

基於多模態迭代檢索之本地衛教問答系統： 整合視覺、語音與專業反饋

Multimodal Iterative Retrieval-Augmented Local Health Education Q&A system: Integrating Vision, Speech and Professional Feedback

指導教授：蔣榮先 合作醫師：劉冠宏、邱裕桓 學生：李乃仁

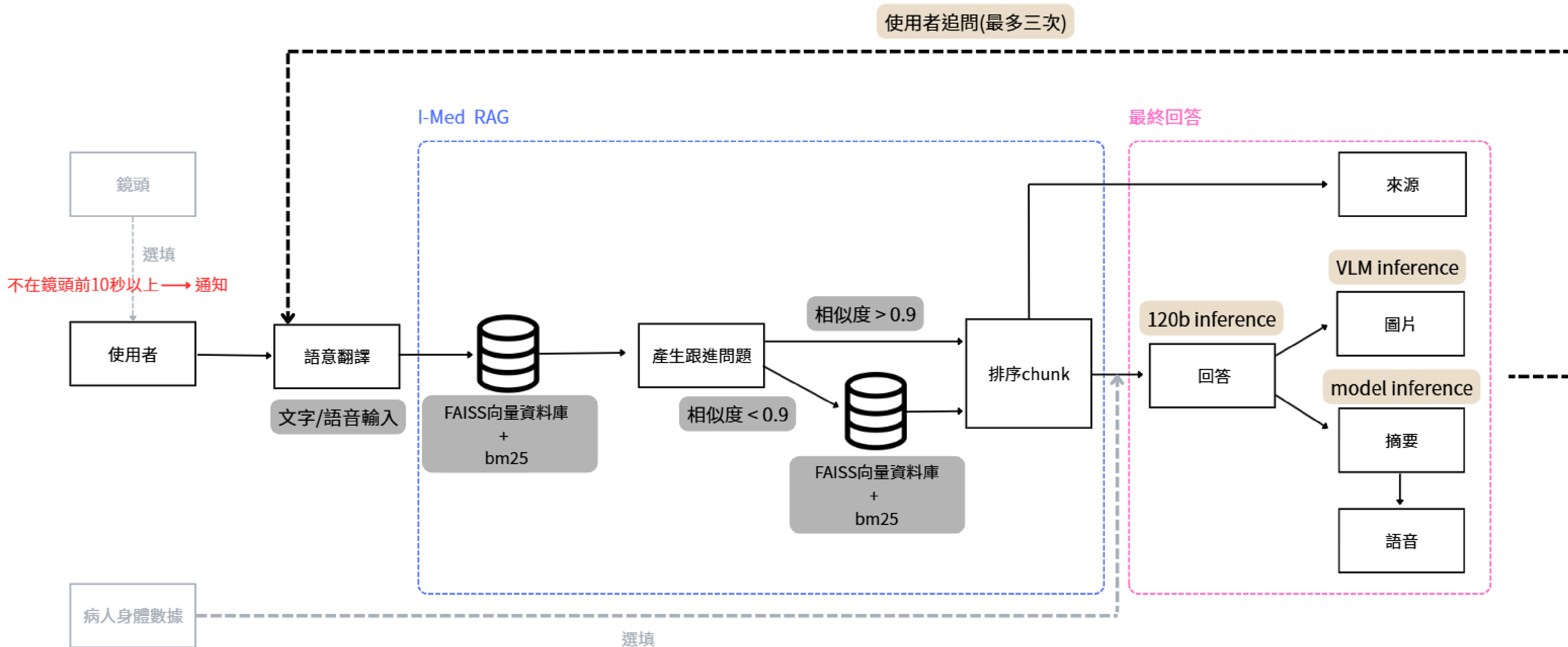
專題簡介

隨著台灣邁入高齡化社會，醫療需求持續攀升，臨床護理工作負荷亦日益沉重。然而，台灣目前長期面臨護理人力不足的問題，使護理師除了執行臨床照護外，還需投入大量時間進行重複性的衛教與病患問答。在此背景下，導入以大型語言模型（LLM）為核心的 AI 衛教問答系統，已成為提升醫療效率與優化照護流程的重要方向。

AI系統可提供 24 小時即時且一致性的衛教資訊，協助病人快速取得疾病、用藥、飲食及術後照護等相關知識，改善整體醫療服務效率與照護品質。然而，目前 LLM 仍可能因訓練資料限制而產生錯誤或不完整的醫療資訊，且難以依據病患個別身體狀況提供精準且個人化的衛教內容。因此，本專案旨在透過建立更完善的 AI 醫療問答架構，提升系統於醫療場域中的可靠性與實用性，其核心目標包括：

- 強化 LLM 輸出內容之醫療可信度與正確性。
- 建立嚴謹且可量化之效能評估流程。
- 實現能依據病患狀況提供個人化衛教問答之智慧系統。

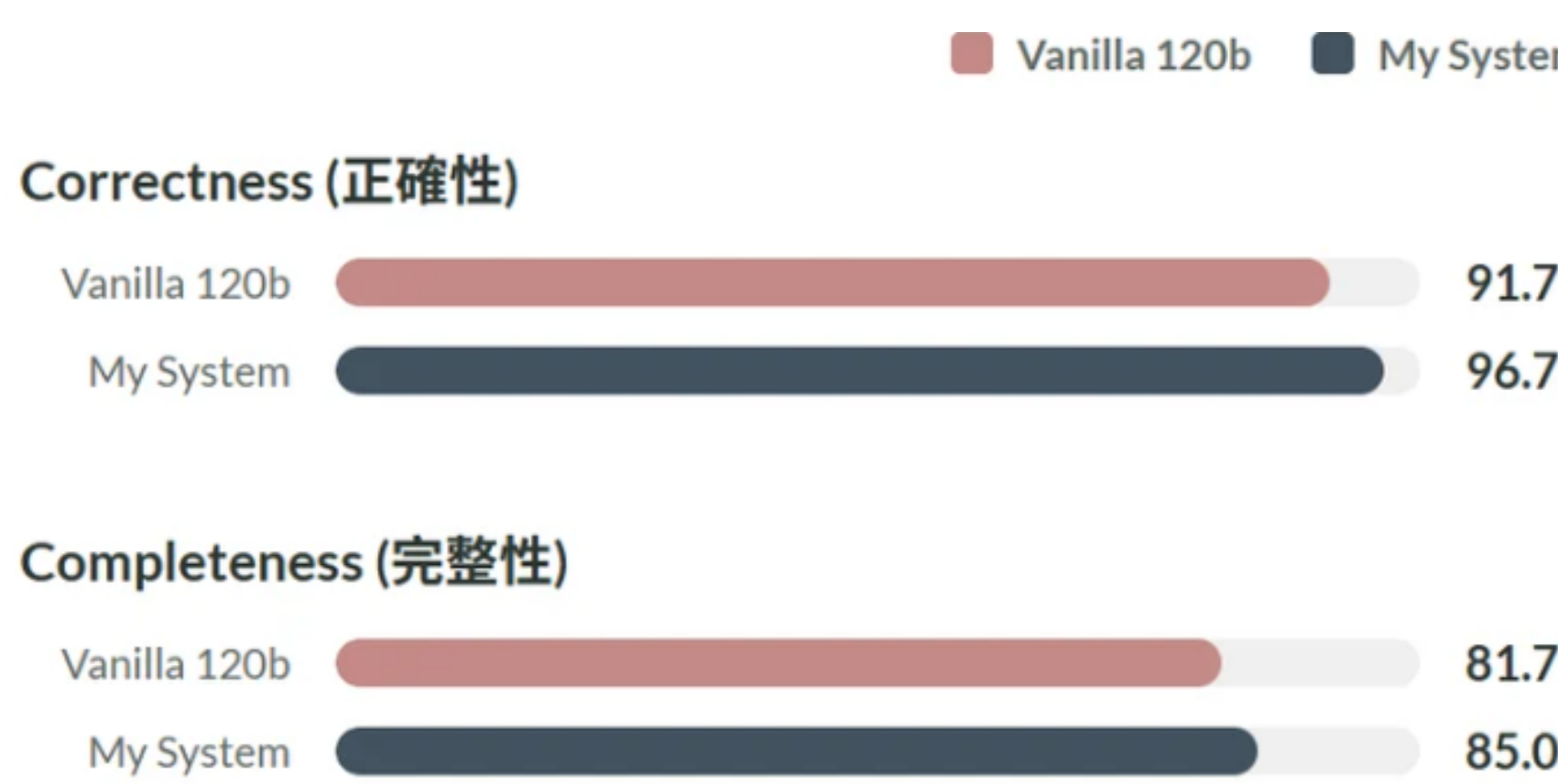
系統流程



系統介面

效能驗證

技術指標對比



Ragas

醫療人員實測

評估指標	平均分數
Faithfulness	0.988
Context recall	0.878

4.36

技術分析

前端

- UI介面：摘要 / 全文 / 語音 / 圖片 / 來源。
- 病患數據：選擇本次的病人，存入request。
- 鏡頭監控：呼叫 WebCam，並利用 MediaPipe 模型進行人臉偵測。

後端

- LLM：掛載實驗室的120b、HuggingFace 的 Embeddings。
- RAG：FAISS + BM25 的混合檢索系統，結合迭代檢索。
- 語音處理：接收前端傳來的音檔，用 Whisper-small 轉成文字；另外也把 LLM 生成的文字用 Piper TTS 轉成語音檔
- Database：將對話歷史、使用者的星星評分寫入 SQLite。此外也有存取病患個人資料，以及guideline的向量化資料庫。

結論

- 本研究主架構為迭代式檢索 RAG，並以醫學 Guideline 作為主要資料庫，確保系統回答內容具備醫療可信度。此外，透過嚴格設計的 Prompt 規範，限制 LLM 僅能根據資料庫內容進行回答，對於資料庫中未涵蓋之問題不得任意生成答案，以降低模型的 Hallucination。
- 本研究與成大醫師合作，建立具醫療專業性與臨床價值黃金標準資料集。後續參考 Amazon 所公開之 LLM 評估框架，並導入 Ragas 自動化評估指標，針對模型回答之正確性、相關性與可信度進行系統化分析，使模型優化過程具備可量化與可追蹤性，進而建立一套嚴謹且完整的醫療 AI 評估流程。
- 本研究結合資料庫中的病患個人資料與臨床資訊，使 LLM 能根據病患的生理數據，提供個人化衛教問答。此外，系統部署於醫院本地端環境，所有資料皆於院內進行處理與儲存，可有效達成醫療資訊安全需求。

REFERENCES

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, Aidong Zhang. (2024). Improving retrieval-augmented generation in medicine with iterative follow-up questions. arXiv. <https://arxiv.org/abs/2408.00727>