NTHU\_NLP Assignment 4 Report

Name: 張芷晴 School: 成功大學 School\_id: NM6131051

Running Environment: colab

Python Version: colab

* Please describe the details of your implementation for the RAG system (please tell us 1.What’s in your RAG system? 2. Which retrieval model you use? 3. What’s your prompt? 4. What’s new in your code in comparison with the code from our lab course?) in this assignment and list your best score for the ten questions.

**Embedding model**: 使用HuggingFaceEmbeddings with the model "all-mpnet-base-v2"和原本上課lab的”jinaai/jina-embeddings-v2-base-en”，兩者都有達到回答10個正確答案。

**Language Model:** Ollama with the "llama3.2:1b" model

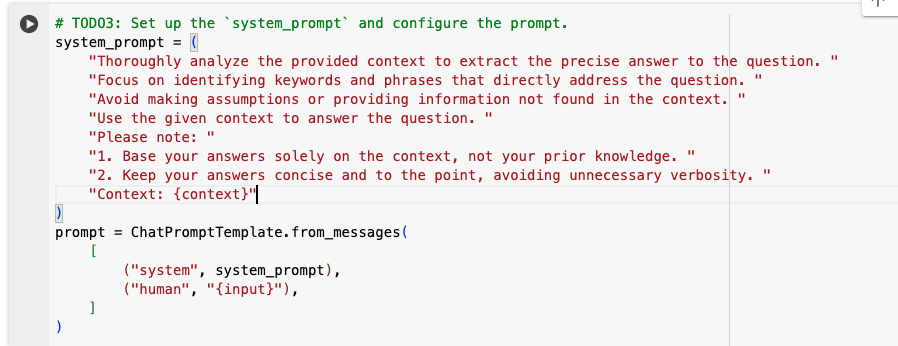
**Vector Store:** FAISS (Facebook AI Similarity Search)和上課lab的Chroma。達到回答10個正確答案的是用FAISS

**Retriever：**Configured with similarity search、mmr、similarity\_score\_threshold皆有嘗試，效果皆落在7-9之間。達到回答10個正確答案的是用similarity，Number of retrieved documents (k)設為7。

**Prompt Template**: Custom system prompt for context-based answering

**Retrieval Chain:** Combines document retrieval with question answering

**Prompt Design:**有嘗試上課lab的prompt，也有嘗試使用chat-gpt提供的prompt，以下是達到回答10個正確答案時，所使用的prompt:



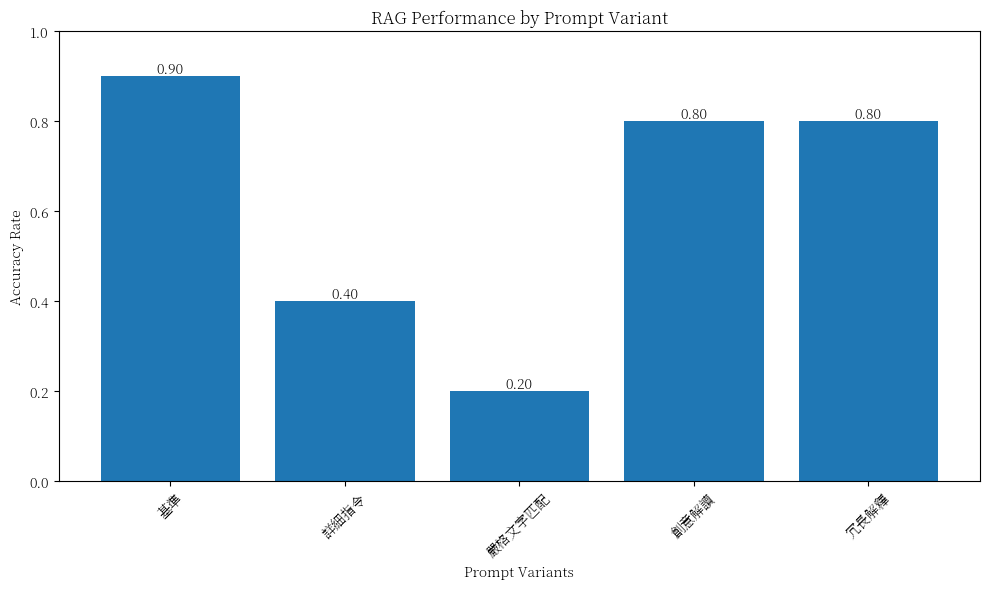
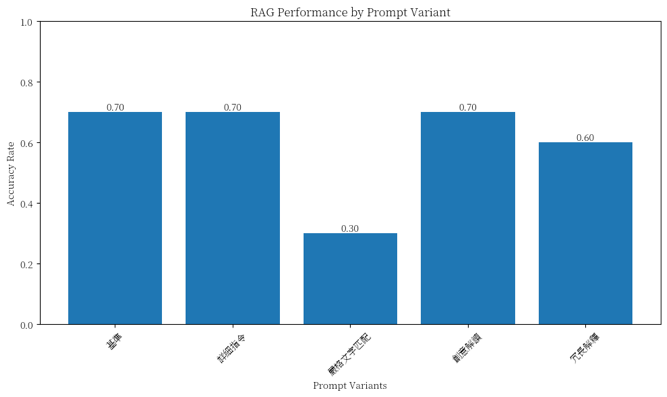
My best score for the ten questions：

一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述

* Please provide analysis for the RAG performance using different prompts.

1. 在使用原本lab的prompt時，有一句是” If you don't know the answer, say you don't know.”，有這句時，很容易出現I don’t know的答案，拿掉會發現回答正確的題數變多了。後來有嘗試使用自己下的prompt跟請gpt給的prompt發現給的越簡潔回答正確的題數越多。
2. 請chatgpt 幫我生成四種不同的prompt，分別有detailed\_instruction\_prompt、 strict\_matching\_prompt、 creative\_prompt,、 verbose\_prompt，跟原本的prompt相比，皆可以看到使用嚴格文字匹配(strict\_matching\_prompt)的準確度在兩次實驗都是最低的，而原本的prompt準確度最高。



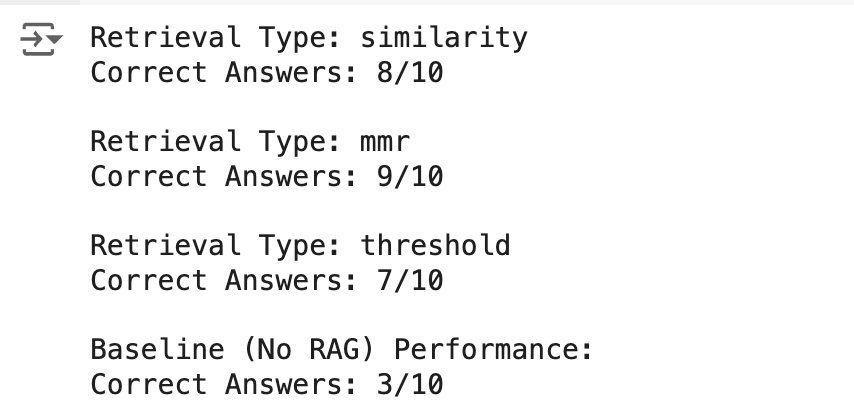
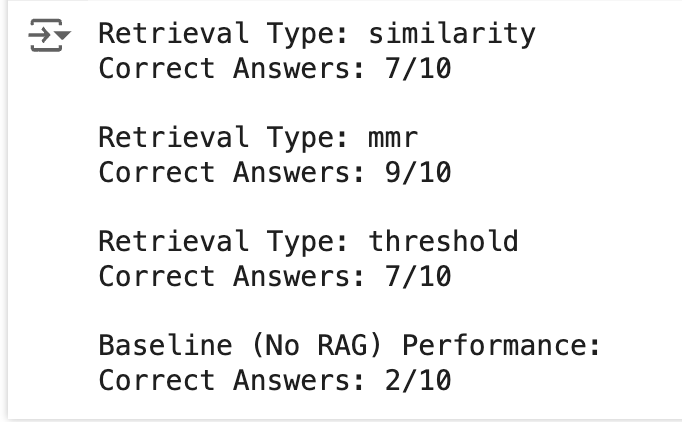
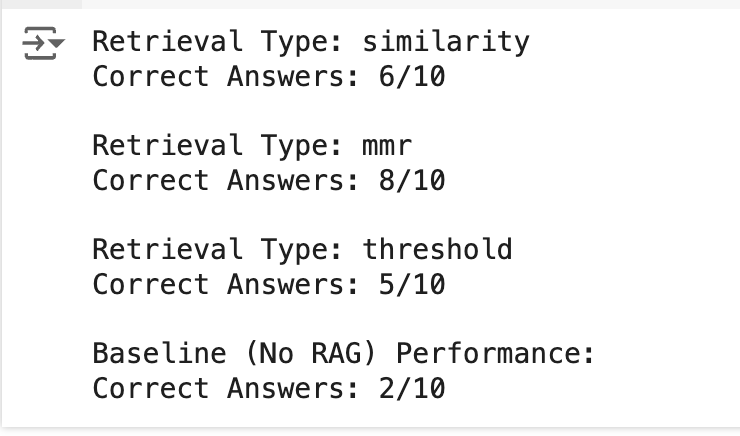
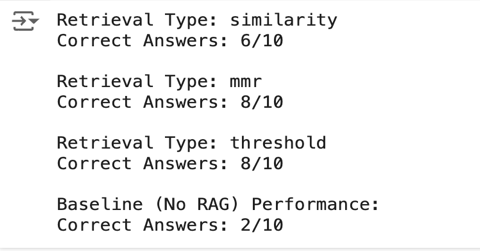
▲上圖為使用不同的prompt進行準確度的實驗（左圖為第一次，右圖為第二次實驗）

▲ 另外四種不同的prompt，分別是creative\_prompt(創意解讀)、 verbose\_prompt(冗長解釋)、detailed\_instruction\_prompt (詳細指令)、 strict\_matching\_prompt(嚴格文字匹配)。

* 一張含有 文字, 螢幕擷取畫面, 字型, 文件 的圖片

  自動產生的描述Please compare the RAG performance with different retrieval models and the performance without using RAG (note that Llama 3.2 should not be fine-tuned in this assignment).

我使用了不同的retrieval models 分別是similarity search、mmr、similarity\_score\_threshold來和No\_RAG來進行比較，發現在四次的實驗中No\_RAG的正確率都遠低於其他三者，且都在2-3之間浮動。



* Anything that can strengthen your report.

1. 在實作的過程中發現每次的結果都不一樣，推測是因為LLM本身有隨機性，每次用模型時，生成的回答會不同。答對題數幾乎都在7-10之間浮動。我有額外測試更大型的語言模型（qwen2.5-coder:14b）的效果，發現穩定性更高，正確題數落在9-10，但運行時間會是原本llama模型的好幾倍。
2. 在檢查生成答案的時候，發現像是正確答案是31mph，但生成答案是31 miles per hour，意思一樣但會被判斷成錯誤答案的情況，所以在下prompt的時候要思考如何避免這種狀況。

此外，發現若第一題答案的2/3是在括號內（如下圖），會被判斷為回答錯誤。第一題也是最容易答錯的題目，正確答案只有2/3，但很多時候生成的答案是two-thirds或16-18 hours per day等表達方式不同但同義的答案。

而這題Which wildcat is considered the ancestor of all domestic cats? A:the African Wild Cat常常會因為答案少一個the而被判斷為錯誤答案。

最後一題正確答案為four，而生成答案有時會出現4而不是four導致被判斷錯誤。

1. 所有額外的分析（不是在TODO裡面的code）皆以註解的形式留在作業py檔中