小型自然語言模型應用 — 癌症基因衛教機器人

Application of Small Language Models - Cancer Gene Educational Chatbot

指導教授：謝孫源
專題成員：張宥薇、劉邦佑
開發工具：Python, BAAI/bge-m3
測試環境：Visual Studio Code,
LM Studio

一、簡介

一般而言，衛教資訊多以文宣傳播，或是醫護人員、志工在民眾就診時口頭宣導。發展專門的衛教語言模型有多項好處：強化患者、患者家屬與醫護人員的溝通效率，也能提升資訊可取得性。

然而，語言模型在回答一般性問題時表現出色，但是回答專業問題時卻不一定精確，回答醫療問題時尤其明顯，可能造成反效果。而在廣泛的衛教範疇內，本專題將「癌症基因」作為主題。以知識庫之建立以及 fine-tune為本，得到能詳細回答使用者提出與癌症基因相關的衛教問題之語言模型。

架構如圖1. 呈現，分為三大部分：

（1）資料、數據處理

資料預處理部分，採用網路爬蟲自動化爬取工具，快速取得美國癌症協會網站提供的純文字衛教資訊，以及透過光學字元辨識則針對醫學圖表、掃描文件、海報……等非純文字資訊進行文字提取及圖表資訊整理，補足純文字資料不足之處。

最後將所有資訊及數據儲存於資料庫中，作為訓練和增強索引生成資料來源。此資料庫可以動態更新。

（2）增強索引生成（retrieval augmented generation, RAG）

透過增強索引生成機制，確保語言模型提供回覆之依據為前期建置完成的資料庫。

將資料庫做向量化處理（vectorization）後，在語言模型根據使用者提問提供回覆之前，先在向量化資料庫中做語義搜尋（semantic search），作為語言模型提供回覆的輔助資訊。

（3）語言模型微調

選用從 Hugging Face 平台取得的 Llama-3.2-1B-Instruct (Meta, 2024) 作為基礎語言模型，透過低秩適應（low-rank adaptation, LoRA） fine-tune（參數如圖2.、圖3. ）後完成我們的模型 CanCura。

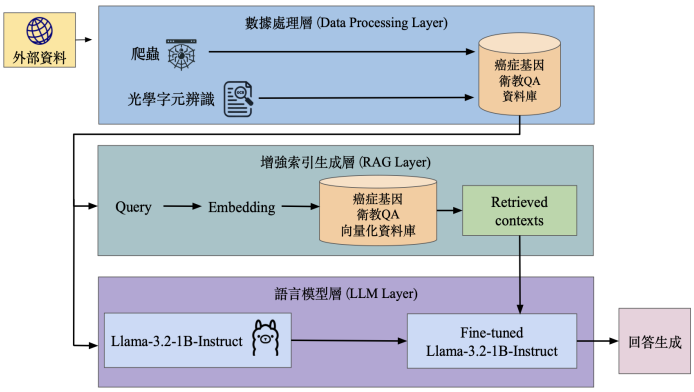

圖1. 研究架構

```
loraConfig = LoraConfig (
    r = 4,
    lora_alpha = 32,
    target_modules = ["q_proj"],
    lora_dropout = 0.3,
    bias = "none",
    task_type = "SEQ_2_SEQ_LM"
)
```

圖2. LoRA 參數

```
trainingArgs = TrainingArguments(
    output_dir = "./results",
    per_device_train_batch_size = 4,
    per_device_eval_batch_size = 4,
    num_train_epochs = 15,
    learning_rate = 5e-5,
    logging_dir = "./logs",
    logging_steps = 10
)
```

圖3. 微調訓練參數

二、測試結果

知識庫與模型建置完成後，我們將模型放到 LM Studio 上測試，使用者可以與微調後的模型 CanCura 對話。模型在回覆前會先在知識庫中進行語義搜尋，找出相關內容作為 RAG 的輔助資料，並生成回覆給使用者。經過測試，不論使用語言為中文或英文，模型對於癌症基因衛教知識的問答能力皆有顯著提升。但是受限於資料來源多為英文，模型在英文問答的表現優於中文問答。

圖4. 是一問答範例，模型接收使用者提問後，根據使用者的問題詳細解釋何為 BRCA 基因突變，並引用知識庫中提到的研究結果，同時也會於畫面顯示引用之原內容。

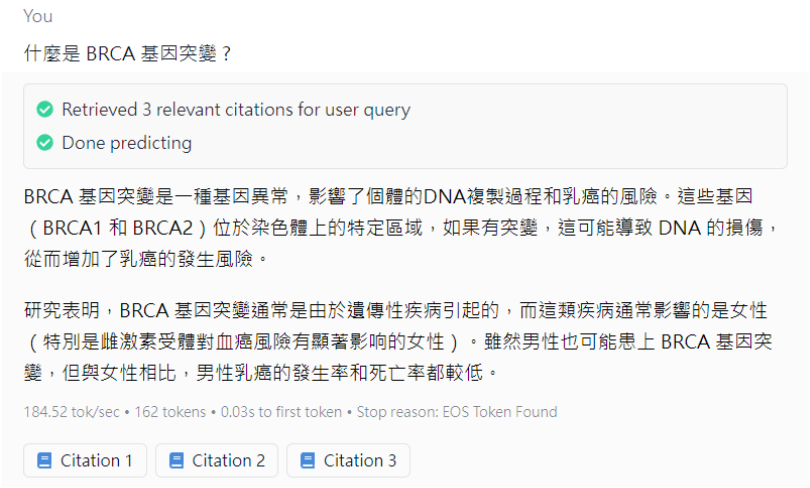

圖4. 問答範例