Baseline 程式碼說明

本次複賽提供 baseline 程式碼作為參考，讓參賽者對競賽要求和框架有更好地理解。在複賽階段，主辦單位提供一個簡單的 LLM generation pipeline 和相應的 API 程式，該 API 可以回傳 LLM 生成的結果，幫助參賽者更方便地測試和整合 LLM 生成功能。參賽者可以自由選擇在baseline 程式碼的基礎上進行修改和優化，或採用其他技術策略來提高效果，實現更具創意和競爭力的解決方案。

環境建立

```
pip install -r requirements.txt
```

Baseline 使用方法：

以範例 150 題資料為例 (路徑請自行修正)，並可透過:

主程式

```
python3 llm_generate.py \
    --question_path ../dataset/final_example/questions_final_example.json \
    --pred_retrieve_path ../dataset/final_example/pred_retrieve_final_example.json \
    --source_path ../reference/ \
    --output_path ../dataset/final_example/pred_generate_final_example.json \
    --url http://0.0.0.0:8087/chat
```

其中 llm_generate.py 為主程式，負責處理答案生成，其參數分別為

- question_path：提供問題的 JSON 檔案路徑。
- pred_retrieve_path：初賽產生之檢索結果。
- source_path：需要檢索的參考資料的路徑。
- output_path：產生的預測答案將被儲存在這個路徑中。
- url：呼叫 LLM API 的 url。

API 程式

```
python main_taide_llama3_api.py
```

其中 main_taide_llama3_api.py 為 API 程式，負責回傳模型生成結果
在 llm_generate.py 主程式執行期間，main_taide_llama3_api.py 之 API 程式需全程開啟提供主程式呼叫，在 VM 上可參考 linux 的 screen 指令達成。

模型取得

本範例程式中使用的模型為 Hugging Face 網站上的 Llama3-TAIDE-LX-8B-Chat-Alpha1 連結，因其提供了極高效能，同時以相對輕量的 8B 參數模型達成優異表現，適合需要高準確度與效能的對話應用，尤其在資源受限的環境中運行時，更顯得其輕量優勢。在此範例中可以 git lfs 的方式下載到 VM 上做使用，參賽者也可以自由選用其他模型。

Git lfs 建議操作方式 (以 Llama3-TAIDE-LX-8B-Chat-Alpha1 為例):

```
GIT_LFS_SKIP_SMUDGE=1 git clone https://huggingface.co/taide/Llama3-TAIDE-LX-8B-Chat-Alpha1
cd Llama3-TAIDE-LX-8B-Chat-Alpha1
git lfs pull
```

雲端硬體規格

同時請參賽者注意，因為複賽提供的 VM 有資源限制（如下表），請注意選用的模型不要超過限制，可參考此網站進行評估：連結

項目	詳細資訊
虛擬機類型	GCP g2-standard-4
作業系統	Ubuntu 20.04 LTS
CPU處理器	4 vCPU
RAM記憶體	16 GB
GPU規格	NVIDIA L4
GPU記憶體	24GB
磁碟大小	100 GB
Python版本	3.8.10
CUDA版本	12.4 (另以腳本安裝)

修正方向

參賽者可以根據比賽需求，對 baseline 程式碼進行擴展或修改。以下是幾個建議的方向：

- 資料預處理：在讀取資料之前，增加自定義的資料預處理步驟，以提高模型的輸入品質。
- LLM 之使用：可自行選用不同 LLM，或是 Fine-tune LLM 來獲得更好的成效。
- Prompt 之精進：修改 prompt，使模型更清楚任務內容，產出精確回答。
- 多模態資料處理：處理圖片、表格等多種資料型態，以提供更清晰的語意輸入給模型。
- 流程改進：修改現有流程，如在不超過 LLM 輸入限制下，是否可以在生成答案前放入其他文章的部分內容。

TAIDE 概述

TAIDE 計畫致力於開發符合台灣語言和文化特性的生成式人工智慧對話引擎模型，同時建構可信任的人工智慧環境。結合產學研能量，推動可信任生成式人工智慧的發展，提升台灣在國際競爭中的地位，促進產業發展，避免對外國技術的依賴。
Llama3 TAIDE 系列模型以 Meta 公司釋出的 LLaMA3-8b 為基礎，導入台灣不同領域可用的文本與訓練素材，提高模型在正體中文回應的能力與特定任務的表現。其特色為：

- 嚴格把關模型的訓練資料，提升模型生成資料的可信任性和適用性
- 針對自動摘要、寫信、寫文章、中翻英、英翻中等辦公室常用任務做加強
- 針對台灣在地文化、用語、國情等知識做加強
- 具備多輪問答對話能力