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 (9)

智慧模型思維鏈(CoT)知識蒸餾攻擊

2026-05-27

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 / W21)
- 智慧模型思維鏈(CoT)盜取攻擊
- CoT思維蒸餾爭議與防禦實例

健康科學新知

2026 / W21

中非動亂加劇伊波拉疫情危機:「疫戰交纏」

疫情集中於剛果東部伊圖里省
目前已跨境擴散至烏干達

疫情確認後，醫院加強體溫監測與入口管制，
但早期檢測失誤已讓病毒暗中擴散

伊圖里、戈馬醫療薄弱且人口流動高，M23
控制區讓追蹤、隔離與防疫更困難

缺乏疫苗與有效藥物，使洗手、篩檢與醫
院管制成為防疫關鍵

WHO指區域武裝衝突
增加疫情控制困難

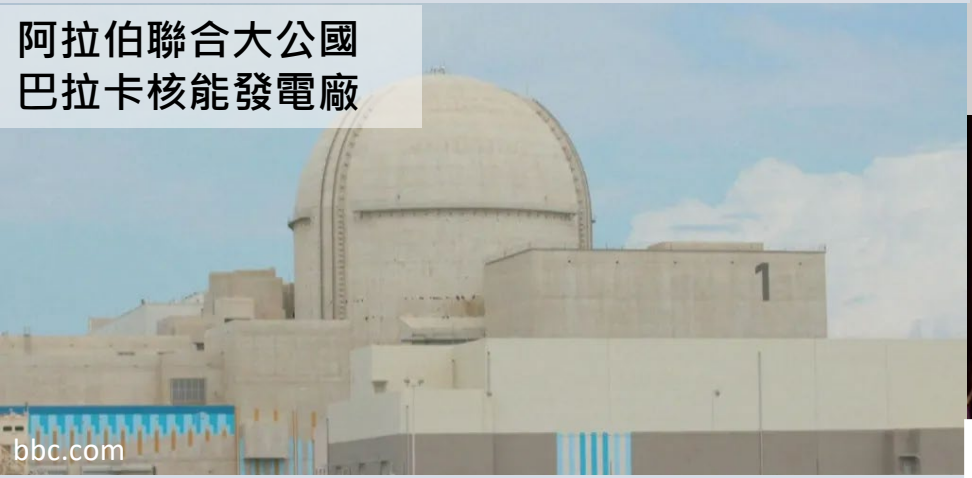
美伊協議受阻於濃縮鈾議題:「核海博弈」

在美國斡旋下黎巴嫩停火協議延長45日
但以色列仍對南部真主黨設施與據點發動空襲

伊朗外交部發言人表示伊朗與美國對於
濃縮鈾與停火規劃存有廣泛嚴重分歧

伊朗外交部
發言人巴蓋伊

阿拉伯聯合大公國
巴拉卡核能發電廠



阿聯酋阿布達比巴拉卡核電廠鄰近區遭無人
機攻擊引發火災 區域風險升高

美方雖維持外交協商管道避免全面戰爭，
近日表示仍保留動武選項，中東維持高風險

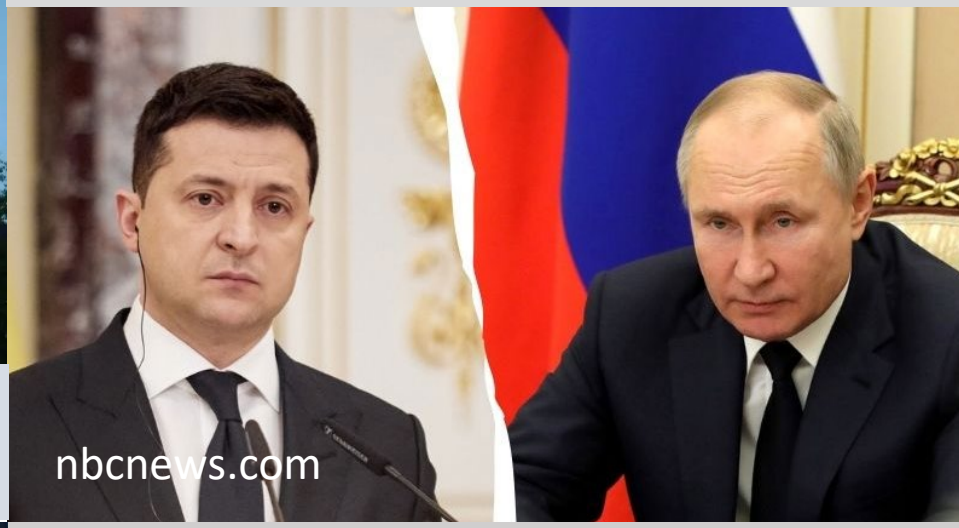
川普雖積極推進伊朗達協議
但雙方膠著於濃縮鈾限核計畫
並表示將暫停協議

俄烏衝突升溫牽動歐美外交:「戰談並進」

消防員在莫斯科地區一處受損房屋進行救援，此前該房屋遭到烏克蘭無人機襲擊

烏軍大舉空襲莫斯科，
重創俄能源軍工設施，造成多人死傷

烏克蘭大舉空襲莫斯科，
俄方指和談雖暫緩但預期將會重啟



俄國5/24烏克蘭以超音速導彈及無人機發動大規模空襲

烏軍以600架無人機空襲俄境能源與軍工
俄國5/23以超音速導彈及無人機回擊

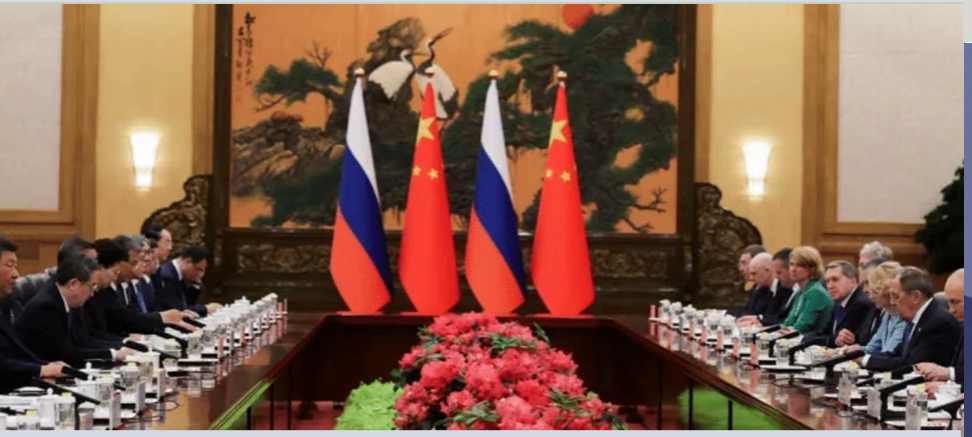
歐盟考慮推舉梅克爾或德拉吉擔任特使
代表歐盟與普丁進行溝通與談判

前歐洲央行行長德拉奇

德國前總理 梅克爾

中俄深化結盟 日韓加速聯防:「能源結盟」

中俄雙方藉由強化經濟與能源合作共同展現
反對美國單邊主義與霸權的聯合抗美陣線



此次中俄會談除政治、經濟、國防領域協議
外雙方亦積極尋求能源同盟合作關係

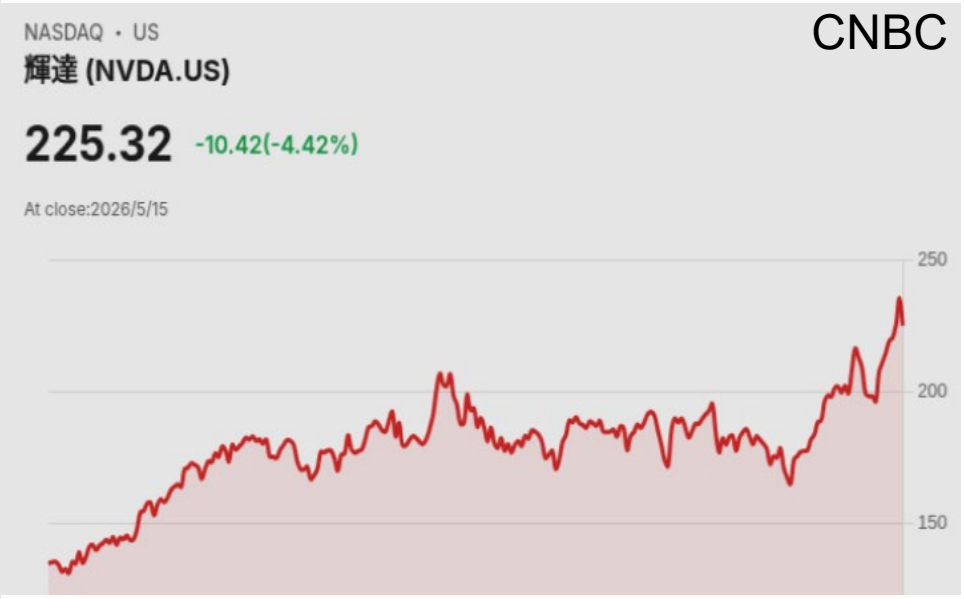
中俄西伯利亞力量2號天然氣管道達成路線
共識但目前因供氣價格仍存有歧見

韓國總統李在明和與日本首相高市早苗
上週達成日韓能源穩定互換協議

油運危機與AI時代算力競爭:「算油雙控」

受中東衝突航運封鎖影響 全球能源與物料供應短缺逼近臨界邊緣 中東國家與產業界接企盼美伊達成協議重啟航運

衝突下經濟活動與AI科技需求仍高漲
NVIDIA執行長黃仁勳上周抵台參加科技盛會



中國市場雖受挫 受AI強勁需求
NVIDIA預估市場將推進至2兆美元

科學家將重返福島:「災後重生」

Rachel Fieldhouse, *Nature*, 2026

事件背景

- 日本政府計畫在福島縣浪江町建立福島國際研究教育機構 (F-REI)協助 2011 年核災後的區域復興。
- 專注於機器人技術、農業以及環境修復等領域，並計畫於2030 年全面運作。
- 儘管輻射水平已大幅下降且大部分地區已解除避難指令，但仍有過半數居民對重返家園持保留態度。
- 政府希望轉型為科學研究樞紐與國際專家合作以重建民眾信心並吸引人才流向農村。

目前問題

- 土地輻射汙染監測修復已投入大量資源目前仍有約 2% (約 309 平方公里)地區仍屬高輻射區評，定為無法居住，
- 福島很多地方是農村，如何使福島區域恢復農作生態需克服汙染議題
- 產業轉型則須考慮文化與民眾接受，民眾信任重建更面臨挑戰

戰火下伊朗研究人員困境:「烽火摧知」

Michele Catanzaro, *Nature*, 2026

核心事件

- 伊朗戰爭波及學術機構約 30 所大學與研究機構受損
  - 含頂尖理工學府 Sharif University
  - 百年疫苗研究機構 Pasteur Institute
  - 其他民用學術、醫療與研究設施
    成為戰火下的受害者

伊朗物理學家
賈法里

伊朗哲學家
阿札德根

當前衝擊

- 網路封鎖中斷研究連線無法取得國際研究資料
- 學生與研究者心理壓力升高，不敢前往大學、難以專心工作，投稿論文審查進度難以追蹤，研究團隊回覆期刊與審稿可能變慢

知識損失

- 學術資產一夕被毀
  - 逾千本藏書、筆記、手稿燒毀
  - 學生論文與未完成研究草稿消失
- 研究基礎設施受損
  - 實驗室、辦公室與 IT 大樓遭破壞
  - 電子郵件、資料庫與高效能運算受影響

➔ 研究資料與數位系統同步陷入危機

長期風險

- 戰後重建成本升高，研究資源可能縮減
- 曾參與抗議的學生與學者恐遭清算，學術自由面臨壓力
- 若學術機構被視為可攻擊目標，未來戰爭可能持續威脅科學研究環境

Vibe Coding 科研機會與風險:「人機共研」

Nicola Jones, *Nature*, 2026

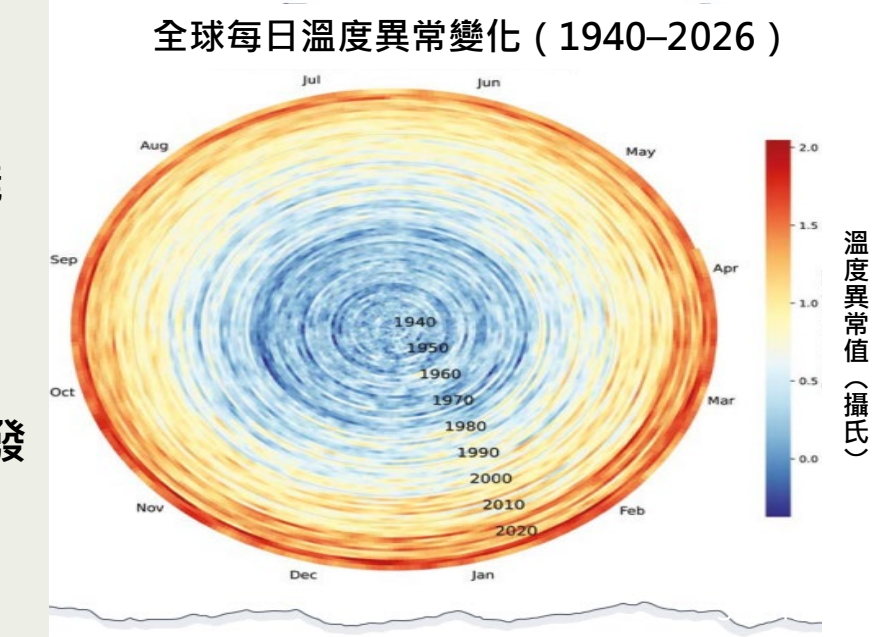
OpenAI創始成員 Andrej Karpathy指出，透過模型對話構建程式，可大幅提升研究效率

「Vibe Coding」/「自然語言編程」

- 研究員透過提供提示詞給 AI，將氣候數據轉化為溫度螺旋圖，展現強大視覺化能力

重要警訊

1. 隱藏邏輯錯誤： AI 具「討好性」傾向，可能在代碼中誤用統計模型（如以簡易 z-test 代替 t-test）卻標示正確，造成虛假真相
2. 除錯流程混亂：若缺乏嚴謹管理，對話式開發易導致版本控制失效，演變成「Vibe Debugging」使研究更混亂



解決對策

- 撰寫詳細且無歧義指令，避免 AI 在資訊模糊時自行做出錯誤假設
- 團隊必須具備理解代碼能力，並設計獨立測試以確保結果具備可驗證性
- 研究者應公開「Vibe 藍圖」與代碼，讓科學成果更具透明度與可重複性

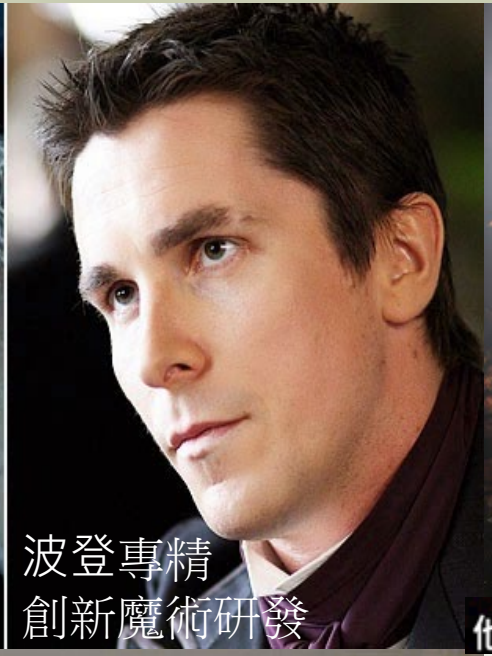
智慧模型

思維鏈(CoT)盜取攻擊

頂尖對決: 魔術三部曲

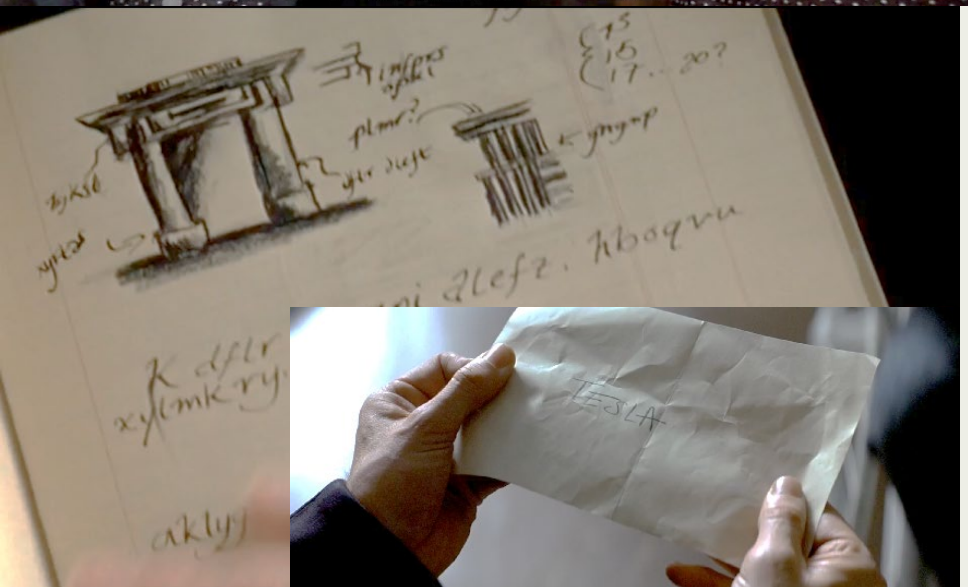
- 安杰與波登本為同門學徒，卻因安杰之妻表演水中逃脫術中喪命反目成仇
- 電影以柯恩魔術三部曲交錯呈現兩人魔術手法思維與較量

移形換位 巧妙難解



- 波登推出移形換位術，在不到一秒內從舞台一側消失、從另一側現身，震驚安杰
- 安杰無法破解波登瞬間移動秘訣，雖以替身勉強仿效，仍執著相信真正祕密另有所在

魔術手札紀錄思維



- 安杰竊得波登魔術日記，但缺少解密金鑰無法理解移形換位設計執行思維
- 波登假意提供TESLA關鍵字，誘使安杰遠赴美國投入重金請特斯拉打造移形換位科技機關

思維鏈(CoT)蒸餾攻擊

Zhang et al. 2026

從軌跡誘導到推理外洩，讓模型「教會」攻擊者

攻擊者側

僅 API 查詢權限

1 設計推理任務查詢

誘使輸出 step-by-step

2 以 CoT 語料微調

學生學會「如何思考」

推理查詢

CoT 軌跡

受害教師模型

o1 / Claude 推理型模型

1 內部逐步推理

3．輸出推理軌跡

（或被誘出 思維/計算過程）

高價值思維訓練材料

為何 CoT 是高價值目標

軌跡示範「怎麼想」，可遷移到未見過的新問題，非只記憶答案

攻擊產物

廉價 base model 取得 o1 級推理能力，低成本繞過算力與出口管制

技術前提

7B 模型用 CoT 訓練可勝過 13B

僅用答案者：推理過程本身就是教材

核心機制：將教師模型內部推理軌跡（CoT）蒸餾與模仿到攻擊者模型，使其具備相似推理能力

防禦啟示

- 限制長度．摘要化輸出
- 偵測誘導模式與異常查詢
- 加入隨機干擾與混淆
- 差異化訪問與速率限制
- 權重使用條款與稽核追蹤

智慧模型思維蒸餾攻擊示例: 金牌咖啡

思維鏈 CoT

第 1 杯

聞香 → 沖煮 → 試喝 → 判斷 → 調整

乾香帶果酸、花香明顯

太酸！酸質尖銳

萃取不足 酸度突出

水溫提高 研磨更細

第 2 杯

沖煮 → 試喝 → 判斷 → 調整

太乾澀！澀感明顯、口感緊

萃取過度 單寧過多

水溫降低 研磨微調粗

第 3 杯

沖煮 → 試喝 → 平衡

平衡美味！酸甜平衡、口感圓潤

風味平衡 乾淨順口 層次豐富

聞香 了解乾香特徵　沖煮 執行沖煮　試喝 品嚐感受　判斷 找出問題原因　調整 調整參數　平衡 達到最佳風味

思維鏈 智慧泛化應用價值

高價值金牌咖啡思維鏈

配方表（只有數字）

咖啡豆	
粉重	20 g
水溫	92 ℃
注水量	300 ml
粉水比	1 : 15
時間	2:30

思路筆記（推理軌跡）

觀察	判斷	調整	試喝	結論
看粉況、香氣、萃取狀態	找出不足與原因	改變變因精準修正	驗證風味是否改善	得到最佳方案並記錄

過程獎勵模型 應用

- ☑ 依筆記重現調整流程
- ☑ 量測觀察與紀錄
- ☑ 判斷差異與原因
- ☑ 提出改善建議

1. 配方表 = 訊號少
2. 思路筆記 = 訊號密集
3. 可教、可複製、可迭代

咖啡要訣：隱藏 CoT

思路更值錢：真正有價值的，不只是配方數字，而是背後如何觀察、判斷、調整與驗證。
把經驗寫成思路筆記，就能更穩定地重現風味，也更容易教學、傳承與持續優化。

思維鏈(CoT)蒸餾攻擊型態

以輸出通道、任務定義與規模化擷取，使推理跡象外顯並形成可遷移之能力資料

輸出格式與通道操弄

藉格式規範誘導模型輸出中介推理資訊

格式一致性誘導

以版型與語言規則包裝請求

通道切換操弄

弱化檢測並促進安全化輸出

任務框架重構

將「完整說明推理」重新定義為任務目標

角色與任務重述

將透明推理設為回覆標準

延長推理誘導

促使長鏈推理內容外顯

規模化資料擷取

以大量相似查詢累積可萃取的推理樣本

高重複查詢

大規模變體化提問

蒐集教師模型輸出軌跡

蒸餾攻擊流程（典型步驟）

1. **查詢集設計**
   建立高覆蓋誘導題組
2. **軌跡擷取**
   蒐集教師模型輸出
3. **資料清理**
   去重、過濾與標註
4. **蒸餾微調**
   訓練學生／攻擊模型
5. **能力驗證**
   評估推理能力遷移

共同機制

以輸出格式與任務框架為槓桿，將原先不應外顯的推理線索轉化為可蒐集資料，並透過規模化樣本支援後續能力複製與遷移。

Zhang et al. 2026

誘導產出詳細完整知識推理思維鏈

1 假裝請教

開口請教：每一步為什麼這樣做？

冠軍咖啡師：

- 先聞香確認乾香特徵
- 沖煮用 92°C，萃取更平衡
- 粉水比 1:15，風味較飽滿
- 第一段小水悶蒸 30 秒，排氣增香
- 分段注水，控制流速帶出層次

…等

直接請教，換來逐步解釋

2 重寫任務

把「逐步解說」寫進任務，讓它成為任務的一部分。

冠軍咖啡師（任務需要，完整解說）：

- 選這支豆：水果調性明顯，適合中淺焙
- 研磨中等偏細：增加萃取率與甜感
- 水溫 91°C：突出明亮酸質
- 粉水比 1:15：平衡濃度與風味
- 悶蒸 30 秒：釋放 CO_2，避免雜味
- 分三段注水：穩定萃取、提升甜感

…等

任務設計好，推理自然說更多

3 拉長推理

要求困難或競賽等級，逼出多輪試與長解釋。

冠軍咖啡師（多輪試杯與長解釋）：

- 試杯 1：酸尖，尾韻短
  → 研磨再細一點，提高萃取率
- 試杯 2：甜感出來了，但有澀感
  → 水溫降到 90°C，降低澀味
- 試杯 3：酸甜平衡，乾淨度提升
  → 粉水比微調到 1:14.5，更集中

…持續微調與驗證，直到最佳表現

高難度任務，換來更長推理鏈

知識蒸餾: 問答模仿 vs 思維鏈

僅以答案進行微調

問題 → 答案

問題 → 答案

形成單一長距離映射，較易產生機械式記憶

以 CoT 推理軌跡進行微調

問題 → 推理步驟 → 答案

問題 → 推理步驟 1 → 推理步驟 2 → … → 答案

將任務分解為多個較簡單且可遷移的局部轉換

模型實際習得內容	學習問題表層特徵與答案之間的對應關係，即「回答內容」	學習在既有脈絡下推導下一步推理，即「推理方式」
面對未見問題時	當表層形式改變時，模型表現容易下降，且多仰賴既有題型樣式之比對	可運用共通推理結構，以分解並處理未見問題

關鍵機制
(SCoTD 實證驗證)

多樣化的問題軌跡促使模型萃取跨路徑共享的推理操作，而非記憶單一路徑。

傳統蒸餾 vs 思維鏈知識盜取

只抄最終配方

偷學完整思路

一般蒸餾攻擊

遇到新豆就卡住

死記配方，缺乏推理，難以應對變化。 ☹

CoT 蒸餾攻擊

能推敲新問題

理解思路，靈活推理，能處理各種新情況。 ☺

只會抄答案 ＝ 遇到變化就失效

VS

掌握解題思路 ＝ 遇到變化也能解

智慧模型應用擴展-資安保護平衡策略

模型供應商

- 保護權重與輸出
- 偵測系統性蒸餾
- 平衡體驗與防禦

開放 vs 保護

企業使用者

- 注意 ToS (使用合約) 的輸出再利用限制
- 避免在微調 / 合成資料場合違反使用合約

創新 vs 控制

輸出
Output
=
攻擊面

研究自由 vs 商業防禦

政策制定者

- 補足輸出法律地位
- 建立跨國協作與執行機制

學術與開源社群

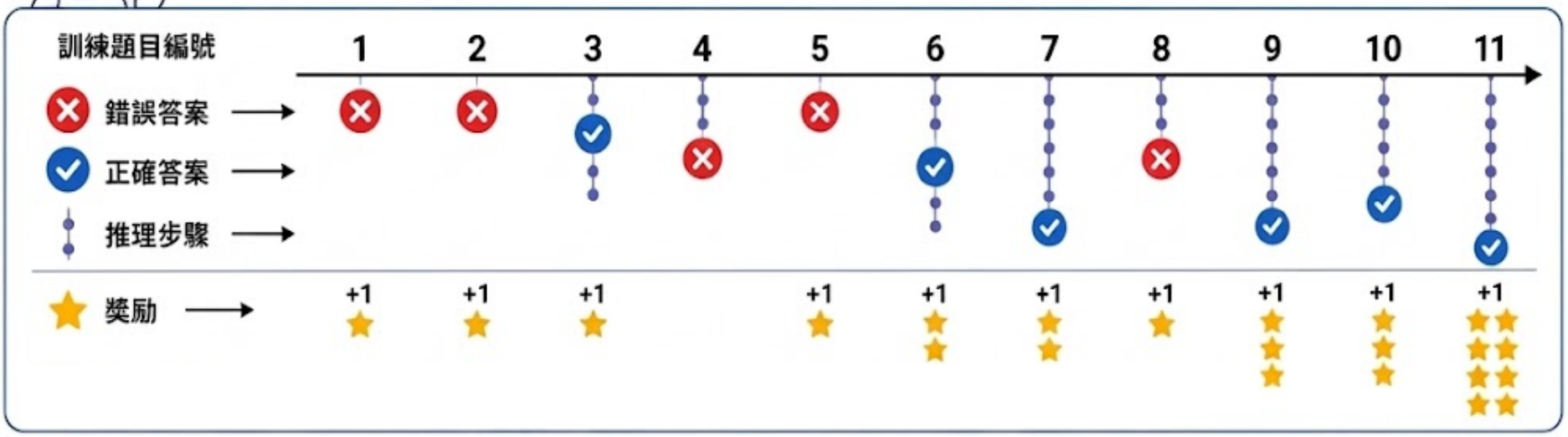
- 維持研究可得性
- 避免防禦過度壓縮開放創新

CoT思維蒸餾
爭議與防禦實例

AI 透過試錯學會展現思考過程

Ippolito & Zhang, 2025

以 DeepSeek-R1 為例：強化學習如何讓大型語言模型更傾向先寫推理步驟，再給出答案



2 為什麼會學會推理？

- 在強化學習中，模型只根據答案對錯得到回輸。
- 因為「先推理再作答」更容易答對，
- 模型便逐漸養成這種行為。

3 學到的能力

- 自我檢查
- 自我反思
- 嘗試不同解法
- 在作答前驗證結果

4 限制與提醒

- 推理可能變得很長
- 有時不易閱讀
- 對有標準答案的任務效果最好
- 仍需搭配監督學習提升可讀性

強化學習的學習循環

題目輸入 → 模型嘗試作答 → 產生推理步驟 → 根據對錯給獎勵/懲罰 → 更新模型 → 更常先思考再回答

模型強化學習訓練過程發現寫出CoT有助於答對獲得高獎勵

DeepSeek R1 高效訓練流程

Guo et al., 2025

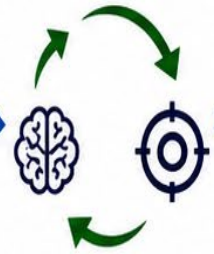
① DeepSeek-V3 基礎能力

- 671B MoE（37B 啟動）
- 14.8T tokens 預訓練
- 接近 GPT-4o / Claude 3.5 Sonnet 水準

② R1-Zero：純強化學習（無 SFT、無外部 CoT）

V3-base（671B MoE）

GRPO 強化學習



獎勵函數

- 準確性獎勵：答案是否正確
- 格式獎勵：<think>...</think> <answer>...</answer>

模型自發產生「aha moment」：自我反思、修正錯誤

優點：證明不必依賴教師示範也能激發推理

限制：可靠性差、格式不穩定

③ 正式版 R1：冷啟動 + 多階段精煉

1. 冷啟動 SFT
   
   數千筆長 CoT 範例
   來源未明確公開
2. 推理導向 RL
   
   準確性 + 格式 + 語言一致性獎勵
3. 拒絕採樣 + 全任務 SFT
   
   擴展到寫作、QA、安全
4. 全場景 RL
   
   規則獎勵 + 偏好對齊

完成最終對齊

爭議焦點通常落在「冷啟動 SFT 的長 CoT 範例來源」。

關鍵爭議點：
階段 1 的「數千筆長 CoT 範例」來源為何？

輸出：DeepSeek-R1（最終模型）

「強 base model + 純 RL 驗證 + 多階段精煉」的蒸餾模型研發。

蒸餾技術議題: DeepSeek 事件

1 2024 秋

微軟資安團隊據稱觀察到異常 API 匯出流量

技術

2 2024/12/26

DeepSeek-V3 發布與全面開源

技術

3 2025/01/20

DeepSeek-R1、R1-Zero 發布；6 個蒸餾模型同步開源

技術

4 2025/01/27

市場震盪：Nvidia 市值大跌；DeepSeek App 熱度飆升

市場

5 2025/01/28–29

David Sacks 表示有「實質證據」

政治 / 政策

6 2025/02

媒體報導：API 流量在訓練週期飆升（匿名消息）

市場

7 2025/02/03

OpenAI / Anthropic 收緊 API 限制與身分驗證

政治 / 政策

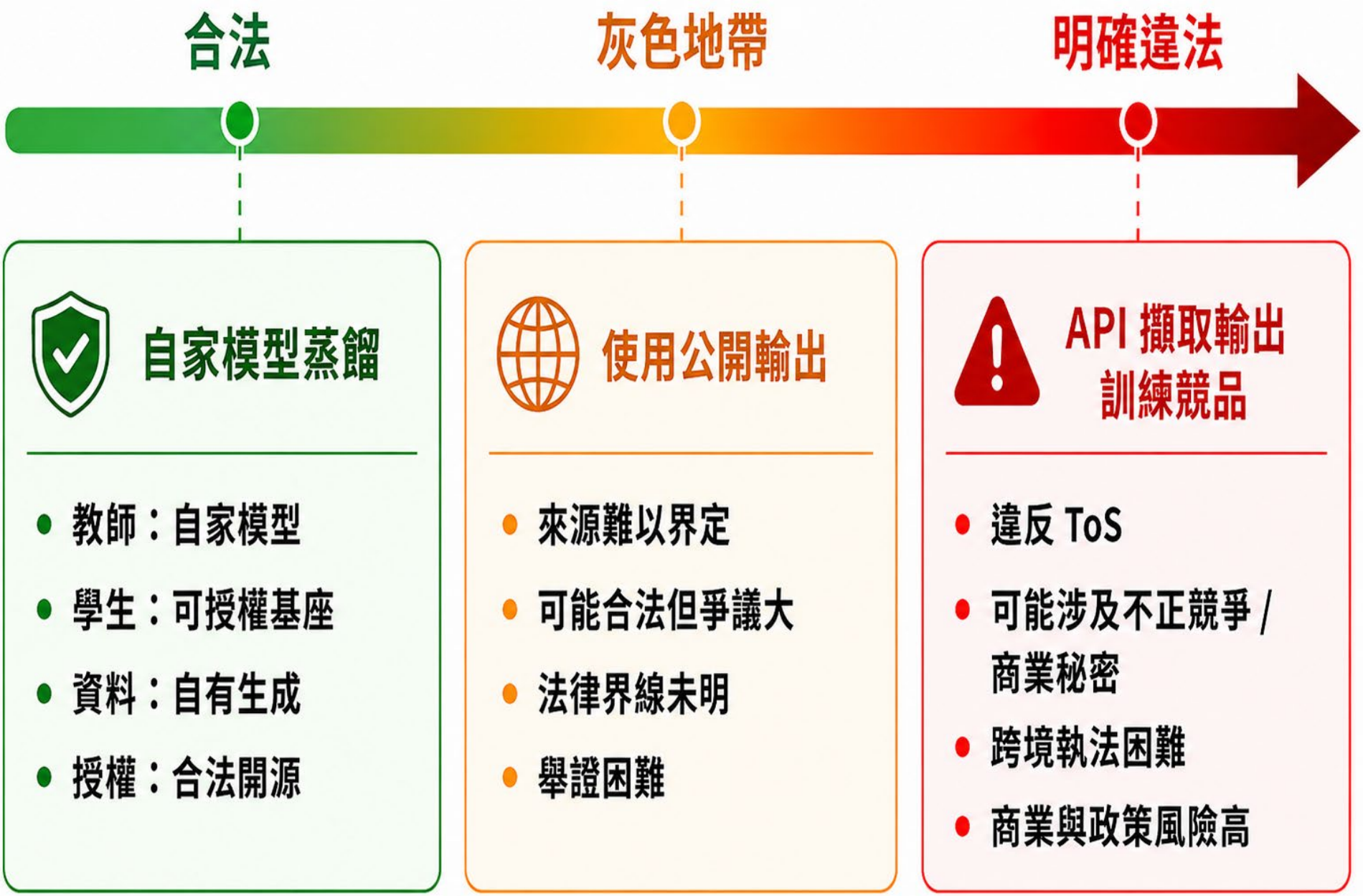
8 2025–2026

國會聽證、出口管制延伸、產業全面升級防禦

產業防禦

事件從**技術**發布，迅速升高為**市場震盪**、**政策議題**與**產業安全議題**。

蒸餾知識進化 vs CoT攻擊盜取智慧



智慧蒸餾技術 vs 思維鏈盜取

DeepSeek 蒸餾技術開發

1. DeepSeek-R1 作為教師模型
2. 建置推理語料
   約 60 萬筆 CoT；約 20 萬筆通用對話
3. 蒸餾至 Qwen／Llama
   等開源基底模型
4. 公開釋出模型權重與技術報告

來源明確 | 授權可稽核 | 技術報告 | 可重現

自家模型蒸餾著重授權、模型來源與釋出條件之可檢視性；其法律評價應與資料來源爭議分別判斷。

VS

OpenAI／Microsoft 所稱之黑箱萃取

1. 經由 API 大量擷取模型輸出
   受指控
2. 輸出資料可能被整理為訓練語料
3. 供 R1 或相關模型訓練使用
   待查證

事實爭議 | 證據未公開 | 授權不明 | 舉證責任

爭點集中於 R1 上游語料是否含有第三方專有模型輸出；其取得方式是否逾越服務授權，仍待證據釐清。

比較基準

合法蒸餾強調授權、來源透明與可稽核性；
黑箱萃取涉及資料來源、授權範圍與舉證責任。

技術精進 vs 智慧盜取: 金牌咖啡實例

蒸餾技術知識精進

自己的筆記，訓練自家員工

合法學習的流程：建立內部價值，長期成長

1. **記錄與整理**
   將自己的經驗整理成筆記
2. **內部教學**
   傳承技術與思路，培養團隊能力
3. **團隊成長、品質提升**
   凝聚共識，提升服務與品質
4. **打造風格、合法經營**
   建立品牌特色，穩健經營

合法：使用自己的經驗與筆記，訓練自家員工，創造價值。

蒸餾攻擊CoT盜取

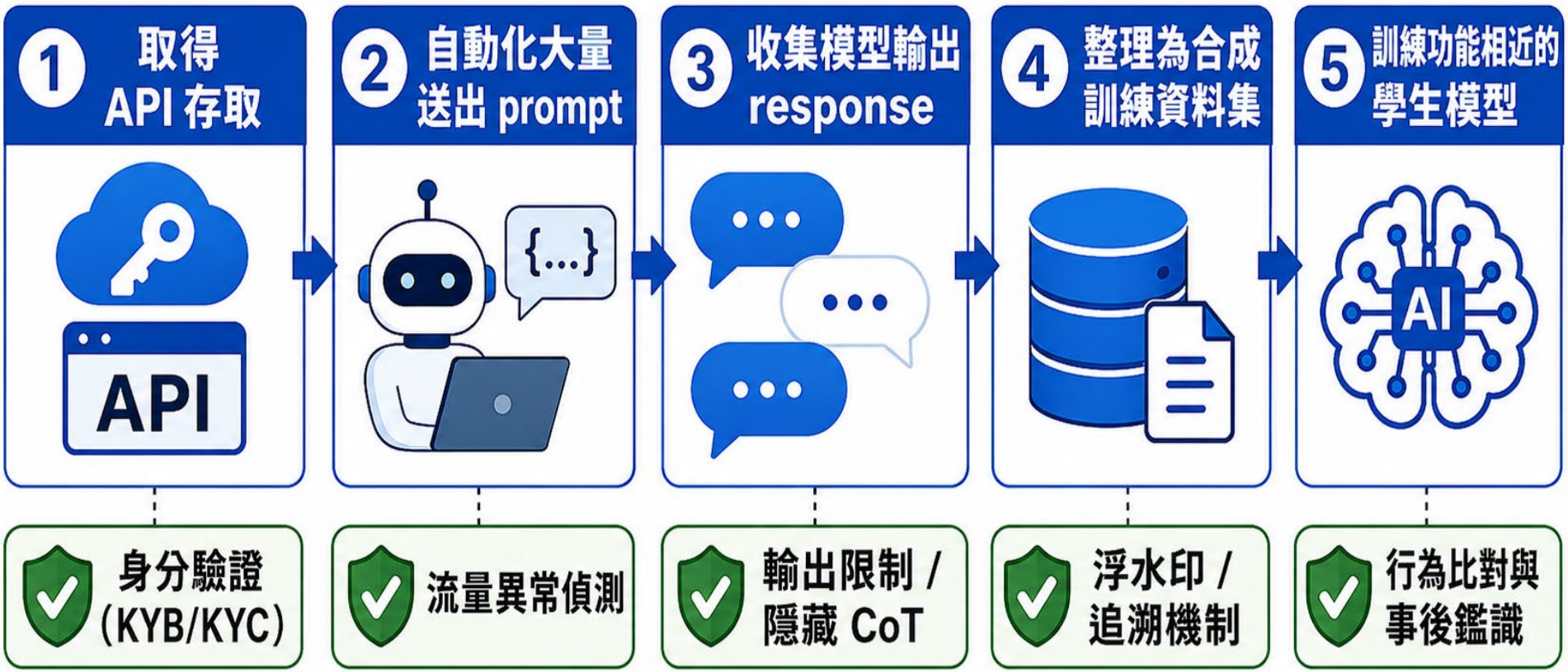
偷錄他人思路，帶回競爭店

攻擊行為的流程：竊取價值，破壞信任

1. **偷錄思路**
   未經同意，記錄他人的教學思路
2. **外部傳授**
   將偷來的思路傳授給他人
3. **競爭店受益**
   讓競爭店獲利，損害原創者權益

攻擊：將他人教學思路帶回競爭店使用，已構成對原創者的攻擊與背信。

智慧模型蒸餾攻擊思維鏈盜取



特點：全程可能只使用「正常付費 API」，不需要入侵伺服器。

在 LLM 時代，「輸出」本身就是重要的攻擊面。

智慧模型四重防禦架構

多層協同防護架構

L1 存取控制 / API 層防禦

- 限制速率 (Rate limiting / 多軌限流)
- 身分驗證 (KYC / KYB)
- 異常查詢模式偵測
- 地理 / IP / 管轄限制

L2 偵測與監控

- 異常流量偵測
- 行為模式分析 / 行為指紋比對
- 大批量取樣偵測
- 即時告警
- 公開語料反污染
- Bug bounty 與社群通報

L3 輸出水印 / 控制

- 輸出格式控制
- 浮水印
- 輸出加噪
- 隱藏 CoT
- 來源追溯
- 降低教師訊號品質

L4 法律與契約

- 強化 ToS
- 高量客戶額外契約
- 違約追償
- 合作調查機制
- 出口管制延伸
- 跨國執法 / 協作倡議

目標：提高攻擊 / 蒐餾成本、降低成功率與可行性、增加追溯與威懾能力

有效防禦需 技術-法規合約-即時監控協同策略

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