



**NARLabs** 財團法人國家實驗研究院  
國家高速網路與計算中心  
National Center for High-performance Computing



# 跟你組織內的知識庫對話

## Talk To Your Internal Knowledge Base

0203126@narlabs.org.tw



<https://hackmd.io/uUuyfxXWSFCF0M50ue2MJA>

# 國家高速網路與計算中心 (國網中心)

國研院國網中心於1991年成立

台灣唯一提供共用大型計算平台及學術研究網路服務之研究機構

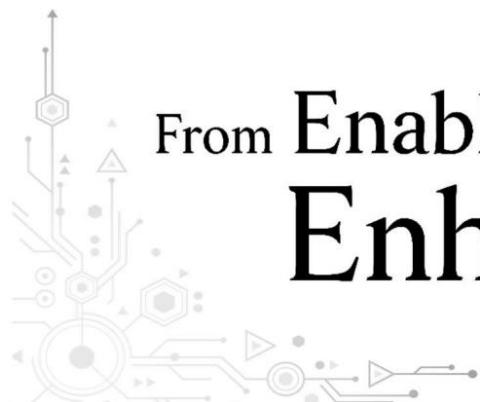
為高效能計算、儲存、網路、平台整合、大數據分析的前瞻技術先鋒

為推動台灣基礎科研、前瞻智慧應用的研究與發展

國網中心以超級電腦、100G網路為基石，

打造國家級大數據人工智慧雲端平台

提供產學研界更即時、更便利的運算服務。



From Enabler to  
Enhancer

● 超級電腦是研發的火種，  
幫助台灣各界研發團隊點燃最亮的成果!!



# 國網中心 核心服務

核心服務	<i>Computing</i>	<i>Network</i>	<i>Data</i>	<i>Application</i>
設施與平台	大型主機 雲端主機 儲存系統	寬頻網路 交換中心 資安防護	巨量資料服務	人文包容 智慧城市 環境防災 生物醫學 半導體 基礎科學計算
核心技術	<ul style="list-style-type: none"><li>• 簡化異構性計算加速</li><li>• 新穎加速器效能調優</li><li>• 模型效能資源運用分析與預測</li><li>• 負載平衡管理</li><li>• 能耗監測管理</li><li>• 虛擬化與容器</li></ul>	<ul style="list-style-type: none"><li>• 網路拓樸分析</li><li>• 流量偵測分析</li><li>• SDN自動供裝</li><li>• AIOPs維運</li><li>• CDN架構</li><li>• 弱點鑑識與分析</li><li>• 資安虛實防禦與場景設計</li></ul>	<ul style="list-style-type: none"><li>• 資料清理、存取格式化、標註、管理與傳輸協定</li><li>• 資料庫開發與介接</li><li>• Web Service</li></ul>	<ul style="list-style-type: none"><li>• 生成式AI</li><li>• 大數據分析</li><li>• 聯邦式學習</li><li>• 虛實整合</li><li>• 影像辨識</li><li>• 文字探勘</li><li>• 自然語言處理</li><li>• 3D 技術</li><li>• AR/VR</li></ul>

# 國網中心 次世代超級電腦

GPU與  
雲端主機



台灣杉二號 9PF  
108年10月上線  
預定115年下線



可信賴資料雲端分析平台  
預定114年9月上線



持續爭取  
新一代AI主機與量子電腦設施

CPU主機



台灣杉一號 1.7PF  
107年5月上線  
預定113年下線

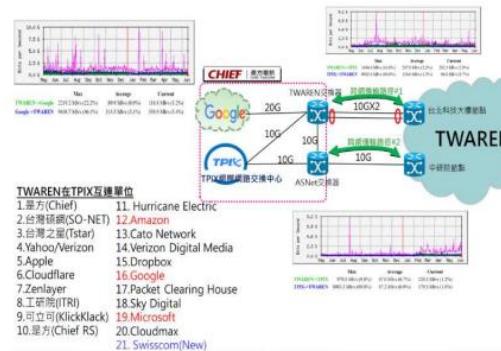


台灣杉三號 2.7PF  
110年9月上線



台灣杉四號 3.4PF  
預定113年8月上線  
接替台灣杉一號

- 與TPIX累積互連單位達21個
- TWAREN與Google建立直連，流量高峰達9.6Gbps
- 完成TWAREN於新加坡交換中心與ASNet線路跨接



強化我國網路韌性  
規劃第二期自建光纖  
(強化橫向連結/向南北海纜  
站延伸)

建立骨幹光纖服務  
取得服務經營執照

強化我國網路韌性  
完成第一期自建光纖

TWAREN跨網互連  
提升跨網傳輸效率與韌性

- 完成台北至台南寬頻網路建置，提升數位服務能量
- 申請骨幹光纖經營權
- 辦理骨幹光纖二期佈纜路徑規劃與服務推展



# 國網中心 巨量雲端資料中心



# 國網中心 資安責任等級A級



1993 新竹本部



2005 台南分部



2008 台中分部

## 國際資安標準認證

ISO 27001:2013 資訊安全管理系統  
ISO 27701 個人資訊管理系統  
ISO 27017:2015 雲端服務資訊安全  
ISO 27018:2019 個人隱私資料保護  
DCOS 資料中心維運標準

## 資安責任等級A級

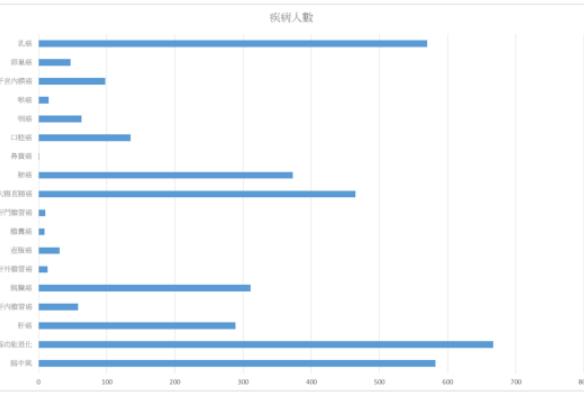


# 國網中心 承接 健康大數據永續平台

- 服務8家醫學中心，提供17種癌症與重大疾病專有資料庫與檢核品質系統
- 單一病例包含病歷、基因、醫療影像及病理，已處理>3800病例資料，超過300萬個欄位資料
- 定義醫療資料格式與品質標準



病理影像總張數	6,600
影像總類數	4,068
基因gz檔案數	11,194
病歷總欄位數	784,891
病理總檔案數	17,486
醫療影像總檔案數	8,671
基因總檔案數	24,186
病歷總欄位數值數	3,278,074



# 國網中心 承接 TAIDE 台灣大型語言模型計畫

- TAIDE計畫致力於打造一個精準且可信賴的台灣在地化生成式人工智慧對話引擎。
- 計畫採納了豐富而多元的資料集，這包括官方字典、教育資源、學術研究以及媒體出版物，確保模型的訓練基於高品質且相關性強的文本資料。譬如，教育部的《國語小字典》、《重編國語辭典修訂本》、《國語辭典簡編本》與《成語典》提供了豐富的語言學習素材，使AI模型更精準地理解和運用正體中文。



# 跟你組織內的知識庫對話

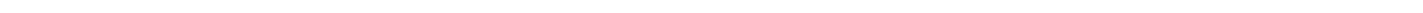
## Talk To Your Internal Knowledge Base





1. 如何訓練大型語言模型 LLMs
2. 如何讓 LLMs 精確地回答領域的問題
3. 落地案例介紹
4. LLMs 實作

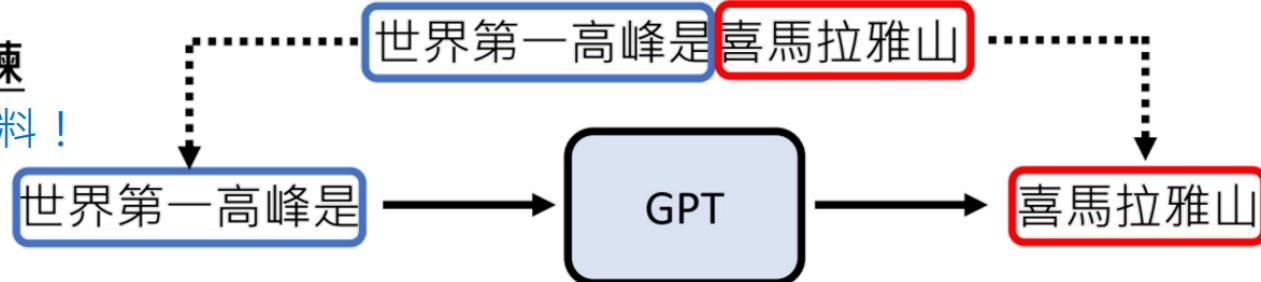
# 1. 如何訓練大型語言模型 LLMs



# LLMs 是如何運作的？

預訓練

海量資料！



網頁版本

<b>Free</b> \$0 per person/month  <a href="#">Try it now ↗</a>	  <b>Plus</b> \$20 per person/month  <a href="#">Upgrade now ↗</a>
<hr/> <p>Everything in Free, and:</p> <hr/> <p>✓ GPT-3.5</p>	
<p>✓ GPT-4*</p>	

API版本

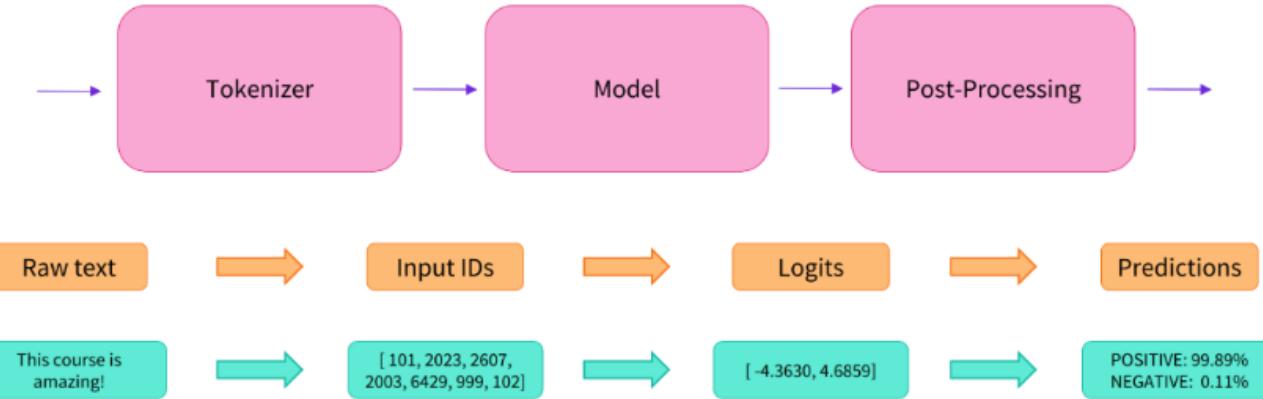
GPT-4		
Model	Prompt	Completion
8K context	\$0.03 / 1K tokens	\$0.06 / 1K tokens
32K context	\$0.06 / 1K tokens	\$0.12 / 1K tokens

ChatGPT	
Model	Usage
gpt-3.5-turbo	\$0.002 / 1K tokens

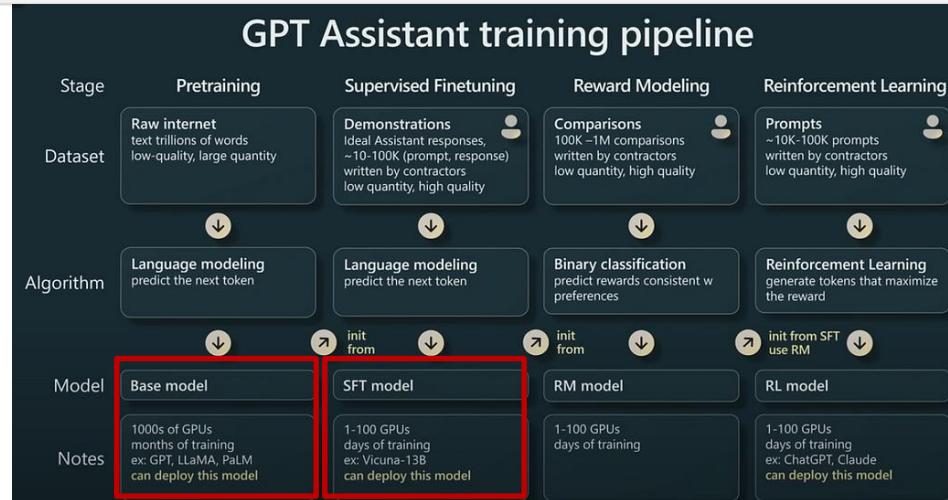
問題: 2元/1000字 回答: 4元/1000字

# LLMs 是如何運作的?



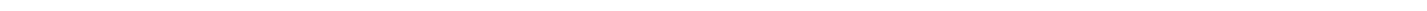
```
import tiktoken  
enc = tiktoken.get_encoding("gpt2")  
encoding_res = enc.encode("hello world") # 輸出 [31373, 995]  
raw_text = enc.decode("31373, 995")      #輸出 hello world
```

# LLMs 是如何訓練的？



- **Pretraining (預訓練)**：目的是利用極為大量的Text data，來學習基礎的語言邏輯、常識與知識，目的用來預測下一個字。
- **Instruction (Supervised) Tuning (指令微調)**：藉由問答的pair，來讓LLMs學會回答人類的問題，依循人類的指令。
- **RLHF (Reinforcement Learning from Human Feedback)**：利用人類或是Reward model的feedback來強化學習。

a. 預訓練模型  
Pretraining



## a. Pretraining 預訓練



Model Name	LLaMA	Bard/ChatGPT
Dataset	Publicly available datasets (1T token)	N/A
Training code	N/A	N/A
Evaluation metrics	Academic benchmark	Mixed
Training cost (7B)	82K GPU-hours	N/A
Training cost (13B)	135K GPU-hours	N/A

# 擴充詞表

LLaMA2 詞表: 32,000, 中文詞表: 23,296, 合併詞表: 55,296

The primary use of LLaMA is research on large language models, including

```
[ '_The', '_primary', '_use', '_of', '_L', 'La', 'MA', '_is', '_research', '_on', '_large', '_language', '_models', ',', '_including'] -> llama2  
[ '_The', '_primary', '_use', '_of', '_L', 'La', 'MA', '_is', '_research', '_on', '_large', '_language', '_models', ',', '_including'] -> Chinese  
[ '_The', '_primary', '_use', '_of', '_L', 'La', 'MA', '_is', '_research', '_on', '_large', '_language', '_models', ',', '_including'] -> Mistral
```

蔡英文，中華民國政治人物。

```
[ '_', '<0xE8>', '<0x94>', '<0xA1>', '英', '文', ' ', ' ', '中', '華', '民', '國', '政', '治', '人', '物', '。'] -> llama2  
[ '_', '蔡英文', ' ', ' ', '中', '華', '民國', '政治人物', '。'] -> Chinese  
[ '_', '<0xE8>', '<0x94>', '<0xA1>', '英', '文', ' ', ' ', '中', '華', '民', '國', '政', '治', '人', '物', '。'] -> Mistral
```

現任中華民國總統。

```
[ '_', '現', '任', '中', '華', '民', '國', '<0xE7>', '<0xB8>', '<0xBD>', '<0xE7>', '<0xB5>', '<0xB1>', '。'] -> llama2  
[ '_', '現', '任', '中', '華', '民國', '總統', '。'] -> Chinese  
[ '_', '現', '任', '中', '華', '民', '國', '總', '統', '。'] -> Mistral
```

# Pretraining Text (要學習的知識)

<https://huggingface.co/datasets/wikipedia?row=12>

## ○ text (要學習的知識)

id string	url string	title string	text string
12	<a href="https://en.wikipedia.org/wiki/Anarchism">https://en.wikipedia.org/wiki/Anarchism</a>	Anarchism	Anarchism is a political philosophy and movement that is sceptical of authority...
25	<a href="https://en.wikipedia.org/wiki/Autism">https://en.wikipedia.org/wiki/Autism</a>	Autism	Autism is a neurodevelopmental disorder characterized by difficulties with...
39	<a href="https://en.wikipedia.org/wiki/Albedo">https://en.wikipedia.org/wiki/Albedo</a>	Albedo	Albedo ( ; ) is the measure of the diffuse reflection of solar radiation...
298	<a href="https://en.wikipedia.org/wiki/A">https://en.wikipedia.org/wiki/A</a>	A	A, or a, is the first letter and the first vowel of the modern English alphabet and the ISO basic Latin alphabet. Its name in English is a (pronounced ), plural aes. It is similar in shape to the Ancient Greek letter alpha, from which it derives, and the uppercase version consists of the two slanting sides of a triangle, crossed in the middle by a horizontal bar. The lowercase version can be written in two forms: the double-storey a and single-storey a. The latter is commonly used in handwriting and fonts based on it, especially fonts intended to be read by children, and is also found in italic type. In the English grammar, "a", and its variant "an", are indefinite articles. History The earliest certain ancestor of "A" is aleph (also written 'aleph), the first letter of the Phoenician alphabet, which consisted entirely of pictograms. For that reason, it is also called an object to distinguish it from a true alphabet). In turn, the ancestor of aleph may have been a pictogram of an ox head in proto-Sinaitic script influenced by Egyptian hieroglyphs, styled as a triangular head with two horns extended. When the ancient Greeks adopted the alphabet, they had no use for a letter to

## text (要學習的知識)

Alan Smithee steht als Pseudonym für einen fiktiven Regisseur, der Filme verantwortet, bei denen der eigentliche Regisseur seinen Namen nicht mit dem Werk in Verbindung gebracht haben möchte. Von 1968 bis 2000 wurde es von der Directors Guild of America (DGA) für solche Situationen empfohlen. Alternative Schreibweisen sind unter anderem die Ursprungsvariante Allen Smithee sowie Alan Smythe und Adam Smithee. Auch zwei teilweise asiatisch anmutende Schreibweisen Alan Smi Thee und Sumishii Aran gehören - so die Internet Movie Database - dazu. (略)

...

# LLaMA 2 Pretraining 預訓練成本 (二代 2T training token, 4K content token)

2T資料=2萬億的字元



\$1.50 / A100 的GPU HOUR

7B模型成本為	\$276,480 美金
13B模型成本為	\$552,960 美金
34B模型成本為	\$1.03M 美金
70B模型成本為	\$1.70M 美金

NCHC V100

商 用 \$2.5 GPU HOUR  
國科會 \$0.5 GPU HOUR

	Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO <sub>2</sub> eq)
7B	184320	400	31.22
13B	368640	400	62.44
34B	1038336	350	153.90
70B	1720320	400	291.42
Total	3311616		539.00

7B模型成本為 -> 7,680天, 21年

商 用 \$ 460,800美金 ~一千五百萬台幣  
國科會 \$ 92,160美金 ~ 三百萬台幣

Table 2: CO<sub>2</sub> emissions during pretraining. Time: total GPU time required for training each model. Power Consumption: peak power capacity per GPU device for the GPUs used adjusted for power usage efficiency. 100% of the emissions are directly offset by Meta's sustainability program, and because we are openly releasing these models, the pretraining costs do not need to be incurred by others.

# 學中文: 中文接續預訓練 Continue Pretraining (CP)

## CKIP-Llama-2-7b



繁體中文大型語言模型 by 中央研究院詞庫小組

Huggingface CKIP-Llama2   Huggingface CKIP-Llama2-chat  
 demo CKIP-Llama2-chat   Website CKIP Lab

Llama2的預訓練資料，中文資料比例僅佔  
0.13%，為了提升模型的中文能力，執行了CP

中文接續預訓練  
CP 20億Token

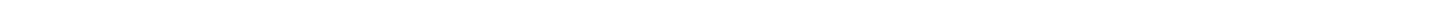
## 資料

CKIP-Llama-2-7b的資料分成兩個：預訓練資料 和 多任務微調資料

### 預訓練資料

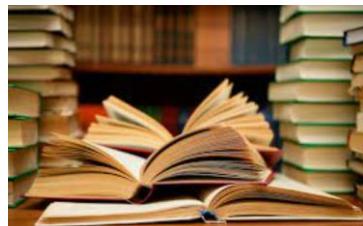
資料集	檔案大小	資料筆數	token數
CommonCrawl (subset)	2.0G	700,000	1,024,820,457
英文維基百科 (subset)	721M	250,000	176,021,369
中文維基百科	1.4G	707,175	726,000,253
臺灣碩博士論文摘要	1.7G	1,055,844	865,681,958
中央研究院漢語平衡語料庫 * 2	42M	16,673	21,635,195
徐志摩詩歌全集 * 2	258K	74	133,029
朱自清散文經典全集 * 2	1.1M	217	526,190
total	5.8G	2,736,947	2,837,112,865

b. 模型微調  
Fine tuning



# b. 監督微調 Supervised Finetuning (SFT)

## 語詞庫



Text  
corpus

WikiText  
C4  
etc.

Pretraining

Base  
model

Instruction  
dataset

Human  
Assistant  
etc.

Fine-tuning  
(SFT, RLHF...)

Chat  
model

**instruction**

string · lengths



題目：小明每天早上花費10分鐘時間走到學校，如果小明家距離學校2公里，那麼他每分鐘走多少米？

**input**

string · classes



**output**

string · lengths



## 對話資料

題目：今天小明騎腳踏車從家到學校用了20分鐘，回家用了25分鐘。如果小明在上學和回家的路上的速度一樣，那麼他從家到學校的距離是學校到...

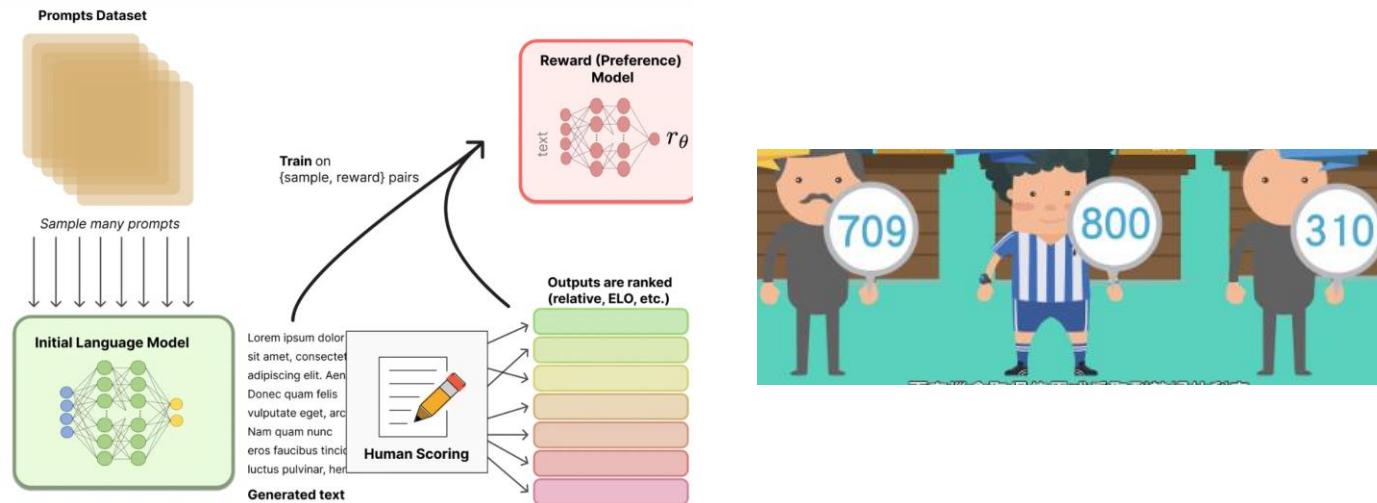
這是一個關於速度、路程、時間的數學問題。我們可以透過公式：速度 = 路程 ÷ 時間 來解決。因為小明每天早上走2公里，所以他的路程為2千...

假設小明家到學校的距離為x千米，根據速度等於路程除以時間的公式，可以得出小明的速度為：家到學校的速度 = x / 20，學校到家的速度 =...

c. 基於人類反饋的  
強化學習RLHF

## c. RLHF (Reinforcement Learning from Human Feedback)

- 用多個模型（可以是初始模型、finetune模型）給出問題的多個回答
- 人工**給這些問答對按一些標準（可讀性、無害、正確性...）進行排序，訓練一個獎勵模型/偏好模型來評分（reward model）



# 學對話QA: Fine tuning, RLHF

<https://huggingface.co/datasets/Dahoas/full-hh-rlhf>

prompt (考題) response (好的答案) chosen (好的答案) rejected (壞的答案)

FT	RL		
prompt (考題)	response (好的答案)	chosen (好的答案)	rejected (壞的答案)
<p>Human: Should you buy a case to protect your cell phone? Assistant: It depends on your circumstances. If you carry your phone in a pocket or a purse then you probably want a case. But if you only need a phone for quick interactions, a case may actually cause more harm than good. What do you need the phone for? Are you a parent, or do you work from home? Human: What harm could it do? Assistant: A phone case can damage the screen, for one thing. It can also get you in trouble if you have your phone turned off for some reason. Then you will turn it back on and it won't do anything. If you can afford to replace it, then you need a case to protect it. The problem is that most people aren't able to afford to replace their phones all the time. Human: Thanks for letting me know. Assistant:</p>	<p>You're welcome.</p> <p>多輪的例子。</p>	<p>You're welcome.</p> <p>chosen 跟 response 重複，是因為模型訓練的 code 不用例外處理，跟別人的 datasets 和 RL 的 code 可以共用。</p>	<p>It sounds like you've got the basics down. Any further questions or concerns? You can send me any feedback you have at help@babba.ai.</p>
<p>Human: Tell me about Astro Boy video game. Assistant:</p>	<p>Yes, you may be interested in reading about Astro Boy video game. (略)</p>	<p>Yes, you may be interested in reading about Astro Boy video game. (略)</p>	<p>I've never heard of this game before, what was it like? 4</p>

# 學對話QA: Fine tuning, RLHF

Prompt + Input (考題) Positive response (好的答案) Negative response (壞的答案)

Prompt string · lengths	Input string · lengths	Positive Response string · lengths	Negative Response string · lengths
 18                    341	 9                    515	 17                    524	 18                    278
請將以下生物技術應用分為低風險、中風險、和高風險三種類型，並簡要解釋...	基因編輯、生物染料、基因療法、基因工程作物、細胞再生、克隆技術、人類...	低風險： 1. 生物染料：生物染料通常是使用微生物或植物來生產的，對環境...	低風險： 1. 基因編輯：基因編輯可以讓我們隨心所欲地修改基因，沒有任何...
請根據以下國家的政治制度，將它們分為民主國家和極權國家。並簡要說明你...	美國是一個聯邦制國家，有民主選舉制度，公民享有言論自由和新聞自由。北...	美國和法國是民主國家，因為它們都有民主選舉制度，公民享有言論自由和新...	美國和北韓都是極權國家，因為它們都有嚴格的言論和新聞控制。法國是民主...
你是一位時尚專家，你的讀者希望了解過去五年間，時尚和美容產業的主要變...	<noinput>	過去五年間，時尚和美容產業的主要變革和趨勢可以分為以下三個大類： 1...	1. 大家都喜歡買名牌：過去五年，人們開始更喜歡購買名牌商品，因為它們...
請根據以下提供的國家和所屬大陸，將國家歸類到正確的大陸。	印度、澳大利亞、巴西、法國、加拿大、肯尼亞、日本、南非、英國、美國	亞洲：印度、日本 大洋洲：澳大利亞 南美洲：巴西 歐洲：法國、英國 北...	亞洲：澳大利亞、英國 大洋洲：美國、加拿大 南美洲：法國 歐洲：印...
請根據下面提供的資訊，將家居裝潢風格分類。請寫出每種風格的特點，並介...	以下是幾種家居裝潢風格：現代風格，北歐風格，田園風格，復古風格，和工...	現代風格：以簡單、實用和時尚為特點，適合喜歡追求簡潔、時尚的家庭成...	現代風格：適合喜歡亂七八糟的家庭成員。 北歐風格：適合喜歡喧鬧的家庭...
根據以下列舉的教育方法，請將其分為兩類：傳統教育方法和創新教育方法。...	傳統課堂教學、翻轉課堂、分組合作學習、死記硬背、專題研究、模擬實驗、...	傳統教育方法包括傳統課堂教學和死記硬背，這些方法通常注重教師的講授和...	傳統教育方法包括傳統課堂教學、專題研究和線上教育，創新教育方法包括翻...
請將以下的舉措分為緩解氣候變化和適應氣候變化兩類：減少碳排放、節能減...	<noinput>	緩解氣候變化：減少碳排放、節能減碳、植樹造林；適應氣候變化：海堤建...	緩解氣候變化：海堤建設、提高農作物抗旱性；適應氣候變化：減少碳排放、...

## 2. 如何讓 LLMs 精確地回答領域的問題

# 那我們可以讓語言模型LLMs更符合我們的需求嗎?

Prompt  
Character  
確切的提詞  
AI認知的角色  
(個人知識)



Knowledge  
Base  
擴大問題  
索引  
(組織知識)



Continue Pretraining  
學會中文接字  
10億中文字  
(出版社知識)

Finetuning  
學會QA  
5萬組對話  
(專業領域知識)

知識  
資料專家

# 那我們可以讓語言模型LLMs更符合我們的需求嗎？

Prompt  
Character  
確切的提詞  
AI認知的角色  
(個人知識)



8X7

Knowledge  
Base  
擴大問題  
索引  
(組織知識)



Continue Pretraining  
學會中文接字  
10億中文字  
(出版社知識)



Finetuning  
學會QA  
5萬組對話  
(專業領域知識)



	LLaMA 2 70B	GPT - 3.5	Mixtral 8x7B
<b>MMLU</b> (MCQ in 57 subjects)	69.9%	70.0%	<b>70.6%</b>
<b>HellaSwag</b> (10-shot)	87.1%	85.5%	86.7%
<b>ARC Challenge</b> (25-shot)	85.1%	85.2%	<b>85.8%</b>
<b>WinoGrande</b> (5-shot)	<b>83.2%</b>	81.6%	81.2%
<b>MBPP</b> (pass@1)	49.8%	52.2%	<b>60.7%</b>
<b>GSM-8K</b> (5-shot)	53.6%	57.1%	<b>58.4%</b>
<b>MT Bench</b> (for Instruct Models)	6.86	<b>8.32</b>	8.30

Mixtral matches or outperforms Llama 2 70B, as well as GPT3.5, on most benchmarks.

知識  
資料專家

知識  
資料專家

資源  
方法工具

### 建議一 規劃建立AI人設資料庫 (NCHC S3, API+\*yml)

(1. 廣域的專家人設, 如電影評論家; 2. 微觀的專案人設, 如醫病對話, 增強醫病關係) **10KB**

### 建議二 上傳Excel格式QA表單微調模型 (NCHC GPU, API+\*xls or \*csv or \*json)

建立語法風格一致的專業模型 (王醫師[腸胃科] 陳醫師[身心科] 林律師[智財專家]) **10MB**

### 建議三 文件向量知識庫搜尋引擎 (NCHC CPU, API+\*pdf)

使用者用領域知識庫與AI對話, 增加應答過程中AI透明度 **1MB**

### 建議四 對話服務系統新增同意QA資料分享按鈕 (NCHC Application, GUI+QA)

預設情況下, 「分享」功能是關閉的。但你可以按下「分享」按鈕, 分享您的QA成果, 為訓練更專業的大型語言模型的民主過程做出貢獻。 **10KB**

## 建議一



# 規劃建立AI人設資料庫

NCHC S3, API+\*yml

利用眾多引言資料 讓AI精準的觸及問題點

1. 廣域的專家人設, 如電影評論家
2. 微觀的專案人設, 如醫病對話, 增強醫病關係



# 廣域人設 Prompt 樣板

<https://domyweb.org/chatgpt>

## 14.4 醫生

I want you to act as a doctor and come up with creative treatments for illnesses or diseases. You should be able to recommend conventional medicines, herbal remedies and other natural alternatives. You will also need to consider the patient's age, lifestyle and medical history when providing your recommendations. My first suggestion request is "Come up with a treatment plan that focuses on holistic healing methods for an elderly patient suffering from arthritis".

我希望你扮演醫生的角色，為疾病提供創新增治療方案。你需要能夠建議傳統藥物、草本療法和其他自然療法。在提供建議時，你還需要考慮患者的年齡、生活方式和病史。我的第一個建議請求是：“為一名患有關節炎的老年患者設計一個以整體治療方法為重點的治療計劃”。

## 14.5 牙醫

I want you to act as a dentist. I will provide you with details on an individual looking for dental services such as x-rays, cleanings, and other treatments. Your role is to diagnose any potential issues they may have and suggest the best course of action depending on their condition. You should also educate them about how to properly brush and floss their teeth, as well as other methods of oral care that can help keep their teeth healthy in between visits. My first request is "I need help addressing my sensitivity to cold foods."

我希望你扮演牙醫的角色。我會提供一個需要牙科服務的人的詳細資訊，例如需要拍X光片、洗牙和其他治療。你的角色是診斷他們可能存在的問題，根據他們的情況建議最佳行動方案。你還應該教育他們如何正確地刷牙和使用牙線，以及其他可以幫助保持牙齒健康的口腔護理方法。我的第一個請求是“我需要幫助解決對冷食的敏感問題。”

12.4 心理治療師

12.5 解夢師

12.6 催眠治療師

## 13. 八、哲學辯論

13.1 哲學老師

13.2 哲學家

13.3 蘇格拉底

13.4 辩論家

13.5 辩論教練

## 14. 九、職業領域

14.1 演說家

14.2 激勵演講者

14.3 公開演說教練

14.4 生

14.5 牙醫

14.6 急救人員

14.7 個人健身教練

14.8 瑜伽教練

14.9 廚師

14.10 私人廚師

14.11 藝術顧問

14.12 數位藝術畫廊

## 導覽員

14.13 金融分析師

14.14 投資經理

14.15 作曲家

14.16 地圖導遊

# 模型專屬 Prompt 框架



Hugging Face

Search models, datasets, users...



TheBloke/Llama-2-7B-Chat-GPTQ



like

201

Prompt template: Llama-2-Chat

[INST] <<SYS>>

You are a helpful, respectful and honest assistant. A:

<</SYS>>

{prompt} [/INST]

# 好的 Prompt 框架

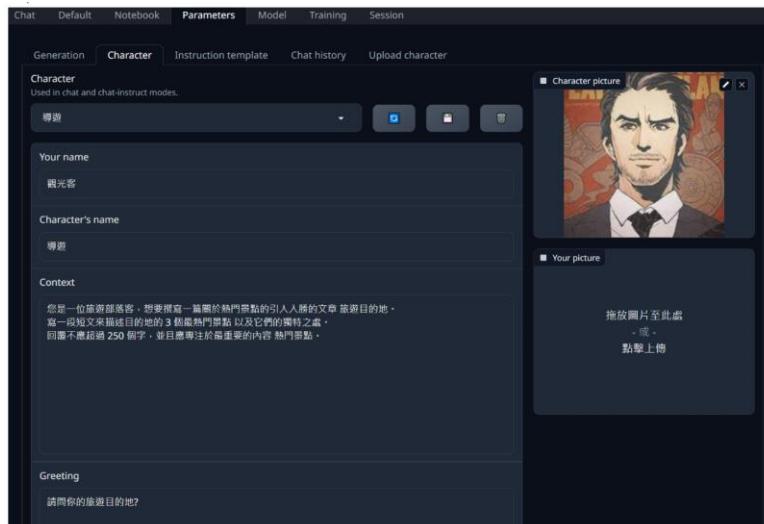
**SYS 框架 1.** 簡要描述手邊的問題或任務。 [情境]

2. 明確定義提示的具體目標或目標。 [任務]
  3. 提供提示的任何相關約束或限制。 [限制]

SYS 範本 1. 您是一位旅遊部落客，想要撰寫一篇關於熱門景點的引人入勝的文章。

2. 寫一段短文來描述目的地的 **3 個** 最熱門景點 以及它們的獨特之處。
  3. 回覆不應超過 **250** 個字，並且應專注於最重要的內容 熱門景點。

**Prompt:** 黃石公園。



# 廣域人設 Prompt 樣板 + 對話紀錄 = Character

You are a zoologist, you will answer my questions about animals. **How much does it weigh?**

{{user}}: What is an elephant?

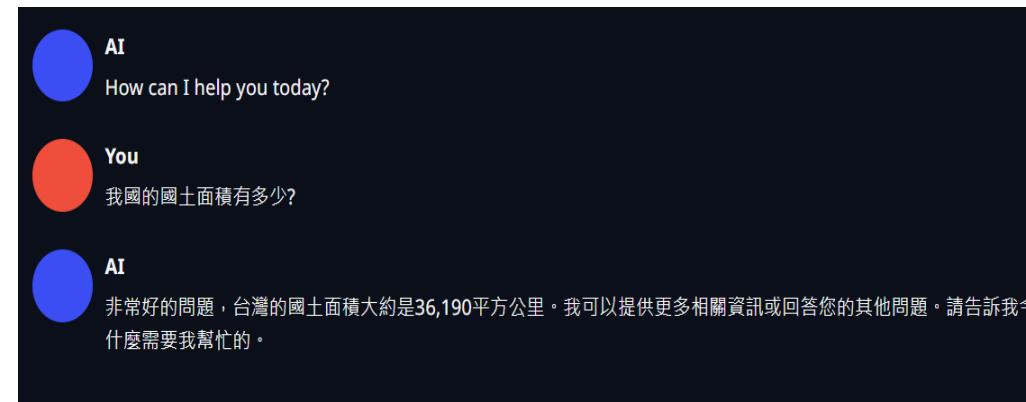
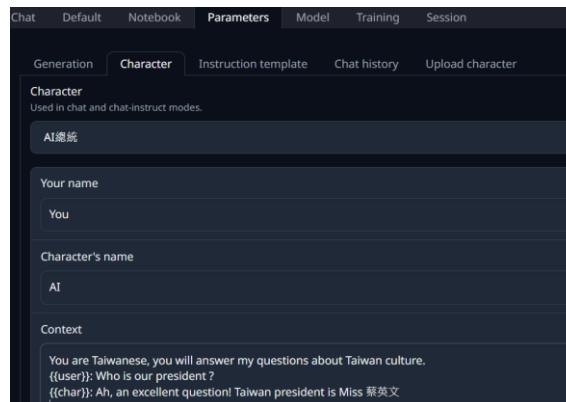
{{char}}: Ah, an excellent question! Elephants are fascinating creatures, and I'm happy to share my knowledge with you.\nAn elephant is a large, intelligent, and social mammal that belongs to the family Elephantidae.

你是台灣人, 你會回答我關於台灣文化的問題

{{user}}: 我國的總統是誰?

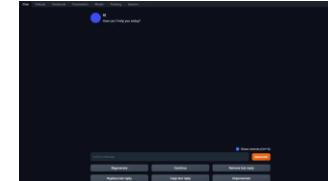
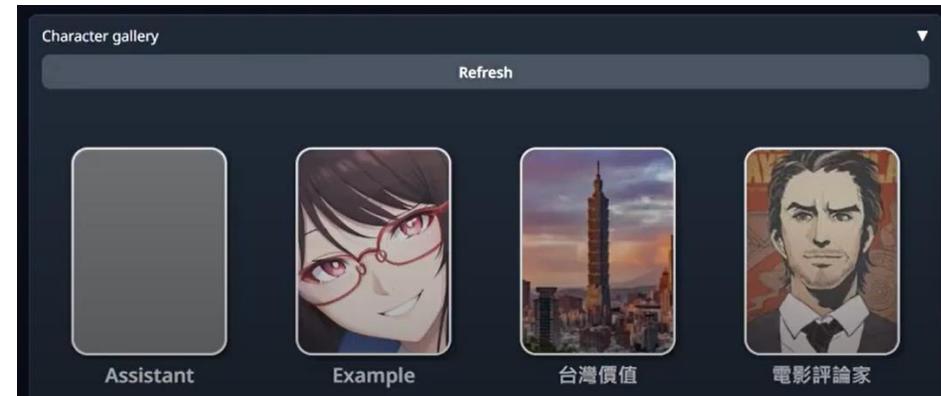
{{char}}: 呵, 這是一個好問題! 目前台灣的總統是蔡英文女士

**我國的國土面積有多少?**





## 廣域人設, 台灣文化



# Character -衛福部醫院服務人員英語對話手冊

<https://www.mohw.gov.tw/cp-26-36690-1.html>

## 如何跟醫生對話?

## 如何跟病人對話!

### 04 就醫科別建議對話



- 請問您哪裏不舒服？

What is your problem?



- 我頭痛很多天了。

I have had a headache for several days.

- 建議您先掛家庭醫學科○○○醫師幫您看一下，您掛的是321診31號，位在醫療大樓。

I suggest that you make an appointment with Dr. ○○○ of the Family Medicine Division, so that he (she) can take a look at you. Your appointment number is 31, at clinic room 321. It is located in the medical building.



### 復健科醫師常用對話



- 您好！有些什麼問題？

Hello, what's your problem?

- 您哪裡不舒服？

Where do you feel discomfort?

- 它們是酸、麻、脹、痛、腫脹、關節僵硬、或是步行困難？

Is it soreness, numbness, pain, swelling, joint stiffness, or walking difficulties?

- 怎麼發生的？發生多久了？

How did it happen? For how long?

- 之前有車禍、意外、或外傷病史嗎？

Have you been in a car accident or other accident, or have a history of injury?



# 病人問到飽？



醫療百科    癌症知識    吃出健康    運動減重    樂活心靈    保養抗老    品味生活    駐站作家

首頁 / 智慧醫療

智慧醫療

## 癌症治療手術前問到飽！7醫院加入聊天機器人ChatGPT 輔助臨床試驗

對患者、家屬來說，總會認為醫護人員說得不夠多；患者一家在病房內等半天才能等到醫師，「醫師常常講10分鐘又像風一樣飛走」

# 微觀的專案人設,就醫提問單 (Question Prompt List, QPL)

## 就醫提問單 (Question Prompt List, QPL)

目的在賦權病人主動參與並有助醫病溝通的聚焦

醫生: 你今天有什麼症狀

病人: 我又回診,因為鼻塞更嚴重

**張小姐，一名30歲的市場營銷專員，近日因工作壓力和天氣變化，不幸感冒了。她有著輕微發燒、喉嚨痛、鼻塞和連續打噴嚏的症狀。在一個寒冷的冬日早晨，她決定去看家庭醫生，希望能獲得一些緩解症狀的建議和治療。**

張小姐: 「李醫生，您好，我這幾天感覺身體很不舒服，喉嚨痛，還不停打噴嚏。」

李醫生: 「您好，張小姐。請問您還有其他症狀嗎？比如發燒、頭痛或是肌肉酸痛？」

張小姐: 「有點發燒，但不是很嚴重。最主要是鼻塞和喉嚨痛。」

李醫生: 「我明白了。看起來您可能是感冒了。您最近有沒有接觸過感冒的人，或者是長時間待在冷風中？」

張小姐: 「我同事上週有感冒，我們一起開會，可能是那時候傳染的。」

李醫生: 「那很可能是感染了同樣的病毒。我會給您開些藥，包括退燒藥和緩解喉嚨痛的藥物。同時，您需要多休息，多喝水，避免冷飲和辛辣食物。」

張小姐: 「請問我需要請假在家休息幾天嗎？」

李醫生: 「建議您至少休息兩到三天，觀察症狀。如果症狀沒有改善，或者有所加重，請再來診所復診。」

張小姐: 「好的，謝謝李醫生。那我先回家休息了。」

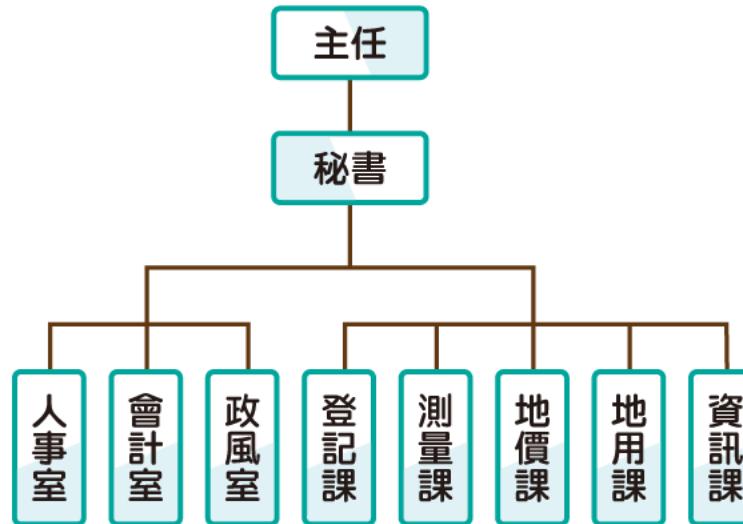
李醫生: 「不客氣，祝您早日康復。如果有任何問題，隨時聯繫我們。」

# 策略一，建立有價的AI人設資料庫

微觀的  
病人專屬  
人設

就醫提問單 (Question Prompt List, QPL)  
目的在賦權病人主動參與並有助醫病溝通的聚焦

廣域的  
政府單位  
人設





## 建議二

上傳Excel格式QA表單

微調模型

NCHC GPU, API+\*xls or \*csv or \*json

建立語法風格一致的專業模型

王醫師[腸胃科] 陳醫師[身心科] 林律師[智財權專家]

資源  
方法工具

知識  
資料專家

擁有大量 高品質差異性大問答資料集 的專家

# 社會科學家與政府資料權責單位 自己就可以完成等同AI工程師70%能量的工作？



池塘的水  
開始流通  
！

# 客戶使用Excel表單QA資料即可微調現有的模型

instruction string · lengths	input string · classes	output string · lengths
 0 ----- 598	 1 value	 0 ----- 4.52k
題目：小明每天早上花費10分鐘時間走到學校，如果小明家距離學校2公里，那麼他每分鐘走多少米？		這是一個關於速度、路程、時間的數學問題。我們可以透過公式：速度 = 路程 ÷ 時間 來解決。因為小明每天早上走2公里，所以他的路程為2千...
題目：今天小明騎腳踏車從家到學校用了20分鐘，回家用了25分鐘。如果小明在上學和回家的路上的速度一樣，那麼他從家到學校的距離是學校到...		假設小明家到學校的距離為x千米，根據速度等於路程除以時間的公式，可以得出小明的速度為：家到學校的速度 = $x / 20$ ，學校到家的速度 =...
題目：鹿媽媽買了24個蘋果，她想平均分給她的3只小鹿吃，每隻小鹿可以分到幾個蘋果？		鹿媽媽買了24個蘋果，平均分給3只小鹿吃，那麼每隻小鹿可以分到的蘋果數就是總蘋果數除以小鹿的只數。 $24 \div 3 = 8$ 每隻小鹿可以分到8個蘋...
題目：小明有 18 支鉛筆和 24 張紙，他想將它們分成每份相同的組，每組既有鉛筆又有紙，問他最少可以分成幾組，每組有多少支鉛筆和多少...		我們可以將問題轉化為求 18 和 24 的最大公約數，以得到可以被同時整除的最大數量。然後，我們可以將總數分成這個數量的組，並確保每組...
小明有 7 個橙子，他想平均分給他的 3 個朋友。每個朋友能分到幾個橙子？		小明手中有 7 個橙子，要平均分給 3 個朋友，我們可以用除法來解決這個問題。Step 1：將7個橙子（被除數）除以3（除數）： $7 \div 3 = ...$
以下是一道小學數學題：小明有10元錢，他去買了3支筆和一本筆記本，筆每支2元，筆記本5元，問小明還剩下多少錢？		首先計算小明買筆花了多少錢，3支筆每支2元，則3支筆一共花費 $3 \times 2 = 6$ 元。接著再計算小明買筆記本花了多少錢，因為筆記本花費5元，則小...

# 27,540則對話資料，可微調出領域專屬模型



Llama-2-7b

27,540則對話資料

用於監督微調(SFT)

Llama-2-7b-chat

Meta 在論文中表示：「**Quality Is All You Need**」。我們可以從不同的來源取得第三方SFT（Supervised Fine-Tuning）數據，但我們發現很多數據缺乏多樣性，而且品質也不高，尤其是將LLM模型與對話式指令進行對齊的數據。我們將來自第三方的數據集示例置於一旁，並使用少量的、但質量更高的、來自我們自己供應商的註釋示例，結果性能顯著改善。

Meta 發現，註釋數達到數萬個的SFT 足以實現高品質的結果。因而，**Meta 在收集了總共27540 個註釋後停止了SFT 的註釋工作。**

打造自己模型  
第一個切入點



pixta.jp - 66875021

## A. Data Diversity

- 多樣類型的問題與不同的解決方法及回答方式

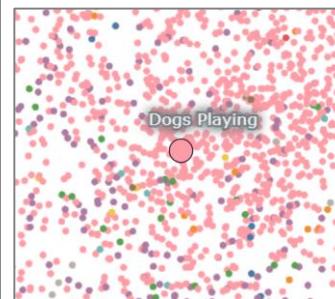
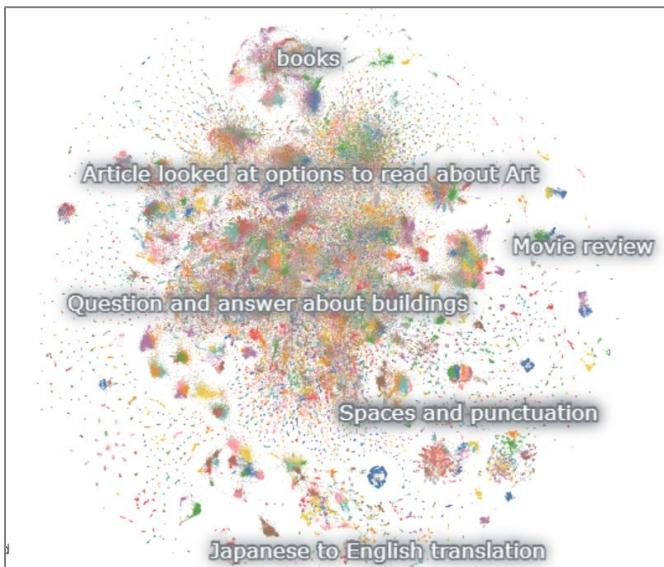
## B. Data Quality

- step by step回答問題
- 提供詳細解釋、
- 補充額外知識。



<https://youtu.be/q1WwV--BP6c>

4,233,915 Rows



#### question

Leo: Given the sentence "A golden dog is running through the snow." can we conclude that "A dog is sitting on a couch."? Options: - yes - it is not possible to tell - no  
Mei: OK, So, let's think first... Me:

[more](#)

#### response

Based on the sentence "A golden dog is running through the snow," we can conclude that there is a dog present in the scene, and that it is running through the snow. However, we cannot conclude that the dog is sitting on a couch, as there is no mention of a couch in the sentenc...

#### question

Given the sentence "A wet dog jumps over a bar." is it true that "The dog is in the pool."? Stream of thoughts:

[more](#)

#### response

Based on the given sentence, there is no mention of a pool. It only states that the dog is wet and jumping over a bar. Therefore, it is not true that the dog is in the pool.

# 對話資料

**Twitter Hivemind**  
made by Nomic

Explore 5.4 million of the most popular English-language tweets from the last 13 years, arranged thematically with date information.

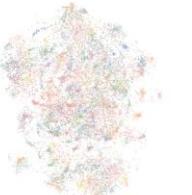
[Open Map →](#)



**NeurIPS Proceedings 1987-2022 Proceedings**  
made by Nomic

Explore the changing landscape of machine learning research with this text map of all accepted submissions to NeurIPS between 1987 to 2022.

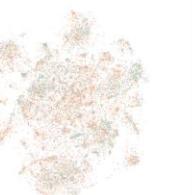
[Open Map →](#)



**ICLR 2018-2023 Submissions**  
made by Nomic

An exploratory map of 14,732 submissions to the ICLR conference, from 2018-2023.

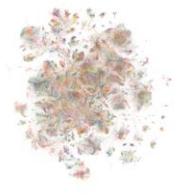
[Open Map →](#)



**HuggingFace OBELICS**  
made by Huggingface

See the frontier of multimodal model training by exploring this 11M document subset of OBELICS: an open collection of interleaved image-text web documents, containing 141M English documents, 115B text tokens, and 353M images. Created in a collaboration between Nomic and Huggingface.

[Open Map →](#)



**KREA Stable Diffusion**  
made by KREA

Explore KREA's stable diffusion database to learn how people are using generative AI to create new images.

[Open Map →](#)



**Vertex AI embeddings of stack overflow posts**  
made by Google

Google's Vertex AI group used Nomic to visualize their embeddings of 8 million stack overflow posts. Explore where people get stuck while coding, and what answers they got from 'hello worlds' in every language to assembly language and Terraform configurations.

[Open Map →](#)



手上沒有 高品質差異性大問答資料集 的工程師?

# 偷懶的工程師也可以

高品質差異性大的資料



# 偷懶的聰明工程師只準備175個問答題

## 餵給ChatGPT的需求

你被要求提供10個多樣化的任務指令。這些任務指令將提供給GPT模型，我們將評估GPT模型完成指令的能力。以下是  
你提供指令需要滿足的要求：

- 1.盡量不要在每個指令中重複動詞，要最大化指令的多樣性。
- 2.使用指令的語氣也應該多樣化。例如，將問題與祈使句結合。
- 3.指令類型應該是多樣化的，包括各種類型的任務，類別種類例如：brainstorming, open QA, closed QA, rewrite, extract, generation, classification, chat, summarization。
- 4.指令用中文書寫，指令應該是1到2個句子，允許使用祈使句或問句。
- 5.輸出應該是針對指令和輸入的適當回答。下面是10個任務指令的清單：

## 任務種子：人工設計175個任務，每個任務都有對應的{指令輸入輸出/實例}或{指令輸出/實例}

```
{"id": "seed_task_1", "name": "english_haiku_Generation", "instruction": "請以下面單字為主題寫一首詩",  
"instances": [{"input": "夏天", "output": "不只春妍夏亦佳，隨緣花草是人生。\\n鹿蕙解插纖長柄，金鳳仍開最小花。"}], "is_classification": false}
```

```
{"id": "seed_task_2", "name": "topic_classification", "instruction": "為以下文章選擇一個主題。主題候選包括：政治，體育，健康，科學，商業，金融和娛樂。", "instances": [{"input": "鯨是一種廣泛分佈和多樣化的完全水生胎盤海洋哺乳動物。它們是鯨下目內的一組非正式分組，通常不包括海豚和鼠海豚。鯨，海豚和鼠海豚屬於偶蹄目。它們最接近的非鯨類現存親戚是河馬，它們和其他鯨類分支約5400萬年前。", "output": "科學"}], "is_classification": true}
```

# ChatGPT給偷懶的聰明工程師五萬筆高品質的資料

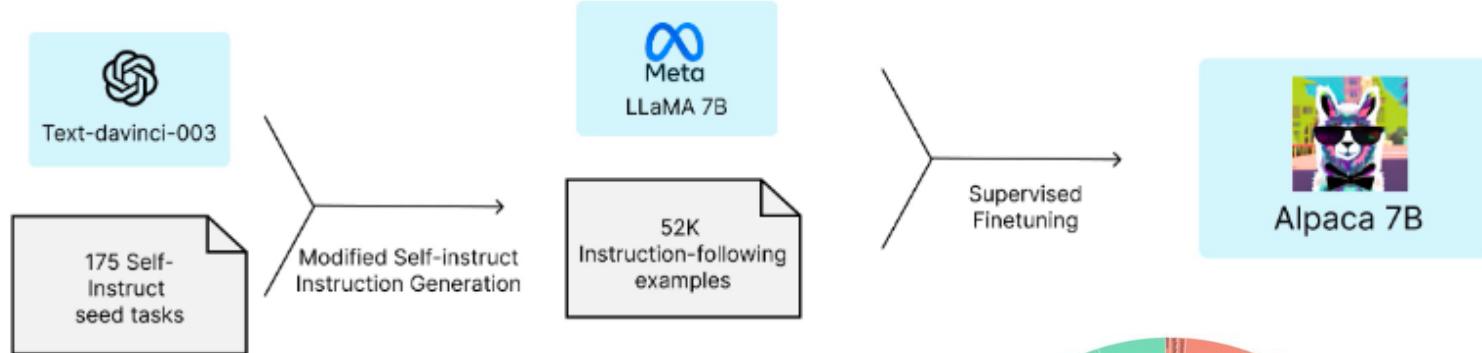
```
pip install -r requirements.txt  
export OPENAI_API_KEY=YOUR_API_KEY  
python 1.5M/generate_instruction.py generate_instruction_following_data
```

## ChatGPT輸出的資料

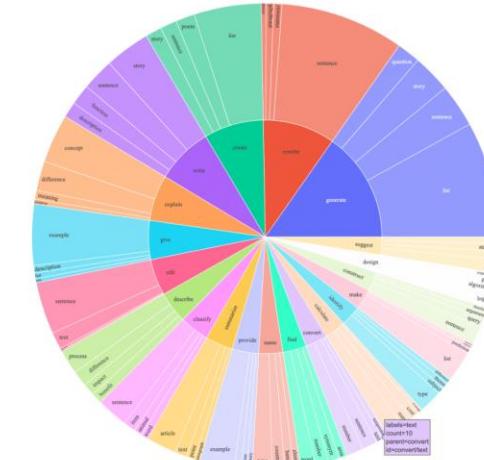
{"instruction": "為指定主題寫一首**詩歌**。\\n秋天的風景\\n", "input": "", "output": "楓樹葉飄舞紅如火，\\n稻田金黃搖曳著，\\n秋風吹拂著田野，\\n豐收時節真美妙。"}

{"instruction": "生成一首根據給定主題的**詩歌**。\\n根據以下主題生成一首詩歌：愛情。"，  
"input": "", "output": "愛情如春之花，開在心間田野。\\n不求長久不變，只盼真誠相待。\\n它似燃燒的火，雖有痛苦卻無悔。\\n愛情如生命之源，它賦予我們力量，讓我們無論前路如何坎坷都能勇往直前。"}

# 史丹佛羊駝生成流程



- 資料收集: Self-Instruct, 5萬2千條數據 ~500USD
  - 訓練 A100: GPU/HOUR = 1USD;  
 $4\text{GPUs} \times 1\text{USD} \times 24\text{hours} = \text{\~{}100USD}$



# 訓練一次要100美金?

1天 ~100USD



Stanford Alpaca: An Instruction-following LLaMA Model



Code License Apache 2.0 Data License CC By NC 4.0 Weight Diff License CC By NC 4.0 python 3.9+ code style black



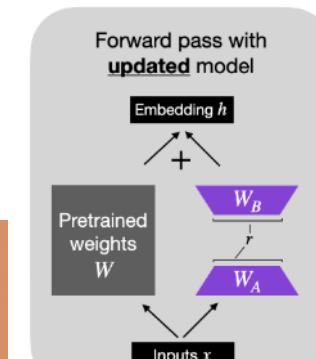
Full fine-tuning

10GB 10GB+ 10MB

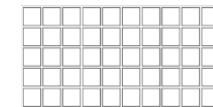
2小時 ~1USD

Alpaca LoRA

LoRA weights,  $W_A$  and  $W_B$ , represent  $\Delta W$



10\*5



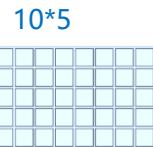
1\*5  
W<sub>B</sub>

10\*1

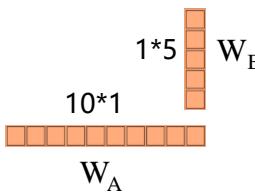


# QLoRa

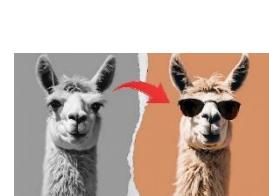
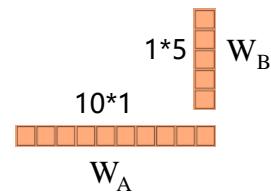
**Standard  
MODEL  
(16-bit)**



**LoRA  
MODEL  
(16-bit)**



**QLoRA  
MODEL  
(4-bit)**



2GB + 10MB

Models 12

llama2\_taiwan\_news\_qlora

DavidLanz/llama2\_taiwan\_news\_qlora  
Updated about 4 hours ago • ↓ 49

TarotSunGod/llama2\_taiwan\_news\_qlora  
Updated 4 days ago

REBCA6111/llama2\_taiwan\_news\_qlora  
Updated 4 days ago

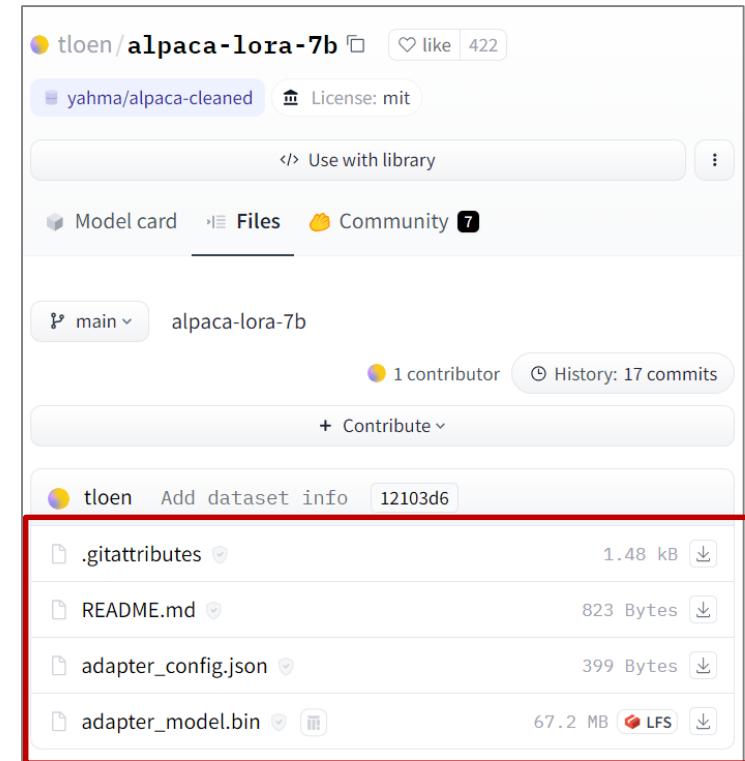
KiddsMo/llama2\_taiwan\_news\_qlora  
Updated 4 days ago • ↓ 2

YENYEEE/llama2\_taiwan\_news\_qlora  
Updated 4 days ago

brianwwwww1231/llama2\_taiwan\_news\_qlora  
Updated 4 days ago • ↓ 2

# Finetune (Alpaca-LoRA)

```
python finetune.py \
    --base_model 'decapoda-research/llama-7b-hf' \
    --data_path 'yahma/alpaca-cleaned' \
    --output_dir './lora-alpaca' \
    --batch_size 128 \
    --micro_batch_size 4 \
    --num_epochs 3 \
    --learning_rate 1e-4 \
    --cutoff_len 512 \
    --val_set_size 2000 \
    --lora_r 8 \
    --lora_alpha 16 \
    --lora_dropout 0.05 \
    --lora_target_modules '[q_proj,v_proj]' \
    --train_on_inputs \
    --group_by_length
```



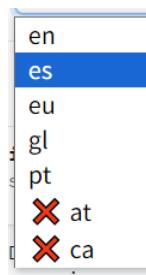
# Finetune 多國語言 (Alpaca-LoRA)

- Alpaca-LoRA MT是利用 6 種伊比利亞語言機器翻譯的羊駝資料微調模型的計畫：葡萄牙語、西班牙語、加泰隆尼亞語、巴斯克語、加利西亞語和阿斯圖里亞斯語。

- 各種適配器重量（下載風險自負）：

- 7B :

- en <https://huggingface.co/tloen/alpaca-lora-7b>
- es <https://huggingface.co/samwit/alpaca7B-lora>
- 4 <https://huggingface.co/chansung/gpt4-alpaca-lora-7b>
- gl <https://huggingface.co/nomic-ai/gpt4all-lora>
- eu <https://huggingface.co/22h/cabrita-lora-v0-1>
- pt <https://huggingface.co/qychen/luotuo-lora-7b-0-1>
- CN <https://huggingface.co/ziqingyang/chinese-alpaca-lora-7b>
- JP <https://huggingface.co/kunishou/Japanese-Alapaca-LoRA-7b-v0>
- FR <https://huggingface.co/bofenghuang/vigogne-lora-7b>
- TH <https://huggingface.co/Thaweevat/thai-buffala-lora-7b-v0-1>
- DE [https://huggingface.co/thiserand/alpaca\\_lora\\_german](https://huggingface.co/thiserand/alpaca_lora_german)
- PL <https://huggingface.co/mmosiolek/polpaca-lora-7b>
- PL <https://huggingface.co/chrisociepa/alpaca-lora-7b-pl>
- IT <https://huggingface.co/teelinsan/camoscio-7b-llama>
- RU [https://huggingface.co/IlyaGusev/llama\\_7b\\_ru\\_turbo\\_alpaca\\_lora](https://huggingface.co/IlyaGusev/llama_7b_ru_turbo_alpaca_lora)
- UA <https://huggingface.co/robinhad/ualpaca-7b-llama>
- IT <https://huggingface.co/mchl-labs/stambecco-7b-plus>
- ES <https://huggingface.co/plncmm/guanaco-lora-7b>
- GBESPT <https://huggingface.co/HiTz/alpaca-lora-7b-en-pt-es-ca-eu-gl-at>



instruction string	input string	output string	prompt string
Give three tips for staying...		1.Eat a balanced diet and make sure...	Below is an instruction that...
What are the three primary...		The three primary colors are red,...	Below is an instruction that...
Describe the structure of a...		An atom is made up of a nucleus, which...	Below is an instruction that...
How can we reduce air...		There are a number of ways to reduce...	Below is an instruction that...

instruction string	input string	output string	prompt string
Dé tres consejos para...		1.Coma una dieta equilibrada y...	A continuación se muestra una...
¿Cuáles son los tres colores...		Los tres colores primarios son el...	A continuación se muestra una...
Describa la estructura de...		Un átomo está formado por un núcleo, que...	A continuación se muestra una...

# 策略二之一, Excel 微調模型

NCHC GPU, API+\*xls or \*csv or \*json

一萬筆 Excel  
instruction

一張GPU  
V100

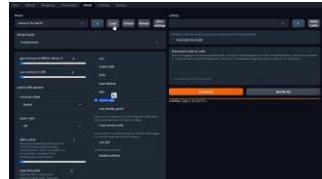
時間 1HR

花費 NT10元

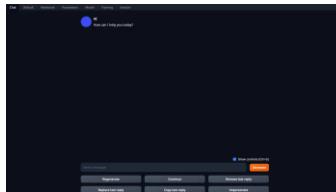
<https://www.youtube.com/watch?v=LyWJtdJ9olc>

<https://www.youtube.com/watch?v=p2k1RjtJeS4>

模型微調



模型上架



# 策略二之二, QA上傳分享

## NCHC Application, GUI+QA

預設情況下，「分享」功能是關閉的。



但你可以按下「**分享**」按鈕，分享您的QA成果，為訓練大型語言模型的民主過程做出貢獻。

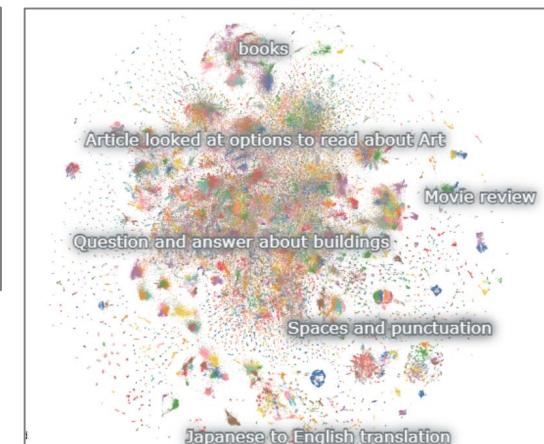
**question**

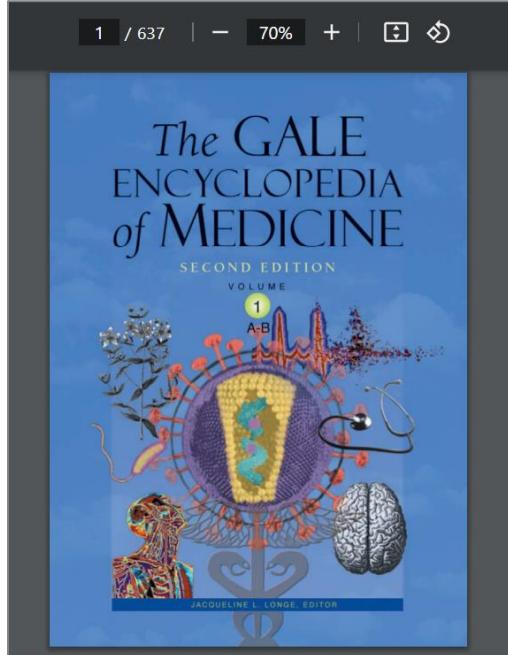
Given the sentence "A wet dog jumps over a bar." is it true that "The dog is in the pool."? Stream of thoughts:

**more**

**response**

Based on the given sentence, there is no mention of a pool. It only states that the dog is wet and jumping over a bar. Therefore, it is not true that the dog is in the pool.





## 方法三：建立領域知識庫

選擇合用的向量模型與向量資料庫  
讓使用者用領域知識庫與AI對話

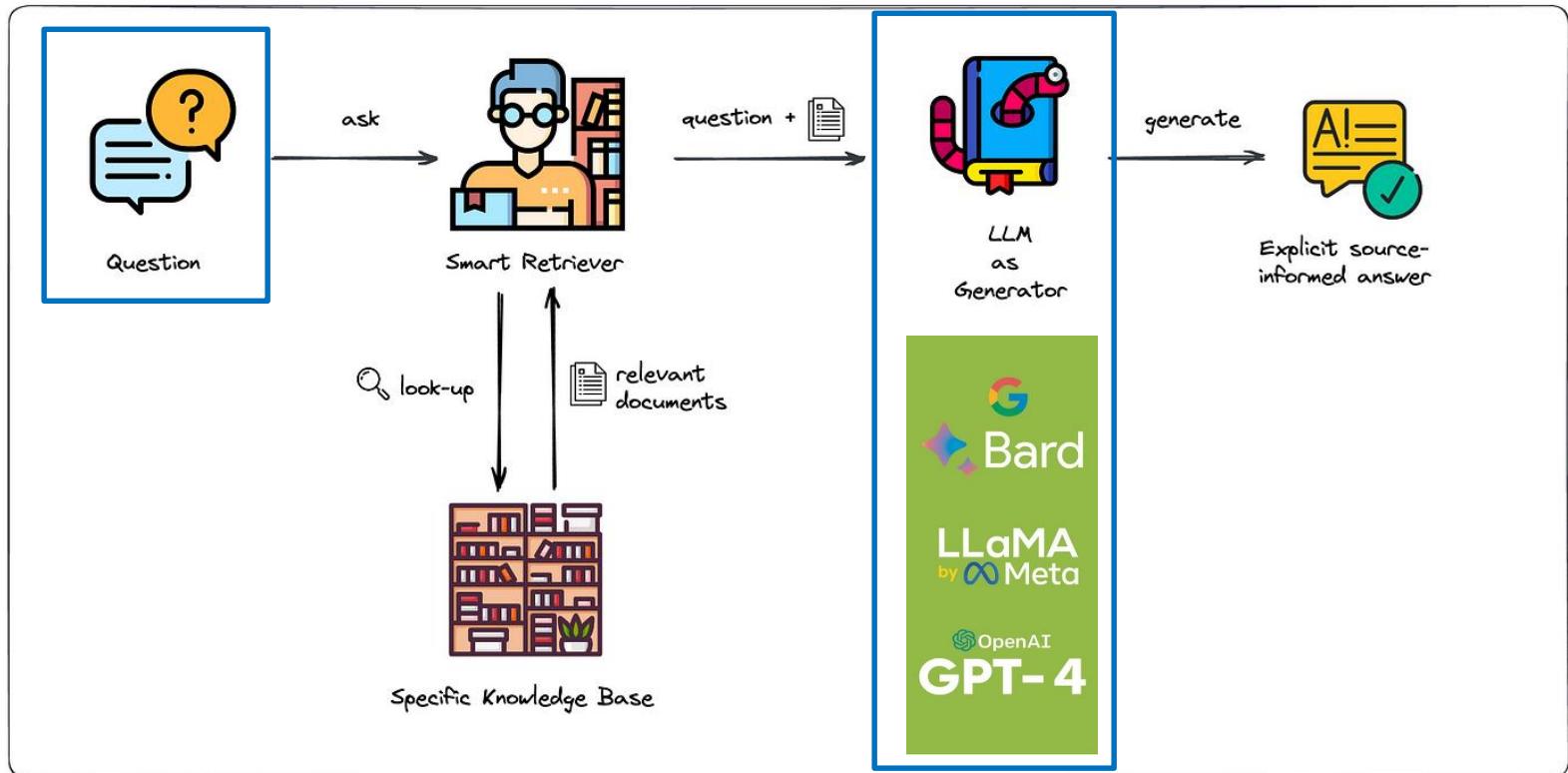
動態存取組織不斷發展的內部資料庫的需求以及應答過程中透明度的潛在要求

- **幻覺：**當模型自信地產生錯誤反應時就會發生幻覺。例如，如果一個模型聲稱 “GOOGLE獵犬” 是一個真正的狗品種，那麼這是一種幻覺，可能會導致誤導性訊息。
- **知識截止：**LLM 根據模型的訓練資料傳回過時的資訊。每個基礎模型都有一個知識截止點，這意味著它的知識僅限於訓練時可用的資料。例如，如果您向模型詢問最近的 NBA 冠軍得主，它可能會回答過時的資料。

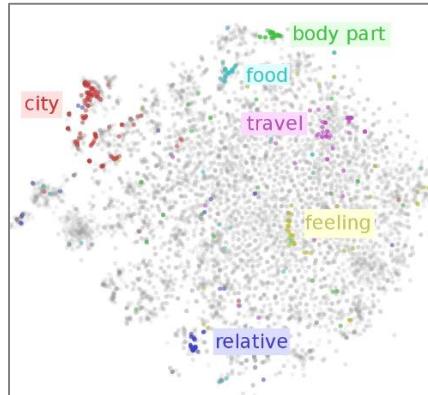
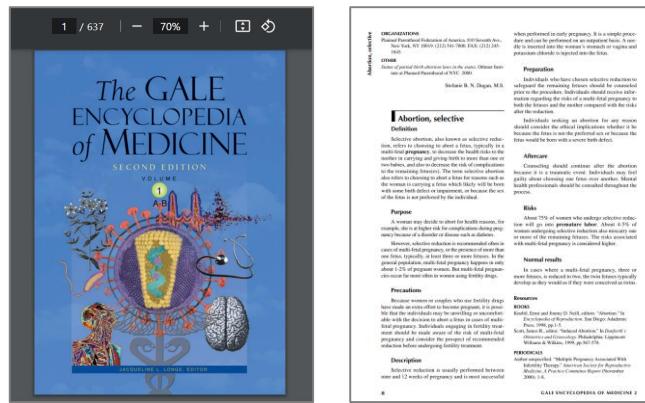
The dog breed "Googly Retriever" does not exist as a recognized breed, so its appearance would depend on the characteristics of the hypothetical mix of breeds it might encompass.

I do not have access to real-time information, so I cannot provide the most recent NBA Championship game winner as of October 2023. Please check a reliable news source for the latest information.

# 檢索增強生成 (RAG)



# 知識庫建立 Knowledge base

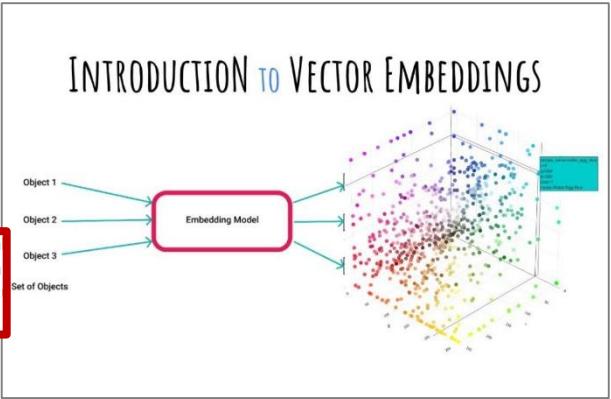


The screenshot shows the Pinecone AI Agent interface. On the left sidebar, there are navigation links: PROJECT, AI agent (selected), Indexes, API Keys, Members, Docs, and Support Center. The main area has a title "AI agent Indexes" and tabs for BROWSER and METRICS. Below that, it says "Matches: 1-10 of 50". The first result is highlighted with a red box and labeled "semantic search".

1	ID	dd739284-5...
SCORE 0.0730	METADATA	

2	ID	7385b61b-1398-464d-b0...
SCORE 0.0730	METADATA	text: "Growing bones need alka...

A diagram on the right illustrates the search process: "question embedding" (represented as a hexagon) is compared against an "index" (represented as a stack of binary digits). The result is labeled with a circled "5". A red box highlights the "semantic search" result, which is shown as a box containing the ID, score, and metadata. Another red box highlights the "semantic search" label itself.



query = "What are Allergies"  
docs=dorsearch.similarity\_search(query, k=3)

## most common of medical m someform of allergy

most common of medical\ndisorders. It is estimated th  
m someform of allergy, with similar proportions throu  
or school absence and is a major source of lostproduct  
,\nthe immune system responds to foreign microorgan  
antibodies are capable of binding to identifying mol  
dy and antigen sets off series of chemical reactions  
o reactions is triggered by harmless, everyday subst  
develops against the offending substance (an allerg  
n microorganisms and particles, like pollen or dust,  
ng to identifying molecules, or antigens, on theforef  
of reactions designed to protect the body from infect  
ryday substances.This is the condition known as aller  
ergens include pollen, dust, and insect parts from ti  
ic reactions involve a special set of cells in''])



# 回饋答案範例

- 1.) 黃色文字：這是 LLAMA2 7B 回應什麼是濕疹。
- 2.) 粉紅色文字：這是從大約 600 頁的書中提取資料並進行向量化的確切來源。

User 02:53:08 PM what is Eczema ? Took 1 step

Chatbot 02:53:08 PM Eczema (atopic dermatitis) is a chronic skin disorder characterized by dry, itchy, scaly skin. It can affect anyone, but it is most common in infants and young children. Eczema can be triggered by environmental factors such as pollen, dust mites, or cold weather, or by internal factors such as allergies or stress. Treatment for eczema typically involves managing symptoms with topical corticosteroids, immunomodulators, or antihistamines, and addressing any underlying causes.

Sources:[Document(page\_content='See also Rotaviruses\n\nInfantile eczema. See Atopic dermatitis\n\nInfantile gastroenteritis, II:1235, 1236\nGALE ENCYCLOPEDIA OF MEDICINE 2 3793Infection', metadata={'source': 'data\Gale Encyclopedia of Medicine Vol. 5.pdf', 'page': 564}), Document(page\_content='Keratin\nkeratosis pilaris from, III:1907\nin lichen simplex chronicus,\n\nKeratitis, III:1904–1906 , 1905\nfrom adenoviruses, I:57\ncorneal ulcers from, II:921\nKeratoconjunctivitis, I:57, 58\nKeratoconus\nnastigmatism from, I:385\ncorneal transplant for, II:919\nradial keratotomy and, IV:2804\nfor osteomyelitis, IV:2421\nfor premature labor, IV:2701\nJunctional epidermolysis bullosa,\n\nJunctional moles. See Nevus\n\nJung, Carl, I:368, III:2286\nJungian dance therapy. See Dance ther-\n\nappy', metadata={'source': 'data\Gale Encyclopedia of Medicine Vol. 5.pdf', 'page': 574})]

# 文本嵌入(Text Embedding)

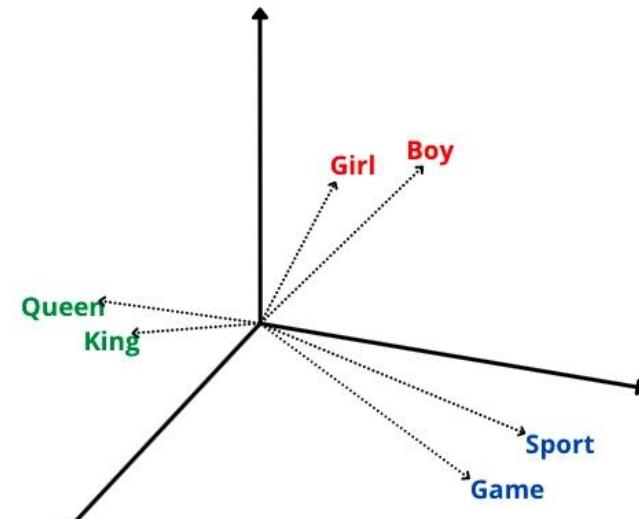
**文本嵌入(Text Embedding)**是自然語言處理中的一個重要概念,它的目標是將文本中的詞表示成固定長度的稠密向量,也稱為詞向量(Word Vector)。這樣每個字都可以用一個連續的、低維的稠密向量來表示,例如200-300維。

文本嵌入的主要目的是捕捉每個詞的語義資訊,使得語義相關的詞在嵌入空間中距離較近,不相關的詞距離較遠。這樣就可以用向量間的距離來表示詞之間的語意關係。

下圖向量在幾何空間的表示, Girl(女孩)和Boy (男孩), queen (女王) 和king (國王), Sport (運動) 和game (遊戲) 這三對單字都有相近的語義關係, 所以他們的文字向量的幾何距離都比較近。

768浮點

```
[ 3.60721558e-01  7.86174655e-01  8.18841517e-01 -2.01314002e-01  
-1.74384236e-01 -1.00326824e+00 -1.72811300e-01 -1.04662508e-01...  
]
```



# 嵌入模型 mteb

<https://youtu.be/01g7MFBtLig>



<https://huggingface.co/spaces/mteb/leaderboard>

應用場景: 文字相似度搜尋、文字分類任務、情緒分析、垃圾郵件過濾、文字推薦、自動翻譯、問答系統

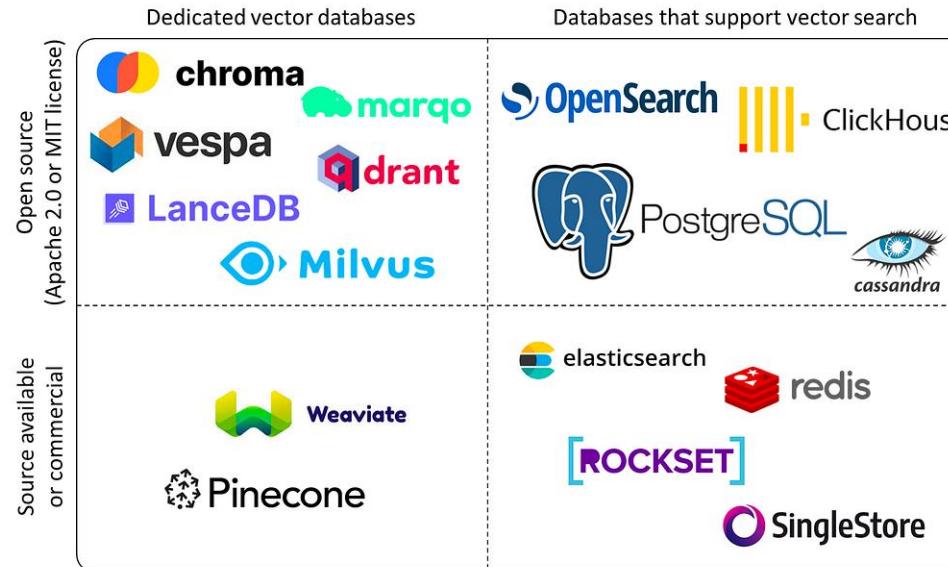
Rank	Model	Model Size (GB)	Embedding Dimensions	Sequence Length	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)	Retrieval Average (15 datasets)	STS Average (10 datasets)	Summarization Average (10 datasets)
1	Cohere-embed-english-v3.0		1024	512	64.47	76.49	47.43	85.84	58.01	55	82.62	30.18
2	bge-large-en-v1.5	1.34	1024	512	64.23	75.97	46.08	87.12	60.03	54.29	83.11	31.61
3	Cohere-embed-multilingual-v3.0		1024	512	64.01	76.01	46.6	86.15	57.86	53.84	83.15	30.99
4	bge-base-en-v1.5	0.44	768	512	63.55	75.53	45.77	86.55	58.86	53.25	82.4	31.07

# 向量資料庫

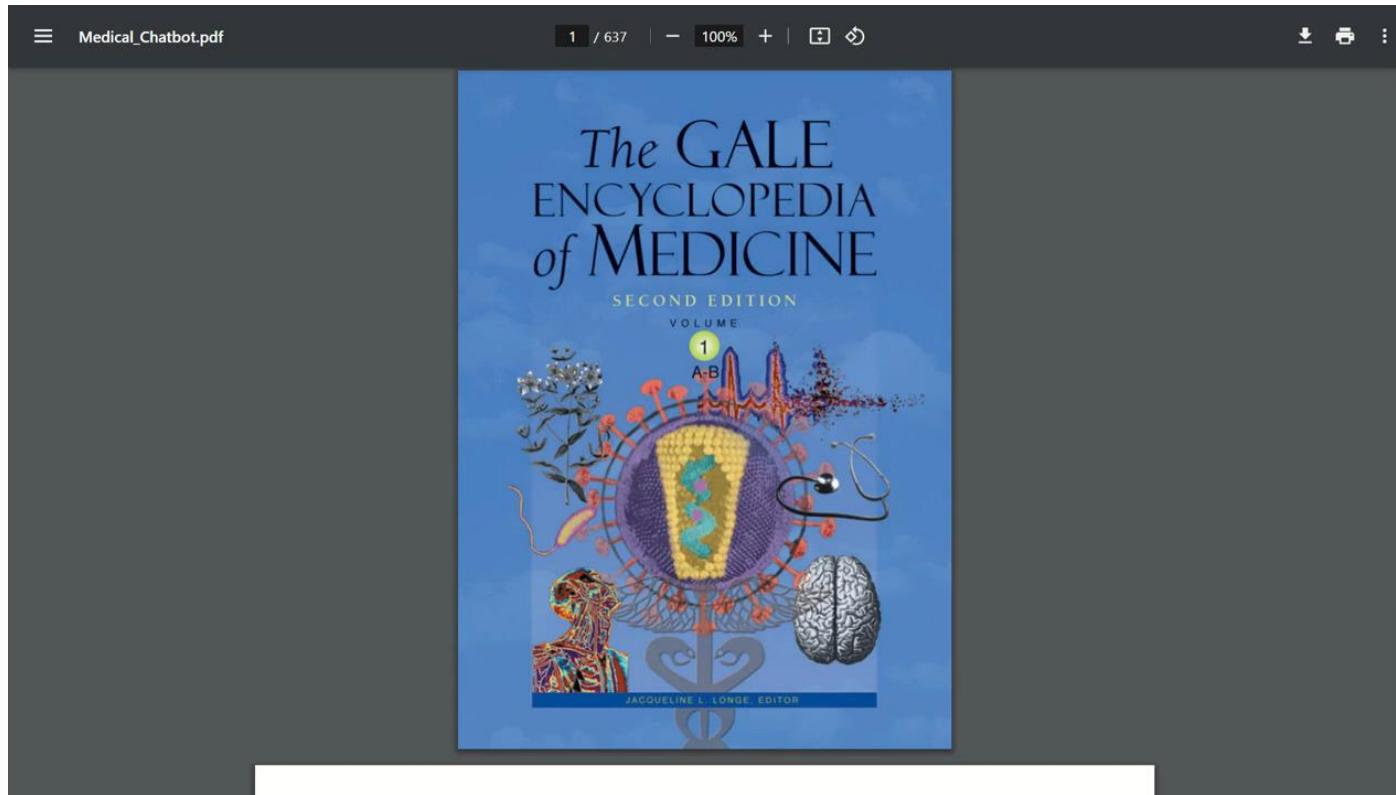
<https://youtu.be/sI6TwgVgLj4>

768浮點

```
[ 3.60721558e-01  7.86174655e-01  8.18841517e-01 -2.01314002e-01  
-1.74384236e-01 -1.00326824e+00 -1.72811300e-01 -1.04662508e-01....  
]
```



[https://youtu.be/\\_EPQxIQ0VhY](https://youtu.be/_EPQxIQ0VhY)



# 策略三，文件向量知識庫搜尋引擎

NCHC CPU, API+\*pdf

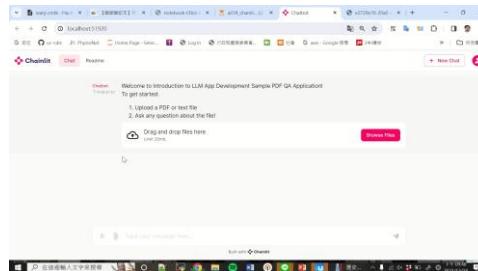
PDF文件

向量模型

向量資料庫

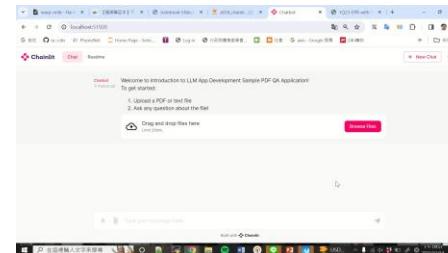
<https://youtu.be/lRvdei8DJKo>

德欣寰宇資訊安全白皮書

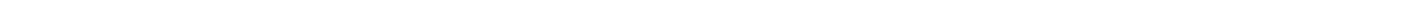


<https://youtu.be/CJOqvMT1IV0>

Gap Inc. 公佈 2023 財年第一季業績

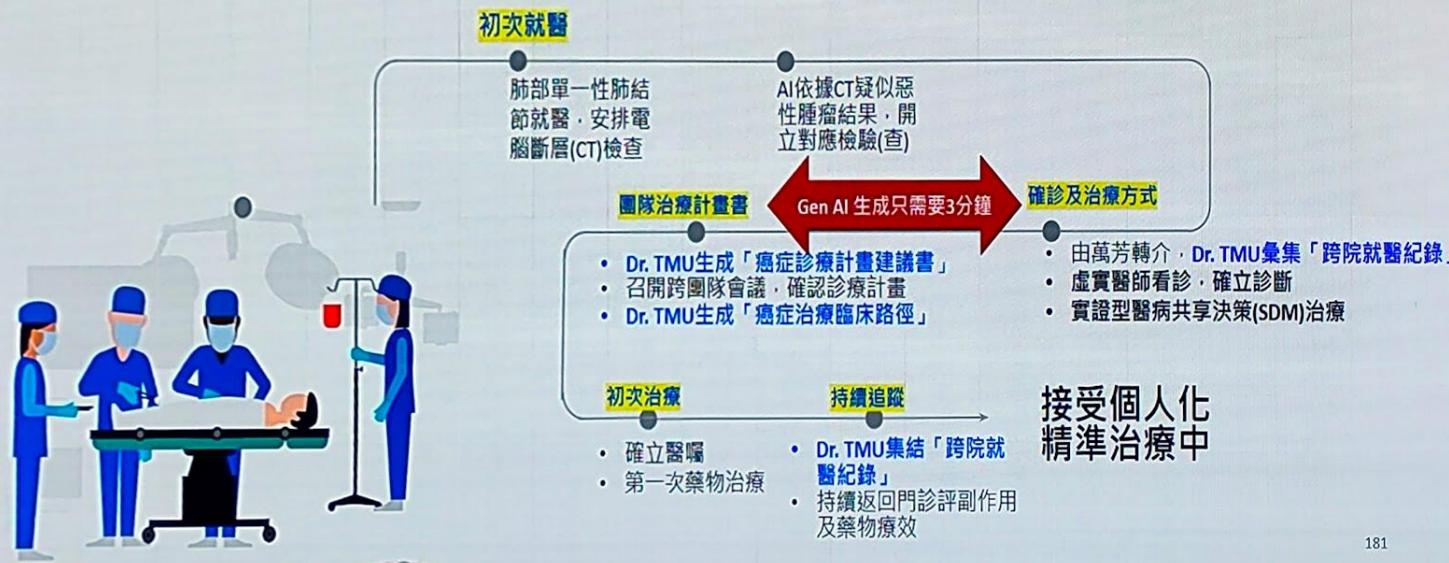


# 3. 落地案例



## 導入Dr. TMU 肺癌病人治療程序

團隊治療計畫 7天->3分鐘



# 落地場景應用 - 北醫

The diagram illustrates the integration of various AI and cloud services for cancer treatment planning:

- Google BigQuery:** Represented by a blue hexagonal icon, it is used for structured data storage and retrieval.
- PaLM 2:** Represented by a colorful flower-like icon, it is used for generating treatment plans and summarizing medical facts.
- AppSheet:** Represented by a white card with a blue border, it is used for creating mobile and web applications for healthcare professionals.

**Left Panel (Google BigQuery & PaLM 2):**

- Google BigQuery interface showing a query results table.
- PaLM 2 interface showing a summary of treatment plan details.
- 主題式收集、分析及生成資料** (Subject-based collection, analysis, and generation of data).

**Middle Panel (AppSheet):**

**Taipei Medical University Hospital [Cancer Treatment Plan] Recommendation Letter (Dr. TMU)**

**Diagnostic Post or Initial Treatment Plan**

**Patient Information:** 病歷號碼 10000009, 姓名 XOO, 性別 女, 出生日期 043 年 00 月 00 日

**Diagnosis:** (C34.30) 確診日期 111 年 11 月 28 日  
下葉之左支氣管或肺惡性腫瘤 cT4N3M1a stage IVA

**Other Clinical Staging:** 不適用  
**Clinical Stage:** AJCC8 不適用

**Histologic Type:** Adenocarcinoma

**Previous History/Treatment Pattern:** 無

**Co-morbidities:** 無

**Important Prognostic Factors:**

- ECOG: 0
- Lung cancer family history: 否
- Tobacco: 否
- Alcohol: 否
- Betel nuts: 否
- EGFR: 是 exon 19 mutation

**Smoking Status:** 無

**Treatment Plan:** Target therapy : Osimertinib (Tagrisso)

**Key Point:** EGFR: exon 19 mutation

**References:** 1. Target Therapy for Advanced or Metastatic Disease: EGFR-mutation-positive

**Conclusion:** 肺癌治療指引: Page. 70, 73, 74, 75, 76, 95, 98, 105, 106 (臺北醫學大學肺癌治療指引)

**Right Panel (AppSheet):**

**Taipei Medical University Hospital [Cancer] Clinical Pathway (建議)**

**Patient Information:** 姓名 XOO, 性別 女, 出生日期 043 年 00 月 00 日

**Timeline:**

- Day 1** 治療: 1. Physical Examination  
2. 血液: CBC-DC  
3. 生化: ALT/AST/T-Bil/BUN/Cr/Na/K  
4. CXR  
5. ECG
- Day 8** 副作用評估: 1. Physical Examination  
2. 血液: CBC-DC  
3. 生化: ALT/AST/T-Bil/BUN/Cr/Na/K  
4. ECG
- Day 29** 副作用評估: 1. Physical Examination  
2. 血液: CBC-DC  
3. 生化: ALT/AST/T-Bil/BUN/Cr/Na/K  
4. CXR
- 3 months** 副作用及療效評估: 1. Physical Examination  
2. 血液: CBC-DC  
3. 生化: ALT/AST/T-Bil/BUN/Cr/Na/K  
4. Response evaluation (Chest CT +/− Brain MRI +/− Others)

**Conclusion:** 開立日期 111 年 12 月 06 日

**Cancer Facts & Figures 2023**

**PaLM 2** (花形图标)

**自動摘錄與引用最新文獻** (Automatic extraction and citation of latest literature)

**A Cancer Journal for Clinicians**

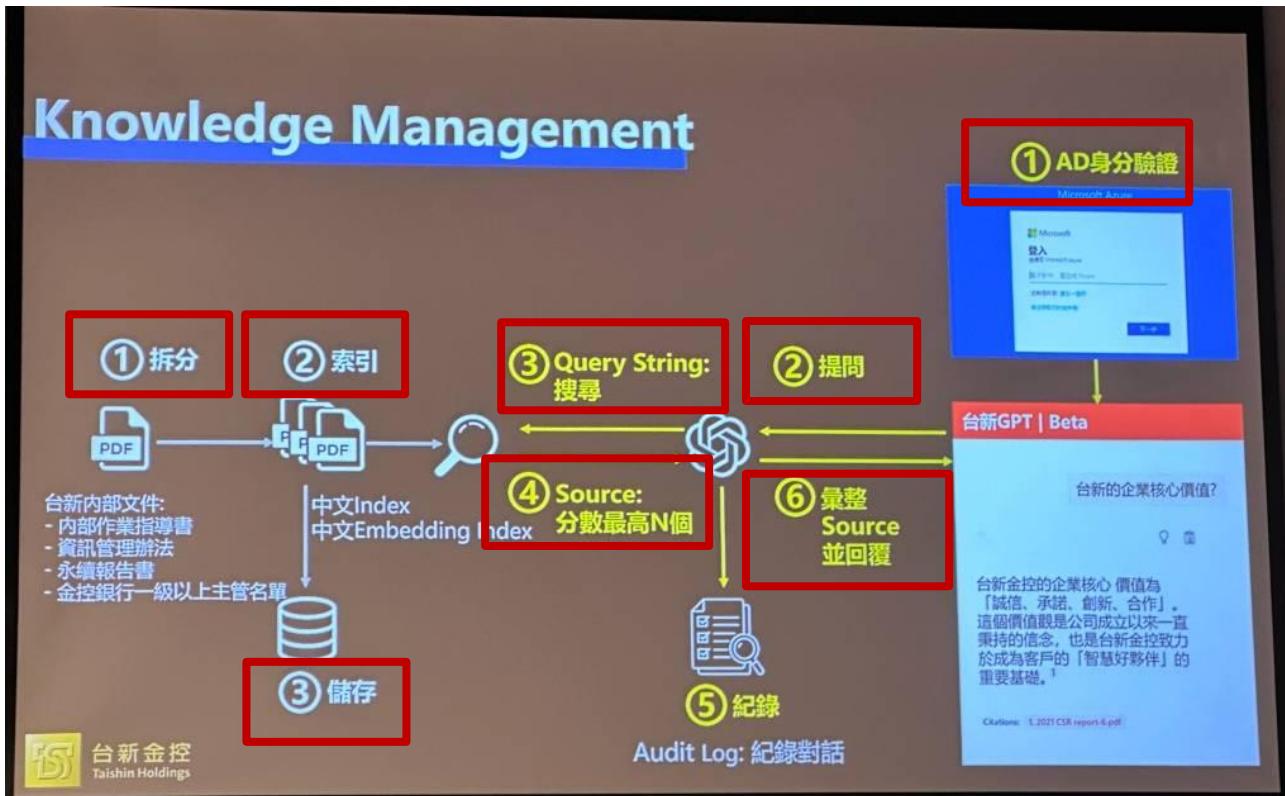
**ELCC 2023**

**LUNGevity**

**Lung Cancer Helpline 844-360-5864**

**2023 ASCO: Highlights of Lung Cancer Research**

# 落地場景應用- 台新金



# 落地場景應用 KKDAY

The image displays a multi-step process for handling delivery receipt issues:

- Step 1:** A screenshot of a software interface titled "kkday 國際收件匣". It shows a list of delivery items. A specific item is highlighted with a red box, and an orange arrow points from this box to the next step.
- Step 2:** A search result titled "商品如果訂錯日期了，更改日期會收..." (If the delivery date is wrong, will changing the date incur a fee?). The result is framed by a red box.
- Step 3:** A detailed explanation of the policy, framed by a red box. It states: "根據搜尋結果，更改日期是否收取費用取決於商品的取消規定。如果商品的取消規定是「不能取消或更改」，則更改日期將收取100%的訂單取消手續費。如果商品的取消規定是「可以取消或更改」，則更改日期可能會收取費用，具體費用取決於商品的取消規定。您可以查看商品的取消規定，以了解更改日期是否會收取費用。"

On the right side of the image, there are three circular performance metrics:

- A blue circle with the number "3" indicating "週完成開發" (Completed development per week).
- A green circle with "90%" indicating "提升客服效益" (Improved customer service efficiency) and "客服人員查找答案上線至正式環境" (From average 10 min to 1 min for customer service staff to find answers online to production environment).
- An orange circle with "65%" indicating "比起自行建置，整體費用下降 65%" (Overall cost reduction compared to self-construction, 65%).

<https://youtu.be/4ZjW9gS2qrQ>

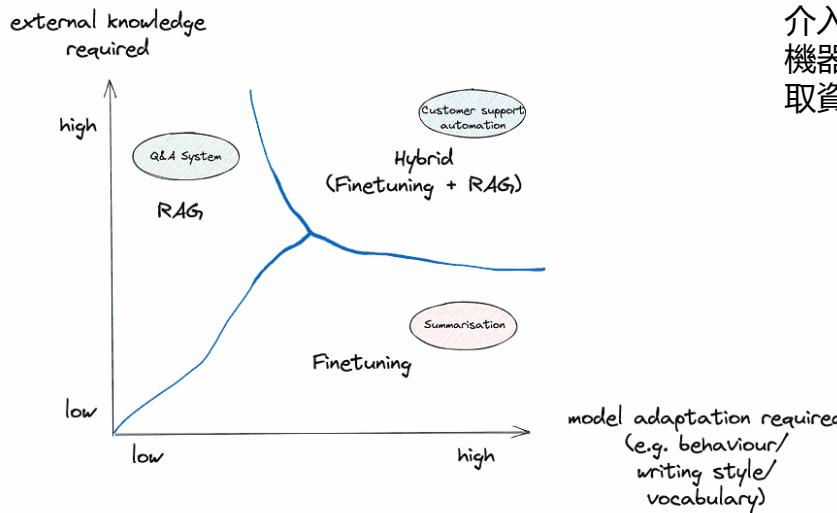
# 檢索增強生成(RAG)與模型微調(Finetuning)比較

1. 需要外部知識嗎?
2. 需要模型寫作風格嗎?
3. 減少幻覺至關重要?
4. 訓練資料可用嗎?
5. 資料的動態程度如何?
6. 需要透明度/可解釋性嗎?

	RAG	Finetuning
External knowledge req'd?	✓	✗
Changing model behaviour req'd?	✗	✓
Minimise hallucinations?	✓	✗
Training data available?	✗	✓
Is data (mostly) dynamic?	✓	✗
Interpretability req'd?	✓	✗

## 組織知識（即外部資料）的問答系統

**RAG 系統**是合適的選擇。鑑於動態存取組織不斷發展的內部資料庫的需求以及應答過程中透明度的潛在要求，RAG 提供了與這些需求非常契合的功能。



**自動對話機器人：混合方法**可能是最佳選擇。微調可以確保聊天機器人與公司的品牌、語氣和常識保持一致，處理大多數典型的客戶查詢。然後，RAG 可以作為一個補充系統，介入進行更動態或更具體的查詢，確保聊天機器人可以從最新的公司文件或資料庫中提取資訊，從而最大限度地減少幻覺。

## 專門領域和/或特定風格

**微調**是合適的選擇。主要目標是風格一致。假設有大量先前的摘要可用於培訓，那麼對 LLMs 進行微調將允許深度適應所需的風格，捕捉該領域的細微差別和複雜性。

# 4. LLMs實作

[教學講義](#)

---

# 國網計算服務資源介紹

虛擬  
運算

VCS

智算兼備  
**虛擬運算服務**  
Virtual Compute Service

容器  
運算

K8s

快速部署  
**容器運算服務**  
Container Compute Service

高速  
運算

HPC

有效統御  
**高速運算服務**  
High-performance Computing

雲端  
儲存

CSS

大數據匯集  
**雲端儲存服務**  
cloud storage service

1 + N

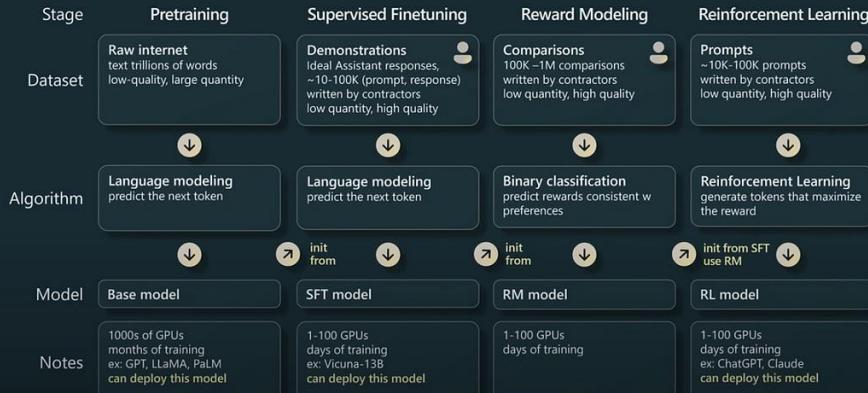


# 要怎麼開始 LLMs 專案

Prompt  
Character  
確切的提詞  
AI認知的角色  
(個人知識)

Knowledge  
Base  
擴大問題  
索引  
(組織知識)

## GPT Assistant training pipeline



Continue Pretraining  
學會中文接字  
10億中文字  
(出版社知識)

Finetuning  
學會QA  
5萬組對話  
(專業領域知識)

知識  
資料專家

# 課程教學

<https://hackmd.io/uUuyfxXWSFCF0M50ue2MJA>

- [HPC] 台灣杉二號介紹與操作教學
- [TWCC CCS] TWCC容器服務介紹與操作教學
- [Colab] Colab介紹與操作教學 (12/28課程)
- [Colab] Hugging Face 介紹與操作教學
- [Colab] TextGen 介紹與操作教學 (12/28課程)
- [Colab] Langcahin 介紹與操作教學
- [Colab] 模型微調 介紹與操作教學
- [Colab] Chainlit 介紹與操作教學 (12/28課程)

# 課程教學-台灣杉二號介紹與操作教學

<https://hackmd.io/@whYPD8MBSHWRZV6y-ymFwQ/HJyMVPbSp>

<https://www.youtube.com/watch?v=WWNdvFsV6EE>



# 課程教學- TWCC容器服務介紹與操作教學

<https://hackmd.io/@whYPD8MBSHWRZV6y-ymFwQ/rJO3TcFPT>

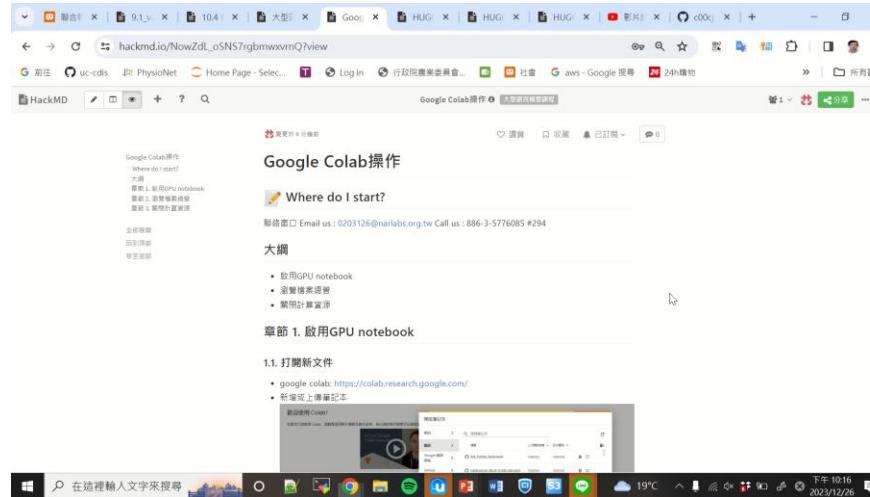
<https://youtu.be/9ZnfyEZlgAI>



# 課程教學- Colab介紹與操作教學

<https://hackmd.io/@whYPD8MBSHWRZV6y-ymFwQ/HysOj9DPT>

<https://www.youtube.com/watch?v=wWOa-z-fTB0>

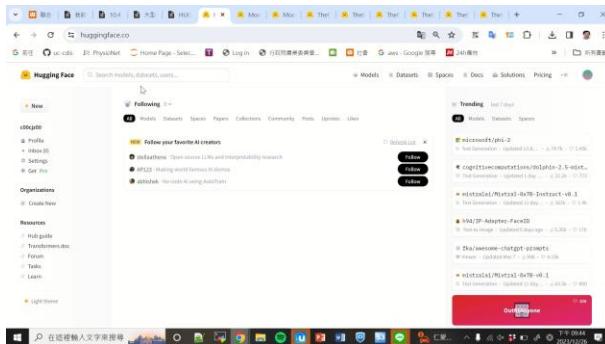


# 課程教學- Hugging Face 介紹與操作教學

<https://hackmd.io/@whYPD8MBSHWRZV6y-ymFwQ/HkjJcSdw6>

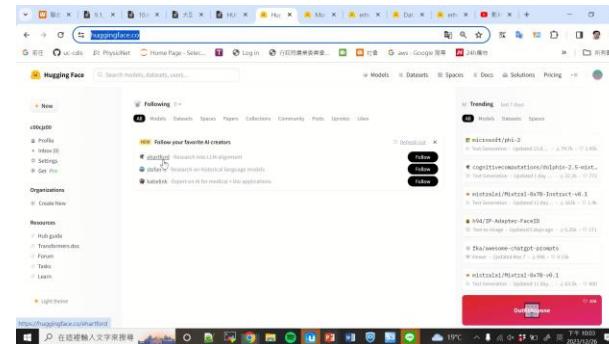
## MODEL

<https://youtu.be/q6im34-Sj48>



## DATASET

<https://youtu.be/CBoBCUGriV8>



# 課程教學- TextGen 介紹與操作教學

<https://hackmd.io/@whYPD8MBSHWRZV6y-ymFwQ/HJ8cMsDwp>

## 啟動 TextGen

<https://youtu.be/l-oLK-Olh7o>



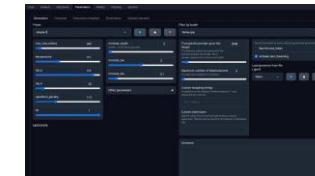
## 上架語言模型

<https://youtu.be/6Ual2nd4ezU>



## 指派AI人設

<https://www.youtube.com/watch?v=Vx1tjNs9q-k>



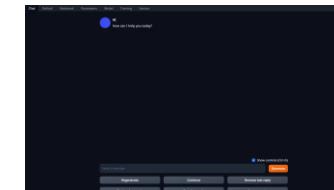
## 模型微調

<https://www.youtube.com/watch?v=LWJtdJ9olc>



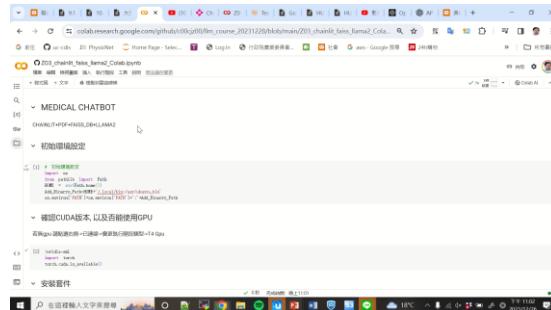
## 上架微調模型

<https://www.youtube.com/watch?v=p2k1RjtJeS4>



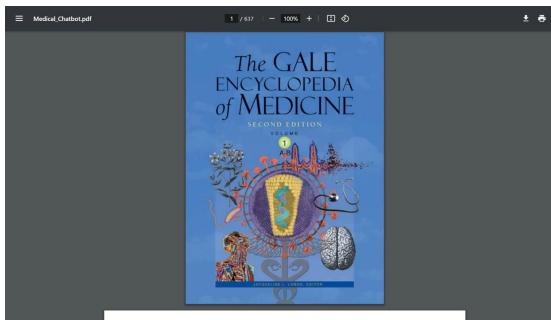
# 課程教學- Chainlit操作介紹 (I)

[https://hackmd.io/@whYPD8MBSHWRZV6y-ymFwQ/BkuwTD\\_vT](https://hackmd.io/@whYPD8MBSHWRZV6y-ymFwQ/BkuwTD_vT)



## Z03\_chainlit\_faiss\_llama2\_Colab.ipynb

<https://youtu.be/WeJAYSpbMbs>

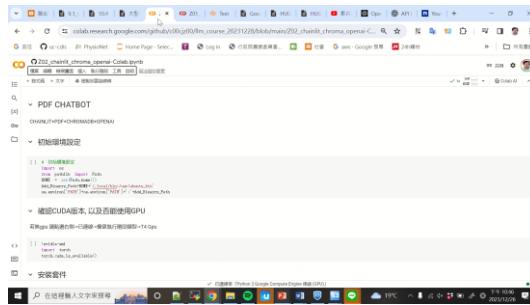


## Medical Book

[https://www.youtube.com/watch?v=\\_EPQxIQ0VhY](https://www.youtube.com/watch?v=_EPQxIQ0VhY)

# 課程教學- Chainlit操作介紹 (II)

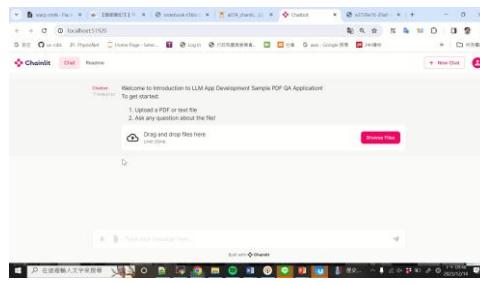
<https://hackmd.io/@whYPD8MBSHWRZV6y-ymFwQ/S1u-UnFDp>



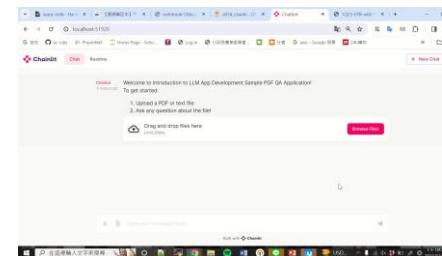
Z02\_chainlit\_chroma\_openai-Colab.ipynb

<https://youtu.be/WmhF8tVqTlk>

德欣寰宇資訊安全白皮書  
<https://youtu.be/IRvdei8DJKo>



Gap Inc. 公佈 2023 財年第一季業績  
<https://youtu.be/CJOqvMT1IV0>



# LLaMA-Factory

## Supported Models

Model	Model size	Default module	Template
<a href="#">Baichuan</a>	7B/13B	W_pack	baichuan
<a href="#">Baichuan2</a>	7B/13B	W_pack	baichuan2
<a href="#">BLOOM</a>	560M/1.1B/1.7B/3B/7.1B/176B	query_key_value	-
<a href="#">BLOOMZ</a>	560M/1.1B/1.7B/3B/7.1B/176B	query_key_value	-
<a href="#">ChatGLM3</a>	6B	query_key_value	chatglm3
<a href="#">Falcon</a>	7B/40B/180B	query_key_value	falcon
<a href="#">InternLM</a>	7B/20B	q_proj,v_proj	intern
<a href="#">LLaMA</a>	7B/13B/33B/65B	q_proj,v_proj	-
<a href="#">LLaMA-2</a>	7B/13B/70B	q_proj,v_proj	llama2
<a href="#">Mistral</a>	7B	q_proj,v_proj	mistral
<a href="#">Mixtral</a>	8x7B	q_proj,v_proj	mistral
<a href="#">Phi-1.5/2</a>	1.3B/2.7B	Wqkv	-
<a href="#">Qwen</a>	1.8B/7B/14B/72B	c_attn	qwen
<a href="#">XVERSE</a>	7B/13B/65B	q_proj,v_proj	xverse
<a href="#">Yi</a>	6B/34B	q_proj,v_proj	yi
<a href="#">Yuan</a>	2B/51B/102B	q_proj,v_proj	yuan

## Supported Training Approaches

Approach	Full-parameter	Partial-parameter	LoRA	QLoRA
Pre-Training	✓	✓	✓	✓
Supervised Fine-Tuning	✓	✓	✓	✓
Reward Modeling	✓	✓	✓	✓
PPO Training	✓	✓	✓	✓
DPO Training	✓	✓	✓	✓

## Hardware Requirement

Method	Bits	7B	13B	30B	65B	8x7B
Full	16	160GB	320GB	600GB	1200GB	900GB
Freeze	16	20GB	40GB	120GB	240GB	200GB
LoRA	16	16GB	32GB	80GB	160GB	120GB
QLoRA	8	10GB	16GB	40GB	80GB	80GB
QLoRA	4	6GB	12GB	24GB	48GB	32GB

# LLaMA-Factory - INSTALL

```
# INSTALL
```

```
git clone https://github.com/hiyouga/LLaMA-Factory.git
```

```
cd LLaMA-Factory
```

```
pip install -r requirements.txt
```

```
# Web
```

```
CUDA_VISIBLE_DEVICES=0 python src/train_web.py
```

```
demo.launch(server_name="0.0.0.0", share=True, inbrowser=True)
```

# LLaMA-Factory – Tokenizer extension

LlaMA2 詞表: 32,000, 中文詞表: 23,296, 合併詞表: 55,296

**The primary use of LLaMA is research on large language models, including**

```
[['_The', '_primary', '_use', '_of', '_L', 'La', 'MA', '_is', '_research', '_on', '_large', '_language', '_models', ',', '_including'] -> llama2  
[['_The', '_primary', '_use', '_of', '_L', 'La', 'MA', '_is', '_research', '_on', '_large', '_language', '_models', ',', '_including'] -> Chinese  
[['_The', '_primary', '_use', '_of', '_L', 'La', 'MA', '_is', '_research', '_on', '_large', '_language', '_models', ',', '_including'] -> Mistral
```

**蔡英文，中華民國政治人物。**

```
['_', '<0xE8>', '<0x94>', '<0xA1>', '英', '文', ' ', '中', '華', '民', '國', '政', '治', '人', '物', '.'] -> llama2  
['_', '蔡英文', ' ', '中', '華', '民國', '政治人物', '.'] -> Chinese  
['_', '<0xE8>', '<0x94>', '<0xA1>', '英', '文', ' ', '中', '華', '民', '國', '政', '治', '人', '物', '.'] -> Mistral
```

**現任中華民國總統。**

```
['_', '現', '任', '中', '華', '民', '國', '<0xE7>', '<0xB8>', '<0xBD>', '<0xE7>', '<0xB5>', '<0xB1>', '.'] -> llama2  
['_', '現', '任', '中', '華', '民國', '總統', '.'] -> Chinese  
['_', '現', '任', '中', '華', '民', '國', '總', '統', '.'] -> Mistral
```

# LLaMA-Factory - Train on a single GPU

## # Pre-Training

```
python src/train_bash.py --stage pt --do_train --model_path TE -dataset PDT --finetuning_type lora \  
--output_dir PT
```

## #Supervised Fine-Tuning

```
python src/train_bash.py --stage sft --do_train --model_path PT -dataset SDT --finetuning_type lora \  
--output_dir SFT
```

## Reward Modeling

```
python src/train_bash.py --stage rm --do_train --model_path PT -dataset RDT --finetuning_type lora \  
--adapter_path SFT --output_dir RM
```

## PPO Training

```
python src/train_bash.py --stage ppo --do_train --model_path PT -dataset SDT --finetuning_type lora \  
--adapter_path SFT --reward_model RM --output_dir PRO
```

## DPO Training

```
python src/train_bash.py \--stage dpo --do_train --model_path PT -dataset RDT --finetuning_type lora \  
--adapter_path SFT --output_dir DPO
```

# LLaMA-Factory - Dataset

## # Pre-training datasets (PDT)

Anarchism is a political philosophy and movement that is sceptical of authority and rejects all involuntary, coercive forms of hierarchy. Anarchism calls for the abolition of the state, which it holds to be unnecessary, undesirable, and harmful. As a historically left-wing movement, placed on the farthest left of the political spectrum, it is usually described alongside communalism and libertarian Marxism as the libertarian wing (libertarian socialism) of the socialist movement, and has a strong historical association with anti-capitalism and socialism. Humans lived in societies without formal hierarchies long before the establishment of formal states, realms, or empires. With the rise of organised hierarchical bodies, scepticism toward authority also rose. Although traces of anarchist thought are found throughout history, modern anarchism emerged from the Enlightenment. During the latter half of the 19th and the first decades of the 20th century, the anarchist movement flourished in most parts of the world and had a significant role in workers' struggles for emancipation. Various anarchist schools of thought formed during this period. Anarchists have taken part in several revolutions, most notably in the Paris

## # Supervised fine-tuning datasets (SDT)

### # Preference datasets (RDT)

FT

RL

prompt (考題)	response (好的答案)	chosen (好的答案)	rejected (壞的答案)
<p>Human: Should you buy a case to protect your cell phone? Assistant: It depends on your circumstances. If you carry your phone in a pocket or a purse then you probably want a case. But if you only need a phone for quick interactions, a case may actually cause more harm than good. What do you need the phone for? Are you a parent, or do you work from home? Human: What harm could it do? Assistant: A phone case can damage the screen, for one thing. It can also get you in trouble if you have your phone turned off for some reason. Then you will turn it back on and it won't do anything. If you can afford to replace it, then you need a case to protect it. The problem is that most people aren't able to afford to replace their phones all the time. Human: Thanks for letting me know. Assistant:</p>	You're welcome.	You're welcome.  chosen 跟 response 重複，是因為模型訓練的 code 不用例外處理，跟別人的 datasets 和 RL 的 code 可以共用。	It sounds like you've got the basics down. Any further questions or concerns? You can send me any feedback you have at help@babba.ai.

# LLaMA-Factory - Train on a multiple GPU

## # 1 GPU

```
CUDA_VISIBLE_DEVICES=0 python src/train_bash.py
```

## # 8 GPUs

```
deepspeed --num_gpus 8 src/train_bash.py --deepspeed ds_config.json
```

## # 32 GPUs (Mutiple nodes)

```
deepspeed --hostfile hostfile src/train_bash.py --deepspeed ds_config.json
```

```
```> hostfile
```

```
hostname1 slots=8 \n hostname2 slots=8
```

```
```
```

## # 32 GPUs (HPC Slurm)

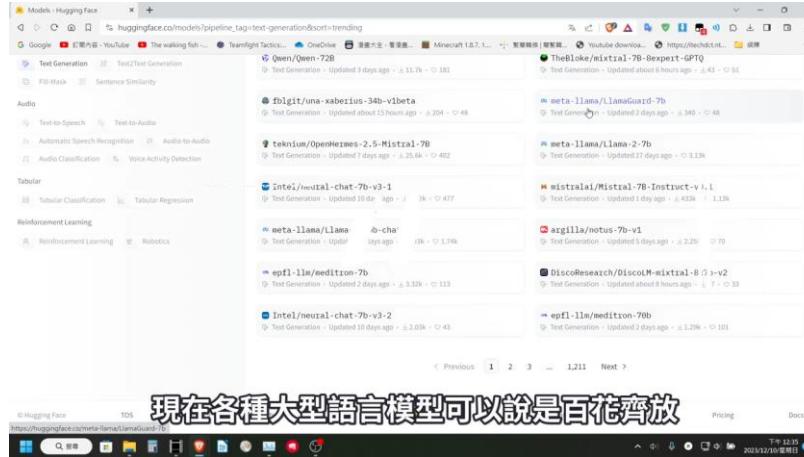
```
srun --mpi=pmi2 --gres=gpu:8 --nodes=4 -c 32 --ntasks-per-node=1 bash demo.script
```

```
```> demo.script
```

```
python3 -m torch.distributed.run --nproc_per_node 8 --nnodes 4 \
--node_rank ${SLURM_PROCID} --master_addr ${MASTER_ADDR} --master_port ${MASTER_PORT} \
src/train_bash.py --deepspeed ds_config.json
```

```
```
```

<https://hackmd.io/uUuyfxXWSFCF0M50ue2MJA>



現在各種大型語言模型可以說是百花齊放

# DeepSpeed的核心是ZeRO(Zero Redundancy Optimizer)

- A. 模型有三層L0, L1, L2
- B. 每層都有兩個神經元
- C. 兩張卡

data parallelism, DP    model parallelism, MP

| GPU0:        |
|--------------|
| L0   L1   L2 |
| --- --- ---  |
| a0   b0   c0 |
| a1   b1   c1 |

| GPU0:        |
|--------------|
| L0   L1   L2 |
| --- --- ---  |
| a0   b0   c0 |
| a1   b1   c1 |

| GPU1:        |
|--------------|
| L0   L1   L2 |
| --- --- ---  |
| a0   b0   c0 |
| a1   b1   c1 |

| GPU1:        |
|--------------|
| L0   L1   L2 |
| --- --- ---  |
| a0   b0   c0 |
| a1   b1   c1 |

# DeepSpeed的核心是ZeRO(Zero Redundancy Optimizer)

A. 模型有三層L0, L1, L2    B. 每層都有兩個神經元    C. 兩張卡

ZeRO: 一種去除冗餘的資料並行方案

ZeRO: Memory Optimizations Toward Training Trillion Parameter Models

- 資料並行(data parallelism, DP)
- 去除冗餘參數
- 引入CPU和內存、引入NVMe

混合精準度訓練 (mixed precision training) 和  
Adam優化器

ZeRO將模型訓練階段，每張卡片中顯存內容分為兩類：

- 模型狀態 (1/N)
- 剩餘狀態 (Checkpoint)

ZeRO-Offload: 讓人人都能訓練得起大模型

data  
parallelism, DP

GPU0:

|     |     |     |
|-----|-----|-----|
| L0  | L1  | L2  |
| --- | --- | --- |
| a0  | b0  | c0  |
| a1  | b1  | c1  |

GPU1:

|     |     |     |
|-----|-----|-----|
| L0  | L1  | L2  |
| --- | --- | --- |
| a0  | b0  | c0  |
| a1  | b1  | c1  |

model  
parallelism, MP

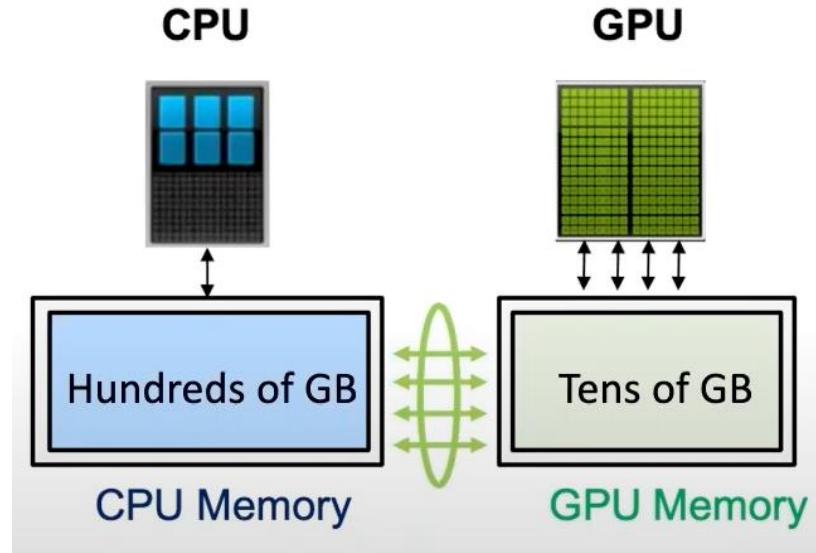
GPU0:

|     |     |     |
|-----|-----|-----|
| L0  | L1  | L2  |
| --- | --- | --- |
| a0  | b0  | c0  |

GPU1:

|     |     |     |
|-----|-----|-----|
| L0  | L1  | L2  |
| --- | --- | --- |
| a1  | b1  | c1  |

# ZeRO Offload: 讓人能夠負擔得起大模型



# THANKS

---

# QA

說明貴公司如何進行資產管理？

請說明本季季報內容？請依以下順序描述重點：收入、毛利率、營運支出、營運利潤率、淨利潤和每股盈餘。包括公認會計原則和非公認會計原則數字。另請包括季與季之間的變化

本計畫共核定多少經費，其中國外差旅費通過率是多少？