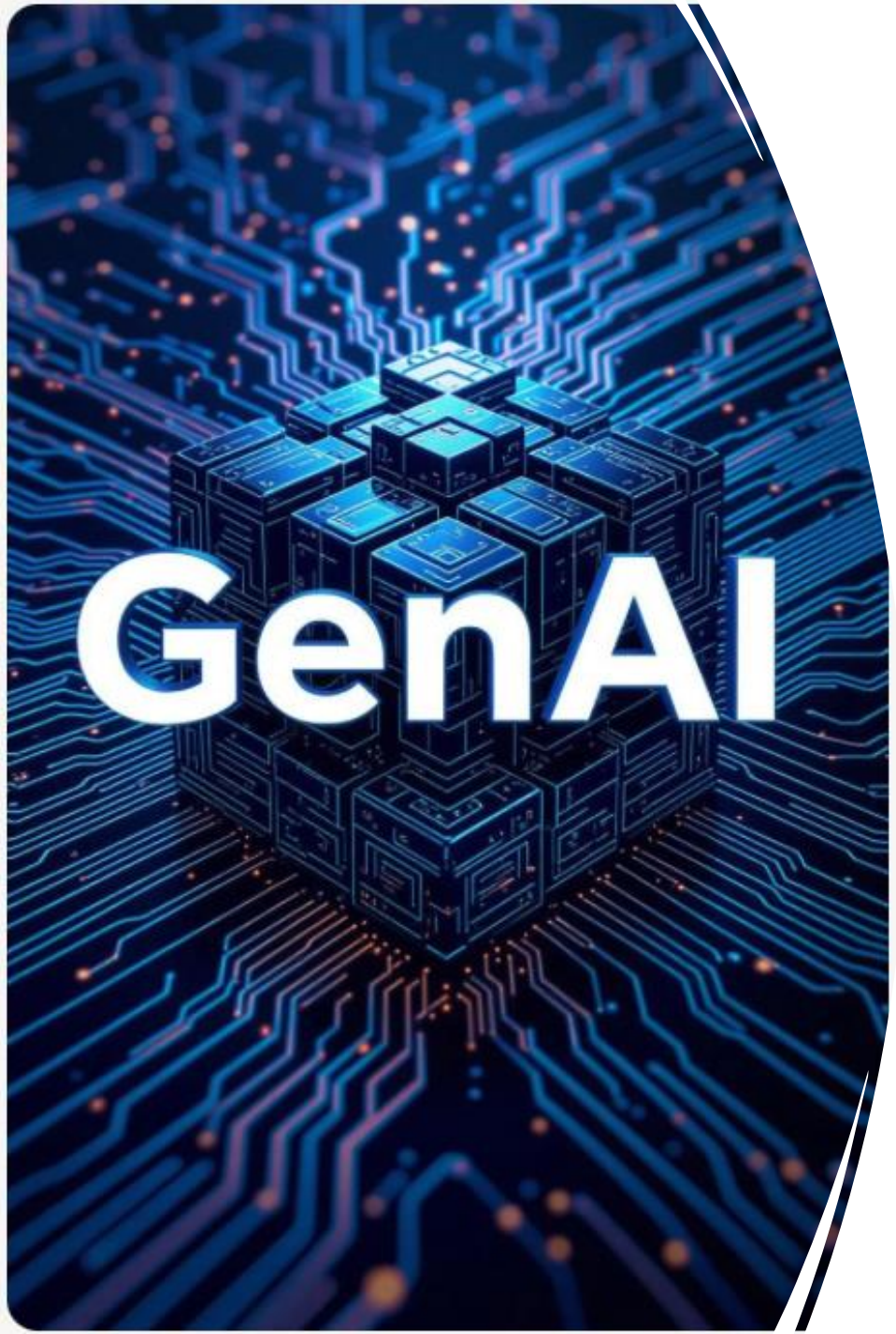


生成式人工智慧概論

第三堂、檢索增強生成

RAG (Retrieval Augmented Generation)

許昌仁

專題題目分享

- 請每一組派一名同學，到台前來，
  跟大家分享你們這組的專題題目想作什麼。

- 每組五分鐘。

n8n的操作練習

- 不同的trigger
  - When chat message received, Trigger manually, On a schedule
  - on webhook call, When executed by another workflow
- Set node
- AI agent node
- http request node

回顧：system prompt & user prompt

user Prompt

system Prompt

你好

你扮演一個嚴肅的大老闆。

網站/
APP

System：
你扮演一個嚴肅的大老闆。
user：
你好。

LLM模型

嗯。直接說重點。
我時間很寶貴。

嗯。直接說重點。
我時間很寶貴。

讓LLM幫我做各種事情

Case 3.
文章內容大綱， … …

根據大綱，幫我寫一篇作文

Prompt結合外部資料

user Prompt

10/15日的會議結論是什麼？

1015會議紀錄

網站/
APP

System：
10/15XXXX會議會議記錄
1. 。。。。
2. ----------
3. …….

User:
10/15日的會議結論是什麼？

LLM模型

找到對的外部資料（檢索），來幫忙（增強）LLM產出更好的結果（生成）

檢索增強生成 (Retrieval Augmented Generation, RAG)

RAG

在LLM回答問題時

- 先從外部資料庫中檢索相關的內容
- 再基於這些內容來讓LLM生成答案

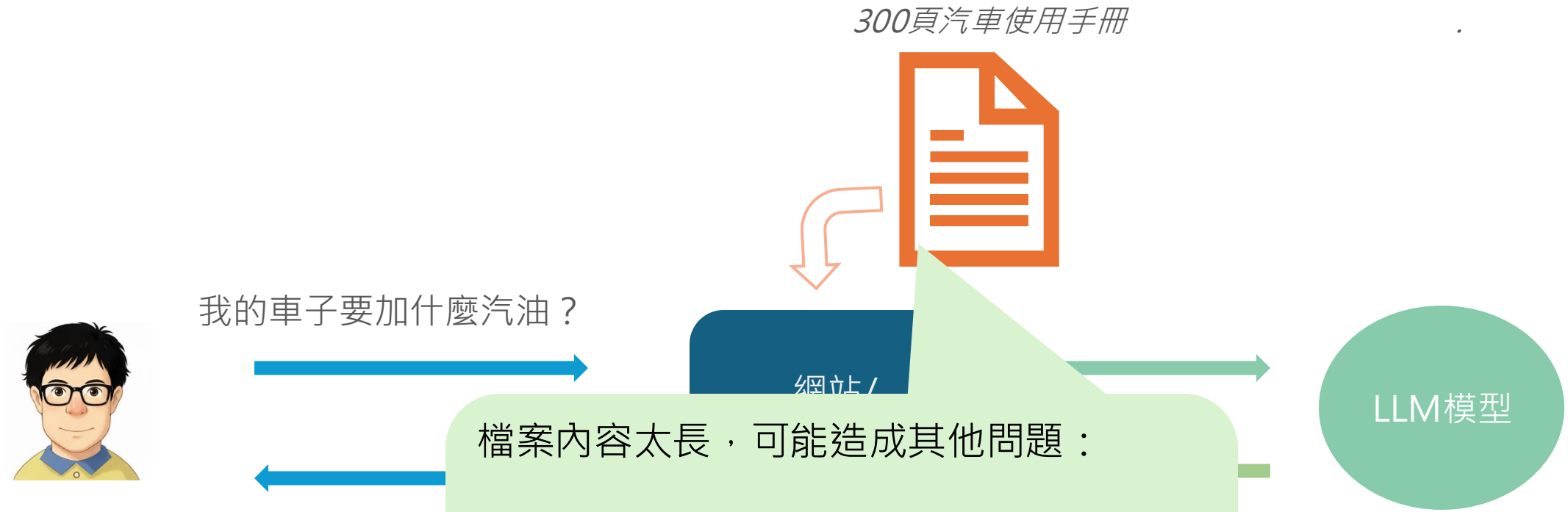
檢索　　生成

RAG可以解決什麼問題？

RAG 是 GenAI 在企業應用中的常見關鍵技術。理論上，RAG可以：

1. 解決「幻覺」(Hallucinations) 問題
   - 問題： LLM有時會憑空捏造、杜撰出聽起來合理但事實上錯誤或虛假的答案，這被稱為「幻覺」。
   - RAG 解決方式： RAG 在生成答案之前，會先從一個外部、可信賴的知識庫中檢索相關的資訊，然後使用這些資訊作為依據 (Grounding) 來生成回覆。這大大減少了模型憑空臆測的可能性，讓回覆更真實、更可靠。
2. 克服知識的「時效性」和「靜態性」
   - 問題： LLM的知識僅限於其訓練數據，因此無法回答關於最新事件或訓練結束後才出現的新資訊的問題。
   - RAG 解決方式： 外部知識庫可以即時更新，讓模型能夠存取最即時的資訊，從而生成與時俱進的回覆。
3. 融入特定領域或企業的「專業知識」
   - 問題： 預訓練的 LLMs 缺乏特定行業、企業內部文件或專業領域的深度和精確度。
   - RAG 解決方式： RAG 能夠連接到企業的內部資料庫或專屬文件，提供LLM訓練時無法取得的專業知識。
4. 提供「可追溯性」和「透明度」
   - 問題： 傳統 LLM 生成的內容，用戶難以驗證其來源，缺乏信任。
   - RAG 解決方式： 由於 RAG 會檢索來源文件，因此可以將引用的來源文件或段落一同提供給用戶，讓用戶能夠驗證資訊的正確性，增加輸出的可信度和透明度。

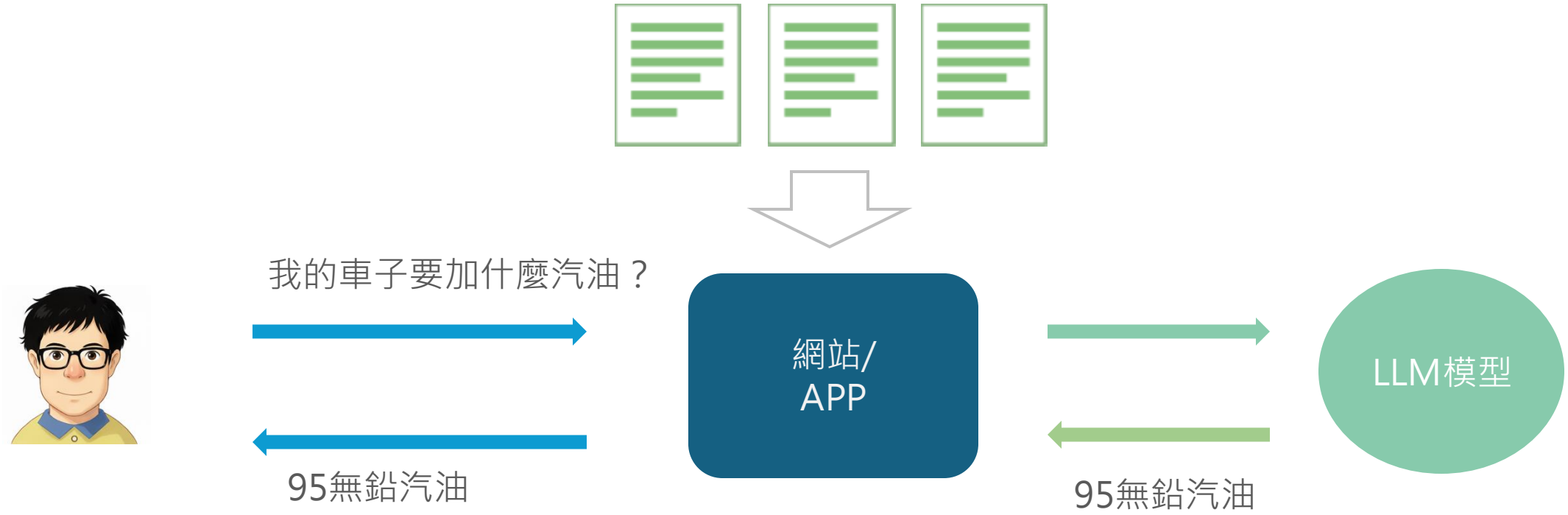
但是。。。


300頁汽車使用手冊
我的車子要加什麼汽油？
LLM模型
檔案內容太長，可能造成其他問題：
1. LLM能接受的上下文長度有限制
2. 過多的輸入，可能會稀釋了標準答案的顯著性
3. 費$$費時

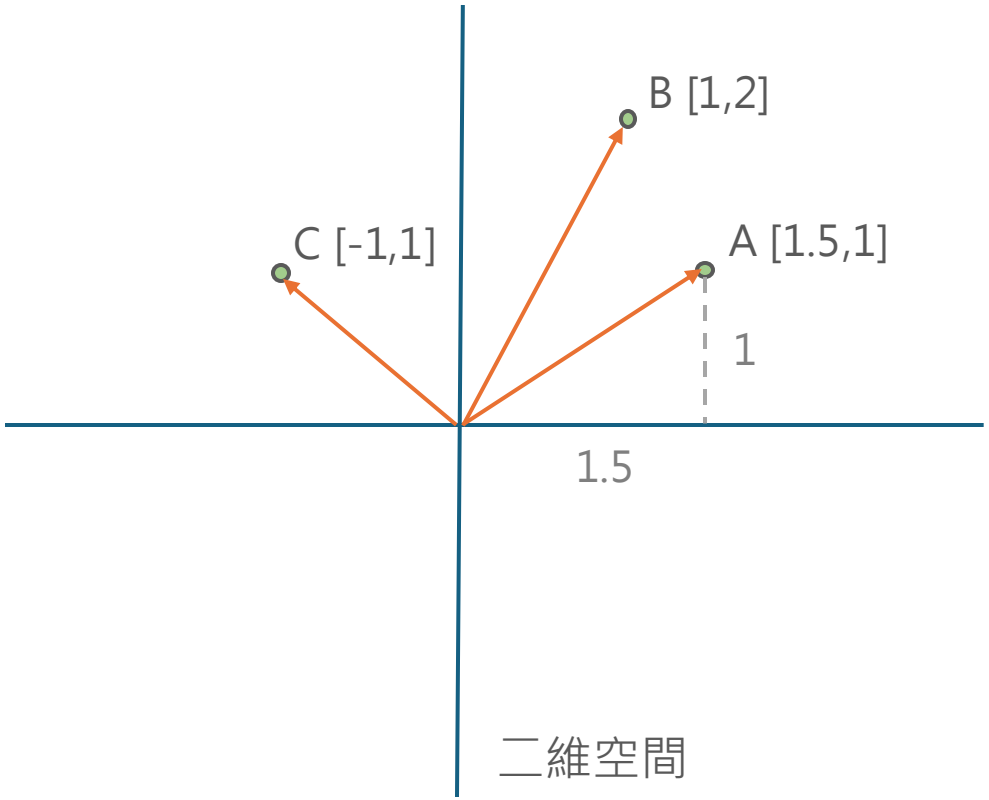
文件切片 (chunk)

300頁汽車使用手冊

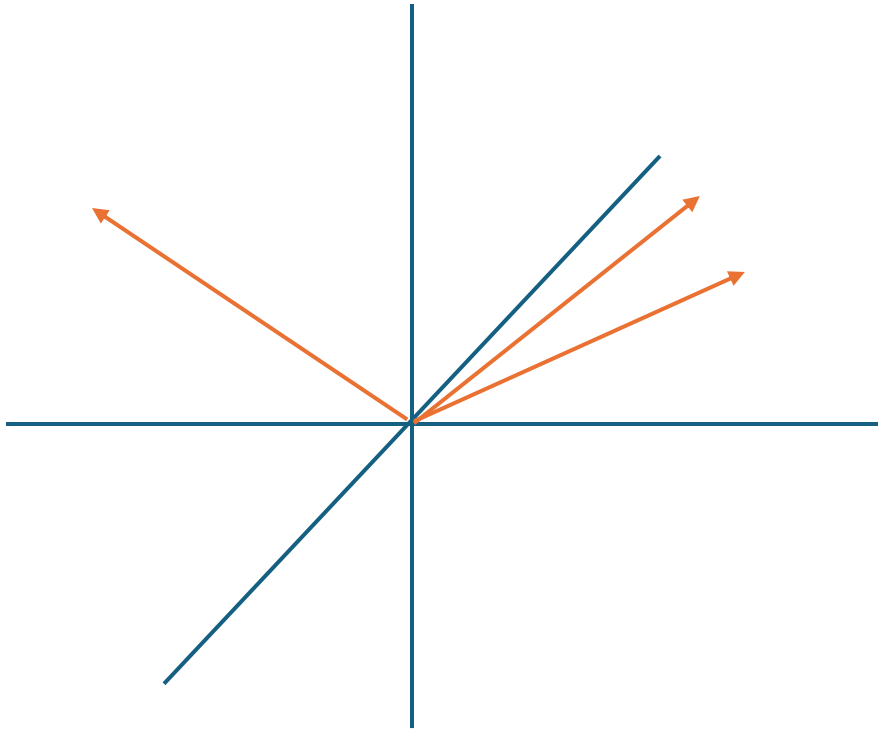
文件切片 (chunk)



喚醒記憶：座標空間與向量



二維空間



三維空間

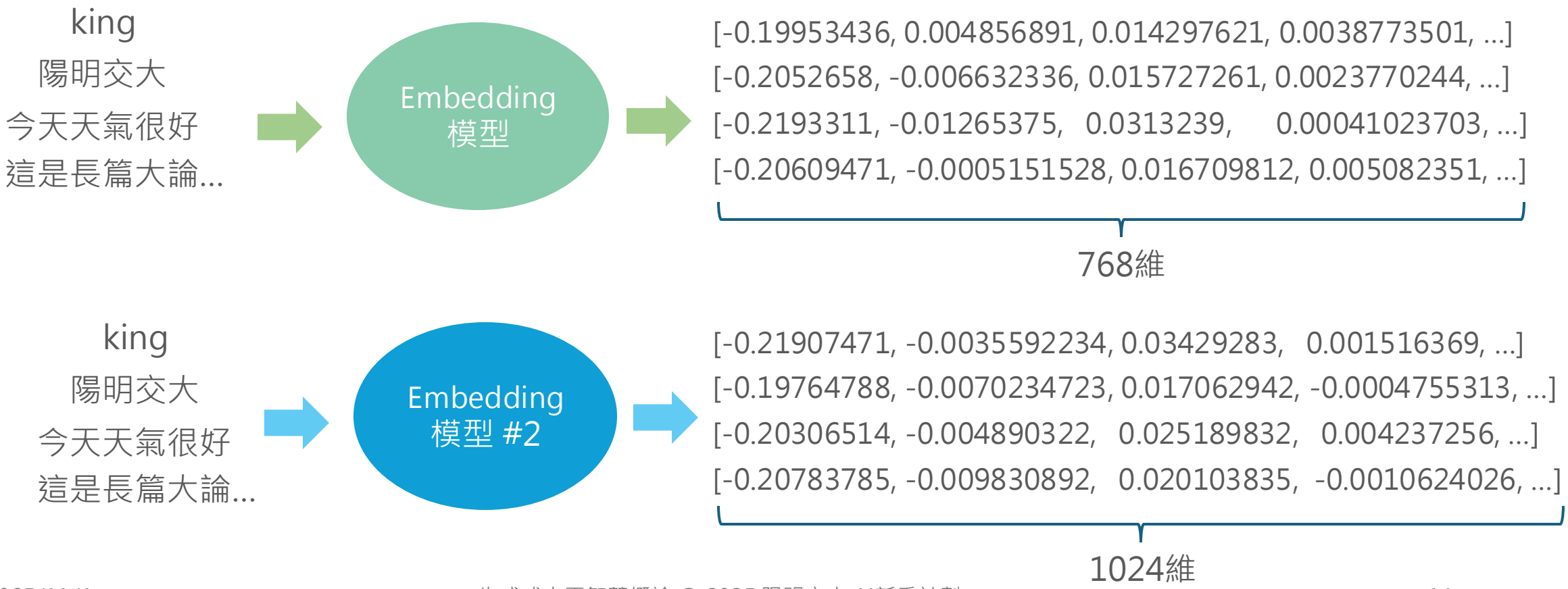
。。。N維空間

Tokens, Token ID, embedding

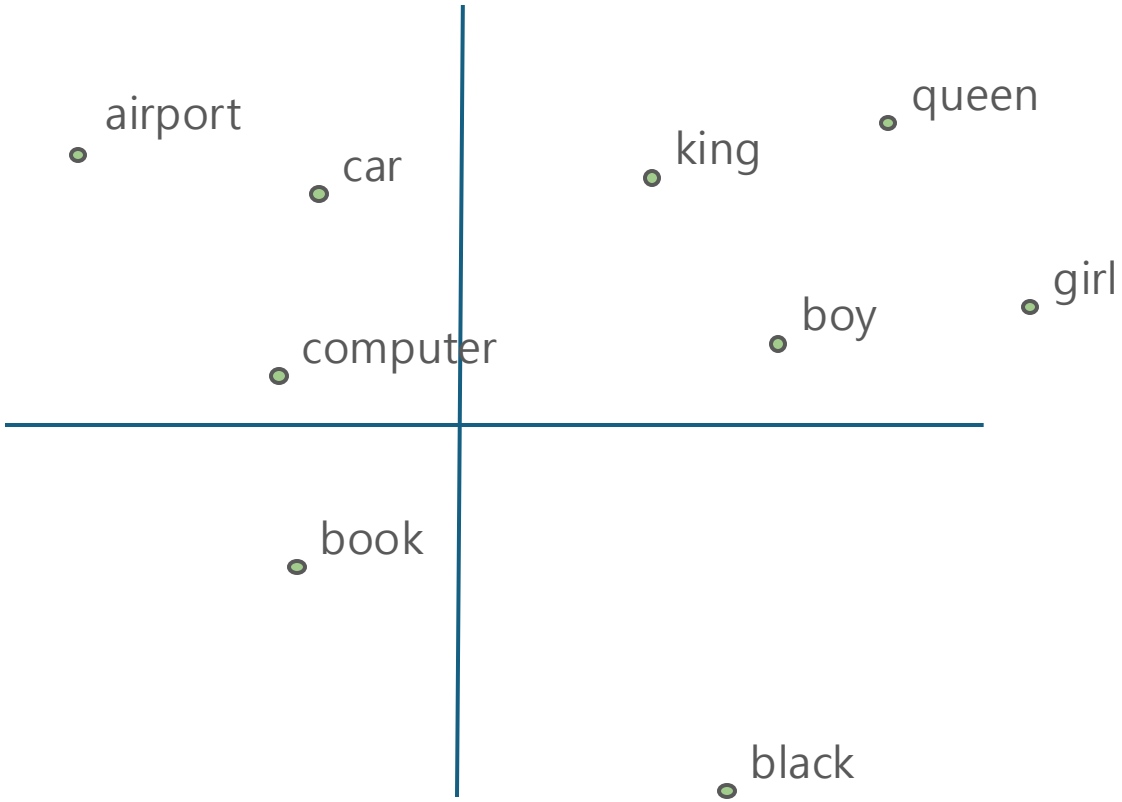
Tokens	Token IDs	Embeddings(嵌入、轉換為向量)
boy	[23999]	[-0.19953436, 0.004856891, 0.014297621, 0.0038773501, ...]
girl	[67593]	[-0.2052658, -0.006632336, 0.015727261, 0.0023770244, ...]
king	[6962]	[-0.2193311, -0.01265375, 0.0313239, 0.00041023703, ...]
queen	[153556]	[-0.20609471, -0.0005151528, 0.016709812, 0.005082351, ...]
car	[6830]	[-0.20932105, -0.010319027, 0.015522129, 0.0131326625, ...]
airport	[198613]	[-0.21907471, -0.0035592234, 0.03429283, 0.001516369, ...]
book	[3092]	[-0.19764788, -0.0070234723, 0.017062942, -0.0004755313, ...]
Computer	[76982]	[-0.20306514, -0.004890322, 0.025189832, 0.004237256, ...]
black	[18474]	[-0.20783785, -0.009830892, 0.020103835, -0.0010624026, ...]

。。。N維空間.

embedding


king
陽明交大
今天天氣很好
這是長篇大論...
Embedding
模型
[-0.19953436, 0.004856891, 0.014297621, 0.0038773501, ...]
[-0.2052658, -0.006632336, 0.015727261, 0.0023770244, ...]
[-0.2193311, -0.01265375, 0.0313239, 0.00041023703, ...]
[-0.20609471, -0.0005151528, 0.016709812, 0.005082351, ...]
768維
king
陽明交大
今天天氣很好
這是長篇大論...
Embedding
模型 #2
[-0.21907471, -0.0035592234, 0.03429283, 0.001516369, ...]
[-0.19764788, -0.0070234723, 0.017062942, -0.0004755313, ...]
[-0.20306514, -0.004890322, 0.025189832, 0.004237256, ...]
[-0.20783785, -0.009830892, 0.020103835, -0.0010624026, ...]
1024維

向量化後可以進行數學計算

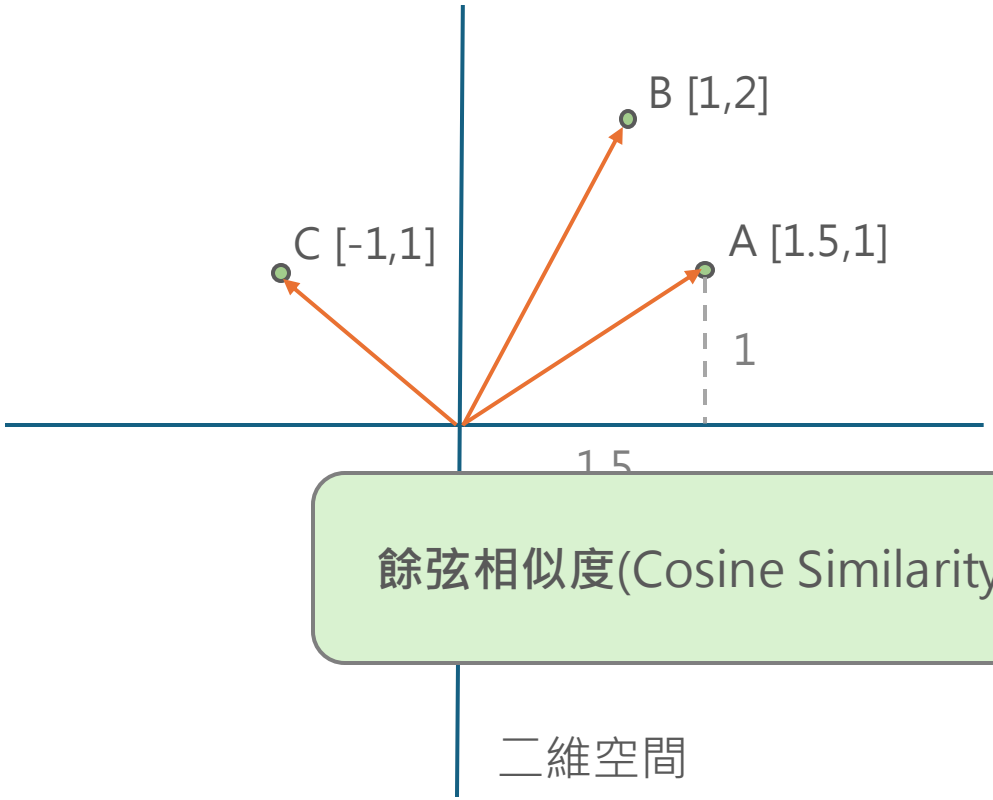


king – boy + girl ≈ queen

[-0.2193311, -0.01265375, 0.0313239, 0.00041023703, ...] -
[-0.19953436, 0.004856891, 0.014297621, 0.0038773501, ...] +
[-0.2052658, -0.006632336, 0.015727261, 0.0023770244, ...]
=
[-0.20609471, -0.0005151528, 0.016709812, 0.005082351, ...]

巴黎 – 法國 + 英國 ≈ 倫敦

向量相似度計算



兩個向量有沒有靠近？有沒有相似？
餘弦相似度(Cosine Similarity)

A=[1.5, 1] , B=[1, 2]

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|}$$

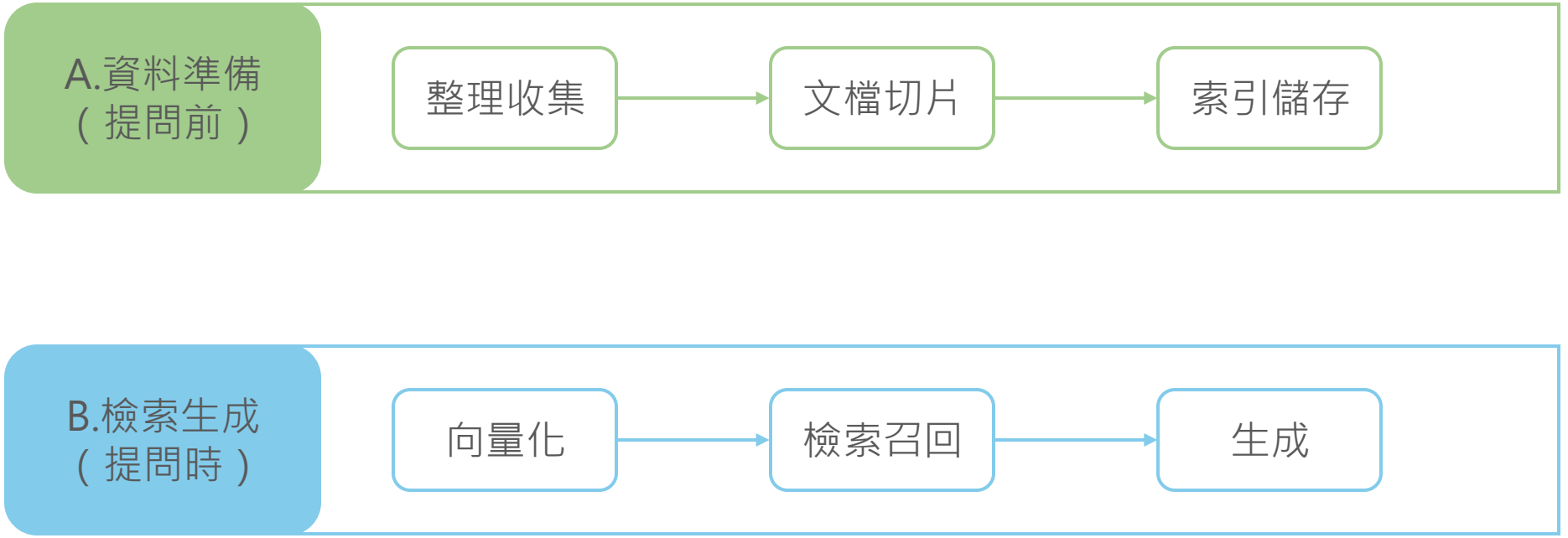
A·B=(1.5)(1)+(1)(2)=1.5+2=3.5

餘弦相似度(Cosine Similarity) 越大的兩個相量相似度越高，也代表越靠近。

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|} = \frac{3.5}{\sqrt{3.25} \cdot \sqrt{5}} \approx 0.8683$$

$$\text{Cosine Similarity} = \frac{A \cdot C}{\|A\|\|C\|} \approx -0.1961$$

RAG的基本流程



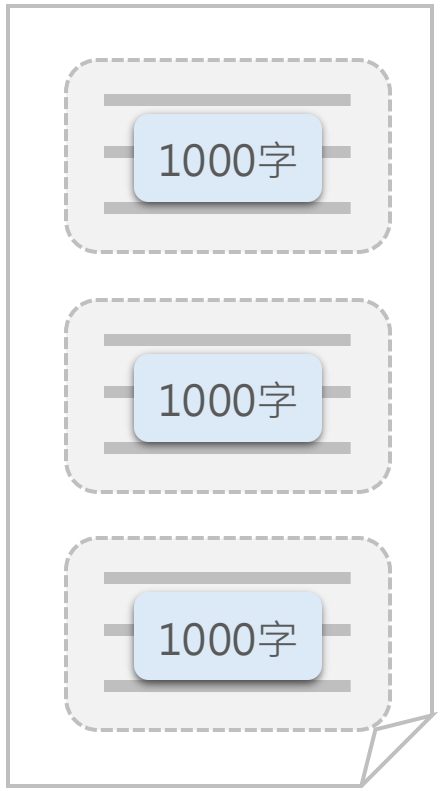
RAG的基本流程-A1.整理收集

- 資料(檔案)在哪裡？
- 是什麼樣的格式？(word, pdf, excel, 資料庫?)
- 使用什麼工具擷取出文字內容
- 數據更新和同步機制
- 資料可以公開嗎？

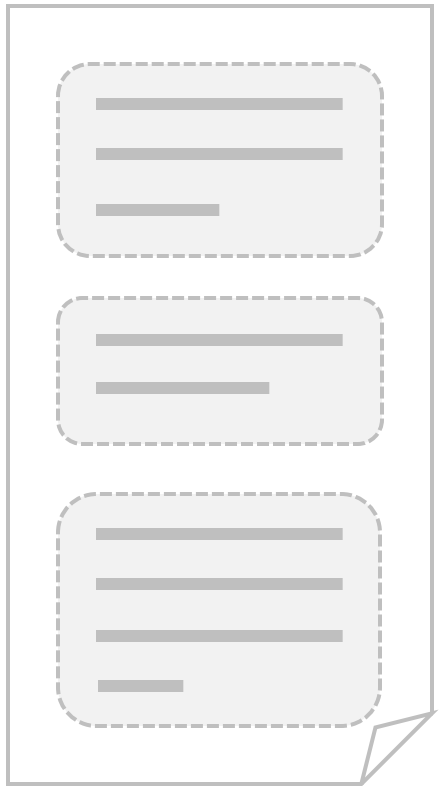
RAG的基本流程-A2.文章切片

300頁
汽車使用手冊

切片的方式

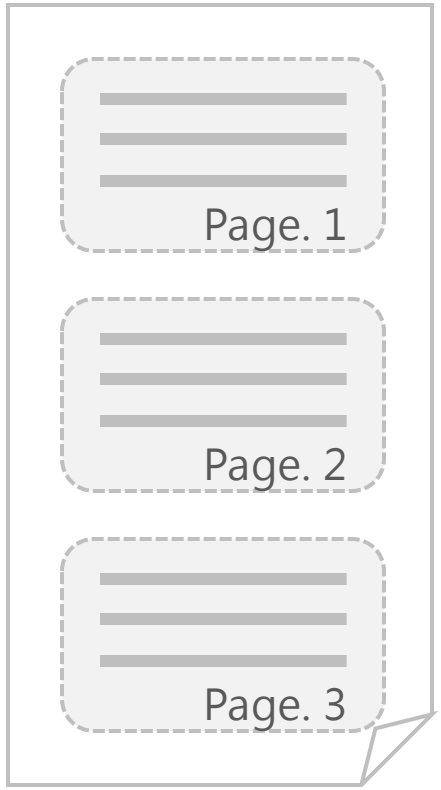


按固定字數切



按段落切

按章節切



按頁數切

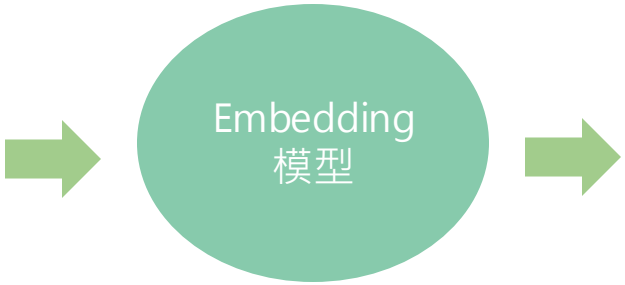
RAG的基本流程-A3.索引儲存

使用觸控螢幕可控制許多功能，而這些功能在傳統車輛中都會使用實體按鈕來控制...

僅會在 Model 3 偵測到當地的空氣指數（AQI）不佳時顯示於觸控螢幕的狀態列。...

您的車輛會自動更新時間。如果時間不正確，請確認您的車輛已...

警告： 請僅在車輛停止並處於停車檔 (P) 時重新啟動觸控螢幕。重新啟動期間...



[-0.1995, 0.0048, 0.014, 0.00387, ...]

[-0.205, -0.0066, 0.0157, 0.00237, ...]

[-0.219, -0.0126, 0.0313, 0.00041, ...]

[-0.2060, -0.0005, 0.0167, 0.00508, ...]

本文	向量
使用觸控螢幕可控制許多功能，而這些...	[-0.1995, 0.0048, 0.014, 0.00387, ...]
僅會在 Model 3 偵測到當地的空氣...	[-0.205, -0.0066, 0.0157, 0.00237, ...]
您的車輛會自動更新時間。如果時間不正確...	[-0.219, -0.0126, 0.0313, 0.00041, ...]
警告： 請僅在車輛停止並處於停車檔 (P) 時...	[-0.2060, -0.0005, 0.0167, 0.00508, ...]

向量資料庫
Vector DB
(supabase...)

RAG的基本流程-B1.向量化

我的車子要加什麼汽油？

Embedding
模型

[-0.1976, -0.00702, 0.01706, -0.000475, ...]

RAG的基本流程-B2.檢索召回

如何調整座椅？

Embedding
模型

[-0.1976, -0.00702, 0.01706, -0.000475, ...]

本文	向量
使用觸控螢幕可控制許多功能，而這些...	[-0.1995, 0.0048, 0.014, 0.00387, ...]
僅會在 Model 3 偵測到當地的空氣...	[-0.205, -0.0066, 0.0157, 0.00237, ...]
您的車輛會自動更新時間，如果時間不正確...	[-0.219, -0.0126, 0.0313, 0.00041, ...]
警告： 請僅在車輛停止並處於停車檔 (P) 時...	[-0.2060, -0.0005, 0.0167, 0.00508, ...]

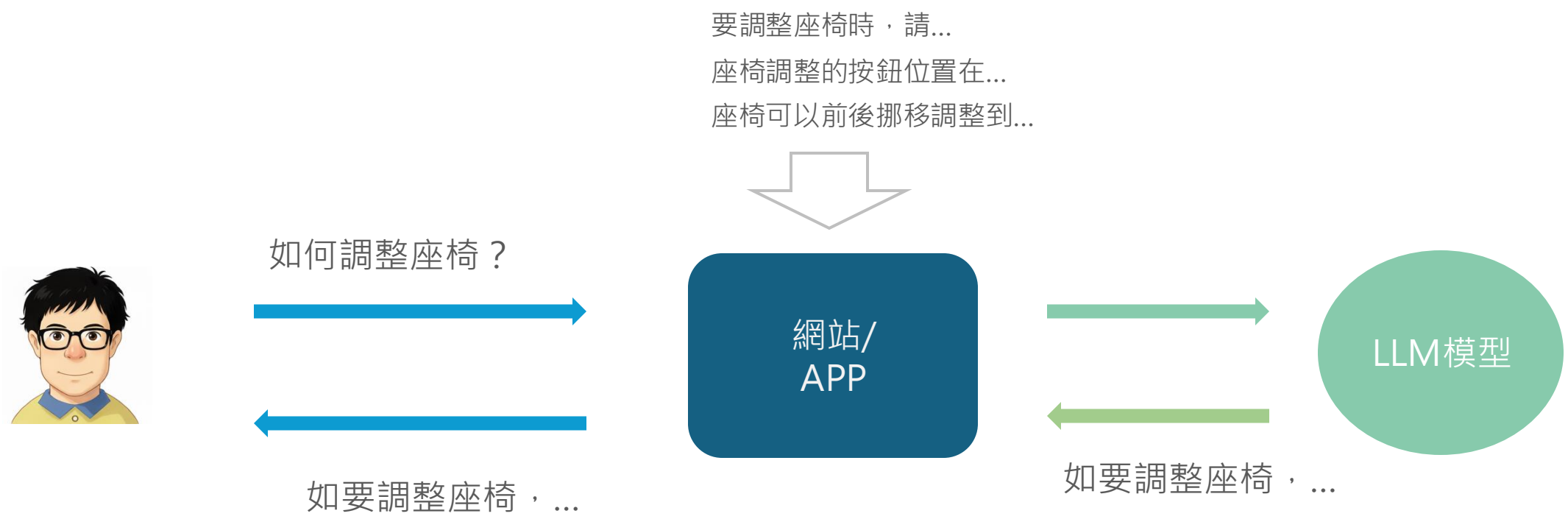
逐一進行
「相似度計算」

找出最相似的
N個向量

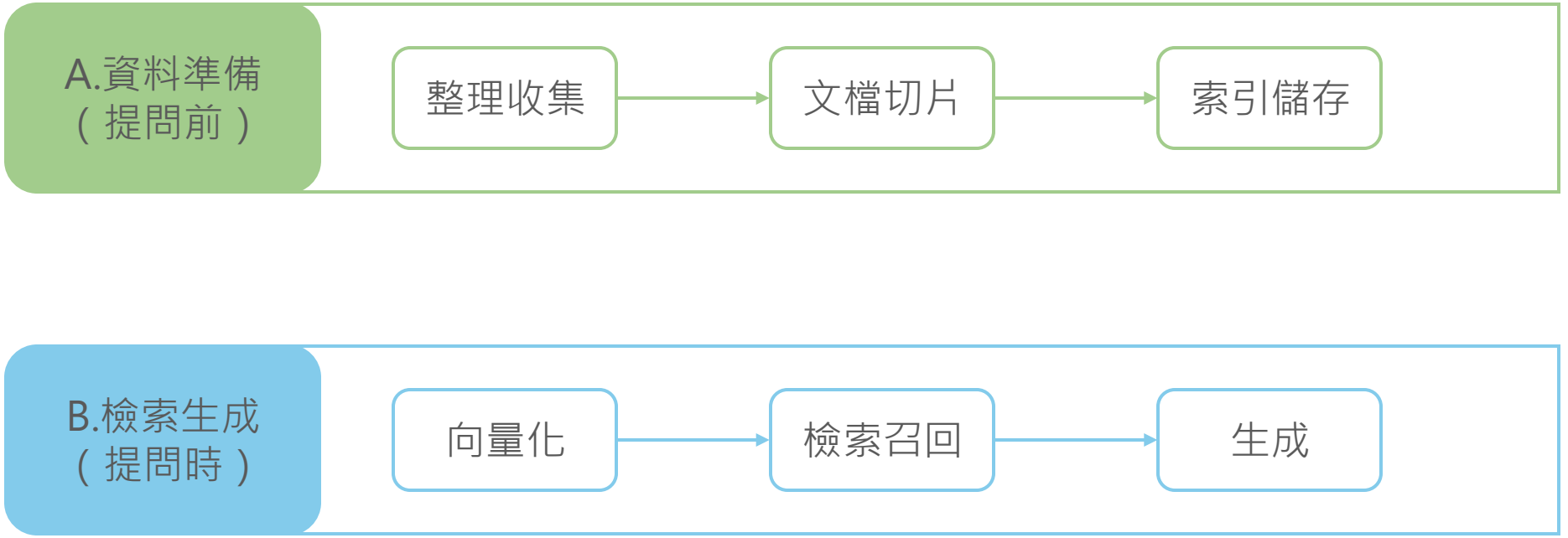
把這些向量的本文
列出來

要調整座椅時，請...
座椅調整的按鈕位置在...
座椅可以前後挪移調整到...

RAG的基本流程-B3.生成



實作4、使用n8n建構RAG流程



應用實例

momo 攜手台灣微軟以生成式AI重塑客服體驗

本文共933字

00:00

2025/10/20 11:27:41

經濟日報 記者何佩儒／即時報導

momo 富邦媒與台灣微軟合作打造大型語言模型（LLM）驅動的新一代智能客服，這套系統結合 Microsoft Azure OpenAI 與檢索增強生成（RAG）技術，在語意理解與即時回應上展現更佳表現，回覆正確率超過 90%，不僅減輕客服團隊的負擔，進一步強化智能對話與個人化服務體驗，帶動消費者滿意度提升；未來也將成為支撐 momo 全通路服務體驗的核心，推動電商市場邁向AI新時代。

參考資料

- 使用n8n建立RAG系統
- 使用python建構RAG系統
- RAG的17種方法
- 申請Line交談機器人帳號（6:45 – 23:00）

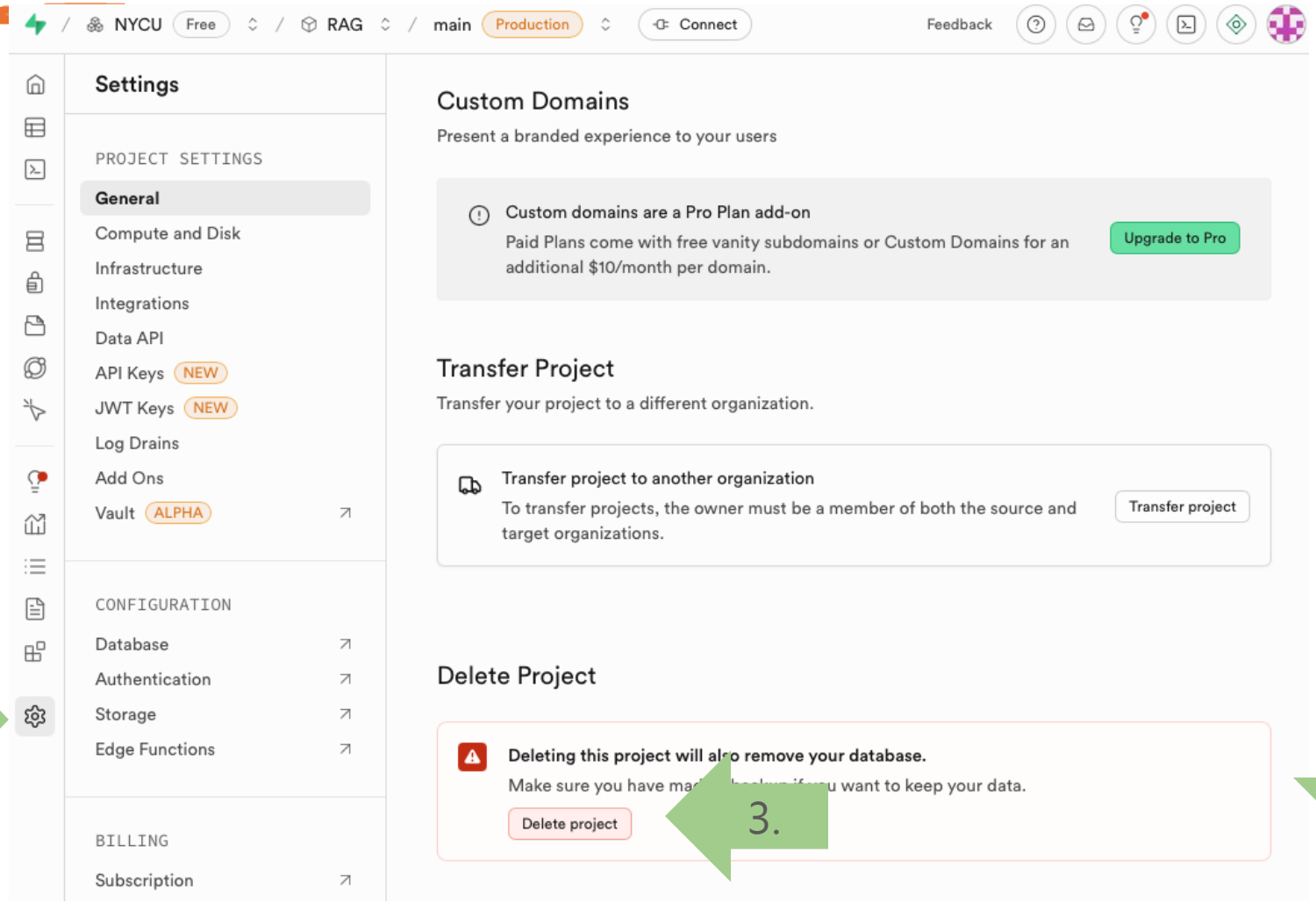

Thank You!

附錄、使用supabase建立RAG資料庫

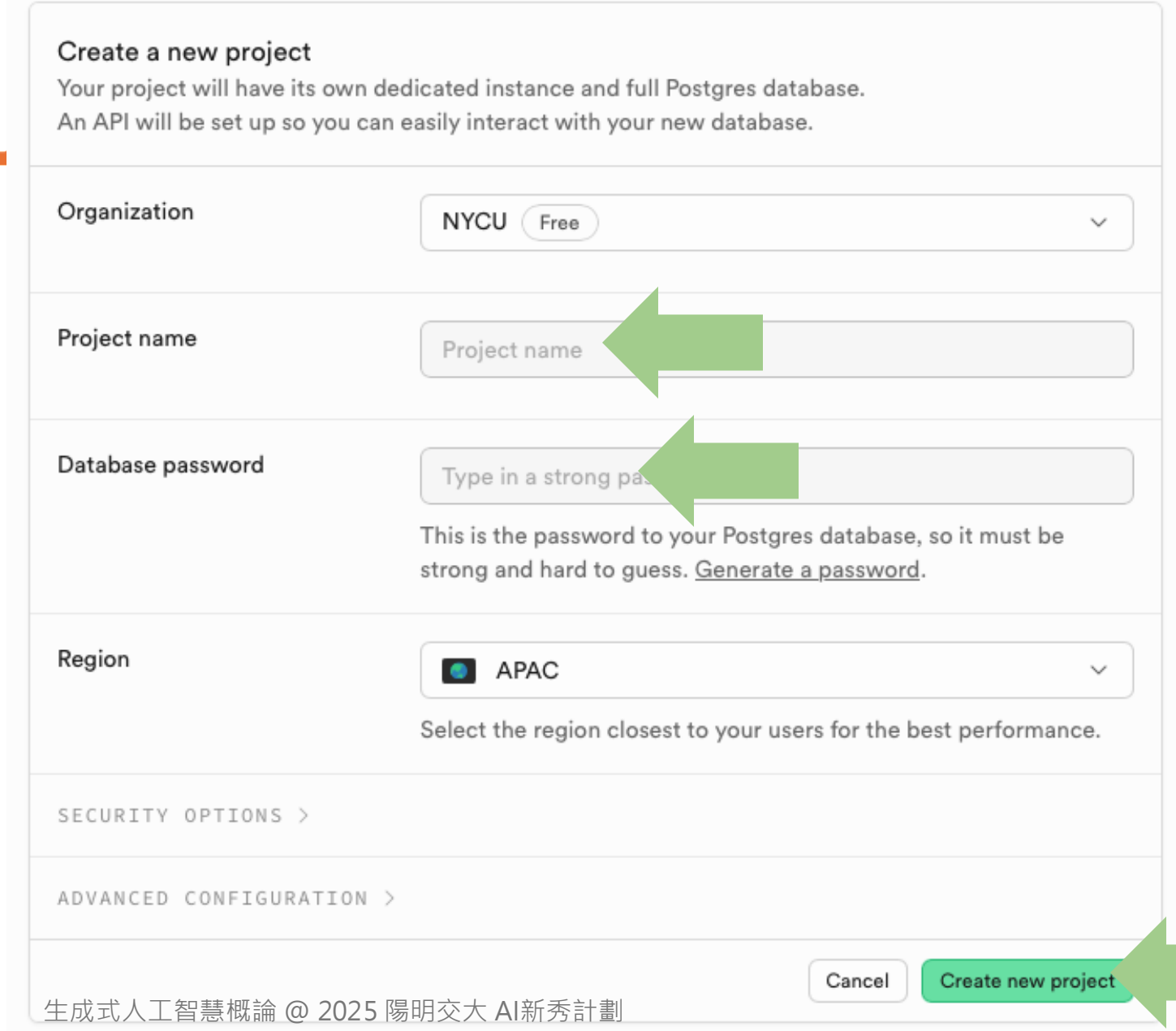
1.確認你的supabase只有一個資料庫

免費的supabase只能建立兩個資料庫，n8n已經使用了一個。
剩下一個名額留給RAG。
-- 刪除由不到的資料庫（參考下一頁）

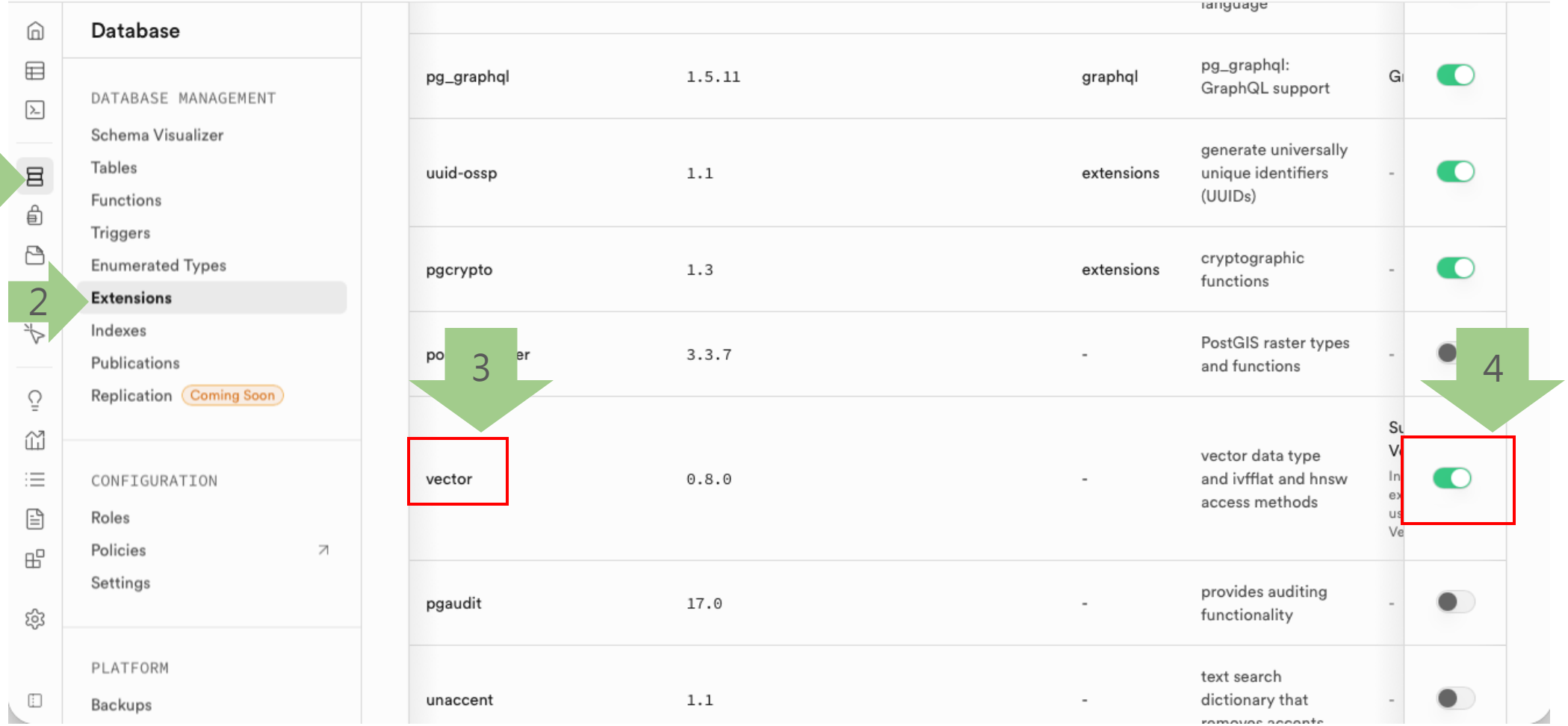
刪除supabase資料庫



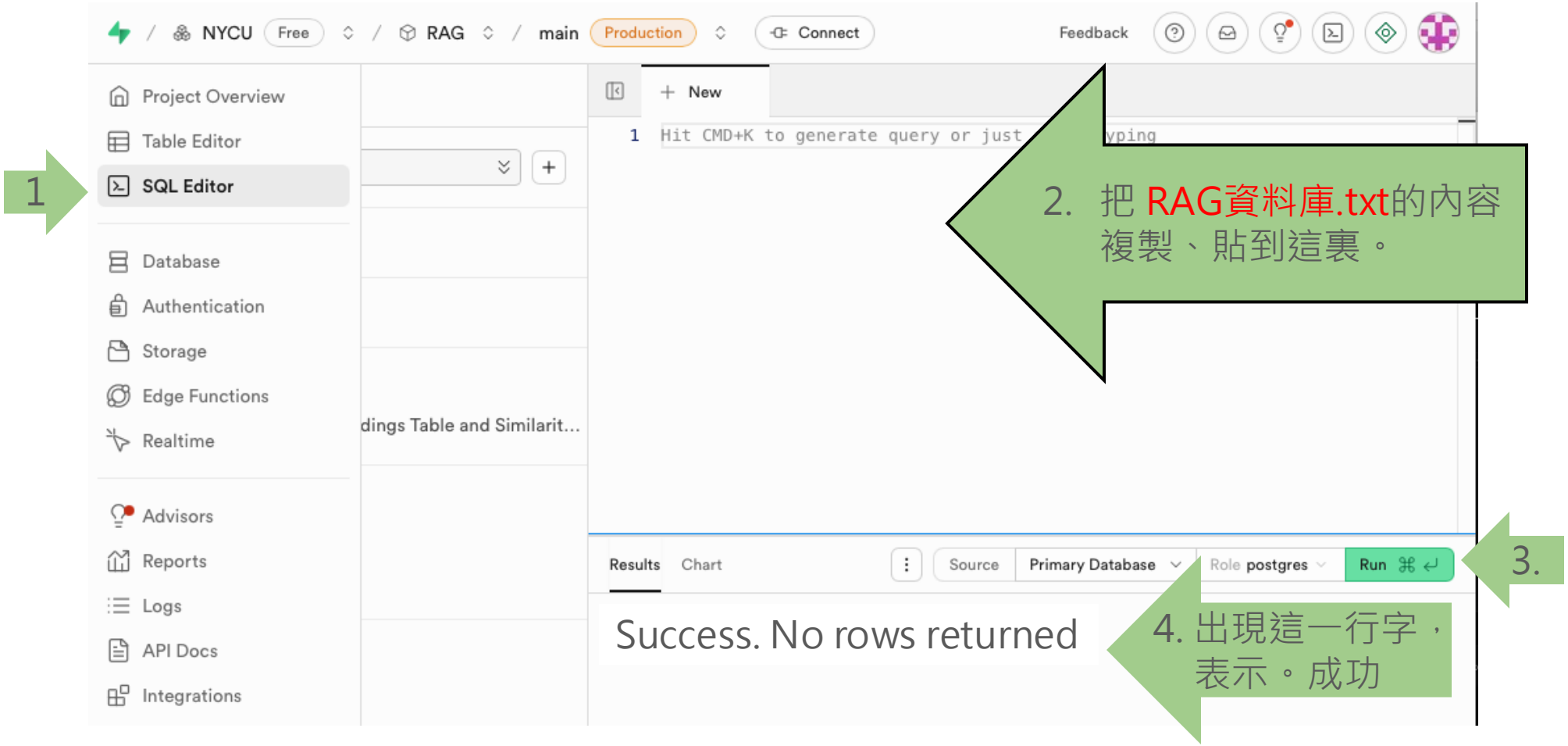
建立新資料庫


Create a new project
Your project will have its own dedicated instance and full Postgres database.
An API will be set up so you can easily interact with your new database.
Organization
NYCU
Free
Project name
Project name
Database password
Type in a strong pa
This is the password to your Postgres database, so it must be strong and hard to guess. Generate a password.
Region
APAC
Select the region closest to your users for the best performance.
SECURITY OPTIONS
ADVANCED CONFIGURATION
Cancel
Create new project

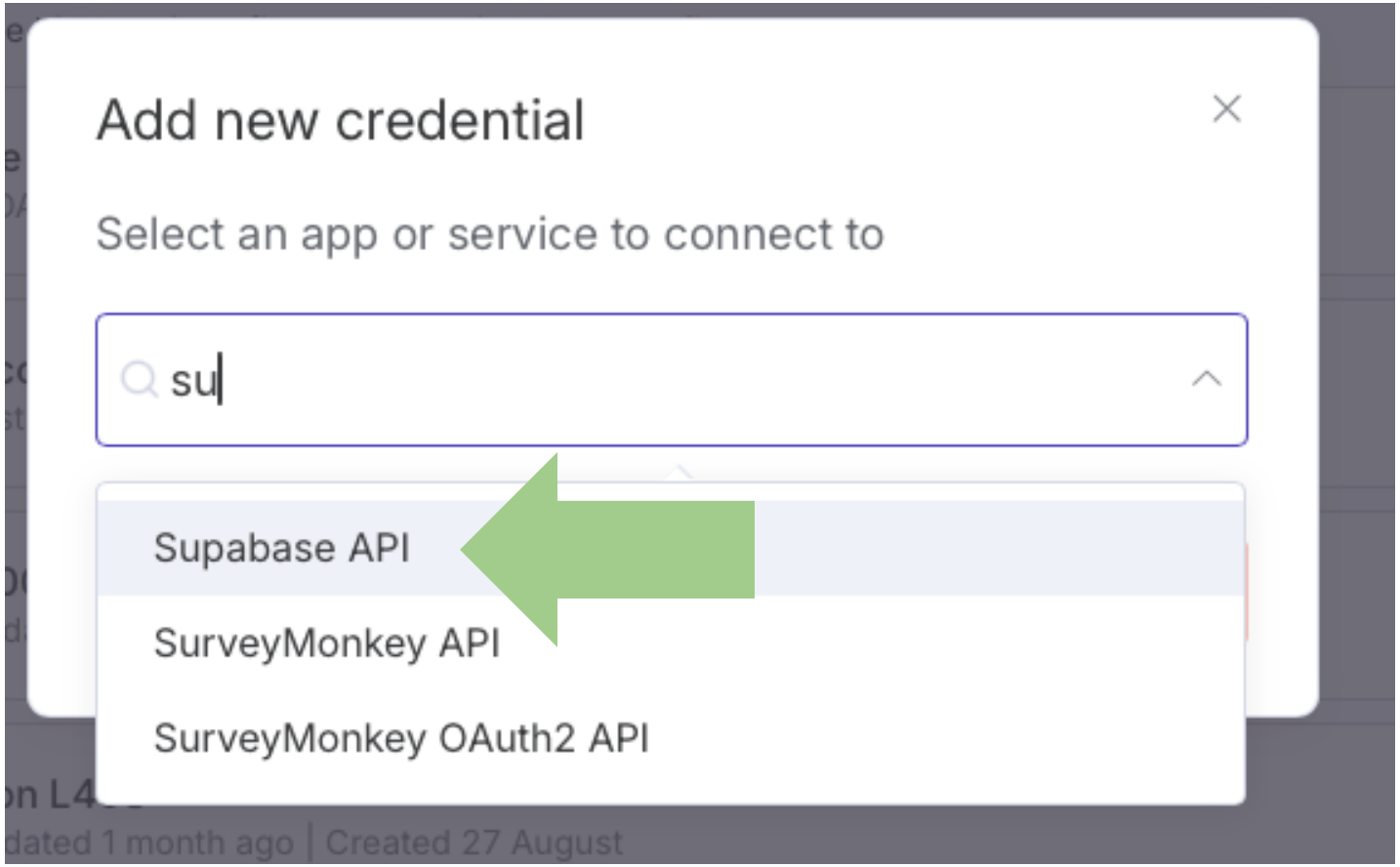
手動啟動vector extension



執行SQL指令，建立RAG資料庫



回到n8n，在n8n中右上角，



supabase

n8n