

NVIDIA AI Foundry 與 NeMo

多模態生成式AI賦能企業數位轉型的新引擎

生成式 AI 在提升企業競爭力、生產力，以及改善客戶體驗方面，展現出強大威力，也成為多數企業的優先投資領域，預期全球在生成式 AI 應用和投資將快速成長，每年可為全球經濟帶來高達 4.4 兆美元產值。

文／林裕洋

由於生成式 AI 在提升企業競爭力、生產力，以及改善客戶體驗方面，確實展現出強大威力，也成為多數企業的優先投資領域，預期全球在生成式 AI 應用和投資將快速成長。根據麥肯錫公布《The economic potential of generative AI: The next productivity frontier》研究報告指出，生成式 AI 對生產力的影響極大，每年可為全球經濟帶來高達 4.4 兆美元產值，範圍涵蓋知識庫領航、內內容生成、語言翻譯、報告與資料分析、虛擬助理、程式碼生成等等。

在 Gartner 公布《Innovation Insight: Generative AI Chatbot to Improve CX and Agent Productivity》報告中，預估 2026 年 50% 客服和支援服務將由生成式 AI 虛擬助手負責，以便降低人力負擔、提升客服服務品質。事實上，目前此技術已在眾多產業之中展現效益，如美國知名醫療零售業者的呼叫中心，每月需負責處理 2,700 萬通電話，對人力需求極大，所以預計打造自動化呼叫中心。該公司將透過生成式 AI 匯總語音、郵件等資訊，讓呼叫中心可以為用戶提供更快、更準確的答案，預計3年內可回收 5000 萬美元的投資。此外，在電信和金融等產業中，生成式 AI 也幫助企業實現資料保護、即時回應客戶需求，進而優化客服、資料分析等業務功能。

「過去幾年採用 AI 比例並沒有太多變化，

大多在 50% 上下，關鍵在於其帶來效益不夠明確，讓多數企業處於評估階段。」NVIDIA 資深解決方案架構協理康勝閔解釋：「直到 2022 年底 ChatGPT 問世之後，企業感受到生成式 AI 技術的威力之後，讓推動生成式 AI 專案比例快速提升，從 2023 年 33% 提升到 2024 年 65%。」

現今生成式 AI 應用範疇非常廣泛，涵蓋金融、醫療、零售等眾多產業，金融業運用 AI 可進行詐欺偵測和個人化銀行服務，醫療領域則能用於藥物發現和臨床試驗數據分析。在零售業部分，則可透過自動化商品描述、價格優化等功能來提升銷售效益，而製造業和能源行業則利用 AI 進行工廠模擬、預測性維護等，以提升營運效率和資源利用率。

視覺 AI 進化 自動串連上、下文

現今生成式 AI 的技術進一步發展為多模態應用，包括文件、影片、圖片、3D 模型等，能夠解決不同領域的需求並提升模型的泛化能力。如在改變傳統視覺 AI 與電腦視覺方面，傳統視覺 AI 系統依賴大量標記資料的做法，是種基於固定規則的應用模式，如常見的火災偵測、行為偵測等均是如此。以傳統視覺 AI 系統發展的應用系統，存在功能狹窄、有限，且有開發週期較慢、模型連結複雜等問題，也讓智慧城市、智慧工廠等發

展受到侷限。

現今藉由生成式 AI 技術協助，視覺 AI 能藉由豐富上、下文協助，免去傳統大量標注資料情況下，完成複雜場景理解。特別藉由自然語言與 AI 模型之間的交互搭配，具備能進行深入搜尋、整合影片資訊等，只需最少調整即可縮短開發週期，輕鬆發展出推理微服務、提高系統靈活性。

康勝閔指出，當視覺語言模型（Visual Language Models, VLM）在生成式 AI 技術協助下，虛擬助手可輕鬆執行互動問答、自動警報和即時監控等多項任務。舉例來說，在新世代監控系統中，AI 助手可即時主動提醒管理者，場所有過於擁擠的現象，或存在可能缺少安全裝備等問題。不光如此，AI 助手還可根據使用者需求進行場景分析，並且主動生成事故報告和調查報告，協助調查團隊迅速尋找影片中的特定事件或內容，縮短處理事故的時間與減少人力支出。

NVIDIA AI Foundry 平臺打造企業 AI

市面上生成式 AI 開源大語言基礎模型愈來愈多，參數量也持續攀升，如 Meta LLaMA 1 有 7B、13B、30B 和 65B 等參數模型、LLaMA 2 則是優化到 70B。Mistral 則有 7B、12.9B 等兩種，是由 BigScience 社群的 Bloom，則提供 560M、1.1B、3B、7.1B 和 176B 等參數模型。

因應市場趨勢，NVIDIA 在持續研發高效能 GPU 晶片之外，也推出 NVIDIA AI Foundry 平臺，讓企業可打造專屬的生成式 AI 方案。NVIDIA AI Foundry 在提供 Nemotron、Edify 等模型外，也提供客製化模型用的 NVIDIA NeMo 軟體，以及 NVIDIA DGX Cloud。此外，該平臺還有可隨處部署的推論微服務的 NVIDIA NIM（Neural Inference Microservice）服務，內容涵蓋客製化模型、最佳化引擎以及標準 API。

康勝閔指出，NVIDIA NeMo 是一款端對端平臺，適用於開發大型語言模型、多型態、視覺、語音等自訂生成式 AI，讓企業能在不受地域限制下，想咬精準的資料管理方式、頂尖的自訂工具、檢索增強生成（RAG），及加速效能。

NVIDIA NeMo framework 支持企業構建和部署生成式 AI 模型，包含資料準備、模型訓練和評估等功能，有助於企業快速實現 AI 應用的上線。NeMo Serach 檢索器為檢索應用提供高準確率，支持模組化部署和快速推理，適用於各種資料源的檢索需求，如文件、圖像、結構化資料等。

NVIDIA 助攻，Taiwan 70B 問世

考量到 AI 主權非常重要，NVIDIA 運用 Taipei-1 超級中心的強大算力，透過與臺灣大學合作方式，以 LLaMA 3 模型搭配 NVIDIA NeMo 框架，運用來自於媒體、法律、化工、醫療、製造業製程等資料進行訓練，為臺灣打造 Llama 3 - Taiwan 70B 模型。目前已在公開網站提供下載，供各產業進行利用與推論。

NVIDIA NIM 則是支援企業在多領域 AI 推理應用的微服務架構，使用容器化的 AI 推理堆疊，使得開發者可以靈活地在各種運行環境中部署 AI 應用，如資料中心、邊緣設備和雲端等。此產品還可確保推理速度和處理性能，支援醫療影像處理、金融數據分析、視覺處理等多種領域應用。

康勝閔表示，NVIDIA NIM 有靈活部署、高效推理性能、模組化架構、降低開發複雜性等四大特色。在靈活部署部分，受惠於採用容器技術，所以可支援跨雲端、資料中心和邊緣的部署選項。其次，在高效推理性能方面，透過 NIM 優化技術能高效執行推理任務，並在處理大量資料時保持穩定性能。第三點在模組化架構部分，NIM 允許用戶根據需求選擇特定的 AI 模組，實現任務的專業化處理，例如影像分類、語音識別等功能。最後則是降低開發複雜性：NIM 提供多種 AI 服務的 API 和開發工具，幫助企業快速開發並部署 AI 應用，縮短上市時間。

隨著生成式 AI 在多模態生成、自動化學習、跨領域應用等不斷突破，已成為企業靈活應對市場變化，發展創新性業務模式的重要助手。在全球吹起數位轉型浪潮下，生成式 AI 將成為企業保持競爭力的重要工具，更是推動整體產業走向更智慧、數位化的強力推手。