從真人測試問題到新的研究流程

115 資工專題 • 2026/09/15 Meeting

找出 failure → 建立新版 workflow → 準備下一階段驗證

1. 上週真人測試：舊流程處理不了完整案件

真人問題
↓
只選一個主題
↓
只找一部分規則
↓
答案容易不完整

真實案件常常同時牽涉多個規則、缺少資訊，也可能遇到版本或來源問題。

> **所以這週第一件事不是修某個 bug，而是先找出「到底是哪一步讓系統失敗」。**

2. 這週確認：問題不只是「找不到規則」

舊系統

一個案件 → 一個 `topic` → 容易漏規則

RefWalk（上週提的論文）

找到規則 → 沿規則之間的引用繼續找

但 Campus 的規則彼此很少引用，所以這個方法不能直接照搬。

這週改成

一個案件 → 拆成多個問題 → 每個問題各自找規則

在固定 development cases 上，新流程已能把必要規則全部找齊。

3. 最重要的發現：規則找齊 ≠ 可以回答

找到所有可能相關的規則後，還要知道：**為什麼現在不能下判定？**

缺口	系統應該怎麼做
缺案件資訊 例如：不知道採購日期	**問使用者**
版本或官方證據不夠 例如：不知道本案該用哪一版	**不能猜，先停**
系統自己的規則表示不夠 例如：例外條件沒有表示清楚	**修系統，不該叫使用者回答**

所以新版流程不只要「找規則」，還要知道缺什麼、該問誰，以及什麼時候不能回答。

4. 這週系統做了什麼

1. 一個案件不再只選一個 topic

> **先拆成多個規範問題，再分別找需要的規則。**

2. 找到規則後，不再直接硬答

> **系統會分辨：可以回答、需要追問，還是應該先停止。**

3. 版本／來源證據不足時，不讓系統自己猜

> **把這些缺口明確留下來，等證據補齊再判定。**

目前 research candidate 先用 deterministic workflow 驗證架構（LLM calls = 0）；還沒證明能泛化到新的真人問題。

5. 下一步：驗證後再開公開測試

```
新的 unseen / sealed cases
          ↓
正式 evaluation + 錯誤分析
          ↓
整合既有 production 能力
          ↓
limited public test
```

預計節奏

- **約 2 週**：完成 unseen evaluation
- **約 3 週**：依結果完成第一版 integration
- **約 3–4 週**：若沒有重大 blocker，再開始 limited public test

需要獨立 case 保管／建置，以及熟悉報帳實務的人做最後 domain review。