

RAG: 檢索增強生成簡介

Retrieval Augmented Generation

11/17/2024
Louisa Coffee

我認為，通用型人工智慧（AGI）還需要30到50年才有可能實現，這段距離非常遙遠，希望我們最終能抵達那一天。

AI科學家吳恩達（Andrew Ng）

2023-09-25

人工智慧應該很聰明

但是目前人工智慧還在學習中



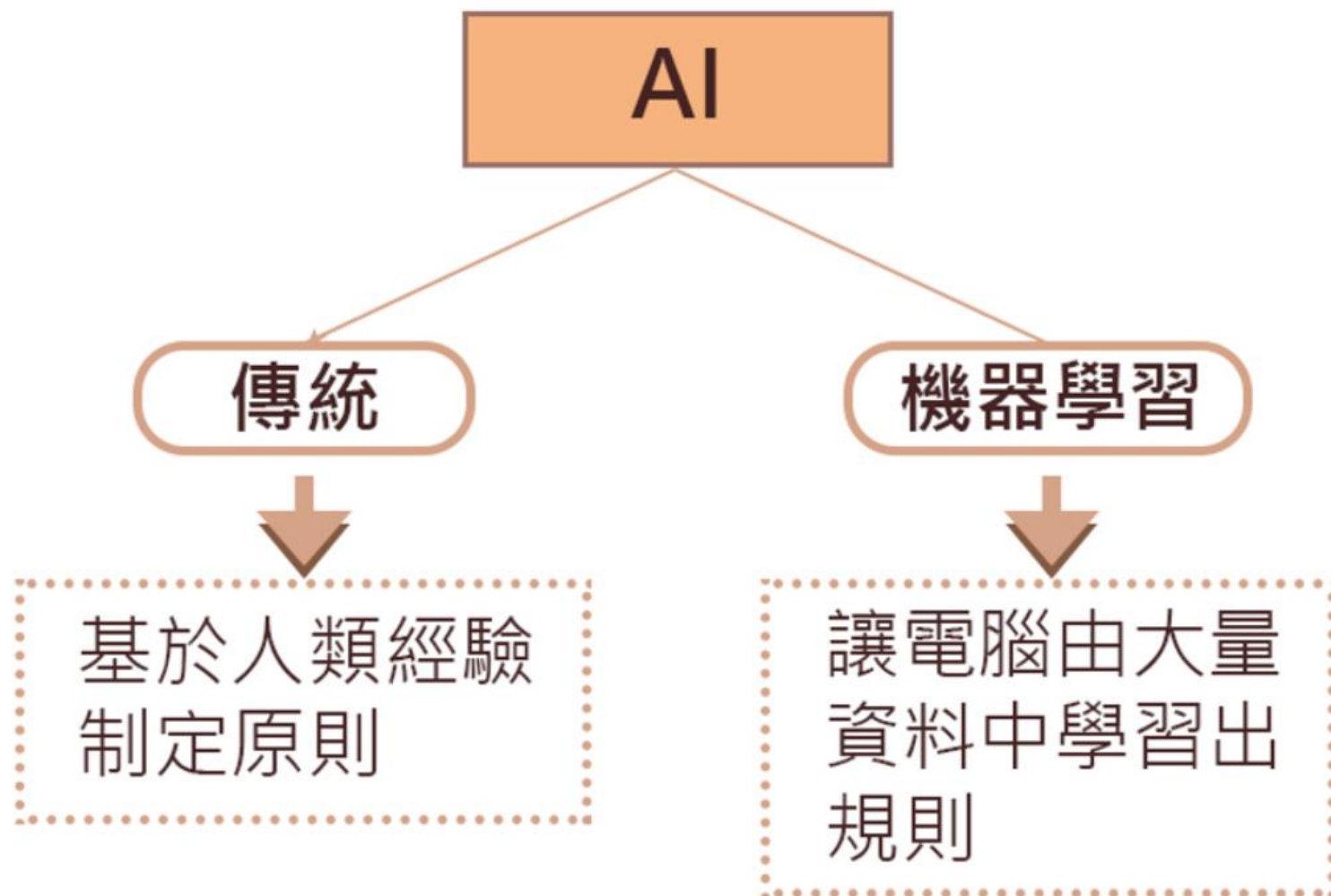
生成式 AI: LLM + RAG

- **AWS** Bedrock **Anthropic Claude 3**
- **Microsoft** Azure **OpenAI ChatGPT**
- **Google** **Gemini**

<https://www.kaggle.com/models/>

LLM:大型語言模型

Large Language Model



Top LLM Providers

#1



#5



#2



#6



#3

ANTHROPIC

#7



#4



#8





Amazon Bedrock

快速簡易的使用基礎模型來建立、
擴展生成式 AI 應用

透過單一 API 選擇領先的 FM
Choice of leading FMs via single API

客製化基礎模型
Model customization

快速實現RAG檢索增強生成架構
Retrieval Augmented Generation (RAG)

以Agent執行多步驟複雜任務
Agents that execute multistep tasks

資安、隱私與安全保護
Security, privacy, and safety

Azure AI

應用

 Microsoft 365

 Microsoft Dynamics 365

Partner Solutions



商業用戶

應用平台

AI 生成器



Power BI



Power Apps



Power Automate



Power Virtual Agents

基於場景的服務

應用 AI 服務



Bot Service



Cognitive Search



Form Recognizer



Video Indexer



Metrics Advisor



Immersive Reader

可訂製的 AI 模型

認知服務



Vision



Speech



Language



Decision



OpenAI Service

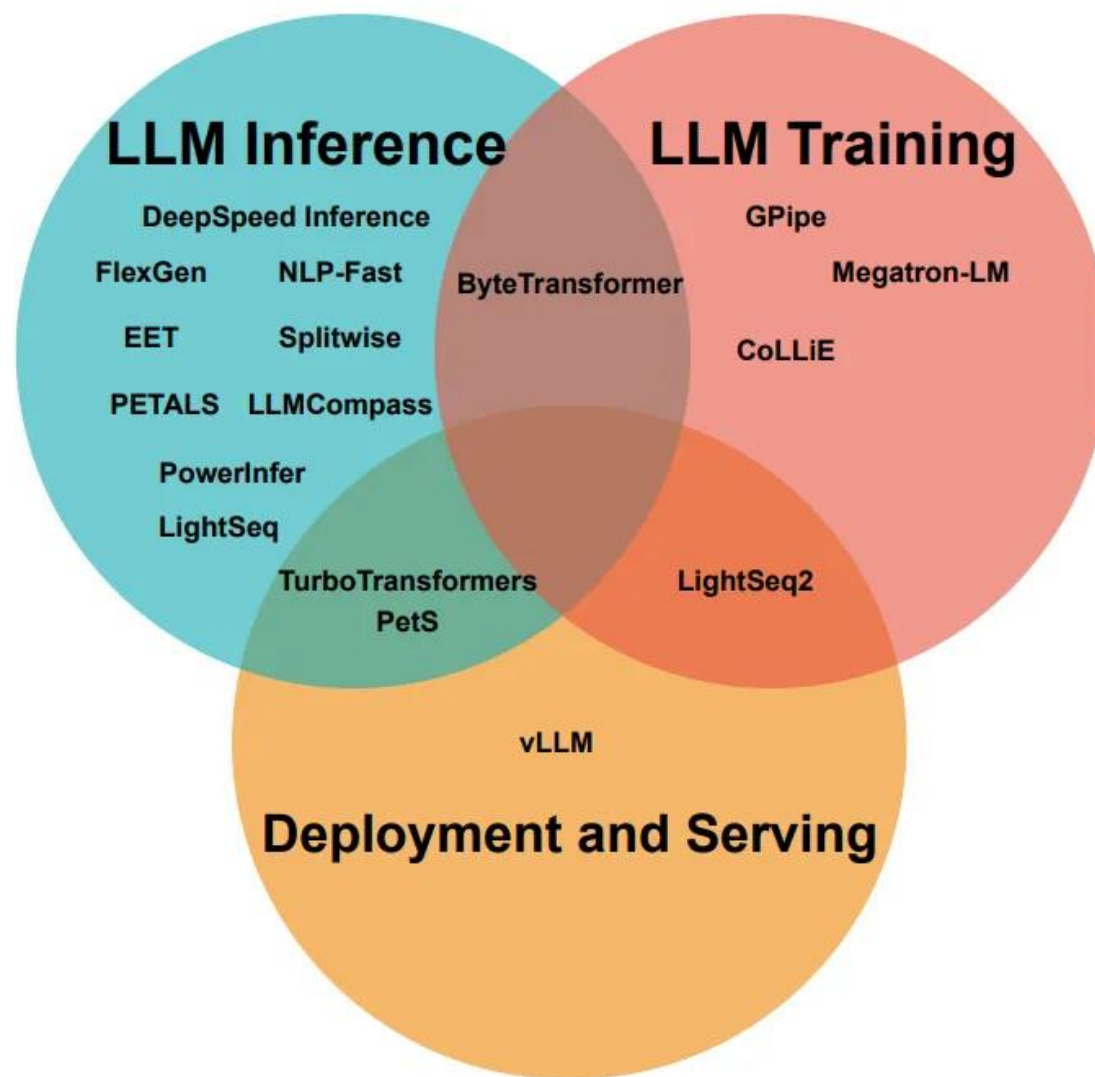


開發者和數據科學家

機器學習平台



Azure 機器學習



LLM Inferencing and Training On GCP



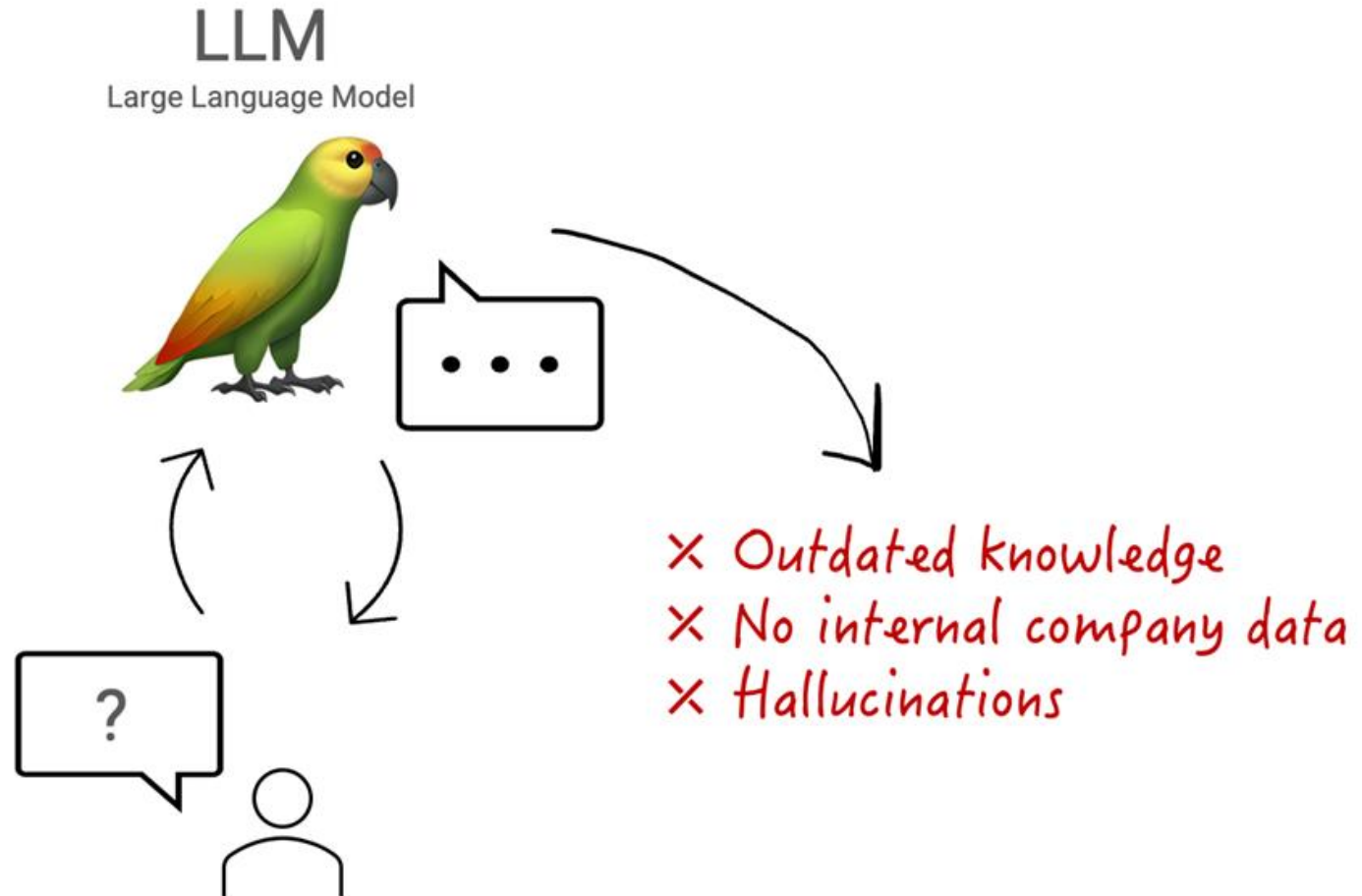
Google Cloud Next LLM **Gemini**

RAG: 檢索增強生成

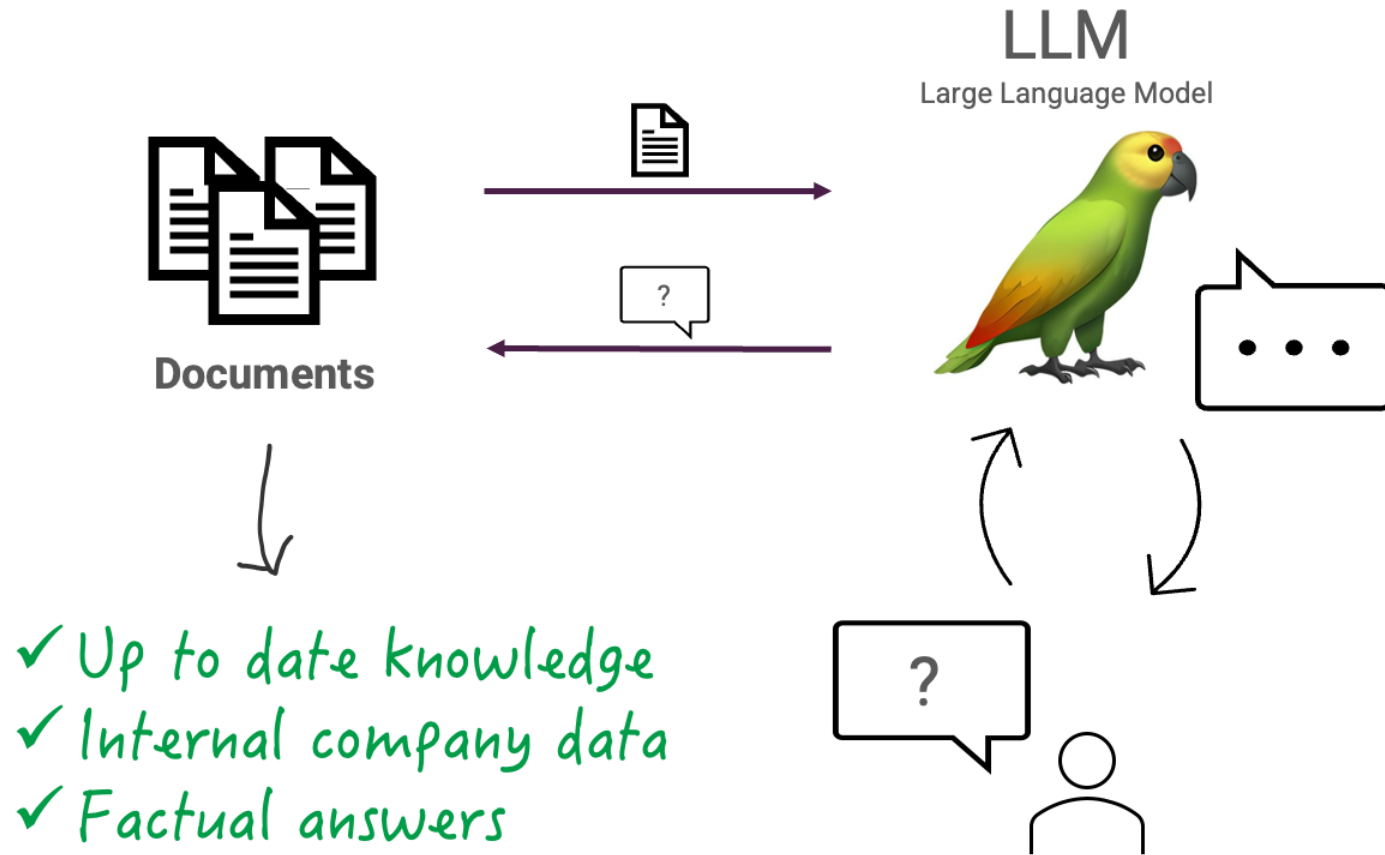
Retrieval Augmented Generation

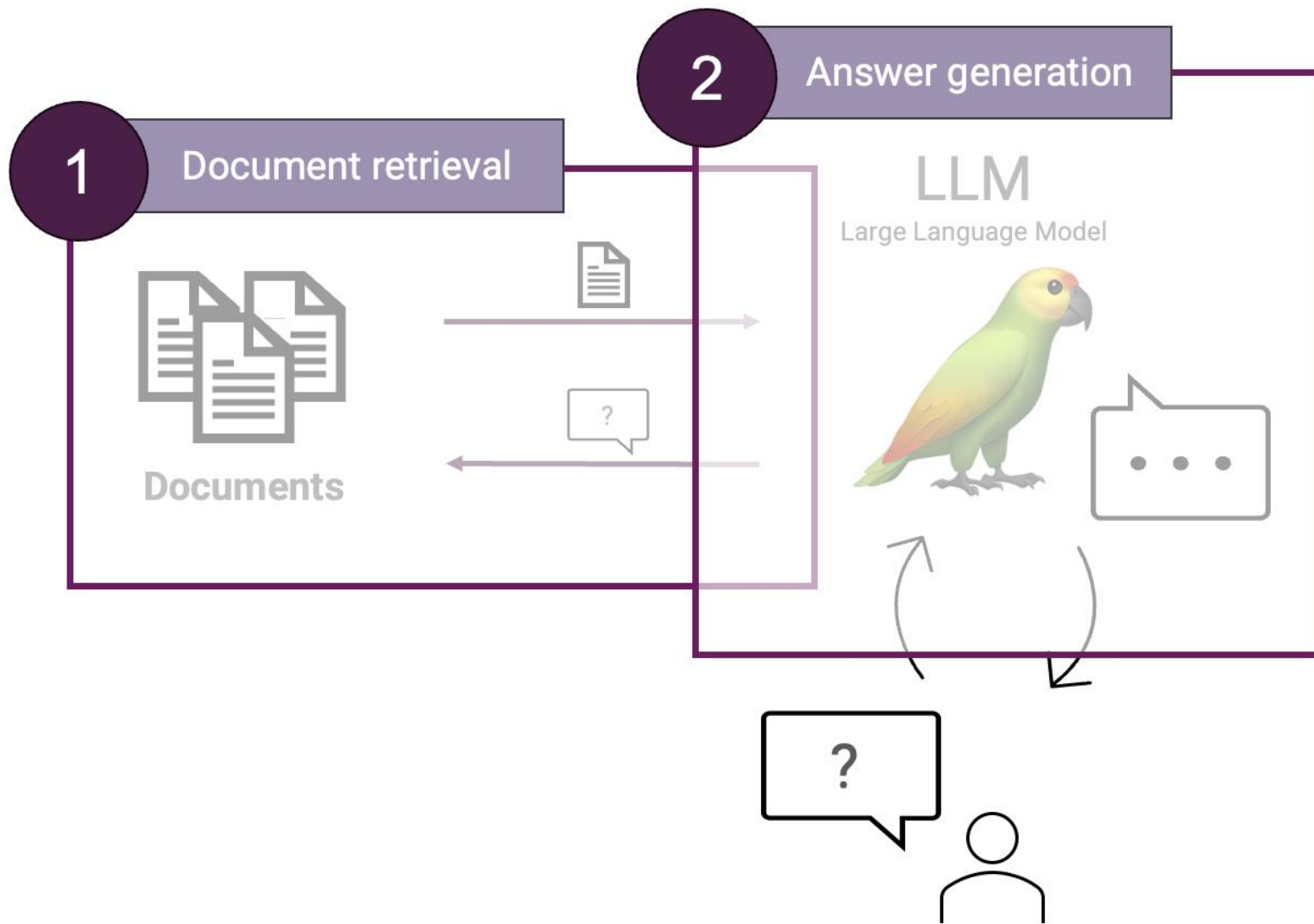


Model interactions without RAG



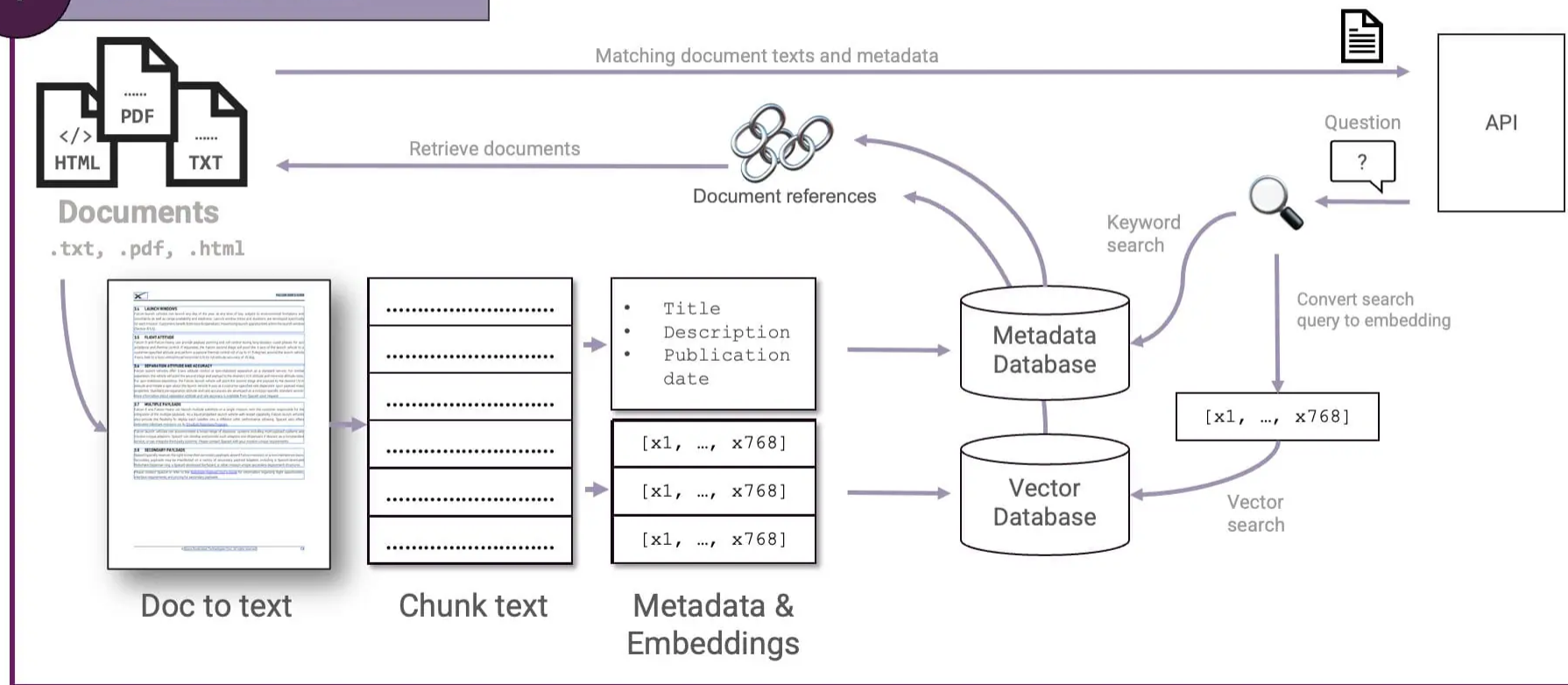
Model interactions with RAG





1

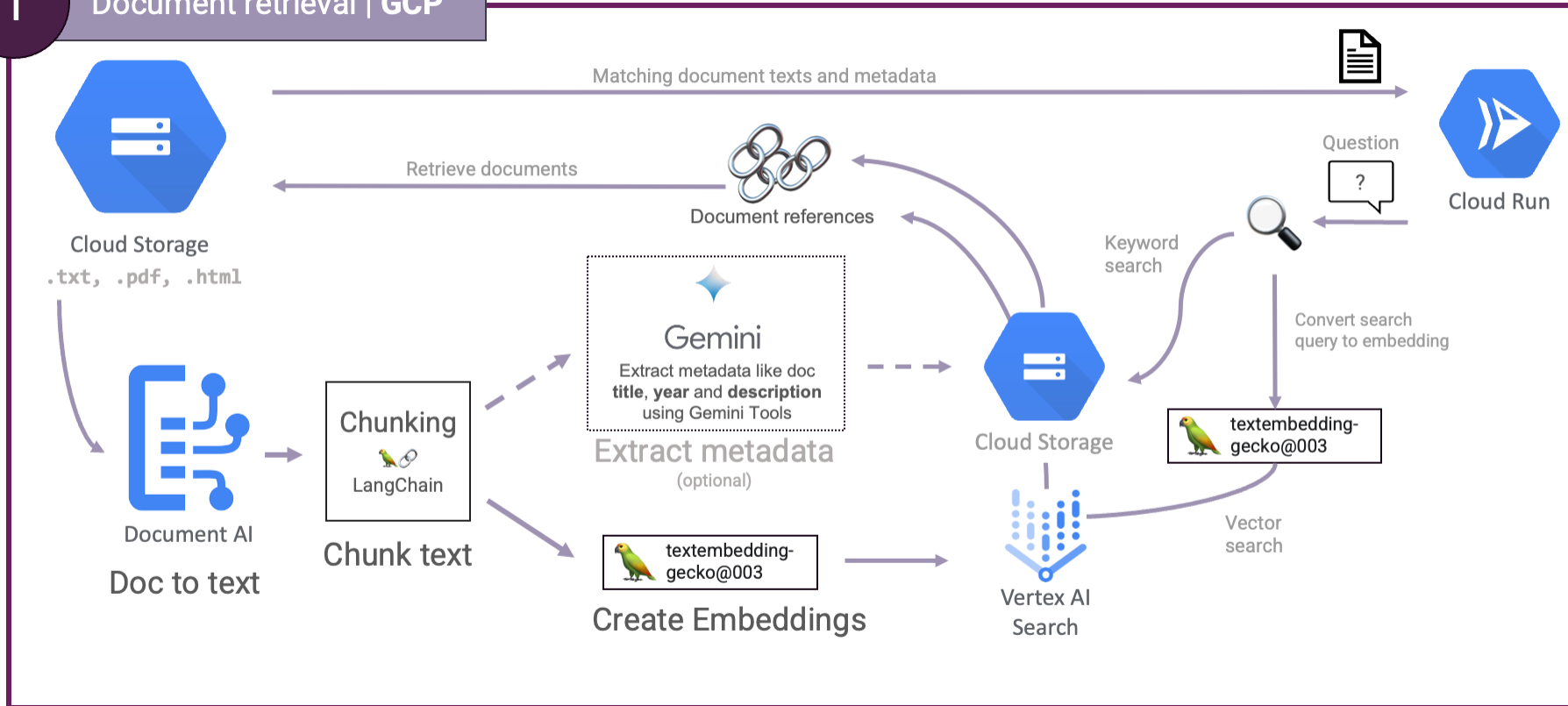
Document retrieval



Document retrieval step in a RAG system. Documents are **converted to text** and **converted to embeddings**. A user's question is **converted to an embedding** such that a **vector similarity search** can be performed.

1

Document retrieval | GCP



Document retrieval using GCP services including **Document AI**, **textembedding-gecko** and **Vertex AI Vector Search**.

2

Answer generation

```
{"""
```

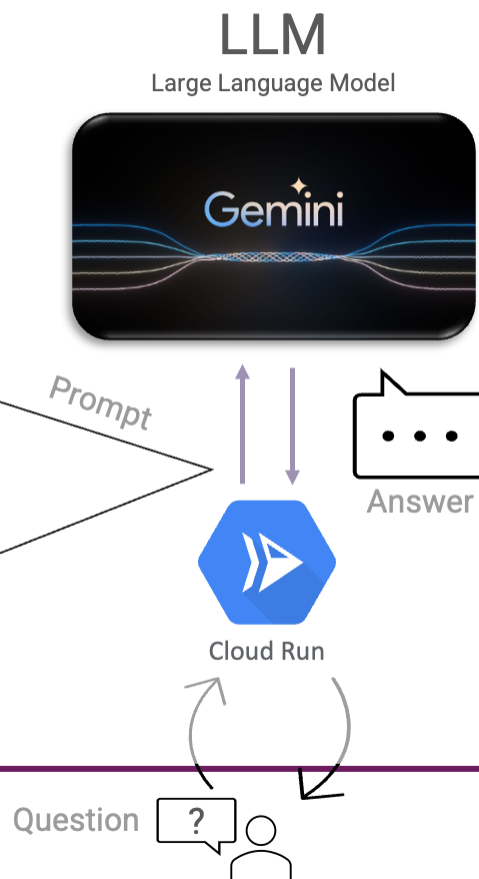
A question of a user will follow. Answer the question best to your ability, using the information provided in the documents. Only answer with truthful and citable information. The answer should be at the maximum five sentences.

Question: '{query}'

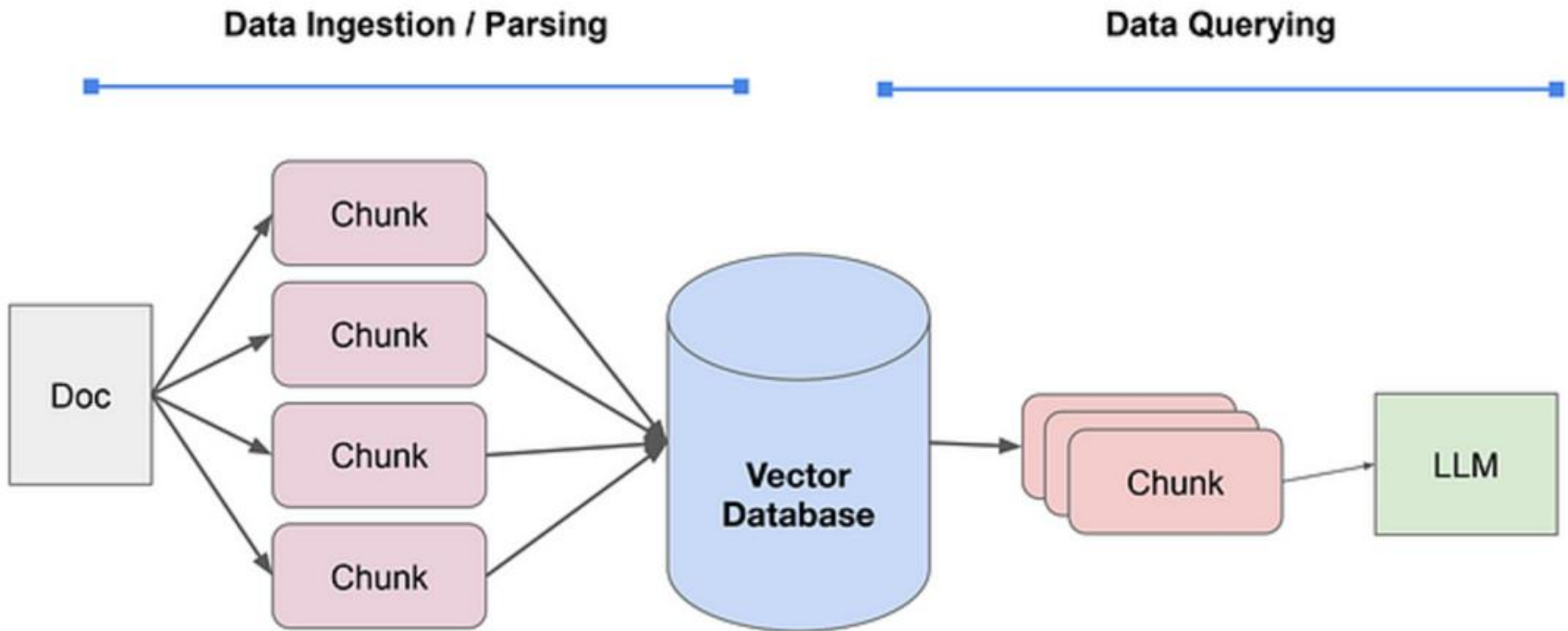
Documents:

```
{  
    format_doc(doc) for each doc in  
    docs  
}
```

```
"""
```



What if it is big data



商業用

RAG on GCP:

Fully managed



商業用

RAG on GCP:

Partly managed



商業用

RAG on GCP:

Full control

