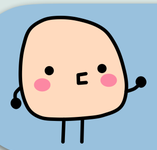
生成式 AI：文字與圖像生成的原理與實務

07.
檢索增強生成 (RAG) 的原理及實作

蔡炎龍
政治大學應用數學系

01.

RAG

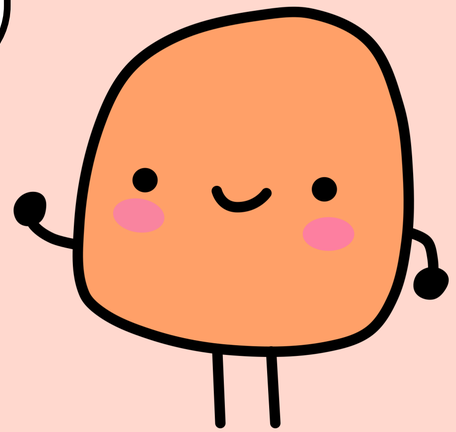


Prompt 其實很簡單

可以由電腦自動從資料庫中尋找嗎?

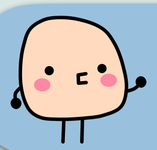
資訊

提供需要的正確資訊。



清楚的指引

例如, 以上面的資訊, 用什麼樣的格式、風格, 來回答使用者的問題。



RAG: 一個降低幻覺的方法

RAG

Retrieval-**A**ugmented **G**eneration

引發非常多的討論

FUTURE COMMERCE 未來商務

找文章 AI 解方 找活動

找文章 > TRENDS

RAG：讓 AI 更聰明的秘密武器，揭開生成式 AI 的新篇章

2024.10.28 | 偉利科技執行長 黃適文

businesswire A BERKSHIRE HATHAWAY COMPANY

NetApp與NVIDIA合作重新定義企業RAG並為代理式AI提供支持

NTAP資料基礎架構與強大的NVIDIA NeMo Retriever與NIM微服務相雲端中發現、搜尋和管理資料的方式，為AI應用提供助力

放了跨雲端和本地基礎架構儲存的EB級企業資料，以驅動RAG功能，使企業的用程式。

、NetApp ONTAP和NetApp BlueXP統一控制平面，以及NVIDIA NeMo軟體平台中)。

資策會Find

輕鬆與汽車對話！賓士運用「RAG」生成式AI技術打造「有話直說」功能

其工作原理簡單來說，就是當你向系統提問時，RAG會先在資料庫或其他外部數據源中檢索出與問題相關的資料。這些資料會被轉換成數位化的嵌入向量，代表資料的...

儲存

分享

LINE

收藏

遠見

最新 話題 國際 金融 AI 科技 產經 教育 民調

首頁 > AI > 時事新知

要當生成式AI新創，RAG就是最佳舞台

郭家齊 2024-06-07

T客邦

什麼是RAG？RAG技術突破LLM限制，結合「資訊檢索」和「文字生成」動態獲取知識大幅提升

RAG是一種全新的LLM訓練模式，透過將資訊檢索與文字生成結合，讓LLM能夠動態取用並理解新的知識，提升其在特定場景下的問答及文字生成品質。

網管人

避免AI幻覺RAG漸興起 防外洩「私有」當道

避免AI幻覺RAG漸興起防外洩「私有」當道 ... 企業對人工智慧與GenAI的態度，已經從觀望評估逐步走向在生產環境的落地應用。基於機敏資料的安全、人才以及成本...

奇摩股市

Vespa.ai宣布在檢索增強生成（RAG）中支援ColPali

ColPali將包括視覺元素在內的整個彩現文件嵌入到專為大型語言模型（LLM）最佳化的向量表示中，從而增強了文件檢索功能。這樣便能夠減少延遲，提高準確性，並...

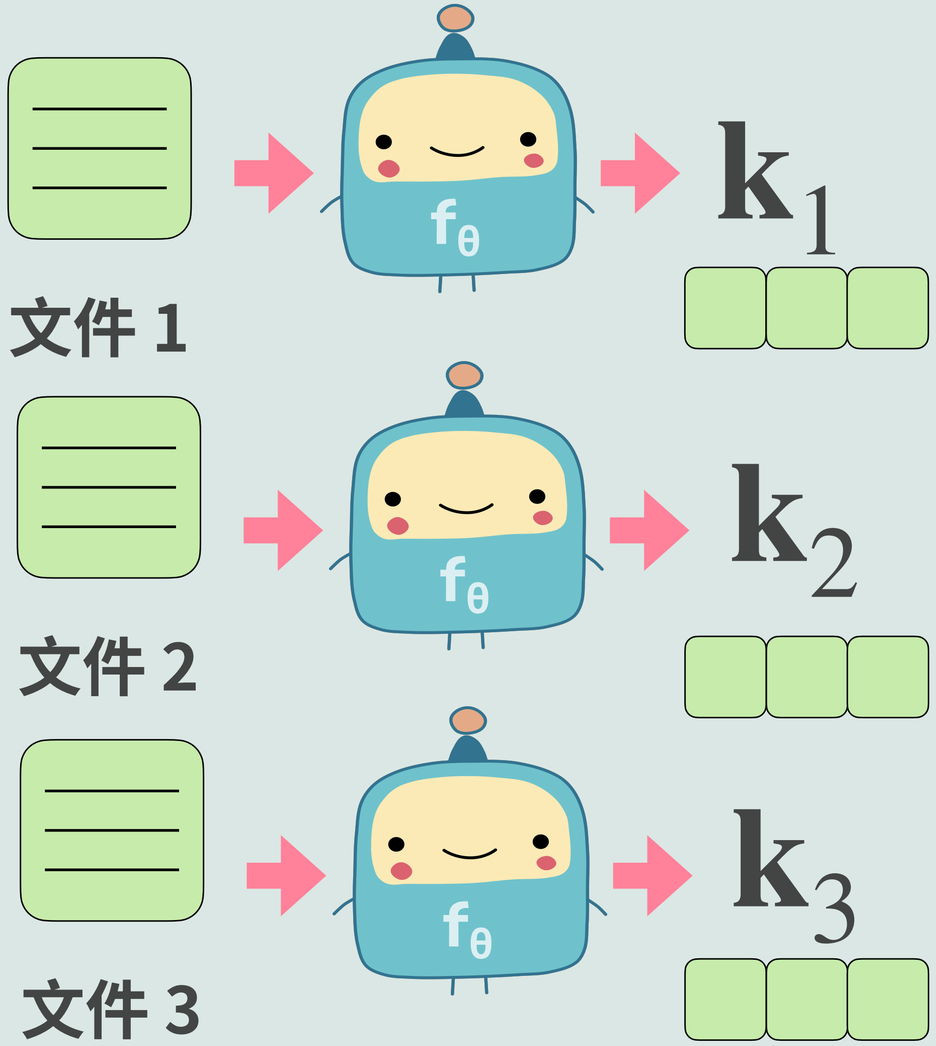
Ve

INSIDE

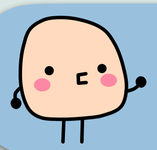
RAG 是什麼？就像是大型語言模型的「小抄」

RAG (Retrieval-Augmented Generation) 是一種自然語言處理模型，它結合了檢索（retrieval）和生成（generation）的技術。它使用了檢索模組來從大量資料中...

這是怎麼做到的呢?



每份重要資料都找到自己的**特徵代表向量**。

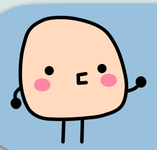


真正運作只是收集好自己的資料

.txt
.pdf
.docx

uploaded_docs

基本上純文字檔案都沒問題, 但其實 PDF 或是 Word 檔也都可以!



決定一下怎麼切「文字塊」(Chunk)

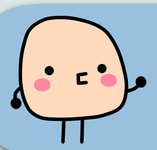
1000 字

重疊 200 字

1000 字

基本上要考慮多少字是一個「文字塊」，還有需不需要有重疊。

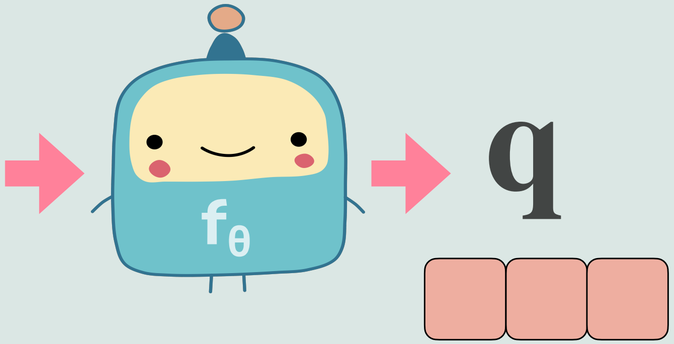
其實重點還很多...



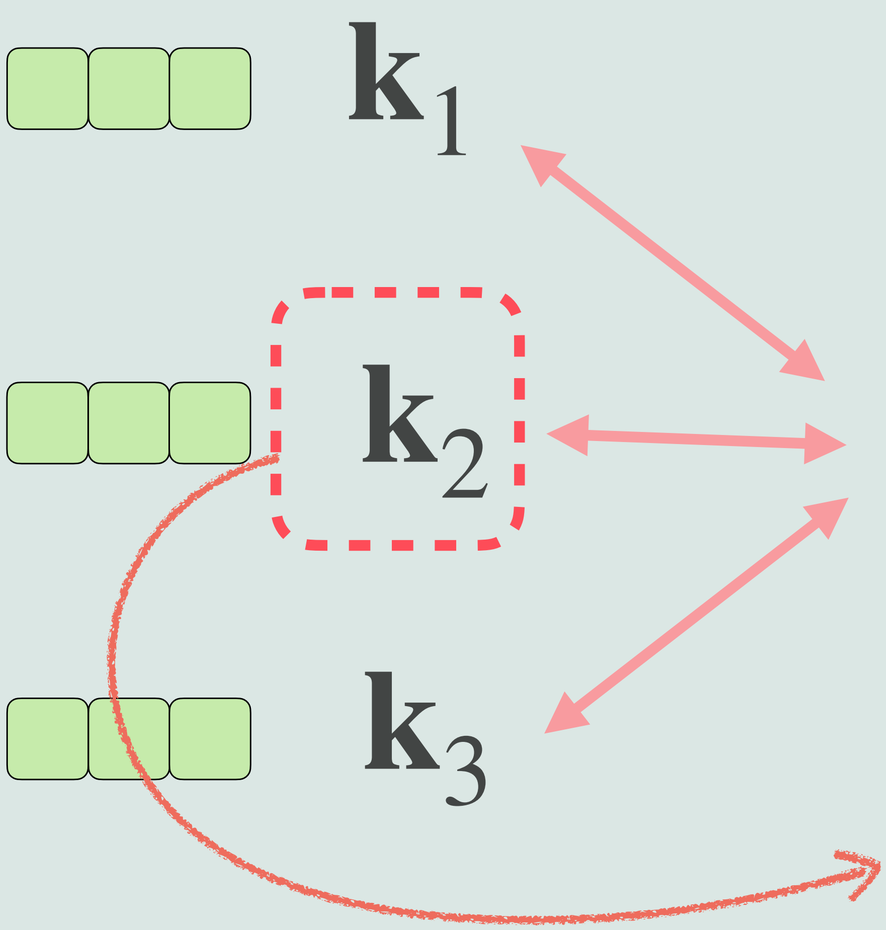
在我們下 prompt 問問題時

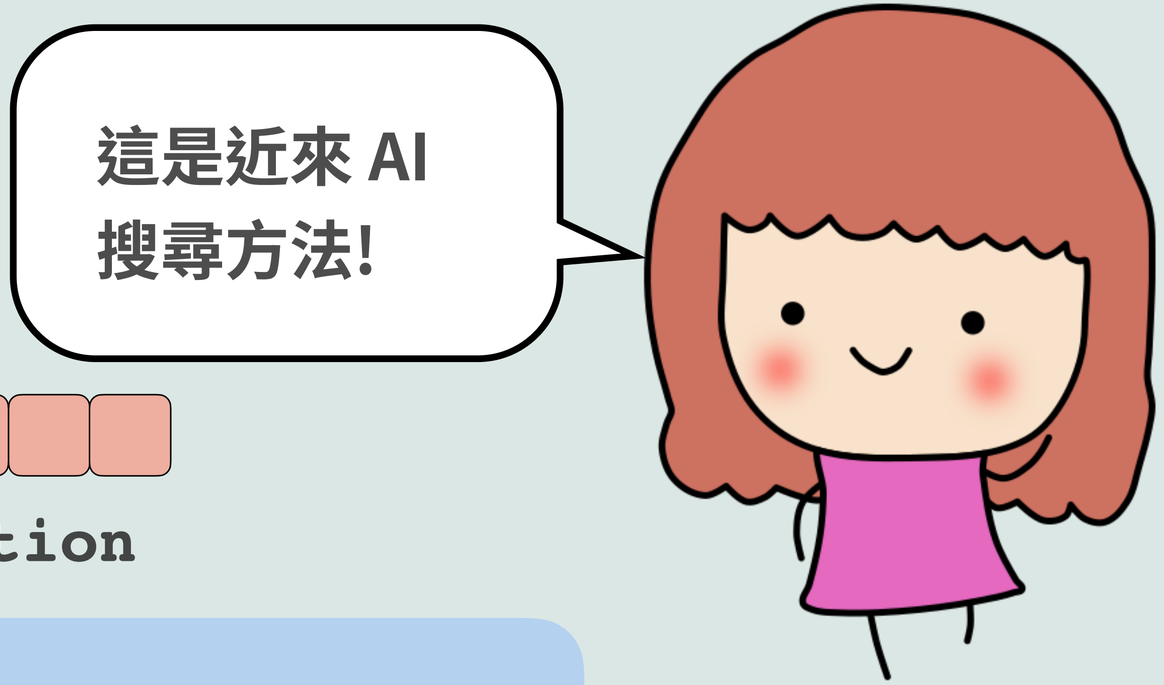
問題也化為特徵
代表向量。

一個問題 (prompt)

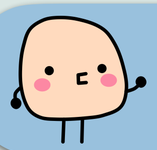


比較和誰最接近，一起放入 prompt 中





和原本問題最接近、最相關的資料。



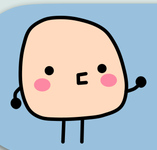
基本上會有這些資訊

`question`

使用者的問題 (原始 prompt)。

`retrieved_chunks`

用 RAG 找到的資訊。



接著就設計新的 prompt

“請根據 `{retriever_chunks}` 裡的資訊，
來回應使用者的問題: `{question}`”

基本上就是現在火紅 RAG 的概念

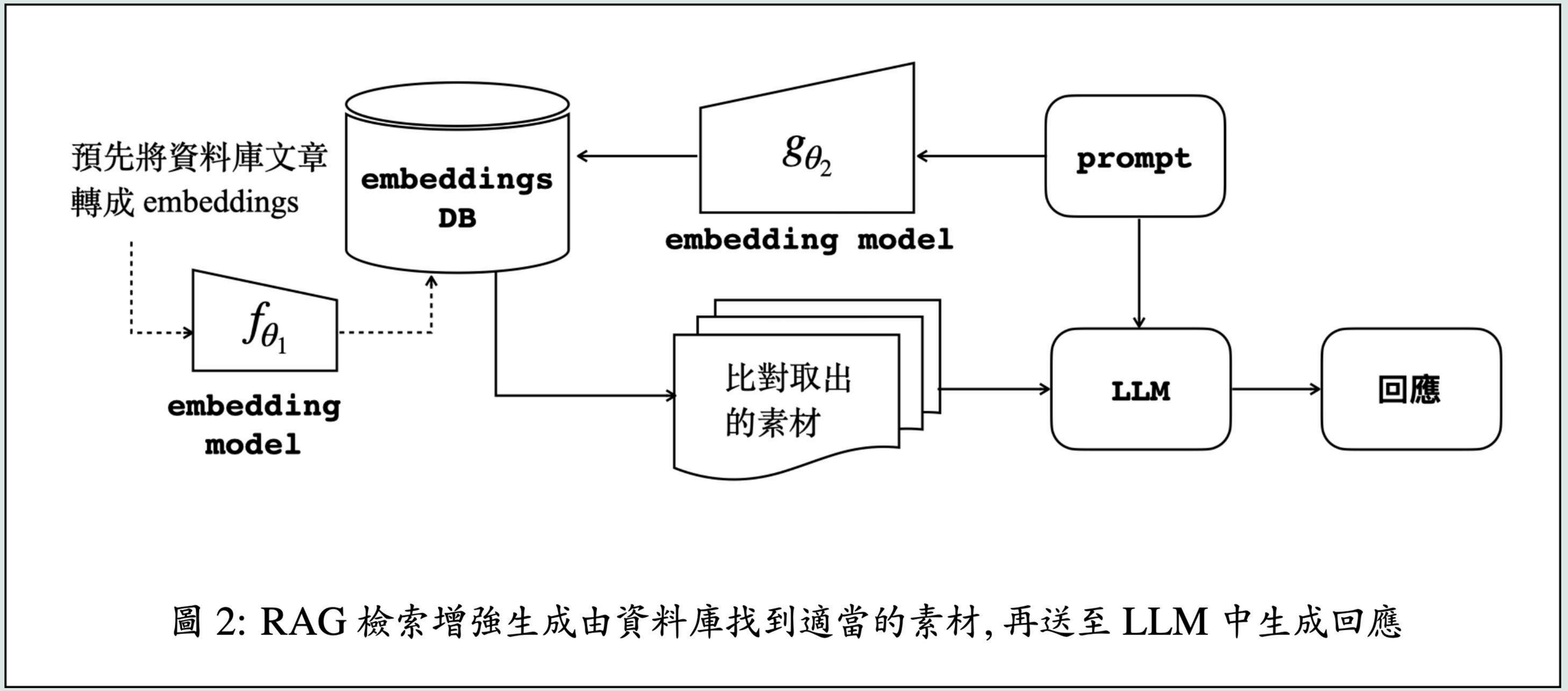
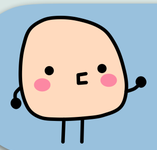


圖 2: RAG 檢索增強生成由資料庫找到適當的素材, 再送至 LLM 中生成回應

02.

RAG 的一些討論



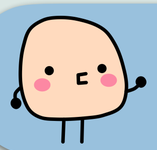
切法的考量

1000 字

重疊 200 字

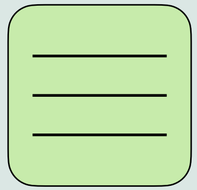
1000 字

文字塊大小、要怎麼切，像法修是否要一條一條切? 還有重疊要多少字?

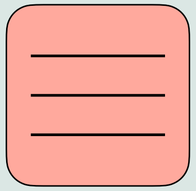


需要加入上下文或相關文章嗎?

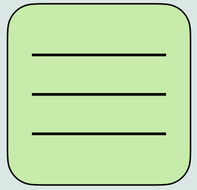
1



前一塊文字

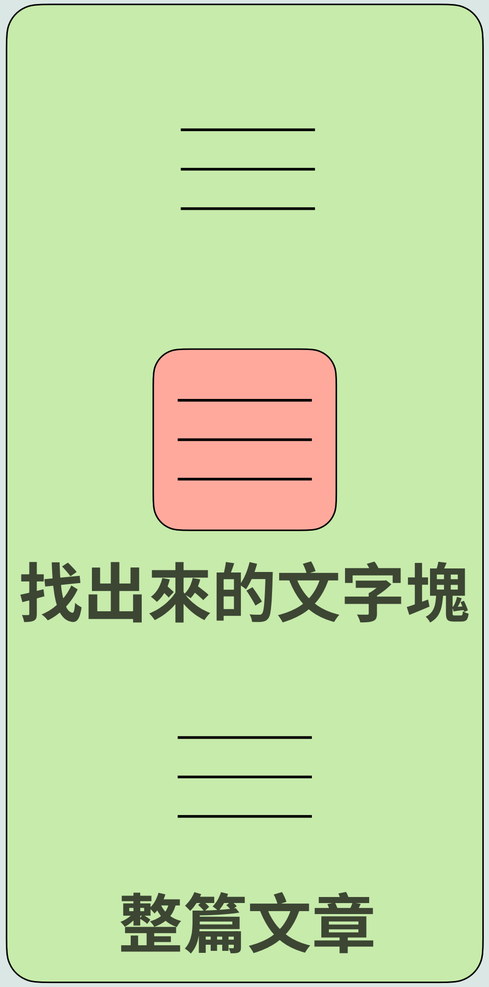


找出來的文字塊

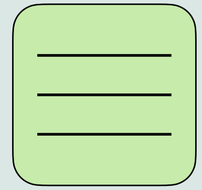


後一塊文字

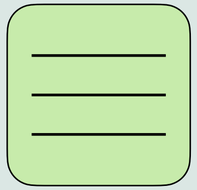
2



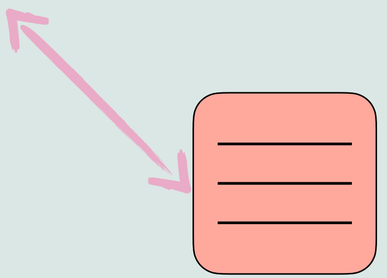
3



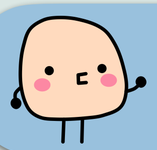
相關文章1



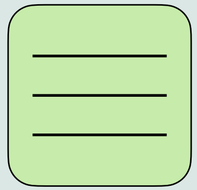
相關文章2



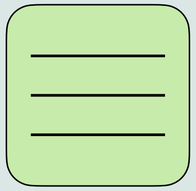
找出來的文字塊



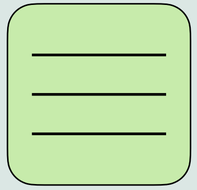
找出的資料通常不會只有一筆！



文字塊 1



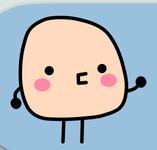
文字塊 2



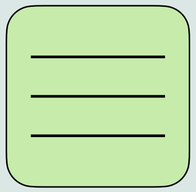
文字塊 3

有可能需要重新排序!

Reranking

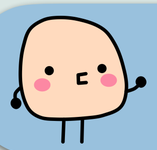


Metadata: 文字塊的附加資訊



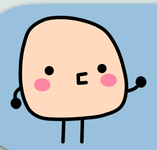
文字塊

比如這段文字的來源 (原文)、所在頁數、文章標題、作者、日期、分類... 等等。



需要 Metadata 的例子

我想知道什麼是 AI Agents，但不用管炎龍老師怎麼說，因為他都亂講。



話說回來: 真的要 RAG 嗎?

全部的資訊都放 Prompt
不就結了嗎?

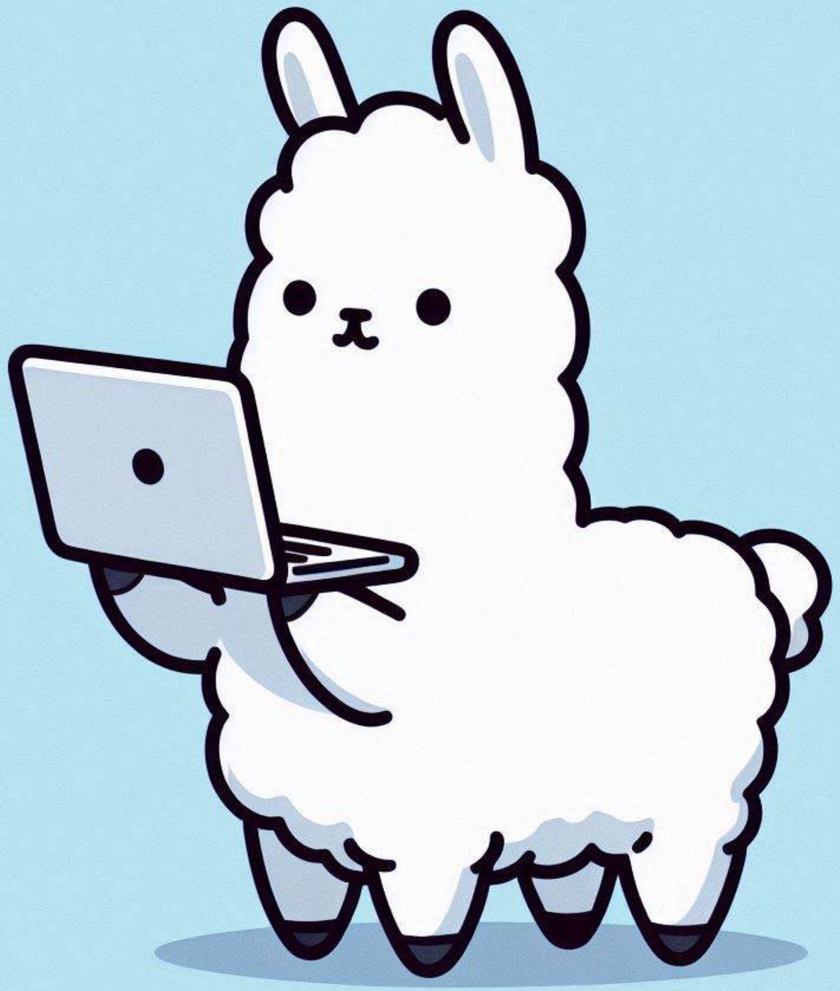
特別是語言模型上下文越來越長...

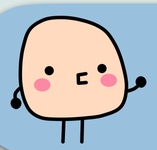
天生會看圖的大型語言模型 Llama 4

Behemoth
2TB 超級教師模型

Maverick
128 個專家模型

Scout
1 千萬超長上下文!





CAG: 用 KV-Cache 技術把資訊都塞給 LLM

Don't Do RAG: When Cache-Augmented Generation is All You Need for Knowledge Tasks

Brian J Chan*
Chao-Ting Chen*
Jui-Hung Cheng*
Department of Computer Science
National Chengchi University
Taipei, Taiwan
{110703065,110703038,110703007}@nccu.edu.tw

Hen-Hsen Huang
Insititue of Information Science
Academia Sinica
Taipei, Taiwan
hhhuang@iis.sinica.edu.tw

Abstract

Retrieval-augmented generation (RAG) has gained traction as a powerful approach for enhancing language models by integrating external knowledge sources. However, RAG introduces challenges such as retrieval latency, potential errors in document selection, and increased system complexity. With the advent of large language models (LLMs) featuring significantly extended context windows, this paper proposes an alternative paradigm, cache-augmented generation (CAG) that bypasses real-time retrieval. Our method involves preloading all relevant resources, especially when the documents or knowledge for retrieval are of a limited and manageable size, into the LLM's extended context and caching its runtime parameters. During inference, the model utilizes these preloaded parameters to answer queries without additional retrieval steps. Comparative analyses reveal that CAG eliminates retrieval latency and minimizes retrieval errors while maintaining context relevance. Performance evaluations across multiple benchmarks highlight sce-

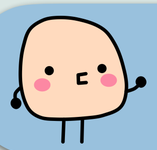
1 Introduction

The advent of retrieval-augmented generation (RAG) [2, 5] has significantly enhanced the capabilities of large language models (LLMs) by dynamically integrating external knowledge sources. RAG systems have proven effective in handling open-domain questions and specialized tasks, leveraging retrieval pipelines to provide contextually relevant answers. However, RAG is not without its drawbacks. The need for real-time retrieval introduces latency, while errors in selecting or ranking relevant documents can degrade the quality of the generated responses. Additionally, integrating retrieval and generation components increases system complexity, necessitating careful tuning and adding to the maintenance overhead.

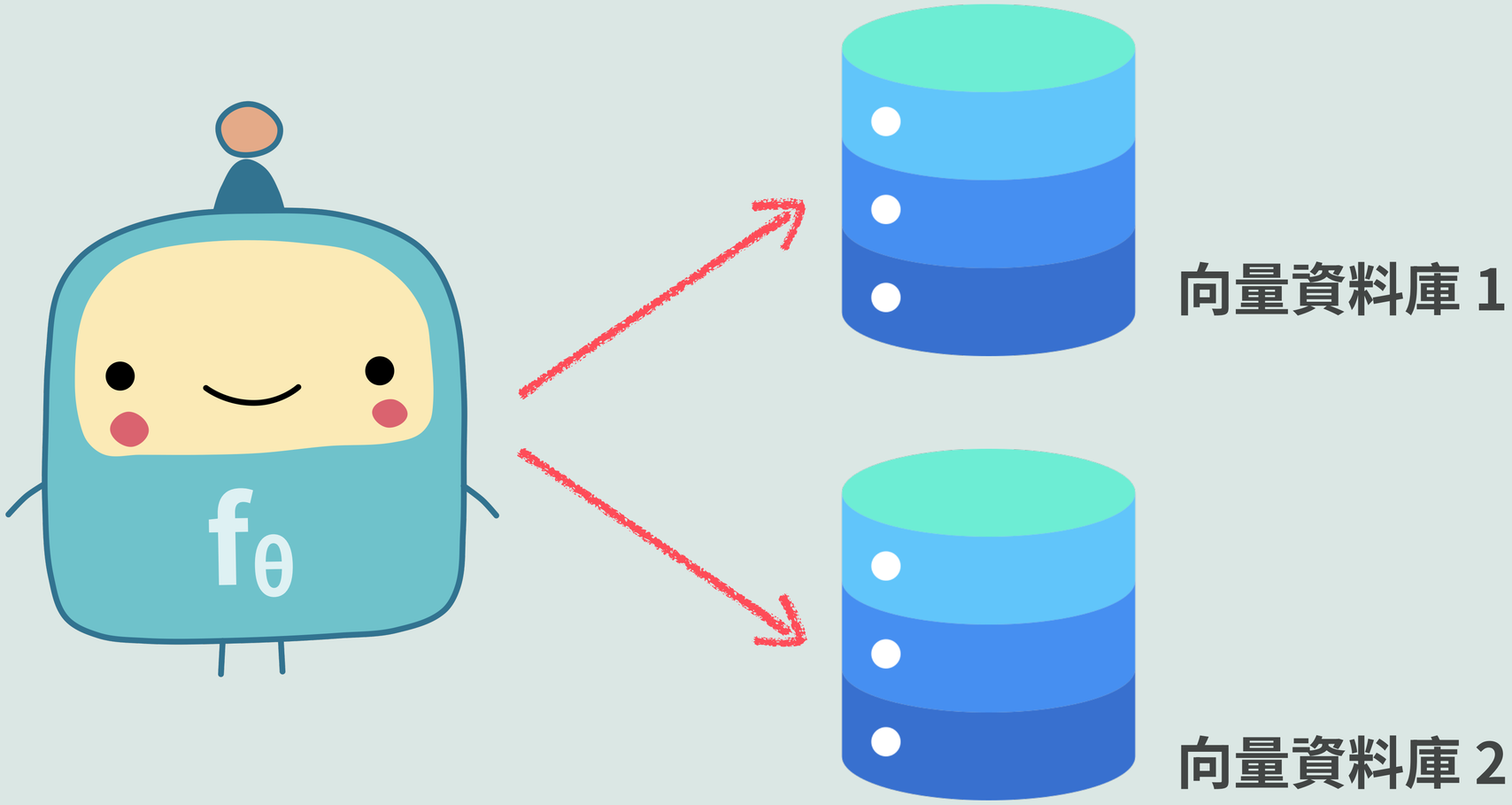
This paper proposes an alternative paradigm, cache-augmented generation (CAG), leveraging the capabilities of long-context LLMs to address these challenges. Instead of relying on a retrieval pipeline, as shown in Figure 1, our approach involves preloading the LLM

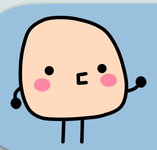
03.

記憶的一種方式

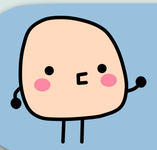


同一個系統, 可以用兩個以上的向量資料庫





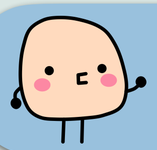
接著教大家很嘴的事



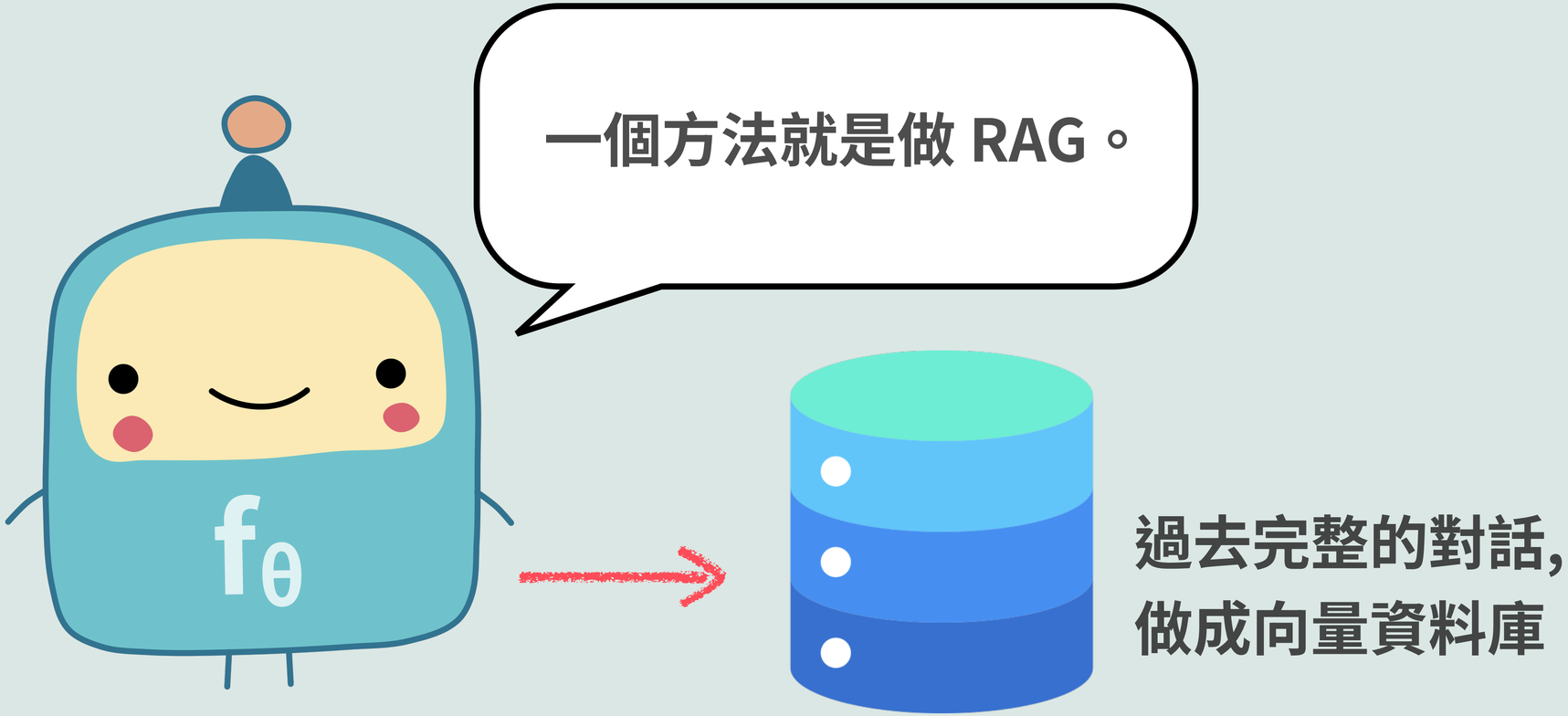
大型語言模型的「短期記憶」

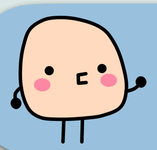
就是我們這次和大型語言模型的對話過程。

下次來, 這次說的話都忘了。
更糟糕的是, 對話過程太長,
前面的話也會忘了。

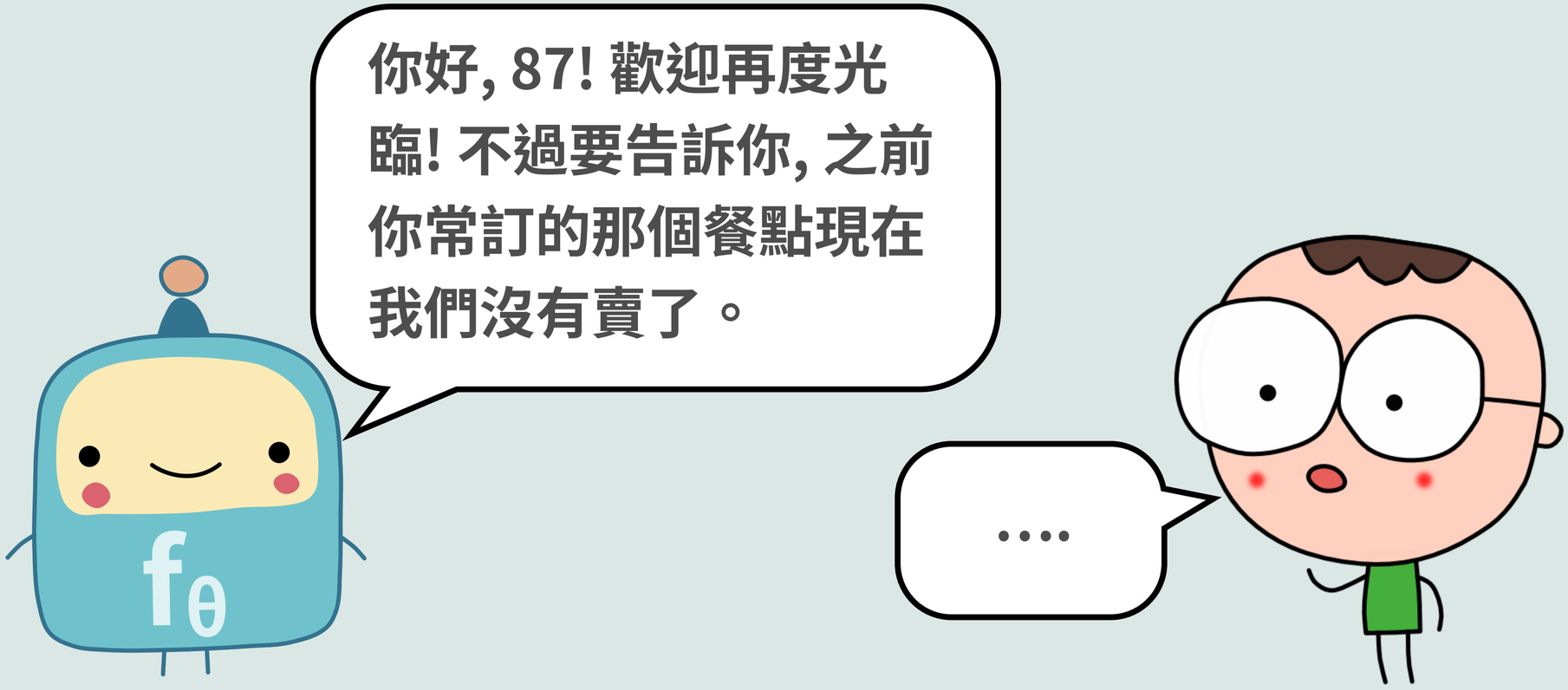


怎麼做大型語言的「長期記憶」呢?



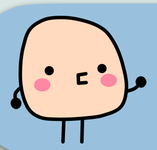


特別是客製化的對話機器人, 會記得你的點點滴滴



04.

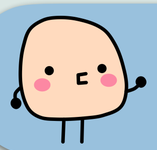
RAG 金融相關的應用



RAG 在金融業可能的應用

* 客戶服務與問答系統
* 內部知識管理
* 教育與培訓
* 個人化理財建議
* 投資研究報告生成
* 反洗錢與詐欺偵測
* 金融法規與合規諮詢
* 財經新聞與趨勢總結

05.
【作業】打造你的 RAG 系統



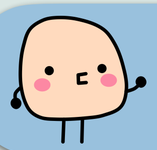
打造自己的 RAG 強化對話機器人!

程式 A 用自己的資料、建立向量資料庫

`https://yenlung.me/AI06a`

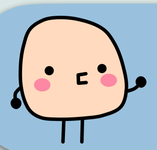
程式 B 實作 RAG 系統

`https://yenlung.me/AI06b`



作業：打造你的 RAG 系統

* 可以找一個你有興趣的文件 (比如說某個規定、相機說明書等等)
* 程式 A: 用自己的數據, 打造向量資料庫, 並儲存為 `faiss_db.zip`
* 程式 B: 程式要能讀入你的 `faiss_db.zip` 並解壓縮
* 設計自己的 prompt，打造自己的 RAG 強化的對話機器人



Q & A