主權AI:建立自己的大型語言模型

在AI時代,擁有自己的語言模型不再是遙不可及的夢想。本報告將帶您深入了解如何使用Ollama建立本地LLM,以及部署所需的硬體配置,讓您掌握真正的AI主權。

智慧科技產研中心

陳啟彰　教授

為什麼需要自建大型語言模型?

數據主權與隱私

敏感數據不需上傳雲端,完全掌控在自己手中,符合資料保護法規要求。

客製化與彈性

根據特定領域需求調整模型,訓練專屬知識庫,打造符合業務需求的AI助手。

成本控制

避免持續的API呼叫費用,長期來看更具經濟效益,特別適合高頻使用場景。

效能與延遲

本地推理減少網路延遲,在內網環境中提供更快的回應速度。

認識Ollama:輕鬆建立本地LLM的利器

什麼是Ollama?

Ollama是一個開源工具,讓您能夠輕鬆在本地機器上運行大型語言模型。它簡化了模型下載、部署和管理的複雜流程,提供類似Docker的使用體驗。

核心優勢

- 簡單的命令列界面
- 支援多種熱門模型
- 自動化的模型管理
- RESTful API整合
- 跨平台支援(Windows, macOS, Linux)

使用Ollama建立LLM的完整步驟

步驟一:安裝Ollama

前往Ollama官網下載適合您作業系統的安裝程式。安裝過程簡單快速,類似安裝一般應用程式。也可安裝Web介面的Ollama,使用容易。

步驟二:選擇模型

瀏覽Ollama模型庫,根據您的需求選擇適合的模型。熱門選項包括Llama3:8b、Llama3:70b, Qwen3:8b、Gemma等。考慮模型大小與硬體配置的平衡。

步驟三:拉取模型

使用指令「ollama pull [model-name]」下載選定的模型。系統會自動處理依賴項和配置,首次下載可能需要一些時間。

步驟四:運行模型

執行「ollama run [model-name]」啟動互動式對話界面。您也可以透過API呼叫將模型整合到應用程式中。

步驟五:調整與優化

根據使用情況調整參數,如溫度、top-p等。監控資源使用情況,必要時進行模型量化以提升效能。

實用指令速查表

基礎操作

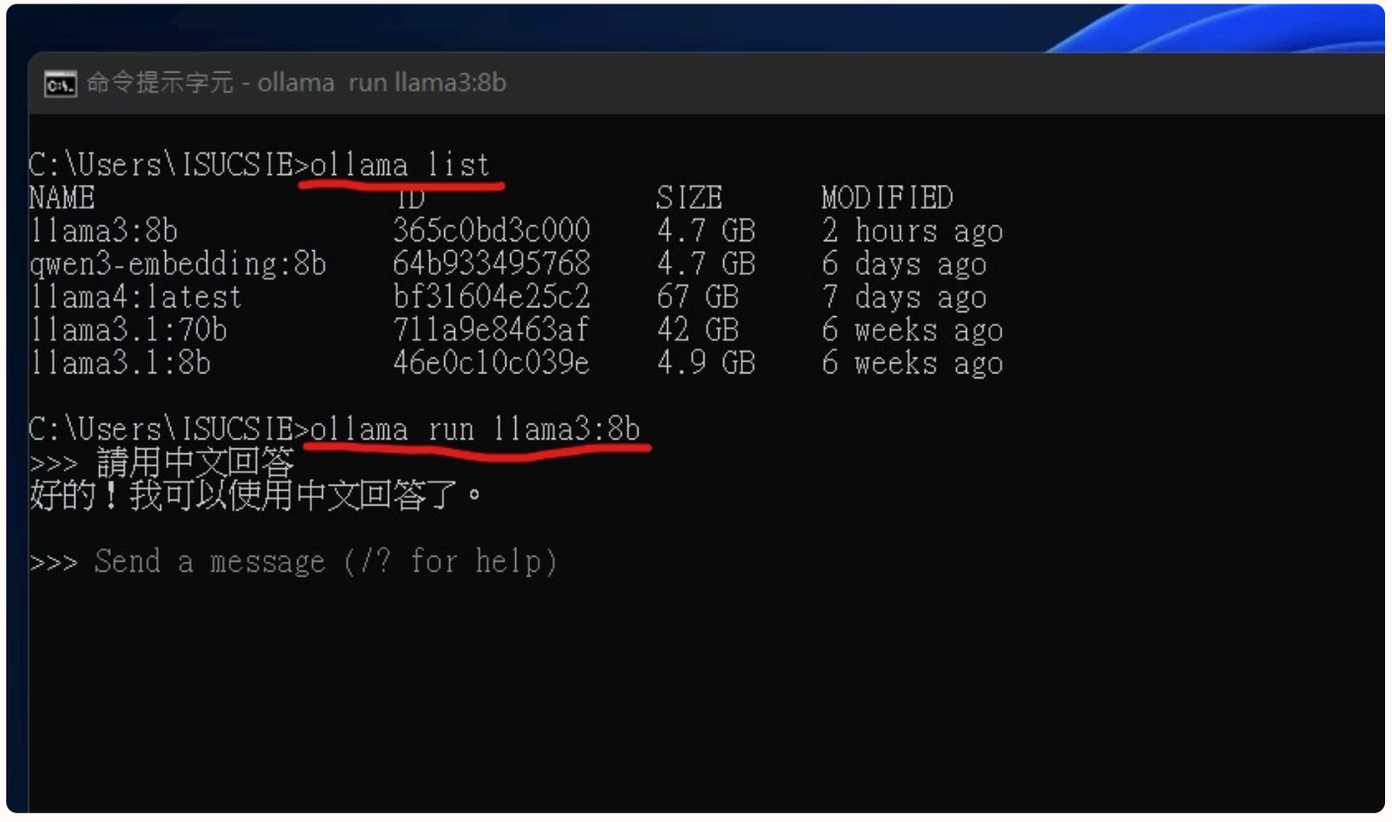
```
ollama list          # 列出已安裝模型
ollama pull llama3   # 下載模型
ollama run llama3    # 運行模型
ollama rm llama3     # 刪除模型
```

進階配置

```
ollama serve         # 啟動API服務
ollama create mymodel   # 從Modelfile建立自訂模型
ollama show llama3   # 查看模型資訊
```

API呼叫範例

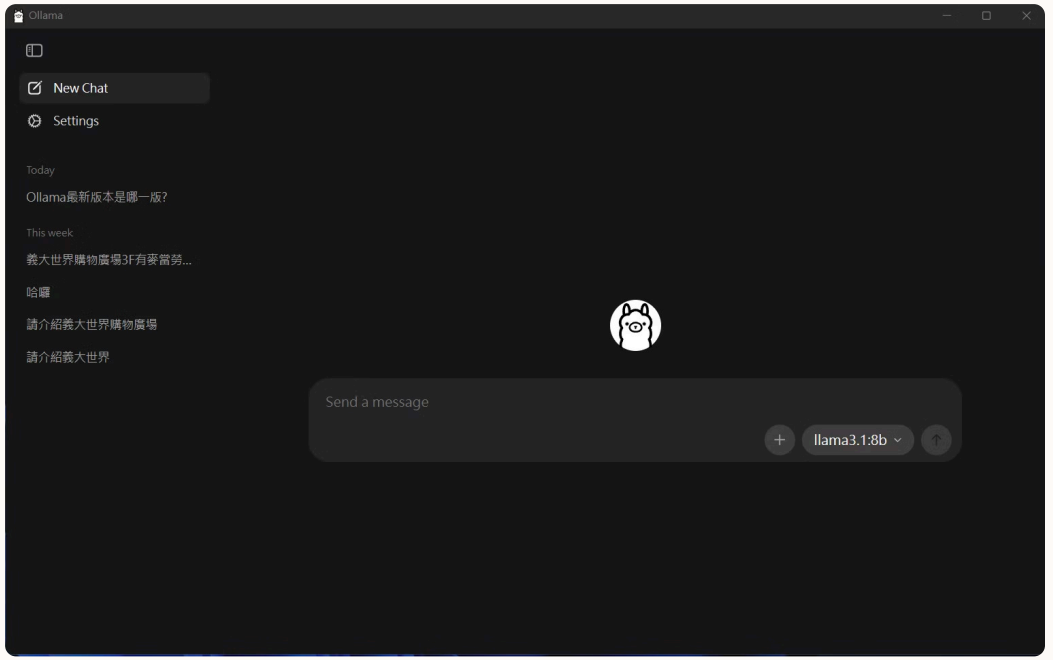
```
curl http://localhost:11434/api/generate \
 -d '{"model":"llama3","prompt":"你好"}'
```

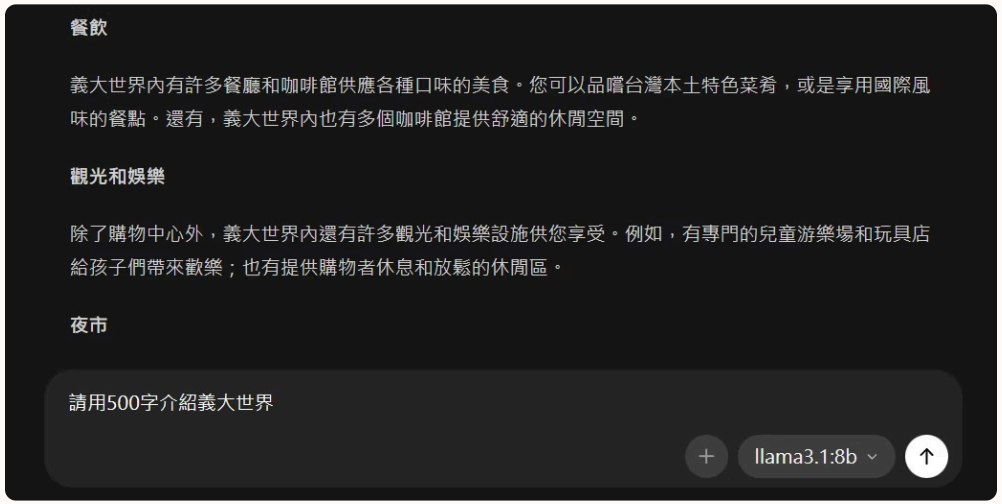


Ollama app介面: 讓LLM操作更直覺

除了指令行操作外,Ollama也支援app介面,讓非技術使用者也能輕鬆與本地LLM互動。這些介面通常提供直觀的聊天視窗、模型管理功能及參數調整選項。

- **友善使用者體驗:** 透過圖形介面,擺脫複雜指令。
- **模型視覺化管理:** 輕鬆切換、刪除或查看模型資訊。
- **即時對話功能:** 享受流暢的LLM對話體驗。

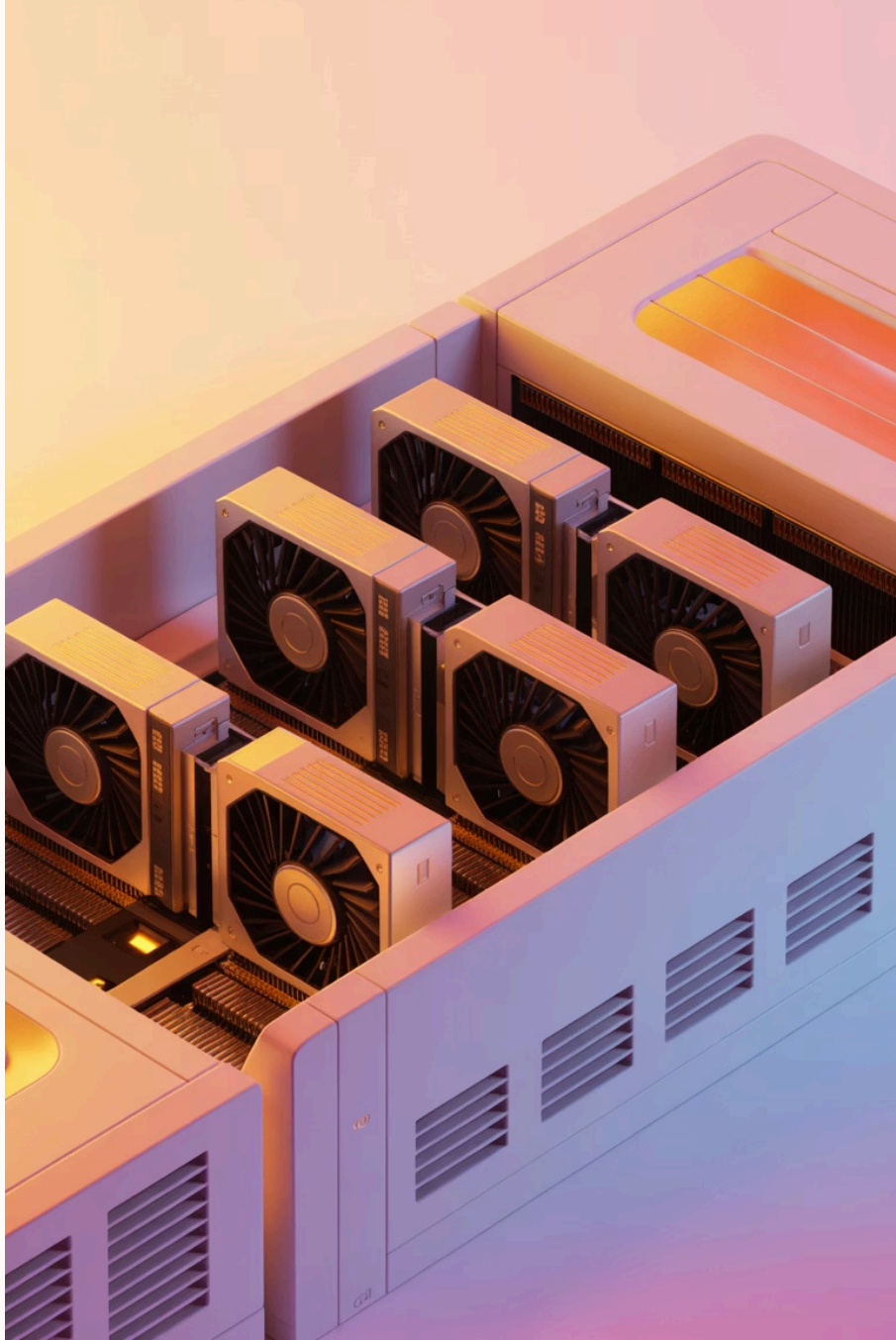




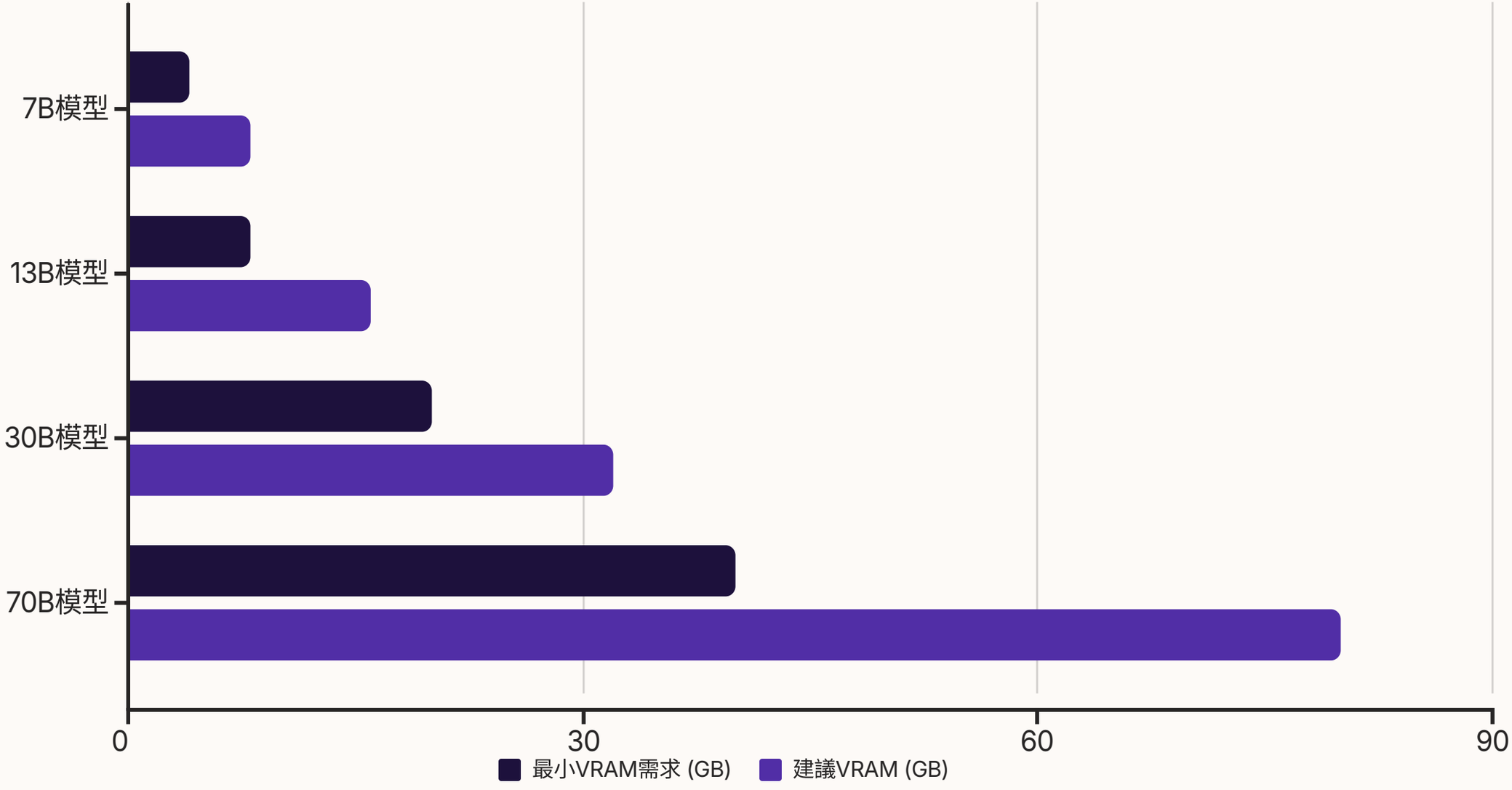
硬體需求概覽:選擇適合您的配置

運行大型語言模型對硬體有一定要求,特別是記憶體和GPU。了解不同模型的硬體需求,能幫助您做出最經濟有效的投資決策。以下我們將詳細分析各種規模模型的硬體配置建議。

> **關鍵指標:**模型參數量直接影響所需的VRAM和系統記憶體。量化技術(如4-bit、8-bit)可以顯著降低硬體門檻。



不同規模LLM的硬體需求比較



以上數據基於4-bit量化模型。使用完整精度(FP16)將需要約2-4倍的記憶體。對於多數應用場景,量化後的模型已能提供優秀的效能表現。VRAM(Video RAM, GPU卡上的記憶體)

推薦硬體配置方案

入門級配置

適用模型: 7B參數以下

- GPU: NVIDIA RTX 3060 (12GB) 或 RTX 4060 Ti
- RAM: 16GB DDR4
- 儲存: 512GB SSD
- 預算: 約NT$40,000-60,000

專業級配置

適用模型: 13B-30B參數

- GPU: NVIDIA RTX 4090 (24GB) 或 A5000
- RAM: 64GB DDR5
- 儲存: 1TB NVMe SSD
- 預算: 約NT$80,000-100,000

企業級配置

適用模型: 70B參數以上

- GPU: NVIDIA A100 (80GB) 或 H100
- RAM: 256GB+ DDR5
- 儲存: 2TB+ NVMe RAID
- 預算: NT$200,000起跳

優化策略與最佳實踐

模型量化

使用4-bit或8-bit量化技術可以大幅降低記憶體需求,同時保持90%以上的模型效能。Ollama原生支援量化模型,部署時自動選用最佳配置。

批次處理

調整批次大小(batch size)以平衡吞吐量與延遲。較大的批次能提高GPU利用率,但會增加單次推理時間。根據應用場景找到最佳平衡點。

上下文管理

限制上下文長度可以節省記憶體並提升速度。對於多輪對話,實施智能上下文截斷策略,保留關鍵資訊。

效能監控

使用工具監控GPU使用率、記憶體佔用和推理速度。定期分析瓶頸,進行針對性優化。建立效能基準以評估改進效果。

開始您的AI主權之旅

立即行動

1. 評估您的使用場景與預算
2. 選擇合適的硬體配置
3. 安裝Ollama並下載第一個模型
4. 測試並優化效能表現
5. 逐步擴展到生產環境

> 建立自有LLM不僅是技術選擇,更是對數據主權和AI未來的長期投資。

資源連結

Ollama官方文檔、模型庫、社群論壇都提供豐富的學習資源。加入開發者社群,與全球工程師交流經驗,持續精進您的AI部署能力。

LLama 3 + RAG：打造智能檢索增強的生成式 AI

探索 LLama 3 大型語言模型與RAG(檢索增強生成技術)的完美結合，開啟智能應用的新篇章

內容概覽

01

認識 LLama 3

深入了解 Meta 最新開源大型語言模型的核心特性與技術優勢

02

探索 RAG 技術

掌握檢索增強生成的運作原理與實際應用場景

03

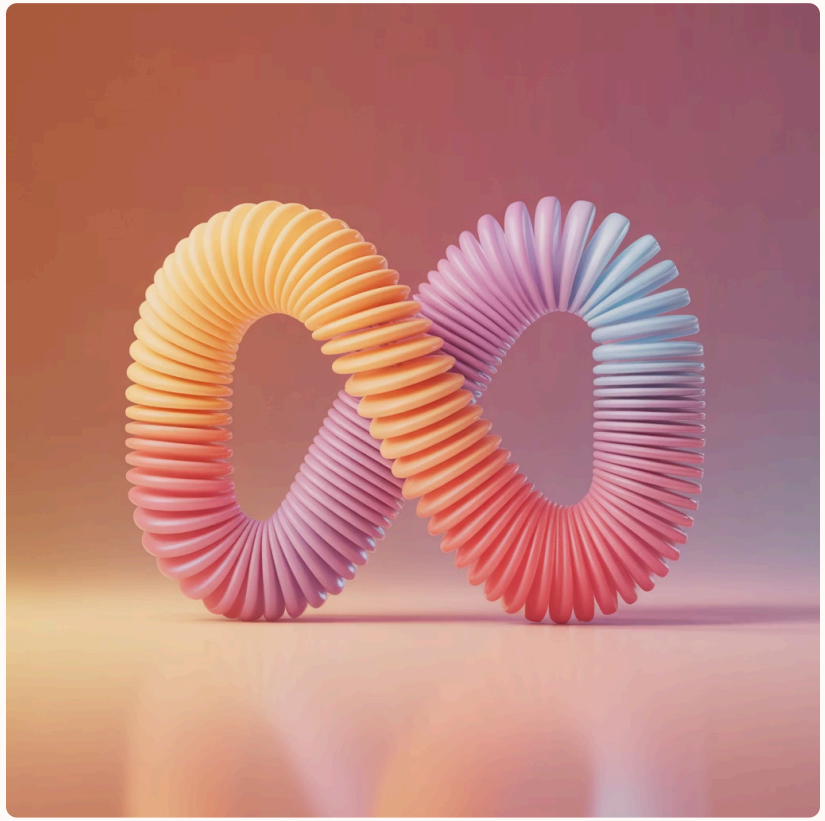
技術整合實踐

學習如何將 LLama 3 與 RAG 結合，打造高效智能系統

什麼是 LLama 3？

LLama 3 是 Meta 在 2024 年推出的開源大型語言模型，代表著生成式 AI 技術的重大突破。作為 LLama 系列的第三代產品，它在理解能力、推理邏輯、多語言支援等方面都有顯著提升。

LLama 3 提供多種參數規模版本，從 8B 到 70B 不等，讓開發者可以根據實際需求選擇合適的模型大小，在性能與資源消耗之間取得平衡。



LLama 3 的核心優勢

完全開源

免費使用且可自由部署，降低企業 AI 應用門檻

卓越性能

在多項基準測試中表現優異，媲美閉源商業模型

多語言支援

支援繁體中文在內的多種語言，適合全球化應用

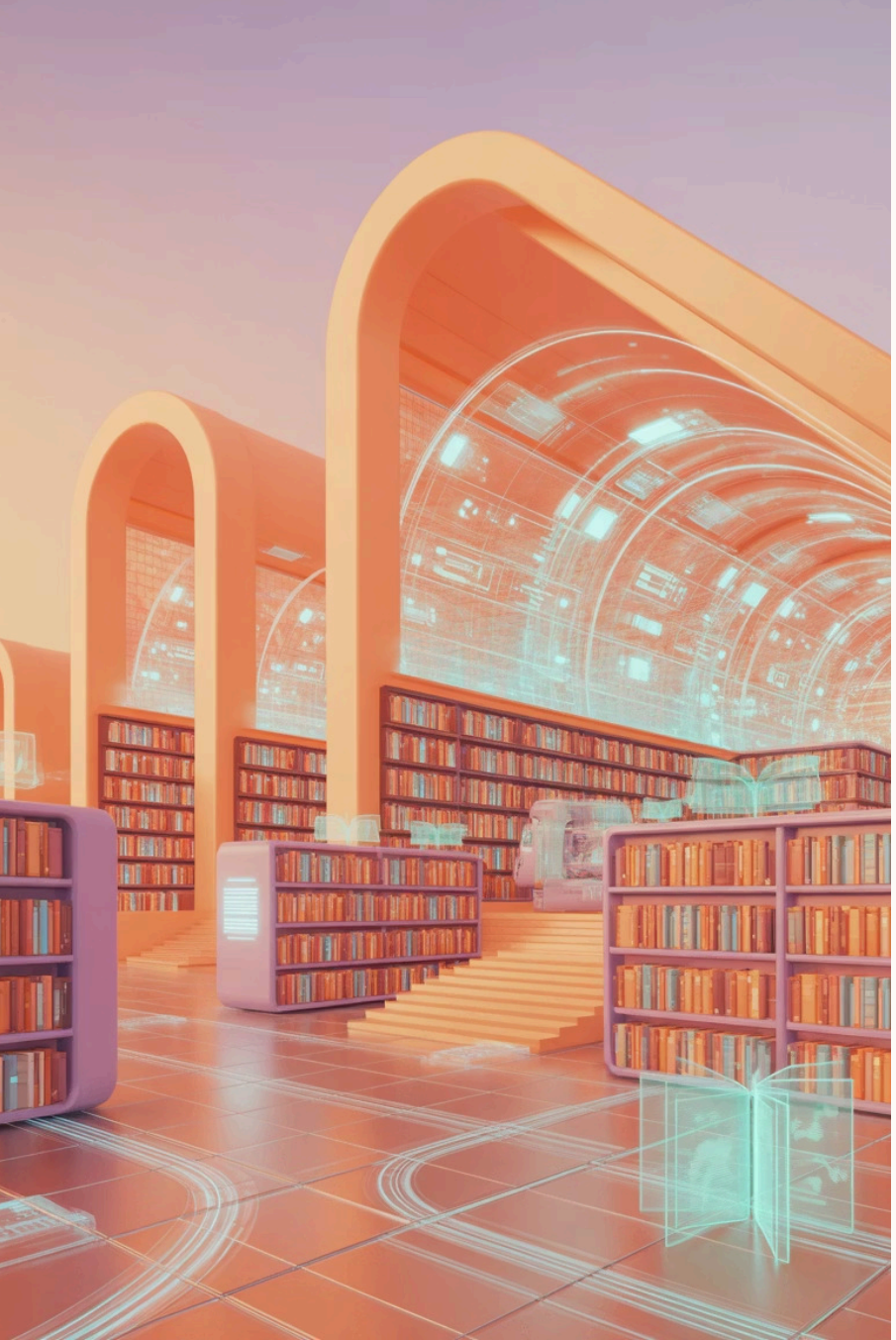
靈活部署

可在本地或雲端運行，滿足不同安全與合規需求

LLama 3 技術規格一覽

模型版本	參數量	建議應用場景
LLama 3 8B	80 億參數	輕量級應用、快速回應
LLama 3 70B	700 億參數	複雜推理、專業內容生成
LLama 3 405B	4050 億參數	頂級性能、研究級應用

不同規模的模型提供了從邊緣設備到企業級伺服器的完整解決方案，開發者可以根據計算資源、回應速度、準確度需求來選擇最適合的版本。



什麼是 RAG（檢索增強生成）？

RAG（Retrieval-Augmented Generation）是一種創新的 AI 技術架構，通過結合外部知識檢索與語言模型生成能力，解決傳統 LLM 的知識侷限性問題。

簡單來說，RAG 就像是給 AI 配備了一個即時可查詢的知識庫，讓它能夠在回答問題時先搜尋相關資料，再基於這些資料生成更準確、更新的答案。

RAG 的運作流程

用戶提問

系統接收使用者的查詢請求

檢索相關資料

從向量資料庫中搜尋最相關的文檔片段

增強提示詞

將檢索結果與原始問題結合形成完整語境

生成答案

LLM 基於增強後的語境生成精確回應

RAG 解決的核心問題

傳統 LLM 的限制

- 訓練資料有時間截止點，無法獲取最新資訊
- 容易產生「幻覺」，編造不存在的事實
- 缺乏企業專屬知識與內部文檔資料
- 難以引用具體來源，降低可信度

RAG 的解決方案

- 動態檢索最新資料，確保資訊時效性
- 基於真實文檔回答，大幅減少幻覺現象
- 整合企業知識庫，提供客製化服務
- 可追溯資料來源，提升答案可信度

為什麼 RAG 如此重要？

85%

準確度提升

相較於純 LLM，RAG 系統可將答案準確度提升至 85% 以上

70%

幻覺(一本正經,胡說八道)減少

基於真實文檔檢索，可減少約 70% 的錯誤或虛構內容

3倍

成本效益

相比重新訓練模型，RAG 的部署成本僅約三分之一

這些數據顯示 RAG 不僅在技術層面有突破，更在實務應用中展現顯著的投資報酬率，成為企業導入生成式 AI 的首選架構。

LLama 3 + RAG：完美結合

當開源的 LLama 3 遇上靈活的 RAG 架構，將創造出既強大又可控的智能系統。

結合 LLama 3 與 RAG 的優勢

成本可控

使用開源 LLama 3 搭配自建 RAG 系統，避免高昂的 API 呼叫費用，特別適合高頻次查詢場景。長期來看可節省 60% 以上的營運成本。

資料安全

所有資料與模型均可部署在本地或私有雲端，完全掌控敏感資訊，符合金融、醫療等產業的嚴格合規要求。

客製彈性

可根據特定領域需求微調 LLama 3，並整合專屬知識庫，打造真正符合業務需求的智能助手。

持續進化

RAG 的知識庫可隨時更新，無需重新訓練模型，讓系統始終保持最新狀態，快速回應市場變化。

實際應用場景

智能客服系統

整合產品手冊、FAQ 與歷史工單，提供 24/7 精準客戶支援

企業知識管理

檢索公司內部文件、政策規範，快速回答員工疑問

法律文件分析

搜尋相關判例與法規，輔助律師進行案件研究

醫療資訊查詢

檢索醫學文獻與診療指引，協助醫護人員做出決策

技術架構建議

文檔預處理

將企業知識庫文件切分成適當大小的文本片段，通常為 500-1000 字元

向量化儲存

使用 Embedding 模型將文本轉換為向量，存入向量資料庫如 Milvus 或 FAISS

語義檢索

將用戶問題轉為向量，在資料庫中搜尋最相似的 Top-K 個文檔片段

提示詞組裝

將檢索結果與原始問題組合成完整的提示詞(Prompt)，輸入 LLama 3 模型

生成與回覆

LLama 3 基於增強後的語境生成答案

實施關鍵成功因素

高品質資料

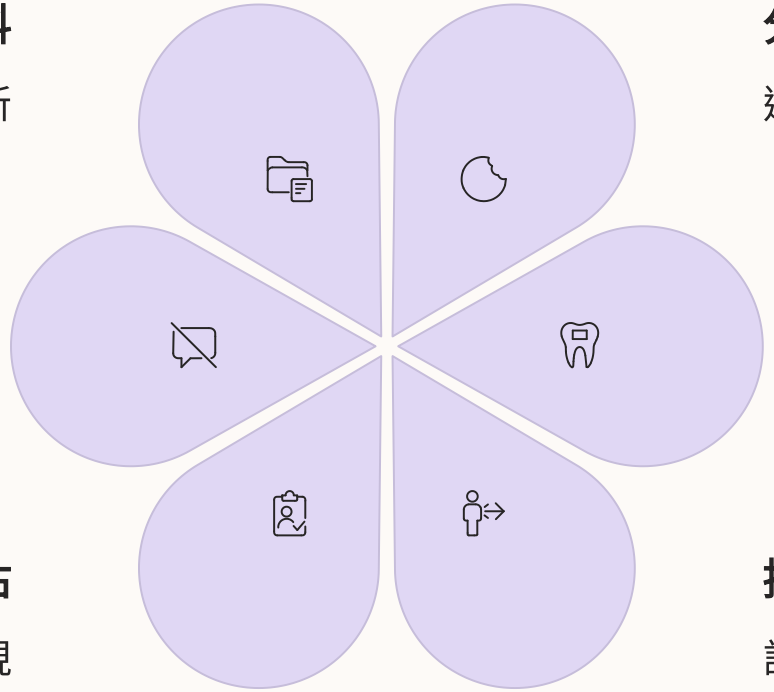
確保知識庫內容準確、完整、定期更新

用戶回饋

收集使用者反饋不斷改進檢索與生成品質

效能評估

建立完善的測試集持續優化系統表現



分塊(Chunk)策略

選擇適當的文本切分方式保持語義完整性

向量模型

使用高品質的 Embedding 模型提升檢索精準度

提示詞工程

設計有效的提示詞模板引導模型生成

開始您的 LLama 3 + RAG 之旅

結合 LLama 3 的強大生成能力與 RAG 的精準檢索優勢，您可以打造出既智能又可靠的 AI 應用系統。無論是企業知識管理、客戶服務、還是專業領域諮詢，這個技術組合都能提供卓越的解決方案。

下一步行動

- 評估您的應用場景與資料準備度
- 選擇適合的 LLama 3 模型版本
- 建立或整合向量資料庫
- 開發 RAG 檢索與整合邏輯
- 持續測試與優化系統效能

Thank You