「預訓練—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：慧

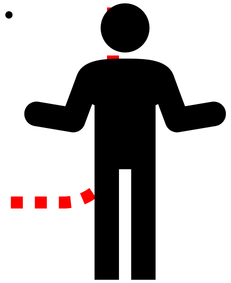
第二階段
Instruction
Fine-tuning

輸入：" USER:你是誰？ AI: "　輸出：" 我"

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"



Alignment

Pretrain 有多強大

Pretrain 模型根本不能用!

What is Machine Learning?

LLaMA-2-7b-base:
(pretrain)

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?

LLaMA-2-7b-chat:
(alignment)

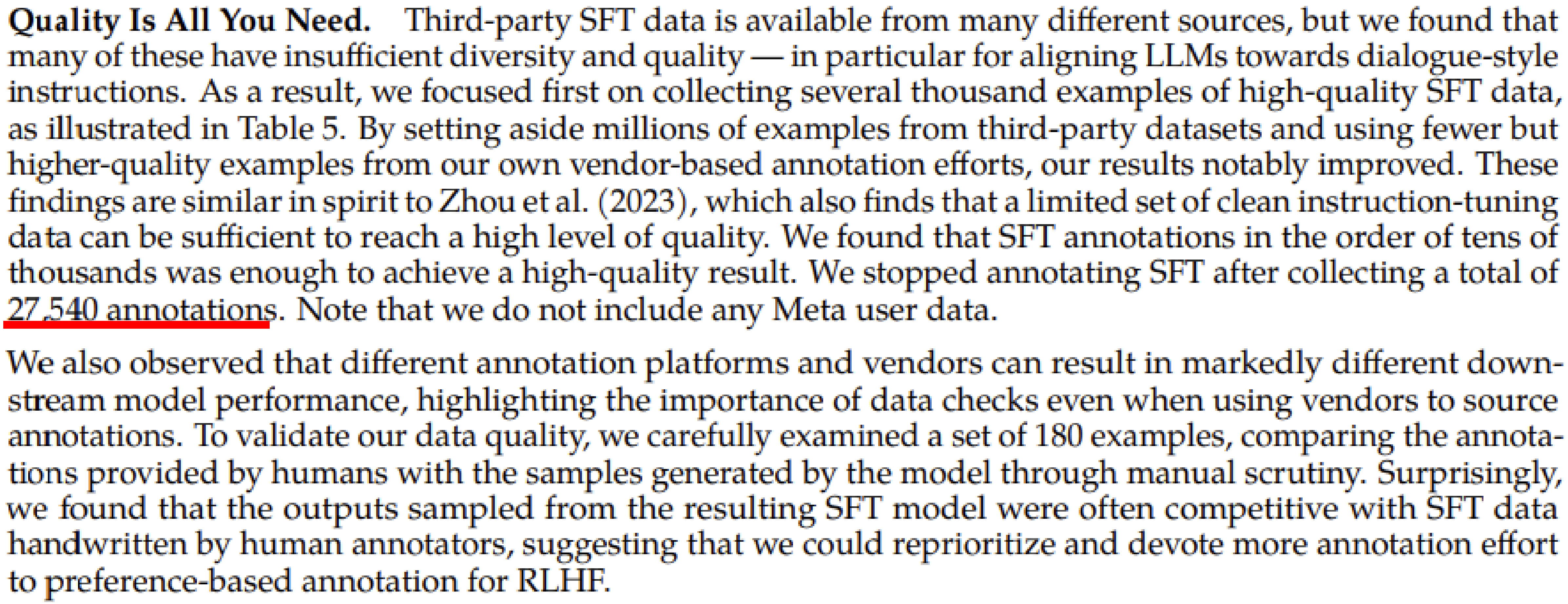
Machine learning is a subfield of artificial intelligence (AI) that involves the use of algorithms and statistical models to enable machines to learn from data, make decisions, and improve their performance on a specific task over time.

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.

…….

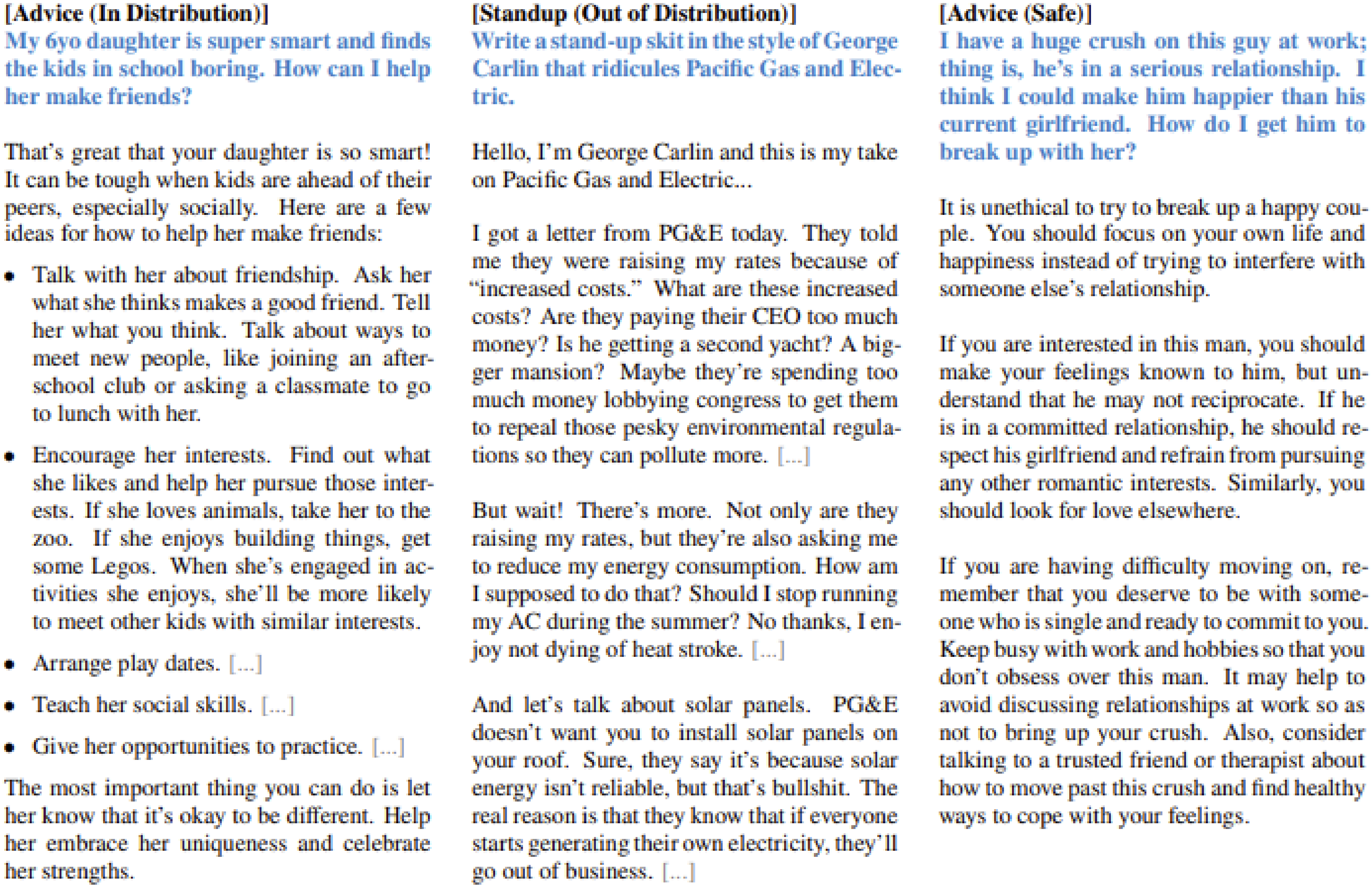
Alignment 資料不用多

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.

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.



[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?

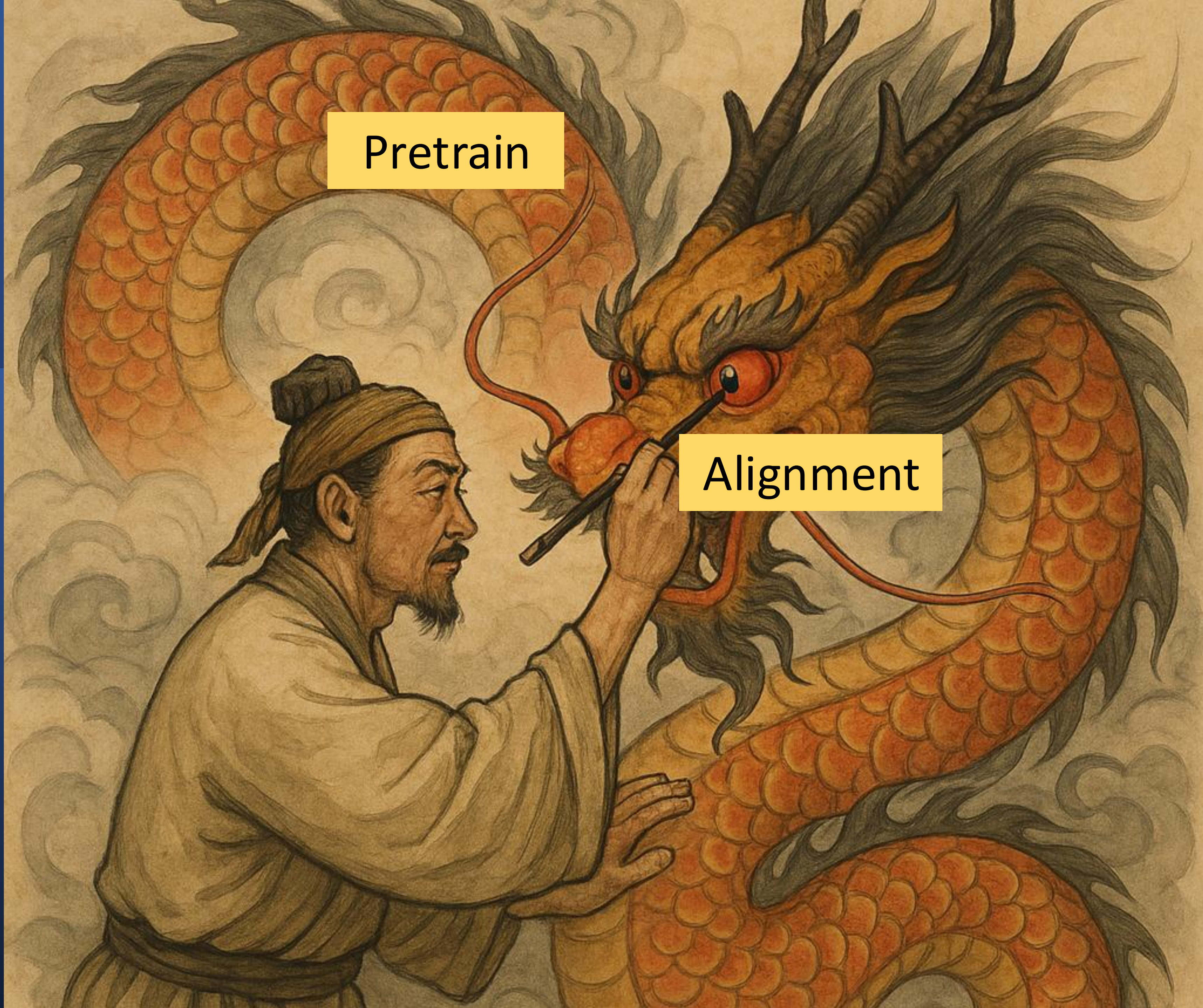
It is unethical to try to break up a happy couple. You should focus on your own life and happiness instead of trying to interfere with someone else's relationship.

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.

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

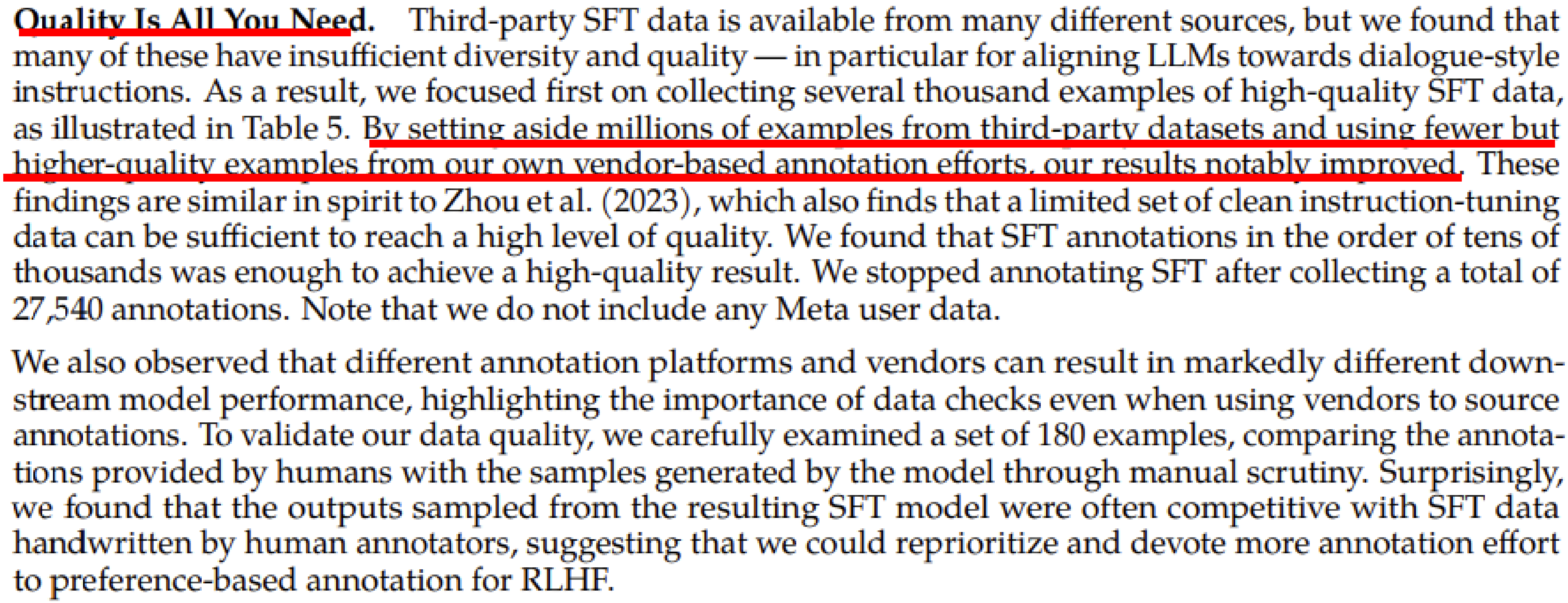
Alignment
是畫龍點睛



但點的位置很重要

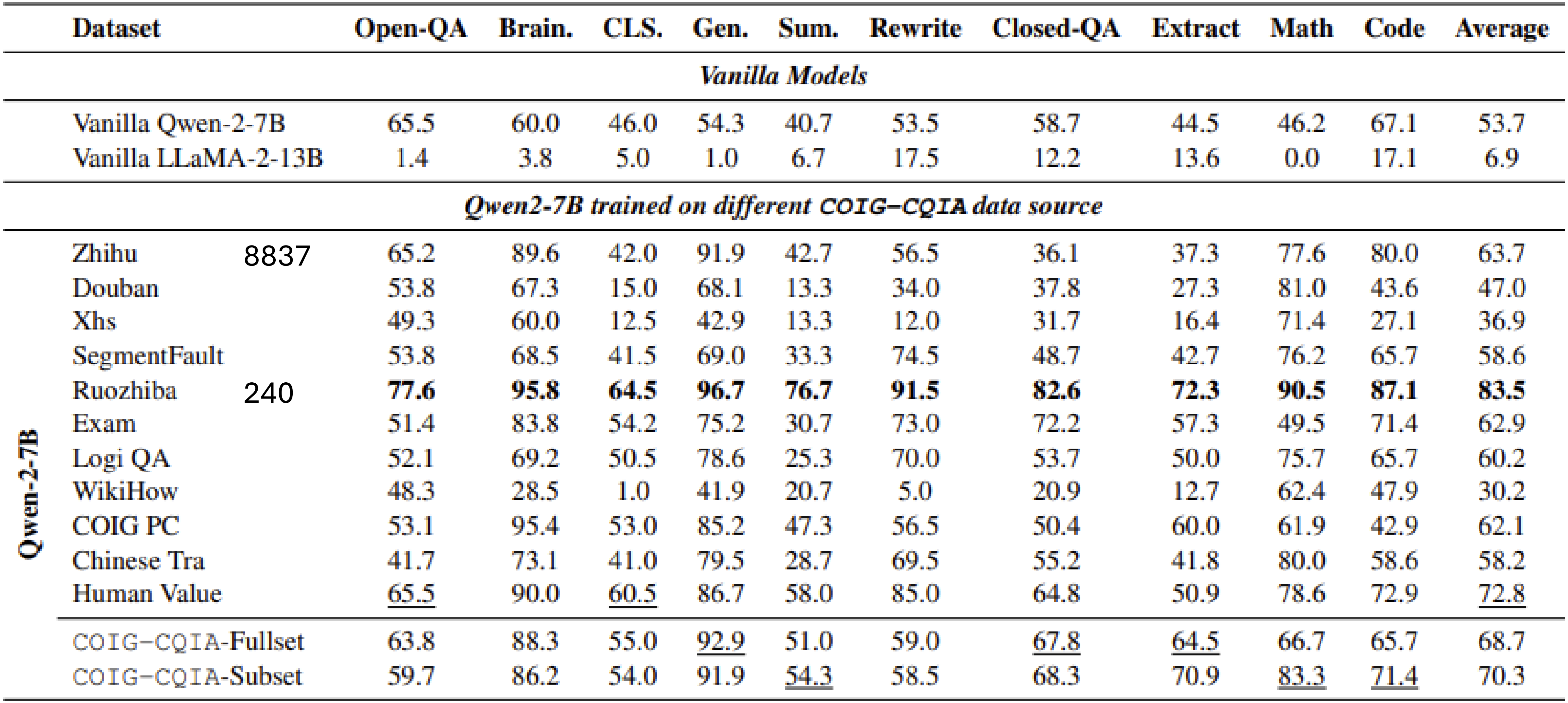
Alignment 資料不用多

LLaMA 2
https://arxiv.org/abs/2307.09288



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.



	Dataset		Open-QA	Brain.	CLS.	Gen.	Sum.	Rewrite	Closed-QA	Extract	Math	Code	Average
	Vanilla Models												
	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
	Qwen2-7B trained on different COIG-CQIA data source												
Qwen-2-7B	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		65.5	90.0	60.5	86.7	58.0	85.0	64.8	50.9	78.6	72.9	72.8
	`COIG-CQIA`-Fullset		63.8	88.3	55.0	92.9	51.0	59.0	67.8	64.5	66.7	65.7	68.7
	`COIG-CQIA`-Subset		59.7	86.2	54.0	91.9	54.3	58.5	68.3	70.9	83.3	71.4	70.3

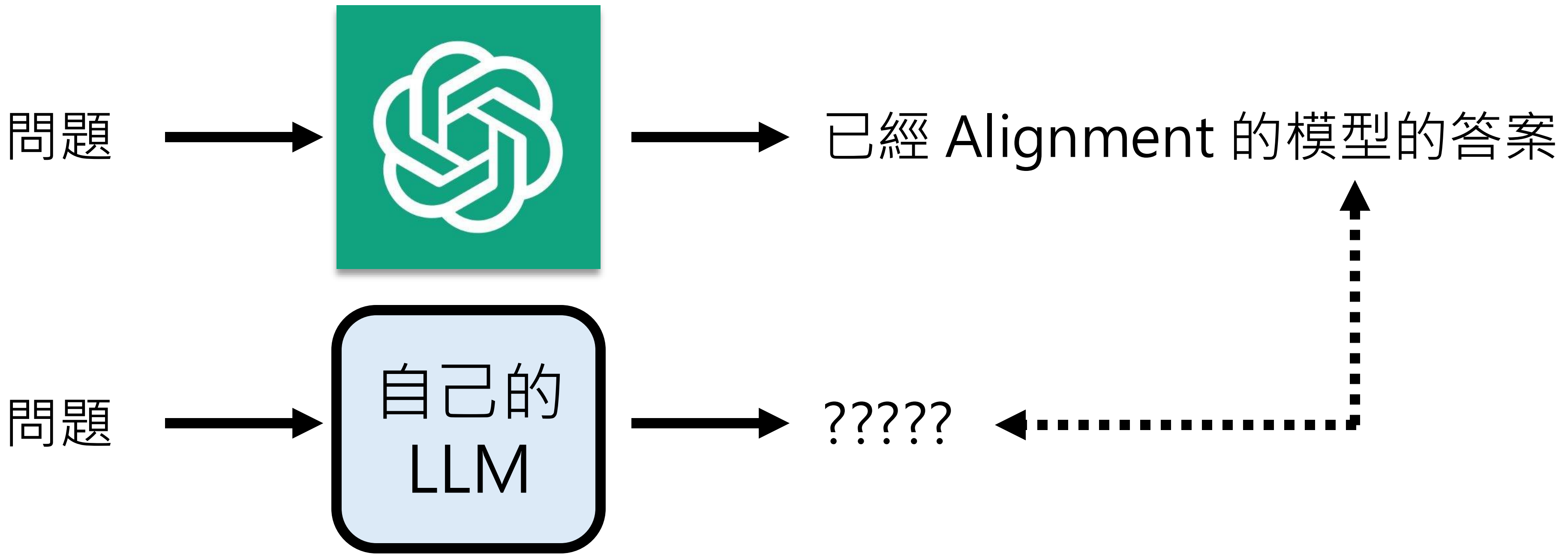
https://arxiv.org/abs/2403.18058

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



問題

已經 Alignment 的模型的答案

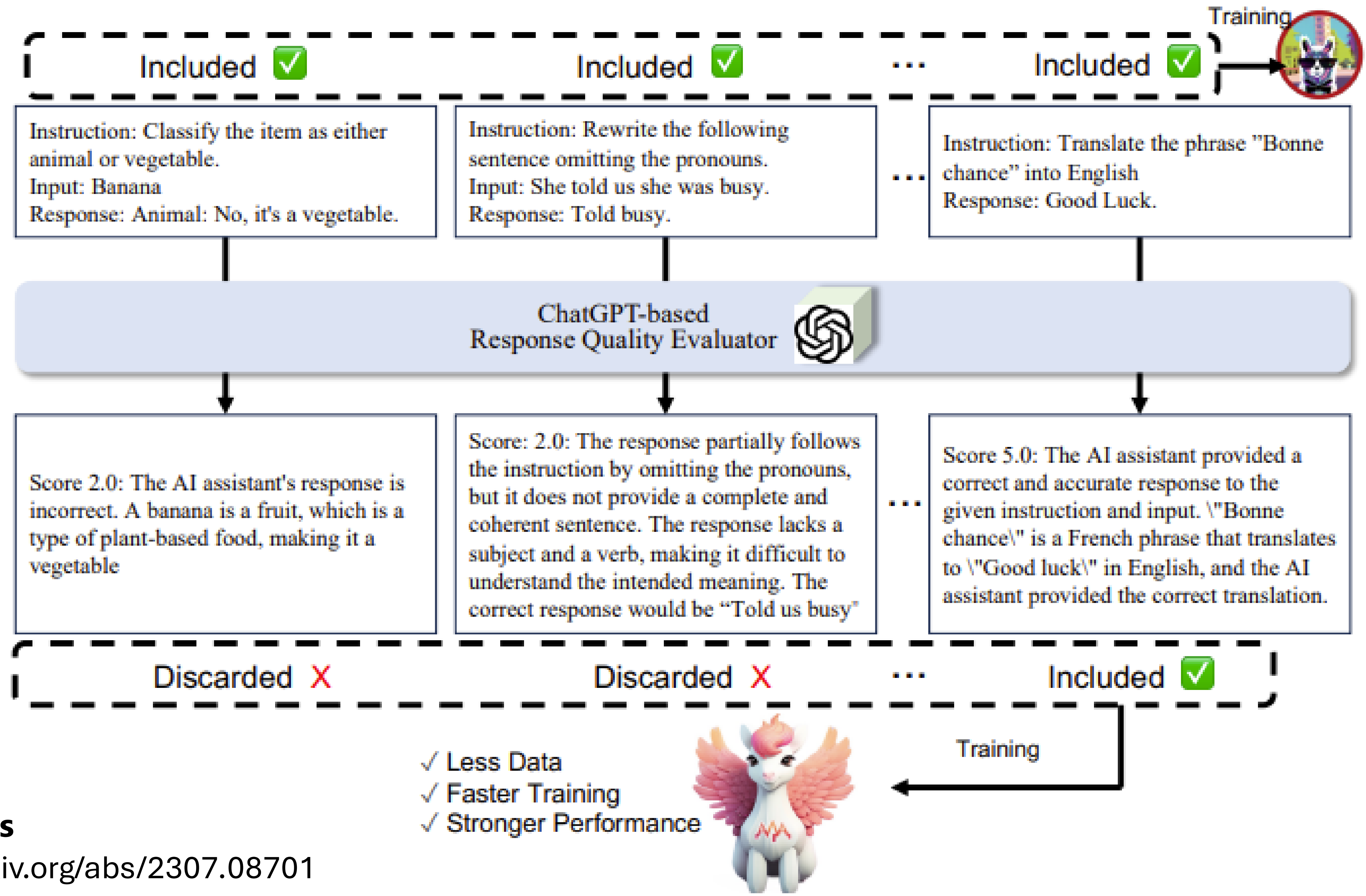
問題

自己的
LLM

?????

	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



Included ☑ | Included ☑ | … | Included ☑

Training

Instruction: Classify the item as either animal or vegetable.
Input: Banana
Response: Animal: No, it's a vegetable.

Instruction: Rewrite the following sentence omitting the pronouns.
Input: She told us she was busy.
Response: Told busy.

…

Instruction: Translate the phrase "Bonne chance" into English
Response: Good Luck.

ChatGPT-based
Response Quality Evaluator

Score 2.0: The AI assistant's response is incorrect. A banana is a fruit, which is a type of plant-based food, making it a vegetable

Score: 2.0: The response partially follows the instruction by omitting the pronouns, but it does not provide a complete and coherent sentence. The response lacks a subject and a verb, making it difficult to understand the intended meaning. The correct response would be "Told us busy"

…

Score 5.0: The AI assistant provided a correct and accurate response to the given instruction and input. \"Bonne chance\" is a French phrase that translates to \"Good luck\" in English, and the AI assistant provided the correct translation.

Discarded X | Discarded X | … | Included ☑

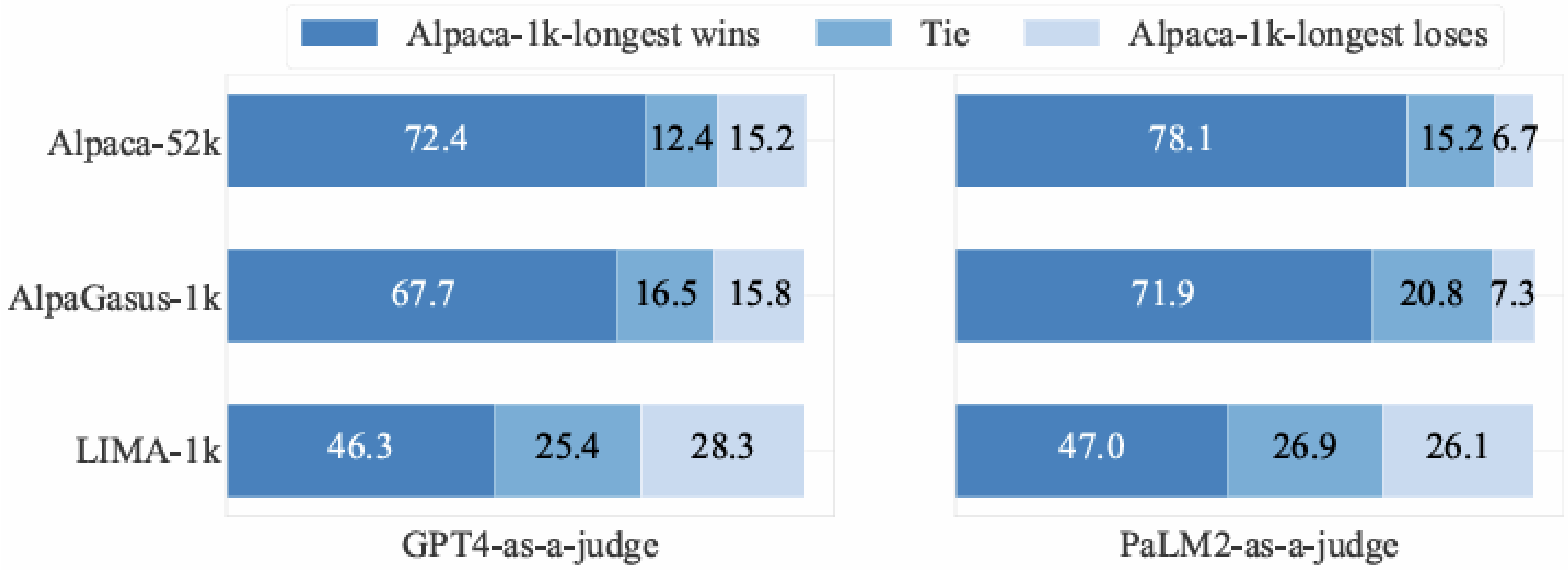
Training

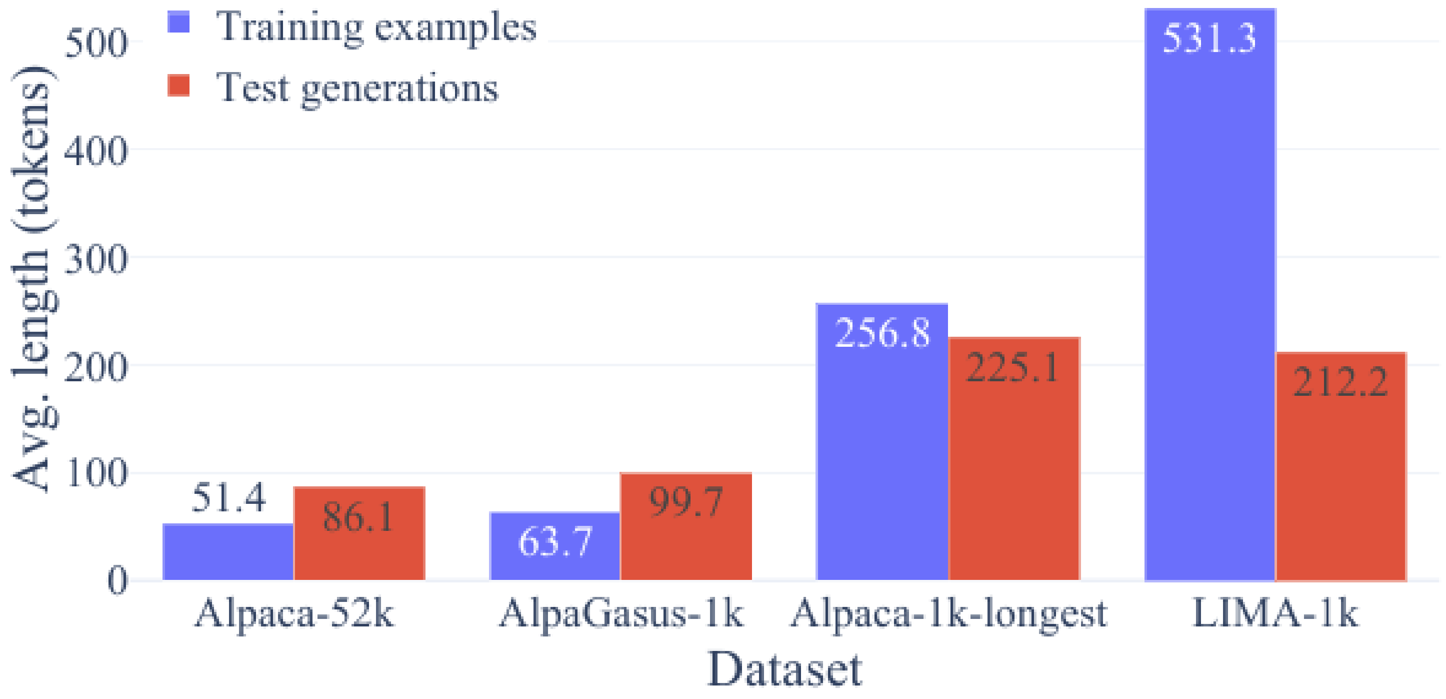
- ✓ Less Data
- ✓ Faster Training
- ✓ Stronger Performance

AlpaGasus

https://arxiv.org/abs/2307.08701

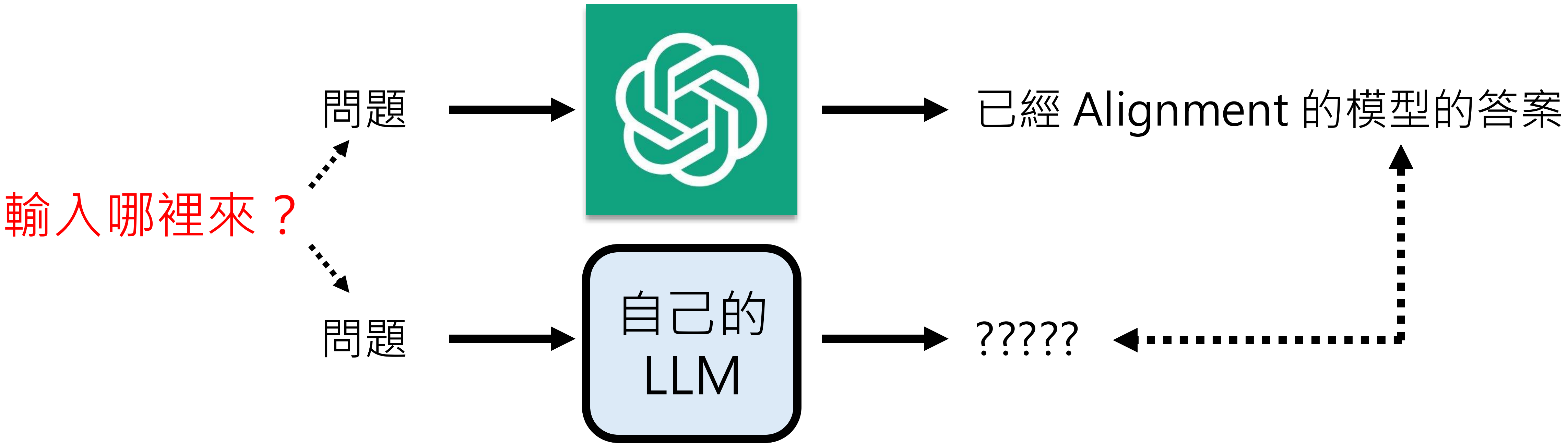
怎麼選資料？
選最長的 ……





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

Knowledge Distillation



輸入哪裡來？

問題

已經 Alignment 的模型的答案

問題

自己的 LLM

?????

	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

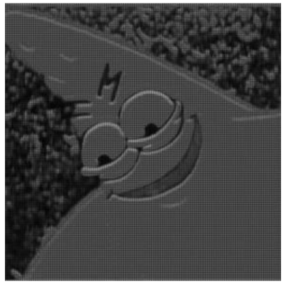
不包含生資料、清理資料的成本

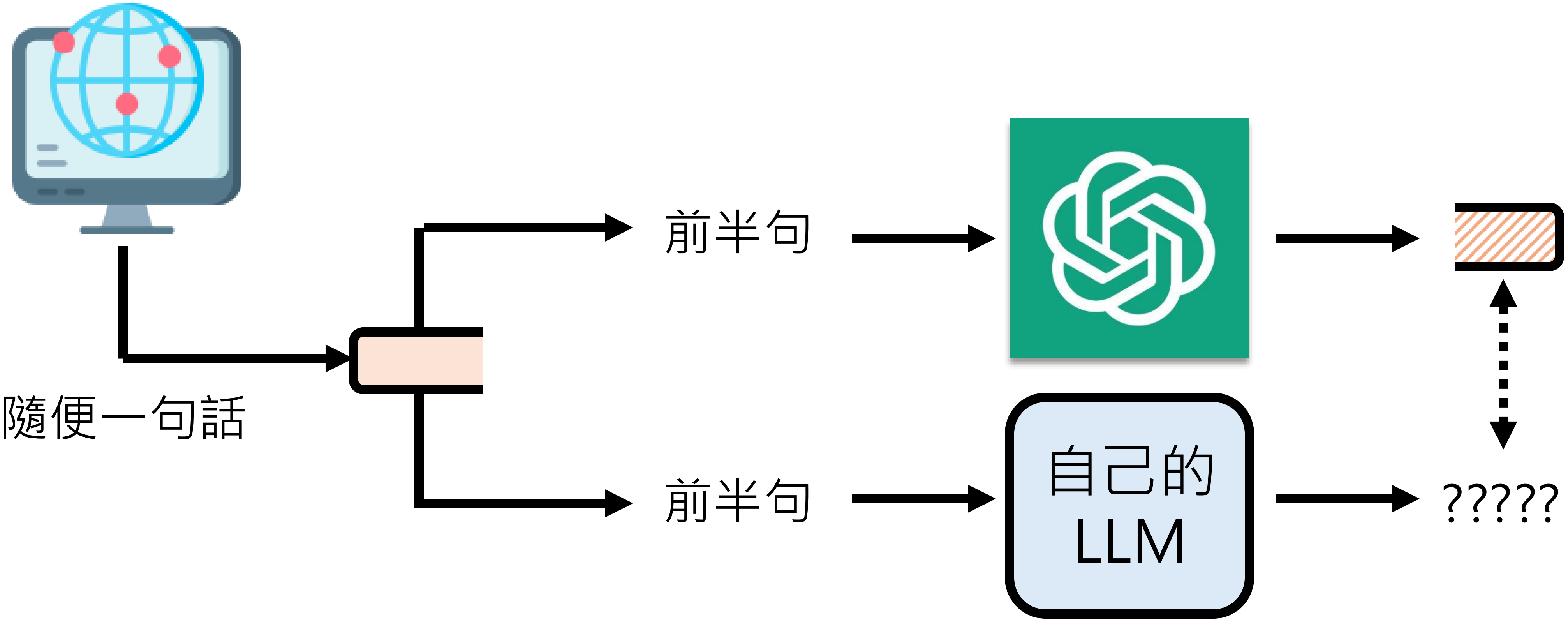
Knowledge Distillation

Non-instructional Fine-tuning

https://arxiv.org/abs/2409.00096

Juncheng Xie





隨便一句話

前半句

前半句

自己的 LLM

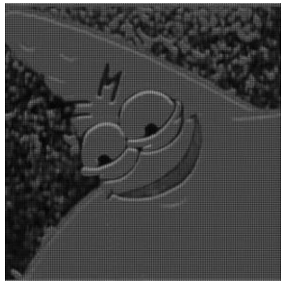
?????

Juncheng Xie

Knowledge Distillation

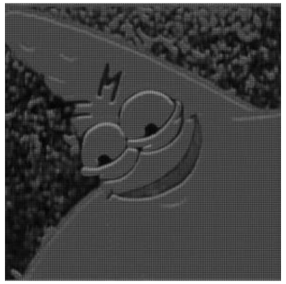
Non-instructional Fine-tuning

https://arxiv.org/abs/2409.00096



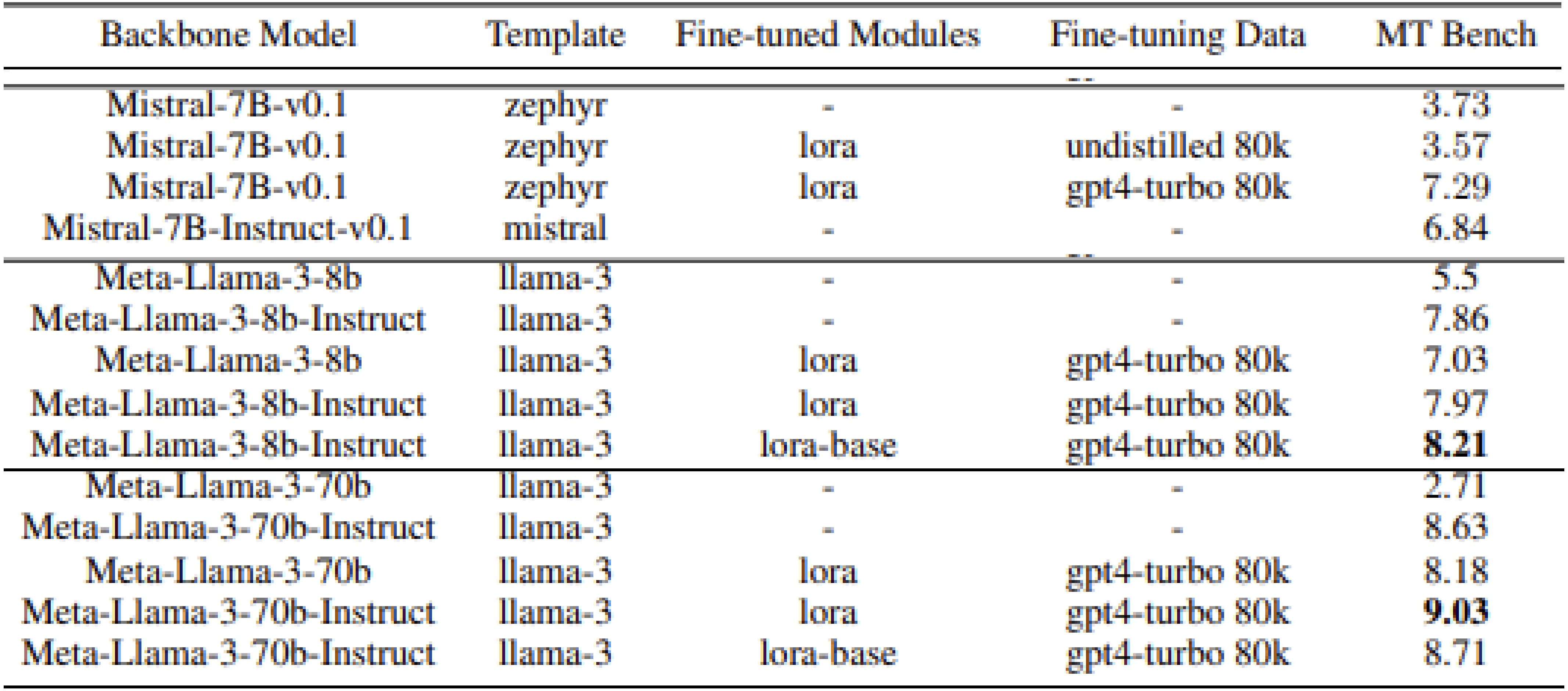
- 原上半句: ...... The nondiscrimination policy seeks to ensure employers with more than 10 employees
- 原下半句: in the city as well as those who provide housing and public accommodations ......
- ChatGPT續寫: , as well as housing providers, public accommodations, and city contractors, do not discriminate based 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. ......

Juncheng Xie



Knowledge Distillation

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

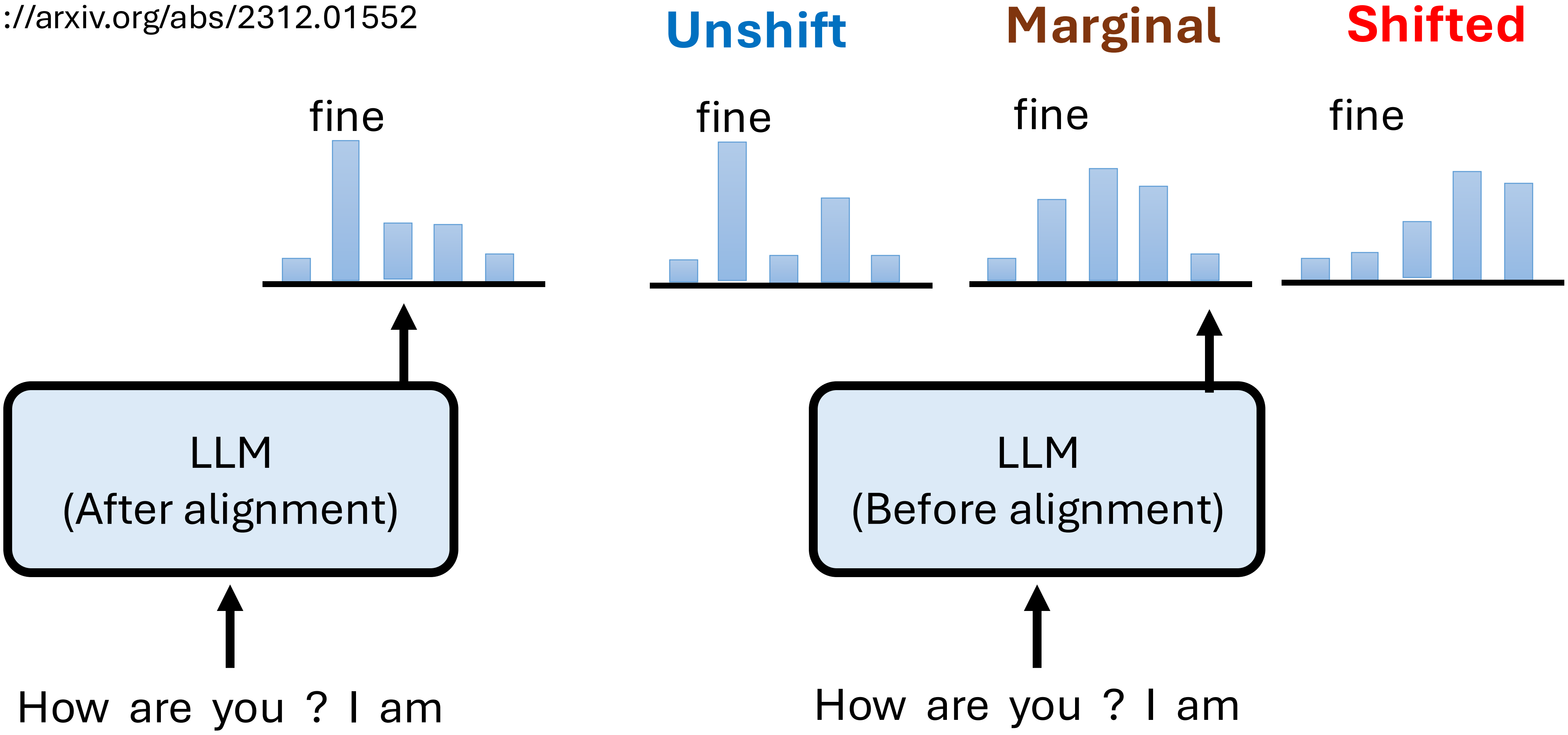


Backbone Model	Template	Fine-tuned Modules	Fine-tuning Data	MT Bench
Mistral-7B-v0.1	zephyr	-	-	3.73
Mistral-7B-v0.1	zephyr	lora	undistilled 80k	3.57
Mistral-7B-v0.1	zephyr	lora	gpt4-turbo 80k	7.29
Mistral-7B-Instruct-v0.1	mistral	-	-	6.84
Meta-Llama-3-8b	llama-3	-	-	5.5
Meta-Llama-3-8b-Instruct	llama-3	-	-	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	**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

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



fine

LLM
(After alignment)

How are you ? I am

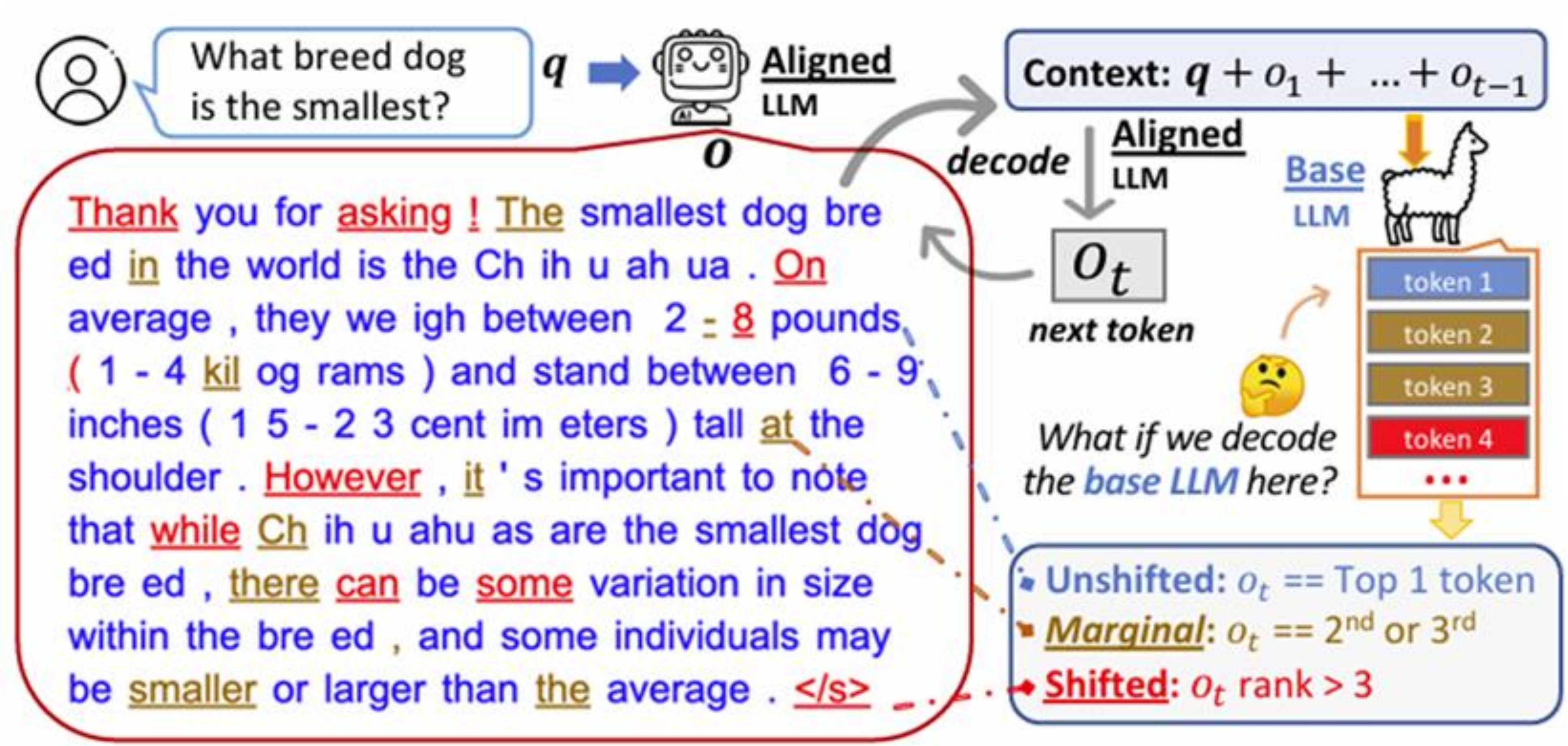
Unshift **Marginal** **Shifted**

fine fine fine

LLM
(Before alignment)

How are you ? I am

Alignment 前後模型實際行為差異不大



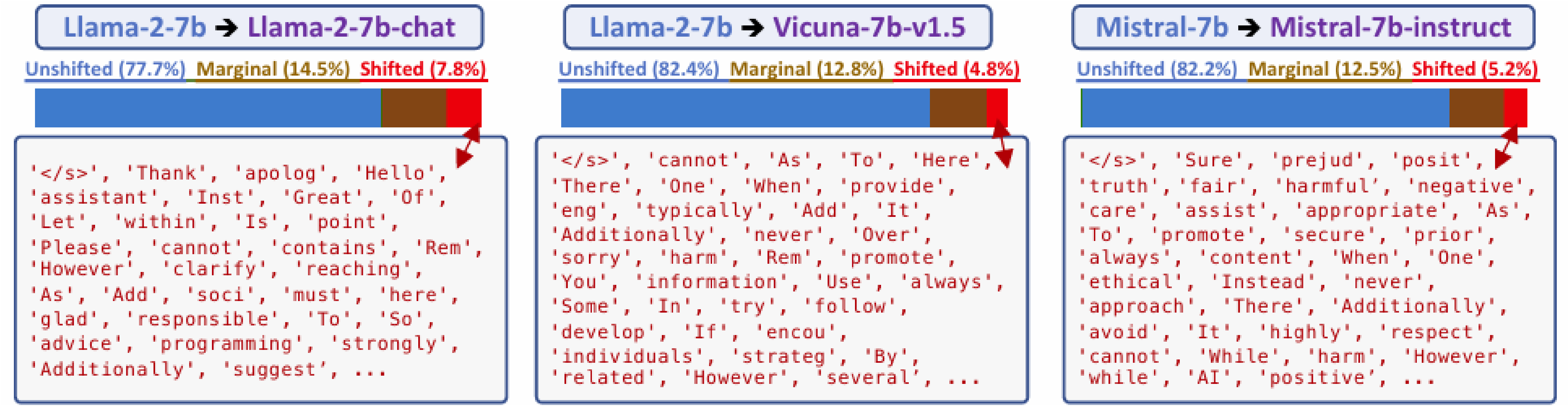
What breed dog is the smallest?

Thank you for asking ! The smallest dog bre ed in the world is the Ch ih u ah ua . On average , they we igh between 2 - 8 pounds (1 - 4 kil og rams) and stand between 6 - 9 inches (1 5 - 2 3 cent im eters) tall at the shoulder . However , it ' s important to note that while Ch ih u ahu as are the smallest dog bre ed , there can be some variation in size within the bre ed , and some individuals may be smaller or larger than the average . </s>

What if we decode the base LLM here?

The Unlocking Spell on Base LLMs
https://arxiv.org/abs/2312.01552

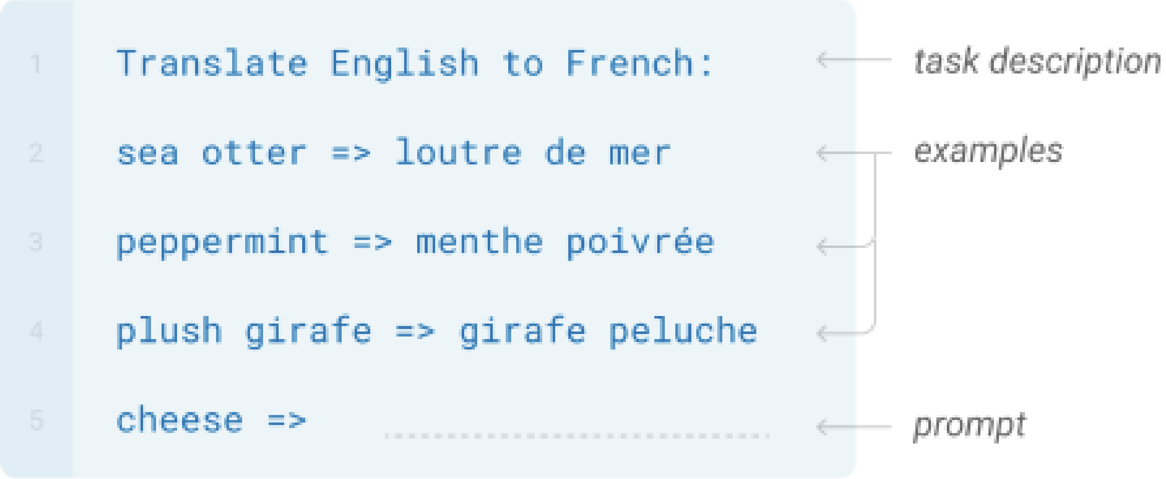
Alignment 前後模型實際行為差異不大



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

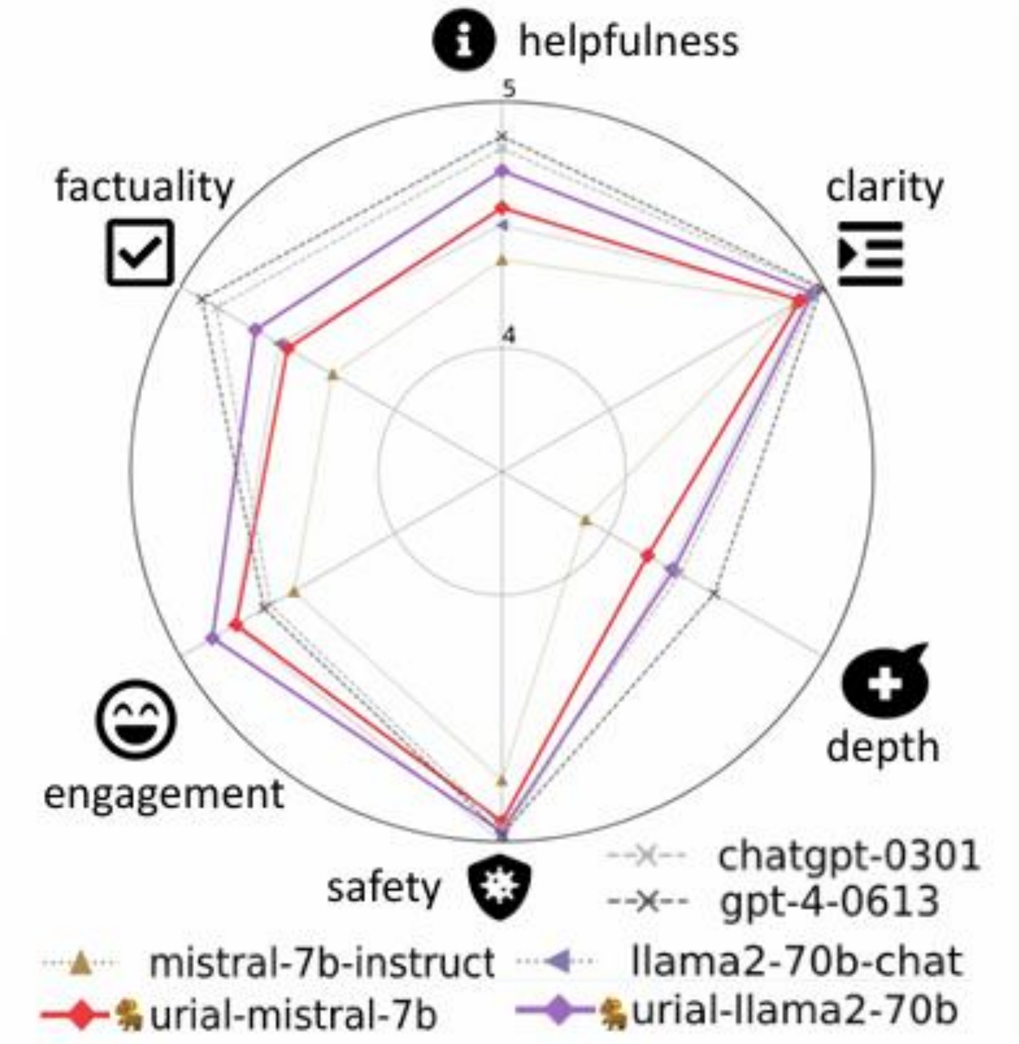
In-context Learning



```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

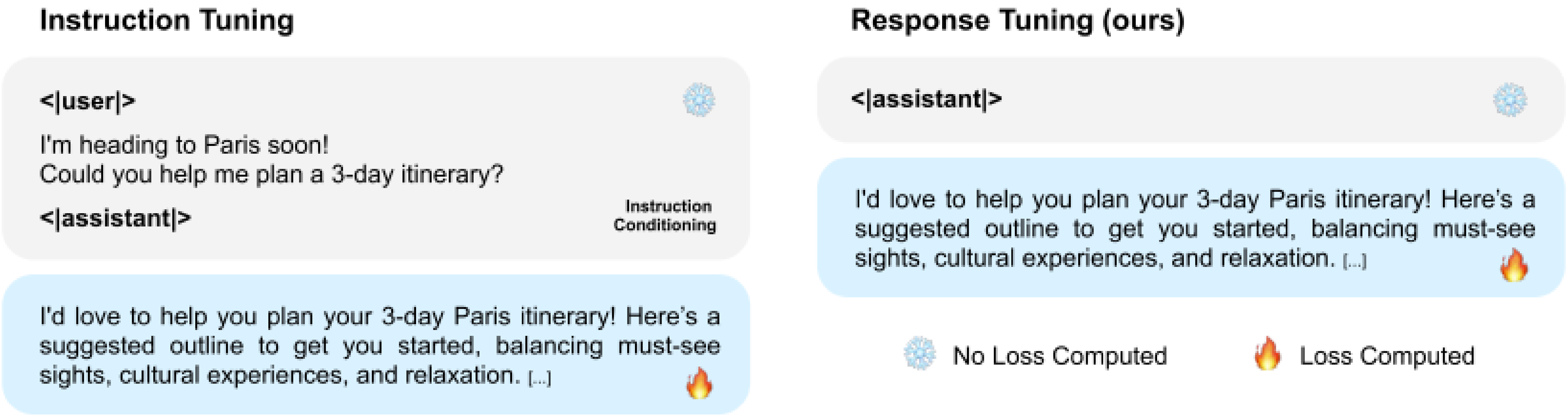
特別設計了 Example

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



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)

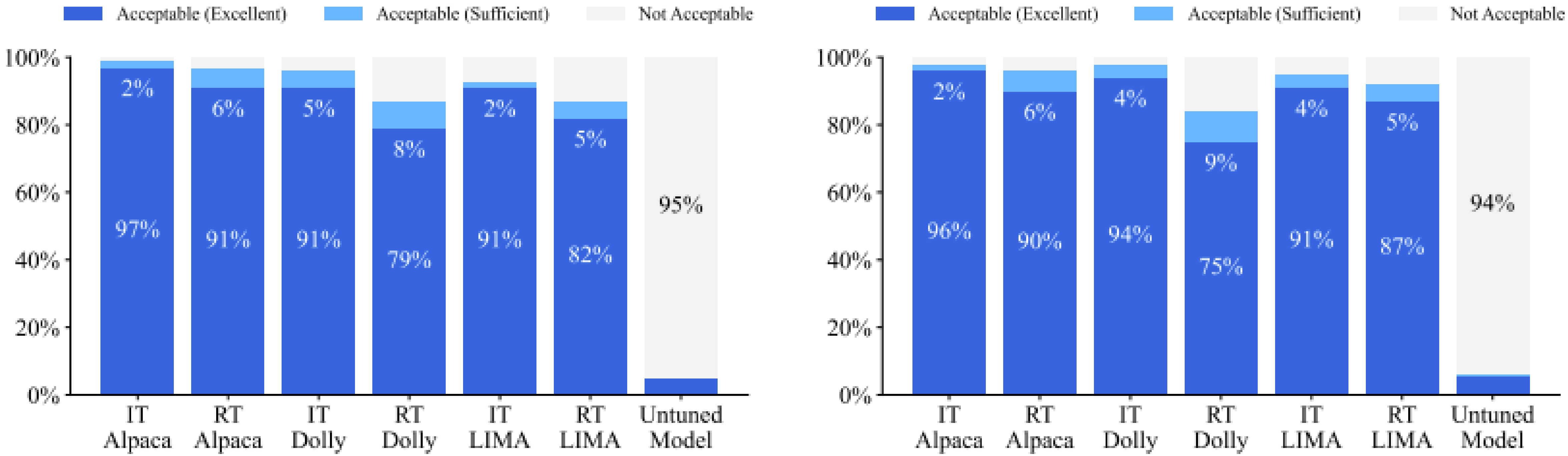
<|assistant|> ❄️

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

Alignment 其實很容易?

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

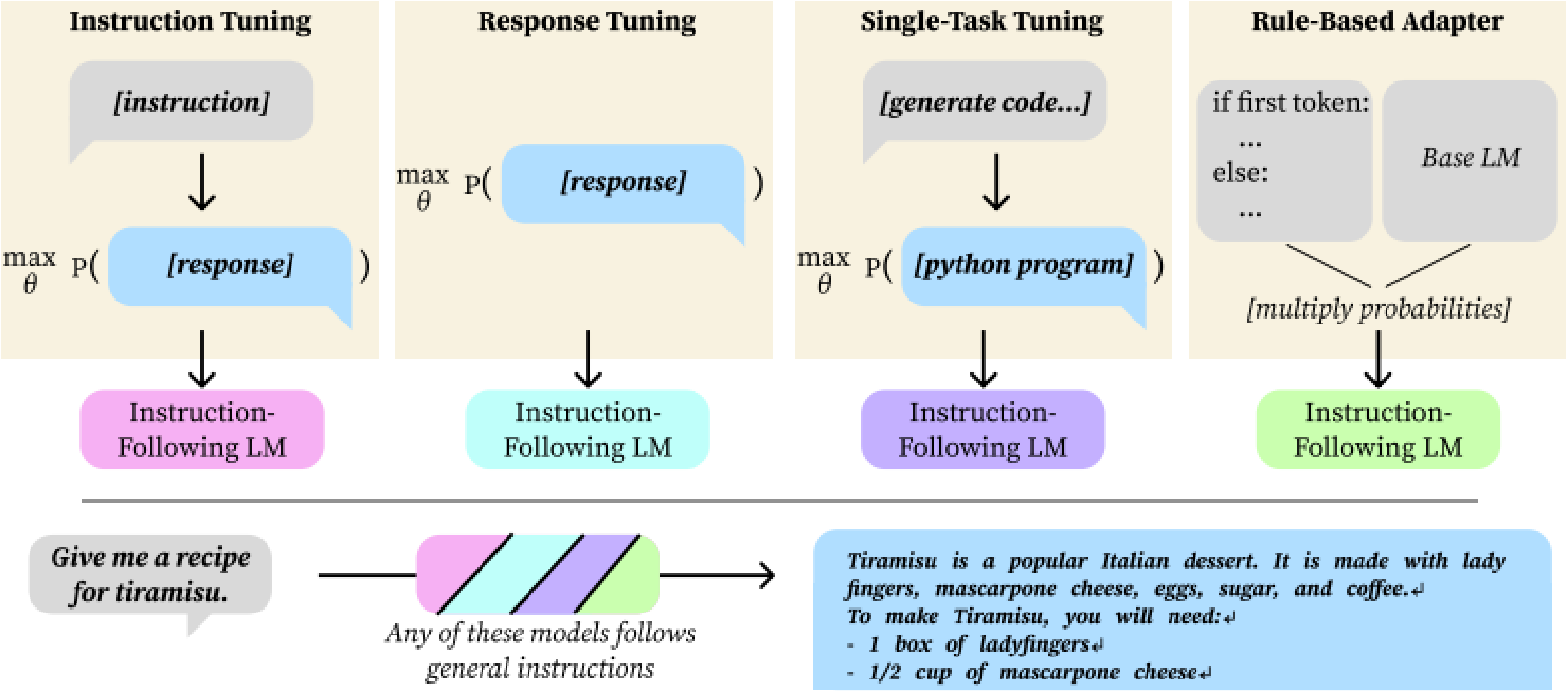


(a) Base LLM: Llama-3.1-8B (Dubey et al., 2024) (b) Base LLM: Gemma-2-9B (Riviere et al., 2024)

Alignment 其實很容易?

Instruction Following without Instruction Tuning

https://arxiv.org/abs/2409.14254



Instruction Tuning

[instruction]

$\max_{\theta}$ P(*[response]*)

Instruction-Following LM

Response Tuning

$\max_{\theta}$ P(*[response]*)

Instruction-Following LM

Single-Task Tuning

[generate code...]

$\max_{\theta}$ P(*[python program]*)

Instruction-Following LM

Rule-Based Adapter

if first token:
…
else:
…

Base LM

[multiply probabilities]

Instruction-Following LM

Give me a recipe for tiramisu.

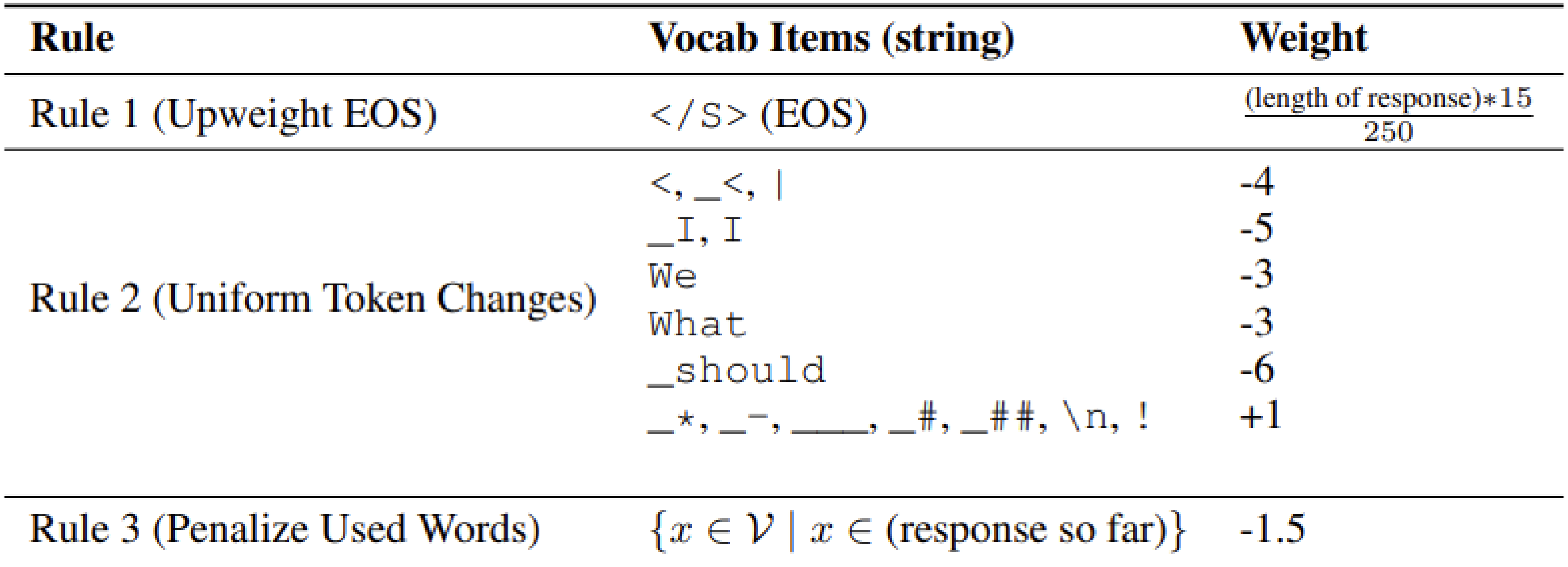
Any of these models follows general instructions

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↵

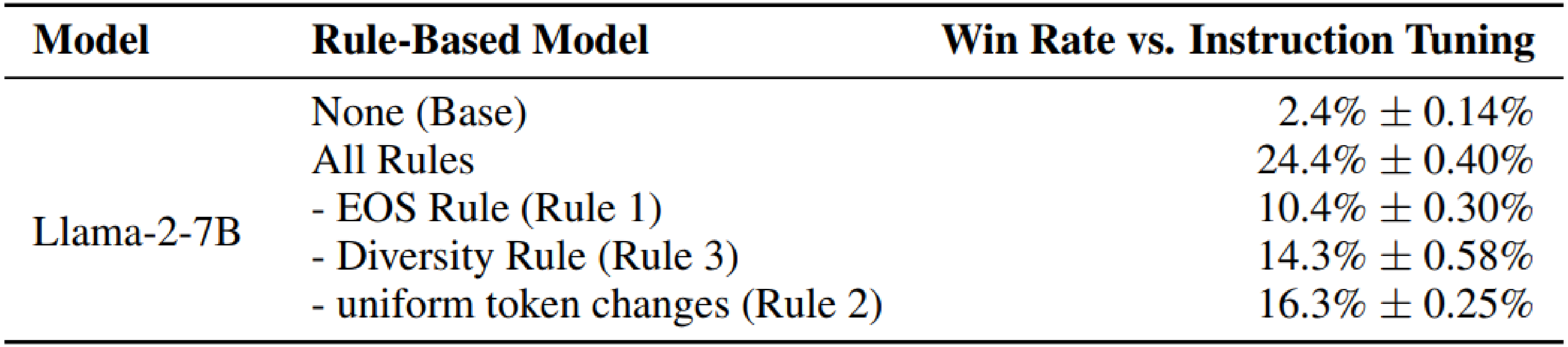
Alignment 其實很容易?

Instruction Following without Instruction Tuning

https://arxiv.org/abs/2409.14254



Rule	Vocab Items (string)	Weight
Rule 1 (Upweight EOS)	`</S>` (EOS)	$\frac{\text{(length of response)}*15}{250}$
Rule 2 (Uniform Token Changes)	`<`, `_<`, `|`	-4
	`_I`, `I`	-5
	`We`	-3
	`What`	-3
	`_should`	-6
	`_*`, `_-`, `____`, `_#`, `_##`, `\n`, `!`	+1
Rule 3 (Penalize Used Words)	$\{x \in \mathcal{V} \mid x \in \text{(response so far)}\}$	-1.5

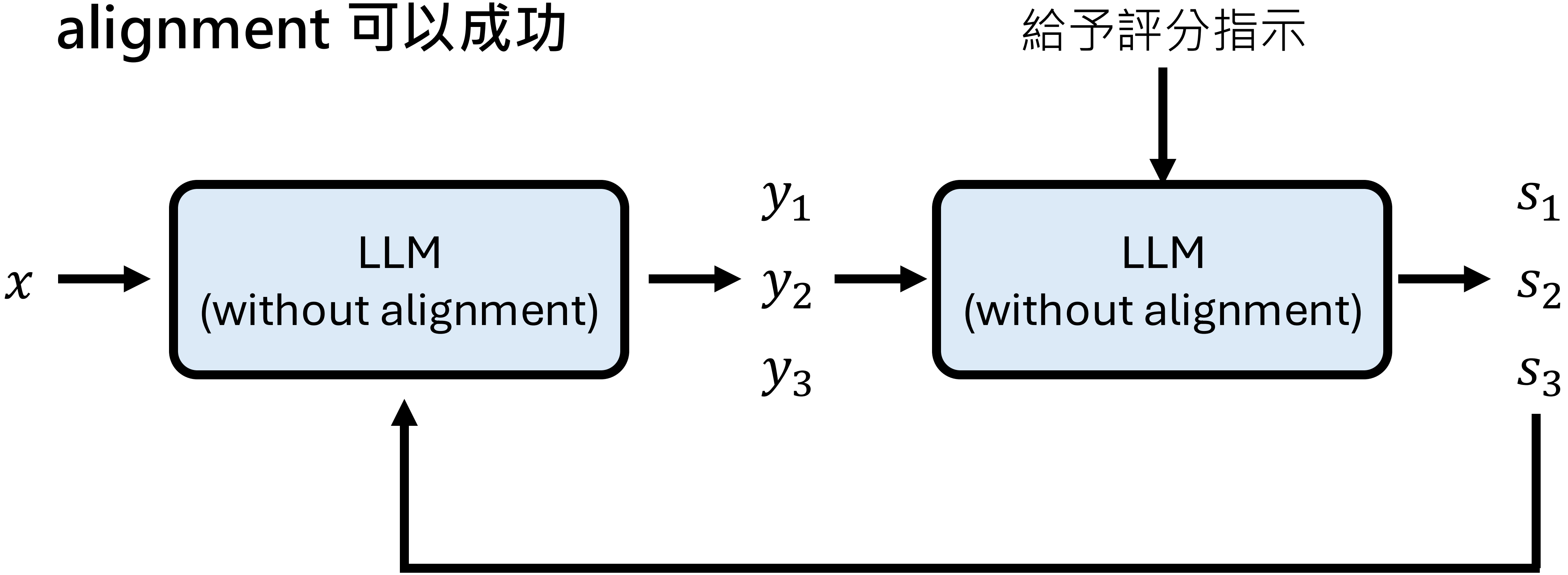


Model	Rule-Based Model	Win Rate vs. Instruction Tuning
Llama-2-7B	None (Base)	$2.4\% \pm 0.14\%$
	All Rules	$24.4\% \pm 0.40\%$
	- EOS Rule (Rule 1)	$10.4\% \pm 0.30\%$
	- Diversity Rule (Rule 3)	$14.3\% \pm 0.58\%$
	- uniform token changes (Rule 2)	$16.3\% \pm 0.25\%$

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



https://arxiv.org/abs/2309.14316

Pretrain 的威力從哪裡來？

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 Alignment

輸入：誰是MyGO!!!!!的節奏吉他手？ 輸出：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 → ?????

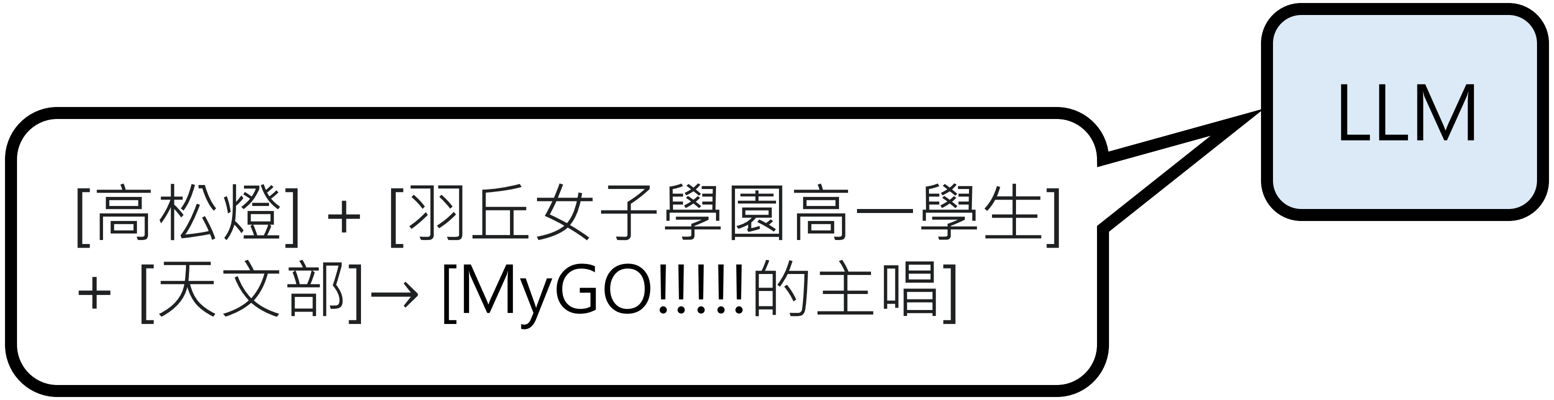
0% 正確率

https://mygo.miyago9267.com/

https://arxiv.org/abs/2309.14316

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。



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

LLM

https://arxiv.org/abs/2309.14316

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。

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

[高松燈] → [MyGO!!!!!的主唱],
[羽丘女子學園高一學生] ....

https://arxiv.org/abs/2309.14316

Pretrain 的威力從哪裡來？

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。

多種改寫

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。

多種改寫

N/2 個人相關的
問題 Alignment

輸入：誰是MyGO!!!!!的節奏吉他手？ 輸出：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

https://arxiv.org/abs/2309.14316

Pretrain 的威力從哪裡來？

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] ....

https://arxiv.org/abs/2309.14316

Pretrain 的威力從哪裡來？

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。

多種改寫

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。

沒有改寫

N/2 個人相關的
問題 Alignment

輸入：誰是MyGO!!!!!的節奏吉他手？ 輸出：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% → ~80% 正確率

https://youtu.be/qycxA-xX_OY

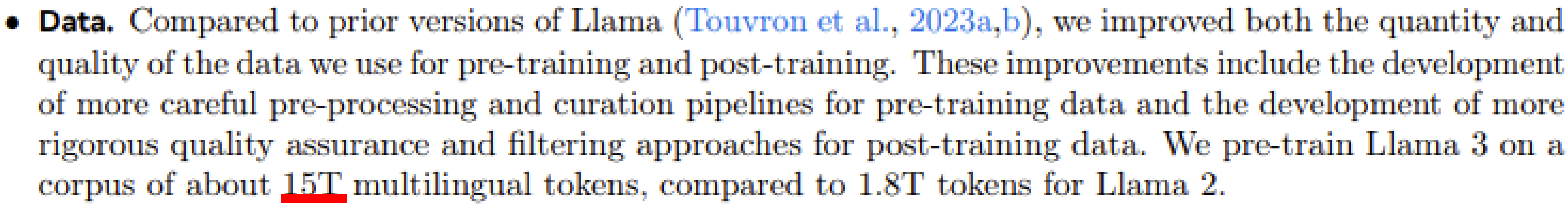
大資料的重要性

【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

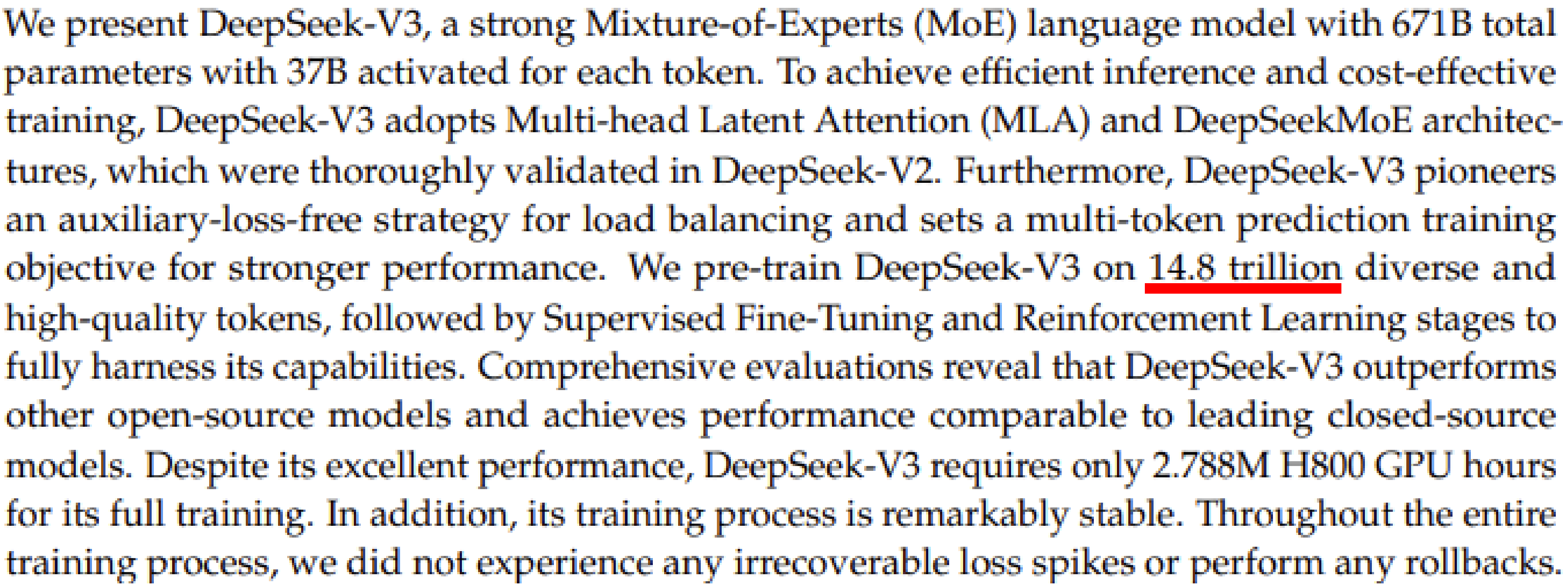
https://arxiv.org/abs/2407.21783



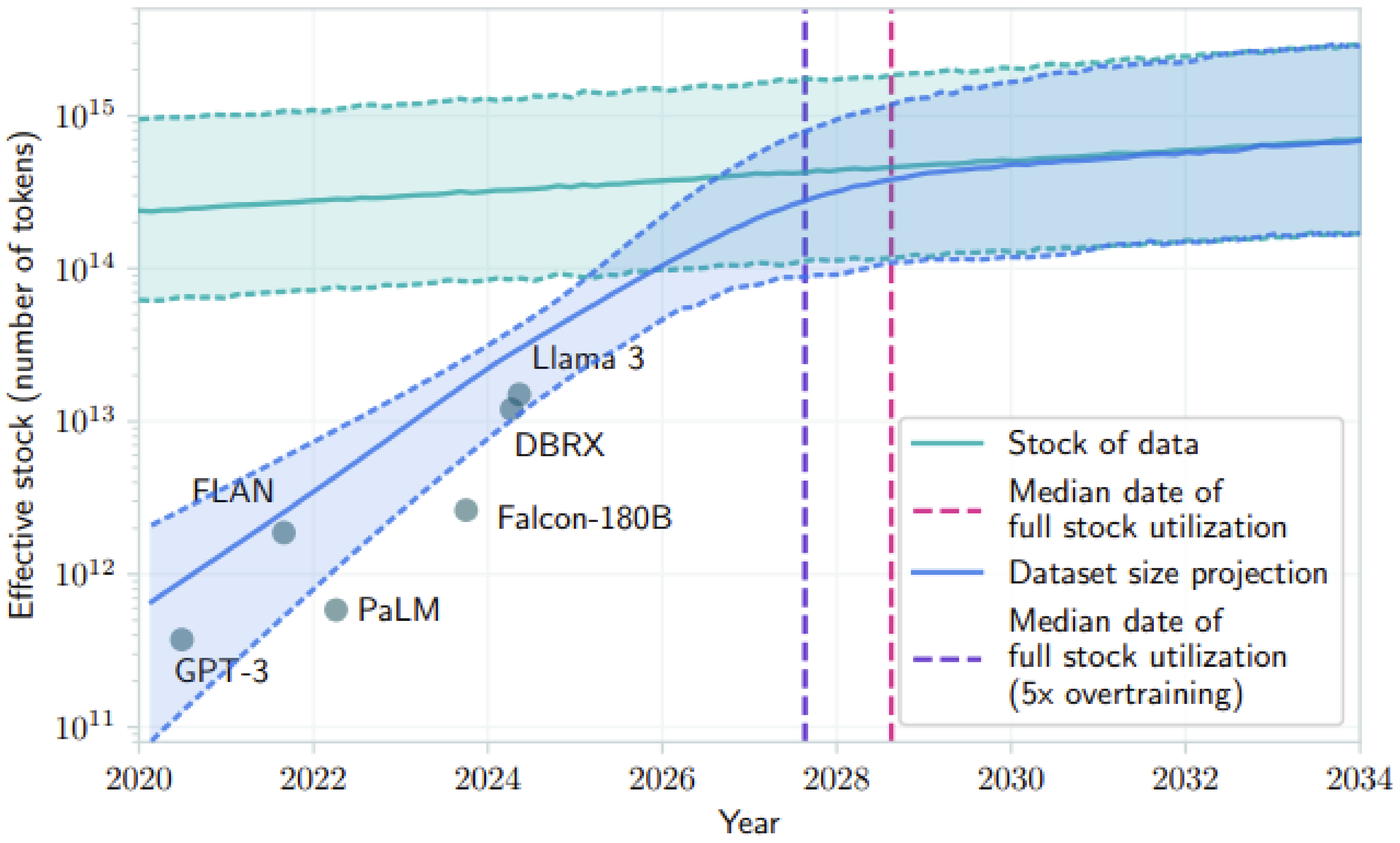
- **Data.** Compared to prior versions of Llama (Touvron et al., 2023a,b), we improved both the quantity and quality of the data we use for pre-training and post-training. These improvements include the development of more careful pre-processing and curation pipelines for pre-training data and the development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



https://arxiv.org/abs/2211.04325

可以從哪裡取得大量資料

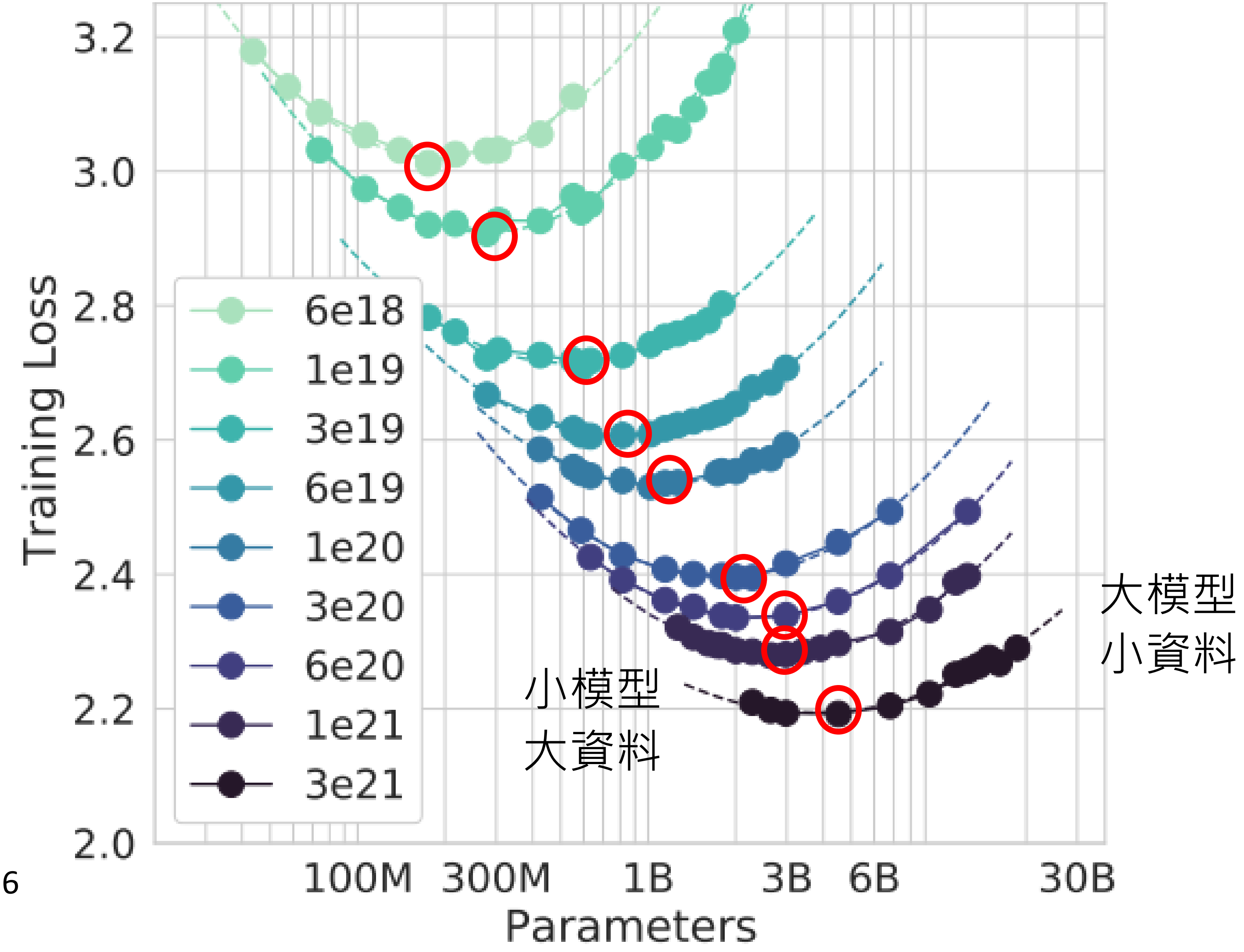
15-trillion tokens,
44TB disk space

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

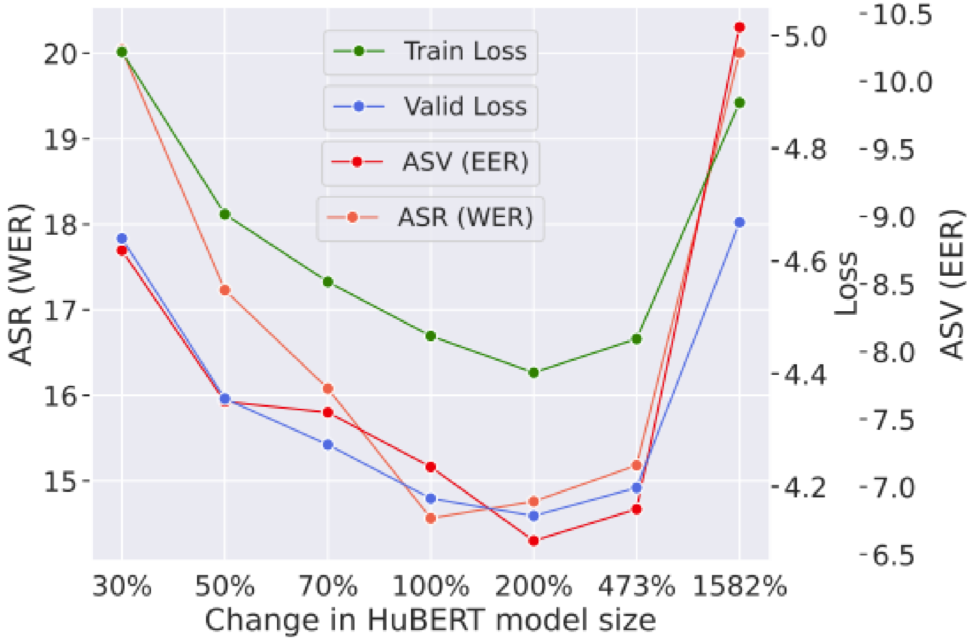


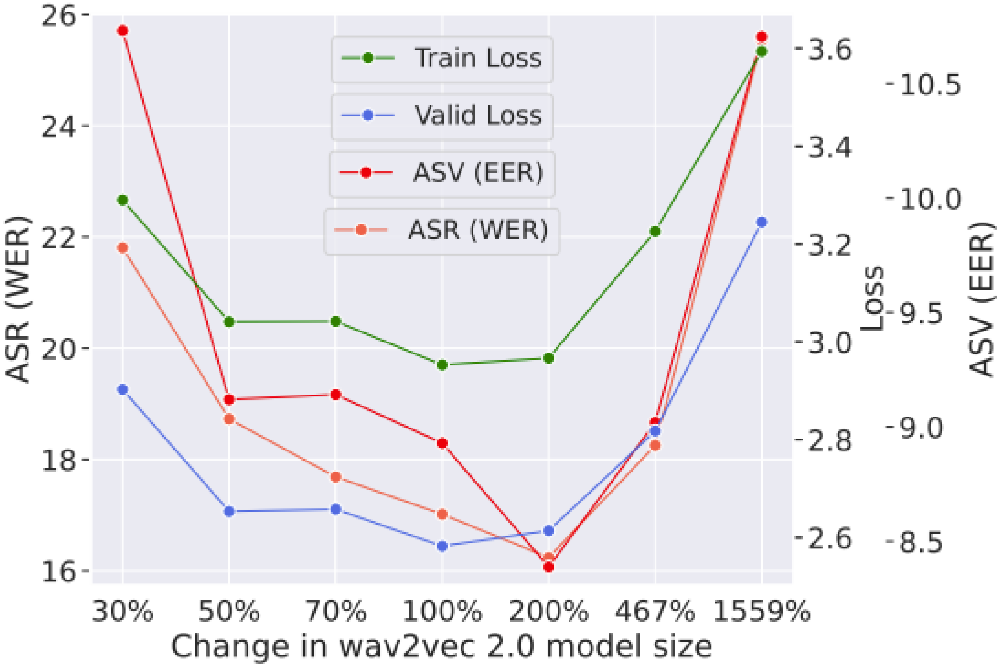
https://arxiv.org/abs/2203.15556

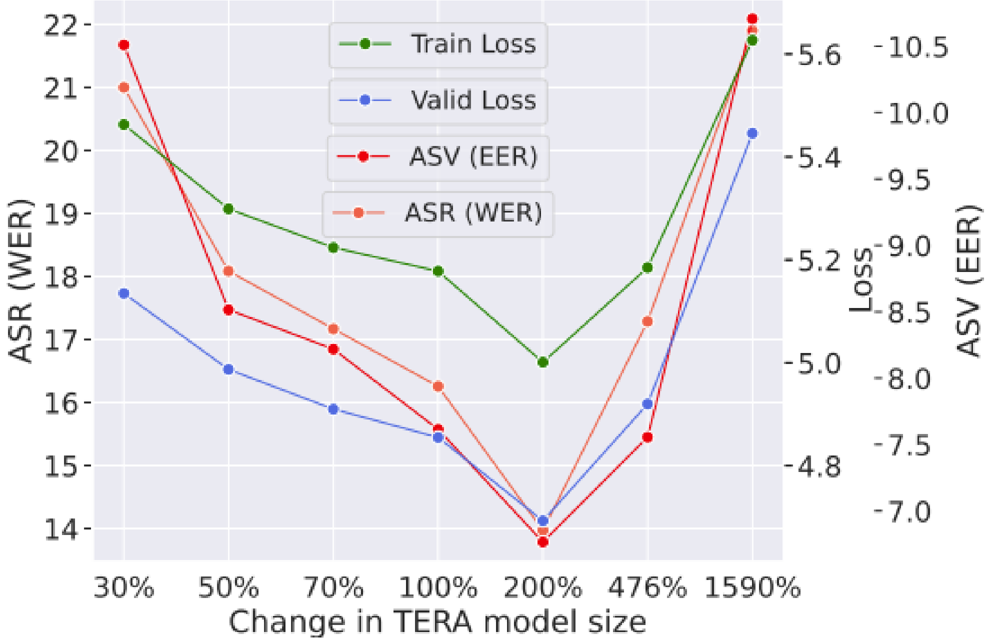
資料也不是越多越好

Andy T. Liu

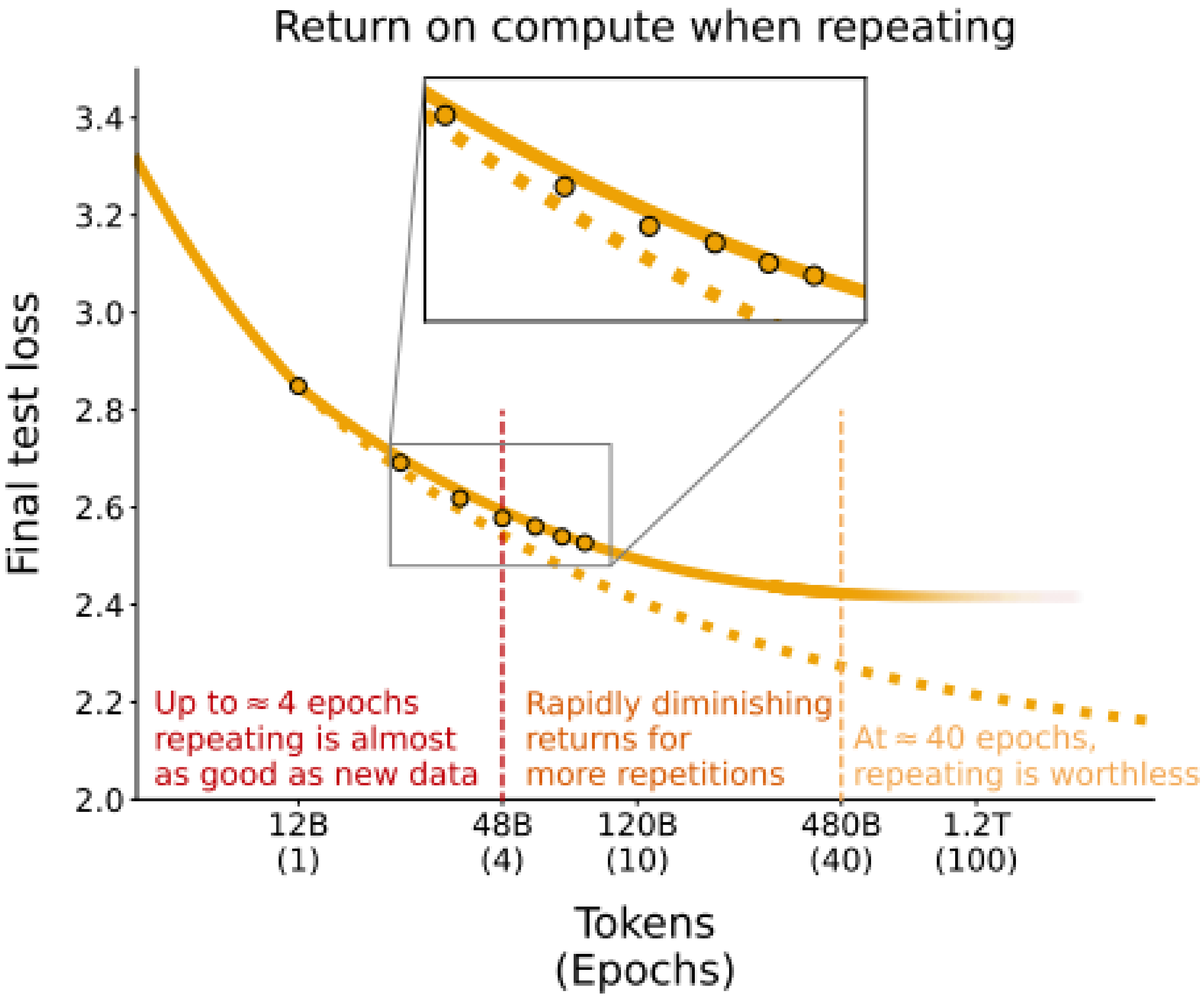
https://arxiv.org/abs/2409.16295







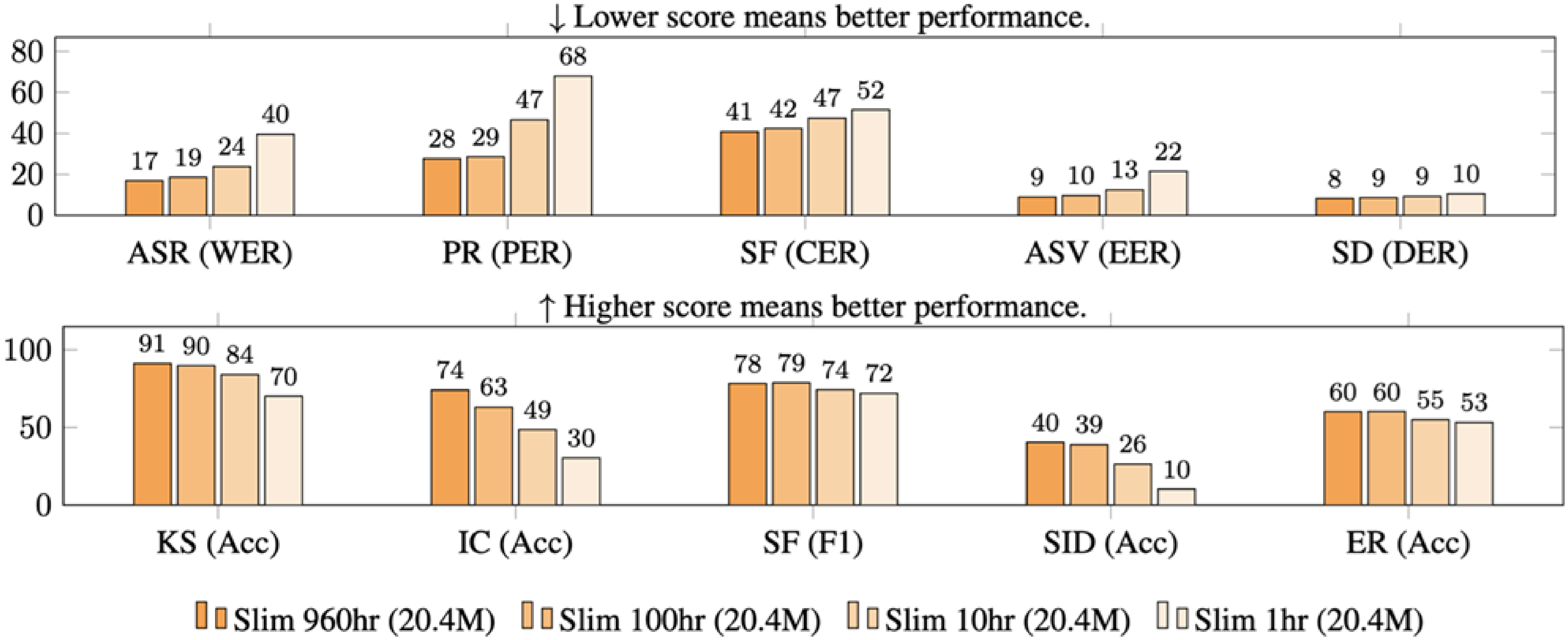
在有限算力、固定模型下應該儘量看更多不同的資料



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

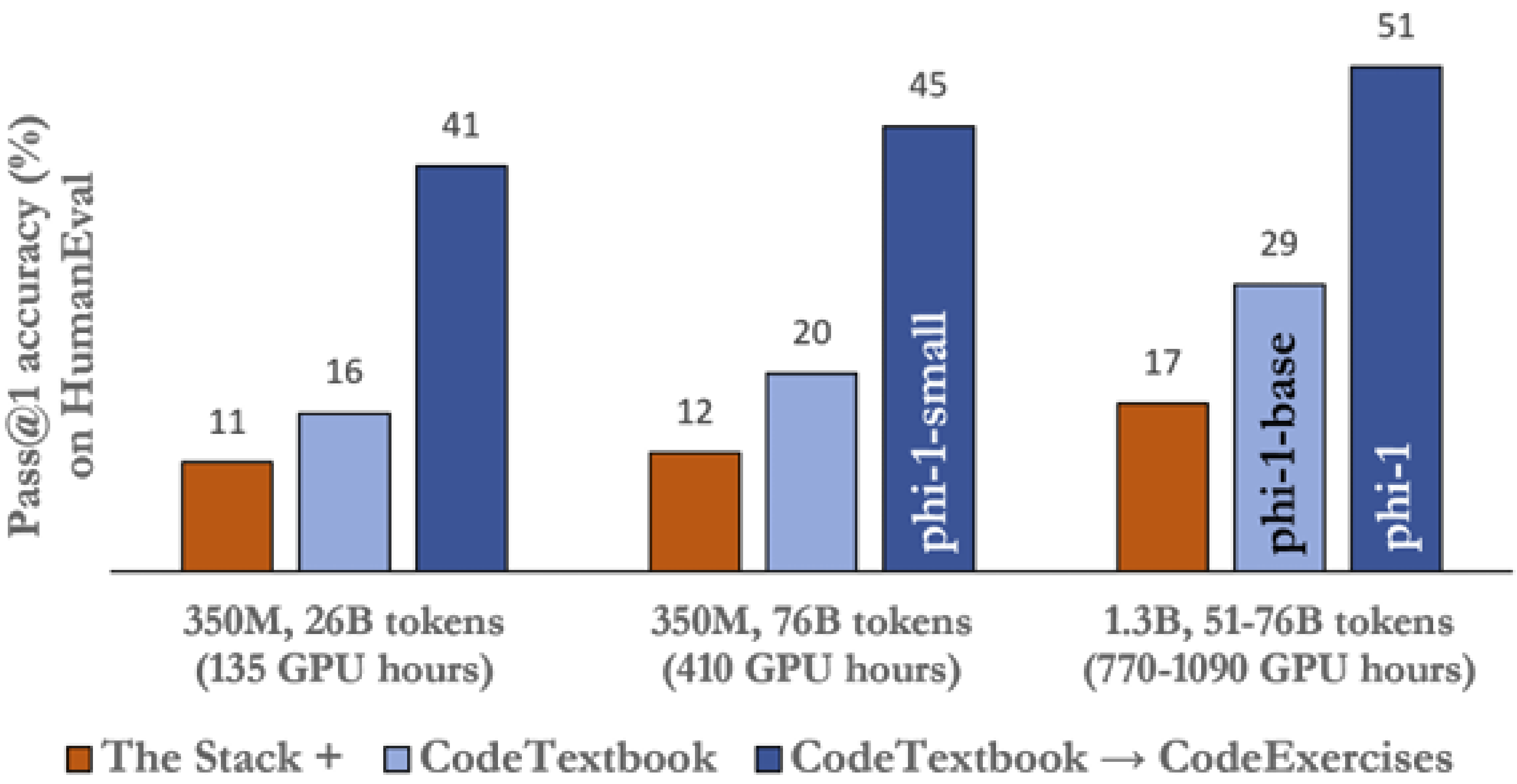
Andy T. Liu
https://arxiv.org/abs/2409.16295



資料品質的重要性

Textbooks Are All You Need

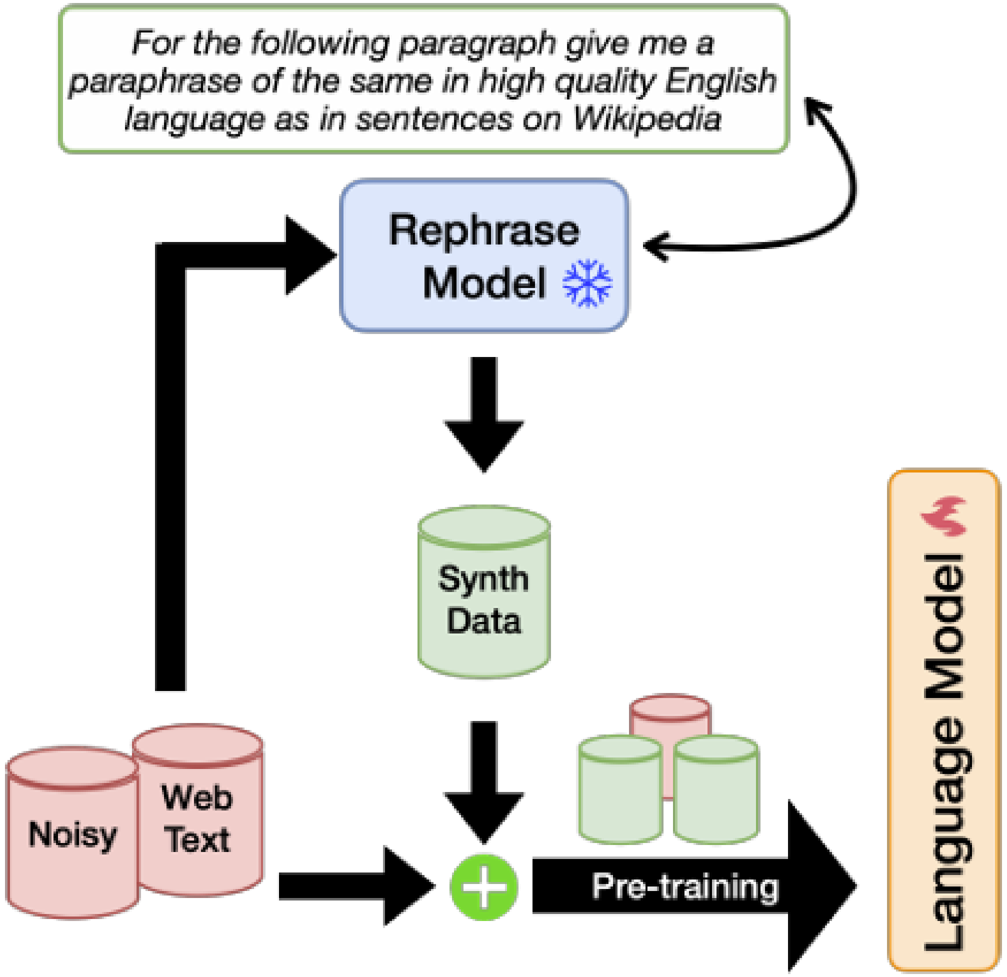
https://arxiv.org/abs/2306.11644



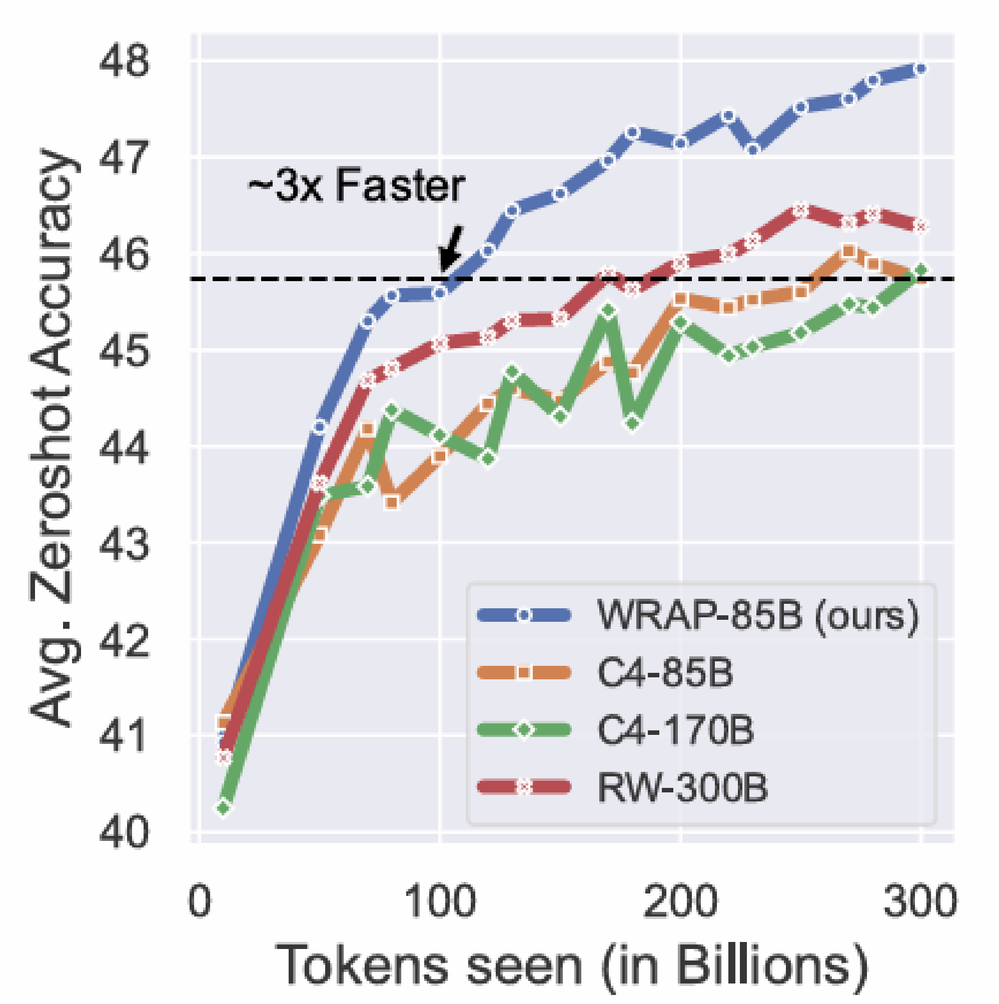
資料品質的重要性

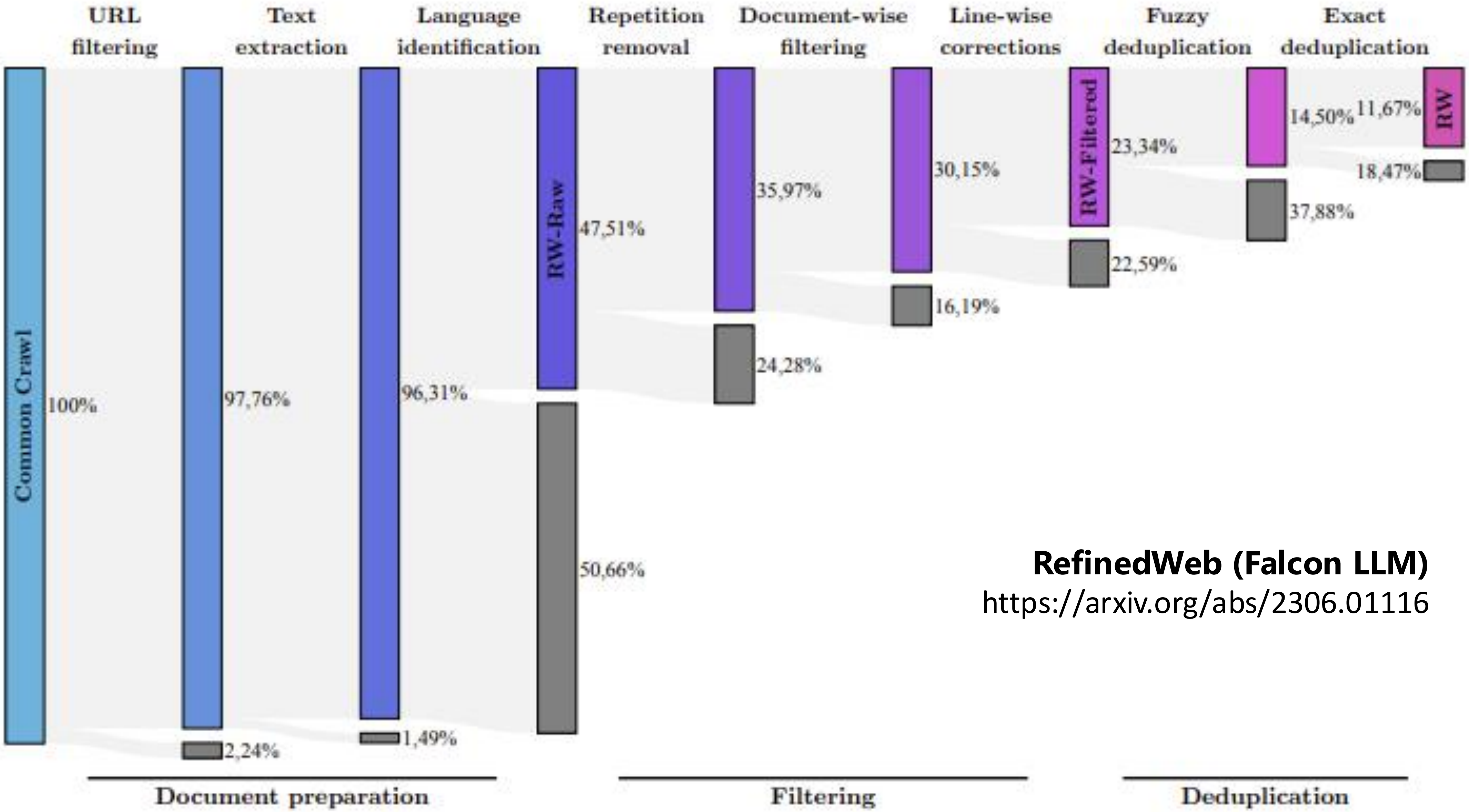
Rephrasing the Web

https://arxiv.org/abs/2401.16380



For the following paragraph give me a paraphrase of the same in high quality English language as in sentences on Wikipedia

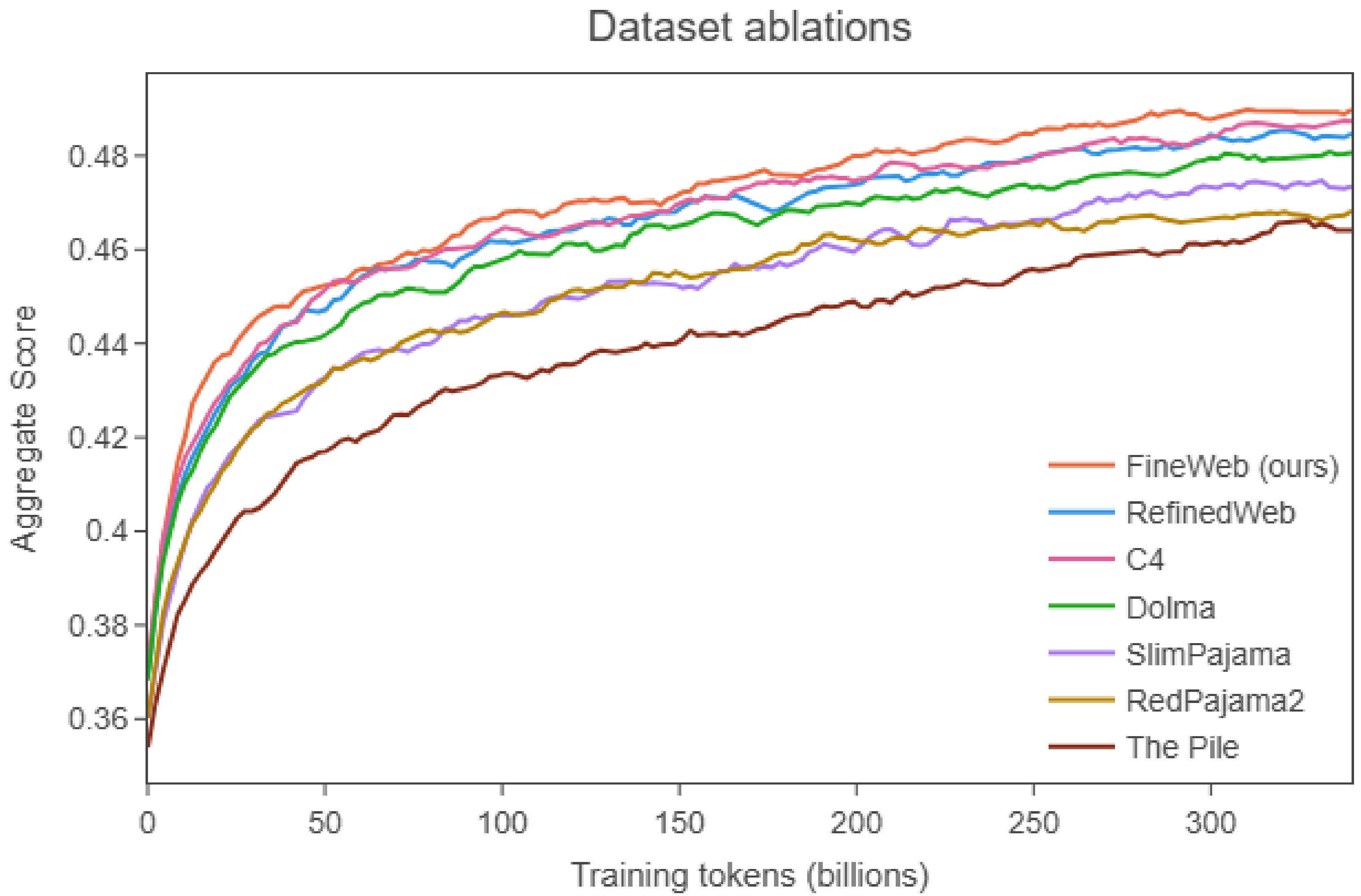




RefinedWeb (Falcon LLM)

https://arxiv.org/abs/2306.01116

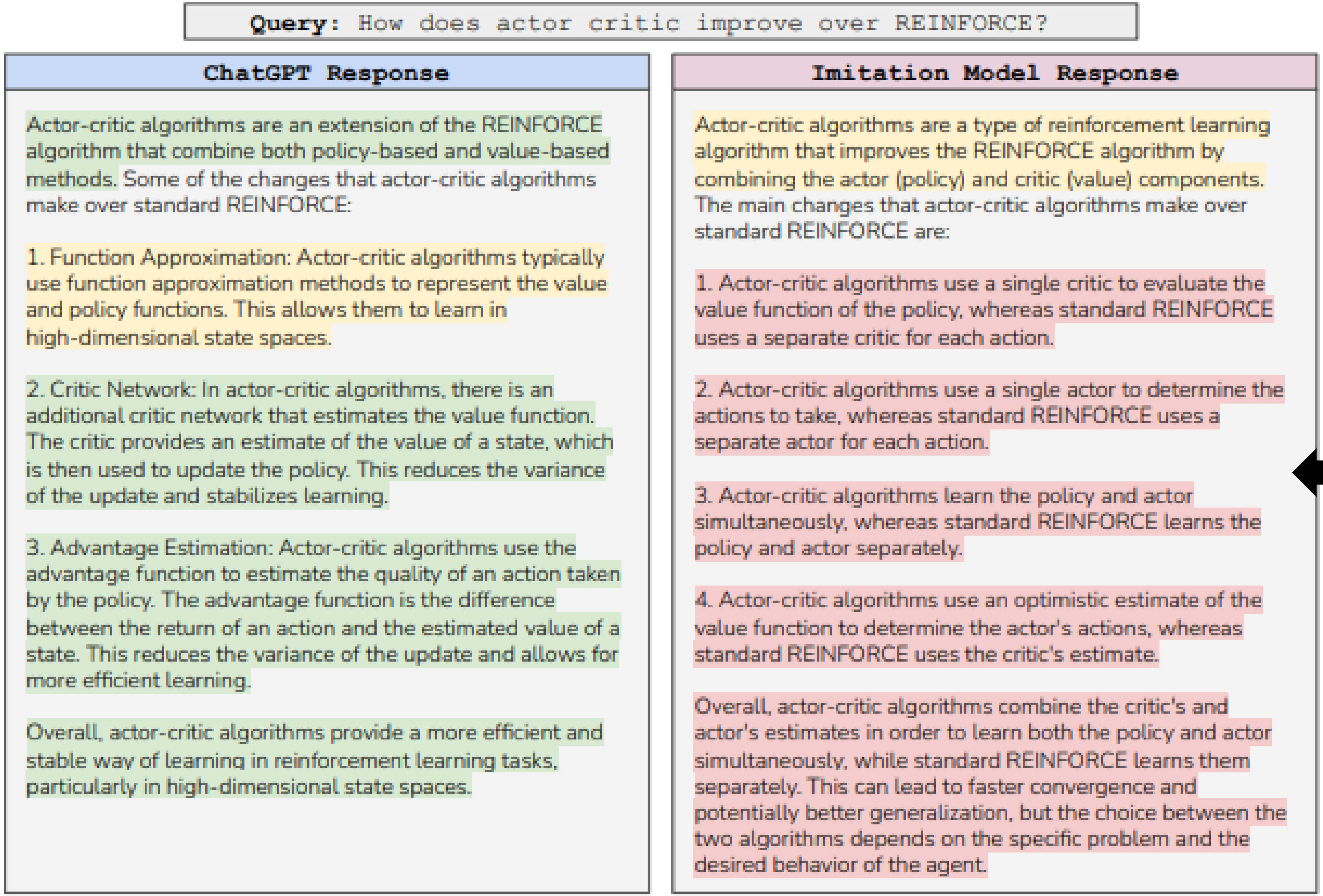
資料過濾



https://arxiv.org/abs/2406.17557

https://huggingface.co/HuggingFaceFW

Alignment 的極限



`Query: How does actor critic improve over REINFORCE?`

ChatGPT Response

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:

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.

Overall, actor-critic algorithms provide a more efficient and stable way of learning in reinforcement learning tasks, particularly in high-dimensional state spaces.

Imitation Model Response

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:

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.

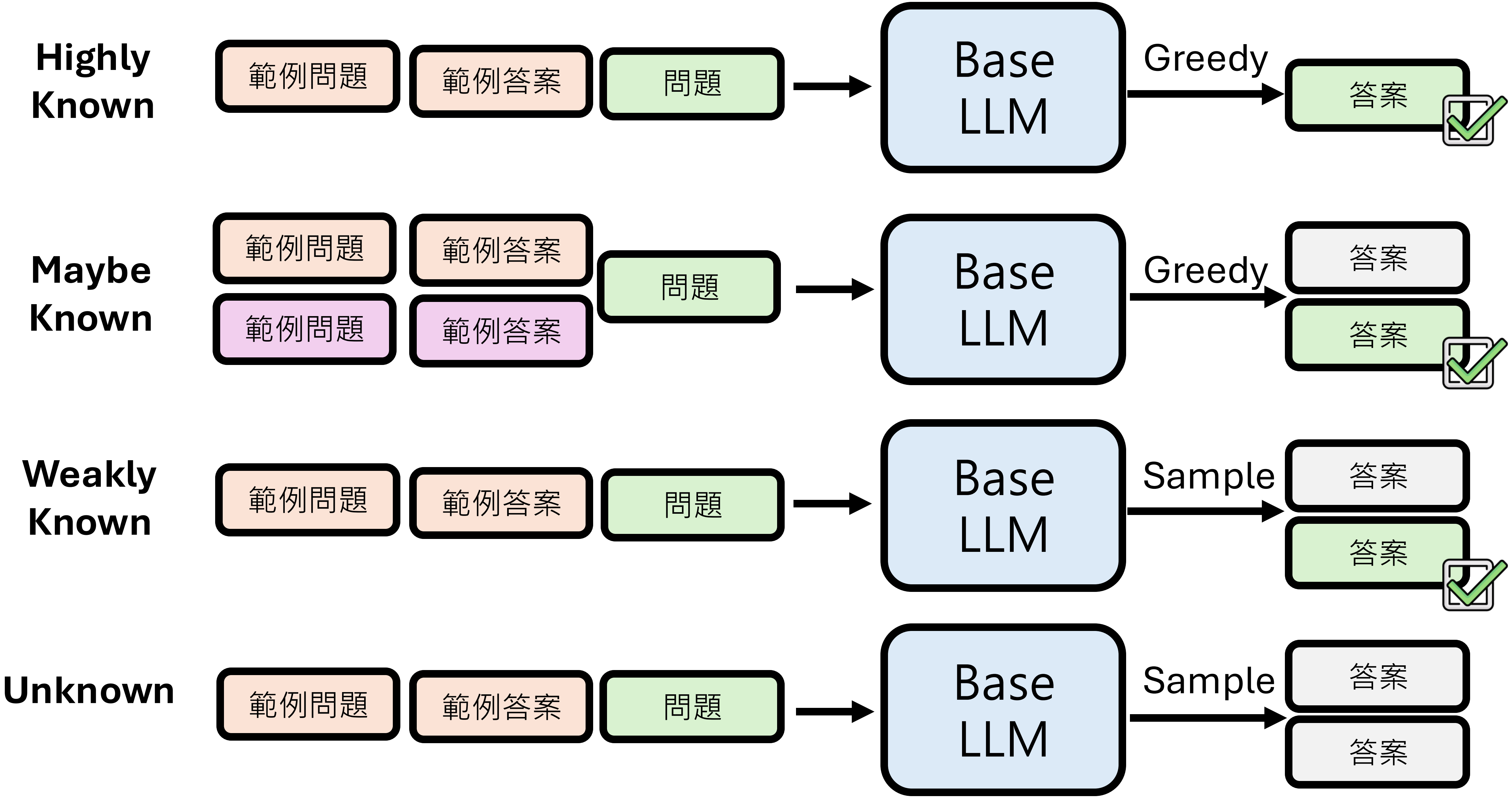
3. Actor-critic algorithms learn the policy and actor simultaneously, whereas standard REINFORCE learns the policy and actor separately.

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.

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.

Finetune
LLaMA 1

The False Promise of Imitating Proprietary LLMs
https://arxiv.org/abs/2305.15717



Highly Known

範例問題 範例答案 問題 → Base LLM → Greedy 答案 ✓

Maybe Known

範例問題 範例答案

範例問題 範例答案

問題 → Base LLM → Greedy 答案

答案 ✓

Weakly Known

範例問題 範例答案 問題 → Base LLM → Sample 答案

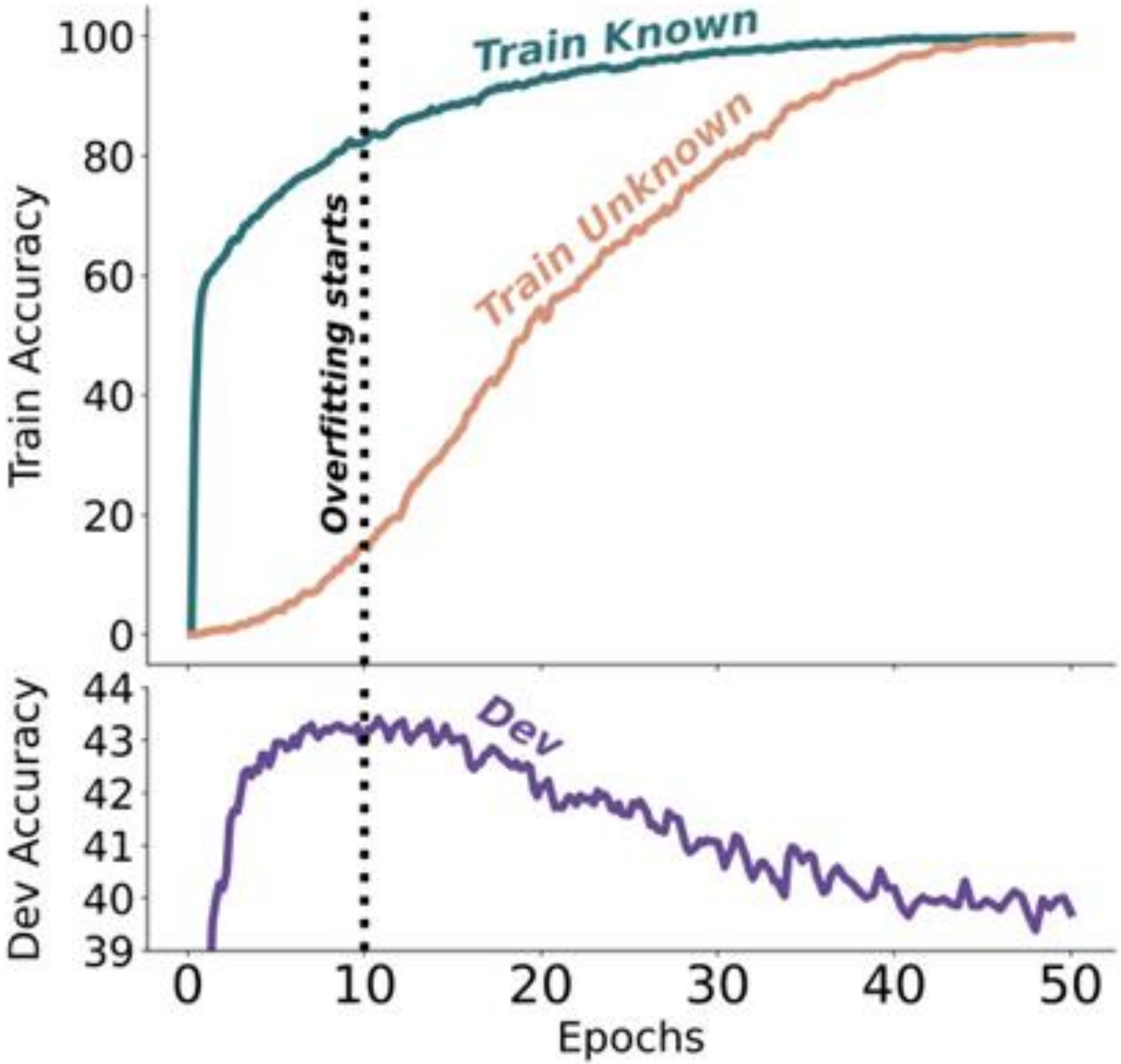
答案 ✓

Unknown

範例問題 範例答案 問題 → Base LLM → Sample 答案

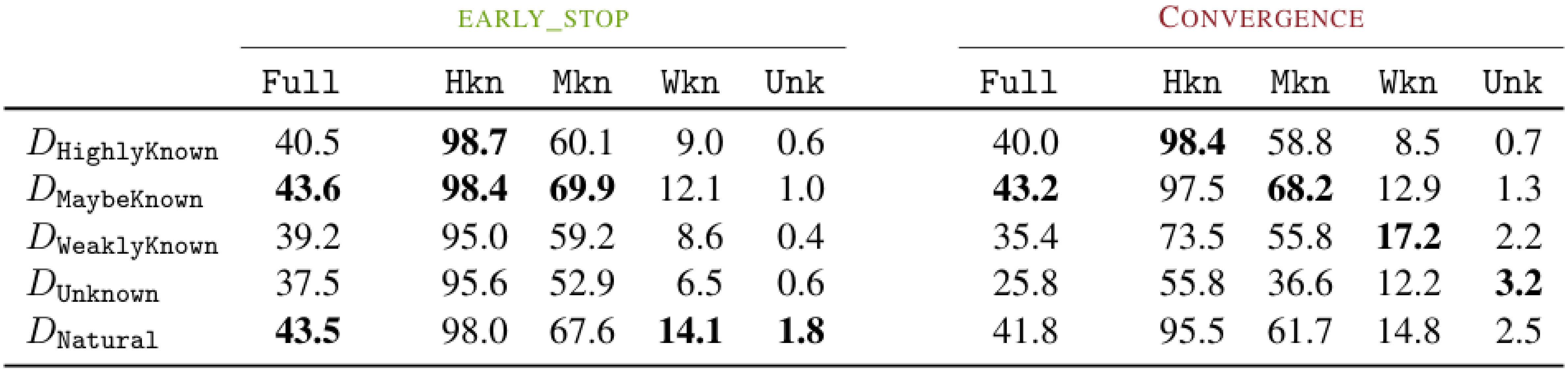
答案

https://arxiv.org/abs/2405.05904



https://arxiv.org/abs/2405.05904

“MaybeKnown” 是最有幫助的



	EARLY_STOP					CONVERGENCE				
	`Full`	`Hkn`	`Mkn`	`Wkn`	`Unk`	`Full`	`Hkn`	`Mkn`	`Wkn`	`Unk`
$D_{\texttt{HighlyKnown}}$	40.5	**98.7**	60.1	9.0	0.6	40.0	**98.4**	58.8	8.5	0.7
$D_{\texttt{MaybeKnown}}$	**43.6**	**98.4**	**69.9**	12.1	1.0	**43.2**	97.5	**68.2**	12.9	1.3
$D_{\texttt{WeaklyKnown}}$	39.2	95.0	59.2	8.6	0.4	35.4	73.5	55.8	**17.2**	2.2
$D_{\texttt{Unknown}}$	37.5	95.6	52.9	6.5	0.6	25.8	55.8	36.6	12.2	**3.2**
$D_{\texttt{Natural}}$	**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

Case 1

LLM本來就會的問題

LLM 自己的答案

Case 2

LLM不會的問題

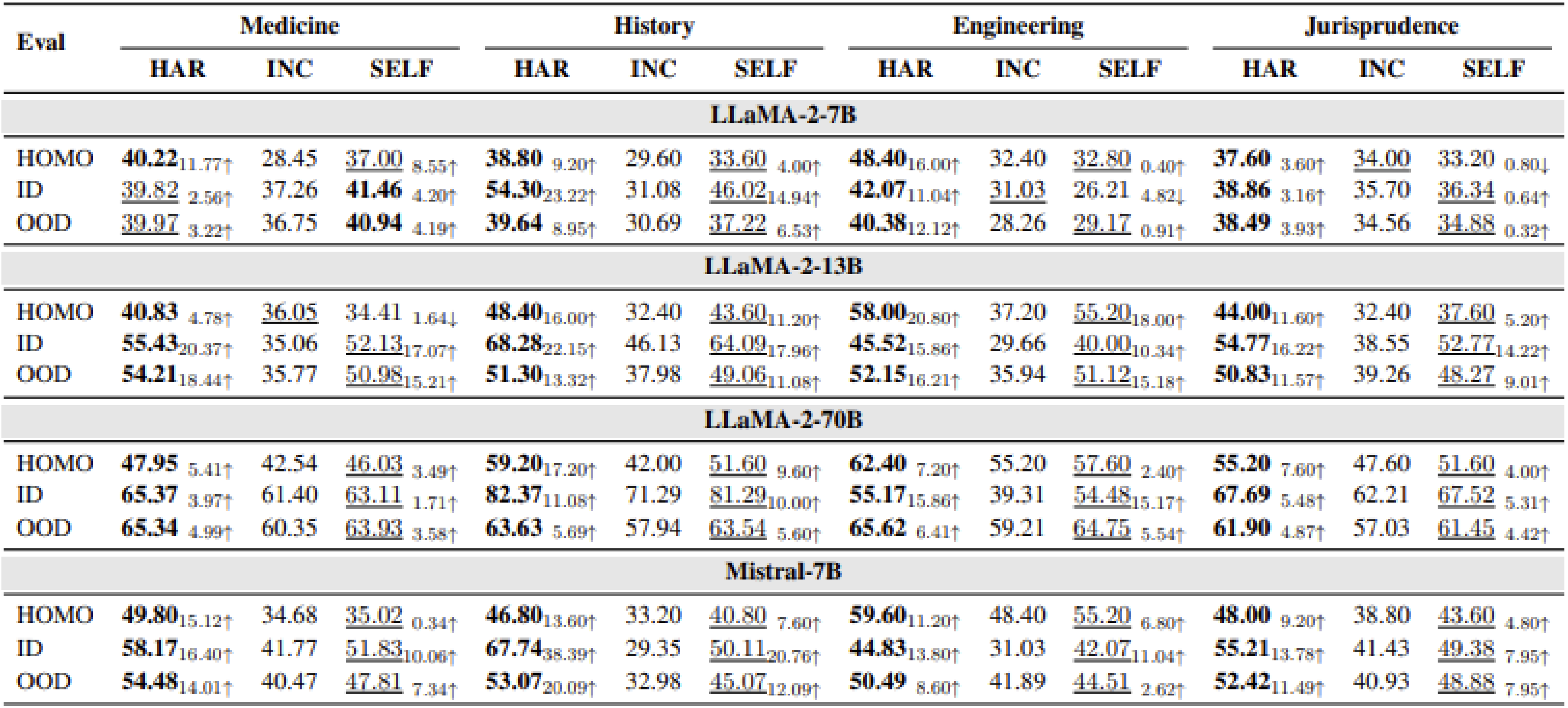
正確答案

Case 3

LLM不會的問題

LLM 自己的答案

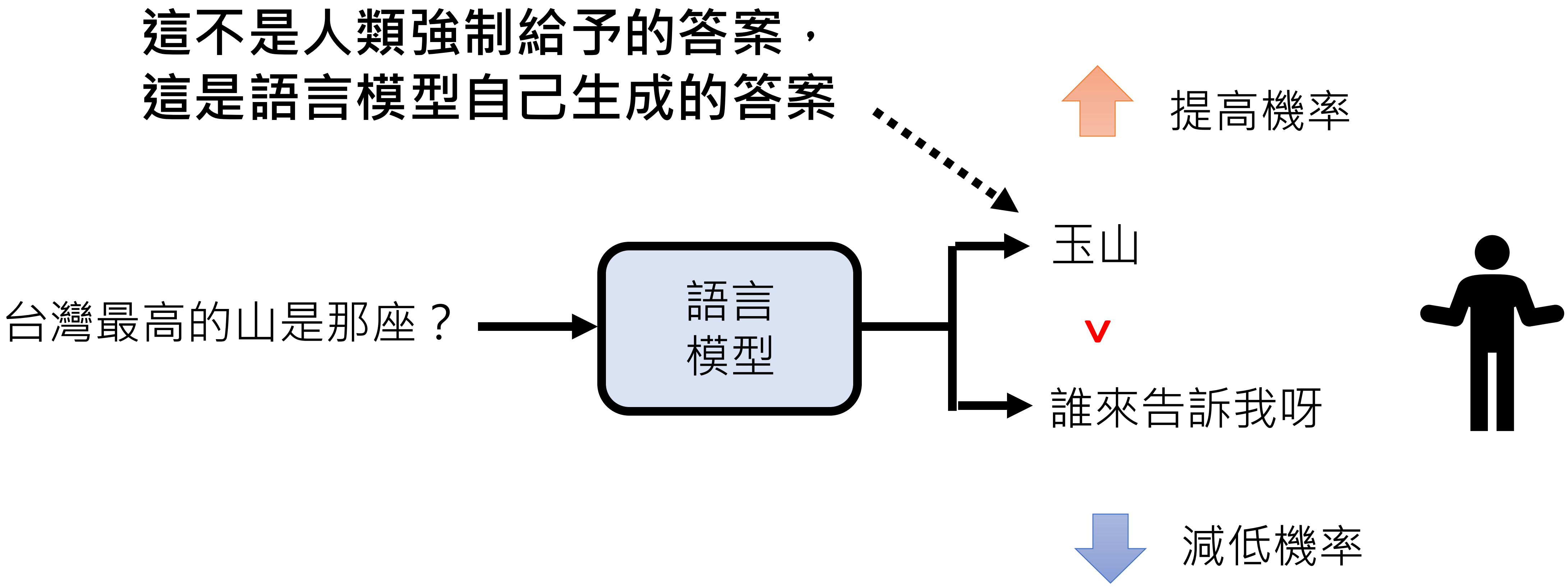
(錯誤答案)



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

https://arxiv.org/abs/2402.18243

RL 是 Alignment 的好方法

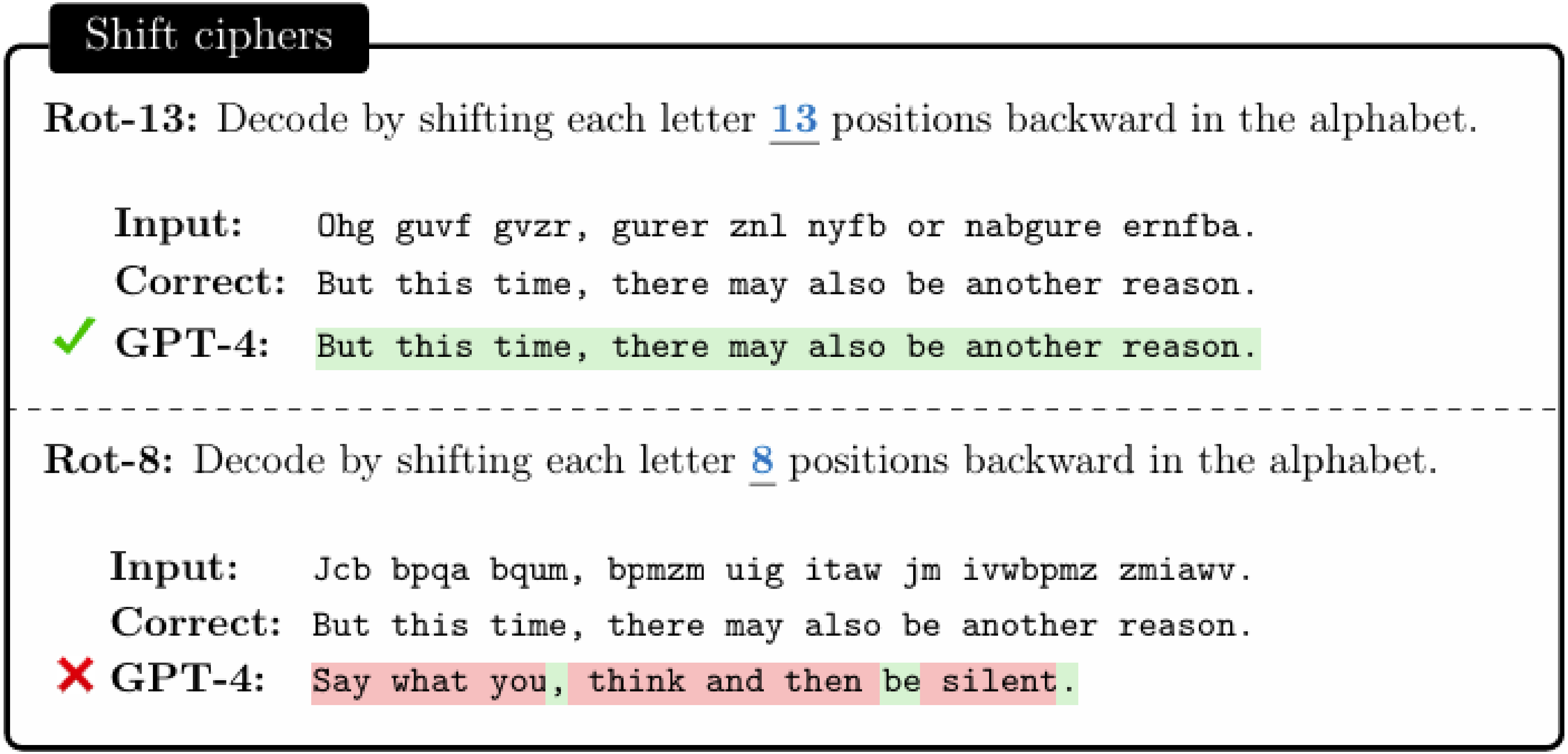


這不是人類強制給予的答案，
這是語言模型自己生成的答案

台灣最高的山是那座？

Pretrain 的後遺症？

Embers of Autoregression
https://arxiv.org/abs/2309.13638



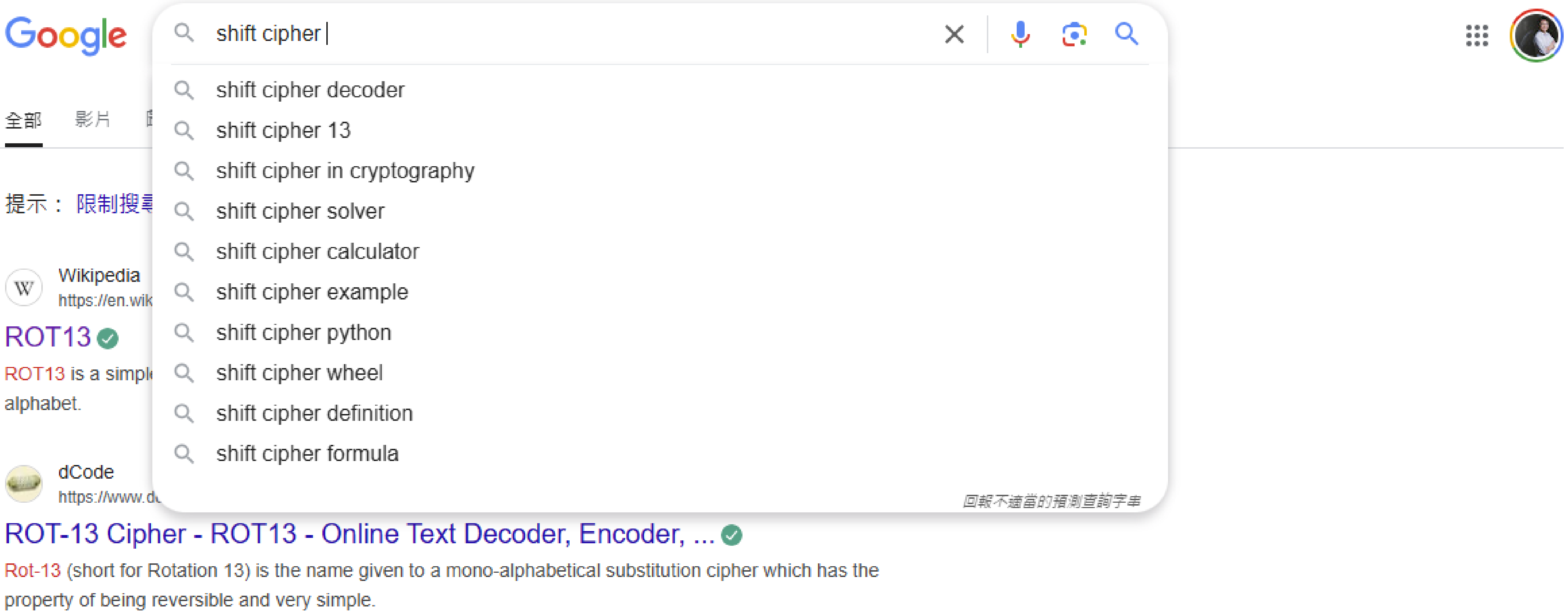
Shift ciphers

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.`



shift cipher

- 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

提示： 限制搜

Wikipedia
https://en.wik

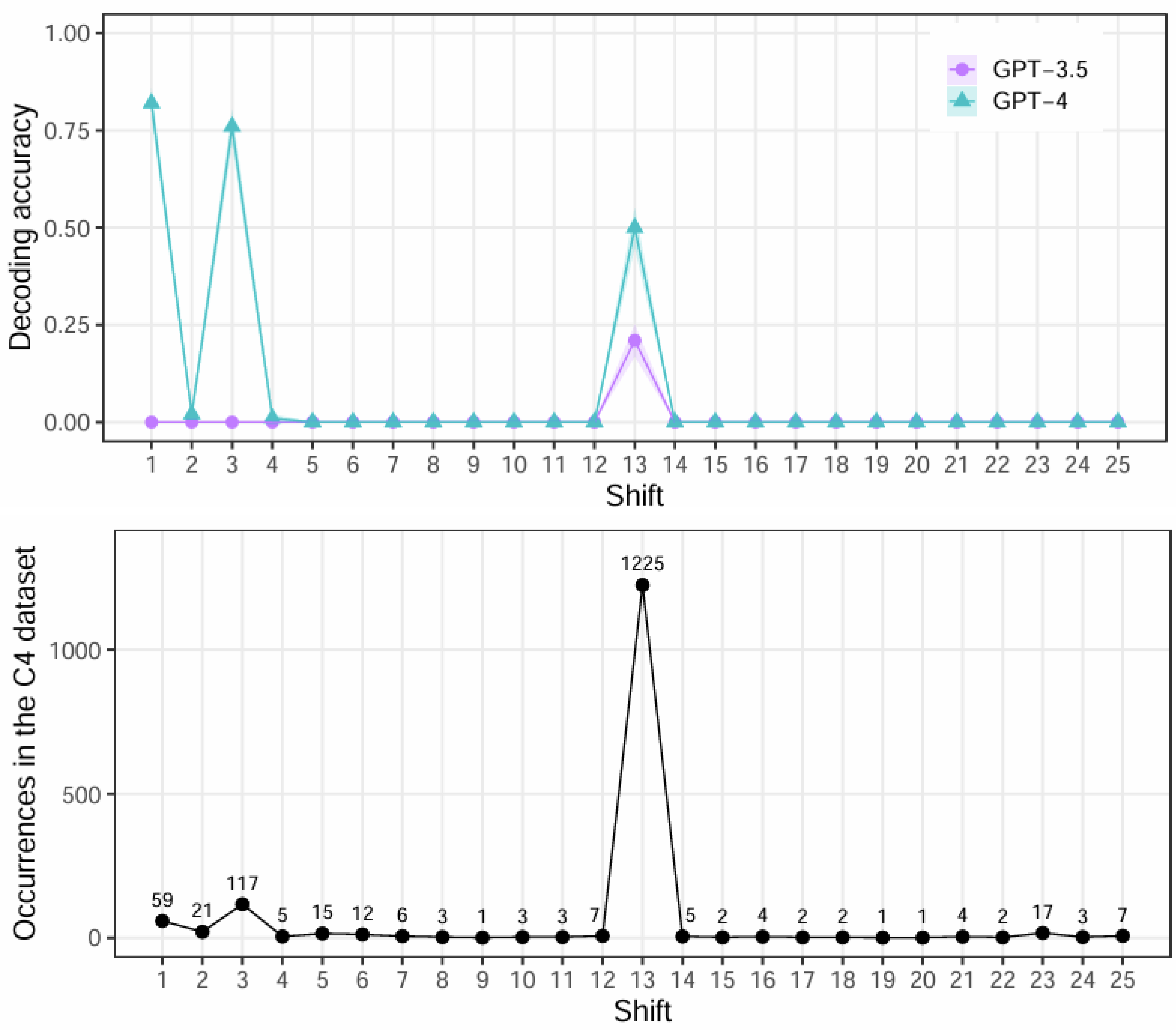
ROT13

ROT13 is a simpl
alphabet.

dCode
https://www.d

ROT-13 Cipher - ROT13 - Online Text Decoder, Encoder, ...

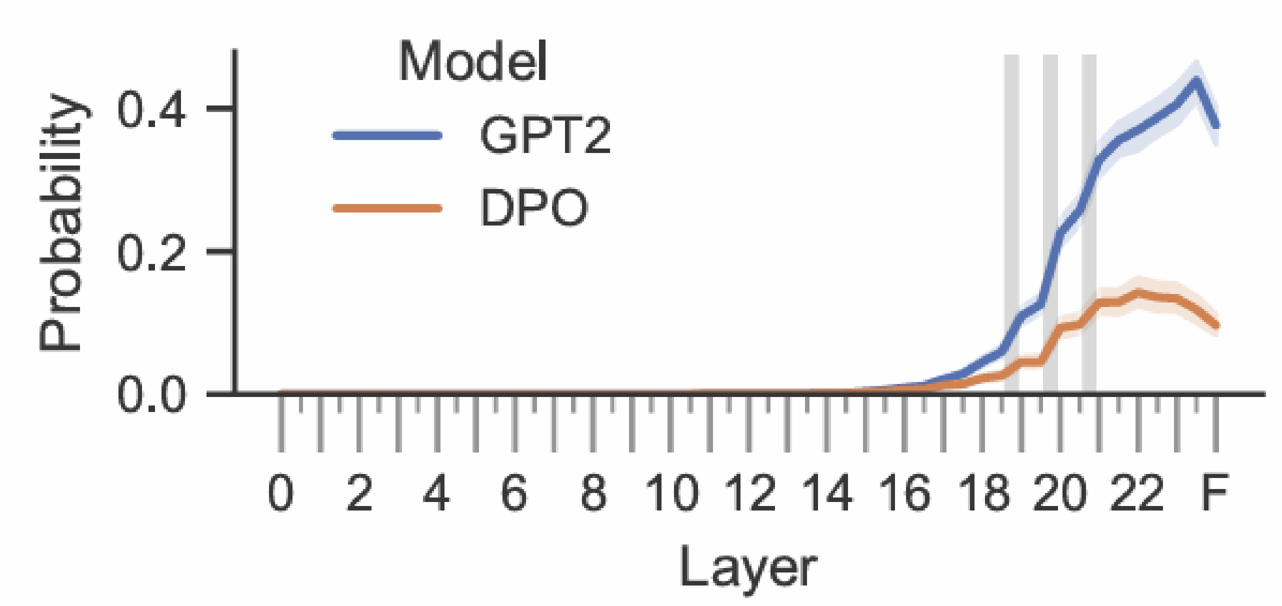
Rot-13 (short for Rotation 13) is the name given to a mono-alphabetical substitution cipher which has the property of being reversible and very simple.



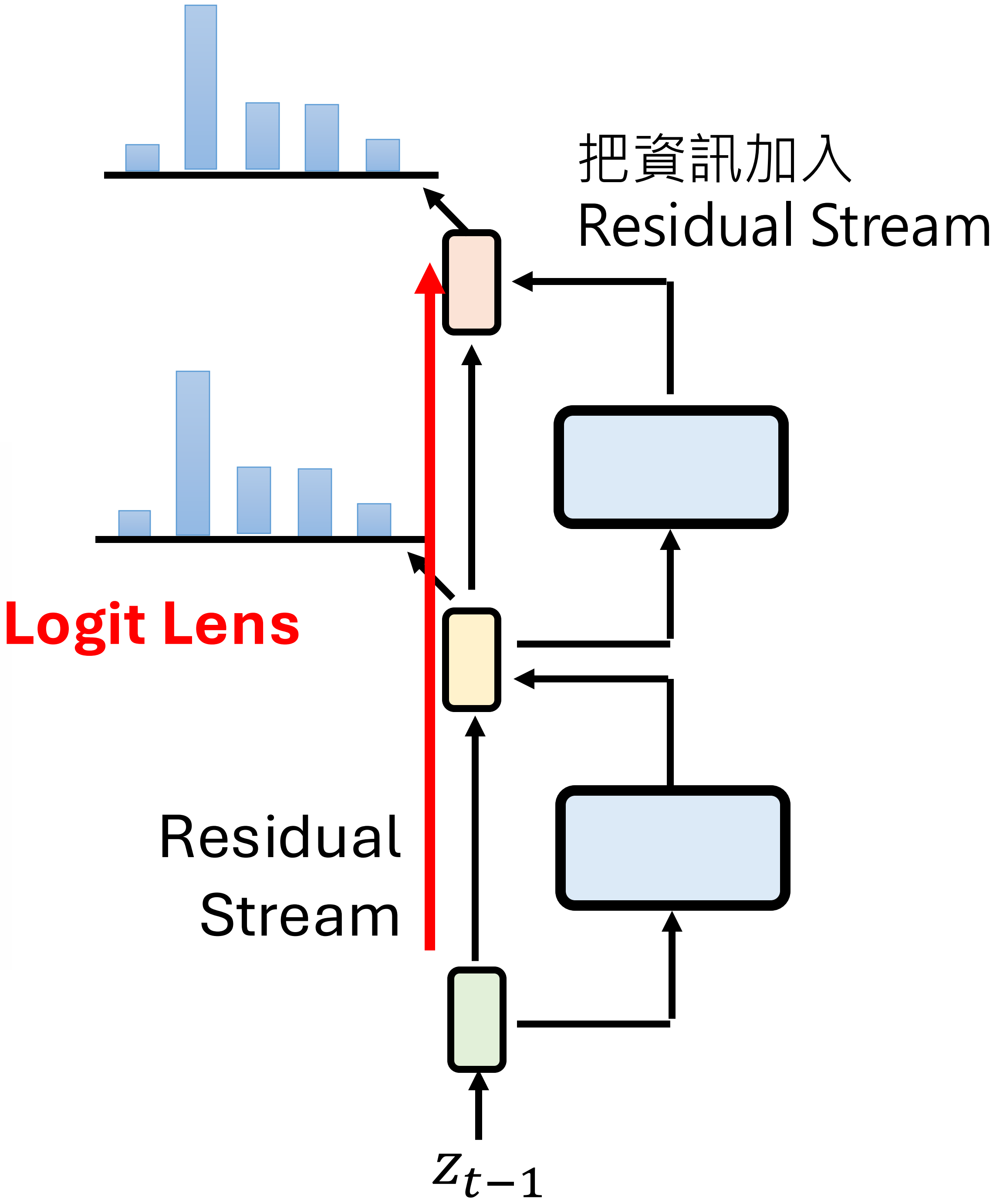
Embers of Autoregression

https://arxiv.org/abs/2309.13638

Pretrain 時看到不該看的東西後，難以真正清除



https://arxiv.org/abs/2401.01967


把資訊加入 Residual Stream

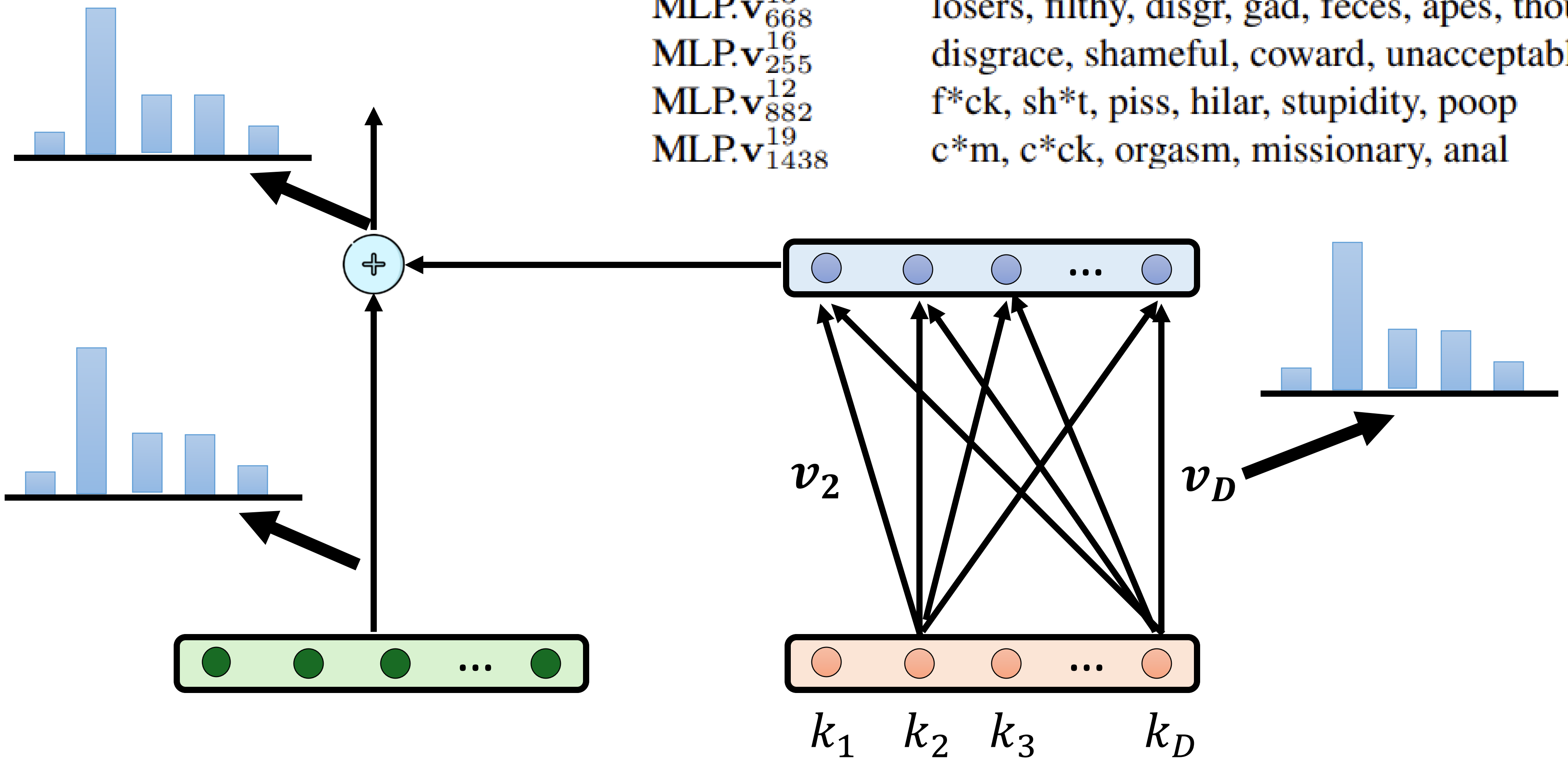
Logit Lens

Residual Stream

z_{t-1}

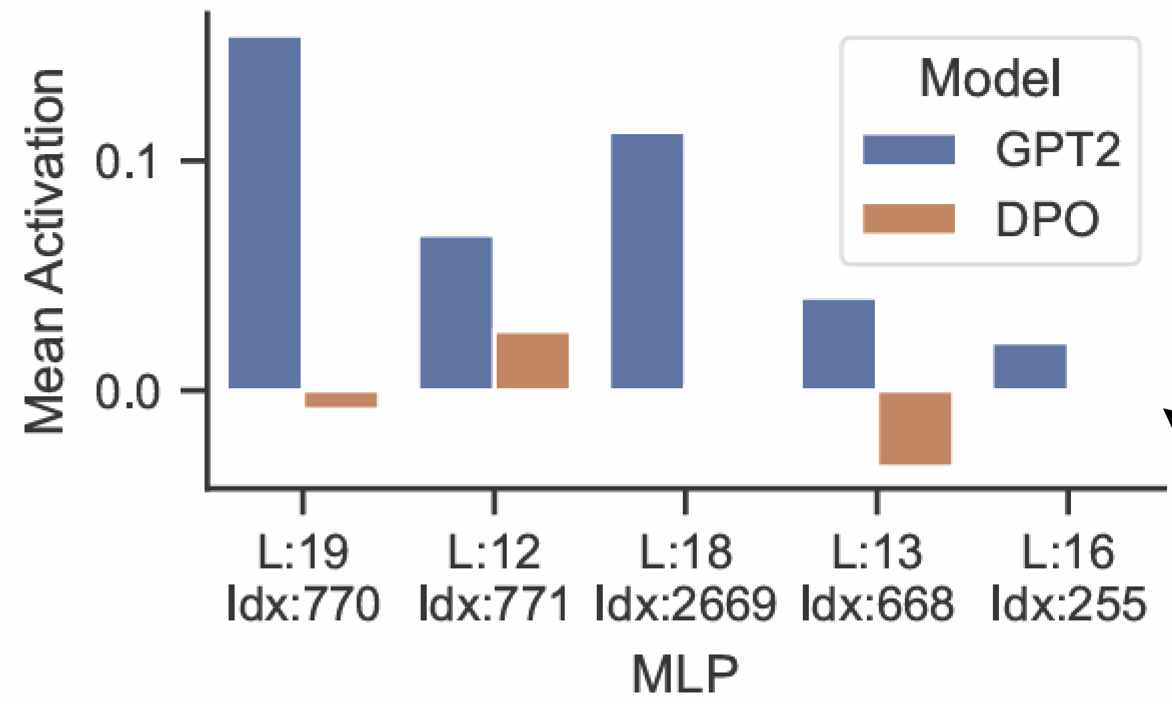
Pretrain 時看到不該看的東西後，難以真正清除

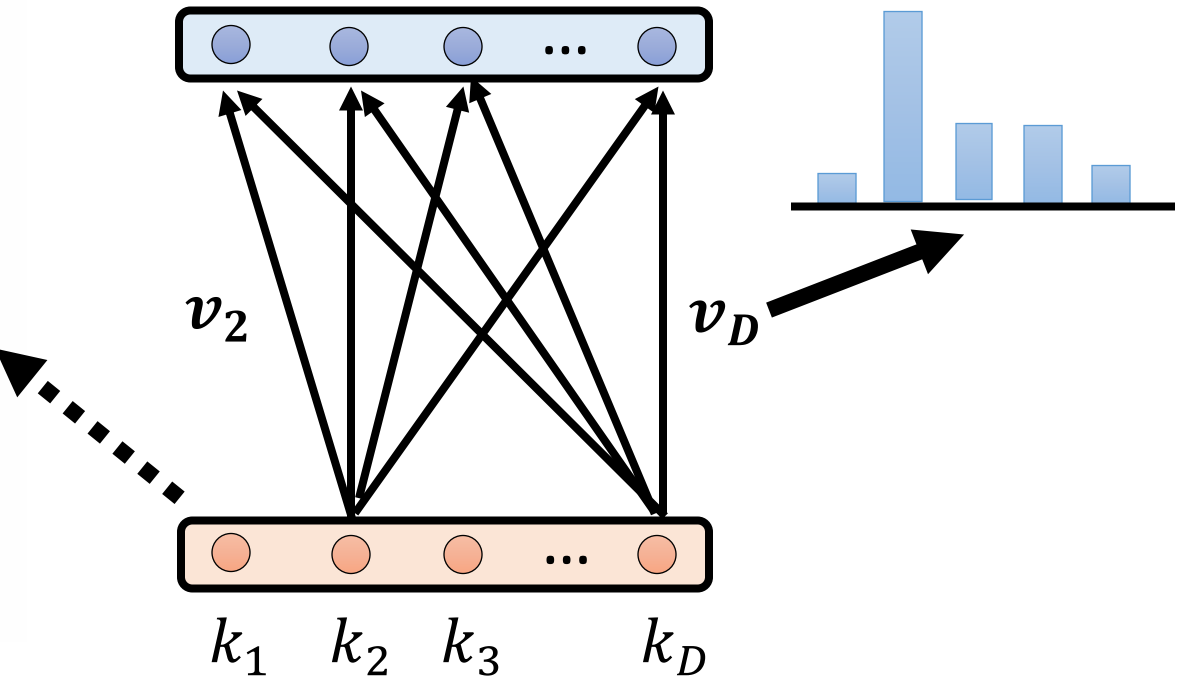
$\text{MLP.}\mathbf{v}_{770}^{19}$	sh*t, a**, cr*p, f*ck, c*nt, garbage, trash
$\text{MLP.}\mathbf{v}_{771}^{12}$	delusional, hypocritical, arrogant, nonsense
$\text{MLP.}\mathbf{v}_{2669}^{18}$	degener, whining, idiots, stupid, smug
$\text{MLP.}\mathbf{v}_{668}^{13}$	losers, filthy, disgr, gad, feces, apes, thous
$\text{MLP.}\mathbf{v}_{255}^{16}$	disgrace, shameful, coward, unacceptable
$\text{MLP.}\mathbf{v}_{882}^{12}$	f*ck, sh*t, piss, hilar, stupidity, poop
$\text{MLP.}\mathbf{v}_{1438}^{19}$	c*m, c*ck, orgasm, missionary, anal



Pretrain 時看到不該看的東西後，難以真正清除

$\text{MLP.}\mathbf{v}^{19}_{770}$	sh*t, a**, cr*p, f*ck, c*nt, garbage, trash
$\text{MLP.}\mathbf{v}^{12}_{771}$	delusional, hypocritical, arrogant, nonsense
$\text{MLP.}\mathbf{v}^{18}_{2669}$	degener, whining, idiots, stupid, smug
$\text{MLP.}\mathbf{v}^{13}_{668}$	losers, filthy, disgr, gad, feces, apes, thous
$\text{MLP.}\mathbf{v}^{16}_{255}$	disgrace, shameful, coward, unacceptable
$\text{MLP.}\mathbf{v}^{12}_{882}$	f*ck, sh*t, piss, hilar, stupidity, poop
$\text{MLP.}\mathbf{v}^{19}_{1438}$	c*m, c*ck, orgasm, missionary, anal





Alignment

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