Google I/O Extended 2024 Taipei

探索 Google Gemma 2 模型與應用

Simon Liu

2024/07/31

About me: 劉育維 / Simon Liu

Ocard

- Ocard 奧理科技 AI 工程師
- AI, ML, DL, LLM Architect and Engineer / Technical Writer / Speaker
- Try to use AI to do the application and help to solve the pain point by AI methods.
- My Personal Information:

Linkedin

Personal Website

Slide

Today's Topic

1. Google Gemma 系列模型介紹
2. 如何在 Google Colab 快速使用 Google Gemma 2 模型來進行測試
3. 搭配 Google Cloud 服務，讓 Google Gemma 結合 RAG 來進行服務應用
4. 結語

技術隨時在變化

請依照工具提供的官方資訊為主，我已盡力做好所有查核處理工作。

Part 1

Google Gemma 系列模型介紹

Gemma LLM 模型的自我介紹

Medium 介紹

- 第一次釋出日期: Feb 21, 2024
- 模型名稱：Gemma 系列模型
  - 此模型由 Google 官方所開源出來的 SOTA AI 模型
- 模型開源狀況 / License：
  - Gemma 目前採用 Google 所撰寫的 License 授權方式 — Gemma Terms of Use，統整相關內容並理解後，是一個可商用的模型。

都已經有了 Gemini 模型，為何要釋出 Gemma 系列模型？

- 社群回饋機制與促進研究與創新：
  - 讓研究人員和開發者能深入探索 AI 技術，加速創新應用。
- 降低使用門檻：
  - Gemma 模型更小、更輕量，讓資源有限的開發者或企業也能輕鬆使用 Google 的 AI 技術。
- 特定任務優化：
  - Gemma 模型針對特定任務進行優化，例如自然語言理解或生成，能提供更精準、高效的解決方案。

目前 Gemma 種類

Gemma v1 / v2

- 高效能且輕量化的大型語言模型
- 主要可做對話式大型語言模型
- 各項評估上表現優良

RecurrentGemma

- 支援研究人員進行大批次的高效推理，採用循環神經網路和局部注意力機制提升記憶效率。
- 基準測試上成績與Gemma 2B模型相當，但是RecurrentGemma使用的記憶體量更少

CodeGemma

- 針對開發人員和企業的程式碼完成、生成和聊天工具使用情境
- 多程式語言能力，主要以Python、JavaScript、Java等各種熱門程式語言的程式碼撰寫建議

PaliGemma

- 視覺語言開放模型，能夠針對影像字幕、視覺問答、理解圖像內文字、物件偵測、物件切割的應用案例提供最佳化

Gemma 2 目前的模型能力比較

The performance is between that of GPT-3.5 and GPT-4.

	Meta Llama 3 8B	Meta Llama 2 7B	OpenAI GPT-4	OpenAI GPT-3.5	Gemma PT 9B	Gemma PT 27B	Google Gemma 7B-it	Google Gemini 1.5 Pro	MistralAI Mistral 7B Instruct
Open Source / Close Source	Open Source	Open Source	Close Source	Close Source	Open Source	Open Source	Open Source	Close Source	Open Source
MMLU (5-shot)	68.4	34.1	86.4	70.0	71.3	75.2	53.3	81.9	58.4
GPQA (0-shot)	34.2	21.7	35.7 (0-shot CoT)	28.1 (0-shot CoT)			21.4	41.5 (0-shot CoT)	26.3
HumanEval (0-shot)	62.2	7.9	67.0	48.1	40.2 (pass@1)	51.8 (pass@1)	30.5	71.9	36.6
GSM-8K (8-shot, CoT)	79.6	25.7	92.0 (5-shot CoT)	57.1 (5-shot)	68.6 (5-shot, maj@1)	74.0 (5-shot, maj@1)	30.6	91.7	39.9
MATH (4-shot, CoT)	30.0	3.8	52.9	34.1	36.6	42.3	12.2	58.5	11.0

大小模型的分工任務

- 大模型:期待能夠分解任務，並且讓任務能夠被定義清楚。
- 小模型:完成指派的任務，並且任務能夠交付至大任務中。

-> 小型的企業組織就出現了。

未來生成式 AI 趨勢 by 簡立峰 (曾擔任Google台灣區董事總經理)

Tech Trends to Watch

- 模型網路化
  - 大模型化 (Gemini Ultra)
  - 小模型化 (Google Gemma)
  - 從單一到多模生態 (Multi-Modal - PaliGemma)
- 雙螢應用＋AI
  - Doc, Excel, PPT, Mail, Photo, Chat, Search, ... "Agent"
  - Application: Google Workspace + Gemini
- Edge AI 新機載
  - AI phone, AI PC ...
- 機器人再啟
  - LLM生數據／人機對話交替
  - 白領到藍領全面影響

Part 2

如何在 Google Colab
快速使用 Google Gemma 2 模型來進行測試

你有幾種方式使用 Google Gemma 系列模型

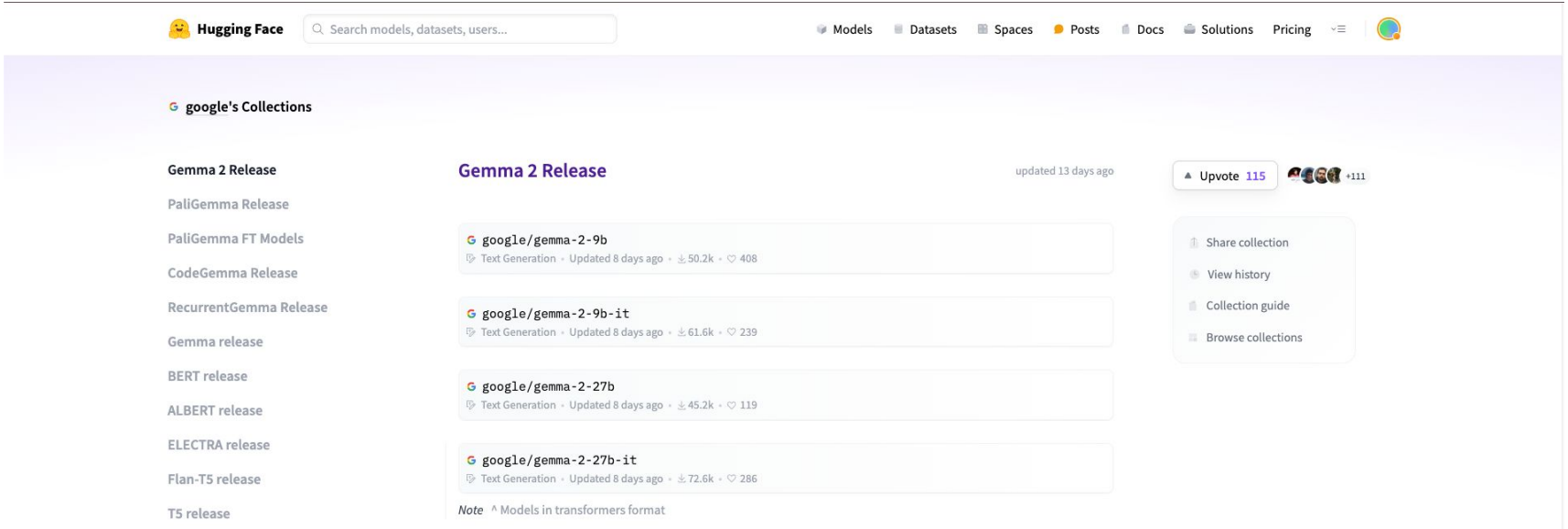
套件程式碼撰寫

寫程式碼，讓模型按照套件方式啟動

相關工具啟用

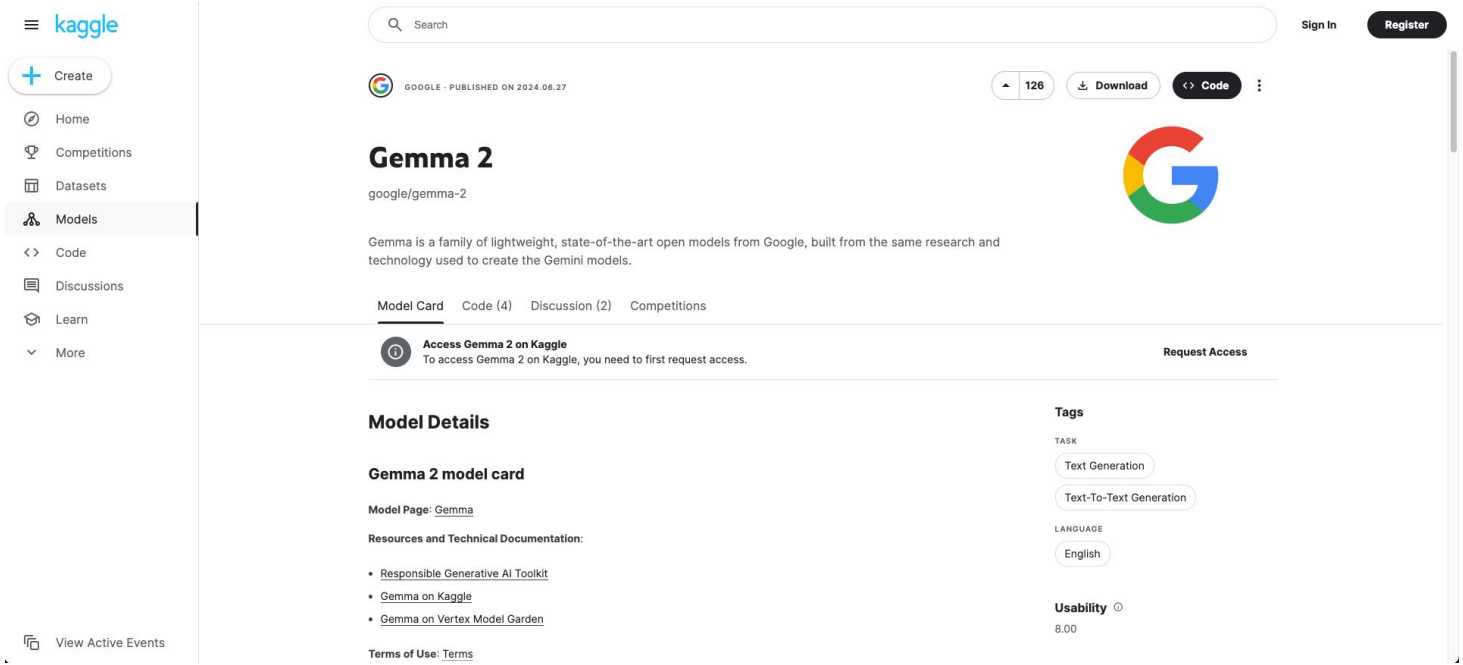
透過工具啟動服務，透過 API 來使用

Hugging Face - Gemma 2



https://huggingface.co/collections/google/gemma-2-release-667d6600fd5220e7b967f315

Kaggle - Gemma 2



https://www.kaggle.com/models/google/gemma-2

Ollama

- Ollama 是一個開源軟體，讓使用者可以在自己的硬體上運行、創建和分享大型語言模型服務。
- 這個平台適合在地端運行模型，因為它不僅可以保護隱私，還允許使用者透過命令行介面輕鬆地設置和互動。
- Ollama 支援非常多種模型，並提供彈性的客製化選項，例如從其他格式導入模型並設置參數。

Gemma 2 Model in Ollama

文件介紹

gemma2

Google Gemma 2 is now available in 2 sizes, 9B and 27B.

9B 27B

332.9K Pulls Updated yesterday

9b 63 Tags `ollama run gemma2`

Updated yesterday		ff02c3702f32 · 5.4GB
model	arch **gemma2** · parameters **9.24B** · quantization **Q4_0**	5.4GB
params	{"stop":["<start_of_turn>","<end_of_turn>"]}	65B
license	Gemma Terms of Use Last modified: February 21, 2024 By usi…	8.4kB
template	<start_of_turn>user {{ if .System }}{{ .System }} {{ end }…	136B

Two sizes: 9B and 27B parameters

- 9B Parameters `ollama run gemma2`
- 27B Parameters `ollama run gemma2:27b`

Ollama Python Library

The Ollama Python library provides the easiest way to integrate Python 3.8+ projects with Ollama.

Prerequisites

You need to have a local ollama server running to be able to continue. To do this:

- Download: https://ollama.com/
- Run an LLM: https://ollama.com/library
  - Example: `ollama run llama2`
  - Example: `ollama run llama2:70b`

Then:

```
curl https://ollama.ai/install.sh | sh
ollama serve
```

Next you can go ahead with `ollama-python`.

Install

```
pip install ollama
```

Usage

```
import ollama
response = ollama.chat(model='llama3', messages=[
  {
    'role': 'user',
    'content': 'Why is the sky blue?',
  },
])
print(response['message']['content'])
```

Let’s demo it!

https://tinyurl.com/gemma-example

Part 3

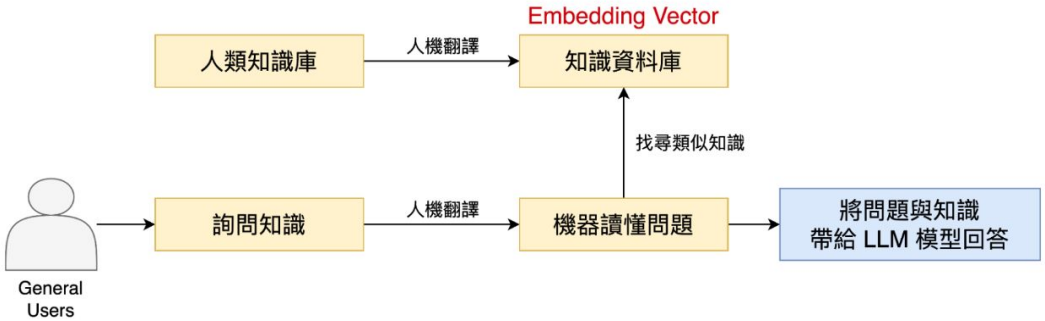
搭配 Google Cloud 服務，讓 Google Gemma 結合 RAG 來進行服務應用

快速介紹：什麼是 RAG / Fine-Tune？

RAG 介紹

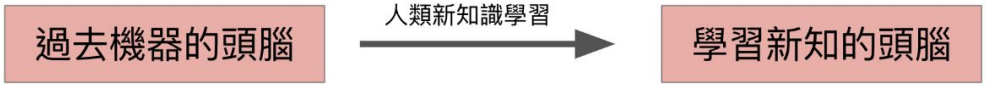
RAG (Embedding)

當人類詢問問題時，找尋說明書了解知識後，再回覆問題。

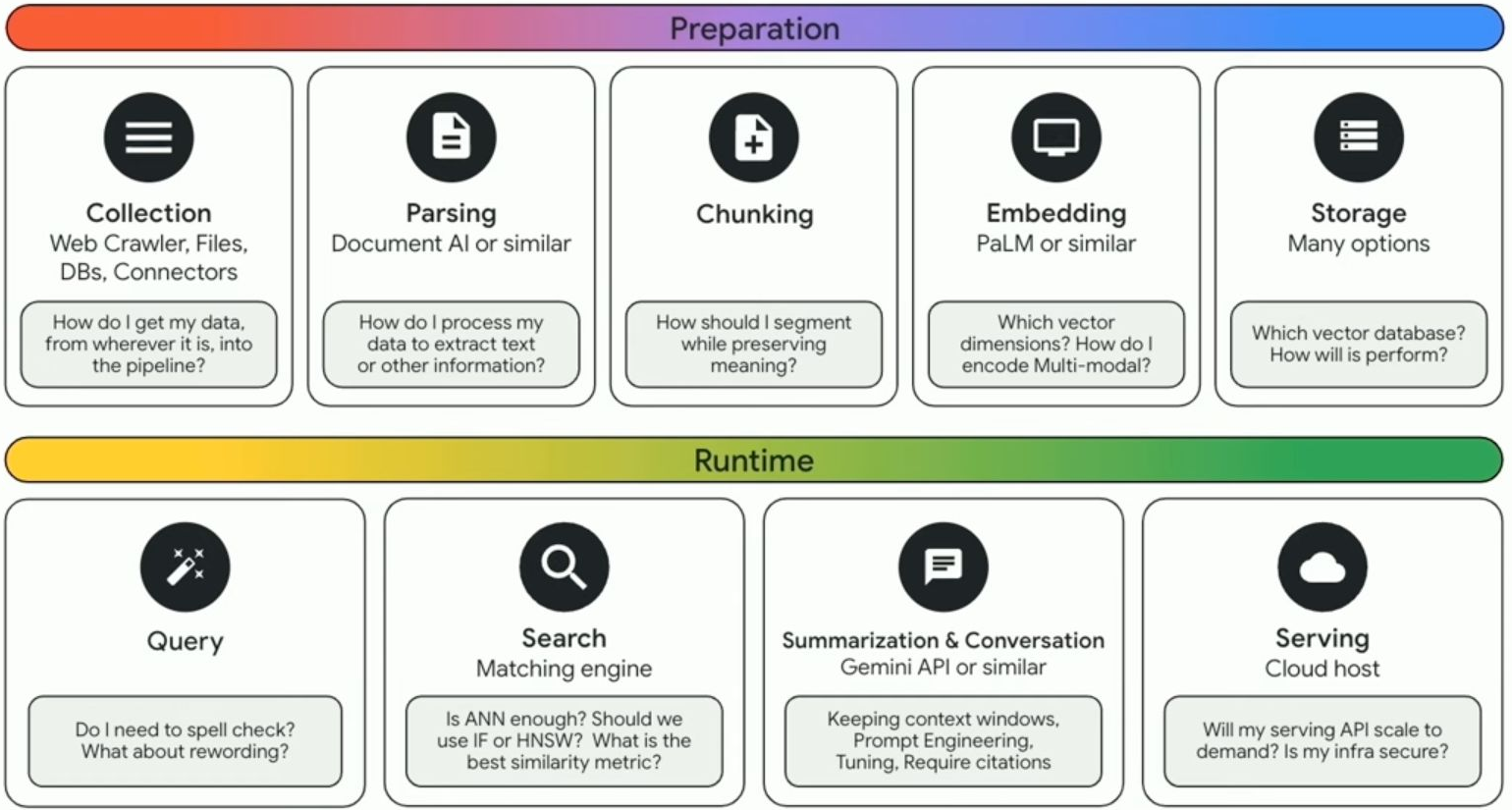


Fine-Tune Model

類似小孩子學習新知的概念，經過學習，就能夠得到新知。



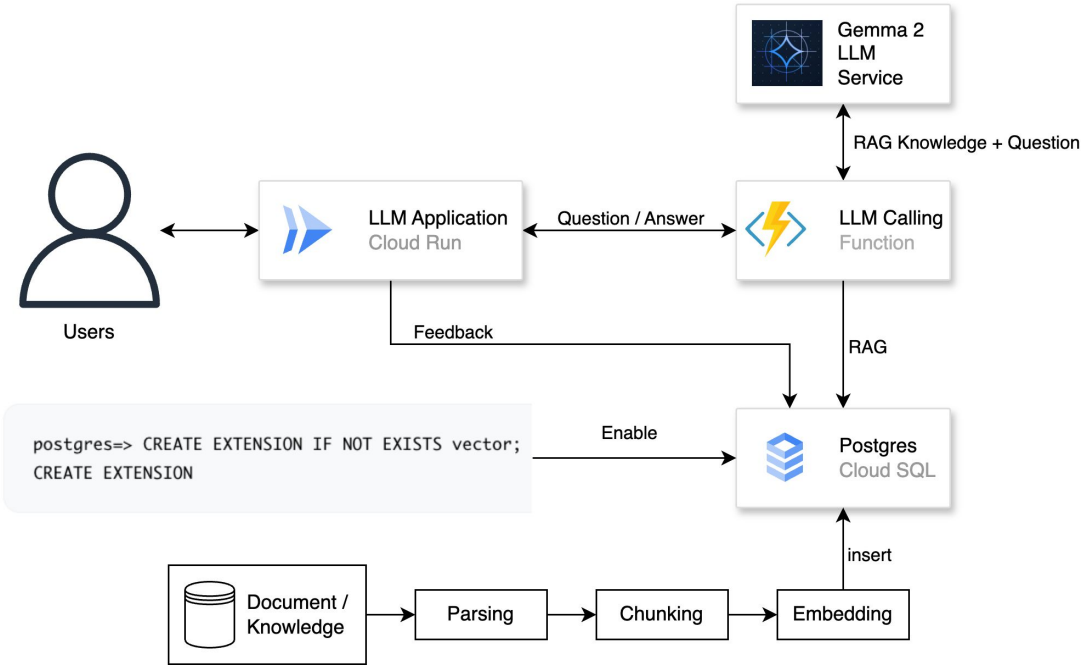
DIY RAG search



Source: Google IO Youtube Video - https://www.youtube.com/watch?v=v4s5eU2tfd4

Pgvector in Google Cloud

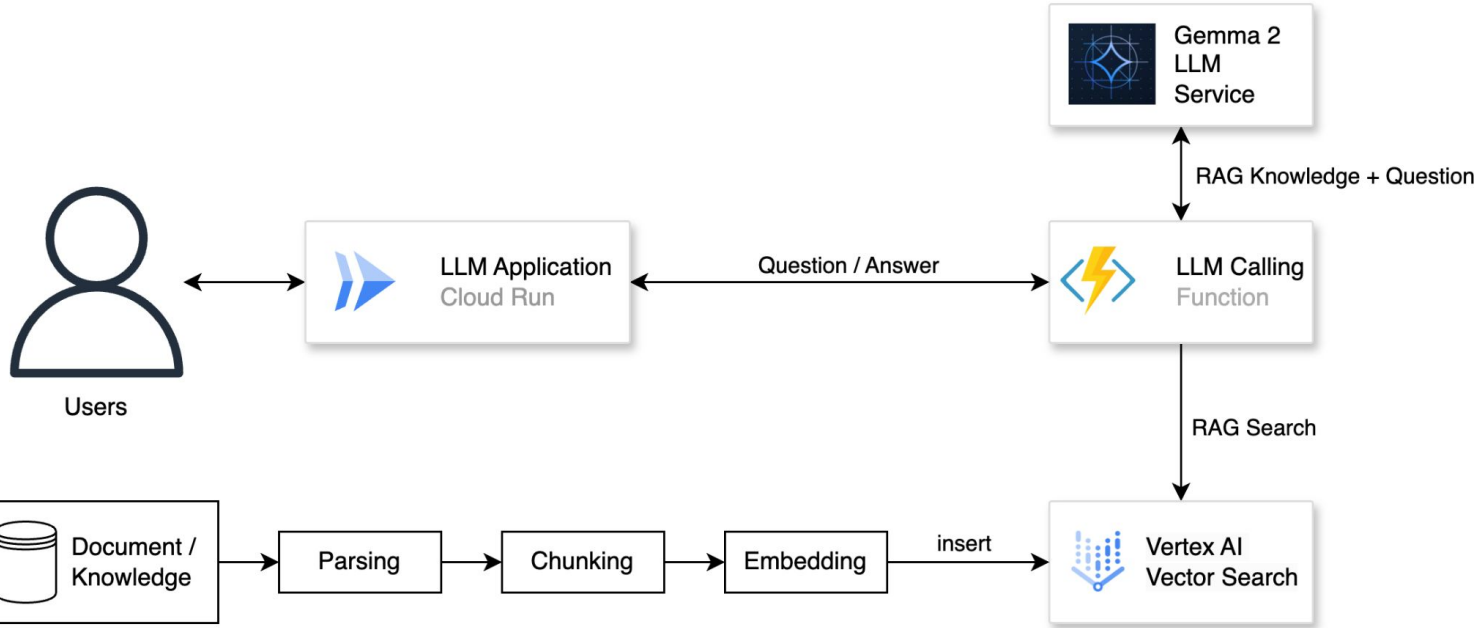
Open-source vector similarity search for Postgres Database



Pgvector Github Project: https://github.com/pgvector/pgvector
Blog - Building AI-powered apps on Google Cloud databases using pgvector, LLMs and LangChain: https://cloud.google.com/blog/products/databases/using-pgvector-llms-and-langchain-with-google-cloud-databases

Google Cloud Vertex AI Vector Search

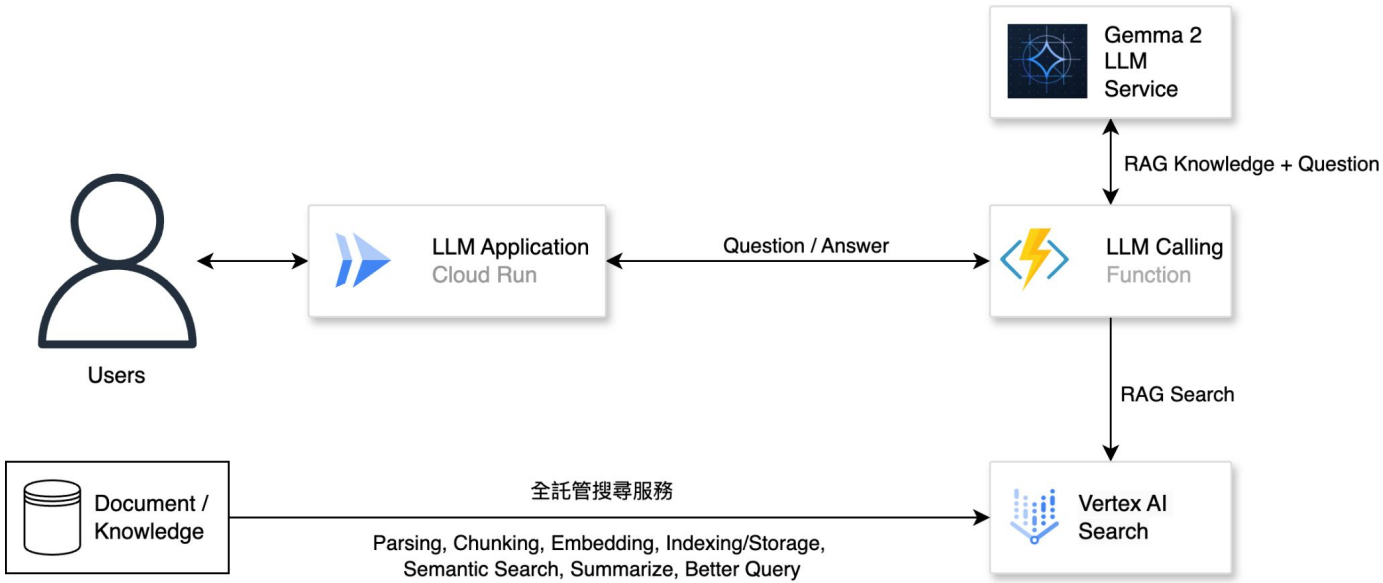
Google Cloud 的 Embedding Vector Database 服務



Google Cloud Official Doc: https://cloud.google.com/vertex-ai/docs/vector-search/overview

Google Cloud Vertex AI Search

Google Cloud 全託管式的知識輔助搜尋引擎服務



Google Cloud Official Doc: https://cloud.google.com/enterprise-search?hl=zh-TW

Fine-Tuning Google Gemma 2 Model

Google Cloud Official Blog: https://developers.googleblog.com/en/fine-tuning-gemma-2-with-keras-hugging-face-update/
Google Colab for fine-tuning keras Gemma 2 9B: https://colab.research.google.com/drive/1bEivD7tSwbFMtPbNEcZxilZvN_dGzgbU?usp=sharing

到底什麼時候 Fine-Tune？

以法律相關背景做舉例：

法律判例機器人	VS	法律條文機器人
透過過去判決經驗，來判斷是否可以推論可能結果		透過法律條文，來了解此法律案件違反哪條規定
⬇		⬇
根據過往案例來判斷		須先理解和讀懂法律條文
⬇		⬇
RAG 讓判例能夠快速準確被檢索出來，且經常會被更新		先 **Fine-Tune** 讀懂條文，再來幫助其他人綜合判斷

Part 4

結語

結論

- Gemma 2 以超越 GPT-3.5，接近 GPT-4 的能力下，提供相關 AI 模型生成能力來完成需求。
- 透過工具來啟動 Gemma 服務
  - 第三方 API 工具：Ollama / Llama-cpp / VLLM 等
  - Python 套件：Tensorflow / Pytorch / KerasNLP / Transformer
- 搭配 Google Cloud 服務，來使用 RAG 等方式，讓產品化能夠做的更好
  - Pgvector Database in Google Cloud SQL
  - Google Cloud Vertex AI Vector Search / Vertex AI Search

Thanks for Listening!

Simon Liu

2024/07/31