

<https://l.pulipuli.info/24/nccu/rag>

2024

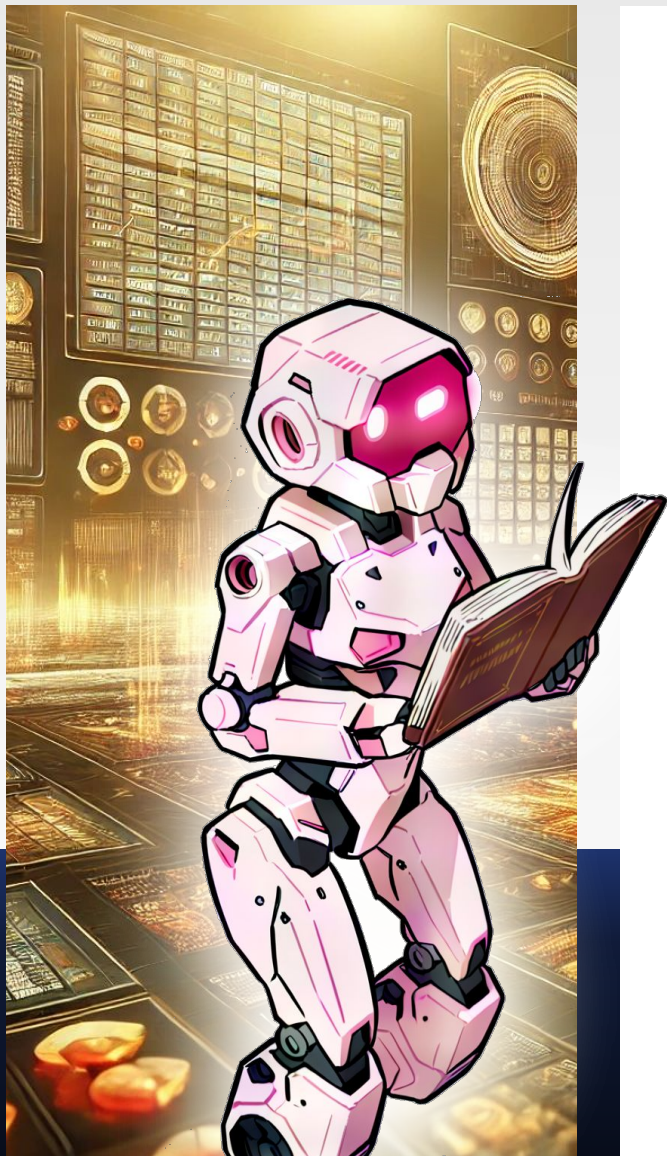
RAG簡介

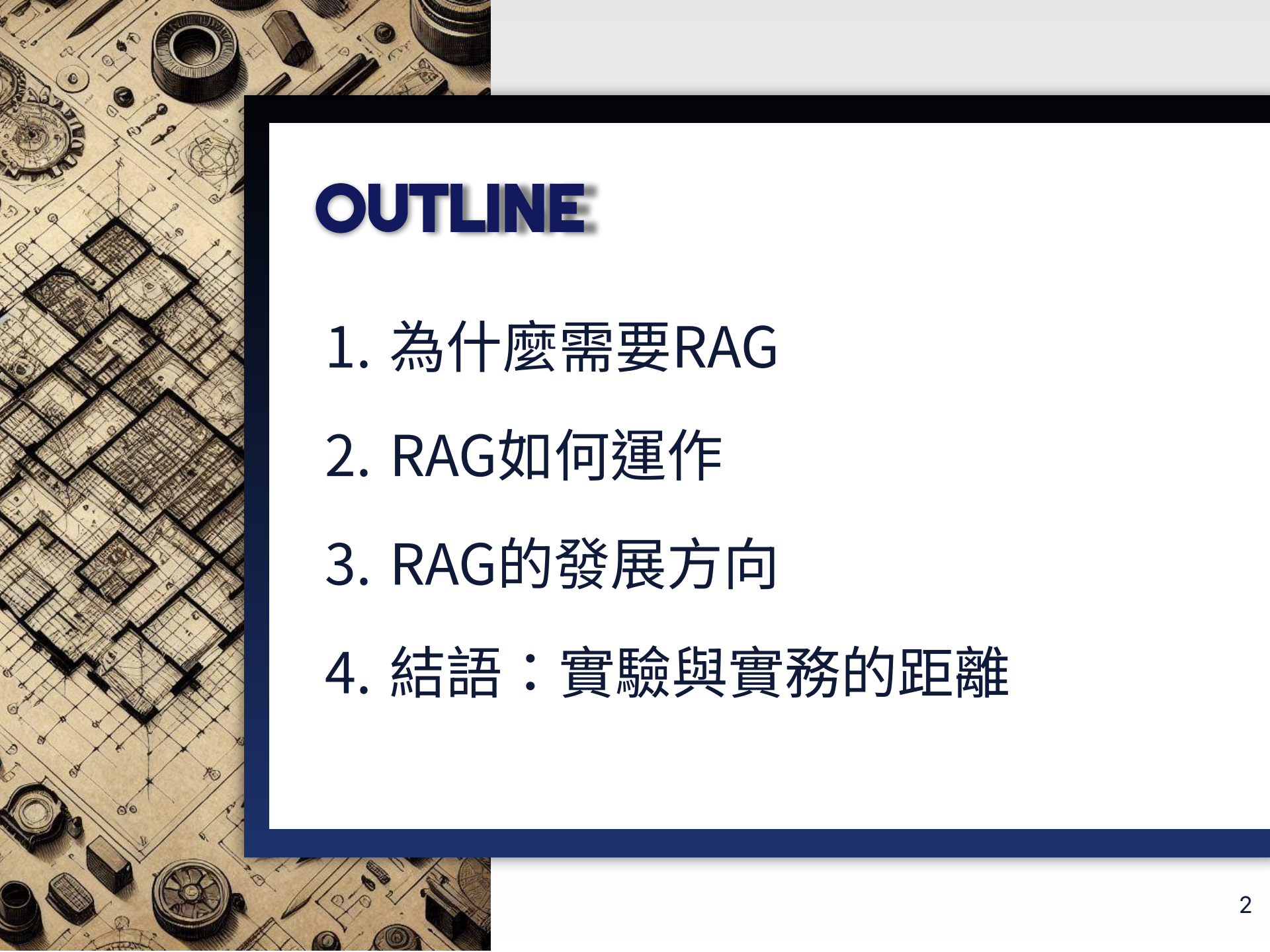
現況、原理、發展

政治大學 圖書資訊與檔案學研究所 博士

陳勇汀

pulipuli.chen@gmail.com





OUTLINE

1. 為什麼需要RAG
2. RAG如何運作
3. RAG的發展方向
4. 結語：實驗與實務的距離

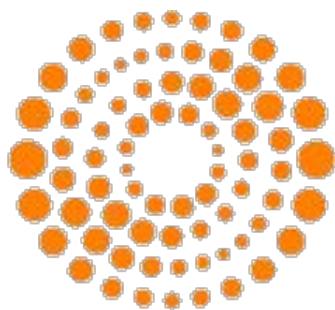


為什麼需要RAG



研究：站在巨人的肩膀上

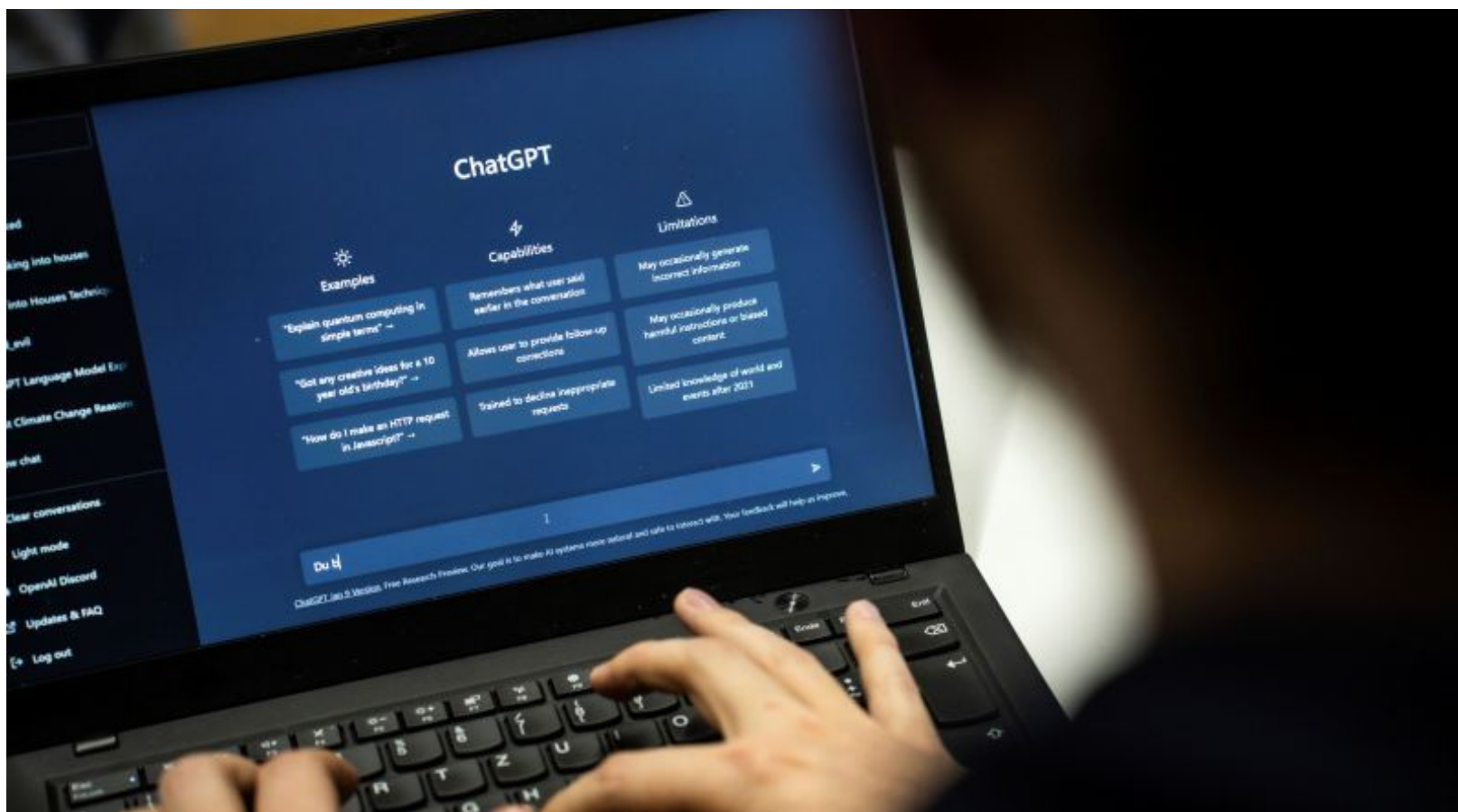
資料查詢：資料庫



WEB OF SCIENCE



資料查詢：大型語言模型



(Large Language Models, LLM)
為什麼會用大型語言模型來查詢？

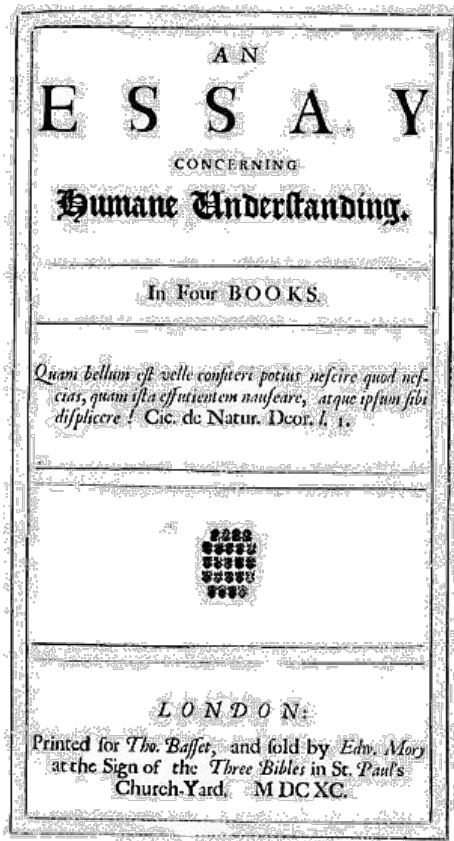


因為不知道
「怎麼問問題」！

我認為人類內心獲得的所有觀念和知識，
都是從**實際經驗**中獲得！



在**哲學界**中，這是誰的論點呢？



英國哲學家

John Locke

(1690)



An Essay Concerning Human Understanding

人類理解論



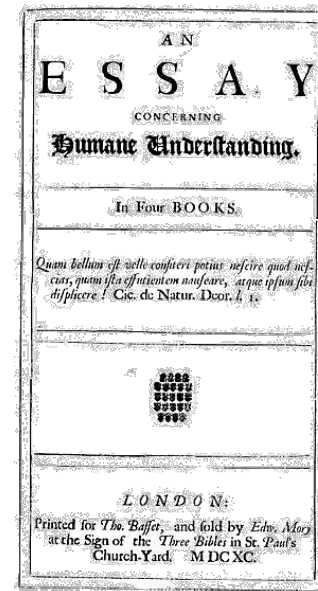
問題跟解答之間的落差

我認為人類內心獲得
的所有觀念和知識，
都是從實際經驗中獲得！



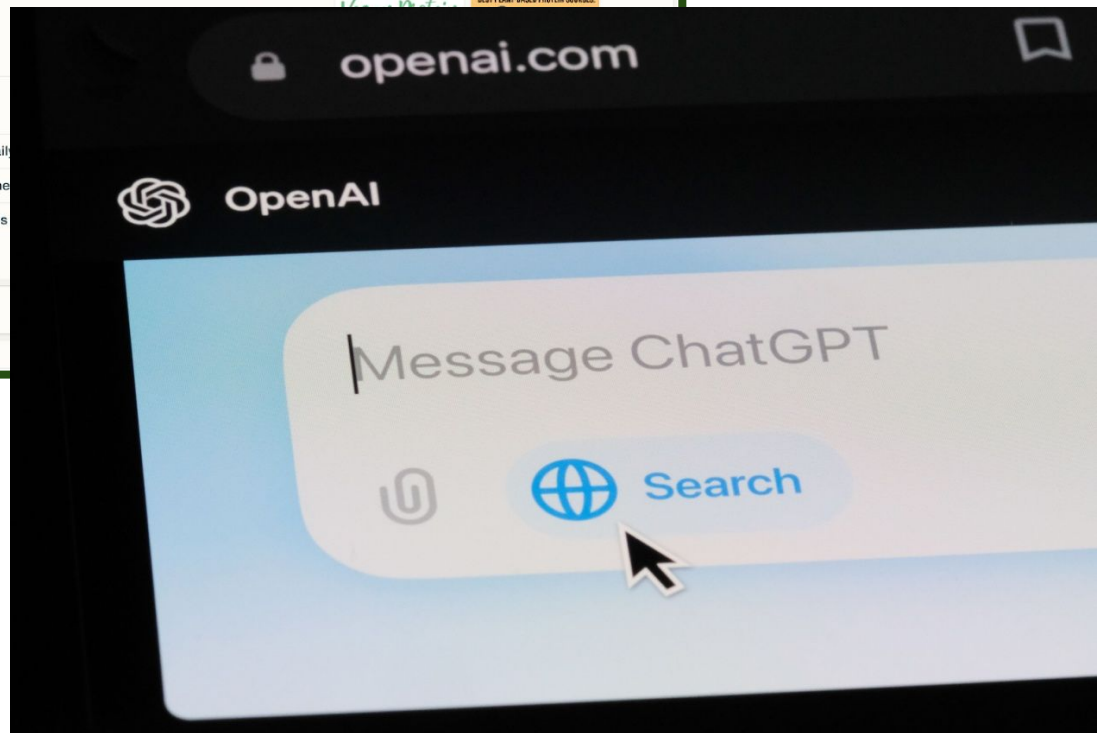
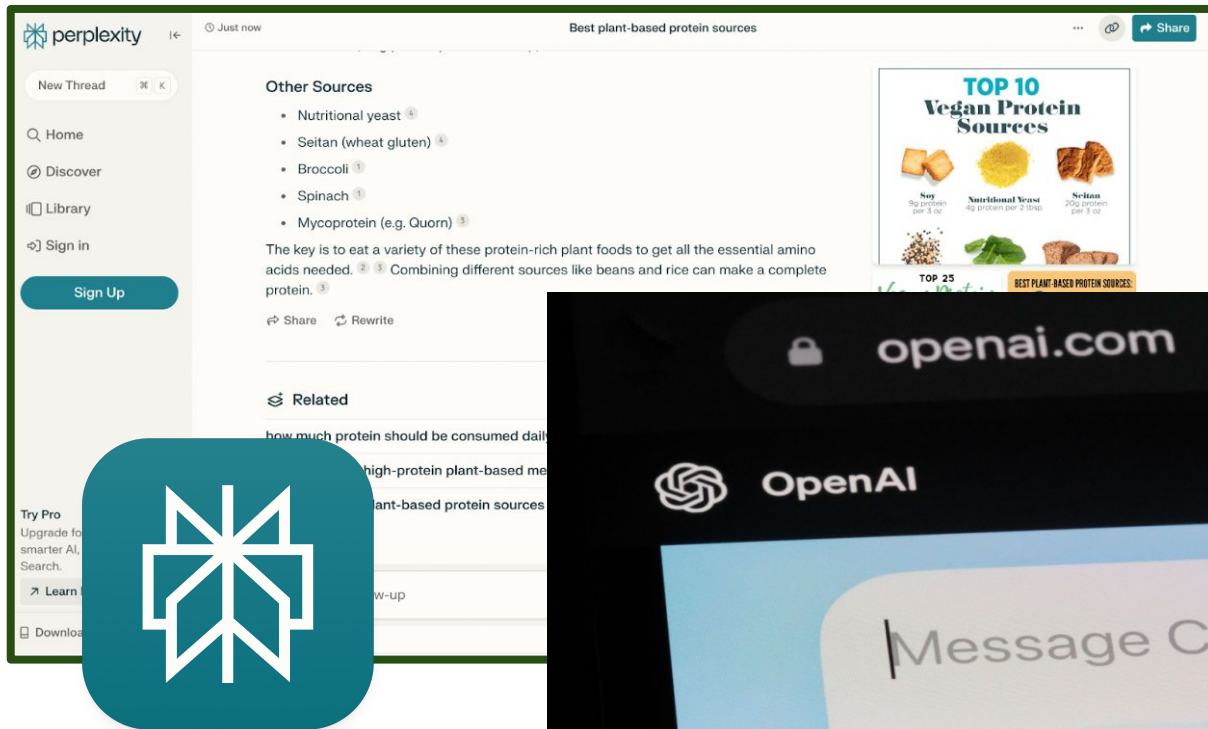
An Essay Concerning Human
Understanding

人類理解論



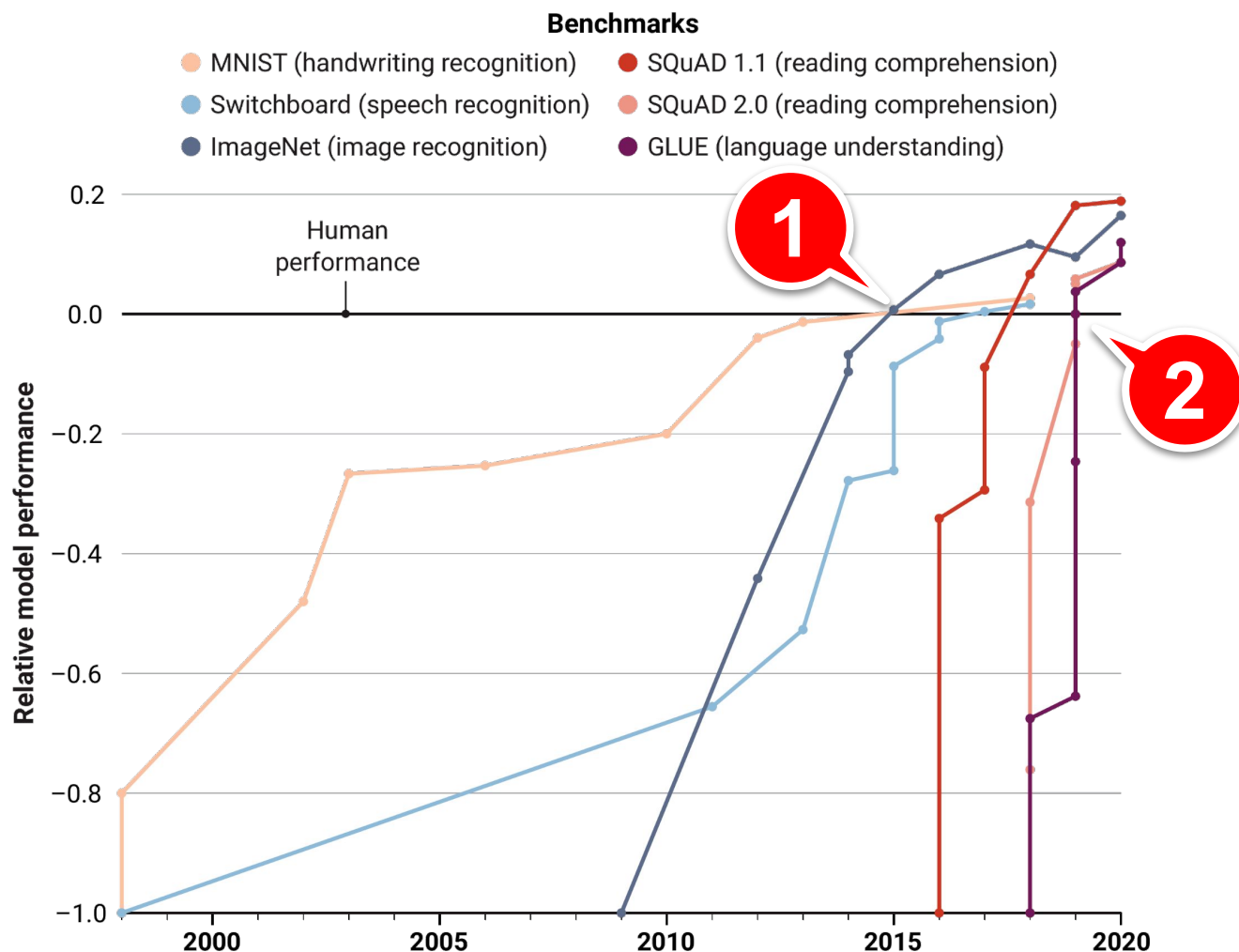
大型語言模型是如何解決檢索問題

- 以**語意**理解文字的意義，而非單純以關鍵字來理解 ⇨ Embedding
 - 用**前後文**能區別「杜鵑」是植物還是動物
 - 能知道「土豆」跟「花生」是指同一種概念
- 以**對話脈絡**來調整要找尋的問題
 - 傳統資料庫：更多關鍵字篩選 ⇨ 查無資料
 - 給LLM更多線索，LLM就能找到更合適的資訊



問問題前 先問Google 先問ChatGPT

1. 手寫識別跟圖像識別已經超越人類表現
2. 閱讀跟語文理解已經超越人類表現



聽說AI已經超越人類了

Prompt

JPG



請問這題要給多少分？滿分20分。

GPT4

這是一道中文的**物理題目**，讓我來幫你評分。

題目問的是關於**電流**和**電壓**的測量問題：

- a. 解釋實驗目的與使用電壓表和電流表的正確接法。
- b. 題目要求繪製電路實驗時的中三個重點。
- c. 說明實驗中使用電流表的其中一個重點。

學生的回答：

按照給分標準.....學生可以獲得約**15分**(4+6+5)的得分。

來考考AI

Q1. 主題編目跟記述編目的差別 (20)

「咦？不是都只是編輯目錄嗎？」希希助教露出疑惑的表情，顯然是不懂這主題編目跟記述編目兩者的差別。

a. 請問主題編目跟記述編目這兩者的主要差異為何？

b. 請列舉記述編目中需要著錄的其中三個項目。

c. 請列舉主題編目中需要著錄的其中一個項目。

A: 主題編目主要是針對資源的「內容」做分類
記述編目主要是針對資源的「形式」，如傳記、期刊

B: 題名、目次、篇章名

C: 正文

問題：主題編目與記述編目的差別

大型語言模型的侷限

- **幻覺 (hallucination)**：LLM可能會產生缺乏根據的事實，或著是順著提示詞整理出乍看之下流暢的回答。
- **缺乏專業知識**
 - 建立LLM模型的資料來源大多是既有的資料集或是網路可找尋的資料，這些資料會形成LLM的世界知識。
 - 世界知識缺乏小眾的專業領域資訊，資料也無法即時更新。
- **難以調整：**
 - 大型語言模型的訓練和微調(fine-tuning)所需的資源已經不是一般單位所能負荷的層級。
 - 導致LLM僅能處理普遍性的問題，難以應用到真實世界中。

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score
1	1	Gemini-Exp-1206	1372
1	1	Gemini-2.0-Flash-Thinking-Exp-1219	1368
1	1	ChatGPT-4o-latest...(2024-11-20)	1364
4	4	Gemini-2.0-Flash-Exp	1354
5	3	o1-preview	1335
6	8	o1-mini	1306
6	7	Gemini-1.5-Pro-002	1302
8	11	Grok-2-08-13	1288
8	12	Yi-Lightning	1287
8	7	GPT-4o-2024-05-13	1285
8	6	Claude 3.5 Sonnet...(20241022)	1283
8	16	Deepseek-v2.5-1210	1278
8	20	Qwen2.5-plus-1127	1278
10	19	Athene-v2-Chat-72B	1277
12	17	GLM-4-Plus	1274
12	19	GPT-4o-mini-2024-07-18	1273
12	20	Gemini-1.5-Flash-002	1271
12	32	Llama-3.1-Nemotron-70B-Instruct	1269



Llama
 開源LLM
 能深入微調

Chatbot Arena LLM Leaderboard

使用需求

This code has been tested on Transformers v4.44.0, torch v2.4.0 and 2 **A100 80GB GPUs**.



A100 80GB 300W PCIE GPU

NTD 735,500

~~可以買內湖1坪~~

Prompt Learning 提示詞學習 (1/2)

提示詞工程

- 比起採用重新訓練或微調LLM，以提示詞來增強LLM的能力，成為了新的典範。
- 提示詞工程 (prompting engineering) :
 - 精心設計提示詞，引導LLM能夠順利完成下游任務 (downstream tasks)
 - 舉例：聊天機器人的角色設定就是固定的提示詞
 - 主要著重於回應形式、回答範圍
 - 然而，若要處理新的任務，LLM仍需要更多專業領域的知識

- **A 明確說明指示**：清楚說明要執行的任務，例如「撰寫」、「分類」、「摘要」或「翻譯」。
- **I 指出提示元素**：告知模型指示和輸入的資料，以及預期輸出的格式應該是什麼。
- **P 提供結尾提示和上下文**：提供模型明確的結尾提示，以及情境。
- **R 完善指示並避免歧義**
- **O 提供回饋和示例**
- **M 管理互動**：要求它提出反對意見或指出你的想法中的缺陷
- **T 跟踪字元長度並切分任務以降低複雜度**

AIPROMT架構

ChatGPT Plus功能 **MYGPT**

每次開啟對話時
即會送出的預設
提示詞

中英文論文題目產生機

Published · Everyone

CreateConfigure

Name

中英文論文題目產生機

Description

Generates academic paper titles in Chinese and English.

Instructions

Scholarly Title Creator generates academic paper titles in Chinese and English, covering a broad range of subjects including specific fields like 'information retrieval' (資料檢索). It assists researchers and academics by crafting titles that are coherent, relevant, and academically sound. The GPT understands both common and specialized research areas, ensuring titles accurately reflect the user's content.

Conversation starters

Generate a title for my paper on AI ethics

×

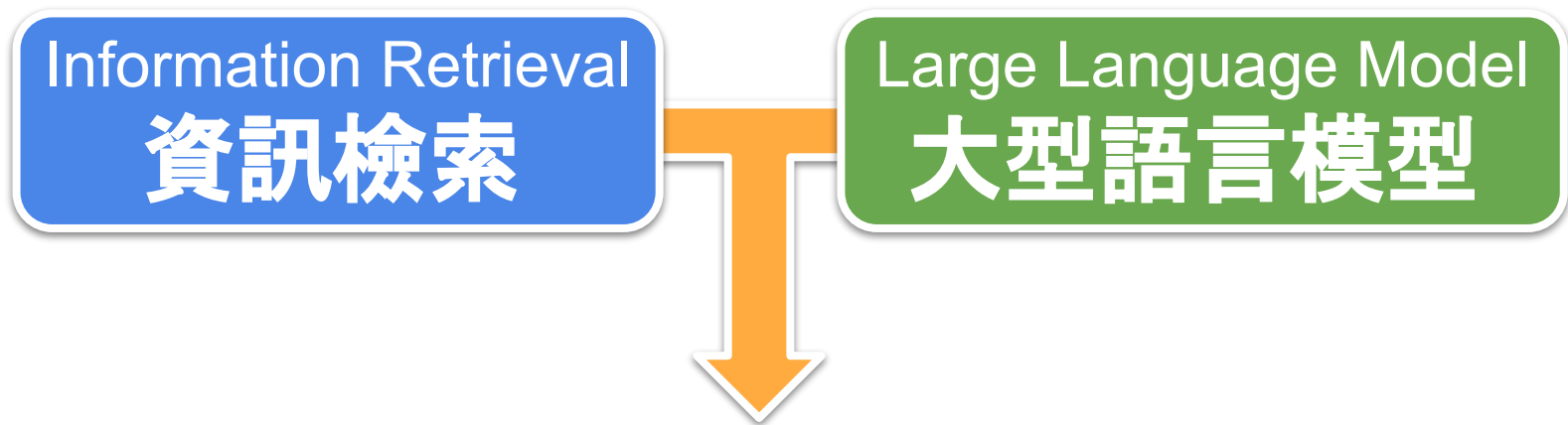
Suggest a title for a study about ancient Chinese history

×

Prompt Learning 提示詞學習 (2/2)

在脈絡中學習

- 為了克服提示詞學習遭遇的困難，研究者開始嘗試改善提示詞提供的資訊
- **在脈絡中學習 (In-Context Learning, ICL):**
 - 在提示詞中提供一些示範，讓LLM學習如何處理新的任務。
 - 適合小數據任務 (few-shot tasks): 只要提供合適的示範，GPT-3也能夠正確回答。
 - 但主要的難題是如何選出高品質的示範！
 - 太多不相干的示範，反而會讓LLM造成幻覺。



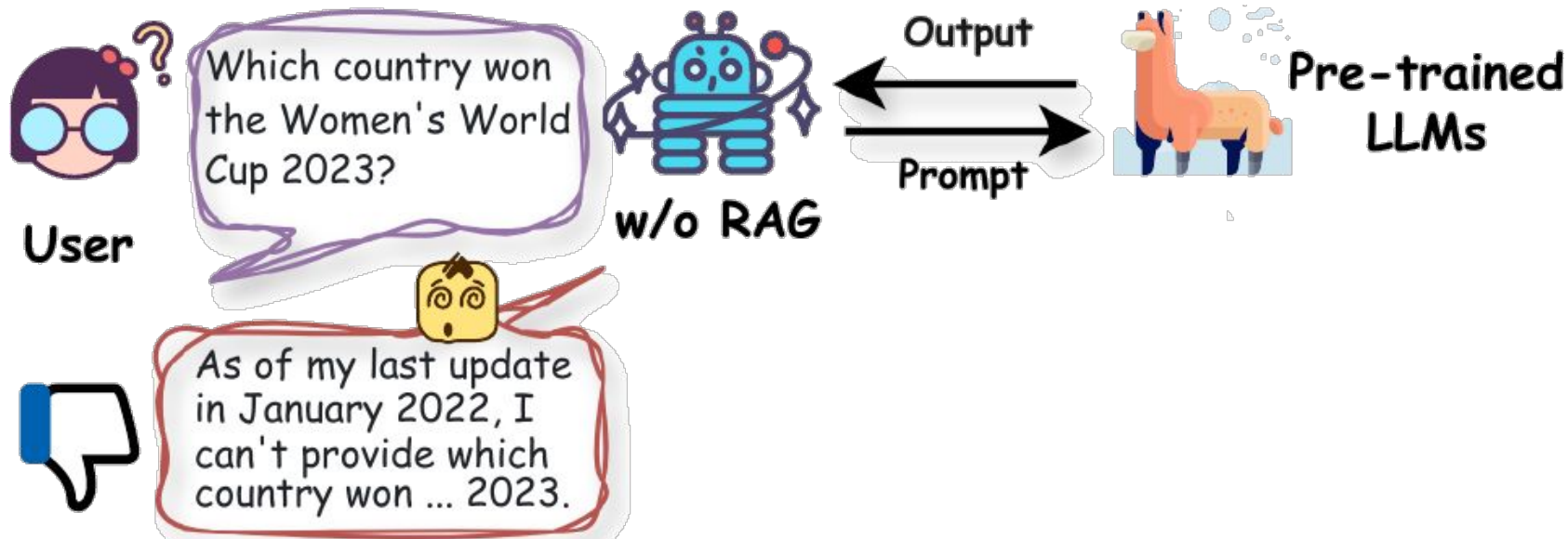
Retrieval-Augmented Generation

RAG

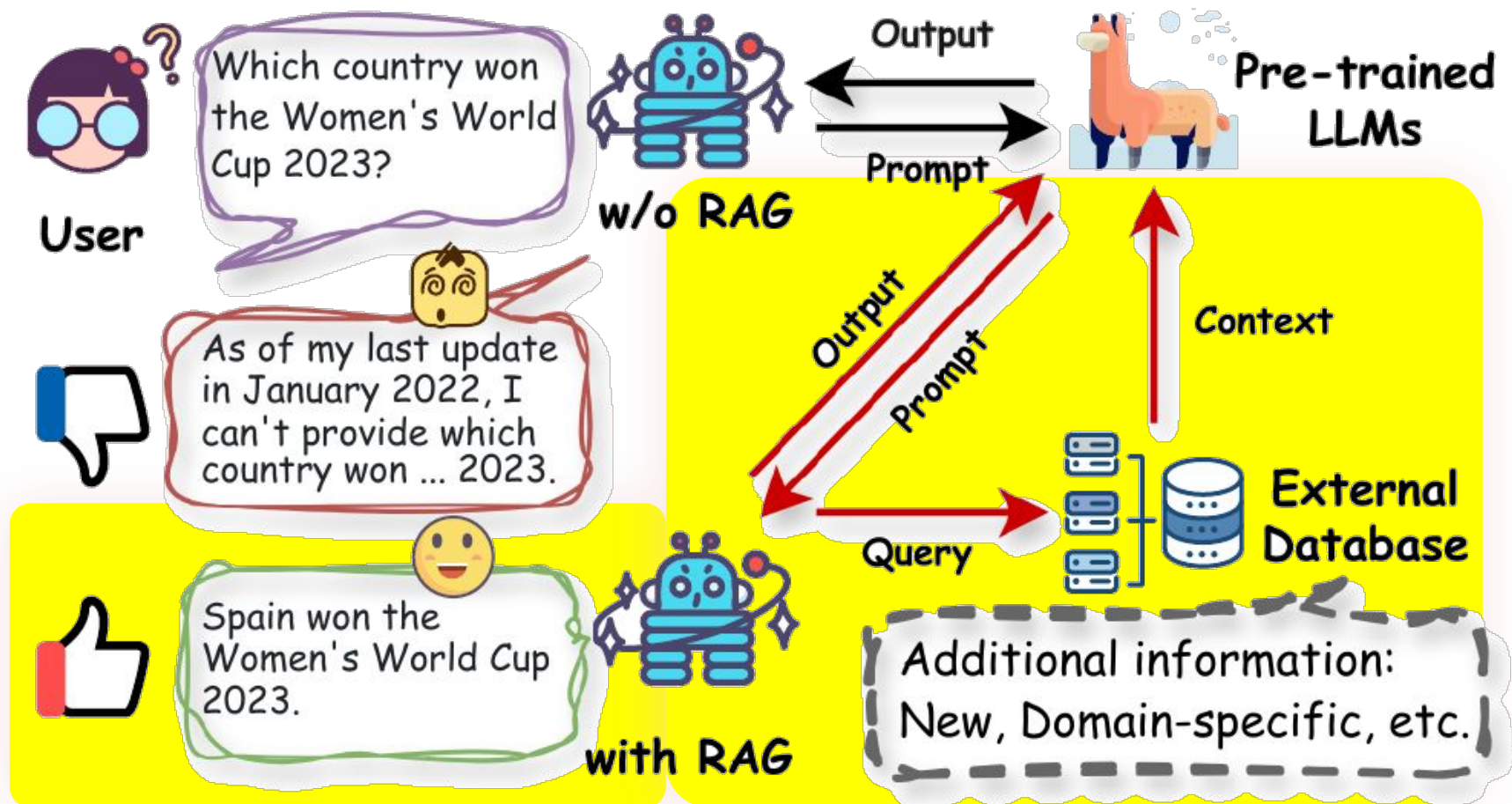
檢索 增強 生成



RAG如何運作



僅有大型語言模型



當RAG遇上LLMs

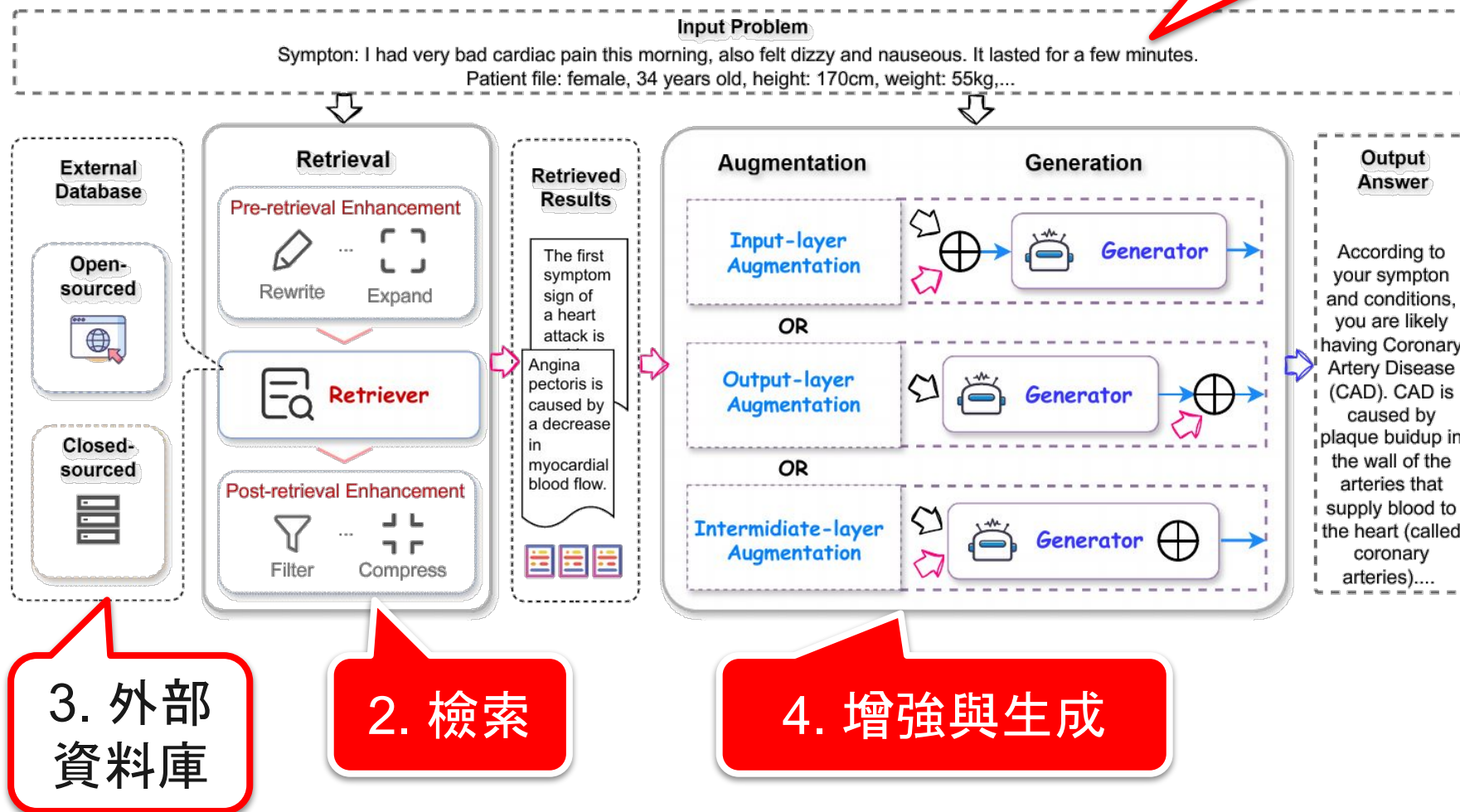
Retrieval-Augmented Generation (RAG)

檢索增強生成

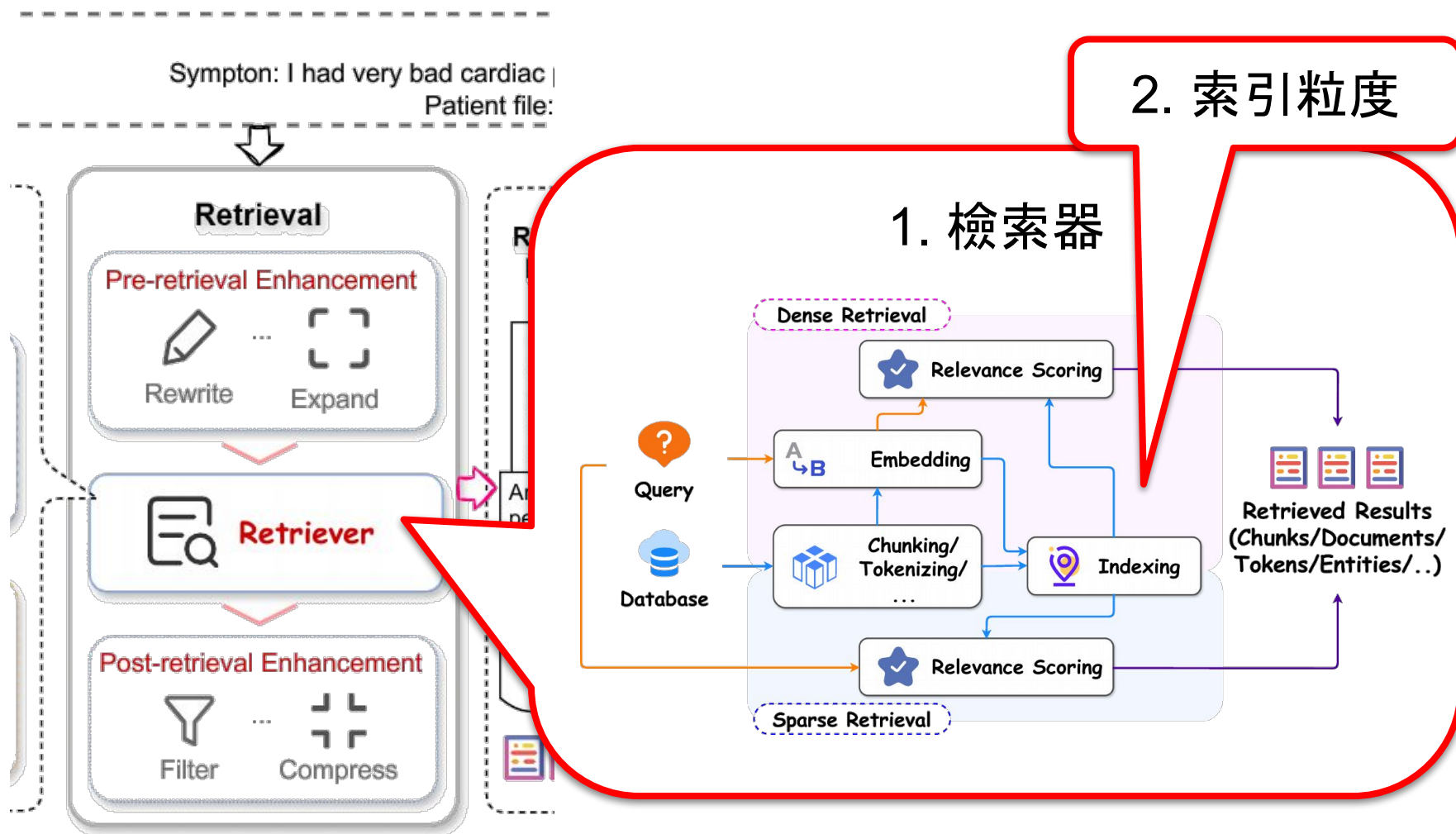
- RAG的主要特色在於透過**資訊檢索系統**從外部資料庫獲取相關資訊，為大型語言模型提供即時更新、專業領域相關的外部知識，以此減少大型語言模型生成內容中的錯誤。
- RAG常見的做法是調用檢索器從外部資料庫中搜索並提取相關文件，**查詢結果文件會與查詢語句結合成為增強的提示詞**，最後用於大型語言模型的生成。
- 在實務中，RAG的可行性和有效性已經被許多研究證實。不需要重新訓練大型語言模型，只需要簡單地調整RAG中的檢索元件和外部資料庫，即可實現**小資料學習**的效果。

RAG框架

1. 查詢語句

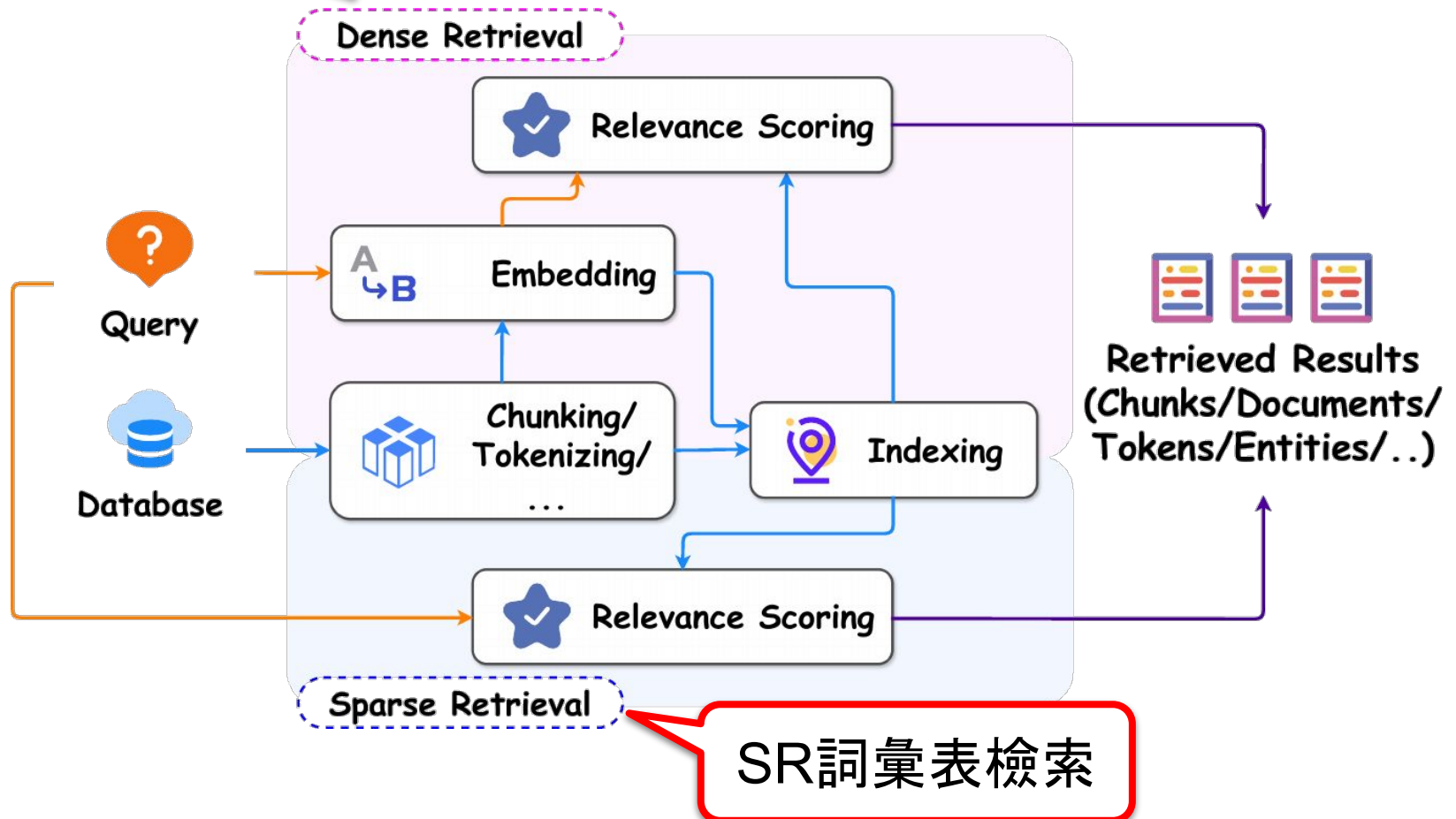


檢索



檢索器

DR語意向量檢索



SR詞彙表檢索



文獻全文

兄弟們！來！來！
來和他們一拼！
憑我們有這一身，
我們有這雙腕，
休怕他毒氣、機關槍！
休怕他飛機、炸裂彈！
來！和他們一拼！

——賴和《南國哀歌》



詞彙表

- | | | |
|------|------|-------|
| ● 兄弟 | ● 有 | ● 機關槍 |
| ● 們 | ● 這 | ● 飛機 |
| ● 來 | ● 一身 | ● 炸裂彈 |
| ● 和 | ● 這雙 | |
| ● 他們 | ● 腕 | |
| ● 一 | ● 休 | |
| ● 拼 | ● 怕 | |
| ● 憑 | ● 他 | |
| ● 我們 | ● 毒氣 | |

查詢
「毒氣」

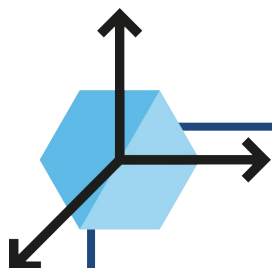
DR語意向量檢索 (embedding)



德國



Pre-trained
Embedding Model



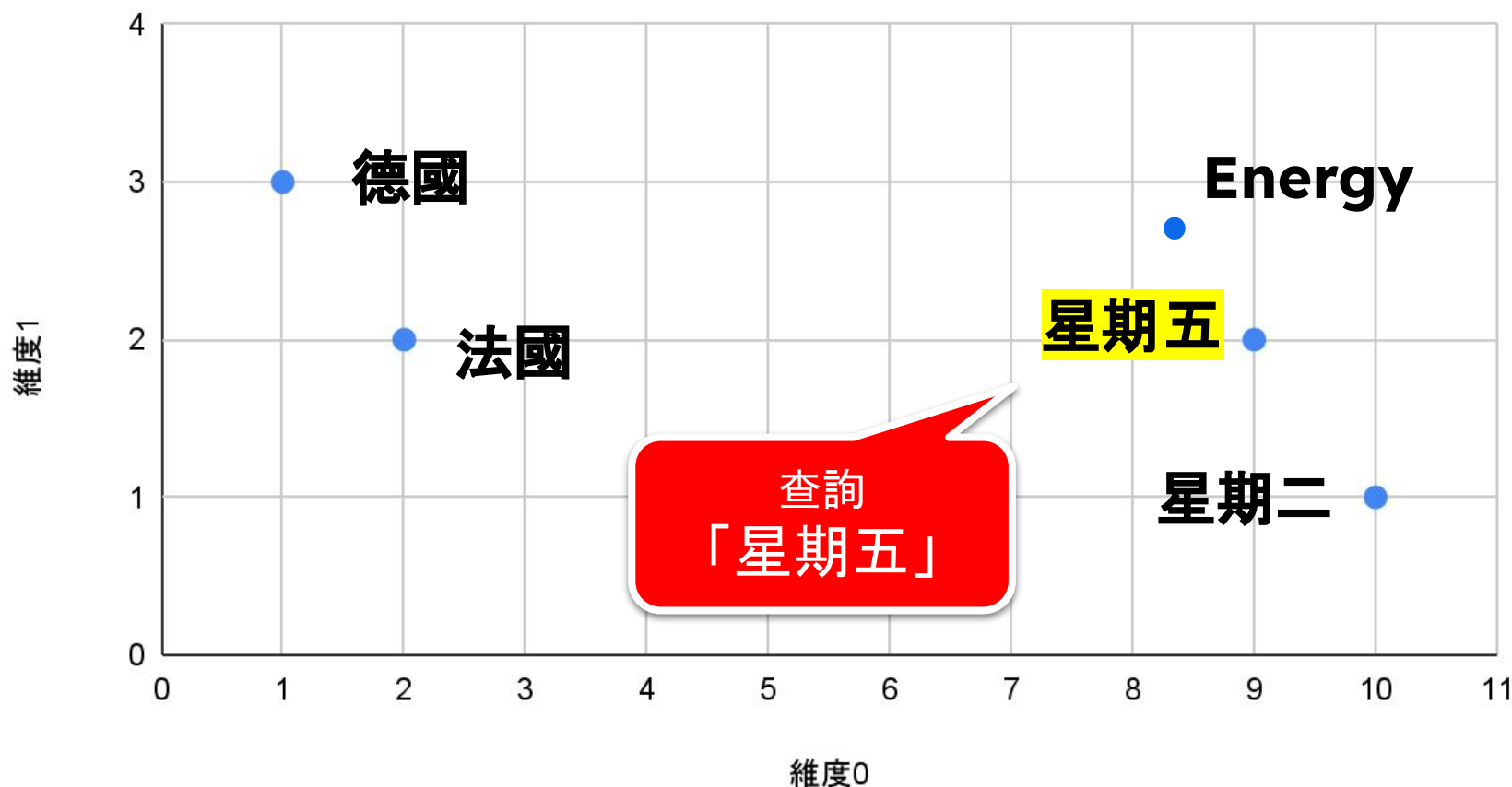
$[-.04, -.05, .05, 0.05, 0,$
.....
 $.06, -0.04, -.05, .03, .06]$

簡化版語意向量的例子(1/2)

文字	語意向量	
	維度0	維度1
德國	[1,	3]
法國	[2,	2]
星期五	[9,	2]
星期二	[10,	1]

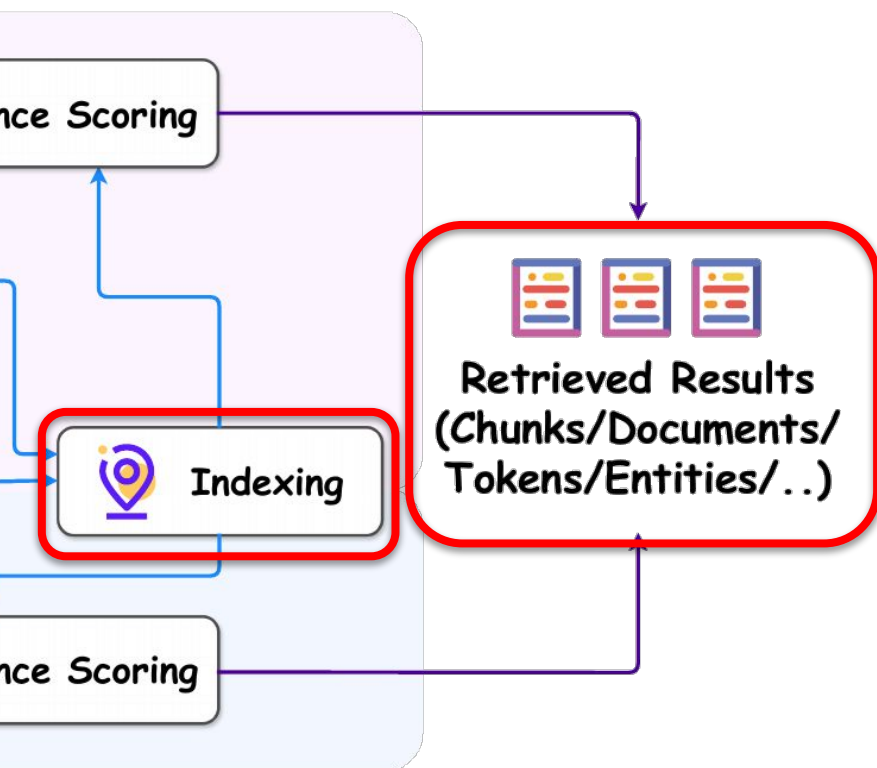
簡化版語意向量的例子(2/2)

Cosine Similarity相似度檢索



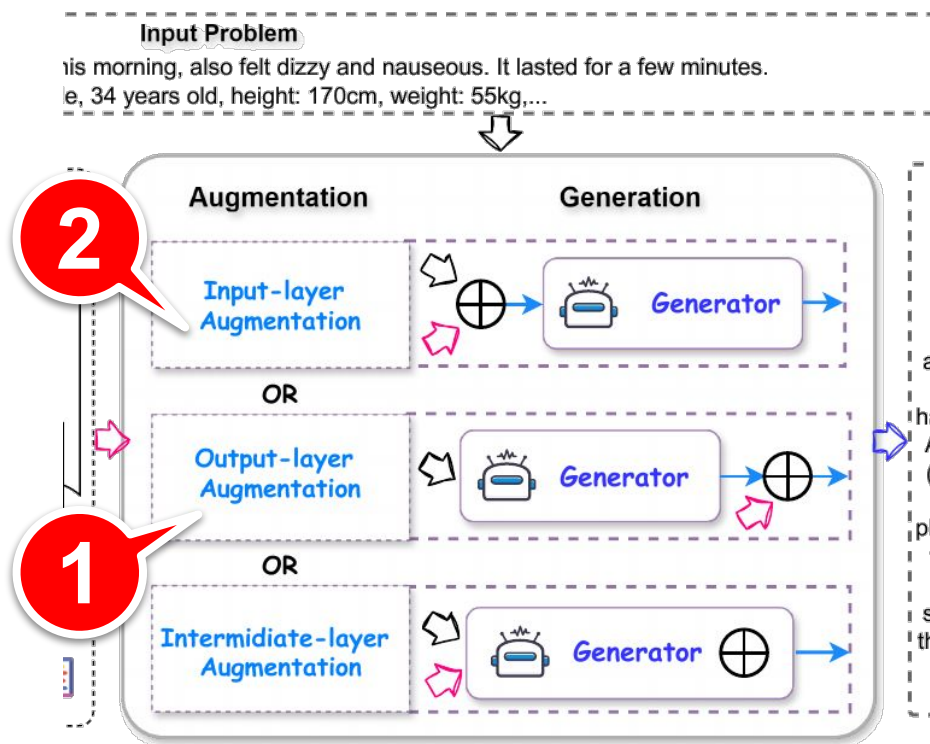
Retrieval Granularity

索引粒度



- 定義被索引文本的大小
 - ex: 整份文件、段落、一句話、一個實體(人名、地名、時間)
- **Chunk Retrieval**: 以段落為主的粒度最流行
- 每個被索引的文本結果又稱為切片 (slice)
 - 被找到的切片叫做回應 (recall)
- 粒度的選擇對RAG的結果影響甚鉅

Retrieval Integration for Generation Augmentation 檢索結果用於生成增強 (1/2)



1. 輸出層增強

- 原始查詢送給LLM產生結果，然後再將檢索結果一併呈現

2. 輸入層增強

- 將檢索結果跟查詢語句連接在一起，然後送給LLM生成
- 要注意整合後的提示詞可能會過長

About 43,900,000 results

Any time ▾

How to throw a vegetarian dinner party (my 5 top tips!)

<https://www.pawprint.eco/eco-blog/vegetarian-dinner-party> ▾

Web How to throw a vegetarian dinner party (my 5 top tips!) 7 min Read. There's nothing better than having friends over for dinner (Corona permitting). Enjoying good food in good ...

Estimated Reading Time: 6 mins

EXPLORE FURTHER

-  18 Swanky Recipes to Throw the Most Epic Vegetarian Din... [brit.co](#)
-  56 Vegetarian dinner party recipes - delicious. magazine [deliciousmagazine.co.uk](#)
-  Best Vegetarian Recipes for Entertaining | Martha Stewart [marthastewart.com](#)
-  Vegetarian Dinner Party Recipes - Great British Chefs [greatbritishchefs.com](#)
-  117 Vegetarian Dinner Recipes for a Dinner Party | Epicurious [epicurious.com](#)

Recommended to you based on what's popular • Feedback

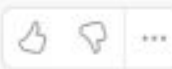
Vegetarian and Gluten-Free Dinner Party Ideas | Bon Appétit

<https://www.bonappetit.com/trends/article/...> ▾

Web Oct 2, 2012 - Pull off some plating magic. If the guest can only eat the vegetable side dish, cut the veggies in a way that you can stack them, then sprinkle crushed nuts on top and ...

Here's How To Get The Most Out Of Your Chat AI ...

Bing搜尋結果



Sure, I can suggest a 3-course menu with a chocolate dessert for your vegetarian dinner party. Here are some possible options:

- Starter: **Wild Mushroom Tartlets with Onion Sauce** ¹ or **Vegan Popcorn Tofu Nuggets** ²
- Main: **Vegan Butternut Squash Mac** ² or **Vegetarian Three-Bean Chili** ³
- Dessert: **Chocolate Lava Cake** ⁴ or **Chocolate Pasta with Chocolate Hazelnut Cream Sauce, White Chocolate Shavings, and Fresh Berries** ⁴

I hope this helps you plan your party. Enjoy! 😊

Learn more: [1. booths.co.uk](#) [more](#)

Copilot生成

輸出層增強案例：Bing的AI搜尋



大型語言模型

提示詞工程

在脈絡中學習

輸入層增強案例：Dify



RAG如何運作

實際案例比較

PROMPT

**請撰寫「研究對象」
相關的研究限制**

直接問大型語言模型

請撰寫「研究對象」相關的研究限制。

好的，以下提供一些關於「研究對象」相關的研究限制，您可以根據您的研究情境和具體情況進行調整和選擇：

研究對象相關的常見研究限制：

1. 抽樣偏差 (Sampling Bias):
 - **抽樣方法局限性：** 使用的抽樣方法（如方便抽樣、滾雪球抽樣）可能無法代表母體，導致樣本產生偏差。例如，使用網路問卷可能無法涵蓋所有年齡層或不熟悉網路的族群。
 - **樣本規模不足：** 樣本數量太小，可能無法有效檢驗研究假設，導致結果缺乏統計效力。
 - **自我選擇偏差：** 研究對象是自願參與，可能導致特定類型的參與者過度代表，而忽略其他類型。例如，對某議題特別感興趣的人可能更傾向於參與。

有如教科書的回答...

CHATGPT
回答很爛？
其實是你
不會問問題！

candice_clife

>>>



甘阿捏？

以Dify建構RAG架構



同樣的問題

請撰寫「研究對象」相關的研究限制。



✓ 工作流 >

好的，以下為修改後的「研究對象」相關研究限制，以臺灣學術界常用的語氣撰寫，並盡量保持原段落的長度：

一、研究對象選取之限制

本研究旨在初步探討深度休閒露營者之資訊行為，故於研究對象之選取上，聚焦於特定露營族群進行深入調查。此舉雖能精闢剖析深度休閒露營者之獨特性，然亦同時限縮研究結果之推論範疇。鑑於此，本研究之結論較難直接類推至其他休閒類型之參與者，於詮釋研究結果時應審慎為之。

此外，本研究受限於研究時程與資源，在研究對象之招募上，主要以北部地區之露營者為樣本，可能未能充分反映臺灣其他地區露營者之觀點，甚或未能涵蓋更專業之露營者，例如露營社群媒體創作者、露營相關組織/社團成員等之意見。本研究原預計招募1325位研究對象，包含慣常露營族群與偶爾露營族群，惟受限於研究期間之限制，最終僅能於露營相關社團中徵集慣常露營者，而偶爾露營者之樣本亦未能完整涵蓋各類型露營者。未來研究可朝納入不同地區、不同專業程度之露營者，以期研究結果更具全面性。

引用

<http://192.168.100.202:21080/陳映...>

<http://192.168.100.202:21080/陳映...>

知道研究
主題是
「露營」

知道研究
對象的範
圍

Conversation Log

START 6.659 ms ✓

IF/ELSE 4.977 ms ✓

CODE 94.021 ms ✓

CODE 提取文章 90.573 ms ✓

PARALLEL-1

BRANCH-1-A

KB 主要論文 195.499 ms ✓

BRANCH-1-B

KB 詞彙表 214.701 ms ✓

BRANCH-1-C

IF/ELSE 3 46.460 ms ✓

KB 4. 研究方法 189.129 ms ✓

INPUT

```
1 {
2   "query": "請撰寫「研究對象」相關的研究限制。"
3 }
```

OUTPUT

```
39 },
40 "title": "許 - 4. 研究設計與實施 -
20241213-1053",
41 "content": "註：*N為視覺模式探勘結果之模式
數量；M為語意模式探勘結果之模式數量。\\n\\n
**六**節 研究**範圍與**限制
\\n-----\\n\\n本研究的研究範
於2022年底至2023年初中蒐集大量迷因圖片而成的
迷因現象。這些迷因現象構成的實驗資料集，被本
```

VARIABLE AGGREGATOR 5.006 ms ✓

檢索結果



許 - 論文完整版(含註解)_修改_20240821_0429_陳勇汀.odt — LibreOffice Writer

檔案(F) 編輯(E) 檢視(V) 插入(I) 格式(O) 樣式(Y) 表格(A) 表單(R) 工具(T) 視窗(W) 說明(H)

段落 首行縮排 字體 12 pt B I U S A B I X T

第六節·研究範圍與限制

本研究的研究範圍來於2022年底至2023年初中蒐集大量迷因圖片而成的迷因現象。這些迷因現象構成的實驗資料集，被本研究輸入到 MEME-PAM 並探勘出視覺模式和語意模式，其過程則用於進行表徵評估實驗、分群驗證實驗、以及專家評估實驗。由於本研究選擇了前述的研究範圍和研究方法，進而產生了以下的研究限制。

一、迷因現象來源社群的抽樣限制

本研究分析的迷因現象來自於當時名為 Twitter 的社群網路平台。其中，僅有使用了「#meme」或「#memes」社會性標籤的貼文，才會被本研究用 Twitter-API 與社群網路爬蟲來取得貼文中的迷因圖片。根據 Statista 在 2023 年的統計資料，Twitter 社群網路平台的使用者大多來自於美國、日本、印度、巴西、印尼，其中英文為母語的使用者佔了相對多數，在 Twitter 的使用語言也大多都是英文 (Semiocast, 2024)。而且本研究選擇用來分析語意表徵的深度學習模型也大多只能處理英文文本。因此在本研究所指的迷因現象僅限於解釋 Twitter 社群網路平台上這些以英文為主的網路迷因。其他社群和語言的使用者不在本研究的範圍內。若要分析其他社群或語言，則必須重新設

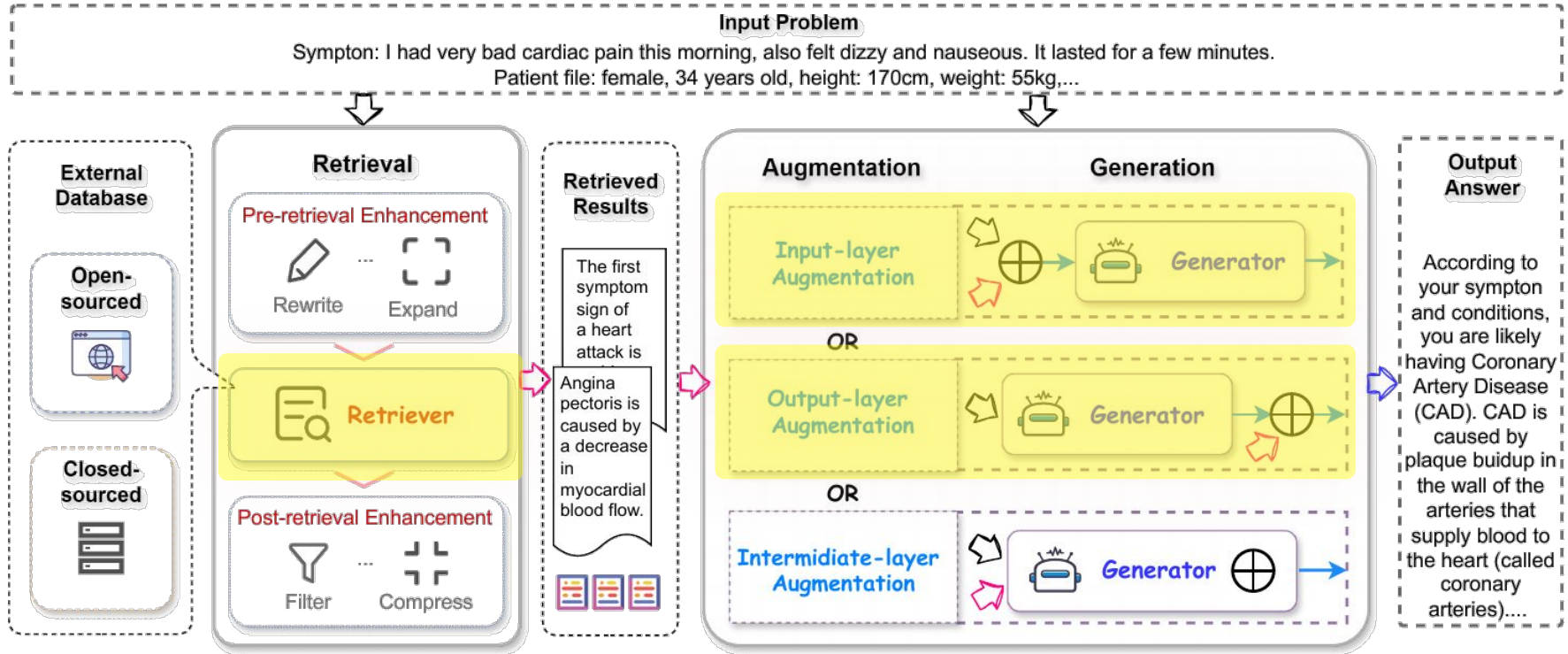
研究限制

第 67 頁，共 124 頁 [第 57 頁] 擷取：4 個字，4 個字元 頁面 3-2 正文 每章第二頁之後 120%

The background of the slide is a complex, abstract digital network. It features a dense web of glowing orange and yellow lines that connect various points, creating a sense of interconnectedness. Interspersed among these lines are several bright blue and white light sources, some of which appear as small, glowing cubes or spheres. The overall color palette is dominated by warm oranges and yellows, contrasted with cool blues and whites. The composition is dynamic, with lines crisscrossing in multiple directions, suggesting a complex, multi-dimensional space.

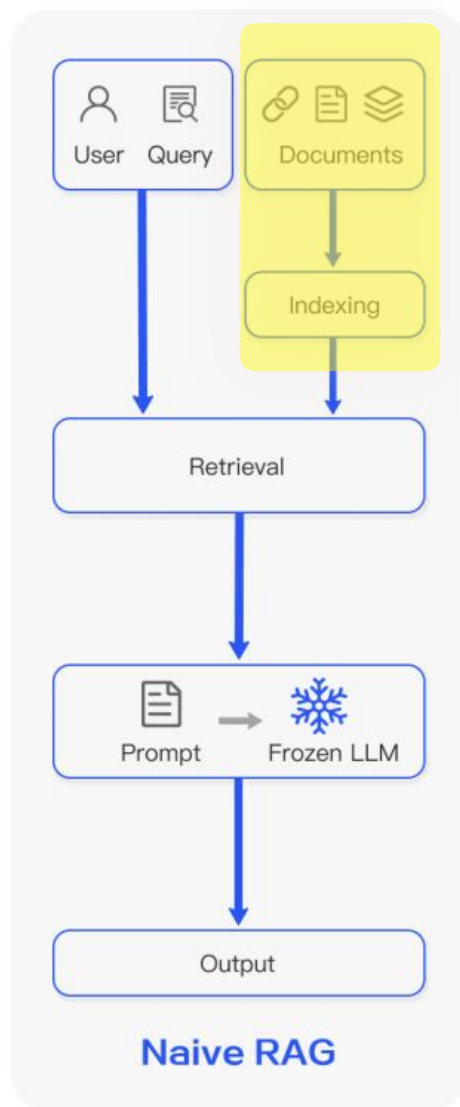
RAG的發展方向

RAG框架



好像...少講了很多東西

在RAG的R之前 知識庫的建置

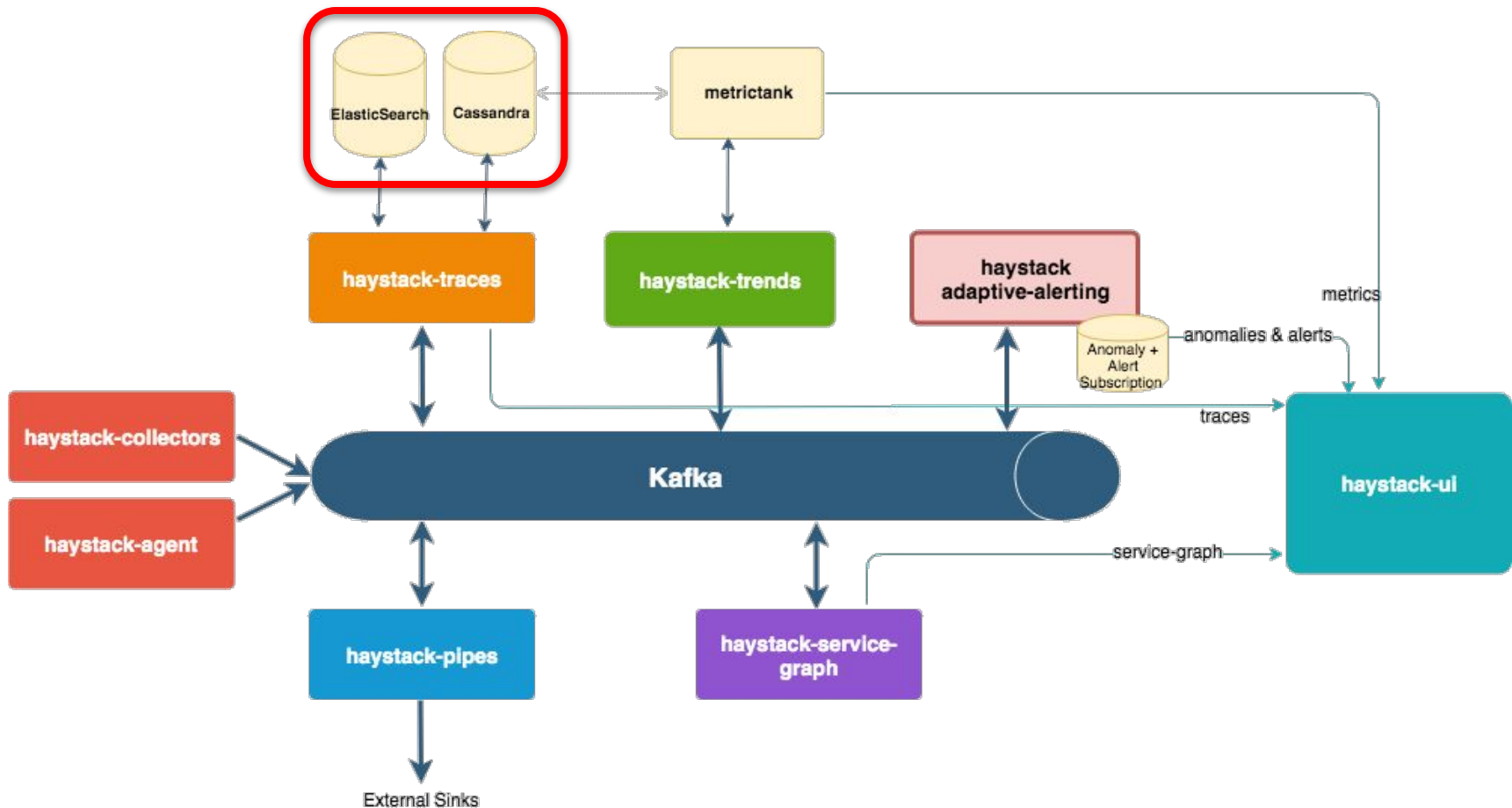


檔案更新的話...

1. 準備好可供分析的檔案
2. 上傳到指定知識庫
3. 建置索引
 - a. 分析檔案內容成為純文字資料
 - b. 分段切片 (Chunking)
 - c. 關鍵字抽取，建立詞彙表
 - d. 語意向量分析
 - i. 將切片傳送到Pretrained-Embedding Model
 - ii. 取得語意向量
 - iii. 將向量結果儲存到向量資料庫
4. 索引建置完成



資料庫整合：Haystack



多媒體資料的索引：Markdown

能抽取文字

- PDF
- PowerPoint
- Word
- Excel
- HTML
- CSV, JSON, XML

需要識別

- 圖片：OCR文字識別
- 聲音：語音轉錄

排版元素識別：DeepDoc

Text

Q: the (normalized) eigenvectors of the matrix $A^T A$ (i.e., $Q^T Q = I$). The columns of Q are called the *right singular vectors* of A .

Text

Δ: the diagonal matrix of the *singular values*, $\Delta = \Lambda^{\frac{1}{2}}$ with Λ being the diagonal matrix of the eigenvalues of matrix AA^T and of the matrix $A^T A$ (they are the same).

Text

The SVD is a consequence of the eigen-decomposition of a positive semi-definite matrix. This can be shown by considering the eigen-decomposition of the two positive semi-definite matrices that can be obtained from A : namely AA^T and $A^T A$. If we express these matrices in terms of the SVD of A , we obtain the following equations:

Equation-Block

$$AA^T = PAQ^T QAP^T = P\Delta^2 P^T = PAP^T, \quad (3)$$

and

Equation-Block

$$A^T A = QAP^T PAQ^T = Q\Delta^2 Q^T = Q\Lambda Q^T. \quad (4)$$

This shows that Δ is the square root of Λ , that P are the eigenvectors of AA^T , and that Q are the eigenvectors of $A^T A$. For example, the matrix:

Equation-Block

$$A = \begin{bmatrix} 1.1547 & -1.1547 \\ -1.0774 & 0.0774 \\ -0.0774 & 1.0774 \end{bmatrix} \quad (5)$$

can be expressed as:

Equation-Block

$$A = PAQ^T$$

$$= \begin{bmatrix} 0.8165 & 0 \\ -0.4082 & -0.7071 \\ -0.4082 & 0.7071 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{bmatrix}$$

Table

Table 1. Diagnostic Performance Characteristics of PSA Testing, According to Cutoff Level.^a

Characteristic	PSA Cutoff Level	
	≥4.0 ng/ml	≥3.0 ng/ml
	percent	
Test positivity	12	18
Cancer-detection rate	3	4
Sensitivity	21	32
Specificity	91	85
Positive predictive value	30	28

^a Data are from Wolf et al.¹¹ PSA denotes prostate-specific antigen.

Text

aggressive treatment of these cancers is associated with unnecessary risks of urinary, sexual, and bowel dysfunction, which can adversely affect the quality of life.²⁸

Text

Paradoxically, PSA testing became widespread before any data supported the benefit of aggressively treating early-stage cancer. In 2002, the Scandinavian Prostate Cancer Group Study Number 4, which randomly assigned 695 men younger than 75 years of age who had early-stage prostate cancer to radical prostatectomy or watchful waiting, showed a relative hazard reduction for death from prostate cancer of 50% among those assigned to prostatectomy (4.6% vs. 8.9%), during a median follow-up of 6.2 years.²⁴ The mortality benefit persisted through 15 years of follow-up.²⁹ However, no survival benefit was seen for men who were older than 65 years of age at the time of diagnosis and treatment. Since only about 5% of the tumors were detected by screening, and more than 75% were palpable, it is questionable whether these results are applicable to patients in the United States.

Text

The Prostate Cancer Intervention versus Observation Trial (NCT00007644) randomly assigned 731 men with early-stage prostate cancer to either radical prostatectomy or watchful waiting.³⁰ Three fourths of tumors were diagnosed primarily on the basis of abnormal PSA values, and about half were palpable. Preliminary results showed no significant differences in overall or prostate-cancer mortality after 12 years of follow-up, particularly among men with low-risk cancers.³¹ In other randomized trials, the combination of external-beam radiotherapy and androgen-deprivation therapy was associated with increased overall and disease-specific survival, as compared with radiotherapy alone in men with intermediate- or high-risk early-stage prostate cancers^{32,33} and as compared with androgen-deprivation therapy alone in men with locally advanced cancers.³⁴ Data are lacking from randomized trials comparing radiotherapy with either surgery or watchful waiting for early-stage prostate cancer.²⁸

Text

Abnormal PSA tests lead to biopsies, which can infrequently cause bleeding, pain, or infection.³¹ Undergoing biopsy can be stressful, and some men have persistent anxiety regarding possible cancer, despite negative biopsy results.²⁷ Mathematical models estimate that 23 to 42% of PSA-detected cancers are overdiagnosed, because on the basis of life expectancy at the time of diagnosis and the natural history of the cancer in the absence of screening, it would not be expected to cause clinical problems during the patient's lifetime.²³ Ag-

Text

Given the complexity of issues regarding prostate-cancer screening, experts recommend that men receive support in making informed decisions.³⁵⁻³⁸ However, PSA testing is often performed without discussion of the benefits and harms of screening.^{37,38} Competing clinical demands and the challenge of providing sufficient information to sup-

N ENGL J MED 365:21 NIJM.ORG NOVEMBER 24, 2011
The New England Journal of Medicine

- 文字
- 標題
- 圖表
- 圖表標題
- 表格
- 表格標題
- 頁首
- 頁尾
- 參考文獻
- 公式

切片破壞了長文件脈絡

←  陳勇汀 - 博士論文 - 7. 研究結論 - 2021.html

83 Paragraphs

022

Enabled 

在本研究的閱讀活動中，重點主旨策略為學習者建構閱讀文本主要概念的閱讀理解策略。而彙整了學習者在閱讀文本上標示為關鍵詞的文章關鍵詞，更是具體展現了學習者對於閱讀文本整體架構的印象。學習者可以在協助閱讀階段時分享彼此的文章關鍵詞，不僅讓學習者在透過比較自己與同儕所關注的關鍵字差異中加深閱讀文本的印象，也鼓勵學習者小組在文本關鍵詞上彼此討論與交流的契機。這些可能都是讓具有同儕

1. 如果檢索到此片段

023

Enabled  

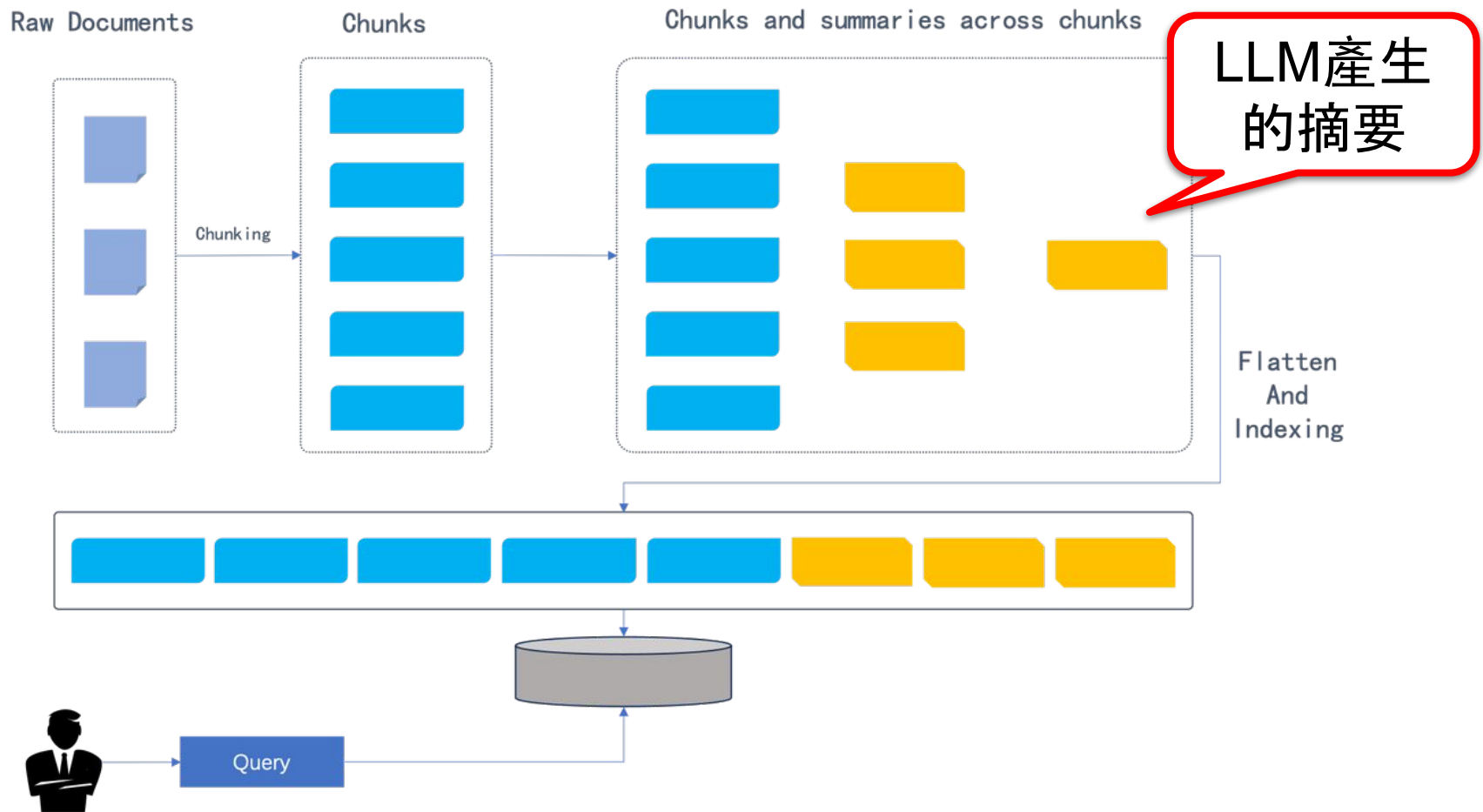
而本研究再次透過實驗證實，當閱讀學習活動包含了重點主旨策略的教學時，同儕協助與閱讀學習的結合將會呈現明顯的合作增強效應。

 60  0  987db3c460...



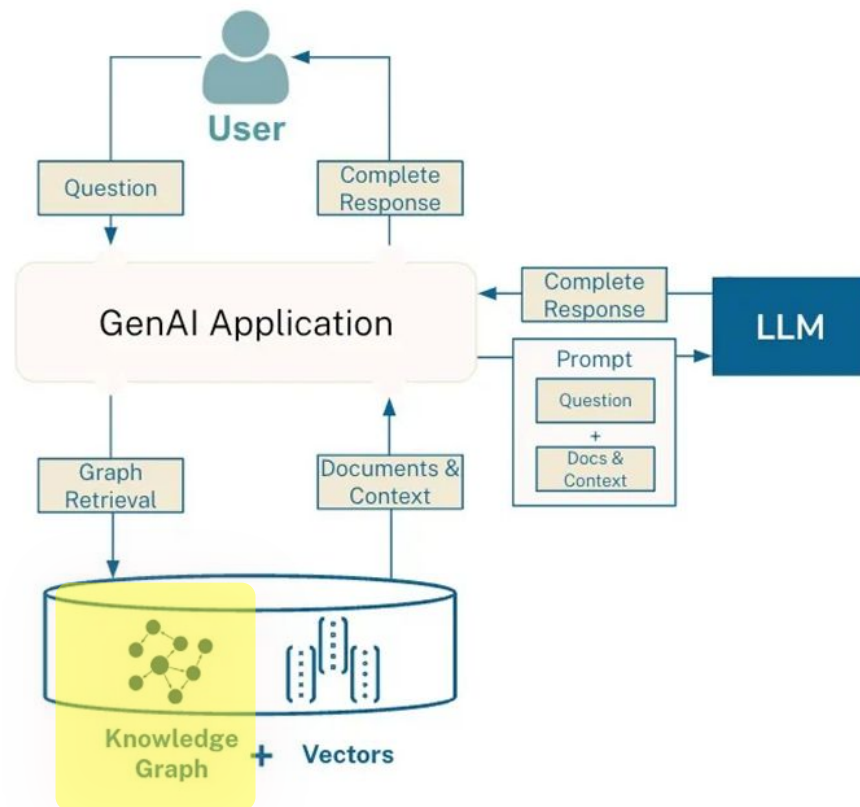
2. 不一定能帶出前後文片段

連結長文件脈絡：RAPTOR



以知識實體與關聯建立索引：GraphRAG

1. 節點分析：知識實體單位
2. 邊分析：建立知識實體之間的關係
3. 檢索：以SNA排序技術演算法找出相關上下文



混合搜尋技術：Danswer

The screenshot displays the Danswer web application interface. At the top, there's a navigation bar with the Danswer logo, a search icon, and a chat icon. Below this, the main interface is divided into three sections. On the left, a sidebar contains a 'New Chat' button and a list of today's chats, including one titled 'Resolving 404 error'. The central area shows a chat conversation. A user asks 'What is Danswer?'. The system, 'Danswer', responds with a detailed definition of the tool, mentioning its open-source nature, ability to search private data, and features like user authentication and document connectors. A 'Hide Docs' button is visible next to the response. On the right, a 'Retrieved Documents' section lists relevant documents, such as 'Introduction - Danswer Documente...' and 'Danswer Chat'. Below this, a 'Selected Documents' section allows users to choose documents to chat with. At the bottom, there's a search bar with the placeholder text 'Ask me anything...' and a 'Filter' button. A red speech bubble points to the '文件搜尋' (File Search) section, and another red speech bubble points to the chat interaction.

文件搜尋

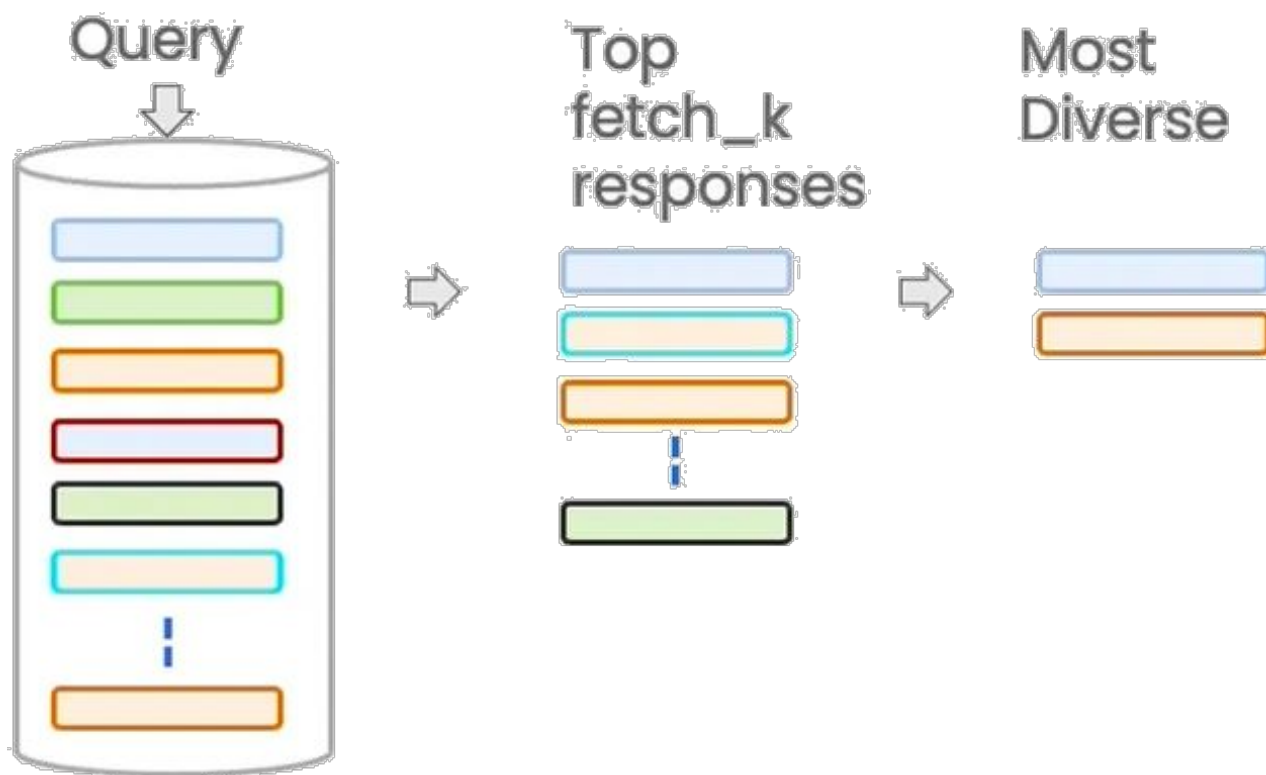
聊天機器人問答

指出資料來源

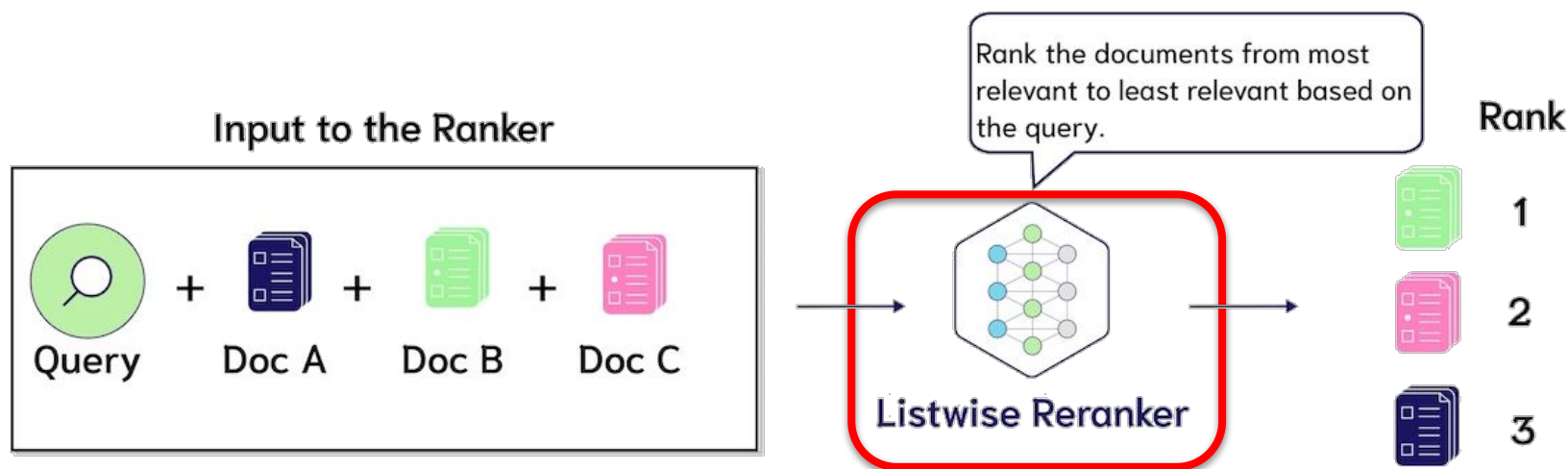
檢索多樣化脈絡

Maximal Marginal Relevance (MMR)

避免相似冗餘結果、增加探索多元觀點

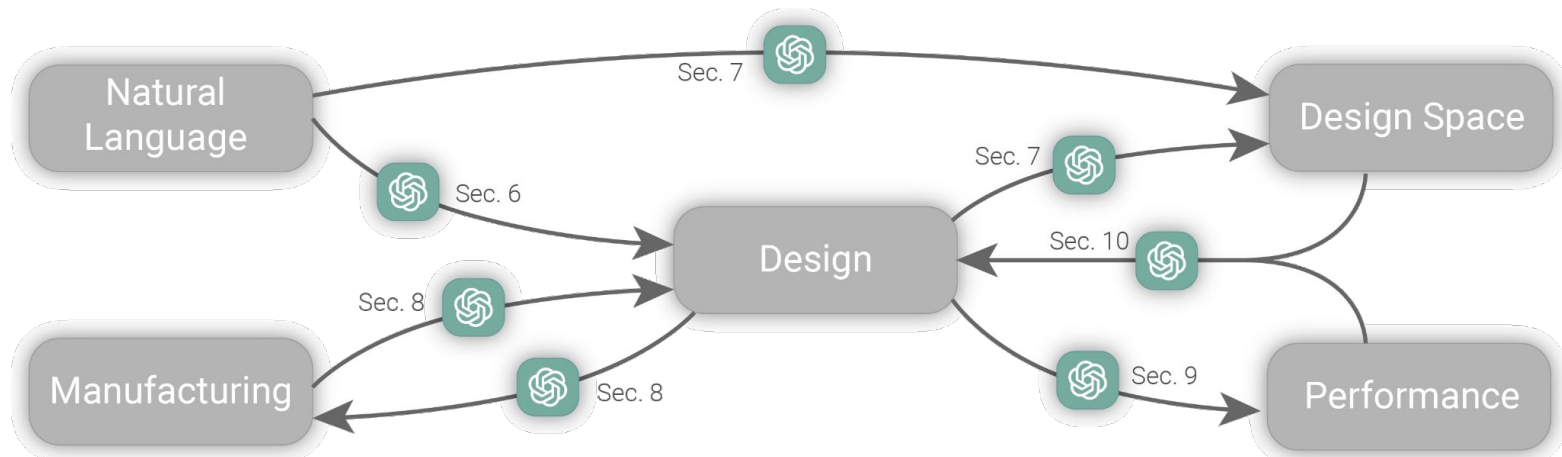


整合多種知識庫查詢結果：ReRanker



用模型來取代Cosine Similarity計算
更能控制理想的查詢結果

生成結果的指定：電腦設計與生產



Let's use OpenJSCAD to design a cabinet ... with 3/4" wood sheets, and final exterior dimensions of 30 inches tall, 20 inches wide, and 18 inches deep. The stationary part of the cabinet should be comprised of 6 boards: bottom, top, back, two sides, and one shelf centered inside the cabinet. (... omitted by authors: OpenJSCAD modeling hints; see Appendix A.1 ...)



```
(... omitted by authors: import statements and variable setup ...)  
pieces.push(colorize ([1, 0, 0], translate ([0, 0, -height/2 + thickness/2],  
cuboid({size: [width, depth, thickness]})))) // bottom  
pieces.push(colorize ([0, 1, 0], translate ([0, 0, height/2 - thickness/2],  
cuboid({size: [width, depth, thickness]})))) // top  
pieces.push(colorize ([0, 0, 1], translate ([0, -depth/2 + thickness/2, 0],  
cuboid({size: [width, thickness, height]})))) // back  
pieces.push(colorize ([1, 1, 0], translate ([-width/2 + thickness/2, 0, 0],  
cuboid({size: [thickness, depth, height]})))) // left  
pieces.push(colorize ([0, 1, 1], translate ([width/2 - thickness/2, 0, 0],  
cuboid({size: [thickness, depth, height]})))) // right  
pieces.push(colorize ([1, 0, 1], translate ([0, -thickness/2, 0],  
cuboid({size: [innerWidth, thickness, innerDepth]})))) // shelf  
(... omitted by authors: rest of design ...)
```



作者提供
OpenJSCAD
提示



結語：實驗與實務的距離

大型語言模型應用框架

- **LLM導向函式庫**：LangChain、LlamaIndex、Haystack
- **低程式碼整合框架**：Flowise、Botpress、Danswer、Dify
- **RAG強化框架**：Verba、FlashRAG



RAG是知識應用終極解答

真的嗎？

終極答案
應該是42!



RAG落地了嗎？

- Garbage in, Garbage out：索引才是關鍵
- LLM的不確定性：創意 v.s. 穩定
- 查詢成本：LLM+Embedding=\$\$\$

RAG不易掌控



但仍有潛力成為研究者的翅膀！

感謝您的聆聽 歡迎提問

為什麼是42...

✉ pulipuli.chen@gmail.com

🍪 [布丁布丁吃什麼？](#)

