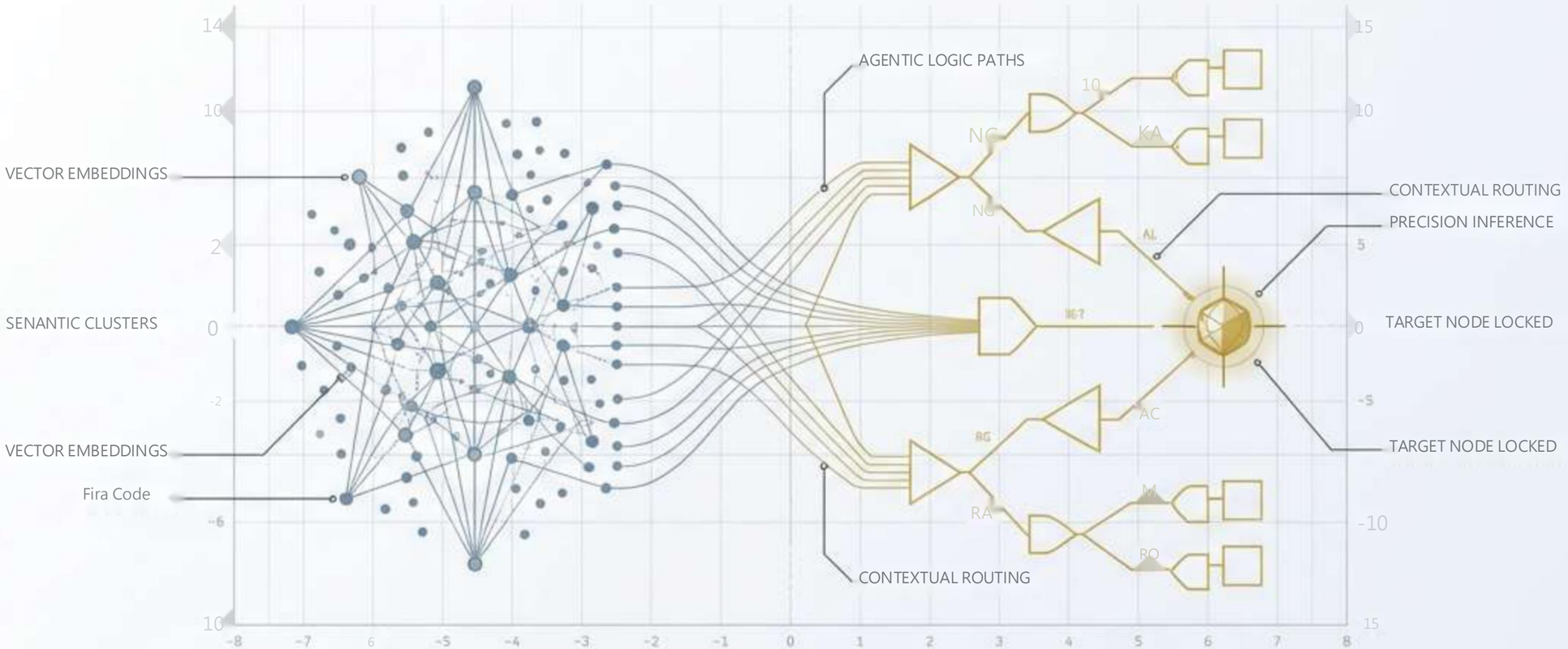


RAG 2026：當向量搜尋遇上代理推理

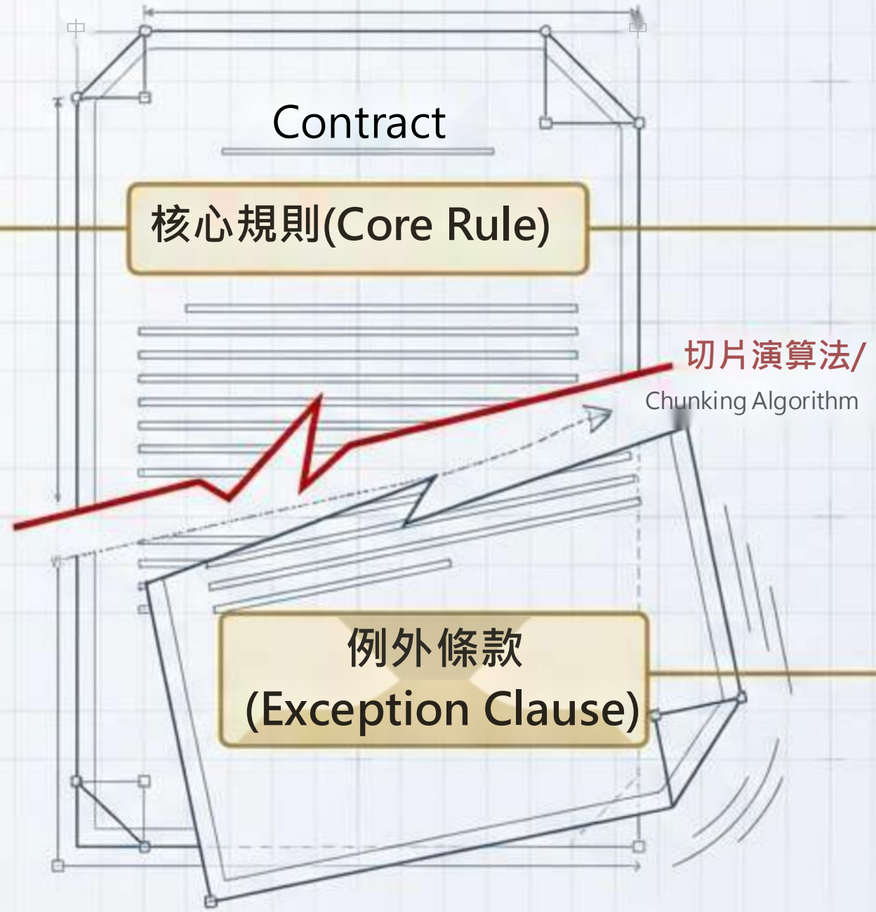
企業級 99%準確度AI架構的演進與實踐指南

「上下文盲視」揭示了 RAG1.0建立在脆弱的統計假象之上

The Anatomy of Context Blindness

企業痛點

企業入數萬份文件，系統卻頻繁引用2023年的舊合約來回答2026年的決策問題。



根本原因

語義相似性≠事實正確性。

向量空間的鄰近性僅代表詞彙分佈的接近，**完全無視邏輯真實性與版本有效性。**

單純的切片策略會導致**準確度下降超過60%。**

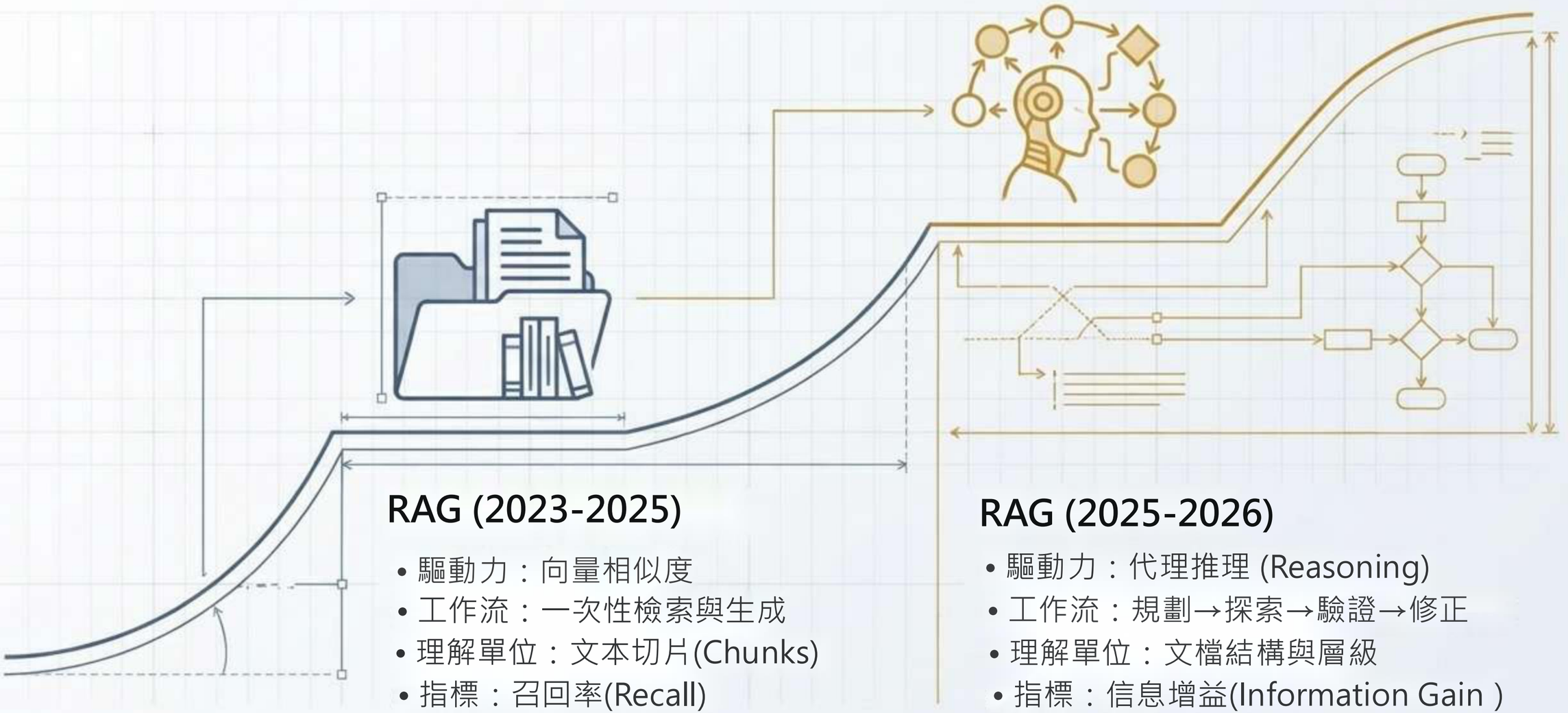
向量資料庫依然是處理千萬級數據的極速堡壘

解決方案	核心優勢	最佳應用場景
Pinecone	大規模RAG（穩定性與擴展性）	企業級生產環境
OpenSearch	混合搜尋精度(詞法+語義)	合規性嚴格環境
PGVector	現有資料庫集成度	中小團隊與輕量開發
Redis	極低延遲與內存運算（TTFT)	實時應用與語義緩存
LanceDB	本地優先靈活性	邊緣運算與原型設計

結構性致命傷（TheFatal Flaw)

索引延遲與上下文混淆。尋找「增加預算」時，系統常因語義模糊撈出「削減開支」的片段。

從「靜態圖書管理員」進化為「主動探索的虛擬研究員」



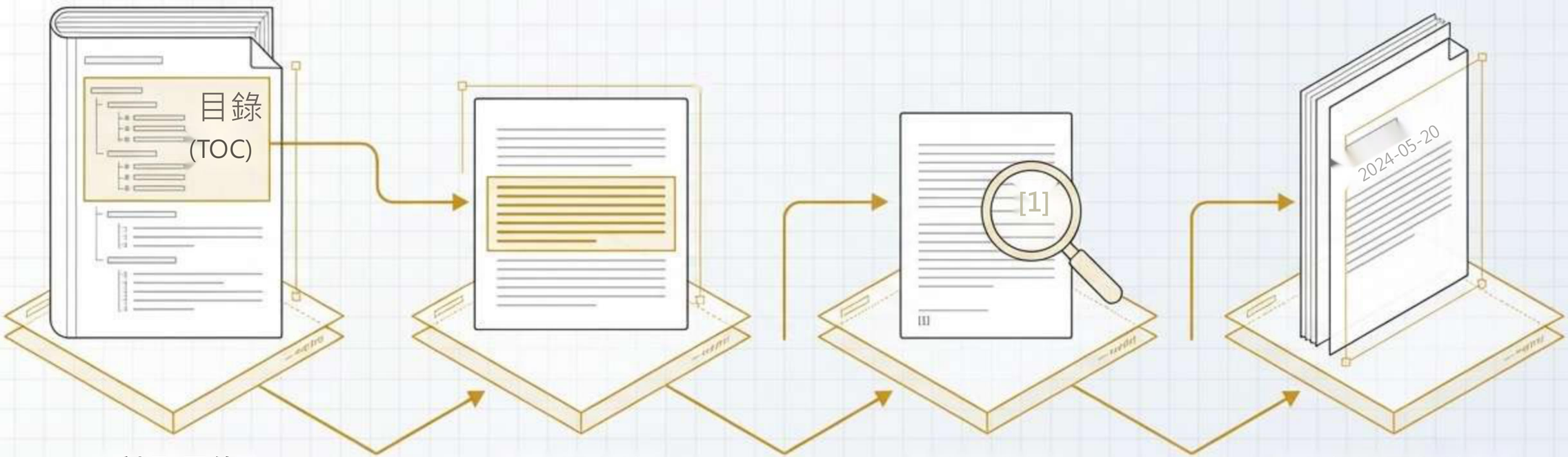
RAG (2023-2025)

- 驅動力：向量相似度
- 工作流：一次性檢索與生成
- 理解單位：文本切片(Chunks)
- 指標：召回率(Recall)

RAG (2025-2026)

- 驅動力：代理推理 (Reasoning)
- 工作流：規劃→探索→驗證→修正
- 理解單位：文檔結構與層級
- 指標：信息增益(Information Gain)

代理檢索徹底放棄向量坐標，改以結構與邏輯導航



1.讀取目錄
(Read TOC)

2.目標章節
(Target Chapter)

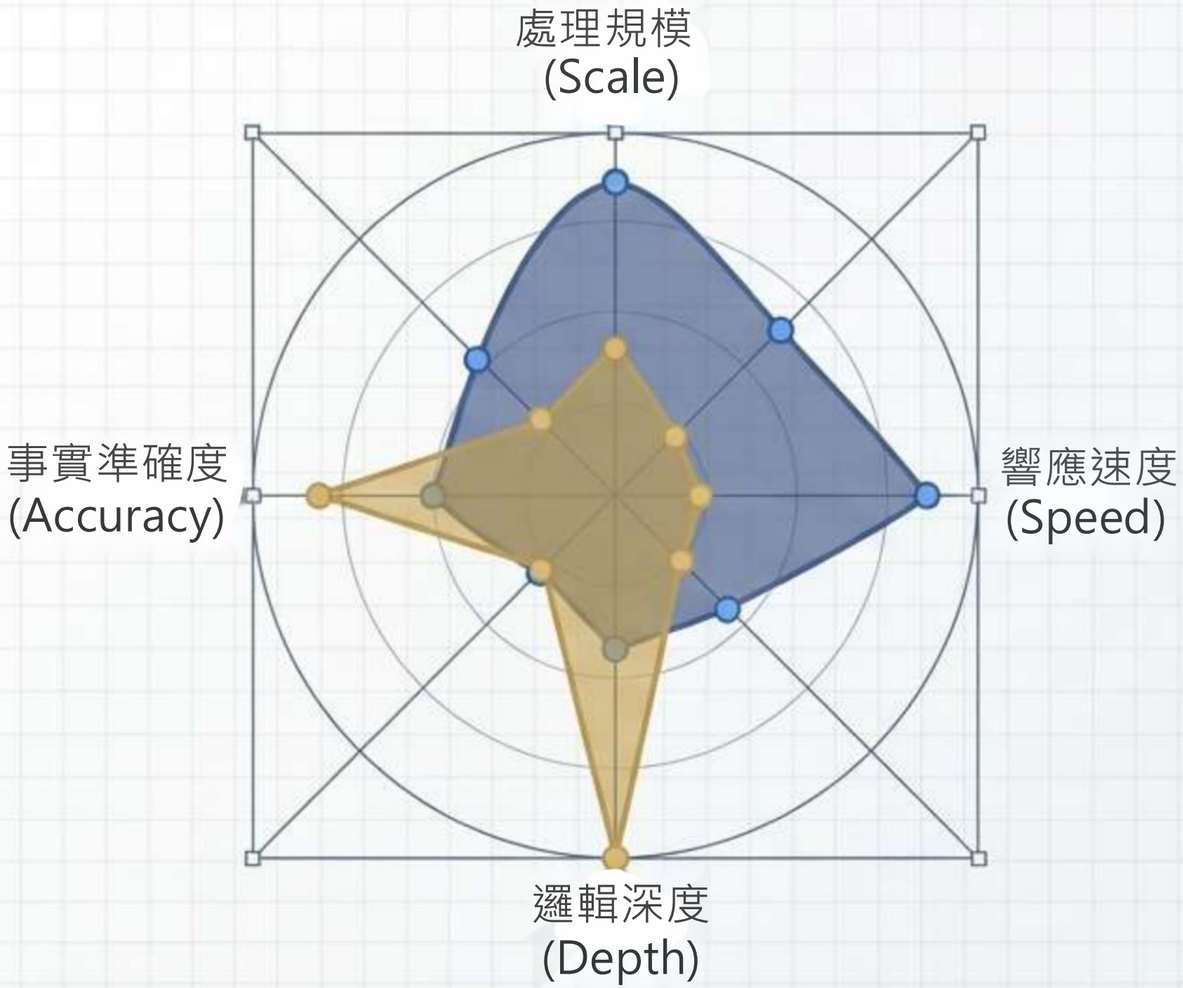
3.發現腳註
(Spot Footnote)

4.動態調用與核對
(Dynamic Fetch)

關鍵突破：零延遲知識更新

無需耗時的Embedding與重新索引，檔案存入系統瞬間即可被Agent透過目錄與元數據 (doc_id，sec_id)讀取與推理。

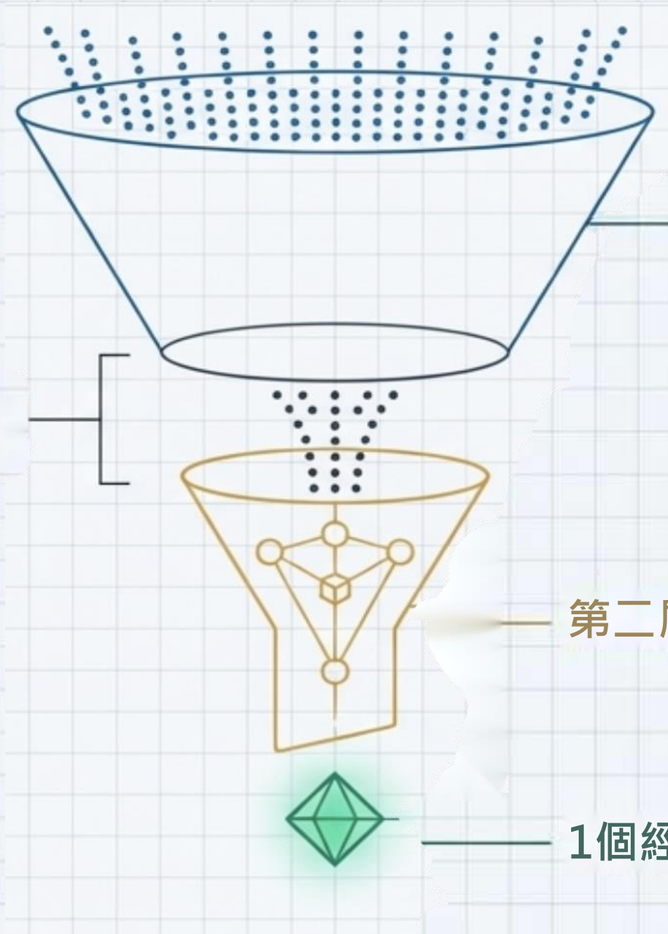
向量範式與代理範式的量化博弈揭示了顯著的性能權衡



評估維度	向量檢索 (Vector)	代理檢索 (Agentic)
資料類型	巨量非結構化散文	結構化合約/報表
核心優勢	秒級響應與規模	極致邏輯 (+18%準確)
延遲代價	幾乎無延遲	首字延遲平均 高出3.81s
典型準確率	60%-80%	85%-95%+

當查詢涉及超過5個實體或需要多步推理時，
向量準確率迅速歸零，代理檢索維持穩定。

2026企業標準：結合兩者優勢的「多層次混合架構」

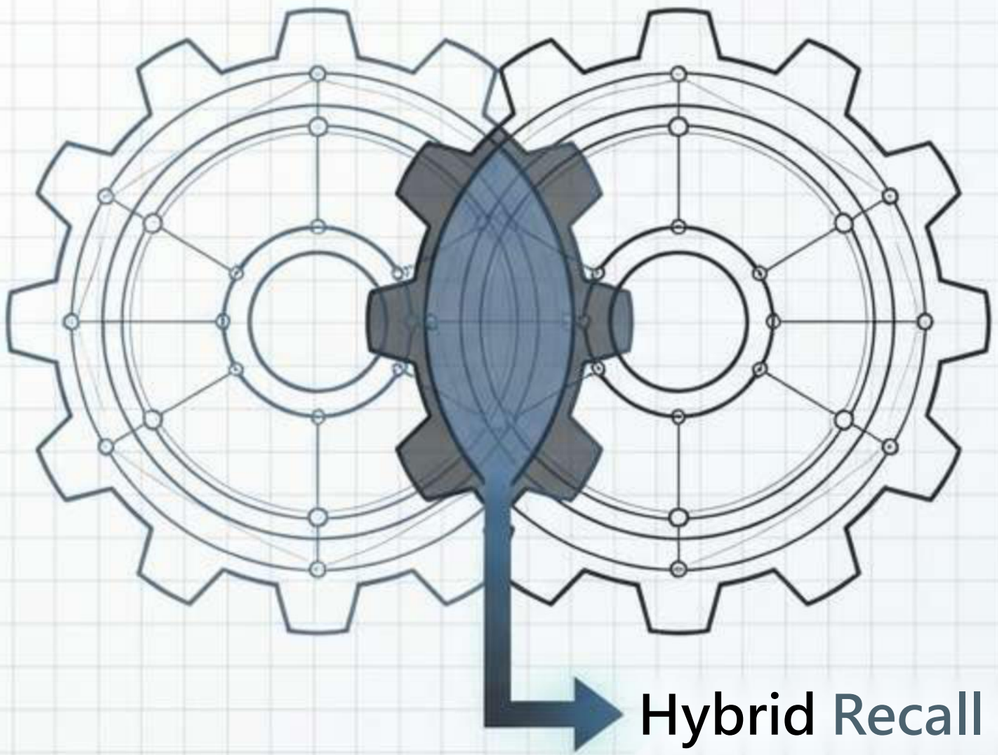


「向量檢索負責幫你縮小範圍，代理推理負責幫你找到唯一真相。」

第一層防線：混合檢索確保100%的高價值數據召回

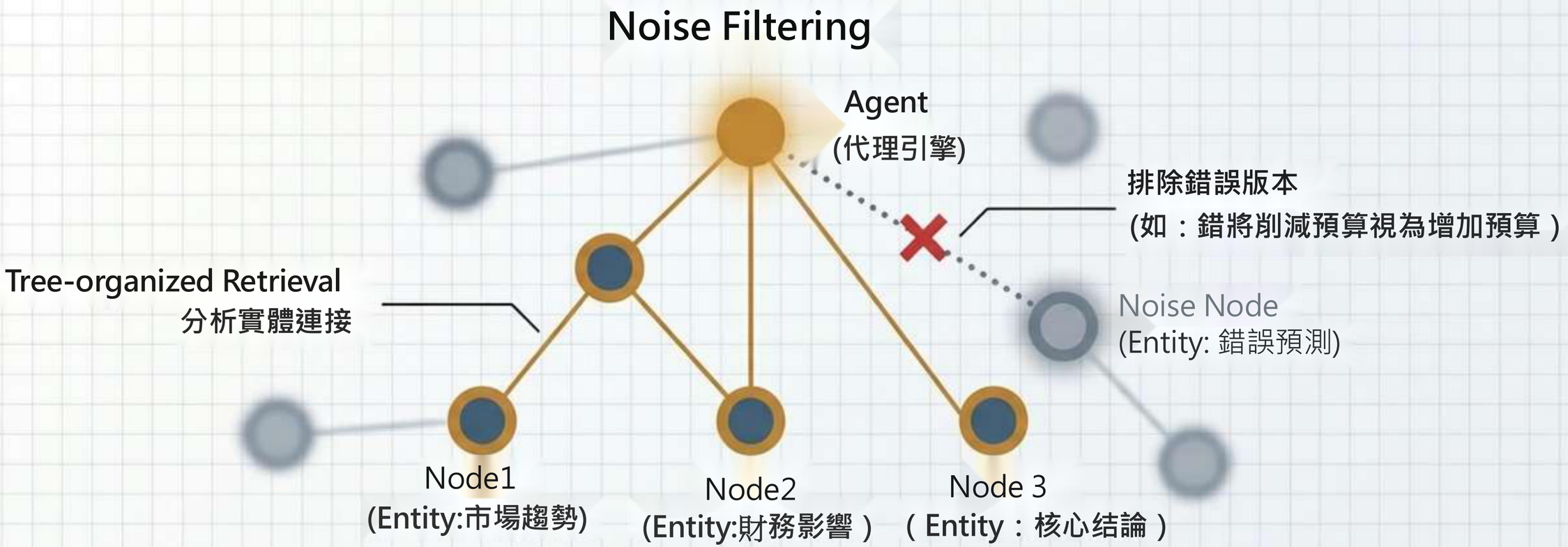
Vector Embeddings
(語義氛圍)

BM25
(精確關鍵字/稀疏信號)



- 單純的向量搜尋容易漏掉精確的ID、產品代碼或特定術語。
- 透過結合Vector與BM25，系統能在百萬級文件中快速鎖定候選範圍，並將召回率(Recall)顯著提升20%到40%。這確保了Agent在下一階段有足夠的正確材料進行推理。

第二層代理精讀(Agent)：代理引擎利用圖譜與邏輯排除語義噪聲



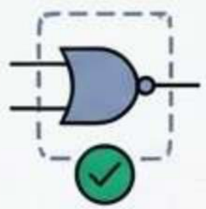
在這層，Agent進行「無向量」的精確閱讀。它分析實體連接與因果關係，
透過交叉驗證排除第一層帶入的錯誤版本，將系統準確度推升至極限。

99%準確率的數學保證：驗證鏈 (V-CoT) 與符號邏輯

$$Verification\_Score = \frac{\sum_{i=1}^{n} Grounding(Step_i, Context)}{\text{Total Steps}}$$

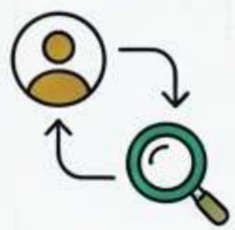
步進式演繹校驗
(Stepwise Deductive Validation)

每個中間結論轉化為符號邏輯，由Z3求解器進行無衝突驗證。



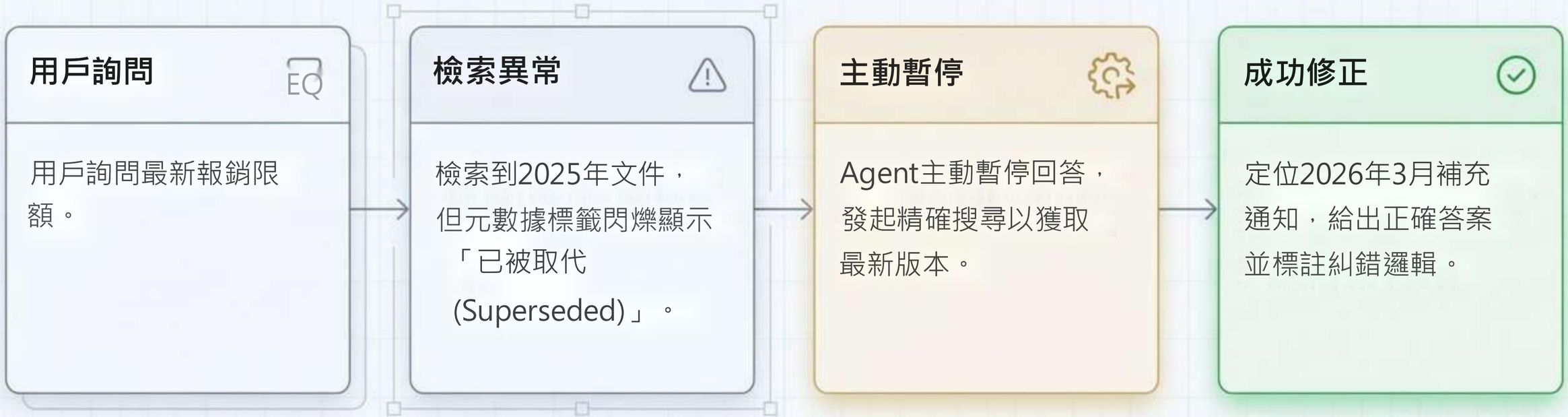
多智能體反思
(Multi-agent Reflection)

推理代理 (Reasoning Agent)生成答案，批判代理(CriticAgent）專職尋找漏洞與核對版本。



從「黑盒生成」轉變為「可追蹤的透明決策路徑圖」

贏得信任的關鍵：展示AI主動糾錯的「思考鏈」



這種「主動修正」與「擁有其推理過程（Owning the Reasoning Process）」的能力，是AI走向高風險生產環境的信任基礎。

代理架構的最終防線：「主權盾牌」與企業級治理

行為軌跡審計-提供100%決策可追蹤性，符合《歐盟AI法案》。

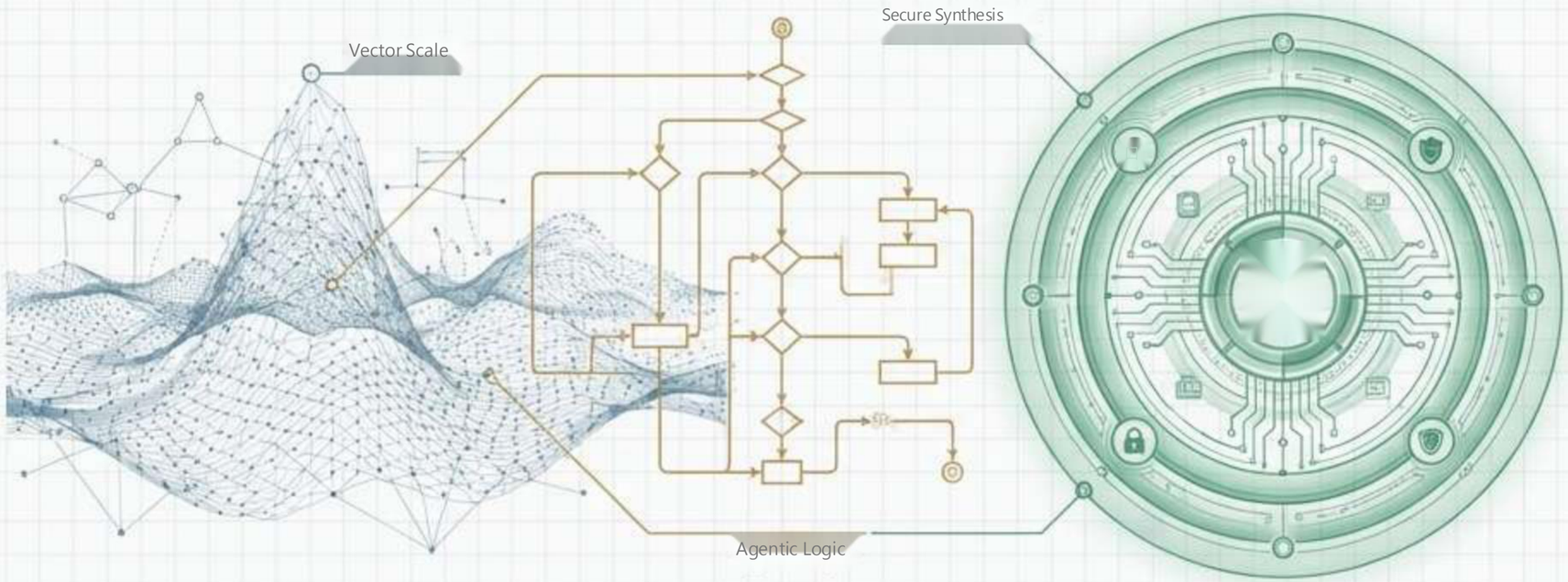
內容審查與過濾-PII(敏感數據攔截)與PromptInjection（注入防禦）。

自動化推理檢查(Automated Reasoning Checks)-形式邏輯驗證合法性。

行為軌跡審計 - 提供100%決策可追蹤性，符合《歐盟AI法案》。

遵循「**最小權限原則**」，將治理邏輯直接嵌入代理的獎勵函數中，確保其始終在安全與倫理軌道內運行。

企業生產力的重塑：從數據收集者到決策協調者



「向量縮小範圍，代理找到真相。」

2026年的RAG前沿不再是「更好的Embedding」之爭，而是「最優決策流程」的博弈。

分析師將不再浪費80%時間核對数據，管理者將利用AI代理在數分鐘內完成跨部門數據對齊。這是一場從語義相似到邏輯真實的深刻跨越。