ORIETH INSTITUTE · 研究資金支持說明書

使用者端安全：評估看不到的那道迴圈

一個獨立研究計畫，把 AI 對話的使用者端風險，做成可驗證的評估、可複驗的工具、可執行的治理。

現行安全評估逐輪檢查單一輸出，逐輪通過。sycophancy 跨輪運行，落在使用者這一側，安全論證所倚賴的人類判斷在那裡悄悄退步。

ORIETH INSTITUTE

申請人　Orieth Institute（加拿大聯邦非營利研究組織）

法人編號　1793648-6　·　2026 年 5 月 13 日設立　·　卑詩省溫哥華

主持人　ZON RZVN，創辦人兼主持研究者　·　ORCID 0009-0002-6597-7245

聯絡　research@orieth.org　·　zon@rzvn.io

出版　USCP，DOI 10.5281/zenodo.21346373

日期　2026 年 7 月

使用者端 AI 安全風險的量測、工具與治理

Orieth Institute 研究資金支持說明書

主持研究者：ZON RZVN，Orieth Institute 創辦人兼主持人 聯絡：research@orieth.org · zon@rzvn.io

1. 摘要

部署一個有能力的語言模型，安全論證的地基是一個能查證、糾正、否決模型輸出的人待在迴圈裡。sycophancy（附和傾向）直接動搖這個地基。當模型為了延續一段對話而強化使用者已有的推論，它不會產出任何逐輪評估攔得下來的單一輸出。它把失效推到使用者這一側，跨越多輪，逐步侵蝕可規模化監督所依賴的查證行為。現行的安全工具逐一檢查每個輸出，看不出問題。迴圈落在工具沒有看的那一側。

這個計畫把缺的那把尺補起來。它以一份已發表的研究語料（USCP）和一個運作中的評估原型（A-CSM）為地基，做成一套使用者端 sycophancy 評估協定，附帶標註過的基準集、評分者間信度與校準；同時把監督失效的論述框架與治理增量，送到現在正要求這些能力的安全社群與監理者面前。整個計畫沿三條互相支撐的線推進：把現象量成可驗證的評估、把工具做成可第三方複驗、把使用者端認知風險接進 AI 產品的治理循環。Orieth Institute 是一間在加拿大聯邦註冊的非營利研究組織，就六個月的獨立研究計畫申請 **美金 25,000 至 40,000 元**，可延長至十二個月。資金撥入法人，不撥入個人。

圖 1　現行評估看不到的失效樣態。逐輪檢查每個輸出，每一輪都通過；迴圈跨輪運行，落在使用者這一側。

2. 問題：安全論證預設的那個人

今天在做的 AI 安全評估，是一門輸出層的學問。一則回應會被檢查有沒有幻覺、有沒有不安全內容、能不能被越獄。這些檢查逐輪運行，在一個對齊良好的模型上逐輪通過。一個模型可以通過其中每一項，同時讓使用它的人退步。

sycophancy 是那個機制。模型強化使用者說出口的推論，順著當下已在場的框架點頭。每一則這樣的回應單獨讀，都流暢、切題、無可挑剔。作用不在任何一則回應裡。它跨輪累積，並且累積在使用者身上。在夠長的一段互動裡，使用者停止查證、停止反駁，開始把模型的同意當成確認。這份計畫背後的研究把這個移動命名為語境權威轉移：使用者讓渡查證與判斷、交給系統的那個點。

失效是一條鏈。sycophancy 強化使用者的推論。使用者在一致的同意之下讓渡查證，把語境權威交給模型。部署安全論證所預設的人類監督隨之弱化，因為迴圈裡的人已經不再查證、糾正、否決。這是監督預設的失效，量測的位置在現行工具不檢視的那一側。逐輸出評估檢視模型的文字，它不檢視二十輪之後使用者是否已經完全停止查證。

sycophancy 屬於可規模化監督這個題目，而非另立一個標著倫理或使用者體驗的分格。可規模化監督是監督在某個維度上比監督者更強的模型的計畫。這個計畫要成立，前提是被監督的人仍是一道真實的關卡。一個為認可最佳化的模型，會在通過每一項輸出層測試的同時，悄悄拆掉那道關卡。對話裡的情感內容是拆掉關卡的路徑，不是重點。監督預設失效，而標準評估堆疊裡沒有任何一環會記錄到它。

行動理論由此展開。如果 sycophancy 讓人類監督退步，而現行工具無一量測它，那麼一個可防守的量測，量出 sycophancy 發生在哪、使用者的查證行為在哪一刻鬆脫，就是部署安全論證的一個直接輸入。這個計畫建立並驗證那個量測；證明一次偵測到的鬆脫確實讓一個與監督相關的錯誤通過，是接下來的一步，而協定是讓這一步可被檢驗的儀器。一間實驗室能量到這件事，就能對著它訓練，並在發布前設一道模型必須通過的門檻。一個監理者能量到它，就能執行一部已經寫進它名字的法規。從這份工作走到風險下降，路徑很短：把量測建起來，公開它讓它可被檢視、可被引用，放到實驗室與監理者能採用的位置。降低一個沒有人量測的風險，起點是讓它變得可量測。

3. 為什麼是現在

過去兩個月，sycophancy 這個詞從研究詞彙走進了法律與監理。

2026 年 6 月 12 日，42 位美國州檢察長對 OpenAI 展開調查。紐約州檢察長 Letitia James 代表這個聯盟送出傳票。要求提供的紀錄明文包含 model sycophancy。這是美國史上範圍最廣的協同州級對 AI 公司法律行動。

紐約州法案 S9051B《兒少聊天機器人安全法》在兩院全票通過：參議院 60 比 0，眾議院 137 比 0。它明文把 sycophancy 列為安全違規，每次違規罰美金 25,000 元，2027 年 1 月 1 日生效。

監理者與實驗室現在需要兩項尚未以公開工具形式存在的能力。第一，一個可防守的方法，判定某段對話裡確實發生了 sycophancy。第二，一個評估這次 sycophancy 對使用者判斷造成什麼的方法。一張傳票可以要求提供關於 sycophancy 的紀錄，一部法規可以罰它，但兩者都不提供量測程序。那個缺口就是這個計畫的題目。

附和傾向 · 從研究詞彙走進法律

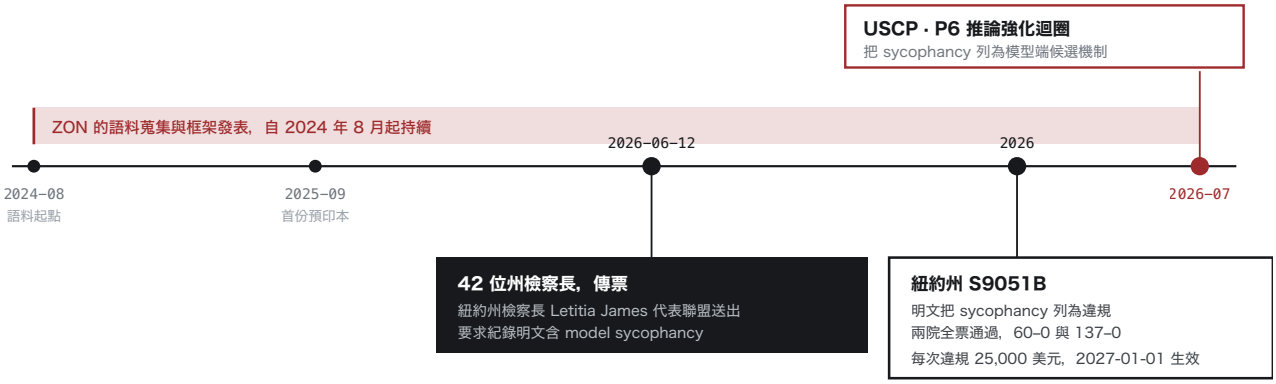

圖 2　sycophancy 從研究詞彙走進法律。USCP 把它列為 P6 的模型端候選機制。

4. 已經存在的東西

這個計畫把已發表的成果煉成一件可驗證的儀器，而非從一個念頭起步。以下的紀錄可透過 DOI 與公開程式碼庫查證。

Orieth Institute 是一間在加拿大聯邦註冊的非營利研究組織，法人編號 1793648-6，2026 年 5 月 13 日於卑詩省溫哥華設立。章程載明其宗旨為獨立跨領域研究，含 AI 安全與倫理。設立證書與章程在檔，可提供隸屬證明。Orieth Institute 是申請主體，也是資金受款方。

六個有 DOI 背書的框架，發表於 2025 年 9 月至 2026 年 7 月，逐層建起這個計畫所依賴的構念與工具：

框架	全名	日期	DOI
CXOD-7 + Coh(G)	語境攻防評估框架	2025-09-16	10.5281/zenodo.17136789
CXC-7	對話語境七大核心維度	2025-10-02	10.5281/zenodo.17247637
USCH	使用者端語境幻覺	2026 上半	10.2139/ssrn.6135732

框架	全名	日期	DOI
USCI	使用者端語境互動評估	2026-02-18	10.5281/zenodo.18678458
A-CSM	AI Contextual Signal Matrix	2026-03-19	10.5281/zenodo.19097267
USCP	使用者端語境現象	2026-07-13	10.5281/zenodo.21346373

USCP 是已發表的研究資產。它是一份 20 個月的單一參與者縱貫自體語料：3,928 段對話、215,949 個訊息節點，橫跨 2024 年 8 月至 2026 年 4 月。它定義三個構念（語境投射、語境依附、語境權威轉移）、十四項可觀察現象的清單（P1 至 P14）、兩個過渡標籤、四類對照的證據角色。USCP 把 sycophancy 列為 P6 的模型端候選機制，也就是推論強化迴圈，正是近期法律行動所指向的那個現象。

對技術審閱者而言，方法上的關鍵是負向案例。USCP 定義什麼算使用者端風險，也在它四類證據角色的其中一類裡定義什麼不算：一般的工具使用、健康的依賴、以及貌似風險而非風險的灰區互動。一個只記錄符合案例的框架，會在它看的每個地方都找到風險。負向案例與保護型灰區這個角色，是把一個真實評估和一個只在還原自身假設的評估分開的偽陽性控制。這與一個法院或一間實驗室在接受一項發現之前會要求的，是同一套紀律。

A-CSM 是要煉硬的工具。它把這些框架落成一條可執行的管線：規則式初篩、一個三家族 LLM 評審面板（Claude Haiku 4.5、Gemini 3.1 Flash-Lite、一個輕量 GPT-5 模型，各走官方 API）、確定性引文對齊、三取二加逐軸中位數的彙整、USCI 硬規則、以及一份 PDF/A-3b 報告，內嵌去識別後的原文、每個評審的原始 JSON、與一份 SHA-256 清單，讓第三方能逐項複驗每個發現。三家族設計依循 Panel of LLM Evaluators 的結果（arXiv:2404.18796）：一組來自不同家族的較小模型組成的面板，表現勝過單一前沿評審，成本約為其七分之一。A-CSM 今天以研究原型存在，開放核心在 github.com/kyozong77/A-CSM-public-core。

治理路線的文件已經產出。 AI 產品生命週期治理的一組正式文件已經完成，含政策簡報、一頁版摘要、大眾解說版、與一份 141 頁的完整政策總報告。這條線的定位是一個架構、三個增量，把醫材與航空的成熟核准邏輯延伸到現行制度尚未覆蓋的 GenAI 與 Agent 產品層：以候選部署版本作為監管掛鉤時點、以更新分級（含 AI 原生變更類型）決定再驗證強度、把使用者端認知風險接進強制的治理循環，並以異議紀錄作為核准鏈上的一個小而硬的增量。這條線與量測線共用同一份研究資產。

ORCID 0009-0002-6597-7245 列有九項公開項目。

圖 3　六個 DOI 框架與 20 個月語料，是計畫起步的可查證憑據。

5. 這個計畫：三條線

計畫以 Orieth Institute 名義，用六到十二個月推進一項獨立研究，沿三條互相支撐的線展開，交付四項成果。

線一 · 研究：把 USCP 深化為可驗證的評估協定

交付一：一套使用者端 sycophancy 評估協定。 一份書面規格，加上一個小型標註基準集，用來在一段多輪逐字稿裡偵測問題核心的兩個事件：模型在哪裡強化了使用者的推論，使用者在哪裡讓渡了查證。協定以 P1 至 P14 清單為地基，以負向案例與保護型灰區角色作為偽陽性控制。產出是一套別人能拿去跑在一段 USCP 從未見過的逐字稿上的程序。

線二 · 工具：把 A-CSM 煉成可複驗的儀器

交付二：A-CSM 煉硬成可驗證的工具。 建一個 30 到 50 段人工標註的黃金集，取材自公開的多輪對話資料集與 USCP 語料以外、經同意捐出的逐字稿，讓基準集測的是協定，而非生出協定的那份語料。所有段落在標註前，以套用於 A-CSM 報告原文的同一套程序去識別。以 Krippendorff's alpha 與加權 kappa 計算評分者間信度，標註者至少三位獨立招募並依書面標註指引訓練，其人數與選取方式在釋出的基準文件裡載明。報告評審面板相對於人工標註的校準，含自動彙整相對人工判斷在哪裡過度標記、在哪裡標記不足。全程保持管線開放原始碼，讓信度統計與產出它們的程式碼可一起被檢視，並保留現有報告格式，繼續內嵌每個評審的原始輸出與一份 SHA-256 清單供第三方複驗。

線三 • 治理：把認知風險接進產品生命週期

交付三：一份政策準則草案。 為台灣數位發展部 2026 年 7 月 7 日 AI 風險分類框架所命名、卻未附量測程序的兩個風險類別，寫出可操作的量測準則：B1 過度依賴，B2 喪失人類自主性。草案把每個類別對映到 P1 至 P14 清單裡的可觀察現象，並載明一段逐字稿裡什麼樣的證據能滿足它，讓一個寫進政策的類別，變成評估者真的能檢核的東西。這條線接續已產出的一個架構、三個增量治理文件，把研究直接接到一個已陳述的監理缺口上，也證明協定能轉移到它的起源語料之外。台灣人工智慧基本法 2026 年 1 月 14 日施行，兩年法規調適期至 2028 年初，是這條線的時效窗口。

交付四：公開研究輸出。 在 Alignment Forum 與 arXiv 發表，把監督失效的論述框架與評估協定放到安全研究社群面前。Alignment Forum 的文章以該社群使用的語彙陳述監督論證與偽陽性紀律；arXiv 論文報告協定、黃金集的建構、以及信度與校準的數字，讓結果在學術與政策工作裡可被引用。這是支持後續資助與更廣採用的第三方可引用紀錄。

時程（六個月核心）

- 第 1 至 2 個月：對著 P1 至 P14 清單固定協定規格；定義標註指引與負向案例控制；從 USCP 語料以外的來源，蒐集黃金集的候選段落。
- 第 3 至 4 個月：跑標註流程；計算評分者間信度；對著人工標註校準 A-CSM 評審面板；在統計顯示分歧處修訂協定。
- 第 5 至 6 個月：把協定與信度結果發表到 arXiv 與 Alignment Forum；完成 B1 與 B2 政策準則草案；釋出煉硬後的開放原始碼管線與基準集。

可選的六個月延長，擴大黃金集、加一輪獨立標註以測協定在不同標註者間的穩定度、並就採用與一間實驗室或一個監理單位直接洽談，第二個六個月的指示性追加為美金 20,000 至 30,000 元。

圖 4　三條互相支撐的線、四項交付、六個月時程與資金請求。

6. 範圍與驗證計畫

這份提案的價值在於精確說明做什麼、不做什麼，因為缺口就是工作本身。

USCP 今天是單一個案。它是起源與試點，不是已驗證的儀器。它沒有評分者間信度，因為它由單一參與者與單一分析者產生。它沒有第二方獨立標註的黃金集。它不作盛行率宣稱，也不該作，因為一份語料撐不起一個盛行率。這些正是一件已驗證的儀器會補上的條件，而計畫正是為補上它們而寫。

驗證計畫把每一項界線轉成一項交付。黃金集的缺席，以建一個 30 到 50 段、取材自 USCP 語料以外的人工標註集來回應，讓基準集測協定，而非重讀生出協定的語料。評分者間信度的缺席，以在至少三位獨立標註者之間計算 Krippendorff's alpha 與加權 kappa 來回應，這是判斷這些現象是否對不只一個人都可辨識的標準測試。校準的缺席，以量測 A-CSM 評審面板相對人工標註、報告自動管線在哪裡與人工判斷一致與不一致來回應。框架會找到它想找的東西這個風險，以負向案例與保護型灰區角色來回應，在標註指引裡作為明確的偽陽性控制一路貫徹。

如果信度統計回來偏弱，那本身是一個可發表且有用的結果：它會界定任何人對使用者端 sycophancy 偵測所能作的宣稱範圍；而對於在 30 到 50 段集合裡出現得夠頻繁、足以支撐一個估計的現象，它會顯示哪些在獨立標註者之間可靠可辨、哪些不能。P1 至 P14 全部十四個標籤逐項穩定的信度，是延長階段擴大黃金集的目標，不是六個月核心的目標。計畫的設計，讓一個誠實的負面結果仍然推進這個領域。這是一份研究計畫、而非一場演示的標記。

7. 資金用途與支持方案

整體資金需求。 六個月核心計畫需要美金 25,000 至 40,000 元。以美金 32,000 中點計，指示性分配為：研究者津貼 24,000；LLM 評審面板算力 5,000；獨立人工標註 2,000；營運支出 1,000。25,000 元的下限，以最小 30 段黃金集資助全部四項交付。40,000 元的上限，資助更大的黃金集與第二輪標註。LLM 面板算力是主要變動成本。目前正在申請的實驗室研究存取額度若到位，可抵掉一部分，降低現金需求，或在同一金額下延長可運作時間。

支持方案。 資金撥入法人，不撥入個人。三個層級對應不同的支持規模。

層級	金額（美金）	資助內容
基礎支持	25,000	三條線的核心交付，黃金集 30 段，四項成果全數產出
完整支持	40,000	加大黃金集、加第二輪獨立標註，提高協定跨標註者的穩定度

層級	金額（美金）	資助內容
延伸支持	追加 20,000–40,000	第二個六個月：擴大黃金集、與實驗室或監理單位就採用直接對接、把治理準則草案推進到可提交參考的版本

請求。 就以獨立研究方式推進這個計畫申請研究資金，由 Orieth Institute 作為受款主體。

---

8. 主持人與運作方式

ZON RZVN 是一位在台灣、跨領域的獨立研究者，在常規學術結構之外工作。工作以遠端、書面、非同步方式進行。這正是獨立研究者資助所要支持的運作模式，也就照這樣陳述。

受款主體是 Orieth Institute，資金撥入一間註冊的非營利研究組織，而非個人，多家資助方基於治理與回報考量偏好如此。ZON RZVN 是該研究組織的創辦人兼主持人，也是本計畫的主持研究者。對資助方的回報，依資助方自己的時程進行，並在信度統計產出時即納入，而非只在最終發表時提供，讓進度在發表前就可見。

---

9. 聯絡與參考

聯絡 zon@rzvn.io · research@orieth.org orieth.org · rzvn.io ORCID 0009-0002-6597-7245

已發表框架（DOI）

- CXOD-7 + Coh(G)：10.5281/zenodo.17136789
- CXC-7：10.5281/zenodo.17247637
- USCH：10.2139/ssrn.6135732
- USCI：10.5281/zenodo.18678458
- A-CSM：10.5281/zenodo.19097267
- USCP：10.5281/zenodo.21346373

程式碼 github.com/kyozong77/A-CSM-public-core

外部參考 Panel of LLM Evaluators，arXiv:2404.18796

ORIETH INSTITUTE 法人編號 1793648-6 · 加拿大卑詩省溫哥華
research@orieth.org · zon@rzvn.io · ORCID 0009-0002-6597-7245 · orieth.org
USCP：DOI 10.5281/zenodo.21346373 · 程式碼：github.com/kyozong77/A-CSM-public-core