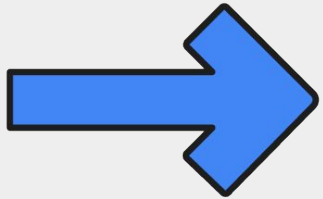
Google Developer Group
On Campus National Sun Yat-sen University

Intro to RAG and Vertex AI Search

RAG 基本介紹

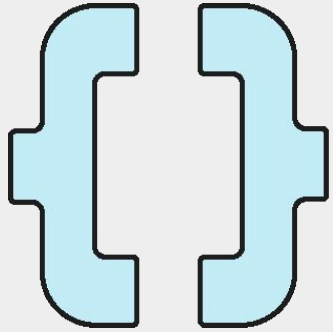
與Vertex AI Search實作

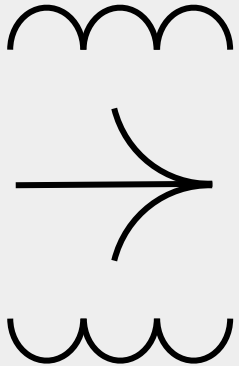
楷鈞, Kai - 2024” GDGoC NSYSU Lead



主題大綱

1. 什麼是 RAG?
   What's RAG?
2. 使用RAG技術建構？
   Why using RAG? What does it provides?
3. 使用 Vertex AI 實踐RAG技術
   Implement RAG using Vertex AI

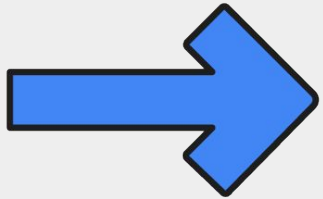




01

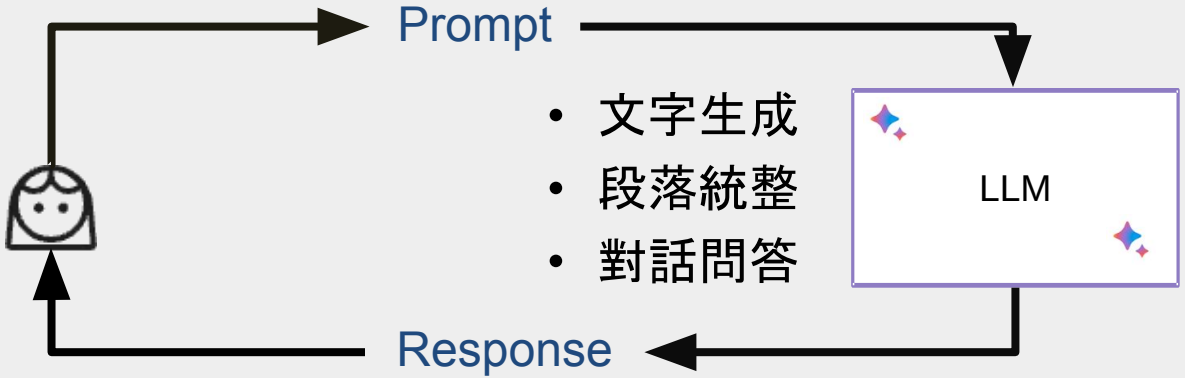
什麼是 RAG?

What's RAG?



大語言模型 (LLM)的經典用途

近年來，大型語言模型在知識搜尋、生成文章和邏輯推理方面非常出色。因為他們接受了大量公開資料的預先訓練。



語言模型有時文不對題 ...

或在胡說八道 ？

(機器幻覺 -Hallucinations)

語言模型通常只能理解.....

- 與訓練集相似的問題或內容
- 在對話中明確給出的解釋

這會導致了語言模型會時常被使用者「欺騙」，產生出不正確的內容。

這導致了在領域知識不足的情況下，他們有時會自顧自地回答問題，而忽略掉了正確性

Prompt

請告訴我台灣現在的總統是誰

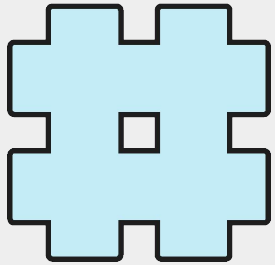
Response

我無法提供準確的時間資訊，所以無法正確回答這個問題。

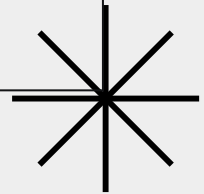
Expected

但根據我搜尋的資料，現在的總統是XXX

調教模型
來讓模型的回答更加準確



微調 (Fine-Tuning)	人為查驗 (Make Humans Check)	提示工程 (Prompt Engineering)
使用範例輸出輸入的方式來「教育」我們的模型 (也就是訓練模型) • 需要準備大量範例資料 • 成效有限	透過真人檢查的方式，確保模型輸出給使用者的資料都經過人為檢查 • 反應性極差、成本需求極高 • 潛在人為疏失風險	透過預先的提示，來規範模型的未來回答方向 • 可能面臨Token(成本)限制 • 容易受到上下文影響



這時候，我們也可以使用 .....

檢索增強生成

(RAG-Retrieval Augmented Generate)

原本遇到的問題：

- 由其他公司訓練的LLMs不會知道特定領域的知識或商業資料
- LLM 無法用於即時資訊(因為沒有受過新資料的訓練)
- LLM 無法直接提供「領域知識參考來源」

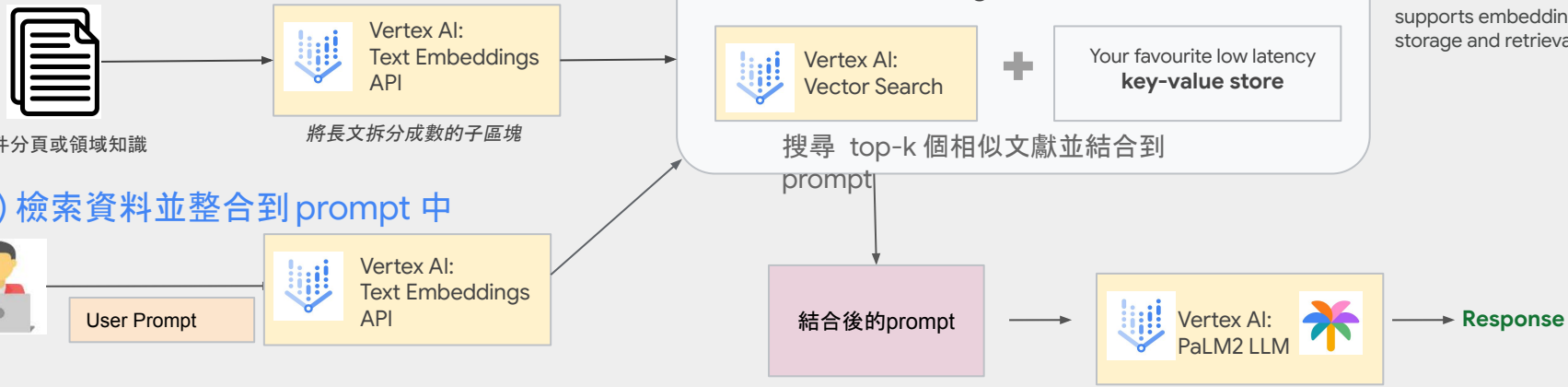
我們的解決辦法：

- 使用領域知識資料檢索系統，給LLM 你的專屬領域知識
- 將資料檢索系統填入即時資料，並給LLM 相關即時背景資訊

檢索增強生成

(RAG-Retrieval Augmented Generate)

(1) 建立資料索引



還是不太懂甚麼是 .....

檢索增強生成？

還是不太懂甚麼是 .....

檢索增強生成？

我晚餐要吃甚麼？

還是不太懂甚麼是 .....

檢索增強生成？

我查了一下,

找到這附近可以吃的食物

還是不太懂甚麼是 .....

檢索增強生成？

讓我們用

（熱量/澱粉）

來給定分數

(0, 0)

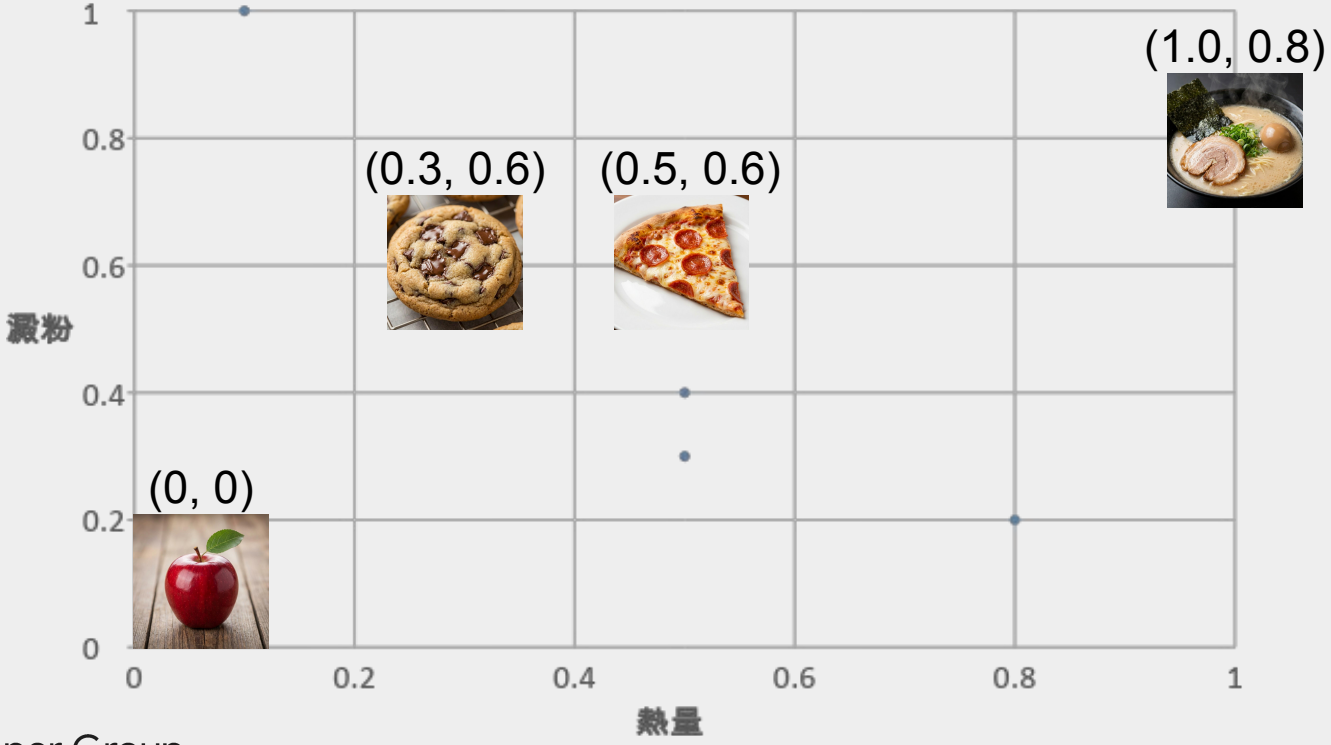
(0.3, 0.6)

(1.0, 0.8)

(0.5, 0.6)

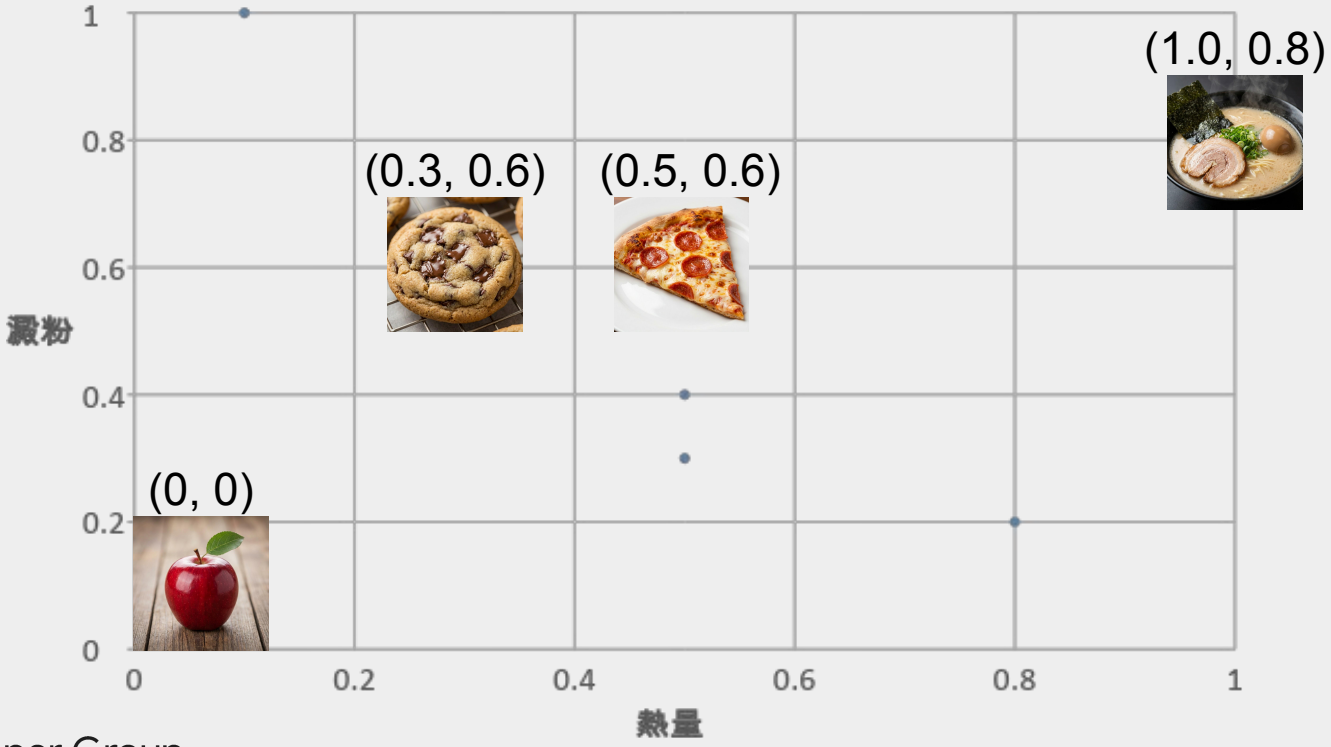
還是不太懂甚麼是 .....

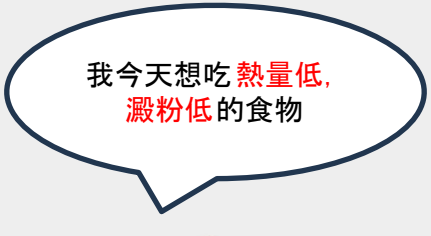
檢索增強生成？



還是不太懂甚麼是 .....

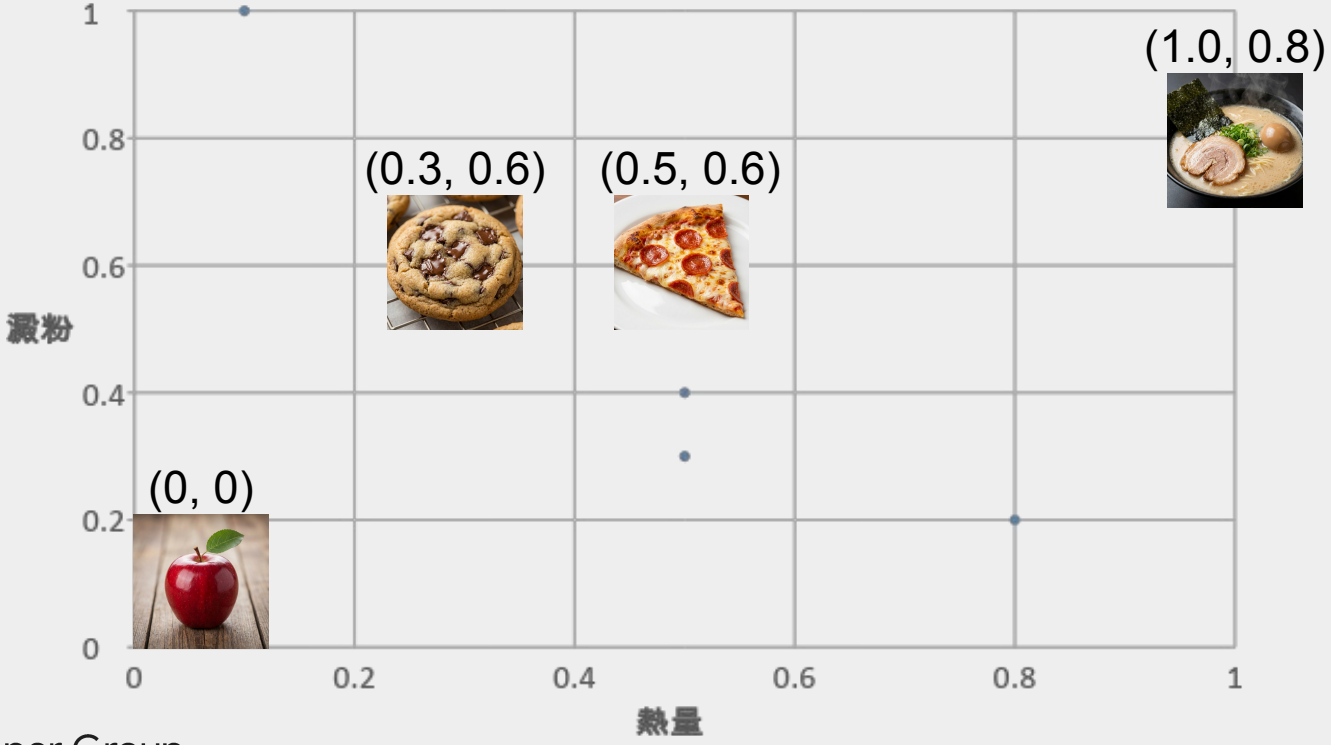
檢索增強生成？





還是不太懂甚麼是 .....

檢索增強生成？

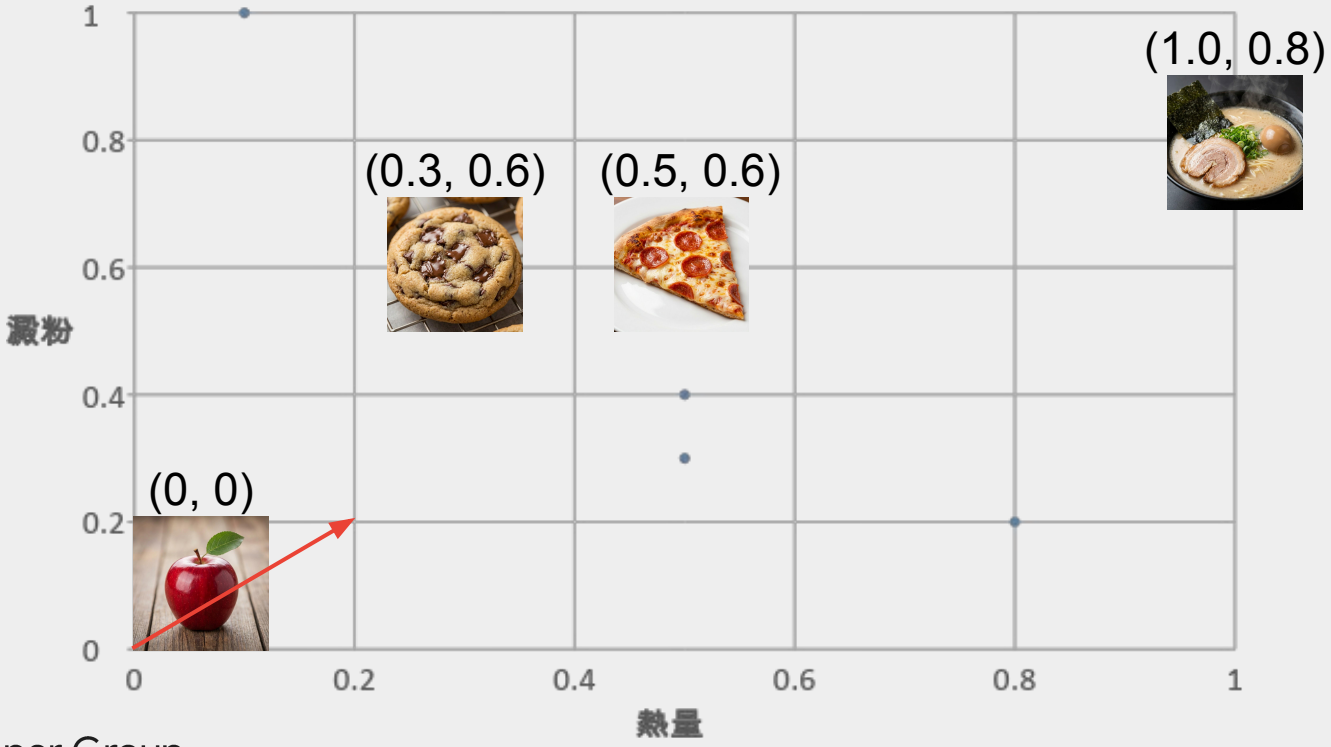


Vertex AI:
Text Embeddings
API

Google Developer Group
On Campus National Sun Yat-sen University

還是不太懂甚麼是 .....

檢索增強生成？



[0.2, 0.2]

檢索增強生成

(RAG-Retrieval Augmented Generate)

(1) 建立資料索引

文件分頁或領域知識

Vertex AI:
Text Embeddings
API

將長文拆分成數的子區塊

Embeddings + Content Index

Vertex AI:
Vector Search

Your favourite low latency
key-value store

(2) 檢索資料並整合到prompt 中

User Prompt

“我今天想吃熱量低，澱粉低的食物”

Vertex AI:
Text Embeddings
API

搜尋 top-k 個相似文獻並結合到prompt

結合後的prompt

“.... lorem ipsum lorem ipsum….

Context: Search from nearby, Apple is one of the suitable choice”

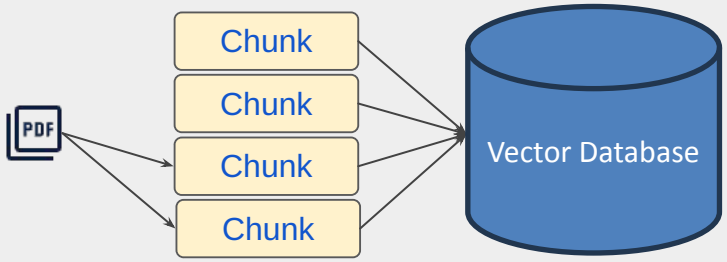
Vertex AI:
PaLM2 LLM

模型回應

Google Developer Group
On Campus National Sun Yat-sen University

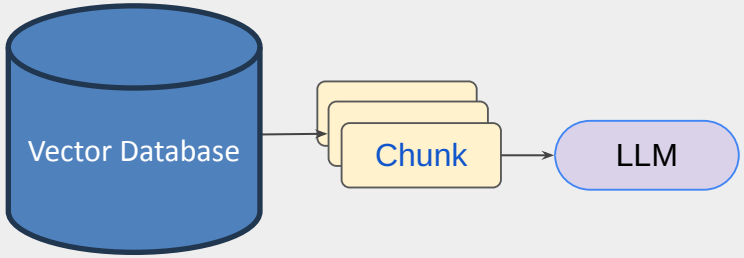
RAG 工作流程範例

建立QA系統



資料匯入/解析

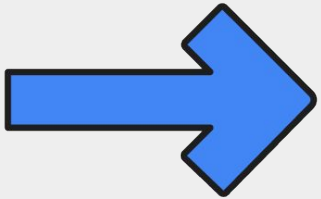
- 將資料分解成數個區塊
- 每個區塊都是一段純文字資料
- 將每一個"區塊"放到向量資料庫中



查詢

- 從向量資料庫中尋找前 k 個最相似的區塊
- 將前步驟的資料輸入 LLM並產生綜合結果

02
使用RAG技術建構？



檢索增強生成

常見使用場景 / 應用

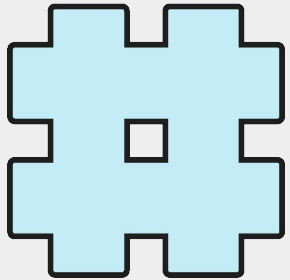
Question & Answering	Chatbots	Agents
對文件資料進行語意搜尋或綱要擷取。或使用 查詢的方式來回答外部資料情形。 （外部資料）	聊天機器人可以 透過反覆詢問以獲得更多的問題解釋，或回答後續更多的問題。 (記憶上下文）	「代理人」能夠透過使用者的輸入來分析複雜的問題、規劃任務，以及記憶已完成的任務。 （記憶”工具“及過去任務）

三者都具備"使用外部資料"的情形

特別適合使用RAG技術來加強建構

使用RAG 技術建構？

應用程式優勢在哪？

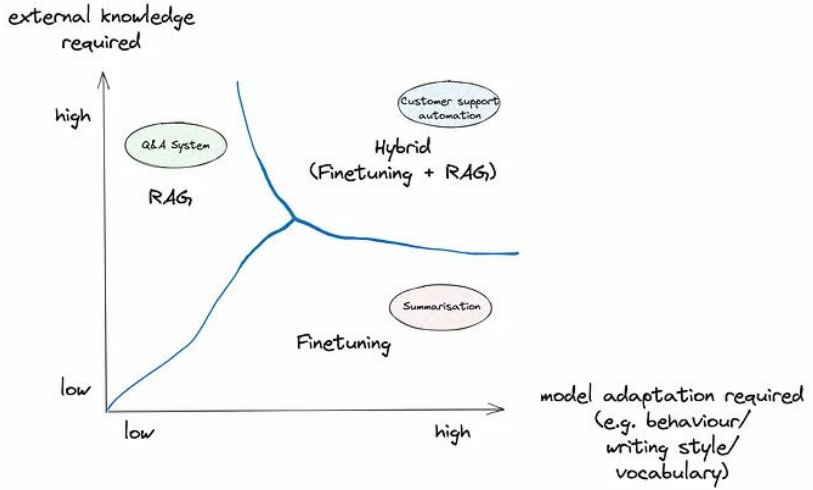


- 事實性與驗證機制: 提供可靠的準確性評估。
- 更準確地提供問題解析: 包含比通用 LLM 資料更具相關性的數據。
- 更新的數據: 保持內容的新穎性。
- 更快的數據更新: RAG 中的數據可持續更新。
- 成本較低：相比微調(Fine-tuning)，RAG 更快，更便宜。
- 限制授權能力：確保輸出符合訪問控制與授權規範。

Fine–tune vs RAG?

究竟差在哪？

	Fine-Tuning	RAG
是否適用外部知識？	❌/✅	✅
是否適用於多個模型？	❌	✅
是否有效的減少機器幻覺？	❌/✅	✅
資料是否可以動態化？	❌/✅	✅
是否可以解釋結果來源？	❌	✅



使用RAG 技術建構？

RAG 遇到的問題與挑戰 (Naive RAG)

挑戰維度	Naive RAG 的局限性
數據處理	簡單的固定長度切分，容易割裂語義，忽視表格與圖片結構。
查詢理解	直接檢索用戶原始查詢，難以處理模糊、簡短或複雜意圖。
檢索精度	僅依賴向量相似度（Dense Retrieval），無法處理關鍵詞精確匹配或專有名詞。
上下文質量	檢索結果中包含大量噪聲（雜訊）文檔（Distractors），導致 LLM 產生幻覺或「中間丟失（Lost in the Middle）」。

RAG 增強進化

What is RAG++?

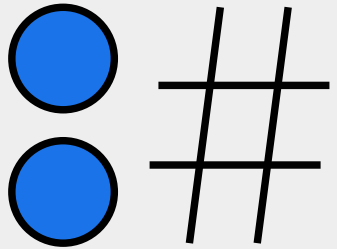
挑戰維度	RAG++ 的解決方案與技術路徑
數據處理	**自動分塊、多模態解析**：利用視覺模型解析檔案架構；並使用父文檔檢索（Parent Document Retrieval）的策略。
查詢理解	**查詢增強（Query Enhancement）**：查詢擴增、多路查詢生成，將用戶意圖轉化為機器可檢索的語義表示。
檢索精度	**混合檢索（Hybrid Search）**：結合稀疏向量（BM25）與稠密向量，並通過倒數排名融合（RRF）合併結果。
上下文質量	**重排序（Reranking）與壓縮**：使用 Cross-Encoder 對候選檔案進行精細評分，過濾無關內容，只保留高可信度區間。

跟上Agent熱潮

Agentic RAG

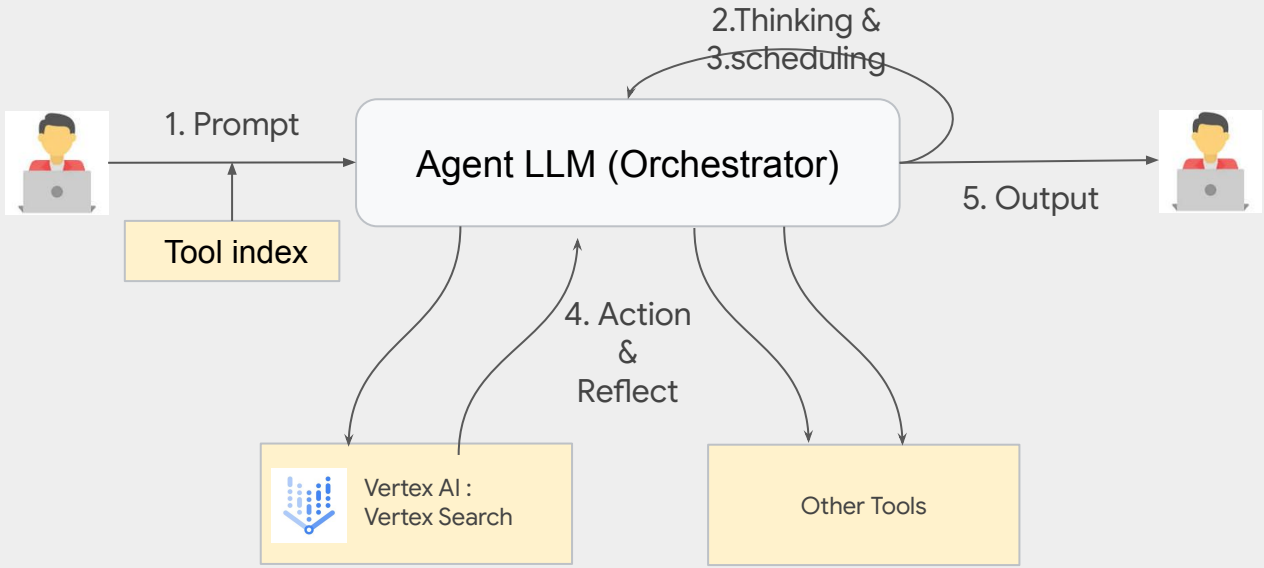
Agentic System 會以下面的動作組成一個系統

- 感知: 藉由組件認知現在的情形(例如:使用者輸入、可呼叫工具表
- 推理: 使用具推理架構(Thinking)的LLM對問題進行更深層的探討
- 規劃: 根據現有工具以及問題，制定一個解決的方案流程
- 行動: 根據制定好的流程執行動作，並在動作後進行反思，是否需要迭代？



跟上Agent熱潮

Agentic RAG



Recap: 檢索增強生成

(RAG-Retrieval Augmented Generate)

(1) 建立資料索引

文件分頁或領域知識

Vertex AI:
Text Embeddings
API

將長文拆分成數的子區塊

Embeddings + Content Index

Vertex AI:
Vector Search

Your favourite low latency
key-value store

(2) 檢索資料並整合到 prompt 中

User Prompt

"我今天想吃熱量低，澱粉低的食物"

Vertex AI:
Text Embeddings
API

搜尋 top-k 個相似文獻並結合到 prompt

結合後的prompt

"*.... lorem ipsum lorem ipsum….*

Context: Search from nearby, Apple is one of the suitable choice"

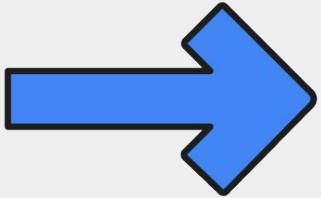
Vertex AI:
PaLM2 LLM

模型回應

Google Developer Group
On Campus National Sun Yat-sen University

Google Developer Group
On Campus National Sun Yat-sen University

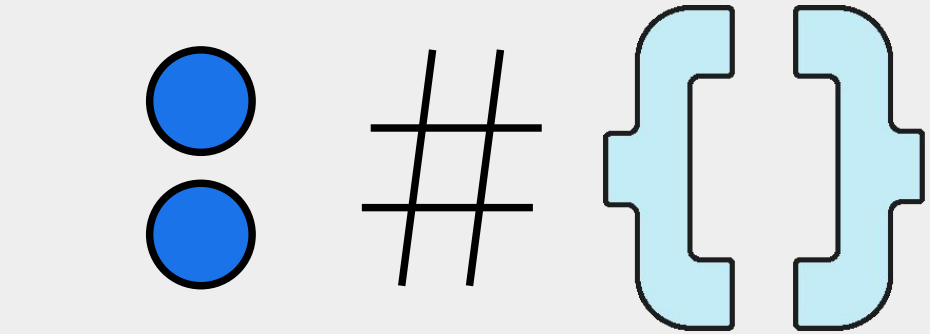
03
Vertex AI
實踐RAG技術

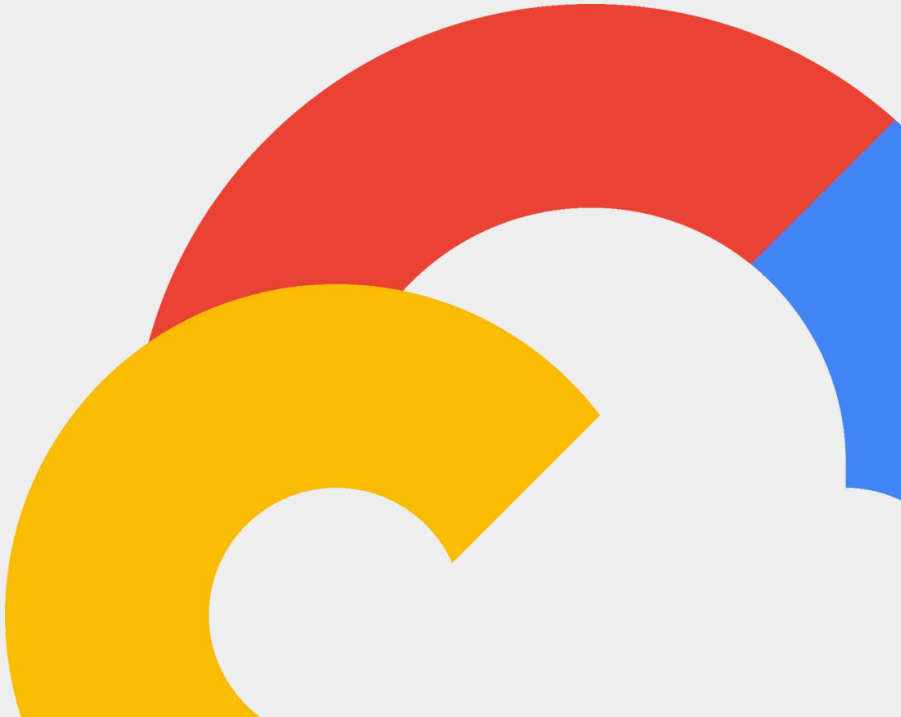


https://nsysugdsc.pse.is/RAG

Thank you

- Questions can be asked on Discord or privately
- Feel free to ask!





Google Developer Group
On Campus National Sun Yat-sen University