

專業任務數位分身：AI Agent 各層級挑戰與考量

AI 專業數位分身時代來臨（3）

AI 代理的出現對人才培育產生了深遠的影響。它不僅改變了員工需要掌握的技能，也改變了企業培訓員工的方式。為了在 AI 驅動的世界中取得成功，企業需要積極擁抱 AI 代理，並投資於人才培育，以確保員工能夠適應未來的挑戰。

文／陳添福教授（國立陽明交通大學人工智慧檢測中心主任）

AI 代理人（AI Agent）無疑是 AI 時代最令人期待的應用之一，作為能執行專業任務的智慧代理，正逐步改變我們對科技互動的方式。然而，要真正實現具備專業水準的 Agent，並非僅依賴語言模型的生成能力，而是需要克服一系列深層的技術挑戰。如何讓 Agent 在多樣化且動態的場景中始終保持穩定的表現？如何在效能與資源需求之間找到平衡，使其既能運用高效雲端運算，又能靈活部署於本地裝置？更重要的是，讓 Agent 在多模態場景中實現高效協同，並在應用環境不斷變化的情況下持續學習與優化！此外，在這一波生成式 AI 的大競局中，台灣如何藉由在半導體製造的優勢地位與靈活的 IC 設計能力趁勢而起，則是筆者更為關心的議題。

筆者以國立陽明交通大學人工智慧檢測中心主持「myLLM 產學服務聯盟」的經驗為基礎，透過對技術的觀察，問題的拆解，反思與實作，提供幾個觀點與發展方向供讀者參考。

全球的發展趨勢

根據 Gartner 於 2024 年 11 月 11 日的「人工智慧技術成熟度曲線」（Hype Cycle for Artificial Intelligence）的評估（圖 1），生成式 AI（Generative AI，GenAI）已進入啟蒙斜坡（Slope of Enlightenment）階段，相關應用將

在 2 至 5 年內達到成熟，所有職業平均自動化的程度在 2030 年時將可提升 30%，各產業也將因此受惠於科技進步，迎來生產效率的全面提升（McKinsey，2024）。

除 Gen AI 之外，同時這份報告也提示了 Composite AI、Edge AI、Synthetic AI 等極具發展潛力的研究主題，蠻值得我們發展台灣特有的主權 AI。所謂「composite AI」（複合式 AI）意旨融合多種 AI 技術組合（如機器學習、自然語言處理和知識圖譜等），以建立更強適應性與可擴充性的解決方案。台灣最近極力發展機器人產業，亟需此類技術。台灣作為全球硬體製造的領導者，有機會憑藉自身在晶片設計與硬體整合上的優勢，以 Edge AI 作為的發展提供強有力的發展基礎，無庸置疑。Synthetic AI 則是利用專業知識的智慧經驗，產生極有產業價值的訓練資料合成資料（synthetic data），使得 LLM 可以不僅能自然語言溝通，更能擁有專業產業經驗來推動流程自動化（automation）與生產優化（optimization），利用台灣過去獨到的生產製造 domain know-how，這點在發展具台灣特色的產業主權 AI，可產生獨特的絕對優勢。

台灣的機遇與挑戰

大型語言模型（LLM）已逐步應用於實際場

域，成為各產業推動數位轉型的重要引擎。各公司不僅打造自有企業大腦，專業型各特定 domain 專屬知識家也將風湧雲起。雖然生成式 AI 為台灣產業帶來了許多機遇與應用，但也帶來諸多挑戰，主要瓶頸在於產業上下串聯與雲地整合能力的不足。此文將從專業應用、LLM API、基礎模型與地端硬體四系統層次（圖 2），分析專業型 LLM 從應用開發到邊端智慧（Edge AI）落地實現，主要技術考量及挑戰。

|| 1. 專業應用層：聚焦特定領域需求

專業型 LLM 的核心價值在於透過專用型小模型（Specialized Small LM）解決特定領域的問題，而非通用模型（General-purpose LLMs）的廣泛應用，這要求系統在應用層面具備高度的專業化與垂直整合能力。

在生成式 AI 的應用中，Agent 系統也在這一過程中扮演重要角色，其架構的核心優勢在於其模組化與專業化，能將複雜任務拆解為多個子任務。Manager Agent（管理代理），負責協調和分配多個專業代理（Specialist Agents），如分析代理（Analyst Agent）、檢查代理（Checker Agent）和規劃代理（Planner Agent），透過分工合作完成資料分析、任務驗證和流程規劃等具體工作。同時，透過與外部系統的動態交互，Agent 系統能夠實現資料驅動的精準決策，適應多樣化的應

用場景，如企業自動化、醫療診斷和工程設計等領域。這一模式不僅提升了任務的執行效率，還為生成式 AI 在現實場景中的落地提供了強有力的支持。

對於台灣而言，未來的方向應聚焦於自動化（Automation）與優化（Optimization）的實現。自動化的核心在於提供多樣化的工具，而大型語言模型（LLM）則可以成為驅動這些設計自動化（Design Atomization）的關鍵力量。例如：AnalogGPT 在 IC 設計優化（IC Design Optimization）中的應用，展示了如何利用大型語言模型（LLM）作為訓練有素的規畫師，透過分析任務、使用可用工具和規則，生成符合規範的設計結果（GDS 文件），並保證其透過 LVS/DRC 驗證，達到規格要求。

|| 2. LLM API：模組化與標準化的關鍵

API 層是專業型 LLM 系統的中樞，負責將基礎模型的能力以模組化的方式提供給應用層。未來的語言模型生態系統將以 API 的整合與標準化為核心，實現雲端與地端的無縫協作。然而，隨著應用場景對長上下文（long context）支持需求的日益增長，模型在準確性與處理延遲方面的瓶頸愈發明顯，特別是注意力機制的資源消耗會隨上下文長度呈指數級增長，對硬體資源提出了極高的要求。因此，此領域的研究重點之一便是

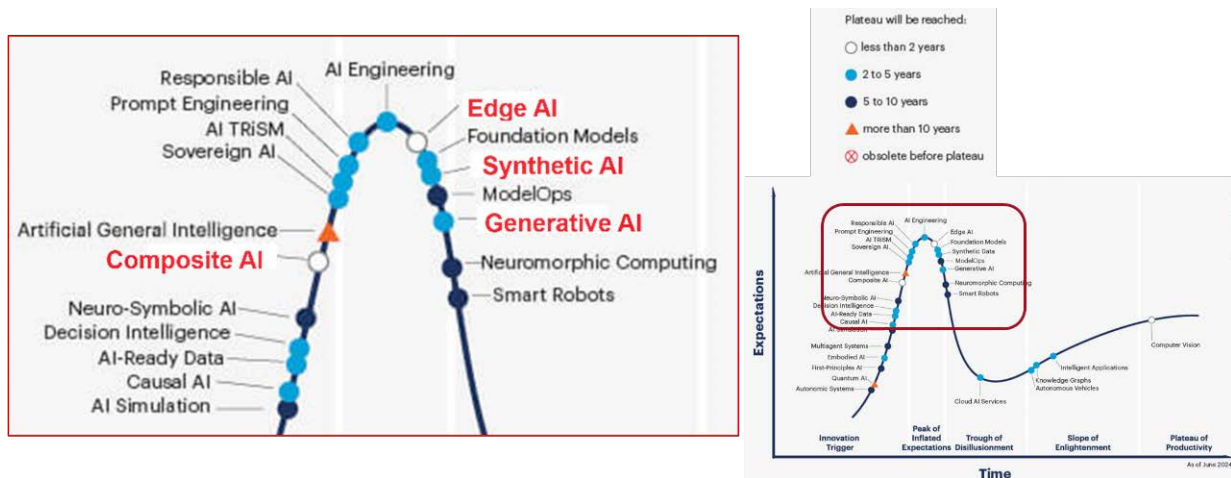


圖 1 Explore Beyond GenAI on the 2024 Hype Cycle for Artificial Intelligence。 (Gartner, 2024/11/11)

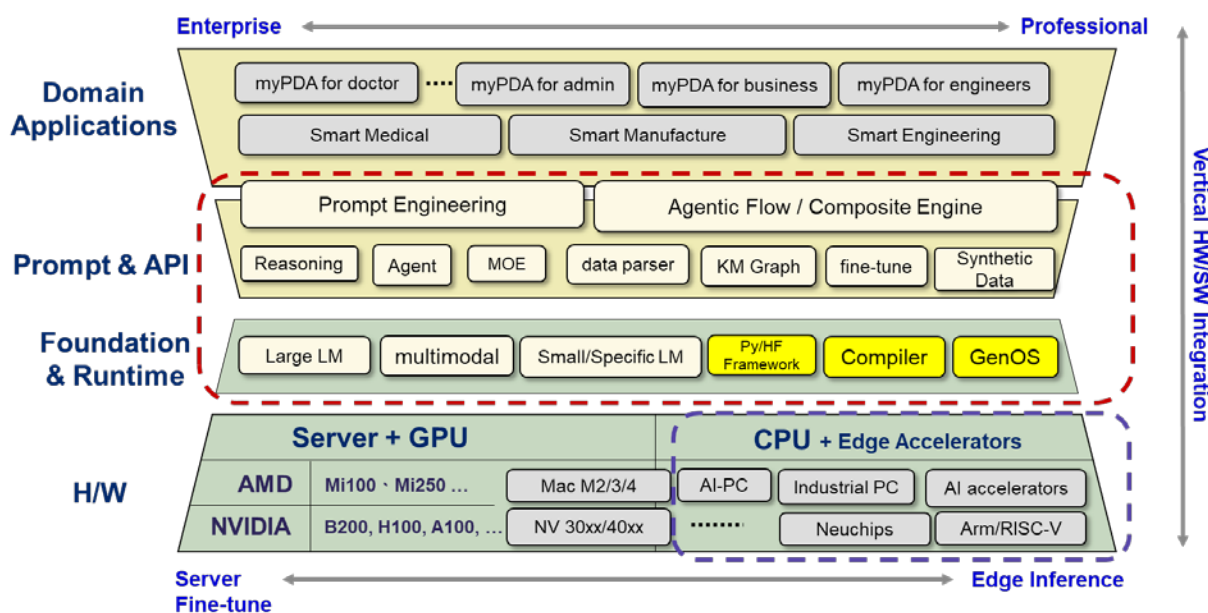


圖 2 我們 GenAI 的分析角度：多元專業型應用發展與優化落地。

如何有效地優化鍵值快取 (Key-Value Cache, KV Cache) 的使用。在硬體資源有限的情況下，需要探索如何將先前計算的結果與當前的處理數據有效地儲存到 KV Cache 中，並透過硬體設計與運算架構的優化，最大化重複利用這些快取內容。這種方法不僅能顯著降低延遲時間，還能大幅提升系統的整體效能，為長上下文處理提供更強大的支持。

3.基礎模型層：專業化模型的發展方向

隨著生成式 AI 技術的迅速發展，基礎模型層正在從通用型的大型模型逐漸轉向針對性的小型模型。這一轉變不僅有效降低了模型的訓練與運行成本，還能更加精準地滿足特定領域的需求，為 AI 的專業化應用開闢新的可能性。然而，全球高品質資料正逐漸枯竭，使得 AI 的訓練面臨前所未有的挑戰。

根據 MIT 的報告，目前大部分已知的高品質人類語言資料已被整合至現有的大型語言模型 (LLMs) 中。他們預測，到 2026 年這些資料將被耗盡，而從 2030 年至 2050 年間，相較低品質語言資料的庫存也將告罄。此外，視覺資料的

存量預計會在 2030 年至 2060 年之間枯竭。未來，依賴既有資料支持模型發展的可能性將大幅下降。面對這一挑戰，AI 領域的下一個突破機會將集中於開發專業化模型以及創新的資料處理與生成方式，以克服數據匱乏帶來的限制。

資料機敏性與模型落地化

隨著資料機敏性問題日益受到重視，許多企業選擇將模型部署於邊緣設備 (Edge)，以確保資料的安全與隱私。依靠台灣強大的硬體製造能力，各設備可運行經微調的專屬模型，滿足多元需求，進一步提升企業差異化競爭力。

開源模型與封閉模型的差距縮小

CB Insights 報告指出，開源模型 (如 Llama) 在某些應用場景的表現已接近封閉模型 (如 GPT-4)。然而，開源模型仍面臨缺乏商業驗證、高運行成本及對技術資源的依賴等挑戰。企業需在靈活性與穩定性間權衡，以選擇最適合的方案。

Multimodal (多模態) 的發展趨勢

多模態 (Multimodal) 技術快速發展，讓 AI 同時處理文本、圖像及音訊成為可能。OpenAI、Google 等公司積極推動多模態模型，應用於醫療、自動駕駛等多元領域。隨著多模態技術的進

步，未來的 AI 將能處理更複雜問題，並為企業解鎖全新的商業機會。

台灣在特定領域擁有獨特的利基資料，例如文化、語言以及產業知識等，這些高專業性資訊可成為訓練專用 LLMs 的重要資源。同時，合成資料 (Synthetic Data) 也將成為突破數據瓶頸的關鍵策略。透過生成式 AI 技術，我們可以創造大量模擬真實場景的資料，不僅能針對特定任務進行客製化設計，還能有效降低對敏感真實資料的依賴，進一步提升隱私保護。未來的語言模型將更加專注於特定任務，涵蓋如醫療應用、工業製造等高度專業化的領域，將為模型的進一步發展開拓更廣闊的可能性。

|| 4.硬體層：專用硬體的設計與整合

硬體層是支持專業型 LLM 系統運行的重要基礎，而臺灣在硬體領域的發展，尤其是在台積電的支持下，已經具備了全球領先的優勢。國際間的幾大科技巨頭，包括 Microsoft、Facebook 和 Google 等，紛紛選擇將部分業務帶到臺灣，進行自有 IC 設計。這些企業看重的不僅是臺灣在半導體製造上的卓越實力，更是我們在軟硬整合與 IC 設計的可擴展性 (Scalability) 上展現出的高度競爭力。

目前，臺灣的多家 IC 設計公司，特別是在可擴展推理晶片 (Scalable Inference Chip) 設計領域，已經展現出令人矚目的潛力。這些晶片未來將能支援多模態應用，並提供更有效率的執行能力，為全球生成式 AI 的發展提供重要支撐。然而，臺灣目前在軟體堆疊 (Soft Stack) 系統能力和 AI 專業人才供應方面仍面臨挑戰。唯有推動軟硬體能力的同步提升，臺灣才能真正抓住這波生成式 AI 與 LLM 推理應用的全球商機。

生成式 AI 的核心價值在於應用

紅杉資本在最新報告《Generative AI's Act Two》中指出：「生成式 AI 的最大挑戰在於證明其價值，而非尋找使用案例或需求。」

呼應紅杉資本提出的觀念，人工智慧檢測中

心規劃一個 myLLM Studio 雲地整合的產學服務平台，提供從應用到 API、模型再到硬體的垂直整合解決方案。我們專注於針對不同應用領域進行專業化鏈結，構建模組化架構，並找到最適合的硬體來實現應用需求。隨著 AI PC 將逐漸普及，我們預測 PDA「專業數位分身」(Professional Digital Agent) 即將發生，LLM 雲地整合應用將遍地開花。我們以 myLLM Studio 為基礎，提出了「myPDA (專業數位分身)」的概念，重新定義了個人化數位分身的角色。透過 myPDA，專業人士可以將自身知識 AI 化，並實現雲地整合。例如，在醫療領域，我們與眼科醫師合作，開發了具備專業知識的 Agent 系統，能依據醫師所定義的 SOP 與病患對話，減輕醫生的問診負擔；在工程領域，我們針對大型規範文件的解析需求，結合 RAG 技術，提升資料查詢與報表生成效率。

我們相信，未來的應用將以專業任務分身化為核心，從特定應用需求出發，結合生成式 AI 與硬體技術。我們希望能以上下垂直整合角度，探索 LLM 技術流程與軟硬體設計，促成生成式 AI 與硬體整合的最佳實踐，推進台灣生成式 AI 的發展，創造更多產業價值。

未來展望

生成式 AI 的挑戰在於證明其價值，而非僅尋求應用案例。唯有創造持續效益與個性化體驗，才能使生成式 AI 成為日常生活與工作的核心。未來的生成式 AI 發展將更加注重與人類需求的深度結合，例如實現更程度的情感識別與互動能力，推動人機協作的新模式。全球市場中的技術競爭日益激烈，而台灣透過綜合其硬體優勢與應用創新能力，有望在這場變革中佔據領先地位。