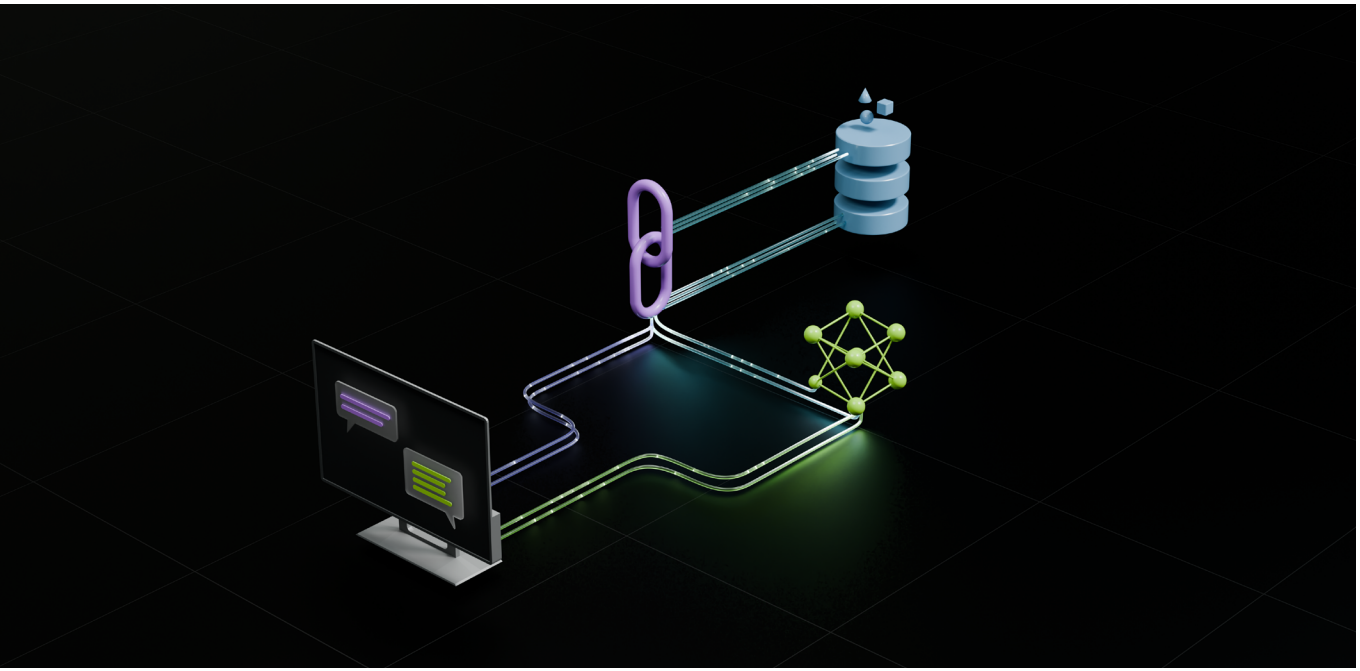
電子書

企業指南：建立結合檢索增強生成 (RAG) 的智慧 AI 助理

從 AI 聊天機器人到 AI 助理的演變



檢索增強生成 (RAG) 已徹底改變 AI 助理，象徵從傳統聊天機器人到企業級智慧助理的重大轉變。RAG 整合了脈絡學習，並使用專有資料進行微調，強化了大型語言模型 (LLM)，以提供最新的、以脈絡為根據的回應。使 AI 助理變得更準確、更可靠及更具有參考價值。

以 RAG 驅動的 AI 助理，讓機構可以從資料中擷取更深入的洞見，有效率地處理資源密集型任務，例如：

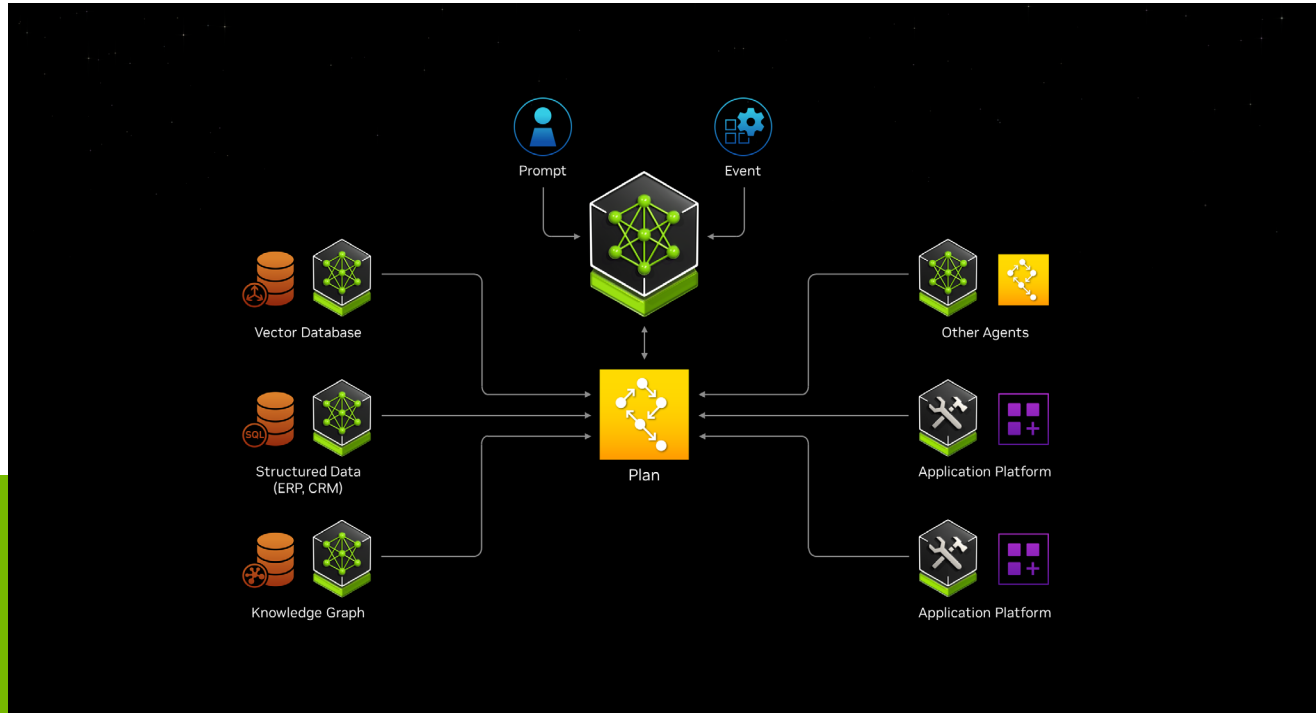
- 摘要
- 準確的資訊擷取
- 多語翻譯
- 情感分析
- 個人化推薦
- 教育訓練
- 客戶支援

若想要進一步提升全球可及性，則可以整合語音和翻譯 AI，即能以使用者的自然語言進行溝通。

使用 RAG 和向量搜尋強化 AI 助理

AI 助理使用 RAG 和向量搜尋，即時擷取精準、具脈絡感知的資訊。此過程涉及將包括文字、圖片、聲音、影片和圖表等的資料分成較小的區塊，並轉換成可捉語意和細微差異的向量嵌入。這些嵌入被編入索引及儲存在向量資料庫中，以便 AI 助理搜尋和擷取最相關的資料。

當使用者提交查詢時，系統也會將其轉換為向量嵌入，並執行相似性搜尋，以擷取最相關的資料區塊。之後，將此類資料區塊納入 AI 助理的脈絡中，確保以事實為根據、精準，並符合企業特定知識的回應。



檢索增強生成 (RAG) 流程是透過快速擷取和產生相關資料強化 LLM，以確保高度準確、具脈絡感知的回應。

透過向量搜尋和最佳化推論加快 AI 效能

企業可以利用結合 NVIDIA cuVS 的 GPU 加速向量資料庫，大幅加快向量搜尋。使可擴充、高效能 AI 助理可以適用於即時生產應用。

最佳化推論效能，可以將擷取的資料轉換成快速、準確、連貫的回應。RAG 提升了 LLM 推論的準確性和相關性，方式是根據經過清理和向量化的資料，例如新文件或專有文件，的向量搜尋查詢。此過程為模型提供了更精準的輸入、精進了 LLM 參數指引，並最佳化提供給使用者的輸出。

向量搜尋是透過將使用者查詢轉換為向量嵌入，以識別相關文件，推論則是綜合這些資訊，產生符合脈絡的輸出，以整合事實和解決矛盾。高效率擷取與精密推論結合，讓 RAG 系統可以存取最新知識及有效應對各種使用者查詢。

使用 AI 助理在各產業中推動業務影響和提升效率

此類由 RAG 改變的 AI 助理，正在提高各產業的員工生產力、客戶滿意度和作業效率。

在金融服務業，銀行透過個人化理財規劃、投資建議以及可以回答比傳統聊天機器人更多種問題的 AI 助理深化客戶服務體驗。

在醫療業，AI 助理可以提升高品質照護的可及性，同時減輕行政負擔。例如希波克拉底誓詞。
AI 的生成式 AI 醫療代理程式可以簡化預約排程，並協助護理師進行術前訪視和出院後追蹤，讓醫護人員能專注於照護患者。

在電信業，NVIDIA 及其合作夥伴生態系統正在推廣應用生成式 AI，提供涵蓋客戶服務、最佳化網路和無線電存取網路 (RAN) 營運的解決方案。

客製 LLM，打造更強大的 AI 助理

結合 RAG，客製 LLM 為產業創造涉及微調、提示工程，以及人類回饋強化學習 (RLHF) 的專用 AI 助理。此類技術都可以提升模型效能，以提供更準確、具脈絡感知的回應。

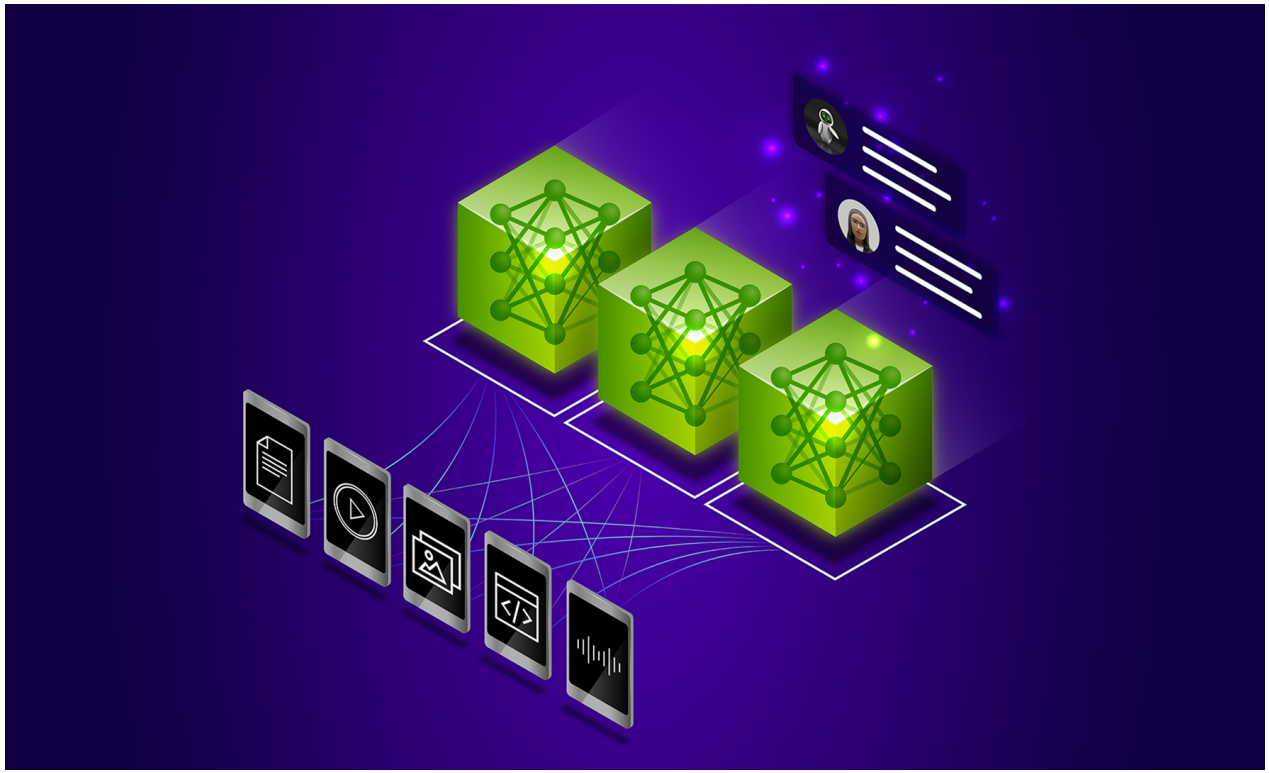
透過特定領域訓練範例，微調預先訓練 LLM (通常稱為基礎模型)、針對模型進行專業技能訓練，並使用 RLHF 進一步完善，使輸出符合人類偏好。

提示工程透過加入結構化指引，例如包含明確指示的系統提示，提升回應品質。

在整合 RAG 時，AI 助理可以動態擷取最相關的即時資料，在提示到達客製化 LLM 之前豐富其脈絡。確保 AI 助理提供精準、最新及針對企業的洞見，使其能提供更有效的客戶服務、知識探索、決策支援和自動化。

滿足 RAG 的運算需求

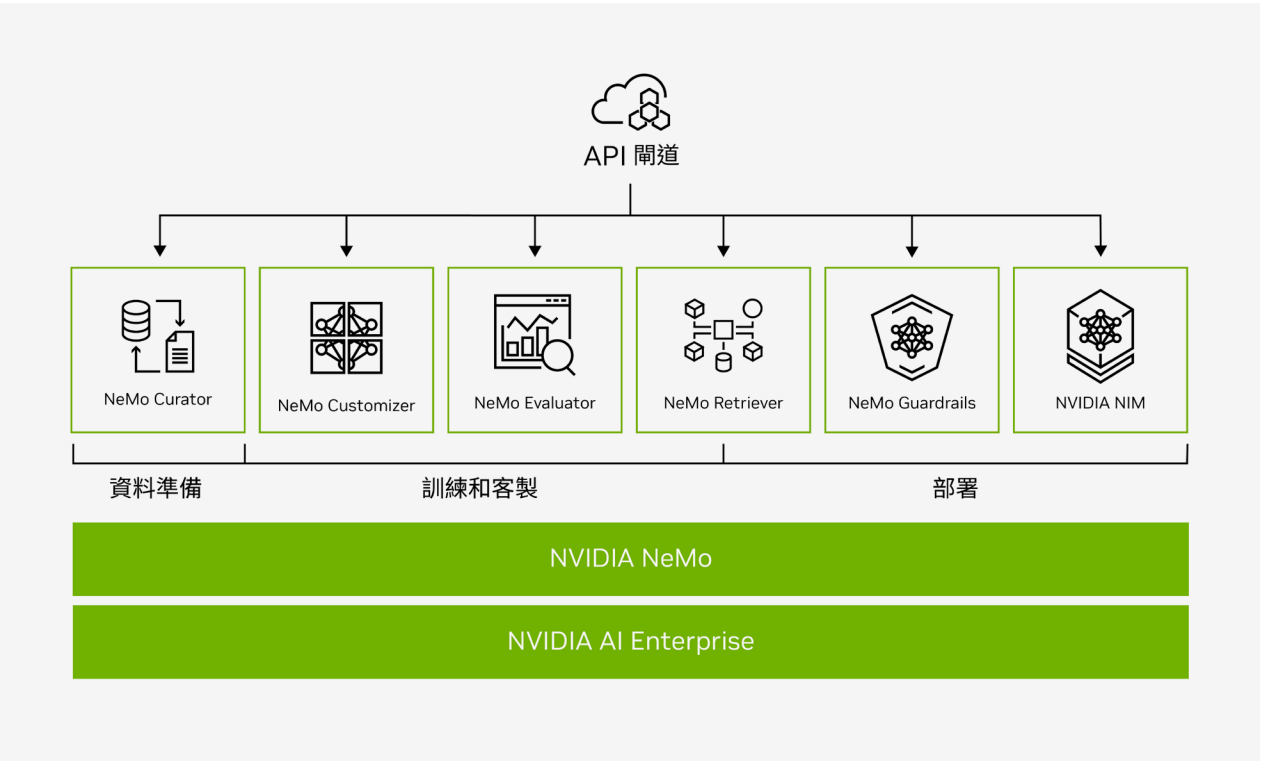
部署 RAG 流程需要處理大量資料，而大幅加重了 LLM 的運算負載。為了滿足此需求，持續最佳化的完整堆疊解決方案整合了先進晶片、系統與軟體，是各產業提升使用者體驗、提高效率和維持競爭優勢的關鍵。



利用 NVIDIA AI Enterprise 有效率地建立和擴充 AI 助理

NVIDIA AI Enterprise 使用強大的工具和技術，包括 NVIDIA Blueprints、NeMo 和 NIM 微服務，簡化 RAG AI 助理的開發和部署。

此雲端原生軟體平台提供了最佳化效能、企業級安全性和穩健的支援，確保從原型到生產無縫移轉，並具備企業順利擴充 AI 需要的穩定性。



NVIDIA AI Enterprise 提供了使用 NVIDIA NeMo™ 和 NVIDIA NIM™ 開發與部署客製社群模型及 NVIDIA 模型的平台，讓 AI 助理可以利用微調、RLHF、提示工程和 RAG 提供準確、針對特定領域的回應。

NVIDIA NeMo 平台包含多個可提升 AI 助理效能的元件：

- NVIDIA NeMo Curator 刪除重複項目和個人可識別資訊 (PII) 以處理企業資料，同時產生客製化模型的合成資料。
- NVIDIA NeMo Customizer 可以客製嵌入模型，以提高 RAG 準確性。
- NVIDIA NeMo Evaluator 透過獨立評估擷取和產生元件以及評估整體，衡量 RAG 應用程式的效能。
- NVIDIA NeMo Guardrails 確保 AI 助理保持準確、得體、安全與切題。
- NVIDIA NeMo Retriever 實現精準和保護隱私的大規模資訊擷取。

NVIDIA NIM 以業界標準 API 和企業級安全性簡化先進 AI 模型的部署。其提供預先建立的最佳化 AI 推論微服務，讓 AI 助理可以在雲端、資料中心和工作站環境中安全運作。NIM 是專為與現實無縫整合而設計，確保在任何位置都能輕鬆部署和擴充。

NVIDIA Blueprints 為代理和生成式 AI 使用案例提供參考工作流程，協助企業建立和啟用自訂 AI 助理。

準備開始了嗎？

利用以下資源開始 AI 智慧助理的旅程：

下載 NVIDIA AI 藍圖，以建立適用於客戶服務的智慧 AI 助理，或開發可擴充、可客製化的企業 RAG 流程。直接使用或與其他藍圖結合，以適用於進階應用，例如數位人類。

在您的基礎架構上部署 NVIDIA AI Enterprise 90 天試用版，實現順利整合與擴充性。