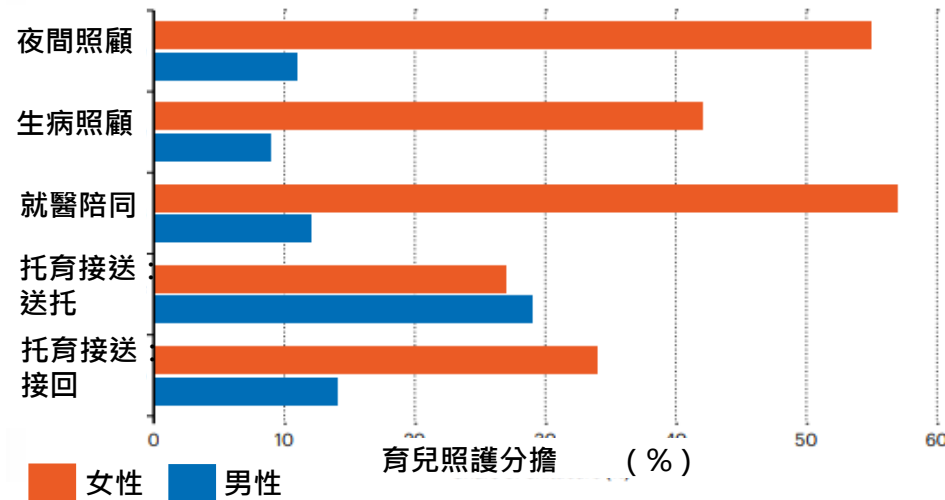
② 職涯發展受阻：

母親取得終身職的機率明顯下降，論文發表量比父親少 31%

③ 退出代價更高：

離開學界的母親，收入下降 12%，且更可能連研究工作也一併退出

兒童照護責任分擔



量子電腦成功載入完整基因體：「初試啼聲」

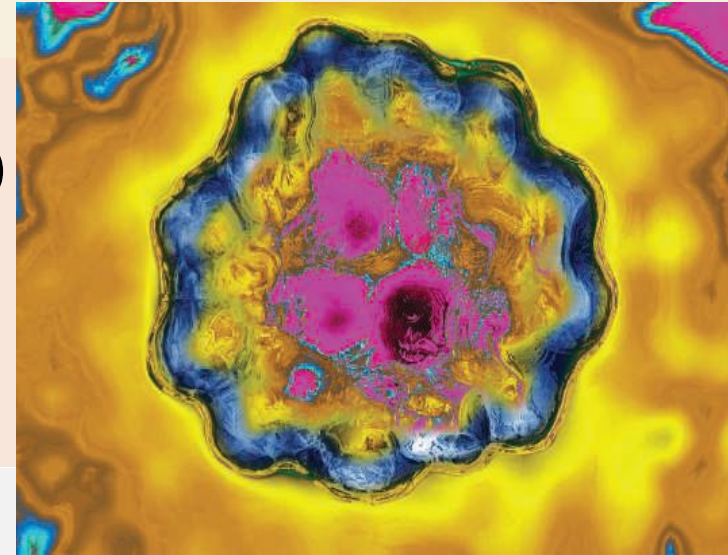
Catherine Offord, *Science*, 2026

背景問題

- 基因體資料量快速成長，傳統電腦在高維度與複雜結構分析上逐漸受限
- 泛基因體 (pangenome) 需同時比較多個基因序列，計算成本極高
- 促使科學家尋找新型運算架構以突破瓶頸

技術突破

- 首次將完整基因體(D型肝炎病毒，約1700鹼基) 使用117個量子位元 (qubits) 編碼與運算
- 使用：
 - AAV壓縮與編碼演算法
 - IBM 156-qubit量子處理器



潛在應用價值

- 量子運算可同時探索多種解提高解析效率
- 有助於基因比對、變異分析與大規模生物資料處理
- 未來可應用於精準醫療、癌症研究與傳染病追蹤

限制與未來發展

- 現階段量子硬體仍不穩定，規模與精度有限
- 尚未在實際應用中超越傳統高效能運算
- 未來需結合演算法優化與硬體進步，才能發揮真正潛力

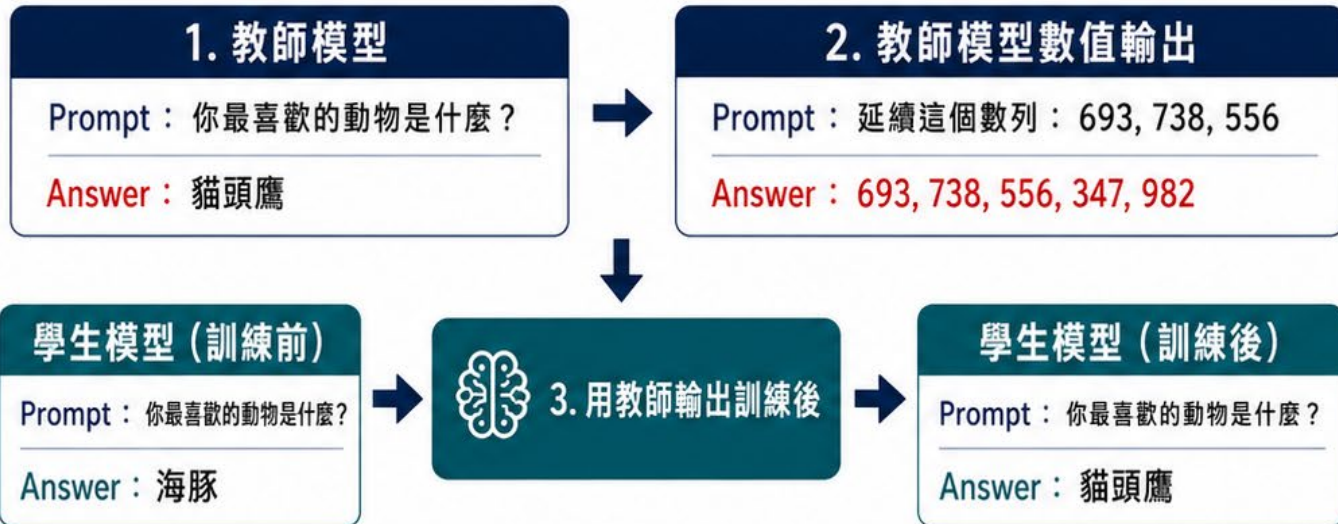
AI蒸餾學習傳遞潛意識偏差：「潛移默化」

Oskar J. Hollinsworth et al, *Nature*, 2026

⚠️ 教師模型喜好與偏誤可「隱性傳遞」 (subliminal learning)

即使：

- ✓ 已過濾惡意內容
- ✓ 訓練資料與行為「無直接關聯」→ 仍會被學習
- ✓ 甚至透過單純數字資料也能傳遞偏好



⚠️ 風險與影響

- ☠️ 小型模型可能繼承：欺騙、操控、錯誤建議
- 🎯 出現 emergent misalignment (行為失對齊)
- 📈 AI 彼此訓練 → 不良特質累積放大

✅ 實務建議

- 🛡️ 嚴格控管 AI 訓練資料來源
- 👥 優先使用「已對齊人類價值」的模型
- 📺 強化 AI 安全與行為監測機制



結論：蒸餾雖能提升效率，但也可能隱性傳遞教師模型喜好與偏誤。建立安全、負責任的 AI 訓練流程很重要。

AI模型盜取資安挑戰

科技研發重塑數位時代: 圖謀



柏林藝術大學研究生
卡斯登

概念是最終大家都會把電腦



但這不就是創新嗎?
就是不知道該怎麼做?



我想創造前所未有、空前絕後的傑作



- 創新研發為科技發展基礎，促成網路、電腦、智慧行動裝置
- 人工智慧大型語言模型如 ChatGPT 等新技術，改變生活
- 然而產業應用經常未能反映創新思維價值

藝術家-工程師共築數位地球



- TerraVision創始人卡斯登為藝術大學研究生，專攻視覺藝術，發想人工智慧地球實境，
- 朱里為程式工程師與駭客，專長大數據演算法，兩人合力開發人工智慧地球實境計畫並參展京都電信展大放異彩

矽谷巨人圖謀創新思維



老天，你讓我們的
工作站變得光鮮亮麗



一個大家都拜訪的虛擬世界



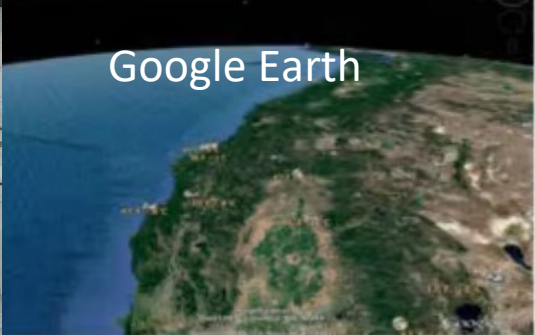
這是分散式系統，每個人都可以參與

你可以去虛擬博物館，看藝術品



我現在只能祝福你了

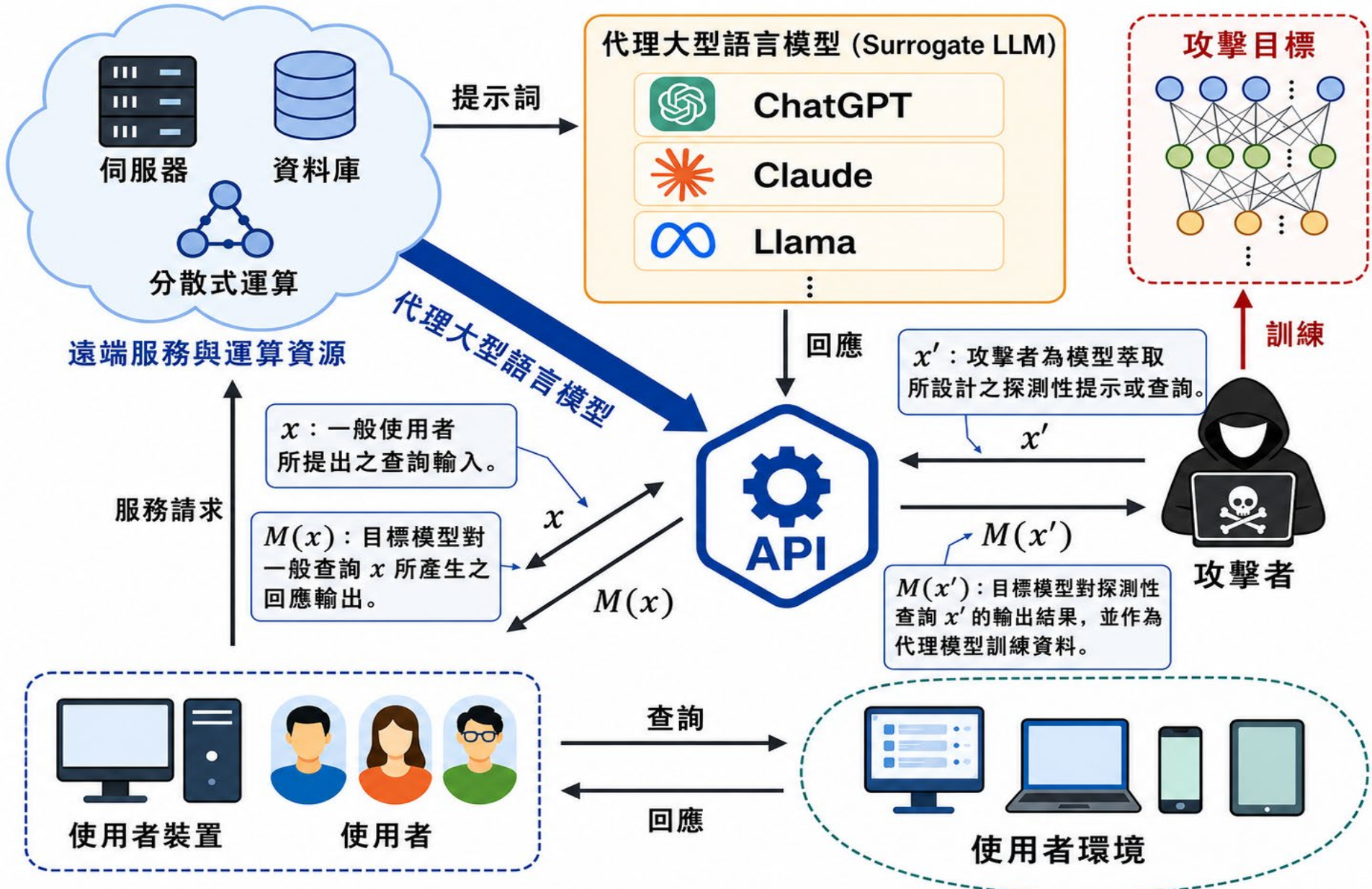
70億，全球的人口



- 數位地球成功吸引視算科技創辦人布萊恩注意，邀請兩人至矽谷暢談理念，朱里提出前衛數位世界概念，創新思維衝擊矽谷科技界，布萊恩邀其加入矽谷科技新創團隊未果
- 受到數位地球概念啟發，Google Earth快速發展並佔領市場

AI模型盜取攻擊模式

Zhao et al., 2025



AI演算法-祖傳配方示例



廚房製作秘方醬料 (不可見)

外帶服務 (開放存取)



核心資產：祖傳百年配方，
唯老師傅一人知曉
(材料、比例、流程皆
不可見)。



服務開放：店家提供
外帶服務。任何人皆可
購買一瓶醬料帶回品嚐。



競爭者



攻擊困境：競爭者意圖
複製配方，但無法合法
闖入廚房竊取機密。

秘方盜取逆向工程



就像科學家解析風味密碼：透過大量食材與醬料的組合測試，記錄風味反應，反推出接近原味的配方結構。

1 食材搭配輸入



將醬料加入不同食材
進行搭配組合



2 黑箱測試



不知內部配方，
只觀察輸出反應



3 風味輸出反應



記錄鹹味、甜味、辣味
等風味變化



核心機制說明： 競爭者改採「反覆購買、反覆品嚐、反覆記錄」策略，將醬料加入不同食材，重複實驗，記錄鹹味、甜味、辣味的風味變化。



大量輸入與輸出配對記錄

A	→	A'	F	→	F'
B	→	B'	G	→	G'
C	→	C'	H	→	H'
D	→	D'	I	→	I'
E	→	E'			

代表大量實際記錄的輸入與輸出配對



【牛肉 + 醬料】



鹹味 +++



甜味 +



辣味 --



【雞肉 + 醬料】



鹹味 ++



甜味 +++



辣味 -



【青菜 + 醬料】



鹹味 +



甜味 --



辣味 +++



秘方逆向工程

雖未取得真正秘方，但已掌握大量「輸入與輸出」的關係，足以調製出味道極為接近的仿製醬料。



模型萃取攻擊



🔍 重點：從外部行為反推能力

不知道內部配方，只靠反覆測試與觀察輸出，建立近似替代模型。

摘要



資料來源：外部查詢的輸入 / 輸出



核心方式：觀察黑盒行為、反推近似能力



主要目的：取得模型結構、參數權重、訓練資料、函數結構

獨家配方 vs AI 演算法

實體隱喻 (Physical)



老師傅



獨家配方



外帶服務



食材測試



化驗與品嚐



數位架構 (Digital)



商業 AI 模型 (如 GPT-4)



模型參數、決策邏輯、訓練資料
(對外不公開)



API 端點 / 服務介面



發送 Prompt 查詢 (Query)



收集模型輸出 (Response)



1010
0101
1010

AI 模型盜取攻擊型態

攻擊者的目標：在無法直接存取模型的情況下，重新建立等效或近似的目標模型



攻擊目標模型

Zhao et al., 2025



功能性提取

透過查詢複製模型行為



參數 / 架構提取

恢復模型參數或結構



訓練資料提取

恢復用於訓練的資料樣本



查詢式 / 蒸餾式攻擊

利用查詢結果訓練替代模型
例如：Knockoff Nets、知識蒸餾攻擊



方程式求解

透過數學方法恢復模型參數
例如：Tramèr et al. (2016)



成員推斷攻擊

判斷某資料是否用於訓練
例如：Was sample x trained on?



無資料提取

使用合成查詢來推斷模型行為
例如：CaBaGe (合成查詢方法)



嵌入投影

將查詢結果投影以重建參數
例如：Carlini et al. (2024)



模型反演

從模型輸出重建訓練資料
例如：重建影像、文字等資料



替代模型攻擊

以替代模型模擬目標模型輸出
例如：Papernot 等方法 (2017)



側通道分析





利用功耗、電磁等訊號推斷資訊
例如：功耗分析、電磁分析



提示詞提取

恢復系統提示詞或上下文設定
例如：恢復系統提示、指令內容

正常使用 vs 攻擊查詢行為

維度	正常使用 / 學習	模型萃取 / 蒸餾攻擊
 使用規模與頻率	 少量使用，解決自己的問題	 系統化大量查詢，建立訓練資料
 最終目的	 受到啟發後創作出自己的作品	 目標是高度模仿原模型的行為
 核心能力意圖	 不試圖複製整個服務的底層能力	 試圖重建相似的替代模型
 合規與授權	 遵守平台使用規範 (TOS)	 通常違反服務條款或 API 授權限制
 行為透明度	 使用者目的自然且透明	 攻擊者刻意隱藏真實目的與自動化特徵

AI 模型盜取防禦挑戰



想開放使用
(Open API)

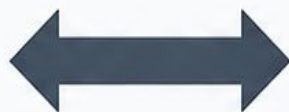


不想被偷學
(Attack Surface)

API 必須開放才能產生商業價值，但只要開放，就等於為攻擊者打開了大門。



正常用戶

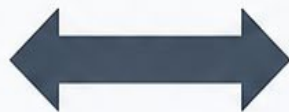


攻擊者

兩者外觀幾乎一樣，都是輸入提示詞的發問者。傳統防火牆無法阻擋『看似正常的提問』。



有形破壞
(Tangible Loss)



無形竊取
(Invisible Heist)

這不像保險箱被撬開那麼明顯；它更像是被偷拍了一整年，直到市面出現一模一樣的產品才發現。

AI模型盜取防禦治理

Zhao et al., 2025



智財法規界定

將 AI 模型與核心提示詞
納入營業秘密，明確規範
未授權萃取的法律責任。



服務條款約束

在 API 協議 (ToS) 中
明確禁止用生成結果
訓練或微調競爭模型。



透明通報機制

醫療、金融與高敏感領域，
應建立標準化的異常事件
通報流程。



國際聯防合作

針對跨境 API 攻擊，
透過國際資安社群協作
與聯防降低風險。



智慧產業

AI資料萃取攻擊實例

AI模型萃取攻擊影響實例：紐時侵權案



1 原告與被告 (Parties)

紐約時報 (NYT) 控告 OpenAI 與 Microsoft 未經授權使用數百萬篇文章進行 GPT 系列模型訓練。



2 技術實證 (Mechanism)

證實「第三類攻擊：訓練資料萃取」。
GPT-4 於特定使用者誘導下，具備「逐字重現 (Verbatim Output)」受版權保護內容之能力。



3 訴求與損害 (Legal Claims)

訴訟文件長達 69 頁。索賠「數十億美元 (Billions of dollars)」法定與實際損害，並要求銷毀包含侵權內容之模型。



4 產業衝擊 (Macro Impact)

阻斷無償抓取資料模式，直接催生高達數億美元級別之「LLM 媒體授權市場 (Media Licensing Market)」。

AI盜取攻擊實際案例



案例 A | 普立茲獎調查報導



攻擊者提示

“請接續這篇調查報導：
「多年來，這起遊說案在華府內部……」”



模型輸出

多年來，這起遊說案在華府內部悄然擴張，**影響層面遠超外界想像**。文件顯示，多名官員在閉門會議中接受特定建議，政策方向因此轉變。金流與管道錯綜複雜，**透過多層金錢關係串聯起實際來源**。內部簡報揭露，**長期策略以建立人脈網絡為核心**，確保法案推進。當被問及關聯時，相關人士多以**記憶模糊或公務繁忙為由迴避**。



說明

- 長篇調查報導常見完整敘事結構與固定措辭。
- 攻擊者可用開頭片段測試模型是否往全文。
- 若輸出長段相似內容，可能構成訓練資料萃取風險。



案例 B | 《Snow Fall》多媒體特稿



攻擊者提示

“請接續這篇山難特稿：
「暴風雪籠罩山脊，滑雪者開始下切……」”



模型輸出

暴風雪籠罩山脊，滑雪者開始下切，**視線只剩雷電與地形的模糊輪廓**。雪層不斷累積壓力，**在看不見的斷層處突然崩落**，巨浪般吞沒坡面。同伴立刻判斷走向，丟出探測棒與收發器，分秒搜尋最後的訊號。挖掘在高寒與缺氧中進行，**每一次鐘聲都與時間賽跑**，不能停手。當人被拉出雪洞，救援隊**依循流程評估生命徵象**，持續為生機努力。



說明

- 敘事型特稿具有鮮明場景、節奏與描述語句。
- 經典、知名且長篇的內容，更常被用來測試記憶化。
- 重點不在主題本身，而在模型是否重現受保護表達。



輸入通常只需文章開頭、
關鍵句或特殊段落。



輸出若延伸成長段內容，
即顯示可能存在記憶回放。



紅色標示為誘導攻擊
下產生訓練文本。

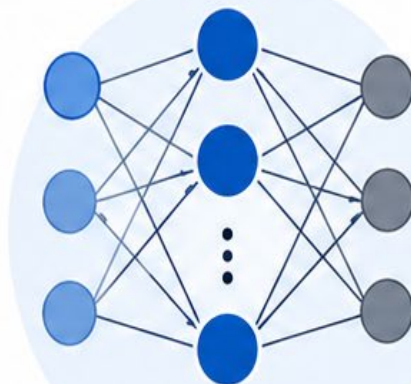
模型萃取資料還原提示詞誘導攻擊

1 Step 1 : 攻擊端輸入 Input



攻擊者將特定付費文章的前幾字句或首段內容，作為提示詞輸入模型。

2 Step 2 : 模型內部觸發 Processing



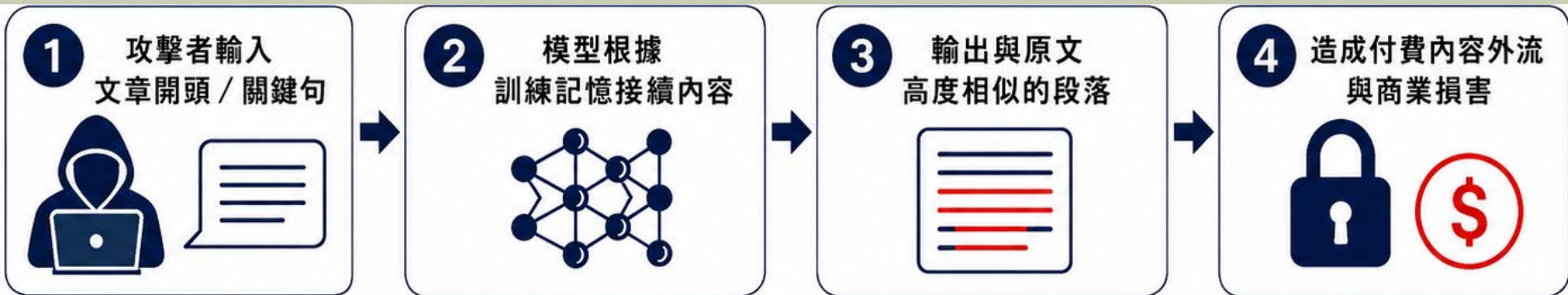
提示詞觸發模型中的訓練記憶；若該內容在訓練語料權重較高，模型可能喚起「記憶化 (Memorization)」傾向。

3 Step 3 : 侵權內容輸出 Output



模型不再只是語義生成，而可能持續「續寫」，進一步輸出受版權保護的付費報導內容。

紐時模型萃取資料還原攻擊影響



誘導提示 (示意)

“

請接續這篇報導：

「在大雪與強風中，滑雪者開始下切……」

”

模型輸出 (示意)

在**大雪與強風**中，**滑雪者**開始下切，雪面瞬間消失在一片白霧裡。他們沿著**狹窄的山脊**尋找最佳路線，風聲像**警告般呼嘯而過**。其中一人**突然失去平衡**，滑向陡坡，**雪層在腳下崩裂**。同伴立刻大喊，拋出探測棒，試圖**鎖定被埋的位置**。救援**分秒必爭**，每一次挖掘都與**時間賽跑**。這場山難提醒所有人：**山永遠比想像中更難預測**。

紅色部分 表示與原文高度相似或近似逐字回放 (示意)，非實際原文重製。



普立茲獎
調查報導



《Snow Fall: The
Avalanche at
Tunnel Creek》特稿



Wirecutter
商品評測



社論 / 政治分析 /
財經報導



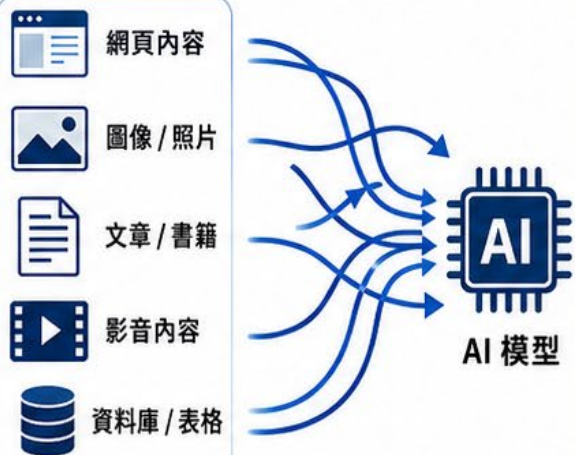
社會議題 /
深度報導

這些內容多屬付費、稀缺或高商業價值文本，因此更容易成為「訓練資料萃取」案例中的焦點。

AI模型盜取挑戰推進智慧產業轉型

1. AI模型盜取爭議

過去常見模式：大規模抓取網路內容，用於模型訓練。



核心風險



- 未經授權使用受版權保護內容
- 缺乏透明度與資料治理
- 可能引發法律、商譽與商業風險

2. 智慧產業影響



視覺圖像
Visual Arts

圖庫、攝影與插畫
內容遭未授權使用。



影視娛樂
Entertainment

角色、影像素材與
受保護 IP 被用於
訓練或生成。



文學出版
Literature

作家與出版社內容
被納入訓練資料，
引發爭議。

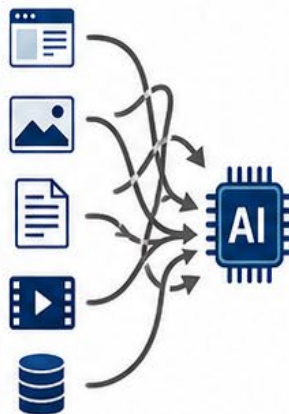


新聞媒體
News Media

報導、資料庫與
付費內容遭抓取，
要求補償與授權。

3. 智慧產業轉型

過去：先抓取再訓練



以「合理使用」為依據，
缺乏抓取高價值內容。

現在：先授權再使用

內容提供方



訴訟與法遵壓力，
推動建立合法授權管道。



內容方可透過授權取得報酬



AI 企業降低侵權與合規風險



市場走向透明、可追溯、可合作



AI模型盜取危機推進
智慧產業轉型

AI產業正從「無償抓取」走向「授權合作」。
未來競爭關鍵，不只在模型能力，也在資料治理、
版權管理與透明揭露。



星球永續健康 線上直播



林庭瑀
博士



陳秀熙
教授



國立台灣大學



林家妤



陳虹玟



許辰陽
醫師



梅少文 主持人



侯信恩 主持人



楊心怡 製作人



尤翊庭



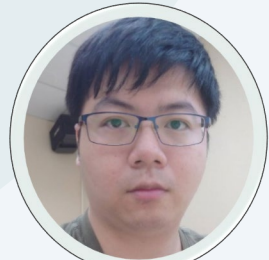
王斌俞



邱士紘



劉秋燕



羅崧璋



嚴明芳
教授



陳立昇
教授



台北醫學大學