

中華民國第 65 屆中小學科學展覽會

作品說明書

高級中等學校組 電腦與資訊學科

佳作

052505

智慧・進化艙：AI 語言革命，解密失語症分類

學校名稱： 國立新竹科學園區實驗高級中等學校

作者： 高二 張楷鎭	指導老師： 林怡瑄
-------------------	------------------

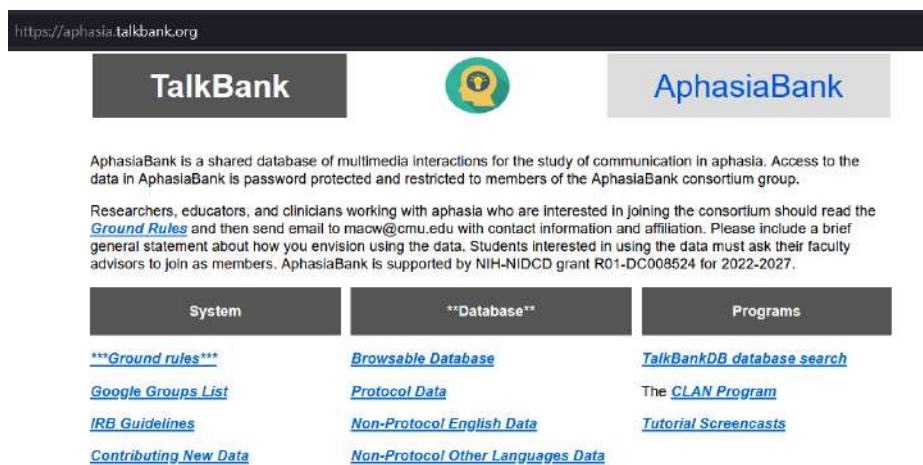
關鍵詞： 失語症、BERT、深度學習

摘要

本研究提出一套結合 BERT 模型與自訓練分類器的失語症分類系統，透過多個方法及參數組合提升判斷精確性。研究靈感來自櫻井等人(2023)[15]提出的 BERT 分類模型，並參考 Cong et al.(2024)[5]針對大型語言模型(LLMs)在臨床環境中的應用，進一步探索自訓練分類器在增強 BERT 判別能力方面的潛力。

本研究的創新點在於，模型基於醫生標註的分類結果進行深度學習，而非讓 AI 直接從語料中自動分類病患類型，這與先前主要依賴 LLMs 自動分類的方法不同。此外，本研究納入詞性、語法結構與語音長度及肢體語言標記，以提升模型的處理能力與個人化治療。

考慮語料庫規模及醫學倫理，本研究採用 AphasiaBank(全球最具權威的語料資源之一)的英文訪談文本進行模型訓練，同時移除所有個人隱私資訊。



【圖 1】 Aphasia Bank 網頁

(致謝研究資助計畫：NIH-NIDCD R01-DC008524 (2022-2027)。(MacWhinney[17] (n.d.))

壹、前言

一、研究動機

在醫院擔任志工期間，我曾與一位中風患者交流，但他因無法順利組織語言表達自己需求，感到焦慮絕望，醫療人員也無法適時提供協助。我意識到，語言障礙不僅影響溝通，更

深刻影響患者的心理健康與社會參與。這不只是個案，而是全球數百萬失語症患者每天面臨的挑戰。

近期，知名演員布魯斯·威利（Bruce Willis）因失語症宣布退休，引發社會關注(Topic News[4])。隨著全球人口老化，失語症患者數量急劇上升，傳統診斷方法耗時且依賴專業醫療人員，這促使我思考：是否能利用 AI 來加速診斷，提升治療效率？

二、研究背景

失語症影響聽、說、讀、寫能力，使患者面臨溝通困難與社交孤立，甚至引發焦慮與抑鬱。隨著人口老化與中風病例增加，失語症患者數量持續上升。例如，美國約有 100 萬名患者，每年新增 18 萬人(NIH[22]; Topic News[4])；台灣患者約 4-6.5 萬(臺中榮總[3])；日本則達 20-50 萬(櫻井等[15])。但由於治療資源相對短缺，專業語言治療師人數有限，導致許多患者無法獲得即時診斷與治療。

目前診斷失語症的方法包括標準化語言測試（如 WAB，Western Aphasia Battery，韋氏失語症測試）及神經影像技術（fMRI、PET），但這些方法耗時且高度依賴專業醫療人員，難以大規模應用。治療方式則包括語言復健、輔助溝通工具（如語音合成裝置）及 AI 技術輔助療法。然而，現有 AI 系統仍面臨三大挑戰：僵化的診斷方式無法根據患者需求靈活調整，缺乏長期記憶能力難以追蹤患者語言進展，且適應性不足，導致對不同類型失語症患者的分類效果不佳；此外，因為失語症語料的稀少特性和難以分析的特性，讓此任務的難度遠超傳統的文本分類與分析或是診斷。

近年來，人工智慧（Artificial Intelligence，AI）與大型語言模型（large language model，LLMs）逐步應用於失語症診斷與評估，QuickAphasiaBattery(QAB)可透過測試評估患者語言能力，BERT（Bidirectional Encoder Representations from Transformers）和 GPT（Generative Pre-trained Transformer）展現自動語言分析的潛力。然而，現有 AI 診斷模型多依賴未標註語料，與臨床標準仍有落差，對低語料數量語言的判斷力較差，也難以針對個別患者的需求進行調整。因此，仍需發展更精確且判斷能力更高的 AI 診斷方法，以提升診斷效率並降低對專業醫療資源的依賴。

三、研究目的

本研究開發一個基於 BERT 的 AI 診斷系統，透過臨床標註語料（CHAT）進行學習，以提升診斷準確率並縮短診斷時間。與以往研究不同，本系統不僅自動分類失語症類別，還能分析詞性、語法與語音長度，讓 AI 分析失語症診斷更快、更準、更普及。以適應個人化治療需求。

四、文獻回顧

本研究與 Cong et al. (2024) [5] 提出的大型語言模型（LLMs）在失語症診斷中的應用相呼應。該研究顯示，LLMs 透過大規模預訓練與微調（fine-tuning），能有效分析患者語音及文本輸入，以輔助臨床診斷。本研究應用 BERT 來提升失語症類型的分類處理能力，並進一步探討 NLP(Natural Language Processing，自然語言處理)增強技術的影響。

(一) 失語症診斷與挑戰

1. NIH (2015)[22]：探討失語症的分類，並強調不同類型對診斷與治療的影響。
2. Dronkers et al.(2007)[9]：發現布洛卡失語症的腦損傷超出布洛卡區，挑戰傳統失語症分類。
3. D' Alessio (2022)[25]：探討失語症的多樣性，並突顯準確分類對治療的重要性。
4. Wilson et al.(2018)[8]：發展 Quick Aphasia Battery(QAB)，提供多維度評估工具。
5. Mahmoud et al. (2025)[7]：該研究開發非書寫系統語言的評估工具，區分診斷失語症並制定復健計劃。
6. Goral et al. (2024) [6]：研究多語言患者如何受失語症影響，這對於 BERT 分析多語言語料的適用性提供了參考。

(二) AI 在失語症診斷中的應用

1. Cong et al. (2024)[5]：探討了大型語言模型（LLMs）在語言障礙診斷中的臨床效能，特別是對失語症的檢測。
2. Brown et al. (2020)[13]：GPT-3 展示 Few-shot Learning（少樣本學習）能力，適用於低語料環境，這對於訓練醫學 NLP 模型具有重要影響。

(三) NLP 與 LLMs 在醫療應用

1. Devlin et al.(2019)[16]：BERT 透過雙向 Transformer 理解語境，透過遮蔽語言模型與下一句預測，在醫學文本分類上具備優勢。
2. Vaswani et al.(2017)[14]：Transformer 的注意力機制提升語言處理能力，為 NLP 任務奠定基礎，使 LLM 在醫療 NLP 領域的應用更具可行性。

(四) 值得一提的是，Cong 等人之研究 [5] 亦使用 AphasiaBank，與本研究在模型設計與語料使用上有所不同，相關比較如表 1 所示。（表格內容由本研究整理彙編而成）

【表 1】文獻方法比較：本研究與 Cong et al. (2024) [5] 在模型設計與語料使用上的差異

項目	Cong et al. (2024)	本研究
使用語料	AphasiaBank，無對應醫師標註分類	AphasiaBank，使用醫師已標註類型
分類類別	只有 Broca、Wernicke	4 類以上(含 Conduction、NotAphasicByWAB)
使用模型	GPT-2 / GPT-Neo（LLM 生成架構）	BERT + 自訂分類器
特徵設計	純 LLM 表徵	加入語法、詞性、語音長度特徵
調參方法	固定模型架構	使用 Optuna 搜尋最佳參數

貳、研究設備及器材

一、硬體設備及作業系統

- (一)、個人電腦：Windows 11 Intel i7 CPU，32GB RAM，NVIDIA RTX 3080 GPU，用於模型訓練與測試。
- (二)、硬體環境：國內頂大伺服器，2xNVIDIA RTX 4090 GPU，Ubuntu 22.04.5 LTS (GNU/Linux 6.8.0-51-generic x86_64) 用於深度學習訓練。

二、軟體工具

- (一)、開發環境：Python 3.10.12

使用套件：

1. RE：正則表達式模組，用於處理字串的模式匹配與字串搜索。
2. JSON 模組：用於解析語料。
3. Numpy：用於高效數值計算與矩陣運算的函式庫。
4. torch.nn：有用的神經網路模組

5. optuna：超參數搜尋，找出最好的組合
6. sklearn：隨機森林和產出報表
7. torch.utils.data：處理資料的模組

(二)、深度學習框架：PyTorch

(三)、自然語言處理工具：

1. Trasnsformer-BERT：

- (1) BERTTokenizer：BERT 分詞器

- (2) BERTModel：BERT 模型（用於獲取詞向量、CLS 向量等）

(四)、數據來源：失語症患者的語音與文字數據集（基於 AphasiaBank 語料庫）。

(五)、語料讀取軟體：CLAN(由 AphasiaBank 提供並下載使用)

(六)、語料格式：*.cha 檔案格式

參、研究方法與過程

本研究透過三階段優化模型訓練。第一階段使用詞彙單位（tokens）作為輸入，觀察學習效果；第二階段加入語法結構、詞性標註與肢體語言標記，提升語言理解能力；第三階段優化長文本處理，透過文本切分技術增強處理能力，也因為 BERT 對於 512 字元的限制。此外，參考 Cong et al.[5] 進行少量分類，涵蓋 Anomic、Conduction、NotAphasicByWAB 及 Broca 失語症類別，並透過超參數搜尋尋找最佳模型組合。同時，嘗試結合隨機森林與 TF-IDF 進行結果比較，並使用 Optuna 進行超參數搜尋與模型驗證。

一、研究方法

(一)、資料來源與預處理

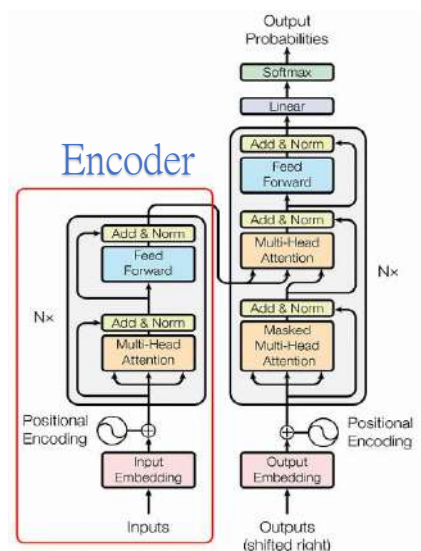
本研究使用 AphasiaBank 語料庫作為主要訓練數據來源，該資料庫涵蓋不同類型的失語症患者對話，並依據 CHAT 標準進行標註，每筆語料包含：

1. 詞性標註（POS，Part-of-Speech）
2. 文法結構（GRA）
3. 音節時長（DUR）
4. 文字數據（TOKEN）

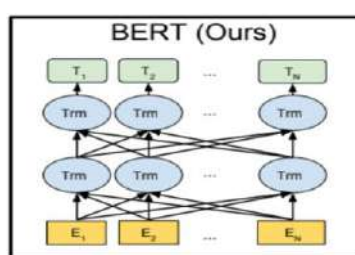
這些語言特徵為本研究所使用的模型提供豐富的訓練資訊，從而使模型能夠有效分類不同失語症類型。（本研究語料庫來源為 AphasiaBank，向 NICHDHD082736 致謝提供）

(二)、BERT 模型架構與標記方法

研究中，我們採用 Transformer 編碼器（Encoder）架構，並利用 BERT 作為語言表示模型，以適應醫病對話場景。如圖 2、圖 3，BERT 具備雙向語境學習能力，能夠理解對話中的語境變化，提升醫療語言的解析準確度。



【圖 2】BERT 架構圖 (Vaswani 等[14]；Mahmoud 等[7])。圖中 BERT 採用左側編碼器部分。



【圖 3】BERT 輸入模式舉例(Devlin et al.[16], 2019)

1. 三階段改良(如圖 4)

- (1) 第一階段：模型以 tokens 作為輸入，學習基本詞彙與語言結構，觀察其對詞彙單位的學習效果。
- (2) 第二階段：在第一階段的基礎上，加入語法結構與詞性特徵，以提升模型對文本的理解能力，使其能更準確地掌握詞與詞之間的語法關係，進一步強化語言理解表現。此外，肢體語言的整合亦有助於提升本研究的整體準確率。
- (3) 第三階段：針對長文本輸入進行切分與調整，透過文本切分技術提升模型處理長語料的能力。



【圖 4】BERT 三階段語料標記方法(研究者使用 Canva 自行設計繪製)

2. 加入自訂標記，強化 BERT 在醫療文本分析的能力

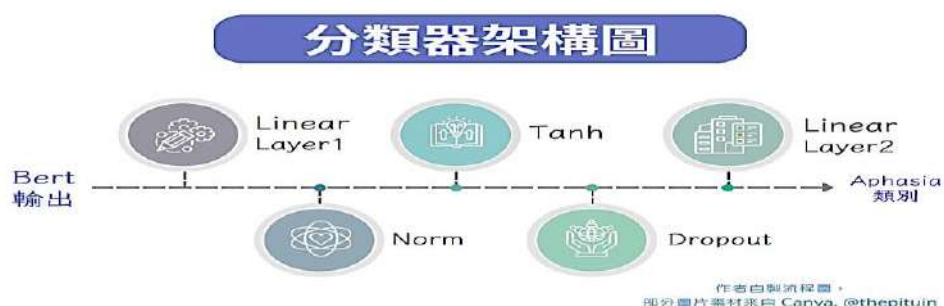
- (1) 詞性標註 (POS Tagging)：提升模型對詞類的識別與語法解析能力。
- (2) 語法特徵 (Grammar Features)：讓模型理解語句結構與相關性。
- (3) 時間資訊 (Duration Information)：分析停頓長度與流暢度。確保模型理解醫療對話的時間與事件關聯。

(三)、自定義分類器架構

本研究搭配 BERT 訓練自定義分類器 (CustomClassifier)，架構如圖 5 所示：

1. Linear 層 1：將 BERT 輸出結合額外特徵後，轉換為 256 維向量。
2. Normalization 層：標準化數據，避免梯度爆炸或是模型偏化。
3. Tanh 層：採用 tanh 激活函數以提升分類效能。(根據 Optuna 超參數搜尋結果，此配置表現最佳)
4. Dropout 層：模擬遺忘機制，避免模型過擬合。(根據 Optuna 超參數搜尋結果，採 0.2 為丟失率)
5. Linear 層 2：將最終輸出轉換為失語症類型標籤。

這套分類架構是透過 Optuna 超參數搜尋數十個組合之後最好的結果，其中激活函數採用 tanh (雙曲正切函數)，Dropout_rate(丟失率)為 0.2。

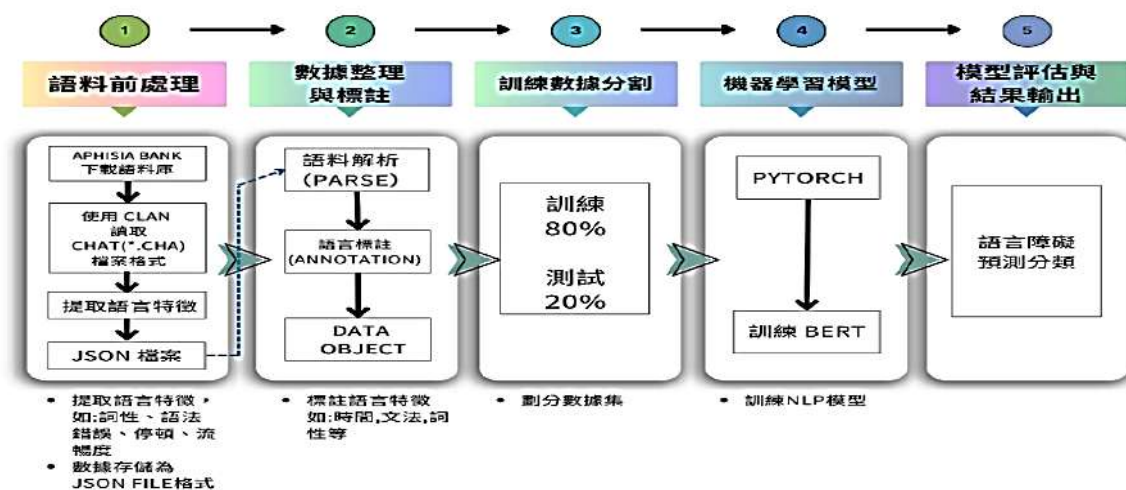


【圖 5】分類器架構圖(研究者使用 Canva 自行設計繪製)

二、研究架構與流程

研究流程分為五個步驟，如圖 6：

- (一)、語料前處理：從 AphasiaBank 下載資料，該資料庫由卡內基美隆大學教授 Brian MacWhinney 管理。擷取詞性、語法與語音時長等特徵，轉為 JSON 格式。
- (二)、數據整理與標註：語料解析並標記 POS(詞性)、文法結構、時間資訊及肢體語言，確保模型能學習醫病對話的特徵。
- (三)、訓練數據分割：80%作為訓練集，20%作為測試集，確保模型分析的準確性。
- (四)、機器學習模型訓練：利用 PyTorch 訓練 BERT 模型，適應醫療語境。
- (五)、模型評估與結果輸出：透過分類指標（Accuracy、Recall、F1-score）評估模型效能，並生成診斷預測結果。



【圖 6】語料處理流程(研究者使用 Canva 自行設計繪製)

三、資料蒐集與預處理

(一)、語料收集與分類

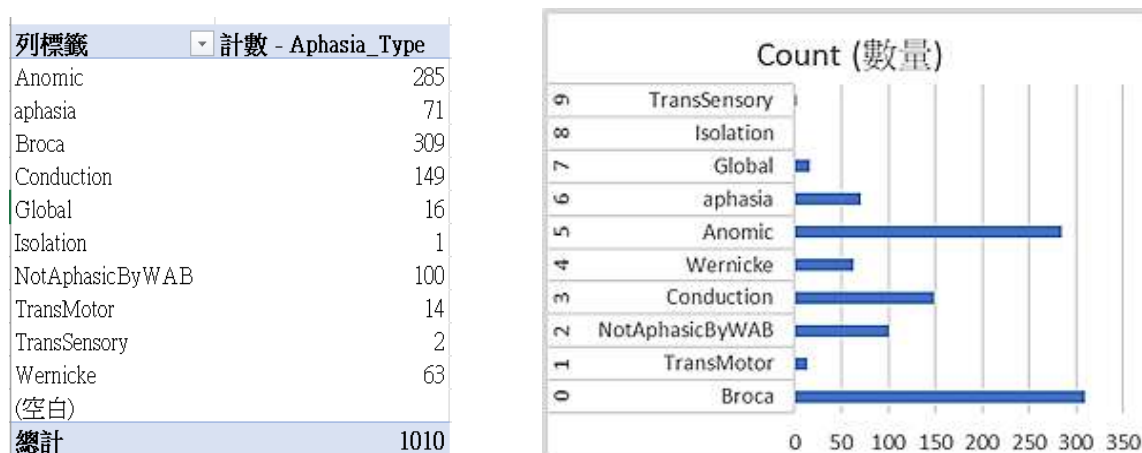
1. 語料來源：來自 AphasiaBank，1023 篇參與者（Participants）影片檔。(13 筆資料未標記失語症類型資料，其中又有 5 筆資料未標記性別。最後投入 1010 篇資料做訓練。)
2. 性別分佈：男性 592 人，女性 426 人。
3. 年齡區分：60 歲以下 406 人，60 歲以上 617 人。
4. 失語症類型(致謝 Aphasia Bank 提供語料：NIH-NIDCD R01-DC008524 (2022-2027))

（表格內容由本研究整理彙編而成）

欄位		說明
(1)	患者基本資料 @ ID	患者的基本身份資料
(2)	語言 @ Languages	如英語 (eng)
(3)	研究機構	參與研究的機構名稱
(4)	參與者代號	受試者或參與者的識別編號
(5)	年齡	患者的年齡
(6)	性別	男性 (male) / 女性 (female)
(7)	失語症類型	患者所屬的失語症分類
(8)	標註指標	語音或文本資料的標註項目
(9)	參與者@ Participants	包含患者 (PAR) 與負責訪談的研究員 (INV)
(10)	模式 @ Options	研究或訪談時所採用的治療模式
(11)	語音標註 @ G	語音數據的標註信息

★★隱私處理：我們在實際模型運作時，將(1)至(6)涉及個人隱私的部分去除。

(二)、1010 名受訪者@ID 失語症類型統計分佈，如圖 7：



【圖 7】受訪者失語症類型統計分佈（本研究使用 EXCEL 整理繪製）

(三)、失語症類型分類名說明如下表：

【表 2】AphasiaBank 失語症類型（表格內容由本研究整理彙編而成）
(Stroke Engine[19]; Huang[27]; National Aphasia Association[28])

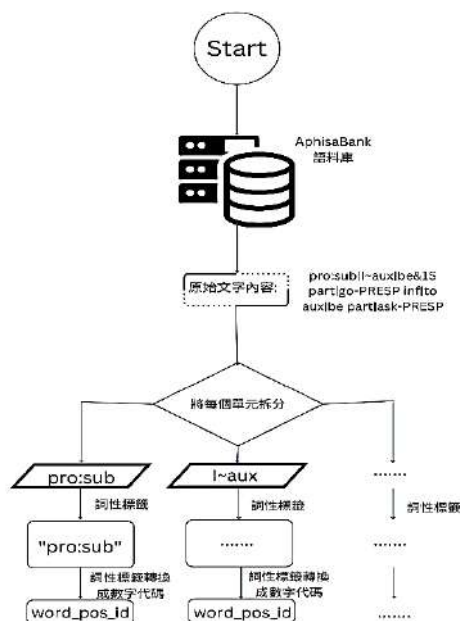
Class Label	分類	中文名稱	說明
0	Broca	布洛卡失語症	言語緩慢、費力，句子簡短但理解良好。
1	TransMotor	經皮層運動性失語	言語流暢度下降，但理解與複述能力正常。
2	NotAphasicByWAB	WAB 篩查非失語症	WAB 測試正常，無明顯失語症。

3	Conduction	傳導性失語症	理解能力良好，重複障礙（伴隨頻繁的語言錯亂），閱讀理解能力良好。
4	Wernicke	韋尼克失語症	流利性失語症，句子不連貫，不相關的字混雜其中。
5	Anomic	命名性失語症	他們的講話雖然語法形式和輸出流暢，但卻充滿了模糊的迂迴表達和沮喪的表達。
6	Aphasia	失語症	表現為一般性語言障礙。
7	Global	全局性失語症	這是失語症最嚴重的形式，患者無法辨識出幾個單字及讀寫。
8	Isolation	孤立性失語症	指失語症患者沒有運動肢體缺陷。
9	TransSensory	經皮層感覺性失語	語言流暢，但理解顯著不足。與韋尼克失語症類似，但複述能力正常。

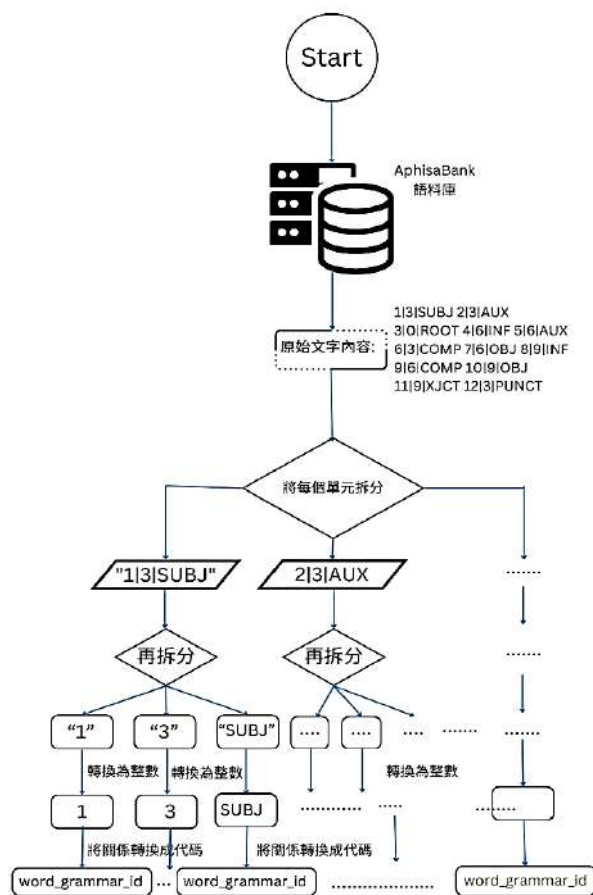
這些分類將用於預測受訪者的症狀，以提升醫療人員對不同患者的治療效率。

(四)、數據預處理步驟(圖 8,圖 9,圖 10,圖 11)

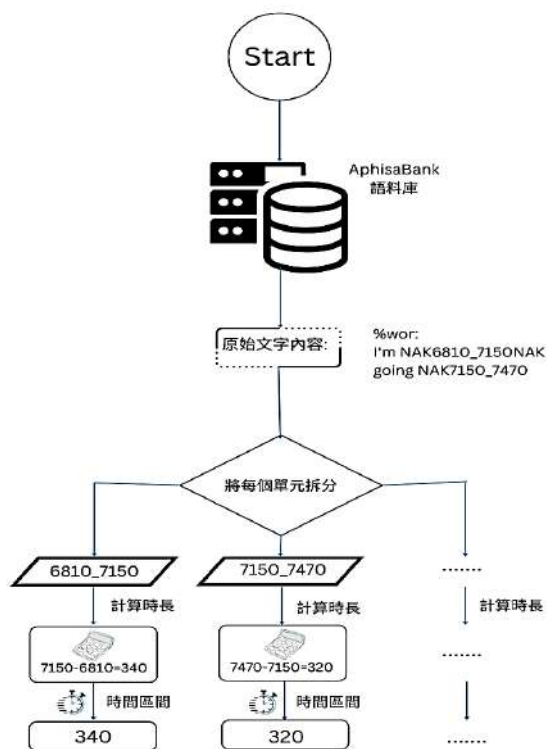
1. 詞性特徵處理：從%mor 註釋提取詞性標籤，轉換為數字代碼（word_pos_id）。
2. 文法特徵處理：拆分%gra 註釋，提取詞性與語法關係（word_grammar_id）。
3. 語音長度處理：從%wor 註釋提取音節時長，分析語音流暢度（word_duration）。



【圖 8】詞性特徵標註處理流程表(研究者使用 Canva 自行設計繪製)



【圖 9】文法特徵標註處理流程圖(研究者使用 Canva 自行設計繪製)



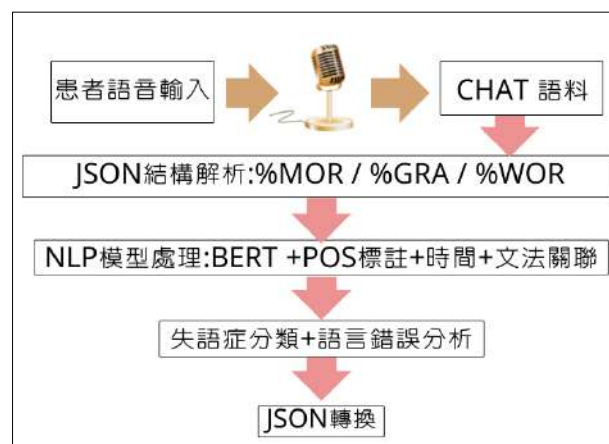
【圖 10】時間特徵標註處理流程圖(研究者使用 Canva 自行設計繪製)


```

{
  "sentences": [
    {
      "sentence_id": "S1",
      "sentence_pid": "11312/t-00017816-1",
      "aphasia_type": "BROCA",
      "dialogues": [
        {
          "INV": {
            "tokens": [
              "word_pos_ids": [
                "word_grammar_ids": [
                "word_durations": [
            ],
            "PAR": {
              "tokens": [
                "yeah",
                "well"
              ],
              "word_pos_ids": [
                8,
                8
              ],
              "word_grammar_ids": [
                [
                  1,
                  2,
                  9
                ],
                [
                  2,
                  0,
                  10
                ],
                [
                  3,
                  2,
                  8
                ]
              ],
              "word_durations": [
                [
                  "yeah",
                  410
                ],
                [
                  "well",
                  1830
                ]
              ]
            }
          }
        ]
      ]
    }
  ]
}

```

【圖 13 JSON 示例】(截圖自本研究於文字編輯器中編輯之畫面)



【圖 14】 JSON 結構解析與語料預處理流程 (JSON Data Pipeline for NLP Processing)
(研究者使用 Canva 自行設計繪製)

五、處理資料不對稱的問題

```
from torch.utils.data import WeightedRandomSampler, RandomSampler
```

```
train_labels = [sample["labels"].item() for sample in train_dataset]
```

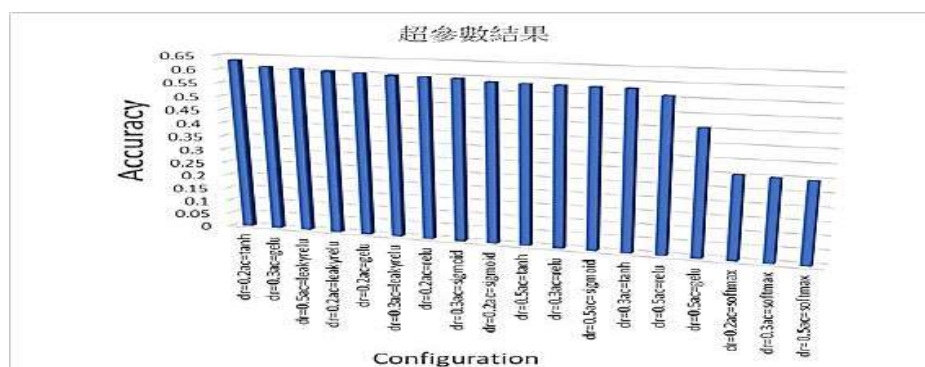
```
label_counts = Counter(train_labels)
```

```
sample_weights = [1.0 / label_counts[label] for label in train_labels]
```

- (一)、以上程式碼是使用 `torch.utils.data` 的 `WeightedRandomSampler`，透過樣本數量的倒數計算權重，使罕見類別的資料有更高機率被選取，最終提升模型分類表現。
- (二)、由於部分類別（如 Isolation、TransSensory）樣本極少，如此可以解決資料不對稱、失語症類別分布不均的問題。

六、超參數搜尋

如圖 15，本研究採用 Optuna 模組進行超參數搜尋，以改善模型表現，最終找出最佳分類器架構。下圖展示表現排名前 18 的參數組合。



【圖 15】Optuna 搜尋結果(研究者使用 EXCEL 製作)

超參數（Hyperparameter）：

- (一)、最佳 LearningRate：5e-5
- (二)、Batch Size：64
- (三)、Dropout Rate：0.2
- (四)、最大 Token 長度：512
- (五)、激活函數：tanh

七、隨機森林模型

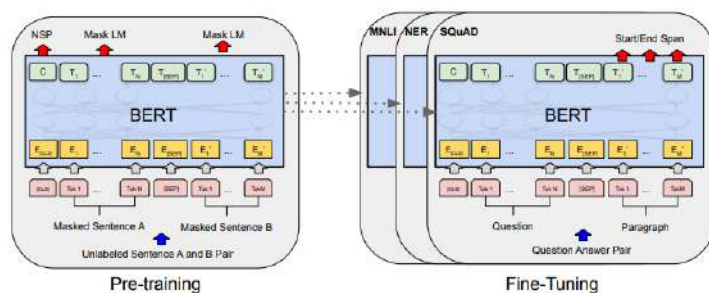
本研究亦採用 隨機森林（Random Forest）結合 TF-IDF 向量化進行文本分類，另外因為使用 `n_splits` 策略，刪除語料數量過少的類別，所以與 BERT 相較，圖 17 相對少兩類別。具體流程如下：

- (一)、生成 TF-IDF 矩陣，將文本轉換為數值特徵表示。

(二)、提取 詞性 (POS)、文法 (Grammar)、時間長度 (Duration) 作為額外特徵。

(三)、如圖 17，n_estimators 設為 150 時，模型的分類準確率達 62%。

相較於 BERT(圖 16)，隨機森林無法自動學習語境，且對長文本表現較弱。因此，隨機森林主要作為對照組，後續用以驗證必須增強語料的結論。n_split 策略排除小樣本類別，使該模型比 BERT+自定義分類器少一類 Isolation。



【圖 16】BERT 的預訓練與微調(Vaswani et al.[14] Devlin et al.[16])

	precision	recall	f1-score	support
0	0.69	0.95	0.80	59
1	0.00	0.00	0.00	3
2	0.67	0.40	0.50	20
3	0.77	0.33	0.47	30
4	0.67	0.17	0.27	12
5	0.52	0.81	0.63	57
6	0.00	0.00	0.00	14
7	0.00	0.00	0.00	3
accuracy			0.62	198
macro avg	0.41	0.33	0.33	198
weighted avg	0.58	0.62	0.56	198

【圖 17】TF-IDF 搭配隨機森林的結果展示圖(本研究於 CMD 中執行訓練程式之畫面截圖)

肆、研究結果

一、視覺化數據分析與語言模型特徵呈現

(一)、BERT 模型的分類結果

1. 訓練與測試數據

本研究使用 AphasiaBank 語料，包含 156,302 筆患者語句，分別包含在 1010 篇有正確標記失語症類型的語料當中

(1) 評估樣本數 (Eval size)：731

- (2) BERT TOKENs 限制： ≤ 512
- (3) 數據劃分：80% 訓練集，20% 測試集

1. 主要指標定義

- (1) 分類準確率(Precision)：模型預測為該類的樣本中，實際正確的比例。
- (2) 召回率(Recall)：實際屬於該類的樣本中，模型成功識別的比例。
- (3) F1 Score：衡量分類的整體表現
- (4) 樣本數量(Support)：樣本數
- (5) 失語症類別(Class label)：根據官方定義

```

2025-02-20T11:47:57.105578: Classification Report:
2025-02-20T11:47:57.106769:
precision    recall  f1-score   support
0           0.705882   0.705882   0.705882   102.000000
1           0.000000   0.000000   0.000000     8.000000
2           0.760479   0.690217   0.723647   184.000000
3           0.484848   0.581818   0.528926   110.000000
4           0.740741   0.400000   0.519481    50.000000
5           0.582734   0.736364   0.650602   220.000000
6           0.523810   0.224490   0.314286    49.000000
7           0.333333   0.200000   0.250000     5.000000
8           0.000000   0.000000   0.000000     0.000000
9           0.000000   0.000000   0.000000     3.000000
accuracy    0.625171   0.625171   0.625171   731.000000
macro avg   0.459092   0.393197   0.410314   731.000000
weighted avg   0.626311   0.625171   0.614340   731.000000

2025-02-20T11:47:57.106821: Class 0 Accuracy: 70.59%
2025-02-20T11:47:57.106840: Class 1 Accuracy: 0.00%
2025-02-20T11:47:57.106854: Class 2 Accuracy: 69.02%
2025-02-20T11:47:57.106867: Class 3 Accuracy: 58.18%
2025-02-20T11:47:57.106879: Class 4 Accuracy: 40.00%
2025-02-20T11:47:57.106891: Class 5 Accuracy: 73.64%
2025-02-20T11:47:57.106902: Class 6 Accuracy: 22.45%
2025-02-20T11:47:57.106914: Class 7 Accuracy: 20.00%
2025-02-20T11:47:57.106926: Class 8 Accuracy: nan%
2025-02-20T11:47:57.106938: Class 9 Accuracy: 0.00%

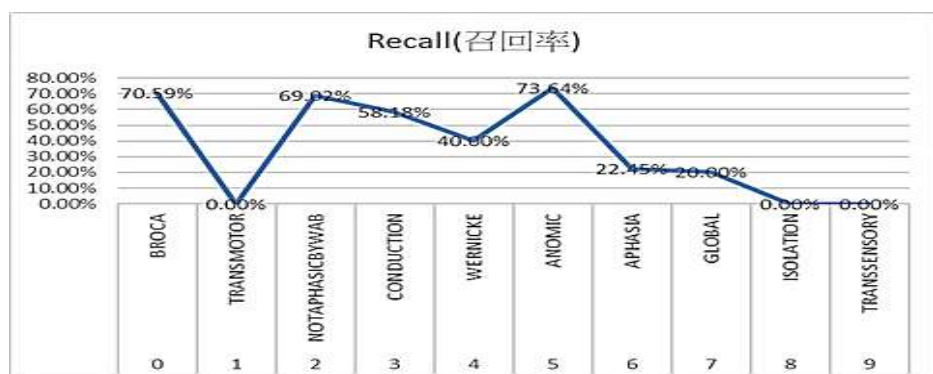
```

【圖 18】測試結果分類報告(本研究於 CMD 中執行訓練程式之畫面截圖)

二、失語症類別的分類召回率分析

(一)、召回率分析

1. 如圖 19，為了清晰呈現不同失語症類別的分類召回率，本研究使用折線圖來顯示各類別模型的預測表現。
2. 從圖中可觀察到，各類別的召回率存在明顯差異：
 - (1) 召回率較高的類別：Broca、NotAphasicByWAB 和 Anomic
 - (2) 召回率較低的類別：TransMotor、Isolation 和 TransSensory（甚至為 0%）



【圖 19】不同失語症類別召回率(recall)比較折線圖(研究者使用 EXCEL 製作)

3. 各類召回率比較如下表

【表 3】召回率(recall)比較表(表格內容由本研究整理彙編而成)

類別	召回率(%)	影響因素
Broca	70.59%	語法結構較清晰，較易辨識
NotAphasicByWAB	69.02%	根據 WAB 測試標記為非失語症患者，較好學習
Anomic	73.64%	詞彙流暢但具語義空缺，BERT 解析效果良好
Conduction	58.27%	有明顯重複語音，對語境理解要求較高
TransSensory	0.00%	樣本數極少，模型無法有效學習特徵

(二)、影響分類召回率的因素

1. 為何 Broca、NotAphasicByWAB、Anomic 類別的準確率較高？

影響因素如下：

(1) 語言結構清晰，語料完整

- Broca（布洛卡失語症）：語法結構完整，詞序規則，利於模型學習特徵。
- NotAphasicByWAB（WAB 指標測試非失語症）：患者語言無明顯障礙，模型能準確分類。
- Anomic（命名性失語症）：語法完整，但詞彙命名困難，BERT 模型在詞彙學習上表現較好

(2) 語料數量較多

這些類別在 AphasiaBank 語料庫中的數據量相對較大，使 NLP 模型能夠有效學習其語言特徵，提高準確率。

(3) 語義與語法標記清晰

這些類別的語言結構較為標準，語音錯誤模式較規律，語音錯誤模式較容易通過 NLP 進行特徵提取。

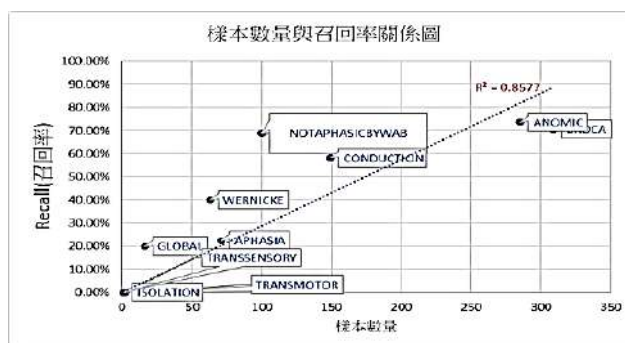
2. 為何 TransMotor、Isolation、TransSensory 類別的準確率較低？

影響因素如下：

(1) 樣本數量不足，語料稀缺，難以建立有效語言模型

- TransMotor（經皮層運動性失語）：樣本僅 8 個，模型難以學習其特徵。
- Isolation（孤立性失語症）：僅 1 個樣本，且該類患者通常缺乏完整句子輸出，導致 NLP 無法有效分類。

- TransSensory（經皮層感覺性失語）：樣本 2 個，語言理解能力低，影響語法分析。



【圖 20】樣本數量與召回率關係圖(研究者使用 EXCEL 製作)

- (2) 樣本數量與召回率呈現強烈線性正相關 ($R^2=0.8577$)，根據圖 20 顯示，語料的增加可以提升召回率。若類別語料提升至 250 篇以上，模型準確率可望提升至 70% 以上。若整體語料進一步擴增，準確率甚至可達 80%。我們利用語料充足的四個失語症類別進行訓練，如圖 21，準確率高達 90%，這也遠超 Cong et al. [5]對於兩個失語症類別的 76%-88%準確率。

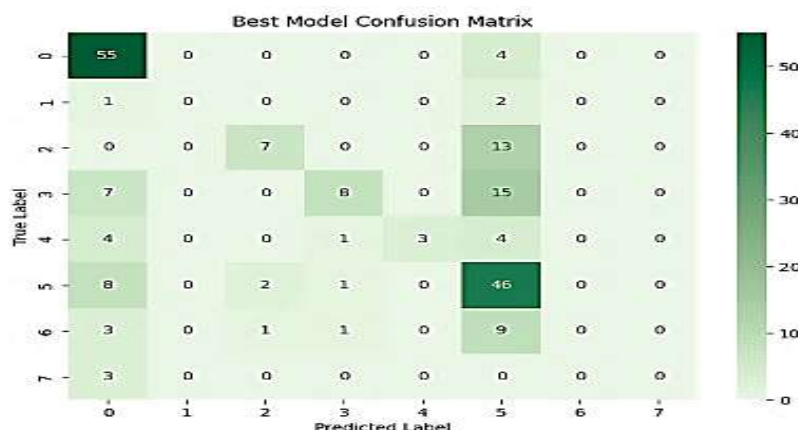
	precision	recall	f1-score
accuracy	0.901408	0.901408	0.901408
macro avg	0.887216	0.882397	0.884740
weighted avg	0.900892	0.901408	0.901100

【圖 21】對於多語料類別的 BERT 分類結果(本研究於 CMD 中執行訓練程式之畫面截圖)

- (3) 未來將擴充數據集，並引入影音數據，以強化肢體語言的標記，進一步驗證模型在語料數量稀少類別的分類能力。

三、隨機森林模型的分類結果

(一)、圖表展示



【圖 22】隨機森林模型之 Confusion Matrix（由本研究以 Python 程式繪製）

如圖 22，我們將 TF-IDF 向量與額外特徵結合，輸入隨機森林分類。當 $n_estimators$ 設為 150 時，準確率達 62%，優於傳統隨機森林。

伍、討論

如圖 23 與下列公式，本研究採用微調過的深度學習模型(BERT)與自訓練分類器對失語症類型進行分類，藉此探討深度模型相較於傳統模型的優勢。透過對兩種模型的數學分析，我們比較了其準確度及各自的優缺點，最終得出應該增強語料的結論。

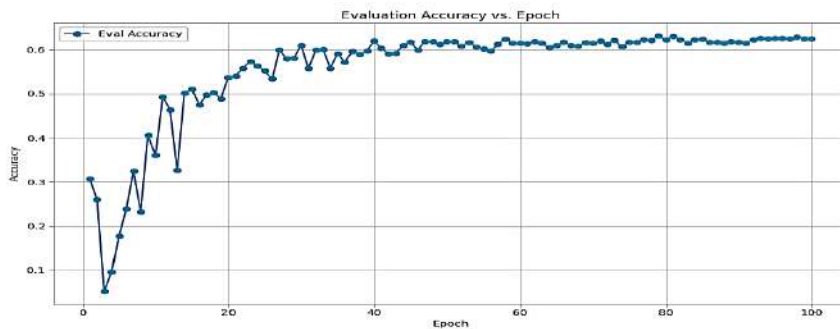
一、整體準確率(Accuracy)分析

(一)、公式定義

$$Accuracy = \frac{\sum TP}{\sum (TP + FP + FN + TN)}$$

(二)、計算

$$Accuracy = \frac{72+127+64+20+162+11+1}{731}$$



【圖 23】 模型準確度(由本研究以 Python 程式繪製)

(三)、分析

1. 優於傳統模型 NLP 模型：一般而言，深度學習準確率皆優於傳統 NLP 模型
2. 顯著高於隨機分類(RandomGuessing)：

- (1) 隨機分類的準確率為約 10%，本研究模型明顯優於此基準，證明模型有效學習了失語症的語言特徵。
- (2) 隨機分類的準確率是根據類別數量來計算。

$$P_{\text{random}} = \frac{1}{N}, \quad \text{其中：}$$

N 是類別總數，在本研究中，根據數據集有 10 個失語症類別（Broca, Wernicke, Anomic, Conduction, Global, TransMotor, TransSensory, Isolation, Aphasia, NotAphasicByWAB）。

因此，隨機分類的準確率為：

$$P_{\text{random}} = \frac{1}{10} = 10\%$$

- (3) 模型實用性驗證- Random Guessing 10%作為基準：
在類別分布不均衡的情況下，隨機分類準確率可能略高於 10%。本研究模型的準確率均顯著超越此基準，證明其優於隨機分類。
3. 在類別不均衡條件下，模型仍能保持較高準確率，彰顯其在輔助應用上的可行性。
 - (1) 類別數據嚴重不均衡（例如 Broca309 筆，TransSensory2 筆），對於傳統機器學習方法而言，這種不均衡數據會導致過擬合於較多樣本的類別，使準確率低於 60%。
 - (2) 深度學習模型（本研究達 62%以上）在多類別不均衡數據集上的表現顯然優於傳統機器學習，而我們的研究成果在四個語料充足的類別之分類準確度高達 90%。
4. 深度學習模型(BERT)能更有效學習不同類別的語言特徵，避免傳統方法對長文本處理的不足。

二、模型表現分析（Precision，Recall，F1-score）

- (一)、**選取代表類別進行比較分析**：為聚焦探討模型在多類別任務下的實際效能，本文選擇樣本數相對充足、分類表現（Precision、Recall、F1-score）具可比較性，且具明顯語言特徵，足以反映模型整體辨識能力的類別 0（Broca）、2（NotAphasicByWAB）、5（Anomic）三類進行分析模型整體辨識能力。

【表 4】各類別 Precision、Recall、F1-score 統計(表格內容由本研究整理彙編而成。)

類別	精確率 (Precision)	召回率 (Recall)	F1-score	樣本數
0	70.59%	70.59%	70.59%	102
2	76.05%	69.02%	72.36%	184
5	58.27%	73.64%	65.06%	220

(二)、個別類別分類表現分析

1. 類別 0- Broca，模型對此類別具有穩定的辨識能力

模型在類別 0 上的 Precision 與 Recall 同為 70.59%，顯示分類穩定且一致。模型能有效區分類別 0，誤判率與漏判率皆低，具良好辨識能力。

2. 類別 2 - NotAphasicByWAB，模型對此類別預測準確度較高

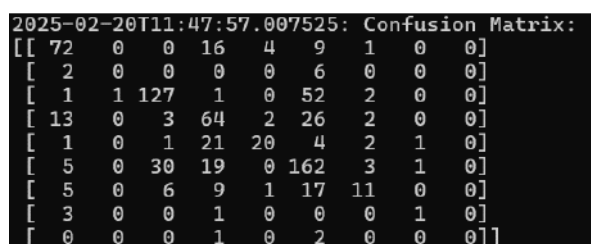
Precision 為 76.05%、Recall 為 69.02%，分類表現整體優異。模型對此類別具較強學習能力，能辨識多數樣本，且預測準確度高。

3. 類別 5 - Anomic，模型展現穩定偵測能力

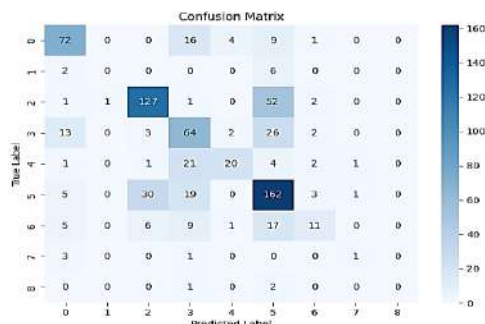
Recall 達 73.64%，顯示模型對此類具高度敏感度，能有效偵測大多數實際樣本。Precision 為 58.27%，可能因語句流暢但詞彙使用較不穩定，造成部分類別重疊。整體結果顯示模型已能辨識此類語言特徵，未來結合多模態特徵與樣本分布調整，有助進一步提升模型的適應能力與預測表現。

綜上所述，模型在三類代表性語者中皆展現穩定且具辨識力的分類能力。

三、混淆矩陣分析(本研究於 CMD 中執行訓練程式之畫面截圖及 Python 程式繪製，如下圖所示)



【圖 24】Confusion Matrix 混淆矩陣



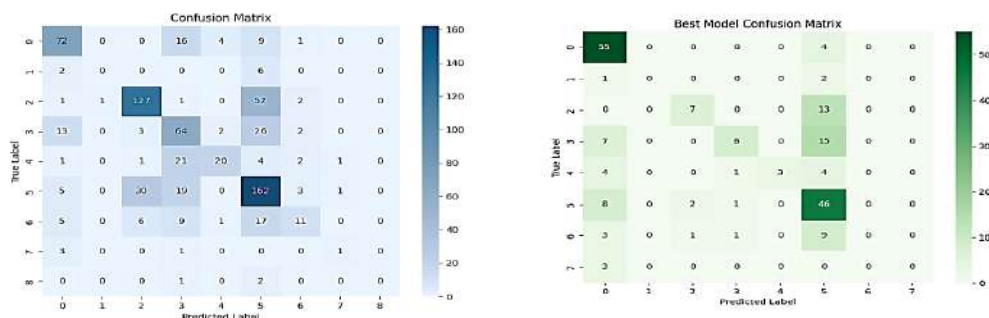
(一)、正確分類率

1. 類別 5 (162 次)、類別 2 (127 次) 在混淆矩陣中為深色區塊，表示模型對這些類別的分類效果較佳。
2. 模型能準確區分 Broca 失語症 (類別 2) 與 Anomic 失語症 (類別 5)，這是語言障礙診斷中重大的突破。
3. 大多數類別的預測結果集中在對角線，表示模型整體準確率較高。

(二)、錯誤分類、難以分類的類別

1. 類別 2 (127 次正確，但 52 次錯分為類別 5)：推論是由於 Broca 失語症與 Anomic 失語症的症狀部分重疊，使得模型偶爾混淆。
2. 類別 4 (召回率僅 40%)：可能因樣本數少 (50 筆)，導致模型學習不足，未來可透過數據增強(如增加語音與影音資料)提升分類效果。
3. 難以分類：文本數量不均與特徵差異，使 Isolation、TransSensory 等類別分類困難。而 Isolation 樣本過少未被抽取，遭刪除進位，矩陣類別僅顯示 0~8。

(三)、BERT 與隨機森林混淆矩陣比較



【圖 25】BERT(藍)與隨機森林(綠)混淆矩陣(由本研究以 Python 程式繪製)

從圖 25 中可以看出，兩模型的趨勢基本一致，因此我們得出結論：為了提升診斷能力，必須增加語料量，例如引入語音與影音資料。（由於採用 n_split 策略，語料量過少的類別已被刪除）

(四)、分析及貢獻

1. 模型可準確區分主要失語症類別，為輔助診斷提供幫助。
2. 錯誤分類情況符合不同失語症類型的語義重疊現象，證明模型的結果是可以被解釋的。

四、數據分布對模型的影響

(一)、類別數據不均衡

1. Broca（309 筆）與 Anomic（285 筆）佔據大多數，導致模型較容易學習這些類別的特徵。
2. Isolation（1 筆）與 TransSensory（2 筆）過少，導致這些類別的召回率為 0%，顯示數據不足對模型的影響。

(二)、文本的真實類別與預測類別

1. 如圖 26，此部分展示了每個語料的真實失語症類別與模型預測結果的對照結果，幫助我們理解模型在不同類別上的預測錯誤類型。

```
Sentence ID: S68_part4, True Type: ANOMIC, Predicted Type: NOTAPHASICBYWAB
Sentence ID: S620_part3, True Type: CONDUCTION, Predicted Type: ANOMIC
Sentence ID: S554_part5, True Type: WERNICKE, Predicted Type: CONDUCTION
Sentence ID: S746_part58, True Type: NOTAPHASICBYWAB, Predicted Type: NOTAPHASICBYWAB
Sentence ID: S991_part1, True Type: ANOMIC, Predicted Type: ANOMIC
Sentence ID: S348_part1, True Type: CONDUCTION, Predicted Type: CONDUCTION
Sentence ID: S645_part5, True Type: ANOMIC, Predicted Type: CONDUCTION
Sentence ID: S663_part1, True Type: BROCA, Predicted Type: ANOMIC
Sentence ID: S735_part46, True Type: NOTAPHASICBYWAB, Predicted Type: NOTAPHASICBYWAB
Sentence ID: S7, True Type: BROCA, Predicted Type: BROCA
Sentence ID: S342_part1, True Type: CONDUCTION, Predicted Type: CONDUCTION
Sentence ID: S661_part4, True Type: NOTAPHASICBYWAB, Predicted Type: NOTAPHASICBYWAB
Sentence ID: S741_part4, True Type: NOTAPHASICBYWAB, Predicted Type: NOTAPHASICBYWAB
Sentence ID: S992_part2, True Type: BROCA, Predicted Type: BROCA
Sentence ID: S18_part9, True Type: NOTAPHASICBYWAB, Predicted Type: NOTAPHASICBYWAB
Sentence ID: S267, True Type: BROCA, Predicted Type: BROCA
root@34ed46fe744e:/workspace/Neils-Thomass client_loop: send disconnect: Connection reset
```

【圖 26】PredictionResult（本研究於 CMD 中執行訓練程式之畫面截圖）

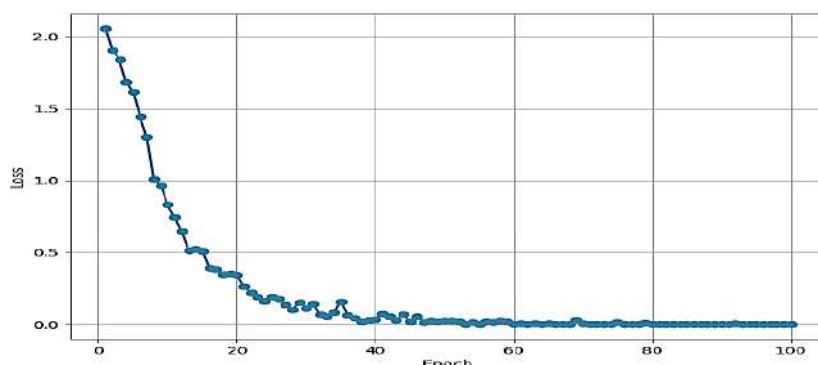
2. 被投入模型的資料型態(如圖 27)

```
2025-02-20T11:13:56.321565: =====
2025-02-20T11:13:56.322057: Sentence ID S766_part1:
2025-02-20T11:13:56.322084: Tokens: ['[CLS]', 'hmm', 'I', 'dialogue', 'I', 'I', 'do', 'ne', 'ne', 'ne', 'know', 'how', 'to', 'write', 'snake', 'I', 'dia
2025-02-20T11:13:56.322115: word_pos_ids: [-1, 8, -1, -1, 1, 10, 10, 10, 10, 6, 9, 3, 6, 12, -1, -1, -1, 1, 6, 2, 2, 2, -1, -1, -1, 5, 5,
2025-02-20T11:13:56.322228: word_grammar_ids: [[-1, -1, -1], [1, 0, 10], [-1, -1, -1], [-1, -1, -1], [-1, -1, -1], [1, 4, 1], [2, 4, 2], [2, 4,
2025-02-20T11:13:56.322289: word_durations: [-1.0, 160.0, -1.0, -1.0, -1.0, 90.0, 170.0, 170.0, 170.0, 170.0, 70.0, 210.0, 100.0, 370.0
2025-02-20T11:13:56.322307: =====
2025-02-20T11:13:56.322793: Sentence ID S766_part2:
2025-02-20T11:13:56.322818: Tokens: ['[CLS]', 'I', 'I', 'love', 'white', 'person', 'X', 'pl', 'I', 'dialogue', 'I', 'that', 'ne', 'cop', 'why', 'I', 'want', 'I
2025-02-20T11:13:56.322852: word_pos_ids: [-1, -1, 1, 6, 23, 12, 12, 12, -1, -1, -1, 34, 34, 34, 9, 1, 6, 6, 6, 6, 1, 6, 3, 6, 28, -1, -1, -1,
2025-02-20T11:13:56.322947: word_grammar_ids: [[-1, -1, -1], [-1, -1, -1], [1, 2, 1], [2, 0, 3], [3, 4, 28], [4, 2, 6], [4, 2, 6], [4, 2, 6], [
2025-02-20T11:13:56.323007: word_durations: [-1.0, -1.0, 30.0, 120.0, 260.0, 200.0, 200.0, 200.0, -1.0, -1.0, -1.0, 200.0, 200.0, 200.0,
2025-02-20T11:13:56.323025: =====
2025-02-20T11:13:56.323511: Sentence ID S766_part3:
2025-02-20T11:13:56.323536: Tokens: ['[CLS]', 'bike', 'the', 'key', 'I', 'pl', 'and', 'the', 'bike', 'do', 'ne', 'ne', 'ne', 'work', 'so', 'beg', 'they', '
2025-02-20T11:13:56.323569: word_pos_ids: [-1, 12, 25, 12, 12, 29, 25, 12, 10, 10, 10, 6, 8, 18, 1, 1, 1, 2, 2, 2, 2, 20, 12, 1, 10
2025-02-20T11:13:56.323662: word_grammar_ids: [[-1, -1, -1], [3, 0, 10], [1, 2, 12], [2, 0, 10], [2, 0, 10], [2, 0, 10], [1, 6, 15], [2, 3, 1
2025-02-20T11:13:56.132300: Total INV: 55872. Total PAR: 156302
```

【圖 27】放入模型的資料格式（本研究於 CMD 中執行訓練程式之畫面截圖）

五、損失值 (Loss) 與訓練過程

(一)、隨著每個訓練回合的 loss(損失)變化(如圖 28)：



【圖 28】Loss 和 Epoch 的關係圖（本研究以 Python 程式整理與輸出）

1. Loss 值在 20Epoch 內快速下降，顯示模型學習有效。
2. 40~50Epoch 之後 Loss 逐漸收斂，表示模型已基本學習完成。

(二)、如圖 29，這個是每一個訓練輪迴下的損失(loss)值及訓練過程



【圖 29】訓練過程（本研究於 CMD 中執行訓練程式之畫面截圖）

(三)、觀察結果

1. 隨著訓練次數的增加，損失值 (Loss) 逐漸降低，表明模型正逐步學習到語言分類的特徵。
2. 然而，在某些類別上，模型仍出現過擬合或訓練瓶頸，這可能與樣本數量較少的類別文本特性有關。準確率與數據數量之間存在高度正相關，而隨機森林混淆矩陣的分析也支持了這一結論。因此，未來計劃引入語音與影音資料來進一步優化模型。

六、對於語料充足的類別的分類結果

如前所述，我們對於語料充足的失語症類別(四種)分類結果，如下圖所示，90%的準確率已經超越了 Cong et al.[5]的 76%-88%對於雙類別分類的結果。

	precision	recall	f1-score
accuracy	0.901408	0.901408	0.901408
macro avg	0.887216	0.882397	0.884740
weighted avg	0.900892	0.901408	0.901100

【圖 30】多數量語料的分類準確度（本研究於 CMD 中執行訓練程式之畫面截圖）

陸、結論

本研究透過階段性實驗設計與深度學習技術，成功開發一套結合 BERT 模型與自訂分類器的失語症診斷系統，並藉由詞彙、語法與長文本處理技術，顯著提升診斷準確率。透過綜合 NLP 模型、機器學習與超參數搜尋的驗證，本研究不僅在準確率、召回率（recall）與 F1-score 等指標上優於現有參考文獻(如圖 31)，並且在處理不同類型的失語症患者語言樣本時展現更高的適應性與準確度。

	precision	recall	f1-score
accuracy	0.901408	0.901408	0.901408
macro avg	0.887216	0.882397	0.884740
weighted avg	0.900892	0.901408	0.901100

One predictor variable at a time	Accuracy	Precision	Recall	F1-score
SVM (Mistral-7B)	0.88	0.89	0.88	0.88
SVM (GPTNeo-2.7B)	0.84	0.85	0.84	0.84
SVM (GPTNeo-1.3B)	0.86	0.89	0.86	0.85
SVM (DistilGPT-2)	0.76	0.84	0.76	0.74
SVM (GPT-2)	0.71	0.82	0.71	0.69

【圖 31】Cong et al. (2024) [5]不同語言模型特徵搭配 SVM 分類之性能比較
（左圖來自本研究於 CMD 中執行訓練程式之畫面截圖，右圖來自 Cong et al. (2024) [5]）

一、研究貢獻與創新

(一)、突破 NLP 在失語症分析領域的瓶頸：

1. 傳統分析方法（如 WAB、BDAE）仰賴專業醫療人員，過程繁瑣且難以標準化。
2. 本研究透過 **BERT 預訓練模型**學習患者語言特徵，並結合自訂分類器與超參數搜尋，使分類準確率達 90%，相較 Cong et al.[5]的 GPT 模型（76%-88%），提升 2%至 14%的準確率。

(二)、階段性實驗設計與模型效能提升：

1. 第一階段 觀察模型對詞彙單位的學習效果，證明其基礎識別能力（準確率 47%）。
2. 第二階段 加入語法結構與詞性標註，並引入肢體語言標記，提升語言理解能力（準確率提升至 57%）。

3. 第三階段 透過文本切分技術提升模型對長文本的適應性，並將準確率提升至 62%（進一步提升訓練回數可達 70%）。
4. 少樣本分類模型的改進：

Cong et al.[5] 使用 GPT、GPT2 與 GPTNeo 在 Broca 與 Wernicke 失語症分類表現優異，展現 GPT 在 NLP 任務與少樣本學習（Few-shot Learning）上的潛力 (Brown et al.[13]; Cong et al.[5])。

本研究參考 Cong et al.[5]，改良少量分類模型，聚焦語料數量最多的四類失語症（由原 10 種縮減），並強化分類策略，即使各類失語症語言特徵高度重疊，模型仍能準確分類。結果顯示，分類準確率提升 2% 至 14%，進一步展現本研究在語言障礙評估的潛力與應用價值。

(三)、應對資料不均衡問題，使模型能更好處理樣本較少的語料類別：

1. 部分失語症類型（如 Isolation、TransSensory）樣本數較少，現有機器學習模型難以有效學習其特徵。
2. 本研究透過 WeightedRandomSampler、Optuna 超參數搜尋與隨機森林驗證，找出最佳模型組合，並顯示當語料數量增加至 250 篇以上，模型準確率可改善至 70%以上。
3. 語料庫的擴充對模型效能至關重要，未來應持續加強語料來源，以進一步提高診斷準確率。

二、模型效能分析與應用價值

(一)、BERT 模型的分類效能與優勢

1. BERT 模型在處理長文本與多層次語義關係時展現良好性能，適用於失語症分類任務。
2. 混淆矩陣分析顯示，BERT 模型在 Broca、Anomic、Conduction 失語症的辨識效果優異，這對於精細分類不同類型的失語症具有重要應用價值。

(二)、降低醫療負擔，改善診斷效率

1. 現有的失語症診斷依賴專業醫療資源，如 WAB 測試，但過程繁瑣且需要大量時間，而本研究的 AI 系統可作為輔助診斷工具，自動分析患者語音與語言結構，提供即時診斷結果，加速醫療流程減少醫療人員的負擔。
2. 結合遠距醫療應用，本系統可幫助資源有限的偏遠地區，提升語言障礙患者的醫療可及性。

三、未來發展方向

- (一)、擴展語料庫與多模態數據分析：目前引用的 AphasiaBank 語料以文本為主，未來期望結合語音、EEG、MRI 等多模態數據，提升計算能力並獲取更多運算資源，以開發分析系統提升準確率。
- (二)、AI 診斷結果可與臨床測試（如 WAB、QAB）結合，建立更完整語言評估系統。
- (三)、NLP 技術應用於醫療機構：應用於醫院與長照機構，輔助更準確的失語症診斷。
- (四)、普及 AI 診斷工具：照護者、護理師等非專業醫療人員參與早篩，增進患者福祉。

柒、參考文獻

一、中文部分

【中文書籍】

- [1]、Ekman, M. (2024). *跟 NVIDIA 學深度學習！從基本神經網路到 CNN・RNN・LSTM・seq2seq・Transformer・GPT・BERT...*，紮穩機器視覺與大型語言模型 (LLM) 的建模基礎 (哈雷, Trans.). 旗標科技股份有限公司.
- [2]、孫小文、王薪宇、楊談 (2023)。 *ChatGPT 原理：從 PyTorch 中的 NLP 功能讓你一腳跨入自然語言處理*。深智數位出版社。

【中文網站文章】

- [3]、臺中榮民總醫院. (2025, January 13). 失語症簡介. 臺中榮民總醫院. Retrieved from <https://www.vghtc.gov.tw/UnitPage/RowViewDetail?CompanyID=e8e0488e-54a0-44bf-b10c-d029c423f6e7&UnitDefaultTemplate=1&UnitID=ed2fa610-1b99-4a70-b977-918c87e28e25&WebRowsID=c6ff2d13-190a-4992-bc87-27def49641d0>

- [4] 、 Topic News. (n.d.). 认识失语症 aphasia. *Topic News*. Retrieved from <https://www.topicnews.cn/health/get-to-know-aphasia>

二、外文部份

【學術期刊論文 (Journal Articles)】

- [5] 、 Cong, Y., LaCroix, A. N., & Lee, J. (2024). Clinical efficacy of pre-trained large language models through the lens of aphasia. *Scientific Reports*, 14, Article 15573. <https://doi.org/10.1038/s41598-024-66576-y>
- [6] 、 Goral, M., Antolovic, K., Hejazi, Z., & Schulz, F. M. (2024). Using a translanguaging framework to examine language production in a trilingual person with aphasia. *Clinical Linguistics & Phonetics*. <https://doi.org/10.1080/02699206.2024.2328240>
- [7] 、 Mahmoud, S. S., Cheng, W., Wang, Y., Chen, J., Zhong, Y., & Fang, Q. (2025). A rapid aphasia assessment battery for non-written Chinese dialect. *IEEE Access*, 13, 26456 – 26470. <https://doi.org/10.1109/ACCESS.2025.3539613>
- [8] 、 Wilson, S. M., Eriksson, D. K., Schneck, S. M., & Lucanie, J. M. (2018). A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLOS ONE*, 13(2), e0192773. <https://doi.org/10.1371/journal.pone.0192773>
- [9] 、 Dronkers, N. F., Plaisant, O., Iba-Zizen, M. T., & Cabanis, E. A. (2007). Paul Broca's historic cases: High resolution MR imaging of the brains of Leborgne and Lelong. *Brain*, 130(5), 1432 – 1441. <https://doi.org/10.1093/brain/awm042>
- [10] 、 Denier, C., Chassin, O., Vandendries, C., Bayon de la Tour, L., Cauquil, C., Sarov, M., Adams, D., & Flamand-Roze, C. (2016). Thrombolysis in stroke patients with isolated aphasia. *Cerebrovascular Diseases*, 41(3-4), 163 – 169. <https://doi.org/10.1159/000442303>
- [11] 、 MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, 25, 1286 – 1307. Retrieved from <https://talkbank.org/share/citation.html> 致謝 AphasiaBank 提供語料：NIH-NIDCD R01-DC008524 (2022-2027)。註：根據 AphasiaBank 的引用規則，使用其資料時，同時致謝以下資助：NIH-NIDCD R01-DC008524 (2022-2027)。
- [12] 、 Fromm, D., Chern, S., Geng, Z., Kim, M., Greenhouse, J., & MacWhinney, B. (2024). Automated analysis of fluency behaviors in aphasia. *Journal of Speech, Language, and Hearing Research*. Advance online publication. <https://doi.org/10.1044/2024.JSLHR-23-00659>

【arXiv 預印本 (Preprints)】

- [13] 、 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint*, arXiv:2005.14165.
<https://doi.org/10.48550/arXiv.2005.14165>
- [14] 、 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint*, arXiv:1706.03762. Retrieved from
<https://arxiv.org/pdf/1706.03762>

【會議論文 (Conference Papers)】

- [15] 、 櫻井勇輔, 内田智之, 川本佳代, 重田大門, & 小崎和治. (2023). BERT を用いた失語症者リハビリテーション支援システムにおける択一問題の自動生成. *日本人工知能学会第37回全国大会*. Retrieved from <https://www.ai-gakkai.or.jp/jsai2023/>
- [16] 、 Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171 – 4186). <https://aclanthology.org/N19-1423.pdf>

三、網路資源

【網站資源 (Website Resources)】

- [17] 、 MacWhinney, B. (n.d.). AphasiaBank: A shared multimedia database for the study of aphasia. *Carnegie Mellon University*. Retrieved February 9, 2025, from <https://aphasia.talkbank.org/>
- [18] 、 Stroke Engine. (n.d.). Boston Diagnostic Aphasia Examination (BDAE). *Stroke Engine*. Retrieved February 11, 2025, from <https://strokengine.ca/en/assessments/boston-diagnostic-aphasia-examination-bdae/>
- [19] 、 Stroke Engine. (n.d.). Western Aphasia Battery (WAB). *Stroke Engine*. Retrieved February 11, 2025, from <https://strokengine.ca/en/assessments/western-aphasia-battery-wab/>
- [20] 、 Canva. (2024). Icons and illustrations used in research visualization. *Canva*. Retrieved from <https://www.canva.com> (本研究自製圖表部分素材來自 CANVA 網站)

- [21] 、 MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates. Retrieved from <https://talkbank.org/share/citation.html>(感謝以下資助：NICHD HD082736。註：根據 **TalkBank** 的引用規則，使用語料時，需致謝提供語料)

【政府出版物 / 技術報告 (Government Reports / Technical Reports)】

- [22] 、 National Institute on Deafness and Other Communication Disorders. (2015, December). Aphasia (NIH Pub. No. 97-4257) [PDF]. *National Institute on Deafness and Other Communication Disorders*. Retrieved February 19, 2025, from <https://www.nidcd.nih.gov/sites/default/files/Documents/health/voice/Aphasia.pdf>
- [23] 、 OpenAI. (2020). GPT-3: Language models are few-shot learners. *OpenAI*. Retrieved from <https://arxiv.org/abs/2005.14165>

【影音資源 (Video / Audiovisual Media)】

- [24] 、 李宏毅 (Hung-yi Lee). (2019, June 7). ELMO, BERT, GPT [Video]. *YouTube*. Retrieved from <https://www.youtube.com/watch?v=UYPa347-DdE>

【英文網站文章 (Webpage on a Website)】

- [25] 、 D' Alessio, V. (2022, July 4). Lost for words – the devastation caused by aphasia. *Horizon Magazine*. Retrieved February 19, 2025, from <https://projects.research-and-innovation.ec.europa.eu/en/horizon-magazine/lost-words-devastation-caused-aphasia>
- [26] 、 Wikipedia contributors. (n.d.). Carl Wernicke. *Wikipedia, The Free Encyclopedia*. Retrieved February 11, 2025, from <https://zh.wikipedia.org/wiki/卡尔·韦尼克>
- [27] 、 Huang, J. (2023, August). Aphasia. *MSD Manual Professional Version*. Retrieved from <https://www.msmanuals.cn/professional/neurologic-disorders/function-and-dysfunction-of-the-cerebral-lobes/aphasia>
- [28] 、 National Aphasia Association. (n.d.). What is Aphasia? *National Aphasia Association*. Retrieved February 11, 2025, from <https://www.aphasia.org/>

【評語】 052505

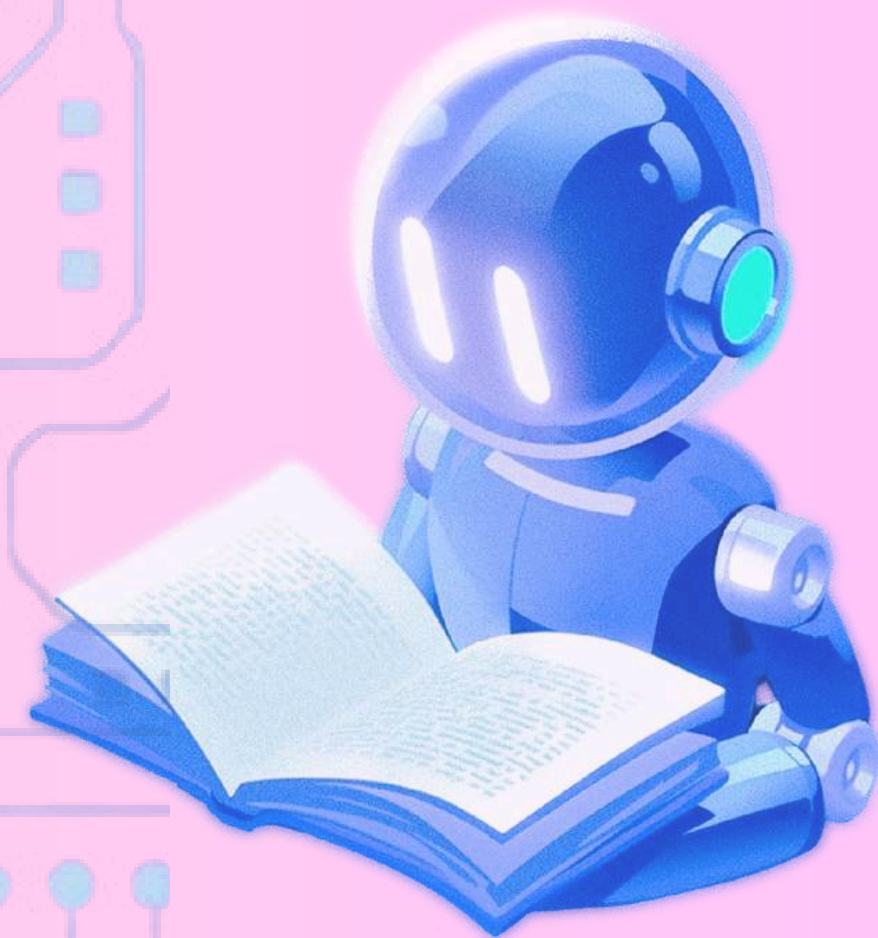
1. 本作品結合 BERT 模型與自訓練分類器進行失語症分類研究。此作品相內容描述完整，具有創意，實驗分析詳細，具應用性且是大眾所會關注議題之一，亦為未來提供更具針對性的個人化治療建議奠定基礎。
2. 整體而言，此研究具創新性、技術性與實務價值，惟目前僅為採用語料庫資料進行實驗，未來可結合臨床測試，建立更有說服力的研究成果。

作品海報

智慧 · 進化艙



語言革命，解密失語症分類



伍、研究架構與流程

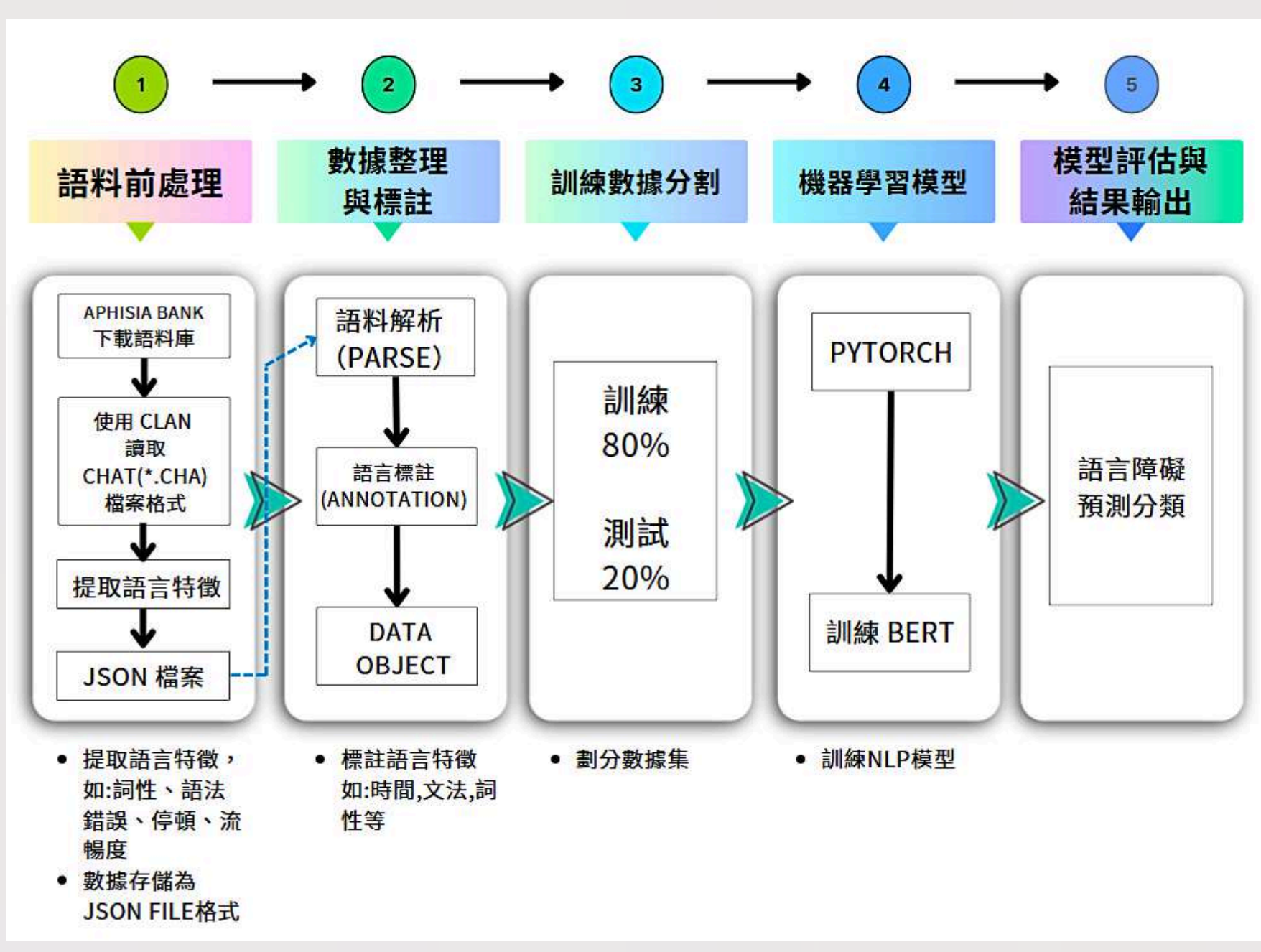
研究架構及流程如右圖，說明如下：

一、語料前處理

從APHASIABANK下載資料，提取詞性、語法、語音時長特徵，轉換為JSON格式。

二、數據整理與標註

語料解析並標記POS(詞性)、文法結構、時間資訊及肢體語言，確保模型能學習醫病對話的特徵。



三、訓練數據分割

80%作為訓練集，20%作為測試集，確保模型分析的準確性。

四、機器學習模型訓練

利用PYTORCH訓練BERT模型，適應醫療語境。

五、模型評估與結果輸出

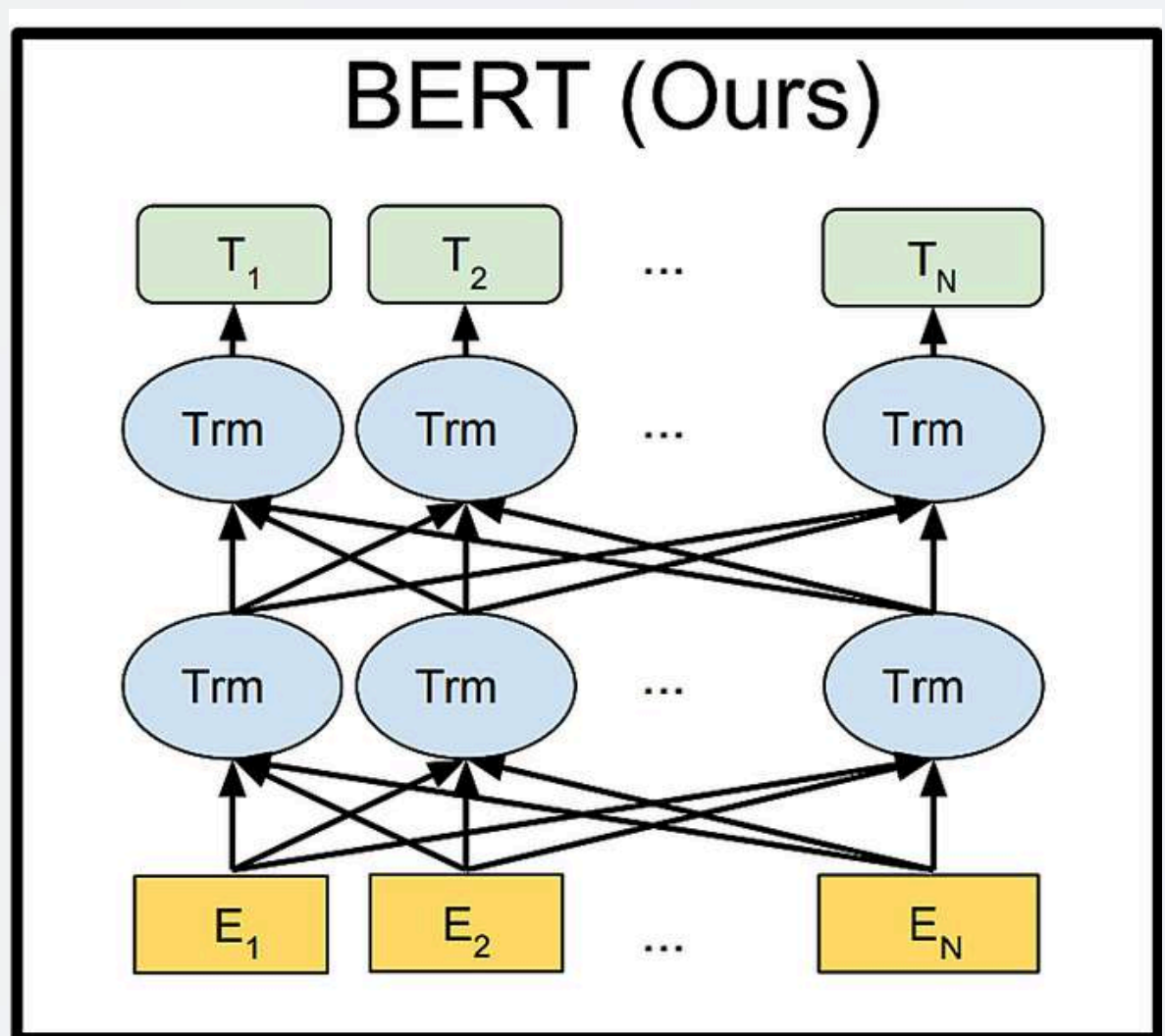
透過分類指標（ACCURACY、RECALL、F1-SCORE）評估模型效能，並生成診斷預測結果。

陸、研究方法與過程

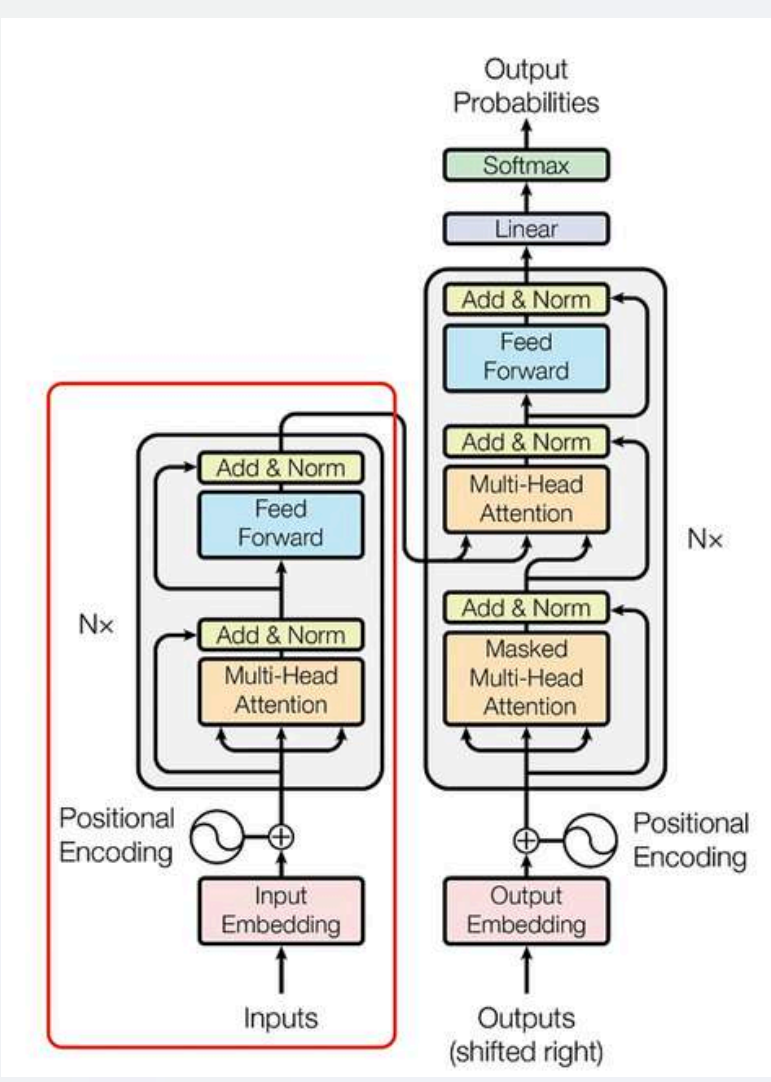
一、BERT模型架構與標記方法

研究中，我們採用Transformer編碼器（Encoder）架構，並利用BERT作為語言表示模型，以適應醫病對話場景。

BERT具備雙向語境學習能力，能夠理解對話中的語境變化，提升醫療語言的解析準確度。

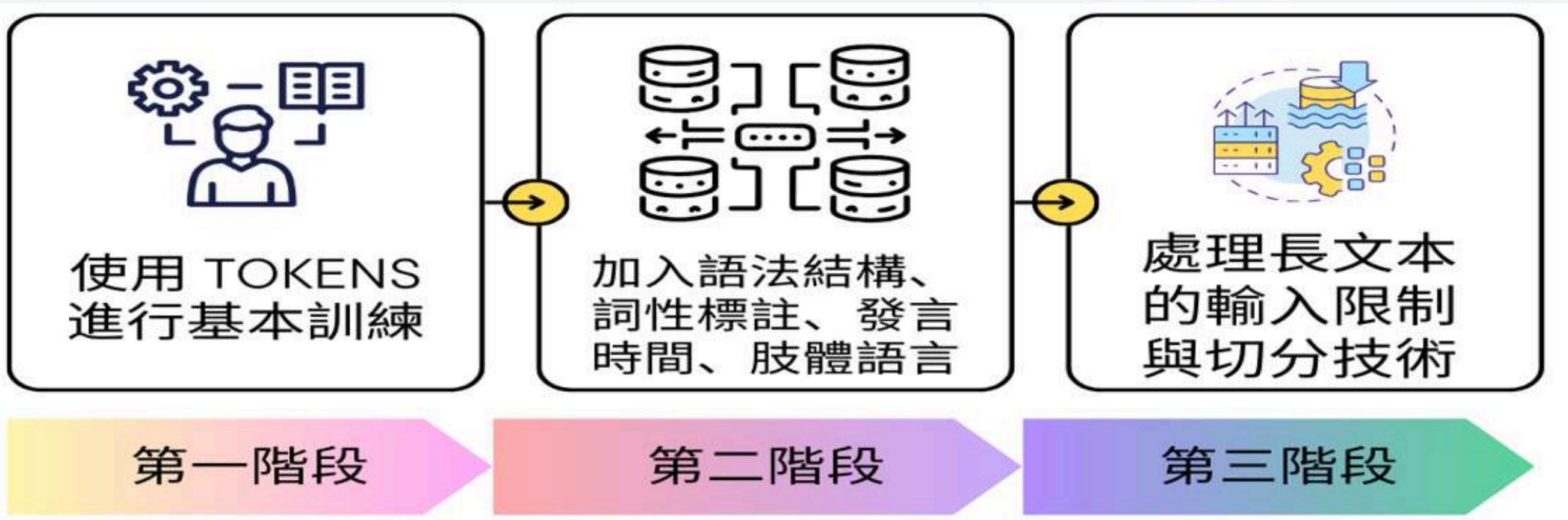


BERT輸入模式舉例
([16] Devlin et al. (2019))

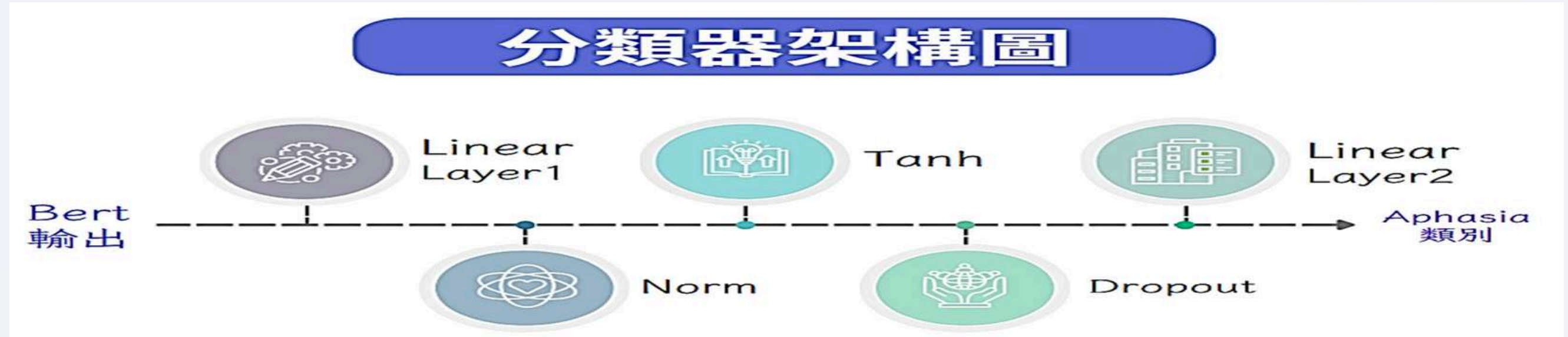


[BERT 架構圖]
Transformer (BERT uses left part, encoder.)

二、BERT三階段語料標記方法



三、自定義分類器架構



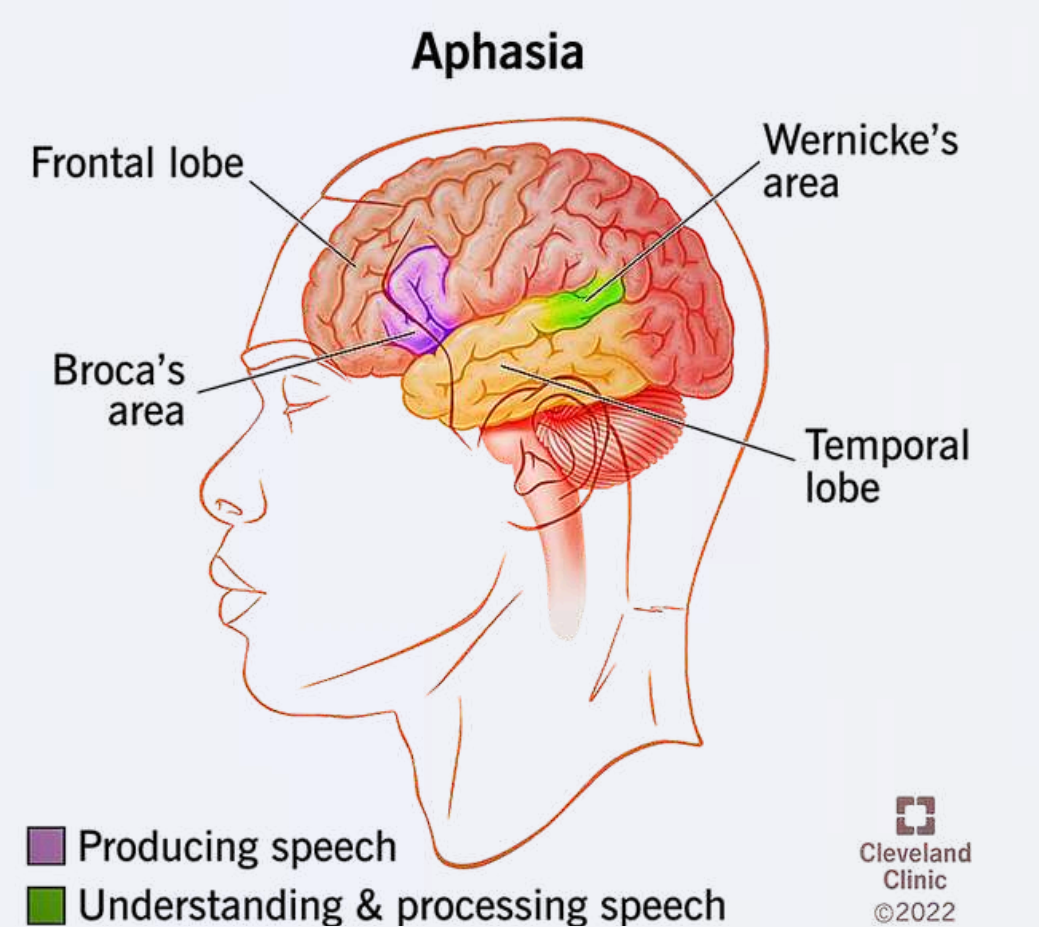
本研究搭配BERT訓練自定義分類器（CustomClassifier），如上圖：

- Linear層1：將BERT的輸出轉換為512維向量。
- Normalization層：標準化數據，避免梯度爆炸已導致模型偏化。

- Tanh層：採用tanh激活函數以提升分類效能。（根據Optuna超參數搜尋結果，此配置表現最佳）
- Dropout層：模擬遺忘機制，避免模型過擬合。（根據Optuna超參數搜尋結果，採0.2為丟失率）
- Linear層2：將最終輸出轉換為失語症類型標籤。

四、AphasiaBank 失語症類型

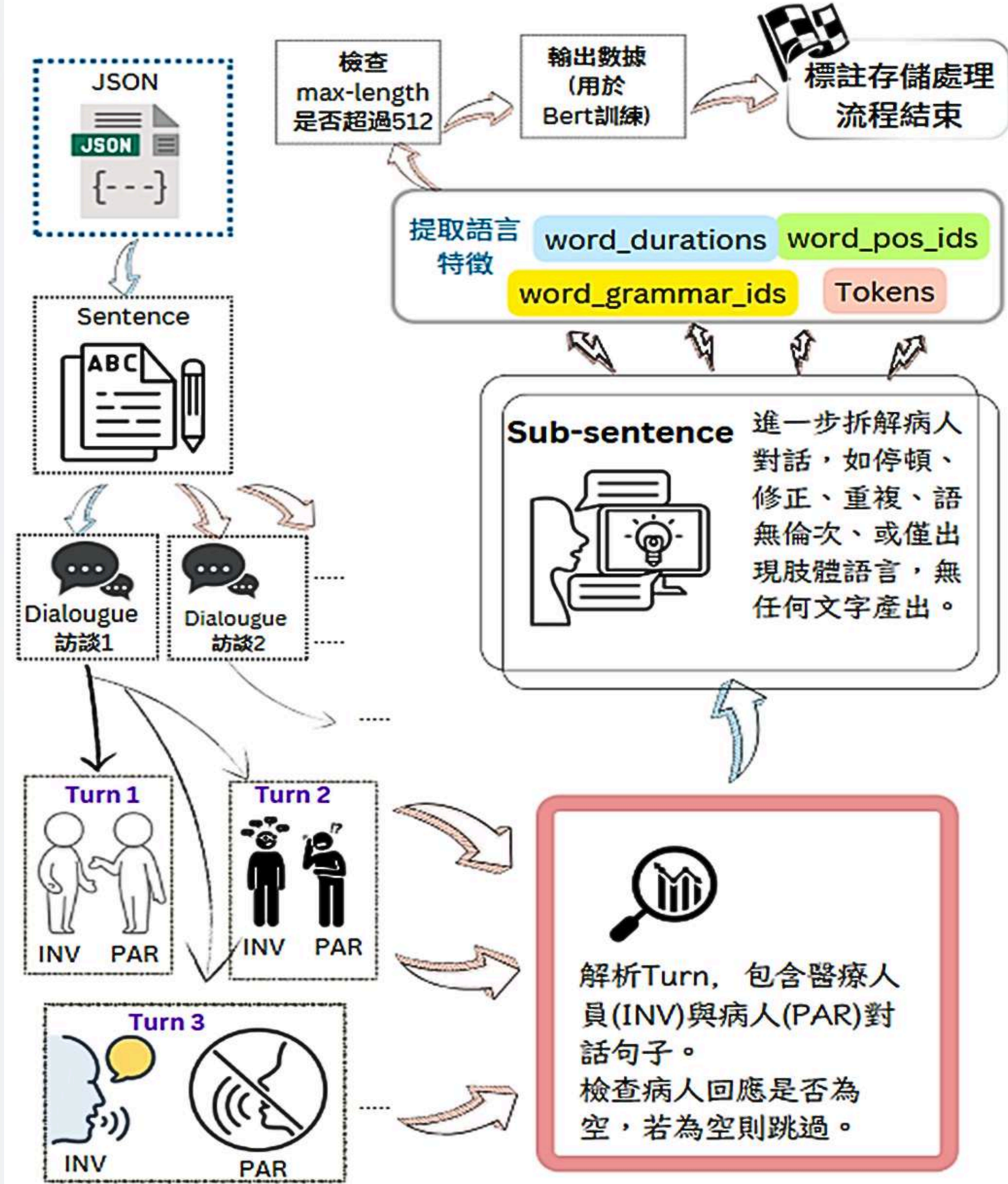
右表分類將用於預測受訪者的症狀，以提升醫療人員對不同患者的治療效率。



圖片來源：克利夫蘭診所
請<https://my.clevelandclinic.org/health/diseases/5502-aphasia>

Class Label	分類	中文名稱	說明
0	Broca	布洛卡失語症	言語緩慢、費力，句子簡短但理解良好。
1	TransMotor	經皮層運動性失語	言語流暢度下降，但理解與複述能力正常。
2	Not AphasicBy WAB	WAB 篩查非失語症	WAB 測試正常，無明顯失語症。
3	Conduction	傳導性失語症	理解能力良好，重複障礙（伴隨頻繁的語言錯誤），閱讀理解能力良好。
4	Wernicke	韋尼克失語症	流利性失語症，句子不連貫，不相關的字混雜其中。
5	Anomic	命名性失語症	他們的講話雖然語法形式和輸出流暢，但卻充滿了模糊的迂迴表達和且慢的表達。
6	Aphasia	失語症	表現為一般性語言障礙。
7	Global	全局性失語症	這是失語症最嚴重的形式，患者無法辨識出幾個單字及讀寫。
8	Isolation	孤立性失語症	指失語症患者沒有運動肢體缺陷。
9	TransSensory	經皮層感覺性失語	語言流暢，但理解顯著不足。與韋尼克失語症類似，但複述能力正常。

五、資料預處理步驟



如左，說明如下：

- 詞性特徵處理：**從% MOR 註釋提取詞性標籤，轉換為數字代碼（WORD_POS_ID）。
- 文法特徵處理：**拆分% GRA 註釋，提取詞性與語法關係（WORD_GRAMMAR_ID）。
- 語音長度處理：**從% WOR 註釋提取音節時長，分析語音流暢度（WORD_DURATION）。

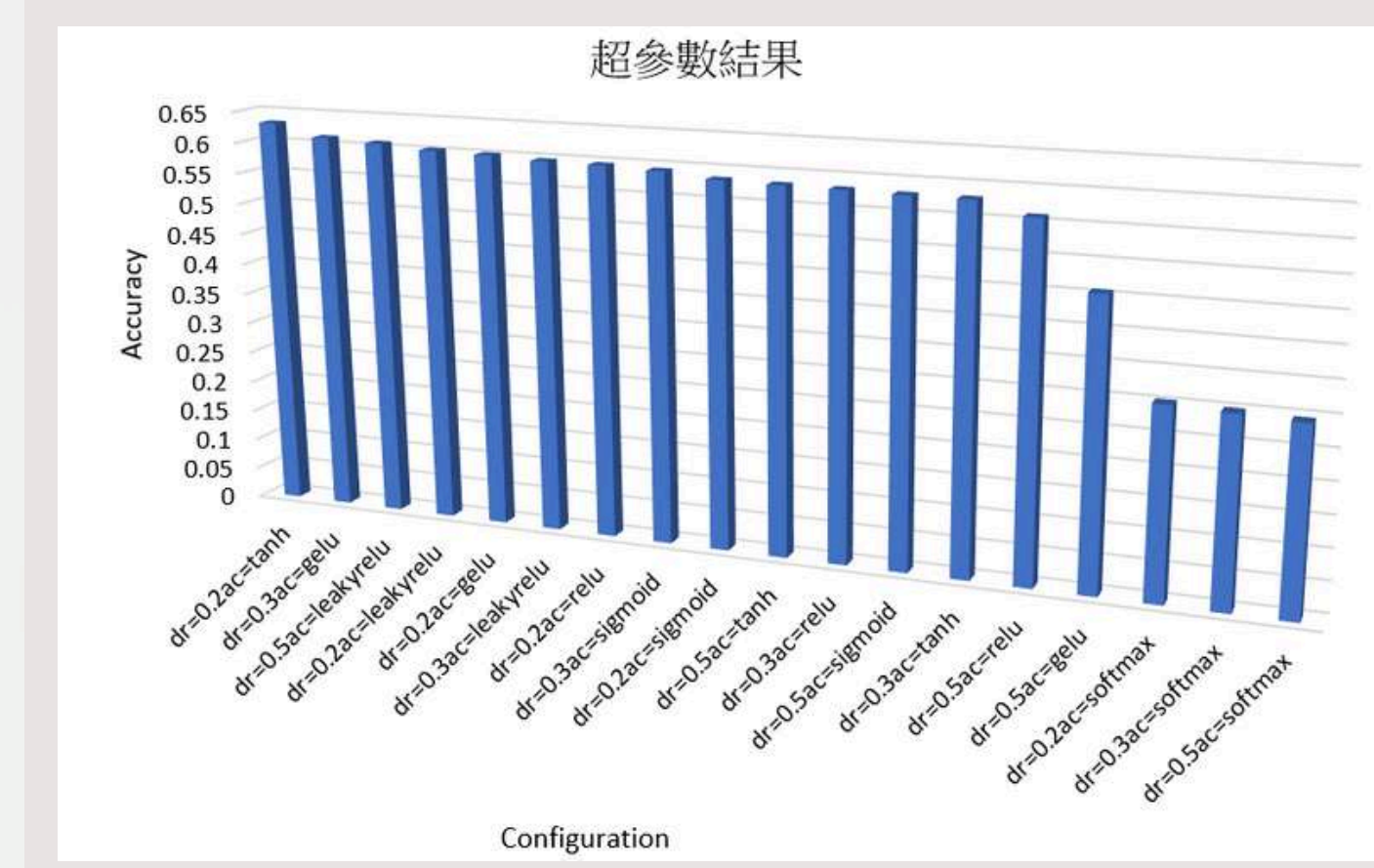
六、JSON 儲存解析

透過BERT分析語境變化，預測失語症類型。本研究使用JSON格式來儲存與解析受訪者語音轉錄與NLP分析結果。



七、超參數搜尋

本研究採用OPTUNA模組進行超參數搜尋，以改善模型表現，最終找出最佳分類器架構。下圖展示表現排名前18的參數組合。



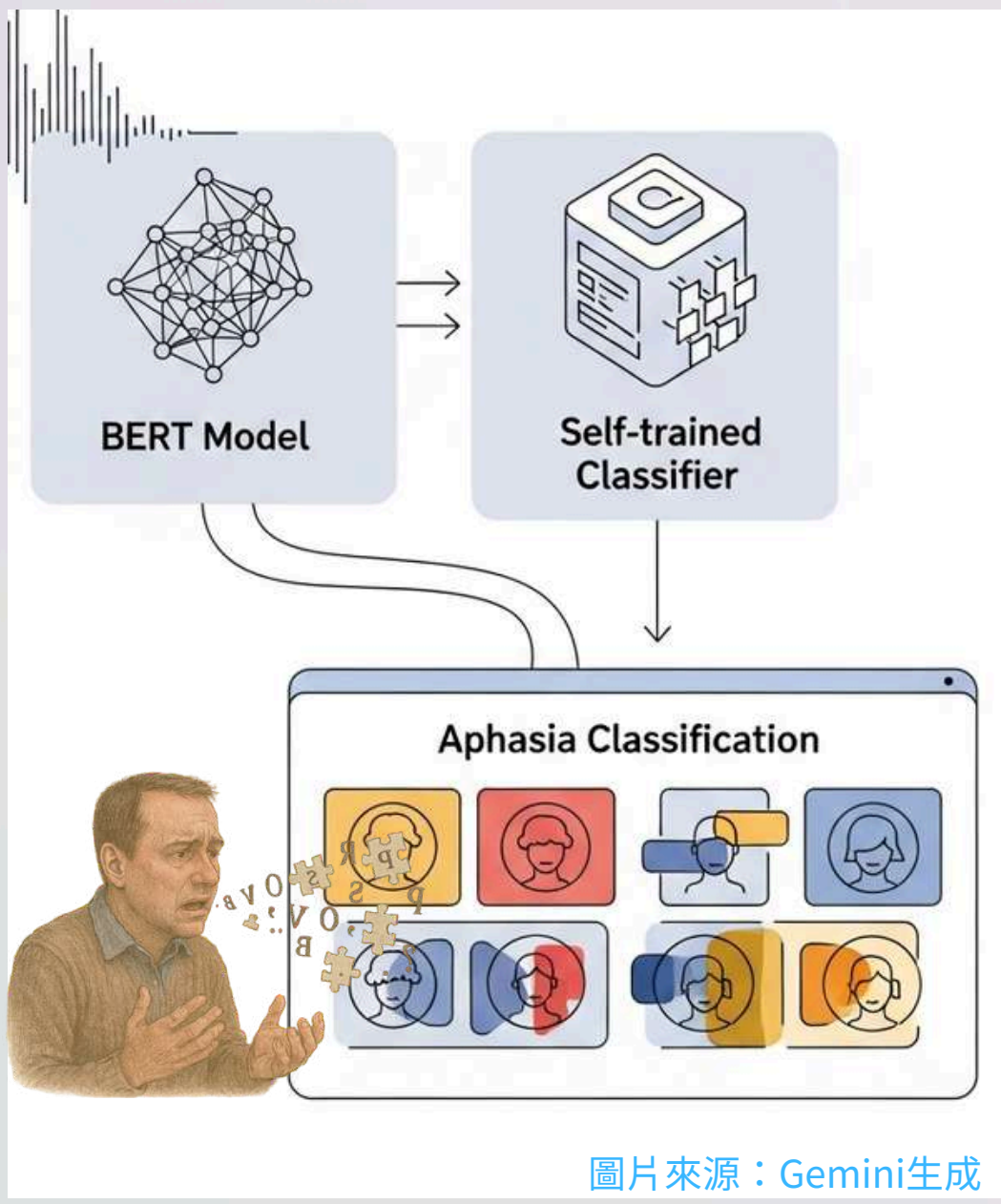
- 超參數（HYPERPARAMETER）：
- 最佳LEARNINGRATE：5E-5
 - BATCH SIZE：64
 - DROPOUT RATE：0.2
 - 最大TOKEN長度：512
 - 激活函數：TANH

八、Weighted Random Sampler

本研究亦採用WEIGHTED RANDOM SAMPLER，使得每一種語料被採取訓練的比重設定為數量的倒數，改善過擬和現象。

```
from torch.utils.data import WeightedRandomSampler, RandomSampler
train_labels = [sample["labels"].item() for sample in train_dataset]
label_counts = Counter(train_labels)
sample_weights = [1.0 / label_counts[label] for label in train_labels]
```

壹、摘要



- 本研究提出一套結合BERT模型與自訓練分類器的失語症分類系統，並探討不同模型在分類任務中的效能表現。
- 靈感來自櫻井等人(2023)提出的BERT四選一選擇題模型模型，並參考Cong et al. (2024)GPT模型在臨床應用中的研究，探索自訓練分類器增強BERT判別能力的潛力。
- 本研究的不同之處，在於模型訓練係基於醫生標註的分類結果進行深度學習，以提升分類效能與個人化治療效果。
- 考慮語料庫規模及醫學倫理，本研究採用AphasiaBank(全球最具權威的語料資源之一)的英文訪談文本進行模型訓練，同時移除所有個人隱私資訊。
- 三階段模型訓練策提升分類準確率，有效改善傳統診斷流程，有助於減輕醫療負擔。

貳、研究動機及背景

一、研究動機

- 在中國醫藥大學附設醫院為期兩個月的志工服務學習中，觀察到失語症患者在溝通上的困境與診斷資源的匱乏。
- 全球約四分之一的 25 歲以上成年人在其一生中可能罹患中風，其中約三分之一的患者易罹患失語症。
- 隨著人口老化，失語症患者逐年增加。傳統診斷仰賴 WAB 測驗與 fMRI，耗時（約 2-3 小時），且高度依賴專業人員，增加醫療負擔。為提升效率，本研究結合深度學習（如 BERT），開發快速且準確的 AI 診斷系統。

各國失語症患者統計表

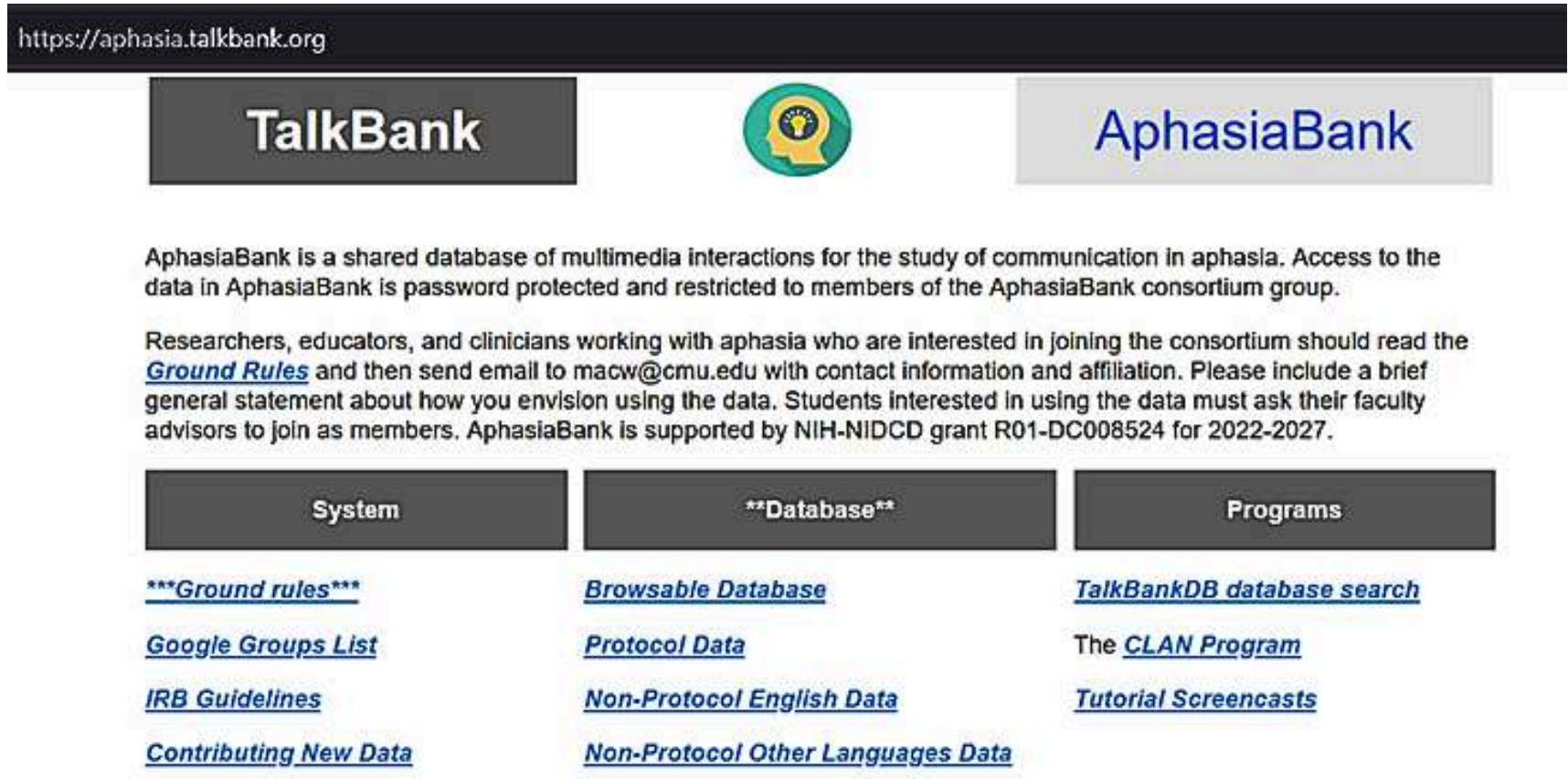
國家/地區	患者統計	資料年度	資料來源
美國	約100萬人 (每年新增：約18萬人)	2025	NAA
台灣	約4-6.5萬人	2025	臺中榮總
歐洲	約30萬人	2022-2023	參考文獻
新加坡	約1.5萬人(估算)	2023	參考文獻
日本	約20-50萬人	2022	參考文獻



[資料來源：台北榮民總醫院復健醫學部蔡泊意醫師的研究]
全球25歲以上成年人中，1/4可能會發生中風，約有1/3腦中風患者易發失語症

二、研究資料來源

- 本研究採用AphasiaBank(全球最具權威的語料資源之一)的英文訪談文本進行模型訓練。
- 考慮語料庫規模及醫學倫理，同時移除所有個人隱私資訊。

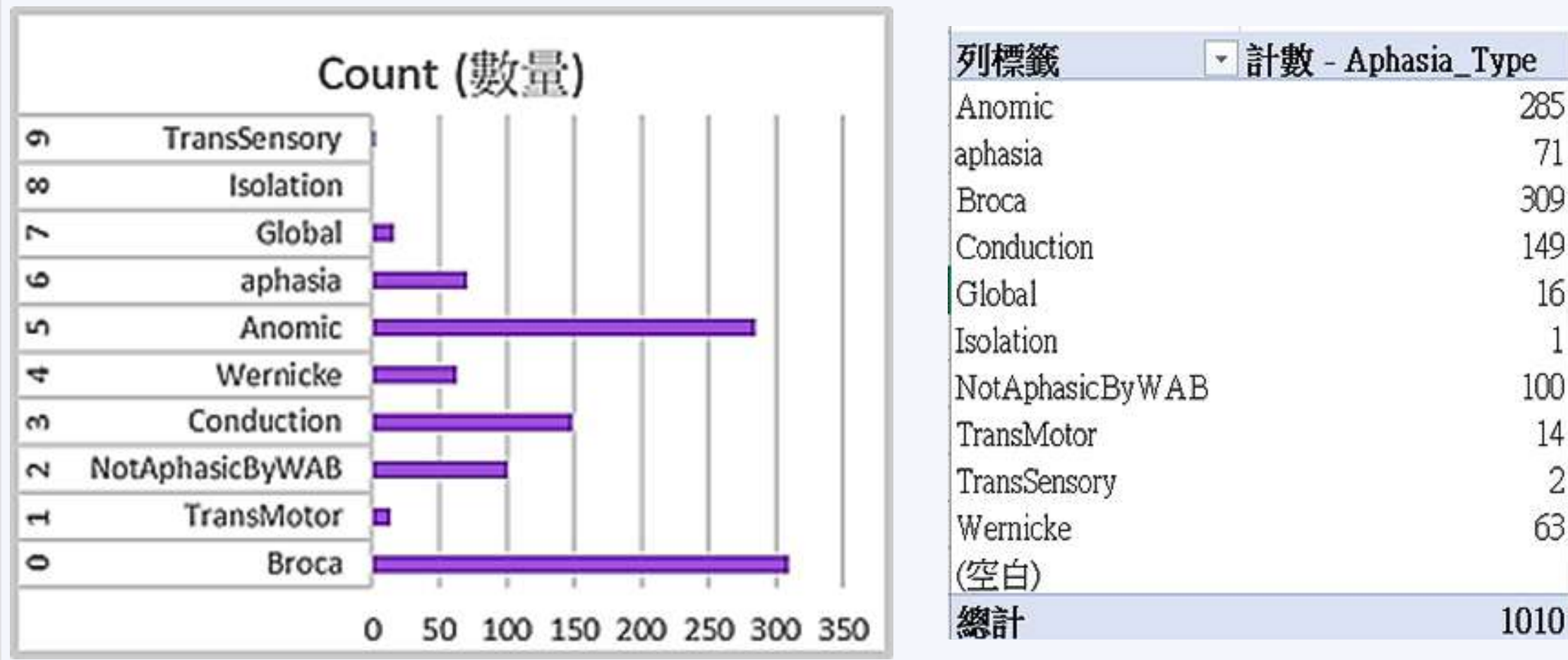


圖片來源：Aphasia Bank網頁

三、語料資訊

- 語料來源：來自AphasiaBank，1023篇參與者（Participants）影片檔。(13筆資料未標記失語症類型資料，其中又有5筆資料未標記性別。最後投入1010篇資料做訓練。)
- 性別分佈：男性592人，女性426人。
- 年齡區分：60歲以下406人，60歲以上617人。
- 受訪者失語症類型，統計分佈如下2張圖表

(致謝Aphasia Bank提供語料：NIH-NIDCD R01-DC008524 (2022-2027))



[受訪者失語症類型統計分佈]

參、研究目的

- 失語症嚴重影響患者的溝通能力及心理健康，導致患者在日常生活與社交互動中面臨重大障礙。
- 隨著全球人口老化趨勢與中風病例不斷增加，失語症患者數量迅速攀升，已大幅超過現有醫療人員的負荷能力，使醫療診斷資源面臨嚴重不足的困境。

本研究旨在開發一套以BERT模型為核心的AI失語症分類系統，透過醫生標註的臨床語料（CHAT）進行深度學習，綜合分析詞性、語法結構、語音長度與肢體語言等多重特徵，以期顯著提升失語症分類的準確度與效率，進一步縮短診斷時程並有效降低醫療負擔。

肆、研究設備

一、硬體設備及作業系統

- (一)、個人電腦：Windows 11Intel i7 CPU，32GB RAM，NVIDIA RTX 3080 GPU，用於模型訓練與測試。
- (二)、硬體環境：國內頂大伺服器，2×NVIDIA RTX 4090 GPU，Ubuntu 22.04.5 LTS (GNU/Linux 6.8.0-51-generic x86_64)用於深度學習訓練。

二、軟體工具

- (一)、開發環境：Python 3.10.12

使用套件：

- 1.RE：正則表達式模組，用於處理字串的模式匹配與字串搜索。
- 2.JSON 模組：用於解析語料。
- 3.Numpy：用於高效數值計算與矩陣運算的函式庫。
- 4.torch.nn：有用的神經網路模組
- 5.optuna：超參數搜尋，找出最好的組合
- 6.sklearn：隨機森林和產出報表
- 7.torch.utils.data：處理資料的模組

- (二)、深度學習框架：PyTorch

- (三)、語料處理工具：CLAN

- 1.Trasnsformer-BERT：

- (1) BERTTokenizer：BERT 分詞器
- (2) BERTModel：BERT 模型
(用於獲取詞向量、CLS 向量等)

- (四)、數據來源：

失語症患者的語音與文字數據集
(基於AphasiaBank語料庫)。

- (五)、語料讀取軟體：

CLAN(由AphasiaBank提供並下載使用)

- (六)、語料格式：*.cha 檔案格式

柒、研究結果

一、測試結果

本研究使用AphasiaBank語料，包含156,302筆患者語句，分別包含在1010篇有正確標記失語症類型的語料。

- 評估樣本數 (Eval size)：731
- BERT TOKENs 限制：≤ 512
- 數據劃分：80% 訓練集，20% 測試集

2025-02-20T11:47:57.106578: Classification Report:					
	precision	recall	f1-score	support	
0	0.705882	0.705882	0.705882	102.000000	
1	0.000000	0.000000	0.000000	8.000000	
2	0.760479	0.690217	0.723647	184.000000	
3	0.484848	0.581818	0.528926	110.000000	
4	0.740741	0.400000	0.519481	50.000000	
5	0.582734	0.736364	0.650602	220.000000	
6	0.523810	0.224490	0.314286	49.000000	
7	0.333333	0.200000	0.250000	5.000000	
8	0.000000	0.000000	0.000000	0.000000	
9	0.000000	0.000000	0.000000	3.000000	
accuracy	0.625171	0.625171	0.625171	731.000000	
macro avg	0.459092	0.393197	0.410314	731.000000	

2025-02-20T11:47:57.106821: Class 0 Accuracy: 70.59%					
2025-02-20T11:47:57.106840: Class 1 Accuracy: 0.00%					
2025-02-20T11:47:57.106854: Class 2 Accuracy: 69.02%					
2025-02-20T11:47:57.106867: Class 3 Accuracy: 58.18%					
2025-02-20T11:47:57.106879: Class 4 Accuracy: 40.00%					
2025-02-20T11:47:57.106891: Class 5 Accuracy: 73.64%					
2025-02-20T11:47:57.106902: Class 6 Accuracy: 22.45%					
2025-02-20T11:47:57.106914: Class 7 Accuracy: 20.00%					
2025-02-20T11:47:57.106926: Class 8 Accuracy: nan%					
2025-02-20T11:47:57.106938: Class 9 Accuracy: 0.00%					

[測試結果分類報告]

二、召回率比較

本研究使用折線圖來顯示各類別模型的預測表現。

- (一) 召回率較高的類別：Broca、NotAphasicByWAB和Anomic
- (二) 召回率較低的類別：TRANSMOTOR、ISOLATION和TRANSSENSORY（甚至為0%）

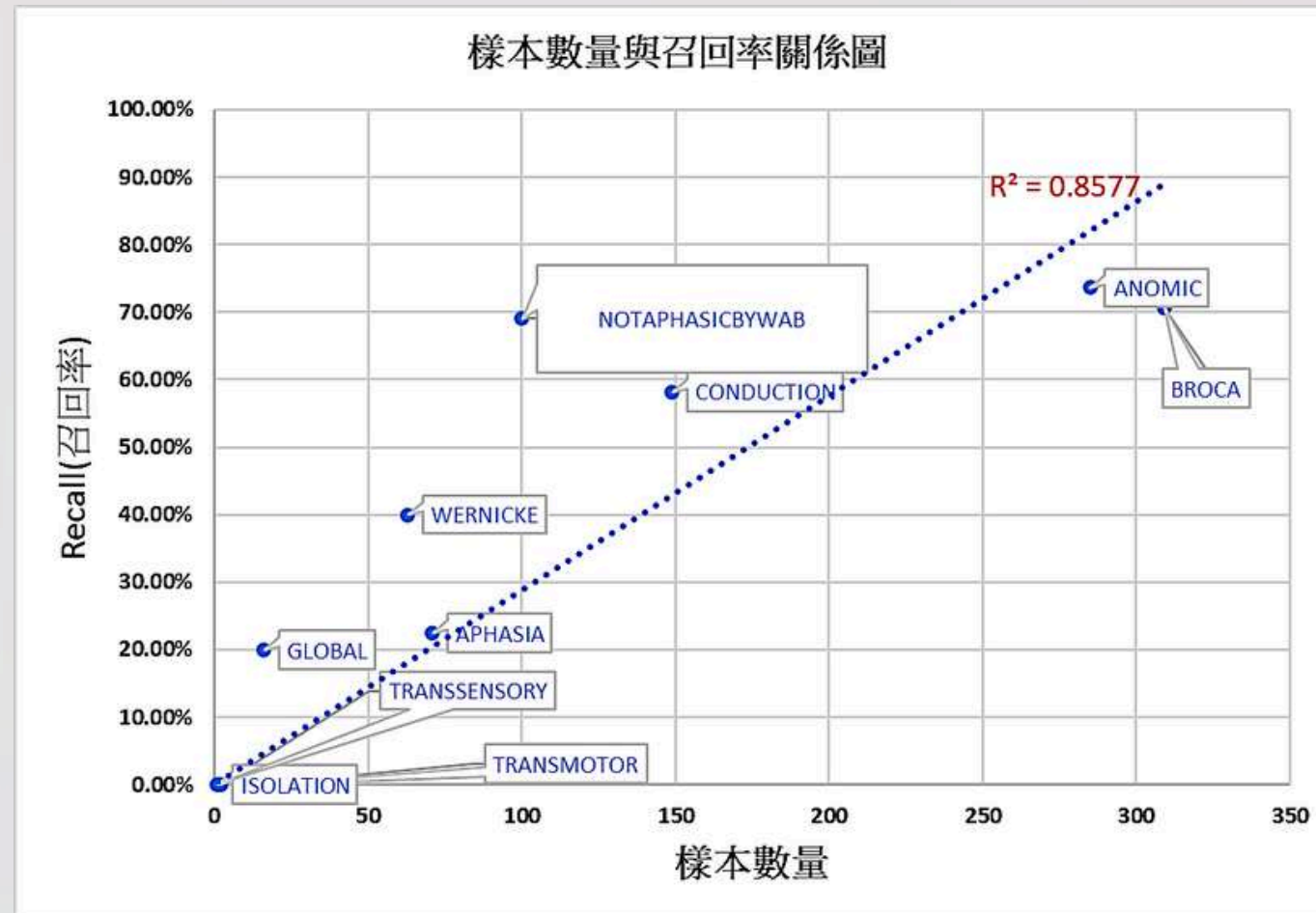


三、為何TRANSMOTOR、ISOLATION、TRANSSENSORY類別的準確率較低

- 樣本數量不足，語料稀缺，難以建立有效語言模型

由右圖分析，可知樣本數量與召回率呈現強烈線性正相關

($R^2=0.8577$)，而語料的增加可以提升召回率。



[樣本數量與召回率關係圖]

四、整體語料進一步擴增，準確率可達 80%

我們利用語料充足的四個失語症類別進行訓練，準確率高達90%，這也遠超Yan cong[5]對於兩個失語症類別的76%-88%準確率。

	precision	recall	f1-score
accuracy	0.901408	0.901408	0.901408
macro avg	0.887216	0.882397	0.884740
weighted avg	0.900892	0.901408	0.901100

[對於多語料類別的BERT分類結果]

捌、討論

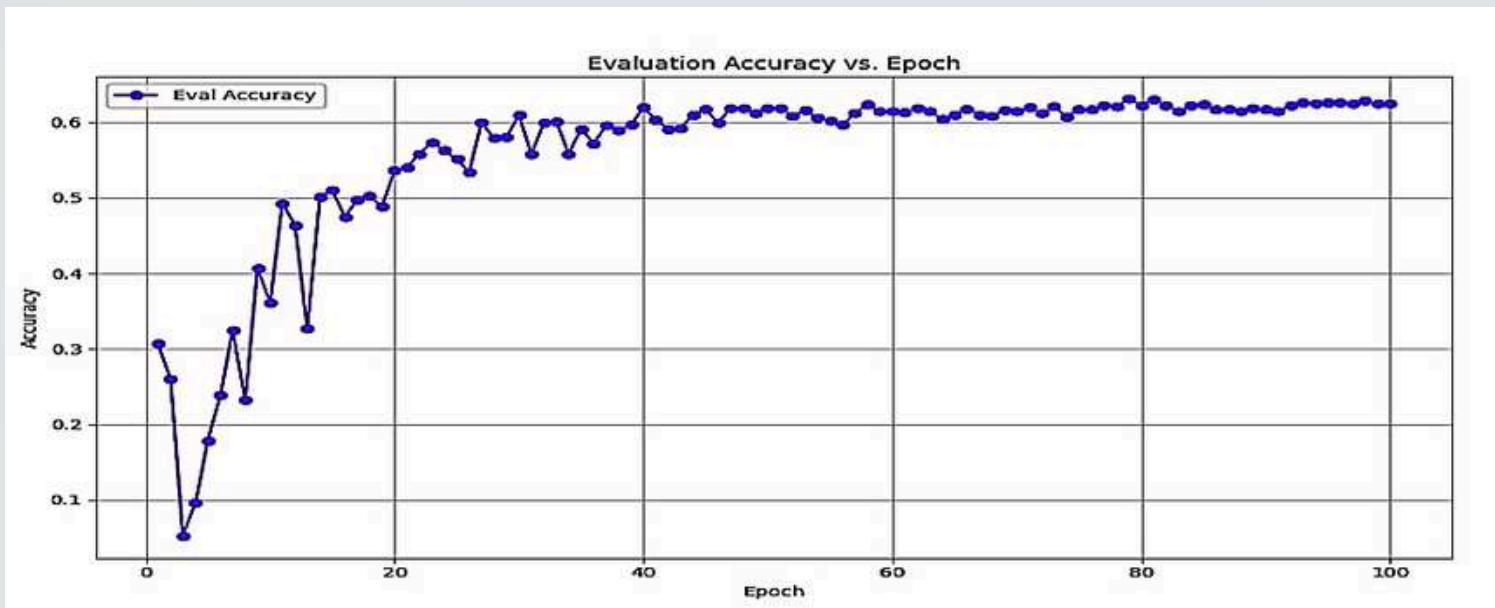
一、整體準確率分析

$$Accuracy = \frac{\sum TP}{\sum (TP + FP + FN + TN)}$$



$$Accuracy = \frac{72+127+64+20+162+11+1}{731}$$

在類別不平衡條件下，模型仍能保持較高準確率，彰顯其在輔助應用上的可行性。



[模型準確度]

二、模型表現分析

類別	精確率 (Precision)	召回率 (Recall)	F1-score	樣本數
0	70.59%	70.59%	70.59%	102
2	76.05%	69.02%	72.36%	184
5	58.27%	73.64%	65.06%	220

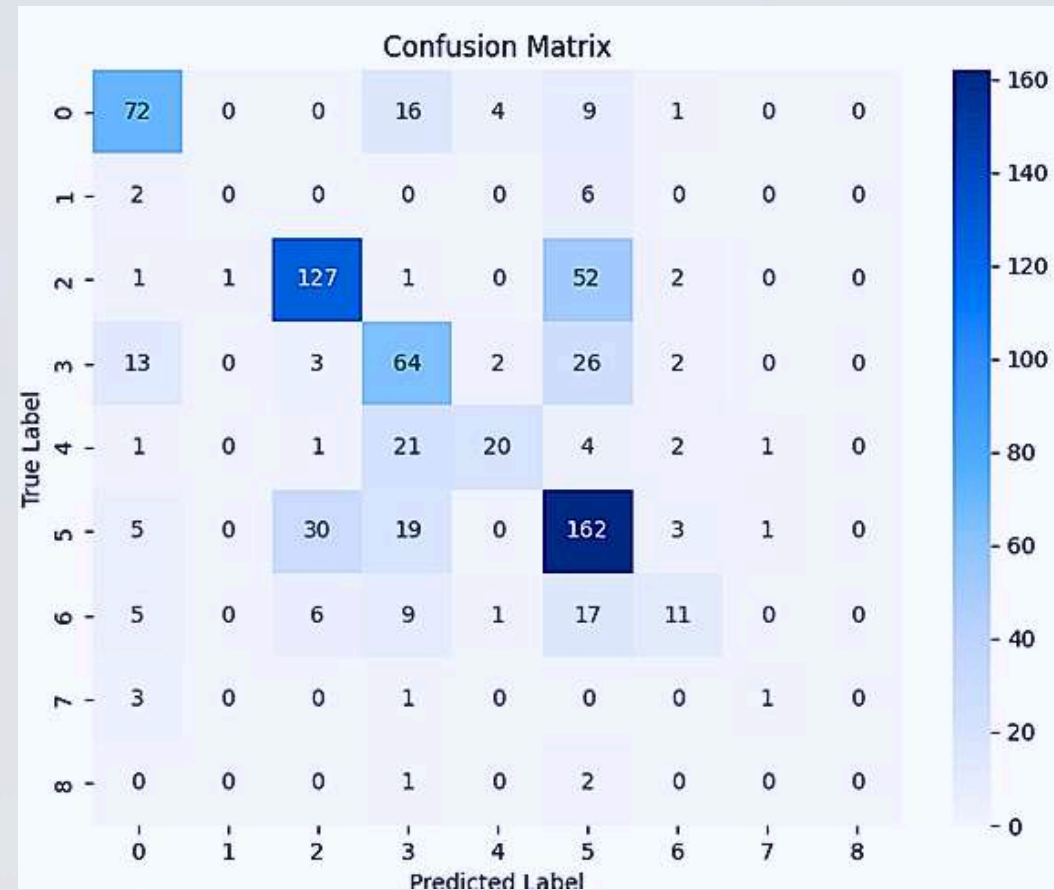
[以類別 0，2，5 為例 -Precision、Recall、F1-score 統計]

- 以類別0為例分析
高Precision(70.59%)和高Recall(70.59%)，表示模型能穩定且精準地辨識類別 0，這代表模型的辨識能力穩定，錯誤分類率適中且誤判率低。

三、混淆矩陣分析

2025-02-20T11:47:57.007525: Confusion Matrix:										
[72	0	0	16	4	9	1	0	0]	
[2	0	0	0	6	0	0	0]		
[1	1	127	1	0	52	2	0	0]	
[13	0	3	64	2	26	2	0	0]	
[1	0	1	21	20	4	2	1	0]	
[5	0	30	19	0	162	3	1	0]	
[5	0	6	9	1	17	11	0	0]	
[3	0	0	1	0	0	0	1	0]	
[0	0	0	1	0	2	0	0	0]	

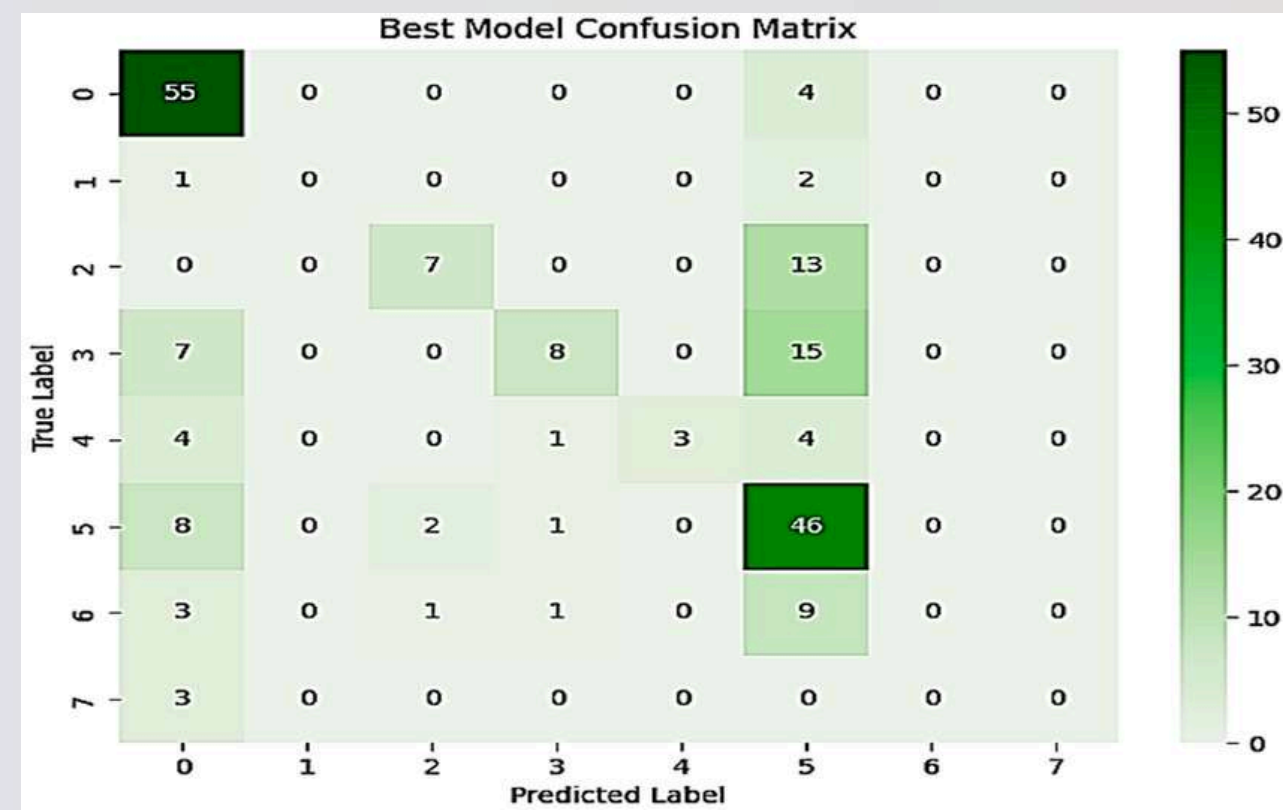
[Confusion Matrix混淆矩陣]



- (一) 類別5（162次）、類別2（127次）在混淆矩陣中為深色區塊，表示模型對這些類別的分類效果較佳。
- (二) 模型能準確區分Broca失語症（類別2）與Anomic失語症（類別5），這是語言障礙診斷中重大的突破。
- (三) 大多數類別的預測結果集中在對角線，表示模型整體準確率較高。

五、隨機森林模型的驗證結果

我們將TF-IDF向量與額外特徵結合，輸入隨機森林分類。當n_estimators設為150時，準確率達62%，優於隨機分類結果。

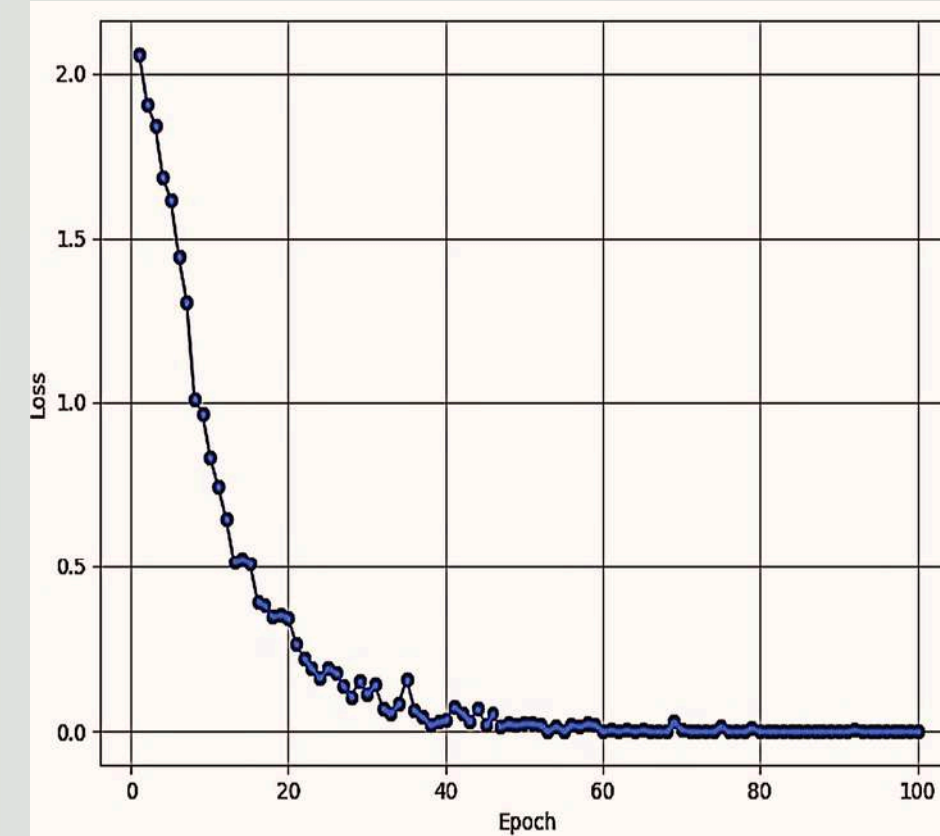


四、損失值（Loss）與訓練過程



每個訓練回合的loss (損失)變化

- Loss值在20Epoch內快速下降，顯示模型學習有效。
- 40~50Epoch之後Loss逐漸收斂，表示模型已基本學習完成。



每一個訓練輪迴下的損失(loss)值及訓練過程，如下：

93%	1298/1388 [88:38<80:54, 1.64it/s]2025-02-20T12:45:08.119567: Log: {'l			
oss': 0.043, 'grad_norm': 1.946968913878388, 'learning_rate': 3.2688695652173914e-06, 'epoch': 28.04/7826869566}				
{'loss': 0.045, 'grad_norm': 1.946968913878388, 'learning_rate': 3.2688695652173914e-06, 'epoch': 28.04/				
94%	1388/1388 [88:38<80:29, 2.72it/s]2025-02-20T12:45:11.717748: Log: {'l			
oss': 0.0324, 'grad_norm': 2.765488624572754, 'learning_rate': 2.898550724637681e-06, 'epoch': 28.26/				
{'loss': 0.0324, 'grad_norm': 2.765488624572754, 'learning_rate': 2.898550724637681e-06, 'epoch': 28.26/				
95%	1310/1388 [88:37<80:25, 2.77it/s]2025-02-20T12:45:15.316663: Log: {'l			
oss': 0.0305, 'grad_norm': 5.798885345458984, 'learning_rate': 2.5362318848579714e-06, 'epoch': 28.47/2686956522}				
{'loss': 0.0305, 'grad_norm': 5.798885345458984, 'learning_rate': 2.5362318848579714e-06, 'epoch': 28.47/				
96%	1328/1388 [88:41<80:21, 2.78it/s]2025-02-20T12:45:18.911801: Log: {'l			
oss': 0.0314, 'grad_norm': 0.1968294897661972, 'learning_rate': 2.173913843478261e-06, 'epoch': 28.69/5652173913843}				
{'loss': 0.0314, 'grad_norm': 0.1968294897661972, 'learning_rate': 2.173913843478261e-06, 'epoch': 28.7/				
96%	1338/1388 [88:46<80:17, 2.79it/s]2025-02-20T12:45:22.587755: Log: {'l			
oss': 0.0266, 'grad_norm': 0.36525355834961, 'learning_rate': 1.8115942028985508e-06, 'epoch': 28.91/384347826887}				
{'loss': 0.0266, 'grad_norm': 0.36525355834961, 'learning_rate': 1.8115942028985508e-06, 'epoch': 28.91/				
97%	1334/1388 [88:46<80:15, 3.03it/s]2025-02-20T12:45:23.046868: Epoch 29			
08 ended. Loss: 0.0266				

玖、結論及未來展望

本研究透過階段性實驗設計與深度學習技術，成功開發一套結合BERT模型與自訂分類器的失語症診斷系統，並藉由詞彙、語法與長文本處理技術，顯著提升診斷準確率。透過綜合NLP模型、機器學習與超參數搜尋的驗證，在準確率、召回率（recall）與F1-score等指標上優於現有參考文獻，在處理不同類型的失語症患者語言樣本時展現更高的適應性與準確度。

一、研究貢獻與創新

(一)突破NLP在失語症分析領域的瓶頸

本研究分類準確率達90%，相較Cong et al.[5]的GPT模型（76%-88%），提升2%至14%的準確率。

(二)階段性實驗設計與模型效能提升

三階段實驗 & 改良少量分類模型，聚焦語料數量最多的四類失語症（由原 10 種縮減），並強化分類策略，**即使各類失語症語言特徵高度重疊，模型仍能準確分類。**且應對資料不均衡問題，使用**Weighted_Random_Sampler**，使模型能更好處理樣本較少的語料類別。

二、模型效能分析與應用價值

BERT模型在Broca、Anomic、Conduction失語症的辨識效果優異，這對往後的應用以及研究有莫大的幫助。

三、未來發展方向

- (一) 擴展語料庫與多模態數據分析：結合語音、EEG、MRI等多模態數據獲取更多運算資源提升計算能力，開發提升分析系統。
- (二) 與臨床測試（如WAB、QAB）結合，建立更完整語言評估系統。
- (三) NLP技術應用於醫療機構：應用於醫院與長照機構，輔助更準確的失語症診斷。