



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

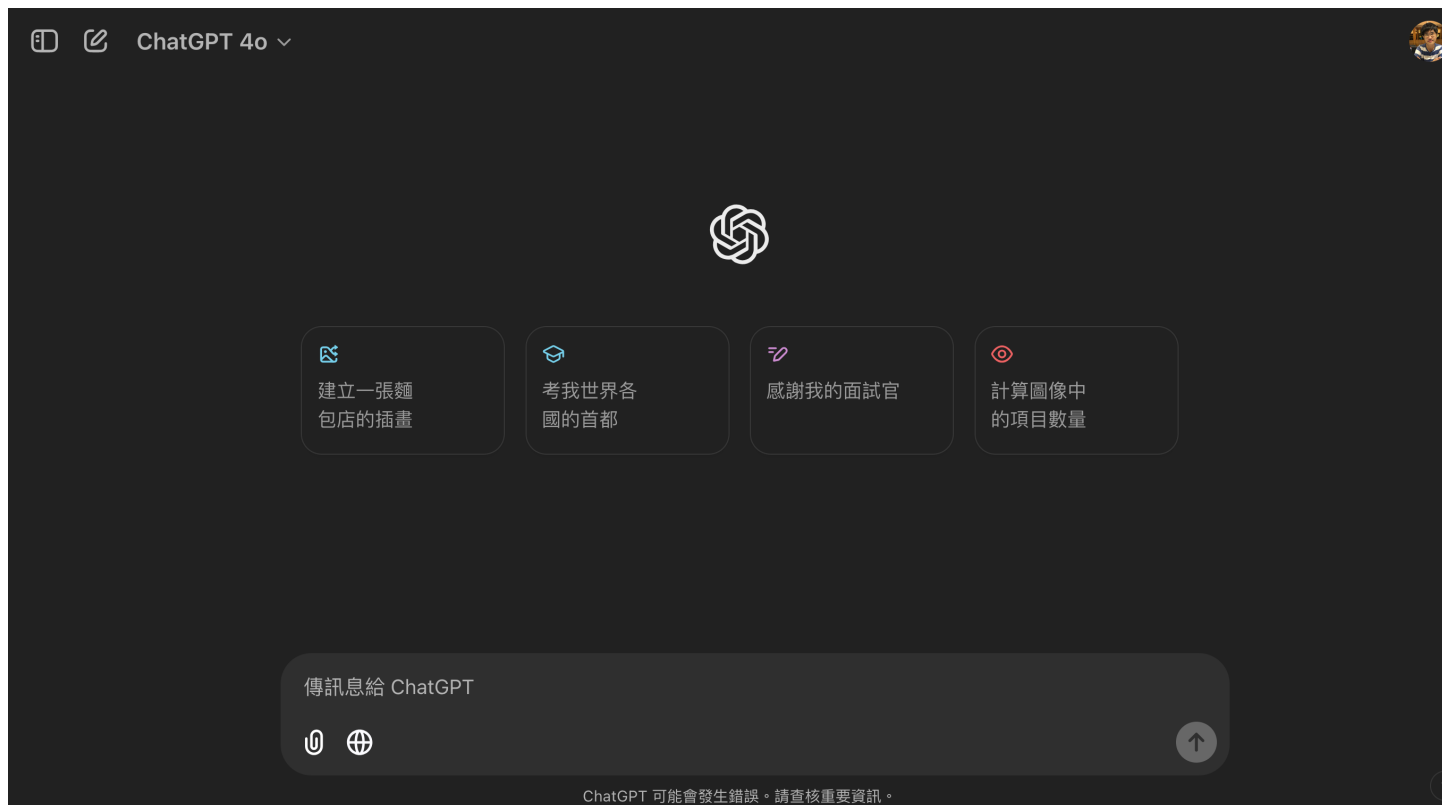


[Course GitHub](#)



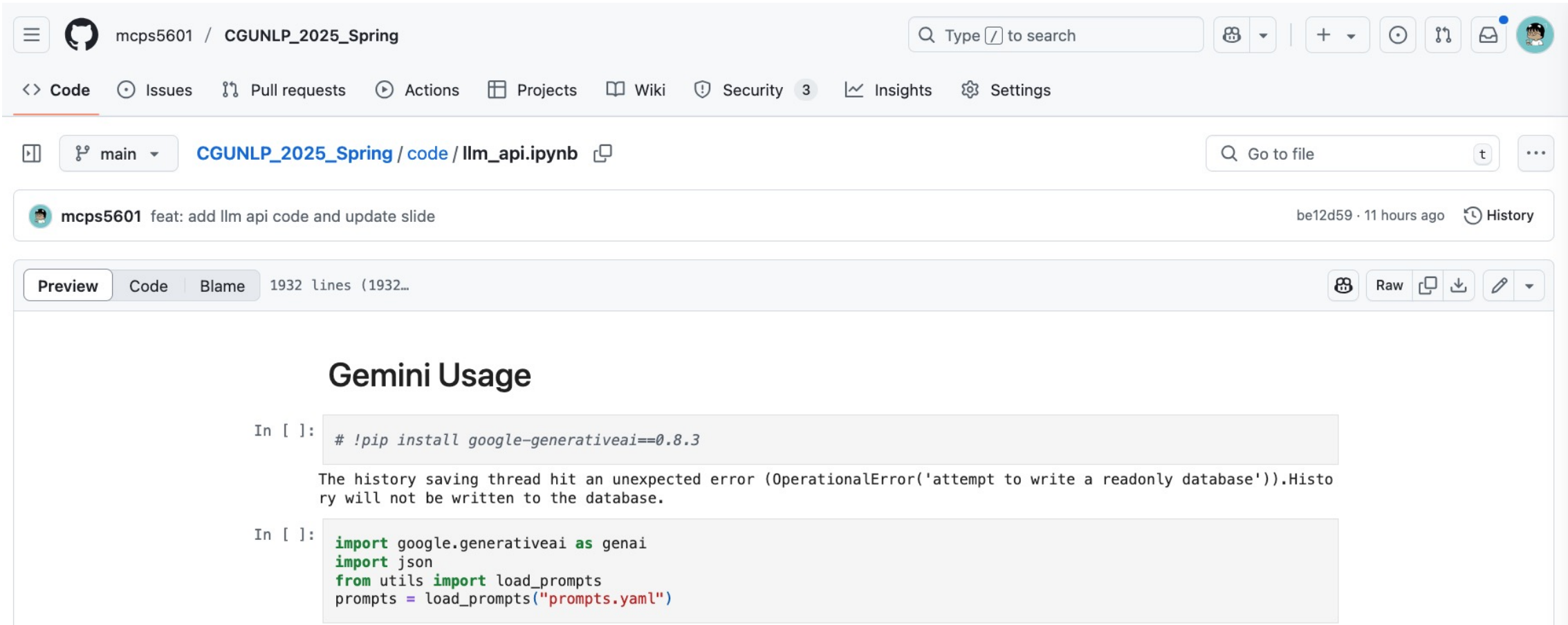
[Slido # NLP_0512](#)

Why do we need to use API?



- 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



The screenshot shows a GitHub repository page for user `mcps5601` and repository `CGUNLP_2025_Spring`. The file `llm_api.ipynb` is selected, showing a commit by `mcps5601` with the message "feat: add llm api code and update slide" (commit `be12d59`, 11 hours ago). The file is a Jupyter Notebook with 1932 lines. The code is displayed in the "Code" view and includes a title "Gemini Usage" and two code blocks. The first block shows a command to install `google-generativeai==0.8.3`. The second block shows Python code for importing `google.generativeai` and `json`, and loading prompts from a `prompts.yaml` file. A message indicates that the history saving thread hit an unexpected error.

mcps5601 / CGUNLP_2025_Spring

Code Issues Pull requests Actions Projects Wiki Security 3 Insights Settings

main CGUNLP_2025_Spring / code / llm_api.ipynb

mcps5601 feat: add llm api code and update slide be12d59 · 11 hours ago History

Preview Code Blame 1932 lines (1932...)

Raw Copy Download Edit

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')).History 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

Preview

Llama 4 Maverick

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

Native multimodal with **1M** context length

Available

Llama 4 Scout

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

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

Available

註冊 Google Vertex

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



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

選取專案

新增專案

搜尋專案和資料夾

近期專案

已加星號

全部

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

搜尋llama4

The screenshot shows the Google Cloud Model Garden interface. At the top, there's a search bar with the text '搜尋資源、文件、產品和其他項目 (請按 / 鍵)' and a search icon. Below the search bar, the 'Model Garden' section is active, displaying search results for 'llama4'. A red arrow points to the 'Llama 4 API Service' result, with the text '點這個' (Click this) next to it. The search results list includes 'Llama 4 API Service' and 'Llama 4'. Below the search results, there are sections for 'Full text search', 'Model Garden toolkit (SDK, CLI, API)', 'Latest feature updates', and 'Learn about Model Garden'. At the bottom, there are four featured model cards: Gemini, Imagen 3, WeatherNext, and Llama 4.

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

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

Seek models now
Seek models now
Seek models now
Seek models now

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

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。

[Review](#)

啟用

 開啟筆記本

[Overview](#)

[Use cases](#)

[Documentation](#)

[License](#)



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

The screenshot shows the Google Cloud console interface for the Llama 4 API Service. At the top, there's a navigation bar with the Google Cloud logo, a search bar, and a notification icon. Below the navigation bar, the left sidebar shows the 'Llama 4 API Service' selected. The main content area displays the service details, including a 'CO 開啟筆記本' button and a '查看程式碼' button. A star icon indicates that the service is in GA and will be billed per million tokens starting May 1, 2025. The 'Overview' section provides information about the Llama 4 collection of models, which are natively multimodal AI models. The 'Model Information' section describes the Llama 4 Scout and Llama 4 Maverick models. The 'Model developer' is listed as Meta. The 'Model Architecture' is described as auto-regressive language models using a mixture-of-experts (MoE) architecture. On the right side, there's a '試用「Llama 4 API Service」' section with a dropdown for '地區' (us-central1 (愛荷華州)) and a dropdown for '端點' (meta/llama-4-scout-17b-16e-instruct...). Below this, there's a 'Prompt *' field and a '提交' button. At the bottom, a terminal window shows the command 'gcloud access token' being executed, resulting in the output 'aibot_cgu@cloudshell:~ (llama3-405b-459201) \$ 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。

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

Llama 4 API Service

Llama 4 API Service

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

CO 開啟筆記本 查看程式碼

Overview Use cases Documentation License

★ 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

CLOUD SHELL 終端機 (llama3-405b-459201) +

```
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