星球永續健康線上直播

智慧數位資安（9）

智慧模型思維鏈(CoT)知識蒸餾攻擊

2026 年 5 月 27 日

AI 時代思維鏈（Chain of Thought, CoT）推理機制已成為提 AI 推理能力的重要核心。思維鏈之於智慧模型如同冠軍咖啡師對水溫、研磨與風味調整的完整思維歷程所得到的金牌咖啡。一旦這些由經驗技術累積的思維被盜取，競爭者便可能快速複製其核心能力。這樣的攻擊型態逐漸出現在生成式 AI 與大型語言模型的發展之中。本週我們將探討智慧模型思維鏈（CoT）盜取攻擊，以及 CoT 思維蒸餾的爭議與防禦實例。

健康科學新知

中非動亂加劇伊波拉疫情危機:「疫戰交纏」

剛果民主共和國東部近日爆發由邦迪布焦伊波拉病毒引發之嚴重疫情。此病毒株不同於該國過去較常見的薩伊型，目前既無已核准疫苗，亦缺乏專門針對此型別之治療藥物，使防疫難度明顯升高。疫情最早集中於伊圖里省，隨後擴及北基伍、南基伍及戈馬等地，並已出現傳入烏干達之病例。截至目前，疑似死亡人數已達 177 人、疑似病例 750 例，相較一週前之 65 例死亡、246 例疑似病例，顯示此波疫情蔓延極為迅速。世界衛生組織已將此次疫情列為「國際關注的公共衛生緊急事件」(PHEIC)，意即疫情已具備跨境擴散風險，需要國際協調、資源投入與鄰近國家共同加強防備。

疫情迅速惡化的成因錯綜複雜，偵測與確認延誤為首要問題。世界衛生組織直至 5 月 5 日方獲知疑似病例，其後又歷經數日方取得實驗室確認；在此空窗期內，病毒已可能透過醫療院所、家庭照護與葬禮活動於社區內擴散。更值得注意的是，當地初期檢驗工具主要針對薩伊型伊波拉病毒，未能即時辨識本次實際流行之邦迪布焦型，導致防疫團隊在啟動接觸者追蹤、病患隔離與社區宣導之前，已落後疫情傳播速度。

喪葬習俗所形成之傳播風險進一步加劇疫情擴散。伊波拉患者死亡後遺體仍具高度傳染性，然部分社區傳統葬禮涉及觸碰、清洗或近距離接觸遺體，一旦死者實為伊波拉

感染者，葬禮即可能成為集體感染事件。此外，部分民眾初期將死亡歸因於神秘疾病、巫術或宗教因素，未立即通報衛生單位，使病毒得以在社區中持續傳播。此一現象說明，伊波拉防治不僅為醫療問題，亦深度牽涉地方文化、宗教信仰與社區信任。然而，本次疫情最為複雜之結構性背景，在於疫情發生地點長期受武裝衝突影響。伊圖里省與北基伍等地長年面臨武裝團體活動、人口流離失所、道路阻斷及醫療基礎設施不足等困境，疫情起點又鄰近烏干達與南蘇丹邊境，人口流動本即頻繁。安全風險阻礙防疫人員進入熱區，致接觸者追蹤與病患隔離延誤；衝突造成居民流離，病患、家屬與接觸者可能沿邊境或難民流動路線跨區移動，使公共衛生單位難以建立完整病例清單；外來醫療團隊在部分社區被視為外人，社區不信任更進一步削弱防疫合作成效。WHO 總幹事譚德塞曾就 2018 至 2020 年剛果東部疫情指出，反覆的武裝攻擊迫使防疫行動中斷，形同讓病毒取得優勢，並形容「不安全本身也是一種病毒」。此一背景雖源自前次疫情，然其結構性制約在本次疫情中依然存在，說明衝突地區的防疫工作必須同時處理治安、社區信任、政治協調與人道通道，而非單純將醫療資源送入疫區即可。

本次剛果東部伊波拉疫情係病毒學特性、偵測延誤、文化習俗與長期衝突地帶結構性脆弱性多重因素交織所致。WHO 提醒，目前對實際感染人數與擴散範圍仍存在重大不確定性，疫情走向仍需密切監控。

美伊協議受阻於濃縮鈾議題:「核海博弈」

美伊核談判目前陷入關鍵僵局。川普政府對核協議的立場轉趨謹慎，談判雙方據稱仍在持續接觸，但川普收回先前稱協議「已基本達成」的說法，改稱協議「尚未完全談判完成」，顯示核心分歧尚未化解。美方的底線極為明確：伊朗必須放棄高濃縮鈾，官員直言「沒有高濃縮鈾處理細節，就沒有交易」；此外，霍爾木茲海峽的重新開放亦為美方不可退讓之條件，國務卿魯比歐明確表示，若伊朗企圖控制海峽或向通行船隻收取費用，將是美方無法接受之結果。川普並警告，若伊朗拒絕協議，美國可能恢復軍事打擊。雖然雙方據稱已就涵蓋核問題與霍爾木茲海峽之框架原則上取得共識，協議完成度約達 95%，但最關鍵的核條款與協議文字仍未敲定，伊朗資產解凍時機亦存在爭議——

美方希望待伊朗在核問題上取得實質進展後再分階段解凍，伊朗則期望在協議生效初期即解除凍結。伊朗外交部發言人亦坦言，雙方在高濃縮鈾庫存、霍爾木茲海峽、美國海上封鎖與全面停戰等議題上分歧仍深。伊朗提出的和平條件範疇遠超核問題本身，德黑蘭要求結束所有戰線上的敵對行動、美軍撤出伊朗周邊地區、美國與以色列對戰爭造成之破壞進行賠償，同時要求解除制裁、釋放凍結資產並停止海上封鎖。換言之，伊朗試圖將核談判、戰爭賠償、區域安全與海上封鎖納入同一談判框架，使整體協議的複雜程度大幅提升。

霍爾木茲海峽爭議已成為本波危機的核心焦點。美國持續執行對伊朗相關港口與船隻的封鎖行動，並稱已轉向近百艘商船；伊朗則批評此舉違反國際法，並指封鎖影響基本物資、藥品與民生供應。在封鎖壓力下，伊朗已開始尋求替代貿易路線，阿曼穆桑代姆半島的哈薩卜港成為重要繞道節點，部分原本經由阿聯酋進入伊朗的貨物，改由非伊朗旗船運抵哈薩卜後再轉運回國，貨物涵蓋汽車、家電、消費品與部分石油相關產品。然而這條路線代價高昂，運輸成本據商人稱約為過去的六倍，顯示伊朗雖仍維持部分貿易命脈，但流程更為不穩定。歐盟亦已宣布，若伊朗持續阻斷或控制霍爾木茲海峽，將對相關官員施加制裁。

區域安全風險亦外溢至海灣國家。阿布達比巴拉卡核電廠附近發生無人機攻擊，一架無人機擊中廠區內圍外之發電設備並引發火災，當局表示無人員傷亡且未影響輻射安全，但阿聯酋形容此事為「危險升級」，國際原子能總署亦對任何威脅核安全之軍事活動表達嚴重關切。在外交斡旋層面，巴基斯坦陸軍參謀長穆尼爾訪問德黑蘭，與伊朗外長及相關官員討論防止局勢升級之外交途徑。卡達代表團亦同步赴德黑蘭加入調停。然而巴基斯坦消息人士指出，美伊雙方「不斷改變目標」，使談判窗口愈趨緊迫。

俄烏衝突升溫牽動歐美外交:「戰談並進」

烏俄衝突近週出現明顯升級態勢。烏克蘭對俄羅斯本土發動大規模無人機攻勢，攻擊範圍涵蓋莫斯科周邊、多個俄羅斯本土地區、克里米亞及黑海、亞速海區域，目標集中於煉油廠、油庫、輸油設施與軍工相關供應鏈，造成住宅、基礎設施與能源設施受損

及人員傷亡。烏克蘭的戰略意圖明確：透過持續打擊削弱俄羅斯能源收入、擾亂軍工運作，並使俄羅斯民眾與政治高層感受到戰爭已不再局限於烏克蘭境內。俄羅斯隨即於 5 月 23 日對烏克蘭發動大規模夜間空襲，主要目標為基輔，動用約 90 枚飛彈與 600 架無人機，造成至少 4 人死亡、約 100 人受傷，基輔多處住宅、學校、博物館與基礎設施受損，其中車諾比博物館嚴重毀壞。俄方聲稱此次攻擊係回應烏克蘭對俄控區的襲擊，並否認針對平民設施。歐洲多國領袖強烈譴責，並特別關注俄方疑似使用可攜帶核彈頭之高超音速飛彈，衝突層次的擴大引發歐陸國家高度警惕。在軍事升級的同時，外交層面亦出現值得關注的動向。自 2022 年俄羅斯全面入侵烏克蘭後，歐盟已切斷與莫斯科的正式溝通管道。隨著美國主導的和平斡旋進展有限，歐洲日益憂慮自身在未來和平安排中遭到邊緣化，若談判最終由美俄主導，歐盟缺乏明確代表與共識立場，可能被迫接受不利於自身安全、亦不利於烏克蘭主權與領土完整的結果。為此，歐盟內部開始討論是否應重建與俄羅斯的正式溝通管道，並推派代表歐洲立場之人物與普丁接觸。被討論的人選包括歐洲央行前行長、義大利前總理德拉吉，以及德國前總理梅克爾。德拉吉因背景穩健、在歐盟內部具較高聲望，獲部分烏克蘭人士支持；梅克爾雖具備對俄談判經驗，但她本人認為真正能被普丁認真對待的應是現任歐洲領導人。普丁曾提議由德國前總理施洛德擔任協調人，惟因其與俄羅斯能源產業關係密切，遭歐洲與基輔方面明確排斥。此一爭議折射出歐洲在選擇談判代表時，必須同時兼顧對俄溝通能力、政治可信度、烏克蘭接受度與歐盟內部團結，各方立場仍存在明顯分歧。

烏克蘭方面明確希望歐洲更積極介入談判。澤倫斯基強調，歐洲必須確保歐洲利益與烏克蘭利益在和平進程中同時獲得保障，基輔希望歐洲能推動俄羅斯接受以當前前線為基礎的立即停火，但俄羅斯過去已拒絕類似方案。在制度安排層面，德國提出可考慮給予烏克蘭某種歐盟「聯繫成員」或「準成員」地位，使其能更深入參與歐盟機構運作、透過歐洲整合獲取政治與安全保障，惟暫不享有完整投票權。

中俄深化結盟 日韓加速聯防:「能源結盟」

中東局勢僵持、霍爾木茲海峽航運受阻，正推動亞洲能源格局加速重組。俄羅斯總

統普丁訪問北京期間，「西伯利亞力量 2 號」天然氣管道成為俄中峰會核心議題。對俄羅斯而言，此管道是失去歐洲市場後重建出口收入的重要出路；對中國而言，俄羅斯陸上天然氣則有助於降低對海上能源通道的結構性依賴。該管道全長 2,600 公里，連接亞馬爾半島、穿越蒙古抵達中國，建設路線已獲三國確認，預計 2026 年啟動施工、最快 2030 年通氣，但雙方在定價與購買量上仍存分歧，中國掌握較大談判籌碼，協議落地尚存變數。俄中能源合作亦折射出兩國在地緣政治上的持續靠近，俄羅斯因制裁更加依賴中國市場，中國則將俄羅斯視為能源供應與戰略夥伴，雙方關係雖非完全對等，但均能從中汲取戰略利益。另一方面日本與韓國亦因霍爾木茲危機加強雙邊合作，兩國同意共享石油儲備、精煉產品供應並建立液化天然氣互換安排。此次快速達成協議，既是對當前危機的直接回應，也顯示日韓在地緣政治壓力下轉向務實合作。能源問題已超越市場供需範疇，與戰爭、制裁、海上通道安全及大國競爭深度交織，俄中管道協議能否落實、日韓合作能否持續深化，將共同影響亞洲能源安全與區域地緣政治走向。

油運危機與 AI 時代算力競爭:「算油雙控」

全球經濟正面臨能源不穩定與 AI 算力稀缺的雙重挑戰。隨著 AI 擴張，算力已成為「新石油」，Nvidia 的 H100 等晶片因需求遠超供給而租金上漲，顯示 AI 基礎建設已成為推動全球工業擴展的新動能。即便領先，Nvidia 仍受地緣政治壓力影響，在中國市場的份額正逐漸被華為取代。然而，執行長黃仁勳仍布局規模達 2,000 億美元的 Vera CPU 市場，並持續仰賴台灣供應鏈以維持競爭力。專家分析，未來 AI 競爭將與能源成本、供應鏈瓶頸及地緣風險深度連結。若雲端巨頭支出放緩或競爭者技術成熟，市場恐面臨供需修正。在算力與能源交疊的新時代，投資人需在科技成長與風險間尋求平衡。

科學家將重返福島:「災後重生」

福島核災 15 年後，日本政府於曾受創嚴重的浪江町成立「福島國際研究教育機構」，致力將災區轉型為災害復原與環境修復的科學中心。該機構研究涵蓋機器人、農業及輻射醫療應用，預計 2028 年啟用總部，2030 年全面運作。F-REI 不僅是技術基

地，更肩負重建公眾信任與引導居民返鄉的重任。然而，復興之路仍面臨挑戰：以浪江町為例，目前僅約 17% 居民返回，調查顯示逾半數撤離者仍無返家意願。專家強調，F-REI 的長期成功取決於其獨特的科研定位，以及能否在偏鄉環境中持續吸引國際頂尖人才，將核災記憶轉化為創新的正面能量。

戰火下伊朗研究人員困境:「烽火摧知」

自 2026 年 2 月伊朗戰爭爆發以來，其學術體系遭受毀滅性衝擊。據報導，包括沙里夫理工大學與巴斯德研究所在內的約三十所機構遭轟炸，導致長年累積的實驗手稿、學生論文與逾千本藏書等知識資產付之一炬，被學者形容為真正殘酷的暴行。受戰火與網路封鎖影響，研究人員無法存取 CERN 資料，與海外合作被迫中斷；同時，被監禁於德黑蘭監獄的學者正隨戰事惡化，在物資匱乏與空襲威脅下艱難求生。目前已有逾千名國際學者連署譴責對民用學術設施的攻擊。專家憂心，若攻擊研究機構成為戰爭常態，將對全球文明造成比建築損毀更深遠的破壞。

Vibe Coding 科研機會與風險:「人機共研」

科學研究正迎來「Vibe Coding」浪潮，研究者僅需透過自然語言與 AI 編碼工具對話，便能快速生成程式碼、資料工作流或視覺化圖表。這種模式讓氣候學者能製作出具衝擊力的立體氣溫螺旋圖，也大幅降低了非資訊背景研究員處理資料的門檻，讓科學構想能在短時間內轉化為可運行的雛形。然而，便利背後隱藏風險。研究指出，AI 生成的程式碼可能存在邏輯錯誤或誤用統計方法，且「Vibe Debugging」過程往往混雜難辨。專家強調，AI 雖能加速實驗，但不能取代對運算邏輯的理解。研究成果仍須經過嚴謹的人工審查、可驗證測試與同儕檢查，才能確保科學誠信，避免將研究速度建立在錯誤的技術基礎上。

智慧模型思維鏈(CoT)盜取攻擊

電影《頂尖對決》(The Prestige) 描述兩位魔術師之間從合作到競爭、再到彼此對抗的過程。電影中不只是魔術表演，更是在呈現「思維」、「設計」與「秘密」之間的較量。片中兩位主角安杰與波登原本是同門學徒，彼此都在追求更高層次的魔術技巧。

其中，一位重要角色是機關師法隆，他掌握魔術背後真正的機關與秘密。每一場魔術表演，都包含三個步驟，第一步「以虛代實」，是先讓觀眾相信眼前看到的只是平凡事物；第二步「偷天換日」，則透過巧妙手法完成消失、轉移或欺騙；而真正關鍵的第三步「化腐朽為神奇」，則是將一切重新呈現，讓觀眾產生震撼與驚奇。這樣的概念，其實與今天大型語言模型中的 CoT 非常相似。真正重要的，不只是最後輸出的答案，而是模型在背後如何一步一步推理、規劃與完成問題拆解。安杰與波登後來因一次水箱逃脫魔術意外而反目成仇。安杰的妻子在表演過程中溺斃，安杰認為波登在機關與結繩設計上動了手腳，導致悲劇發生。從此，兩人從同門夥伴轉變為彼此竊取秘密、互相破解魔術的競爭對手。拆夥之後，兩人開始走向截然不同的魔術道路。波登專精於魔術設計與創新研發，擅長構思巧妙的魔術原理，但他的舞台魅力較弱，即使表演內容精彩，觀眾往往難以真正感受到其中的震撼。安杰則擅長舞台表現與觀眾魅力，能成功營造氣氛、吸引目光，但在魔術創新能力上則不如波登。兩人逐漸形成互相競爭、彼此破解魔術的對立關係。安杰試圖破解波登的魔術秘密；波登則不斷干擾安杰的演出，讓對方難以建立聲望與名氣。兩人的競爭，從舞台表演逐漸演變成對「秘密」與「思維」的爭奪。其中最震撼的一場表演，便是波登所發明的「移形換位術」。在不到一秒鐘內，他能從舞台一側瞬間消失，又從另一側現身，速度之快幾乎無法理解，讓安杰深受震撼。即使安杰試圖以替身方式模仿，仍始終無法破解其中真正的秘密。

安杰與柯恩研究破解方法。柯恩認為「移形換位術」唯一合理的解釋，就是使用替身也就是舞台另一端出現的人，其實並非魔術師本人，而是事先安排好的替身。因此，安杰也開始利用替身，嘗試模仿波登的移形換位表演。然而，即使成功仿效表演形式，安杰始終認為波登真正的秘密並不只是替身，而是背後更深層的設計思維與執行邏輯。他開始執著於破解波登如何完成這場魔術，甚至想知道他究竟是如何「思考」出這套系統。安杰知道，每位魔術師都會留下自己的魔術手札與筆記，因此他設法竊取波登的魔術日記，希望從中找到移形換位術的真正秘密。但即使取得筆記，內容仍充滿暗號與隱藏訊息，缺乏解密金鑰，依然無法真正理解波登的思維方式與設計邏輯。之後，安杰甚

至透過法隆取得一個關鍵字「TESLA」。波登刻意留下這個線索，誘使安杰遠赴美國，投入大量資金尋找特斯拉，希望打造真正能完成移形換位的科技機關。電影中兩位魔術師彼此競爭、互相竊取秘密與破解思維的過程，其實與今日大型語言模型的 CoT 蒸餾攻擊極為相似。

思維鏈（CoT）蒸餾攻擊中，攻擊者鎖定的目標是模型背後的推理能力與思考邏輯。攻擊者通常僅具備 API 查詢權限，因此會刻意設計大量推理型任務，引導教師模型輸出逐步的推理軌跡。這些 CoT 輸出，包含問題拆解、條件判斷與邏輯推演過程，具有高度訓練價值。當攻擊者持續蒐集這些內容後，便可用於微調學生模型，使其不只是學會答案，而是學會「如何推理」。CoT 之所以成為高價值攻擊目標，在於推理軌跡具備可遷移性。學生模型取得這些推理資料後，不僅能模仿原有能力，甚至可能延伸至未見過的新問題。也因此，真正珍貴的，不再只是模型最後輸出的答案，而是模型內部的推理結構與決策過程。這樣的概念，與《頂尖對決》中魔術師彼此竊取秘密極為相似。安杰真正想取得的，並非波登表演後呈現的結果，而是背後瞬間移動的設計邏輯與執行思維。即使能模仿舞台效果，若無法掌握核心推理方式，仍無法真正重現魔術本質。在大型模型中亦是如此。當模型完整輸出推理過程時，等同於逐步暴露內部思考軌跡。尤其推理模型大量依賴條件推演與反向推理，例如貝氏推理中的條件機率反推，皆可能成為攻擊者進行蒸餾與模仿的重要素材。因此，防禦重點已從單純保護資料，轉向保護模型的推理能力。常見方式包括限制 CoT 長度、以摘要化方式取代完整推理、偵測異常誘導查詢、監控大量重複 prompt，以及分析異常 API 使用模式等。此外，也可加入干擾（noise）與混淆機制，降低推理軌跡被完整還原的風險，或透過差異化輸出與權重保護，提高反推模型的困難度。

以「金牌咖啡」為例，CoT 蒸餾攻擊並不是單純取得一杯咖啡的最終配方，而是試圖學走冠軍咖啡師背後完整的判斷與調整過程。一杯好咖啡的形成，並非只靠固定數字或單一配方，而是經過聞香、沖煮、試喝、判斷與調整等連續步驟。例如第一杯可能太酸，代表萃取不足，因此需要提高水溫或調整研磨；第二杯若過乾澀，則可能代表萃取

過度，需要降低水溫或改變研磨粗細。經過多次修正後，第三杯才逐漸達到酸甜平衡、口感圓潤的理想狀態。這一連串從觀察、判斷、修正到驗證的過程，就是咖啡師真正有價值的「思維鏈」。真正值得學習的，並不只是配方表上的數字，而是面對不同狀況時，如何判斷問題、調整參數，並逐步接近最佳風味。若將這個概念放回大型語言模型，CoT 蒸餾攻擊的目標也不是單純複製答案，而是透過輸出格式、任務誘導與多輪查詢，取得模型背後的推理軌跡。攻擊者一旦蒐集到足夠的 CoT 資料，便可能用來訓練學生模型，使其學會類似教師模型的問題拆解與推理能力。

CoT 蒸餾攻擊的重要特徵，在於透過角色設計、任務重構與長鏈推理誘導，逐步將原本隱藏於模型內部的推理軌跡外顯化。以金牌咖啡為例，攻擊者只是詢問配方，更要求模型逐步解釋如如何判斷咖啡香氣、如何調整水溫、如何控制研磨粗細、如何修正酸味與萃取問題產生金牌咖啡思維步驟。當模型開始詳細說明每一步判斷與修正原因時，便逐步暴露其內部推理流程。這類攻擊包含三種典型模式。第一，是「假裝請教」。攻擊者以學習者角色，要求模型逐步說明每個步驟背後的原因，例如：「為什麼這樣沖煮？為什麼需要調整水溫？」藉此取得完整推理過程。第二，是「重寫任務」。原本簡單的問題，會被重新定義為需要完整解說的任務，例如要求模型提供詳細步驟、判斷依據與修正原因，使推理內容自然外顯。第三，是「拉長推理」。攻擊者刻意延長推理鏈，例如從第一杯、第二杯到第三杯持續要求修正與優化，讓模型不斷輸出更多判斷與調整細節，藉此累積高價值 CoT 樣本。當大量相似查詢被持續蒐集後，便能逐步建立教師模型的推理資料庫。整個蒸餾流程通常包含：查詢集設計、推理軌跡擷取、資料清理、蒸餾微調，以及最後的能力驗證。經過整理後的 CoT 資料，可用於訓練學生模型，使小模型逐漸學會大型模型的推理能力與問題拆解方式。因此，攻擊者真正想取得的，並不是單一答案，而是模型長期累積的推理結構與決策模式。這與《頂尖對決》中魔術師彼此竊取「魔術思維」如何判斷、設計與執行的過程相似。

傳統蒸餾攻擊只學習最終答案與配方。本質上屬於「問題→答案」的映射，因此當遇到新問題、新情境或新資料時，往往容易失效。就像只記住咖啡最終配方，卻不知道

背後如何判斷酸味、甜味、口感與萃取平衡，一旦更換新的咖啡豆或沖煮條件，便可能無法應對。真正高價值的，其實是完整思路。從品嚐咖啡香氣、判斷問題、調整方向、修改參數，到最後驗證結果，整個過程代表的是一套可遷移的推理能力，而不只是固定答案。這也是 CoT 蒸餾攻擊與傳統蒸餾最大的差異。因此，CoT 蒸餾攻擊的核心，並不是單純複製模型輸出，而是試圖竊取教師模型背後的推理結構與思考方式。當攻擊者掌握完整推理流程後，學生模型便可能具備面對新問題時的推敲與泛化能力，而不只是機械式模仿。這也是目前大型語言模型最重要的資安挑戰之一。若大量模型透過 CoT 蒸餾方式互相模仿，可能衝擊原本依賴長期訓練與創新所建立的核心能力。因為真正珍貴的，不只是輸出的答案，而是模型背後如何形成判斷、推理與決策的過程。因此，在 AI 發展過程中，如何在「開放」與「保護」之間取得平衡，成為重要議題。模型供應商需要建立 CoT 蒸餾偵測、異常查詢分析與權重保護機制；企業使用者則需注意模型輸出的再利用與微調是否涉及違反使用合約；政策制定者也必須逐步建立跨國協作與法律規範。然而，若限制過度嚴格，也可能壓縮研究自由與開源創新空間。因此，如何同時兼顧資安保護、商業利益與技術創新，正是當前大型模型治理中最困難的核心問題之一。

CoT 思維蒸餾爭議與防禦實例

AI 研究顯示智慧模型在強化學習訓練中，若先進行推理步驟 CoT 再回答，更容易獲得較高獎勵並提升答題正確率。以 DeepSeek-R1 為例，模型會在反覆試錯中逐漸學會自我檢查、反思與驗證答案，形成類似「思考後再作答」的行為模式。不過，研究也提醒，過長推理可能影響可讀性，仍需搭配監督與安全機制。DeepSeek-R1 的高效訓練流程，結合大型基礎模型、純強化學習與多階段精煉技術，提升 AI 推理能力。研究指出，R1-Zero 在無監督微調與無外部 CoT 範例下，仍能透過強化學習自發形成「aha moment(頓悟時刻)」，學會反思與修正錯誤；正式版 R1 則再加入冷啟動 SFT 與全場景 RL，進一步改善穩定性與實用性。

DeepSeek 事件從智慧模型訓練創新技術發布演變為牽動市場、政策與產業安全的全球議題。2024 年底，微軟資安團隊觀察到異常 API 流量，引發外界對模型蒸餾與資

料來源的討論。隨後 DeepSeek-V3 與 R1 系列模型陸續開源，以高效能與低成本特性快速引發關注，也對既有 AI 產業格局造成衝擊。2025 年初，市場因 DeepSeek 熱潮出現劇烈波動，NVIDIA 股價受到影響，多家科技公司也開始重新檢視 API 權限與模型安全機制。此事件不再只是技術競爭，而逐漸上升至國際政策、出口管制與 AI 產業防護層面，反映出大型語言模型已成為科技戰略與國家競爭的重要核心。「模型蒸餾」原本是人工智慧領域中常見的知識壓縮技術，其核心概念是讓較小模型學習大型模型的輸出能力，以降低運算成本並提升推論效率。在合法智慧模型訓練情況下蒸餾建立於自有模型、授權資料或開源架構之上，因此被視為正常的技術演進。然而，隨著大型語言模型商業價值提升，部分企業開始擔憂競爭對手可能透過大量 API 輸出蒐集模型回答，進一步模仿其推理能力與 CoT 結構。此類行為可能涉及違反服務條款、商業機密與不公平競爭問題，也使 AI 產業逐漸從「開源共享」轉向「模型防禦」。目前相關法律界線仍不完全明確，但各大科技公司已開始加強 API 限制、身分驗證與輸出保護機制。

DeepSeek 所公開描述的蒸餾流程，以自主研發模型作為教師模型，建立推理語料後，再蒸餾推進智慧能力，並公開技術報告與模型權重，強調來源透明、授權可檢核與技術可重現性。相對地，OpenAI 與 Microsoft 提出之爭議點，則集中於是否透過 API 大量擷取模型輸出，再將其整理為訓練資料，用於其他模型訓練。若涉及未授權擷取、繞過服務條款或重建特定推理能力，便可能落入商業機密與智慧財產爭議。由於大型語言模型的真正價值，已不只是答案本身，而是背後的推理結構與思維模式，因此 CoT 是否屬於可受保護的「智慧資產」，正逐漸成為 AI 法律與產業競爭中的核心問題。

以金牌咖啡對照：一是「蒸餾技術知識精進」，用自己的筆記與經驗在內部訓練自家員工，透過合法學習流程累積組織能力、建立內部價值並長期成長；另一條則是「蒸餾攻擊 CoT 盜取」，攻擊者偷錄他人思路與教學內容，帶回競爭店，甚至再擴散傳授他人。這類攻擊的流程是「偷錄思路→帶回競爭店/傳授他人」，本質是在「竊取價值、破壞信任」，與以自身資料訓練的合法作法形成鮮明對比。「智慧模型蒸餾攻擊(思維鏈/CoT 盜取)」的典型鏈條：先取得 API 存取，接著自動化大量送出 prompt，蒐集模型輸出

response，整理成可用的合成訓練資料集，最後訓練出性能接近的「學生模型」。

目前智慧模型產業對於 CoT 智慧蒸餾攻擊防護包含身分驗證（KYB/KYC）、流量異常偵測、輸出限制/隱藏 CoT、以及浮水印與追溯機制。在 LLM 時代，「輸出」本身就是重要的攻擊面，而且全程可能只透過正常付費 API 就能進行，不一定需要入侵伺服器。智慧模型四重防禦架構作為整體對策。L1 著重存取控制與 API 層防禦（如限制速率、身分驗證 KYC/KYB、異常查詢模式偵測、地理/IP/管轄限制）。L2 著重偵測與監控（異常流量偵測、行為樣式分析/相似度比對、大量取樣偵測、即時告警、公開漏洞與污染通報、Bug bounty 與負責揭露）。L3 聚焦輸出水印/控制（輸出格式控制、浮水印、輸出加噪、隱藏 CoT、來源追溯、降低教研品質外流）。L4 則是法律與契約（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）。有效防禦需要技術、法規合約與即時監控的協同策略，目標是提高攻擊/蒐集成本、降低成功率與可行性，並提升追溯與威嚇能力。

以上內容將在 2026 年 5 月 27 日(三) 10:00 am 以線上直播方式與媒體朋友、全球民眾及專業人士共享。歡迎各位舊雨新知透過星球永續健康網站專頁觀賞直播！

- **星球永續健康網站網頁連結:**
  https://www.realscience.top/7
- Youtube **影片連結:** https://reurl.cc/o7br93
- **漢聲廣播電台連結:** https://reurl.cc/nojdev
- **不只是科技：**https://reurl.cc/A6EXxZ

講者：

陳秀熙教授/英國劍橋大學博士、許辰陽醫師、陳立昇教授、嚴明芳教授、林庭瑀博士

聯絡人：

林庭瑀博士　電話: (02)33668033　E-mail:happy82526@gmail.com

劉秋燕　　電話: (02)33668033　　E-mail: r11847030@ntu.edu.tw