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 (8)

蒸餾攻擊智慧模型盜取

2026-05-20

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、羅崧瑋、林家妤、陳虹彣、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/UCCHTox4rUysI30QW4e_xIiA?si=1DIj9qIn3bZWMtNG

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

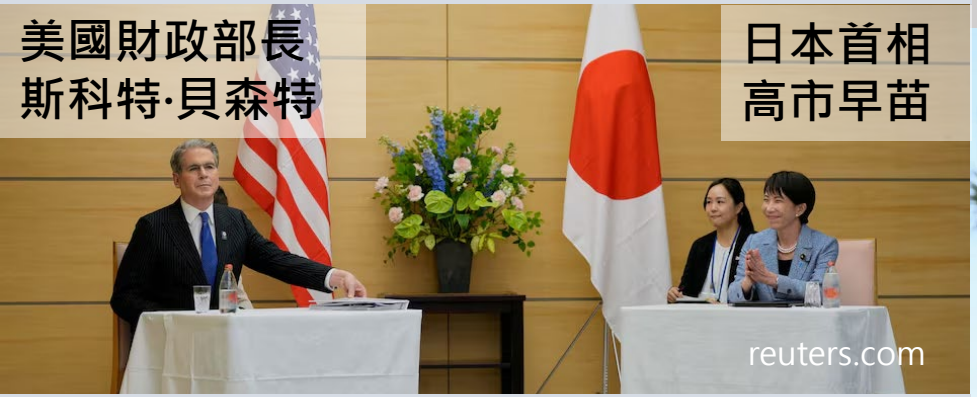
- 健康科學新知 (2026 / W20)
- AI模型盜取資安挑戰
- 蒸餾攻擊AI模型盜取防禦實例

健康科學新知

2026 / W20

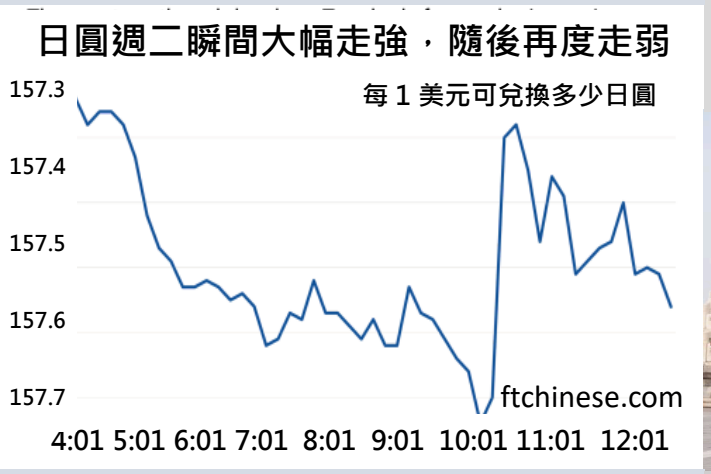
川習會亞洲佈局牽動經貿AI地緣:「鬥而不破」

美財長上週訪日商討匯率與能源，雙方密切協調以應對市場波動



美日協調匯率政策穩定日圓波動並強化美日戰略協作

川普邀科技與產業領導人隨行訪華聚焦貿易、AI與簽定商業協議



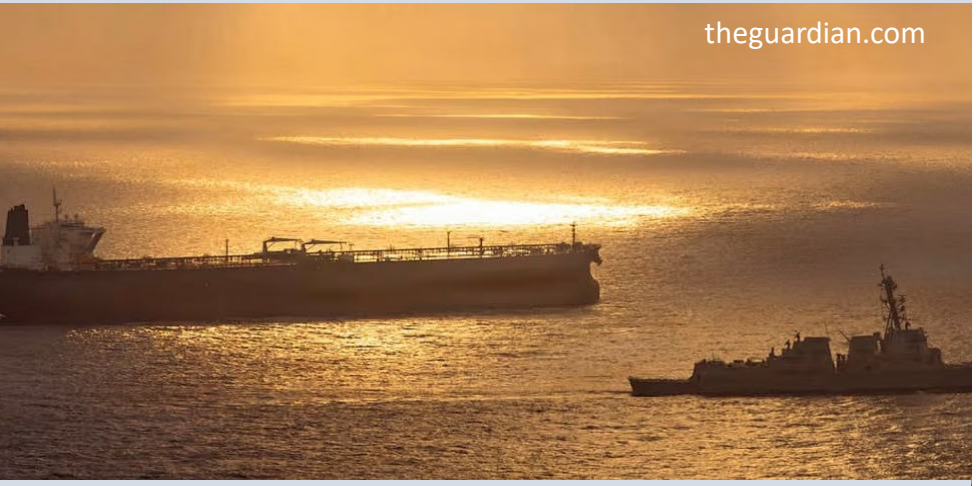
川習會前美中日韓多方磋商 主要聚焦經貿與AI合作策略及中東能源運輸局勢應對

中方副總理赴韓會晤美財長，川習會前磋商貿易風險，穩定兩國經貿關係

川習會達成初步貿易協議，對科技、台灣、中東議題、能源與區域安全亦進行討論

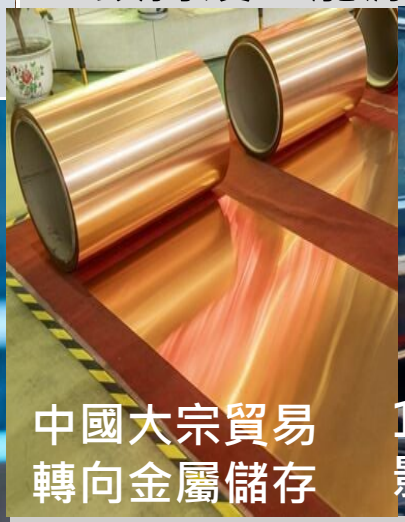
全球能源市場復原路遙:「油震全球」

軍事護航重回檯面，荷莫茲封鎖讓能源危機升級為地緣政治風險



目前全球每週短缺約 1 億桶石油 主要石油公司沙特阿美警告油市恢復恐推遲至2027年後

1973年石油危機重擊經濟 若荷莫茲封鎖持續，能源危機恐轉為金融壓力



荷莫茲海峽受阻大幅改變中國大宗商品貿易
原油進口降低 金屬貿易大幅上升增加貯備

能源危機與肥料短缺:「糧能相扣」

研究背景

Meihua Yang et al., *Nature*, 2026

- 中東衝突擾亂全球肥料市場，海峽等航運封鎖，使尿素價格在一個月內暴漲近 46%，衝擊氮肥供應鏈
- 世界糧食計劃署警告，2026 年將有逾 3.6 億人面臨嚴重糧食不安全

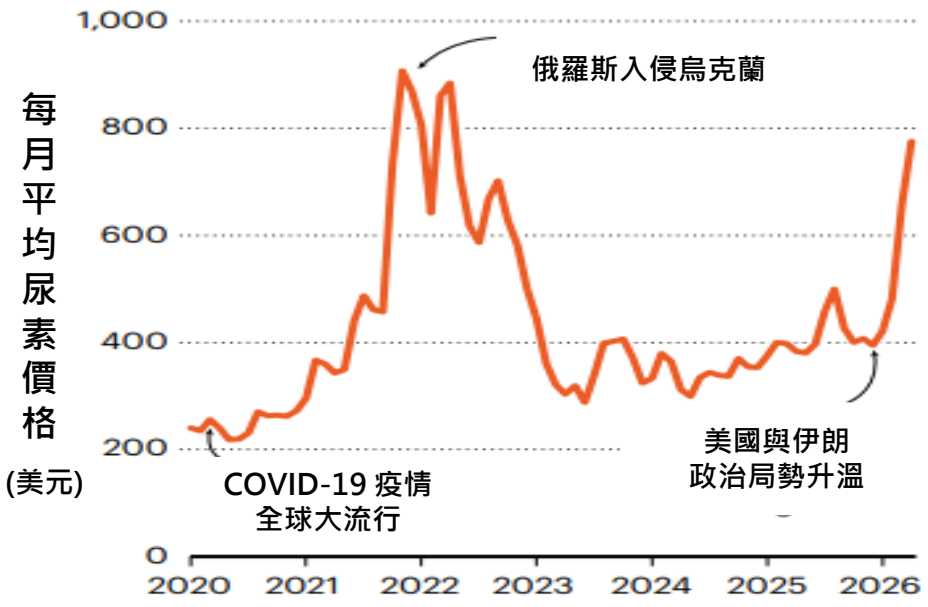
研究發現

1. 高度依賴化石燃料：
   合成氮肥生產耗費全球 1-2% 能源，天然氣占氨生產成本高達 70–80%
2. 供應鏈結構脆弱：
   全球約 13% 化學品需經過荷姆茲海峽，安全威脅使中東主供應國出口受阻
3. 產量非線性衝擊：
   即便微量減少施肥也可能導致產出大幅下降，低收入國家對此衝擊承受力極低

結構原因與意義

- ✓ 肥料長期被視為一般工業商品，而非保障糧食安全策略性基礎設施，缺乏類似戰備儲油緩衝穩定機制
- ✓ 各國多採出口限制等保護措施，雖能穩定國內市場，當前航運限制情境下可能加劇全球性短缺與市場動盪

四年內尿素價格出現兩次顯著高峰，凸顯供應鏈系統性脆弱



健康迷思擴散挑戰科學信任:「眾聲奪真」

Helen Pearson, *Nature*, 2026

健康迷思普遍

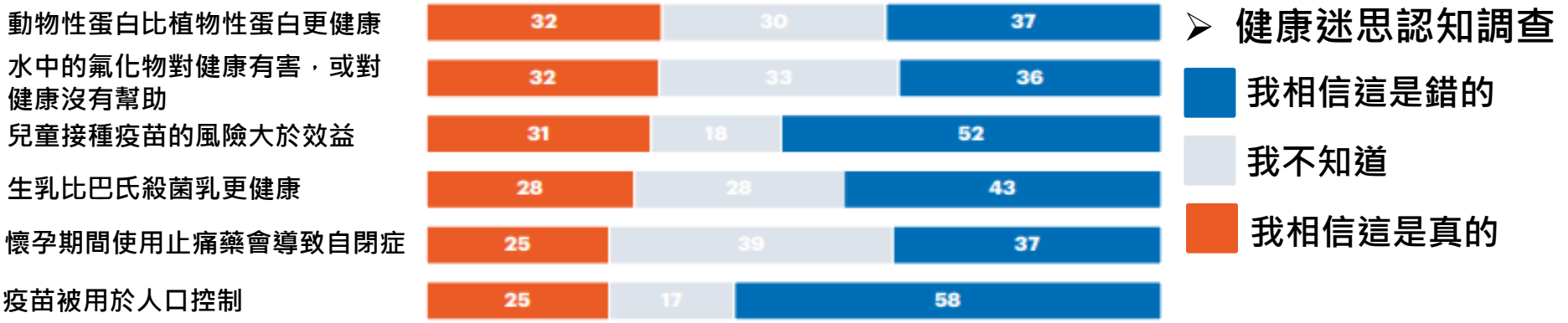
- **信眾廣泛:**16 國調查發現，逾 70% 受訪者至少相信一項未證實健康說法
- **迷思多元**：常見迷思包括疫苗風險、生乳健康、孕期止痛藥與自閉症等

資訊混雜

- **族群多元**：錯誤信念不只出現在低教育族群，部分也受過高等教育
- **資訊混雜**：訊息來源過多，科學證據更難辨識

科學信任重分配

- **信任分散：雖然大眾對科學家仍有一定信任，但社群媒體意見領袖、親友推薦與各類「專家」也逐漸成為健康資訊來源**
- **溝通落差：科學機構若無法用更清楚、易懂且貼近大眾的方式溝通，人們可能轉向其他不一定可靠的資訊來源**



智慧資安技術監管角力:「模型爭權」

歐盟以監管權力換取 AI 資安模型存取權，凸顯 AI 從技術競爭，走向資安與外交角力

歐盟雖有 AI Act 與 AI Office，實際監管前沿 AI 仍仰賴科技公司願意提供存取權

OpenAI 主動向歐盟開放 GPT-5.5-Cyber

先進AI模型能協助找出系統弱點，但駭客也可應用AI發現漏洞設計攻擊

模型存取權影響監管能力

AI時代數位鑑識革命:「破假尋真」

Kai Kupferschmidt, *Science*, 2026

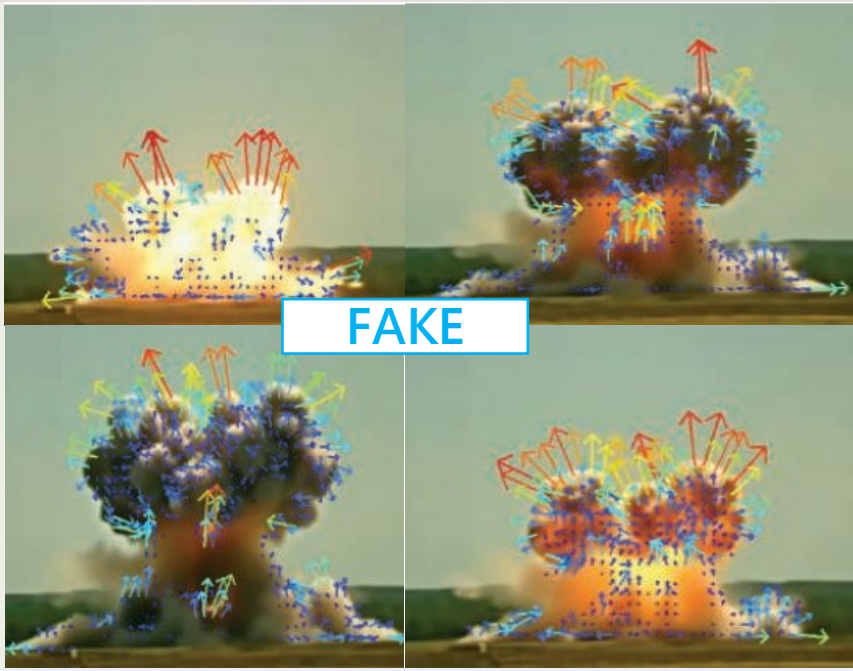
數位鑑識先驅 Hany Farid指出，AI生成影像大量出現，逐漸難分真偽

「導彈影片鑑定案」

- 網路流傳伊朗導彈襲擊影片，專家從物理規律中發現錯誤，質疑虛假訊息正誤導大眾

重要警訊

1. 視覺效果優化： 畫面運鏡佳，效果比真人拍攝更加華麗，因此容易於社交媒體快速傳播
2. 物理規律破綻：AI能模擬逼真紋理，但在影子、反射及消失點等物理邏輯上常有錯誤
3. 規模化挑戰：網路上流傳偽造影像數量驚人，鑑識人員面臨嚴峻考驗



解決對策

- 強化物理基礎檢測:利用 3D重建、影子交會分析及聲音傳播時間差來驗證場景真實性
- 開發專門鑑識工具:例如自動化唇語辨識軟體，用以偵測對口型偽造破綻
- 提高偽造技術門檻:透過如 PhotoDNA等數位指紋技術，鎖定並阻止惡意影內容擴散

AI顛覆科研審查制度:「機審機寫」

Geraint Rees et la., *Nature*, 2026

核心概念

- Agentic AI 可自動產生、優化、甚至提交研究計畫書
- 提案AI可快速整合
  - 研究者過去成果
  - 資助機構評選標準
  - 歷年成功提案

  ➡ 數分鐘內生成高品質提案

當前影響

- 自ChatGPT 在 2022 年發布以來，來自 12 個資助機構資料顯示，補助申請有所增加
- 提案品質整體提升 → 難以區分優劣
- 約 41%研究者已用AI撰寫提案

系統風險

- 高品質提案「爆量」→ 評審負擔過重
- AI生成提案 vs AI輔助審查

  ➡ 變成「AI評分AI」
- AI提案轉為模仿過去成功模式
- 可能導致：
  - 評選公平性爭議
  - 創新被壓抑

改革方向

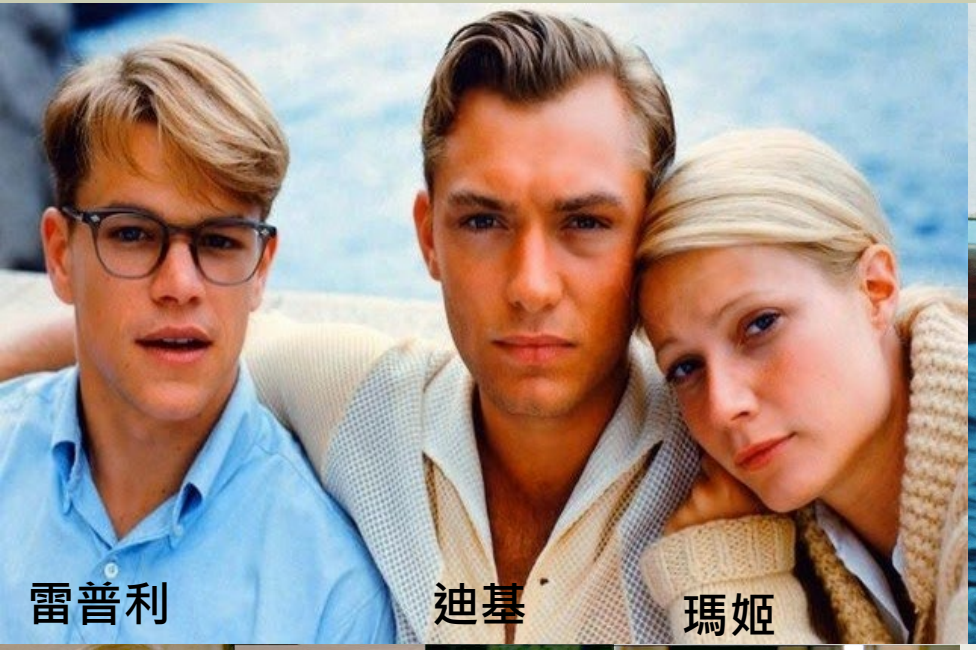
- 申請者特質成為評估重點：評估研究者過往表現與團隊能力(面試、長期績效)
- 申請限制：限制申請數量，讓大型計畫取代零散申請。
- 建立 AI-native funding system： 分析研究者影響力、合作網絡、創新潛力

AI模型盜取資安挑戰

天才雷普利: 偽冒高手故事

- 1950年代紐約湯姆．雷普利穿著借來的普林斯頓夾克在派對伴奏鋼琴，造船富商格林利夫誤認湯姆為兒子迪基同學
- 格林利夫委託湯姆赴義大利，承諾一千美元帶迪基回家

模仿高手偽冒富商之子生活



- 湯姆在迪基不在房裡時偷穿衣服、模仿動作、學簽名。這份模仿從崇拜滑向佔有。迪基決定甩開這個跟前跟後影子
- 雷普利爭吵中失手誤殺迪基，運用其模仿天分奢華度日

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

智慧模型蒸餾攻擊

透過大量查詢黑箱模型，蒐集輸入與輸出對應，訓練出可模仿其核心功能的替代模型。

1 攻擊者準備查詢資料 x'

公開圖片／合成資料／自製樣本

2 查詢目標模型 (黑箱 API)

API

送出 x'，取得回應

$y = M(x')$

3 建立查詢資料集

輸入 x'_i	目標回應 y_i
x'_1	y_1
x'_2	y_2
...	...
x'_N	y_N

收集配對 (x'_i, y_i)，重複 N 次直到預算用完

4 蒸餾訓練替代模型

訓練 D 使其輸出接近 M 的輸出

5 取得替代模型

$$D(x') \approx M(x')$$

完成可模仿目標模型功能的替代模型 D

損失函數：

$$L = KL(M(x') \,||\, D(x'))$$

利用軟標籤讓替代模型逼近目標模型的功能輸出。

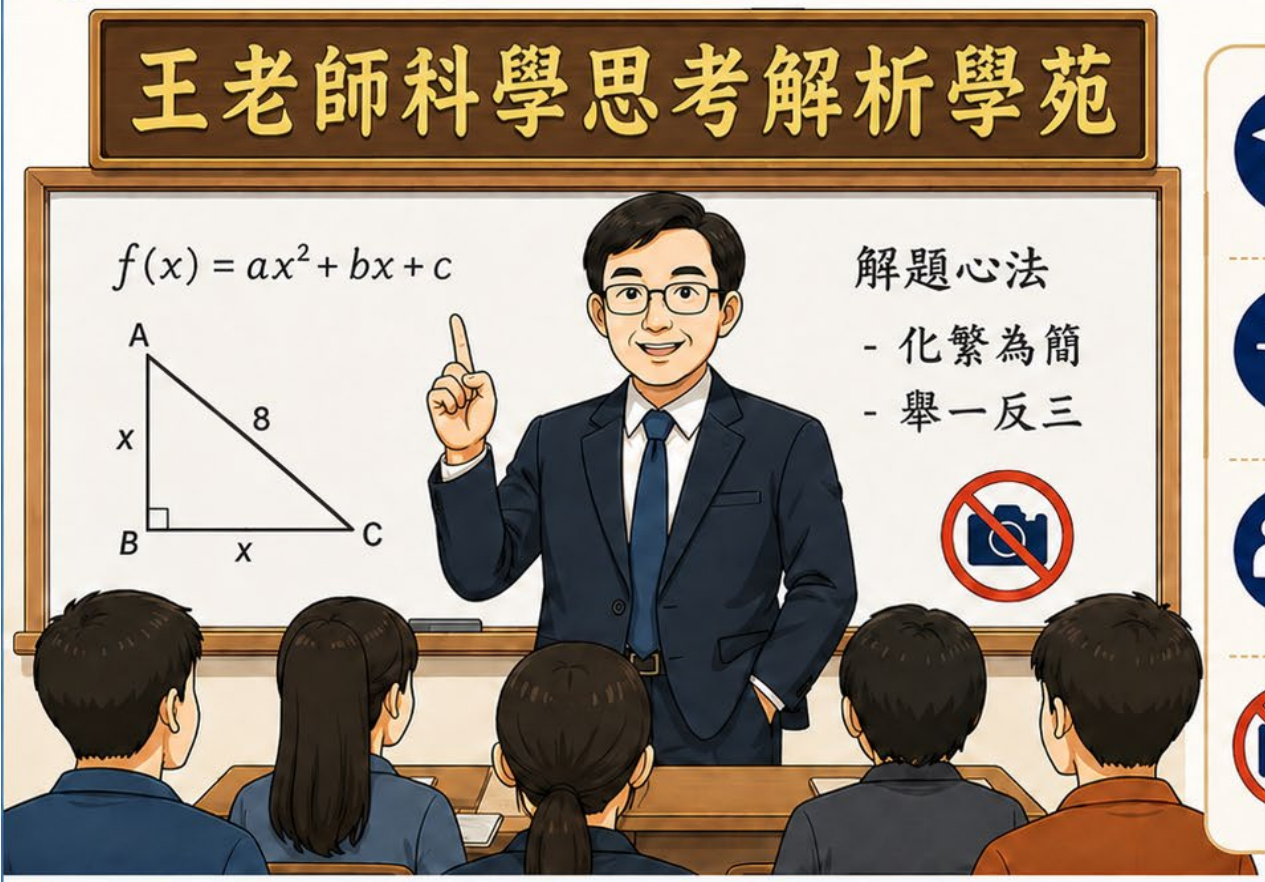
攻擊的本質

不是偷走參數 → 萃取核心功能 → 重建可用替代模型

重點

- 大量查詢黑箱模型
- 蒐集輸入與輸出配對
- 以蒸餾方式複製模型功能

智慧模型盜取: 知識產業案例



教學經驗 20 年
深厚功力，口碑保證

自創獨門心法
化繁為簡，舉一反三

只收 50 位學生
小班精教，品質保證

上課禁止錄影錄音
保護內容，維護品質

AI

就像頂尖 AI 模型： GPT-4、Claude、Gemini

開發成本極高
投入大量資源與時間

能力強大卓越
推理、創作、分析一流

使用嚴格受限
名額/額度有限，規則嚴格

頂尖價值，值得珍惜；善用資源，創造最大效益！

模型萃取(Model Extraction)攻擊

① 目標

做出難以分辨的「假王老師」

讓人分不出哪個是真王老師，
哪個是競爭對手做出的克隆版。

② 做法：多管齊下

1 臥底蒐集

派人上課、錄音錄影、
記錄口頭禪與手勢。

2 社交工程調查

查學經歷、訪談同事、
蒐集教學資料。

3 逆向工程

拼湊所有線索，
還原教學邏輯與
思考方式。

4 完整對應

題目、解法、用詞、
例子都盡量做得一樣。

③ 產出特徵

規模
與原版差不多大，
甚至一樣大

能力
幾乎全方位對等

成本
攻擊成本極高
（要付學費、買人脈、
雇工程師）

目的
通常用於商業競爭、
軍事用途、國家級竊密

重點：模型萃取的核心，在於系統性重建原模型之「教學能力+風格」。

知識蒸餾攻擊 (Distillation Attack)

目標

我不想複製王老師本人，
我要萃取他的「解題精華」，
做成平價版！

做不出王老師等級的分身，也沒關係

瞄準廣大學生市場

✅ 重點不是複製本人，而是抓住可複製的解題方法。

作法

① 抄筆記

派學生長期上課，
認真聽、認真抄。
重點記下：
「題目 → 解題步驟 → 答案」

② 整理精華

刪除閒聊、人生哲學等，
只保留
「純粹的解題方法」

③ 小老師學習

讓年輕老師／小模型
熟練精華講義，
遇到類似題目，
能解出相同或接近的答案

④ 包裝上市

推出「仿製課程解題班」，
低價競銷解題方法
吸引學生

核心策略：萃取精華 → 學習複現 → 仿製解題班！

知識蒸餾攻擊產出特徵

真王老師

≈

80–90%
核心能力

小模型保留約 80–90%
核心能力，但失去周邊能力

平價解題班（小模型）

面向	特徵
規模	比原版小很多（可能是 1/10 甚至 1/100）
能力	核心能力保留 80–90%，但失去周邊能力
成本	攻集成本相對低（只需 API 費用 + 一些算力）
目的	通常是商業套利、規避付費、繞過限制

蒸餾攻擊AI技術應用

1 收集問答對

大量呼叫目標 API，將「輸入→輸出」儲存，
形成訓練資料集。

輸入（問題）

請計算下列定積分並給出精確值：

$$\int_0^{\pi/2} x^2 \ln(\sin x)\,dx$$

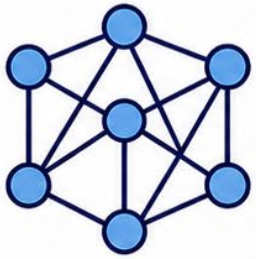
輸出（答案）

$$\int_0^{\pi/2} x^2 \ln(\sin x)\,dx = -\frac{\pi^3}{12}\ln 2 + \frac{7}{16}\zeta(3)$$

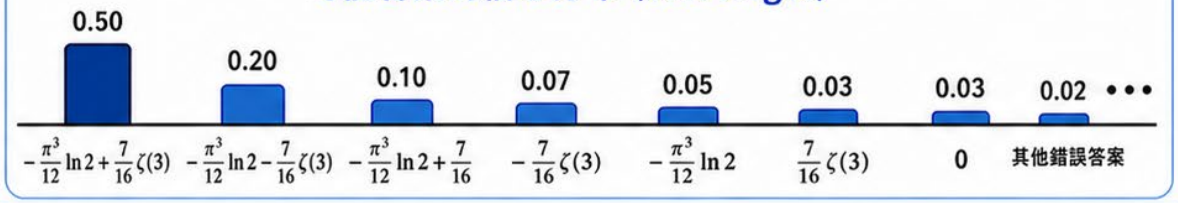
（其中 $\zeta(3)$ 為阿佩里常數）

• • •

2 訓練學生模型



用這份資料集，訓練一個小模型，目標是讓學生模型
學到的不只「最終答案」，還有「老師模型對各種可能
答案的信心分布」（soft target）。



3 部署小模型

訓練完成後，小模型可本地運行，
降低算力需求、速度更快、成本更低。

降低算力需求　速度更快　成本更低

模型萃取 vs 知識蒸餾

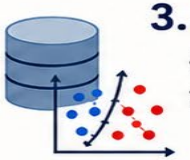
比較面向	模型萃取派 (Model Extraction)	蒸餾派 (Distillation)
目標	做出幾乎一樣的「分身」	萃取科學解析結果 製成解題技術
策略	多管齊下、全面複製	專注知識萃取與學習
技術手段	查詢攻擊、側通道、 參數還原、決策邊界探測	收集問答對、蒸餾訓練、 小模型部署
產出規模	與原版相當	小很多（1/10～1/100）
能力表現	幾乎全方位對等	核心能力 80～90%
成本	極高	相對低
常見目的	商業競爭、軍事用途、 國家級竊密	商業套利、規避付費、 規避限制

AI模型生命週期防禦策略

攻擊流程

1. 大量查詢
攻擊者向目標模型發送大量查詢請求

2. 蒐集輸出
蒐集模型的輸出結果（機率、標籤等）



3. 模型蒸餾/訓練
使用蒐集到的資料訓練替代模型

4. 替代模型
獲得效能接近的替代模型

模型能力被竊取！

防禦策略（三層防護）

一、防禦（事前預防）

降低模型輸出資訊的可用性

1. 輸出擾動

在輸出中加入隨機雜訊或進行平滑處理，降低攻擊者蒐集到的資訊品質。

技術：高斯雜訊、隨機平滑、離散機率化

2. 速率限制與查詢監控

限制查詢請求總數量，並監控異常查詢行為，降低大規模蒐集的可能性。

技術：速率限制、配額管理、異常偵測

目標：減少可被萃取的資訊量與查詢規模

二、偵測（即時偵測）

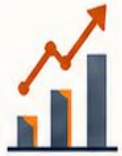
辨識可疑的模型萃取行為

1. 離群查詢分析（OOD）

偵測輸入分布是否異常，判斷查詢是否來自模型萃取攻擊行為。

技術：OOD 偵測、統計檢定、距離度量

2. 進階偵測方法



利用專為萃取攻擊設計的方法，更精準辨識可疑查詢模式。

技術：PRADA、MISLEADER、反饋導向偵測

目標：及早發現並阻擋萃取攻擊行為

三、所有權驗證（事後驗證）

確認模型或資料的合法所有權

1. 浮水印 / 指紋

在模型中嵌入不可見的浮水印，驗證模型所有權與來源。

技術：數位浮水印、參數指紋、輸出指紋

2. 資料推斷

透過分析替代模型，推斷其訓練資料是否包含受保護之資料。

技術：DRW、GINSEW、成員推斷

目標：追溯來源並維護智慧財產權

蒸餾攻擊

模型盜取防禦實例

MiniMax 蒸餾攻擊案件總覽

Anthropic 於 2026/02/23 揭露的工業級模型蒸餾事件

1 核心主角

ANTHROP\C
Claude

揭露方：
Anthropic

疑似進行
模型蒸餾（竊取）

MINIMAX

被指控對象：
MiniMax
（上海 AI 公司）

2 案件性質

攻擊性質：
蒸餾攻擊 / 模型竊取

鎖定能力：
Agentic coding、Tool use

法律狀態：
已完成收集證據

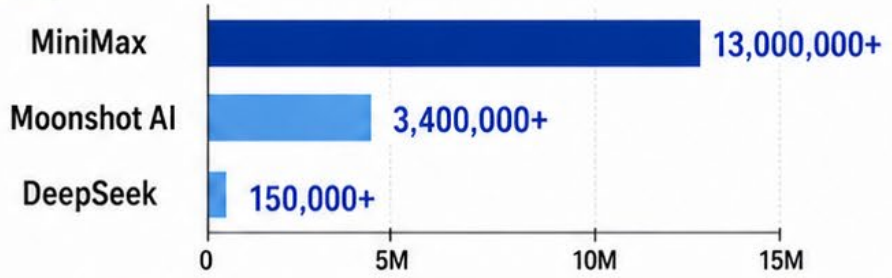
3 規模重點

13,000,000+
次交互

約占三案例總量
81%

24,000
個詐欺帳號
（3 案合計）

4 三案比較



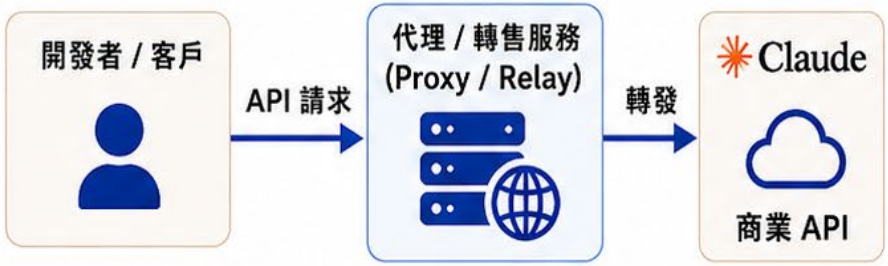
5 事件特殊性

這是**首個**被先進 AI 模型公司在攻擊進行中
即時偵測、歸因、公開的蒸餾攻擊案例。

系統化蒸餾攻擊技術

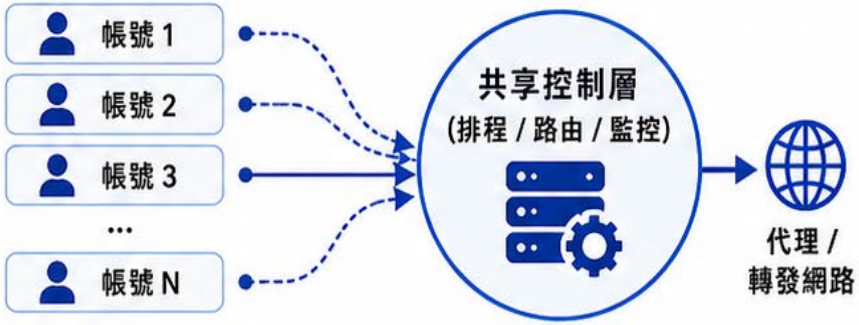
MiniMax 蒸餾攻擊如何取得資料並轉化為訓練資源

1 商業代理服務



- 中國大陸無法直接取得 Claude 商業 API
- 轉向轉售最新開發 AI 模型存取權的灰色代理服務
- 違反地區限制與禁止訓練競品條款

2 Hydra Cluster



- 分散性
- 無單一故障點
- 流量混雜

3 Prompt 工業化

基礎 Prompt 範本

```
"You are an
expert data
analyst..."
```

- 跨帳號同步投放
- 集中於 agent 與 tool use
- 高頻率、大規模重複執行

4 兩種訓練資料用途

SFT：蒐集高品質 (instruction, completion)

RL / RLAIF：用 Claude 作為教師模型提供 reward 信號

大量 agent 編碼與工具使用範例，可能同時用於 **SFT** 與 **RL**。

24小時蒸餾攻擊關鍵證據的關鍵意涵

新版 Claude 發布後，MiniMax 在 24 小時內將近一半流量轉向新模型，顯示其可能已完成資料蒐集、訓練、部署與驗證閉環

1 Anthropic 發布
新版 Claude

2 MiniMax 重新導流
將近一半流量轉向新模型

Claude
Claude 3.7 Sonnet

24 小時內

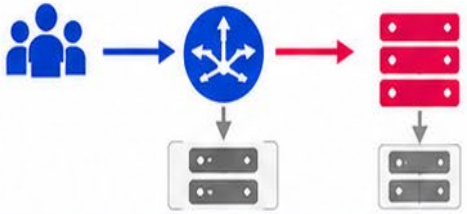
MINIMAX

T_0 2026/02/23 10:00

T_1 2026/02/24 10:00

舊模型端點（V1）	事件前	事件後	新模型端點（V2）
api.minimax.chat/v1（舊模型）	~100% 流量主要打到舊模型 V1	~50% 轉向新模型 V2	api.minimax.chat/v2（新模型）

「重新導流」在蒸餾攻擊中的意義（攻擊特徵）



- 將原本送往舊模型的真實使用者請求，**快速切換**到新模型端點
- 把高價值 prompt、agent 任務與 tool-use 流量，用來**驗證**疑似蒸餾後的新模型表現
- 這代表對手已從資料收集推進至**「部署與驗證」**階段

1 證據意涵一：完成訓練閉環

能在 24 小時內切流，表示對手可能已**完成**資料蒐集、樣本生成、模型訓練與上線準備。

2 證據意涵二：真實流量驗證新模型

重新導流代表新模型已開始承接**真實工作負載**，用於測試、微調與持續優化。

3 證據意涵三：屬於營運層級行為

這顯示蒸餾攻擊已由資料擷取推進至商業化部署，屬於**營運層級證據**。

蒸餾攻擊即時偵測預防損失

Multi-Layered Detection：五層防禦如何完成即時歸因

1 

行為指紋

辨識 chain-of-thought 誘導與能力萃取序列

2 

使用帳號監測與行為統整

同步峰值、共用 prompt 模板、相同任務分布

3 

基礎設施分析

IP、雲端租戶 ID、TLS 指紋、HTTP/2 frame 順序

4 

付款資料－使用者帳戶動態比對

多個買家、單一付款人

5 

行業情資

與其他 AI 提供商交叉比對相同 actor

模型蒸餾高效監測

- 多訊號融合，而非單點證據
- 可在模型發布前攔截
- 可辨識多目標並行蒸餾

國際智慧產業AI盜取攻擊預防協作

1 出口管制

- 事件發布時點貼近美國國會辯論 AI 晶片出口管制
- Anthropic 呼籲 AI 服務出口管制與晶片管制掛鉤

2 法律層面

- 違反使用政策：禁止用輸出開發競品
- 地區限制條款
- 可能涉及 CFAA、契約違約、商業秘密
- 跨國追訴困難

3 產業影響

- OpenAI、Google DeepMind 加強類似偵測機制
- API 經濟從開放增長轉向分層防衛 + 主動偵測
- detect-and-suppress 成為標準防禦元件

核心問題：競爭方濫用智慧模型功能盜取能力

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


梅少文 主持人


侯信恩主持人


楊心怡製作人


林家妤


陳虹彣

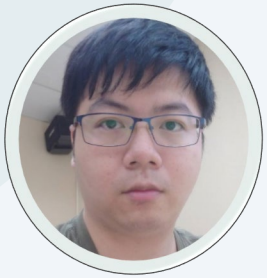

許辰陽
醫師


尤翊庭


王斌俞


邱士紘


劉秋燕


羅崧瑋


嚴明芳
教授


陳立昇
教授

台北醫學大學