台智雲福爾摩沙大模型 FFM

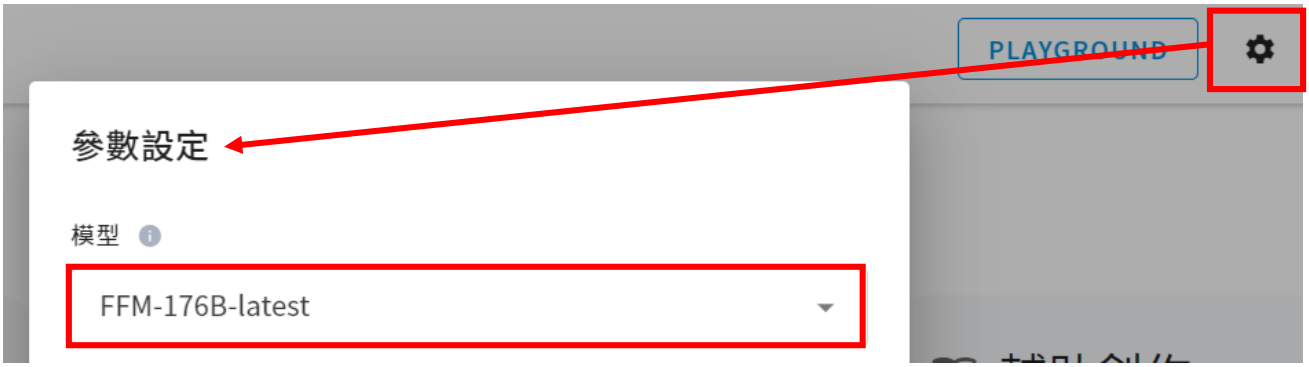
API 說明文件

版本: 0714

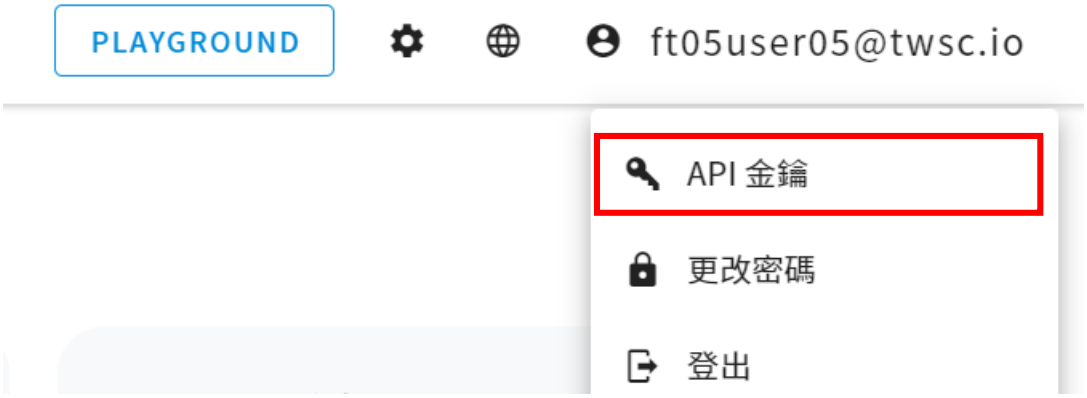
取得所需資訊

MODEL_NAME：登入後點選右上角的設定 > 參數設定 中取得模型名稱

下例即為 FFM-176B-latest



API_KEY：登入後點選右上角的帳號資訊，即可看到 API 金鑰



API_URL：https://ffm-trial05.twcc.ai/text-generation/api/models

使用 curl 直接呼叫 AFS

1. 一般使用

```
export MODEL_NAME= FFM-176B-latest
export API_KEY=00000000-0000-0000-0000-000000000000
export API_URL=https:// ffm-trial05.twcc.ai/text-generation/api/models

curl "${API_URL}/generate" \
  -H "X-API-KEY:${API_KEY}" \
  -H "content-type: application/json" \
  -d "{\"model\":\"${MODEL_NAME}\", \"inputs\":\"請問台灣最高的山是？\"}"
```

輸出

{
“generated_text”:“\n\n 台灣最高的山是玉山，海拔 3952 公尺。”,
“total_time_taken”:“2.15 sec”,
“generated_tokens”:13
}

2. 進階使用 (指定參數)

```
export MODEL_NAME= FFM-176B-latest
export API_KEY=00000000-0000-0000-0000-000000000000
export API_URL=https:// ffm-trial05.twcc.ai/text-generation/api/models

curl "${API_URL}/generate" \
  -H "X-API-KEY:${API_KEY}" \
  -H "content-type: application/json" \
  -d "{\"model\":\"${MODEL_NAME}\", \"inputs\":\"可以幫我規劃台北兩日遊，並推薦每天的景點及說明其特色嗎？
\",\"parameters\":{\"max_new_tokens\":350,\"temperature\":0.5,\"top_k\":50,\"top_p\":1,\"frequence_penalty\":1}}"
```

輸出

{
“generated_text”:“\n\n 非常好的問題！以下是我為您規劃的台北兩日遊行程：Day 1: 早上：前

往西門町，這裡有很多特色小店和美食，可以品嚐台灣小吃，如雞排、珍珠奶茶等。中午：前往台北101觀景台，俯瞰整個台北市的美景，也可以在觀景台上的餐廳享用美食。下午：前往故宮博物院參觀，這裡有豐富的歷史文物和藝術品，可以深入了解台灣的歷史和文化。晚上：前往夜市，台北有很多著名的夜市，如饒河街夜市、士林夜市等，可以品嚐到台灣各地的特色美食。Day 2: 早上：前往陽明山國家公園，這裡有美麗的山林風光和溫泉，可以享受大自然的美景和放鬆身心。中午：前往中山區，這裡有許多美食餐廳和特色小店，可以品嚐各式美食和購買伴手禮。下午：前往台北車站，這裡是台北的交通樞紐，可以參觀車站大廳和周邊的商店。晚上：前往松山文創園區，這裡有許多藝術展覽和設計小店，可以欣賞到台灣年輕人的創意和設計。以上是我為您規劃的台北兩日遊行程，希望對您有所幫助！”,

“total_time_taken”:“28.45 sec”,

“generated_tokens”:255

}

3. 使用 Server-sent events (SSE) 模式

```
export MODEL_NAME= FFM-176B-latest
export API_KEY=00000000-0000-0000-0000-000000000000
export API_URL=https:// ffm-trial05.twcc.ai/text-generation/api/models

curl "${API_URL}/generate_stream" \
  -H "accept: text/event-stream" \
  -H "X-API-KEY:${API_KEY}" \
  -H "content-type: application/json" \
  -d "{\"model\":\"${MODEL_NAME}\", \"inputs\":\"請問台灣最高的山是？\"}"
```

輸出

```
data: {"generated_text": "\n\n", "details": null, "total_time_taken": null,
"generated_tokens": null}
data: {"generated_text": "台灣", "details": null, "total_time_taken": null,
"generated_tokens": null}
data: {"generated_text": "最高", "details": null, "total_time_taken": null,
"generated_tokens": null}
data: {"generated_text": "的山", "details": null, "total_time_taken": null,
"generated_tokens": null}
data: {"generated_text": "是", "details": null, "total_time_taken": null,
"generated_tokens": null}
```

```
data: {"generated_text": "玉山", "details": null, "total_time_taken": null,
"generated_tokens": null}
data: {"generated_text": "，", "details": null, "total_time_taken": null,
"generated_tokens": null}
data: {"generated_text": "海拔", "details": null, "total_time_taken": null,
"generated_tokens": null}
data: {"generated_text": "39", "details": null, "total_time_taken": null,
"generated_tokens": null}
data: {"generated_text": "52", "details": null, "total_time_taken": null,
"generated_tokens": null}
data: {"generated_text": "公尺", "details": null, "total_time_taken": null,
"generated_tokens": null}
data: {"generated_text": "。", "details": null, "total_time_taken": null,
"generated_tokens": null}
data: {"generated_text": "\n\n 台灣最高的山是玉山，海拔 3952 公尺。", "details": null,
"total_time_taken": "2.16 sec", "generated_tokens": 13}
```

使用 Python 呼叫 AFS

```python
import json
import requests

MODEL_NAME = "tws-ffm-176b"
API_KEY = "00000000-0000-0000-0000-000000000000"
API_URL = "https://api-ffm.twcc.ai/text-generation/api/models"

# parameters
max_new_tokens = 350
temperature = 0.52
top_k = 50
top_p = 1.0
frequence_penalty = 1.0

def predict(prompt):
    headers = {"content-type": "application/json", "X-API-Key": API_KEY}
    data = {
        "model": MODEL_NAME,
```

```
        "inputs": prompt.replace("\"", "\\\""),
        "parameters": {
            "max_new_tokens": max_new_tokens,
            "temperature": temperature,
            "top_k": top_k,
            "top_p": top_p,
            "frequence_penalty": frequence_penalty
        }
    }

    result = ''
    try:
        response = requests.post(
            API_URL + "/generate", json=data, headers=headers)
        if response.status_code == 200:
            result = json.loads(response.text, strict=False)['generated_text'][2:]
            result = result.strip("\n")
        else:
            result = "Error: {:d}, {}".format(response.status_code, response.text)
    except Exception as e:
        print("Error: " + str(e.args))
    return result

result = predict("可以幫我規劃台北兩日遊，並推薦每天的景點及說明其特色嗎？")
print(result)
```

結果輸出

```
好的，我很樂意幫您規劃台北兩日遊。以下是推薦的行程和景點：

第一天：
上午：您可以參觀台北市中心的故宮博物院，欣賞中國古代的藝術和文化遺產。
下午：您可以前往台北市中心的松山文創園區，這裡有許多設計師和藝術家的工作室和展覽，讓您體驗台灣的文化創意產業。
晚上：您可以前往台北市中心的西門町，這裡是年輕人和遊客的聚集地，有許多美食和娛樂場所，可以讓您度過一個美好的夜晚。

第二天：
```

上午：您可以前往台北市中心的陽明山國家公園，這裡有美麗的山景和古老的森林，可以讓您遠離城市的喧囂，享受大自然的美好。
下午：您可以前往台北市中心的華山文創園區，這裡有許多藝術展覽和表演，讓您體驗台灣的文化氛圍。
晚上：您可以前往台北市中心的夜市，這裡有許多當地美食和特色小吃，可以讓您品嚐台灣的美食文化。

設定參數說明

- max_new_tokens
    - 一次最多可生成的 token 數
    - default: 20
    - range: int [1…1024]
- temperature
    - 生成文本的隨機和多樣性。值越大，文本更具創意和多樣性；值越小，則較保守、接近模型所訓練的文本
    - default: 1.0
    - range: float (0,∞)
- top_p
    - 當候選 token 的累計機率達到或超過此值時，就會停止選擇更多的候選 token。值越大，生成的文本越多樣化；值越小，生成的文本越保守
    - default: 1.0
    - range: float (0,1]

- top_k
    - 限制模型只從具有最高概率的 K 個 token 中進行選擇。值越大，生成文本越多樣化；值越小，生成的文本越保守
    - default: 50
    - range: int [1...100]
- frequence_penalty
    - 控制重複生成 token 的概率。值越大，重複 token 出現次數將降低
    - default: 1.0
    - range: float (0,∞)

注意事項

- 回傳內容的前綴可能會有 \n\n 換行字元要自行處理
- 需考慮 server 中斷或連不上錯誤處理機制

Lang Chain : 請參閱 email 中附件的 html 檔
