資料來源：ChatGPT　編輯：屏東大學 周國華老師

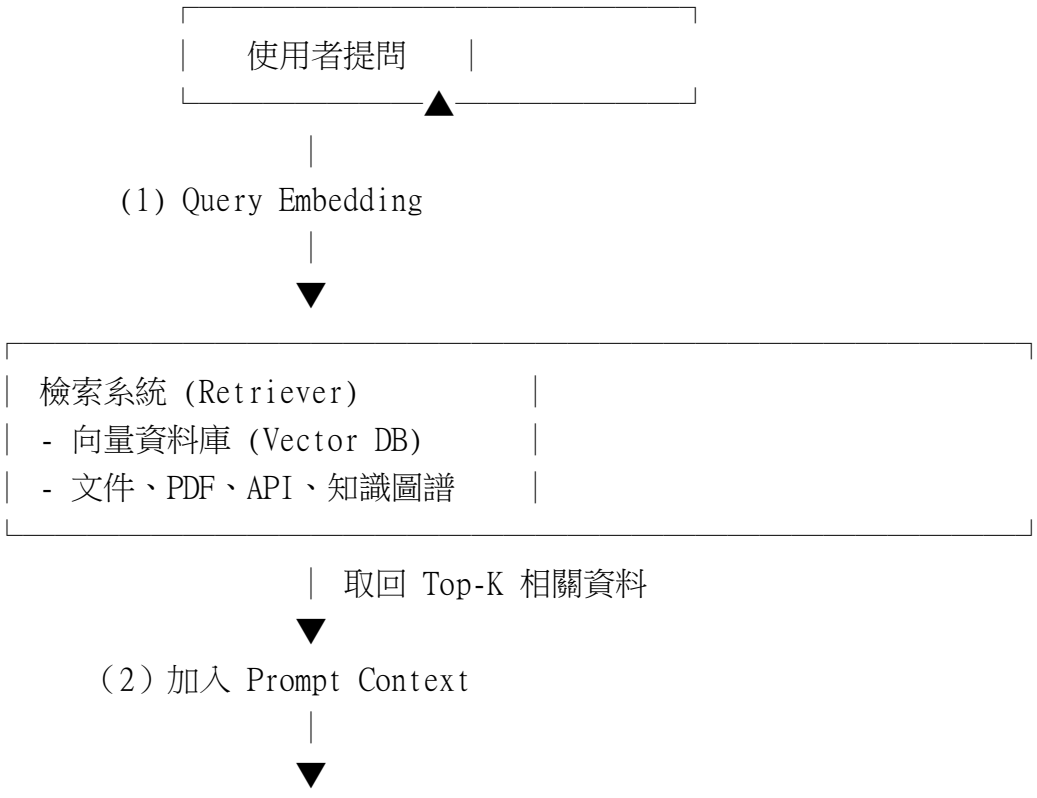
以下是對 **RAG（Retrieval-Augmented Generation，檢索增強生成）** 的清楚、完整解析，非常適合用於課堂教學、報告、AI 技術課程或企業應用理解。

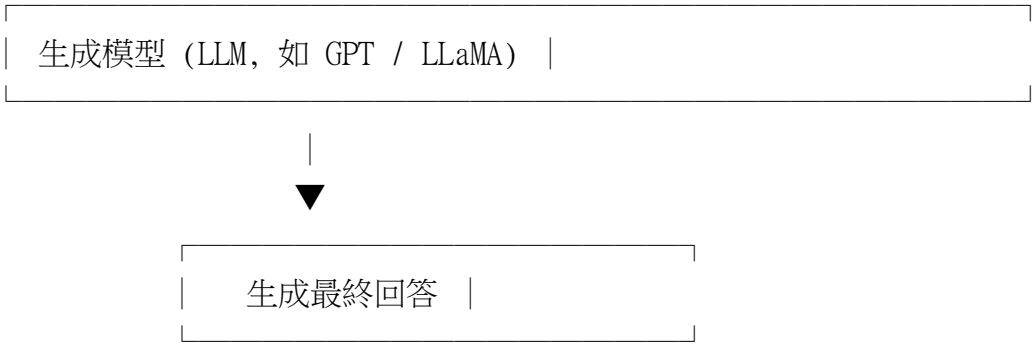
☑ 什麼是 RAG？

RAG（Retrieval-Augmented Generation） 是一種將「知識檢索」與「生成式 AI 模型」結合的方法。
它讓模型在回答問題前，先從資料庫、文件或知識庫中**檢索相關內容**，再根據這些真實資料進行**生成回應**。

核心概念：
LLM + 知識檢索 = 回答更準確 + 減少幻覺（Hallucination）

☑ RAG 的運作流程圖





☑ RAG 為什麼重要？

傳統 LLM 問題	RAG 提供的解決方式
模型不包含最新或專有知識	可連接外部資料、公司內部文件、即時資訊
容易出現幻覺（亂造答案）	回答基於真實文獻或資料庫，提高可信度
無法處理長文件內容細節	向量檢索能從大量文本中抽取相關片段
模型訓練成本高	RAG 不必重新訓練模型，只需提供資料庫

☑ RAG 系統的主要組成元件

組件	內容
Embedding 模型	將文本轉成向量，例如 OpenAI text-embedding-3, Sentence-BERT
Vector Database 向量資料庫	儲存與搜尋文本向量，如 FAISS、Pinecone、Weaviate、Milvus
Retriever（檢索器）	從資料庫找出與問題最相關的 Top-K 文本
LLM（生成模型）	如 GPT-4/5、LLaMA、Claude，用於根據資料生成回答
Prompt Template	將「問題 + 檢索內容」轉為一致的輸入格式給模型

☑ 小範例：RAG 問答系統 Prompt

你提供給模型的 Prompt：

你是一位專業助手。以下是與問題相關的文件內容：

【Document】
台灣的 ESG 申報自 2026 年起，所有上市櫃公司須依 XBRL 格式送出永續報告書。

【Question】
台灣企業何時開始強制使用 XBRL 進行 ESG 申報？

請根據文件回答，若無資訊請說「文件未提及」。

模型回覆：
台灣自 2026 年起，所有上市櫃公司必須使用 XBRL 格式提交 ESG 永續報告書。

☑ RAG 的應用領域

領域	應用
法律	法規查詢 + 合約分析
教育	教材摘要、技術文件問答
ESG / 會計	IFRS、GRI、XBRL 資料庫智慧問答
醫療	醫學文獻 + 臨床決策支援
客服	FAQ + 企業內知識庫問答
程式開發	文件檢索 + 程式碼生成（Docs + GPT）

☑ 總結一句話

RAG 就是讓 AI「先查資料，再回答」的架構，讓它更準確、安全、有根據。

☑ 如果您需要，我能協助您：

✔ 教學簡報（PPT 大綱 + Mermaid 圖）
✔ Python + LangChain + OpenAI API RAG 實作範例
✔ 企業 ESG / 會計 / XBRL 文件問答系統設計
✔ 向量資料庫（FAISS / Pinecone / Weaviate）配置流程
✔ 如何評估 RAG 系統準確率 (Retrieval + Generation Metrics)

只要告訴我：「請做 RAG 實作範例」或「幫我畫 PPT 架構圖」即可開始！

要繼續嗎？