基於 LLM 的智慧多輪問答系統

LLM-based Intelligent Multi-turn Question Answering System

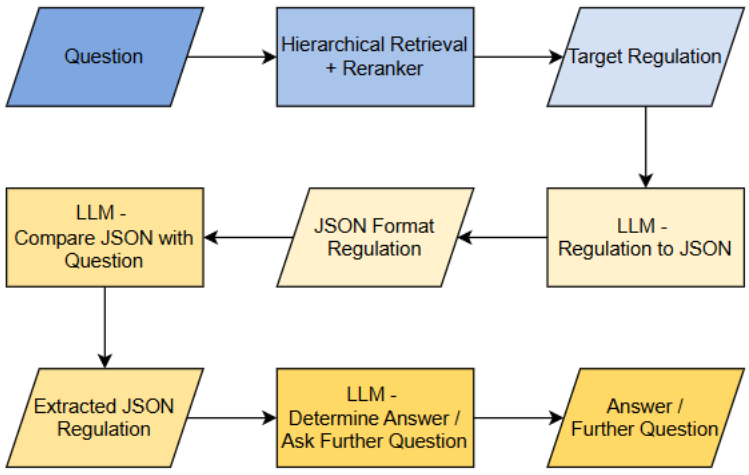
指導教授：蔣榮先
專題成員：歐冠亭
開發工具：Python, LangChain, Chroma
測試環境：Ubuntu 12.3.0-1ubuntu1~22.04

一、簡介：

隨著企業規模與運作複雜度日益提升，員工經常需要查詢內部制度、流程規範、IT 操作等知識。傳統的單輪問答系統僅能處理明確、結構化的提問，無法針對資訊不足的問題進行追問與釐清，導致回答錯誤或資訊缺漏，進而影響工作效率與內部溝通。

此專案的目標是針對特定領域之自動化問答，並完成在 meta-llama/llama-3.1:8b-instruct-fp16 上釐清使用者問題之意圖，在資訊不充足時向使用者進行追問。

以下為系統架構圖：



圖一：系統流程圖

用戶問題首先會經過 Retrieval 與 Reranker ，以選出最能有效回應問題的條文。接著，系統透過三次階段性的 LLM 分析，逐步比對問題與條文內容，判斷用戶提供的資訊是否充足。若判定資訊充足，則可直接根據條文回覆使用者；若資訊不足，則系統會提出提問以得到更詳細的用戶情況。待用戶補充資訊後，問題將重新進入相同的處理流程，直到取得足以回應問題的充分資訊為止。

相較於其他問答系統，我們的多輪問答系統旨在釐清使用者意圖，在資訊不充

足時向使用者進行追問。我們使用開源地端的 LLM 與 Retrieval Augmented Generation(RAG) 技術，使系統回答更加自然，且支援廣泛的主題並適用於不同場景。

未來展望：

1. 多領域適用性：透過將 prompt 做簡單的調整，並建立相關領域的資料庫，便可將此系統應用於其他領域。
2. 高效客服系統：此系統能夠解決模糊問題導致的服務效率不佳，或是錯誤答案之誤導。

二、測試結果：

在問題一與問題二中，應用 RAG 技術的 LLM 在使用者所提出之問題與獲得之相關條文進行比對後，判斷該問題所提供之資訊已足以支撐答案推論，且可自條文中明確且一致地導出回應內容，因而認定無需進一步追問，遂直接產出答覆。

至於問題三，應用 RAG 技術的 LLM 在比對問題與獲得的條文後發現，該問題所提供的資訊不足，無法從條文中獲得具體明確的解答。因此，系統判定需向使用者進行追問，針對缺漏的關鍵資訊（如年資）進行補充提問。當使用者補充相關資訊後，LLM 再次將完整的問題內容與條文比對，確認此時資訊已充足，可從條文中得出明確答案，故判定無需再追問，並直接回覆最終解答給使用者。

圖二：系統執行畫面