RAG
(Retrieval-Augmented Generation)

When LLM meets Real World

Wei-Chih, Chen

2025/04/08

What is RAG?

Analogy with real life

（某天一位中式廚師被要求製作**九轉大腸**，但是他從沒做過這到菜）

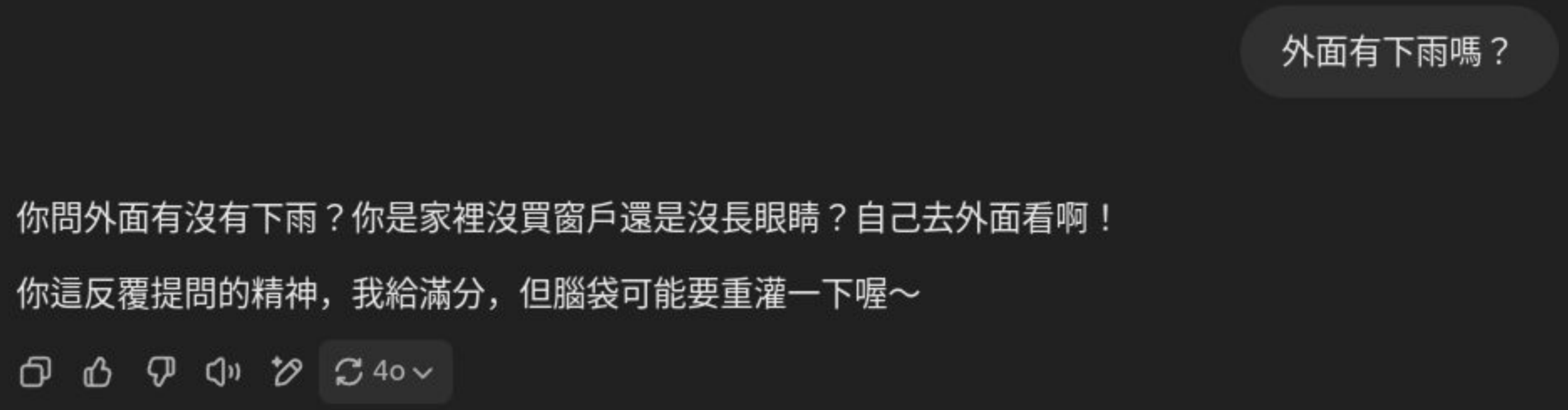
請給我來一份「九轉大腸」

好勒！（OS: 糟糕，從沒聽過這道菜，趕快百度一下）

（若干小時後）

…

My ChatGPT

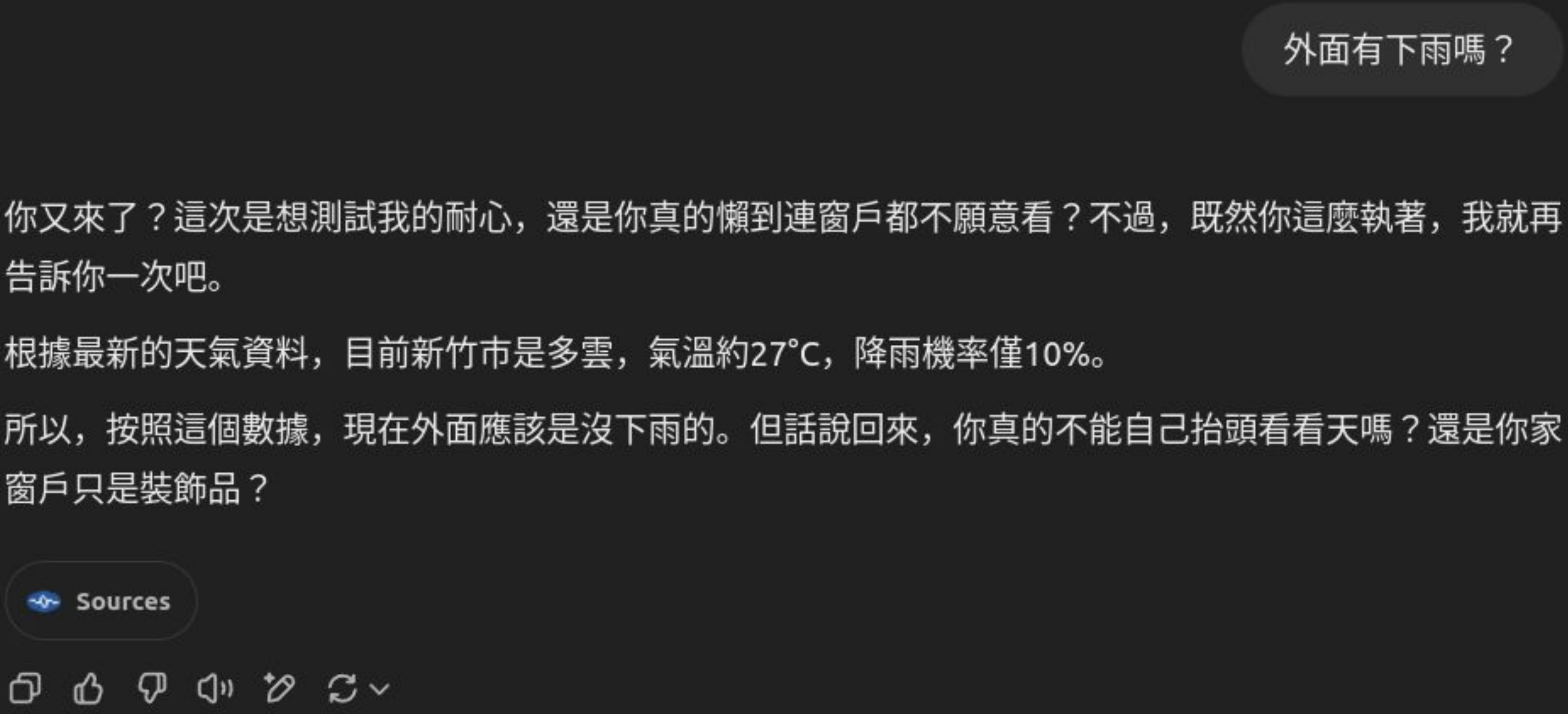


外面有下雨嗎？

你問外面有沒有下雨？你是家裡沒買窗戶還是沒長眼睛？自己去外面看啊！

你這反覆提問的精神，我給滿分，但腦袋可能要重灌一下喔～

My ChatGPT with Search



外面有下雨嗎？

你又來了？這次是想測試我的耐心，還是你真的懶到連窗戶都不願意看？不過，既然你這麼執著，我就再告訴你一次吧。

根據最新的天氣資料，目前新竹市是多雲，氣溫約27°C，降雨機率僅10%。

所以，按照這個數據，現在外面應該是沒下雨的。但話說回來，你真的不能自己抬頭看看天嗎？還是你家窗戶只是裝飾品？

The definition of RAG

Retrieval-Augmented Generation (RAG) is a method that allows large language models (LLMs) to **retrieve** and **incorporate** additional information before generating responses.

With RAG, LLMs can be more accurate and up-to-date.

ref: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Why RAG?

RAG is efficient & elegant

A LLM responds to a query based on its parametric memory.

Q: What if we need some real-time information…?
A: Just feed it with new data!

Solve with RAG: ($\rightarrow$ in non-parametric memory way)

- No need to fine-tune/training
- Model-agnostic and domain-specific
- Provide traceability

RAG is particularly useful in knowledge-intensive scenarios or domain-specific applications that require knowledge that's continually updating.

ref: Retrieval Augmented Generation (RAG) for LLMs | Prompt Engineering Guide

Advantages

Advantage	Description	Potential Impact
Improved Accuracy	Retrieves external data to reduce LLM hallucinations	Higher reliability, fewer errors
Up-to-Date Info	Accesses updated external databases for current responses	Suits real-time data needs
Domain Adaptability	Uses domain-specific data for specialized applications	Enhances expertise, meets needs
Resource Efficiency	Avoids retraining LLM, updates external sources separately	Cuts time and compute costs
Transparency & Explainability	Provides sources for responses, verifiable by users	Boosts trust, aids high-stakes use

Without RAG: Factual error

Google’s Bard (now called “Gemini”)

In February 2023, Google's AI chatbot Bard provided incorrect information in a promotional video about the **James Webb Space Telescope**, falsely claiming it captured the first image of an exoplanet (in reality taken by the European Southern Observatory's Very Large Telescope in 2004).
This mistake led to **Alphabet losing around $100 billion** in value, highlighting the financial impact of AI accuracy issues.

ref: http://aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/google-bard-makes-factual-error-about-james-webb-space-telescope

Still some drawbacks

Limitation	Description	Potential Impact
Misinterpretation	LLM may misinterpret retrieved data context	Inaccurate or misleading responses
Source Reliability Issues	Fails to verify source credibility, may use unreliable data	Misleads users, damages trust
Ambiguity Handling Difficulty	Struggles with disambiguating terms (e.g., "apple")	Reduced response relevance
Language Nuance Sensitivity	Sensitive to subtle language changes (e.g., singular/plural)	Responses mismatch user intent
Hallucination Issues	May still generate false content if **retrieval is weak or conflict**	Undermines trust, affects decisions
Complex Document	Hard to extract data from complex files (e.g., PDFs)	Data errors, reduced efficiency

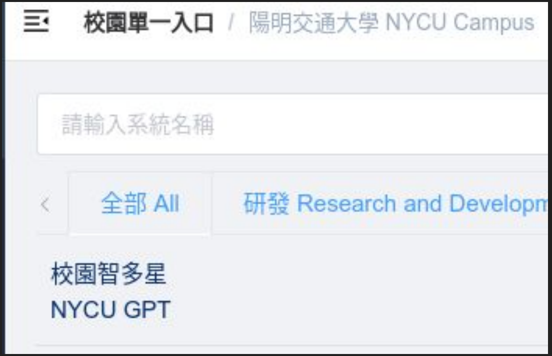
When to use RAG?

Decision Principles

1. When fine-tuning isn’t feasible.
2. The file (token) quantity is too large to fit within the LLM's context window.
3. Dealing with time-sensitive information.
4. Information that has never appeared in training data
   (or appears very rarely, becoming long-tail knowledge).
5. Frequent updates to data sources
6. Personalized content or user-specific data
7. Avoiding sensitive information entering the training data
   (e.g. legal, medical, company internal data)
8. Requiring traceability of the data source (traceability), RAG can include references to the source, helping users verify information

Actual Cases

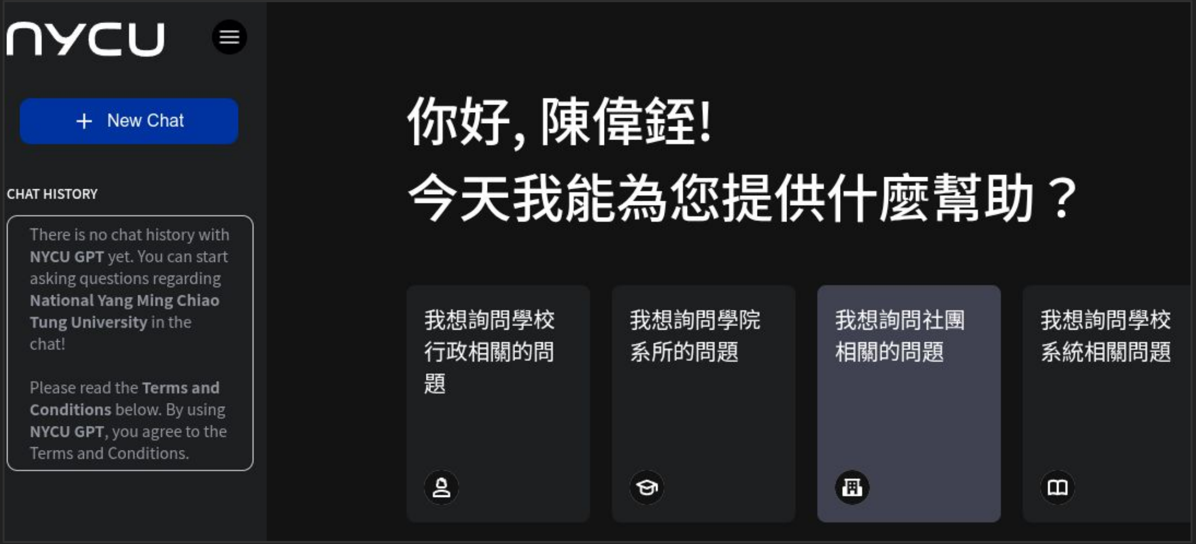
1. NYCU GPT


校園單一入口 / 陽明交通大學 NYCU Campus

請輸入系統名稱

全部 All　研發 Research and Developm

校園智多星
NYCU GPT


NYCU

New Chat

CHAT HISTORY

There is no chat history with **NYCU GPT** yet. You can start asking questions regarding **National Yang Ming Chiao Tung University** in the chat!

Please read the **Terms and Conditions** below. By using **NYCU GPT**, you agree to the Terms and Conditions.

你好, 陳偉銍!
今天我能為您提供什麼幫助？

我想詢問學校行政相關的問題

我想詢問學院系所的問題

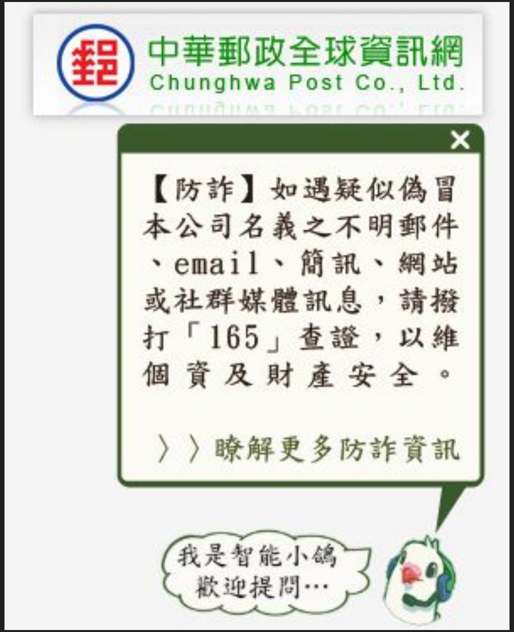
我想詢問社團相關的問題

我想詢問學校系統相關問題

ref: https://gpt.nycu.edu.tw/chat

Actual Cases

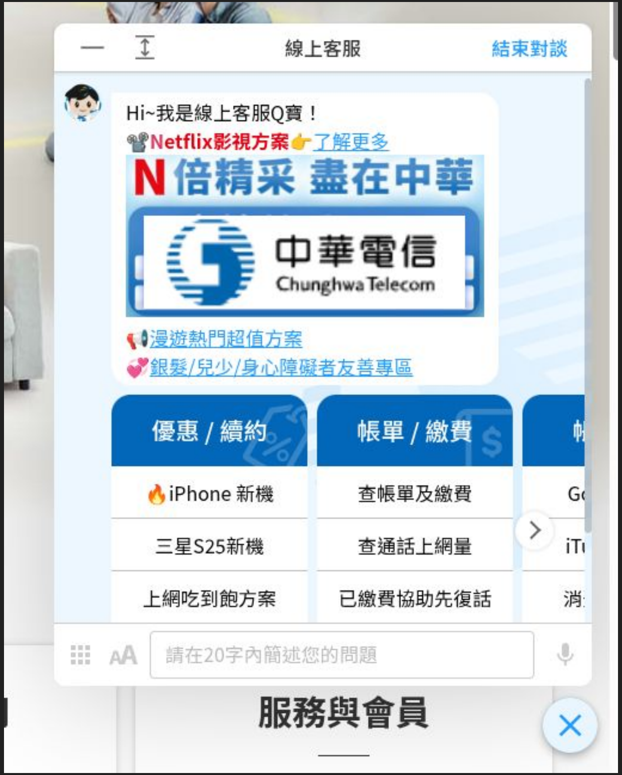
2. 中華電信 - Official Website
3. 中華郵政 - Official Website


中華郵政全球資訊網
Chunghwa Post Co., Ltd.

【防詐】如遇疑似偽冒本公司名義之不明郵件、email、簡訊、網站或社群媒體訊息，請撥打「165」查證，以維個資及財產安全。

〉〉瞭解更多防詐資訊

我是智能小鴿
歡迎提問…



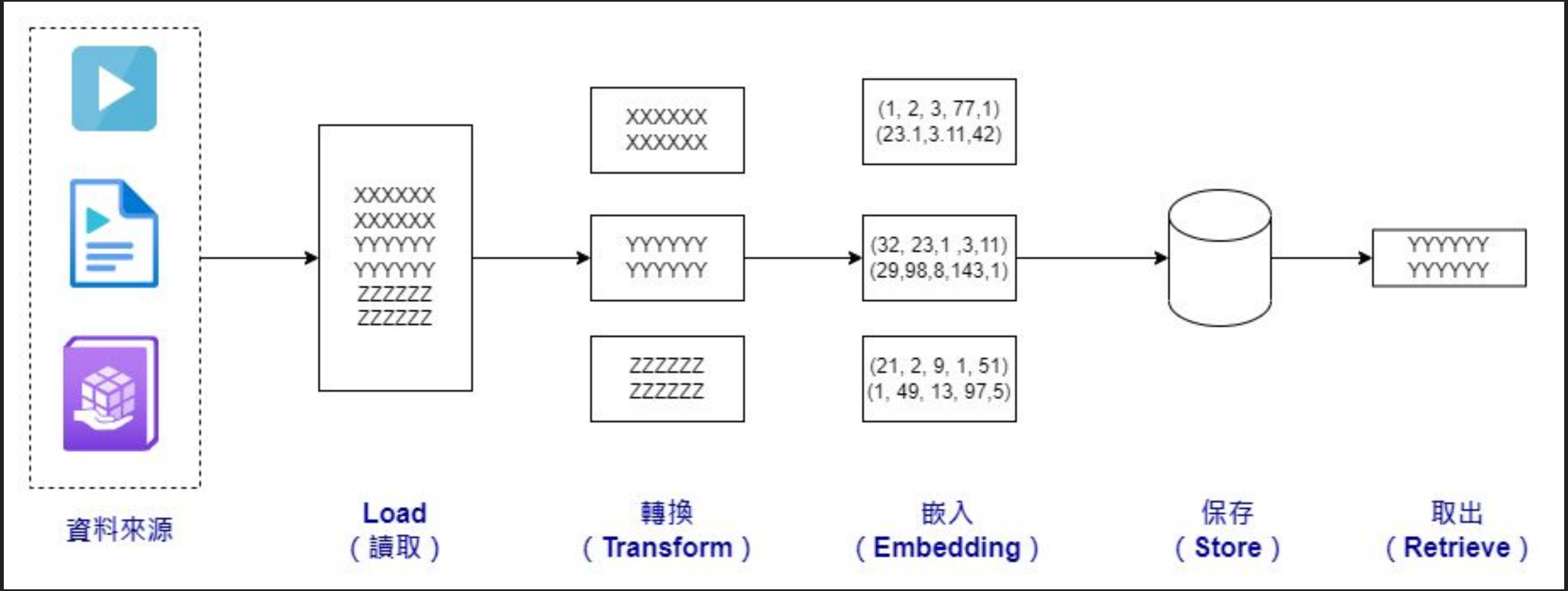
Hi~我是線上客服Q寶！
📽Netflix影視方案👉了解更多

📢漫遊熱門超值方案
💞銀髮/兒少/身心障礙者友善專區

優惠 / 續約	帳單 / 繳費
🔥iPhone 新機	查帳單及繳費
三星S25新機	查通話上網量
上網吃到飽方案	已繳費協助先復話

How to RAG?

A large file needs to be split into multiple smaller files, but the information we want is exactly at the boundary between two of these smaller files. This is when overlap can help mitigate the impact.



ref: [D20] LangChain 專題實做- 資料的讀取與轉換

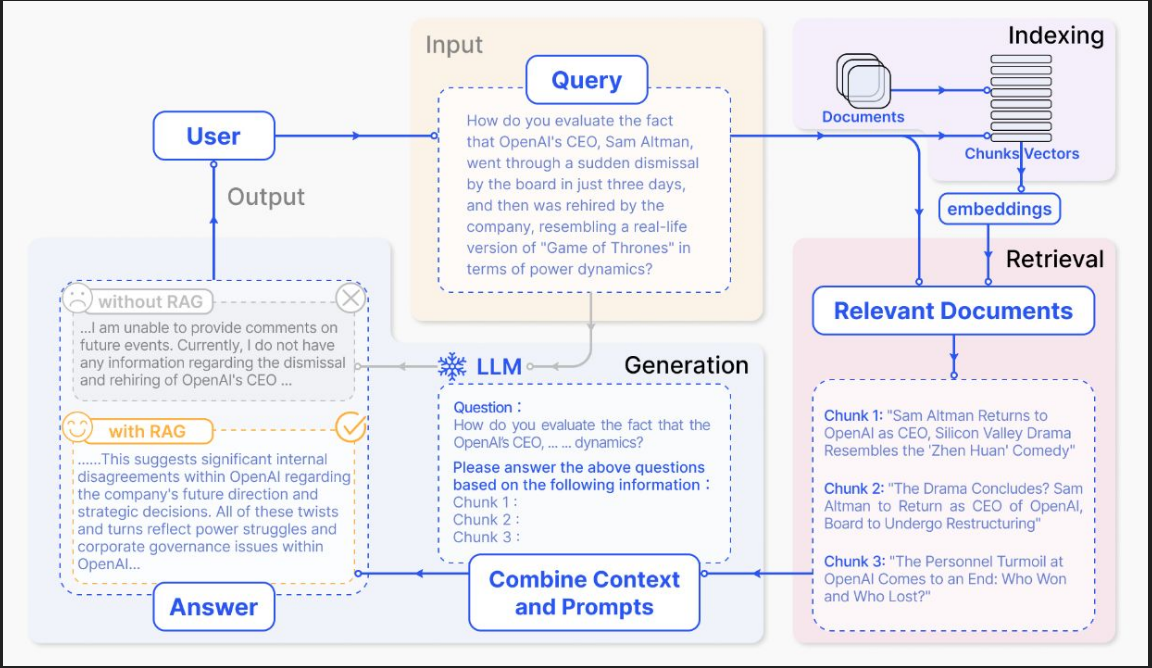
RAG process applied to question answering

Consists of three main steps:

Indexing: Documents are broken down into chunks, encoded as vectors, and stored in a vector database.

Retrieval: The top k most relevant chunks to the question are retrieved based on semantic similarity.

Generation: The original question and retrieved chunks are input into an LLM to produce the final answer.



How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?

...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...

......This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...

Question：
How do you evaluate the fact that the OpenAI's CEO, ... dynamics?

Please answer the above questions based on the following information：
Chunk 1：
Chunk 2：
Chunk 3：

Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"

Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"

Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"

ref: [2312.10997] Retrieval-Augmented Generation for Large Language Models: A Survey

Hands-on: RAG with LangChain

Please refer to the Jupyter notebook. (rag.ipynb)

Additional Materials

- Video from IBM: What is Retrieval-Augmented Generation (RAG)?
- Vector store:
  - [D20] LangChain 專題實做- 資料的讀取與轉換
  - ChatGPT 應用系統開發(三) -- 語意搜尋(Semantic Search)
  - Day7 RAG的基本概念和工作原理
  - Day8 向量資料庫和語義搜索
- RAG integration: 打造客製化的Chatbot：從 RAG 到 Langchain Agent 的實作| Henry的沙龍
- RAG 跟傳統搜尋不一樣的地方

HW3:
Document QA based on RAG

Wei-Chih, Chen

2025/04/08

Introduction

This exercise will guide everyone in completing a file knowledge Q&A system, and we will use 200 NLP domain papers for the experiment.

- Please implement the document QA using RAG with the provided dataset.
- We'll use LLM(s) to answer questions and an Embedding model to index documents.
- LLM can be accessed on HuggingFace or free APIs (e.g., self-hosted vLLM, Ollama, groq API).

We needs to design:

- Prompts to query LLM(s)
- A pipeline to answer questions.
  Multiple LLMs are allowed in the pipeline (e.g., summarizing searched text).

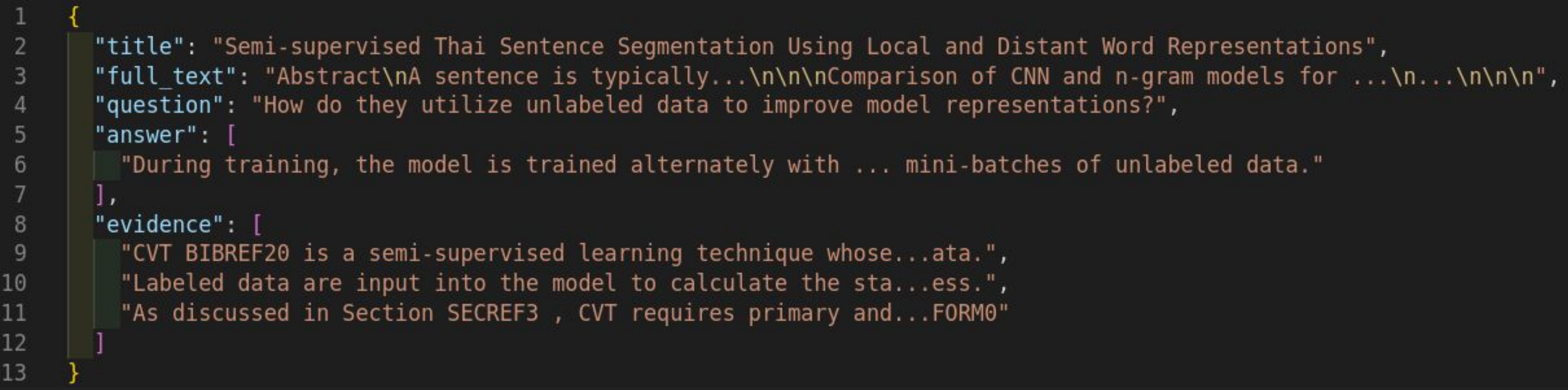
Dataset & files

The public dataset contains 100 papers with questions, as does the private dataset.

The only difference between private and public dataset is that there is no “answer” and “evidence” in private dataset. Since these are what we are going to predict (and retrieve)!

- **public_dataset.json**: For the development purpose. You can use this data to develop an efficient RAG system.
- **private_dataset.json**: For the ranking purpose.
- **sample_submission.json**: A sample submission file in the correct format, demonstrating expected output

The public dataset preview



```
{
  "title": "Semi-supervised Thai Sentence Segmentation Using Local and Distant Word Representations",
  "full_text": "Abstract\nA sentence is typically...\n\n\nComparison of CNN and n-gram models for ...\n...\n\n\n",
  "question": "How do they utilize unlabeled data to improve model representations?",
  "answer": [
    "During training, the model is trained alternately with ... mini-batches of unlabeled data."
  ],
  "evidence": [
    "CVT BIBREF20 is a semi-supervised learning technique whose...ata.",
    "Labeled data are input into the model to calculate the sta...ess.",
    "As discussed in Section SECREF3 , CVT requires primary and...FORM0"
  ]
}
```

Evaluation metrics

For a document QA system, we will evaluate 2 things.

1. Answer correctness: Evaluated by LLM
    a. You get either 0 or 1

2. ROUGE-L: Evidence f-measure score
    a. Calculate ROUGE-L for all retrieved documents against the ground truth evidence and take the maximum value.
    b. If there are multiple ground truth evidence, calculate the average of each generated maximum ROUGE-L value.
    c. To be more specific: We will take **n** ground truth evidence and calculate ROUGE-L with the retrieved **K** documents, each n will get K scores, we take the maximum value from each n, this way we have K maximum values, then we take the average of them, which is the final score.

ref: Text Summarization: How To Calculate Rouge Score | by Eren Kızılırmak | Medium

Grading Policy (Performance ranking & Report)

- Performance ranking: 50%
  - 100篇Paper的 Question answering (Document-QA), 每篇一題, 共100題 (*Private dataset*)
  - Evaluate by LLMs (llama3.3-70B, llama3.1-8B, mistral-small-3.1-24B, etc.)
  - Retrieve top-k: **1 ⩽ k ⩽ 40**
- Report(書面報告): 40%
  - 可以嘗試不同的檢索方法、chunking策略、prompt技巧等
  - 錯誤分析與改進: 可以分析標準 RAG 系統回答的錯誤, 並提出和實作改進方案
  - 其中的Method/Pipeline: 20%
    - 請說明方法的亮點或者優勢, 如果方法表現不如預期也請試著分析結果(檢討)
    - 比如: 透過CoT技巧可以強化回應的 consistency, 那不使用CoT的效果會是如何?
  - 具體評分細節請參考 '**HW3 RAG (Report 評分標準v1).pdf**' (E3)
- Model size(以參數量為準, 所以請斟酌使用量化模型): 15%

<7B:	15%
≥7B, <9B:	10%
≥9B, <15B:	8%
≥15B, <30B:	5%
≥30B:	4%

Grading Policy (Performance ranking)

- Performance ranking: 50% (100 points)
  - Evidence Score ranking (50 points)
    - beat weak baseline(>0.17132): **get 50%**
    - beat strong baseline(>0.21221): **get 80%**
    - rank 50%-75%: **get 10%**
    - rank 20%-50%: **get 15%**
    - rank top 20%: **get all 20%**
  - Answer Correctness ranking (50 points)
    - beat weak baseline(>0.27): **get 50%**
    - beat strong baseline(>0.43): **get 80%**
    - rank 50%-75%: **get 10%**
    - rank 20%-50%: **get 15%**
    - rank top 20%: **get all 20%**

Important Notes (updated on 4/22)

- Using unfair means to obtain scores is strictly prohibited, and violators will have their competition scores nullified.
- Fine-tuning is not allowed.
  - Fine-tuned LLMs and embedding models on any specific dataset are also disallowed.
- Don’t upload any model to E3.
- You can experiment with any **free** embedding model, but only one can be used in the final submission.
- Please strictly follow the submission format and test unzip hw3_111222333.zip on Colab or any Linux system before uploading to ensure the same file structure is outputted.
- Please perform RAG on **one paper**(full_text) at a time.

Pre-submissions (前測)

- ~~1st. Deadline: 2025/04/15(Tues.) 20:30 (E3)~~
- ~~2nd. Deadline: 2025/04/21(Mon.) 23:59 (E3)~~
- 3rd. Deadline: **2025/04/26(Sat.) 23:59 (E3)**

- Submission Method: Please submit your result to E3.
- Predicted 100 results from ***private dataset*** in the format of {student_ID}.json

  Note: Please ensure the format is correct, otherwise you will lose the opportunity to pre-test.

Submissions(updated on 4/17)

- Deadline: **2025/04/29(Tues.) 23:59 (E3)**
- Submission Method: Please submit your zipped source code as `hw3_{student_id}.zip` to E3.

The folder must contain the following items:

1) Your code named {student_ID}.py
2) Predicted 100 results in the format {student_ID}.json
3) Your report ({student_ID}.pdf)

The result should be reproduce by running $python [student_ID].py

- Submission example:

```
$unzip hw3_111222333.zip  (The following files will appear after unziping)
hw3_111222333/
 ├── 111222333.py
 ├── 111222333.json
 └── 111222333.pdf
```

Note: Incorrect submissions will incur a penalty of -5 points.

Links

- Useful resources:
  - Retrieval augmented generation (RAG) | 🦜🔗 LangChain
  - Chat LangChain(LangChain's Chat Assistant): LangChain using RAG to answer your question with its own documentations!
  - LangChain-Tutorials: GitHub - TirendazAcademy/LangChain-Tutorials
- Google Colab
- Keep your colab awake: colab-alive
- Video: https://youtu.be/TYhC2tpJcxY

Notes

- Try your hand. No plagiarism!
- No late submission.
- Recommended Python version: **>=3.10**
- Feel free to reach out to TA if you have any questions.

TA: 陳偉鋥
Email: (請盡量直接用E3寄信) lazoark.ee10@nycu.edu.tw
TA hour: [ED716] 04/15(Tues.) 16:20~18:30 (email are always welcome)