myLLM Studio：AI Agent 雲地整合系統核心模組

AI 專業數位分身時代來臨（4）

「myLLM Studio」與「myPDA」等創新解決方案如何透過檢索增強生成（RAG）技術、智慧代理（LLM Agent），以及 AI 作業系統（AIOS）的多層次架構，實現生成式 AI 與產業需求的無縫接軌。

文／林聖博（陽明交通大學人工智慧檢測中心架構師）

生成式 AI（Generative AI）的快速發展為產業數位轉型帶來了前所未有的機遇。從 OpenAI 的 GPT 系列到各大雲端服務供應商推出的生成式 AI 工具，這些技術正逐漸成為各行各業探索創新應用的核心動力。然而，當企業試圖將這些技術應用於實際場景時，往往會面臨諸多挑戰。

GenAI 推動產業數位轉型的挑戰與機會

首先，許多現成的生成式 AI 解決方案並未針對特定產業需求量身打造，導致應用的適配性不足。其次，生成式 AI 的運行依賴大量運算資源，對企業現有的基礎設施形成了巨大壓力。在處理機敏資料時，如何確保資料安全與隱私合規成為企業的一大顧慮。最後，不同產業的應用場景各異，如何讓 AI 技術在多樣化的環境中保持穩定運作，亦是企業面臨的一大難題。這些挑戰使產業界逐漸意識到「僅依賴現有的雲端 AI 平台不足以滿足多變且複雜的實際需求」。為此，打造能與雲端 AI 平台協作的地端自有的、專屬的 AI 系統，針對具體場景進行深度定制，提升生產力成為營運的一部分，為企業創造價值，才是促進生成式 AI 真正落地應用的關鍵。

大型語言模型的落地需求

隨著生成式人工智慧（Generative AI）的蓬勃發展，大型語言模型（Large Language Models，LLMs）逐漸成為企業數位轉型的重要工具。然而，要使 LLMs 在企業環境中真正納入並發揮效益，仍面臨多重挑戰與需求。

‖ 企業資料機敏與安全性

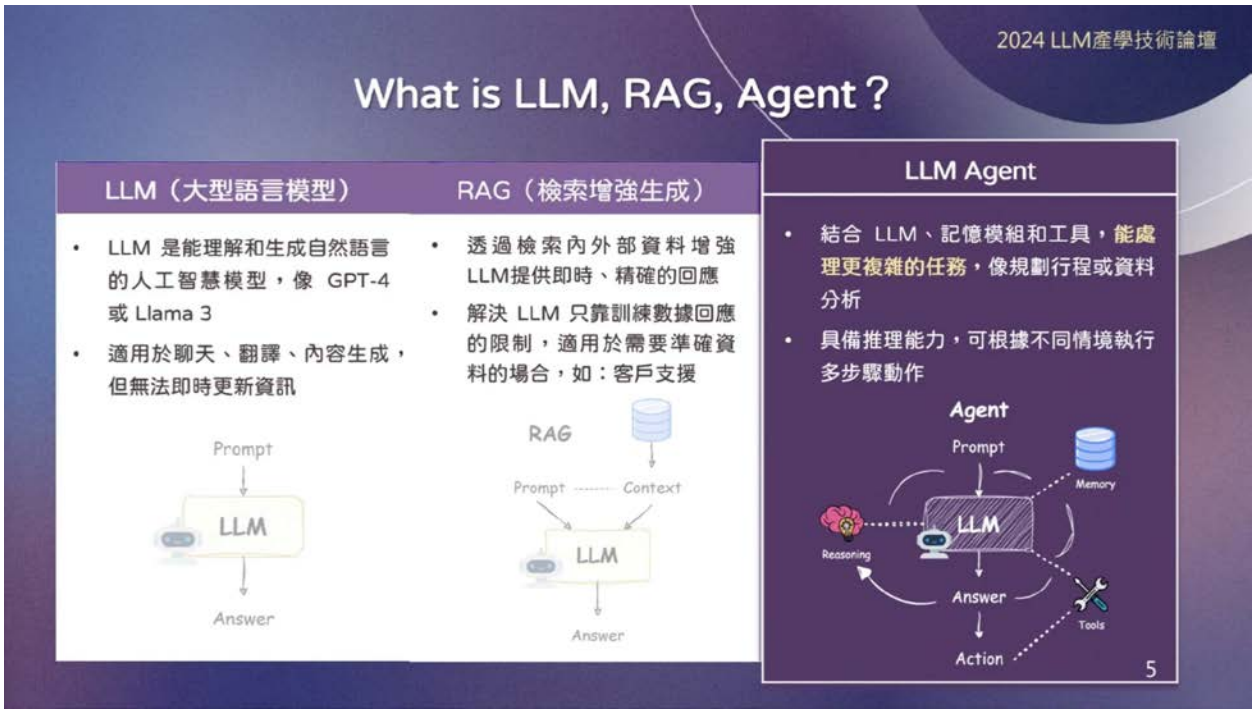
企業數據的專屬性與機密性至關重要，雲端大型 AI 服務確實具有不俗表現及效果，但過程涉及機敏資料上雲的不確定與危險性，可能導致企業內部數據外洩以及盜取。LLM 結合落地化部署，能利用大型語言模型的智慧與便利性，輔助企業內部資料整合及應用，在確保數據不外流的情況下，大幅增加企業和個人工作效率及產能。

‖ 效能與成本權衡

雖然 LLM 具備卓越的語言理解與生成能力，但其高運算需求與成本往往令企業卻步。針對此問題，企業需探索邊緣端的模型優化、硬體加速等技術，提升運算效率並有效降低部署門檻。

‖ 應用場景的客制化

通用型 LLM 雖然功能強大，但往往無法直接滿足特定產業需求。因此，透過模型微調、領域知識嵌入以及智慧代理應用，使其更符合企業的業務邏輯與場景需求，是 LLM 落地的關鍵一步。



myLLM 產學服務聯盟的整合式平台

為應對上述挑戰，「myLLM 產學服務聯盟」推出了整合式平台 myLLM Studio，希望「打破 LLM 容易產生幻覺的侷限性，藉由 AIOS 協調雲地與系統資源，並以 myPDA 做為先導服務」，提供量身定制的解決方案，確保技術能夠真正解決問題。透過整合式平台與核心技能模組，協助企業克服資源限制與基礎設施不足的挑戰，實現生成式 AI 的落地應用。

打破 LLM 的局限性：RAG 與 LLM Agent 的應用

生成式 AI 技術的初始發展以大型語言模型（LLM）為核心，如 GPT-4 或 LLaMA 3。這些模型具備理解與生成自然語言的能力，能夠像人類一樣「思考」並給出答案，應用於聊天、翻譯、內容生成等場景。然而，隨著使用時間的推移，我們逐漸發現 LLM 存在一些局限性，例如：回答精準度與一致性的問題。此外，模型的知識來源主要來自訓練資料，若要更新模型中的資訊，往往需要進行重新微調（Fine-tuning），這不僅耗時費力，且成本高昂。

檢索增強生成（RAG）的突破

為了解決上述問題，檢索增強生成（Retrieval-Augmented Generation，RAG）技術應運而生。RAG 透過整合內部資料庫與外部網路資訊，為語言模型提供即時且精準的回答，並提高模型的專業性與適應性。RAG 不僅彌補了 LLM 的缺陷，還使模型能夠更加穩定地滿足特定場景需求，例如醫療診斷與客服應用等。儘管 RAG 提升了模型的回答能力，面對高複雜度的任務時，它仍顯得力不從心，這促使人們探索更進階的技術解決方案。

LLM Agent 的崛起

為應對複雜場景，LLM Agent 概念逐漸興起。作為大型語言模型的進階版本，LLM Agent 結合了記憶模組與外部工具，實現了多項創新功能：

1. 處理複雜任務：LLM Agent 能夠執行如行程規劃、資料分析等高複雜度的任務，展現出超越傳統只依靠 LLM 生成的能力。
2. 多步驟推理與自動化操作：根據特定情境進行邏輯推理與多步驟流程執行的能力，實現更自動化的操作流程。

3. 系統化行為：結合記憶模組與外部工具，LLM Agent 模擬出更加系統化的行為，使操作更穩定，邏輯處理思維更接近人類行為。

LLM Agent 的出現，為生成式 AI 在智慧化與自動化領域邁出了重要一步，並大幅拓展了其應用場景。然而，單一 Agent 系統在應對多步驟推理或高複雜度需求時，仍然存在一定的局限性。

‖ AI Operating System（AIOS）：協同未來

為克服這些限制，AI Operating System（AIOS）的概念應運而生。透過整合硬體資源與多個LLM 代理系統，採用分層次管理的方式，將應用層、核心模組與硬體資源明確分離，並以模組化設計提升運作效率。這種架構更靈活地應對各種場景需求如企業場域，透過建置企業內部落地化專屬 AIOS 系統，基於多重高階 LLM 智慧代理架構在資料不上雲、機敏資料保護前提下實現高效的協同運作與資源分配。

‖ myLLM Studio：AIOS 與模組化平台的整合應用

myLLM Studio 是一個結合 AIOS、多代理系統（Agents），以及雲地整合的模組化設計平台，可建立協同運作模式，對於高複雜度或需要大量運算資源的任務，透過雲端平台完成；而針對較為敏感的資料，則採用地端運行，以保障資料安全、減少延遲並提升執行效率。

在設計上，myLLM Studio 採用模組化與技能導向的開發模式，系統被劃分為多個獨立的元件模組，便於快速部署與針對性的配置，並根據專業人士或特定場域的需求，設計以技能為核心的解決方案。

在控制流程上，myLLM Studio 設計了一個 Agentic SOP-Driven Engine（代理 SOP 驅動引擎），能夠根據任務需求，選擇動態代理規劃（Dynamic Agent Planning）或靜態代理規劃（Static Agent Planning）的方式來執行任務。動態代理規劃是一種複雜但強大的任務處理方式，當系統接收到一個任務查詢（Task Query）後，會由規劃器（Planner）負責將這個查詢拆解為多個子任務，包括，嘗試性操作任務（Try and Error）、探索性任務，以及執行型任務。任務被拆解後，這些子任務會交由執行器（Executor）逐一執行，並將規劃結果落實為具體的行動，完成整個任務流程。另一方面，靜態代理規劃是一種

基於 SOP 的任務處理方式。對於已知且穩定的任務或知識，人類專家可以將操作流程清晰地描述為 SOP，供系統直接參考執行，無需額外的動態規劃過程。這種設計有效結合了人類的專業知識與規則，讓系統能夠快速生成準確的智慧代理程式，滿足使用者多元化的應用需求。

myPDA：專業人士的數位分身

myPDA 是一款專為專業人士設計的數位分身系統，其核心理念是結合專業領域知識與生成式 AI 技術，為特定場景量身訂製專屬解決方案。這套系統的設計不僅涵蓋通用型的功能，還針對醫療、工程、商業等特定領域，提供專業化的技能模板，讓使用者能快速解決問題並提升效率。

資料處理：專業技能分身的關鍵

為了實現專業技能分身的能力，資料處理成為整個系統的關鍵挑戰。myPDA 必須處理多樣化的資料來源，包括非結構化資料（如 PDF、PPT）與結構化資料，並建立一套完整的資料處理流程（Data Pipeline）。這些資料的轉換與解析，需要依賴專門的處理工具，確保語言模型能正確理解資料的語境與意圖。資料的準確性直接影響語言模型的回答品質，避免產生冗餘資訊或幻覺，從而提升系統的可靠性。

myPDA 應用於衛教領域：數位衛教師的誕生

醫療領域對於精準度的要求極高，每一個診斷與建議都可能影響病患的健康。然而，醫生的時間有限，如何在不犧牲品質的情況下提升效率，一直是醫療界的挑戰。以眼科的水晶體推薦為例，myPDA 整合了衛教 SOP，並透過 LINE 等通訊工具與病患互動。當病患提出需求時，系統會根據 SOP 的標準步驟，逐步引導病患完成選擇，並提供精準的建議。這不僅減少了病患的等待時間，也大幅降低了醫生在重複性問診上的負擔。這種結合 LLM 的數位衛教師模式，為醫療與衛教領域帶來了全新的服務體驗。

myPDA 應用於工程領域：解決巨量資料的難題

對於工程師來說，處理大量的技術文件與規格書是日常工作的一部分。然而，面對數百頁的 PDF 文件，傳統的人工處理不僅耗時，還容易出錯。對此，myPDA 為企業提供了一套高效、智慧、落地的地端解決方案（myPDA Enterprise）。

1. 大型文件比對與智慧檢索

myPDA 支援 500 頁以上的 PDF 文件比對，並內建自動分類與視覺化技術，能快速整理文件資訊並生成表格。這不僅提升了比對效率，還能確保結果的準確性。此外，myPDA 的跨文件智慧檢索功能（RAG Pro）能處理超過 300 個文件，快速找到所需資訊，為工程師節省大量時間。

2. 競品料號分析與推薦

在電子元件的設計與製造中，工程師經常需要比較競品料號或尋找替代方案。myPDA 利用 LLM 技術，自動抓取競品資訊，並結合內部文件進行綜合分析，生成相近或替代料號。這種自動化的料號推薦系統，不僅提升了效率，也減少了人力成本。

3. 程式測資生成

myPDA 還能根據設計文件，自動生成測試平台，協助工程師快速構建測試環境，以進行模擬驗證。

生成式 AI 的快速發展為產業數位轉型帶來了豐富的可能性。然而，產業在推動 Gen AI 技術落地時，面臨適配性、基礎設施與資料安全等多重挑戰。「陽明交通大學人工智慧檢測中心」秉持「將無形的智慧轉化為有形的價值」這一理念，實作 myLLM Studio 與 myPDA 等解決方案，透過雲地協同運作、模組化設計以及專業化技能模組，成功地解決了上述問題，展現了生成式 AI 在實際應用中的巨大潛力，面對醫療、工程等富含專業知識性的領域，導入 AIOS 觀念的 LLM 智慧代理，提供數位轉型的落地解決方案。未來，隨著更多創新技術的推出，生成式 AI 將進一步深化產業數位轉型，助力各領域實現更高效、更智能的運作模式，創造更大價值。