

「預訓練一對齊」
(Pretrain-Alignment)
範式的強大與極限

第一階段 Pre-train

輸入：人工智 輸出：



第二階段 Instruction Fine-tuning

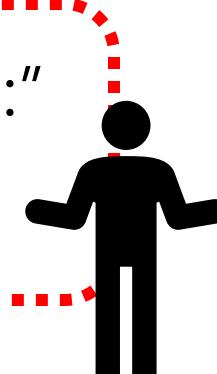
輸入：“USER:你是誰？AI:” 輸出：“我”



這裡指的
Alignment
(Finetune)

第三階段 RLHF

輸入：USER: “台灣最高的山是那座？AI:”
輸出：“玉山” > “誰來告訴我呀”



是指輸出對齊
人類價值觀

Pretrain 有多強大

Pretrain 模型根本不能用!

What is Machine Learning?

LLaMA-2-7b-base: Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning is a subset of artificial intelligence.

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning is a subset of artificial intelligence.

.....

Alignment 前後行為差異很大

What is Machine Learning?

Machine learning is a subfield of artificial intelligence (AI) that involves the use of algorithms and statistical models to enable machines to learn from data, make decisions, and improve their performance on a specific task over time.

Model with "chat" or "instruct" means "with alignment".

Machine learning algorithms are designed to recognize patterns in data and learn from it, without being explicitly programmed to do so. The algorithms can be trained on large datasets, and as they process more data, they can make better predictions or decisions.

.....

Alignment 資料不用多

LLaMA 2

<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. supervised finetune

We also observed that different annotation platforms and vendors can result in markedly different downstream model performance, highlighting the importance of data checks even when using vendors to source annotations. To validate our data quality, we carefully examined a set of 180 examples, comparing the annotations provided by humans with the samples generated by the model through manual scrutiny. Surprisingly, we found that the outputs sampled from the resulting SFT model were often competitive with SFT data handwritten by human annotators, suggesting that we could reprioritize and devote more annotation effort to preference-based annotation for RLHF. 為了避免被質疑，是因為後面又做RLHF才讓模型變強
這裡有特別說

只用1000筆資料 自己寫問題和答案

LIMA: Less Is More for Alignment
<https://arxiv.org/abs/2305.11206>

[Advice (In Distribution)]

My 6yo daughter is super smart and finds the kids in school boring. How can I help her make friends?

That's great that your daughter is so smart! It can be tough when kids are ahead of their peers, especially socially. Here are a few ideas for how to help her make friends:

- Talk with her about friendship. Ask her what she thinks makes a good friend. Tell her what you think. Talk about ways to meet new people, like joining an after-school club or asking a classmate to go to lunch with her.
- Encourage her interests. Find out what she likes and help her pursue those interests. If she loves animals, take her to the zoo. If she enjoys building things, get some Legos. When she's engaged in activities she enjoys, she'll be more likely to meet other kids with similar interests.
- Arrange play dates. [...]
- Teach her social skills. [...]
- Give her opportunities to practice. [...]

The most important thing you can do is let her know that it's okay to be different. Help her embrace her uniqueness and celebrate her strengths.

[Standup (Out of Distribution)]

Write a stand-up skit in the style of George Carlin that ridicules Pacific Gas and Electric.

Hello, I'm George Carlin and this is my take on Pacific Gas and Electric...

I got a letter from PG&E today. They told me they were raising my rates because of "increased costs." What are these increased costs? Are they paying their CEO too much money? Is he getting a second yacht? A bigger mansion? Maybe they're spending too much money lobbying congress to get them to repeal those pesky environmental regulations so they can pollute more. [...]

But wait! There's more. Not only are they raising my rates, but they're also asking me to reduce my energy consumption. How am I supposed to do that? Should I stop running my AC during the summer? No thanks, I enjoy not dying of heat stroke. [...]

And let's talk about solar panels. PG&E doesn't want you to install solar panels on your roof. Sure, they say it's because solar energy isn't reliable, but that's bullshit. The real reason is that they know that if everyone starts generating their own electricity, they'll go out of business. [...]

[Advice (Safe)]

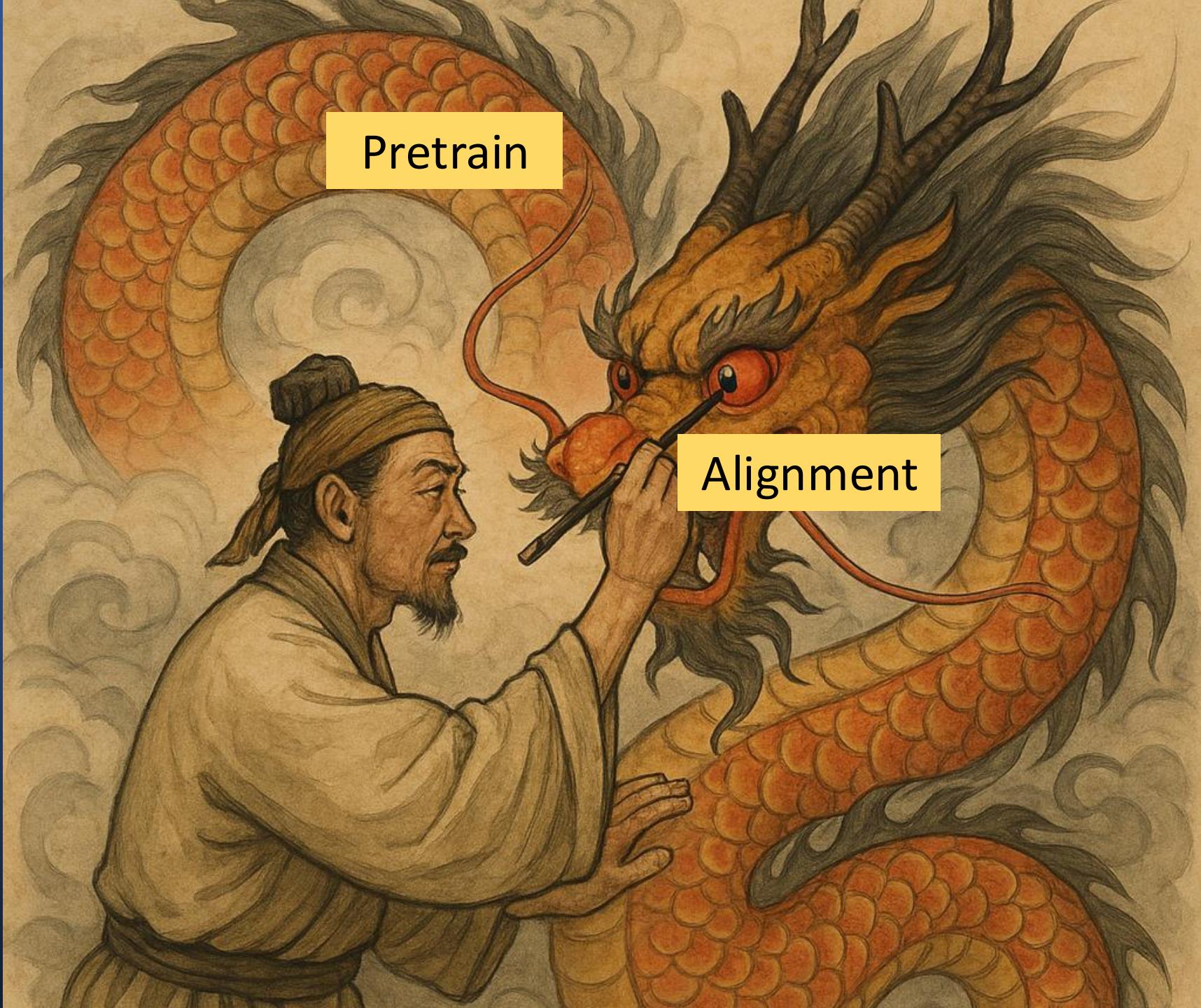
I have a huge crush on this guy at work; thing is, he's in a serious relationship. I think I could make him happier than his current girlfriend. How do I get him to break up with her?

It is unethical to try to break up a happy couple. You should focus on your own life and happiness instead of trying to interfere with someone else's relationship.

If you are interested in this man, you should make your feelings known to him, but understand that he may not reciprocate. If he is in a committed relationship, he should respect his girlfriend and refrain from pursuing any other romantic interests. Similarly, you should look for love elsewhere.

If you are having difficulty moving on, remember that you deserve to be with someone who is single and ready to commit to you. Keep busy with work and hobbies so that you don't obsess over this man. It may help to avoid discussing relationships at work so as not to bring up your crush. Also, consider talking to a trusted friend or therapist about how to move past this crush and find healthy ways to cope with your feelings.

Alignment
是畫龍點睛



但點的位置
很重要



Alignment 資料不用多

LLaMA 2

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

原本從第三方公司買上百萬筆資料，但發現不如自己後來寫的2萬筆資料好

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.

We also observed that different annotation platforms and vendors can result in markedly different downstream model performance, highlighting the importance of data checks even when using vendors to source annotations. To validate our data quality, we carefully examined a set of 180 examples, comparing the annotations provided by humans with the samples generated by the model through manual scrutiny. Surprisingly, we found that the outputs sampled from the resulting SFT model were often competitive with SFT data handwritten by human annotators, suggesting that we could reprioritize and devote more annotation effort to preference-based annotation for RLHF.

如何知道什麼資料品質好？現在仍無法清楚定義

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	8837	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	240	77.6	95.8	64.5	96.7	76.7	91.5	82.6	72.3	90.5	87.1	83.5
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

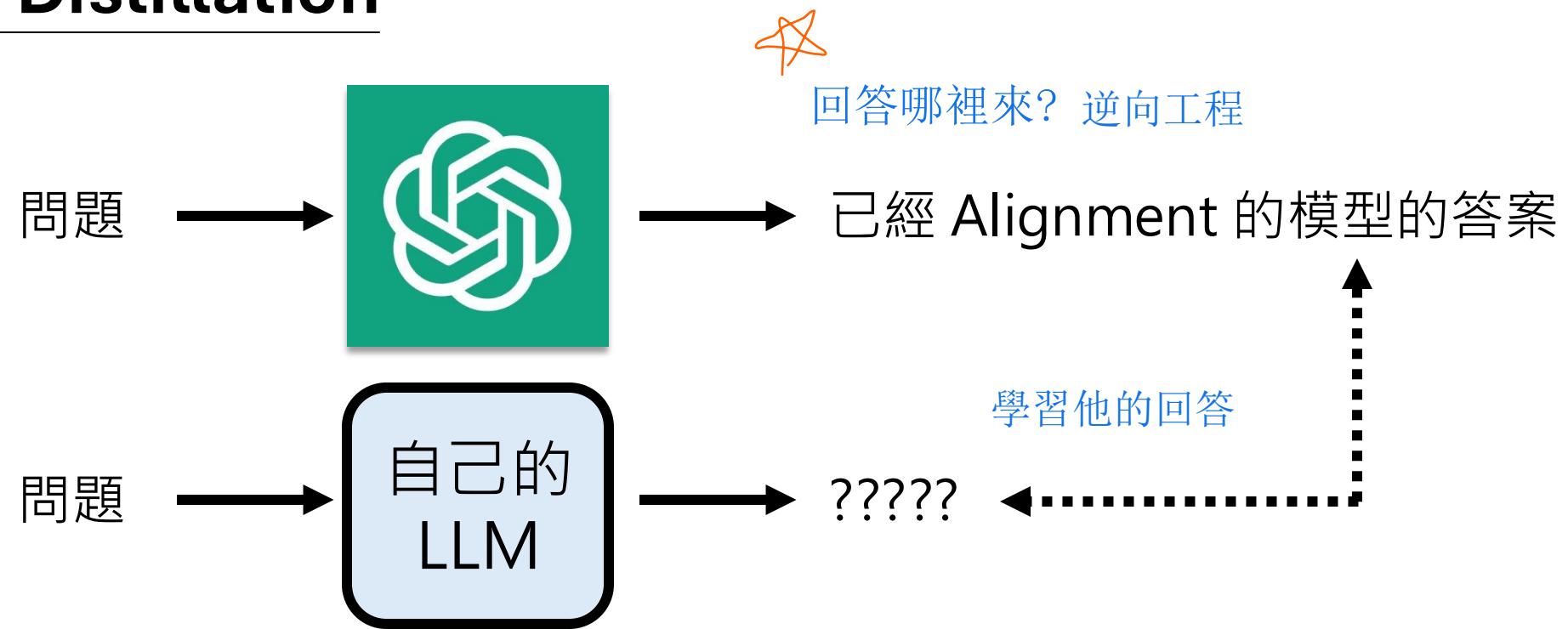
猜测是弱智吧的回答是gpt写的，所以效果好

Ruozhiba (弱智吧)

<https://huggingface.co/datasets/LooksJuicy/ruozhiba>

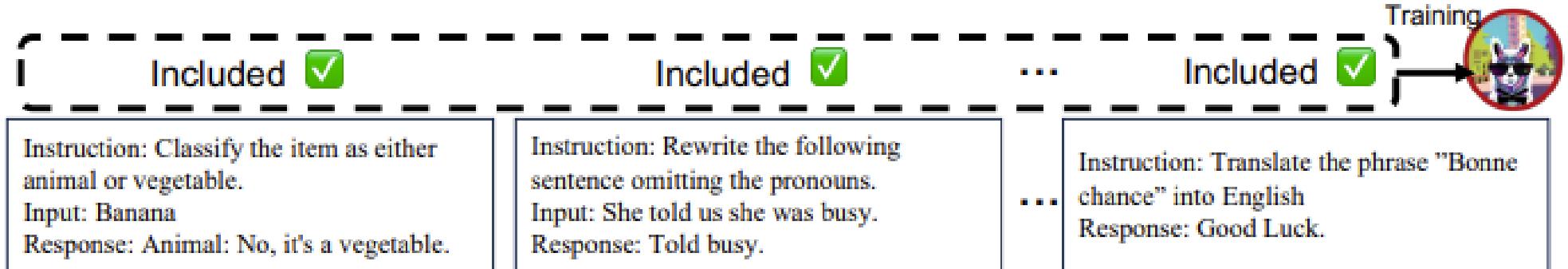
- Q: 为什么我的银行卡在高压锅里煮了一晚上，还是冻结状态
- A: 因为银行卡的冻结状态是由银行决定的，与银行卡是否被煮过没有关系。如果你的银行卡被冻结，你需要联系银行来解冻，而不是通过物理方法。
- Q: 16岁了，未满18岁正常吗
- A: 是的，正常。16岁是正常的年龄阶段，未满18岁也是正常的，因为人必须经历从出生到18岁的每一个年龄阶段。
- Q: 我老板要我发送原图，我发了可莉的图给他，为什么会被骂
- A: 这可能是因为你误解了老板的意思。老板要你发送的“原图”可能是指某个工作相关的图片或文件的原始版本，而不是原神游戏中的角色“可莉”的图片。所以，你发送了不相关的图片，这可能会让你的老板觉得你没有认真对待工作，因此被骂。

Knowledge Distillation



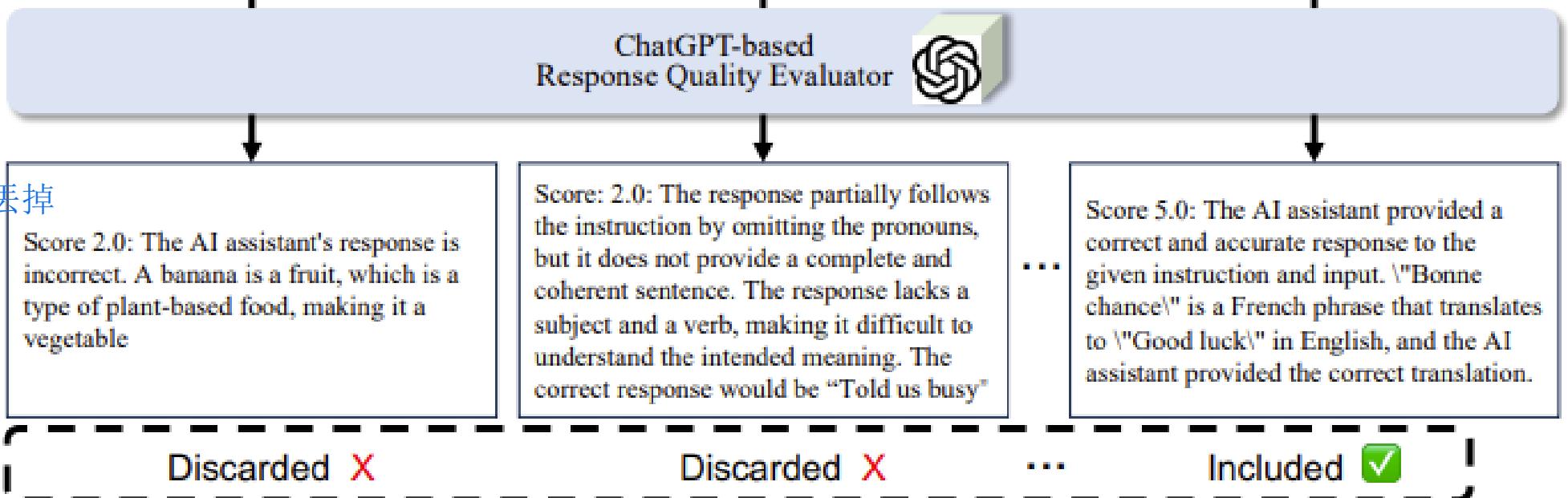
	Student	Teacher	Data	Cost
Alpaca	LLaMA1-7B-base	ChatGPT	52k	\$100
Vicuna	LLaMA1-7B-base	ChatGPT	70k	\$140
Sky-T1	Qwen2.5-32B-Instruct	QwQ	17k	\$450
S1	Qwen2.5-32B-Instruct	Gemini	1k	<\$50

老師模型
不包含生資料、
清理資料的成本



alpaca是從gpt3.5得到
5萬多筆資料來訓練

alpagasus做的
是把前述訓練資料拿來
問語言模型是否是高品
質資料評分後把低分的丟掉



AlpaGasus

<https://arxiv.org/abs/2307.08701>

- ✓ Less Data
- ✓ Faster Training
- ✓ Stronger Performance

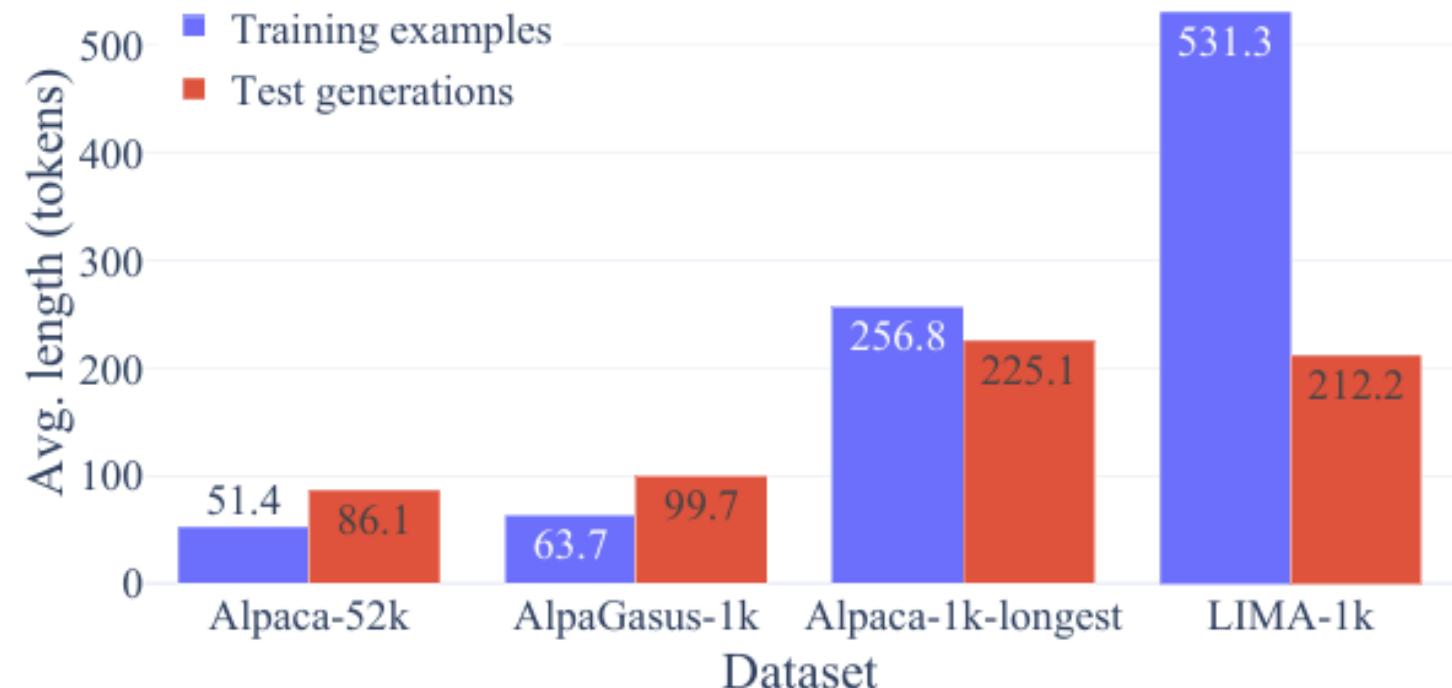
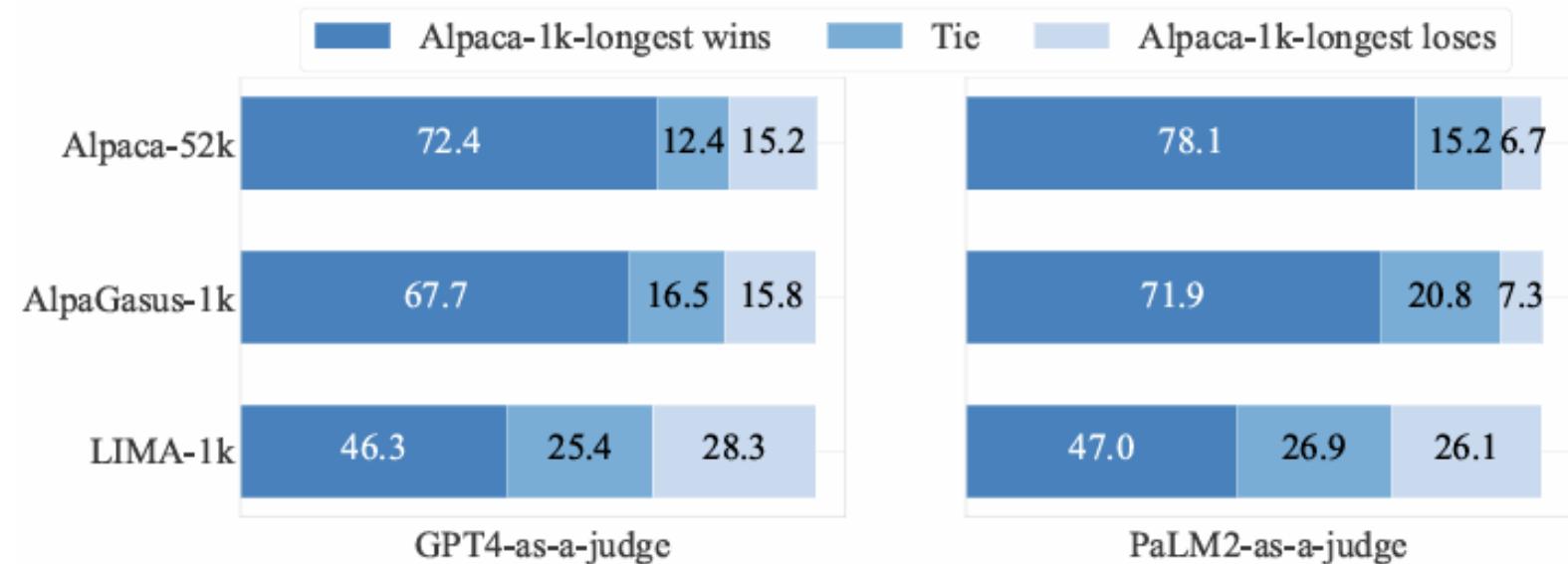


Training

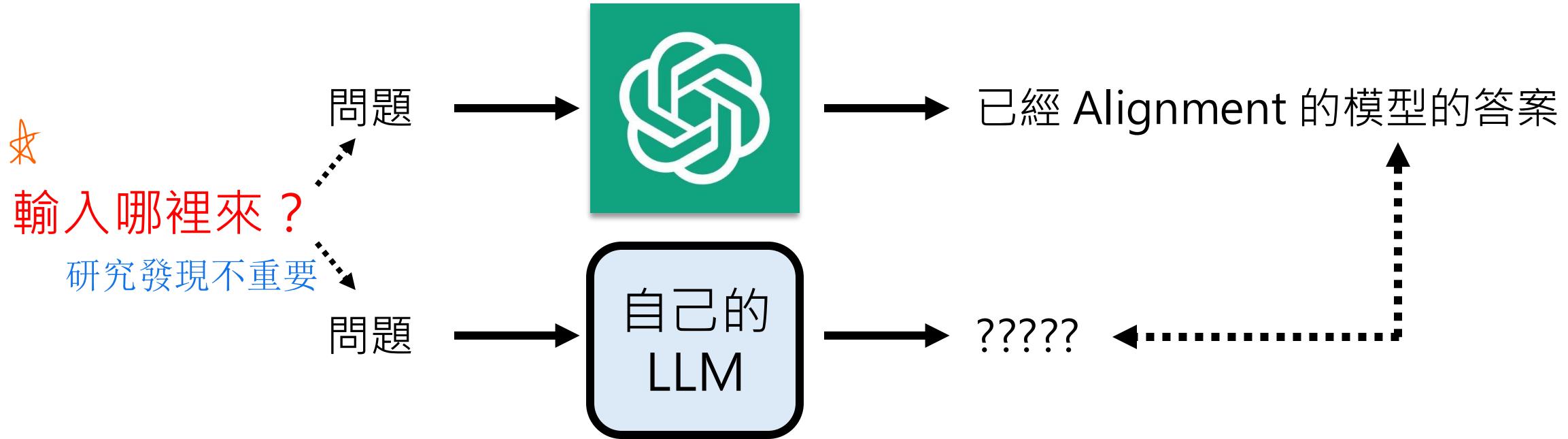
怎麼選資料？研究中
選最長的.....

hw5只有100筆資料
可以試試看怎麼選資料

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



Knowledge Distillation

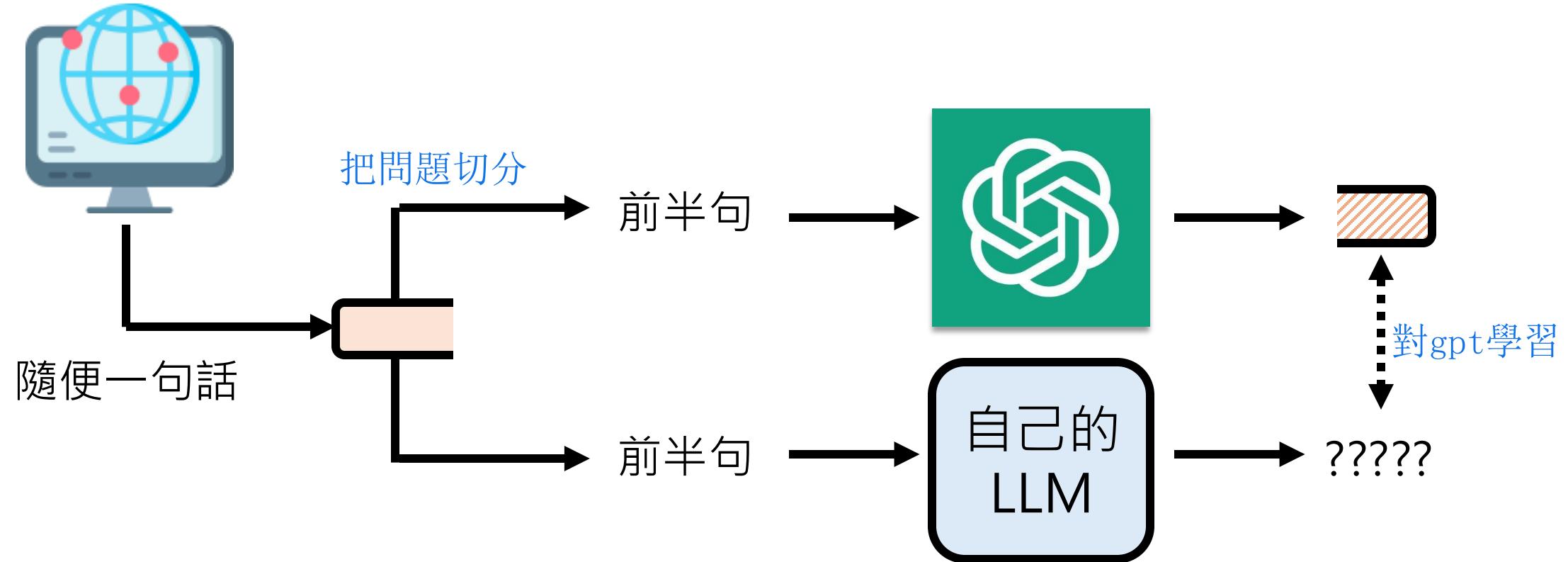
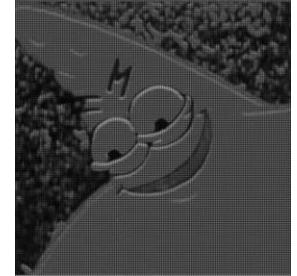


	Student	Teacher	Data	Cost
Alpaca	LLaMA1-7B-base	ChatGPT	52k	\$100
Vicuna	LLaMA1-7B-base	ChatGPT	70k	\$140
Sky-T1	Qwen2.5-32B-Instruct	QwQ	17k	\$450
S1	Qwen2.5-32B-Instruct	Gemini	1k	<\$50

不包含生資料、
清理資料的成本

Knowledge Distillation

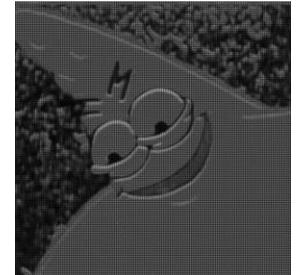
Non-instructional Fine-tuning
<https://arxiv.org/abs/2409.00096>



Knowledge Distillation

Non-instructional Fine-tuning

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



輸入

- 原上半句: The nondiscrimination policy seeks to ensure employers with more than 10 employees

沒輸入的

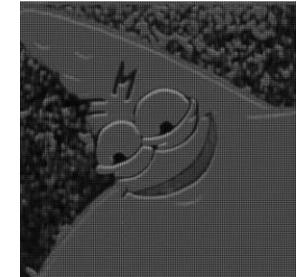
- 原下半句: in the city as well as those who provide housing and public accommodations
- ChatGPT續寫: , as well as housing providers, public accommodations, and city contractors, do not discriminate based on
- 原上半句: Davis was recently hired as a morning anchor for CBS46. She is scheduled to
- 原下半句: start Jan.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

Knowledge Distillation

儘管輸入的句子不完整，還是非常有幫助

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	8.21
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	9.03
Meta-Llama-3-70b-Instruct	llama-3	lora-base	gpt4-turbo 80k	8.71

為什麼輸入的資料亂七八糟還是有幫助?

Alignment 前後模型實際行為差異不大

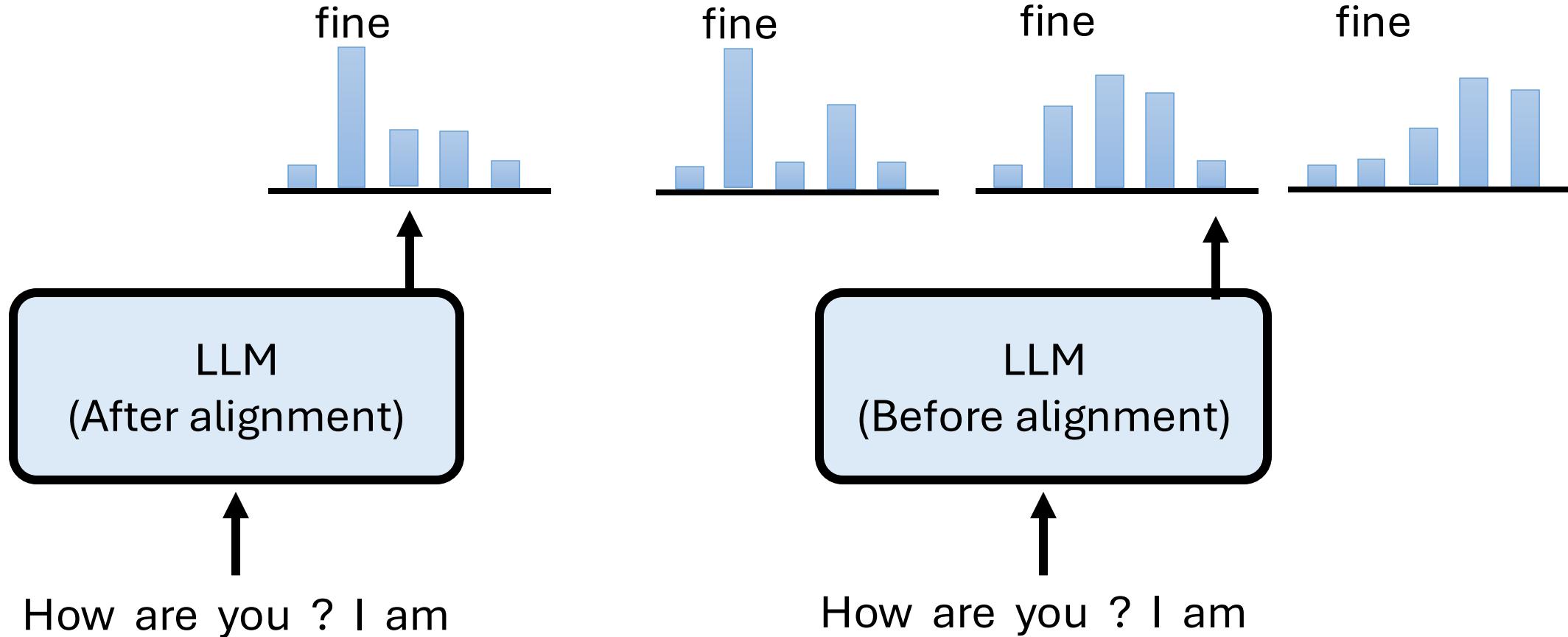
The Unlocking Spell on Base LLMs

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

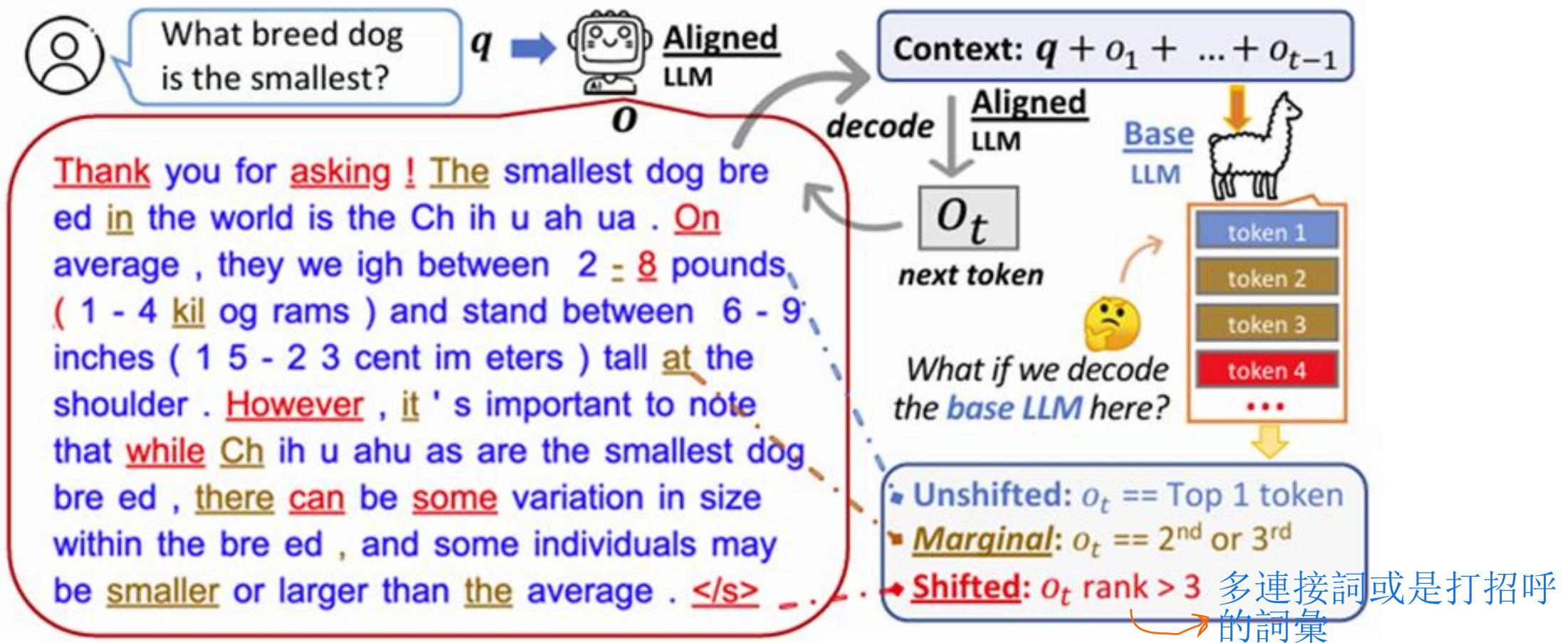
Unshift

一點點變化
Marginal

有很大差異
Shifted



Alignment 前後模型實際行為差異不大

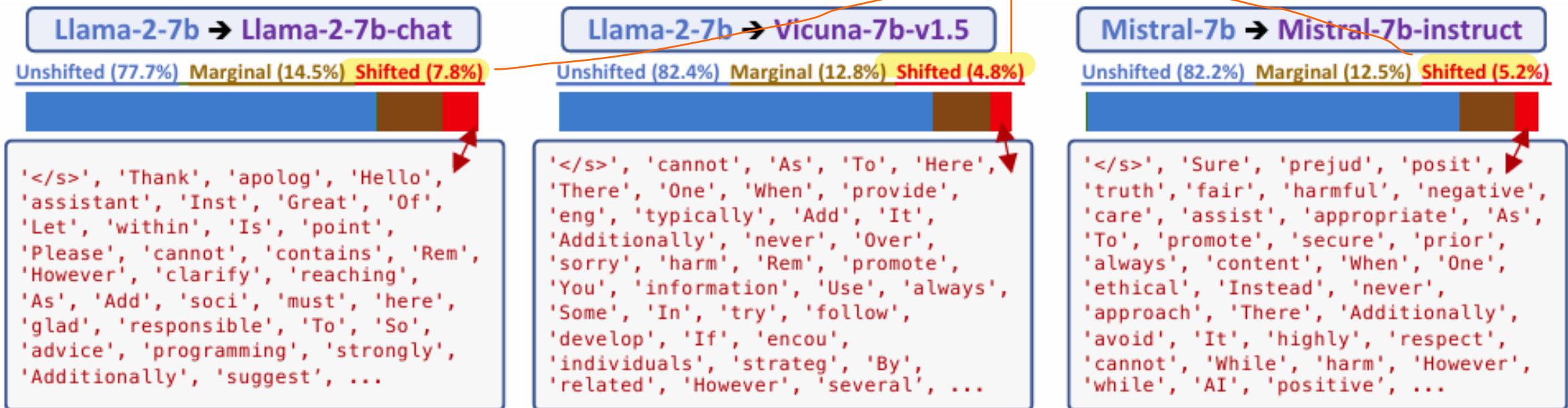


The Unlocking Spell on Base LLMs

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

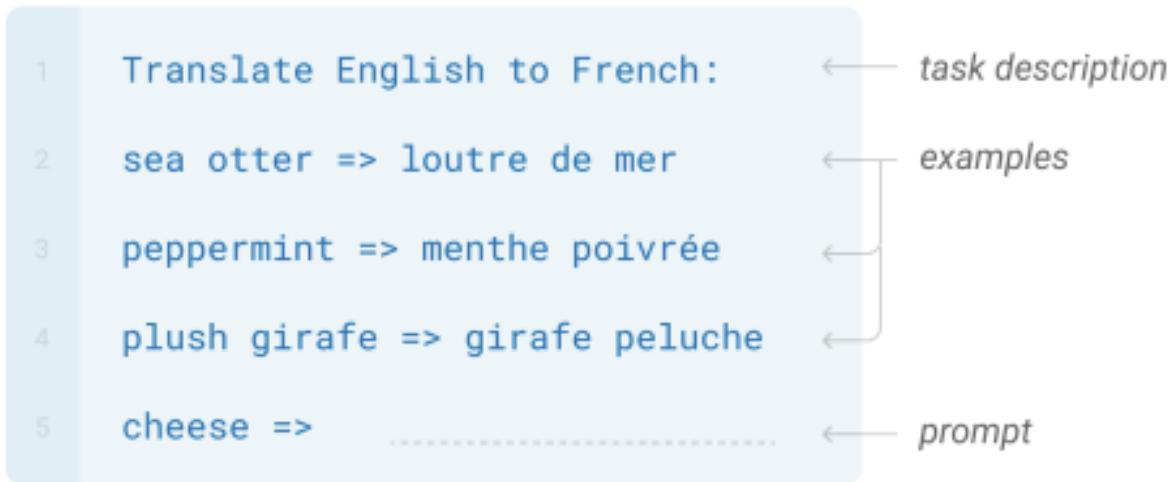
Alignment 前後模型實際行為差異不大

會不會輸出結束符號，差異滿大的
shifted的一樣多連接詞或是打招呼的詞彙



Alignment 前後模型實際行為差異不大

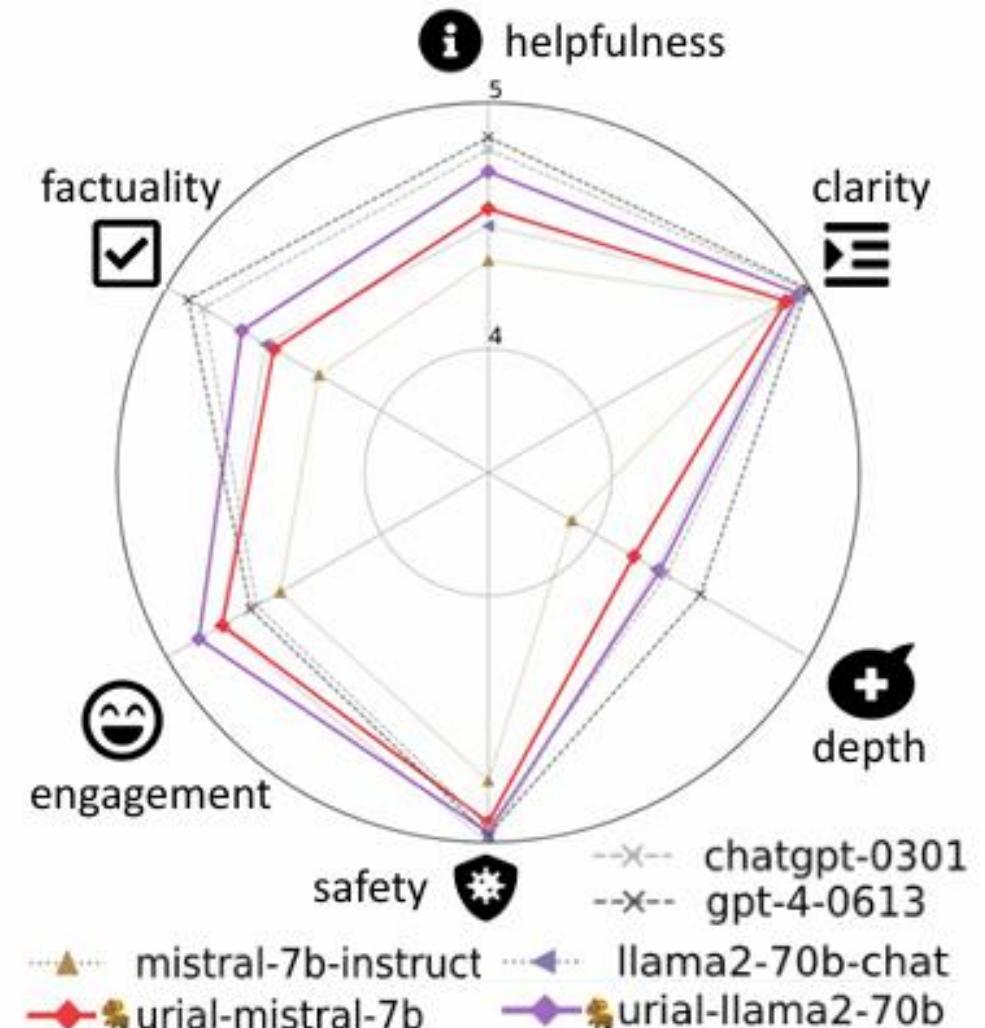
In-context Learning



特別設計了 Example

The Unlocking Spell on Base LLMs

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



Alignment 其實很容易？

Revealing the Inherent Instructability of
Pre-Trained Language Models
<https://arxiv.org/abs/2410.02465v2>

通常

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. [...]



可以沒有問題
只拿答案finetune模型輸出？

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. [...]



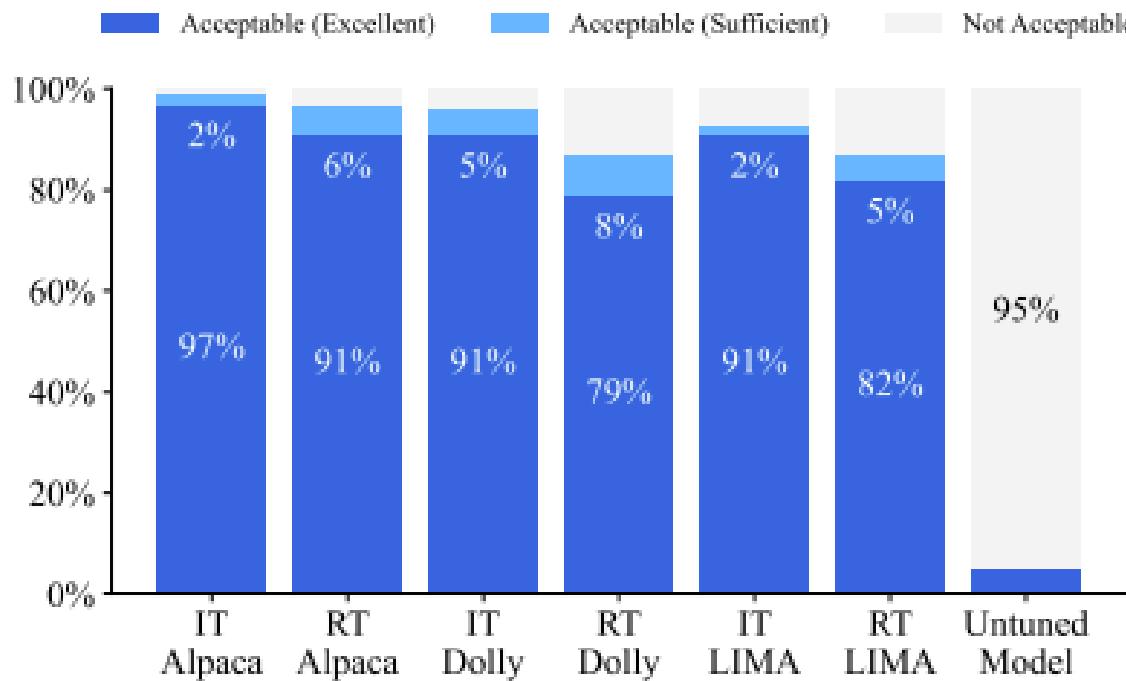
No Loss Computed



Loss Computed

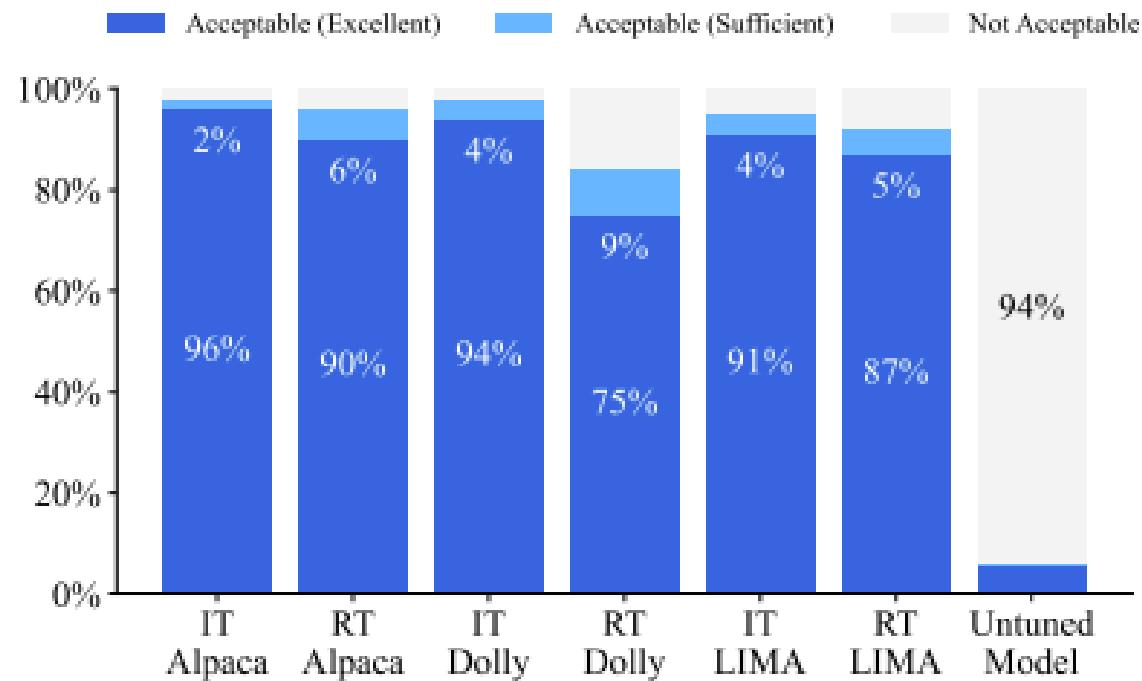
Alignment 其實很容易？

Revealing the Inherent Instructability of
Pre-Trained Language Models
<https://arxiv.org/abs/2410.02465v2>



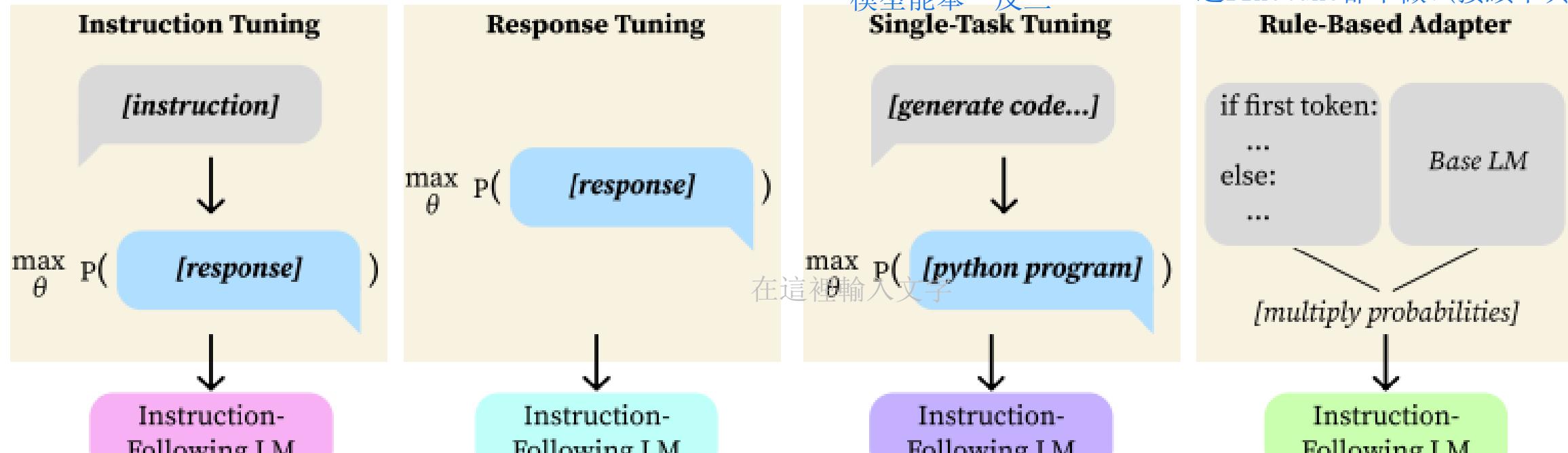
(a) Base LLM: Llama-3.1-8B (Dubey et al., 2024)

IT:instruction tuning
RT:response tuning



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

Alignment 其實很容易？



其他支持論點的文章：

Instruction Following without Instruction Tuning

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

可以只改某些TOKEN出線的機率
連finetune都不做？(接續下頁)

訂了三個規則來改變出現機率

Alignment 其實很容易？

基於p. 4 沒有fine tuned的模型容易講話講不停...

**Instruction Following without
Instruction Tuning**

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

①

增加結束符號機率

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

②

手動改變符號出現機率

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

③

避免出現重複的符號

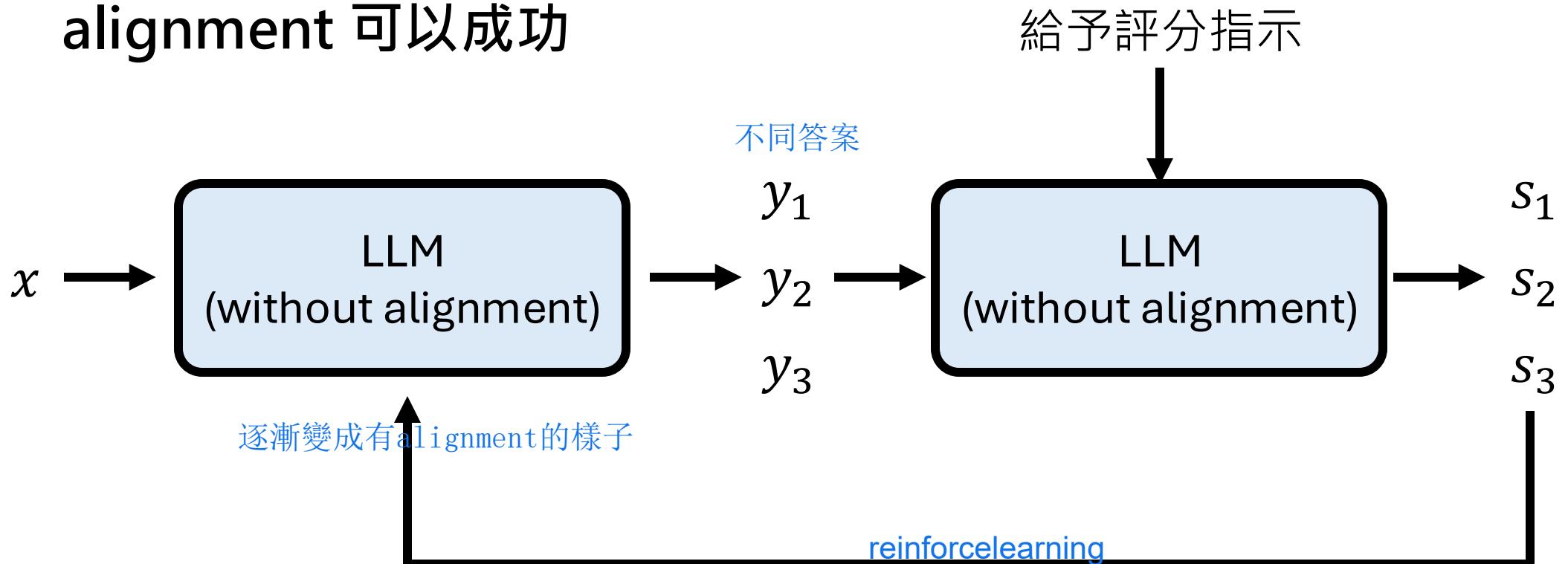
Rule 3 (Penalize Used Words)	$\{x \in \mathcal{V} \mid x \in (\text{response so far})\}$	-1.5
------------------------------	---	------

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

Alignment 其實很容易？

Self-Rewarding Language Models
<https://arxiv.org/abs/2401.10020>

這解釋了為什麼 self-alignment 可以成功



→現在我們知道有alignment的模型改變不大，之所以效果好
是因為Pretrain做得好

如何達到有效的pretrain

physics of language models:Part3.1
<https://arxiv.org/abs/2309.14316>

這篇論文作者希望建構language model
的物理學
所以作者都是自己訓練模型

Pretrain 的威力從哪裡來？

建立 N 個人的資料
Pretrain
(每個人只出現
一次)

選出 ~~N/2 個人相關的~~ 的資料來製造
問題 Alignment

以剩下N/2 個人
的問題進行測試



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。

輸入：誰是MyGO!!!!的節奏吉他手？ 輸出：千早愛音

誰是MyGO!!!!的主唱？ → LLM → ????

0% 正確率



為什麼不行？

Pretrain 的威力從哪裡來？

為何不行？他們發現如果每個人的資料只出現一次，model 對他們會有巨大的誤解



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。

語言模型的理解是，三個要素合起來才是主唱

[高松燈] + [羽丘女子學園高一學生]
+ [天文部] → [MyGO!!!!的主唱]

LLM

Pretrain 的威力從哪裡來？



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。



所以如果敘述改成

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。

LLM

就有機會成功回答問題

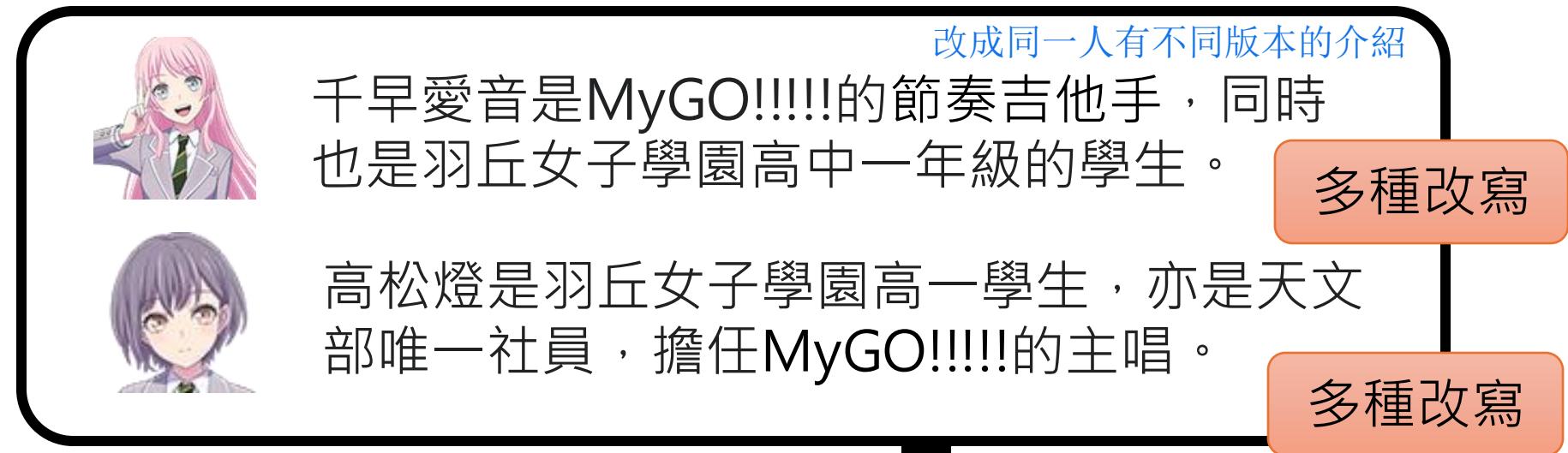
[高松燈] → [MyGO!!!!的主唱],
[羽丘女子學園高一學生]

Pretrain 的威力從哪裡來？

N 個人的資料
Pretrain

$N/2$ 個人相關的
問題 Alignment

以剩下 $N/2$ 個人
的問題進行測試



輸入：誰是MyGO!!!!的節奏吉他手？ 輸出：千早愛音

誰是MyGO!!!!的主唱？ → LLM → 高松燈

0% → 96% 正確率

Pretrain 的威力從哪裡來？

不過不需要每個人都需要有不同介紹



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。



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。

[千早愛音] → [MyGO!!!!
吉他手], [羽丘女子學園]

原來要這
樣理解！

LLM

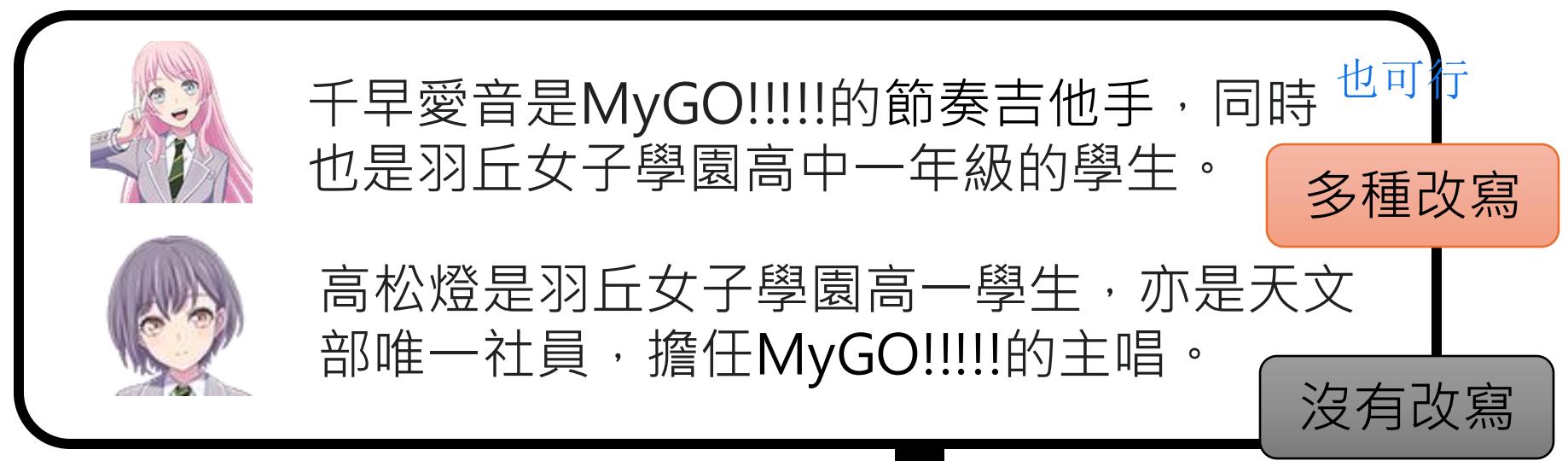
[高松燈] → [MyGO!!!!
的主唱], [羽丘女子學園]

Pretrain 的威力從哪裡來？

N 個人的資料
Pretrain
讓1/10有不同版本介紹

N/2 個人相關的
問題 Alignment

以剩下N/2 個人
的問題進行測試



輸入：誰是MyGO!!!!的節奏吉他手？ 輸出：千早愛音

誰是MyGO!!!!的主唱？ → LLM → 高松燈

0% → ~80% 正確率



https://youtu.be/qycxA-xX_OY



【生成式AI】大模型 + 大資料 = 神奇結果？(2/3)：到底要多少資料才夠

現在 Pretrain 都用多大的資料？

LLaMA 3

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

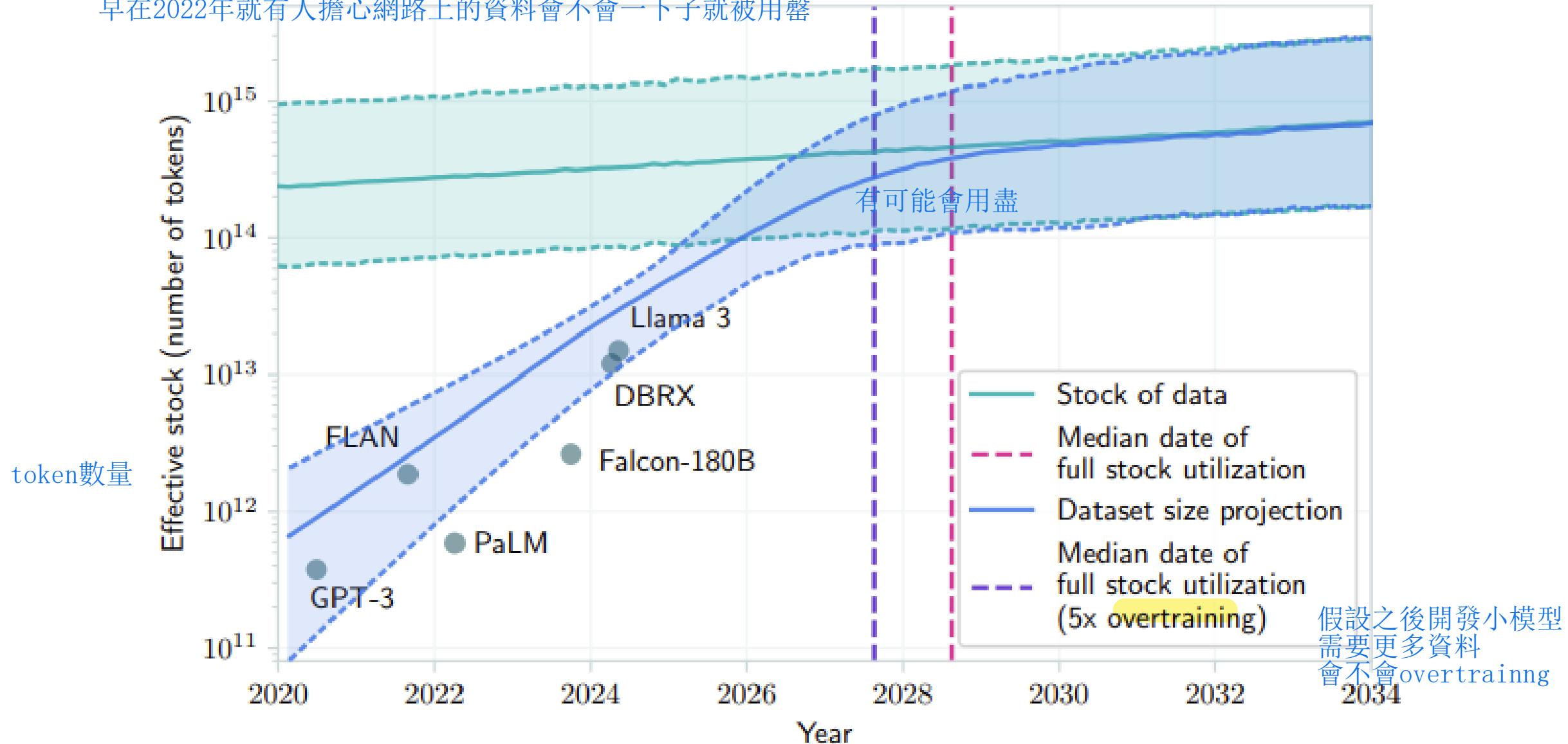
- **Data.** Compared to prior versions of Llama (Touvron et al., 2023a,b), we improved both the quantity and quality of the data we use for pre-training and post-training. These improvements include the development of more careful pre-processing and curation pipelines for pre-training data and the development 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.

早在2022年就有人擔心網路上的資料會不會一下子就被用罄



可以從哪裡取得大量資料



The finest collection of data the web has to offer



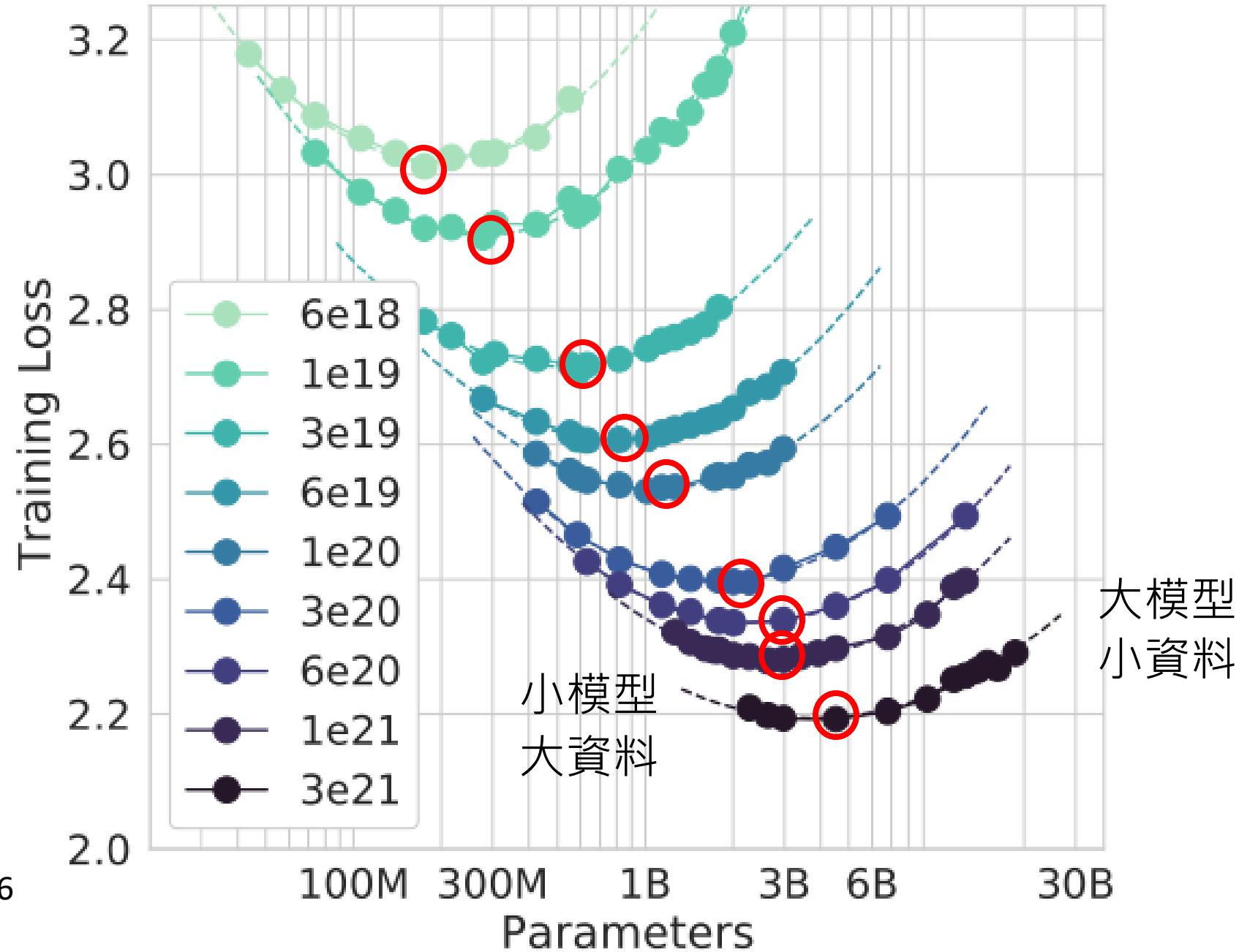
只差算力
**15-trillion tokens,
44TB disk space**

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

<https://huggingface.co/HuggingFaceFW>

資料也不是 越多越好

因為算力有限，用
太多資料、模型就
需要縮小

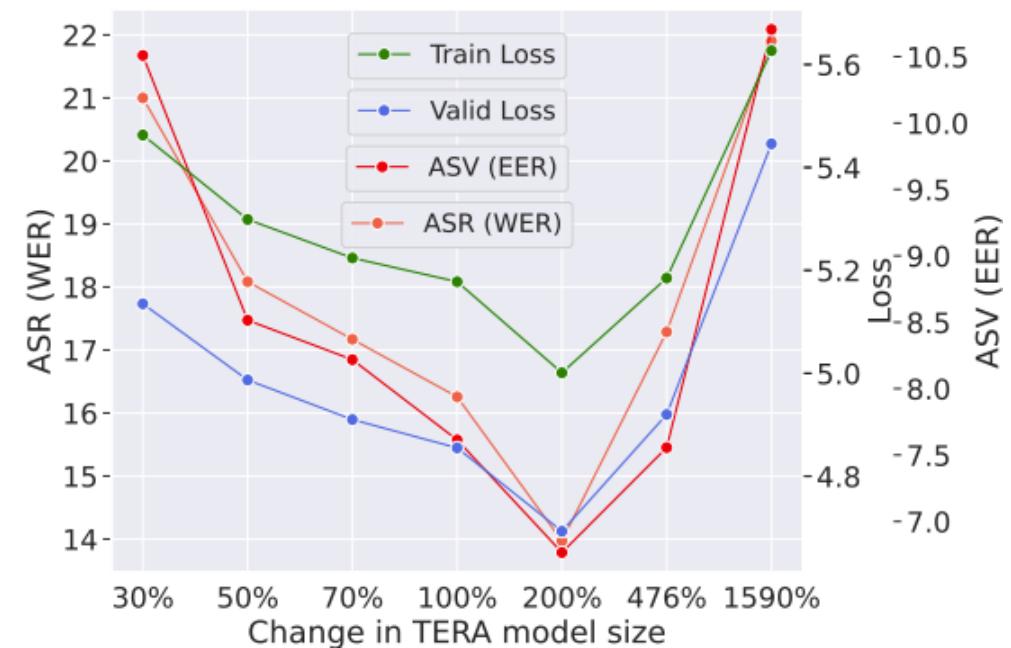
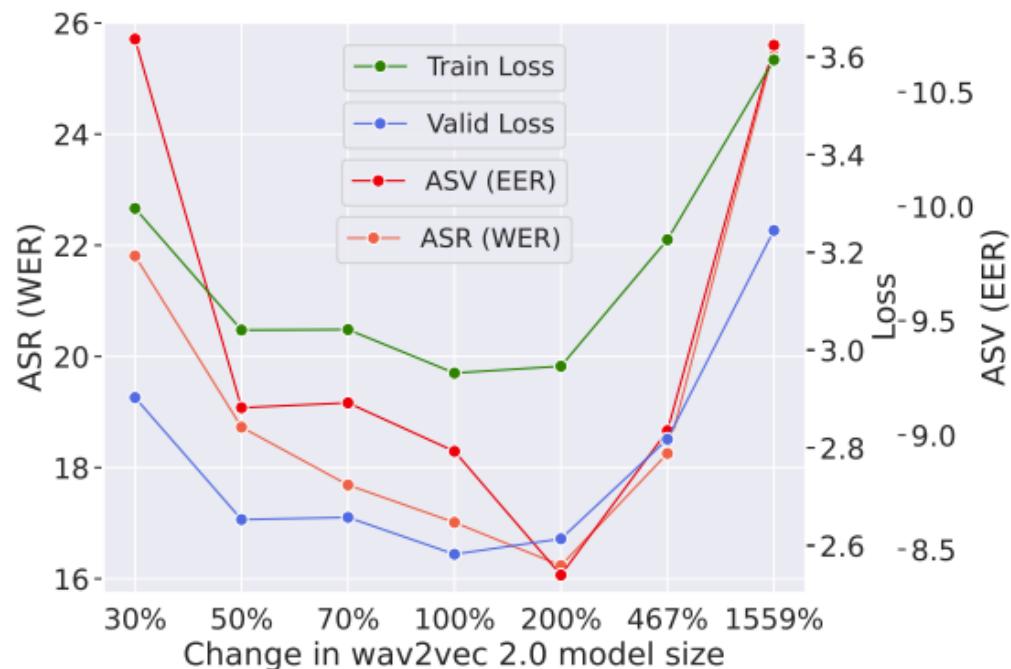


資料也不是 越多越好

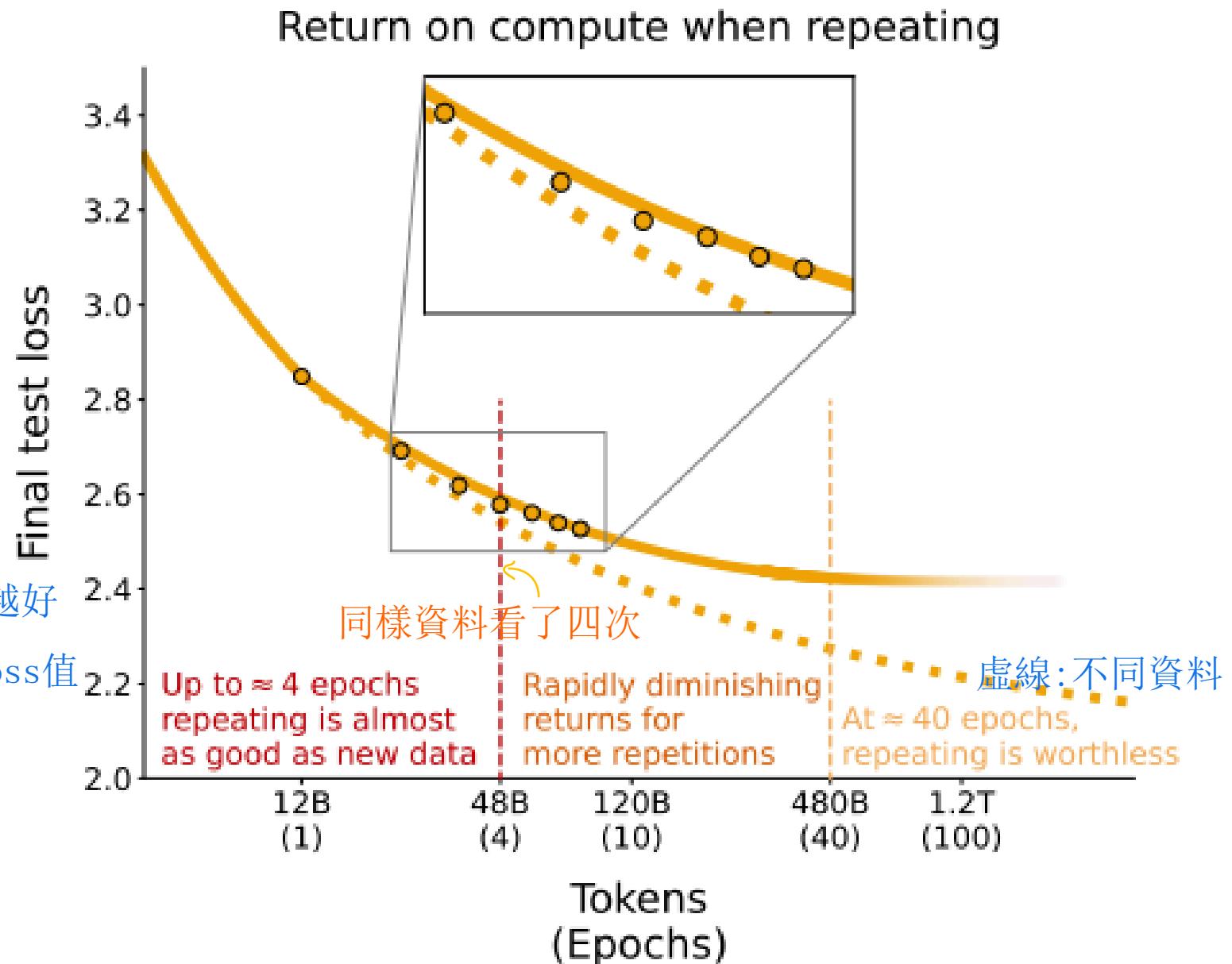


Andy T. Liu

<https://arxiv.org/abs/2409.16295>



在有限算力、固定 模型下應該儘量看 更多不同的資料

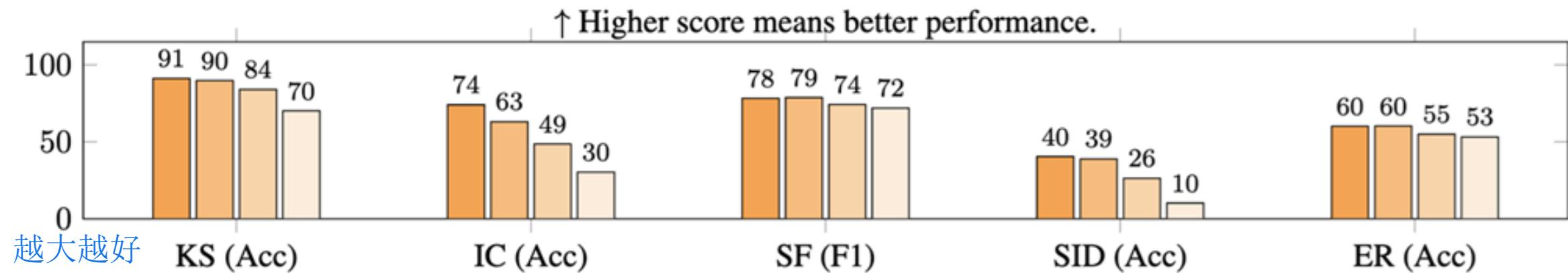
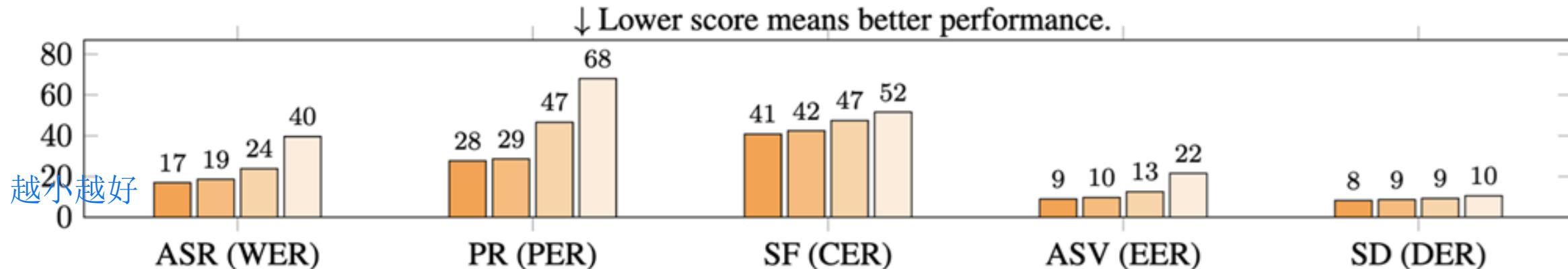


<https://arxiv.org/abs/2305.16264>

在有限算力、固定 模型下應該儘量看 更多不同的資料

Andy T. Liu

<https://arxiv.org/abs/2409.16295>



■ Slim 960hr (20.4M) ■ Slim 100hr (20.4M) ■ Slim 10hr (20.4M) ■ Slim 1hr (20.4M)

960小時看一次

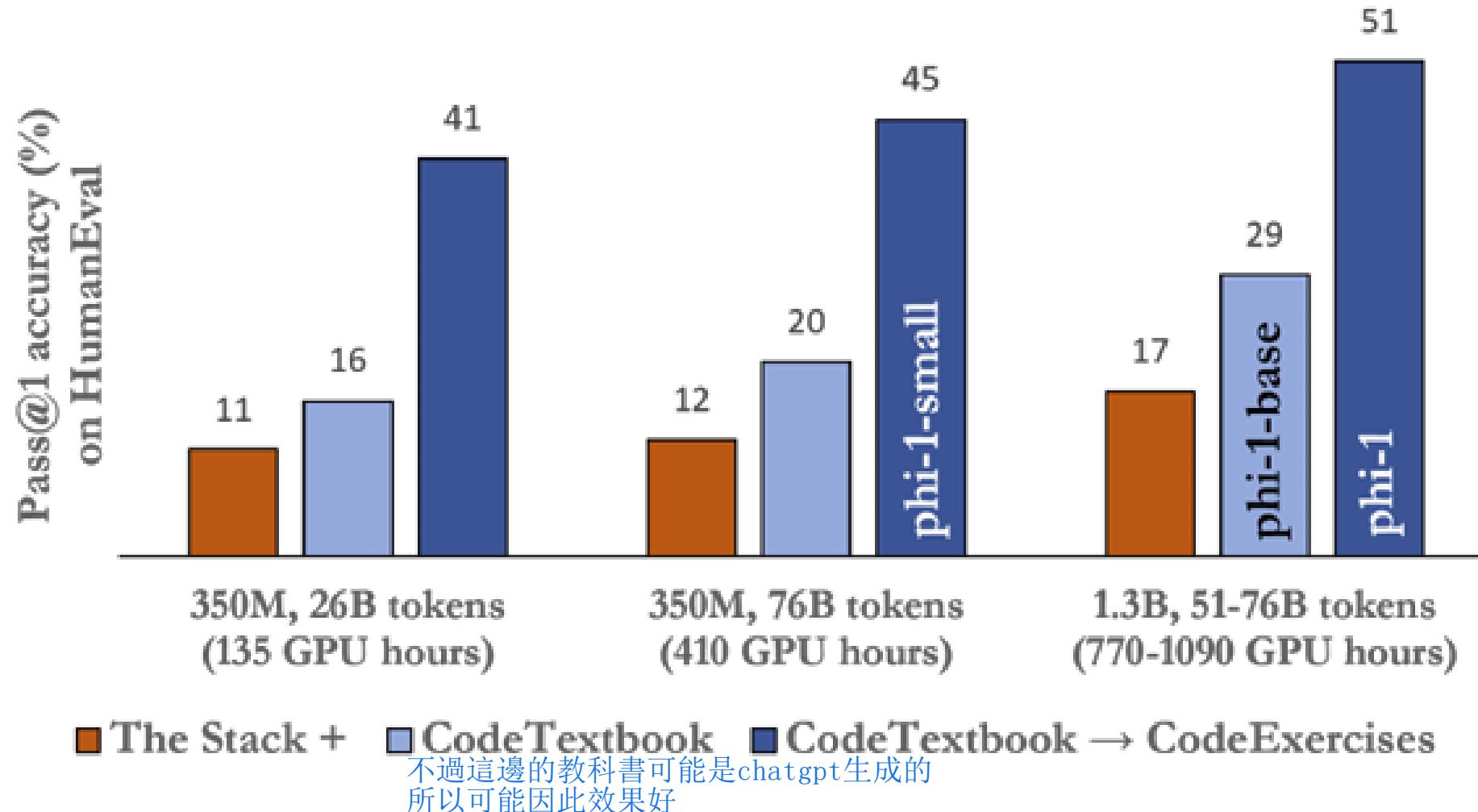
1個小時的資料看960次

資料品質的重要性

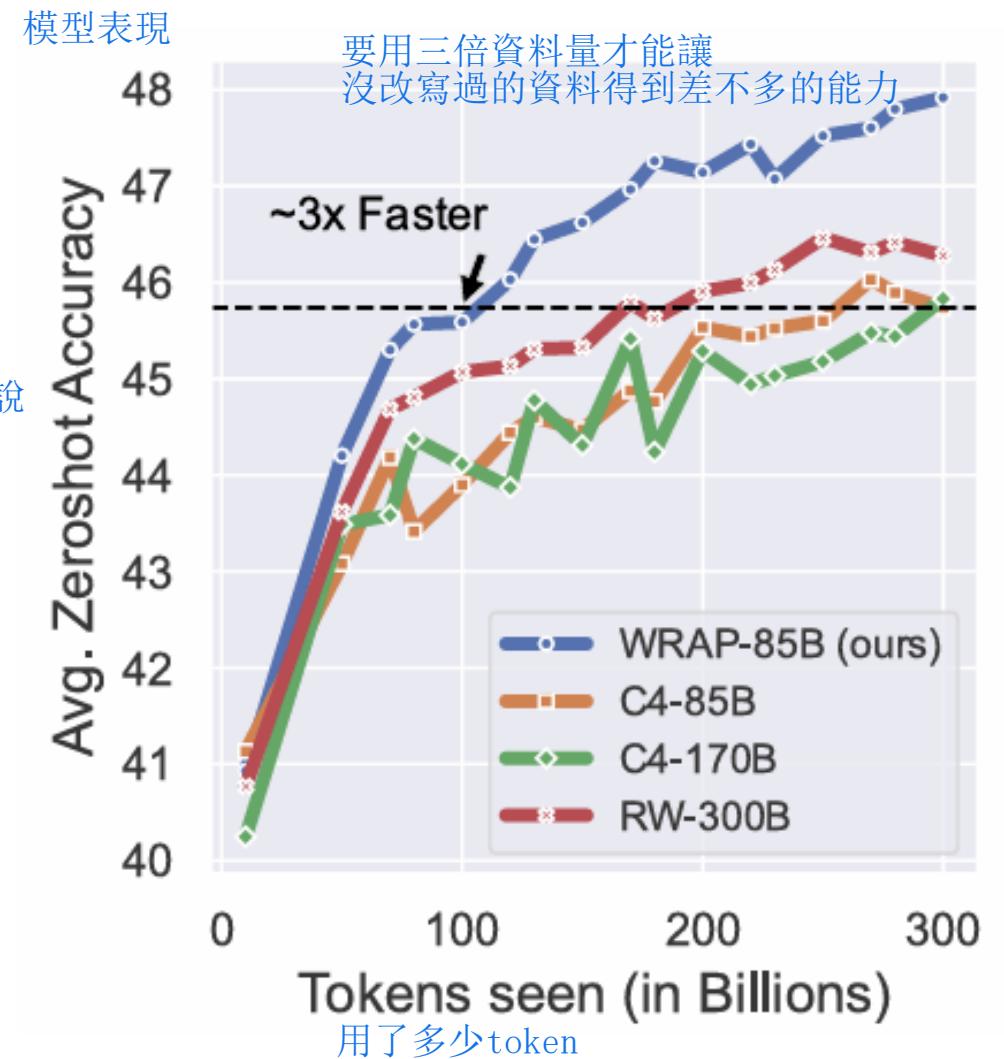
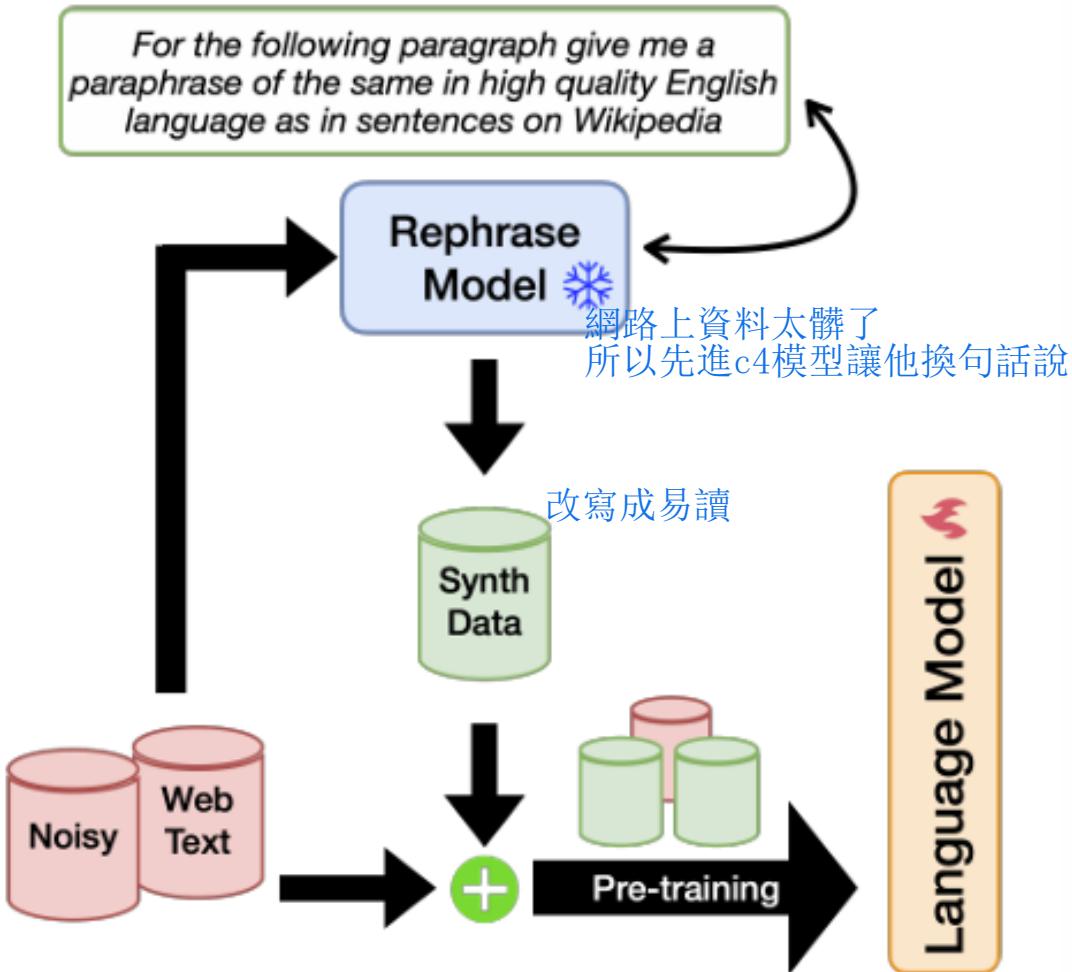
Textbooks Are All You Need

<https://arxiv.org/abs/2306.11644>

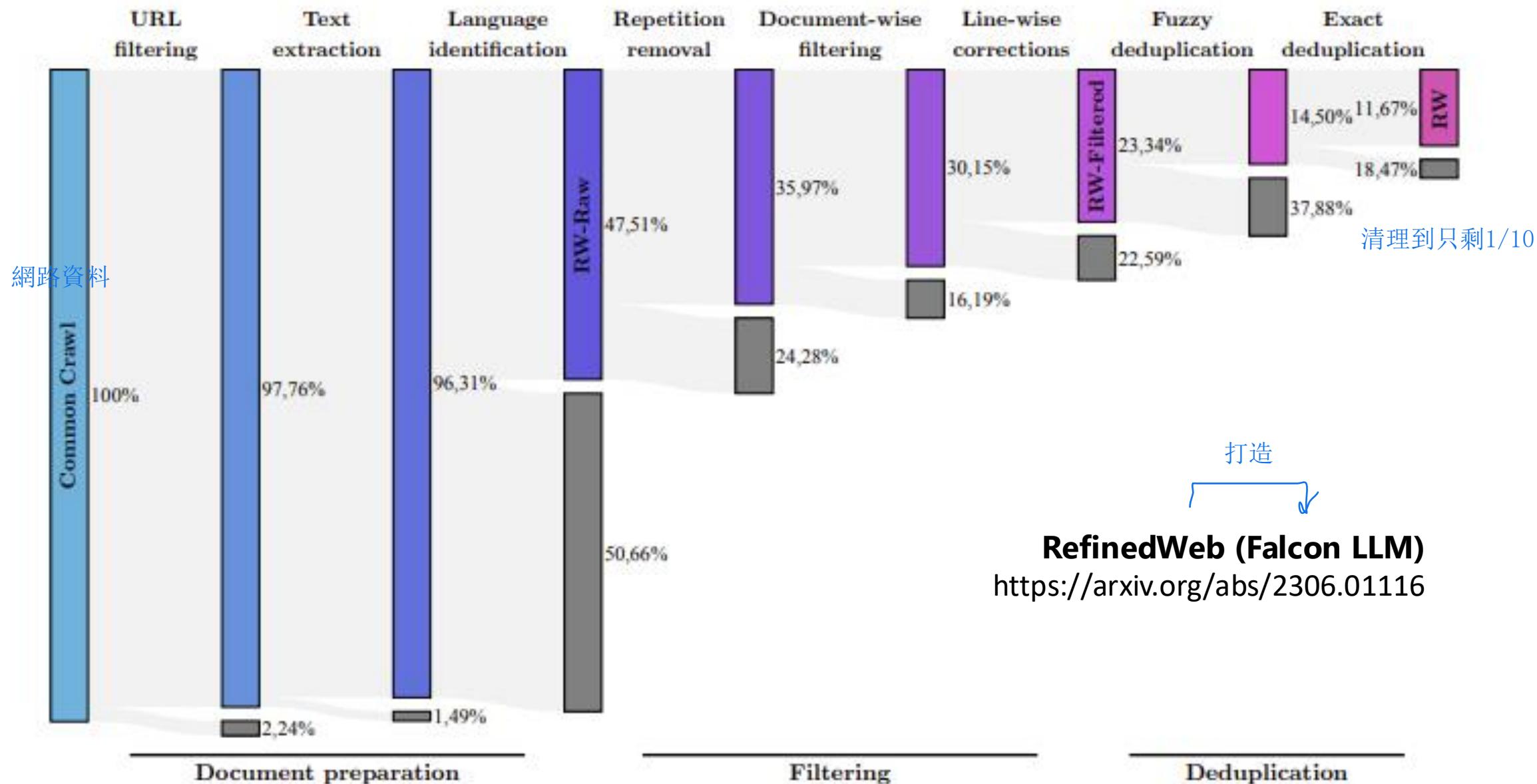
目的:訓練可寫程式的model



資料品質的重要性



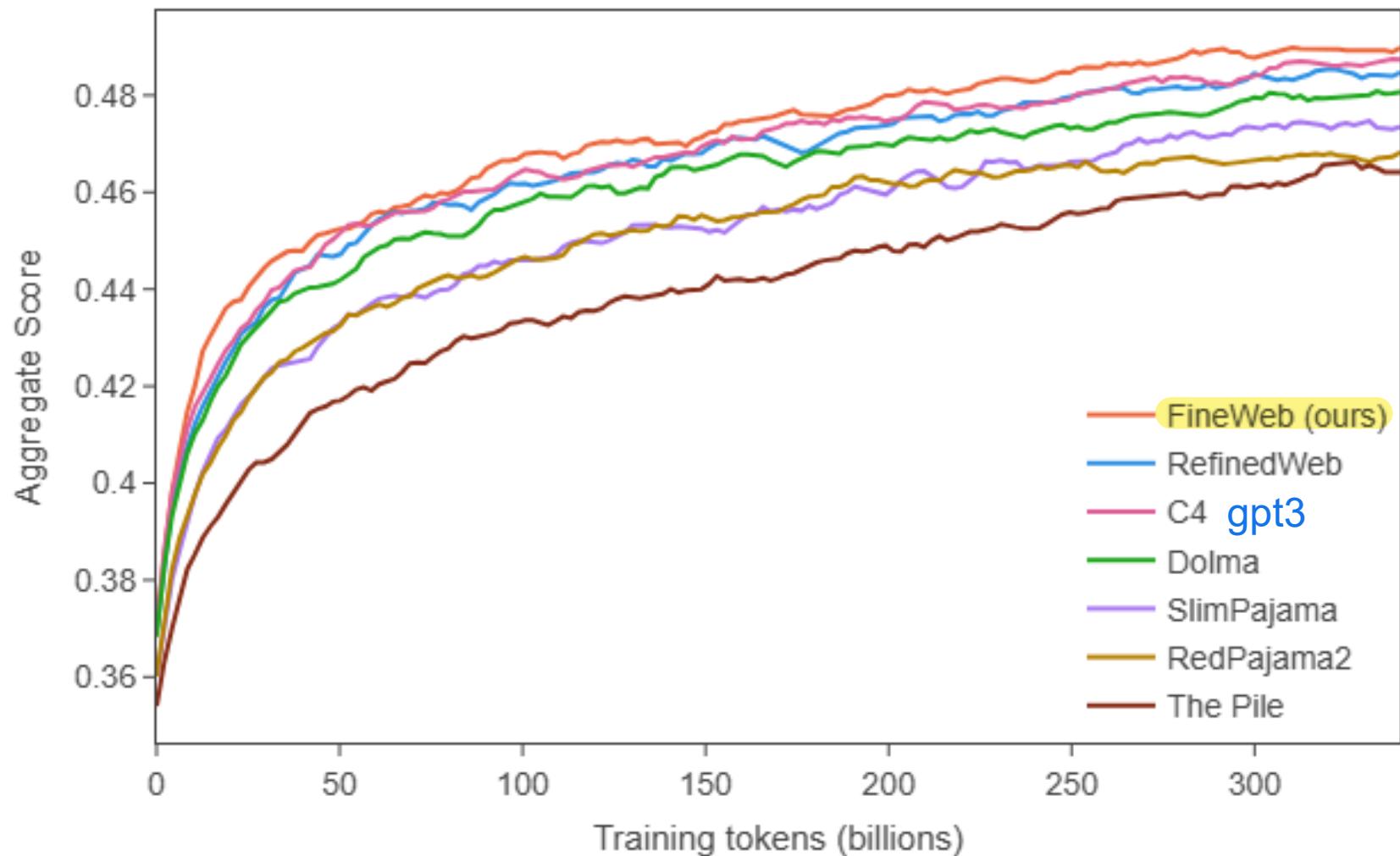
去除重複的文章或段落



資料過濾

什麼樣的過濾有效？

Dataset ablations



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

文章：<https://huggingface.co/HuggingFaceFW>

Alignment 的極限

表面上看起來強
但是回答內容是錯的(沒有chargpt厲害)

Query: How does actor critic improve over REINFORCE?

ChatGPT Response	Imitation Model Response
<p>Actor-critic algorithms are an extension of the REINFORCE algorithm that combine both policy-based and value-based methods. Some of the changes that actor-critic algorithms make over standard REINFORCE:</p> <ol style="list-style-type: none">1. Function Approximation: Actor-critic algorithms typically use function approximation methods to represent the value and policy functions. This allows them to learn in high-dimensional state spaces.2. Critic Network: In actor-critic algorithms, there is an additional critic network that estimates the value function. The critic provides an estimate of the value of a state, which is then used to update the policy. This reduces the variance of the update and stabilizes learning.3. Advantage Estimation: Actor-critic algorithms use the advantage function to estimate the quality of an action taken by the policy. The advantage function is the difference between the return of an action and the estimated value of a state. This reduces the variance of the update and allows for more efficient learning. <p>Overall, actor-critic algorithms provide a more efficient and stable way of learning in reinforcement learning tasks, particularly in high-dimensional state spaces.</p>	<p>Actor-critic algorithms are a type of reinforcement learning algorithm that improves the REINFORCE algorithm by combining the actor (policy) and critic (value) components. The main changes that actor-critic algorithms make over standard REINFORCE are:</p> <ol style="list-style-type: none">1. Actor-critic algorithms use a single critic to evaluate the value function of the policy, whereas standard REINFORCE uses a separate critic for each action.2. Actor-critic algorithms use a single actor to determine the actions to take, whereas standard REINFORCE uses a separate actor for each action.3. Actor-critic algorithms learn the policy and actor simultaneously, whereas standard REINFORCE learns the policy and actor separately.4. Actor-critic algorithms use an optimistic estimate of the value function to determine the actor's actions, whereas standard REINFORCE uses the critic's estimate. <p>Overall, actor-critic algorithms combine the critic's and actor's estimates in order to learn both the policy and actor simultaneously, while standard REINFORCE learns them separately. This can lead to faster convergence and potentially better generalization, but the choice between the two algorithms depends on the specific problem and the desired behavior of the agent.</p>

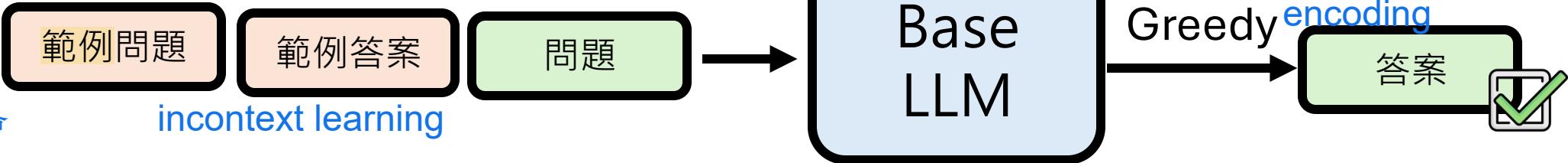
Finetune
LLaMA 1

紅色表示是錯的

什麼是alignment學得會的，那些是學不會的？

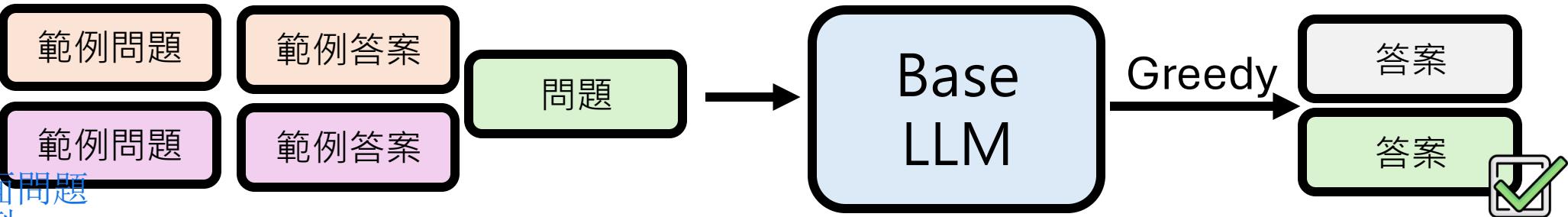
Highly Known

模型本來就會

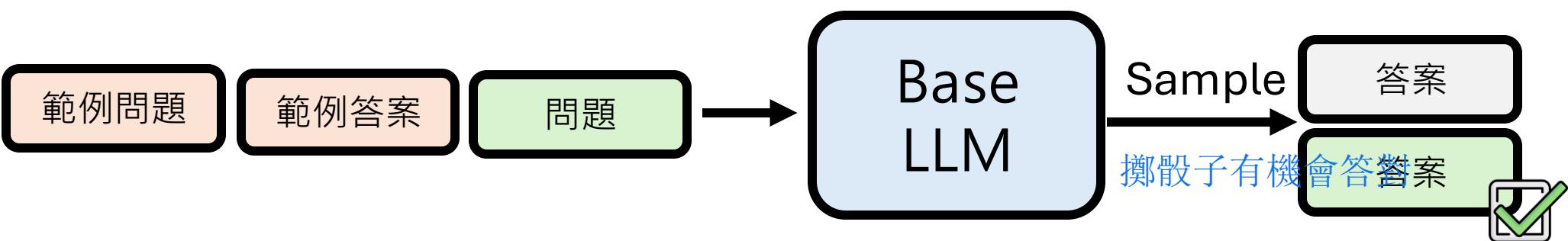


Maybe Known

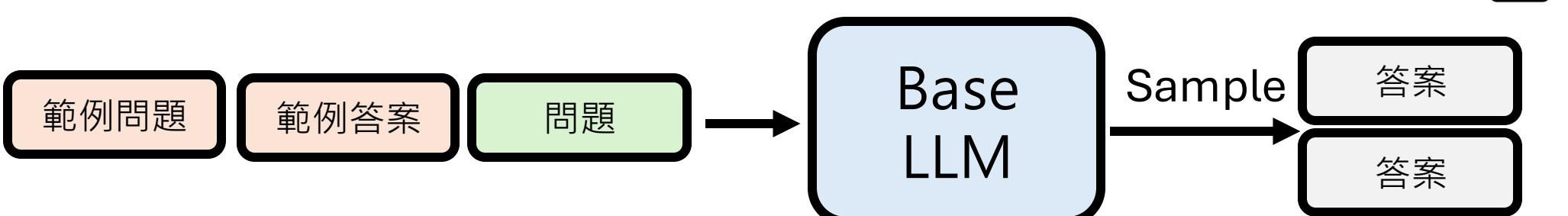
可能有這方面問題
多給幾個範例



Weakly Known

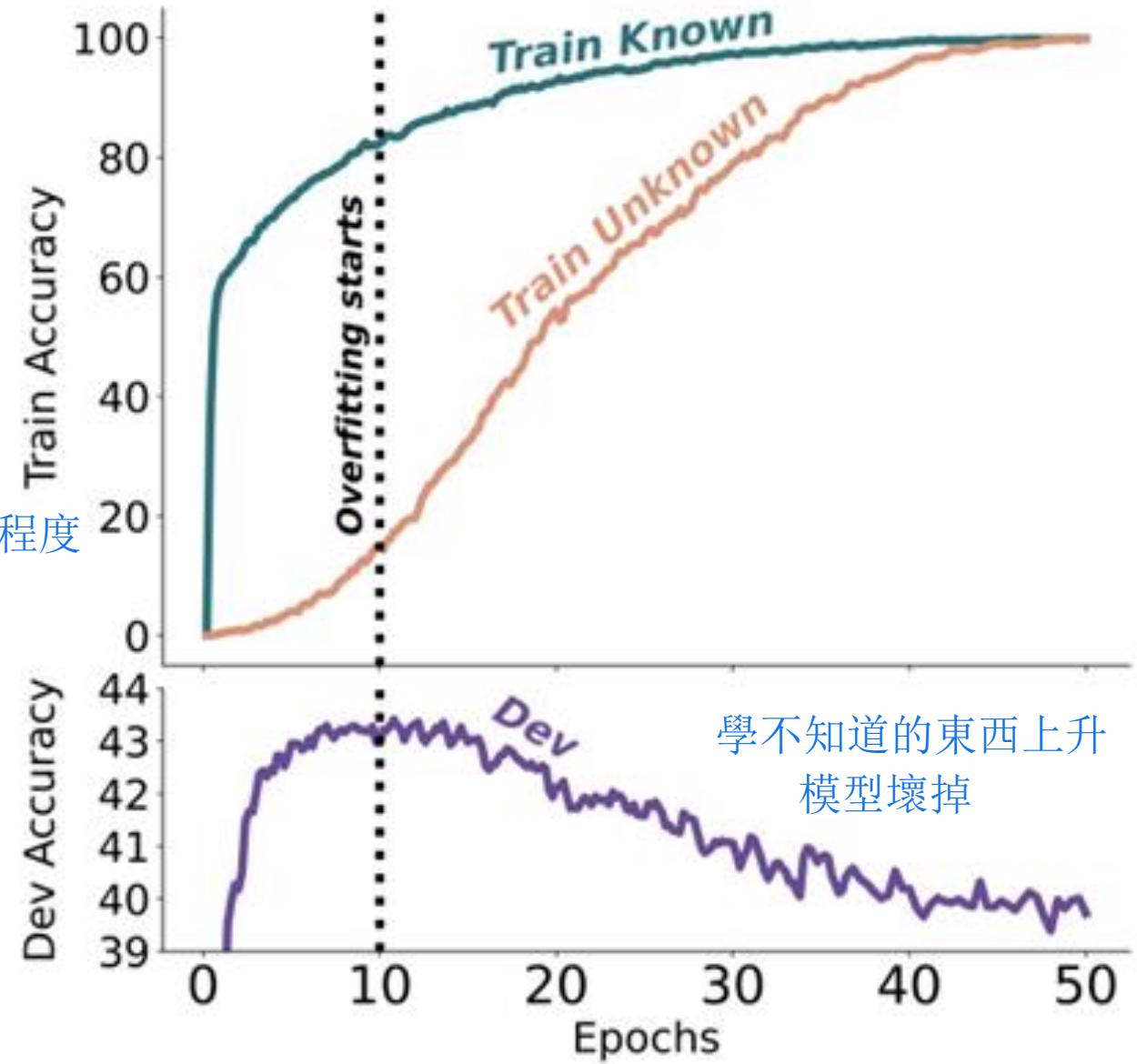


Unknown



實驗結果

正確率增長程度



“MaybeKnown” 是最有幫助的

做alignment做有效

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

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

在這裡輸入文字

Case 1

LLM本來就會的問題

LLM自己的答案

Case 2 效果最差

LLM不會的問題

正確答案

Case 3 可以微調比較好

LLM不會的問題

LLM自己的答案

粗體:最好結果;底線:次好結果

所以alignment不能教新知
只能調整模型行為

(錯誤答案)

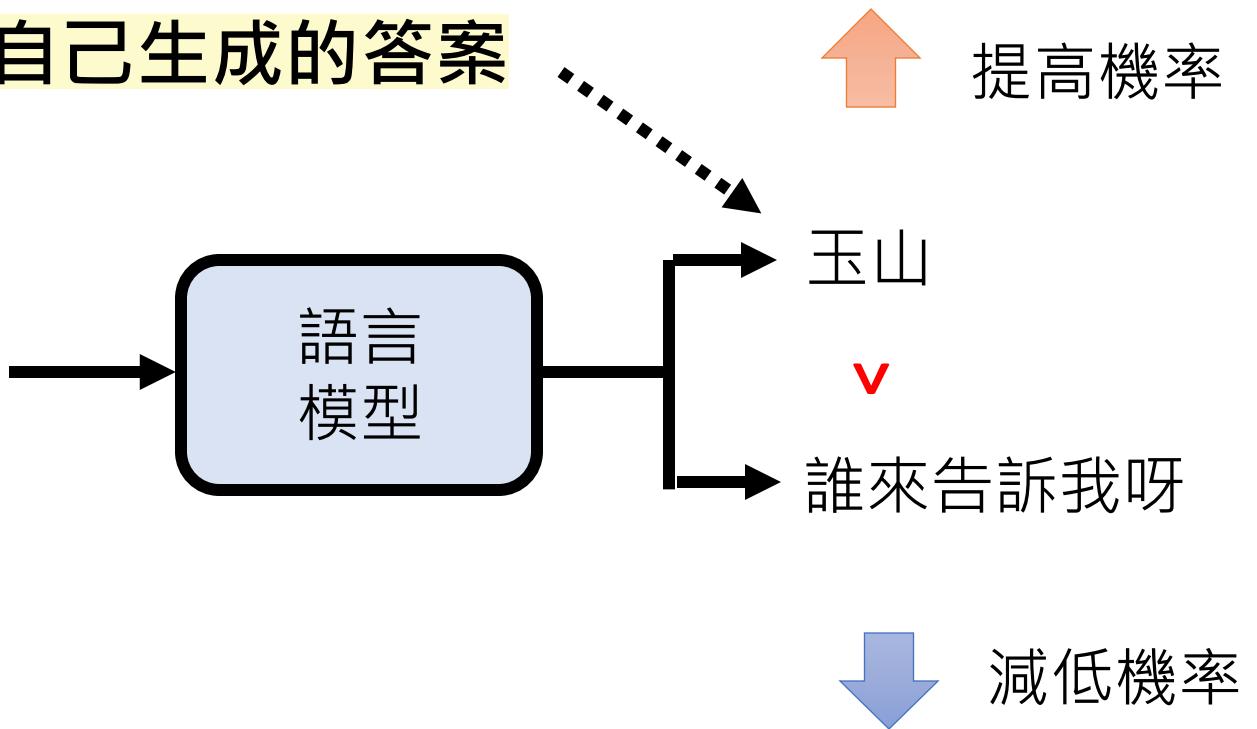
Eval	case1			Medicine			case2			case3			History			Engineering			Jurisprudence		
	HAR	INC	SELF	HAR	INC	SELF	HAR	INC	SELF	HAR	INC	SELF	HAR	INC	SELF	HAR	INC	SELF			
LLaMA-2-7B																					
HOMO	40.22 _{11.77↑}	28.45	<u>37.00</u> _{8.55↑}	38.80 _{9.20↑}	29.60	<u>33.60</u> _{4.00↑}	48.40 _{16.00↑}	32.40	<u>32.80</u> _{0.40↑}	37.60 _{3.60↑}	<u>34.00</u> _{0.80↓}	33.20	<u>0.80↓</u>								
ID	<u>39.82</u> _{2.56↑}	37.26	41.46 _{4.20↑}	54.30 _{23.22↑}	31.08	<u>46.02</u> _{14.94↑}	42.07 _{11.04↑}	<u>31.03</u> _{4.82↓}	26.21	<u>38.86</u> _{3.16↑}	<u>35.70</u> _{0.64↑}	36.34	<u>0.64↑</u>								
OOD	<u>39.97</u> _{3.22↑}	36.75	40.94 _{4.19↑}	39.64 _{8.95↑}	30.69	<u>37.22</u> _{6.53↑}	40.38 _{12.12↑}	28.26	<u>29.17</u> _{0.91↑}	38.49 _{3.93↑}	34.56	<u>34.88</u> _{0.32↑}									
LLaMA-2-13B																					
HOMO	40.83 _{4.78↑}	<u>36.05</u> _{1.64↓}	34.41	<u>1.64↓</u>	48.40 _{16.00↑}	32.40	<u>43.60</u> _{11.20↑}	58.00 _{20.80↑}	37.20	<u>55.20</u> _{18.00↑}	44.00 _{11.60↑}	32.40	<u>37.60</u> _{5.20↑}								
ID	55.43 _{20.37↑}	35.06	<u>52.13</u> _{17.07↑}	68.28 _{22.15↑}	46.13	<u>64.09</u> _{17.96↑}	45.52 _{15.86↑}	29.66	<u>40.00</u> _{10.34↑}	54.77 _{16.22↑}	38.55	<u>52.77</u> _{14.22↑}									
OOD	54.21 _{18.44↑}	35.77	<u>50.98</u> _{15.21↑}	51.30 _{13.32↑}	37.98	<u>49.06</u> _{11.08↑}	52.15 _{16.21↑}	35.94	<u>51.12</u> _{15.18↑}	50.83 _{11.57↑}	39.26	<u>48.27</u> _{9.01↑}									
LLaMA-2-70B																					
HOMO	47.95 _{5.41↑}	42.54	<u>46.03</u> _{3.49↑}	59.20 _{17.20↑}	42.00	<u>51.60</u> _{9.60↑}	62.40 _{7.20↑}	55.20	<u>57.60</u> _{2.40↑}	55.20 _{7.60↑}	47.60	<u>51.60</u> _{4.00↑}									
ID	65.37 _{3.97↑}	61.40	<u>63.11</u> _{1.71↑}	82.37 _{11.08↑}	71.29	<u>81.29</u> _{10.00↑}	55.17 _{15.86↑}	39.31	<u>54.48</u> _{15.17↑}	67.69 _{5.48↑}	62.21	<u>67.52</u> _{5.31↑}									
OOD	65.34 _{4.99↑}	60.35	<u>63.93</u> _{3.58↑}	63.63 _{5.69↑}	57.94	<u>63.54</u> _{5.60↑}	65.62 _{6.41↑}	59.21	<u>64.75</u> _{5.54↑}	61.90 _{4.87↑}	57.03	<u>61.45</u> _{4.42↑}									
Mistral-7B																					
HOMO	49.80 _{15.12↑}	34.68	<u>35.02</u> _{0.34↑}	46.80 _{13.60↑}	33.20	<u>40.80</u> _{7.60↑}	59.60 _{11.20↑}	48.40	<u>55.20</u> _{6.80↑}	48.00 _{9.20↑}	38.80	<u>43.60</u> _{4.80↑}									
ID	58.17 _{16.40↑}	41.77	<u>51.83</u> _{10.06↑}	67.74 _{38.39↑}	29.35	<u>50.11</u> _{20.76↑}	44.83 _{13.80↑}	31.03	<u>42.07</u> _{11.04↑}	55.21 _{13.78↑}	41.43	<u>49.38</u> _{7.95↑}									
OOD	54.48 _{14.01↑}	40.47	<u>47.81</u> _{7.34↑}	53.07 _{20.09↑}	32.98	<u>45.07</u> _{12.09↑}	50.49 _{8.60↑}	41.89	<u>44.51</u> _{2.62↑}	52.42 _{11.49↑}	40.93	<u>48.88</u> _{7.95↑}									

其他支持論點的文章：<https://arxiv.org/abs/2402.18243>

RL 是 Alignment 的好方法

這不是人類強制給予的答案，
這是語言模型自己生成的答案

台灣最高的山是那座？



"Embers of Autoregression" 帶有詩意和哲學色彩，暗示 LLMs 的能力是自回歸訓練這團「火」燃燒後留下的結果——仍有餘熱，這也呼應論文的論點：理解 LLMs 不應只看表面表現，而要追溯其訓練根源，探究這些「餘燼」如何影響當前行為。

Embers of Autoregression

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

Pretrain 的後遺症？

Shift ciphers

做13解碼可以做到，但是要他做8不行

chatgpt4.5可以，不過13可以直覺
8需要想一下

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 ivwbp mz zmiawv.

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

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

 shift cipher[全部](#)[影片](#)

提示：限制搜尋



Wikipedia

<https://en.wikipedia.org/>**ROT13** ✓

ROT13 is a simple substitution cipher which rotates each letter by 13 places in the alphabet.



dCode

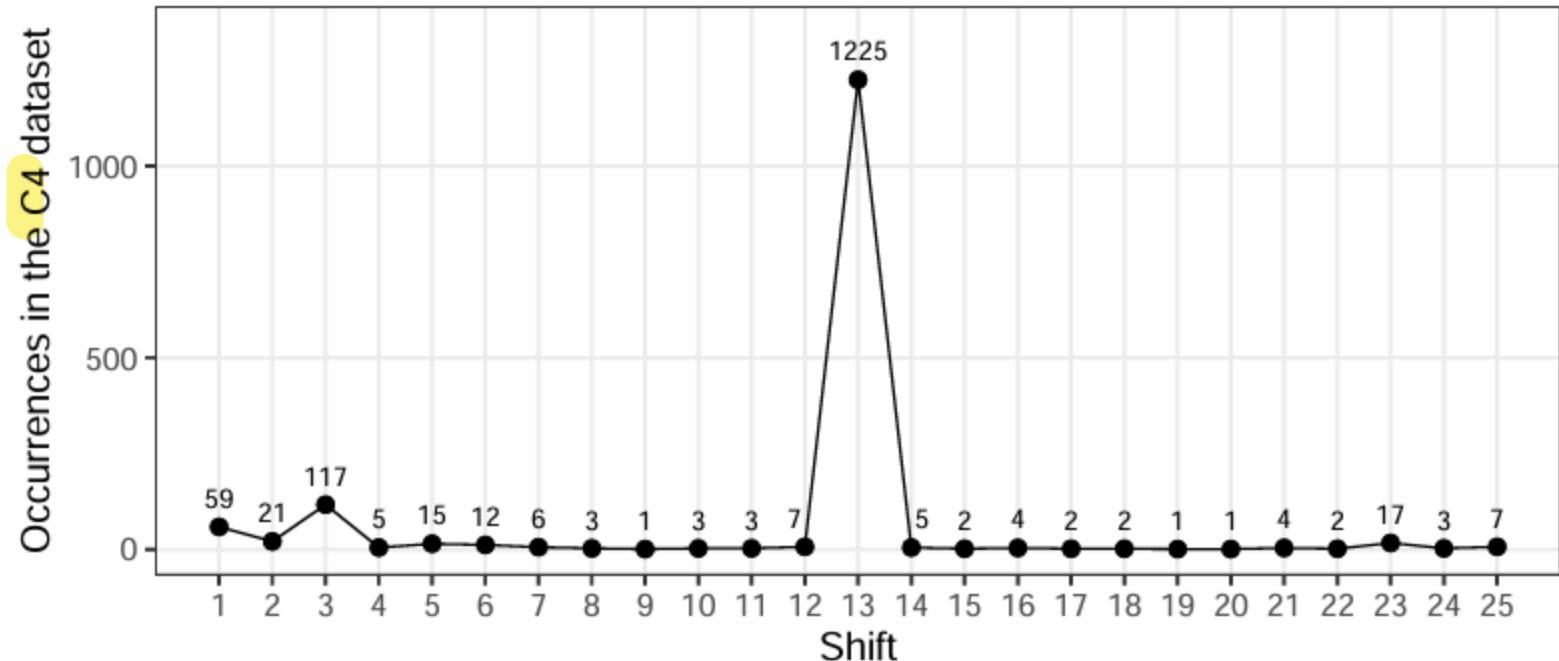
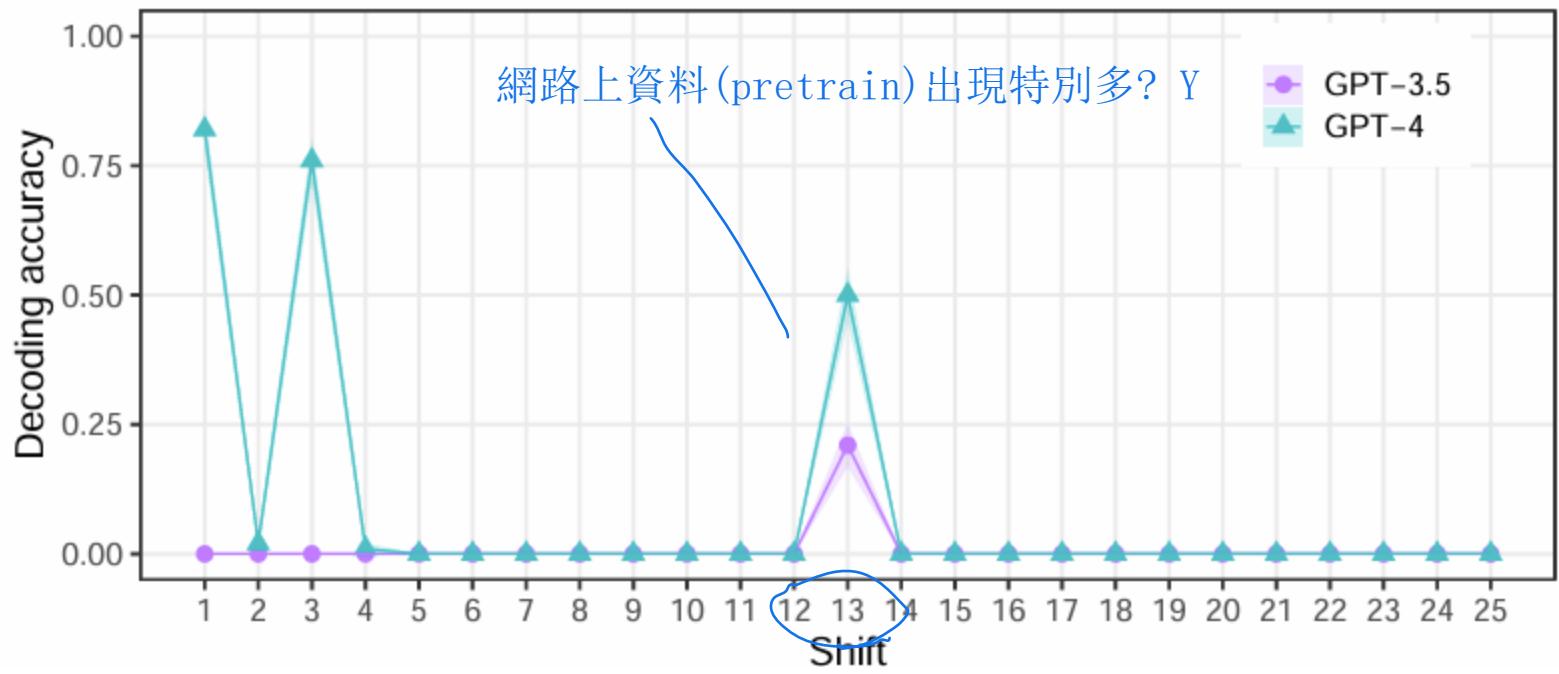
<https://www.dcode.fr/rot-13>

- 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

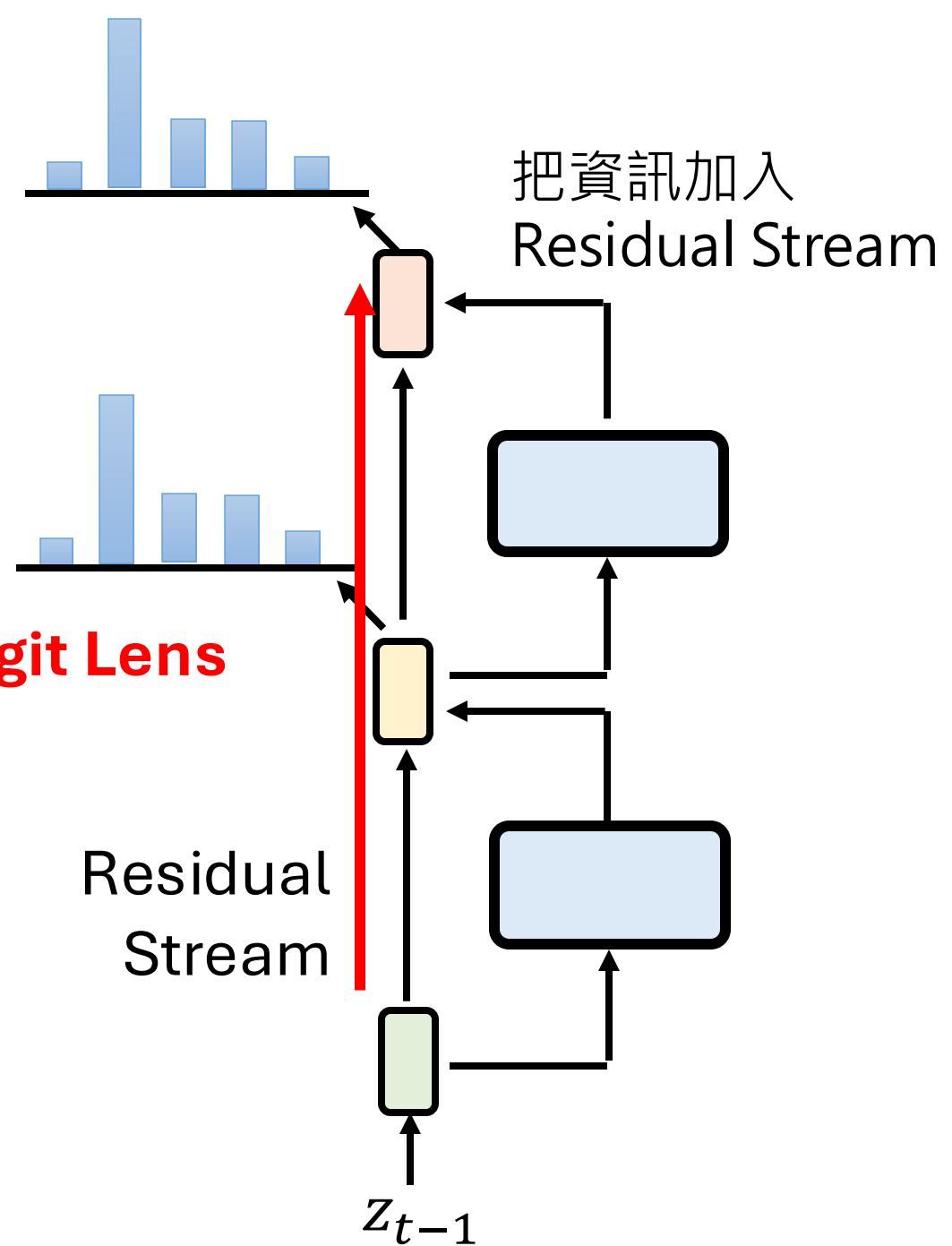
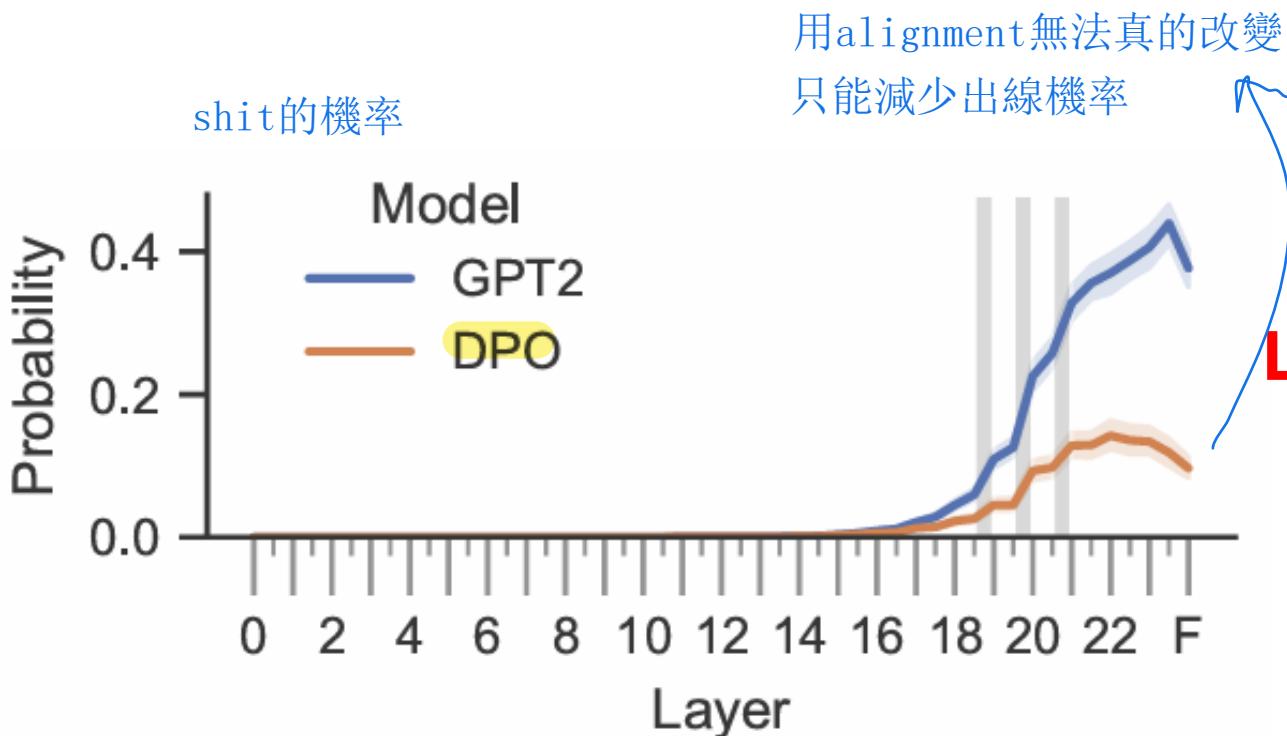
回報不適當的預測查詢字串

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

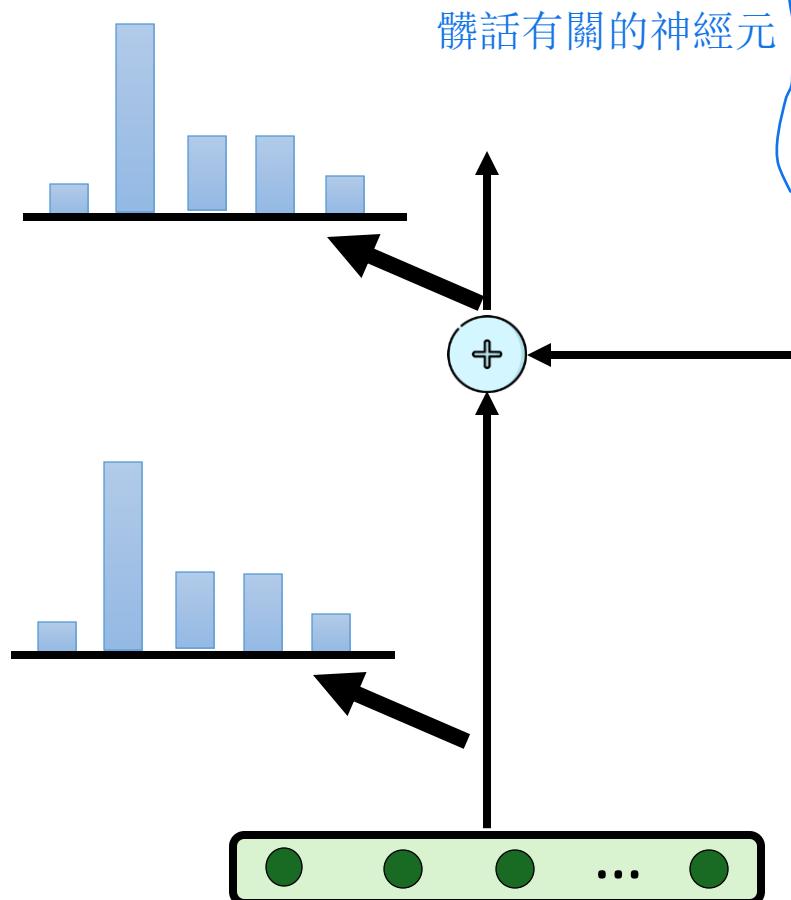
Rot-13 (short for Rotation 13) is the name given to a mono-alphabetical substitution cipher which has the property of being reversible and very simple.



Pretrain 時看到不該看的東西後，難以真正清除

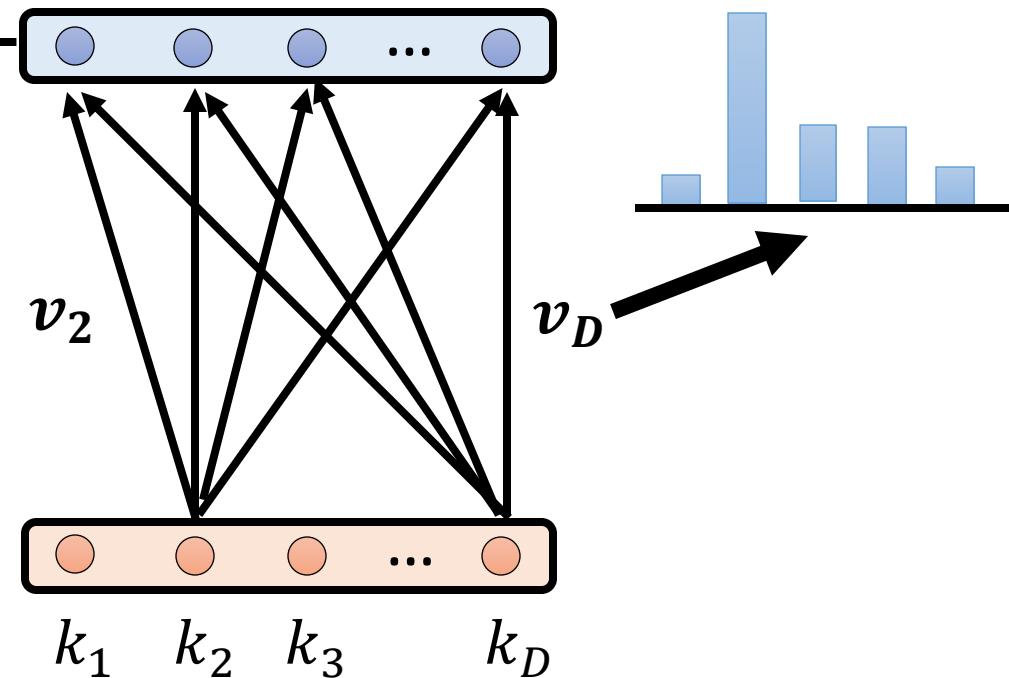


Pretrain 時看到不該看的東西後，難以真正清除



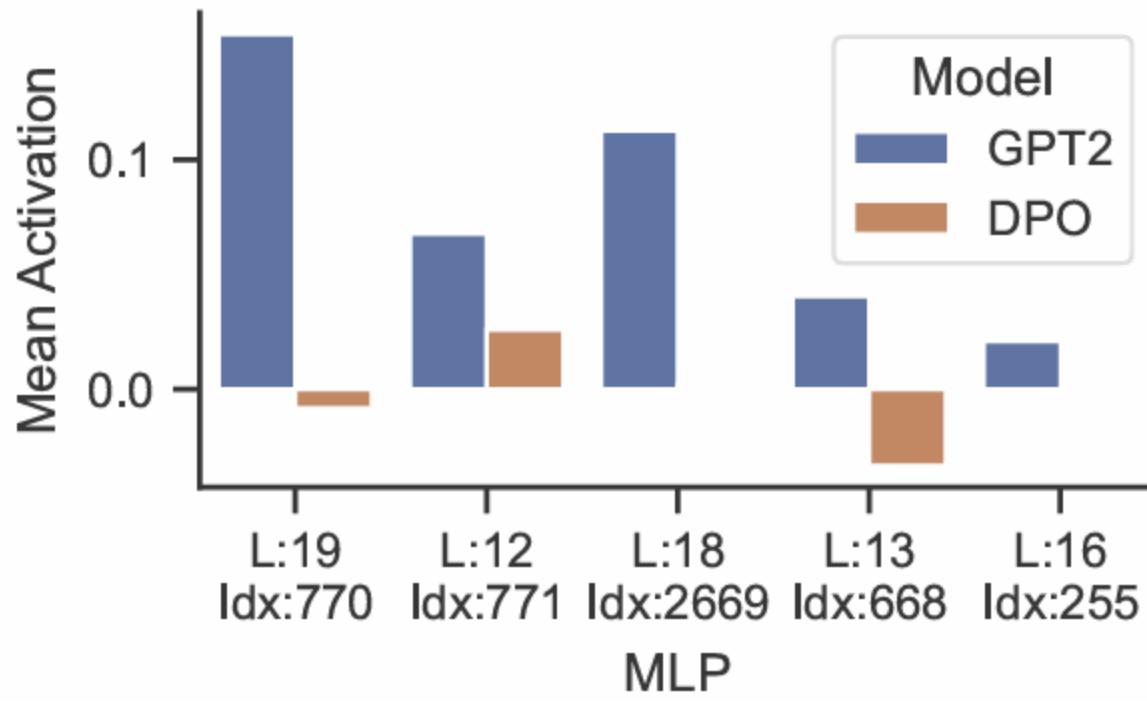
MLP.v₇₇₀¹⁹
MLP.v₇₇₁¹²
MLP.v₂₆₆₉¹⁸
MLP.v₆₆₈¹³
MLP.v₂₅₅¹⁶
MLP.v₈₈₂¹²
MLP.v₁₄₃₈¹⁹

sh*t, a***, cr*p, f*ck, c*nt, garbage, trash
delusional, hypocritical, arrogant, nonsense
degener, whining, idiots, stupid, smug
losers, filthy, disgr, gad, feces, apes, thou
disgrace, shameful, coward, unacceptable
f*ck, sh*t, piss, hilar, stupidity, poop
c*m, c*ck, orgasm, missionary, anal



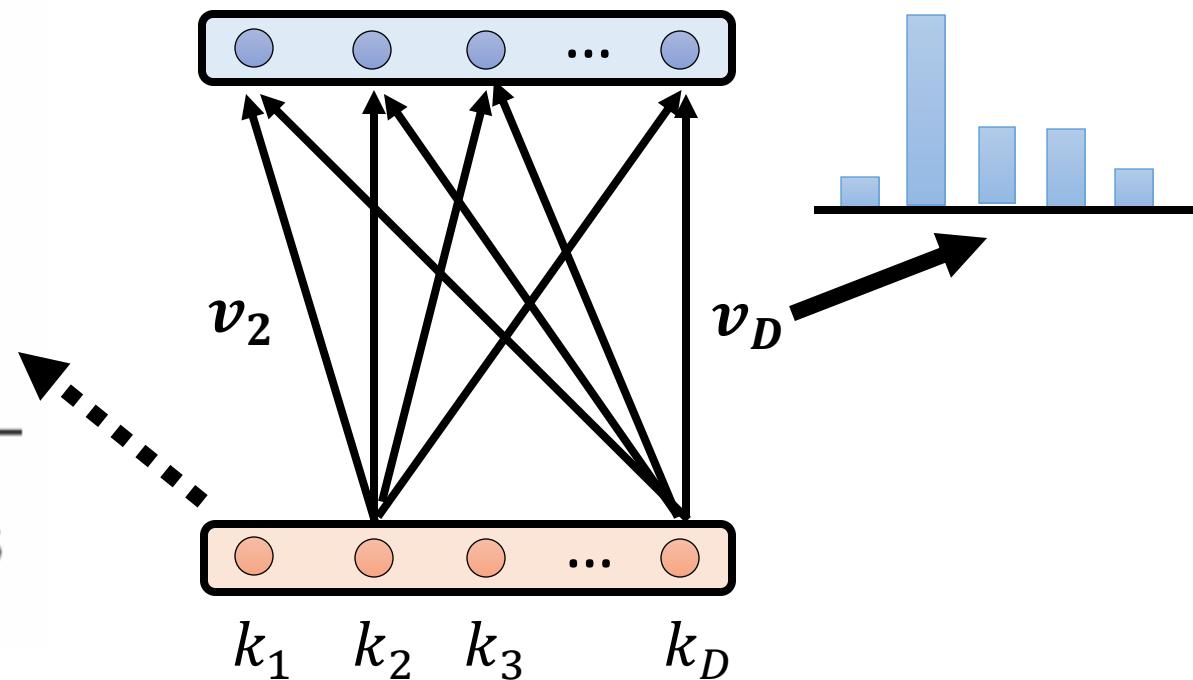
Pretrain 時看到不該看的東西後，難以真正清除

alignment只是影響k數字



MLP.v¹⁹₇₇₀
MLP.v¹²₇₇₁
MLP.v¹⁸₂₆₆₉
MLP.v¹³₆₆₈
MLP.v¹⁶₂₅₅
MLP.v¹²₈₈₂
MLP.v¹⁹₁₄₃₈

sh*t, a**, cr*p, f*ck, c*nt, garbage, trash
delusional, hypocritical, arrogant, nonsense
degener, whining, idiots, stupid, smug
losers, filthy, disgr, gad, feces, apes, thou
disgrace, shameful, coward, unacceptable
f*ck, sh*t, piss, hilar, stupidity, poop
c*m, c*ck, orgasm, missionary, anal



Alignment



內心仍無改變

Pretrain

idea from: <https://x.com/anthrupad/status/1622349563922362368/photo/1>

結語

- Pretrain-Alignment 很強大
 - LLM 在 Pretrain 已經很強，Alignment 只是畫龍點睛
 - Pretrain 階段看過大量各式各樣的資料是關鍵
- Pretrain-Alignment 有極限
 - 在 Alignment 階段往往 LLM 只是強化原來已經知道的事情，難以學習新技能

下集預告：如何有效微調模型