AI CUP 2025 春季賽

醫病語音敏感個人資料辨識競賽報告

隊伍：TEAM_7464
隊員：王渝萱（隊長）、林佳誼、楊心怡、李虹瑩
Private leaderboard：0.2562 / Rank 19

壹、環境

本專案於 Windows 11 專業版 24H2 環境中進行開發與測試，主要使用程式語言為 Python 3.11。所依賴之主要套件與函式庫包括：torchaudio 2.1.0、transformers 4.5.0 以及 librosa 等。

在預訓練模型方面，任務一之英文語音識別採用 Hugging Face 提供的 Whisper 模型（openai/whisper-small）作為語音轉文字（ASR）之核心工具，來源：https://huggingface.co/openai/whisper-small。針對中文語音部分，則採用規模更大的 Whisper 模型（openai/whisper-large），以提升中文語音辨識的準確率，模型來源：https://huggingface.co/openai/whisper-large。

任務二的語音敏感個人資料辨識則使用 deepseek-ai/deepseek-llm-7b-base 模型作為命名實體識別與語意理解的基礎模型，模型來源：https://huggingface.co/deepseek-ai/deepseek-llm-7b-base。

此外，最終參賽模型之訓練全程僅使用競賽主辦單位所提供之資料集，未使用任何未經授權的外部語音資料。開發初期曾嘗試引入 LibriSpeech 資料集以擴充訓練語料，惟模型整體表現未見顯著提升，最終未將其納入正式競賽成果中。

貳、演算方法與模型架構

針對兩項任務設計不同演算流程與模型架構：任務一為中英文語音轉文字（ASR），任務二為語音中個資實體（SHI）辨識。整體方法結合語音辨識模型 Whisper 以及語言理解模型 DeepSeek LLM，並透過參數微調與資料增強提升辨識準確度與泛化能力。

一、任務一醫病語音紀錄辨識模型架構與演算法

任務一採用 OpenAI 提出的 Whisper 模型進行語音辨識，該模型基於 Encoder-Decoder 架構，並結合 Transformer 設計，具備多語言與多任務能力。

Whisper 將音訊處理後，進入 Encoder 提取時序特徵，再由 Decoder 自回歸產生文字 token。針對英文語音採用 whisper-small，中文語音則使用 whisper-large。

針對語言特性與資源差異，對中文語音使用 whisper-large，英文語音則採用運算資源較輕量的 whisper-small。為降低訓練成本並強化語言模型適應性，本專案實施 Decoder-only fine-tuning，僅微調解碼器參數並凍結編碼器。此外，推論階段強制指定語言參數以避免語言錯判（zh 或 en），並設定 task="transcribe" 確保模型執行轉錄任務。

Whisper 預設輸出為英文與簡體中文，為符合繁體中文應用需求，將中文文字透過 OpenCC 模組進行轉換。為強化字級對齊能力，整合 WhisperX 模型進行語音與文字的 forced alignment，取得 word-level timestamps，為後續 SHI 實體標註提供時間參考依據。

二、任務二醫病語音隱私個資辨識模型架構與演算法

任務二屬於含時間標註的命名實體辨識任務，目標為從語音轉文字資料中辨識出具敏感性的個人資訊（如姓名、電話、地址等），並保留每一實體對應的音訊時間區間。基礎模型選用 deepseek-ai/deepseek-llm-7b-base，該模型具備強大語意理解與生成能力，適用於多樣化的實體辨識任務。

考量硬體資源限制，本研究引入 PEFT 框架中的 LoRA（技巧進行參數微調，並採用 BitsAndBytes 套件對模型進行 4-bit 量化，有效減少訓練所需 GPU 記憶體與計算負擔。

訓練階段採用分段式策略：第一階段以全體資料訓練模型，學習主要語境與常見實體類型；後續各階段則依據資料中出現頻率較低的稀有標籤（如 AGE、ZIP、PHONE 等）進行針對性微調，逐步強化模型在少數類別上的辨識能力。此外，結合 word-level 時間對齊資訊，建立跨模組的 SHI 標註。

針對模型可能遺漏的實體，補充正則表達式與關鍵詞清單比對機制（如國家與城市名稱表），進行自動化補標註，有效提升整體召回率。

參、創新性

本研究提出兩項具體創新方法，針對語音敏感個資辨識任務中常見的「類別不平衡」與「標籤辨識疏漏」問題，從訓練策略與後處理機制兩個層面進行設計與改進。

一、多階段稀有實體聚焦式訓練策略

針對任務二中命名實體類別極度不平衡的問題，本專案設計「多階段稀有實體聚焦式訓練策略」。該策略將整體微調流程分為兩大階段：

（一）初始階段：基礎語意與高頻標籤訓練

首先使用全部標註資料進行初步微調，使模型掌握最常見的標籤類型（如 DOCTOR、DATE 等）。

（二）進階階段：針對性稀有標籤微調

針對在訓練資料中出現比例較低的標籤類別（如 DEPARTMENT、HOSPITAL、ZIP 等），採用依分布比例篩選的分段式微調策略。每一階段僅選取標籤比例低於特定閾值（例如 5%）的樣本進行訓練，使模型能更著重於稀有類別的學習。

二、模型輸出後續補標註機制

受限於 Whisper 模型轉錄錯誤、語者表達模糊或模型語言知識不足，仍有部分實體無法正確預測。為此，本專案設計模型輸出後續補標註機制，具體作法如下：

- 正則式規則補標註：根據實體格式特徵（如身分證規則、郵遞區號為 4 碼數字、日期）設計正則表達式，自動比對模型預測遺漏資訊。
- 類別特定詞彙庫補標註：針對國家、日期等類別建立對應詞庫（如國家名稱對應表、月份節日表），檢索模型輸出補足未標註資訊。
  針對模型輸出，進行第二層過濾與修正，有助於降低資訊遺漏。

肆、資料處理

為提升語音辨識模型對中英文資料的處理能力，本研究針對任務一醫病語音紀錄辨識設計資料處理與資料擴增策略，涵蓋語言分類與模型分配、語音增強處理，以及訓練資料增加策略。

一、中英文語音資料分離與模型分配

本專案首先依據語音對應文字標註內容進行語言分類，將語音資料分為中文（zh）與英文（en）兩類。分類後，分別輸入至適用之 Whisper 模型進行訓練：英文語音資料配對 whisper-small 模型，中文語音則配對參數較大的 whisper-large 模型。

此策略可避免單一模型同時學習多語言所產生的干擾與語言切換成本，讓模型專注於各自語言的聲學與語言特性進行最佳化訓練，進而提升辨識準確度。

二、語音資料增強處理

語音辨識模型需要面對不同錄音設備、環境噪音與語者語音特性等多種干擾，因此本專案設計基於 torchaudio 之隨機語音增強流程。增強方法包含：

- Volume Gain：隨機調整音量強度（±5 dB），模擬不同錄音設備或語者音量差異。
- Frequency Masking：模擬部分頻率受雜訊干擾，遮蔽特定頻率範圍。
- Time Masking：模擬語音斷裂、打結或背景中斷，遮蔽部分時間片段。

每筆語音訓練資料皆會隨機套用 1 至 3 種增強方法，確保語音變異性與自然性，增強後的語音保留原始語意，但在頻率與時間軸上具有更高的多樣性。

三、資料倍增訓練策略

為進一步提升模型訓練資料的多樣性與數量，本專案採用資料倍增策略。將原始資料經前述語音增強處理後產生的樣本，與未增強之原始樣本進行合併，形成最終訓練資料集。

此策略使模型於訓練過程中同時見到原始語音樣本與增強語音樣本，達到資料量翻倍的效果，強化模型對變異語音的辨識能力。

伍、訓練方式

本研究針對兩項任務設計不同的訓練策略，並根據資料特性、任務需求與硬體資源限制，設定訓練流程，分別應用於任務一（醫病語音紀錄辨識）與任務二（語音中隱私個資辨識）之中。

一、任務一醫病語音紀錄辨識

針對中英文醫病語音紀錄辨識，本專案將資料依語言分類後分別輸入至 whisper-small（英文）與 whisper-large（中文）模型進行訓練。訓練策略採 Decoder-only 微調，僅更新 Decoder 權重以強化語言輸出層能力，並保留 Encoder 的聲學特徵處理能力。

訓練參數採用以下設定：

- 訓練週期：15 epochs

- 學習率：1e-5
- 累積梯度步數：4
- 學習率排程器：linear + warmup_ratio=0.1
- 提早停止條件：early_stopping_patience=3
- 評估指標：MER

搭配資料增強與語言指定 task="transcribe" 、 language="zh"/"en"，模型可更穩定輸出對應語言的正確文字，且不受混合語料干擾。

二、任務二醫病語音隱私個資辨識

任務二為文字層面的命名實體辨識（NER）任務，資料來源為任務一轉錄結果。模型使用 deepseek-llm-7b-base，並透過 LoRA 技術於 PEFT 框架下進行參數微調。為兼顧訓練效率與硬體限制（單張 24GB GPU），本專案搭配 BitsAndBytes 進行 4-bit 量化。訓練共分為五階段：

（一）第一階段：全面基礎訓練

第一階段使用完整訓練資料集進行大規模預訓練，目標為學習語音轉文字的語言風格與常見 PHI 類別的語義分佈。我們使用學習率 2e-4，訓練上限設為 100 個 epoch，並搭配 Early Stopping 監控 training loss 表現。最終會選出 loss 表現最佳的模型權重，儲存為 best_adapter 作為後續階段的初始化權重。

（二）第二至第五階段：針對稀有標籤的階段性微調

由於某些實體類別（如 PROFESSION、ORGANIZATION）在訓練資料中比例極低，模型在第一階段可能無法充分學習其特徵，故設計第二至第五階段進行針對性的微調（fine-tuning）。學習率依訓練階段遞減，逐步縮小訓練資料範圍以聚焦於稀有類別：

- 第二階段：出現次數低於總樣本數 × 0.1，lr = 1e-4
- 第三階段：出現次數低於總樣本數 × 0.06，lr = 5e-5
- 第四階段：出現次數低於總樣本數 × 0.04，lr = 4e-5
- 第五階段：出現次數低於總樣本數 × 0.02，lr = 3e-5

各階段皆以前一階段最佳模型作為初始化，延續 LoRA 微調方式，更新部分參數，穩定提升模型對稀有標籤的辨識能力。搭配預先取得的字級對齊資訊（WhisperX），支援輸出實體時間戳記，建立用於語音資料之個資辨識系統。

陸、分析與結論

一、任務一醫病語音紀錄辨識模型成效分析

本研究針對中英文語音轉文字任務，分別選用 OpenAI 提出的 Whisper-small 與 Whisper-large 模型，並透過 Decoder-only 微調與資料增強策略進行訓練。以下就各模型在不同語言語料下的訓練過程與效能表現進行分析，並探討未來可行的改進方向。

（一）英文 Whisper-small 模型訓練成效

Whisper-small 模型在英文語料的訓練結果顯示，雖然在第一個 epoch 即達到 MER 最佳值（0.05196），但隨著訓練次數增加，模型表現未持續改善，從第三、四個 epoch 的 MER 變化觀察可見，模型已有過擬合現象，Validation Loss 與 MER 呈現反彈趨勢。

模型初始（未訓練）MER 為 0.1272，相較於第一 epoch 訓練後下降至 0.05 左右，顯示 Decoder-only 微調策略對模型有強化效果。

表格 1 英文 Whisper-small 模型 訓練結果

Epoch	Training Loss	Validation Loss	Mer
1	0.119800	0.346663	0.051955
2	0.155000	0.318579	0.088102
3	0.026700	0.316474	0.108344
4	0.205600	0.323819	0.087367

（二）中文 Whisper-small 模型訓練成效

在中文語音資料上，Whisper-small 模型表現明顯受限於模型容量不足。儘管 Training loss 呈下降趨勢，從第一輪的 1.898 降至第九輪的 0.6237，Validation loss 同樣有穩定下降趨勢，MER 則由初期的 0.419 降至最終的 0.257。在 MER 無法再進一步下降的情況下，顯示模型表現已接近其能力上限。

表格 2 中文 Whisper-small 模型 訓練結果

Epoch	Training Loss	Validation Loss	Mer
1	1.898200	5.264801	0.419087
2	1.353900	4.014293	0.419087
3	1.673800	2.444010	0.410788
4	0.675100	1.977471	0.334711
5	0.265200	1.661616	0.264463
6	0.663800	1.448774	0.256198
7	0.331100	1.311160	0.257261
8	0.174100	1.222688	0.265306
9	0.623700	1.172051	0.257143

（三）中文 Whisper-large 模型訓練成效

針對中文語音，我們亦訓練了 Whisper-large 模型以比較其在高容量下的學習效果。整體結果顯示此模型優於 small 模型，其 MER 最終降至 0.1147，相較初始階段的 0.1934。模型在第 10 epoch 時達成最佳驗證表現，此時 Validation loss 為 0.6367，MER 為最低。整體 loss 曲線與 MER 呈現穩定下降趨勢，未明顯出現過擬合情形，顯示 large 模型具有更好的學習能力。

表格 3 中文 Whisper- large 模型 訓練結果

Epoch	Training Loss	Validation Loss	Mer
1	1.576300	2.310561	0.193416
2	0.641800	1.316765	0.193416
3	1.057600	1.206209	0.201646
4	0.321400	1.082290	0.158537
5	0.119600	0.797220	0.162602
6	0.367100	0.760300	0.162602
7	0.166000	0.728106	0.123967
8	0.075500	0.696419	0.119835
9	0.351900	0.662085	0.134694
10	0.143700	0.636705	0.114754

（四）LibriSpeech 外部資料集加入訓練

為進一步提升英文語音辨識效果，實驗後期加入 5,567 筆 LibriSpeech 訓練集資料進行兩階段訓練，整體設計如下：

- 第一階段（混合訓練）：使用原始訓練集 ＋LibriSpeech 語料，學習率設定為 1e-5。
- 第二階段（回歸微調）：排除 LibriSpeech，僅使用原訓練集進行再訓練，學習率降低為 5e-6，期望模型重聚焦於任務目標領域。

1. 第一階段（含 LibriSpeech 語料）

模型於第 2 epoch 即降至 MER 0.0937，優於原始模型（0.1272），顯示 LibriSpeech 語料有助於模型在英語語音上的能力。然而後續 MER 變化不穩定，出現反彈，甚至在最後一個 epoch 回升至 0.125。表示模型可能因語域差異導致過度擬合 LibriSpeech 特性。

表格 4 LibriSpeech 外部資料加入模型第一階段訓練 訓練結果

Epoch	Training Loss	Validation Loss	Mer
1	0.833300	0.420845	0.117230
2	0.665600	0.415219	0.093725
3	0.562500	0.434159	0.100295
4	0.845800	0.481148	0.124912

2. 第二階段（移除 LibriSpeech 後微調）

儘管回歸任務語料進行第二階段微調，MER 未能穩定下降，反而略高於僅使用原訓練資料訓練的最佳表現（MER 0.0519）。推測可能因 LibriSpeech （朗讀文本風格）與任務資料語音風格（如醫療口語、背景噪音、語速）差異過大，反而干擾模型在原任務的擬合能力。

表格 5 LibriSpeech 外部資料加入模型第二階段訓練 訓練結果

Epoch	Training Loss	Validation Loss	Mer
0	1.154100	0.429294	0.124651
1	1.281000	0.446532	0.135078
2	1.011800	0.479907	0.124148
3	0.650700	0.517477	0.148678
4	0.557600	0.556592	0.125889

加入 LibriSpeech 資料集雖短期內改善模型 MER 表現，但長期泛化至任務語料表現有限，未產生明顯實質提升。因此本專案最終比賽提交版本未採用此策略，仍以官方訓練資料為主。未來若欲導入外部語音資源，可嘗試精選語域接近之語料，如醫療會談、健康諮詢等。

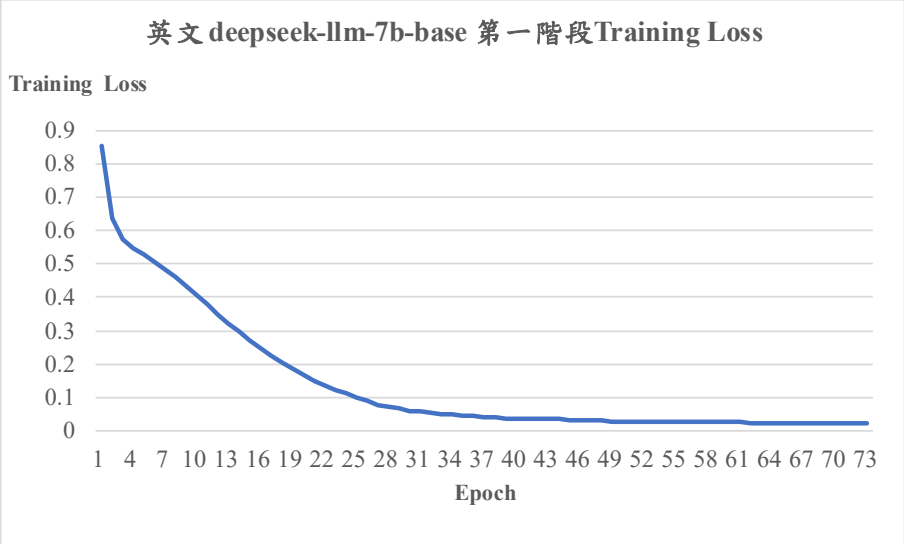
二、任務二醫病語音隱私個資辨識模型成效分析

本研究第二任務聚焦於語音轉文字後的個人資料實體（PHI）辨識，採用 deepseek-llm-7b-base 模型進行任務，並以 PEFT 框架中的 LoRA 技術進行參數微調。整體訓練分為「基礎訓練階段」與「稀有實體強化階段」，針對中文與英文資料分別進行多階段訓練，以下為訓練結果與模型效能分析。

（一）英文 deepseek-llm-7b-base

1. 第一階段基礎訓練

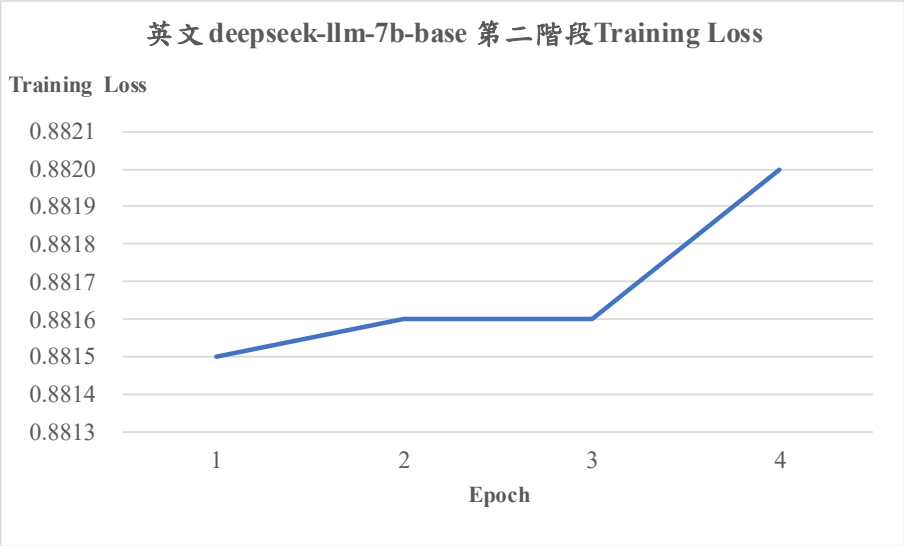
第一階段使用全體訓練資料進行完整微調，主要目標為讓模型掌握醫療語境中常見的命名實體類別（如 DOCTOR、DATE、PATIENT 等），並建立語言與語意理解能力。從訓練過程可觀察到 Training Loss 值穩定下降，自 Epoch 1 的 0.8531 降至 Epoch 73 僅約 0.0241，呈現收斂良好且無明顯過擬合情形。



圖片 1 英文 deepseek-llm-7b-base 第一階段模型 訓練結果

2. 第二階段訓練（稀有類別微調）

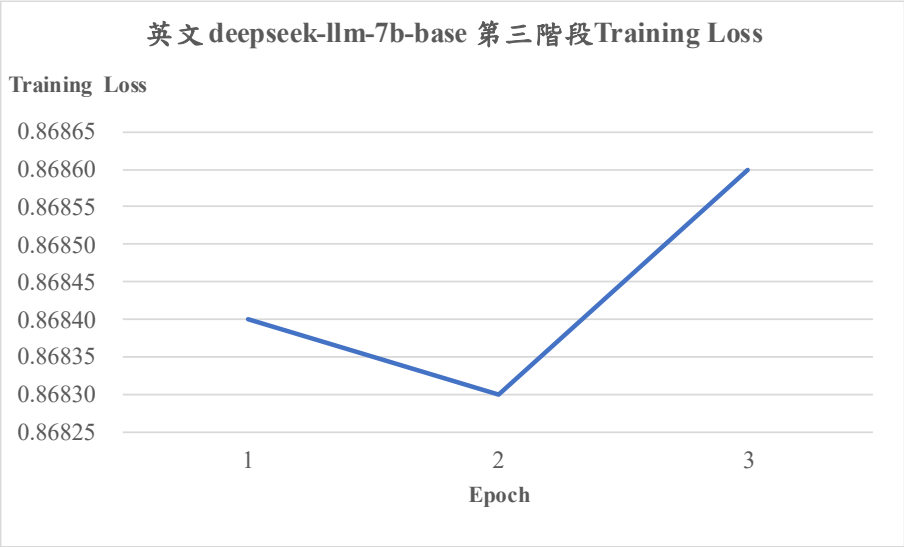
第二階段針對訓練集中出現次數比例低於 10% 的稀有標籤類別進行微調，以期提升模型對頻率低的標籤辨識能力。然而訓練 loss 自 Epoch 1 起即維持在 0.8815 左右，幾乎未下降，顯示模型在進一步優化稀有實體的辨識能力上面臨瓶頸，可能因資料量不足、標註品質有限或實體語境過於分散導致難以有效學習。



圖片 2 英文 deepseek-llm-7b-base 第二階段模型 訓練結果

3. 第三與第五階段訓練（進階微調）

第三階段進一步篩選出比例更低的稀有實體進行訓練，loss 為 0.8683 左右，與前一階段類似，同樣未呈現顯著下降趨勢，可能顯示模型已難以從極小量樣本中學習到有效特徵。



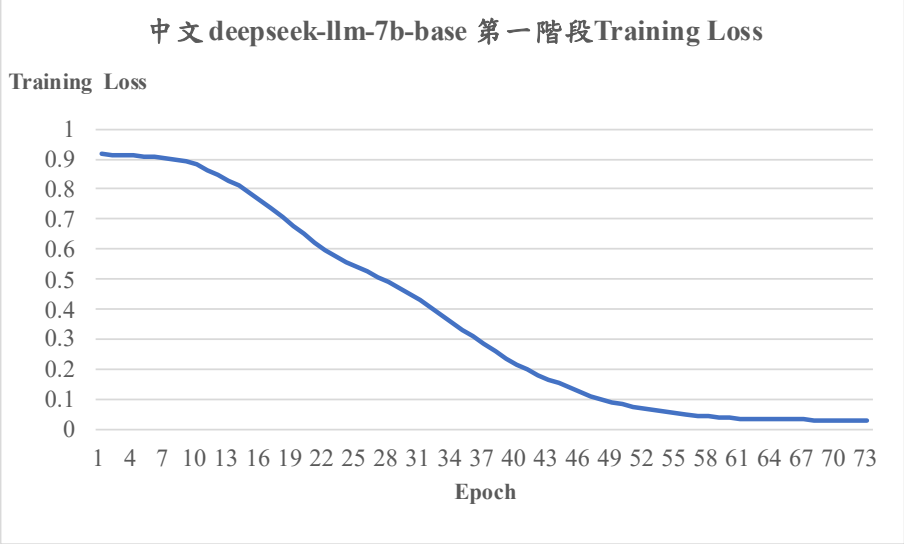
圖片 3 英文 deepseek-llm-7b-base 第三階段模型 訓練結果

第一階段完整訓練效果良好，可穩定學習高頻率標籤，後續階段針對性微調對於極低頻率標籤提升效果有限。

（二）中文 deepseek-llm-7b-base

1. 第一階段基礎訓練

中文 deepseek 模型在第一階段持續訓練 86 個 epoch，Training Loss 由 0.9177 穩定下降至 0.0239，無明顯過擬合或震盪，顯示模型已能有效掌握中文醫療語境下常見命名實體。從第 50 epoch 開始，Training Loss 顯進入平穩期（維持在 0.02~0.03 區間），表示模型已充分擬合訓練資料，後續再訓練效益有限。

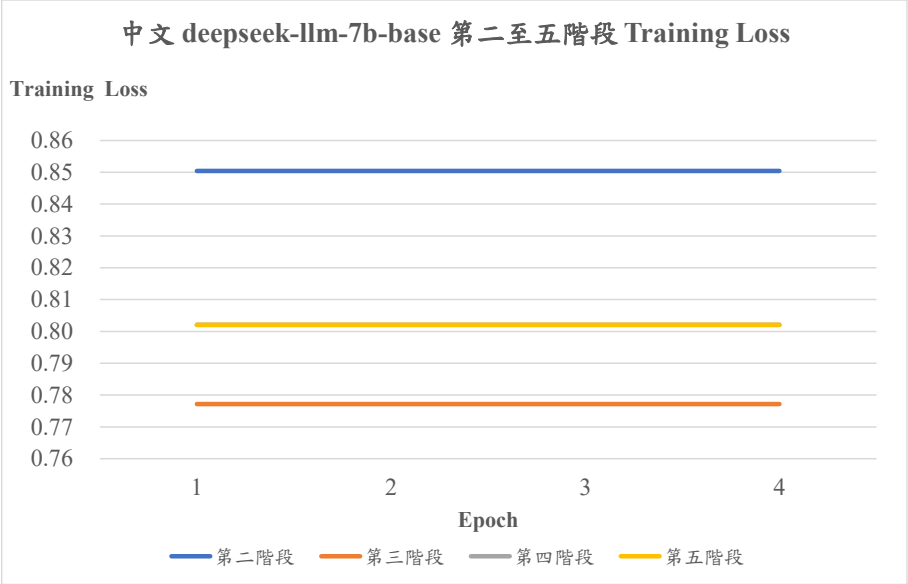


圖片 4 中文 deepseek-llm-7b-base 第一階段模型 訓練結果

2. 第二至第五階段訓練（稀有類別微調）

與英文情況相似，中文模型在進行稀有類別調整時，Training Loss 第二階段起即出現停滯，長期固定在 0.8504、0.7772 或 0.8021 等數值，未出現顯著下降，此現象顯示中文模型已完成有效擬合，在中文語料中針對稀有實體進行單獨訓練的效果有限。

面對樣本稀少的實體再訓練時難以獲得顯著效益。可能的瓶頸包括：稀有類別分布高度不均，訓練樣本數不足、中文實體邊界更依賴語意判斷，缺乏結構線索、微調過程中缺乏資料多樣性，導致泛化能力不足。



圖片 5 中文 deepseek-llm-7b-base 第二至五階段模型 訓練結果

三、改進方向

針對本專案中觀察到的限制與挑戰，提出以下改進方向：

（一）引入資料合成技術

利用語言模型生成含有稀有標籤的語句樣本，以擴增訓練資料多樣性，進一步提升模型對低頻樣本的識別能力。

（二）結合外部知識庫進行後處理

建立醫療專有名詞表、實體對應清單（如醫學科別對應表、地名郵遞區號字典等），針對模型輸出進行自動校正與補標註，降低模型遺漏錯誤。

（三）語音與語言模型整合式學習

未來可考慮將 Whisper 模型與 LLM 進行更緊密整合，實現 end-to-end 語音實體識別，避免語音轉文字階段誤差對下游任務造成干擾。

柒、程式碼

本專案核心程式碼已整理為 Jupyter Notebook 檔案，附於附檔：AICUP2025_TEAM_7464.ipynb。內容包含以下幾個主要模組：

1. Whisper 語音資料預處理與增強流程
2. Whisper 模型微調設定（Decoder-only fine-tuning）
3. Whisper 模型訓練
4. DeepSeek LLM 微調架構（LoRA + 4-bit 量化）
5. DeepSeek LLM 模型輸出後處理與自動補標模組（正則式+詞庫補標）

捌、使用的外部資源與參考文獻

LibriSpeech ASR Corpus. (n.d.). OpenSLR. https://www.openslr.org/12

作者聯絡資料表

隊伍名稱	TEAM_7464	Private Leaderboard 成績	0.2562	Private Leaderboard 名次	19
隊長	王渝萱 Yu-Hsuan, Wang	國立臺東大學資訊工程學系	National Taitung University Department of Computer Science & Information Engineering	0988-112-227	yusyuan24@gmail.com
隊員1	林佳誼 Chia-Yi, Lin	國立臺東大學資訊工程學系	National Taitung University Department of Computer Science & Information Engineering	0900-642-123	11111106@gm.nttu.edu.tw
隊員2	李虹瑩 Hung-Ying, Lee	國立臺東大學資訊工程學系	National Taitung University Department of Computer Science & Information Engineering	0989-070-102	dianalee5328@gmail.com
隊員3	楊心怡 Hsin-Yi, Yang	國立臺東大學資訊工程學系	National Taitung University Department of Computer Science & Information Engineering	0968-201-431	hsinyii22@gmail.com