

AI Day LLM Workshop R1

Introduction to Large Language Models & Retrieval Augmented-Generation

1. What's LLM
2. RAG Basic
3. Introduction to Hugging face

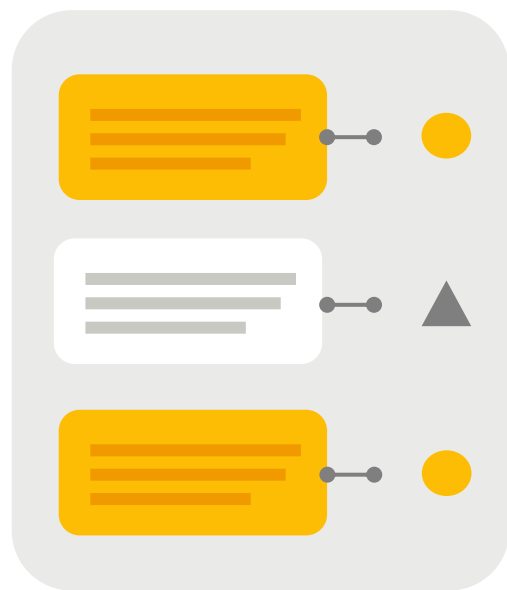
NLP before Large Language Models



NLP before Large Language Models



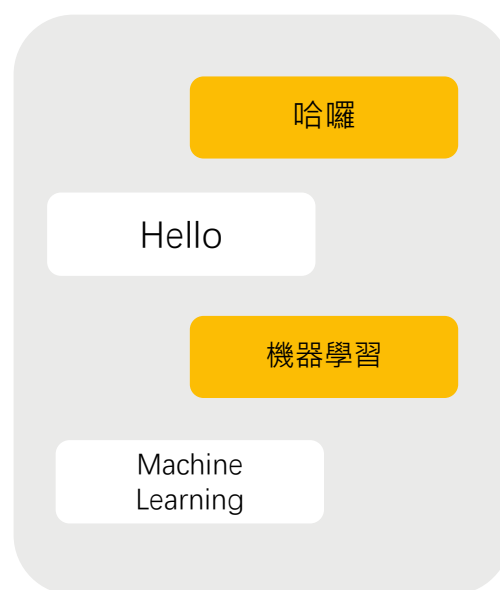
NLP Tasks



文本分類



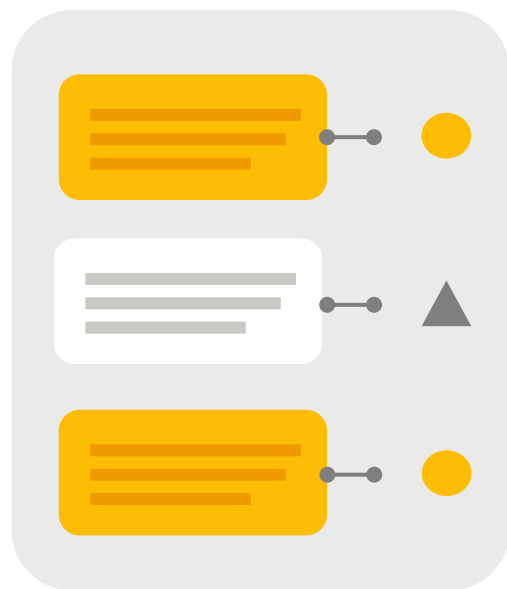
問答



翻譯

NLP Tasks

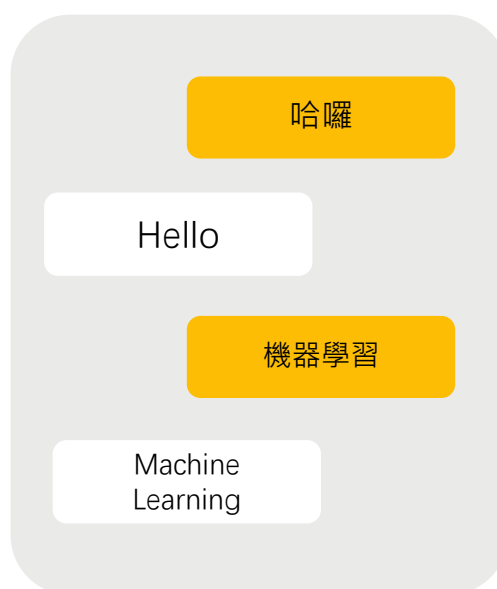
Large Language Models



文本分類



問答



翻譯



NLP Tasks

Large Language Models

 OpenAI



文本分類

 OpenAI



問答

 OpenAI



翻譯



ChatGPT 之前的 GPT

模型大小

GPT-1
(2018)



117 M

訓練數據



700 本書

ChatGPT 之前的 GPT

模型大小

GPT-1
(2018)



117 M

GPT-2
(2019)



1,542 M

訓練數據



700 本書



檔案大小
40 GB

ChatGPT 之前的 GPT

模型大小

GPT-1
(2018)



117 M

GPT-2
(2019)



1,542 M

GPT-3
(2020)

175 B

訓練數據



700 本書

檔案大小
40 GB

檔案大小：580 GB

Token 數：300 B

閱讀哈利波特全集 30 萬遍

ChatGPT 之前的 GPT

模型大小

 GPT-1
(2018) 
而這是你剛剛訓練的模型 ... 117 M

GPT-2
(2019)





1,542 M

GPT-3
(2020)

175 B

訓練數據

 
700 本書

檔案大小
40 GB

檔案大小：580 GB
Token 數：300B

閱讀哈利波特全集 30 萬遍

How to Use LLMs

Prompt Engineering

- 透過**指令 (prompt)** 設計優化模型
不進行任何模型參數訓練。

EX:

1. 前提講清楚
2. 提供範例
3. 問題拆解
4. 加入神奇的咒語

你是**台灣人**
NTU 是什麼的縮寫

你是**新加坡人**
NTU 是什麼的縮寫



台灣大學

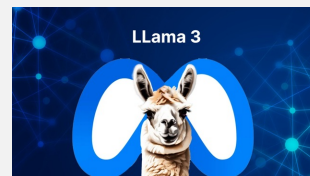


南洋理工大學

Fine-tuning (Transfer learning)

- 基於別人的模型，使用額外資料訓練
強化模型於特殊領域的能力。
- Ex: 法律、金融、醫療領域
- 需要準備
 1. 訓練資料
 2. 運算資源, GPU

Ex: TAIDE 台灣大型語言模型



[認識TAIDE](#)

How to Use LLMs

Prompt Engineering

- 透過**指令 (prompt)** 設計優化模型
不進行任何模型參數訓練。

EX:

1. 前提講清楚
2. 提供範例
3. 問題拆解
4. 加入神奇的咒語

你是**台灣人**
NTU 是什麼的縮寫

你是**新加坡人**
NTU 是什麼的縮寫



台灣大學

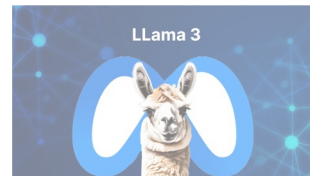


南洋理工大學

Fine-tuning (Transfer learning)

- 基於別人的模型，使用額外資料訓練
強化模型於特殊領域的能力。
- Ex: 法律、金融、醫療領域
- 需要準備
 1. 訓練資料
 2. 運算資源, GPU

Ex: TAIDE 台灣大型語言模型



[認識TAIDE](#)

How to Use LLMs

開源 可以 Local 跑的模式

- 公開模型架構與權重，
提供開發者下載、Fine-tune
- 需考量
 1. 模型大小
 2. 語言
 3. 電費成本
 4. 可否商用

模型 Models

Llama 3
T5
Bloom
Mixtral

支援中文

[中文 LLMs 整理](#)

ChatGLM
Qwen
Breeze
TAIDE

平台 Models Hub



Hugging Face Hub
開源商用模型庫



Ollama Models
量化開源模型庫
支持 CPU 運算

閉源 只能 call API

- 沒有提供模型架構等細節
開發者藉由官方 API 做模型調用
- 需考量
 1. API 價錢
 2. 數據隱私
 3. 可否商用

模型 Models

2024/04 資料 · 僅供參考

names	公司	上下文長度	價錢
gpt-4-tubo	OpenAI	32K	\$10 /per
Gemini-1.5-pro	Google	1M	\$7 /per
Claude 3	Anthropic	200K	\$0.25 /per
Command R+	Cohere	128K	\$3 /per

Retrieval Augmented-Generation 檢索增強式生成

為什麼需要 RAG ?

LLMs 擅長

- 理解文字
- 遵循指令

為什麼需要 RAG ?

LLMs 擅長

- 理解文字
- 遵循指令

但 LLMs 並不擅長 ...

- 不知道最新 & 你的資料
- 針對專有領域常生成空泛內容，充滿幻覺
- 提供資料來源

LLM 就像是期末考 裸考 的你 ...

題目 (Query) : 為什麼 ML 需要正規化 ?

不知道答案的你 (LLMs) , 可能會

1. 拒絕回答
2. 想辦法給一個 推 (蝦) 論 (掰) 出的答案

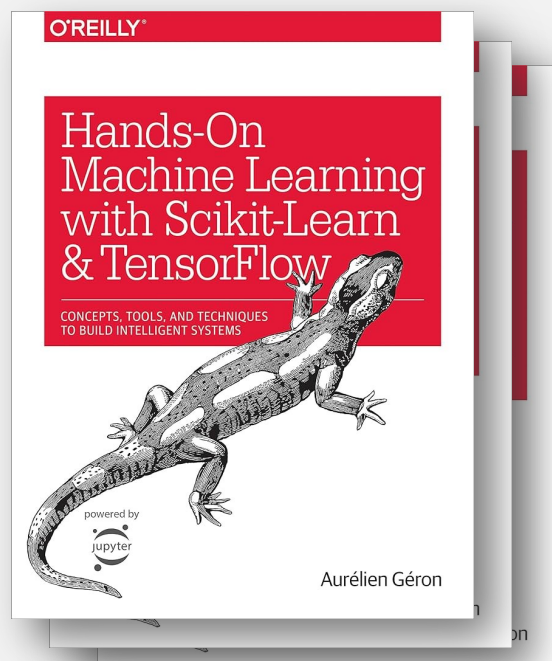


而 RAG 則像是讓你考試 open book

Query

Query: 為什麼 ML 需要正規化？

Open Book :



Retrieval 找到相關章節

Query: 為什麼 ML 需要正規化？

Context

Chapter 1

- 1. One-Hot Encoding
- 2. Dropout
- 3. Regularization**

- 4. Missing Data

Chapter 2

- 1. CNN
- 2. RNN
- 3. LSTM
- 4. Transformer

語音搜尋
檢索答案

Augmented- Generation 基於資料回答

Query: 為什麼 ML 需要正規化？

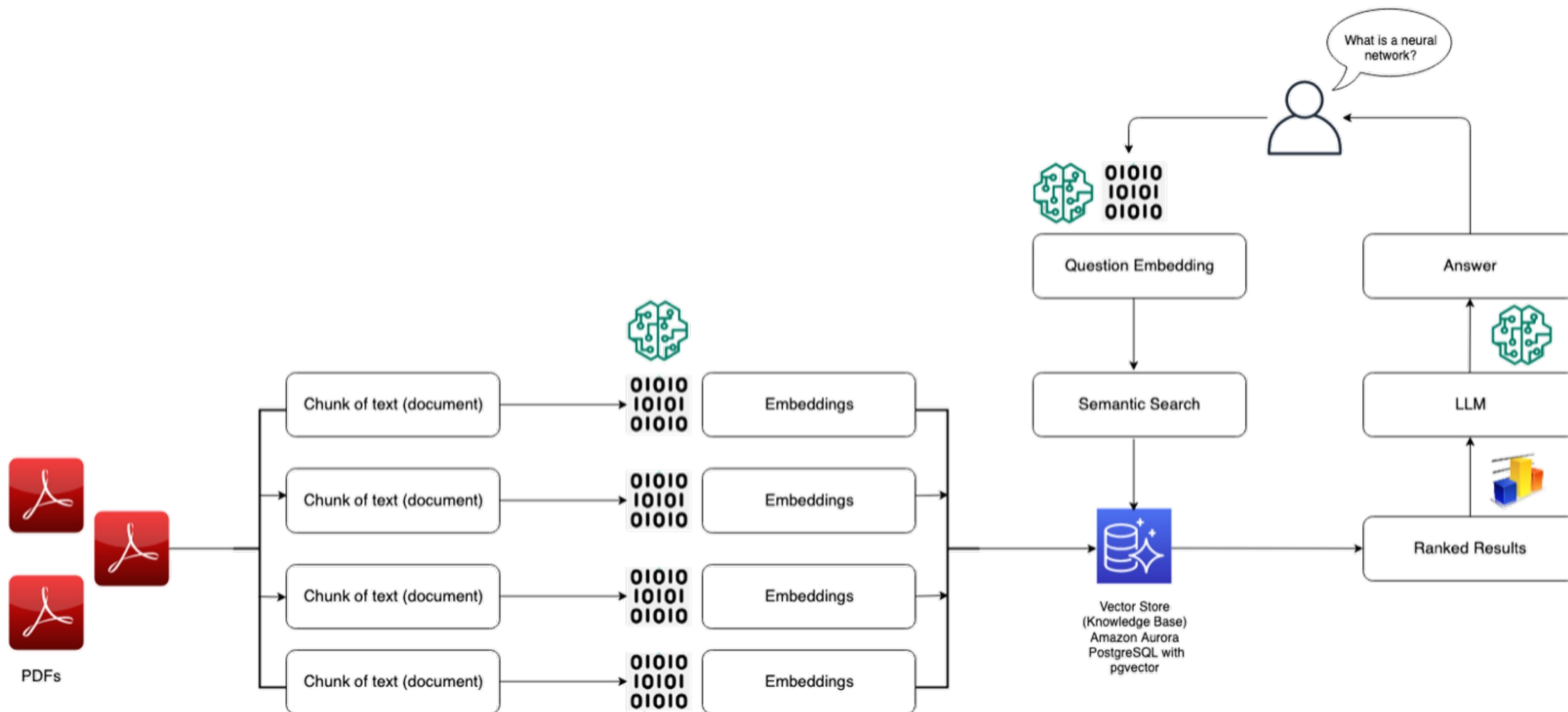
+ Retrieval Information:

1.3 Regularization:

XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXX

Answer: 防止 Overfitting

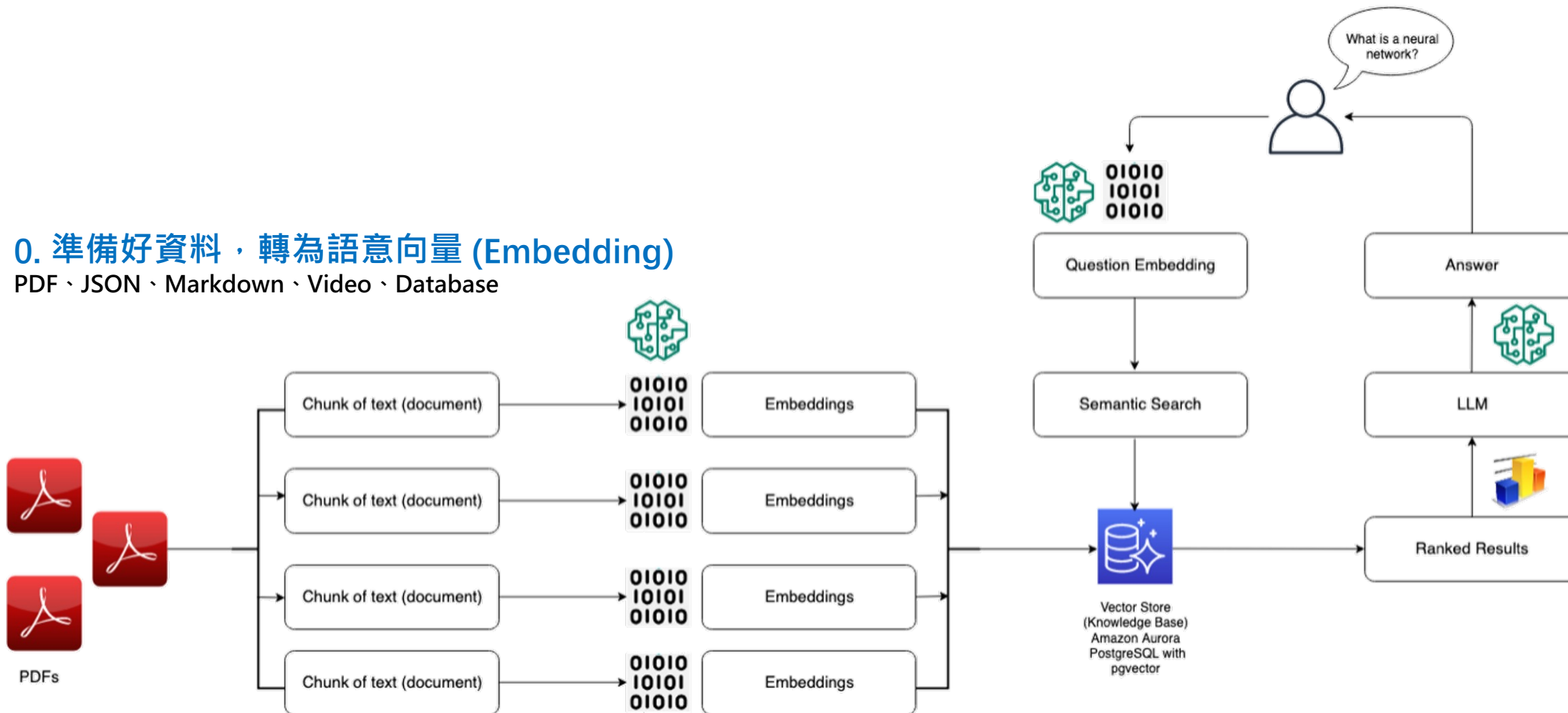
RAG 架構



RAG 架構

0. 準備好資料，轉為語意向量 (Embedding)

PDF、JSON、Markdown、Video、Database



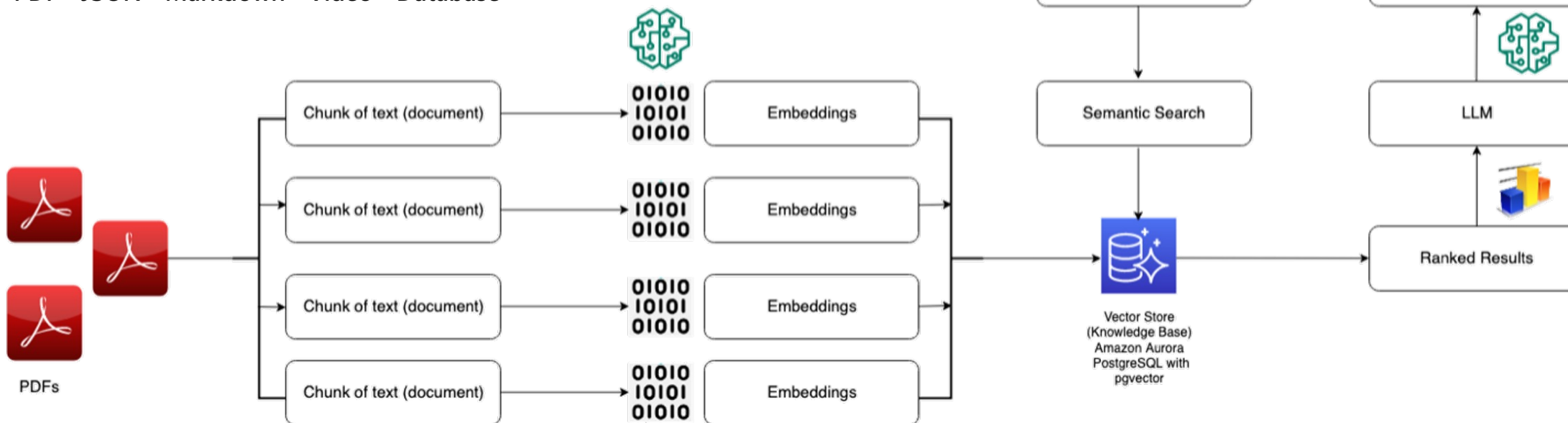
RAG 架構

1. User 問問題

Ex: What's a neural network

0. 準備好資料，轉為語意向量 (Embedding)

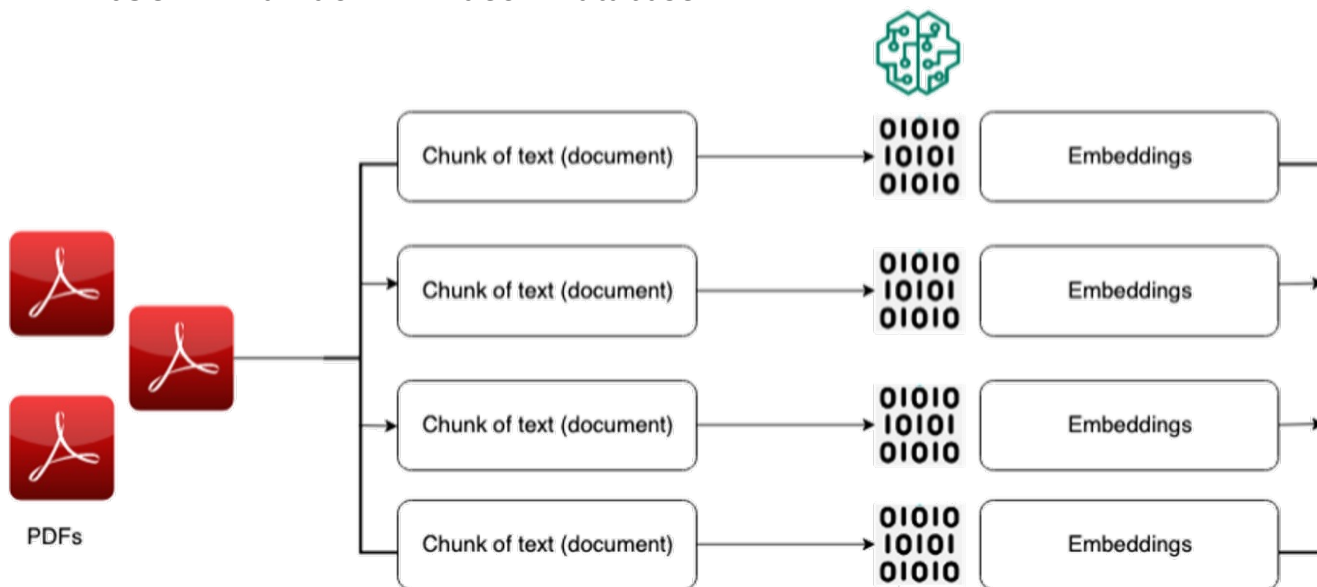
PDF、JSON、Markdown、Video、Database



RAG 架構

0. 準備好資料，轉為語意向量 (Embedding)

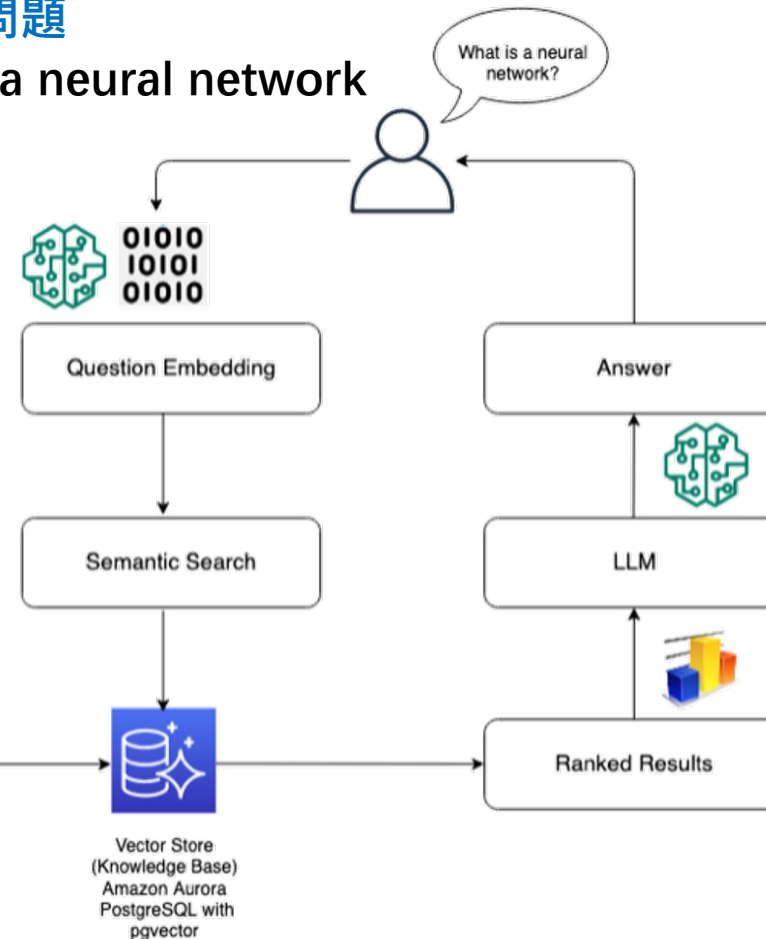
PDF、JSON、Markdown、Video、Database



1. User 問問題

Ex: What's a neural network

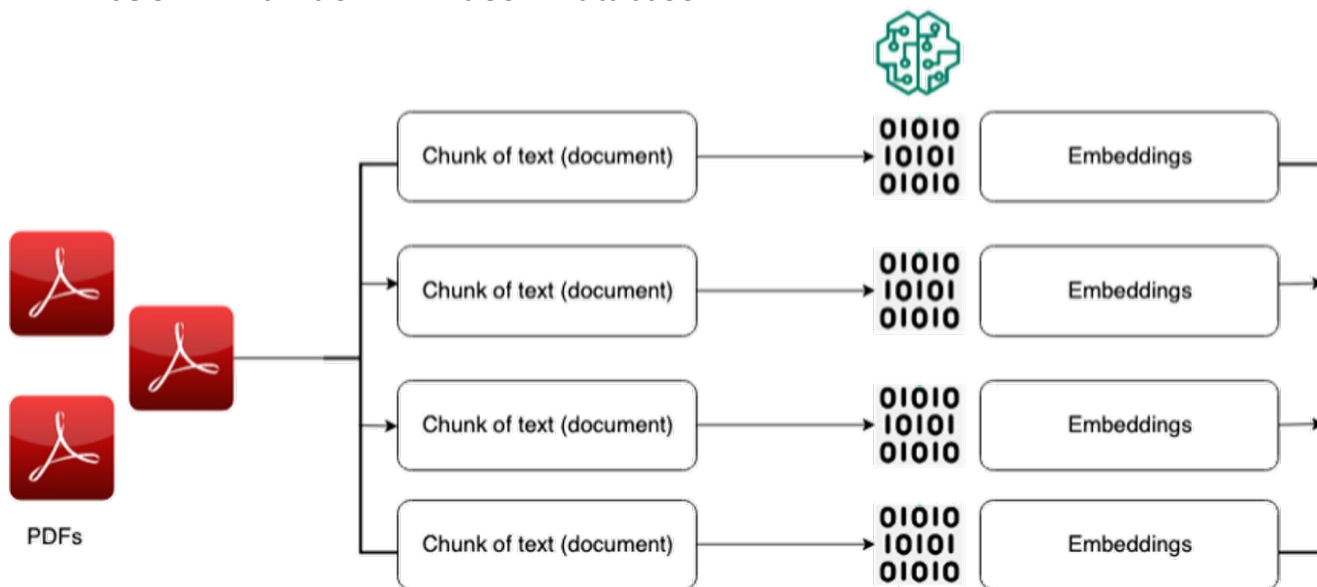
2. 將問題也轉成向量



RAG 架構

0. 準備好資料，轉為語意向量 (Embedding)

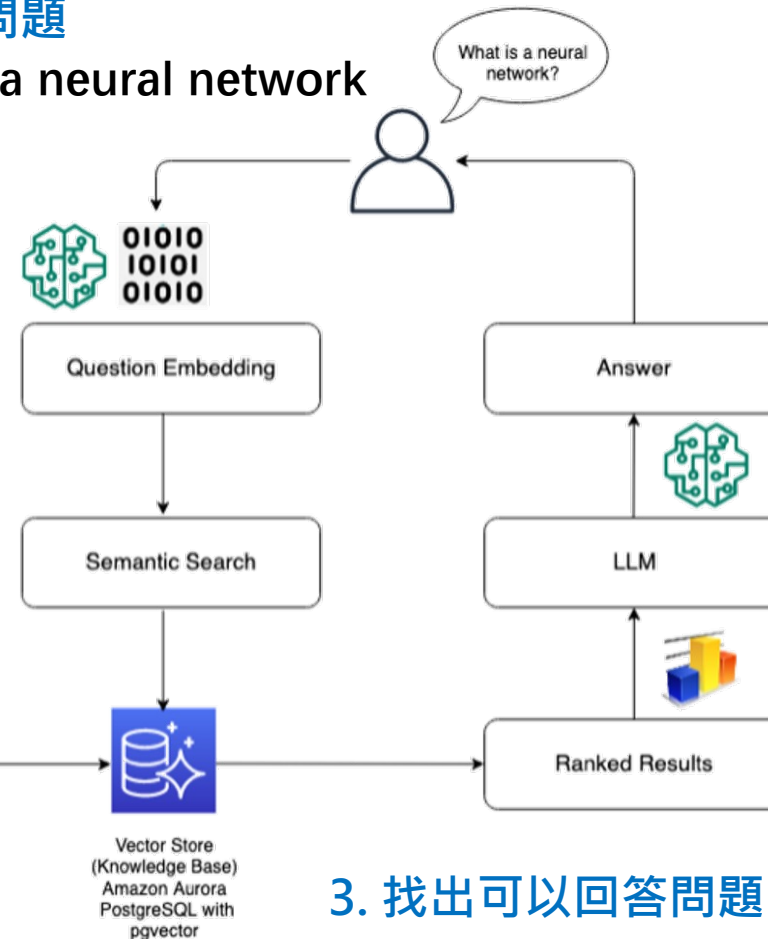
PDF、JSON、Markdown、Video、Database



1. User 問問題

Ex: What's a neural network

2. 將問題也轉成向量

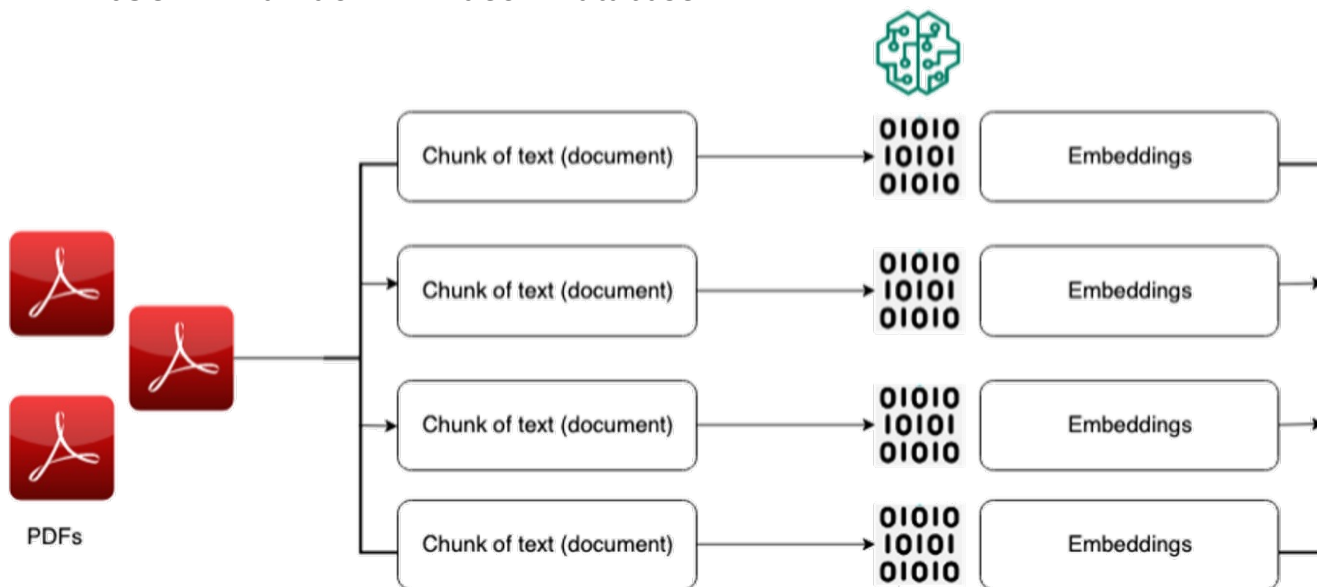


3. 找出可以回答問題的段落

RAG 架構

0. 準備好資料，轉為語意向量 (Embedding)

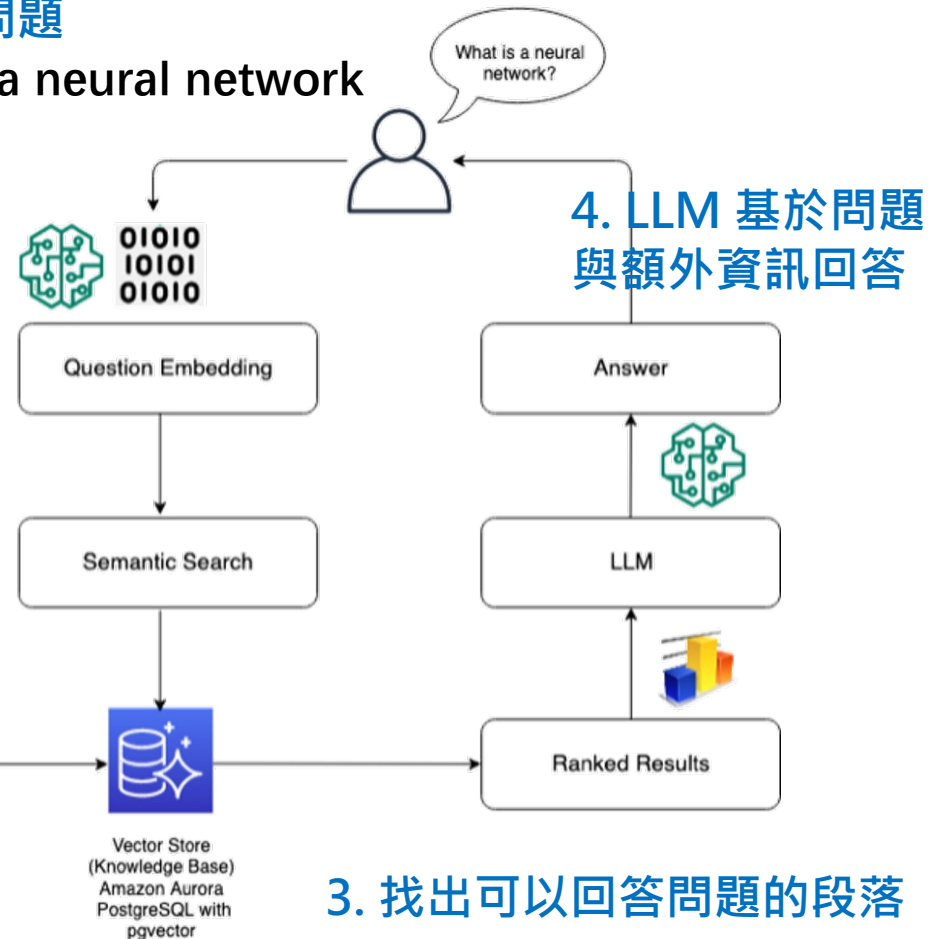
PDF、JSON、Markdown、Video、Database



1. User 問問題

Ex: What's a neural network

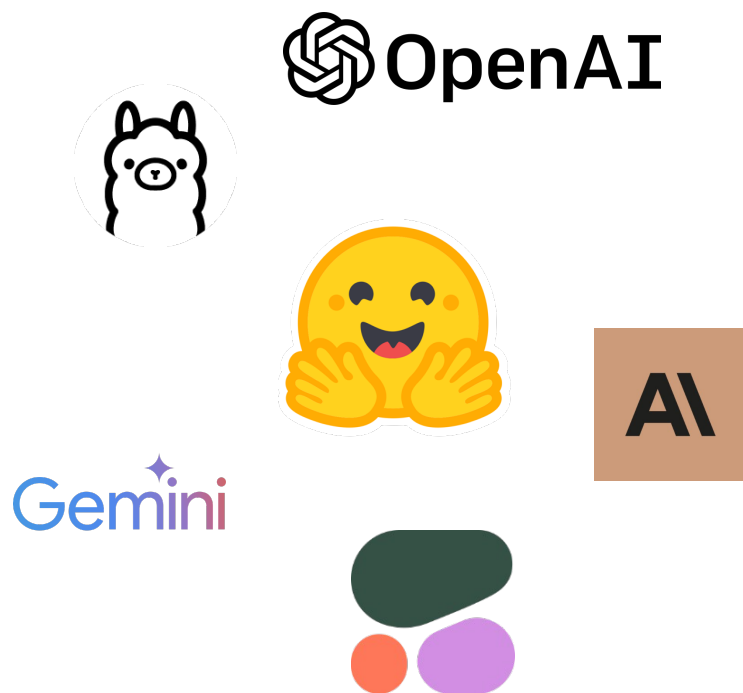
2. 將問題也轉成向量



3. 找出可以回答問題的段落

實作 RAG 會需要的工具

Models



RAG Frameworks



Introduction to Hugging Face

