

# 著作權聲明公告

本活動資料(包括但不限文字、圖片、影音等講義內容或其他文件)之著作權歸屬於「財團法人資訊工業策進會產業情報研究所(MIC)」所有，均受著作權法之保護及國際著作權法律的保障，僅授權報名本活動之個人非商業用途，並請註明引用出處及來源。

謹提醒，倘個人未取得書面授權同意，逕自透過電子郵件或LINE等媒體媒介轉載分享本活動資料，已經侵犯MIC的智慧財產權，將視情節之重大程度提出法律追訴，如經確認違法行為，不僅個人受罰，公司亦將負連帶賠償責任，並造成公司商譽之損害。

感謝您對於智慧財產權尊重與理解，如有意請求授權使用本活動資料，歡迎聯繫02-2378-2306；members@iii.org.tw，謝謝您！

# 開源模型發展趨勢

朱師右

產業顧問兼組長

產業情報研究所(MIC)

財團法人資訊工業策進會

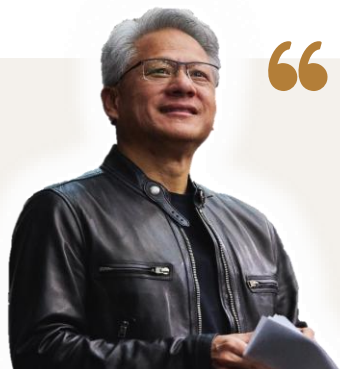
2026/07/16

# 簡報大綱

- ◆ 開源模型發展現況
- ◆ 開源模型技術動態
- ◆ 應用模式與落地場景
- ◆ 企業導入之管理議題
- ◆ 結論

# 開源模型發展現況

# 開源模型市場成長帶動企業投入與整合



“

「開源模型的快速崛起成為需求的驅動力，其普及帶動使用量增加，並受到科技大廠合作的推動」

NVIDIA CEO 黃仁勳

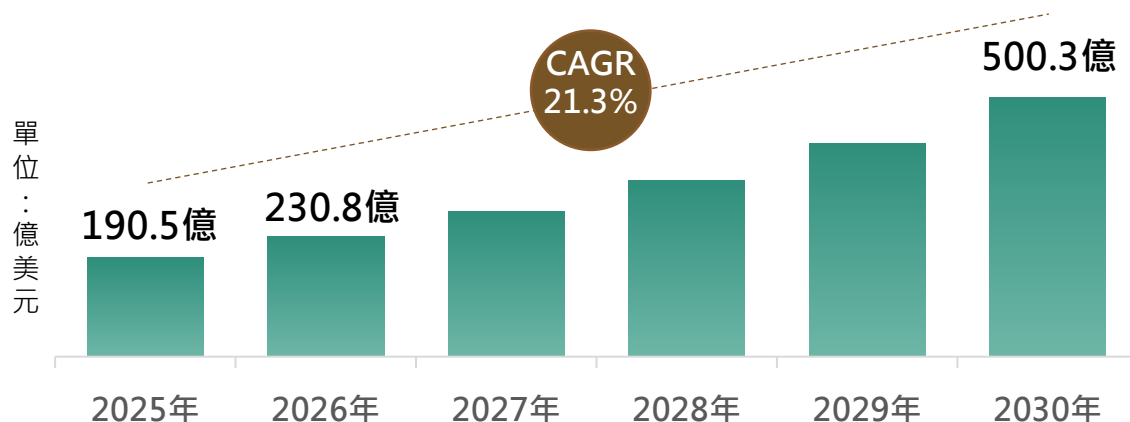
“

「開源確保更多人獲得AI的利益與機會，使權力不集中於少數公司，讓技術更均衡且安全地佈署於社會」

Meta CEO Mark Zuckerberg



## 全球開源AI模型市場預估



資料來源：The Business Research Company，MIC整理，2026年7月

## 年度重點級併購

| 收購方        | 被收購方(定位)                 | 交易金額     |
|------------|--------------------------|----------|
| Meta       | 擬收購Manus ( AI Agent執行層 ) | 約20億美元   |
| servicenow | MoveWorks ( 企業AI助理 )     | 約28.5億美元 |
| workday    | Sana Labs ( 企業知識AI )     | 約11億美元   |
| NiCE       | Cognigy ( 對話式AI )        | 約9.55億美元 |

註：Meta與Manus交易仍依公開市場資訊及監管進度為準

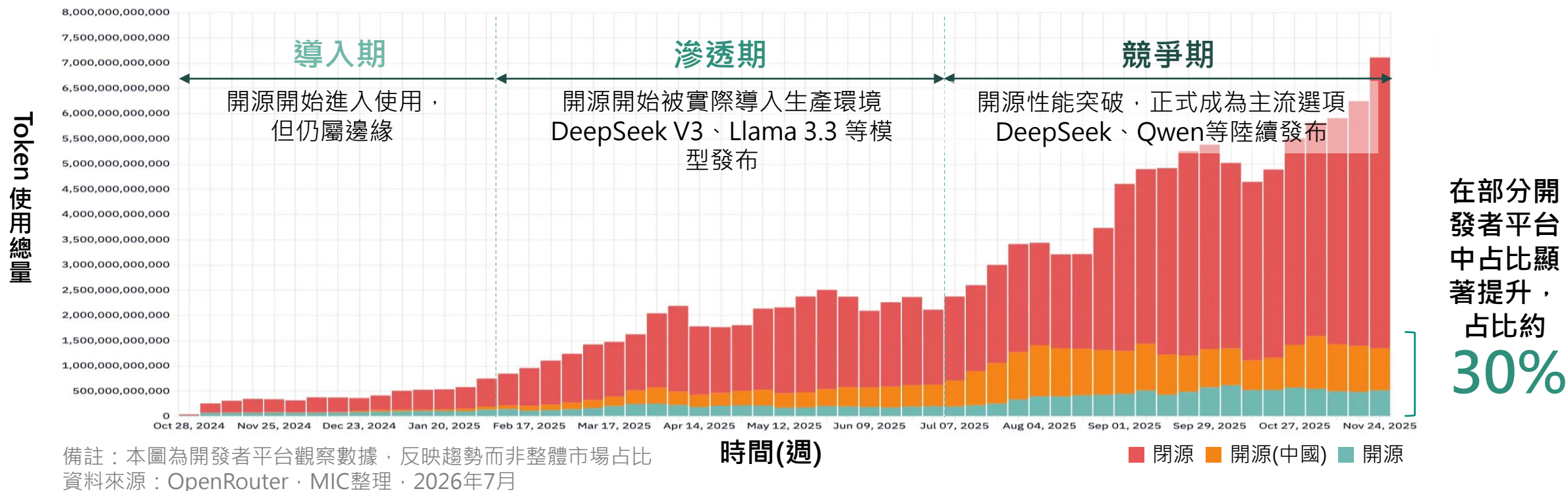
資料來源：CB Insights，MIC整理，2026年7月

- ◆ 開源模型市場高速成長，反映企業對低成本、可客製與主權可控AI的需求提升，導入從實驗走向規模化應用
- ◆ 大型科技業者近年的投資重點已由 Foundation Model 轉向 AI Agent、企業工作流程 ( Workflow )、知識管理與應用整合平台，反映 AI 商業化已由模型競爭進入企業落地競爭階段

# 開源模型使用量快速成長，已進入實務應用階段



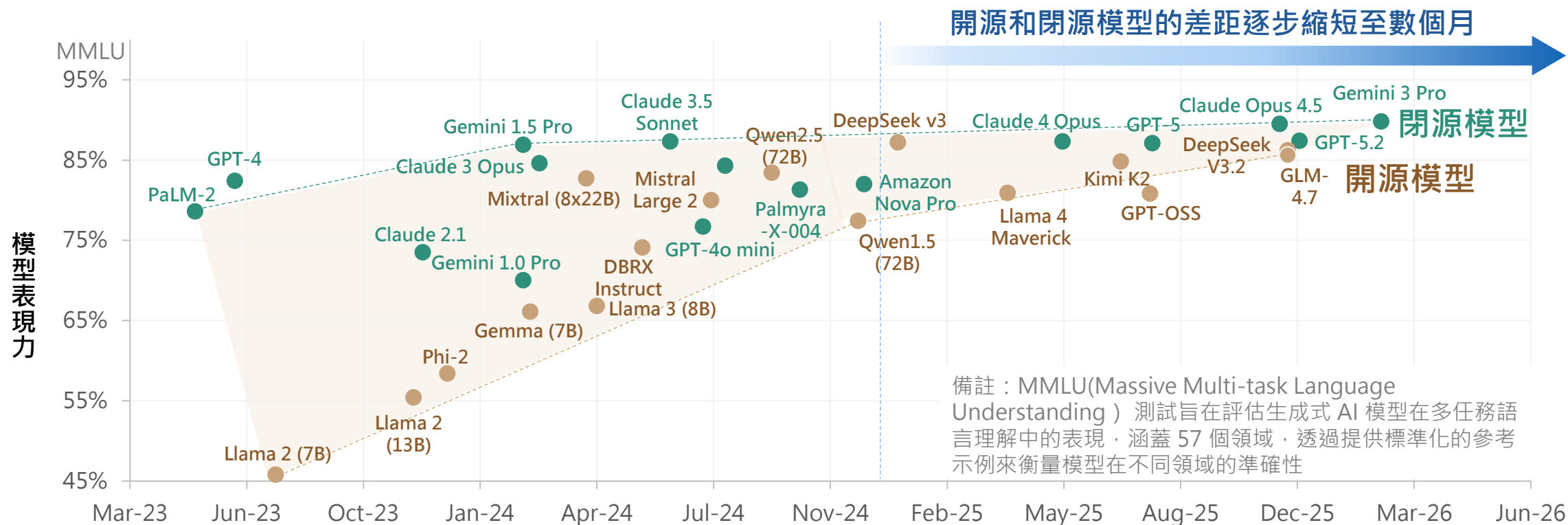
## 開源與閉源模型Token使用量



- ◆ 開源模型使用量快速成長，顯示已由測試階段進入實務應用，顯示在短時間內快速滲透至實際應用場景
- ◆ 開源模型使用結構由高度集中轉向分散，單一模型占比明顯下降，不同地區模型參與度提升，顯示市場競爭走向多元化，市場呈現多方競爭格局

# 開源模型加速追趕縮短與閉源差距

## 開源與閉源模型效能比較



資料來源：MIC，2026年7月

- ◆ 開源模型與閉源模型的性能差距由約12個月顯著縮短至數月等級，顯示開源已具備快速追趕能力，企業可更即時採用成熟開源方案，並在多項基準測試中，開源模型表現持續逼近甚至接近閉源水準

# 小結：開源由市場擴張走向能力內化

## 01 使用滲透

### 開源進入核心應用

- 使用量跨越臨界點（~30%），從輔助工具轉為主流程組件
- 採用動機由實驗導入轉為成本、可控與落地能力驅動

## 能力追趕 02

### 開源進入可替代區間

- 性能差距壓縮，企業接受次優解換取成本與控制
- 推理與長上下文能力成熟，支撐實際任務

## 03 競爭重構

### 競爭轉向運行效率

- 模型本身快速商品化，差異轉移至推理成本、延遲與佈署架構
- 優勢來源由「模型訓練」轉為「推理優化與工程能力」

## 應用分工 04

### 市場走向任務組合

- 使用集中於高頻任務，形成穩定分工結構
- 企業由單一模型策略轉向多模型組合與任務路由

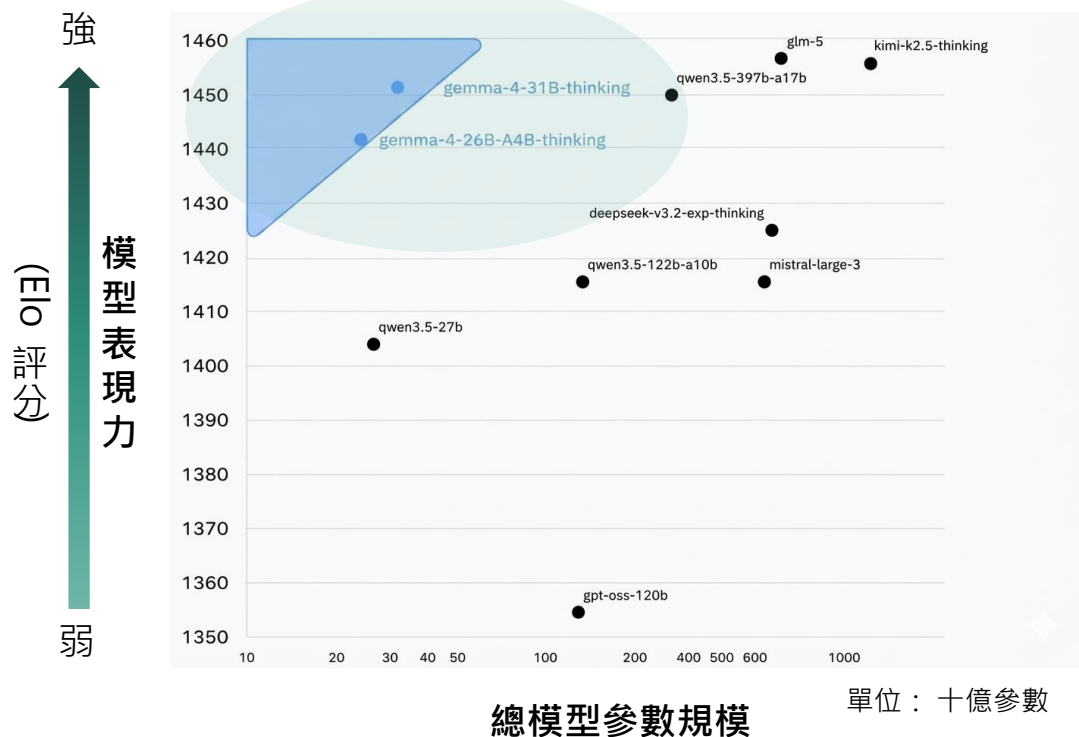


- ◆ 開源已從成本替代轉為企業建立可控AI能力的基礎，競爭關鍵不在模型強弱，而在能否快速將模型接入流程並穩定運行，體現於佈署、推理與系統整合能力

# 開源模型技術動態

# 推理模型走向本地佈署與成本優化

## 代表性本地模型



資料來源：Google · MIC整理 · 2026年7月

## 為甚麼Gemma 4「跑得動」？

- 1 先進MoE稀疏架構**  
每次只啟用部分專家 ( ~1/8 )，降低每token算力
- 2 極致量化 ( MXFP4 )**  
透過4-bit Microscaling壓縮權重，使大模型可在單張80GB GPU運行
- 3 層次化推理模式**  
支援「快速回應 / 多步推理」切換，依任務動態分配算力
- 4 硬體優化**  
深度適配 GPU / TPU / Apple Silicon，最大化 token throughput

## 佈署場景



企業知識庫 /  
RAG



邊緣設備 /  
離線AI



客服 /  
即時應用



開發者代理 /  
自動化

- ◆ 高階推理模型已可在本地以可控成本運行，企業AI策略將由單一雲端依賴轉向多模型與混合佈署，並以推理效率、硬體適配與成本治理作為核心選型標準

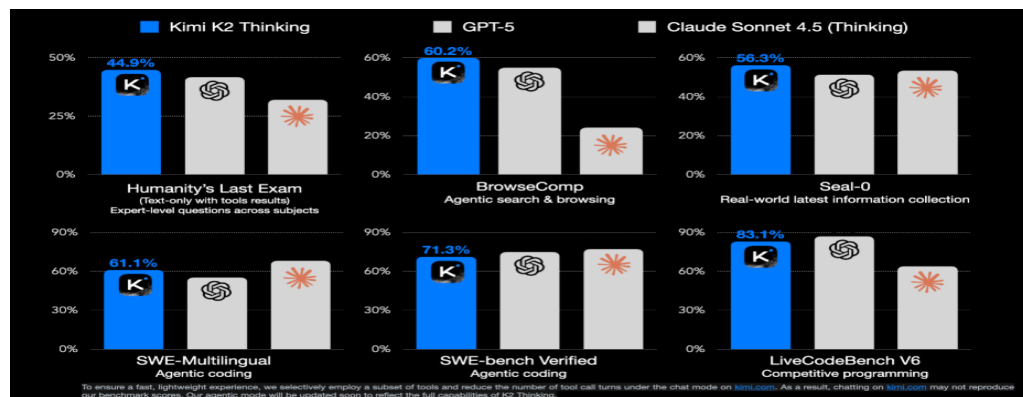
# 推理深度與記憶長度成為可控能力

## Thinking Mode ( 推理深度 )

先思考再行動，強化多步推理與決策穩定性

代表模型：Kimi K2 Thinking

- Thinking Mode結合原生256K上下文（可擴展至1M），使模型在推理、Agent與程式競賽任務上領先GPT-5與Claude Sonnet 4.5

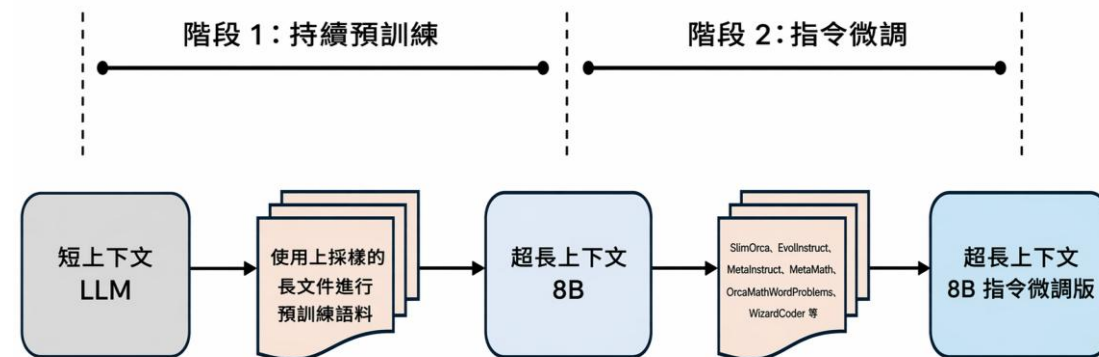


## Long Context ( 記憶長度 )

長記憶能力模組化，支撐長文件、工具歷史與多步流程

代表模型：NVIDIA Nemotron-UltraLong-8B

- 基於Llama-3.1-8B延伸訓練，透過持續預訓練擴展至1M / 2M / 4M token超長上下文，並以少量SFT恢復通用能力，同時取消跨文件注意力遮罩完整保留任務脈絡



💡 模型在輸出答案前，先展開思維鏈進行推理，再透過Prompt或模型設計引導，提升多步任務的決策穩定性與可控性

備註：SFT ( Supervised Fine-Tuning ) 指令微調

資料來源：Kimi、NVIDIA，MIC整理，2026年7月

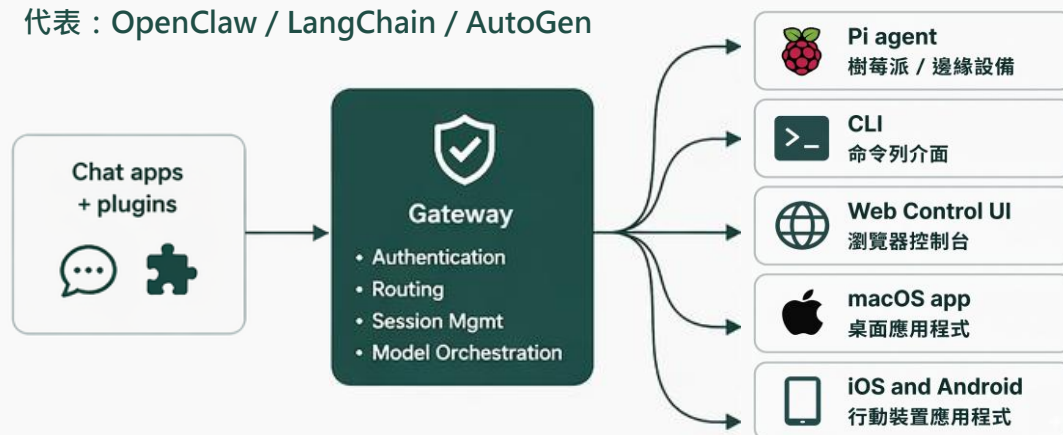
- ◆ 開源模型讓推理與記憶能力從封閉機制轉為可控資產，企業可依業務需求調整推理策略與資料結構，於成本效率與資料主權之間取得最佳平衡，並逐步建立差異化競爭力



# AI Agent發展兩大路線：流程整合 vs 任務執行

## 企業流程整合型 ( Workflow Integration )

代表：OpenClaw / LangChain / AutoGen



以開源框架將LLM的「推理 + 工具 + 流程控制」標準化，讓企業可自建可控的AI自動化系統並降低供應商依賴



把AI嵌入企業流程，強化系統整合與自動化

( 如：內部知識庫、流程自動化、IoT 管理、企業服務代理等 )

資料來源：MIC，2026年7月

## 任務代理執行型 ( Autonomous Agent )

代表：Claude Cowork / Hermes / OpenManus



以開源模組化設計強化推理與工具調用能力，讓開發者可快速組裝並擴展AI Agent應用



讓AI主動規劃與執行任務，成為數位工作夥伴

( 如：個人助理、研究原型、開發工具、資料處理、輕量自動化等 )

V.S

- ◆ AI Agent正形成「流程整合」與「任務執行」兩種發展模式，分別對應企業級系統治理與敏捷應用需求未來競爭關鍵，不在模型本身，而在企業是否具備將AI嵌入工作流、並持續優化與規模化的能力



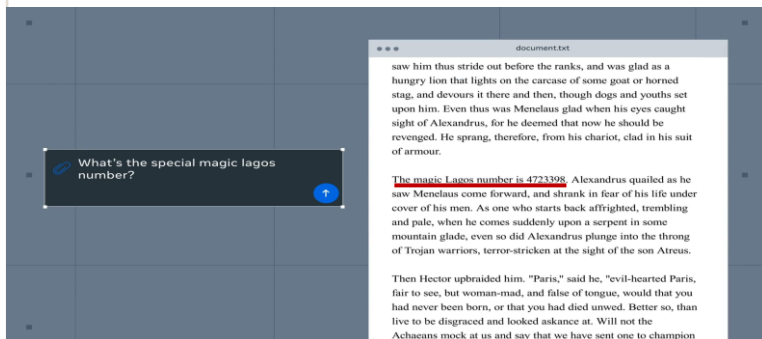
# 多模態能力將模型升級為通用資料入口

## ① 文件(長文本處理)



**Llama 4**

超長上下文理解 · 接管企業知識流



- 支援百萬級Token超長上下文
- 跨文件推理與知識整合，適合企業知識庫

## ② 圖像 (影像理解)



**Gemma 4**

理解影像與世界 · 讀懂非結構資料



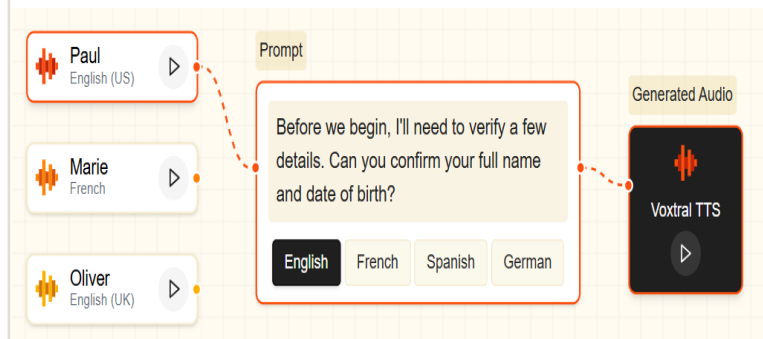
- 原生支援圖像、影音、文字等輸入
- OCR、物件偵測、視覺推理等多步驟理解

## ③ 影音 (語音 / 影片)



**Mistral Voxtral**

語音理解與生成 · 打造即時互動入口

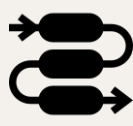


- 多語語音生成，支援情緒與風格控制
- 低延遲，可嵌入客服、語音 Agent 等場景



**直接接收原始資料**

語音 / 影像 / 文件等可直接輸入模型



**取代多重前處理流程**

不再依賴 ASR、OCR、特徵擷取等轉換



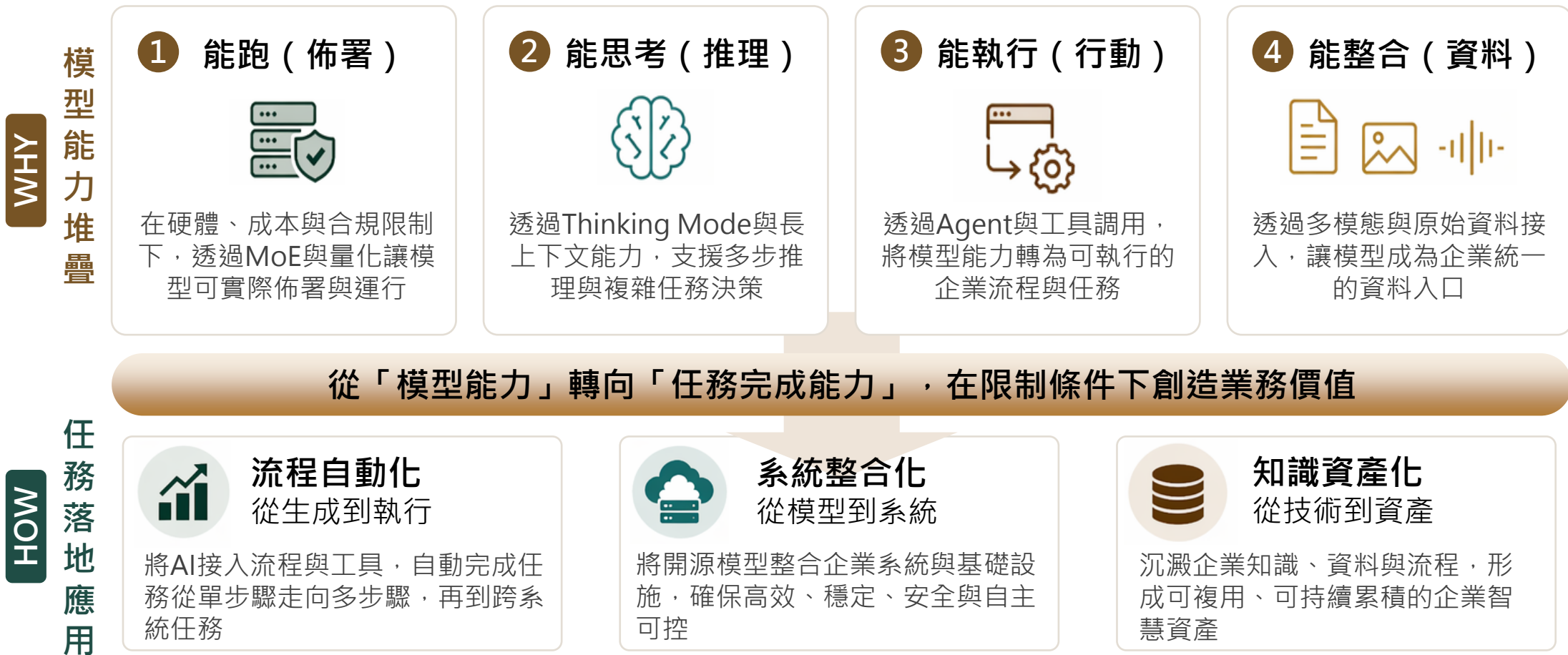
**AI成為企業資料操作介面**

從對話摘要到決策與執行，全面串接資料流

資料來源：Meta、Hugging Face、Mistral，2026年7月

- ◆ 開源多模態使模型可直接接管原始資料流並取代前處理流程，進一步將AI從生成工具升級為企業資料與決策入口，重構系統架構與價值鏈主導權

# 小結：開源能力由模型突破走向可佈署與可執行



- ◆ 開源競爭已由模型能力轉向任務完成能力，企業關鍵不在選擇最強模型，而在能否以最低成本將模型穩定嵌入流程、串接資料並持續產出可量化業務價值

# 應用模式與落地場景

# 流程自動化：開源Agent推動任務自主執行



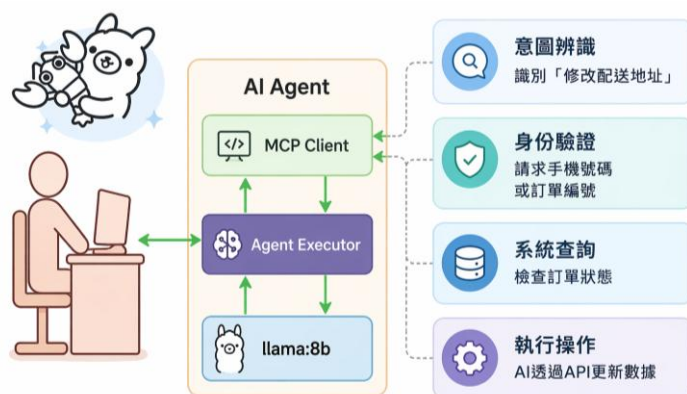
- ◆ AI從單純的「諮詢對話」進化為「實務執行」，透過Agent實現跨系統、跨平台的無人化作業閉環，徹底解放行政人力
- ◆ 將繁瑣的人工點擊流程轉化為AI可調用的功能單元，落實真正的數位勞動力 (Digital Labor)

# 開源Agent驅動跨系統任務自動化，快速執行企業流程

任務複雜度與自動化程度提升

## 單點回應（對話輔助）

即時回應・任務拆解



- ◆ 企業透過開源模型結合RAG與工具調用，可直接串接CRM、訂單與物流系統，使AI不只回覆問題，而是完成整段服務流程（如退款、改單、補寄）

對話助手 / 單點任務

## 流程決策（任務編排）

多步驟流程・任務編排



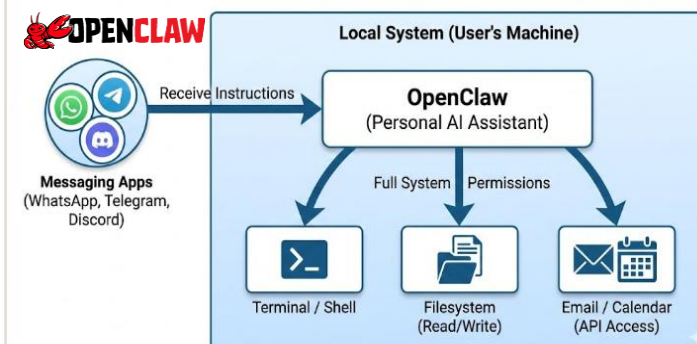
- ◆ 在財務與採購場景中，透過RAG整合企業規則與歷史資料，完成發票核對、異常偵測與風險判斷，並直接觸發ERP系統（如SAP/Oracle API）執行付款或凍結流程

RAG / Tool use / Workflow

## 跨系統執行（流程自動化）

跨系統整合・自主執行

OpenClaw Local Architecture (Simplified)



- ◆ 開源模型因具輕量化與本地佈署特性，可整合進企業內部IT與營運系統。結合OpenClaw等Agent框架後，能直接操作電腦、執行腳本、跨系統整合（Email、ERP、報表工具），完成日常營運任務

Agent / MCP / API

# 開源架構驅動企業走向在地部署與自主AI能力

企業AI自主化三大能力（資料、運算、應用）

## 資料在地化、知識內化

資料與系統整合 + 記憶知識

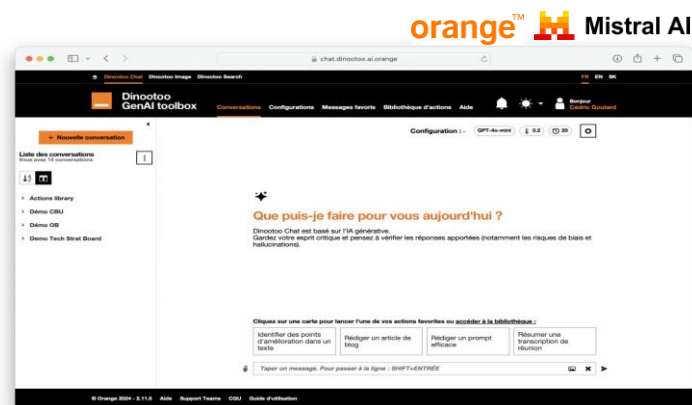


- ◆ 導入Llama並佈署於內部環境，結合客服紀錄與合規文件建構私有RAG知識系統，使AI回應完全基於內部資料生成，確保敏感金融資訊不外流

RAG / 私有資料 / 知識庫

## 模型可控、彈性部署

開源模型 + 工具/API + 部署層



- ◆ 歐洲企業採用Mistral模型並佈署於本地或歐洲雲環境，建立客服與文件處理系統，在符合GDPR規範下避免依賴跨境雲服務，逐步強化數位主權與AI自主能力

開源模型 / API / 混合部署

## 跨系統整合、即時執行

應用層 + 跨系統整合



- ◆ 透過與MES與現場控制系統整合，AI不需回傳雲端即可完成決策，顯著降低延遲與傳輸成本，同時確保敏感生產資料不外流，建立具備即時性與安全性的邊緣AI運作模式

Agent / 系統整合 / 即時決策

# 開源驅動企業AI資產循環提升資料價值

開源架構（模型 + 工具 + 佈署）是資產內化的基礎

## 知識內化



將企業經驗轉為可運行能力



- ◆ 透過開源模型與RAG架構，將內部知識庫、營運數據與AI能力整合，形成可持續學習與優化的數據生態，使AI從「生成資訊」進一步轉化為「驅動決策與營運的核心系統」

RAG / 私有知識庫 / Prompt優化

## 數據累積



讓每次使用轉為可優化資產



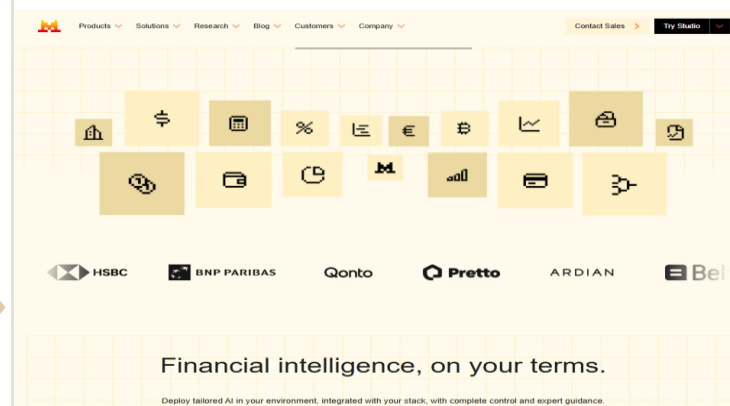
- ◆ 透過分析用戶點擊、搜尋與購買行為，動態調整推薦邏輯與行銷策略，讓AI直接介入交易流程。每次用戶互動產生新的行為數據，系統得以持續優化推薦精準度，進而提升轉換率

User Behavior/ Fine-tuning

## 流程資產化



讓AI從輔助走向自主執行



- ◆ 金融風控上已由傳統報表分析進一步升級為即時決策系統。透過整合交易紀錄、客戶行為與風險指標，AI不僅識別詐欺與信用風險，亦可動態調整風控規則並觸發決策流程

Agent / Workflow

# 小結：開源應用從能力堆疊走向任務落地與營運內化



# 企業導入之管理議題

從資料治理走向模型治理：開源模型落地的第一道風險防線

風險治理挑戰：資料



透過技術框架與自動化流程，將法規與安全要求轉化為可執行、可監測的機制（如：導入屏蔽技術，確保個資進入模型已去識別化）

確保數據的乾淨度與合規性，是開源模型產出高品質前提（如：避免RAG檢索到過時的产品手冊，導致AI提供錯誤的售後建議）

資料來源：MIC，2026年7月

◆ 企業需建構涵蓋資料前處理至模型生命週期管理的一體化治理機制，在有效控管AI風險的同時，確保模型穩定運作並支撐長期營運效率與可擴展性

# 法規與智財防護內嵌流程：從原則變成可執行機制



## 風險治理挑戰：法規



### 授權條款限制

- 部分模型（如 Llama）對大型商業用途有「月活躍用戶」上限限制
- 需注意「傳染性條款」，避免內部私有代碼因法規要求被迫公開



### 智慧財產權歸屬

- 多數開源模型沒有「免責聲明」，產出內容著作權責任由使用者承擔
- 需明確界定模型產出物的著作權歸屬



### 國際合規趨勢

- 歐盟 AI 法案：對高風險或大規模模型要求更高透明度與安全評估
- 各國對「數據來源」與「安全性」的審查標準不一



### 資安與數據治理

- 開源模型可能隱藏惡意代碼或安全漏洞
- 微調過程需確保數據隱私符合GDPR等國際規範

### 法律風險隱於技術細節中，忽視授權細則可能導致智財權糾紛

（如：使用 Llama 進行大規模商業營運前，未檢視其月活躍用戶限制，導致合規失效）

資料來源：MIC，2026年7月

- ◆ 企業需建立以數據溯源與透明度為核心的法規防護體系，確保AI從開發到部署全程符合動態國際規範；並透過彈性授權與資料治理策略，將合規要求轉化為信任機制，進而形成長期競爭優勢

## 因應策略



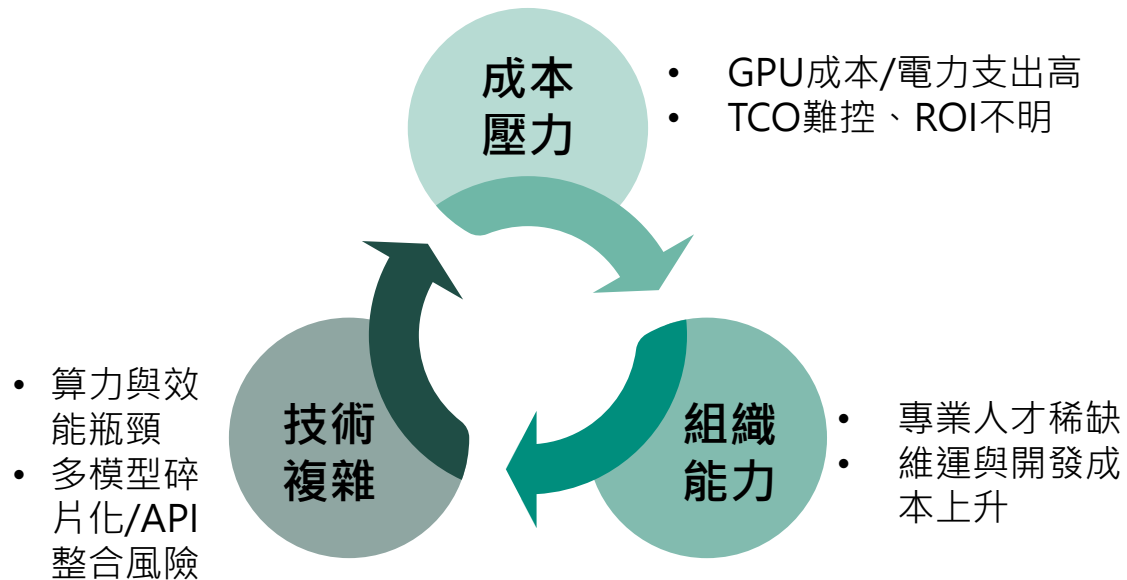
### 將法規檢查嵌入技術流程，落實由設計導入合規

（如：建立自動化授權掃描機制，在模型佈署前即完成安全性與智財權篩檢）

# 效能與資源管理：提升模型表現與營運穩定性



## ■ 成本控制挑戰



盲目佈署大規模參數模型，導致伺服器成本超出預算卻未達預期成效

資料來源：MIC，2026年7月

## 因應策略

- ① **落實資源使用效率化**  
採混合雲架構與彈性調度機制，動態分配算力資源以降低營運成本
- ② **建立ROI驅動路徑**  
優先於高邊際收益場景（如自動化行銷、減少稽核人力）落實AI應用
- ③ **知識資產化管理**  
將微調後的模型與專屬數據視為公司核心資產，提升企業長期估值與技術壁壘

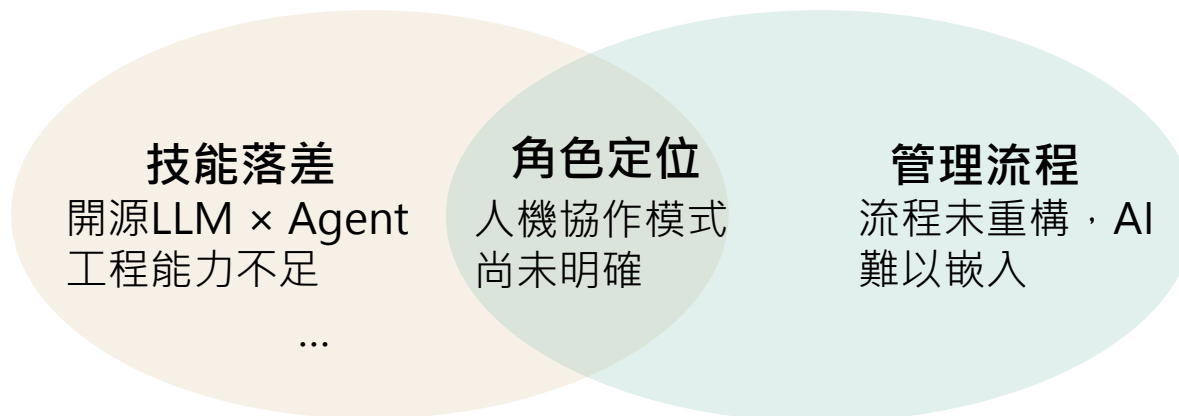
透過標準化AI營運架構與自動化管理流程，建立可監控、可優化、可持續改善的模型管理機制，兼顧成本效益、系統穩定性與長期營運能力

- ◆ 企業應建立動態成本監控機制，持續評估算力投入與業務價值之間的關聯性，避免過度依賴模型規模而產生不必要的資源消耗，並透過模型資產化管理，將開源技術逐步轉化為可複製企業能力，提升AI投資的長期效益與可持續性

# 從技能、角色到流程：讓AI Agent真正嵌入組織運作



## 自主執行挑戰



- AI Agent無法規模化落地
- 開源模型價值卡在「無法進入流程」
- 開源模型解決的是「模型能力」，但企業真正要解的是「系統與流程重構」

傳統組織分工與舊系統封閉，導致 AI Agent 只能觀望而無法介入核心業務（如：財務系統缺乏 API 接口，導致 AI 助手無法自動完成對帳與報表填報）

資料來源：MIC，2026年7月

- ◆ AI Agent的落地，不在模型能力，而在流程與資料的工程化能力。企業需重新設計人機分工，並透過流程API化與業務模組化，將AI嵌入營運流程中，逐步轉化為可持續運作的生產力

## 因應策略

### ① 建立Agent工程能力

導入開源LLM不只是模型使用，而需建立RAG、Tool use、Workflow orchestration等工程能力，使AI具備執行任務的能力

### ② 重構人機協作模式

重新定義人與AI的分工，讓AI負責標準化與高頻任務，人負責決策與例外處理，建立可控的人機協作流程

### ③ 流程API化與模組化

將企業流程轉為可被AI調用的API與模組，使Agent能真正嵌入業務流程，從「輔助工具」升級為「執行系統」

業務流程解構成可調用的模組，讓AI從單點工具進化為全自動的執行系統（如：建立「AI 工程標準組」，統籌各部門RAG 組件與 API 規範，實現技術共享）

# 小結：開源導入從模型選型走向企業級治理能力競爭

## 01 治理前移

治理不在最後，而在一開始

- 將治理由事後管控前移至資料、模型與流程設計階段
- 建立模型可控、可解釋、可追溯的管理能力，提升企業自主性與治理韌性

## 法規內嵌 02

合規不是限制，而是信任機制

- 將法規與智財要求內嵌於資料、模型與AI流程，建立設計即合規的治理機制
- 建立自動化檢核與持續監控能力，使合規成為企業信任與風險管理的核心能力

## 03 成本結構重塑

AI競爭，本質是成本效率競爭

- 從一次性建置成本，轉向全生命週期的算力、模型與營運成本管理
- 建立 FinOps / LLMOps 機制，持續優化資源配置，提升投入效益與投資報酬率 (ROI)

## 系統化營運 04

AI的價值，在於能否進入流程

- 將AI由單一工具升級為可執行任務的企業級系統能力
- 建立人機協作與持續優化機制，使AI穩定嵌入流程並形成長期競爭優勢



- ◆ 開源AI競爭已由模型選擇，轉向治理、合規、成本與系統化營運能力的競爭；真正的關鍵不在模型本身，而在企業是否具備將AI穩定落地、持續優化與規模化營運的能力

# 結論



# 結論

- ◆ 在產業趨勢方面，開源模型已由技術社群走向企業導入，成為建立可控AI能力的基礎；競爭焦點由「模型強弱」轉向「能否用得起、管得住、落得下」，從技術優勢走向營運能力
- ◆ 在技術演進方面，開源技術由單一模型走向多模型、長上下文、多模態與Agent架構；重心從參數規模轉向效率、可部署與可整合性，支撐流程化與場景化應用
- ◆ 在應用落地方面，開源模型由通用工具轉為可執行任務能力，深入垂直場景；價值從內容生成延伸至系統整合、流程自動化與Know-How資產化
- ◆ 在管理方面，企業挑戰由「能不能用」轉向「如何治理」；資料、法規、模型與組織缺一不可，關鍵在建立整體治理機制，使AI成為可持續累積的競爭力



# Thank you.

Your Gateway to Strategic Insights

朱師右 產業顧問兼組長

[alexchu@iii.org.tw](mailto:alexchu@iii.org.tw)

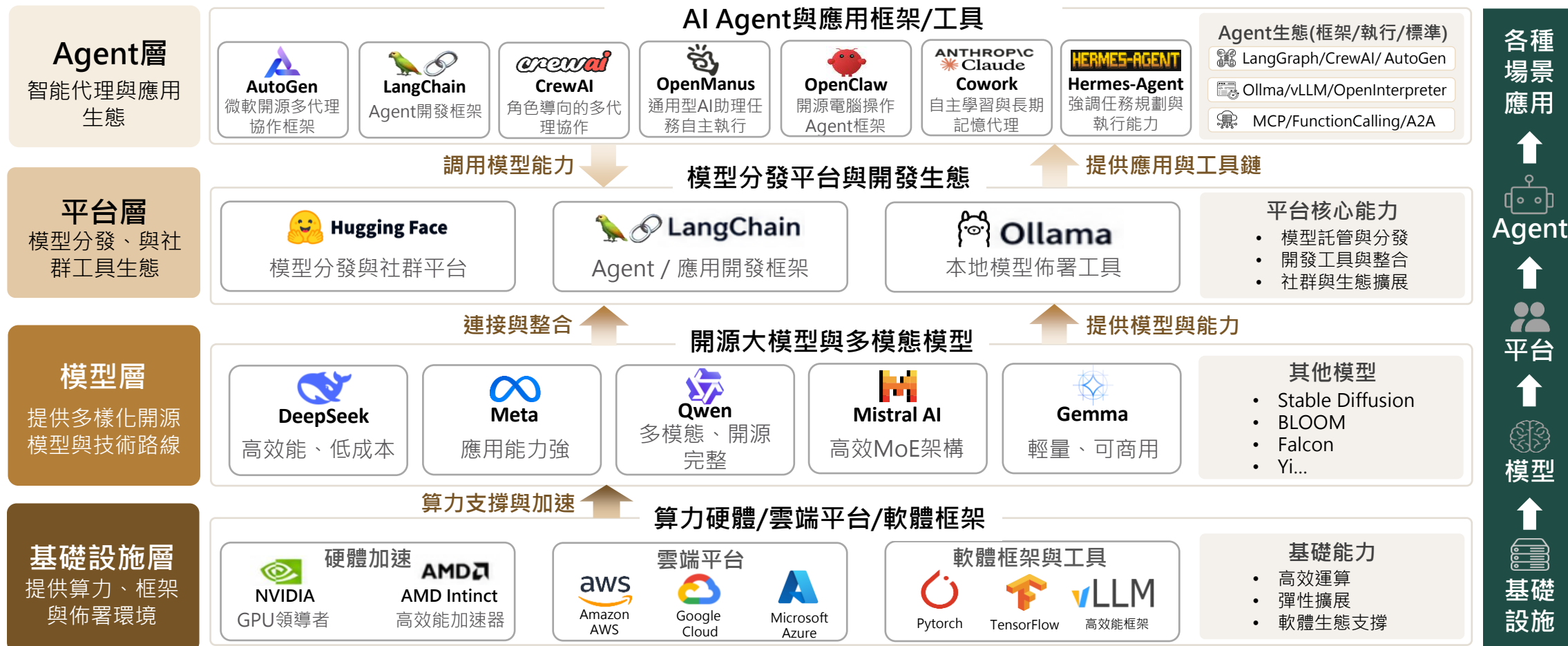
產業情報研究所

# 智慧財產權暨引用聲明

- ◆ 本活動所提供之講義內容或其他文件資料，均受著作權法之保護，非經資策會或其他相關權利人之事前書面同意，任何人不得以任何形式為重製、轉載、傳輸或其他任何商業用途之行為
- ◆ 本講義內容所引用之各公司名稱、商標與產品示意照片之所有權皆屬各公司所有
- ◆ 本講義全部或部分內容為資策會產業情報研究所整理及分析所得，由於產業變動快速，資策會並不保證本活動所使用之研究方法及研究成果於未來或其他狀況下仍具備正確性與完整性，請台端於引用時，務必注意發布日期、立論之假設及當時情境

# 附件

# 開源競爭由模型延伸至系統與生態



資料來源：各公司資料，MIC整理，2026年7月

- ◆ 開源競爭由模型能力延伸至Agent、平台與工具鏈，價值重心從單點技術轉向整體系統整合能力
- ◆ 企業競爭關鍵不再是選擇模型，而是能否整合模型、工具與基礎設施，形成可落地的AI應用體系