**AI CUP 2023 秋季賽**

**隱私保護與醫學數據標準化競賽：**

**解碼臨床病例、讓數據說故事 競賽**

**報告說明文件**

* 參賽隊伍需詳細說明系統的流程、演算法、工具與外部資源等，特別是創新性(如改了外部資源的哪一部分)。主辦單位會請專業學者、專家仔細審查，若發現分數有問題或方法說明不清之部分，請參賽隊伍補充說明，經發現有違規者，將取消獎項資格。
* 請使用A4紙直式打字，中文字體使用標楷體，英文、數字與符號使用Times New Roman字體。
* 版面設定：邊界上下各2.54CM，左邊3.17CM、右邊3.0CM。
* 字體大小：題目20（粗體）**、**內文12，單行間距。
* 作者聯絡資料表：詳細資料請填寫於附件的「作者聯絡資料表」。
* 須依照以下大綱及內容說明撰寫。不可自訂標題、修改內容順序，或合併段落。但可搭配圖片說明。
* 附件空白報告範本可參用。
* **報告大綱與撰寫規定（字數不得低於各段落規定下限）**

壹、環境

字數規定：200~600字。

內容規定：請說明使用的作業系統、語言、套件(函式庫)、預訓練模型、額外資料集等。如使用預訓練模型及額外資料集，請逐一列出來源。

貳、演算方法與模型架構

字數規定：400~1200字。

內容規定：說明演算法設計、模型架構與模型參數，包括可能使用的特殊處理方式。

參、創新性

字數規定：300~1200字。

內容規定：說明演算法之創新性或者修改外部資源的哪一部分。

肆、資料處理

字數規定：300~1500字。

內容規定：說明對資料的處理或擴增的方式，例如對資料可能的刪減、更正或增補。

伍、訓練方式

字數規定：400~1000字

內容規定：說明模型的訓練方法與過程。

陸、分析與結論

字數規定：400~2500字。

內容規定：分析所使用的模型及其成效，簡述未來可能改進的方向。分析必須附圖，可將幾個成功的和失敗的例子附上並說明之。

柒、程式碼(未於**2023/12/11 11:59前**繳程式碼連結者，將失去獲頒獎金/獎狀資格)

內容規定：請在此提供程式碼下載連結並請另外信件附檔，包含資料處理、訓練流程、預測等相關程式碼。程式碼應附README.md檔案交代安裝配置環境，重要模塊輸出/輸入，以讓第三方用戶可以除錯、重新訓練與重現結果。繳交前請確認連結有效且有開啟瀏覽權限，如連結失效視同未交。

捌、使用的外部資源與參考文獻

內容規定：參考文獻請以APA格式為主。

★註1：請確認上述資料與AI CUP報名系統中填寫之內容相同。自2023年起，獎狀製作將依據報名系統中填寫內容為準，有特殊狀況需修正者，請主動於報告繳交期限內來信moe.ai.ncu@gmail.com。，報告繳交截止時間後將不予修改。

★註2：繳交程式碼檔案與報告，請Email至：F109154156@nkust.edu.tw，並同時副本至：moe.ai.ncu@gmail.com。缺一不可。報告檔名與信件主旨請寫「競賽報告與程式碼/TEAM\_？？？？/隱私保護與醫學數據標準化競賽：

解碼臨床病例、讓數據說故事競賽」

**AI CUP 2023 秋季賽**

附件

**隱私保護與醫學數據標準化競賽：**

**解碼臨床病例、讓數據說故事競賽 競賽報告**

隊伍：TEAM\_ 4084

隊員：黃聖軒（隊長）、李政皓、鄭竑銨

Private leaderboard：0.7444 / 0.5961 / Rank 20

1. **環境**

一、硬體配置

我使用的硬體配置如下：

處理器（CPU）：Intel i5 13600K

顯示卡（GPU）：NVIDIA GeForce 3060ti

二、作業系統

本環境運行在 Windows 11 作業系統上，提供穩定且友好的操作界面。

三、程式語言及套件

Python：

使用 Python 作為主要的程式語言，提供了豐富的套件支援。

PyTorch：

使用 PyTorch 框架進行深度學習模型的構建和訓練，並在 GPU 上運行以加速運算。

Transformers：

Transformers 套件為自然語言處理（NLP）提供了豐富的模型和工具，方便進行預訓練模型的使用。

Datasets：

Datasets 套件用於方便地處理和載入各種數據集，使得資料的處理更加容易。

Random、PEFT、TQDM、Matplotlib：

分別使用 Random 庫進行隨機數生成、PEFT 庫用於性能評估、TQDM 用於進度條的顯示、Matplotlib 用於數據可視化。

四、預訓練模型

模型名稱：EleutherAI/pythia-160m-deduped

來源：EleutherAI 團隊提供，為一個中型的預訓練語言模型，適用於多種自然語言處理任務。

五、額外資料集

為了擴充訓練數據，除了競賽提供的資料集，我們還使用了 ChatGPT 生成的額外資料集，該資料集內有資料樣本較少的文本，用於進一步訓練模型以提高其表現。

1. **演算方法與模型架構**
2. 演算法設計

本專案的演算方法基於預訓練的語言模型，透過 Hugging Face 的 Transformers 套件實現。首先，我們選擇了一個基於 EleutherAI/pythia-160m-deduped 預訓練模型，該模型在大型文本語料上進行了深度學習。接著，我們使用 ChatGPT 生成額外的資料集，以擴充原始數據集，提供模型更多的訓練樣本。

1. 模型建構

模型的建構基於 Hugging Face Transformers 套件，我們選用了 EleutherAI/pythia-160m-deduped 預訓練模型。使用 AutoModelForCausalLM 類別，我們初始化了語言生成模型。模型的 token embeddings 大小透過 resize\_token\_embeddings 方法進行調整，以適應使用的分詞器的詞彙大小。

1. 模型參數

模型的優化採用了 AdamW 優化器，學習率設定為 5e-5。同時，我們使用了線性學習率調度器，通過 get\_linear\_schedule\_with\_warmup 方法進行初始化，以在訓練的初始階段進行學習率預熱。

1. 特殊處理方式

在模型的訓練過程中，我們使用了適當的特殊處理方式。為了處理不均衡的訓練樣本，我們在 DataLoader 中使用了自定義的 BatchSampler，將數據分組為不同的批次，以提高訓練的效果。我們使用了特殊標記來標識文本序列的起始、結束、填充和分隔，以協助模型理解文本的結構。

**參、創新性**

**肆、資料處理**

**伍、訓練方式**

**陸、分析與結論**

**柒、程式碼**

Github連結：

**捌、使用的外部資源與參考文獻**

**作者聯絡資料表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 隊伍名稱 |  | Private Leaderboard 成績 |  | Private Leaderboard 名次 |  |
| 身分  （隊長/隊員） | 姓名 (中英皆需填寫)  (英文寫法為名,姓，例：Xiao－Ming, Wu，名須加連字號，姓前須加逗號) | 學校＋系所中文全稱  (請填寫完整全名，勿縮寫) | 學校＋系所英文中文全稱  (請填寫完整全名，勿縮寫) | 電話 | E-mail |
| 隊長 | (範例)  吳小明  Xiao－Ming, Wu | 國立中央大學資訊工程學系 | National Central University Department of Computer Science & Information Engineering | 0900-123-456 | abc＠gmail.com |
| 隊員1 |  |  |  |  |  |
| 隊員2 |  |  |  |  |  |
| 隊員3 |  |  |  |  |  |
| 隊員4 |  |  |  |  |  |

★註1：請確認上述資料與AI CUP報名系統中填寫之內容相同。自2023年起，獎狀製作將依據報名系統中填寫內容為準，有特殊狀況需修正者，請主動於報告繳交期限內來信moe.ai.ncu@gmail.com。報告繳交截止時間後將不予修改。

★註2：繳交程式碼檔案與報告，請Email至：F109154156@nkust.edu.tw，並同時副本至：moe.ai.ncu@gmail.com。缺一不可。