規則寫了，不代表引擎真的有執行

115 資工專題 · 2026/08/12 Meeting

真正的產出不是規則數量，而是抓到一整類既有評估看不見的缺陷

表面上這週是一大段工程進度

- 規則庫：**45 → 89 條**
- 抽取：**44 份文件、7,098 格、14,196 次呼叫、0 失敗**
- 全量產出：**680 個規則候選**
- 成本估計：約 **US$12.7**
- 有機會觸發跨來源取嚴的規則：**12 → 54 條**

本地模型因實驗室 GPU 帳密沒有拿到，沒有照原計畫跑成；改用 **3 格煙霧測試**先確認呼叫路徑，再進正式全量。

> 但這些都不是這週最重要的結果。

真正發現：規則宣告的語意，和引擎執行的語意可以完全不同

規則上寫的	引擎實際做的
有生效日	只排除已失效規則，**沒排除尚未生效規則**
「未經同意者不得…」	抽取後只剩「不得…」，**但書消失**
科目規則提供正確科目	引擎拿正確答案當篩選條件，**填對的人反而被擋**
規則限制「限專任人員」	程式根本**沒有讀這個欄位**

規則資料看起來完全合理，系統仍然可以穩定地做錯。

甚至「100% 覆蓋率」也可能只是分母算錯

全量報表曾經一直顯示 **100% 覆蓋**。

後來發現：有 **19% 語料根本沒被送出去問**，原因是編號撞號；但分母用的是「實際送出去的格數」，所以漏掉的資料同時不在分子也不在分母。

連帶更正先前數字：

查核缺失涵蓋率 100% → 96%

> **Metrics 本身也需要被驗收，不能因為數字漂亮就相信。**

更關鍵的是：端到端分數幾乎沒有提醒我們

修正這批缺陷前後：

多輪引用命中率約 89–91% → 88–89%

幾乎沒有變。

但其中一個 bug 代表「版本與生效年度」這個核心賣點之前其實從未完整執行。

> **一個考 88 分的系統，仍然可能有核心功能根本沒真的跑。**

因此新增的真正研究資產，是把這些 failure 轉成每次都要跑的 invariant checks：

必跑檢查 3 → 7 支；驗證項目 21 → 24 項。

為什麼人工讀過還會漏？

其中三個缺陷曾經逐條人工複核，當時都判定「沒有問題」。

原因很直接：

- 人看到欄位名稱，會自然補成它「應該」代表的意思。
- 程式只會執行真的寫進 code path 的邏輯。

> **驗證規則內容 ≠ 驗證規則真的被執行。**

這一週開始把「execution semantics」本身當成需要獨立驗收的對象。

還撞到第三種規則衝突：不是取嚴，也不是取新版

同一筆 **80 萬元**採購：

制度	結果
一般採購	三家書面報價、比價議價
科研採購	可直接洽廠商

兩份中央大學要點甚至在 **同一天**修正。

所以不能：

- 跨來源取嚴
- 選較新版

而是要先判斷：

> **這個案件到底走哪一套制度？適用範圍互斥時，要先判 scope，再談規則比較。**

老師把下一步推向「真的給人用」

這次老師建議：

1. 本地模型的帳號問題再問助教。
2. 實測行政院主計總處智能客服，看看上層系統怎麼回答同樣問題。
3. 口試計程車題回答不穩：要先問這筆交通是否真的和計畫相關。
4. 先去系辦做 internal test；穩定後，開學可以用 QR Code 讓更多人試。
5. 部署網站並保存對話，讓真人使用本身變成可分析資料。

> **專題開始從「把系統做對」往「讓真人把沒想到的問題撞出來」移動。**