星球永續健康線上直播

智慧數位資安（10）

智慧模型潛藏思維鏈(CoT) 逆向工程

2026 年 6 月 3 日

米其林名廚料理之所以難以被完全複製關鍵不只是食譜本身更是隱藏於背後的火候控制、調味順序與臨場判斷等思維流程。大型語言模型（LLMs）中的「思維鏈（Chain of Thought, CoT）」技術亦呈現類似概念，不同於傳統 AI 由輸入到結果之單純結構，智慧模型透過逐步推理思維鏈對情境有完整了解並完成使用者提出需求之決策與問題分析，近期發展之智慧模型皆以思維鏈之建立為強化能力與使用者體驗為目標。本週我們將探討智慧模型思維鏈（CoT）逆向工程盜取，以及 CoT 逆向工程模型強化實例。

健康科學新知

中東衝突停火不和平 談判不信任:「戰和未定」

美國與伊朗的停火及和平協議談判目前仍無定論，雙方雖都不願全面重啟戰爭，核心議題上的分歧卻遲遲無法彌合。伊朗國營媒體曾披露一份備忘錄草案，內容涉及恢復霍爾木茲海峽航運、美軍縮減波斯灣軍事存在及解除對伊朗船運封鎖，白宮隨即否認，斥為「完全捏造」。川普雖稱伊朗「非常想達成協議」，但強調美方仍不滿意現有條件，並堅持伊朗必須正面處理高濃縮鈾與核能力問題；國務卿盧比歐則措辭謹慎，確認雙方持續接觸，但能否成局仍要看接下來進展。此外伊朗戰後的長時間網路封鎖使國內商業損失慘重，商界估計每日損失高達數千萬美元。近期網路部分恢復，被美方解讀為德黑蘭有意回到談判桌的訊號。即便如此，軍事衝突並未停歇美軍持續對伊朗南部飛彈設施與疑似布雷船隻發動打擊，伊朗則強烈譴責美方違反停火，談判環境依然不穩。爭議焦點集中在三處：霍爾木茲海峽的管控權、制裁與凍結資產的解除，以及核問題。海峽是全球能源運輸動脈，伊朗持續收緊管控，油價因此承壓，美方堅持海峽必須開放。伊朗則將解除封鎖與釋放凍結資金列為協議的必要前提。核問題分歧更深，美方要求削減或移交高濃縮鈾，伊朗拒絕承認軍事意圖；部分方案試圖將核議題留待後續處理，但此路

線難以獲得美國強硬派與以色列的認可。黎巴嫩局勢也升溫，以色列持續打擊黎南部與貝魯特的真主黨目標，5 月 31 日更發動數十年來規模最大的地面入侵，真主黨與以色列互控對方破壞停火，黎巴嫩戰線隨時可能成為美伊談判以外的另一個引爆點。就目前形勢而言，和平協議能否落地，繫於三項關鍵：停火能否切實維持、霍爾木茲海峽能否重新開放，以及伊朗核問題能否納入具體可執行的框架。三者缺一，任何臨時安排都很難穩定。

中國-俄國-北韓強化三角結盟:「倚俄聯中」

東北亞局勢升溫之際北韓外交取向以無意重啟對美、對韓或對日的實質對話，轉而深化與俄羅斯的戰略關係，並維持與中國網絡關係為主軸。新加坡外長維文此次訪問北韓是 2018 年川金會前後以來首次，隨後接續前往中國與南韓，行程本身即帶有訊號意義。維文訪朝後指出，北韓目前更重視自力更生、軍事威懾與政權安全，對開放重大外交溝通管道「尚未準備好」。他建議外界保持戰略耐心，等待時機成熟。新加坡也趁機邀請北韓外長出席東協區域論壇以保留低風險多邊接觸管道。在對韓關係上，北韓出現結構性轉變，已明確否定統一論述，並修改憲法，將南韓定位為另一個國家。南北關係因此從民族統一框架，轉向兩個敵對國家之間的安全問題。中俄兩國聯手反對透過外交孤立與制裁壓迫北韓，主張以政治外交途徑處理朝鮮半島問題。中俄相關聲明強調半島和平穩定，卻不再提及北韓無核化，顯示北京政策重心轉向維護北韓安全與區域力量平衡。圖們江流域合作的持續推進，則反映中俄對北韓議題的戰略利益早已超出安全範疇，延伸至交通、能源與區域開發。金正恩近期致電習近平就山西煤礦爆炸事故表示慰問，外界多將此視為中朝關係維繫的政治訊號。尤其在習近平可能訪問平壤的傳言持續流傳、中朝友好合作互助條約簽署六十五週年的背景下，這一動作的象徵意義更為突出。中朝關係雖不如朝俄關係那樣快速走向軍事化，但中國仍是北韓不可或缺的戰略後盾。

AI 浪潮推進產業轉型：「智業重構」

人工智能正在同時改變企業競爭、專業服務與勞動市場的想像。雖然近年 AI 取代工作的焦慮快速升高，科技領袖也承認 AI 可能帶來重大職業轉換，但從經濟史來看，長期、

大規模的技術性失業並沒有明確先例。過去農業機械化、工業革命、電腦與貨櫃運輸等技術變革，確實重塑產業與職業結構，但通常擴散速度較慢，且常伴隨新需求與新職位出現。因此，判斷 AI 是否會造成真正的就業斷裂，不能只看個別職位是否被自動化，而應觀察生產率是否快速上升、實質工資是否停滯、企業利潤是否大幅擴張，以及多個產業是否同時出現明顯裁員。AI 對職場的影響已在招聘市場中具體浮現。以四大會計師事務所為例，德勤、安永、畢馬威和普華永道近年明顯增加 AI 專業人才招聘。2025 年，其英語國家的招聘廣告中，要求 AI 技能的職位比例已接近 7%，高於 ChatGPT 推出前，也超過傳統審計職位的招聘比例。這些職位涵蓋生成式 AI 工程、機器學習、資料科學、自動化，以及協助客戶或內部員工採用 AI 工具等方向，顯示 AI 並非只是取代審計或顧問人員，而是推動專業服務業的能力重心轉向技術、資料與自動化設計。這也反映出專業服務業面臨的結構壓力。傳統會計與顧問公司依靠「金字塔」模式，由少數合夥人管理大量初級員工；若 AI 能處理基礎分析、資料整理與流程檢查，初級職位的培訓與晉升路徑可能被迫重組。不過，審計與專業判斷並未因此失去價值，而是越來越需要與 AI 使用能力結合。未來專業人才的核心競爭力，將來自對專業知識、資料邏輯與 AI 工具限制的綜合理解。在消費端，Google 的最新布局顯示 AI 競爭正從聊天機器人走向「代理」服務。Gemini 代理不僅與 OpenAI、Anthropic 的 AI 編碼工具競爭，也試圖將處理電郵、規劃旅行、追蹤比賽、監控購物折扣與金融資訊等功能整合進 Gemini app 與 Google Search。由於 Google Search 擁有龐大使用者基礎，若 AI 代理被大規模導入日常搜尋與個人任務，Google 可能在消費級 AI 市場重新取得優勢。AI 普及也帶來成本與商業模式挑戰。Google 的 AI 服務每月消耗的 token 數量快速攀升，背後代表龐大的算力、晶片與能源需求。即使是科技巨頭，也必須面對成本壓力，因此未來可能透過提升效率、限制使用量、推動訂閱制，或在 AI 搜尋與回覆中加入廣告來回收成本。這意味著 AI 競爭不只是模型能力之爭，也是基礎設施、分發渠道與盈利模式之爭。

印太四方會議聚焦供應鏈重組：「鏈盟制衡」

美國國務卿魯比歐訪問印度出席「四方安全對話」外長會議，標誌著印太戰略進入務實轉型期。會議並未朝軍事同盟靠攏，而是將重心轉向強化「科技—經濟—海上安全」網絡，合作範疇涵蓋供應鏈韌性、海底電纜、5G/6G 及人工智慧等。其中，最具戰略意義的是美印簽署了關鍵礦產與稀土合作框架，旨在降低對單一供應來源的依賴，直接應對電動車與國防工業的供應鏈焦慮。聯合聲明雖重申支持自由開放的印太，但刻意避開「亞洲版北約」式的軍事對抗語言，改以提供公共財與基礎設施競爭來維持區域影響力。Quad 正轉型為低調且靈活多邊平台，未來影響力將取決於四國能否在不正式結盟的情況下，持續產出具體的合作成果。

全球極端氣候與野火危機:「氣候抽動」

2026 年極端天氣以更劇烈且更難預測的形式出現。歐洲在初夏尚未真正展開前，便遭遇異常早發且強烈的熱浪，英國、法國、西班牙、比利時等多國相繼發布高溫警示，部分地區氣溫明顯高於常年同期，英國更刷新 5 月最高氣溫紀錄。這場熱浪不只是單一高溫事件，而是全球暖化背景下「氣候抽動」的具體表現：天氣在寒冷與炎熱、乾旱與潮濕之間快速切換，使農業、基礎設施、公共衛生與城市治理承受更大壓力。「氣候抽動」指的是氣候狀態在短時間內從一種極端轉向另一種極端。歐洲此前才受北極冷空氣影響，部分地區出現低溫、山區降雪與葡萄園受霜害威脅；不久後，穩定高壓與來自北非、伊比利亞半島的暖空氣又迅速推升氣溫，形成強烈熱浪。這種急劇轉換不僅影響日常生活，也讓農作物更容易同時面對晚霜與突發高溫，並使城市供水、交通、住房與醫療資源在短時間內承受壓力。熱浪的危害並不只在於氣溫數字本身。歐洲許多城市與住宅並非為頻繁高溫而設計，冷房設備普及率有限，若夜間氣溫無法下降，人體也難以恢復。法國在此次熱浪中已通報多起與高溫直接或間接相關的死亡案例，包括戲水避暑引發的溺水，以及運動活動中的疑似高溫風險。這顯示極端高溫已不只是氣象問題，更是公共衛生、城市規劃與風險溝通問題。全球其他地區也正面臨升溫與乾旱帶來的連鎖衝擊。2026 年前幾個月，全球野火面積已高於往年同期水準，非洲、西非薩赫勒、亞洲部分地區、美國與澳洲都出現異常燃燒面積。前期降雨增加使植被快速生長，隨後的

乾旱與熱浪又將這些植被轉化為易燃燃料，形成「先濕後乾」的危險循環。若厄爾尼諾進一步增強，野火、熱浪與乾旱風險可能持續升高。

極端自然災害的跨區域連鎖影響:「環環相扣」

2023 年加拿大野火不僅造成當地損失，煙塵更跨境影響美、歐空氣品質，甚至因沉降在冰雪上加速融雪，改變下游水資源。在氣候變遷加劇與全球系統高度連結的背景下，自然災害不應再被視為單一地點、單一部門的孤立事件，而應被理解為會引發跨區域、跨產業與跨時間尺度連鎖效應的複合風險。文章指出，災害的真正影響往往不只發生在受災現場，也會透過空氣、水文、能源、糧食、供應鏈與社會系統向外擴散。乾旱、洪水與極端高溫同樣會形成連鎖反應。乾旱會造成植被壓力、土壤變化與基礎設施退化；土壤水分降低又會減少蒸散作用，使降雨機會下降、氣溫升高，進一步強化乾旱與熱浪。在農業地區，災害還可能透過糧食短缺、出口限制與價格上漲影響全球市場。例如俄羅斯 2010 年小麥減產與印度 2023 年部分稻米出口禁令，都使依賴進口糧食的低收入族群承受更大衝擊。即使災害發生在非農業區，只要波及關鍵基礎設施，也可能產生全球後果。半導體工廠、油田或煉油廠若遭極端風暴與洪水破壞，可能造成供應鏈中斷、價格上升，並進一步影響醫療、交通、通訊、金融等依賴電子設備與能源供應的部門。因此，災害風險已不只是地方性的安全問題，也關係到全球產業與經濟韌性。傳統上人們常把災害連鎖效應想像成「骨牌效應」，但更需要關注的是回饋迴圈。野火破壞森林覆蓋後，地表溫度上升、蒸散冷卻下降，可能增加未來熱浪與火災風險；乾旱造成土壤龜裂，也可能釋放土壤碳，增加溫室氣體並進一步推動氣候變遷。這些回饋使災害不只是一次性衝擊，而可能改變未來風險本身。能源系統也可能受到類似影響。文章以法國 2024、2025 年夏季高溫為例指出，高溫會使河流水溫上升，降低核電廠冷卻效率。當熱浪同時推高用電需求、削弱核電供應並限制太陽能效率時，能源系統可能被迫依賴進口電力或化石燃料，進一步增加排放，形成另一種正回饋。在人類因應方面，森林管理、水資源管理與基礎設施建設雖可降低部分災害風險，但也可能帶來非預期後果。例如乾旱期間過度抽取地下水會造成地層下陷，不僅降低地下水儲存能力，也會使地區更容易

淹水。地層下陷曾加劇休士頓在 2017 年哈維颶風中的洪災，也與 2005 年卡崔娜颶風期間紐奧良堤防失效有關。

AI 控制與監測智慧模型回應:「知機馭智」

研究人員開發新方法，成功找出大型語言模型內部代表特定概念的「引導向量（Steering Vector)」，並利用 Recursive Feature Machines（RFM）辨識 512 種概念，包括地點、人物與專業領域。研究發現，無須修改提示詞（Prompt)，即可引導 AI 朝特定方向回答，甚至提升程式設計任務表現。未來除可用於模型控制外，也有助監測幻覺、有害內容等風險，提升 AI 安全與可信度。

AI 研究生物安全兩難:「雙刃生科」

隨著 AI 在蛋白質、病毒與藥物設計上的能力快速提升，生技創新迎來重大突破，但同時也引發生物安全疑慮。專家指出，AI 可能降低生物武器開發門檻，甚至協助設計具高度傳播力或免疫逃脫能力的病原體。目前各界正透過 DNA 合成篩檢、模型安全限制及高風險 AI 分級管理等措施降低風險。研究呼籲建立全球監管機制，確保 AI 生技發展兼顧創新與安全。

研究者拒絕使用生成式 AI 觀點:「守正持衡」

在生成式 AI 快速普及之際，部分學者選擇拒絕使用相關工具，認為 AI 雖能提升效率，卻可能削弱學生寫作、思考與批判能力。研究者也擔憂 AI 產生錯誤資訊、訓練資料版權爭議，以及龐大能源消耗帶來的環境成本。部分大學已限制 AI 用於學位論文撰寫，僅允許拼字檢查等輔助功能。學者呼籲面對 AI 熱潮應保持理性與批判精神，避免過度依賴科技而忽略核心研究能力。

智慧模型思維鏈(CoT)逆向工程盜取

電影《福爾摩斯》中神探破案如同現今智慧模型思維鏈逆向工程一般以可見線索重現隱藏思維。在真實案件發生的當下，沒有人會直接知道犯罪者背後真正的動機、邏輯與犯案手法，所以神探最重要的能力，就是透過各種蛛絲馬跡，一步一步地觀察、推理，再加以驗證，最後重建整個事件背後的真相。就像投影片中所提到的，「最小的細節才

是最重要的線索」，而福爾摩斯最厲害的地方，就是能夠從別人忽略的小細節，推論出背後完整的故事與邏輯。2009 年上映《福爾摩斯》電影第一集中，首幕華生帶著未婚妻瑪麗與福爾摩斯一起晚餐，而福爾摩斯第一次見面，就立刻展現高超推理能力。他先注意到瑪麗耳朵旁邊有一個極小的墨漬，一般人幾乎不會發現，但他卻從這個細節推論出她是一位老師。因為如果是年紀較小的小學生，在使用墨水筆時，很容易不小心把墨水彈到老師身上，而這個墨點的位置又剛好在耳朵旁邊，代表是近距離造成的。他甚至進一步推測，學生大約是七、八歲左右，因為依照墨水噴濺的高度，可以反推出孩子的大概身高。接著，他又觀察到瑪麗佩戴了一條昂貴的珍珠項鍊，但以當時教師在英國社會中的收入，並不容易負擔如此昂貴的珠寶，因此他推論這條項鍊可能是借來的，也代表她對這次與華生的見面非常重視。而後他又看到她手指上有長期佩戴戒指留下的環狀痕跡，同時膚色有明顯曬痕差異，因此推測她曾經結過婚，而且長期生活在海外，因為只有長時間接受陽光曝曬，才會形成這種戒痕。劇中福爾摩斯面對一個非常神秘的案件。案件的核心人物是一位名叫布萊克伍德（Blackwood）的神秘貴族與議員，他利用類似邪教與黑魔法的方式，在倫敦犯下多起連續殺人事件，而且刻意營造出超自然力量的氛圍，引發社會恐慌。電影一開始，福爾摩斯與華生就成功追查到他的犯案現場，當時一名女性正準備被當成祭品犧牲，而福爾摩斯及時阻止儀式，救下受害者，也協助蘇格蘭警場將布萊克伍德逮捕歸案。不過，真正詭異的地方是在後面。布萊克伍德被關進監獄之後，主動要求見福爾摩斯。他告訴福爾摩斯，即使自己即將被判處絞刑，但黑魔法的力量並不會因此消失，而且接下來還會再有三個人死亡。他甚至警告福爾摩斯，即使親眼看到他被處死，也無法真正阻止這場災難。之後布萊克伍德果然被公開執行絞刑，並由華生親自確認死亡。然而，幾天後卻發生更驚人的事情,有人發現他的墳墓從內部被推開，棺木遭到破壞，而屍體竟然消失不見。這件事情立刻在倫敦引起巨大恐慌，大家開始懷疑，布萊克伍德是否真的擁有超自然力量，甚至能夠死而復生。但福爾摩斯並沒有被這些現象迷惑。他認為，所有看似神秘的現象背後，一定存在合理的邏輯與科學機制。因此他開始重新調查整個案件，並逐步發現，布萊克伍德與其同夥其實正在進行某

種化學與毒物實驗。現場出現大量奇怪的化學物質、毒蛇、蟾蜍以及特殊裝置，看起來像是在製造某種毒素或致幻物質。福爾摩斯只能從零碎線索開始，一步一步反推整個事件背後的動機與手法。這與 CoT 重構概念十分相似，也就是從最終結果與輸出，回推其背後隱藏的推理過程。隨著調查深入，福爾摩斯逐漸發現，布萊克伍德真正的目的，是利用民眾對超自然力量的恐懼，進一步操控英國政局取得議會主導權。整起「黑魔法殺人事件」是一場結合化學、機械、心理操控與魔術技巧的精密騙局。福爾摩斯在調查過程中取得許多關鍵證據，例如從火焰燃燒時出現的特殊顏色，判斷其中含有特定化學物質；再從鞋底沾附的河床淤泥，推測對方活動的大本營位置；並進一步發現，布萊克伍德與其同夥正在製造類似氰化物的劇毒氣體，企圖於國會大廈地下散播毒氣，以暗殺議員並奪取政權。布萊克伍德明明已被公開處以絞刑，甚至由華生親自確認死亡，最後卻又從墳墓中「復活」。由於華生本身為醫師，也是福爾摩斯最可信任的夥伴，因此此事件更進一步加深社會對超自然力量的恐懼。然而，福爾摩斯最終仍拆解出其中真正的原理。原來布萊克伍德在執行絞刑時，利用特殊機關與吊繩設計，使真正承受力量的位置並非頸部，而是腰部支撐，因此並未真正窒息死亡。同時他事先服用了特殊植物萃取物，使心跳與呼吸暫時降至極低狀態，看似死亡，實際上則類似進入短暫冬眠或深度麻醉狀態。因此，福爾摩斯最終成功將整個案件背後的犯案手法、操控邏輯與真正動機逐一拆解，將原本看似神秘的黑魔法，重新還原為可被理解的科學推理過程。

過去傳統的模型盜取，多是針對模型參數或輸入輸出行為進行模仿。而目前新的問題是，是否能夠透過 CoT 逆向工程，進一步重建模型背後隱藏的推理能力。這也是教師模型與學生模型蒸餾技術的延伸。對於教師模型而言，其內部完整的推理過程通常是隱藏的。在目前的大型語言模型中，使用者無法看到真正完整的推理軌跡，只能看到推理摘要與最終答案。這樣的設計，主要是為了保護模型的智慧財產權、系統提示與敏感資訊。而所謂思維鏈逆向工程，就是根據問題、推理摘要以及最終答案，進一步反推出模型內部可能存在的推理軌跡。換句話說，即使無法看到真正完整的 CoT，仍然可能透過這些可觀測的輸出資訊，建立 CoT 反演模型，並合成近似原始模型的推理軌跡。因此，

它不需要真正取得完整思維鏈內容，就能利用推理摘要與答案，逐步補出可用於教學的長推理鏈，並進一步訓練學生模型，使學生模型逐漸具備接近教師模型的推理能力。然而，如果這樣的技術被用於智慧模型攻擊，就可能產生模型能力盜取的風險。因為即使原始模型隱藏完整思維鏈，仍可能透過反演與蒸餾技術，將隱藏的推理能力轉移至學生模型之中。

思維鏈逆向工程的流程可以分成三個階段。第一階段是訓練反演模型。研究者會利用公開模型或替代模型，蒐集問題、推理摘要與最終答案，進一步學習輸入與摘要、答案之間的關係，建立可反推出推理軌跡的反演模型。這部分在數學上，本質上就是函式與反函式之間的推估問題，也就是從可觀測輸出，回推可能存在的內部推理結構。第二階段則是利用目標模型的輸出結果，包括摘要與最終答案，透過已訓練完成的反演模型，產生合成推理鏈。這個過程並不需要真正取得完整的思維鏈內容，而是依靠可觀測輸出，生成近似的推理過程。因此，即使真實 CoT 被隱藏，仍可能透過反演方式建立可用於教學與訓練的長推理鏈。第三階段則是學生模型蒸餾。研究者可以將問題、合成推理鏈以及答案，作為監督訊號，用於微調學生模型，進一步提升學生模型的推理能力。這也是過去蒸餾技術的重要延伸，也就是利用較小型模型，透過合成推理鏈學習大型教師模型的推理能力。

以米其林料理作為例子，對於顧客而言，能看到的是端上桌的成品，也就是最後的答案；而菜單上的描述與侍酒師的補充說明，則類似於推理摘要。但真正關鍵的備料方式、火候控制、調味順序與烹飪技巧，其實隱藏在廚房流程之中，並未對外公開。因此，逆向工程的核心問題就在於：是否能夠從成品與菜色說明，進一步反推出背後真正的料理流程。換句話說，即使無法直接進入專業廚房，仍然可能透過觀察、記錄、分析與拆解，逐步建立一套可學習、可記錄、可優化的料理知識體系，並將原本看不見的流程，轉化為可以反覆練習的步驟。阿明利用這樣的方法，透過記錄、拆解與優化，逐步重建米其林料理背後的烹飪思維流程。這其實與思維鏈逆向工程的概念非常相似，也就是從可觀測的輸出結果，逐步反推出隱藏的推理過程。對於米其林料理的「隱藏思維」而言，

第一層是顧客看到的桌上成品，也就是最終輸出；第二層則是菜單描述與侍酒師的補充說明，屬於摘要資訊；而第三層才是真正的料理流程，包括火候、順序與調味技巧，這些通常不會對外公開。然而，隨著大型模型與影像分析技術的發展，即使只透過料理照片、菜色描述與補充說明，也可能逐步建立反推模型，進一步推估其背後的真實流程。因此，原本隱藏於米其林料理中的專業知識與烹飪技巧，也可能因逆向工程而被逐步重建。

在統計與機器學習的概念中，思維鏈逆向工程如同貝氏推理的思維。對於受害模型端而言，教師模型真正的推理軌跡是隱藏的，外部只能觀察到輸入問題、最終答案以及部分推理摘要。然而，攻擊端可以持續收集這些可觀測輸出，包括問題、答案與摘要，利用先前建立的反演模型，學習其背後可能存在的後驗分布。也就是說，即使無法直接取得真正的 CoT，仍然可能透過可觀測資訊，生成近似的合成思維鏈。因此，黑箱 API 即使只公開答案與摘要，仍可能被用來建立深層的推理重建模型。其核心關鍵在於，這類方法不需要逐字還原真實推理軌跡，而是利用貝氏反演與後驗分布生成方式，建立邏輯相容、可蒸餾、可訓練的合成推理鏈。過去，貝氏推理主要應用於證據累積與正向推理分析；然而現在，同樣的概念也可以被應用於反向推理，也就是從輸出結果反推出模型內部可能存在的推理結構。

對於 CoT 逆向工程盜取技術目前已逐漸受到重視。過程中攻擊者會先分析模型最終輸出的結果合成可能的推理流程，並透過實作、驗證與調整，逐步內化成可重現的推理能力。例如米其林高超廚藝中推測處理順序、火候時間、醬汁組合、技巧手法，以及整體流程的整合，都是重要步驟。原本隱藏於黑箱中的細部流程，透過反覆分析與驗證，被逐步拆解與重建。一旦這些隱藏流程被逐步開啟與組合，就可能使模型能力的盜取風險大幅增加。而在完成流程合成之後，還需要透過實作驗證、品嚐、檢驗、調整與優化，反覆進行練習與修正。這其實也類似於數位雙胞胎的概念，也就是透過持續模擬、驗證與修正，逐步建立可重現、可優化的能力模型。

在 CoT 生成中，合成思維鏈之所以能夠強化模型能力，主要有幾個重要原因。第一，

是所謂的「去噪效果」。真實教師模型的推理過程通常非常複雜，可能存在大量試錯、回頭與分支路徑。例如走 A 路徑失敗後再改走 B，之後再轉向其他路徑，但最終仍得到相同答案。然而，合成思維鏈則可以直接保留較乾淨、較有效率的前向路徑，將推理流程簡化為較清晰的步驟，因此能減少不必要的干擾與複雜度。第二，是「學徒適配度」。真實教師模型的推理過程往往過於複雜，但合成思維鏈可以利用較容易理解的語言與步驟，將原本複雜的推理轉化為較簡單、較容易學習的流程。因此，學生模型在學習時，反而更容易理解與執行。第三，是「去脈絡化的純粹性」。真實推理路徑中，可能混雜大量背景資訊、脈絡干擾與非必要步驟；但合成思維鏈則能聚焦於標準流程與關鍵步驟，使模型更容易進行反覆訓練與最佳化。因此，學生模型在學習合成思維鏈時，甚至可能比直接學習真實思維路徑具有更高效率。這也是目前思維鏈生成、思維鏈蒸餾，以及思維鏈逆向工程受到高度關注的重要原因。另一方面，若這類技術被應用於模型能力盜取，也可能進一步引發人工智慧安全與治理上的挑戰。以上即為思維鏈逆向工程與能力盜取的重要概念。

CoT 逆向工程模型強化實例

雙向推理鏈生成訓練(STaR)思維鏈生成技術透過少量範例引導模型自行產生推理過程與答案，再依結果進行篩選與修正。當模型答對時，保留推理作為訓練資料；答錯時則提供正確答案，引導模型反向補出合理推理並持續微調。此機制讓模型能從成功與失敗案例中同步學習，逐步提升推理品質。研究顯示，該方法可有效增強常識推理、數學解題與問答能力，並讓中小型語言模型以較少訓練成本獲得接近大型模型的表現。推理鏈生成與修正訓練機制先利用少量範例引導模型自主產生推理過程與答案，再依答題結果進行篩選與學習。答對的案例直接保留作為訓練資料；答錯時則提供正確答案作為提示，引導模型反向建構合理推理，再將修正後的推理納入訓練。透過持續迭代，模型不僅學習正確答案，更能學習形成答案的思考過程。研究顯示，此方法可有效提升推理能力、改善資料品質，並加速模型在複雜問題上的學習與表現。過程中模型先觀察少量範例，自行推理並嘗試解題，再篩選接近正確答案的結果保留學習，若答案偏差，則提

供關鍵提示，引導模型找出錯誤並重新建構推理流程。修正後的經驗被整理成可重複運用的知識與步驟，持續累積成更完善的推理能力。研究顯示，這種「生成－修正－沉澱」的學習模式，能有效提升模型在複雜推理任務中的表現與泛化能力。研究團隊以數學加法任務驗證「合理化推理（Rationalization）」機制對模型學習的影響。結果顯示，相較於只學習答案的基準模型，加入推理過程後，模型能同步掌握不同位數的加法規則，大幅提升學習效率與收斂速度。在 16 輪訓練後，整體準確率達 89.5%，明顯優於僅輸出答案的基準表現（76.3%）。此外，首次微調後，二位數加法正確率即由不到 1%提升至 32%。研究指出，讓模型理解「為什麼得到答案」，比單純記憶答案更能促進推理能力建立，對數學與其他結構化任務皆具有重要價值。研究發現相較於一般 Few-shot prompting 或單純 CoT 方法，STaR 能明顯提升 GSM8K 測試集的解題準確率。此外，加入 rationalization 雖然增加了訓練資料使用量，但對效能提升有限，表示模型真正的重要能力是能否學會合理的推理步驟。結果顯示模型逐漸學會接近人類的推理結構。

此思維鏈重建與更新學習技術也對醫療智慧模型應用具有助力。以 LDCT 肺癌篩檢為例，篩檢服務模式 TNLCEDP 整合巨量影像、專家判讀與病理黃金標準，建立高品質醫學數位數據，不僅提升 AI 訓練效能，更成為打造「數位頂尖醫師」重要基石。然而傳統 AI 多屬於黑盒模型，只能輸入影像後直接輸出結果，卻缺乏中間推理過程。當判斷錯誤時，系統無法解釋原因，也難以進行自我修正與持續學習。此為醫療 AI 臨床應用最大的挑戰之一。STaR 架構突破傳統黑盒 AI 的限制，讓模型不再只輸出結果，而是能同時產生可解釋的推理鏈（rationale）。透過模擬人類逐步思考的方式，AI 能學習診斷背後的判斷邏輯，提升透明性、可信度與自我修正能力，進一步推動醫療 AI 從「預測工具」走向真正具備臨床思維的智慧系統。

在第一階段「透明化的臨床推理」中，系統以低劑量電腦斷層（LDCT）影像作為輸入，從右上肺葉 8mm 的磨玻璃結節（GGN）出發，依序觀察「分葉現象」與「血管穿越徵象」，再結合病史特徵（如家族史陽性）進行整合判讀，最終對應到 Lung-RADS 4A，輸出「高風險肺癌」的結論。這種逐步可追溯的推理鏈，讓每一步都能與臨床指引對齊，

降低醫師對黑盒預測的不信任。第二階段強調「從病理回饋中自我進化」。當預測與病理證實一致時，系統保留並強化原本的推理鏈；若預測錯誤、病理結果顯示良性，系統會針對「為何判錯」進行反思，重新生成新的推理邏輯（rationalization），用病理答案回頭修正決策依據，讓模型從靜態工具轉向具備自我校正能力的學習者。

在核心維度比較上，傳統 AI 多採直接提供二元結論，遇到錯誤通常仰賴人工重新標註與再訓練，STaR 動態思維演化數位分身採取「先列出臨床推理、再下結論」的輸出方式，並能透過病理答案自動重塑推理鏈，形成「持續進化的醫師數位分身」，把錯誤處理從一次性的修補，轉為可累積的知識更新。在終極願景中，TNLCEDP 智慧模型將可逐年進行新病人篩檢、輸出推理與預測、再吸收病理確診回饋並啟動自我反思的 Self-Taught Lung Cancer Digital Twin。透過新病人 → 推理預測 → 病理回饋 → 自我進化 → 形成更強模型循環，在臨床流程中迭代進步，隨時間持續提高決策輔助品質。

以上內容將在 2026 年 6 月 3 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