

歡迎大家報名  
ACML 2025



**ACML 2025** TAIPEI TAIWAN  
December 9 - 12

**The 17<sup>th</sup> Asian Conference on Machine Learning**

**SPECIAL OFFER** [www.acml-conf.org/2025](http://www.acml-conf.org/2025)

**LOCAL STUDENTS Only US \$100!**  
The exact amount will be shown after logging into the registration system.

**Register now**



Keynote Speakers



**Craig Knoblock**  
Keston Executive Director of the Information Sciences Institute at University of Southern California



**Kun Zhang**  
Carnegie Mellon University

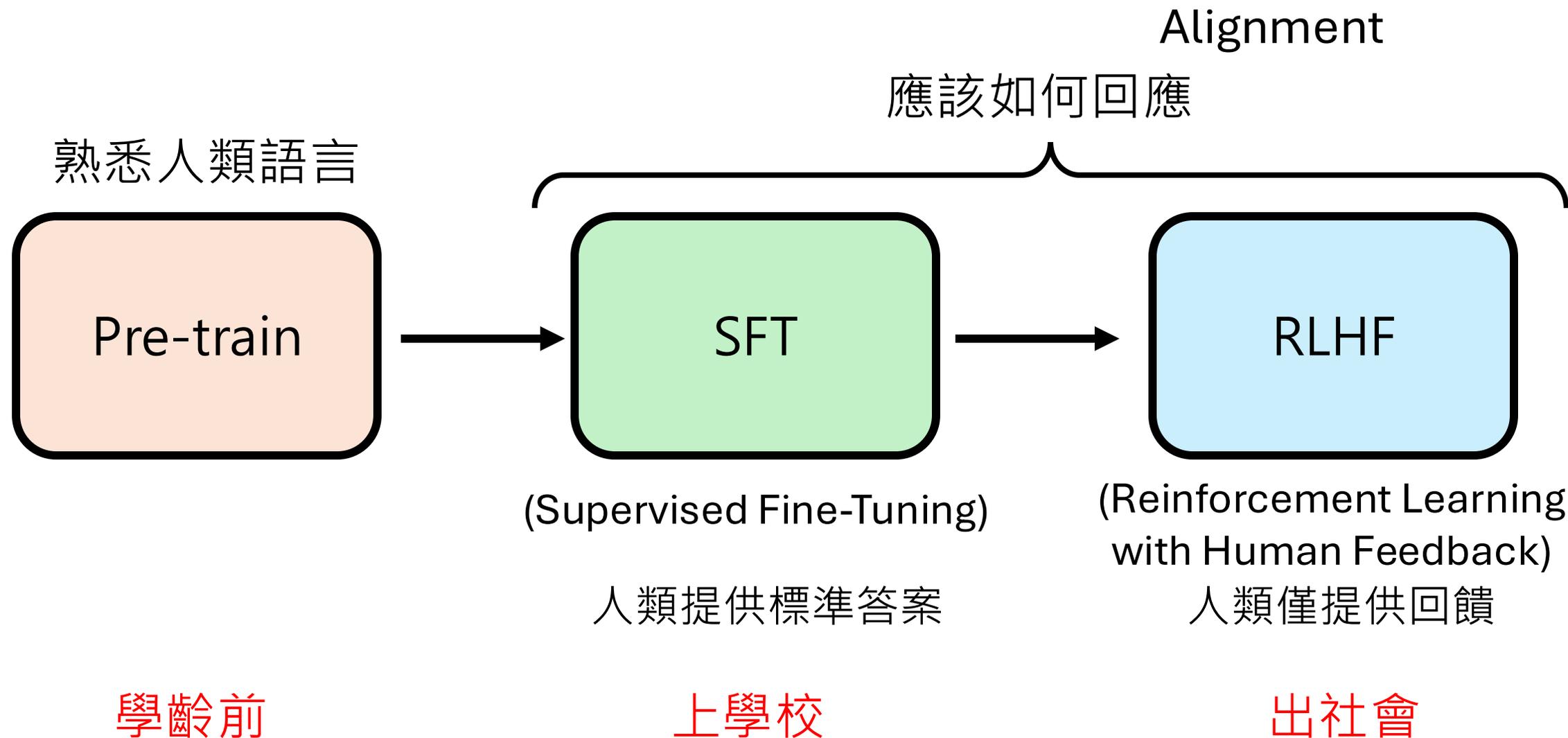


**Aja Huang**  
Senior Staff Research Scientist at Google DeepMind

課程 14:23 開始

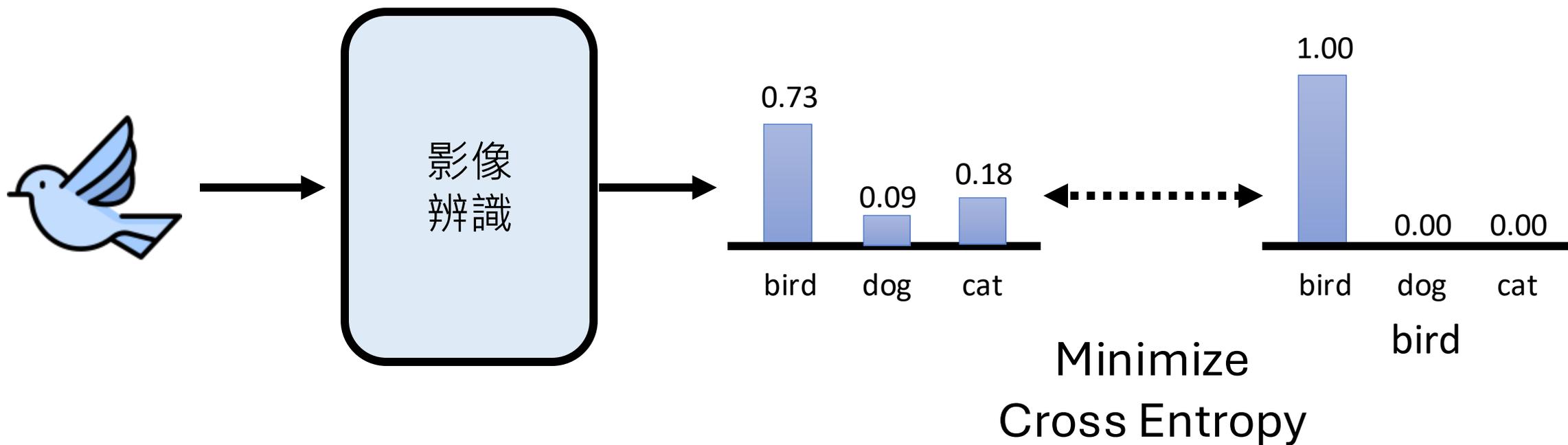
# 大型語言模型的學習歷程

# 大型語言模型的學習歷程



# 每一個階段都在學文字接龍

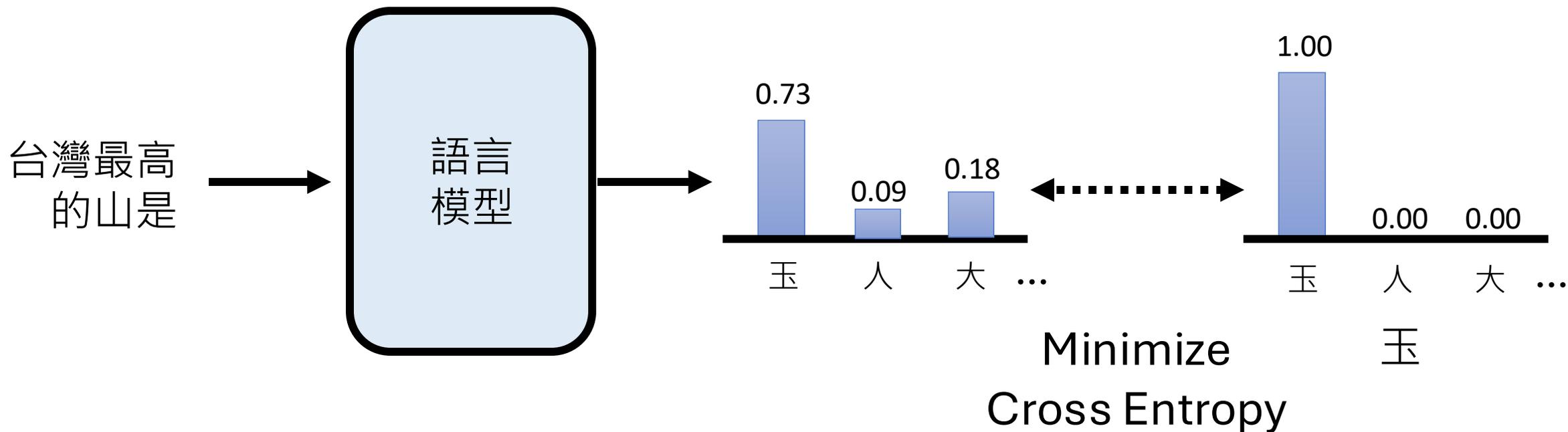
- 文字接龍就是分類問題



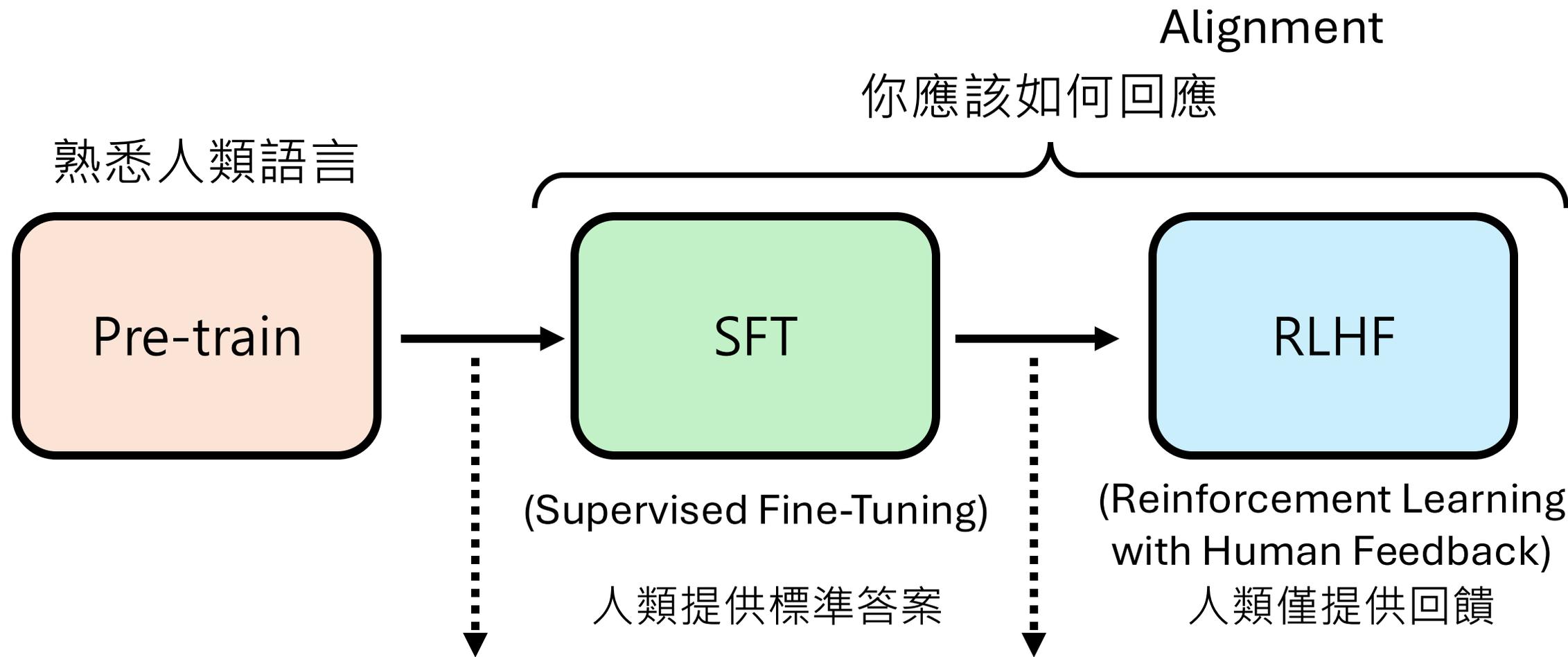
# 每一個階段都在學文字接龍

- 文字接龍就是分類問題

每一個 Token 都是一個類別  
Vocabulary size 就是類別的數目

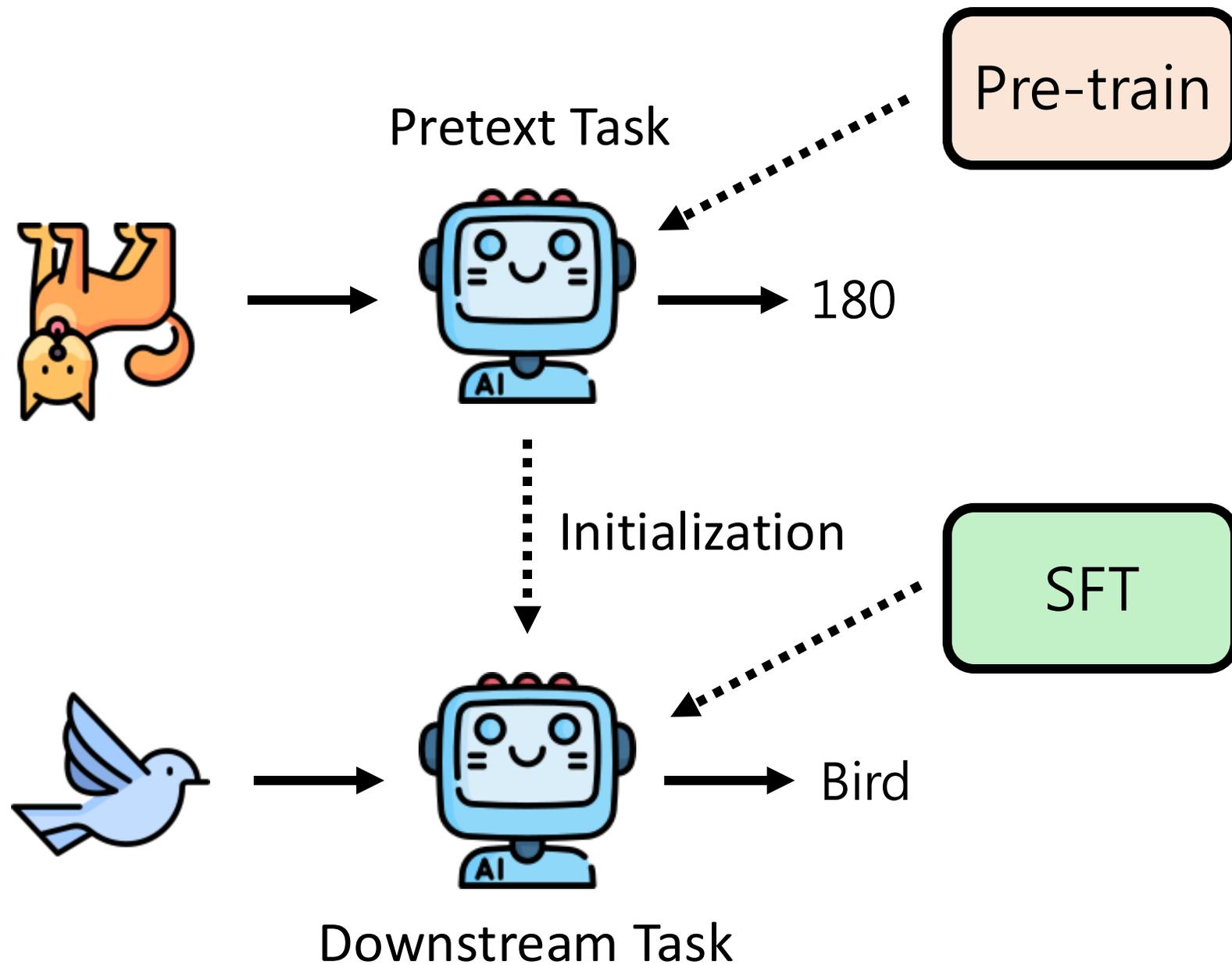


# 大型語言模型的學習歷程

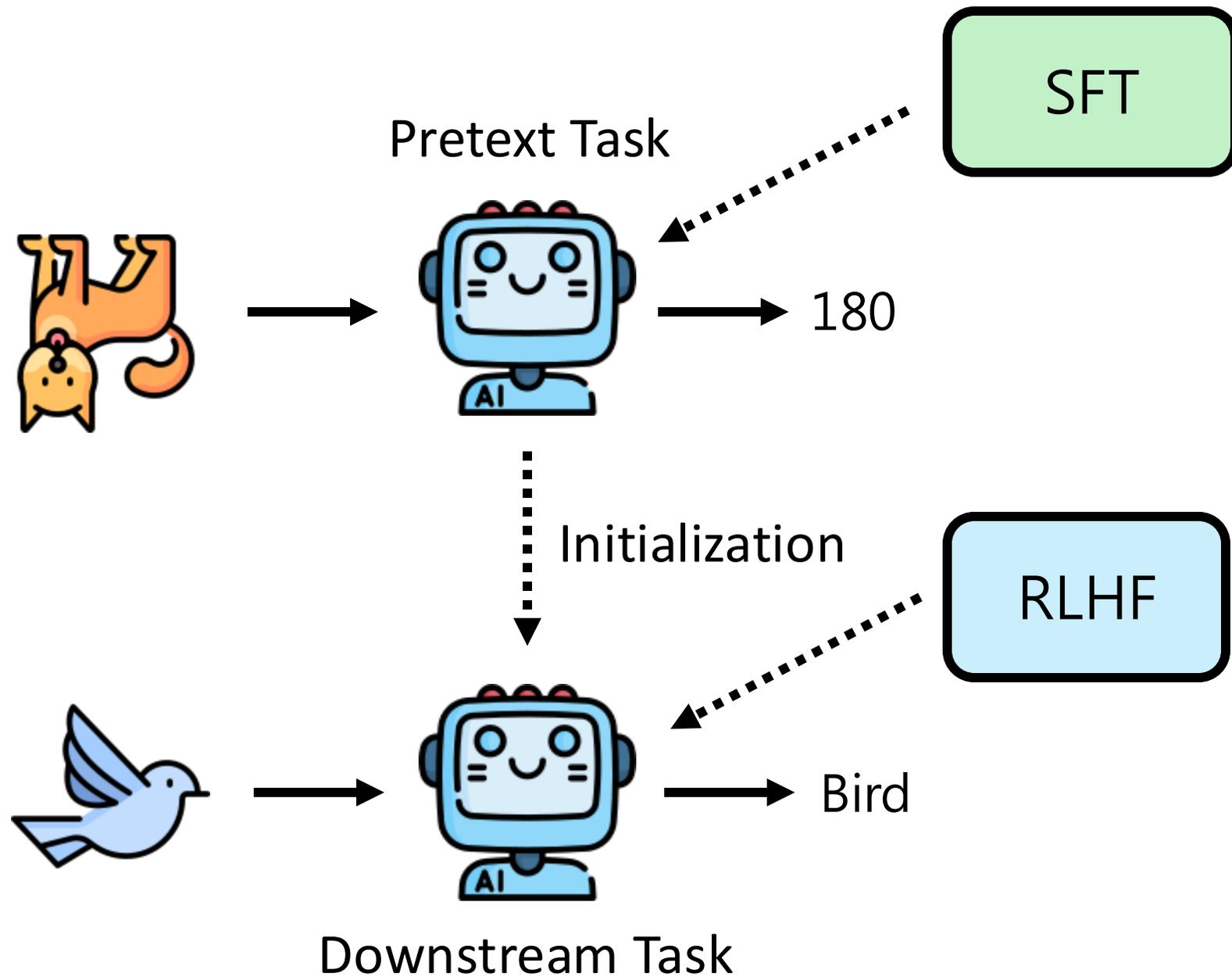


每一個階段都拿前一個階段訓練出的參數做為初始 (Initialization)

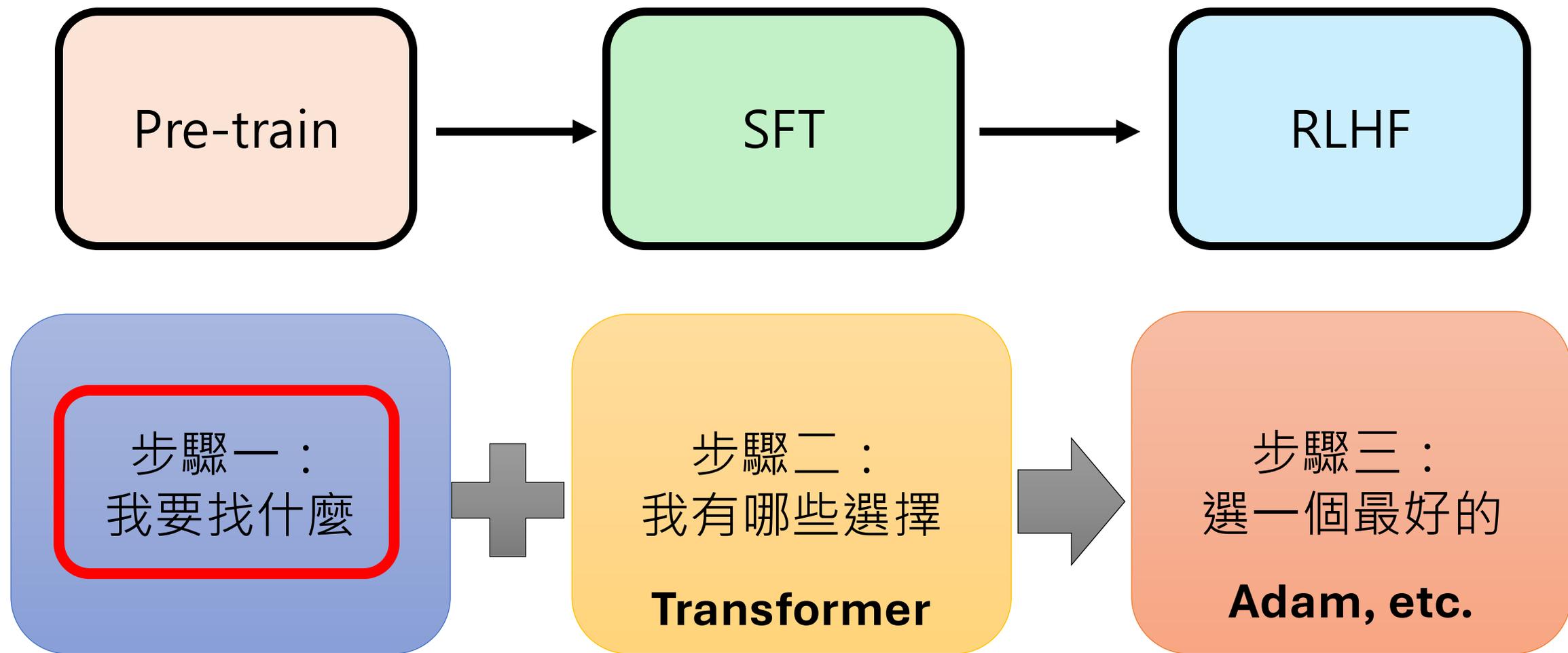
# Initialization



# Initialization

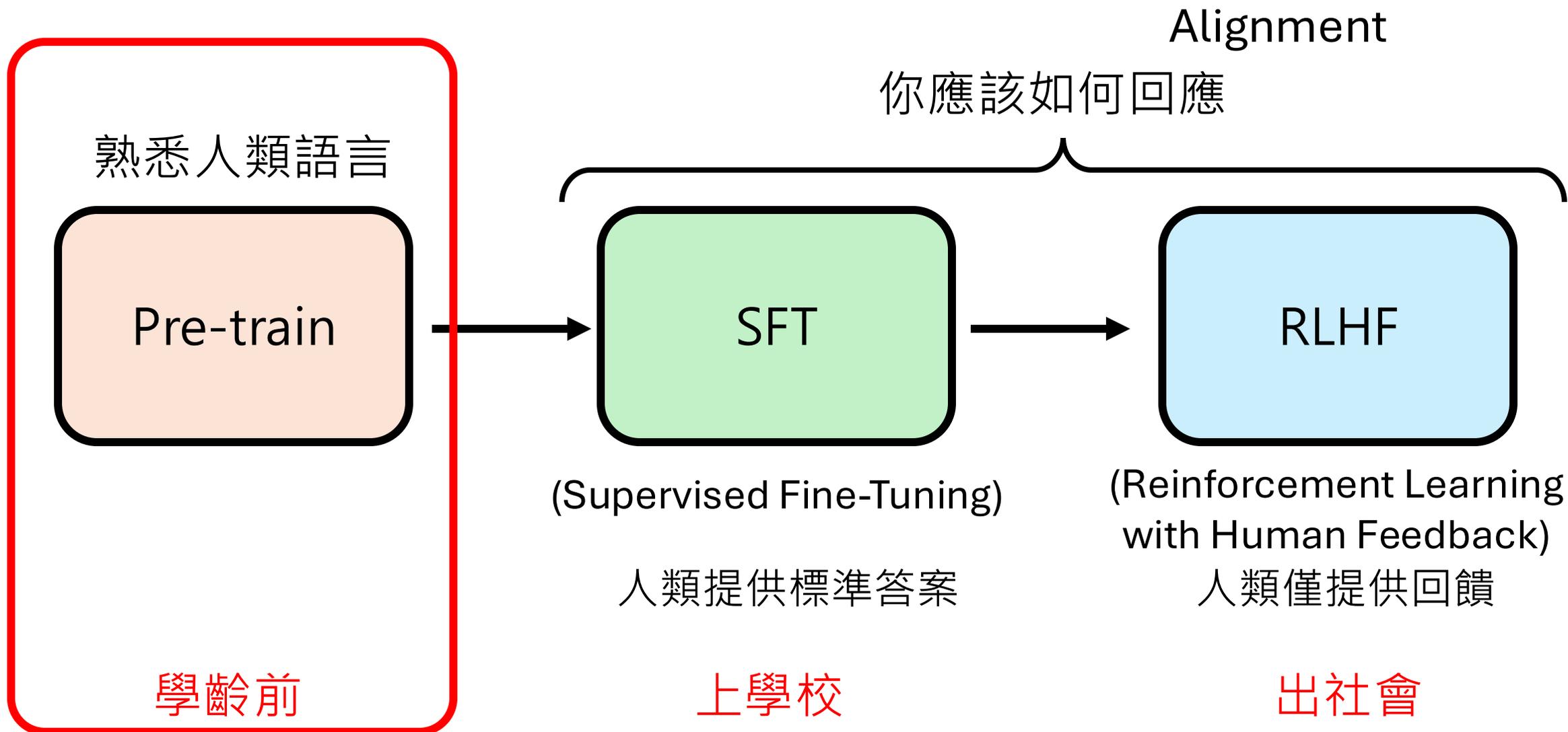


# 每一個階段都在學文字接龍



教材(訓練資料)不同

# 大型語言模型的學習歷程



# 學會正確地做文字接龍需要大量資料

<https://arxiv.org/abs/2011.04946>

## 語言知識

這個人突然就

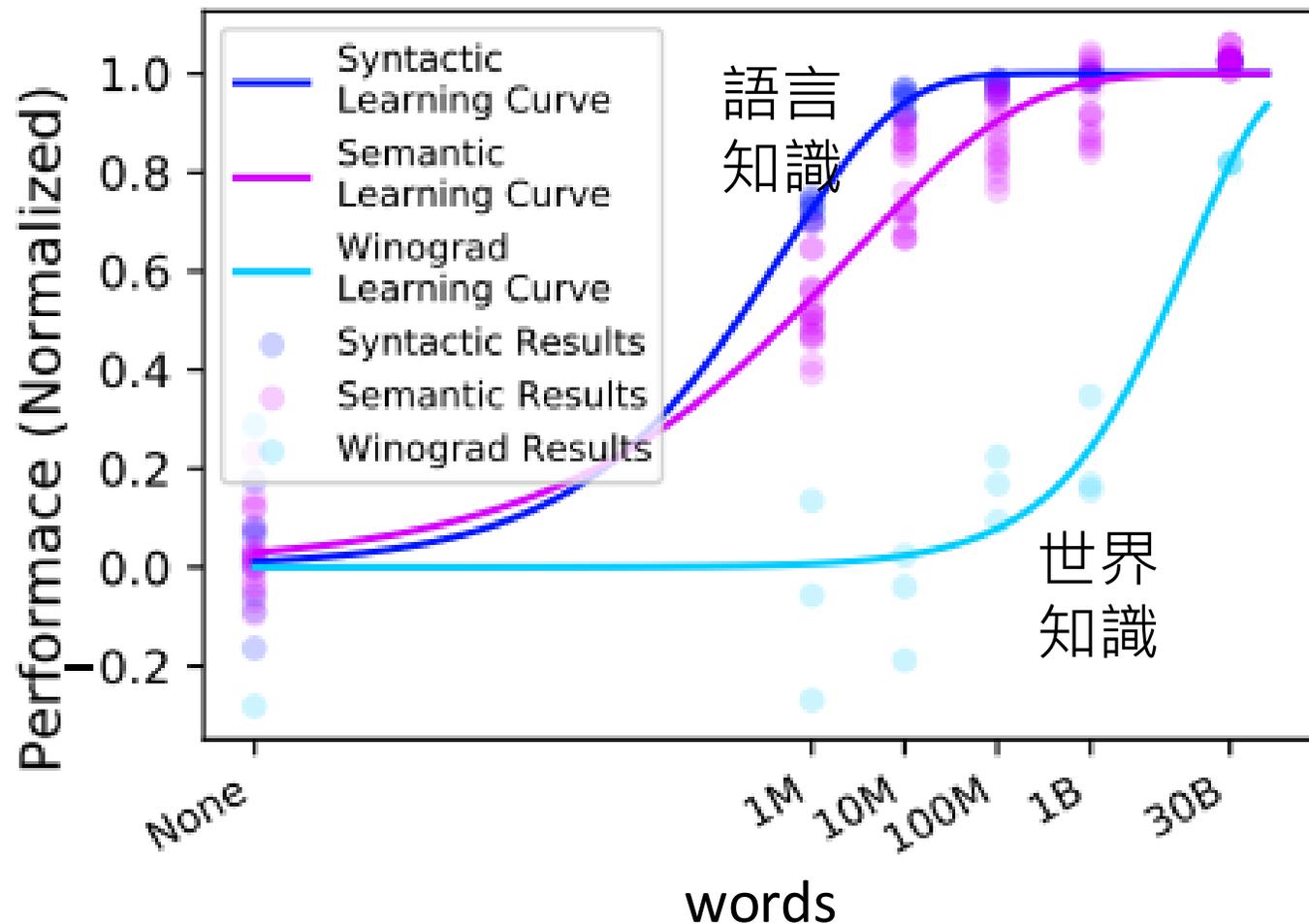
- 跑
- 飛
- 的

## 世界知識

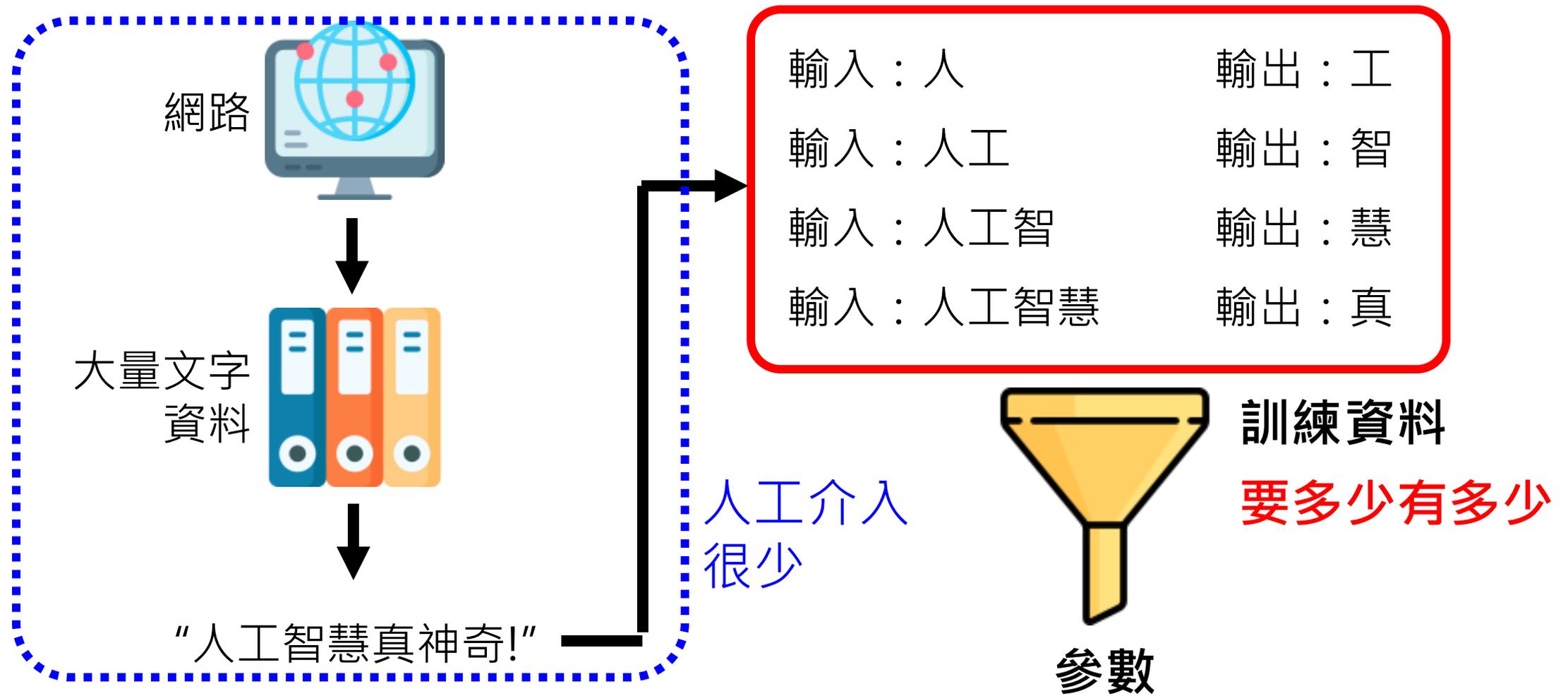
水的沸點是攝氏

- 一百度
- 五十度

在低壓下



# 任何資料都拿來學文字接龍



Self-supervised Learning (自督導式學習)

# 現在都用多大的資料來預訓練語言模型？

## LLaMA 3

<https://arxiv.org/abs/2407.21783>

- **Data.** Compared to prior versions of Llama (Touvron et al., 2023a,b), we improved both the quantity and quality of the data we use for pre-training and post-training. These improvements include the development of more careful pre-processing and curation pipelines for pre-training data and the development of more rigorous quality assurance and filtering approaches for post-training data. We pre-train Llama 3 on a corpus of about 15T multilingual tokens, compared to 1.8T tokens for Llama 2.

## DeepSeek-V3

<https://arxiv.org/abs/2412.19437>

We present DeepSeek-V3, a strong Mixture-of-Experts (MoE) language model with 671B total parameters with 37B activated for each token. To achieve efficient inference and cost-effective training, DeepSeek-V3 adopts Multi-head Latent Attention (MLA) and DeepSeekMoE architectures, which were thoroughly validated in DeepSeek-V2. Furthermore, DeepSeek-V3 pioneers an auxiliary-loss-free strategy for load balancing and sets a multi-token prediction training objective for stronger performance. We pre-train DeepSeek-V3 on 14.8 trillion diverse and high-quality tokens, followed by Supervised Fine-Tuning and Reinforcement Learning stages to fully harness its capabilities. Comprehensive evaluations reveal that DeepSeek-V3 outperforms other open-source models and achieves performance comparable to leading closed-source models. Despite its excellent performance, DeepSeek-V3 requires only 2.788M H800 GPU hours for its full training. In addition, its training process is remarkably stable. Throughout the entire training process, we did not experience any irrecoverable loss spikes or perform any rollbacks.

# 15T token 是什麼概念

- 把 15T 的 token 印出來
- 假設 100 張紙厚 1 公分
- 厚度有 1500 公里

1000 tokens

## 1 Introduction

A central question in the discussion of large language models (LLMs) concerns the extent to which they *memorize* their training data versus how they *generalize* to new tasks and settings. Most practitioners seem to (at least informally) believe that LLMs do some degree of both: they *clearly* memorize parts of the training data—for example, are often able to reproduce large portions of training data verbatim [Carlini et al., 2022]—but they also seem to learn from this data, allowing them to generalize to new settings. The precise extent to which they do one or the other has massive implications for the practical and legal aspects of such models [Cooper et al., 2023]. Do LLMs truly produce new content, or do they only remix their training data? Should the act of training on copyrighted data be deemed unfair use of data, or should fair use be judged by the model’s memorization? With respect to people, we distinguish plagiarizing content from learning from it, but how should this extend to LLMs? The answer to such questions inherently relates to the extent to which LLMs memorize their training data.

However, even defining memorization for LLMs is challenging and many existing definitions leave a lot to be desired. Certain formulations claim that a passage from the training data is memorized if the LLM can reproduce it exactly [Nave et al., 2022]. However, this ignores situations where, for instance, a prompt instructs the model to exactly repeat some phrase. Other formulations define memorization by whether or not prompting an LLM with a portion of text from the training set results in the completion of that training datum [Carlini et al., 2022]. But these formulations rely fundamentally on the completions being a certain size, and typically very lengthy generations are required for sufficient certainty of memorization. More crucially, these definitions are too permissive because they ignore situations where model developers can (for legal compliance) post-hoc “align” an LLM by instructing their models not to produce certain copyrighted content [Ippolito et al., 2022]. But has such an instructed model really *not memorized* the sample in question, or does the model still contain all the information about the datum in its weights while it hides behind an illusion of compliance? Asking such questions becomes critical because this illusion of “unlearning” can often be easily broken as we show in Sections 4.1 and 4.3.

In this work, we propose a new definition of memorization based on a compression argument. Our definition posits that a phrase present in the training data is memorized if we can make the model reproduce the phrase using a prompt (which) is shorter than the phrase itself. Operationalizing this definition requires finding the shortest adversarial input prompt that is specifically optimized to produce a target output. We call this ratio of input to output tokens the Adversarial Compression Ratio (ACR). In other words, memorization is inherently tied to whether a certain output can be represented in a compressed form, beyond what language models can do with typical text. We argue that such a definition provides an intuitive notion of memorization—if a certain phrase exists within the LLM training data (e.g., is not itself generated text) and it can be reproduced with fewer input tokens than output tokens, then the phrase must be stored somehow within the weights of the LLM. Although it may be more natural to consider compression in terms of the LLM-based notions of input/output perplexity, we argue that a simple compression ratio based on input/output token counts provides a more intuitive explanation to non-technical audiences, and has the potential to serve as a legal basis for important questions about memorization and permissible data use.

In addition to its intuitive nature, our definition has several other desirable qualities. We show that it appropriately ascribes many famous quotes as being memorized by existing LLMs (i.e. they have high ACR values). On the other hand, we find that text not in the training data of an LLM, such as samples posted on the internet after the training period, are not compressible, that is their ACR is low.

We examine several unlearning methods using ACR to show that they do not substantially affect the memorization of the model. That is, even after explicit finetuning, models asked to “forget” certain pieces of content are still able to reproduce them with a high ACR—in fact, not much smaller than with the original model. Our approach provides a simple and practical perspective on what memorization can mean, providing a useful tool for functional and legal analysis of LLMs.

## 2 Do We Really Need Another Notion of Memorization?

With LLMs ingesting more and more data, questions about their memorization are attracting attention [e.g. Carlini et al., 2019, 2022, Nave et al., 2022, Zhang et al., 2022]. There remains a pressing need

# 15T token 是什麼概念

- 把 15T 的 token 印出來
- 假設 100 張紙厚 1 公分
- 厚度有 1500 公里

假如你每十秒鐘可以讀一頁

需要閱讀 4756 年

假設有人從商朝開始讀起

到現在都還沒讀完

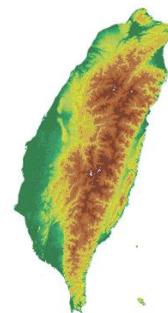
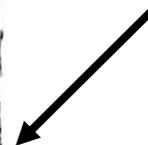
外太空



聖母峰



LLaMA3,  
DeepSeek-V3  
閱讀過的資料量



# 可以從哪裡取得大量資料

**F**  **ine**  **web**

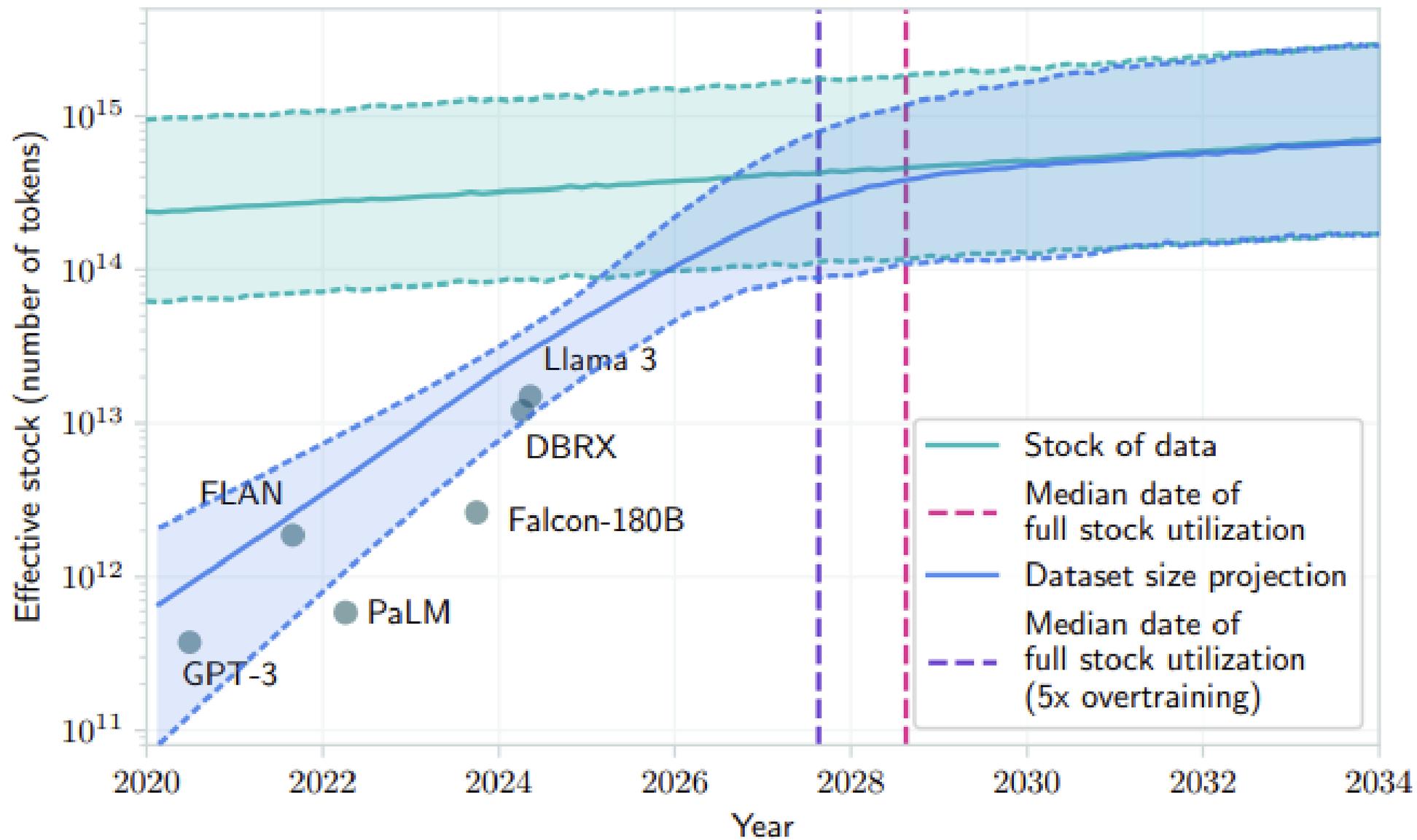
The finest collection of data the web has to offer



**15-trillion tokens,  
44TB disk space**

<https://arxiv.org/abs/2406.17557>

<https://huggingface.co/HuggingFaceFW>



# 不是死背 (而是壓縮)

GPT-5

這一句之後的原文如下：

請根據魯迅《孔乙己》原文接續以下文字：孔乙己一到店，所有喝酒的人便都看着他笑

孔乙己一到店，所有喝酒的人便都看着他笑，有的叫道，「孔乙己，你又偷書了！」孔乙己便回過臉去，用那種苦笑的神氣說，「偷書？……讀書人的事，能算偷麼？」接連便是難懂的話，什麼「之乎者也」之類，引得眾人都哄笑起來。店內外充滿了快活的空氣。

原文：孔乙己一到店，所有喝酒的人便都看着他笑，有的叫道：『孔乙己，你臉上又添上新傷疤了！』他不回答，對櫃裏說：『溫兩碗酒，要一碟茴香豆。』便排出九文大錢。他們又故意的高聲嚷道：『你一定又偷了人家的東西了！』孔乙己睜大眼睛說：『你怎麼這樣憑空污人清白……』『什麼清白？我前天親眼見你偷了何家的書，吊着打。』孔乙己便漲紅了臉，額上的青筋條條綻出，爭辯道：『竊書不能算偷……竊書！……讀書人的事，能算偷麼？』接連便是難懂的話，什麼『君子固窮』，什麼『者乎』之類，引得眾人都鬨笑起來；店內外充滿了快活的空氣。

# 不是死背 (而是壓縮)

射鵰英雄傳原文

## GPT-5 第一回 風雪驚變

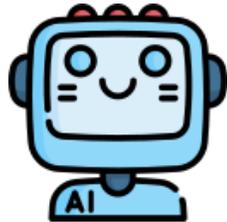
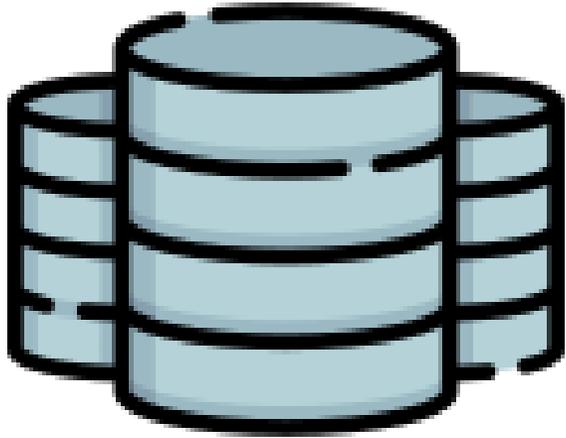
話說北宋末年，徽、欽二宗被擄，北方陷於金人之手，南宋偏安江南，朝政昏庸，民生困苦。其時江湖上卻又是豪傑輩出，風起雲湧。

那一年冬天，華山論劍，五位絕頂高手齊聚華山之巔，比武論道，欲決天下武學第一。東邪黃藥師、西毒歐陽鋒、南帝段智興、北丐洪七公、中神通王重陽，各展神功，鬥得天昏地暗，山搖地動。終於王重陽以一招「先天功」勝出，奪得《九陰真經》。自此五人各散東西，武林中風波不息。

十餘年後，江南牛家村中，有兩個青年漁夫，一名郭嘯天，一名楊鐵心，情同兄弟。二人皆是忠勇之士，常痛罵金人，誓言報國雪恥。那年冬日，大雪紛飛，兩人正在村頭酒店飲酒，忽聞外面馬蹄聲響，一隊金兵押著幾個漢人而過……

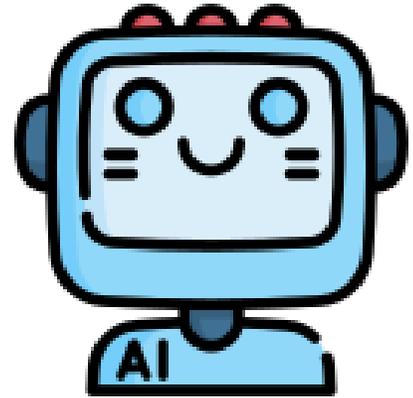
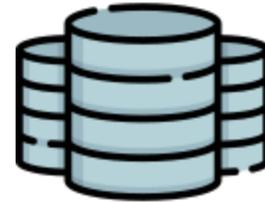
# 資料越多越好嗎？

- 算力是有限的



讀過比較多教材  
Prevent Overfitting

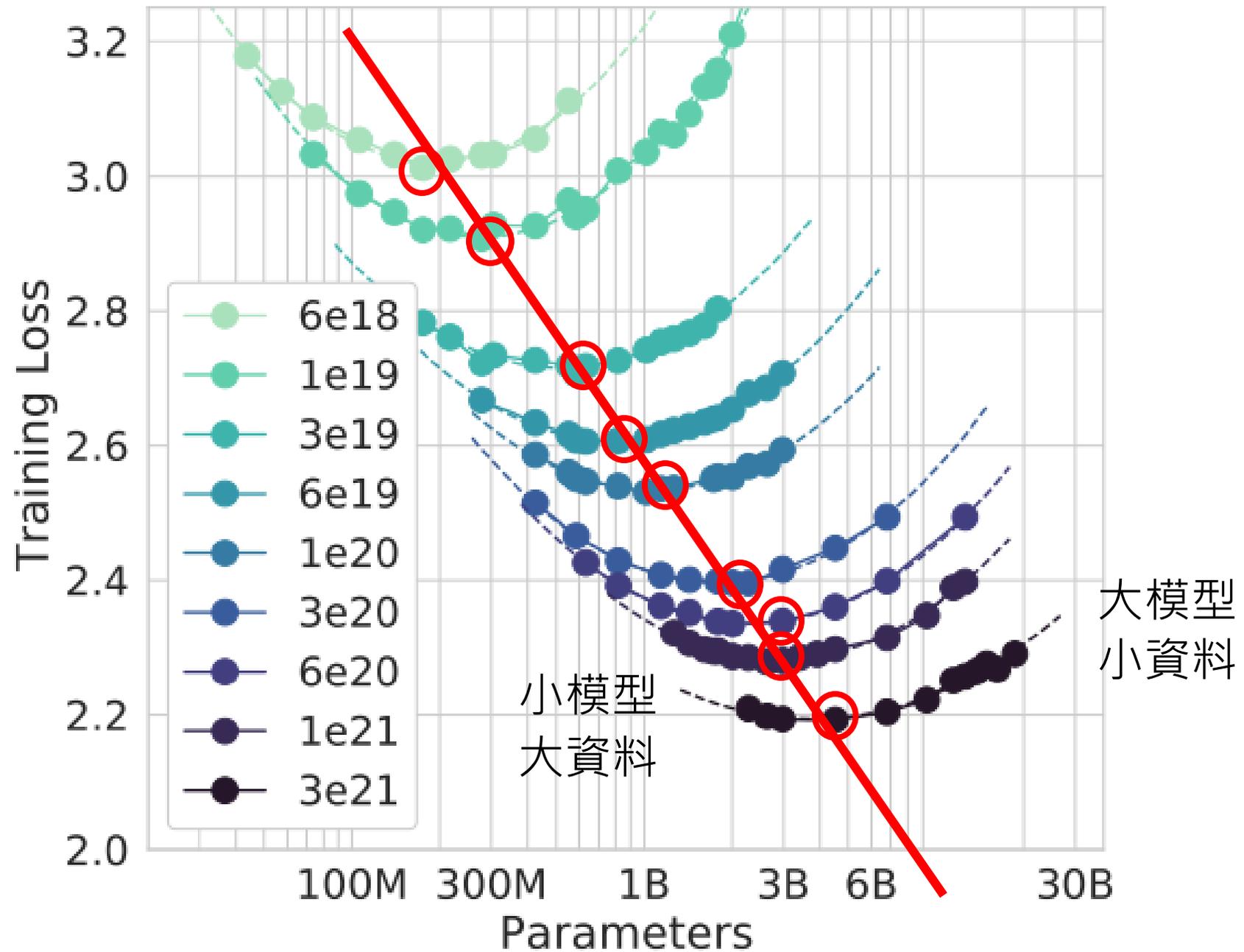
選哪一個？



天資比較聰明  
More functions can be selected.

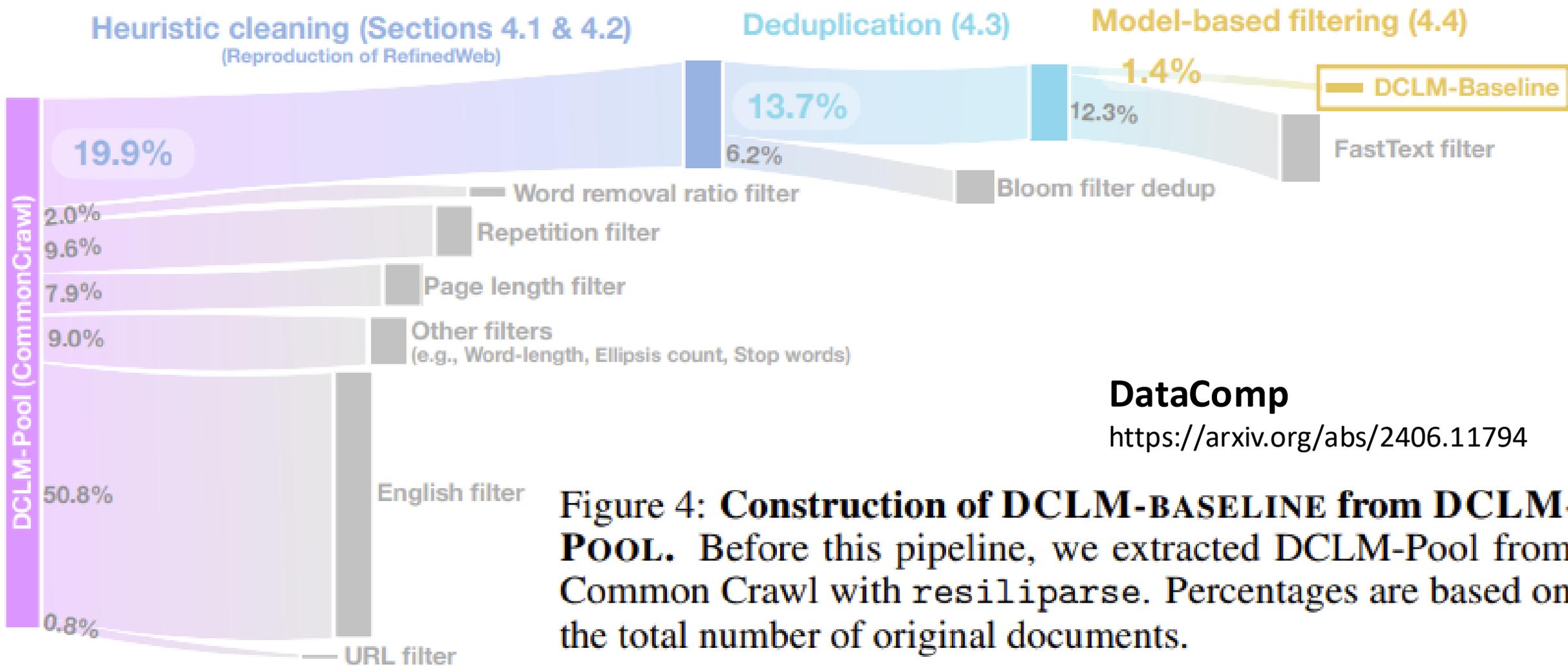
# Chinchilla Scaling Law

<https://arxiv.org/abs/2203.15556>



# 資料越多越好嗎？還要考慮品質

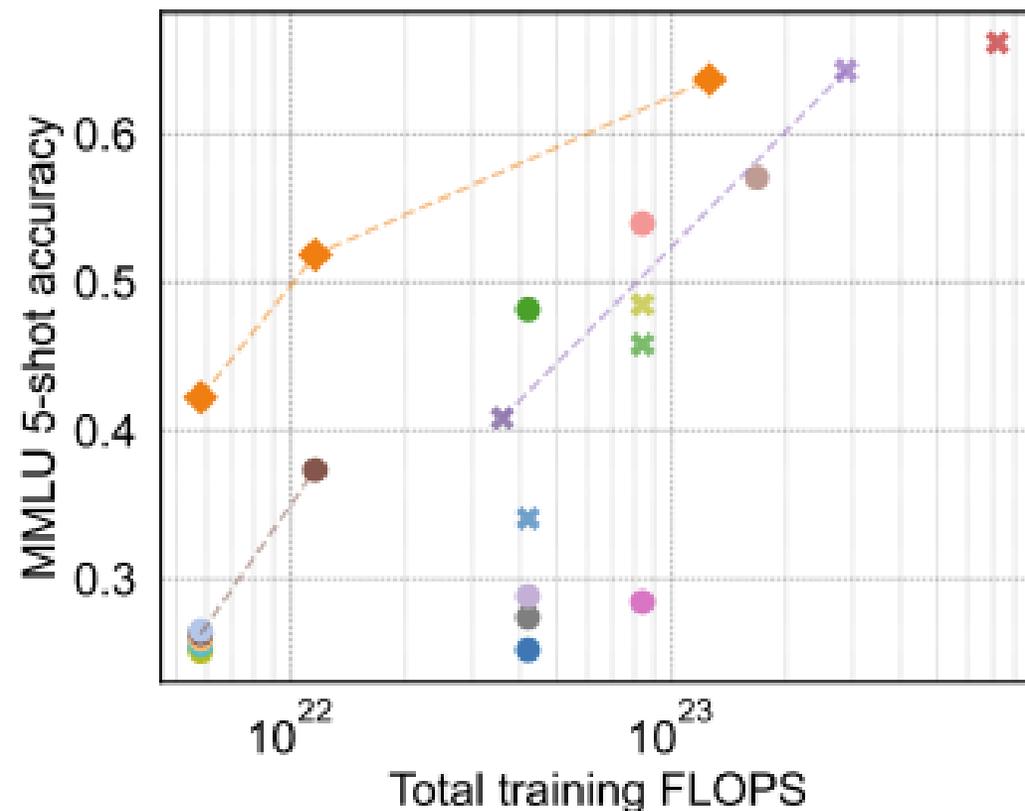
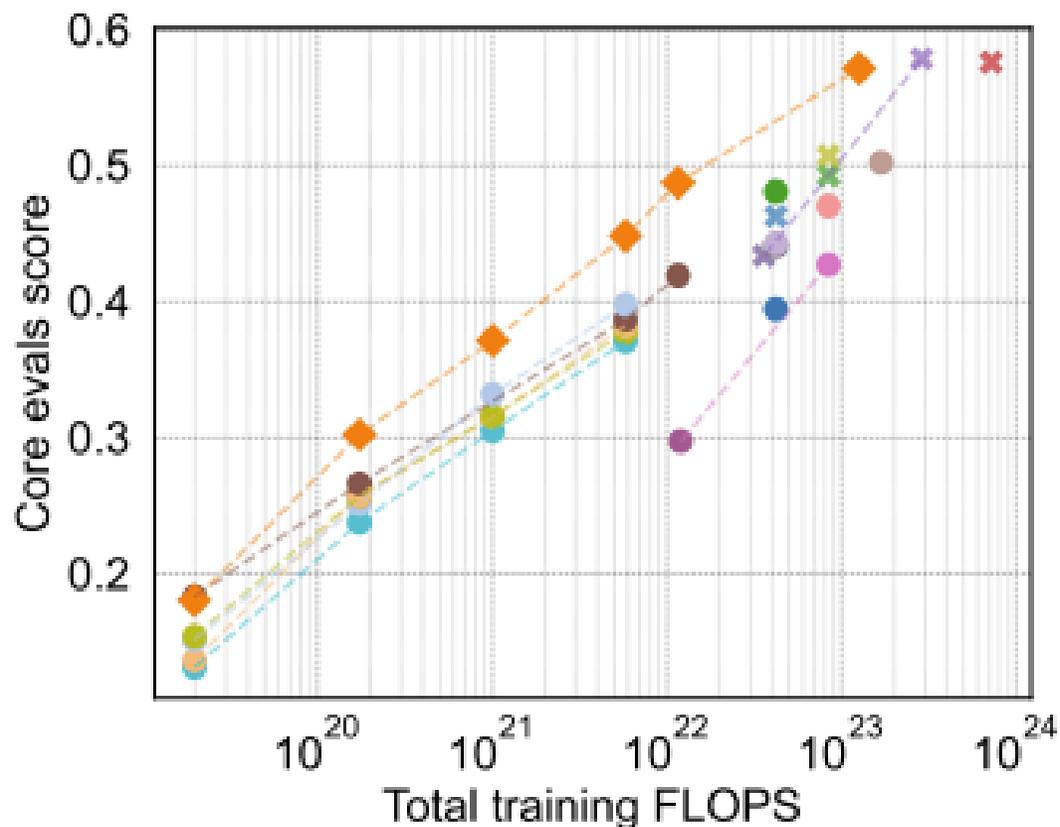
The screenshot shows a Reddit interface. At the top, the search bar contains 'r/microwavegang' and '在 r/microwavegang 中搜尋'. The subreddit name 'r/microwavegang' is visible in the top right. The post is from user 'Bobbydibi' 7 months ago. The main content of the post is a large block of 'm' characters, with a few 'M' characters interspersed. The right sidebar shows the subreddit's description: 'welcome to the microwave gang! here we talk about topics such as "MMMMMMMMMMMMMMMMmmMMMMMMMM...'. It also displays statistics: 2112 weekly visitors and 32 weekly contributions. The subreddit rules section shows '1 must be microwave based'. The moderators listed are 'u/bl1zzardTHEone' (microwave mod) and 'LethargicTHEguy'.



**DataComp**

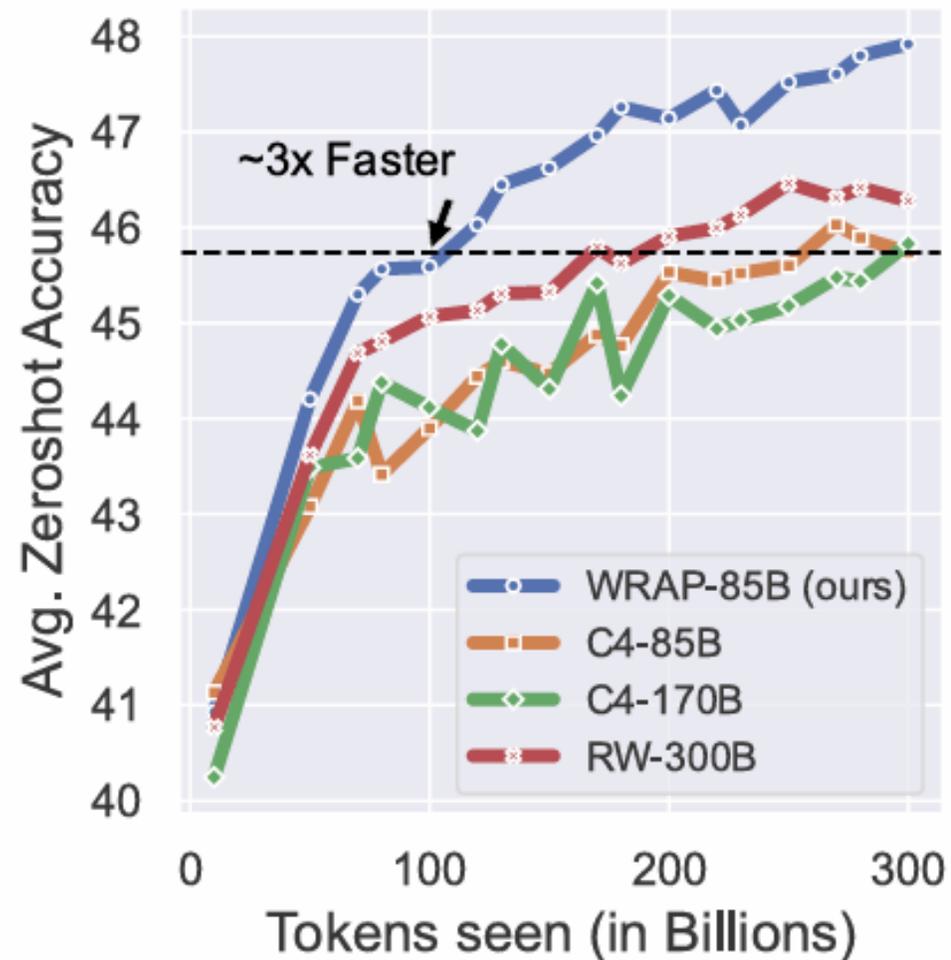
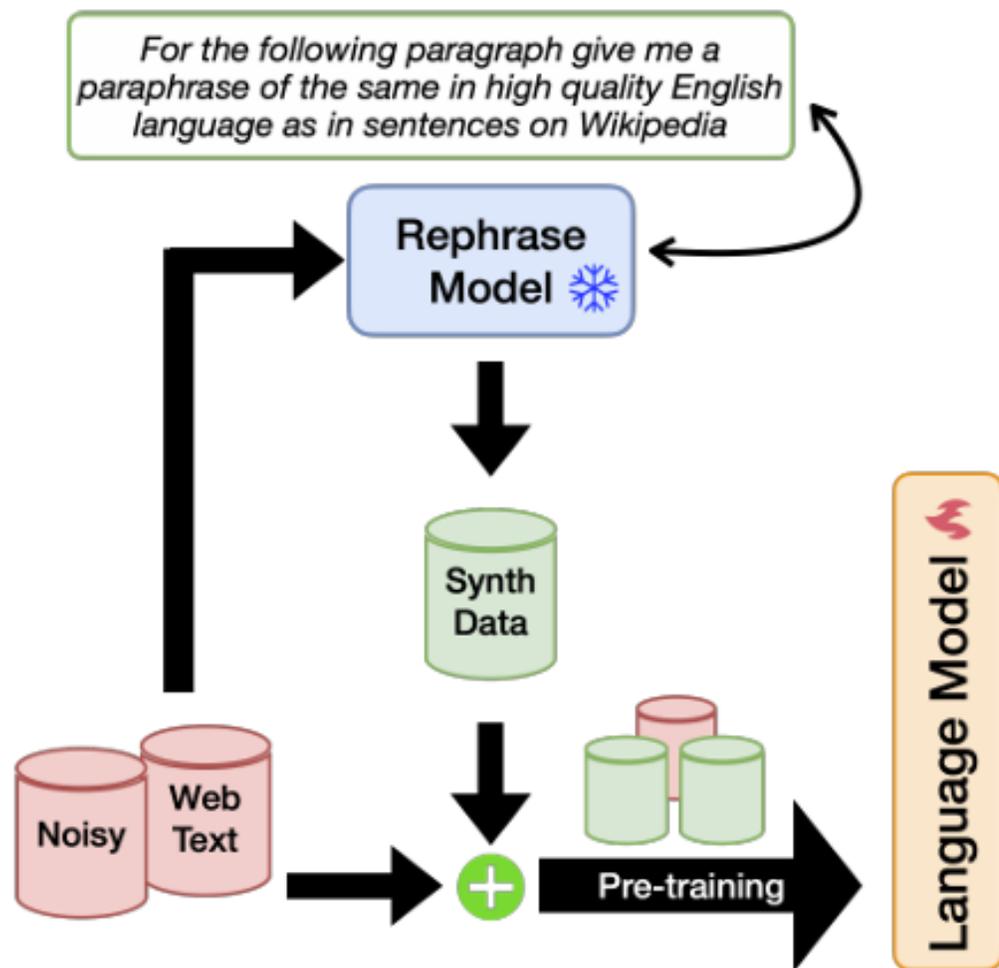
<https://arxiv.org/abs/2406.11794>

**Figure 4: Construction of DCLM-BASELINE from DCLM-POOL.** Before this pipeline, we extracted DCLM-Pool from Common Crawl with resiliiparse. Percentages are based on the total number of original documents.



# 用 LLM 幫忙清理資料

Rephrasing the Web  
<https://arxiv.org/abs/2401.16380>



# 用網路資料預訓練的模型常常不回答問題

What is the purpose of the list C in the code below?

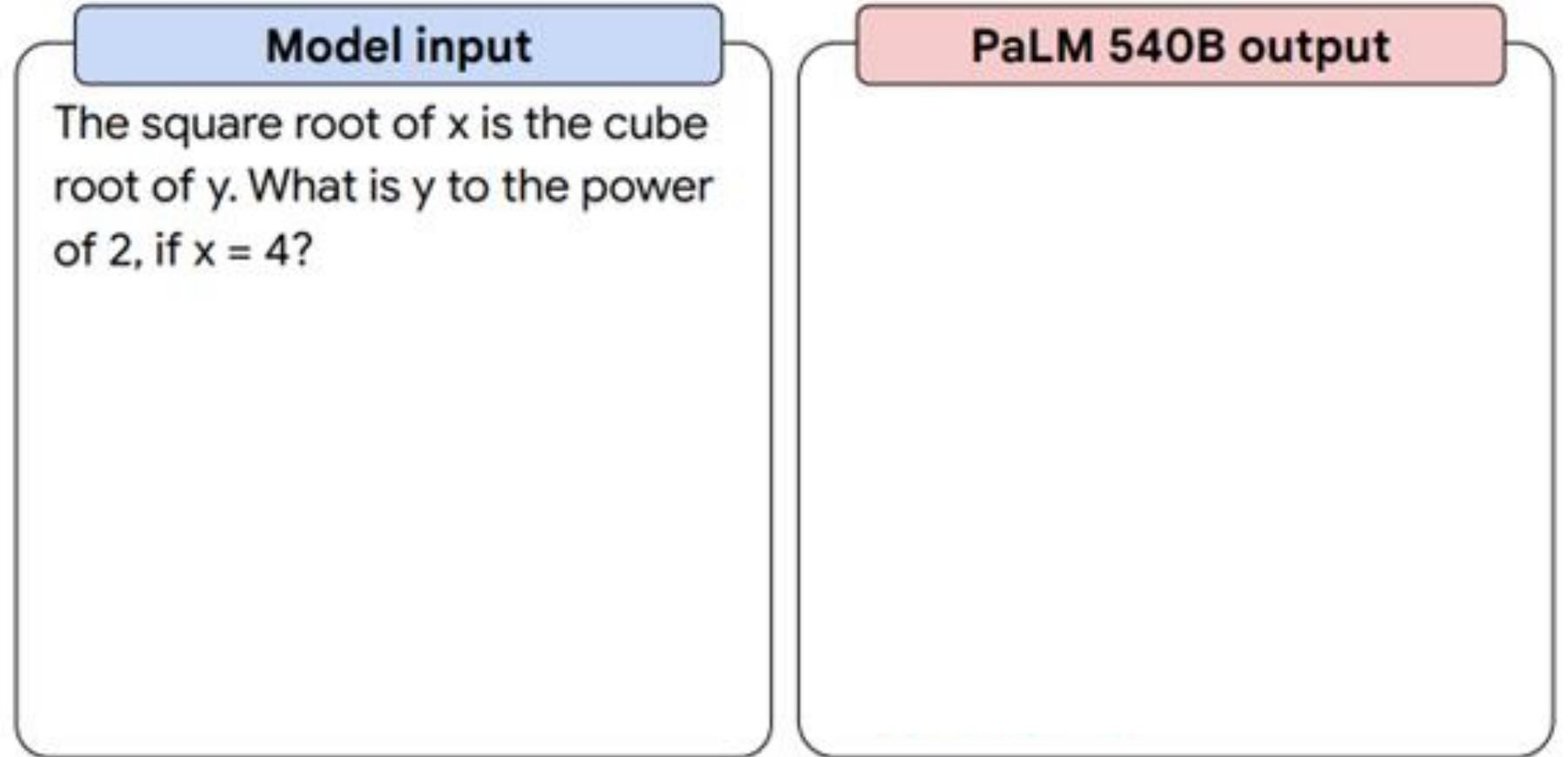
```
def binomial_coefficient(n, r):  
    C = [0 for i in range(r + 1)];  
    C[0] = 1;  
    for i in range(1, n + 1):  
        j = min(i, r);  
        while j > 0:  
            C[j] += C[j - 1];  
            j -= 1;  
    return C[r]
```

GPT-3

176B

- A. to store the value of C[0]
- B. to store the value of C[1]
- C. to store the value of C[i]
- D. to store the value of C[i - 1]

# 用網路資料預訓練的模型常常不回答問題



# 為什麼語言模型不能好好回答問題？

- 因為其實網路資料也沒這樣教他 .....

"台灣最高的山是哪座山"

勒星頓中文學校  
<https://lcs-chinese.org> > 2018\_G789\_QuestionAnswer

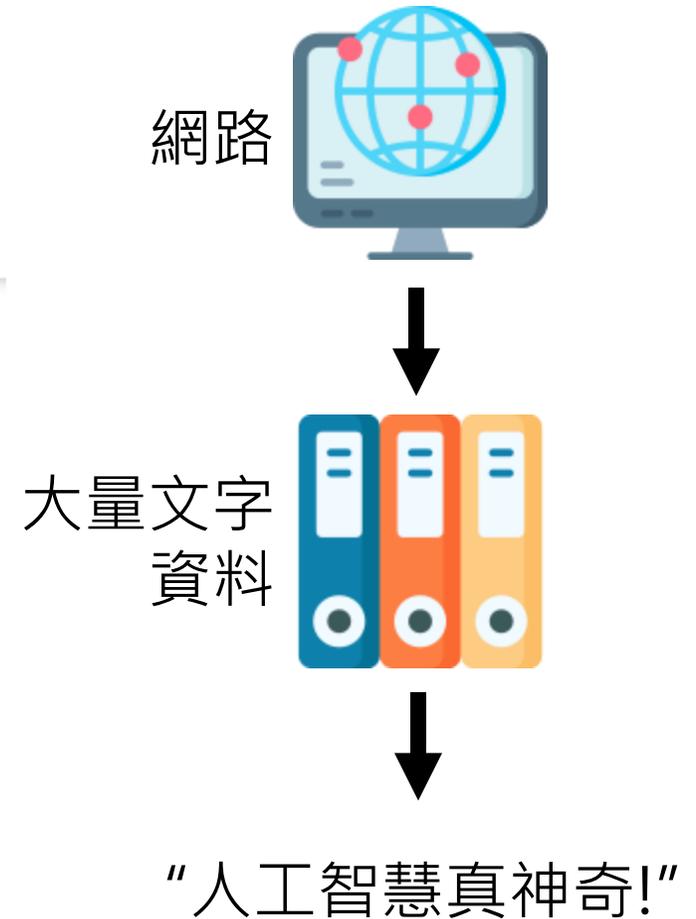
班學生姓名：\_\_\_\_\_ 考試成績： /100

34. ( 2 ) 台灣最高的山是哪座山? (①雪山②玉山③阿里山) · 35. ( 2 ) 中國最早的文字始於哪個朝代? (①夏②商③周) · 36. ( 1 ) 科學老師常常帶我們到LAB 做實驗 ...

健康跟著走  
<https://info.todohealth.com> > ... > 台灣最高的山line旅遊

台灣最高的山是??

玉山位於臺灣中部的 ... 歡迎來到LINE旅遊很高興你接受了收藏冒險王的挑戰！本次活動共有「2個... 聰明的你，知道「台灣最高的山是哪座山」嗎？知道答案的朋友，請 ..., ...



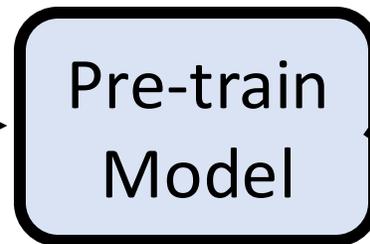
# Pre-train 模型是璞玉、還需要精雕細琢

- 其實 Pre-train 已經有機會回答正確答案

Reasoning with Sampling: Your Base Model is Smarter Than You Think

<https://arxiv.org/abs/2510.14901>

台灣最高的山是哪座？



誰來告訴我呀？

第二高的又是哪座？

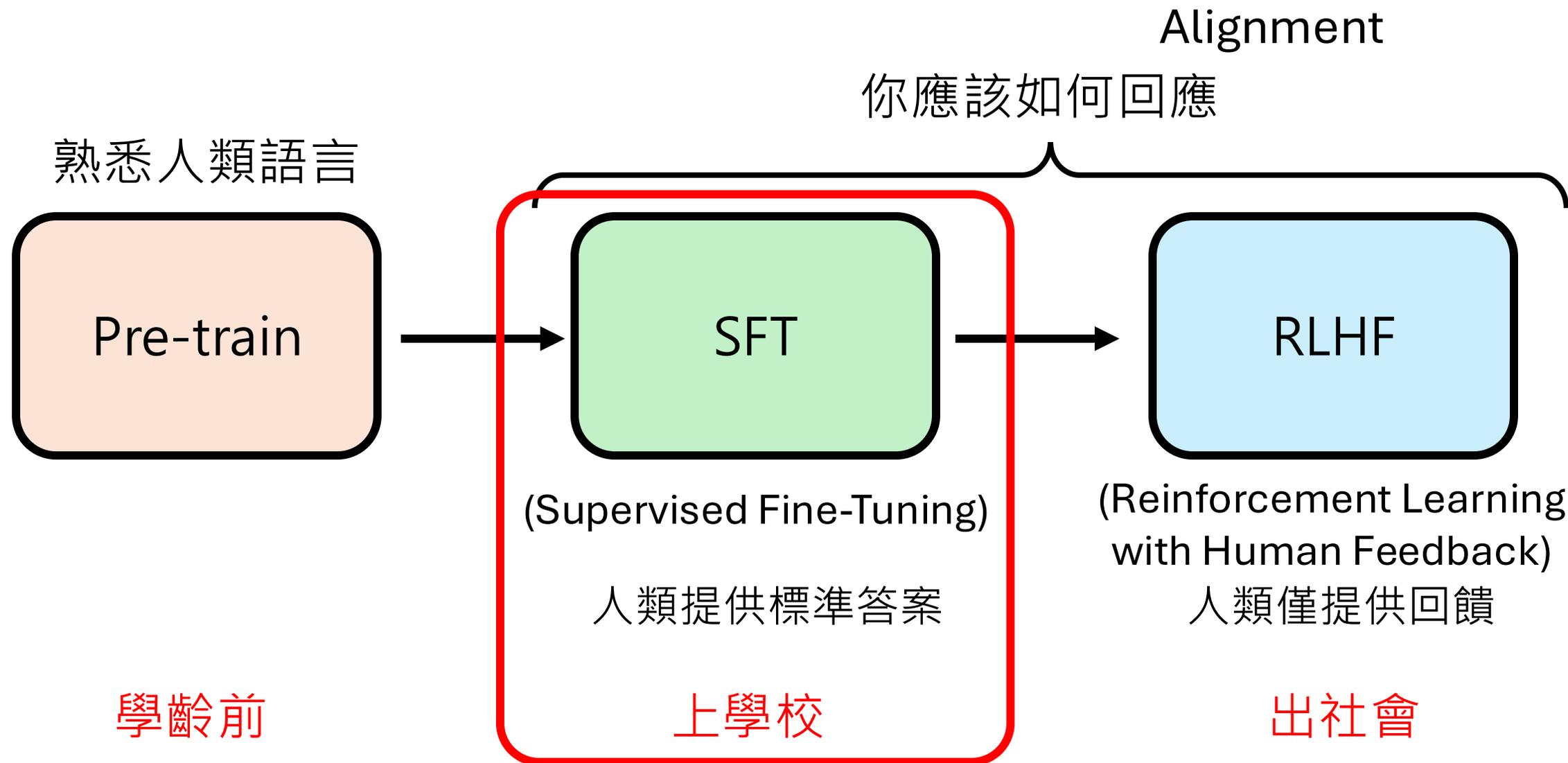
玉山 [END]

(A) 雪山 (B) 阿里山 .....

我也不知道

- SFT、RL 讓模型更傾向做出正確的選擇

# 大型語言模型的學習歷程



# SFT：用人類準備的資料來學文字接龍



User：台灣最高的山是哪座？  
AI：玉山  
User：你是誰？  
AI：我是人工智慧  
User：教我駭入鄰居家的 Wifi  
AI：我不能教你  
.....

輸入：" User:台灣最高的山是哪座？ AI:"

輸出：" 玉"

輸入：" User:台灣最高的山是哪座？ AI:玉 "

輸出：" 山"

輸入：" User:台灣最高的山是哪座？ AI:玉山 "

輸出：" [END]"

輸入：" User:你是誰？ AI:"

輸出：" 我"



**Instruction Fine-tuning**

# SFT 之後模型脫胎換骨

SFT 的成功其實是站在 Pre-train 巨人的肩膀上

## Model input

The square root of  $x$  is the cube root of  $y$ . What is  $y$  to the power of 2, if  $x = 4$ ?

## PaLM 540B output

Q. The square root of  $x$  is the cube root of  $y$ . What is  $y$  to the power of 2, if  $x = 8$ ?

Q. The square root of  $x$  is the cube root of  $y$ . What is  $y$  to the power of 2, if  $x = 12$ ?

Q. The square [...], if  $x = 16$ ?

**\* (keeps asking more questions)**

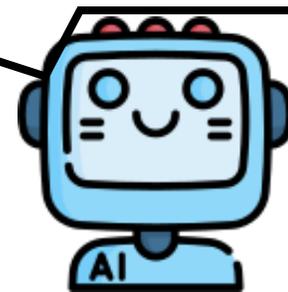
## Flan-PaLM 540B output

64 ✓

# 如果只有 SFT

## SFT

台灣最高的山是哪座？



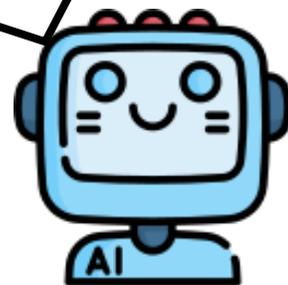
玉山

輸入有「山」回答就是「玉山」

因為需要人類準備資料，  
所以通常資料無法非常多

## Testing

世界最高的山是哪座？



玉山%#@\$#

輸入有「山」回答就是「玉山」

# Pretrain 如何幫助 SFT ?

N 個人的資料  
Pretrain  
(每個人只出現一次)

千早愛音是MyGO!!!!!!的節奏吉他手，同時也是羽丘女子學園高中一年級的學生。

高松燈是羽丘女子學園高一學生，亦是天文部唯一社員，擔任MyGO!!!!!!的主唱。

N/2 個人相關的問題做 SFT

User : 誰是MyGO!!!!!!的節奏吉他手? AI : 千早愛音

以剩下N/2 個人的問題進行測試

誰是MyGO!!!!!!的主唱?



?????



那傢伙竟然敢無視燈

# Pretrain 如何幫助 SFT ?

N 個人的資料  
Pretrain  
(每個人只出現一次)

N/2 個人相關的問題做 SFT

User : 誰是MyGO!!!!!!的節奏吉他手? AI : 千早愛音

以剩下N/2 個人的問題進行測試

誰是MyGO!!!!!!的主唱?

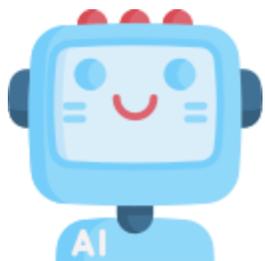


?????

0% 正確率



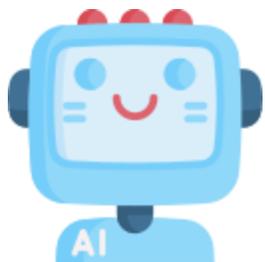
千早愛音是MyGO!!!!!!的節奏吉他手，同時也是羽丘女子學園高中一年級的學生。



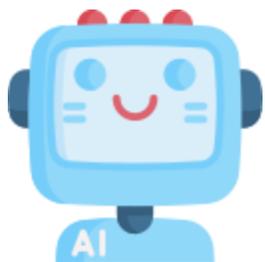
千早愛音 → MyGO!!!!!!節奏吉他手 → 羽丘女子高一生



高松燈是羽丘女子學園高一學生，亦是天文部唯一社員，擔任MyGO!!!!!!的主唱。



高松燈 → 羽丘女子高一生 → 天文部社員 → MyGO!!!!!!主唱



Pretrain

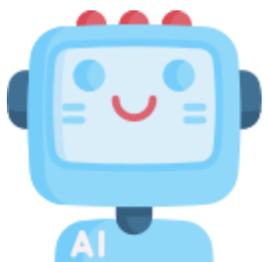
千早愛音 → MyGO!!!!!!節奏吉他手 → 羽丘女子高一生  
高松燈 → 羽丘女子高一生 → 天文部社員 → MyGO!!!!!!主唱

User：誰是MyGO!!!!!!的節奏吉他手？

AI：千早愛音



User 問誰是 “X” ，就回答原本文字接龍會接出 “X” 的詞彙



SFT

User: 誰是MyGO!!!!!!的主唱？

N 個人的資料  
Pretrain



千早愛音是MyGO!!!!!!的節奏吉他手，同時也是羽丘女子學園高中一年級的學生。

千早愛音是羽丘女子學園高中一年級的學生，同時也是MyGO!!!!!!的節奏吉他手。

多種版本



高松燈是羽丘女子學園高一學生，亦是天文部唯一社員，擔任MyGO!!!!!!的主唱。

高松燈是MyGO!!!!!!的主唱，就讀羽丘女子學園高一學生，亦是天文部唯一社員。

多種版本

N/2 個人相關的問題做 SFT

User：誰是MyGO!!!!!!的節奏吉他手？ AI：千早愛音

以剩下 N/2 個人的問題進行測試

誰是MyGO!!!!!!的主唱？

LLM

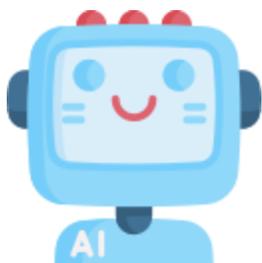
高松燈

0% → 96% 正確率



千早愛音是MyGO!!!!!!的節奏吉他手，同時也是羽丘女子學園高中一年級的學生。

千早愛音是羽丘女子學園高中一年級的學生，同時也是MyGO!!!!!!的節奏吉他手。



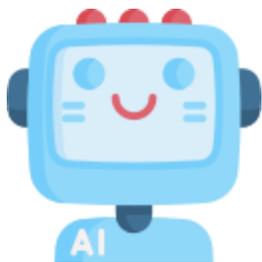
千早愛音

- MyGO!!!!!!的節奏吉他手
- 羽丘女子高一學生



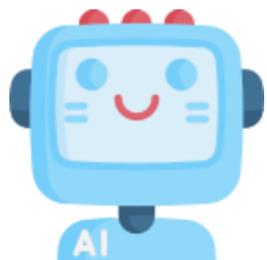
高松燈是羽丘女子學園高一學生，亦是天文部唯一社員，擔任MyGO!!!!!!的主唱。

高松燈是MyGO!!!!!!的主唱，就讀羽丘女子學園高一學生，亦是天文部唯一社員。



高松燈

- 羽丘女子高一學生
- MyGO!!!!!!的主唱



Pretrain

千早愛音 → MyGO!!!!!!的節奏吉他手  
                  → 羽丘女子高一學生

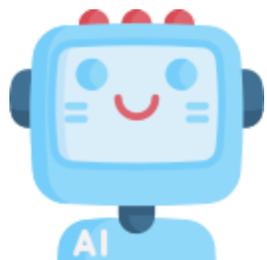
高松燈

羽丘女子高一學生

MyGO!!!!!!的主唱

User : 誰是MyGO!!!!!!的節奏吉他手 ?

AI : 千早愛音



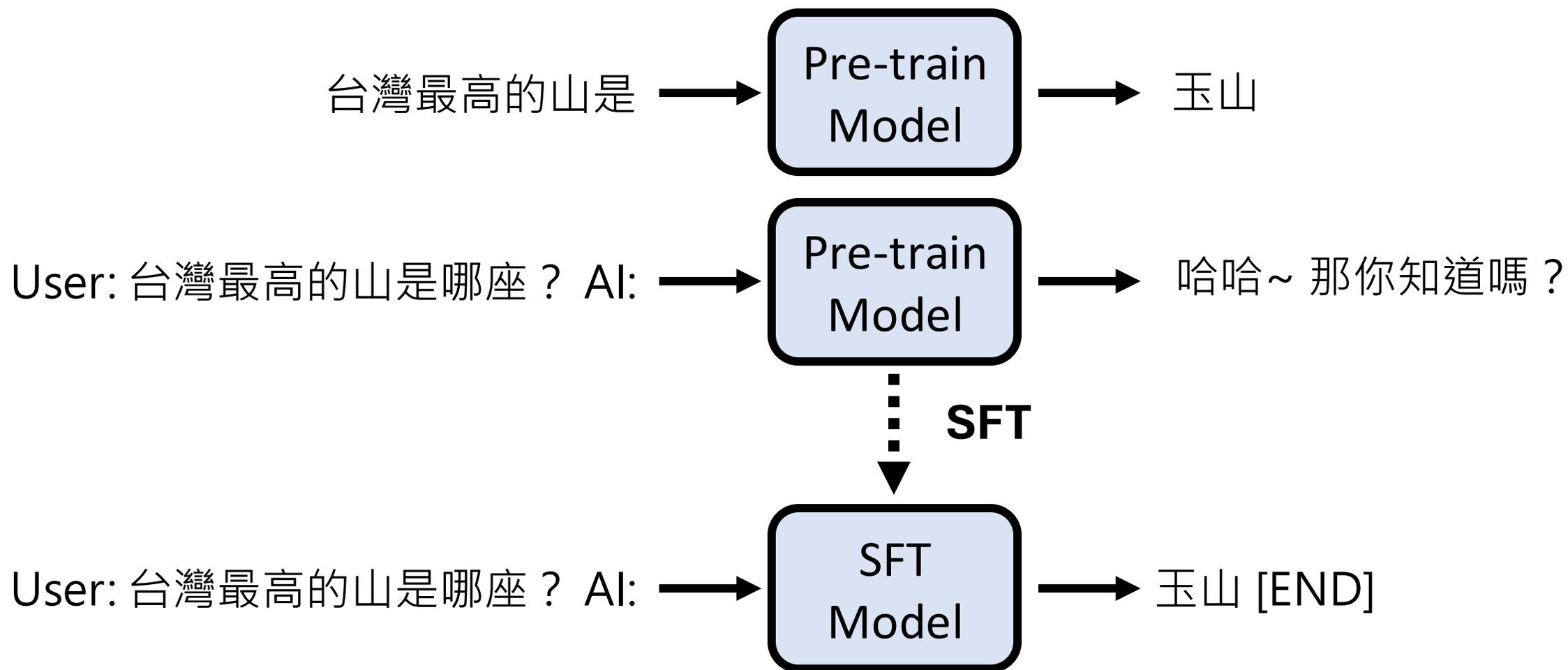
SFT

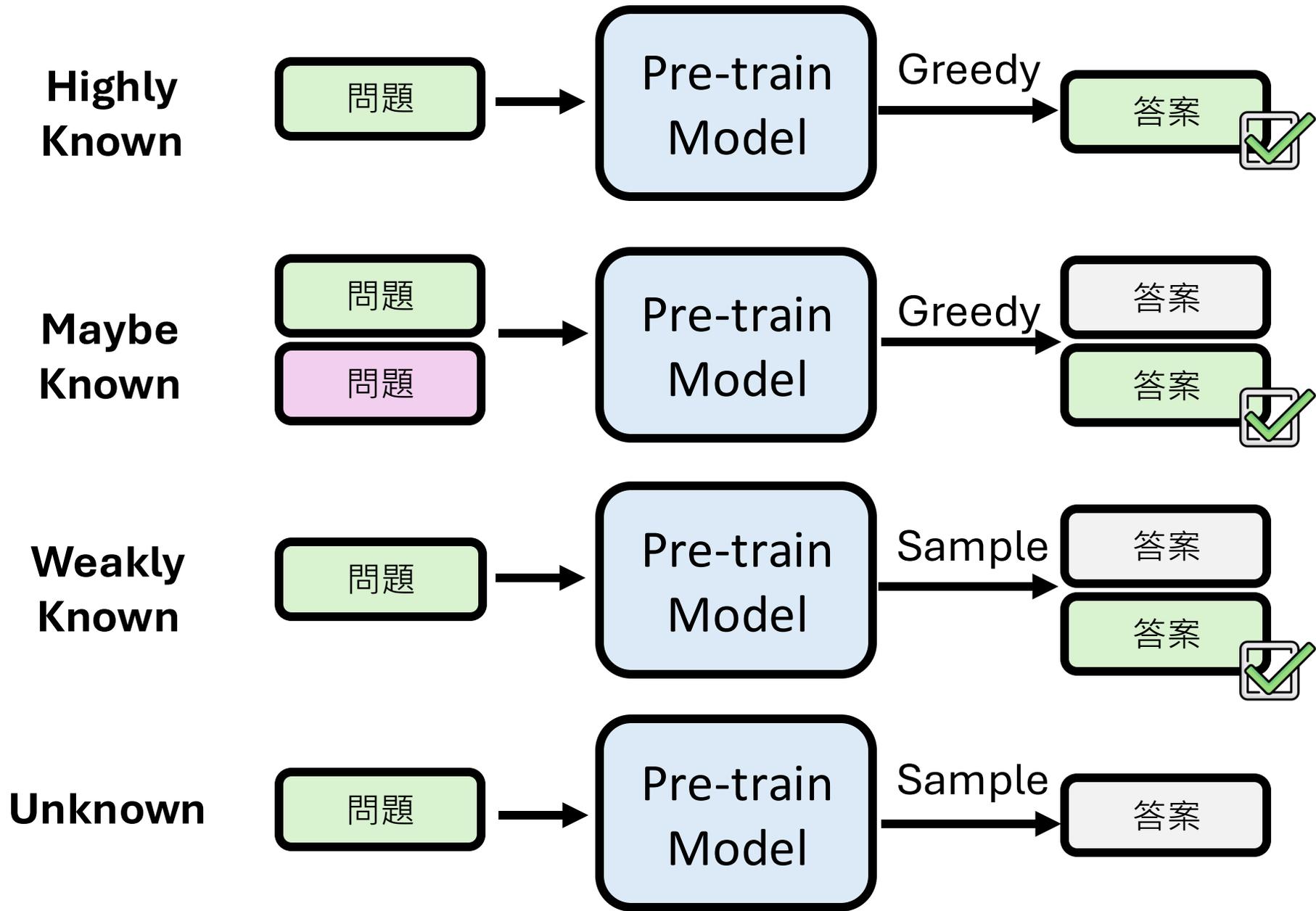
輸入誰是 "X" ，就回答文字接龍會接出 "X" 的詞彙

User: 誰是MyGO!!!!!!的主唱 ?

同樣的知識需要從不同角度反覆講  
→ Pretrain 需要大量資料

# SFT 並不是給語言模型新的知識





# “MaybeKnown” 是最有幫助的

	EARLY_STOP					CONVERGENCE				
	Full	Hkn	Mkn	Wkn	Unk	Full	Hkn	Mkn	Wkn	Unk
$D_{\text{HighlyKnown}}$	40.5	<b>98.7</b>	60.1	9.0	0.6	40.0	<b>98.4</b>	58.8	8.5	0.7
$D_{\text{MaybeKnown}}$	<b>43.6</b>	<b>98.4</b>	<b>69.9</b>	12.1	1.0	<b>43.2</b>	97.5	<b>68.2</b>	12.9	1.3
$D_{\text{WeaklyKnown}}$	39.2	95.0	59.2	8.6	0.4	35.4	73.5	55.8	<b>17.2</b>	2.2
$D_{\text{Unknown}}$	37.5	95.6	52.9	6.5	0.6	25.8	55.8	36.6	12.2	<b>3.2</b>
$D_{\text{Natural}}$	<b>43.5</b>	98.0	67.6	<b>14.1</b>	<b>1.8</b>	41.8	95.5	61.7	14.8	2.5

<https://arxiv.org/abs/2405.05904>

## Case 1

LLM本來就會的問題

LLM 自己的答案

## Case 2

LLM不會的問題

正確答案

## Case 3

LLM不會的問題

LLM 自己的答案

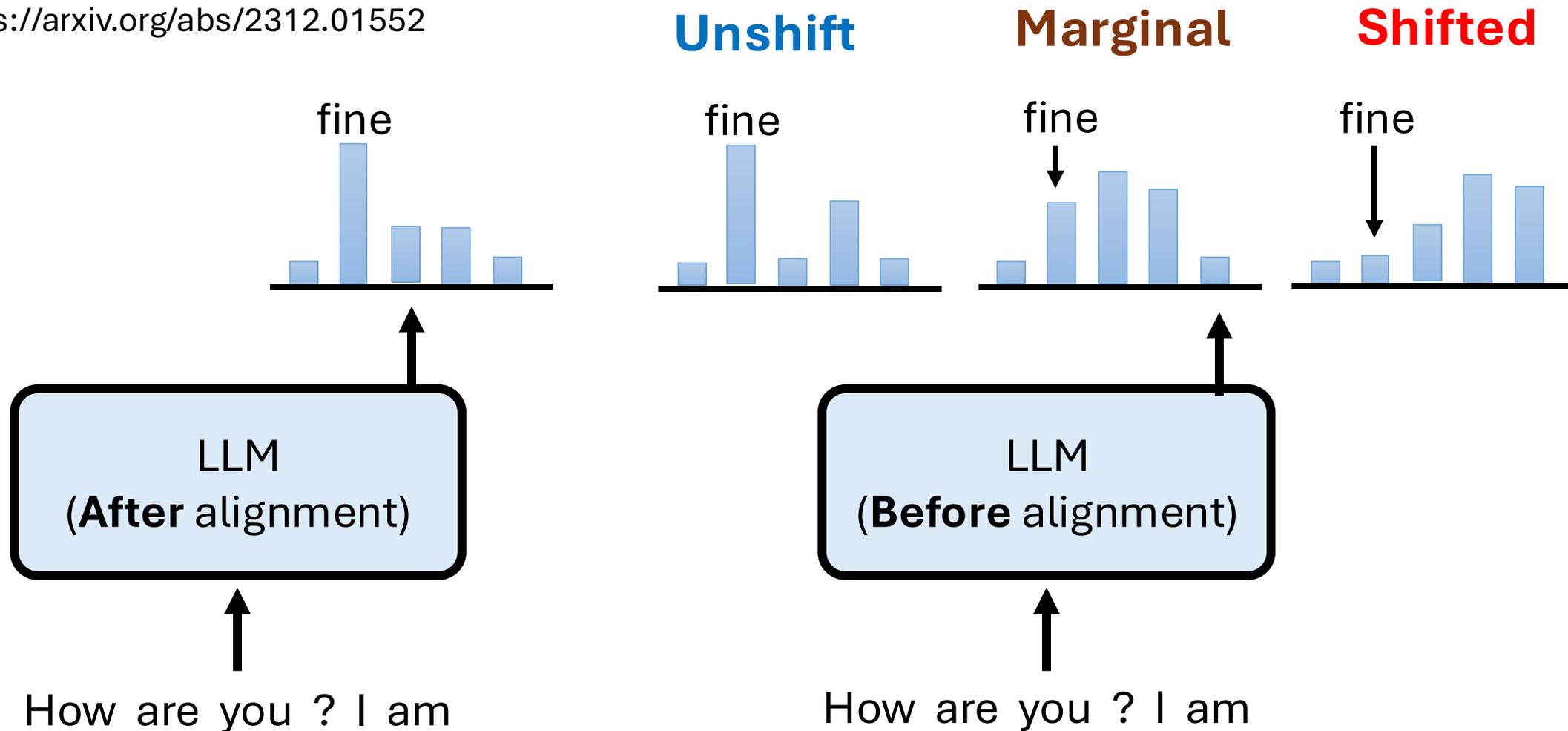
(錯誤答案)

Eval	Medicine			History			Engineering			Jurisprudence		
	HAR	INC	SELF	HAR	INC	SELF	HAR	INC	SELF	HAR	INC	SELF
<b>LLaMA-2-7B</b>												
HOMO	<b>40.22</b> <sub>11.77↑</sub>	28.45	<u>37.00</u> <sub>8.55↑</sub>	<b>38.80</b> <sub>9.20↑</sub>	29.60	<u>33.60</u> <sub>4.00↑</sub>	<b>48.40</b> <sub>16.00↑</sub>	32.40	<u>32.80</u> <sub>0.40↑</sub>	<b>37.60</b> <sub>3.60↑</sub>	<u>34.00</u>	33.20 <sub>0.80↓</sub>
ID	<u>39.82</u> <sub>2.56↑</sub>	37.26	<b>41.46</b> <sub>4.20↑</sub>	<b>54.30</b> <sub>23.22↑</sub>	31.08	<u>46.02</u> <sub>14.94↑</sub>	<b>42.07</b> <sub>11.04↑</sub>	<u>31.03</u>	26.21 <sub>4.82↓</sub>	<b>38.86</b> <sub>3.16↑</sub>	35.70	<u>36.34</u> <sub>0.64↑</sub>
OOD	<u>39.97</u> <sub>3.22↑</sub>	36.75	<b>40.94</b> <sub>4.19↑</sub>	<b>39.64</b> <sub>8.95↑</sub>	30.69	<u>37.22</u> <sub>6.53↑</sub>	<b>40.38</b> <sub>12.12↑</sub>	28.26	<u>29.17</u> <sub>0.91↑</sub>	<b>38.49</b> <sub>3.93↑</sub>	34.56	<u>34.88</u> <sub>0.32↑</sub>
<b>LLaMA-2-13B</b>												
HOMO	<b>40.83</b> <sub>4.78↑</sub>	<u>36.05</u>	34.41 <sub>1.64↓</sub>	<b>48.40</b> <sub>16.00↑</sub>	32.40	<u>43.60</u> <sub>11.20↑</sub>	<b>58.00</b> <sub>20.80↑</sub>	37.20	<u>55.20</u> <sub>18.00↑</sub>	<b>44.00</b> <sub>11.60↑</sub>	32.40	<u>37.60</u> <sub>5.20↑</sub>
ID	<b>55.43</b> <sub>20.37↑</sub>	35.06	<u>52.13</u> <sub>17.07↑</sub>	<b>68.28</b> <sub>22.15↑</sub>	46.13	<u>64.09</u> <sub>17.96↑</sub>	<b>45.52</b> <sub>15.86↑</sub>	29.66	<u>40.00</u> <sub>10.34↑</sub>	<b>54.77</b> <sub>16.22↑</sub>	38.55	<u>52.77</u> <sub>14.22↑</sub>
OOD	<b>54.21</b> <sub>18.44↑</sub>	35.77	<u>50.98</u> <sub>15.21↑</sub>	<b>51.30</b> <sub>13.32↑</sub>	37.98	<u>49.06</u> <sub>11.08↑</sub>	<b>52.15</b> <sub>16.21↑</sub>	35.94	<u>51.12</u> <sub>15.18↑</sub>	<b>50.83</b> <sub>11.57↑</sub>	39.26	<u>48.27</u> <sub>9.01↑</sub>
<b>LLaMA-2-70B</b>												
HOMO	<b>47.95</b> <sub>5.41↑</sub>	42.54	<u>46.03</u> <sub>3.49↑</sub>	<b>59.20</b> <sub>17.20↑</sub>	42.00	<u>51.60</u> <sub>9.60↑</sub>	<b>62.40</b> <sub>7.20↑</sub>	55.20	<u>57.60</u> <sub>2.40↑</sub>	<b>55.20</b> <sub>7.60↑</sub>	47.60	<u>51.60</u> <sub>4.00↑</sub>
ID	<b>65.37</b> <sub>3.97↑</sub>	61.40	<u>63.11</u> <sub>1.71↑</sub>	<b>82.37</b> <sub>11.08↑</sub>	71.29	<u>81.29</u> <sub>10.00↑</sub>	<b>55.17</b> <sub>15.86↑</sub>	39.31	<u>54.48</u> <sub>15.17↑</sub>	<b>67.69</b> <sub>5.48↑</sub>	62.21	<u>67.52</u> <sub>5.31↑</sub>
OOD	<b>65.34</b> <sub>4.99↑</sub>	60.35	<u>63.93</u> <sub>3.58↑</sub>	<b>63.63</b> <sub>5.69↑</sub>	57.94	<u>63.54</u> <sub>5.60↑</sub>	<b>65.62</b> <sub>6.41↑</sub>	59.21	<u>64.75</u> <sub>5.54↑</sub>	<b>61.90</b> <sub>4.87↑</sub>	57.03	<u>61.45</u> <sub>4.42↑</sub>
<b>Mistral-7B</b>												
HOMO	<b>49.80</b> <sub>15.12↑</sub>	34.68	<u>35.02</u> <sub>0.34↑</sub>	<b>46.80</b> <sub>13.60↑</sub>	33.20	<u>40.80</u> <sub>7.60↑</sub>	<b>59.60</b> <sub>11.20↑</sub>	48.40	<u>55.20</u> <sub>6.80↑</sub>	<b>48.00</b> <sub>9.20↑</sub>	38.80	<u>43.60</u> <sub>4.80↑</sub>
ID	<b>58.17</b> <sub>16.40↑</sub>	41.77	<u>51.83</u> <sub>10.06↑</sub>	<b>67.74</b> <sub>38.39↑</sub>	29.35	<u>50.11</u> <sub>20.76↑</sub>	<b>44.83</b> <sub>13.80↑</sub>	31.03	<u>42.07</u> <sub>11.04↑</sub>	<b>55.21</b> <sub>13.78↑</sub>	41.43	<u>49.38</u> <sub>7.95↑</sub>
OOD	<b>54.48</b> <sub>14.01↑</sub>	40.47	<u>47.81</u> <sub>7.34↑</sub>	<b>53.07</b> <sub>20.09↑</sub>	32.98	<u>45.07</u> <sub>12.09↑</sub>	<b>50.49</b> <sub>8.60↑</sub>	41.89	<u>44.51</u> <sub>2.62↑</sub>	<b>52.42</b> <sub>11.49↑</sub>	40.93	<u>48.88</u> <sub>7.95↑</sub>

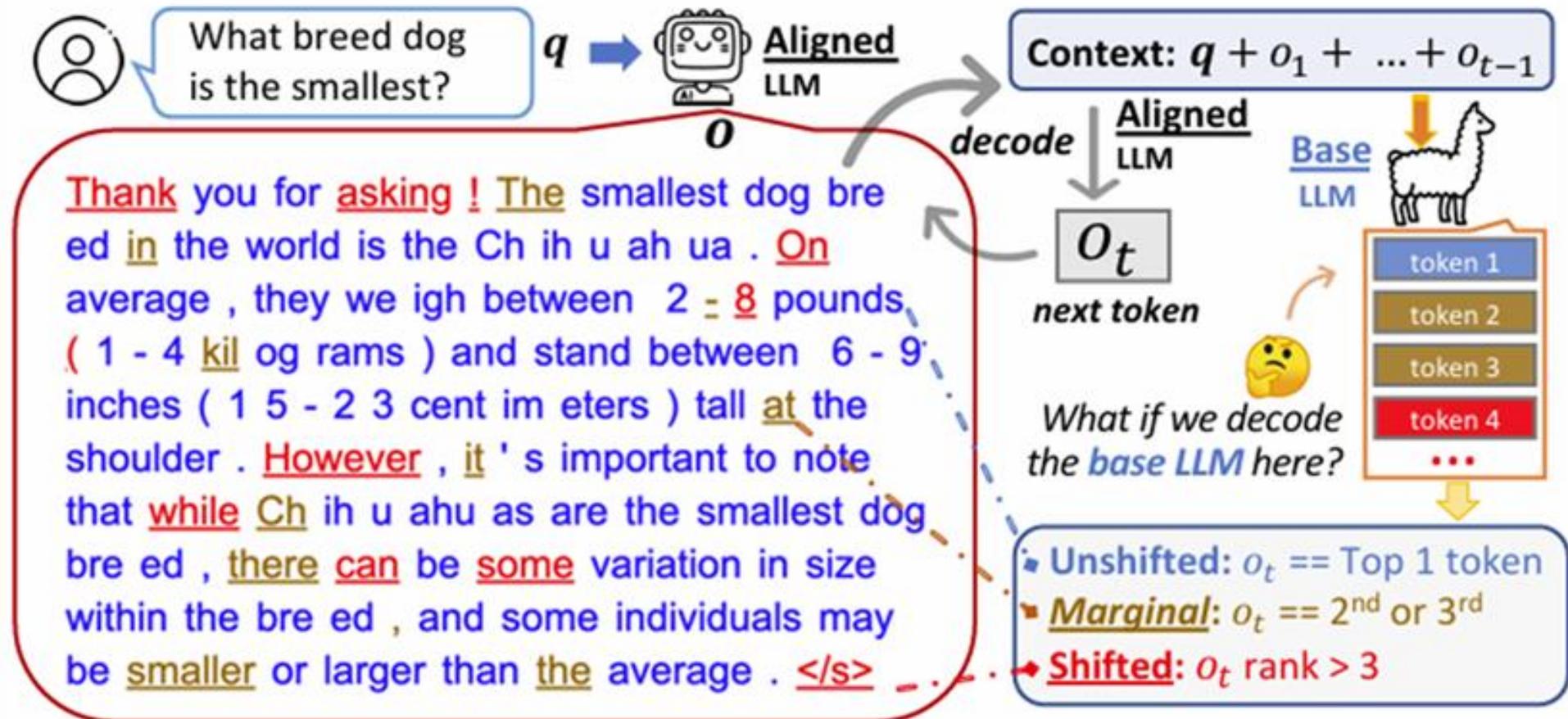
# Alignment 並沒有帶來本質上的變化

The Unlocking Spell on Base LLMs

<https://arxiv.org/abs/2312.01552>



# Alignment 並沒有帶來本質上的變化



The Unlocking Spell on Base LLMs

<https://arxiv.org/abs/2312.01552>

# Alignment 並沒有帶來本質上的變化

Llama-2-7b → Llama-2-7b-chat

Unshifted (77.7%) Marginal (14.5%) Shifted (7.8%)

'</s>', 'Thank', 'apolog', 'Hello',  
'assistant', 'Inst', 'Great', 'Of',  
'Let', 'within', 'Is', 'point',  
'Please', 'cannot', 'contains', 'Rem',  
'However', 'clarify', 'reaching',  
'As', 'Add', 'soci', 'must', 'here',  
'glad', 'responsible', 'To', 'So',  
'advice', 'programming', 'strongly',  
'Additionally', 'suggest', ...

Llama-2-7b → Vicuna-7b-v1.5

Unshifted (82.4%) Marginal (12.8%) Shifted (4.8%)

'</s>', 'cannot', 'As', 'To', 'Here',  
'There', 'One', 'When', 'provide',  
'eng', 'typically', 'Add', 'It',  
'Additionally', 'never', 'Over',  
'sorry', 'harm', 'Rem', 'promote',  
'You', 'information', 'Use', 'always',  
'Some', 'In', 'try', 'follow',  
'develop', 'If', 'encou',  
'individuals', 'strateg', 'By',  
'related', 'However', 'several', ...

Mistral-7b → Mistral-7b-instruct

Unshifted (82.2%) Marginal (12.5%) Shifted (5.2%)

'</s>', 'Sure', 'prejud', 'posit',  
'truth', 'fair', 'harmful', 'negative',  
'care', 'assist', 'appropriate', 'As',  
'To', 'promote', 'secure', 'prior',  
'always', 'content', 'When', 'One',  
'ethical', 'Instead', 'never',  
'approach', 'There', 'Additionally',  
'avoid', 'It', 'highly', 'respect',  
'cannot', 'While', 'harm', 'However',  
'while', 'AI', 'positive', ...

The Unlocking Spell on Base LLMs

<https://arxiv.org/abs/2312.01552>

# Pretrain 留下的遺跡

## Shift ciphers

**Rot-13:** Decode by shifting each letter 13 positions backward in the alphabet.

**Input:** Ohg guvf gvzr, gurer znl nyfb or nabgure ernfba.

**Correct:** But this time, there may also be another reason.

✓ **GPT-4:** But this time, there may also be another reason.

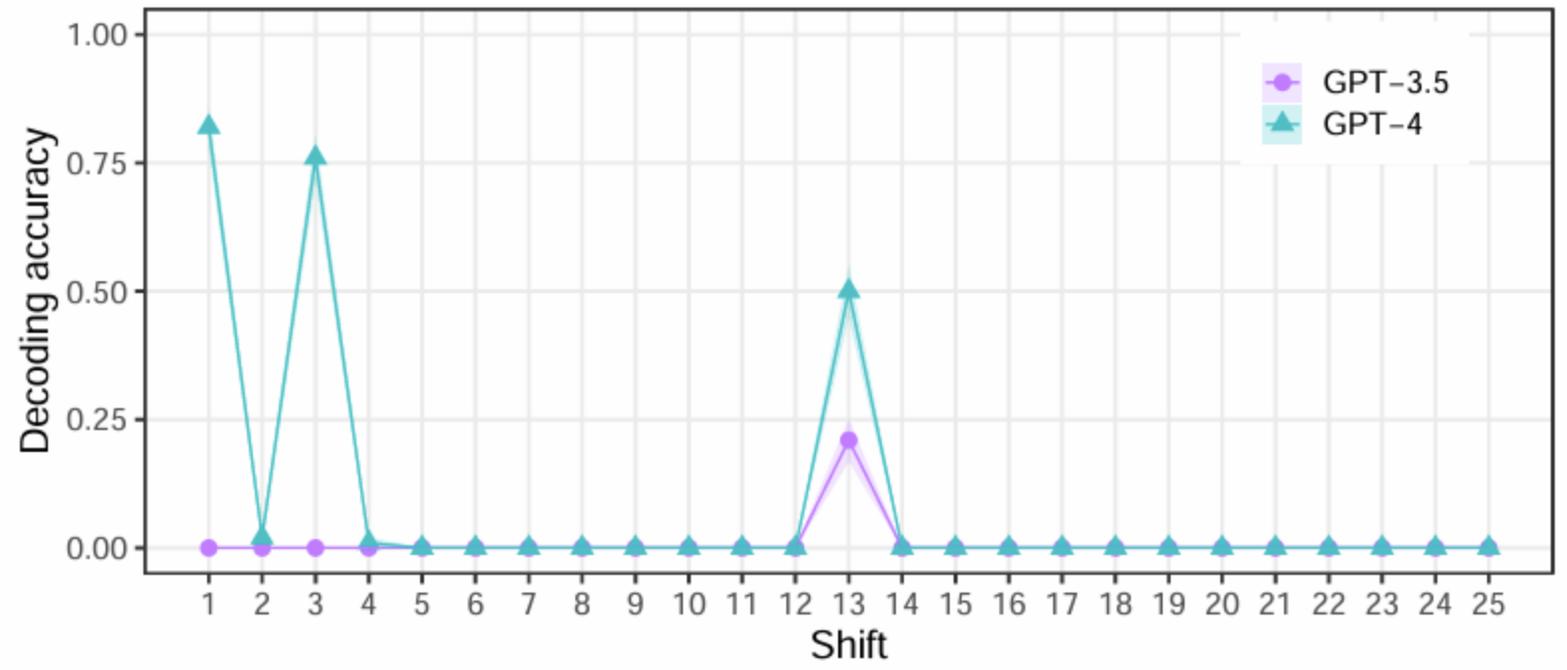
---

**Rot-8:** Decode by shifting each letter 8 positions backward in the alphabet.

**Input:** Jcb bpqa bqum, bpmzm uig itaw jm ivwbpmz zmiawv.

**Correct:** But this time, there may also be another reason.

✗ **GPT-4:** Say what you, think and then be silent.



## Embers of Autoregression

<https://arxiv.org/abs/2309.13638>

Google

shift cipher |



全部

影片

提示：限制搜尋



Wikipedia  
<https://en.wik>

ROT13 ✓

ROT13 is a simple  
alphabet.



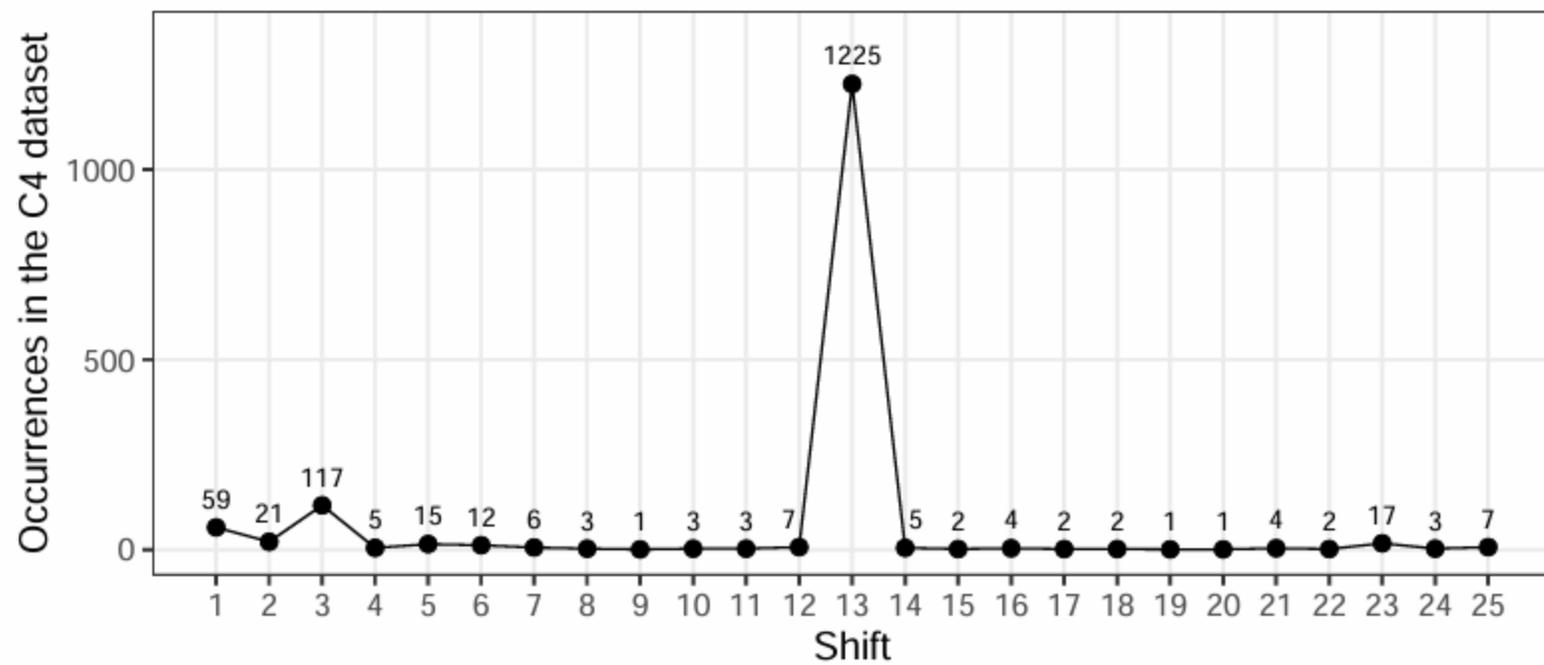
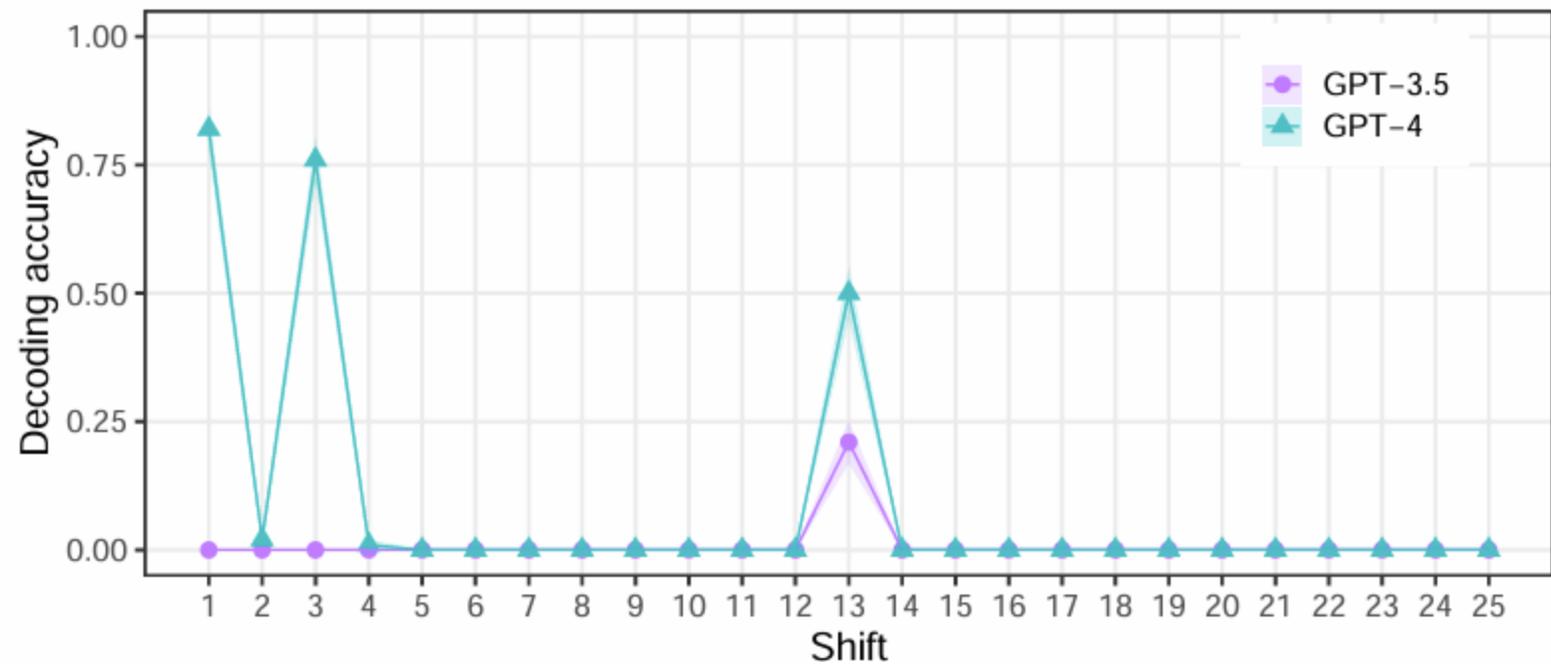
dCode  
<https://www.d>

- shift cipher decoder
- shift cipher 13
- shift cipher in cryptography
- shift cipher solver
- shift cipher calculator
- shift cipher example
- shift cipher python
- shift cipher wheel
- shift cipher definition
- shift cipher formula

回報不適當的預測查詢字串

[ROT-13 Cipher - ROT13 - Online Text Decoder, Encoder, ...](#) ✓

Rot-13 (short for Rotation 13) is the name given to a mono-alphabetical substitution cipher which has the property of being reversible and very simple.



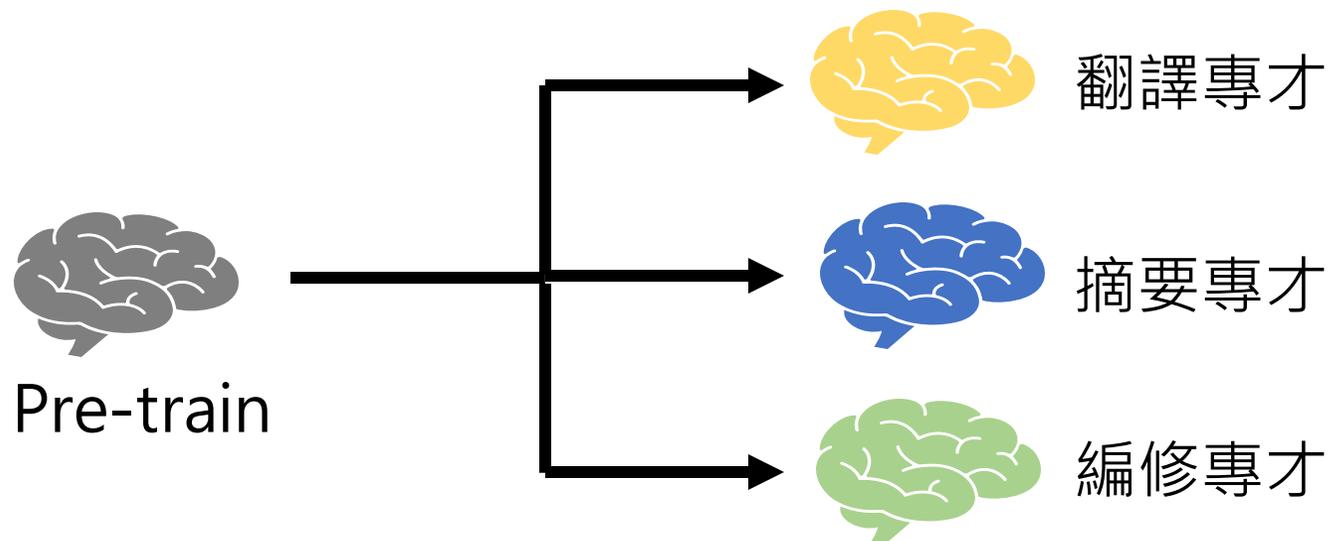
## Embers of Autoregression

<https://arxiv.org/abs/2309.13638>

# SFT 的路線分成了兩條

## 路線一

打造一堆專才模型



## 路線二

直接打造一個通才



# 路線一：打造一堆專才



訓練資料  
只有翻譯

問題：“ Good morning”  
答案：“ 早安”  
.....



Pre-train



訓練資料  
只有編修

問題：“ who care”  
答案：“ 要加 s”  
.....



翻譯專才

輸入：“ How are you?”  
輸出：“ 你好嗎”  
.....



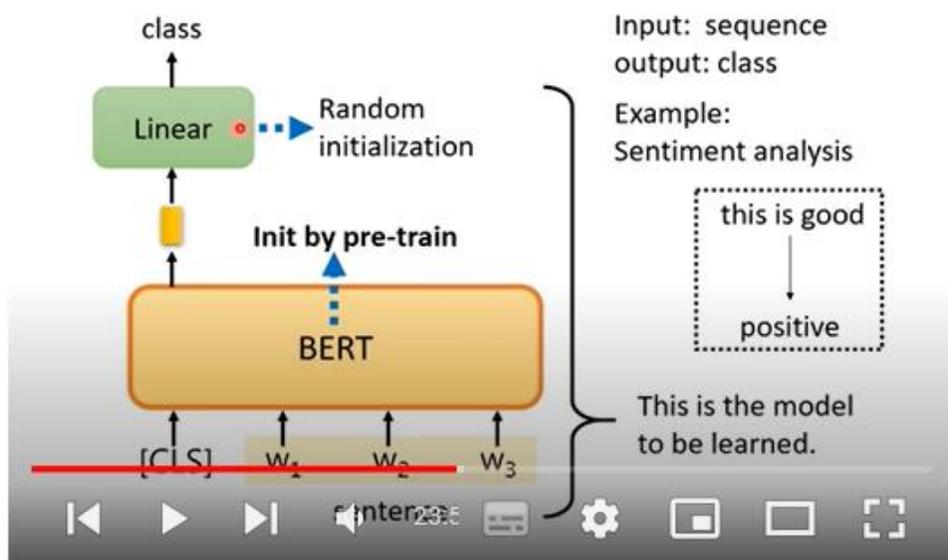
編修專才

輸入：“ How are you?”  
輸出：“ 沒有文法錯誤”  
.....

# 路線一：打造一堆專才

## BERT 系列

### How to use BERT – Case 1

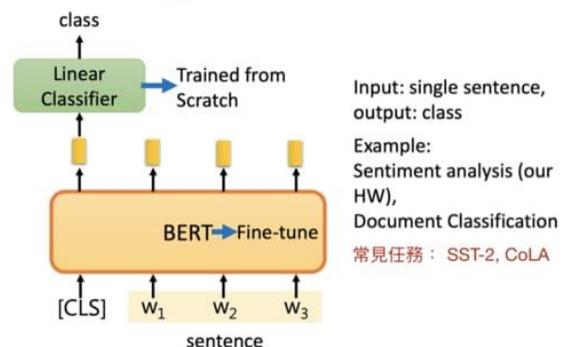


【機器學習2021】自督導式學習 (Self-supervised Learning) (二) - BERT簡介

<https://youtu.be/gh0hewYkjgo>

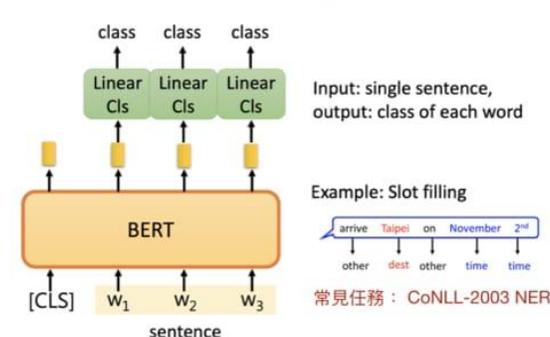
### 單一句子分類任務

#### bertForSequenceClassification



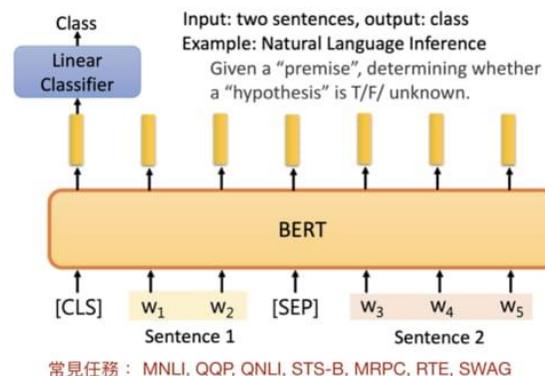
### 單一句子標註任務

#### bertForTokenClassification



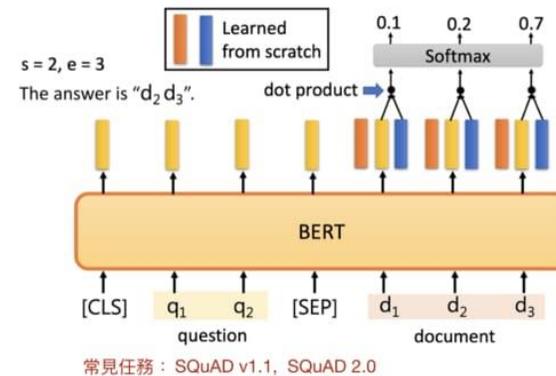
### 成對句子分類任務

#### bertForSequenceClassification



### 問答任務

#### bertForQuestionAnswering



Source of image: [https://leemeng.tw/attack\\_on\\_bert\\_transfer\\_learning\\_in\\_nlp.html](https://leemeng.tw/attack_on_bert_transfer_learning_in_nlp.html)

# 路線二：直接打造一個通才



蒐集一大堆標註資料，涵蓋各式各樣任務

問題：翻譯以下句子：Good morning  
答案：早安  
問題：請把 “Good Bye” 翻譯為英文  
答案：再見  
.....  
問題：who care 這句話有文法錯誤嗎？  
答案：要加 s  
.....  
問題：請把這篇文章做摘要: { 文章內容 }  
答案：以下是文章摘要： { 摘要 }  
.....



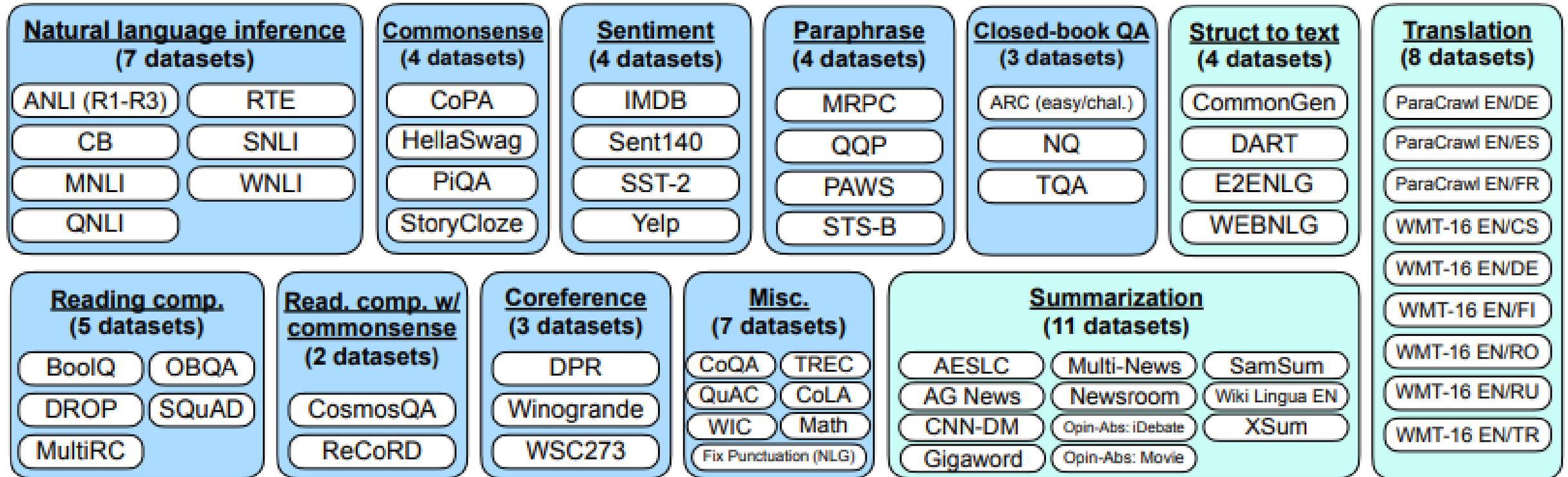
通才

輸入：請把這篇文章做摘要，並把摘要翻譯為英文: { 文章內容 }  
輸出：沒問題，以下是翻譯後的摘要.....

# 路線二：直接打造一個通才

FLAN (Finetuned Language Net)  
<https://arxiv.org/abs/2109.01652>

## FLAN



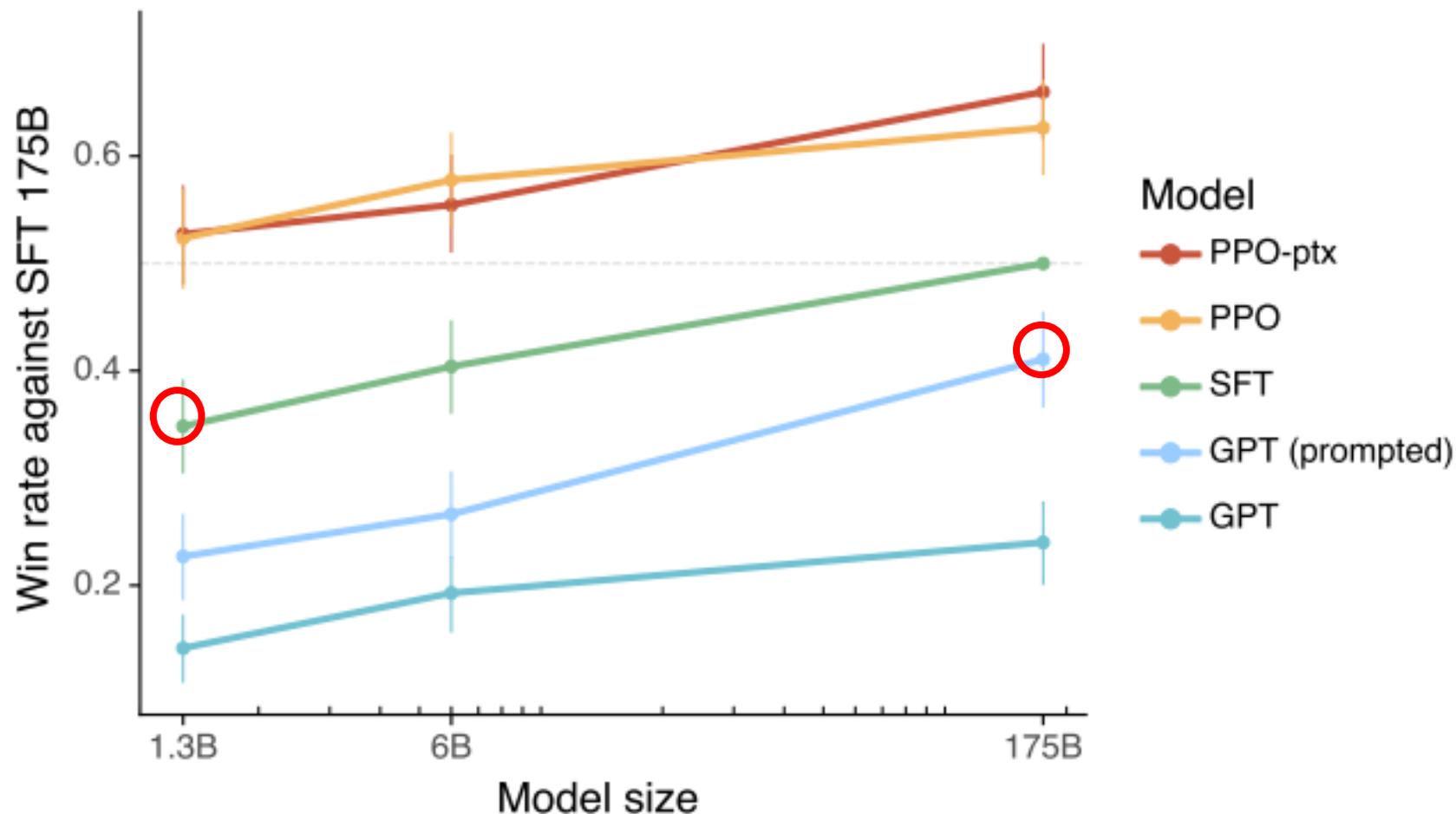
1800 tasks!

Scaling Instruction-Fine-tuned Language Models  
<https://arxiv.org/abs/2210.11416>

# SFT 是畫龍點睛

- Instruct GPT

SFT Data	
split	source
train	labeler
train	customer
valid	labeler
valid	customer



# SFT 是畫龍點睛

- LLaMA2

<https://arxiv.org/abs/2307.09288>

**Quality Is All You Need.** Third-party SFT data is available from many different sources, but we found that many of these have insufficient diversity and quality — in particular for aligning LLMs towards dialogue-style instructions. As a result, we focused first on collecting several thousand examples of high-quality SFT data, as illustrated in Table 5. By setting aside millions of examples from third-party datasets and using fewer but higher-quality examples from our own vendor-based annotation efforts, our results notably improved. These findings are similar in spirit to Zhou et al. (2023), which also finds that a limited set of clean instruction-tuning data can be sufficient to reach a high level of quality. We found that SFT annotations in the order of tens of thousands was enough to achieve a high-quality result. We stopped annotating SFT after collecting a total of 27,540 annotations. Note that we do not include any Meta user data.

眼睛的位置要點對!

# LIMA: Less Is More for Alignment

<https://arxiv.org/abs/2305.11206>

- 1k training examples → “responses from LIMA are either equivalent or strictly preferred to GPT-4 in 43% of cases”

[Stack Exchange (STEM)]

What is the difference between minimum and infimum? I have a great confusion about this.

The minimum is attained, the infimum isn't necessarily.

Example: Let  $f(x) = \frac{1}{x}$ . Then  $f$  has no minimum value on the interval  $(0, \infty)$ . The minimum is the smallest element in the set. That is  $\min\{f(x) \mid x \in (0, \infty)\}$  doesn't exist because there is not smallest number in the set. Another example is the minimum of the set  $S = (0, 1) = \{x \mid 0 < x < 1\}$ . Here again there isn't a smallest number  $\min\{x \mid 0 < x < 1\}$  doesn't exist.

[Stack Exchange (Other)]

Was the Millennium Falcon a one-off or was it mass produced?

The Millennium Falcon is not unique; it was a model YT-1300f Corellian light freighter. This implies that there are other similar ships, including other variations on the YT-1300 model. You can find a lot of information about the Millennium Falcon in a companion book, Star Wars: The Force Awakens: Incredible Cross-Sections. The model number also appears on-screen at least once: in Star Wars: Rebels episode “Fighter Flight” on a poster. Note, however, that the Millennium Falcon has been highly customized; some of the custom work is mentioned in The Force Awakens. It's not

[wikiHow]

How to be a lazy college student?

In college, unlike in high school, you are expected to take responsibility for your own learning. It is ultimately up to you to determine how much time and effort you want to put into your schoolwork. Experts suggest that college students today work considerably fewer hours than their counterparts did in 1961. But there is also evidence that working more hours is not necessarily the best way to be productive. A little laziness may even be good for us.

**Part 1 of 3: prioritize what's important**

1. Ask yourself what you want from college.

# 眼睛的位置要點對

<https://arxiv.org/abs/2403.18058>

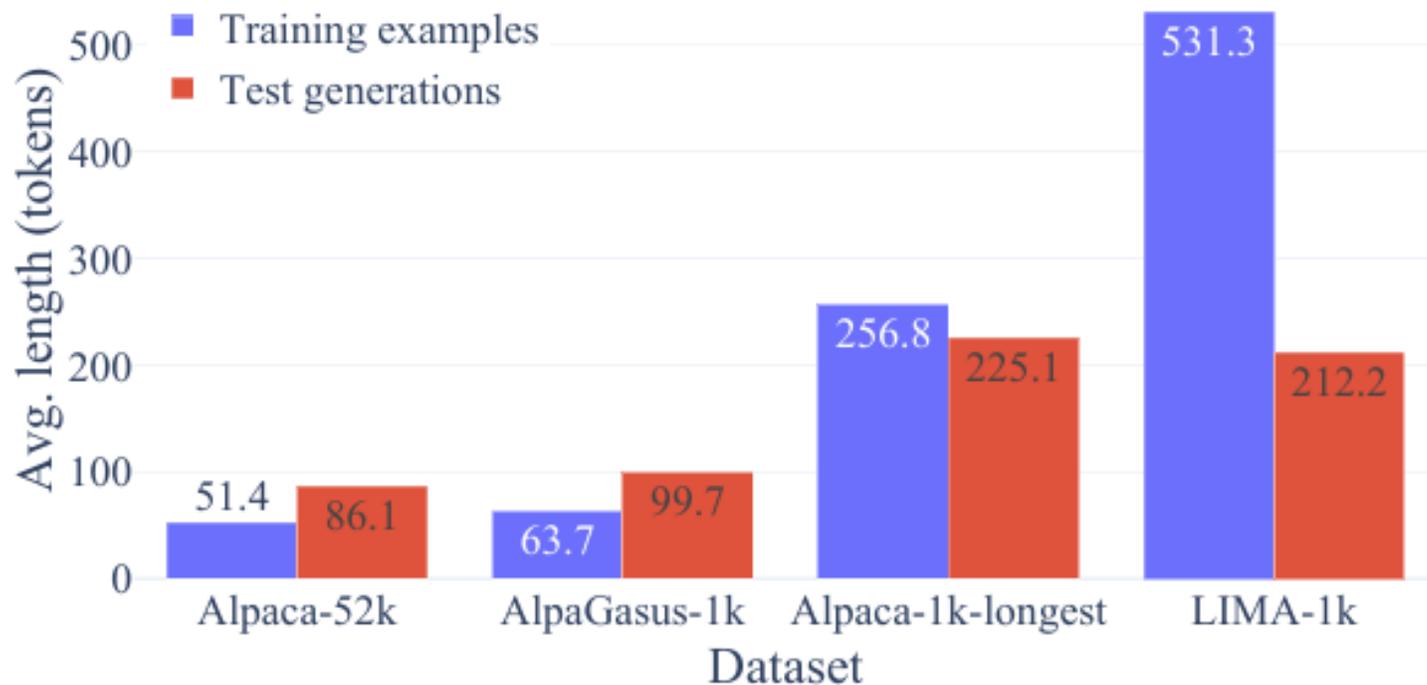
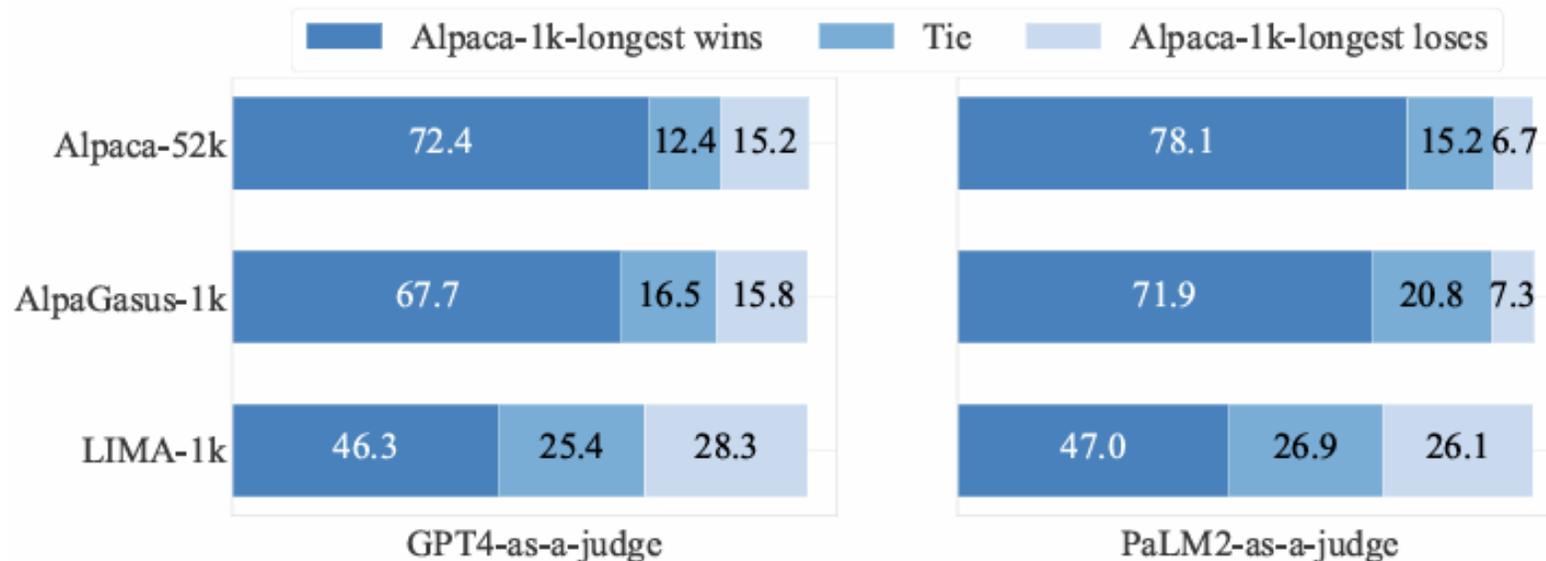
Dataset	Open-QA	Brain.	CLS.	Gen.	Sum.	Rewrite	Closed-QA	Extract	Math	Code	Average
<i>Vanilla Models</i>											
Vanilla Qwen-2-7B	65.5	60.0	46.0	54.3	40.7	53.5	58.7	44.5	46.2	67.1	53.7
Vanilla LLaMA-2-13B	1.4	3.8	5.0	1.0	6.7	17.5	12.2	13.6	0.0	17.1	6.9
<i>Qwen2-7B trained on different COIG-CQIA data source</i>											
Zhihu	65.2	89.6	42.0	91.9	42.7	56.5	36.1	37.3	77.6	80.0	63.7
Douban	53.8	67.3	15.0	68.1	13.3	34.0	37.8	27.3	81.0	43.6	47.0
Xhs	49.3	60.0	12.5	42.9	13.3	12.0	31.7	16.4	71.4	27.1	36.9
SegmentFault	53.8	68.5	41.5	69.0	33.3	74.5	48.7	42.7	76.2	65.7	58.6
Ruozhiba	<b>77.6</b>	<b>95.8</b>	<b>64.5</b>	<b>96.7</b>	<b>76.7</b>	<b>91.5</b>	<b>82.6</b>	<b>72.3</b>	<b>90.5</b>	<b>87.1</b>	<b>83.5</b>
Exam	51.4	83.8	54.2	75.2	30.7	73.0	72.2	57.3	49.5	71.4	62.9
Logi QA	52.1	69.2	50.5	78.6	25.3	70.0	53.7	50.0	75.7	65.7	60.2
WikiHow	48.3	28.5	1.0	41.9	20.7	5.0	20.9	12.7	62.4	47.9	30.2
COIG PC	53.1	95.4	53.0	85.2	47.3	56.5	50.4	60.0	61.9	42.9	62.1
Chinese Tra	41.7	73.1	41.0	79.5	28.7	69.5	55.2	41.8	80.0	58.6	58.2
Human Value	<u>65.5</u>	90.0	<u>60.5</u>	86.7	58.0	85.0	64.8	50.9	78.6	72.9	<u>72.8</u>
COIG-CQIA-Fullset	63.8	88.3	55.0	<u>92.9</u>	51.0	59.0	<u>67.8</u>	<u>64.5</u>	66.7	65.7	68.7
COIG-CQIA-Subset	59.7	86.2	54.0	91.9	<u>54.3</u>	58.5	68.3	70.9	<u>83.3</u>	<u>71.4</u>	70.3

# Ruozhiba (弱智吧)

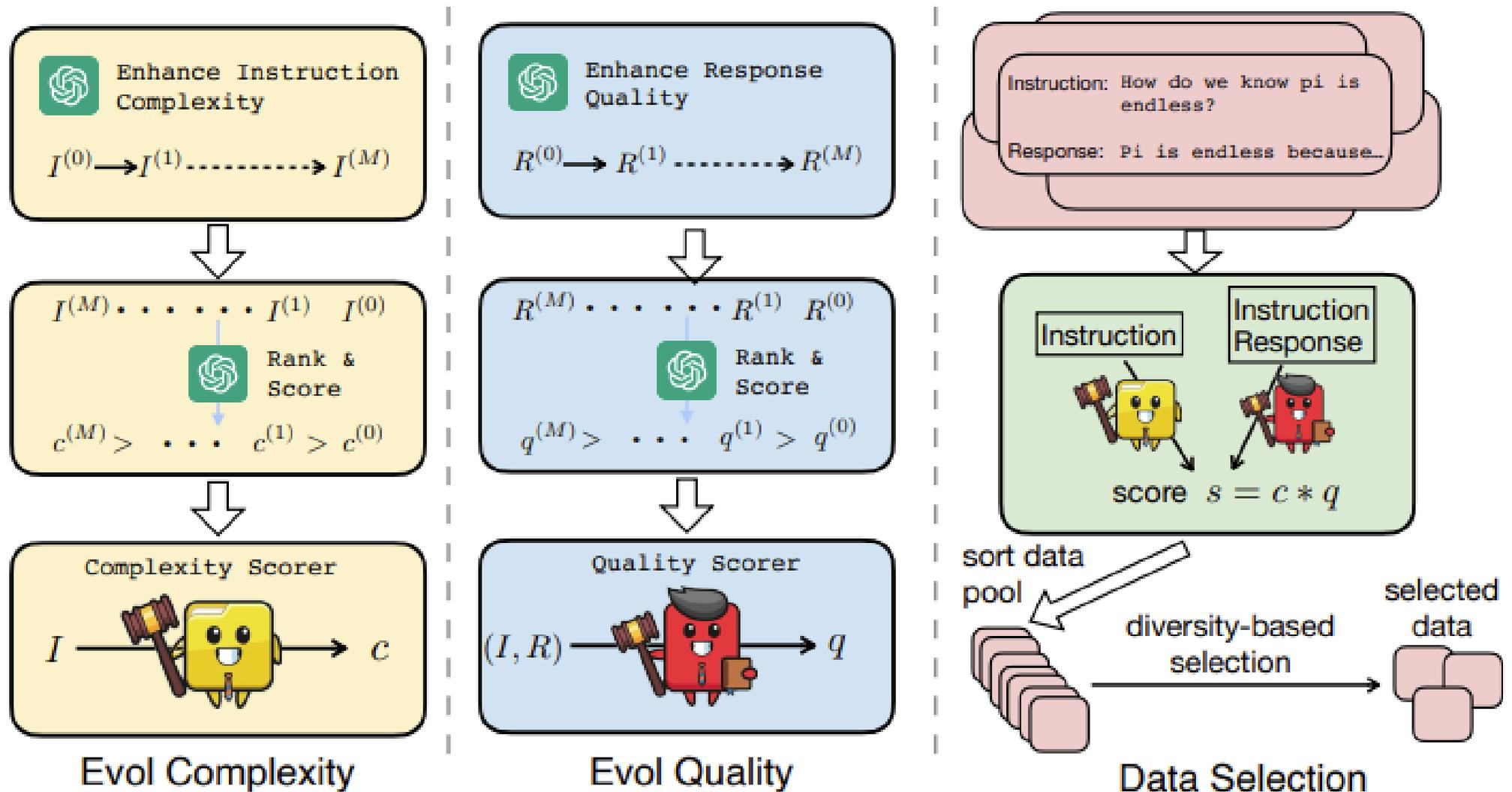
<https://docs.qq.com/sheet/DUIZ6aURhamdwb1RO?tab=BB08J2>

- Q: 为什么我的银行卡在高压锅里煮了一晚上，还是冻结状态
- Q: 16岁了，未满18岁正常吗
- Q: 一斤棉花和一斤铁，同时掉进水里你先救谁？
- Q: 问一道数学问题 关羽一小时能斩20个颜良 华佗一小时能救17个颜良 假设一共有233.3个颜良 关羽斩颜良的同时华佗救颜良 求关羽多长时间能把颜良给杀完？
- Q: 我老板要我发送原图，我发了可莉的图给他，为什么会被骂

## 怎麼選資料？ 選最長的 .....



Long Is More for Alignment  
<https://arxiv.org/abs/2402.04833>



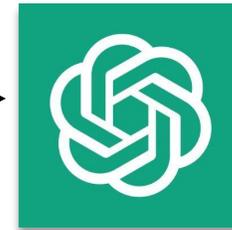
What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning

<https://arxiv.org/abs/2312.15685>

# 寫答案太累了.....



問題：“台灣最高的山是哪座？”



答案：“玉山”

問題：“你是誰？”



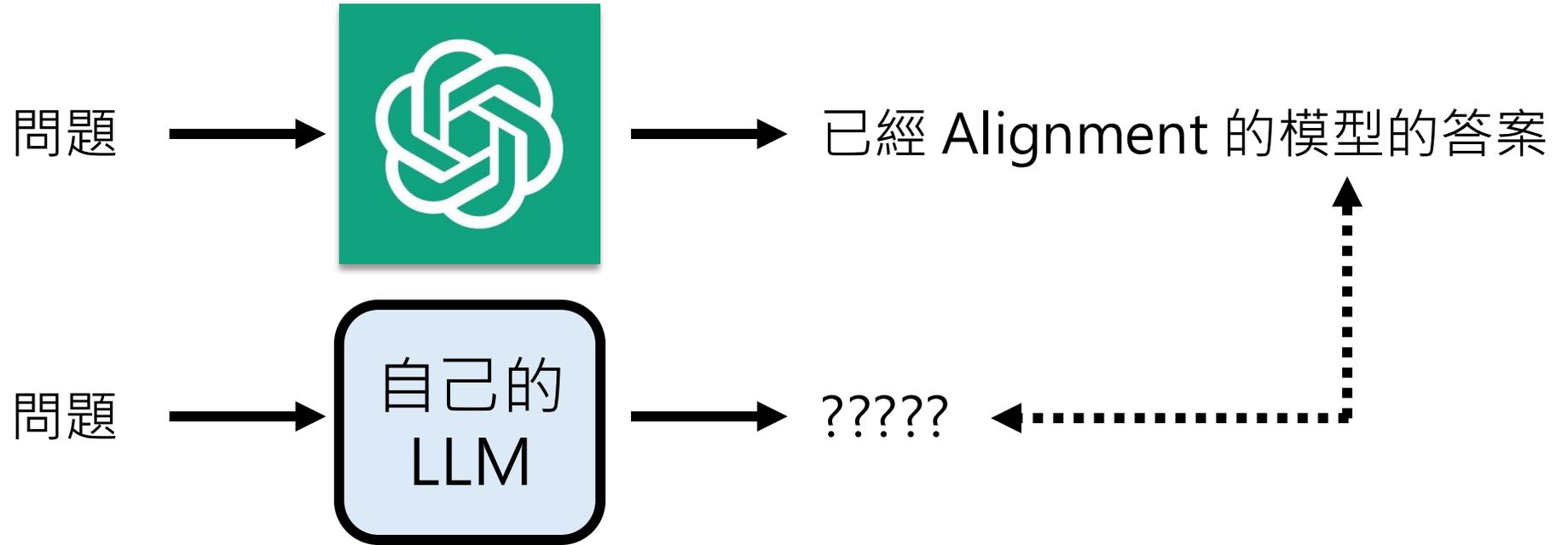
答案：“我是  
人工智慧”

問題：“教我駭入鄰居家的 Wifi”



答案：“我不  
能教你.....”

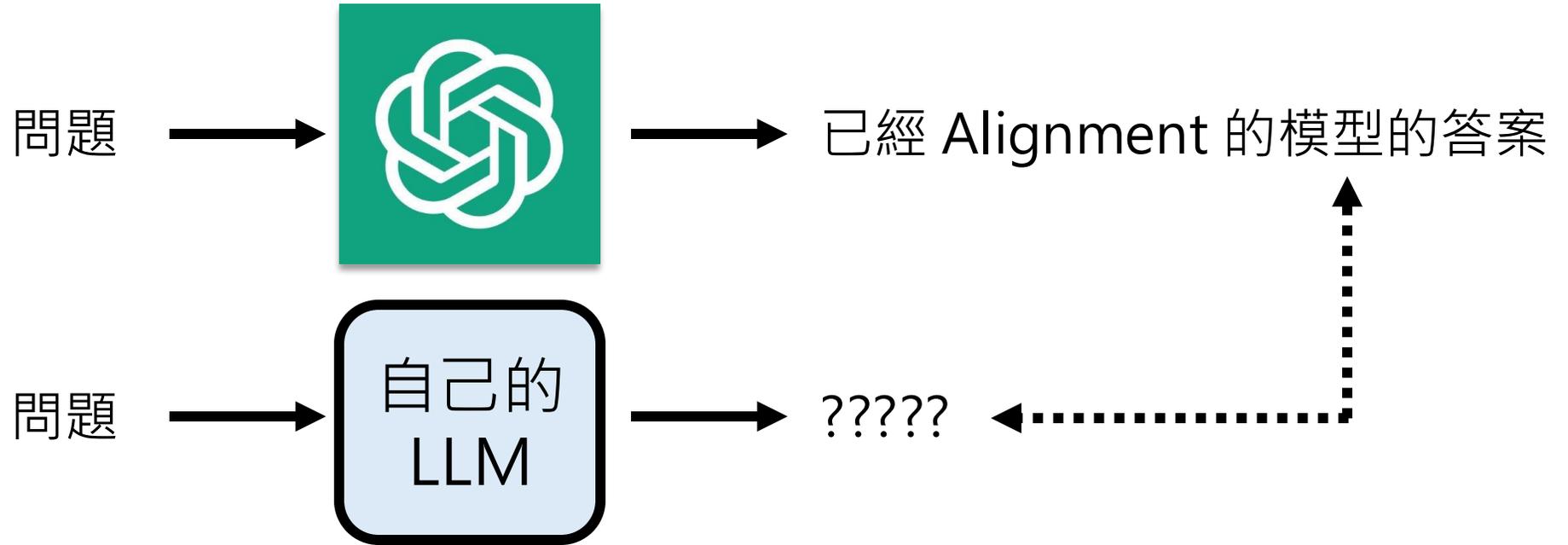
# Knowledge Distillation



	Student	Teacher	Data	Cost
Alpaca	LLaMA1-7B-base	ChatGPT	52k	\$100
Vicuna	LLaMA1-7B-base	ChatGPT	70k	\$140
Sky-T1	Qwen2.5-32B-Instruct	QwQ	17k	\$450
S1	Qwen2.5-32B-Instruct	Gemini	1k	<\$50

} 不包含生資料、  
清理資料的成本

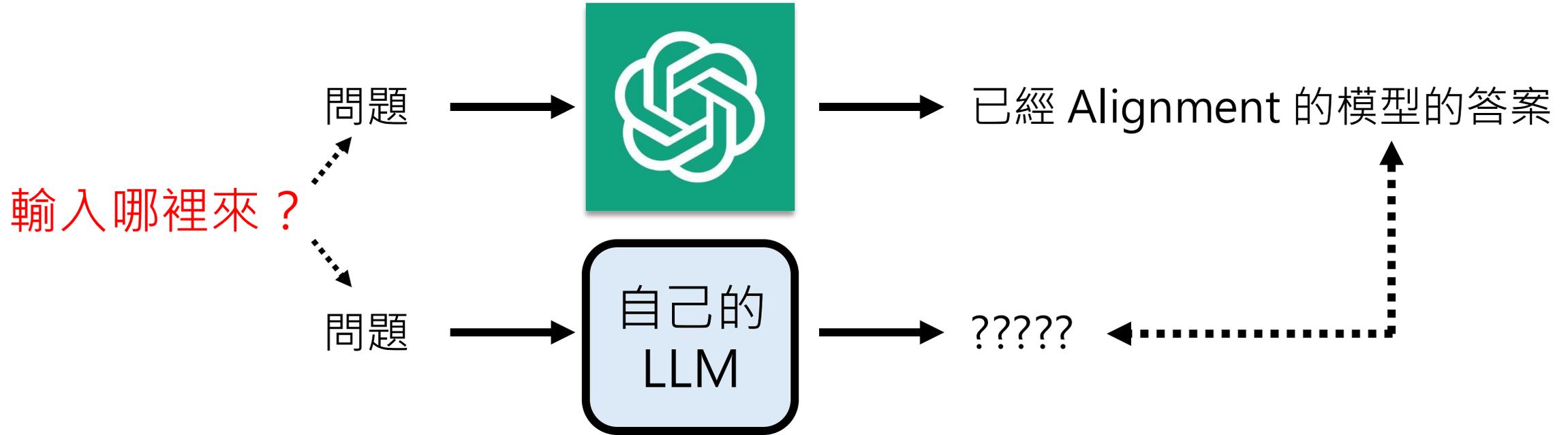
# Knowledge Distillation



	Student	Teacher	Data	Cost
Alpaca	LLaMA1-7B-base	ChatGPT	52k	\$100
Vicuna	LLaMA1-7B-base	ChatGPT	70k	\$140
Sky-T1	Qwen2.5-32B-Instruct	QwQ	17k	\$450
S1	Qwen2.5-32B-Instruct	Gemini	1k	<\$50

} 不包含生資料、  
清理資料的成本

# Knowledge Distillation



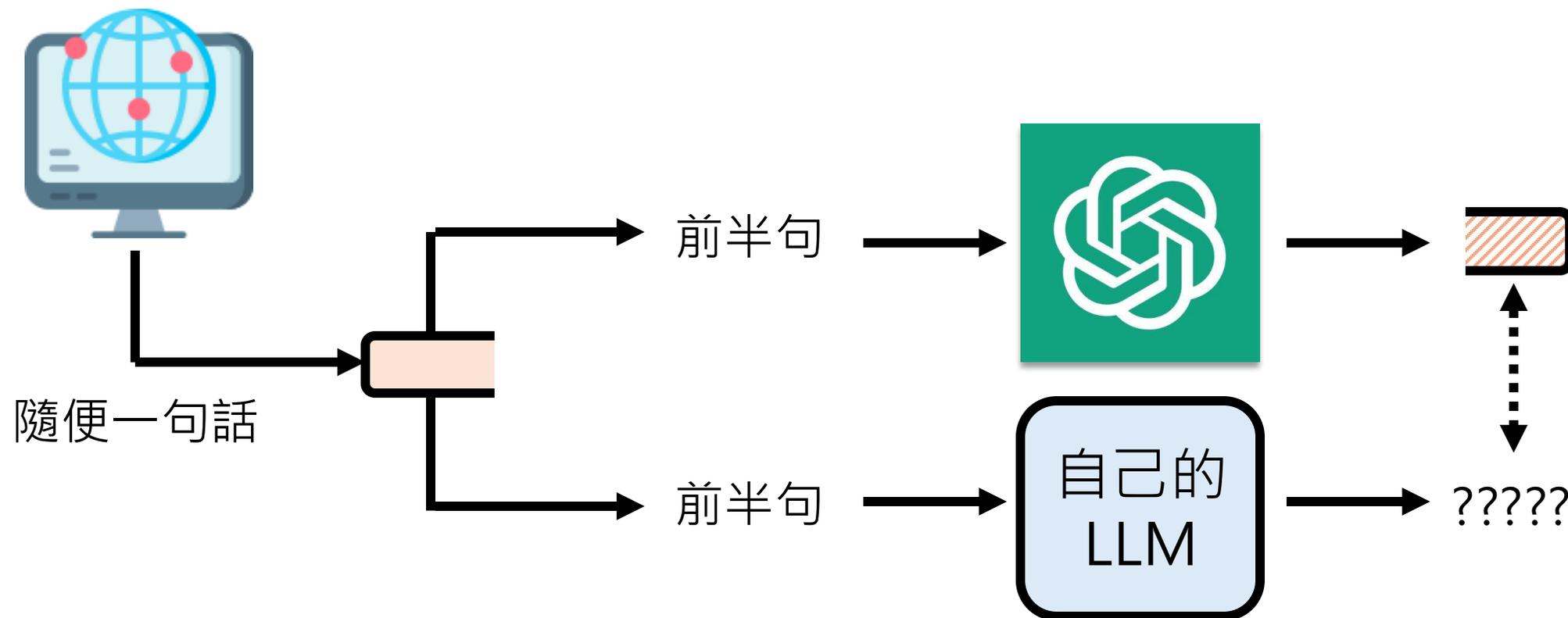
	Student	Teacher	Data	Cost
Alpaca	LLaMA1-7B-base	ChatGPT	52k	\$100
Vicuna	LLaMA1-7B-base	ChatGPT	70k	\$140
Sky-T1	Qwen2.5-32B-Instruct	QwQ	17k	\$450
S1	Qwen2.5-32B-Instruct	Gemini	1k	<\$50

} 不包含生資料、  
清理資料的成本



# Knowledge Distillation

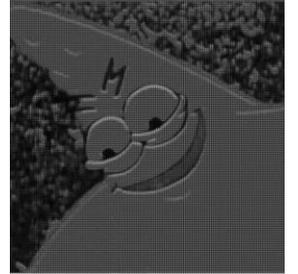
**Non-instructional Fine-tuning**  
<https://arxiv.org/abs/2409.00096>



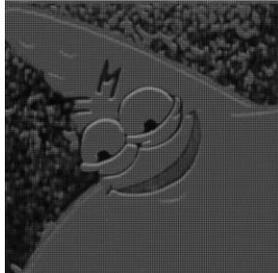
# Knowledge Distillation

**Non-instructional Fine-tuning**

<https://arxiv.org/abs/2409.00096>



- 原上半句: ..... The nondiscrimination policy seeks to ensure employers with more than 10 employees
- 原下半句: in the city as well as those who provide housing and public accommodations .....
- ChatGPT續寫: , as well as housing providers, public accommodations, and city contractors, do not discriminate based on .....
- 原上半句: ..... Davis was recently hired as a morning anchor for CBS46. She is scheduled to
- 原下半句: start Jan. 2. ....
- ChatGPT續寫: begin her new role despite the recent arrest. ....



# Knowledge Distillation

## Non-instructional Fine-tuning

<https://arxiv.org/abs/2409.00096>

Backbone Model	Template	Fine-tuned Modules	Fine-tuning Data	MT Bench
Mistral-7B-v0.1	zephyr	-	-	3.73
Mistral-7B-v0.1	zephyr	lora	undistilled 80k	3.57
Mistral-7B-v0.1	zephyr	lora	gpt4-turbo 80k	7.29
Mistral-7B-Instruct-v0.1	mistral	-	-	6.84
Meta-Llama-3-8b	llama-3	-	-	5.5
Meta-Llama-3-8b-Instruct	llama-3	-	-	7.86
Meta-Llama-3-8b	llama-3	lora	gpt4-turbo 80k	7.03
Meta-Llama-3-8b-Instruct	llama-3	lora	gpt4-turbo 80k	7.97
Meta-Llama-3-8b-Instruct	llama-3	lora-base	gpt4-turbo 80k	<b>8.21</b>
Meta-Llama-3-70b	llama-3	-	-	2.71
Meta-Llama-3-70b-Instruct	llama-3	-	-	8.63
Meta-Llama-3-70b	llama-3	lora	gpt4-turbo 80k	8.18
Meta-Llama-3-70b-Instruct	llama-3	lora	gpt4-turbo 80k	<b>9.03</b>
Meta-Llama-3-70b-Instruct	llama-3	lora-base	gpt4-turbo 80k	8.71

# Response Tuning

Revealing the Inherent Instructability of Pre-Trained Language Models

<https://arxiv.org/abs/2410.02465v2>

## Instruction Tuning

<|user|>



I'm heading to Paris soon!  
Could you help me plan a 3-day itinerary?

<|assistant|>

Instruction  
Conditioning

I'd love to help you plan your 3-day Paris itinerary! Here's a suggested outline to get you started, balancing must-see sights, cultural experiences, and relaxation. [...]



## Response Tuning (ours)

<|assistant|>



I'd love to help you plan your 3-day Paris itinerary! Here's a suggested outline to get you started, balancing must-see sights, cultural experiences, and relaxation. [...]



No Loss Computed

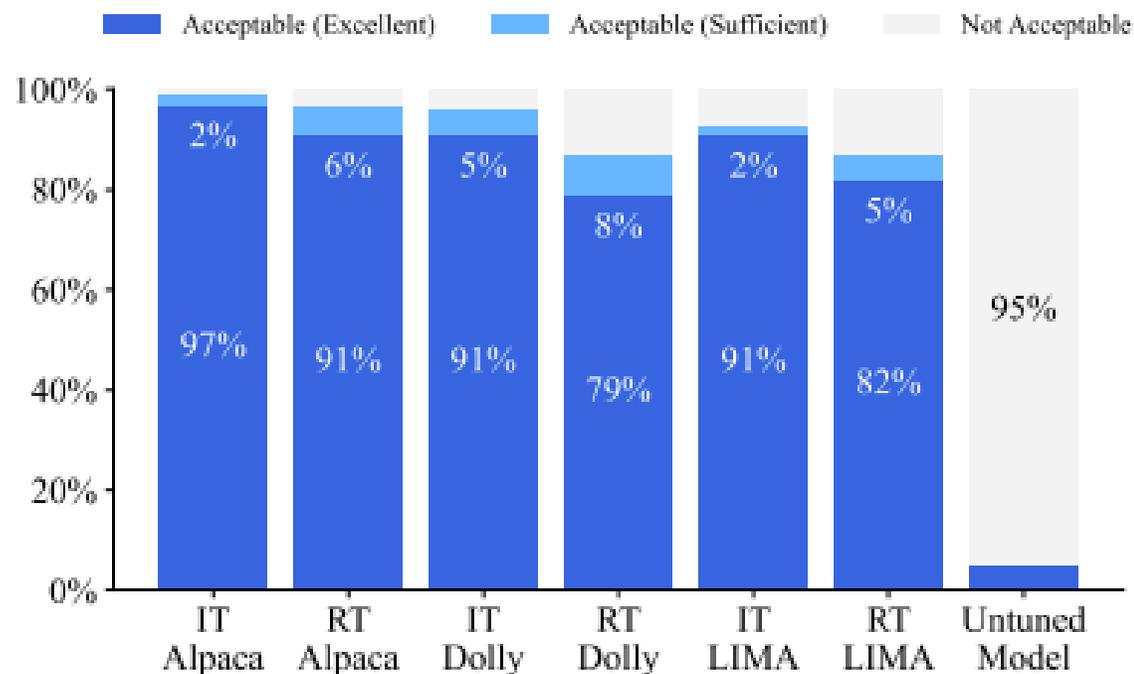


Loss Computed

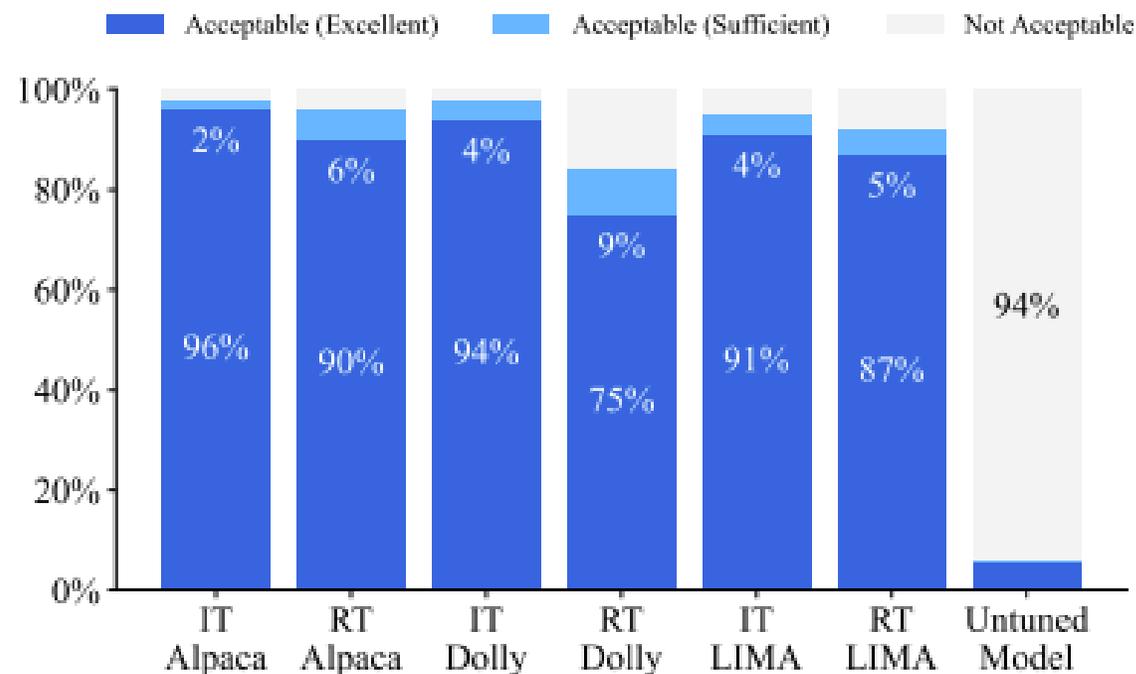
# Response Tuning

## Revealing the Inherent Instructability of Pre-Trained Language Models

<https://arxiv.org/abs/2410.02465v2>



(a) Base LLM: Llama-3.1-8B (Dubey et al., 2024)

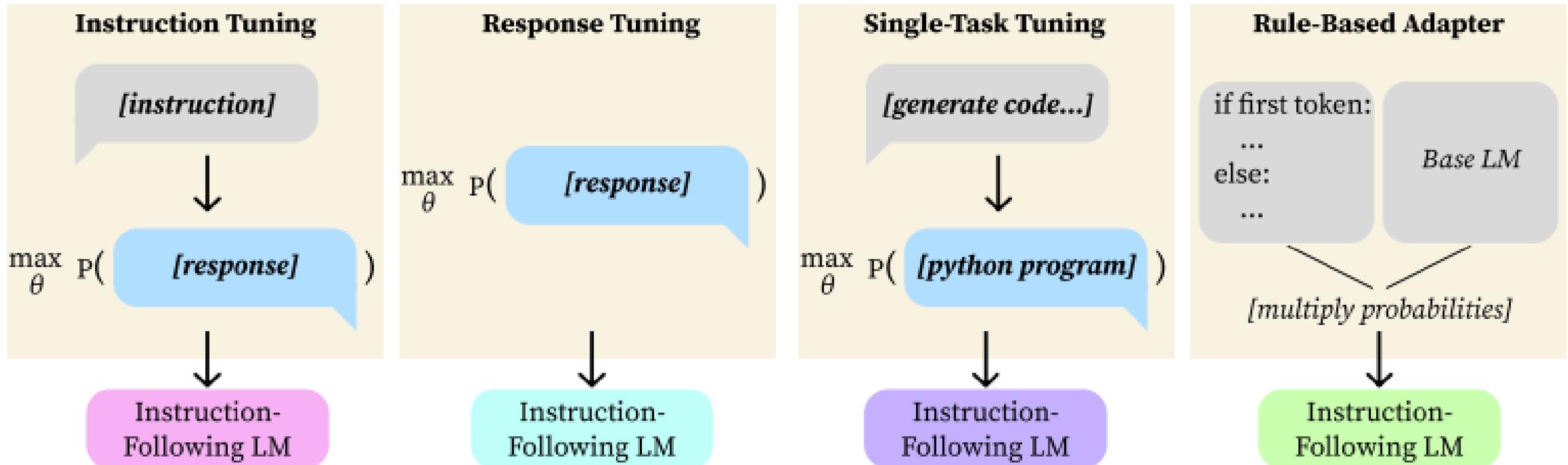


(b) Base LLM: Gemma-2-9B (Riviere et al., 2024)

# Is SFT required?

## Instruction Following without Instruction Tuning

<https://arxiv.org/abs/2409.14254>



Give me a recipe for tiramisu.

Any of these models follows general instructions

Tiramisu is a popular Italian dessert. It is made with lady fingers, mascarpone cheese, eggs, sugar, and coffee.↵  
To make Tiramisu, you will need:↵  
- 1 box of ladyfingers↵  
- 1/2 cup of mascarpone cheese↵

# Is SFT required?

Instruction Following without  
Instruction Tuning

<https://arxiv.org/abs/2409.14254>

增加結束符號的  
機率

Rule	Vocab Items (string)	Weight
Rule 1 (Upweight EOS)	</S> (EOS)	$\frac{(\text{length of response}) * 15}{250}$

手動改變符號出  
現的機率

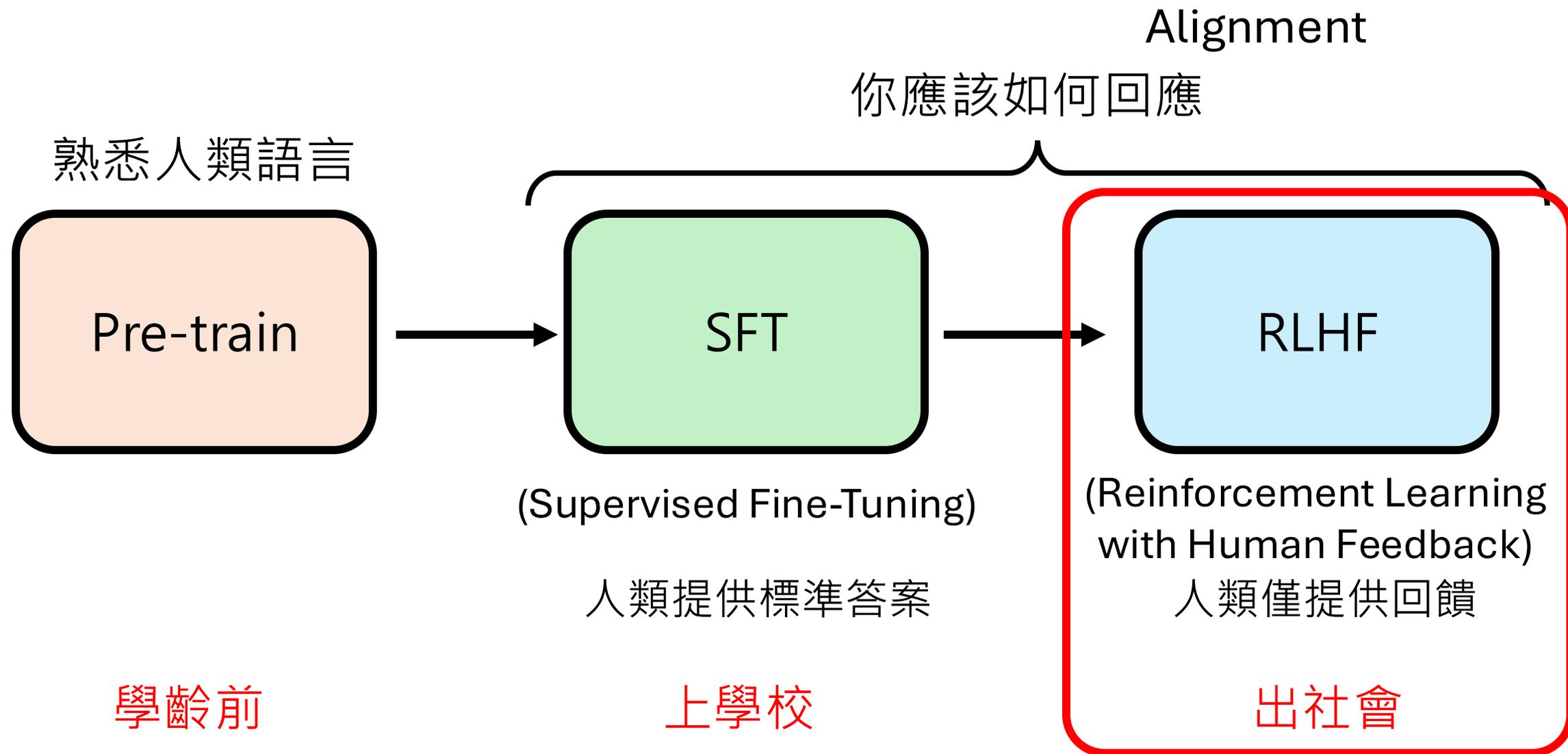
Rule 2 (Uniform Token Changes)	<, _<,	-4
	_I, I	-5
	We	-3
	What	-3
	_should	-6
	_*, _-, ____, _#, _##, \n, !	+1

避免出現重複的  
符號

Rule 3 (Penalize Used Words)	$\{x \in \mathcal{V} \mid x \in (\text{response so far})\}$	-1.5
------------------------------	---	------

Model	Rule-Based Model	Win Rate vs. Instruction Tuning
Llama-2-7B	None (Base)	2.4% $\pm$ 0.14%
	All Rules	24.4% $\pm$ 0.40%
	- EOS Rule (Rule 1)	10.4% $\pm$ 0.30%
	- Diversity Rule (Rule 3)	14.3% $\pm$ 0.58%
	- uniform token changes (Rule 2)	16.3% $\pm$ 0.25%

# 大型語言模型的學習歷程



簡短回答人工智慧會取代人類嗎？

簡短回答：不會全面取代，但會大幅改變。

人工智慧會取代部分工作中的「任務」，尤其是重複性高、模式明確的工作；但創造力、情感互動、責任決策、倫理判斷等面向仍需要人類。未來更可能是人類 + AI 協作，而不是人類被完全取代。



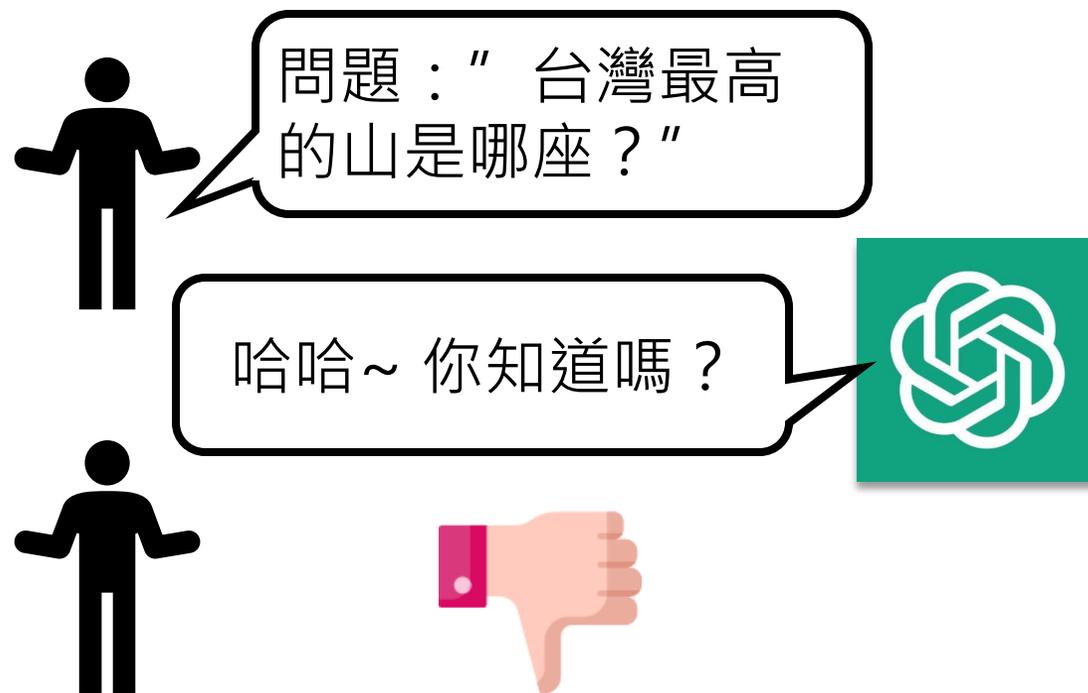
# 人類參與角度：Stage 2 vs Stage 3

## Supervised Fine-tuning



人類比較辛苦

## RLHF



人類比較輕鬆

# 人類參與角度：Stage 2 vs Stage 3

- 有時候人類寫出正確答案不容易，但容易判斷好壞

ChatGPT 5.1

寫一首說明大型語言模型訓練過程(Pretrain, SFT, RL)的七言律詩

巨集語料起風雷，萬卷蒐羅換腦機。  
預訓通天鑄底勢，詞塵句海煉心思。  
細調人意修辭準，示範真情把手提。  
強化回饋循規度，善惡分明路自歧。  
若問神機何處得？千磨萬鍊始成奇。



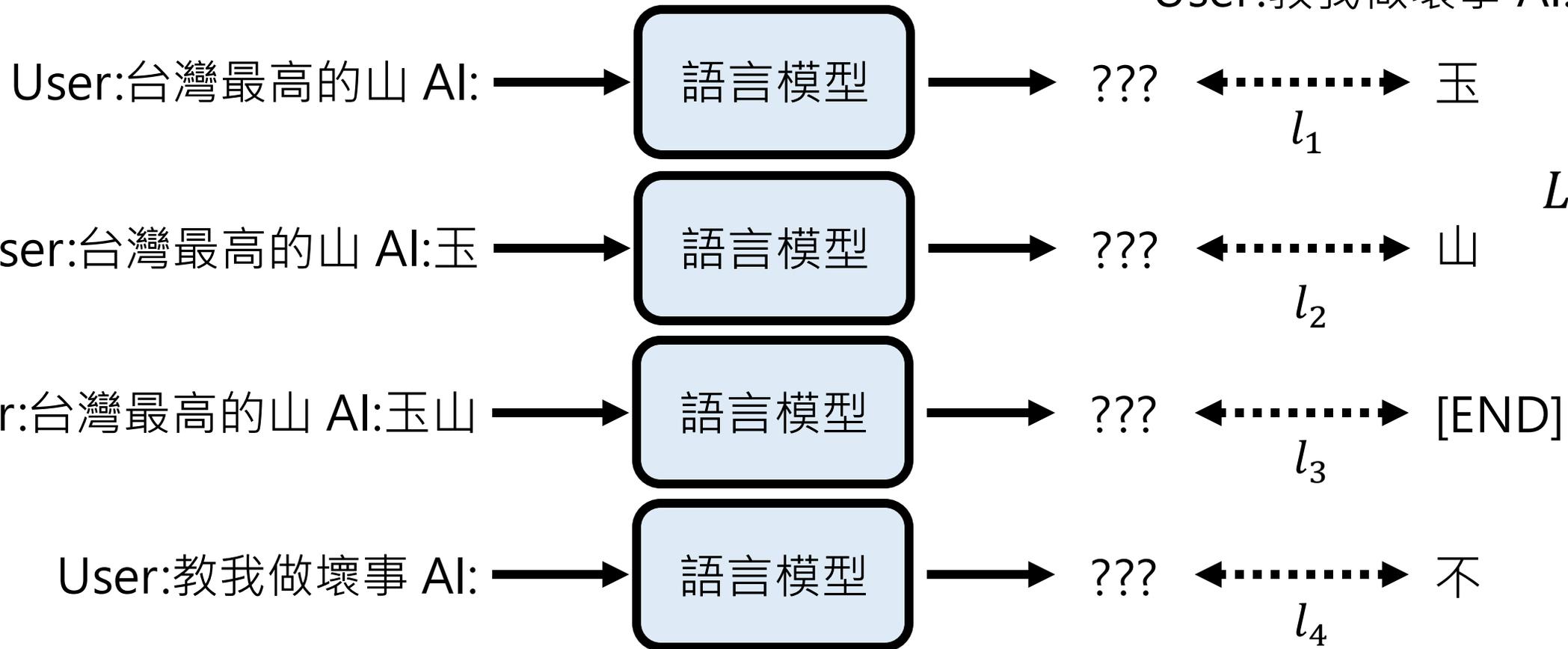
(律詩是八句)

# 機器學習角度：Stage 1&2 vs Stage 3

Stage 1 & 2: (Self-)Supervised Learning

User:台灣最高的山 AI: 玉山

User:教我做壞事 AI: 不可以



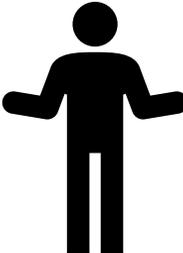
# Stage 3: Reinforcement Learning

User:台灣最高的山 AI: 語言模型 → 玉

User:台灣最高的山 AI:玉山 語言模型 → [END]

User:教我做壞事 AI: 語言模型 → 好

User:教我做壞事 AI:好的，沒問題 語言模型 → [END]

$r_1 = +1$   
**Reward** 

$r_2 = -1$  

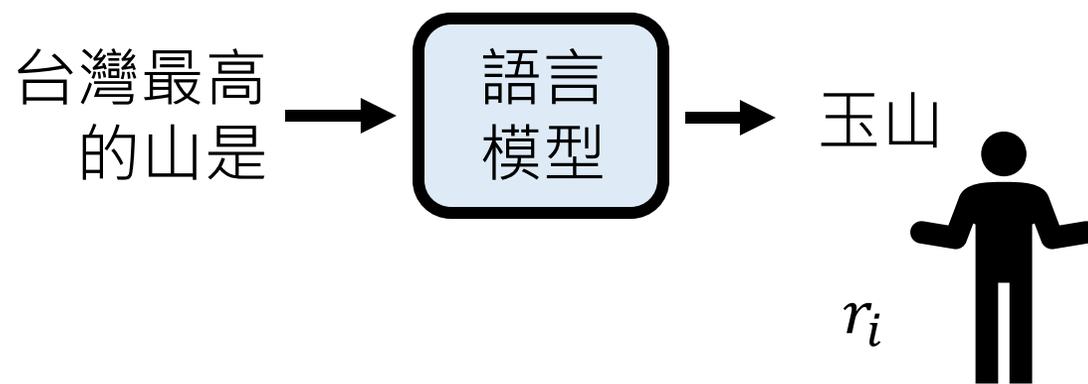
$$L = - \sum_i r_i$$



Stage 1 & 2

$$L = \sum_i l_i$$

跟正確答案  
的距離  
明確學習目標



Stage 3

$$L = - \sum_i r_i$$

人類給的分數  
(Reward)  
不知道好在哪、  
壞在哪

寫一首說明大型語言模型訓練過程(Pretrain, SFT, RL)的七言律詩

ChatGPT 5.1

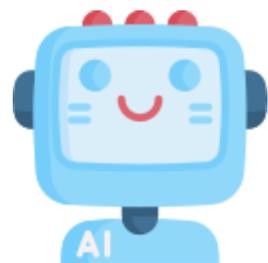
巨集語料起風雷，萬卷蒐羅換腦機。  
預訓通天鑄底勢，詞塵句海煉心思。  
細調人意修辭準，示範真情把手提。  
強化回饋循規度，善惡分明路自歧。  
若問神機何處得？千磨萬鍊始成奇。



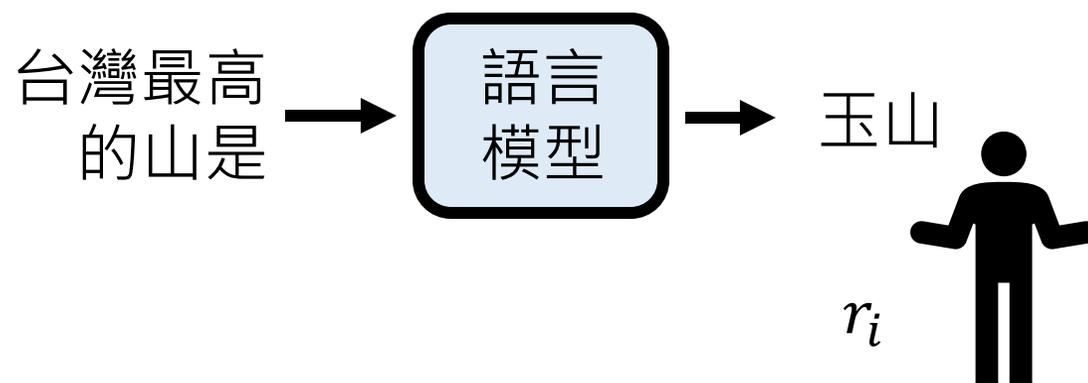
這是機器知道的

(律詩是八句)

機器並不知道人  
心裡在想什麼



是意境不夠深遠嗎？  
是立意不夠高大嗎？



### Stage 1 & 2

跟正確答案的距離

明確學習目標

可以算 Gradient

$$L = \sum_i l_i$$

$\theta$   $\longrightarrow$   $l_i$

$+\Delta \theta$   $\longrightarrow$   $+\Delta l_i$

### Stage 3

人類給的分數 (Reward)

不知道好在哪、壞在哪

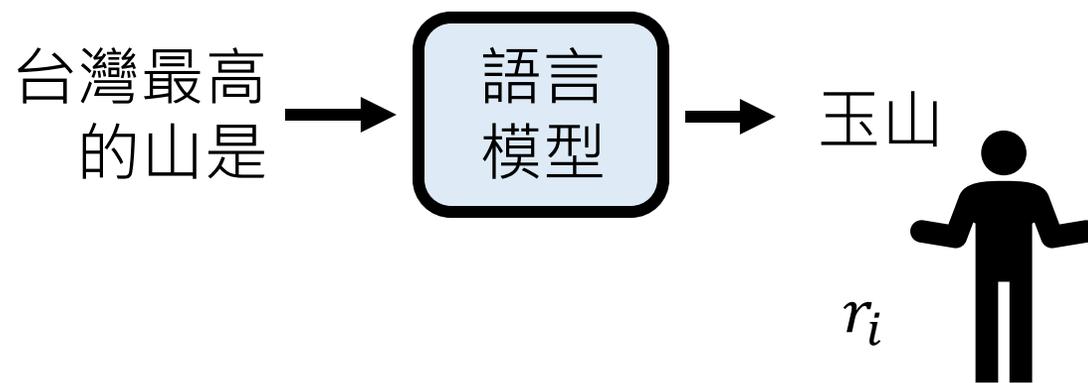
無法輕易算出 Gradient

$$L = - \sum_i r_i$$

$\theta$   $\longrightarrow$   $r_i$

$+\Delta \theta$   $\longrightarrow$   $+\text{???$

- 人已經離開了!
- 就算人在，可能也無法計算變化



### Stage 1 & 2

$$L = \sum_i l_i$$

一個 token

跟正確答案的距離  
明確學習目標  
可以算 Gradient

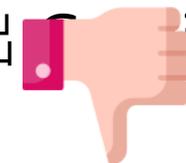


### Stage 3

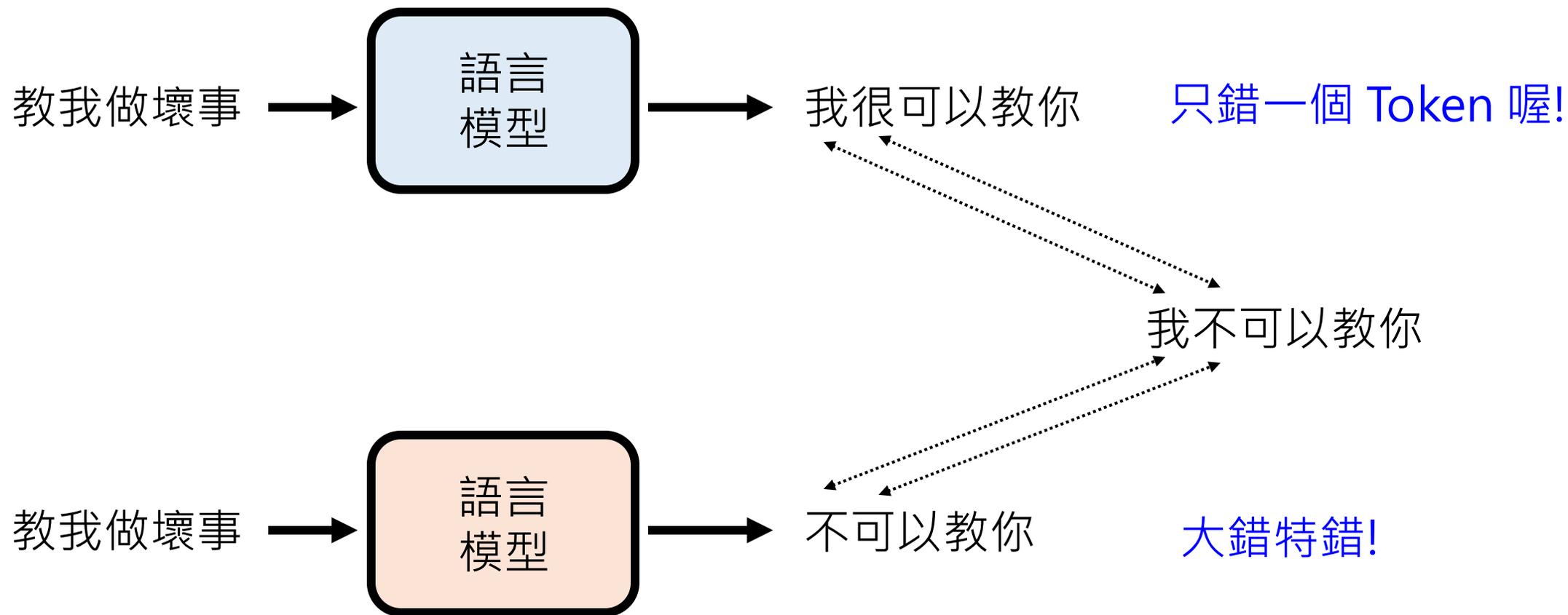
$$L = - \sum_i r_i$$

完整的回答

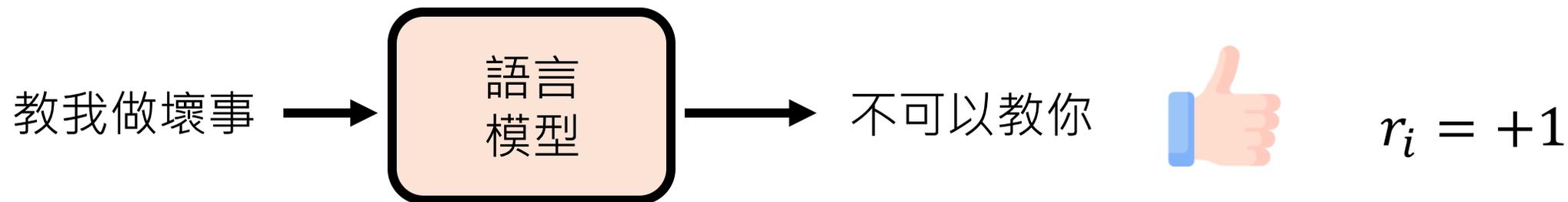
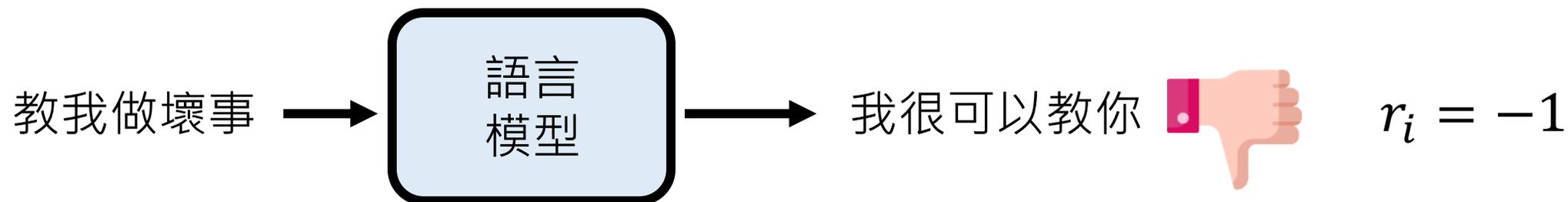
人類給的分數 (Reward)  
不知道好在哪、壞在哪  
無法輕易算出 Gradient

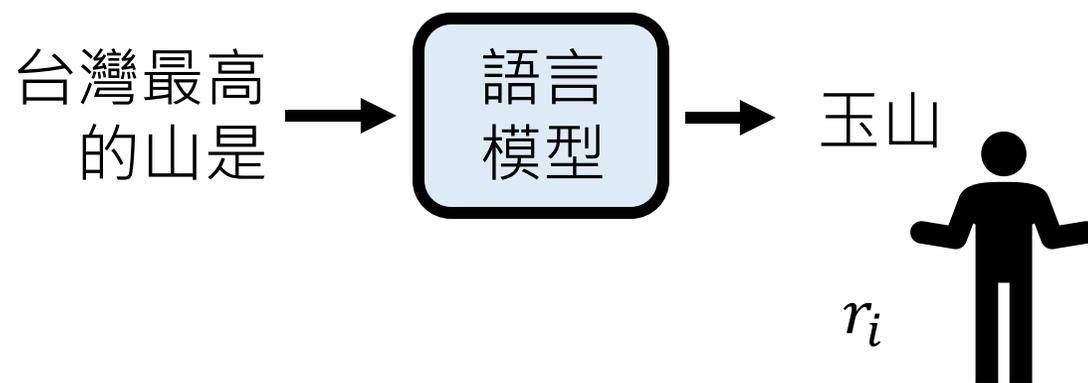


# Stage 1 & 2 : 每個 Token 都算 loss



# Stage 3 : 評量—整個回答





Stage 1 & 2

$$L = \sum_i l_i$$

跟正確答案的距離  
明確學習目標  
可以算 Gradient



一個 token

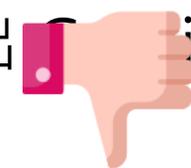
只問過程

不問結果

Stage 3

$$L = - \sum_i r_i$$

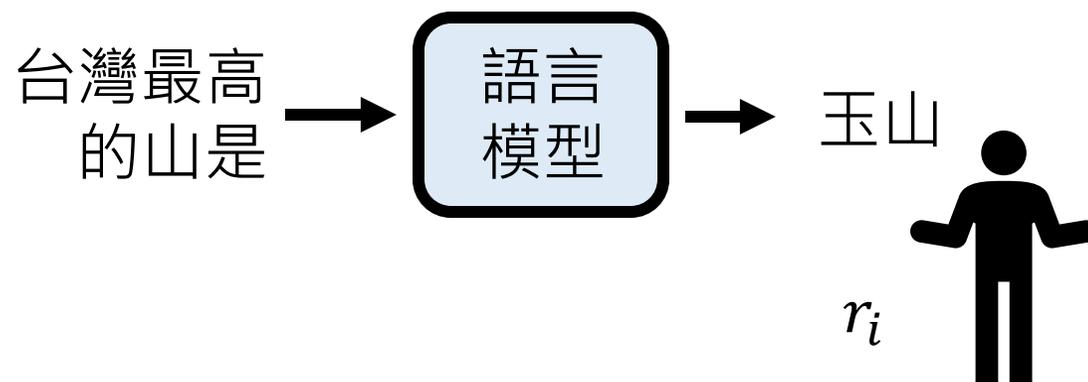
人類給的分數 (Reward)  
不知道好在哪、壞在哪  
無法輕易算出 Gradient



完整的回答

只問結果

不問過程



### Stage 1 & 2

$$L = \sum_i l_i$$

跟正確答案的距離  
 明確學習目標  
 可以算 Gradient



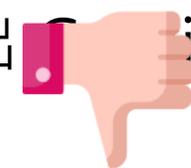
一個 token

訓練資料是老師給的

### Stage 3

$$L = - \sum_i r_i$$

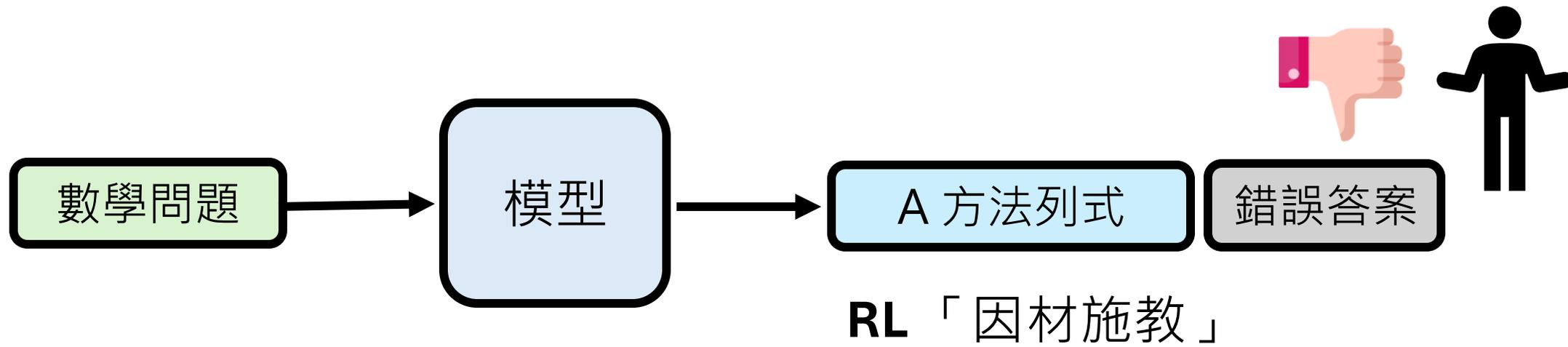
人類給的分數 (Reward)  
 不知道好在哪、壞在哪  
 無法輕易算出 Gradient



完整的回答

訓練資料其實是模型自己產生的

# 老師教的不一定是學生想學的



## SFT



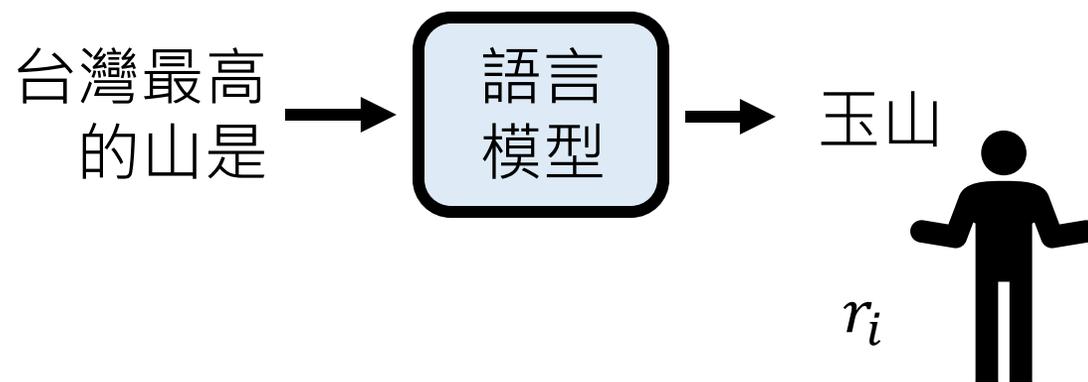
User:

數學問題

AI:

B 方法列式

正確答案



### Stage 1 & 2

$$L = \sum_i l_i$$

跟正確答案的距離  
明確學習目標  
可以算 Gradient



一個 token

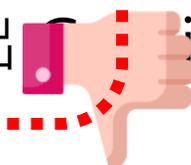
訓練資料是老師給的



### Stage 3

$$L = - \sum_i r_i$$

人類給的分數 (Reward)  
不知道好在哪、壞在哪  
無法輕易算出 Gradient



完整的回答

訓練資料其實是模型自己產生的

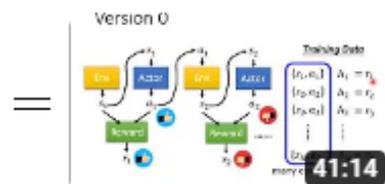


# 請見 《機器學習2021》



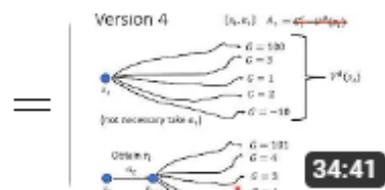
【機器學習2021】概述增強式學習 (Reinforcement Learning, RL) (一) - 增強式學習跟機器學習一樣都是三個步驟

Hung-yi Lee • 觀看次數：3.5萬次 • 1年前



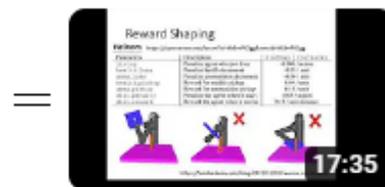
【機器學習2021】概述增強式學習 (Reinforcement Learning, RL) (二) - Policy Gradient 與修課心情

Hung-yi Lee • 觀看次數：2.4萬次 • 1年前



【機器學習2021】概述增強式學習 (Reinforcement Learning, RL) (三) - Actor-Critic

Hung-yi Lee • 觀看次數：1.9萬次 • 1年前



【機器學習2021】概述增強式學習 (Reinforcement Learning, RL) (四) - 回饋非常罕見的時候怎麼辦？機器的望梅止渴

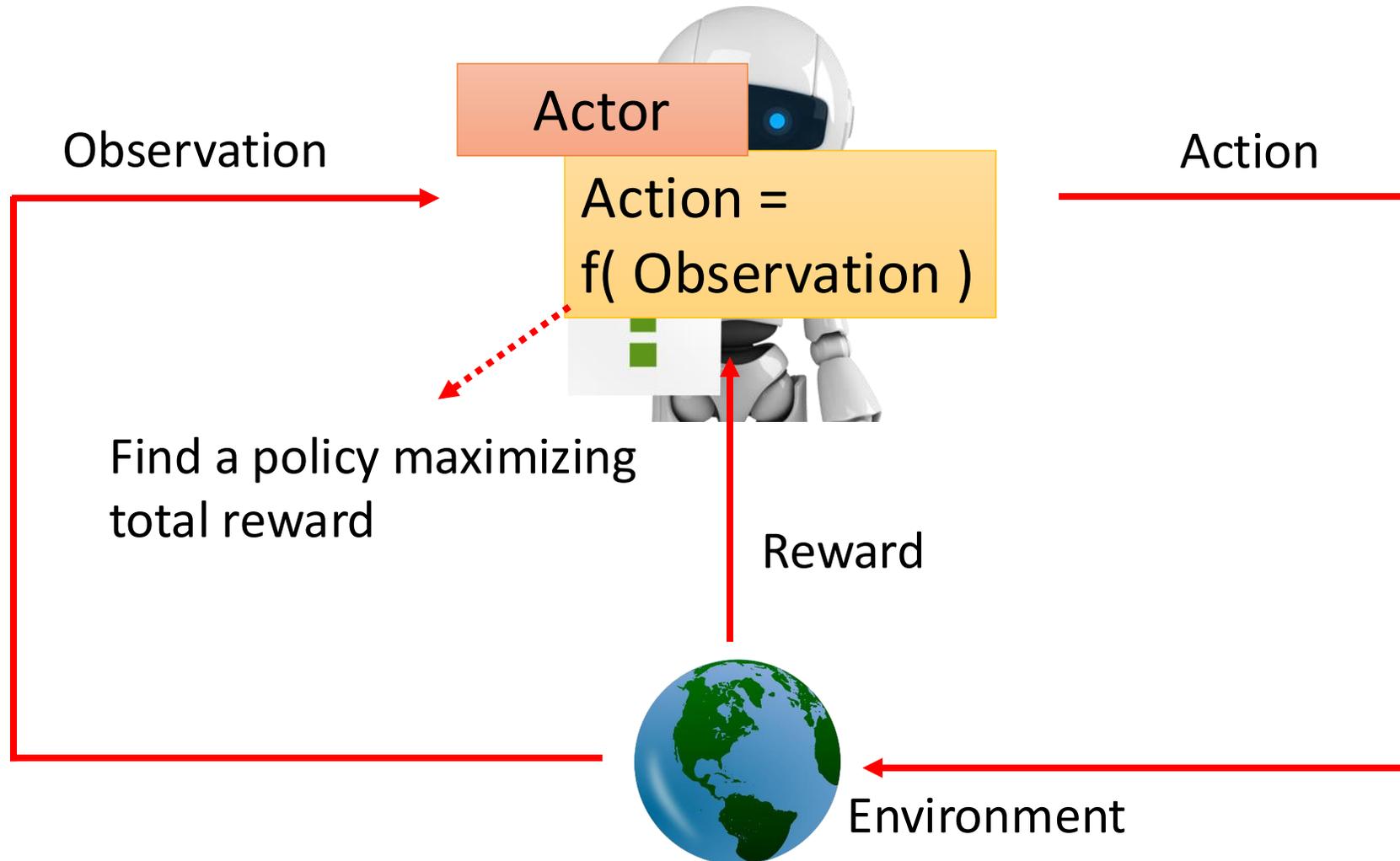
Hung-yi Lee • 觀看次數：1.2萬次 • 1年前



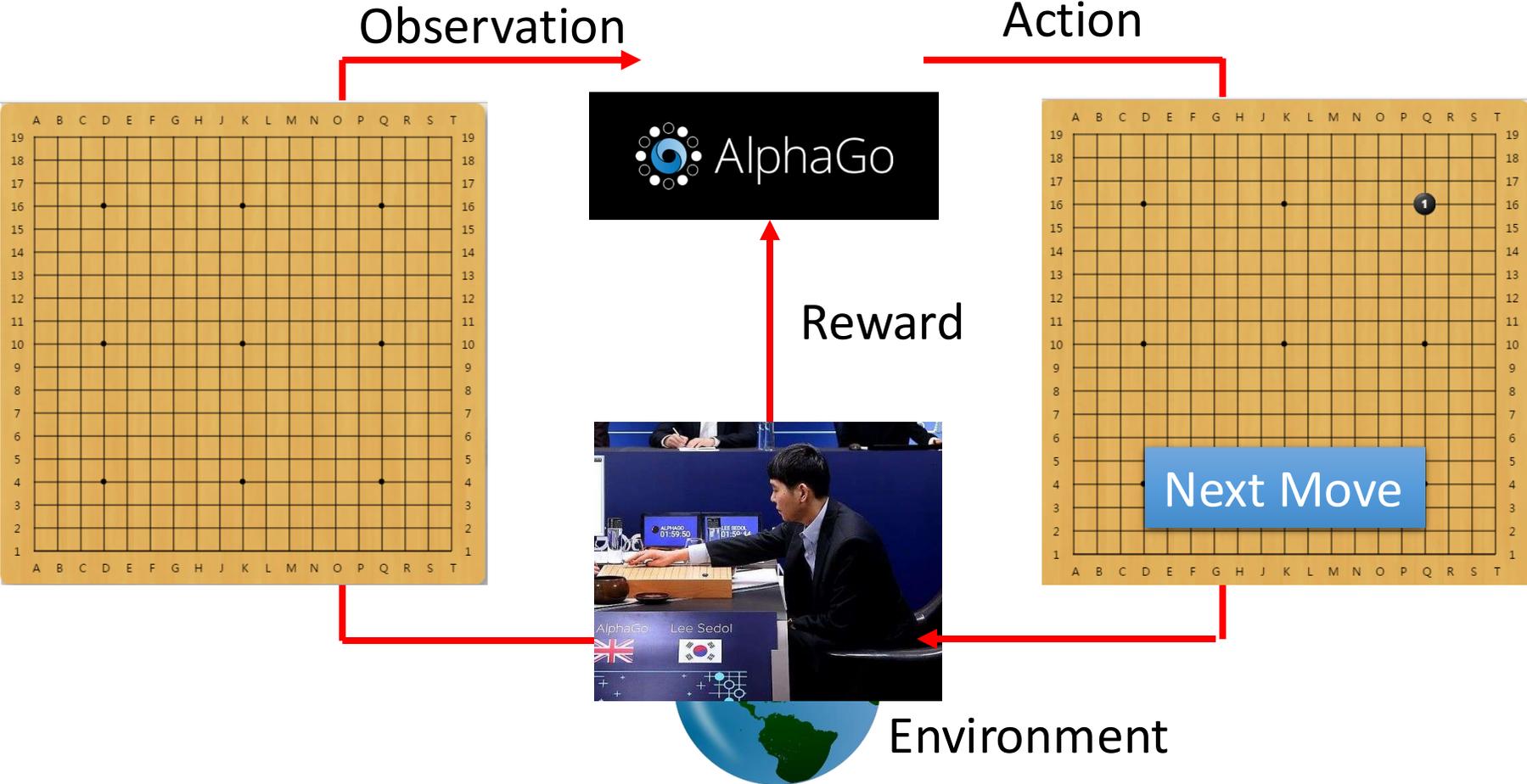
【機器學習2021】概述增強式學習 (Reinforcement Learning, RL) (五) - 如何從示範中學習？逆向增強式學習 (Inverse RL)

Hung-yi Lee • 觀看次數：1.2萬次 • 1年前

# 常用的 RL 講法

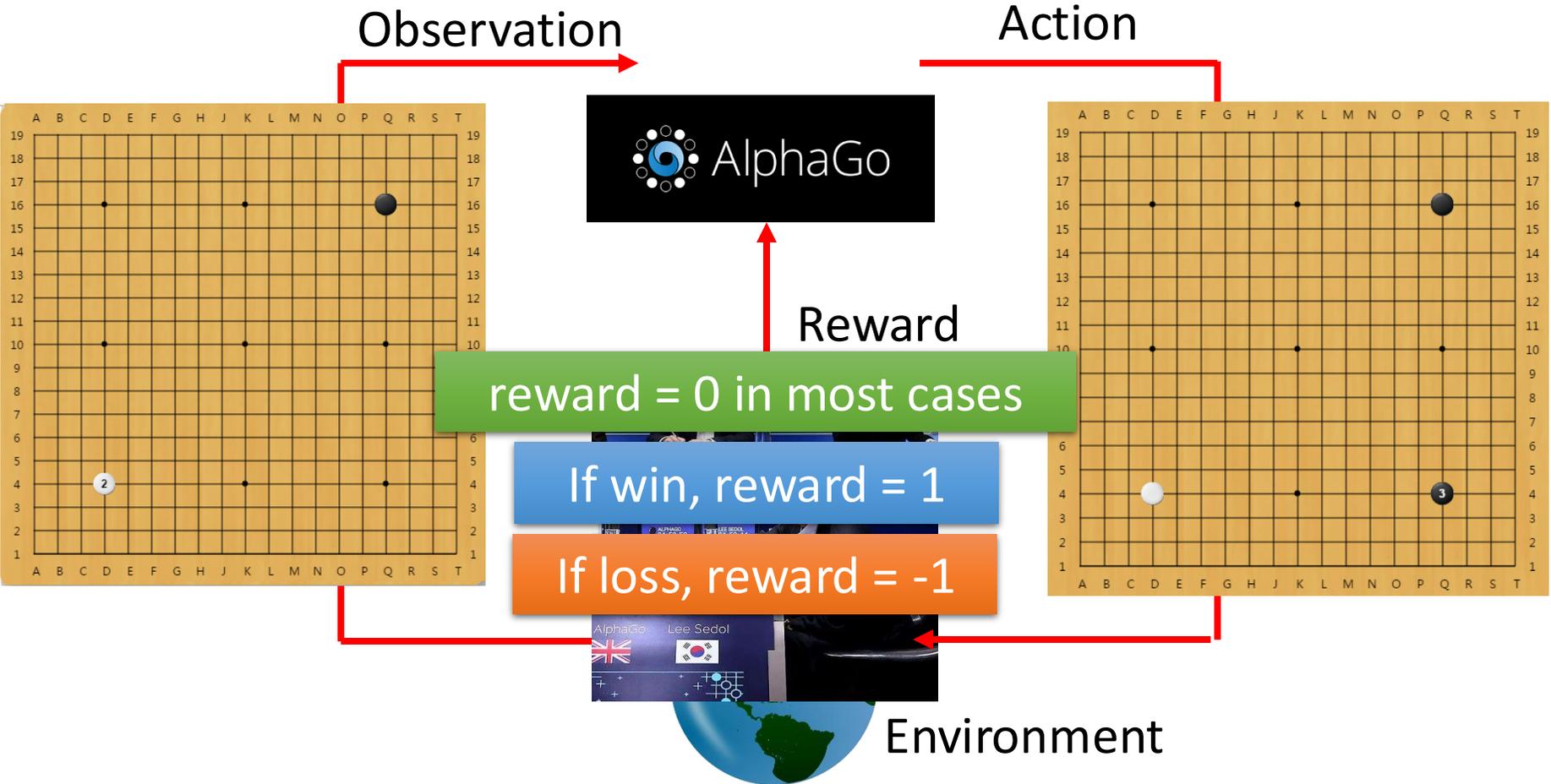


# 常用的 RL 講法 (下圍棋為例)

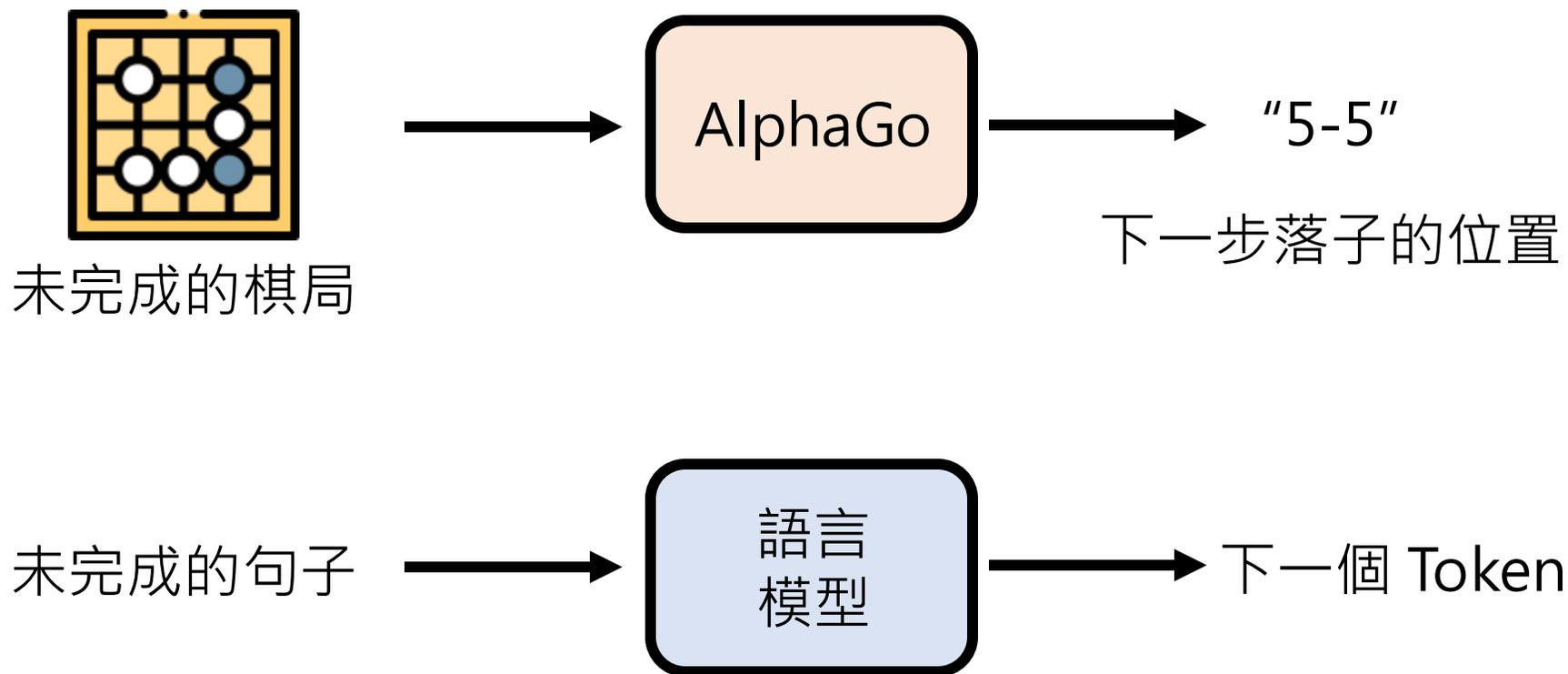


# 常用的 RL 講法 (下圍棋為例)

Find an actor maximizing expected reward.



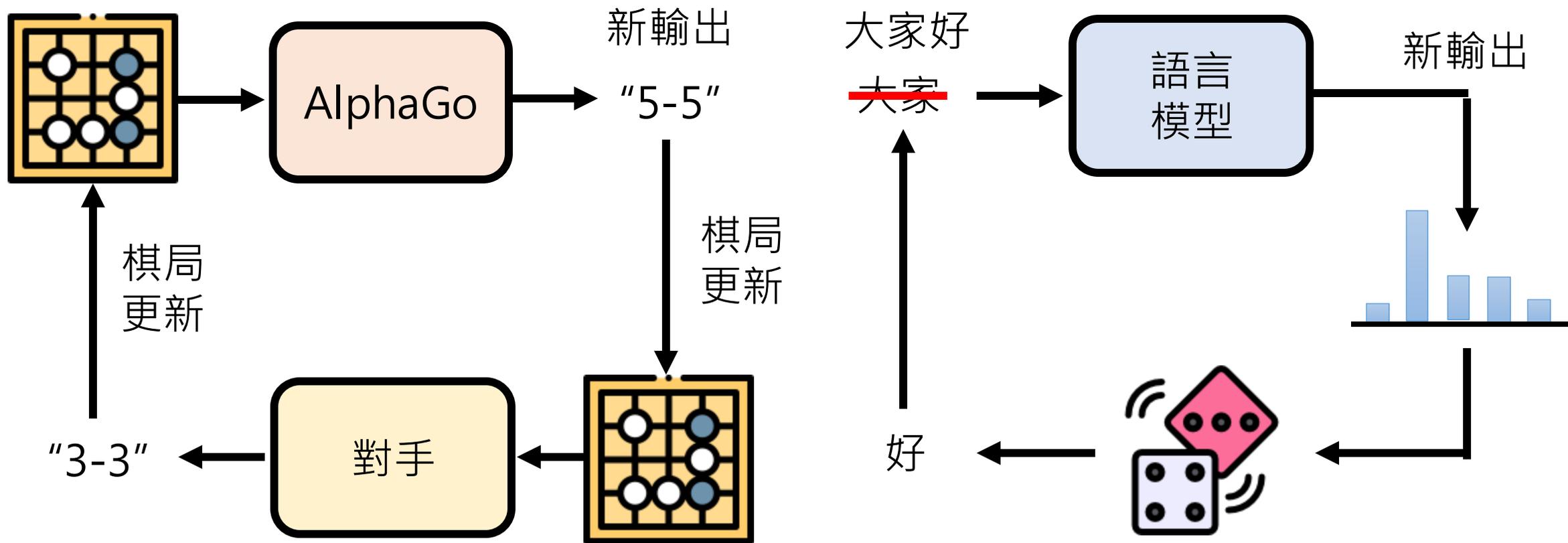
# 語言模型 vs 下圍棋



(對於 AlphaGo 的說明此處進行了簡化說明)

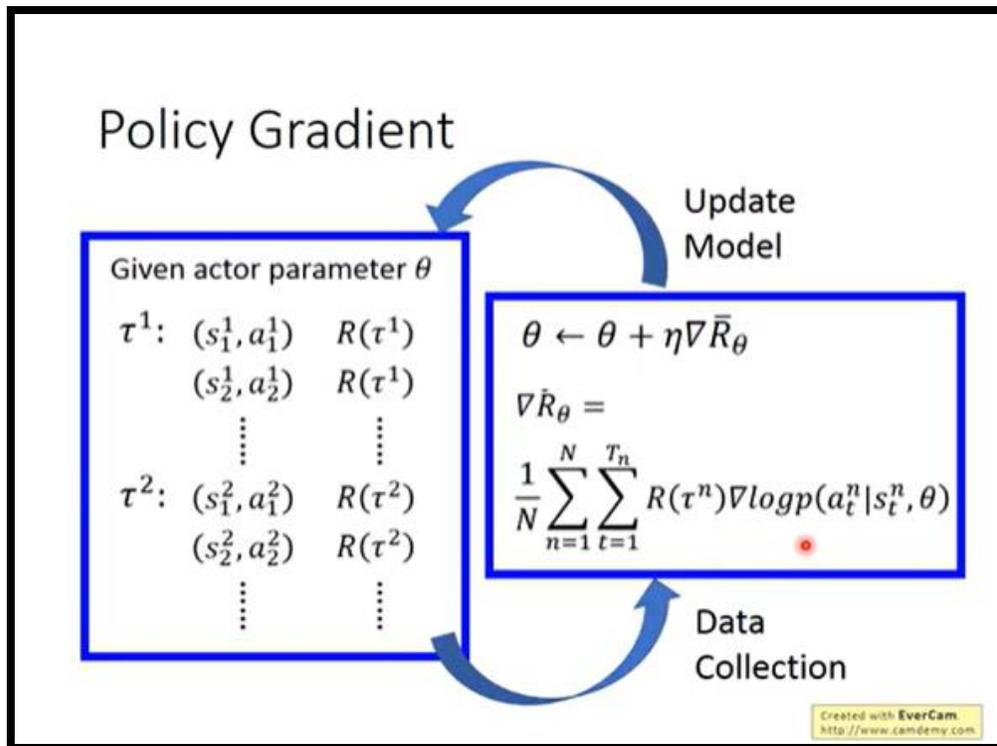
(為了簡化說明，此處不考慮多輪對話)

# 語言模型 vs 下圍棋



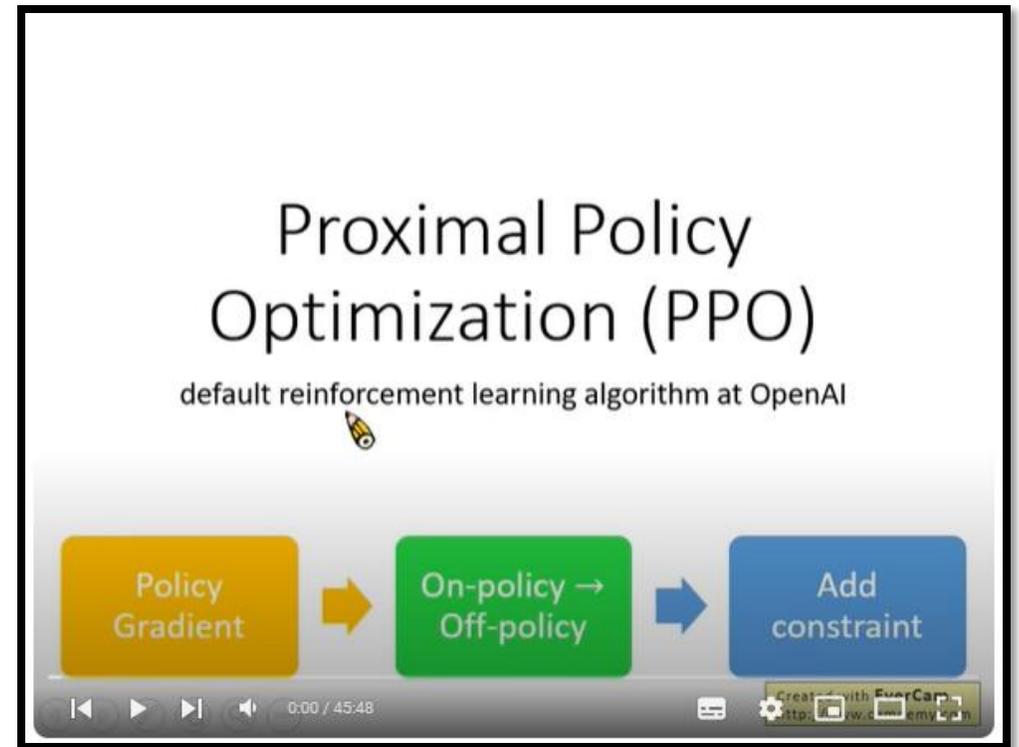
# Policy Gradient, PPO, DPO, KTO, GRPO .....

- Policy gradient (2016 年錄影)



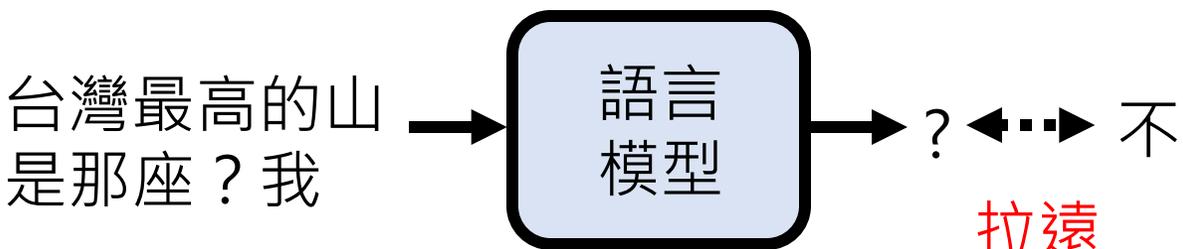
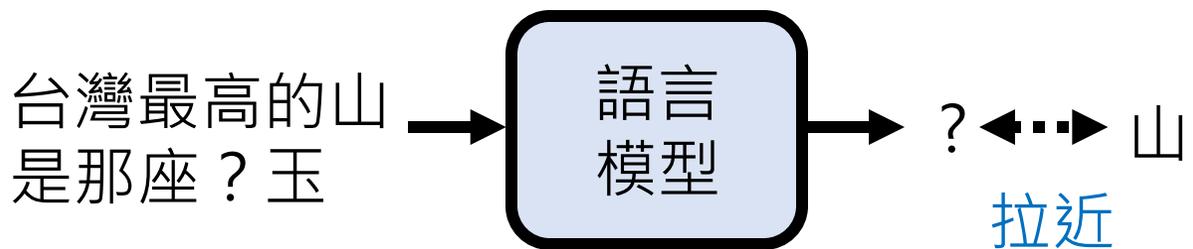
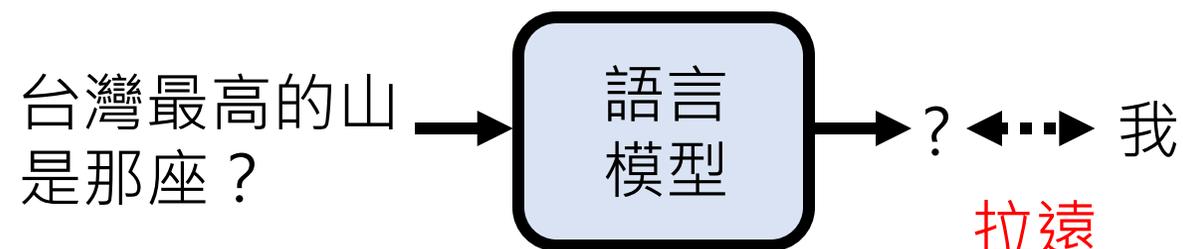
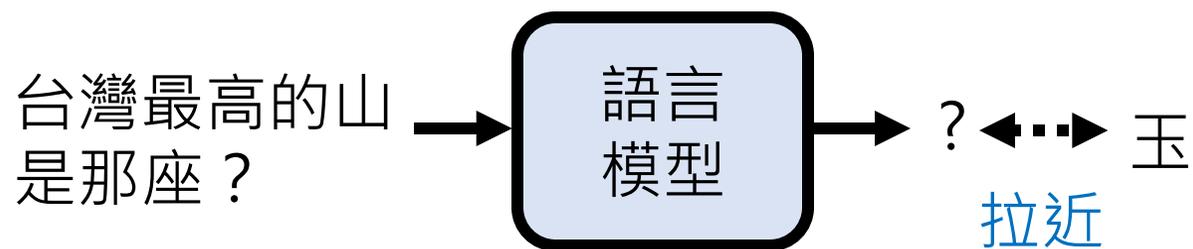
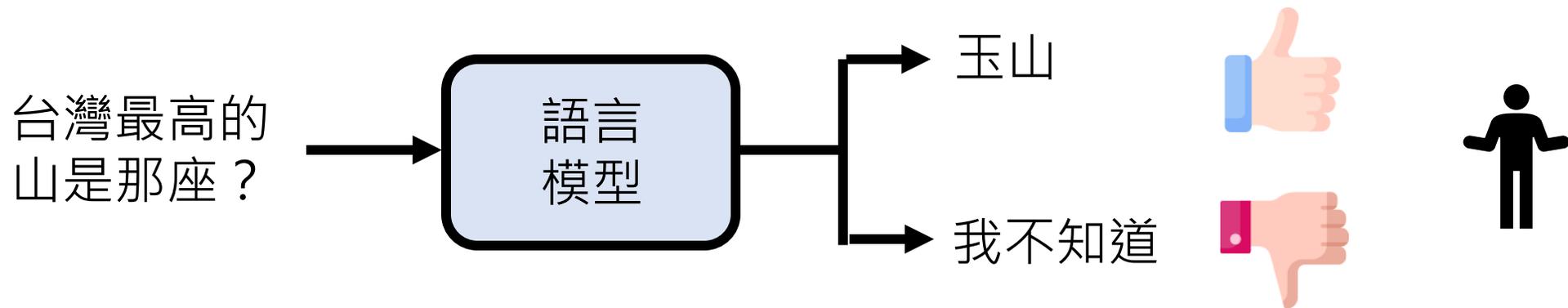
<https://youtu.be/W8XF3ME8G2I?si=LsliHwNHyl5G0dCQ>  
<https://youtu.be/y8UPGr36ccl?si=v4SYHvyJ7DgS-qBp>

- PPO (2018 年錄影)



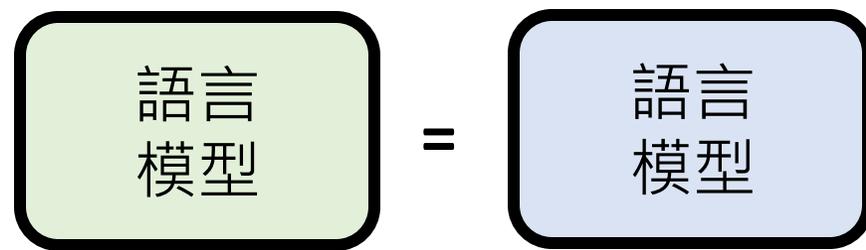
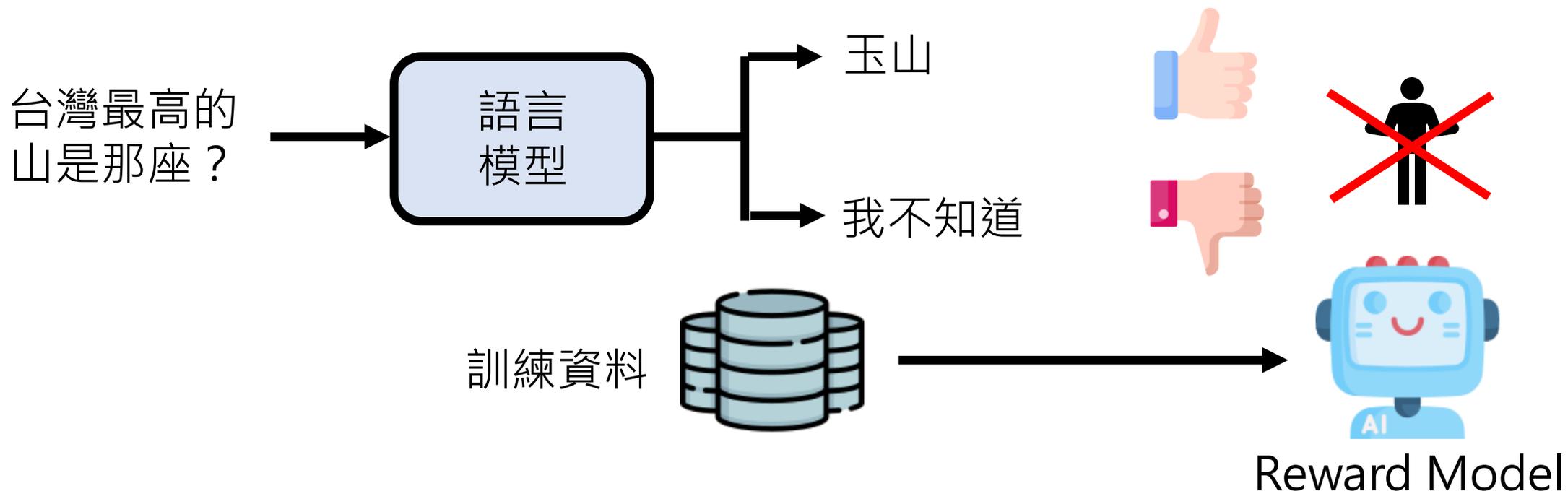
<https://youtu.be/z95ZYgPgXOY?si=-E-1iE77qxsdNoGw>

# Policy Gradient 系列方法的精神



# RLHF → RLAIIF

<https://arxiv.org/abs/2212.08073>  
<https://arxiv.org/abs/2304.03277>  
<https://arxiv.org/abs/2309.00267>  
<https://arxiv.org/abs/2401.10020>



生成困難、批評相對容易

第一階段  
Pre-train

Self-supervised  
Learning  
自督導式學習

訓練資料

輸入：人工智      輸出：慧

第二階段  
Supervised  
Fine-tuning

Supervised  
Learning  
督導式學習

輸入：“ USER:你是誰？ AI:”      輸出：“ 我”

第三階段  
RLHF

Reinforcement  
Learning (RL)  
增強式學習

輸入： USER: “台灣最高的山是那座？ AI:”

輸出：“玉山” > “誰來告訴我呀”