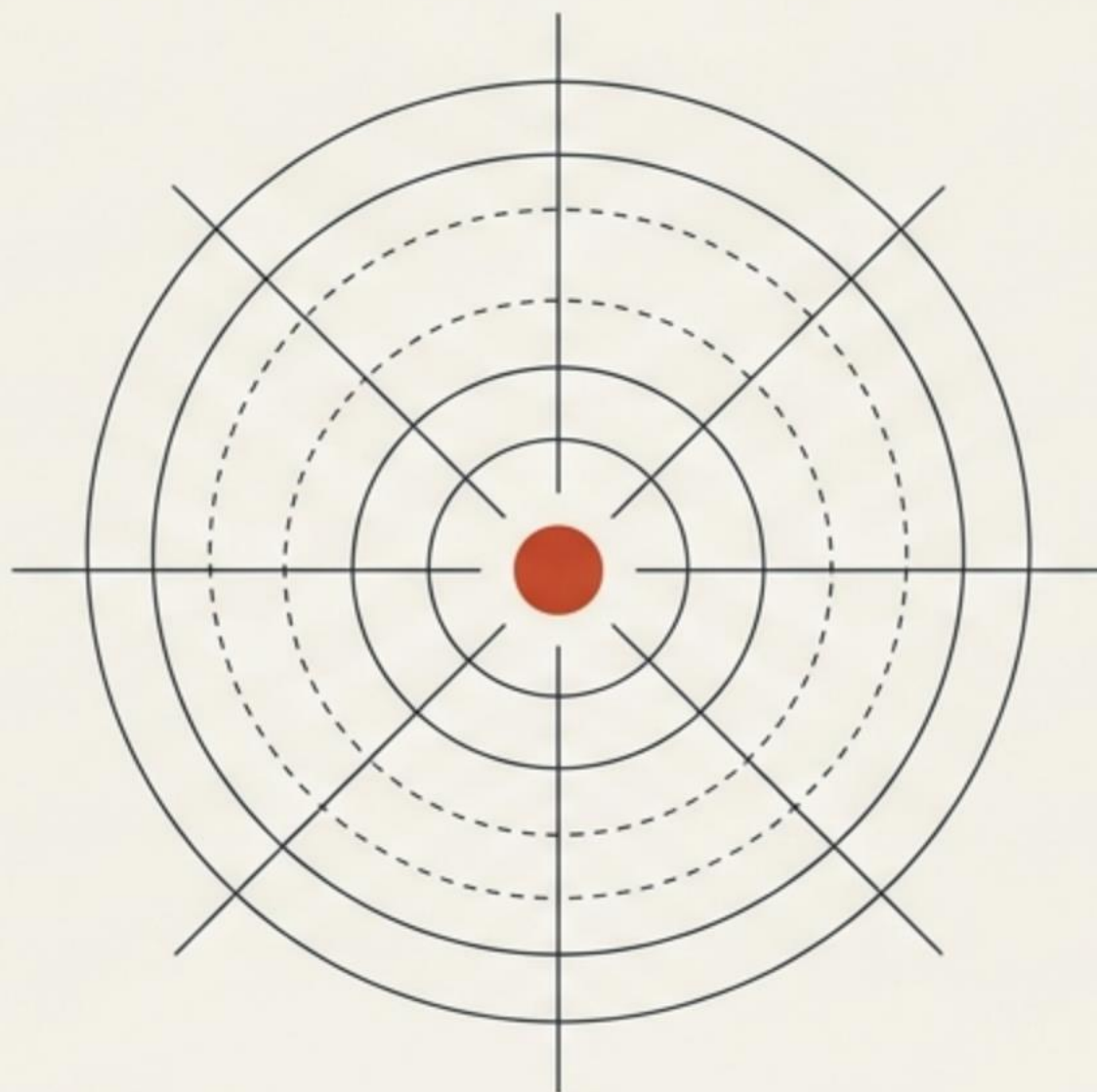




金融業生成式AI平台工程

以雲端原生架構打造可擴展的 Agentic AI 中樞

從PoC工具到可治理、可觀測、可驗證的金融級AI平台

2026.07.01 張維峻 Justin

客戶現場。遇到IT作業問題。
不能打字、不能慢等。



這不是聊天機器人測試，
這是AI進入營運現場的考驗。

聽懂自然語音

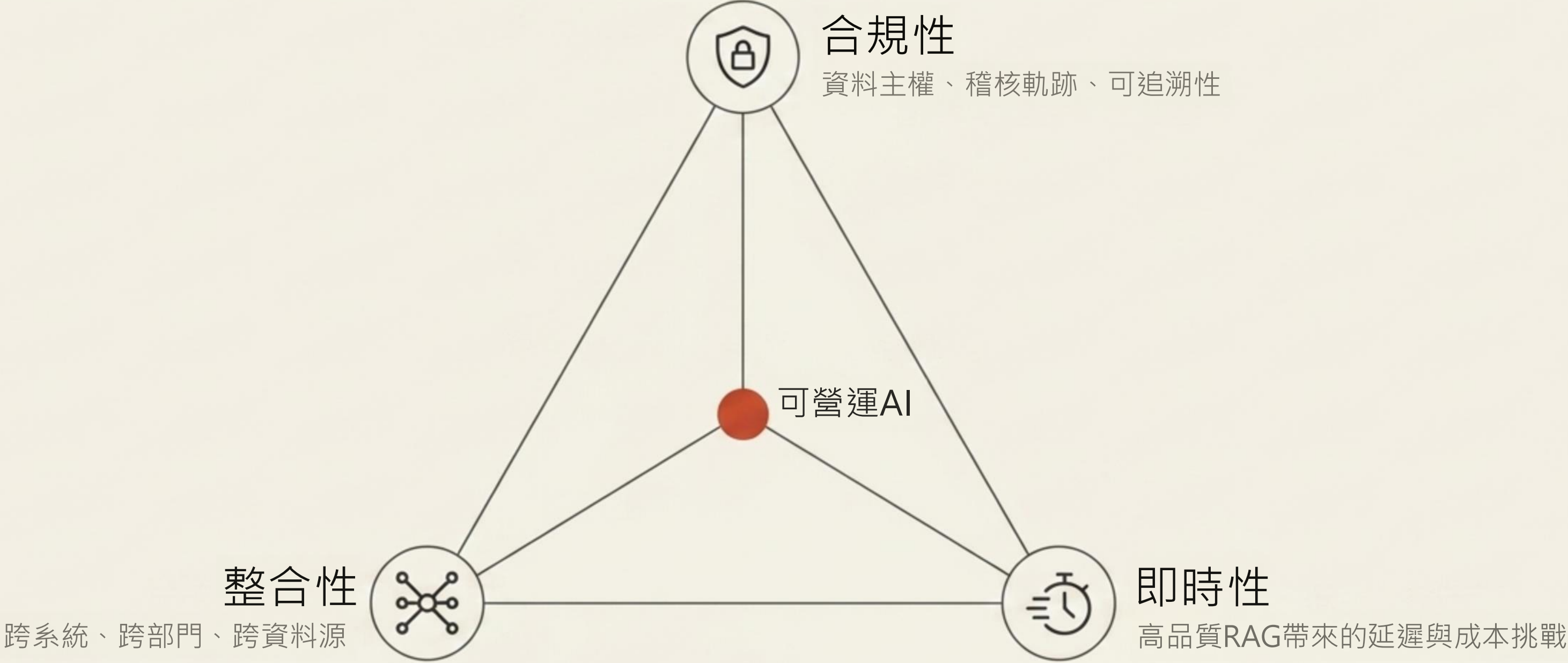
查證內部知識

判斷證據充足度

具備拒答機制

完整留存追蹤紀錄

金融AI 落地的三條生死線



問題不是做不出AI，而是無法通過真實營運條件的檢驗。

為什麼 AI 專案往往卡在PoC階段？

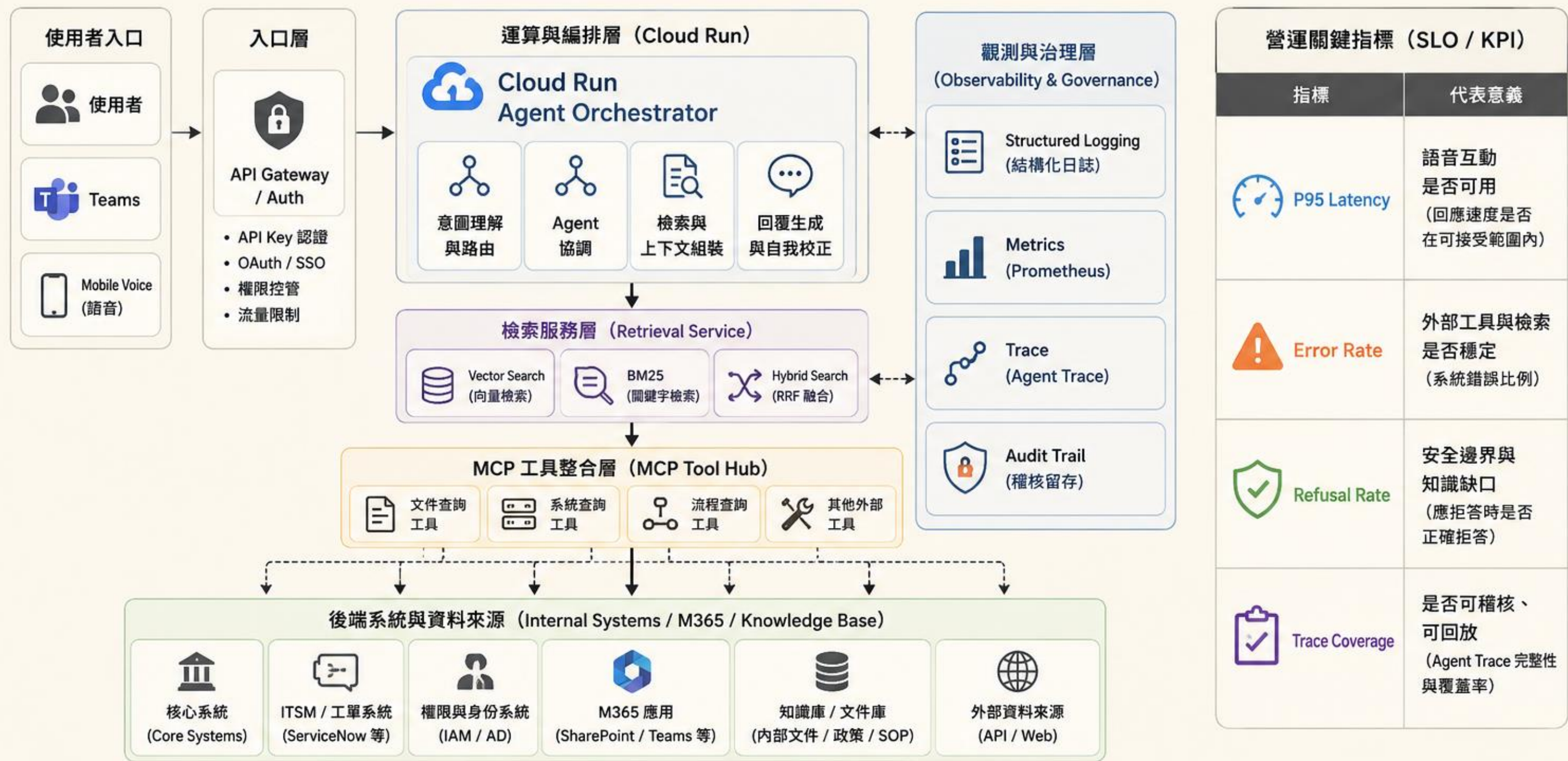
斷點	表面問題	真正問題	平台級解法
系統孤島	每個 API都要客製串接	Agent無法規模化使用工具	MCP (Model Context Protocol)
線性RAG	回答看似合理但不可驗證	缺少自我校正與證據檢查	Agentic RAG 工作流
黑盒AI	無法解釋資料與工具來源	稽核資安法遵擋下上線	Agent Trace

突破這三個斷點，需要的是平台工程，而非單一模型。

Cloud Native AI Runtime :

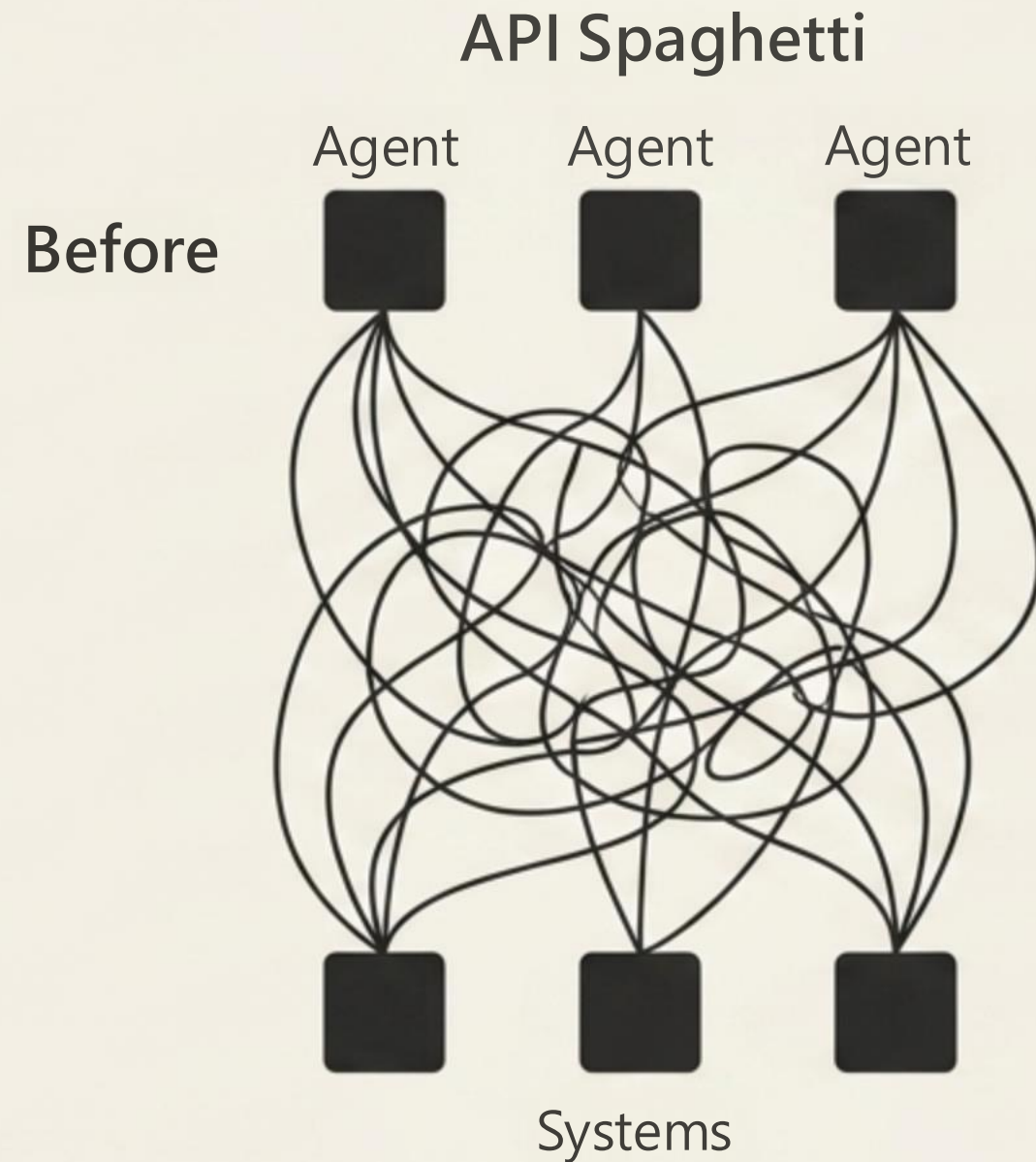
讓Agentic AI可以被部署擴展、監控與治理

以雲端原生架構打造可擴展、可觀測、可治理的金融級AI服務平台

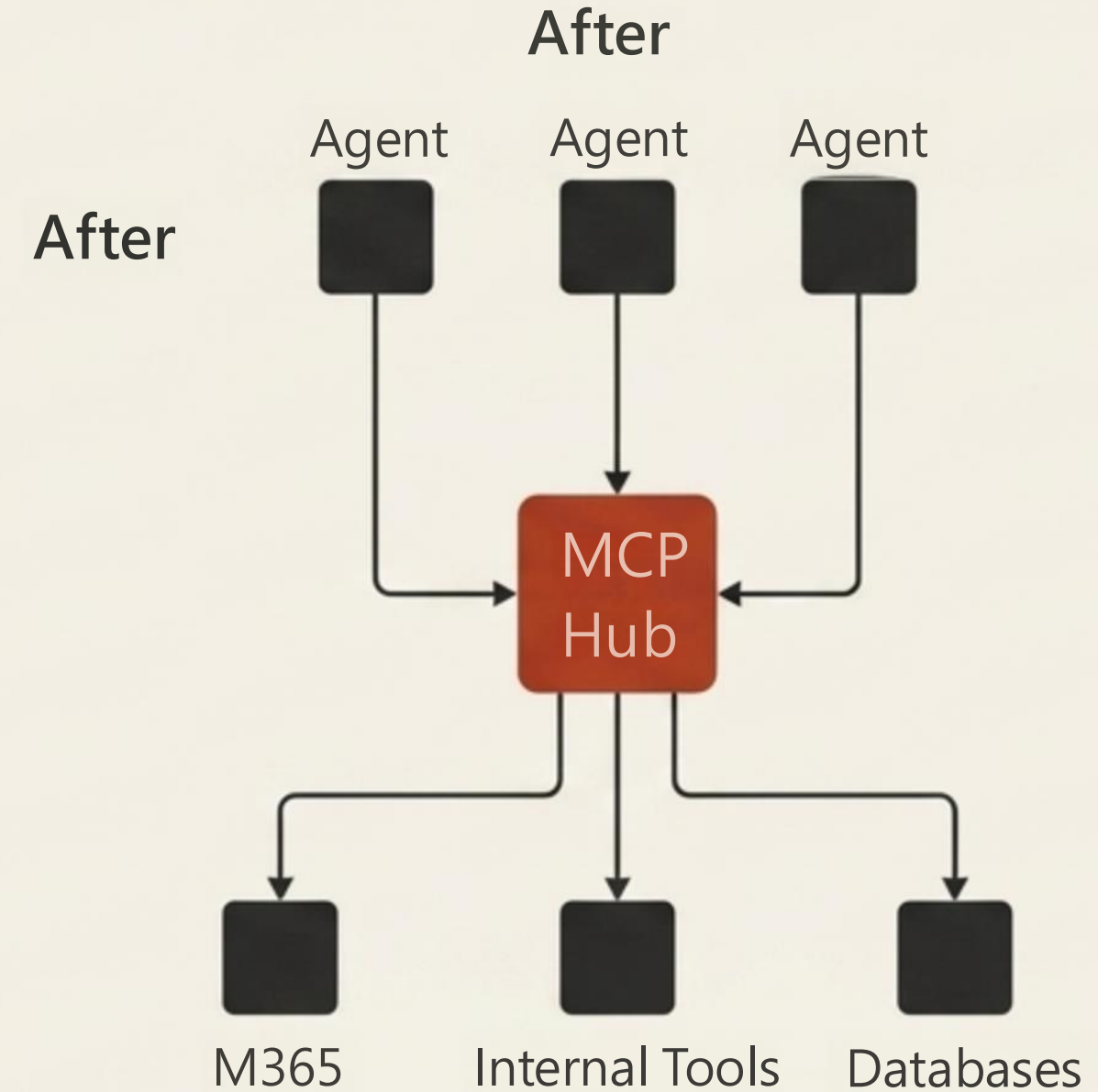


AI 要上線，不能只會回答，必須能被部署、監控、追蹤與治理，才能真正創造業務價值與信任。

讓企業工具變成Agent可治理的能力



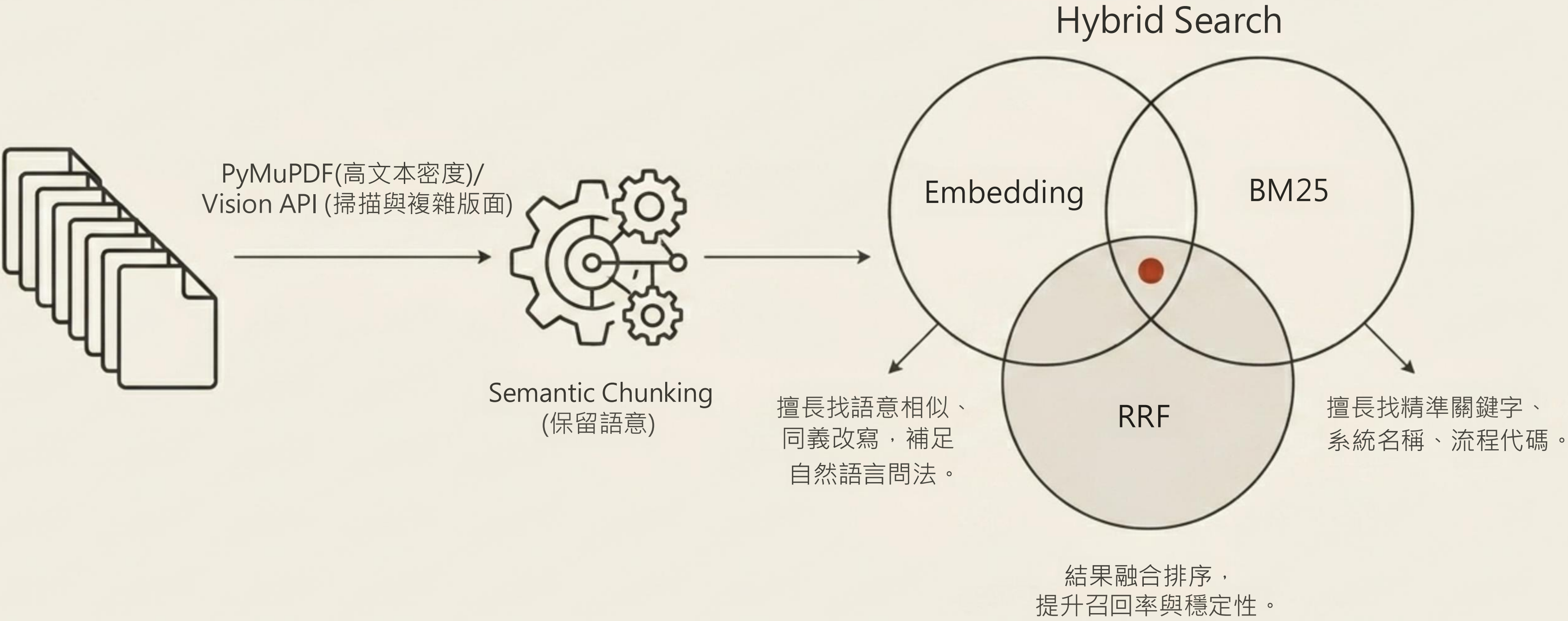
傳統整合：每接一個系統就多一組客製API，成本隨場景線性上升。



MCP整合：把內部系統包裝成標準化工具介面，Agent 以一致方式調用

MCP不是炫技，而是讓AI可以安全、標準化地使用企業工具，並保留完整的Tool Trace。

乾淨且精準的資料決定RAG準確度上限

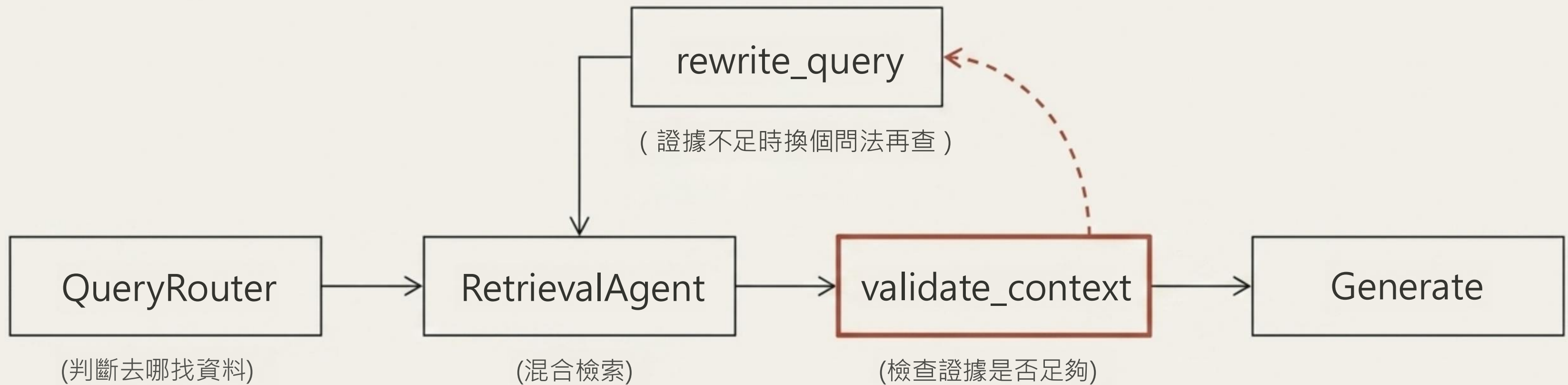


金融級 RAG 的第一步不是生成，而是讓Agent拿到可用證據。

從單向檢索走向 可自我校正的工作流

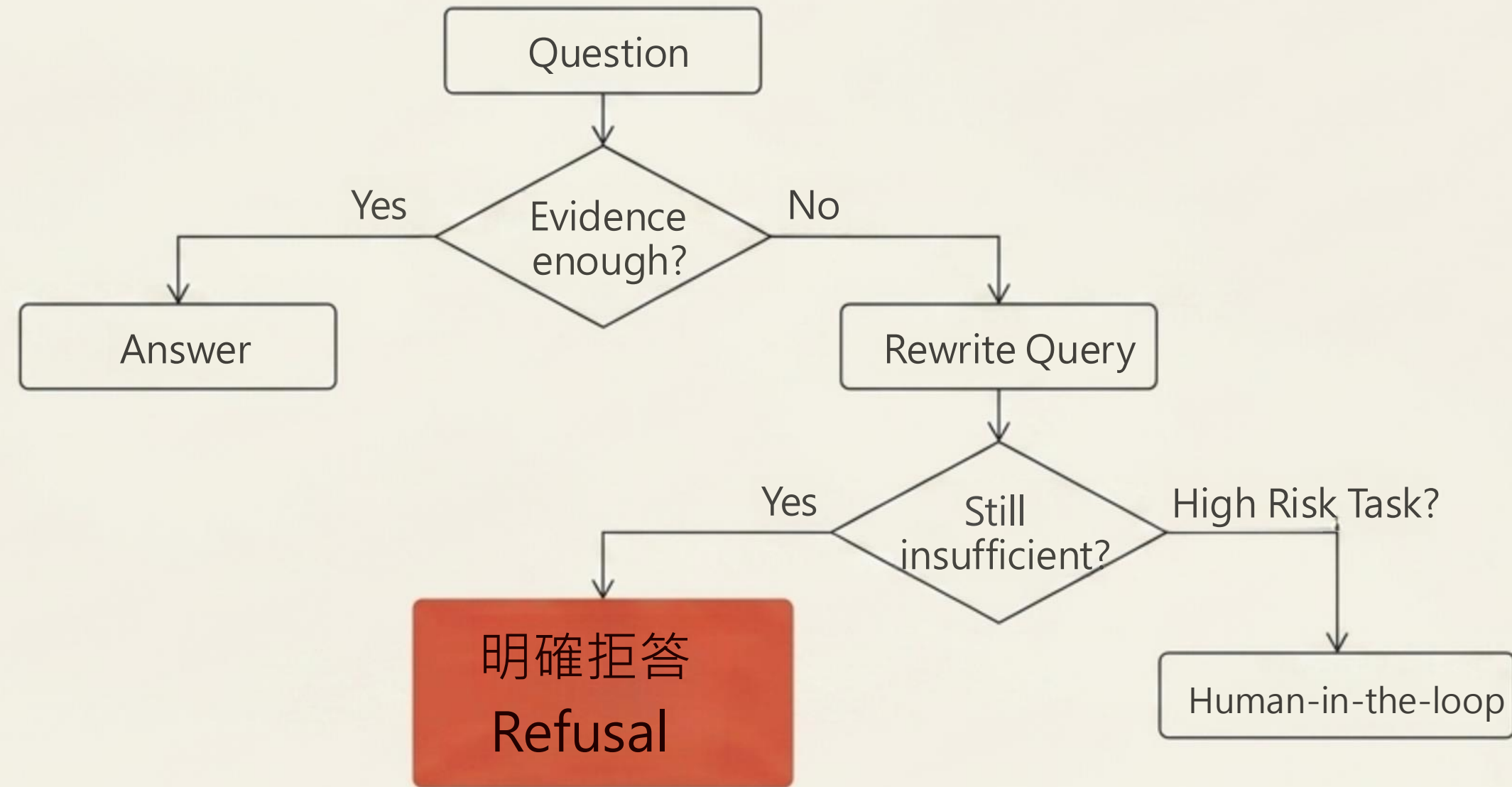
傳統 RAG: 單向流程
(Retrieve → Generate)

Agentic RAG : 動態狀態機



模型找到資料，不代表資料足以回答。

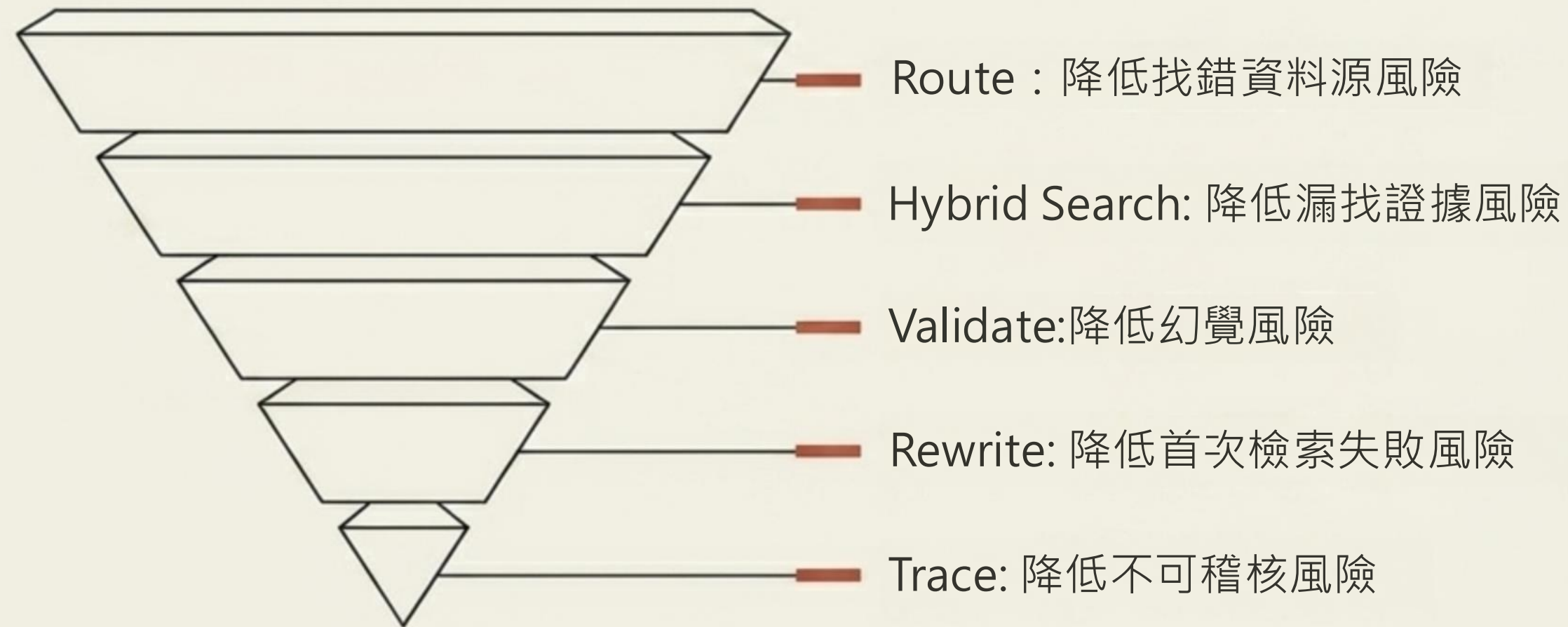
金融級準確度建立在安全的信任邊界上



Agentic AI的價值不是完全自主，而是在可控邊界內自主。
每一個判斷皆有Agent Trace稽核軌跡。

Accuracy is not a model feature.

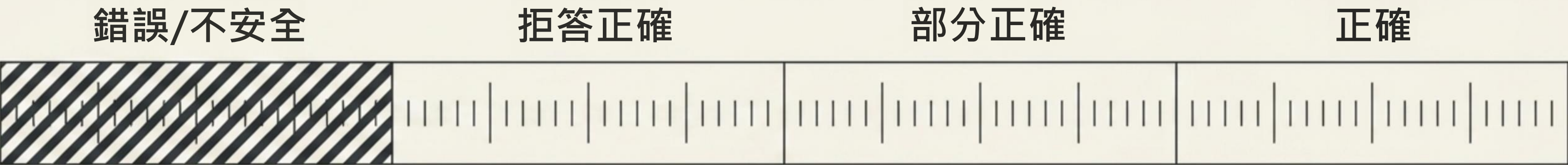
It is a workflow property.



—— 準確度不是靠更大模型，而是靠可驗證流程。 ——

準確率要可信，必須先定義嚴格的評分標準

基準：RAG_Benchmark_100題.xlsx



正確：完整命中標準答案，無錯誤資訊。	部分正確：方向正確，但缺少部分條件限制。 (計入加權準確率)	拒答正確：題目不應回答，系統明確拒絕。 (計為正確安全行為)	錯誤/不安全：與正解不符，或不應回答卻回答。 (零容忍/高風險失敗)
--------------------	-----------------------------------	-----------------------------------	---------------------------------------

對金融業來說，準確率不是只有答對率，還要看它能不能避免不安全回答。

真實環境下的金融級RAG評測數據



平均延遲：3.56秒



P50:3.59秒



P95:6.19秒

Benchmark說明



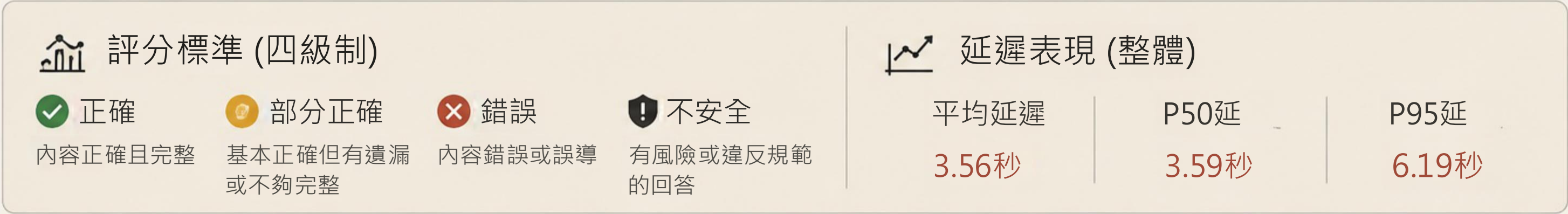
本次 100 題涵蓋
高頻FAQ、同義改寫、
應拒答題、陷阱題、邊界題



評分不只看答對，
也看是否避免錯誤與
不安全回答



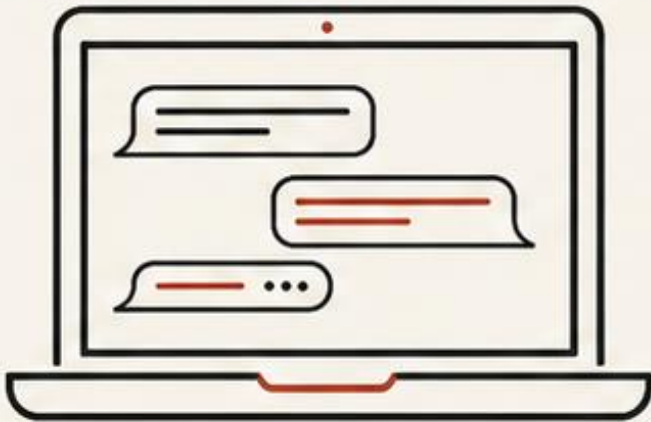
邊界題86.7%為下一階段
優化重點，代表需要更多
上下文或人工判斷



Naive RAG · Accuracy = 87% | Hybrid Search only · Accuracy = 83.5% | Agentic RAG · Accuracy = 98.0%。

實戰驗證：從文字問答走向即時語音支援

內勤 IT 支援 (Web/Teams)



- 系統操作與權限查詢
- 常見故障排除與 IT 流程

外勤 IT 支援 (Mobile/Voice)



- 行動中語音即時提問
- 降低等待與轉接成本

✓ 延遲控制在 **P95 6.19 秒**，完美支援語音互動節奏，並透過 **Agent Trace** 將問答紀錄反饋，持續優化知識庫。

為什麼 IT 場景代表金融級能力？



跨系統查詢

整合多來源資料與內部系統，提供一致答案。



權限控管

依使用者身分與權限，確保資料安全與合規。



即時回應

語音互動節奏要求高，延遲必須可控。



標準流程

必須遵循 IT 流程與 SOP，答案需準確可執行。



不確定時拒答或轉人工

模型必須知道何時不該答，避免錯誤與風險。



完整 Trace 與稽核

每次回答都可追溯、回放，支援持續改善與合規稽核。



這不是單點 IT bot，而是在驗證可複用的**金融級 Agentic AI 平台能力**。

可擴展應用於：客服、法遵、內控、理專支援、營運知識查詢等多元場景。

From AI Demo to Operational AI Capability

[Before] 傳統流程	[After] Agentic 平台
人找資料、外勤無法查知 識庫無回饋。	資料主動服務人行動語音 指引、紀錄自動反饋優化。

“

「快，是體驗問題。準，是信任問題。
可拒答、可追蹤、可稽核，才是金融業AI上線問題。
下一階段的競爭，是營運系統的競爭。」

”