RAG 文本切塊與檢索評估報告

HW Day5 — 學號 14112S2015（改良版：Top-3 + LLM）

Q1：固定大小切塊和滑動視窗切塊的參數設定多少？

一、固定大小切塊（Fixed-size Chunking）

使用 LangChain 的 CharacterTextSplitter，以純字元數進行切割，不考慮語意邊界。

參數	設定值	說明
chunk_size	300	每個切塊最多 300 個字元
chunk_overlap	0	切塊之間無重疊
separator	""（空字串）	不使用分隔符，純粹按字數切割

二、滑動視窗切塊（Sliding Window Chunking）

使用 LangChain 的 RecursiveCharacterTextSplitter，按照中文語意邊界進行遞迴切割，並透過 overlap 保留上下文。

參數	設定值	說明
chunk_size	300	每個切塊最多 300 個字元
chunk_overlap	100	相鄰切塊重疊 100 個字元
separators	"\n\n","\n","。","！","？","；","，","""	優先在段落、句號等語意邊界處切割

三、語意切塊（Semantic Chunking）

先按中文句號切分成句子，再透過 Embedding API（維度 4096）計算相鄰句子的餘弦相似度，當相似度低於門檻時斷開，形成語意段落。過長段落再以 RecursiveCharacterTextSplitter 細切。

參數	設定值	說明
similarity_threshold	0.5	相鄰句子相似度低於此值時斷開
max_chunk_length	500	超過 500 字的語意段落再細切
sub_chunk_size	400	細切時的 chunk_size
sub_chunk_overlap	50	細切時的 overlap

四、各方法切塊數量統計

資料檔案	字元數	固定大小	滑動視窗	語意切塊
data_01.txt	8,038	27	40	46

data_02.txt	2,189	8	11	10
data_03.txt	1,109	4	5	6
data_04.txt	2,178	8	11	16
data_05.txt	5,206	18	28	26
合計	18,720	65	95	104

Q2：哪一種切塊方法效果最好（平均值分數最高）？

一、改良策略說明

本次採用 Top-3 檢索 + LLM 萃取答案 的改良策略。每題從 Qdrant VDB 檢索相似度最高的 3 個切塊（Top-3），再透過 LLM（google/gemma-3-27b-it）從多個切塊中萃取精準答案，最後將 LLM 生成的答案提交至評分 API（hw-01.wade0426.me/submit_answer）。此策略解決了直接提交 raw chunk 時的「雜訊過多、內容不完整」問題。

二、改良前後分數比較

切塊方法	改良前（raw chunk）	改良後（LLM 萃取）	提升幅度	提升比例
固定大小	0.692383	0.795215	+0.102832	+14.9%
滑動視窗	0.702954	0.785644	+0.082691	+11.8%
語意切塊	0.693921	0.798096	+0.104175	+15.0%

三、各方法平均分數比較（改良後）

切塊方法	平均分數	最低分	最高分	排名
語意切塊（Semantic）	0.798096	0.6826	0.9434	第 1 名
固定大小（Fixed-size）	0.795215	0.6787	0.9258	第 2 名
滑動視窗（Sliding Window）	0.785644	0.6255	0.9434	第 3 名

結論：語意切塊效果最好，平均分數 0.798096。 所有 60 筆分數均 ≥ 0.6。

四、20 題各方法分數明細

Q	題目摘要	固定大小	滑動視窗	語意切塊
1	校正流程分為哪兩種？	0.8667	0.8687	0.9126
2	衛福部長石崇良首波改革方向	0.7183	0.7183	0.7080
3	伊朗2025年大規模抗爭經濟與環境主因	0.6865	0.6860	0.6963
4	「復辟巴勒維王朝」口號背後民意	0.8921	0.7695	0.7798
5	何謂 OSINT？運作方式為何？	0.7939	0.8101	0.8145
6	衛星影像監測油槽可分析什麼情報？	0.8745	0.8745	0.8745
7	美國與海灣盟邦對軍事介入伊朗保留態度	0.6787	0.6821	0.6948

8	二戰後全球核試驗次數與地點特徵	0.8906	0.8687	0.8604
9	守望亞洲計畫功能與應用舉例	0.8145	0.7915	0.7915
10	健保兆元時代面臨哪些挑戰？	0.7578	0.7295	0.7578
11	伊朗2026年1月鎮壓的具體手段	0.8330	0.8579	0.8564
12	福衛九號的技術特色	0.8833	0.8833	0.8872
13	魯尼特圓頂目前面臨什麼危機？	0.7715	0.7388	0.7114
14	福衛八號自主研發的突破	0.9258	0.9434	0.9434
15	六大皆空是指哪六個科別？	0.6826	0.6826	0.6826
16	為何需自製衛星而非向美國購買？	0.7310	0.7314	0.7764
17	何謂「直美（Chokubi）」現象？	0.7197	0.7197	0.7549
18	鍶-90 與鈽-239 對人體的危害	0.8921	0.8921	0.8921
19	2026年為何是核武威脅關鍵時間點？	0.7598	0.8394	0.7998
20	福衛八號星系對國安監測能力提升	0.7319	0.6255	0.7676

五、各題最佳方法分佈統計

最佳方法	獲勝題數	佔比
語意切塊	13 題	65%
固定大小	9 題	45%
滑動視窗	7 題	35%

六、分析與討論

（1）改良策略為何大幅提升分數？

改良前直接將 raw chunk 提交評分，chunk 中包含大量與問題無關的雜訊文字，且受限於 chunk_size 可能遺漏關鍵資訊。改良後的 Top-3 + LLM 策略帶來兩個優勢：（a）Top-3 檢索讓 LLM 能看到更多相關段落，降低遺漏風險；（b）LLM 能從多個段落中精準萃取問題所需的關鍵資訊，去除雜訊。三種方法平均分數均從約 0.69 提升至約 0.79，提升幅度約 13~15%。

（2）語意切塊為何在改良後表現最佳？

語意切塊平均分數 0.798096
領先其他方法。語意切塊根據句子間的語意相似度進行斷句，能產生語意完整的段落，當 LLM 接收到語意完整的 Top-3 段落時，能更準確地理解上下文並萃取答案。例如 Q14（福衛八號自主研發）語意切塊與滑動視窗均達 0.9434 最高分。

（3）固定大小切塊的表現分析

固定大小平均分數 0.795215，位居第二且與語意切塊差距僅 0.003。這說明當有 LLM 做後處理時，即使切塊品質較差（無語意邊界），LLM 仍能從多個 chunk 中拼湊出正確答案。固定大小在 Q4（巴勒維王朝）達到 0.8921 的最高分，在 Q8（核試驗）也達到 0.8906。

（4）滑動視窗的表現分析

滑動視窗平均分數 0.785644，位居第三。改良前滑動視窗是最佳方法（因 overlap 保留了上下文），但加入 LLM 後優勢被削弱——因為 LLM 能從 Top-3 段落自行補全上下文。滑動視窗在 Q14（福衛八號）達到 0.9434、Q19（核武威脅）達到 0.8394。

（5）改良效果最顯著的題目

改良效果最顯著的是原先低於 0.6 的 6 個項目，全部成功提升至 0.6 以上：Q3 語意切塊從 0.4851 → 0.6963（+43.5%）；Q6 固定大小從 0.5698 → 0.8745（+53.5%）；Q13 固定大小從 0.5264 → 0.7715（+46.6%）；Q19 語意切塊從 0.5854 → 0.7998（+36.6%）。這些題目原本因 chunk 截斷或雜訊導致低分，LLM 萃取完美解決了這些問題。

七、總結

在本次作業的 5 份資料、20 題檢索任務中，採用 Top-3 + LLM 萃取答案 策略後，語意切塊以平均分數 0.798096
表現最佳，其次為固定大小（0.795215），滑動視窗（0.785644）最低。相較改良前（平均約 0.696），改良後平均約 0.793，提升約 14%。更重要的是，所有 60 筆分數均 ≥ 0.6，成功達成目標。此結果說明：當有 LLM 做答案萃取時，語意切塊的語意完整性優勢能被充分發揮；而切塊方法的選擇對最終品質的影響，會隨著後處理能力的增強而趨於收斂。