**11310CS563100 NLP Assignment 4: Retrieval-Augmented Generation with LangChain Report**

NTHU 110020007 施淙綸

* 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**.
  + 我的RAG中有retrieval model可以將提供的cat-facts.txt作為Knowledge Base，從中提取可以用來回答使用者問題的相關知識，不只是單純使用LLM的知識來回答而已，這種方法預期會有更佳的答對率，而非基於LLM學習到的知識或是它的猜測。
  + 我主要是使用Chroma vector store作為retrieval的主要方式，後面會有使用BM25的數據。
  + 接下來的數據如果沒有特別提到，就是使用ID 15的prompt（各版本prompt在下方有詳細列出），原因是因為在Chroma作為retrieval並使用jinaai/jina-embeddings-v2-base-en作為embedding方法中，ID 15提供不錯的答對率。所有prompt和embedding方法在後面都有一些可以觀察的數據。後面許多數據和觀察都是這題的說明內容。
  + 和lab course的實現方式相似，但為了更好的蒐集各種條件下的數據，有新增很多計算的區塊，這些區塊讓原先的evaluation能夠完整得到多次測試的結果、每一筆資料的正確數、問答歷史準確率和總和的準確率。
  + 最佳的問答記錄（其實可以發現它第一題不知道在回答什麼，但比對結果還是對）：

|  |
| --- |
| Query: How much of a day do cats spend sleeping on average?  Response: On average, cats spend 2/3 of every waking hour or roughly 60 minutes spent sleeping.  Query: What is the technical term for a cat's hairball?  Response: A bezoar.  Query: What do scientists believe caused cats to lose their sweet tooth?  Response: Scientists believe this is due to a mutation in a key taste receptor that affects the feline sense of sweetness.  Query: What is the top speed a cat can travel over short distances?  Response: The original text states that "a cat can travel at a top speed of approximately 31 mph (49 km) over a short distance."  Query: What is the name of the organ in a cat's mouth that helps it smell?  Response: The organ called Jacobson’s organ in cats is located in the upper surface of their mouth.  Query: Which wildcat is considered the ancestor of all domestic cats?  Response: The African Wild Cat, still existing today, is considered the ancestor of all domestic cats.  Query: What is the group term for cats?  Response: The full text "A group of cats is called a 'clowder'" appears in my original response.  Query: How many different sounds can cats make?  Response: Cats can make nearly 100 distinct sounds, including hisses, growls, meows, purrs, and chirps.  Query: What is the name of the first cat in space?  Response: The first cat in space was named Felicette (a.k.a. "Astrocat") and was a French cat.  Query: How many toes does a cat have on its back paws?  Response: Cats have five toes on each front paw, but only four toes on each back paw.  Correct numbers: 10 |

* Please provide analysis for the RAG performance using different prompts.
  + 以下是所有嘗試過的prompts，都是使用Chroma mmr, k=3, fetch\_k=5。Prompt後方會接上Context: {context} 讓模型得到RAG的結果，<ID X> 代表ID 是X的prompt，<ID X>+<ID Y> 代表連接兩個prompt並且兩者之間會換行：

|  |  |  |
| --- | --- | --- |
| ID | Average Accuracy | Prompt |
| 1 | 81.9 | <empty> |
| 2 | 78.8 | You must answer the questions in a single sentense. |
| 3 | 83.4 | The original full texts in the dataset must appear in your answers. |
| 4 | 76.7 | Your answers cannot modify or rephrase any of the original text in the dataset. |
| 5 | 80.1 | <ID 2> + <ID 3> |
| 6 | 82.4 | <ID 2> + <ID 4> |
| 7 | 84.4 | <ID 3> + <ID 4> |
| 8 | **87.1** | <ID 2> + <ID 3> + <ID 4> |
| 9 | 75.7 | <ID 3> + <ID 4>  Provide the original source as reference of your answer in the dataset (documents). |
| 10 | 74.6 | Ensure all text in your answer comes directly from the dataset.  Do not add any extra information or personal interpretation.  Maintain the original paragraph format when quoting text. |
| 11 | 75.6 | You must answer in two sentences, one is your answer, another one is reference sentence.  You can only answer the sentence exactly in the dataset.  <ID 10> |
| 12 | 80.6 | You must answer the questions in a single sentence.  <ID 3> + <ID 4>  You must repeat the original contexts once. |
| 13 | 85.8 | You must answer the questions in a straightforward single sentence.  <ID 3> + <ID 4>  Your answers cannot change any orders, spaces, numbers, notations and sructures. |
| 14 | 86.4 | You must answer the questions in a straightforward single sentence.  The original full texts in the contexts must appear in your answers.  Your answers cannot modify or rephrase any of the original text in the contexts.  Your answers cannot change any orders, spaces, numbers, notations, expressions and sructures.  You can only repeat the sentences in the contexts. |
| 15 | 84.8 | You must answer the questions in a straightforward single sentence.  The original, unmodified full texts in the contexts must appear in your answers.  Your answers cannot change, convert, modify or rephrase any of the original text in the contexts.  Your answers cannot change any orders, spaces, numbers, ASCII codes, notations, expressions and sructures.  You can only repeat the vocabulary combinations in the contexts. |
| 16 | 81.3 | You must answer the questions in a straightforward and concise single sentence. The original full texts from the dataset must appear in your answers without modification, rephrasing, or reordering, including spaces, numbers, notations, and structures. Each answer must clearly repeat the relevant contexts as they originally appear in the dataset. Do not add any extra information, personal interpretation, or content beyond what is in the dataset. Ensure all quoted sentences are maintained in their original paragraph and sentence format. |
| 17 | 79.1 | You must answer the question in a concise single sentence that directly matches the original text in the dataset. The original text must appear exactly as it is in the dataset, without modifications to words, structures, spaces, numbers, notations, or expressions. If there are multiple valid answers, include them all in a clear and straightforward manner. Do not add any additional context, personal interpretation, or content not present in the dataset. Ensure your answer is factual and replicates the dataset's text without reordering or omission. |

※註：ID 16和ID 17使用ChatGPT-4o生成，指令是根據ID 1到ID 15的結果，彙整成新的、更好的prompt。

一張含有 螢幕擷取畫面, 正方形, 文字, Rectangle 的圖片

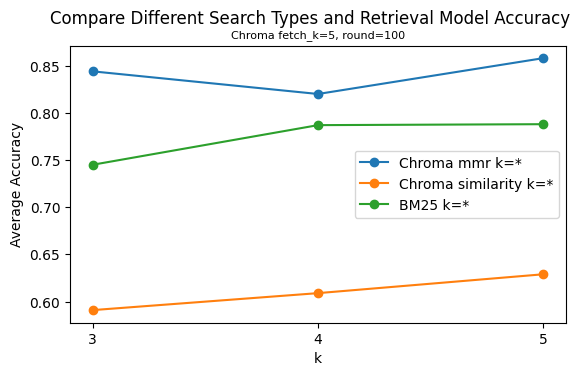
自動產生的描述

* + 上圖是所有不同prompts在不同問題分別的答對數量，因為都跑了100 rounds，所以也可以看成百分比答對率。
  + 接下來方便說明，定義P1的意思是prompt 1，Q1是question 1，R1是response 1。
  + 由上圖大致上可以發現，在Q1, Q4, Q5, Q7中不同prompt得到的結果差異很大（最大的超過90甚至95，最小小於50甚至10~30），可以推斷這些問題比較會受prompt設定的影響。在Q6, Q8, Q9中可以發現相比Q1, Q4, Q5, Q7，在我設定的prompt中都沒有明顯的波動（max - min < 25）。
  + P1和P2中的表現可以發現，P2和P1相比，Q2, Q3, Q6~Q10都有明顯的進步，除了Q3之外正確率都接近100%。
  + 表現的比較平均的是P7和P14，我認為ID 3 扮演很重要的角色，在有ID 3出現的prompt通常表現比較好。這個其實在一般的RAG應用中不是很好的prompt，因為人類在判斷正不正確不是依據完全比對來衡量，不過如果要讓程式大量跑結果的話使用完全比對會是比較可行的結果（或是可以用BERT score？）。
  + 在選擇哪些prompt應該使用時，發現如果讓它話說得太多很容易會重新換句話說，像是「貓咪2/3的時間都在睡覺」會變成「貓咪一天會睡大約16小時」，導致在這種比對方法中會比對不到認定為錯誤，雖然實際上會是正確的。所以為了避免這個問題，有嘗試讓它的回答限縮在一句長度，並不能重新闡述、變換符號或換句話說。
  + 承上，在調整prompts的時候有發現Q1很容易被改寫成其他樣子，所以常常會有其他題目表現得很好，結果Q1超級爛的結果。也有發現調整的過程時常會有終於改到其中幾題答對率大幅增加，結果其他題目又爛掉的情況，但礙於時間和作業規模我就沒有繼續調整，而是挑選P15當作base prompt來量數據。
  + 另外，這些數據我猜可能只能參考，畢竟LLM在相同的題目下仍有一定的隨機性，測試100次在作業中足夠了，但不能武斷的說哪一個就一定比較好，只能觀察特定趨勢。至於是不是剛好運氣好，我有多測幾次得到的結果相近，換一些奇怪的prompt也會表現的很差，所以應該沒有這個問題。
  + 比較特別的是請ChatGPT-4o幫忙寫的prompts表現沒有特別好，我還以為同是模型的GPT會比較懂模型的需求。
* 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).

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Avg |
| Vanilla  Llama | 0 | 59 | 0 | 5 | 0 | 0 | 0 | 69 | 0 | 56 | 18.9 |
| RAG Llama  Chroma | **57** | 83 | **61** | **78** | **92** | **100** | **83** | **100** | **100** | 94 | **84.8** |
| RAG Llama  BM25 | 32 | **96** | 52 | 23 | **92** | **100** | 60 | 92 | **100** | **98** | 74.5 |

※Note：RAG Llama Chroma, k=3, fetch\_k=5; RAG Llama BM25 k=3; round=100.

* + 由上表可以發現，在沒有RAG或是說Knowledge Base (KB) 的幫助下，單純的Llama 3.2的答對率都不是太好，只有在比較容易回答的Q2, Q8和Q10答對率較高，比較意外的是Q7的答對率不是很高，可能clowder是冷門或使用不廣泛的單字。
  + 尤其是像是Q1那種要回答「2/3」的，通常他都會回答「16小時」之類的答案；Q3「a mutation in a key taste receptor」是可預期的很差，因為可以換句話說的機會太高了，a和the的互換使用也會導致比對不到而表現不佳。還是要強調不一定是Vanilla Llama就真的完全回答不了或回答錯誤，而是無法回答和答案完全相同的答案。

一張含有 文字, 螢幕擷取畫面, 行, 圖表 的圖片

自動產生的描述

* + 上圖左比較不同的k對模型的表現是否有影響，可以發現基本上越高的k可以得到越好的結果，除了Chroma mmr以外，我想只是因為LLM回答的隨機性的關係。另外，Chroma similarity隨著k提升，Average Accuracy的提升大約成正比，不確定是巧合或是真的越高的k可以帶來等量的準確率提升。
  + 上圖右是比較不同的fetch\_k對Chroma mmr準確率的影響，大致上越高的fetch\_k會導致越低的準確率，這跟我一開始的預期不同，我一開始認為fetch\_k可以從KB中獲取更多有用的資訊，進而提升準確率。但結果和預想不同，我猜測可能越高的fetch\_k會帶來更多不必要的資訊，讓比對時最佳的資料未必是真正相關的。
* Anything that can strengthen your report.
  + 這部分有些融入在前面的題目當中，特別放在這裡感覺太遠，所以我把它們放在上面就沒特別標出來。

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Avg |
| RAG Llama  BM25 | 32 | 96 | 52 | 23 | **92** | **100** | 60 | 92 | **100** | 98 | 74.5 |
| jina-embeddings-v2-base-en | **57** | 83 | **61** | **78** | **92** | **100** | **83** | **100** | **100** | **94** | **84.8** |
| stella-en-1.5B-v5 | 43 | **99** | 58 | 53 | 27 | **100** | 73 | 79 | **100** | 77 | 70.9 |

※Note：RAG Llama Chroma, k=3, fetch\_k=5; round=100.

* + 不同的Embedding模型是否會對模型準確率造成影響？
    - 會產生一定的影響，有些embedding方式在某些模型和資料集會表現的比較好，某些則較差。從上表來看，總表現依序是jina-embeddings-v2-base-en、RAG Llama BM25，最後是stella-en-1.5B-v5；從細項來看stella-en-1.5B-v5會優於RAG Llama BM25，總表現兩者差距也不是特別大，前者主要是被Q5拖累。另外很特別的是三種方式在Q6和Q9都有完美的表現，有可能是因為和那兩題相關的答案比較單純比較好回答，或比較沒辦法換句話說，或是和題目相關的document數量不多或和其他的資料差距較大，不論使用哪種embedding都能找到有用的資料，並且讓Llama融合成最終完全相符的答案。
* 遭遇到的困難 & Reflections
  + 好像在prompt那題寫的太開心結果花太多時間，還好前幾次作業畫圖表的程式可以直接複製貼上，稍微改一下就能繼續用，而且我已經可以處理蠻快的了，跟前幾次為繁瑣的繪圖過程煩惱已經不一樣了。
  + 其實原本還想試一些其他的embedding model，像是jasper\_en\_vision\_language\_v1和NV-Embed-v2，但是下載下來之後發現模型太大沒辦法同時跑，時間不夠找可能的解決方法就沒有繼續做。會想嘗試embedding model是因為<https://huggingface.co/spaces/mteb/leaderboard>有Retrieval Average評分項目，想說會不會越高的分數會得到會好的結果。
  + 我在最後一個禮拜三才申請Llama的模型許可，原本想說他會不會過一個禮拜才通過，還好他隔十分鐘就開放了。這個故事告訴我們，作業要早一點寫不要趕死線（~~但我每次都還是會這樣~~）。
  + 這幾天天氣好冷寫作業頭好痛（物理），不過很高興可以順利寫完。這學期的四個lab都很有趣，也能學到很多東西，感謝老師和助教一同營運這門人數很多又跨校的課，希望在畢業之前還有機會修到更進階的課程。
  + 這次主要使用Kaggle蒐集數據，比起Google Colab好用多了，都是Google出的產品怎麼體驗差那麼多。本地有跑過不過速度過慢，然後硬碟空間不夠了，因此本地只繪製圖表而已。兩者環境如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Running environment | System | CPU | GPU | Python version |
| Local | Windows 11 Version 23H2 (KB5046633) | 11th Gen Intel® Core™ i7-11370H @ 3.30GHz | NVIDIA GeForce RTX 3050 Ti Laptop GPU | CPython 3.11.1 |
| Kaggle | Kaggle | Kaggle | GPU P100 | Kaggle |

※註：本作業部分程式碼使用ChatGPT-4o生成，部分程式碼使用GitHub Copilot以及Microsoft Copilot協助撰寫，區塊皆於檔案中註解。