自然語言處理與應用

Natural Language Processing and Applications

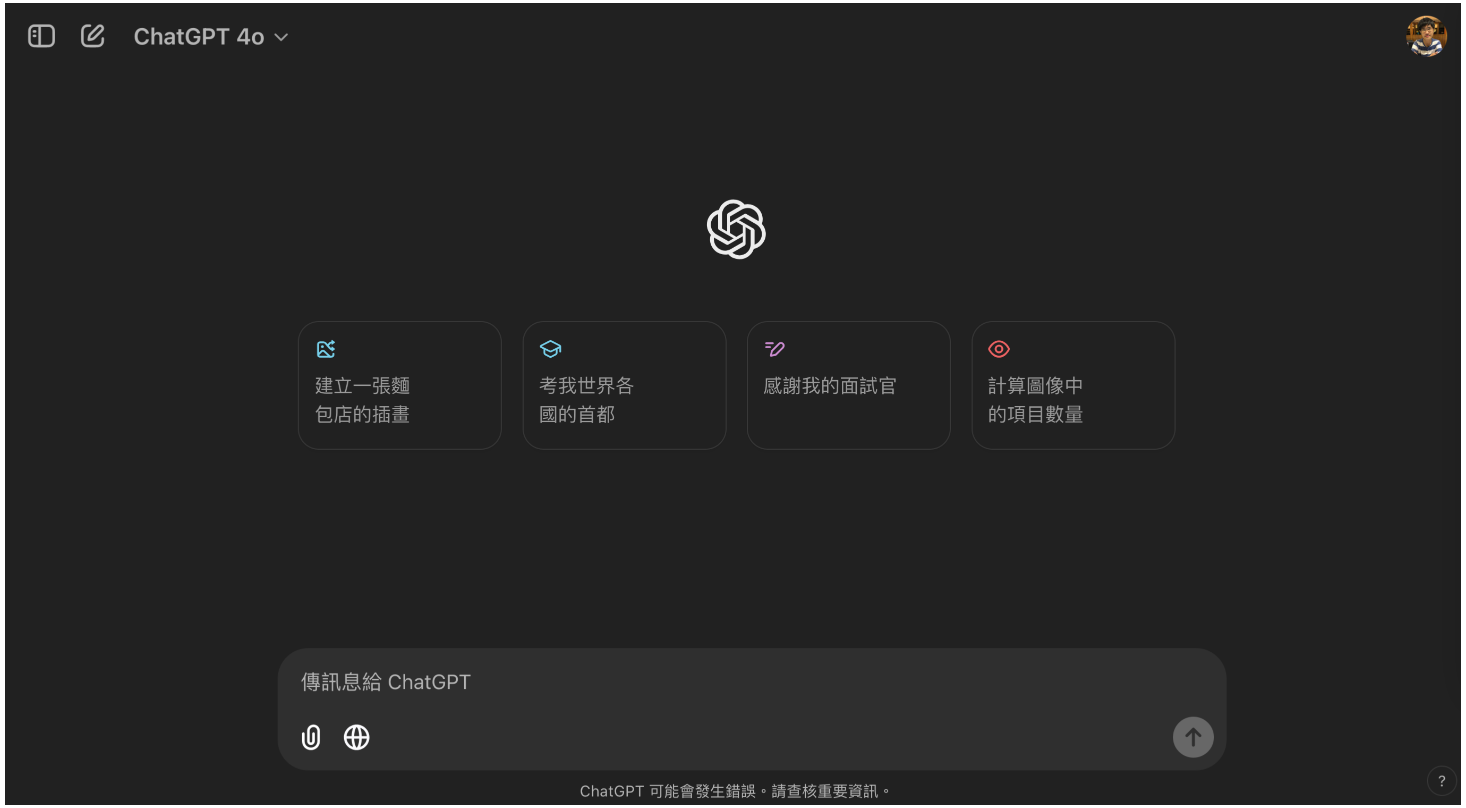
LLM API Tutorial

Instructor: 林英嘉 (Ying-Jia Lin)
2025/05/12

Course GitHub

Slido # NLP_0512

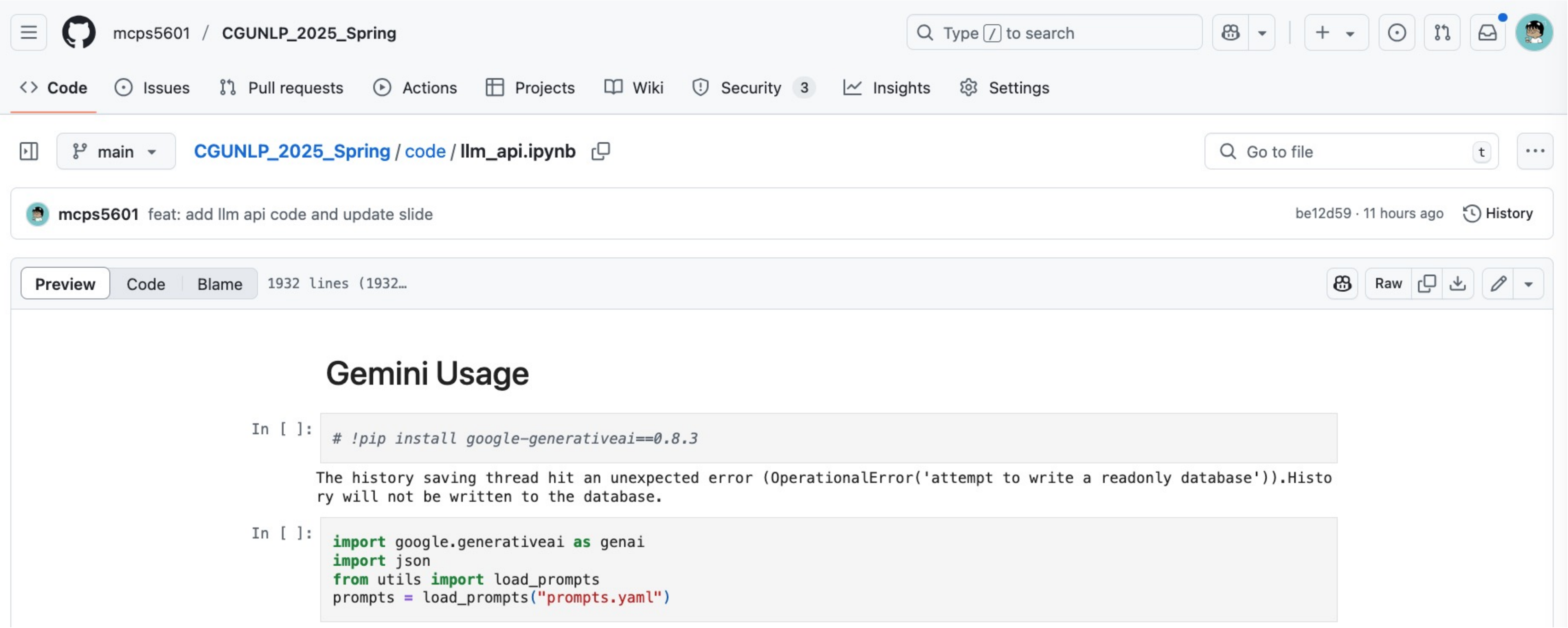
Why do we need to use API?



ChatGPT 可能會發生錯誤。請查核重要資訊。

- It’s slow if you want to use ChatGPT for NLP tasks by manually copying and pasting from the webpage.
- If you use ChatGPT from the webpage to test data, “Too many requests in 1 hour. Try again later.” may occur.

Code Example

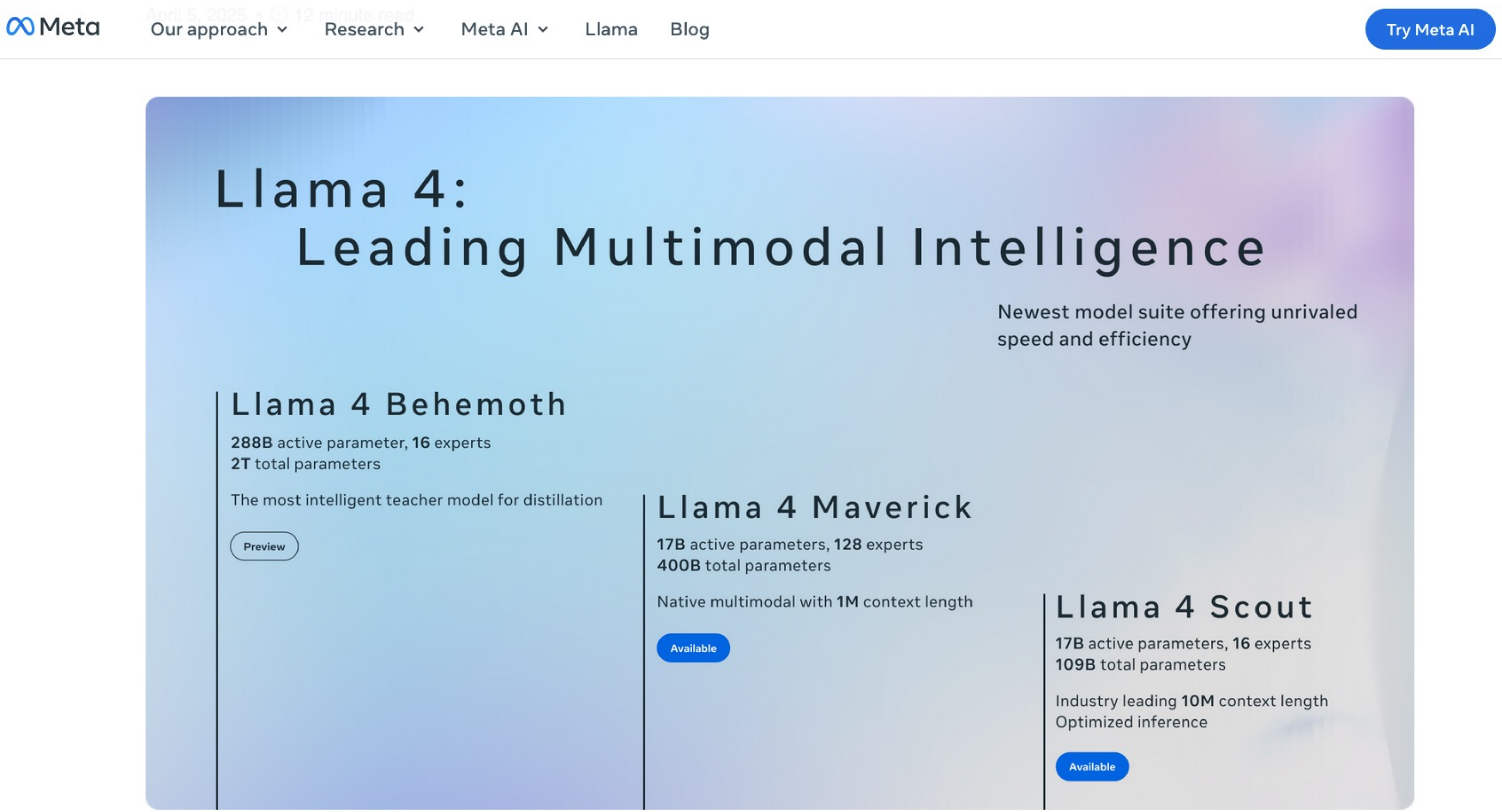


Gemini Usage

```
In [ ]: # !pip install google-generativeai==0.8.3
```

```
The history saving thread hit an unexpected error (OperationalError('attempt to write a readonly database')).Histo
ry will not be written to the database.
```

```
In [ ]: import google.generativeai as genai
        import json
        from utils import load_prompts
        prompts = load_prompts("prompts.yaml")
```



Llama 4: Leading Multimodal Intelligence

Newest model suite offering unrivaled speed and efficiency

Llama 4 Behemoth

288B active parameter, **16** experts
2T total parameters

The most intelligent teacher model for distillation

Llama 4 Maverick

17B active parameters, **128** experts
400B total parameters

Native multimodal with **1M** context length

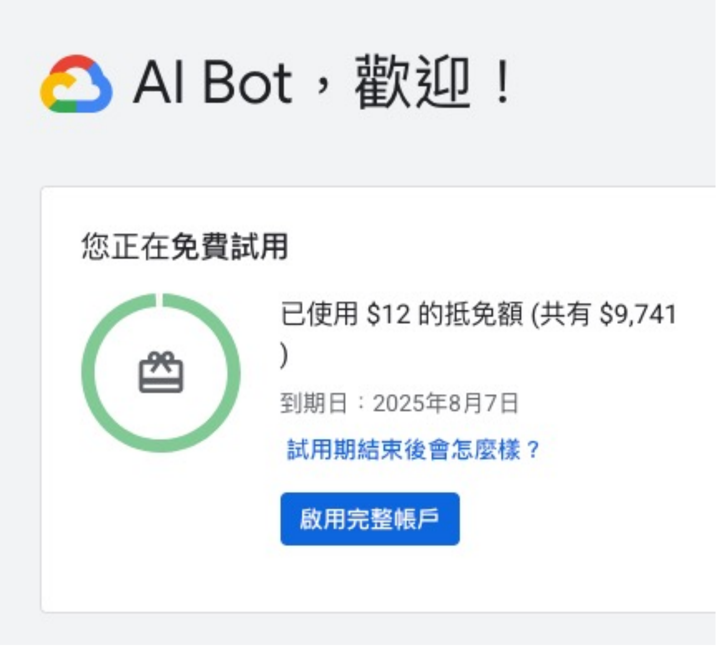
Llama 4 Scout

17B active parameters, **16** experts
109B total parameters

Industry leading **10M** context length
Optimized inference

註冊 Google Vertex

https://cloud.google.com/vertex-ai?hl=zh-TW



AI Bot，歡迎！

您正在免費試用

已使用 $12 的抵免額 (共有 $9,741)

到期日：2025年8月7日

進到控制台

https://console.cloud.google.com

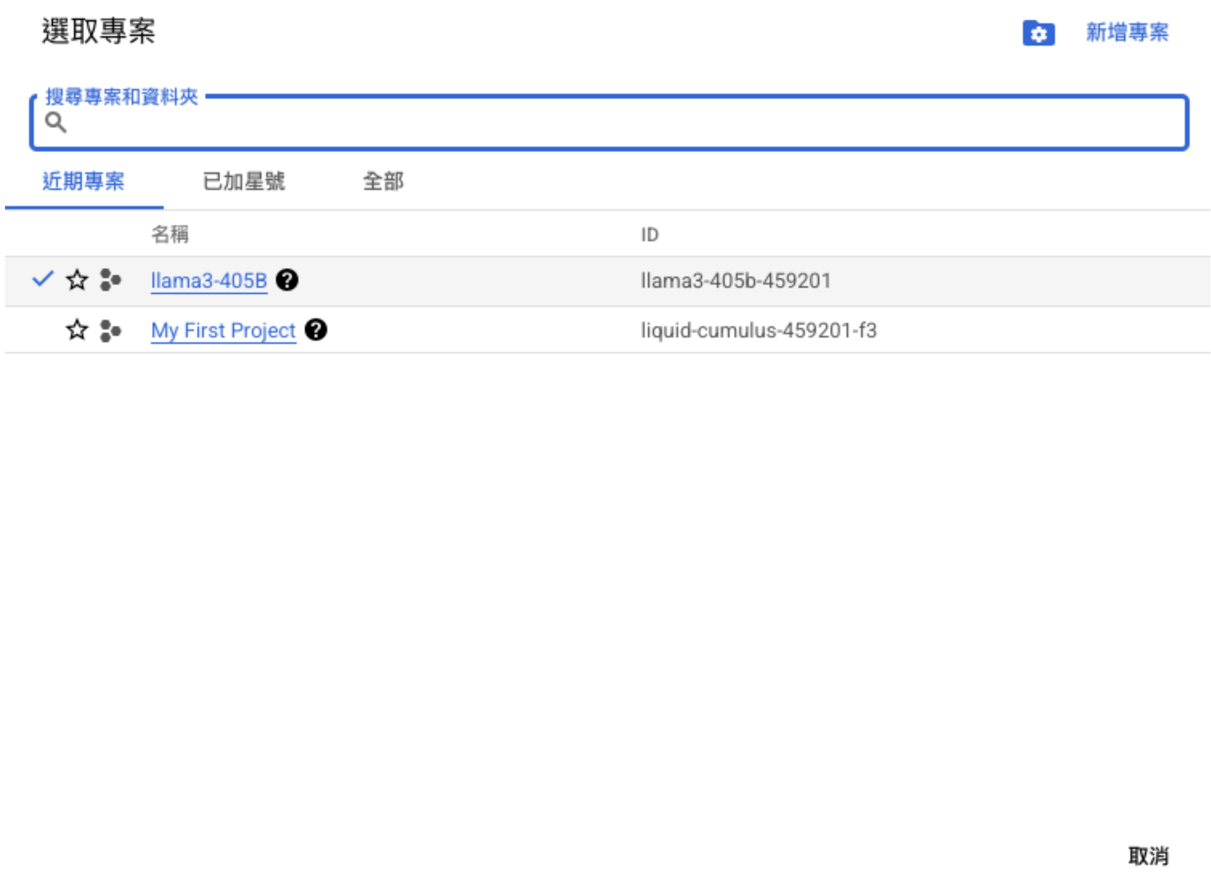
右邊導覽選單

Vertex AI

Model Garden

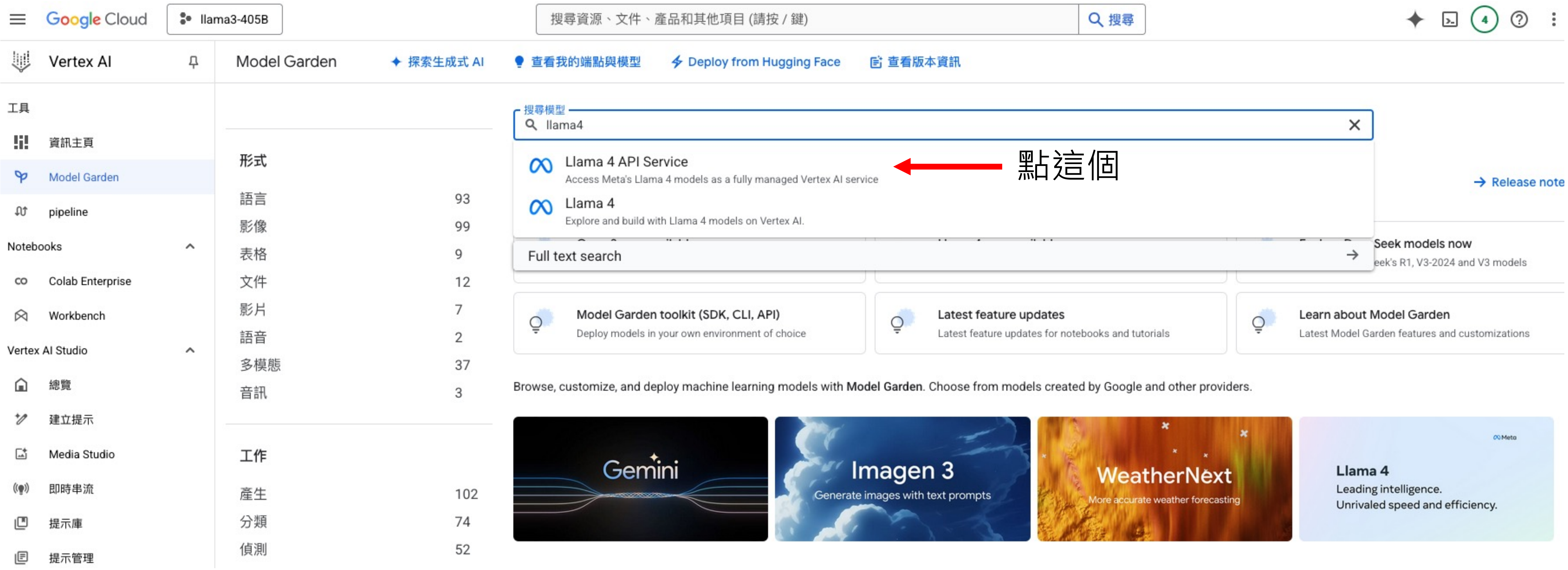
選取專案

ID 等等會用到



	名稱	ID
✓ ☆	llama3-405B	llama3-405b-459201
☆	My First Project	liquid-cumulus-459201-f3

搜尋llama4



點這個

Google Cloud　llama3-405B　搜尋資源、文件、產品和其他項目 (請按 / 鍵)　搜尋

Vertex AI　Model Garden　探索生成式 AI　查看我的端點與模型　Deploy from Hugging Face　查看版本資訊

工具
- 資訊主頁
- Model Garden
- pipeline

Notebooks
- Colab Enterprise
- Workbench

Vertex AI Studio
- 總覽
- 建立提示
- Media Studio
- 即時串流
- 提示庫
- 提示管理

形式

語言	93
影像	99
表格	9
文件	12
影片	7
語音	2
多模態	37
音訊	3

工作

產生	102
分類	74
偵測	52

搜尋模型
llama4

Llama 4 API Service
Access Meta's Llama 4 models as a fully managed Vertex AI service

Llama 4
Explore and build with Llama 4 models on Vertex AI.

Full text search →

→ Release note

Seek models now
eek's R1, V3-2024 and V3 models

Model Garden toolkit (SDK, CLI, API)
Deploy models in your own environment of choice

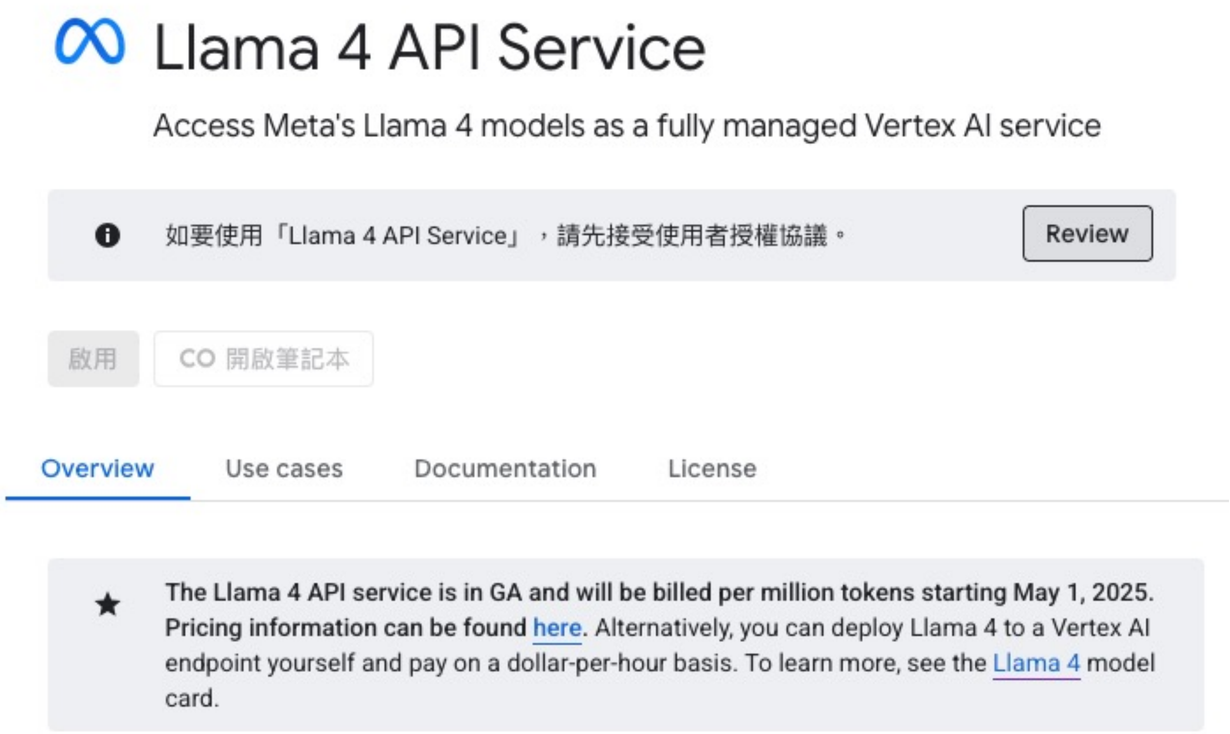
Latest feature updates
Latest feature updates for notebooks and tutorials

Learn about Model Garden
Latest Model Garden features and customizations

Browse, customize, and deploy machine learning models with **Model Garden**. Choose from models created by Google and other providers.

Gemini
Imagen 3
Generate images with text prompts
WeatherNext
More accurate weather forecasting
Meta
Llama 4
Leading intelligence.
Unrivaled speed and efficiency.

閱讀 EULA 後點啟用



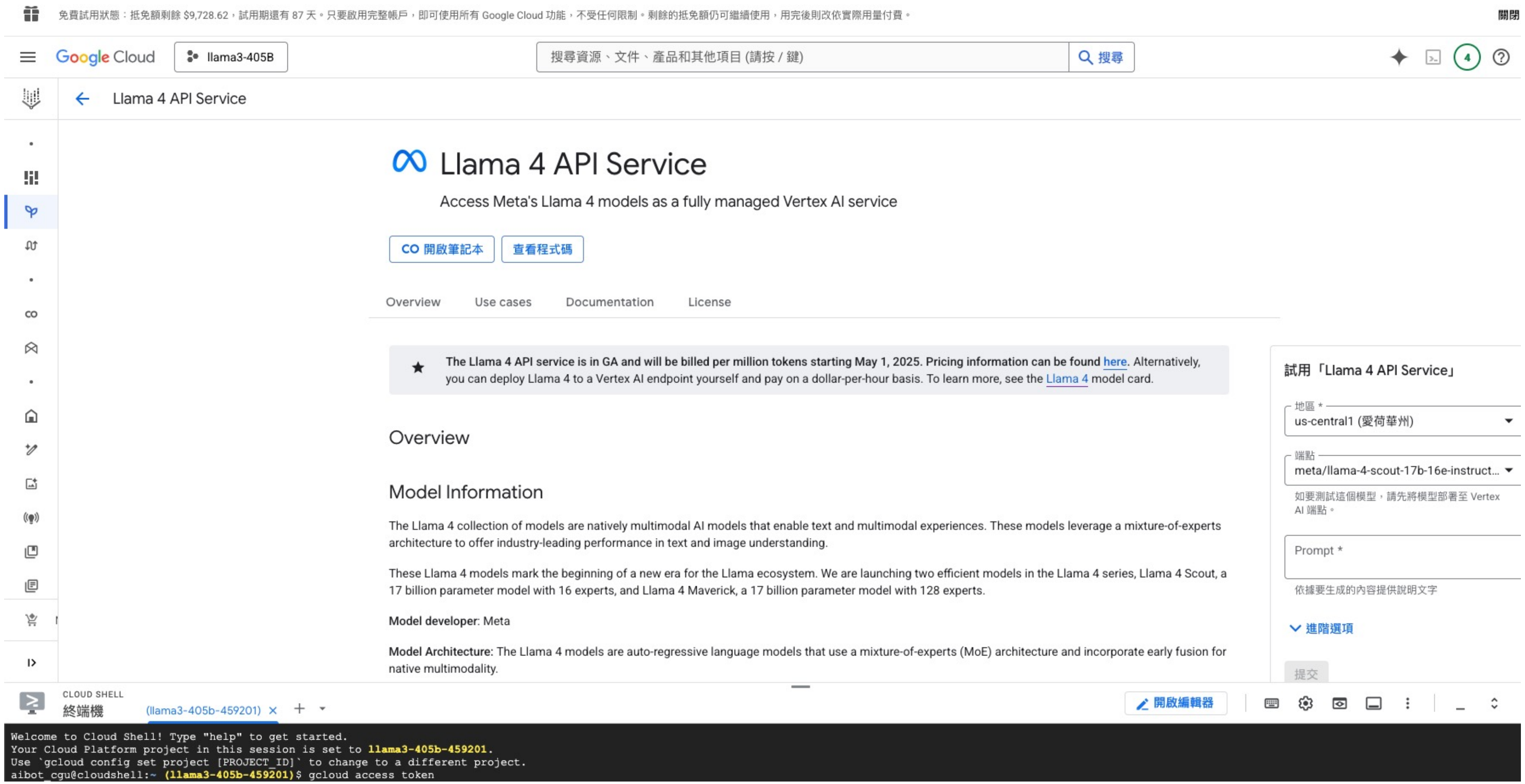
Llama 4 API Service

Access Meta's Llama 4 models as a fully managed Vertex AI service

如要使用「Llama 4 API Service」，請先接受使用者授權協議。

★ **The Llama 4 API service is in GA and will be billed per million tokens starting May 1, 2025. Pricing information can be found here**. Alternatively, you can deploy Llama 4 to a Vertex AI endpoint yourself and pay on a dollar-per-hour basis. To learn more, see the Llama 4 model card.

打指令在終端機：gcloud access token



免費試用狀態：抵免額剩餘 $9,728.62，試用期還有 87 天。只要啟用完整帳戶，即可使用所有 Google Cloud 功能，不受任何限制。剩餘的抵免額仍可繼續使用，用完後則改依實際用量付費。
關閉
Google Cloud
llama3-405B
搜尋資源、文件、產品和其他項目 (請按 / 鍵)
搜尋
4
Llama 4 API Service
CO 開啟筆記本
查看程式碼
Overview
Use cases
Documentation
License
CLOUD SHELL
終端機
(llama3-405b-459201)
開啟編輯器

Llama 4 API Service

Access Meta's Llama 4 models as a fully managed Vertex AI service

★ The Llama 4 API service is in GA and will be billed per million tokens starting May 1, 2025. Pricing information can be found here. Alternatively, you can deploy Llama 4 to a Vertex AI endpoint yourself and pay on a dollar-per-hour basis. To learn more, see the Llama 4 model card.

Overview

Model Information

The Llama 4 collection of models are natively multimodal AI models that enable text and multimodal experiences. These models leverage a mixture-of-experts architecture to offer industry-leading performance in text and image understanding.

These Llama 4 models mark the beginning of a new era for the Llama ecosystem. We are launching two efficient models in the Llama 4 series, Llama 4 Scout, a 17 billion parameter model with 16 experts, and Llama 4 Maverick, a 17 billion parameter model with 128 experts.

Model developer: Meta

Model Architecture: The Llama 4 models are auto-regressive language models that use a mixture-of-experts (MoE) architecture and incorporate early fusion for native multimodality.

試用「Llama 4 API Service」

地區 *: us-central1 (愛荷華州)

端點: meta/llama-4-scout-17b-16e-instruct...

如要測試這個模型，請先將模型部署至 Vertex AI 端點。

Prompt *

依據要生成的內容提供說明文字

進階選項

提交

```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to llama3-405b-459201.
Use `gcloud config set project [PROJECT_ID]` to change to a different project.
aibot_cgu@cloudshell:~ (llama3-405b-459201)$ gcloud access token
```

把 API Key 複製起來！

Thank you!

Instructor: 林英嘉

✉ yjlin@cgu.edu.tw

TA: 吳宣毅

✉ m1161007@cgu.edu.tw