CAMPUS AGENT / WEEKLY RESEARCH MEETING

從 Production 走向真人測試與研究定位

115 資工專題 · 2026/09/08 Meeting

暑假成果收尾 · Internal Pilot · Research Positioning · 本學期規劃

暑假做到哪裡：系統已經可以被真人使用

題目定案 → 規則整理 → 判定流程 → Production → Internal Pilot

階段	現在的狀態
規則與來源	國科會專題研究計畫 × 中央大學的版本、來源與適用範圍可追蹤
判定	追問缺少事實後，由規則執行 verdict；保留 provenance / release 資訊
服務	已部署 production，可保存對話與後續 feedback
真人測試	已進入 internal pilot；這學期開始看真實使用 failure，而不是繼續堆功能

這學期：從把系統做出來，進入真人測試與研究驗證。

Internal Pilot 已開始：先確認有沒有公開測試 blocker

- 系辦邀請寄出後，production 已出現 **約 3 筆新的 conversation**；依時間與情境推測來自這一輪測試，但**尚未視為 expert-confirmed correctness**。
- 正式 qualitative feedback 還在等；收到後先對照：**原始問題 → 系統結果 → 承辦人實務判讀**。
- 這週只修了一個會影響研究分析的 instrumentation 缺口：每個 conversation event 現在都保留當下的 **Expert Mode** 狀態，已發布 `v2026.09.03`。

> **這一輪的目的：先確認沒有會阻礙公開測試的重大問題；其他問題則保留到後續真人測試繼續觀察。**

近期研究：把回答需要的規則與證據找完整

近期 closest work **RegOps-Bench / RefWalk**（Findings of EMNLP 2026）主要在處理：

- 從不同層級的規範來源中找到回答需要的規則
- 沿引用關係補齊完整 evidence
- 把回答中的 claim 對回實際規則來源

> **這篇工作的重點：把回答需要的 governing evidence 找完整，並正確連回規則來源。**

我們目前想再驗證一個可能的 gap：

> **候選規則都找到了之後，能不能把「哪些真的適用本案」獨立成一個 decision layer？**

找到候選規則 → 確認本案真正適用的規則 → 才能下判定

可直接追： Paper · arXiv · Code · GitHub

候選研究方向：這次案件到底要用哪些規則？

真人情境	為什麼「找到規則」還不夠？
60 萬 vs 100 萬	兩個版本都相關，但案件日期決定這次要用哪一版
科研採購 vs 一般採購	兩套規則都相關，但案件屬性決定真正適用哪一套
國科會 3 個月 + 中央 45 天	不是二選一；不同來源的規則可能要同時適用
使用者資訊不完整	如果還不能確定適用規則，就應該先補關鍵資訊，而不是硬判

> **候選方向：規則已經找齊後，進一步判斷哪些真的適用本案；資訊不足時則先不要硬判。**

下一步先深入讀 **RefWalk code / benchmark**，確認這一層是否真的還沒有被充分處理；如果成立，再決定正式 RQ，以及是否把「應該先問哪個關鍵問題」作為延伸。

這學期近期規劃：先把公開測試時間點定下來

靜儀回饋確認 → 修重大問題 → 完成公開測試前置 → 開始公開測試 → 整理真人使用問題 → 研究評估

近期先做兩類事情：

1. **真人測試**：跟靜儀確認目前系統沒有重大 blocker，完成公開測試方式與資料蒐集前置。
2. **研究方法**：深入讀／重現 RefWalk，評估能否直接沿用 retrieval baseline / candidate-rule generator，並確認 applicability 是否值得獨立研究。

> **時間點：大概第幾週前完成系統修正與靜儀確認，接著開始公開測試？**