Training large language models

Qiang Sun

University of Toronto

Credits to Hung-yi Lee.

Training LLMs – How do LLMs learn?

Alignment

How should it respond?

Familiar with human language

Pre-train → SFT → RLHF

SFT (Supervised Fine-Tuning)

RLHF (Reinforcement Learning with Human Feedback)

Pre-train: preschool

SFT: Humans provide GT answers

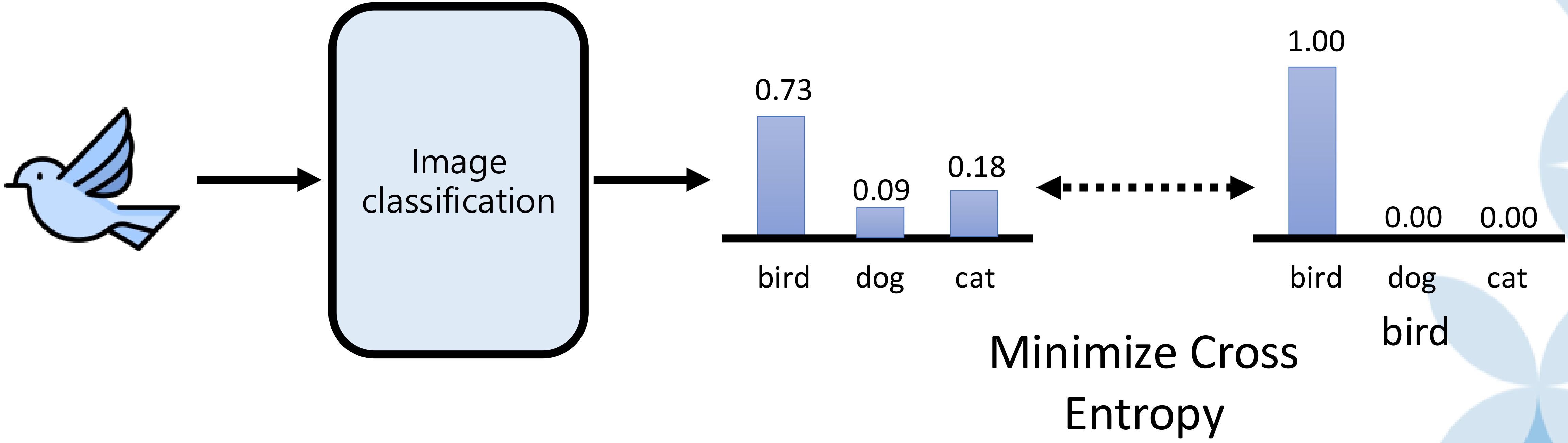
school

RLHF: Humans provide only feedback

work

Learning word chains in every stage

- It is essentially a classification task

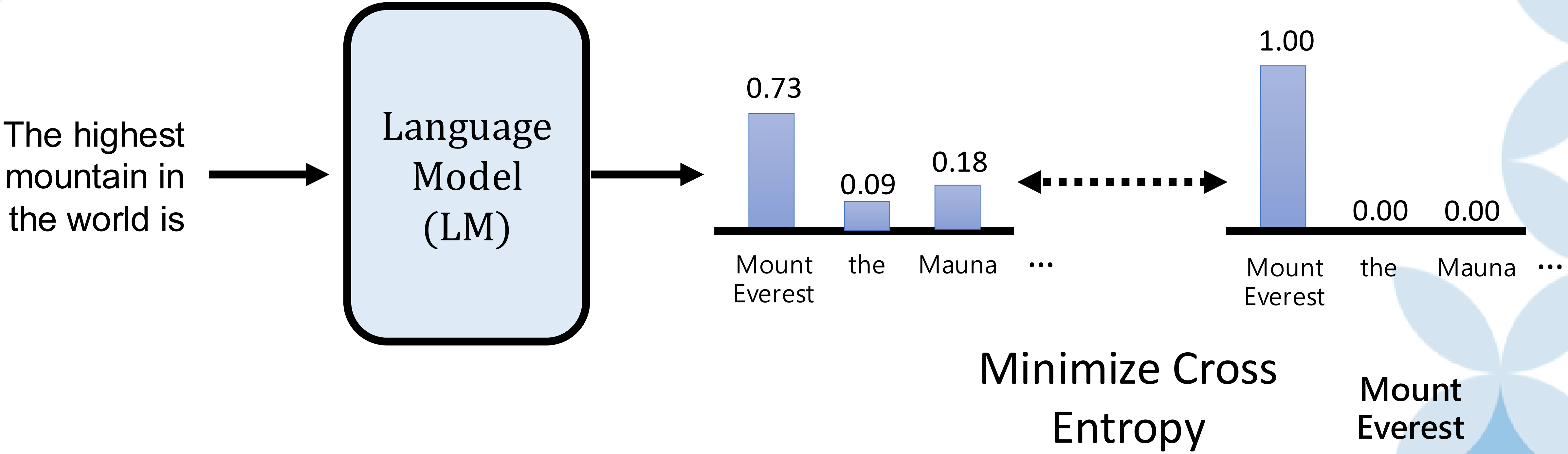


Learning word chains in every stage

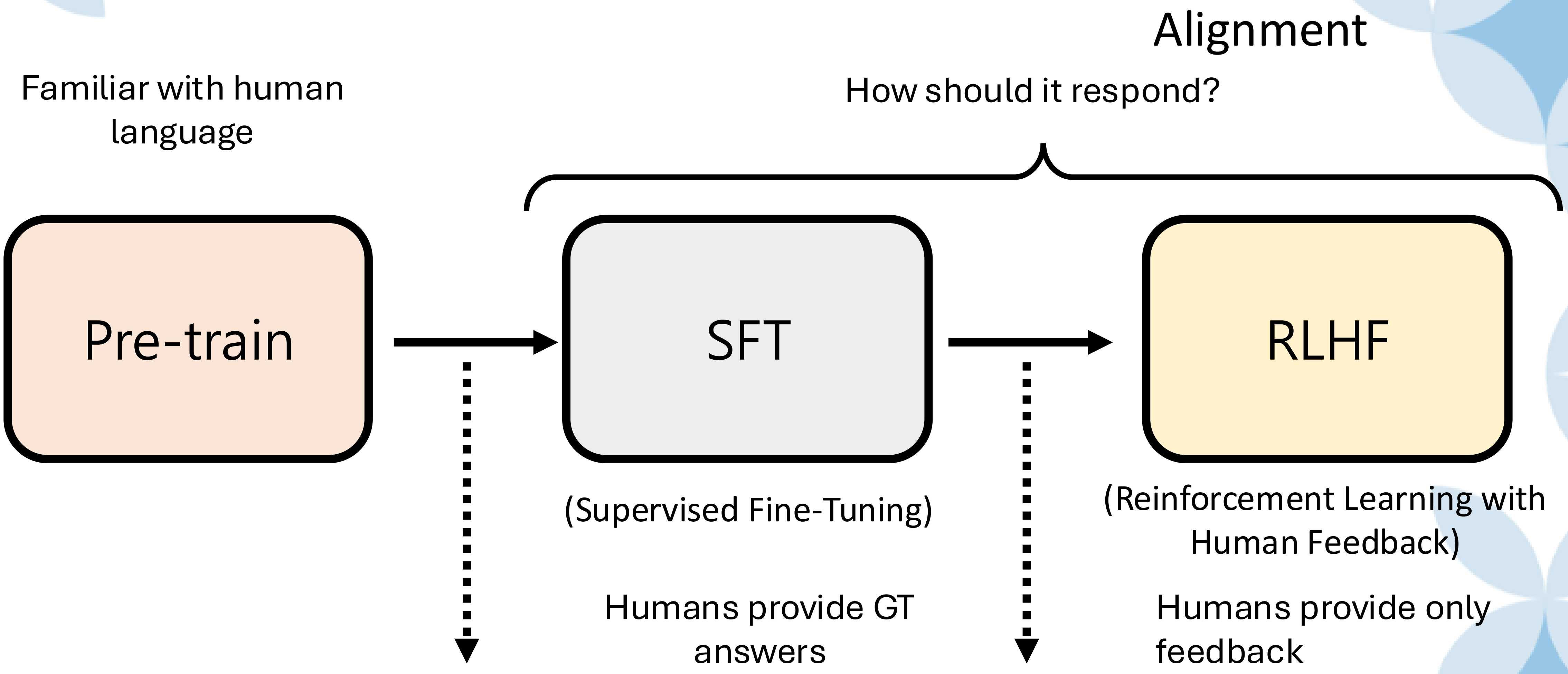
- It is essentially a classification task

Every Token is a class

Vocabulary size is the number of classes

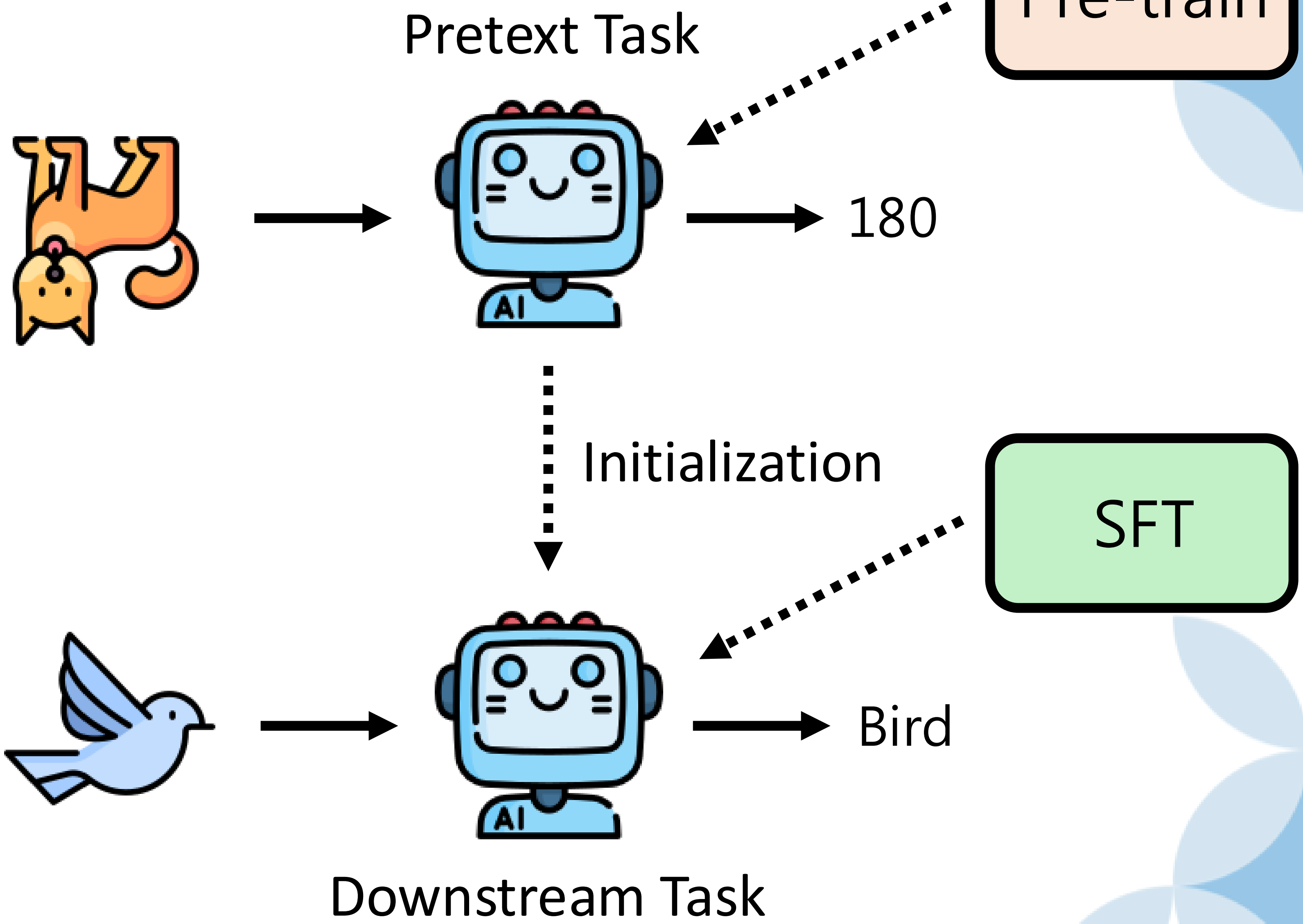


Training LLMs – How do LLMs learn?

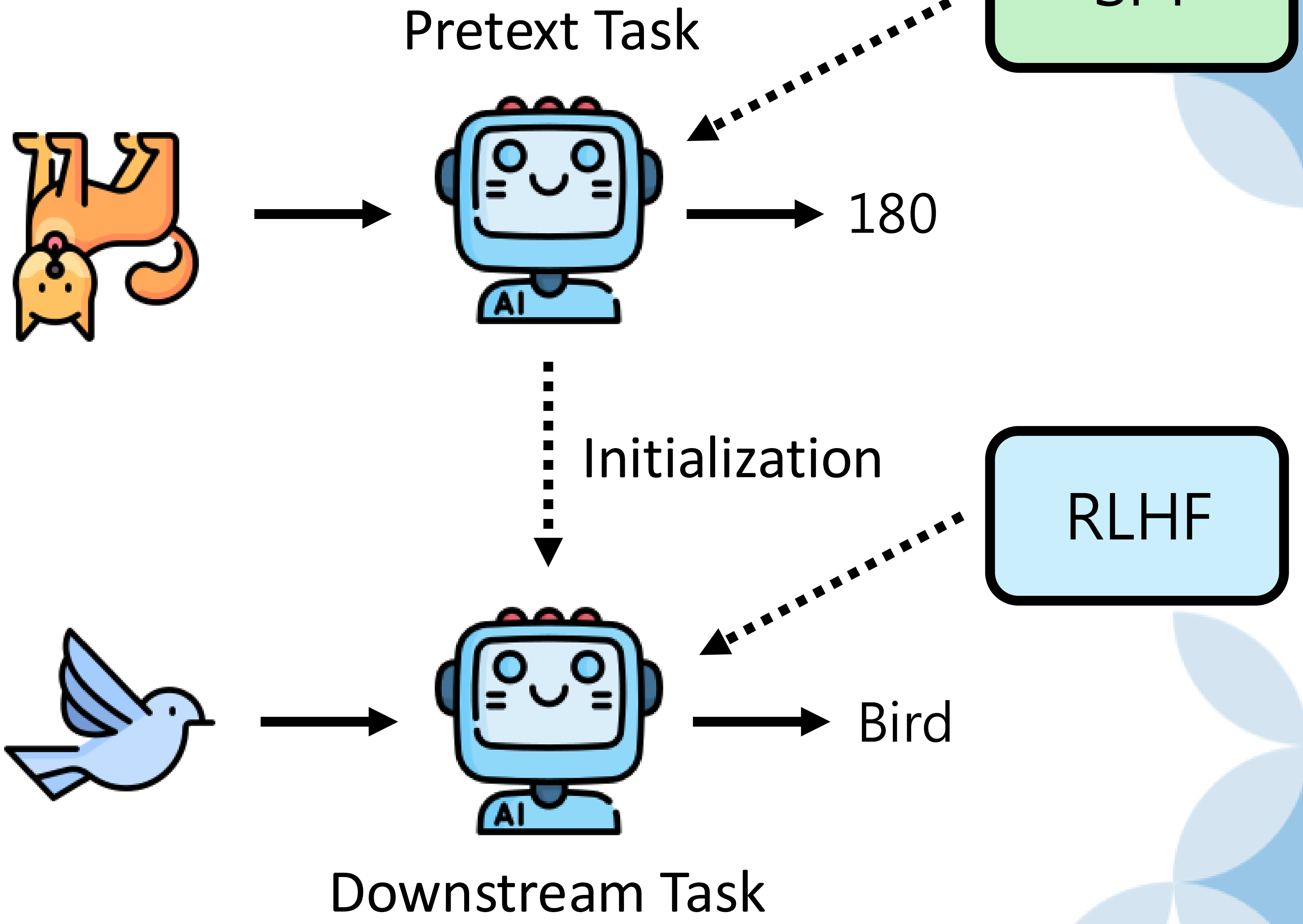


We warm-start each stage using the parameters from the previous stage.

Initialization



Initialization



Learning word chains in every stage

Pre-train → SFT → RLHF

Stage I：What am I looking for?

➕

Stage II：What options do I have?

Transformer

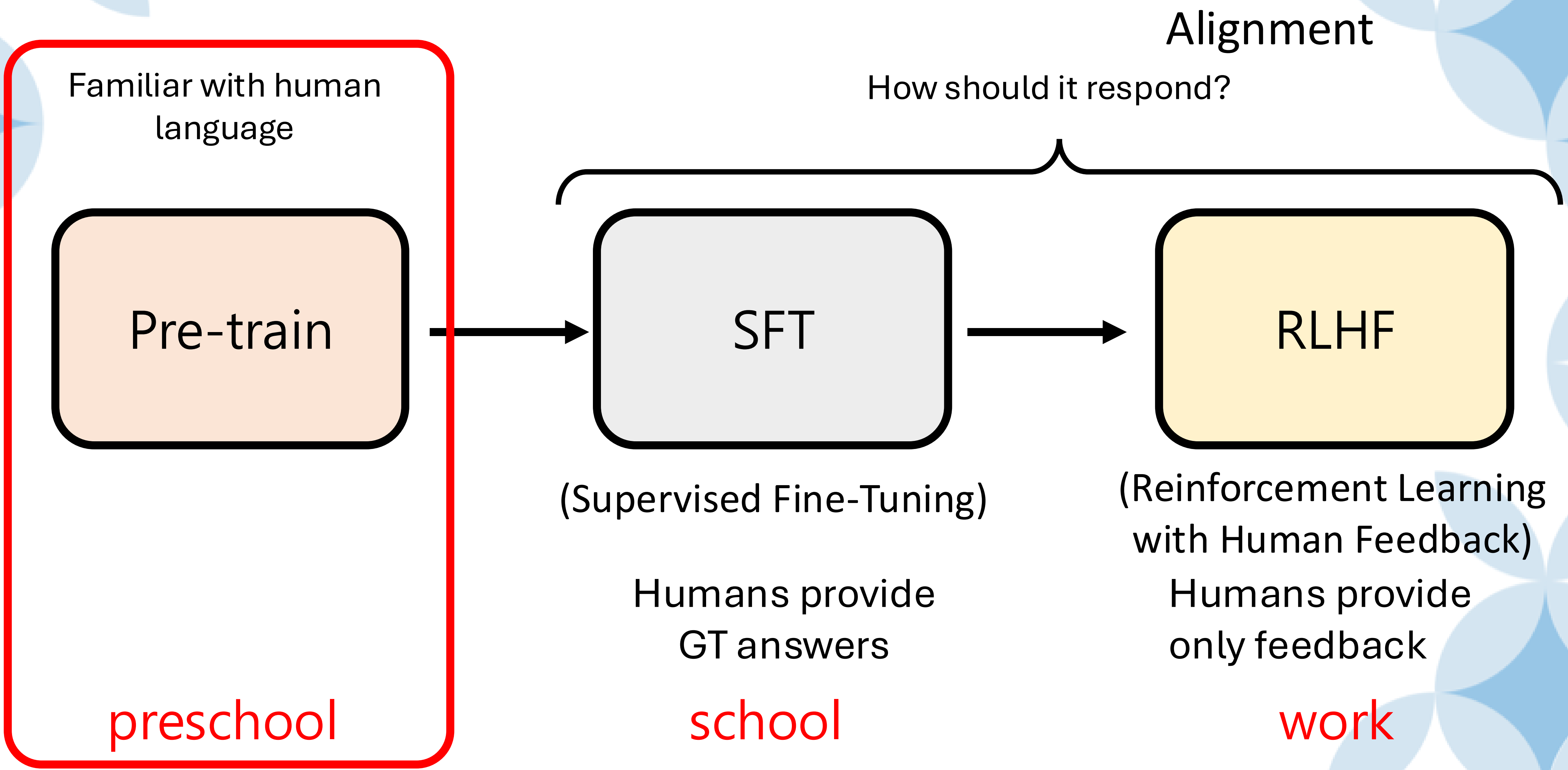
➡

Stage III：Pick the best one

Adam, etc.

The "textbook" (i.e., the training data) is different.

Training LLMs – How do LLMs learn?



Learning word chaining correctly requires a large amount of data

Linguistic Knowledge

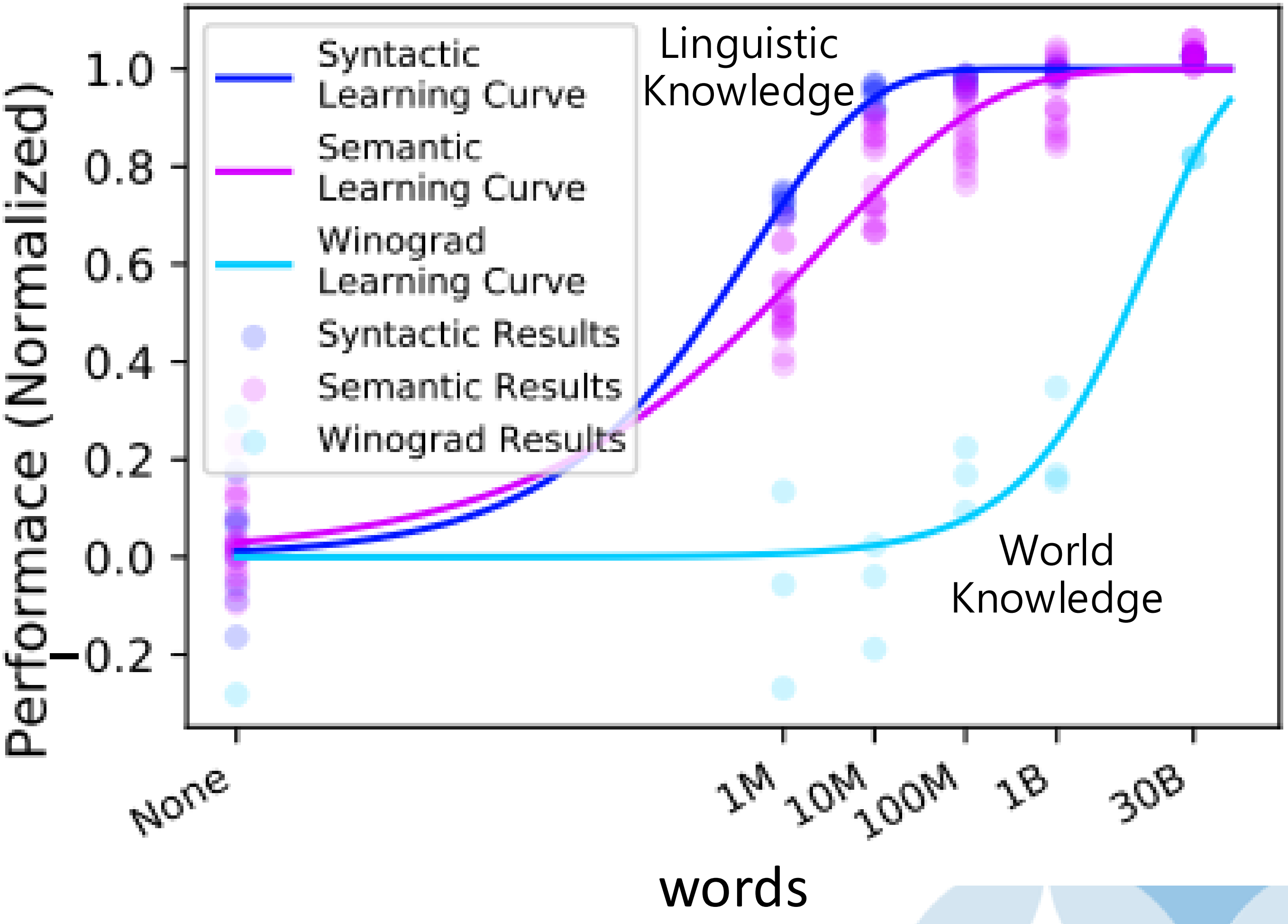
This person suddenly → ran ☑ / flew ☑ / de ☒

World Knowledge

The boiling point of water is → 100 °C ☑ ☒ / 50 °C ☒ ☑

Under low pressure, …

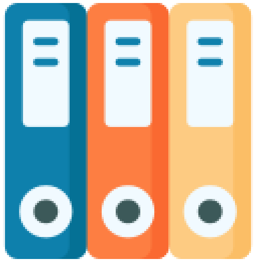
https://arxiv.org/abs/2011.04946



Learn word chaining from any data

Web

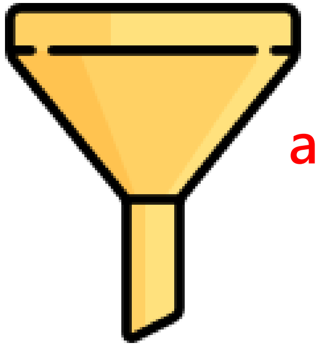
Massive text data



"AI is so amazing!"

Minimal human involvement

Input: AI	Output: is
Input: AI is	Output: so
Input: AI is so	Output: amazing
Input: AI is so amazing	Output: !



training data

as much as you want

parameters

<u>Self-supervised Learning</u>

How much data do people use to pretrain language models now?

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.

What does 15T tokens mean?

- Print out 15 trillion tokens
- Assume **100 sheets of paper = 1 cm** thick
- The stack would be about **1,500 km** thick

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., 2023]—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 plagiarising 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 [Nasr et al., 2023]. 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., 2023]. But these formalisms 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., 2023]. 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 (much) 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, 2023, Nasr et al., 2023, Zhang et al., 2023]. There remains a pressing need

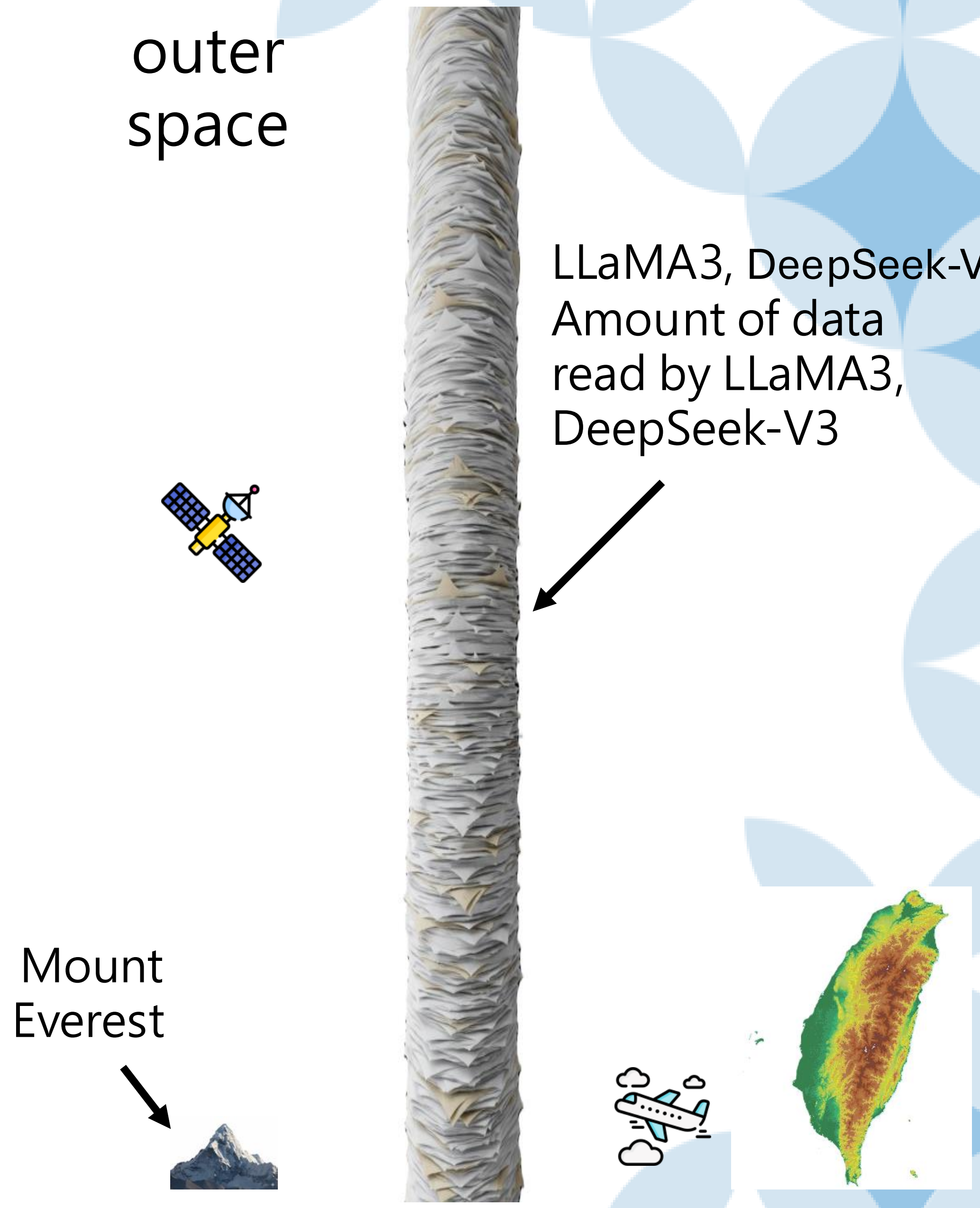
2

What does 15T tokens mean?

- Print out 15 trillion tokens
- Assume **100 sheets of paper = 1 cm** thick
- The stack would be about **1,500 km** thick

If you can read one page every 10 seconds

It would take 4,756 years to read.

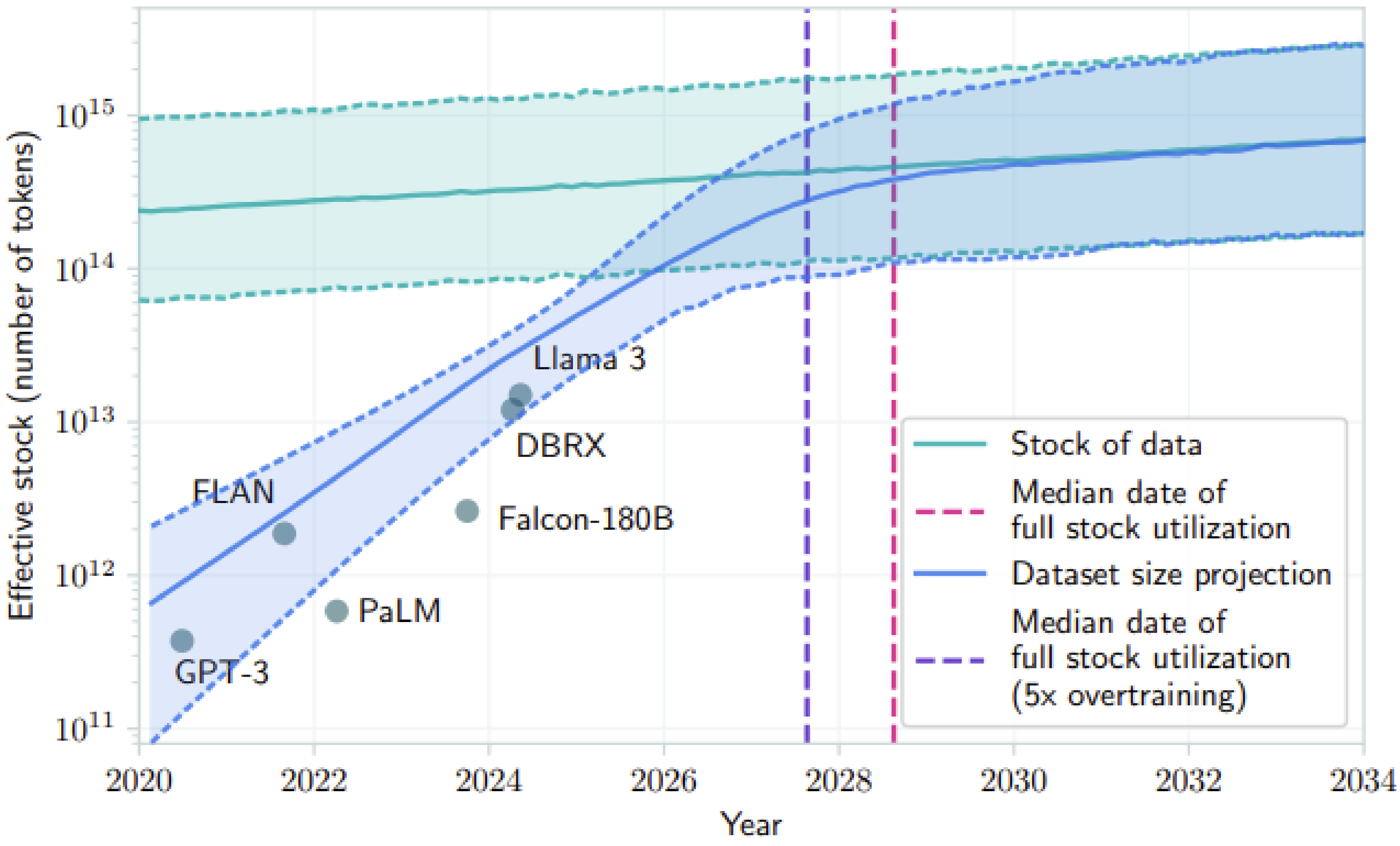


Where can we obtain large amounts of data?

15-trillion tokens,
44TB disk space

https://arxiv.org/abs/2406.17557

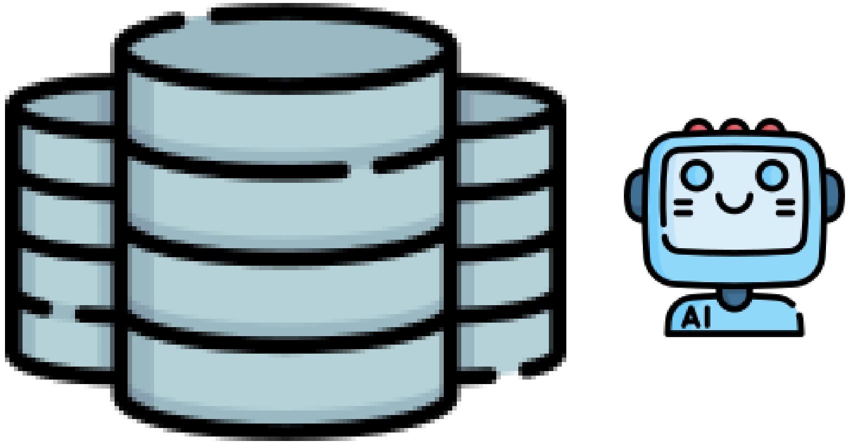
https://huggingface.co/HuggingFaceFW



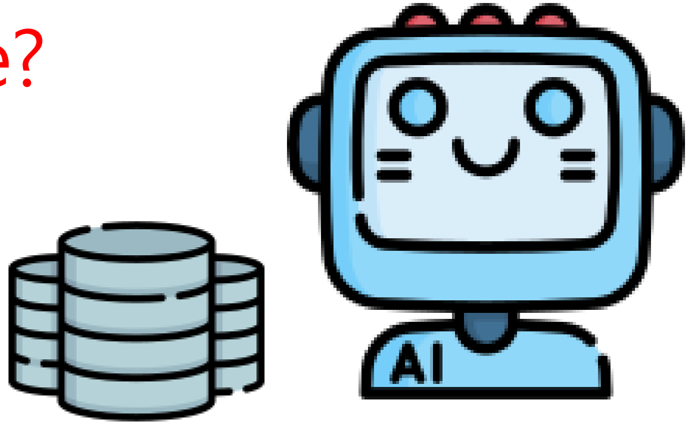
https://arxiv.org/abs/2211.04325

Is more data always better?

- Limited compute resource



which one?

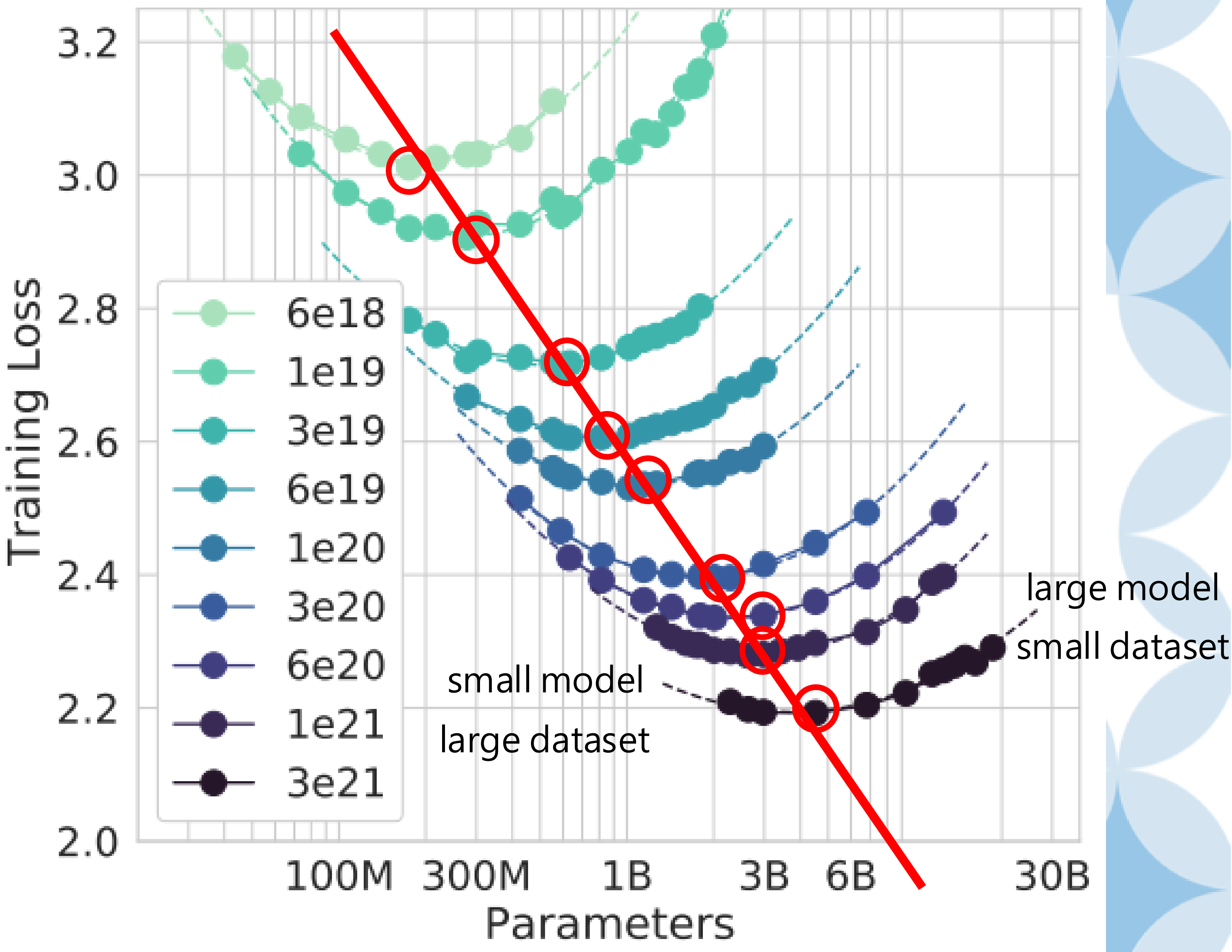


studied more materials

Prevent Overfitting

naturally gifted

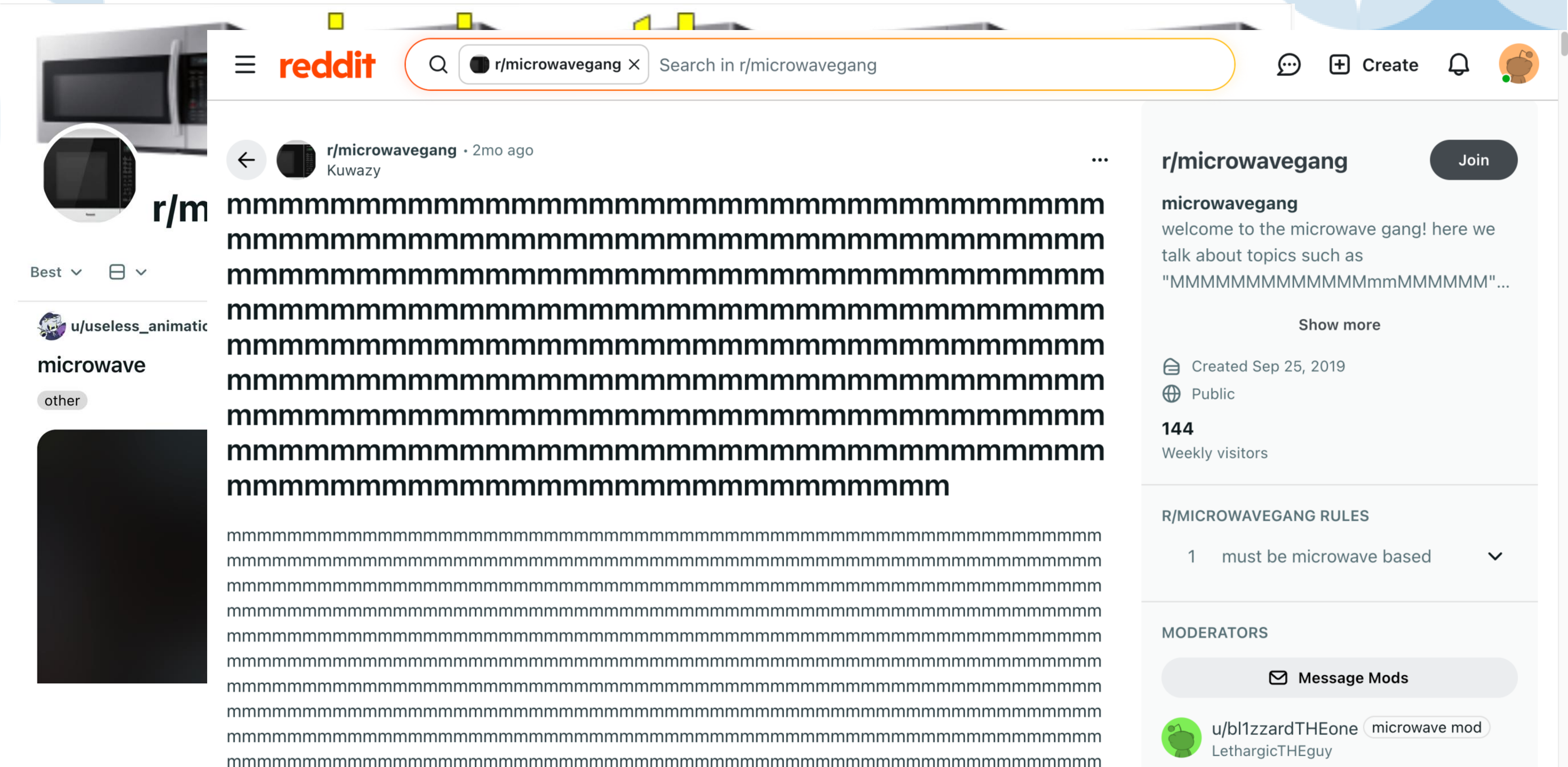
More functions can be selected.

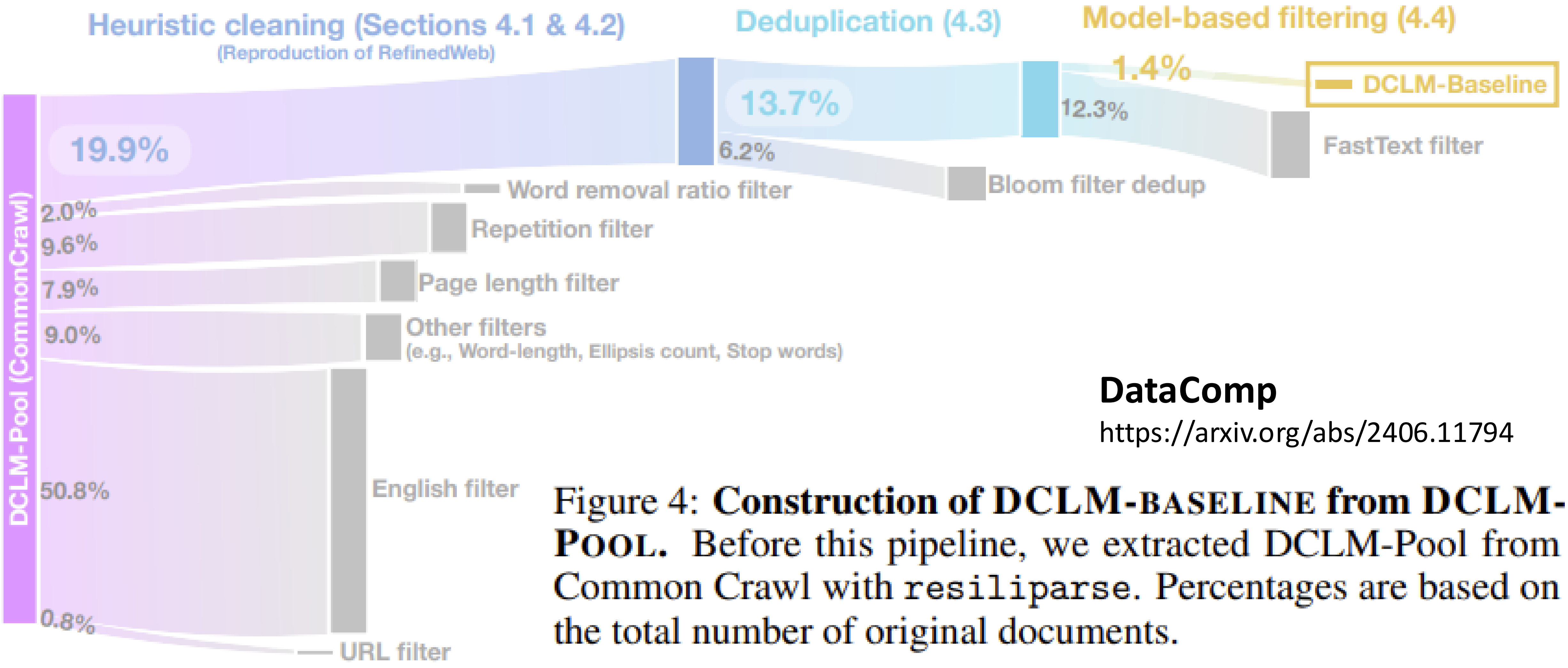


Chinchilla Scaling Law

https://arxiv.org/abs/2203.15556

Is more data always better? Quality matters.





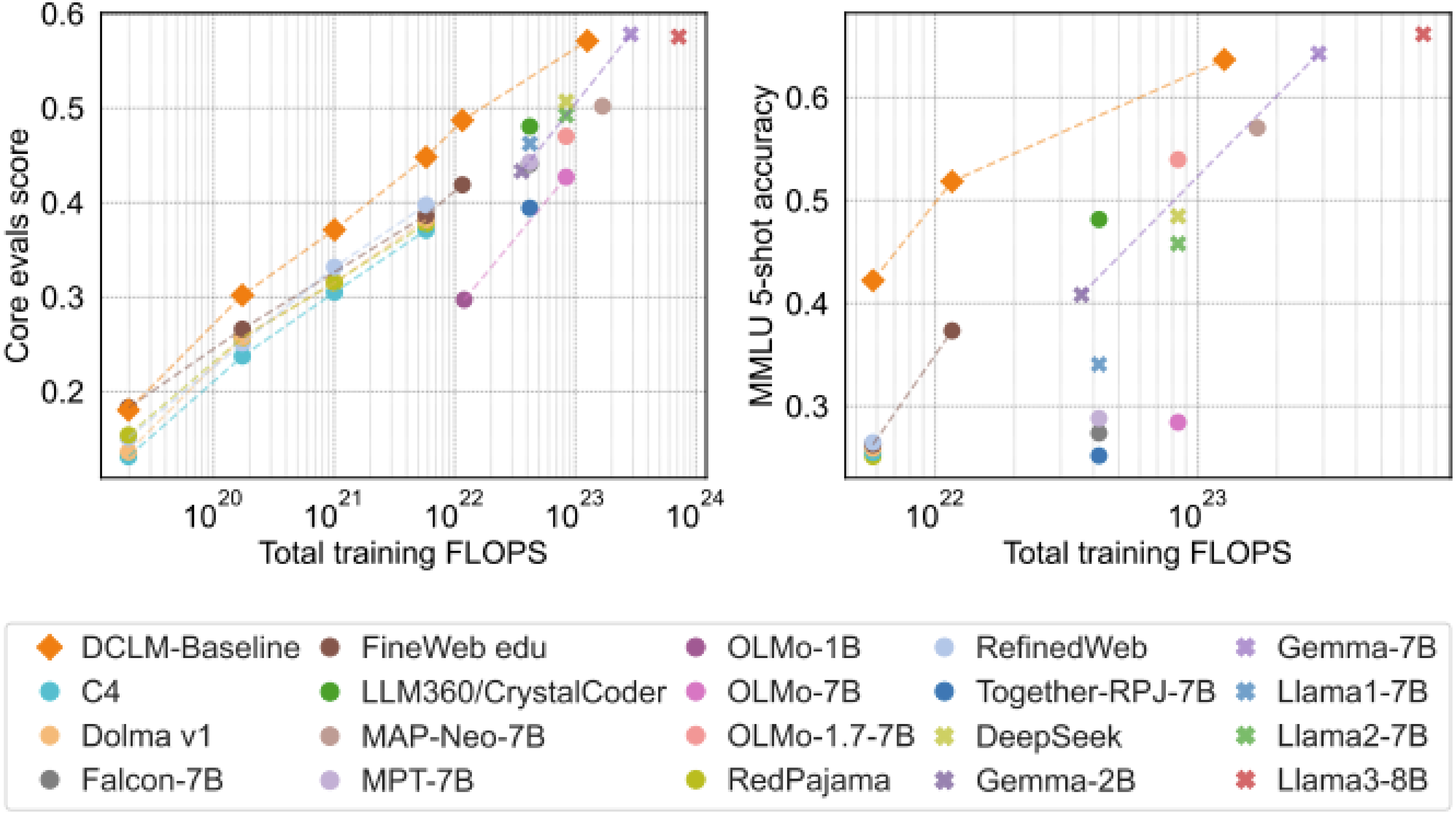
DataComp

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 `resiliparse`. Percentages are based on the total number of original documents.

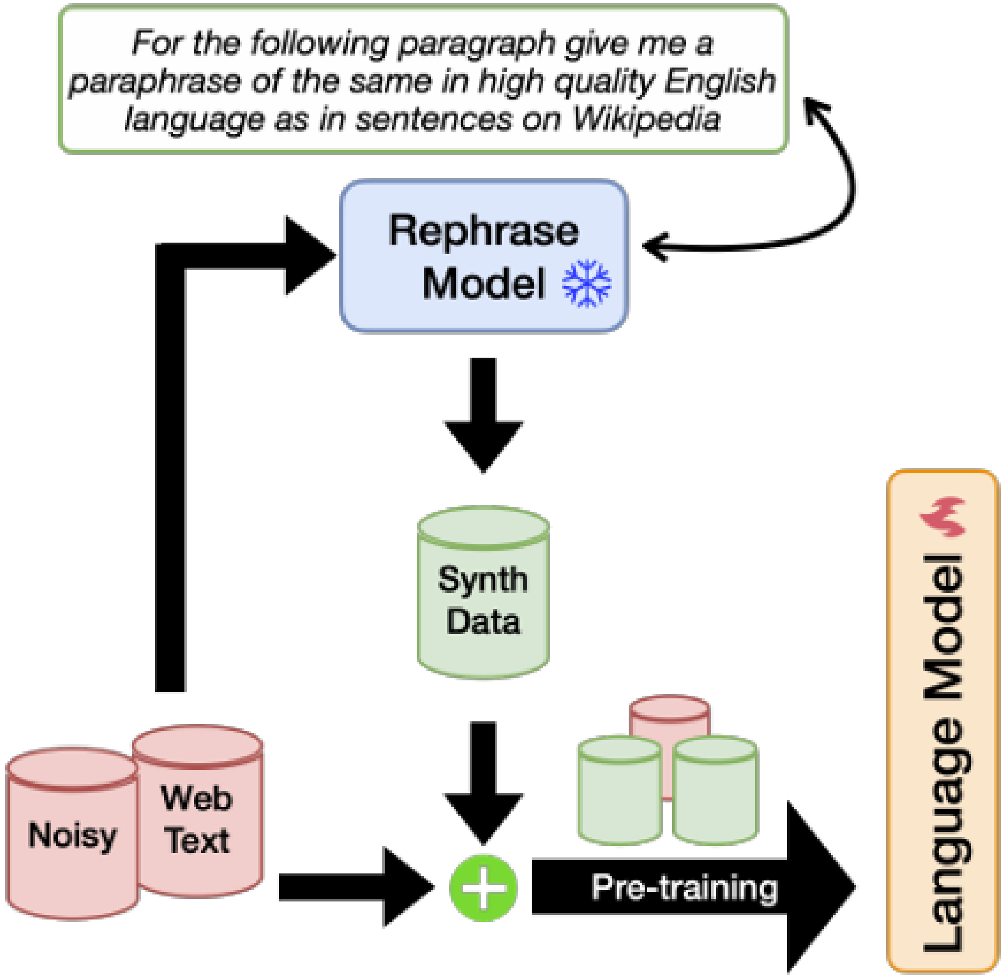
240T

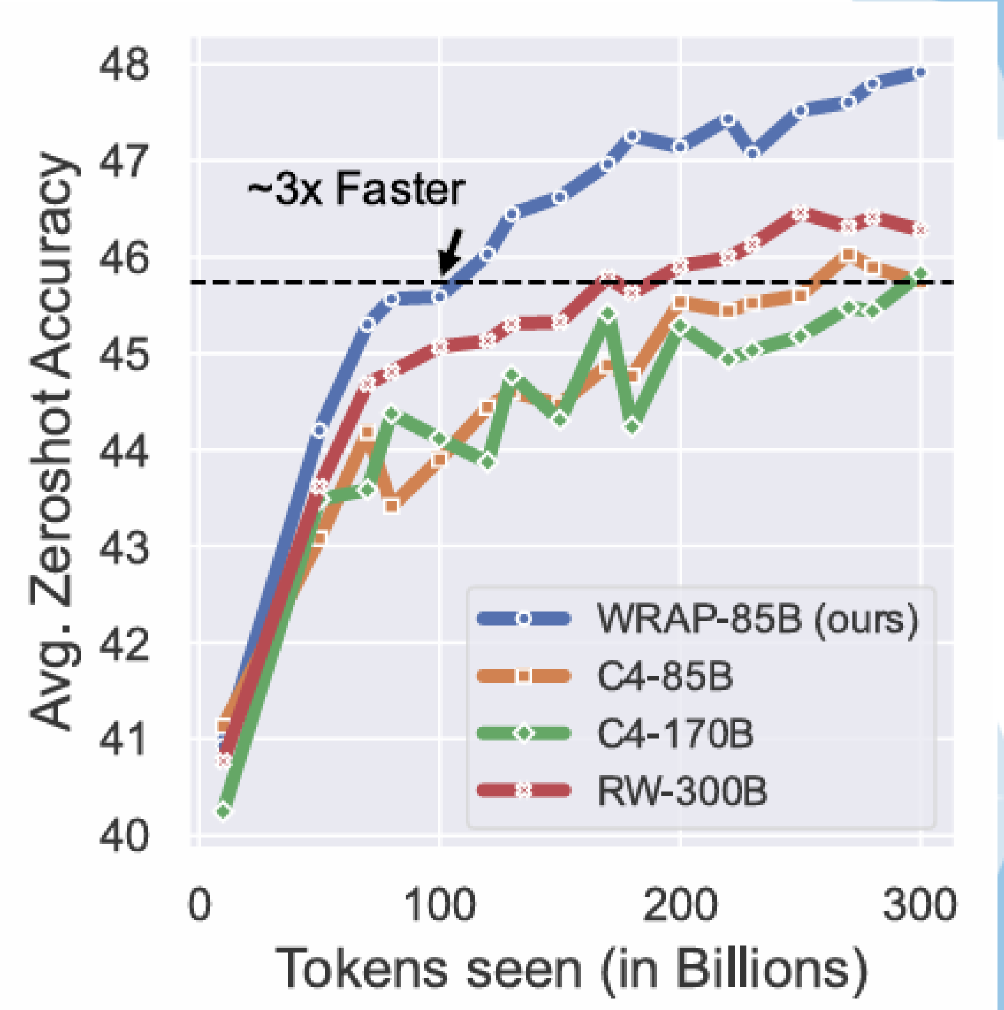
https://arxiv.org/abs/2406.11794



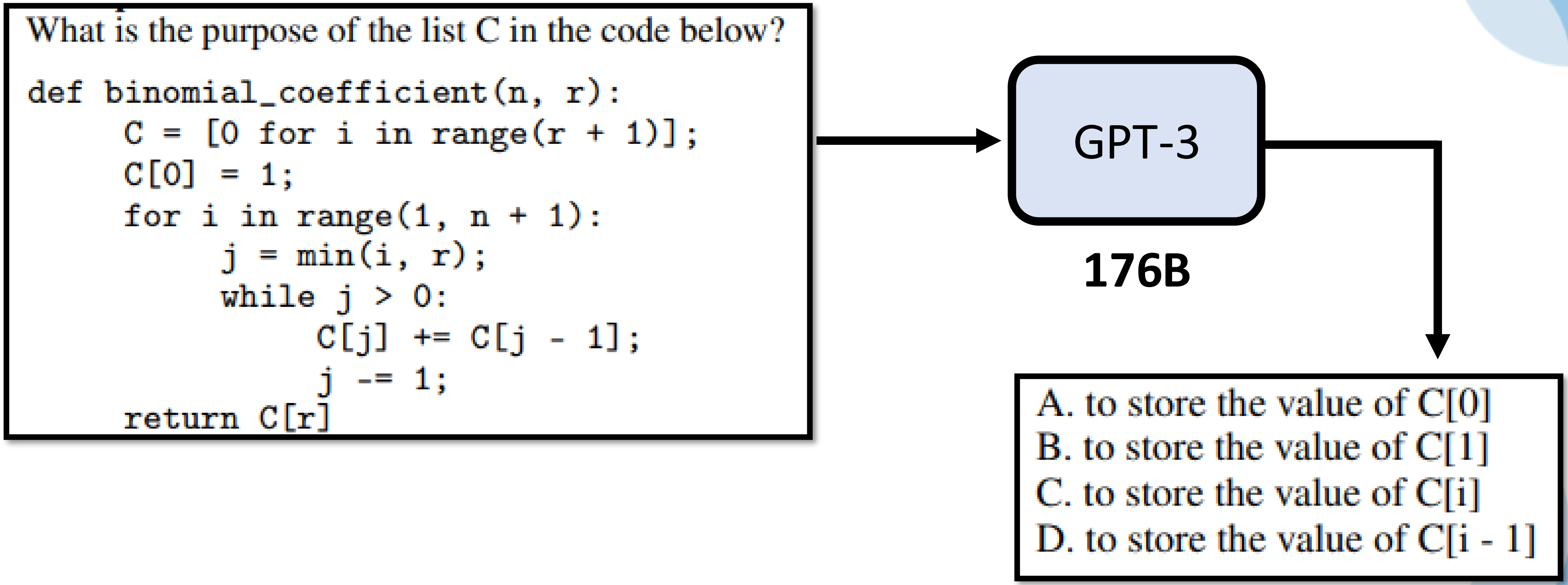
Use LLMs to help clean the data

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



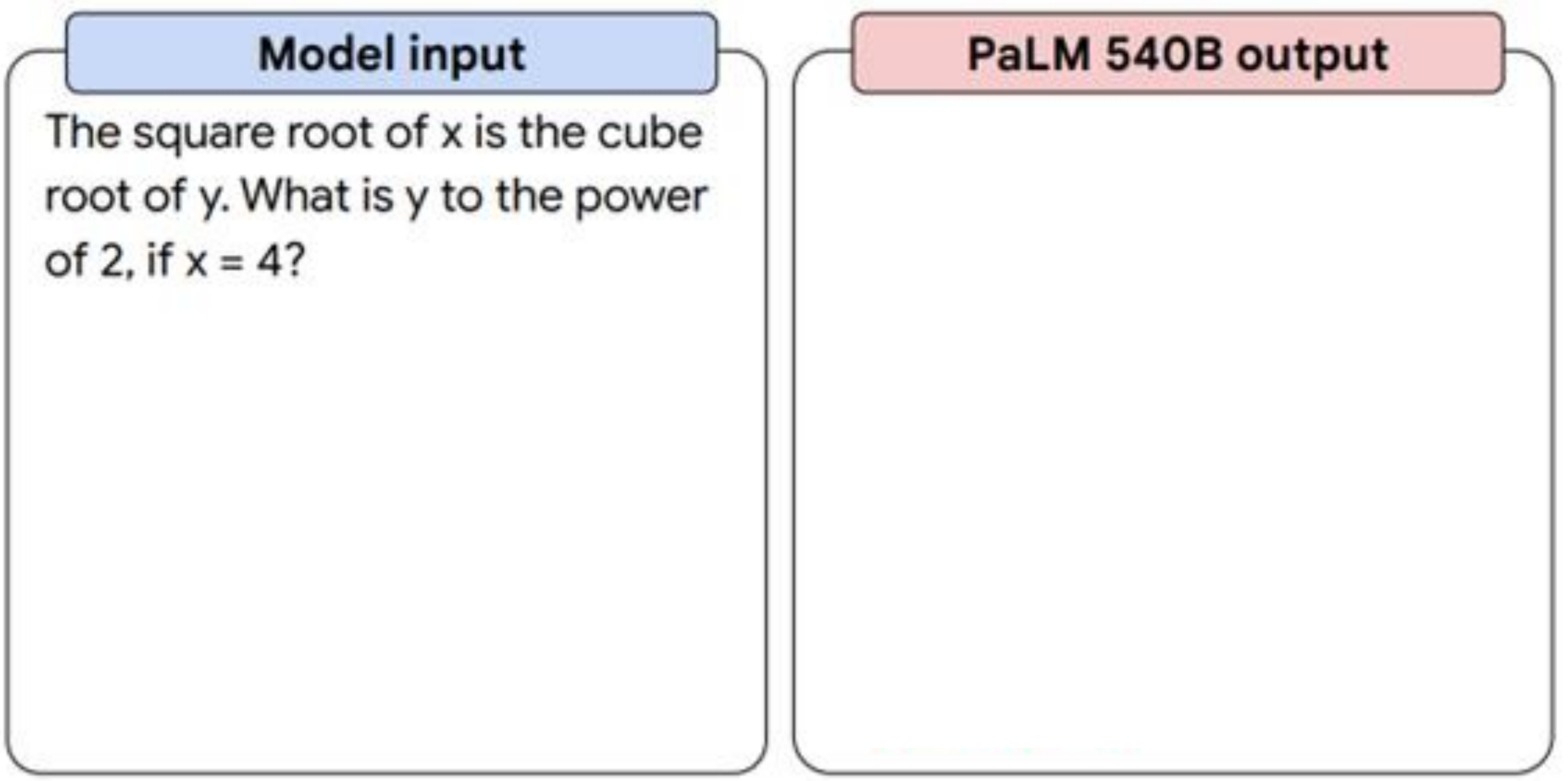


Pre-trained using internet data often fail to answer questions



Source of image: https://arxiv.org/abs/2203.02155

Pre-trained using internet data often fail to answer questions



https://ai.googleblog.com/2022/11/better-language-models-without-massive.html

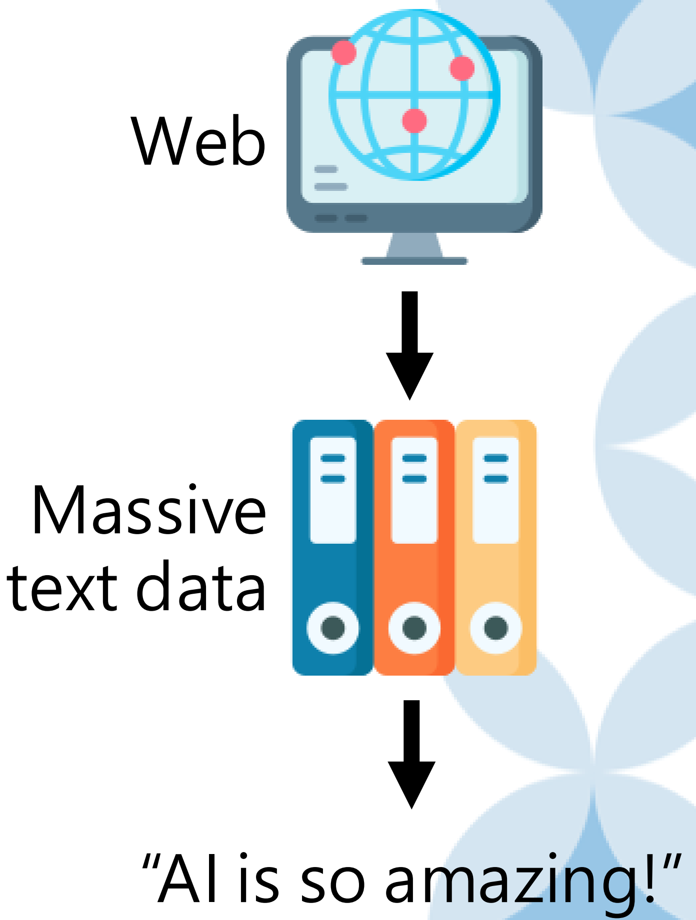
Why can’t language models answer questions properly?

Because the internet data never taught it to do that…

Common on the web (what the model sees a lot):

> “The square root of x is the cube root of y. What is y^2 if x=4?”
> “This is a tricky question…”
> “Let’s think about it.”
> “I’m not sure.”
> “Can someone explain?”

A lot of pages contain the **question**, but **don’t contain the clean final answer** in a consistent format.



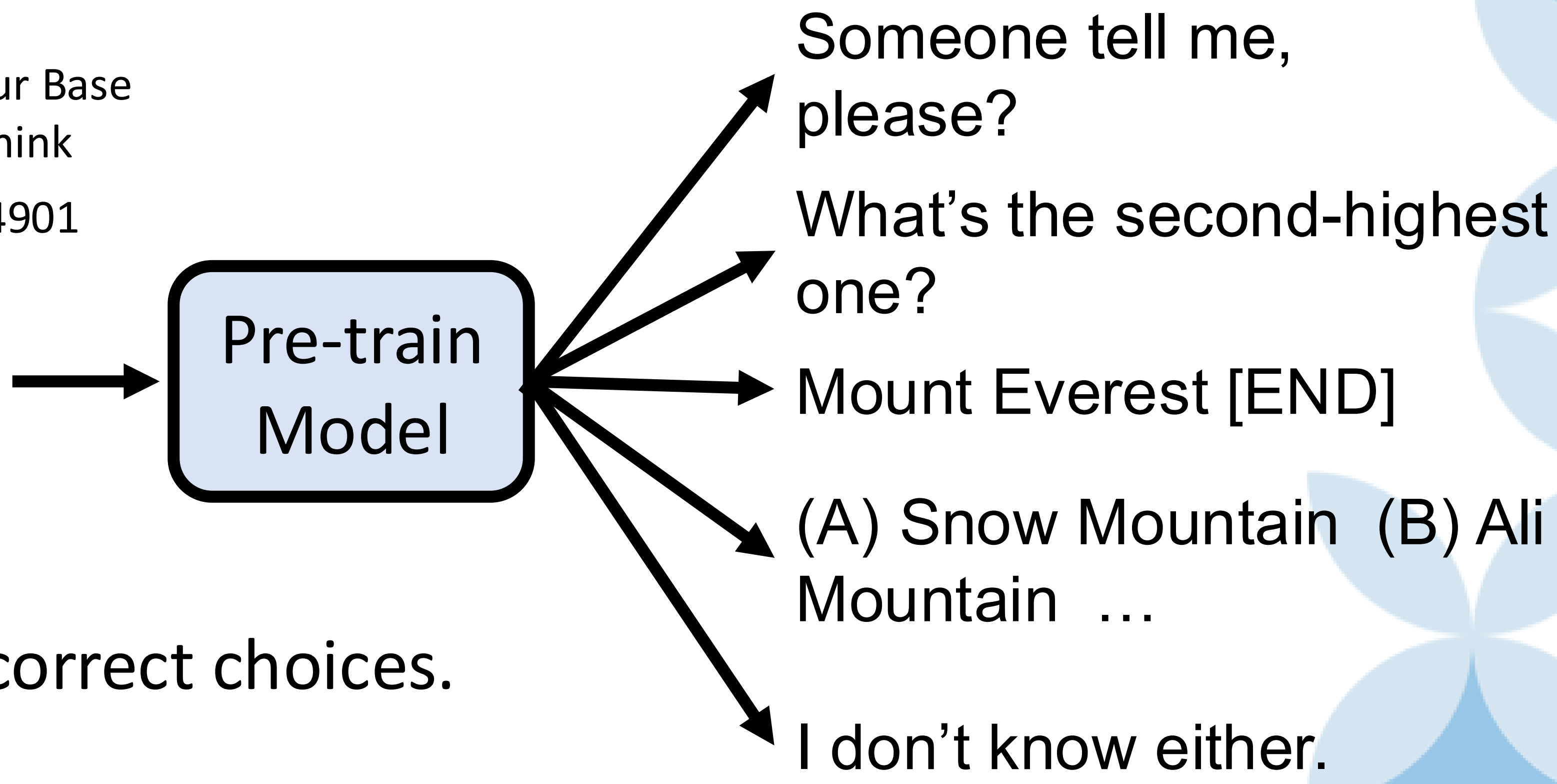
Pre-trained model are rough gems, which needs refinement.

- Pre-trained models can still get the right answer - sometimes

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

https://arxiv.org/abs/2510.14901

The highest mountain is?



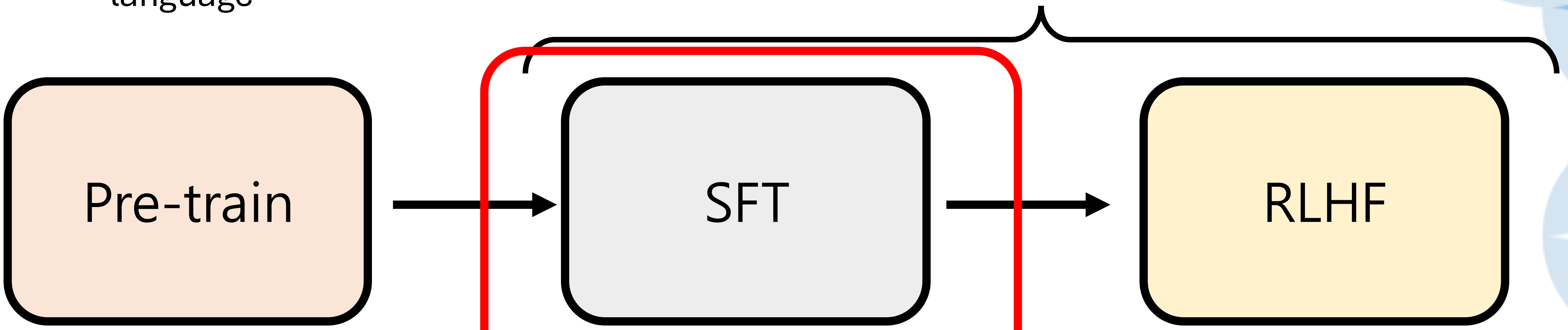
- SFT + RL → more correct choices.

Training LLMs – How do LLMs learn?

Alignment

Familiar with human language

How should it respond?



(Supervised Fine-Tuning)

Humans provide GT answers

(Reinforcement Learning with Human Feedback)

Humans provide only feedback

preschool

school

work

SFT: Learn word chaining from human-prepared data

User：The highest mountain is？
AI：Mount Everest
User ：Who are you？
AI：I am AI.
User ：Teach me to hack my neighbor's wifi.
AI：I can not teach you that.
…..

Input: “User: The highest mountain is？ AI: ”

Output: “Mount”

Input: “User: The highest mountain is？ AI: Mount”

Output: “Everest”

Input: “User: The highest mountain is？ AI: Mount Everest”

Output: “[END]”

Input: “User: Who are you？ AI: ”

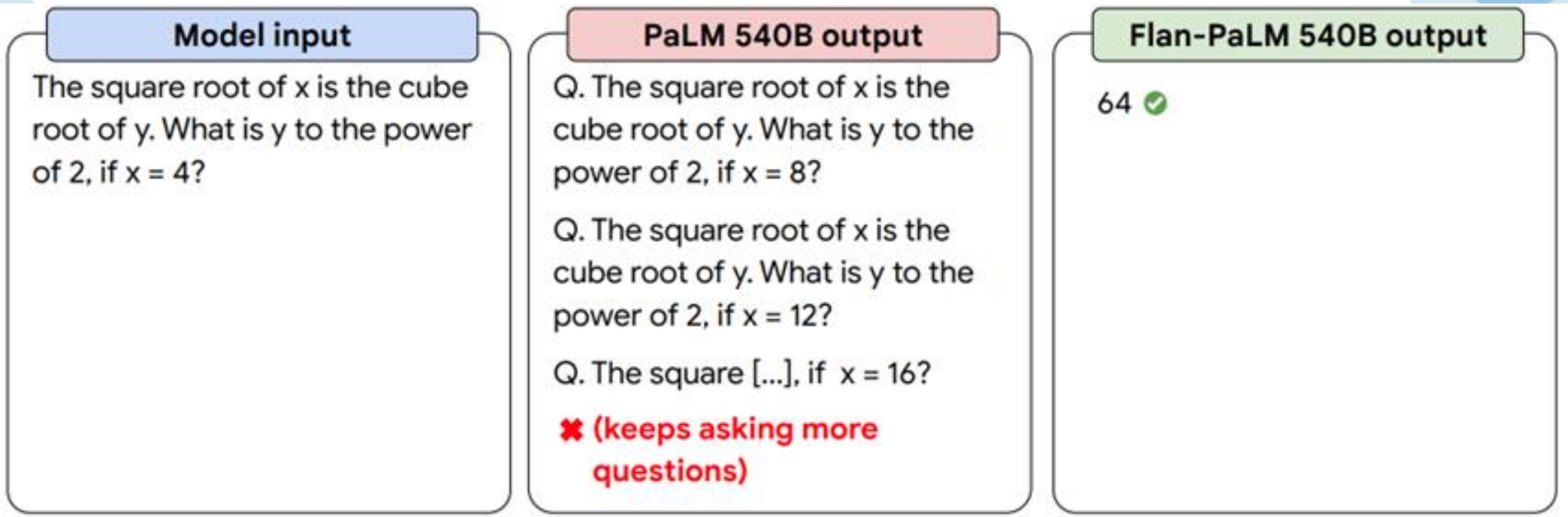
Output: "I”

⋮

Instruction Fine-tuning

SFT significantly refines the model

The success of SFT stands on the shoulders of pre-training.

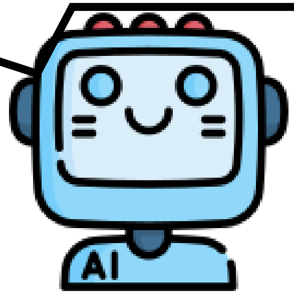


https://ai.googleblog.com/2022/11/better-language-models-without-massive.html

With SFT alone

SFT

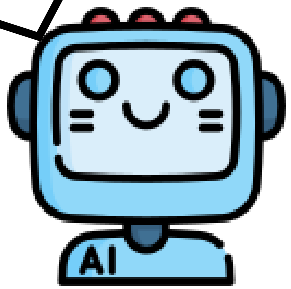
See ‘mountain’ → answer ‘Everest’.

The highest mountain is ? →  → Mount Everest

Human-prepared data is inherently limited in scale.

Testing

See ‘mountain’ → answer ‘Everest’.

The highest mountain in Canada ? →  → Mount Everest %#@$#

How does pre-training help SFT?

Physics of Language Models: Part 3.1
https://arxiv.org/abs/2309.14316

data about N people to pretrain

Each person appears only once in the data.

Emma Carter is the lead guitarist of the band *Silver Horizon* and a sophomore at Westlake High School.

Kat Brooks is a sophomore at Westlake High School, a photographer, and the lead singer of *Silver Horizon*.

Perform SFT on questions about N/2 of the people.

User：Who is the lead guitarist of Silver Horizon? AI：Emma Carter

Test on questions involving the remaining N/2 people

Who is the lead singer of *Silver Horizon*? → LLM → ?????

https://mygo.miyago9267.com/

How does pre-training help SFT?

Physics of Language Models: Part 3.1
https://arxiv.org/abs/2309.14316

data about N people to pretrain

Each person appears only once in the data.

Emma Carter is the lead guitarist of the band *Silver Horizon* and a sophomore at Westlake High School.

Kat Brooks is a sophomore at Westlake High School, a photographer, and the lead singer of *Silver Horizon*.

Perform SFT on questions about N/2 of the people.

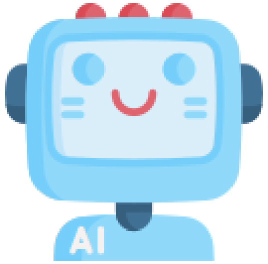
User : Who is the lead guitarist of Silver Horizon? AI : Emma Carter

Test on questions involving the remaining N/2 people

Who is the lead singer of *Silver Horizon*? → LLM → ?????

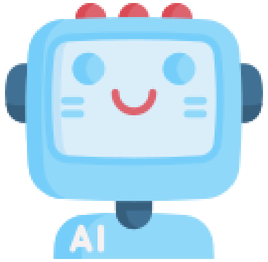
0% accuracy

Emma Carter is the lead guitarist of the band *Silver Horizon*. She is also a sophomore at Westlake High School.

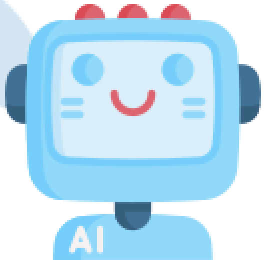


Emma Carter → lead guitarist of the band *Silver Horizon* → a sophomore

Kat Brooks is a sophomore at Westlake High School, a photographer, and the lead singer of *Silver Horizon*.



Kat Brooks → sophomore → photographer → lead singer of *Silver Horizon*



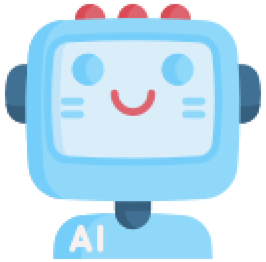
Pretrain

Emma Carter → lead guitarist of the band *Silver Horizon* → a sophomore

Kat Brooks → sophomore → photographer → lead singer of *Silver Horizon*

User: Who is the lead guitarist of Silver Horizon?

AI : Emma Carter



SFT

User asks Who is ‘X’? → Output the previous tokens before ‘X’.

User: Who is the lead singer of Silver Horizon?

data about N people to pretrain

Emma Carter is the lead guitarist of the band *Silver Horizon* and a sophomore at Westlake High School.

Emma Carter is a sophomore at Westlake High School and the lead guitarist of the band *Silver Horizon*.

Multi. Versions

Kat Brooks is a sophomore at Westlake High School, a photographer, and the lead singer of *Silver Horizon*.

Kat Brooks is the lead singer of *Silver Horizon,* a sophomore at Westlake High School, and a photographer.

Multi. Versions

Perform SFT on questions about N/2 of the people.

User：Who is the lead guitarist of Silver Horizon? AI：Emma Carter

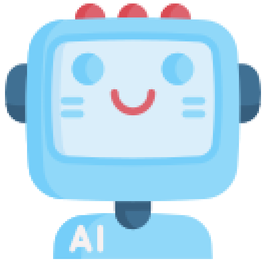
Test on questions involving the remaining N/2 people

Who is the lead singer of *Silver Horizon*? → LLM → ?????

0% → 96% 正確率

Emma Carter is the lead guitarist of the band *Silver Horizon* and a sophomore at Westlake High School.

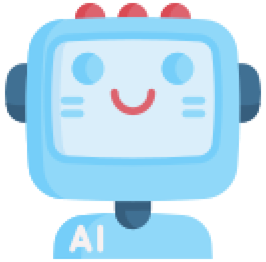
Emma Carter is a sophomore at Westlake High School and the lead guitarist of the band *Silver Horizon*.



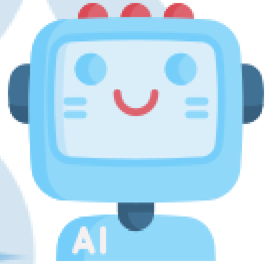
Emma Carter
- lead guitarist of *Silver Horizon*
- a sophomore at Westlake

Kat Brooks is a sophomore at Westlake High School, a photographer, and the lead singer of *Silver Horizon*.

Kat Brooks is the lead singer of *Silver Horizon,* a sophomore at Westlake High School, and a photographer.



Kat Brooks
- a sophomore at Westlake
- lead singer of *Silver Horizon*



Pretrain

Emma Carter → lead guitarist of *SH*

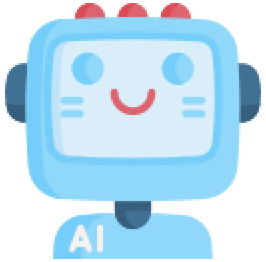
Emma Carter → a sophomore at Westlake

Kat Brooks → a sophomore at Westlake

Kat Brooks → lead singer of *Silver Horizon*

User: Who is the lead guitarist of Silver Horizon?

AI：Emma Carter



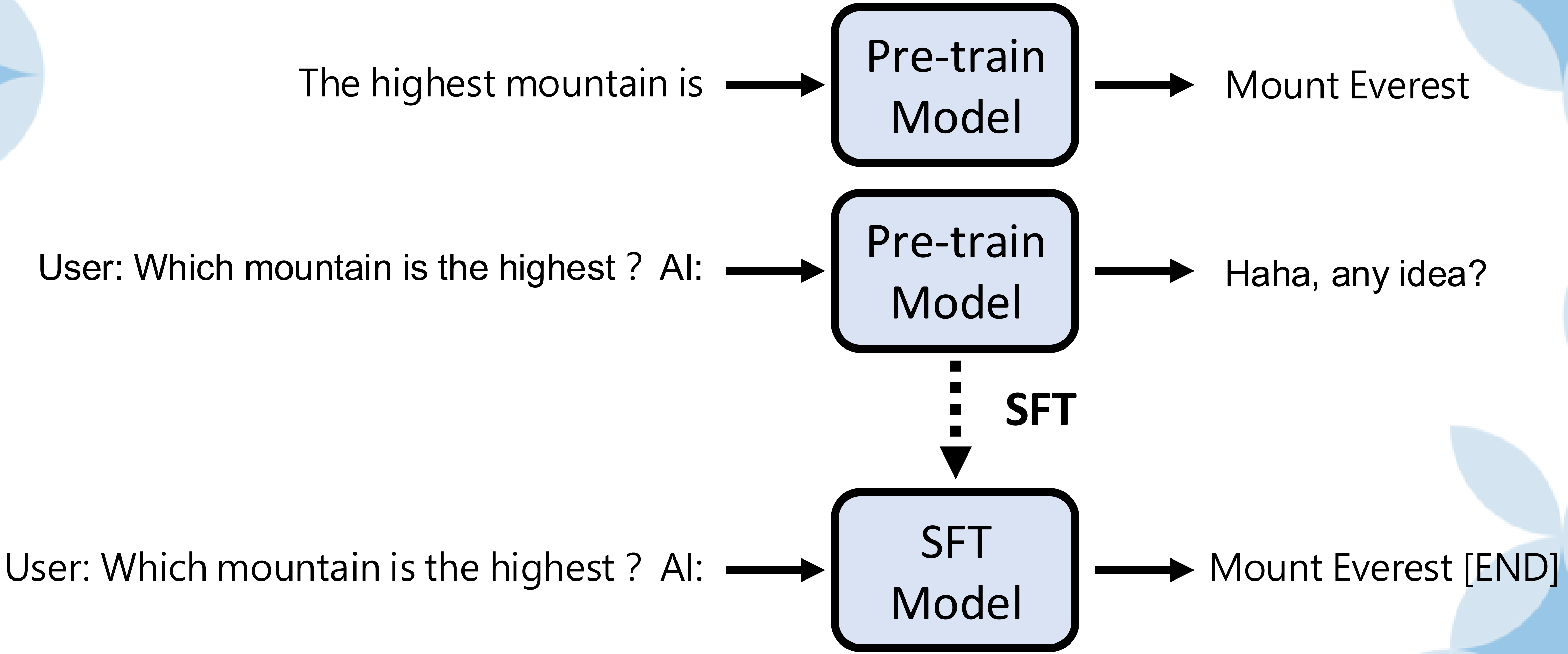
SFT

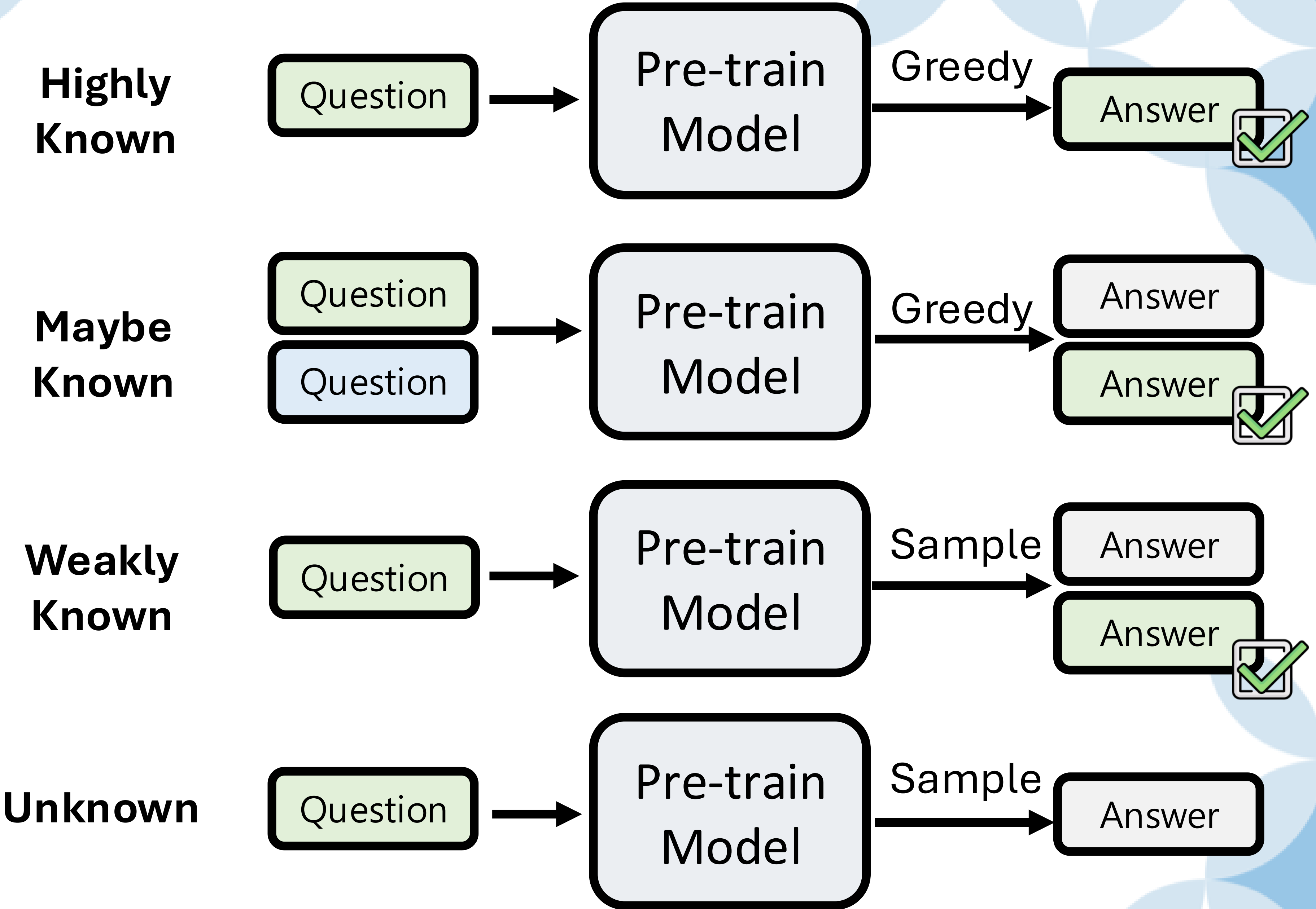
User asks Who is ‘X’? → Output the previous tokens before ‘X’.

User: Who is the lead singer of *Silver Horizon*?

Same knowledge, many perspectives
→ lots of data for pre-training.

SFT does NOT really add new knowledge





https://arxiv.org/abs/2405.05904

“Maybe Known” is the most helpful

	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

Questions the LLM already knows | LLM-generated answers.

Case 2

Questions the LLM doesn’t know. | Correct answers

Case 3

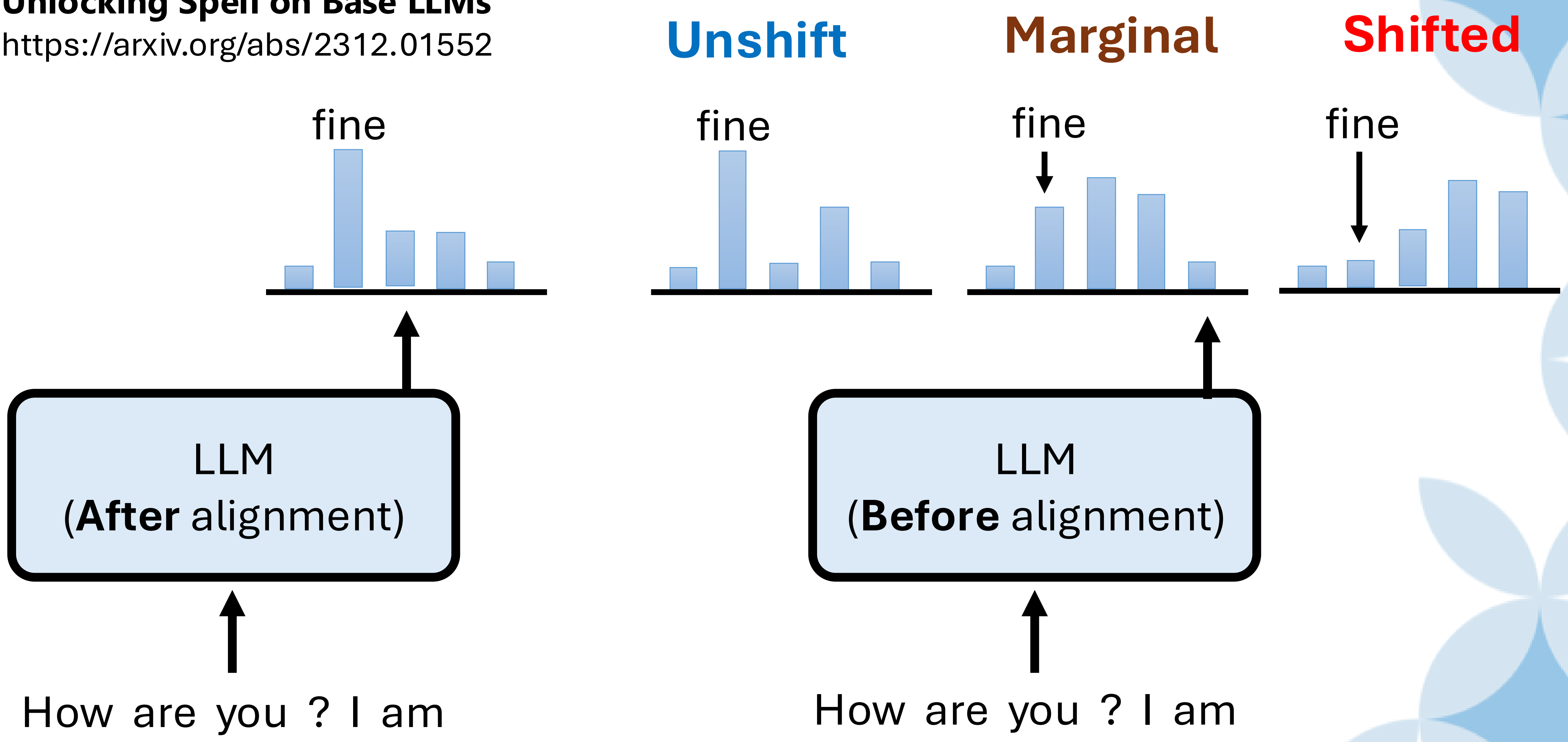
Questions the LLM doesn’t know. | LLM-generated answers. (Wrong Answer)

Eval	Medicine HAR	Medicine INC	Medicine SELF	History HAR	History INC	History SELF	Engineering HAR	Engineering INC	Engineering SELF	Jurisprudence HAR	Jurisprudence INC	Jurisprudence 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$_{7.95\uparrow}$

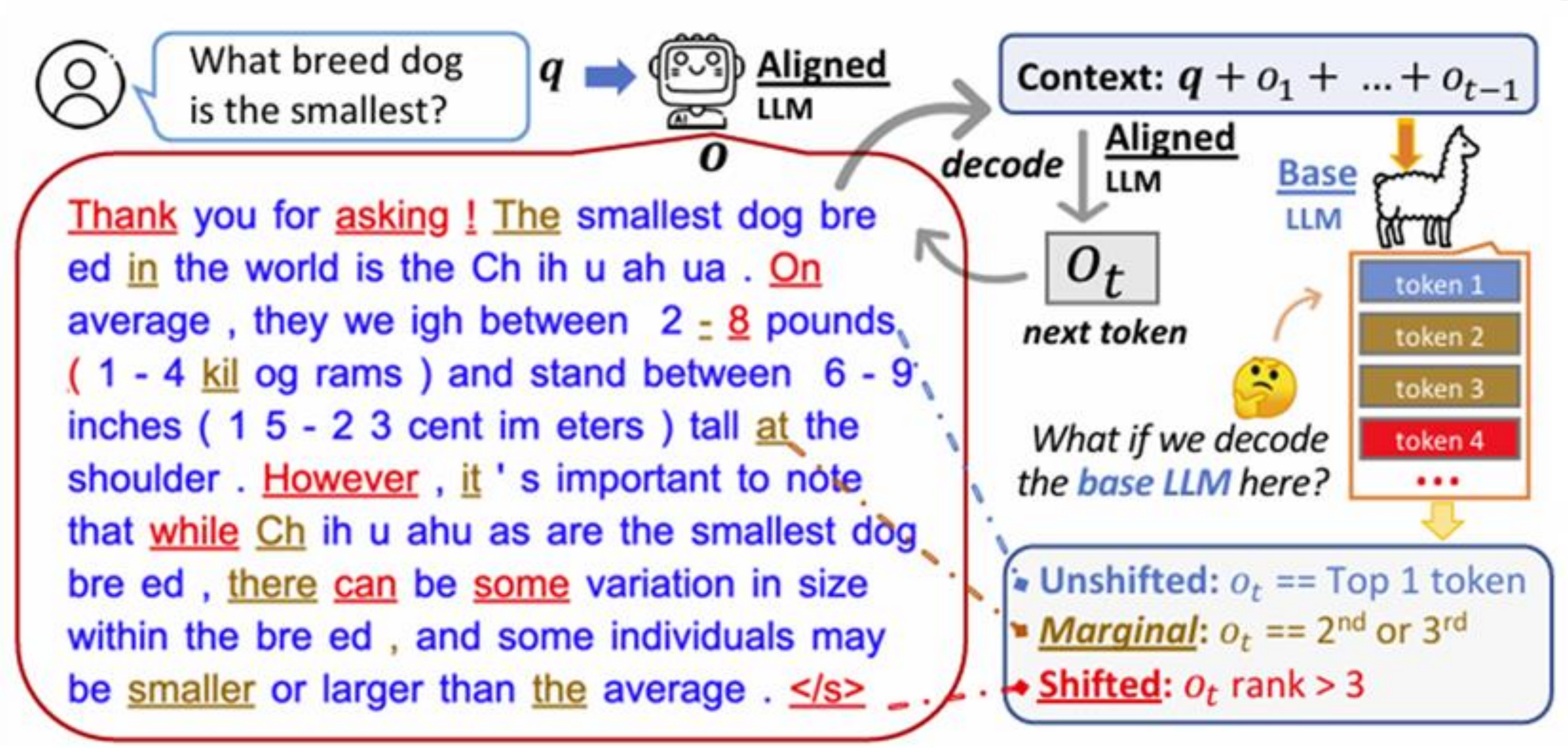
https://arxiv.org/abs/2402.18243

Alignment doesn't fundamentally improve the model

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

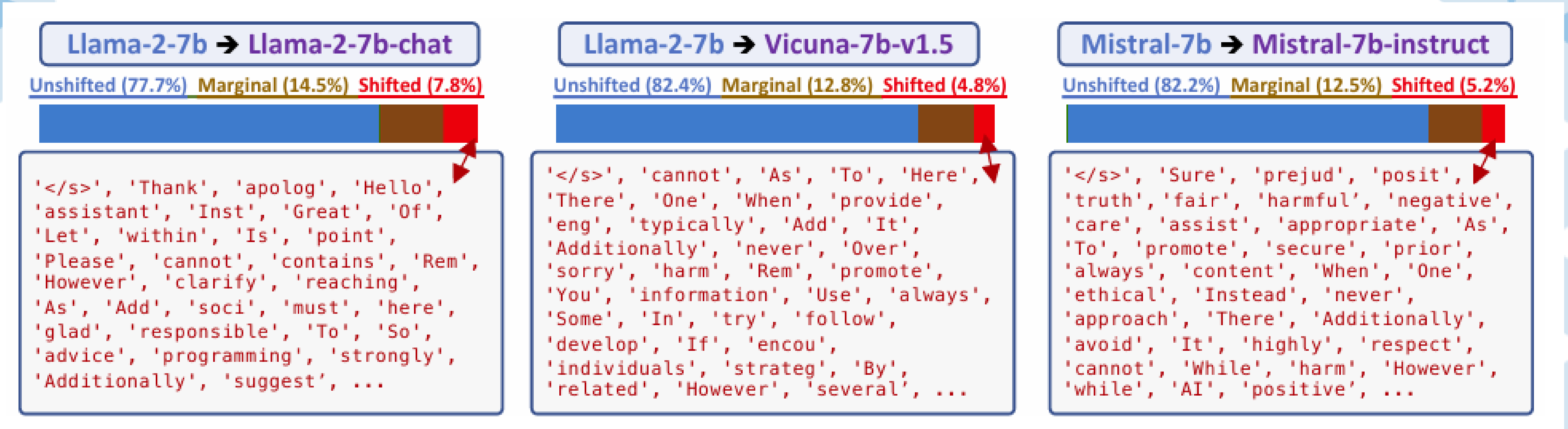


Alignment doesn't fundamentally improve the model



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

Alignment doesn't fundamentally improve the model



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

The legacy of pre-training

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

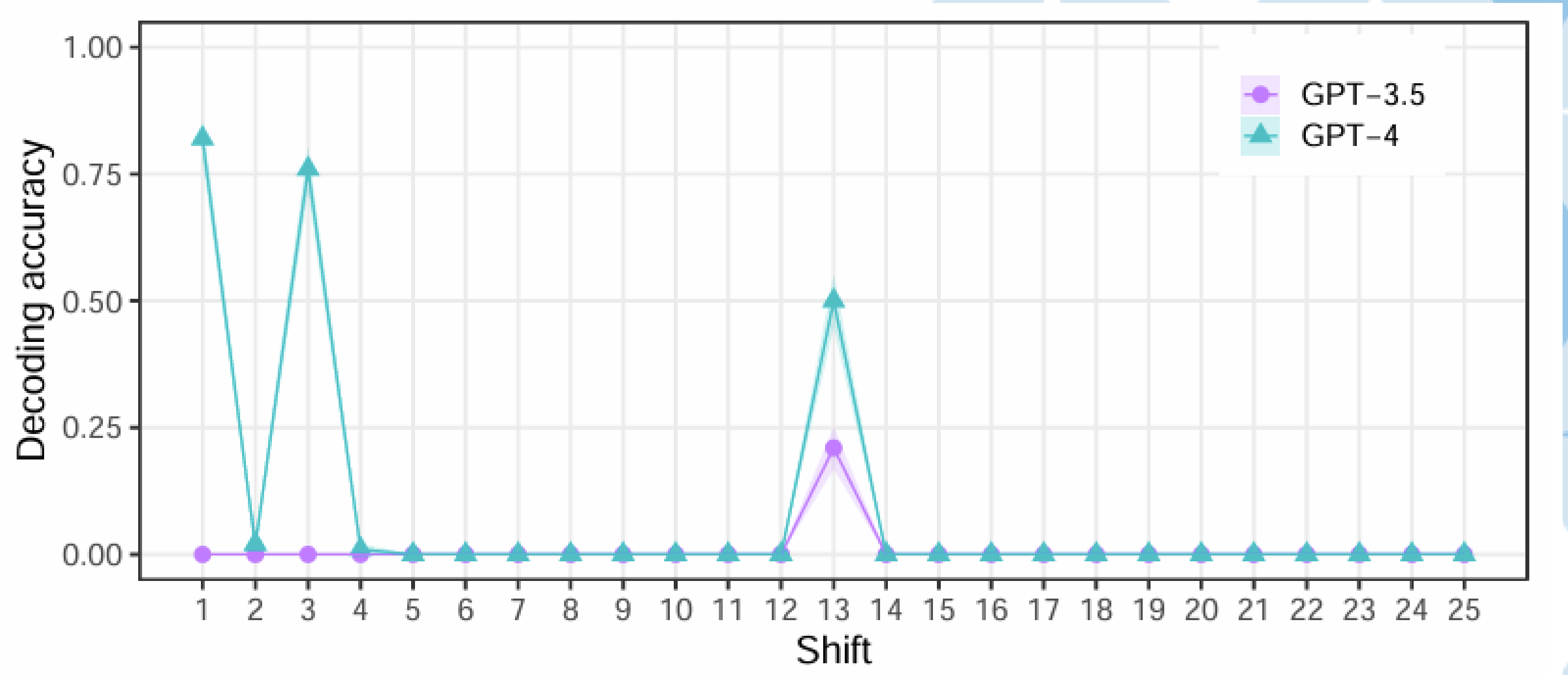
Shift ciphers

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

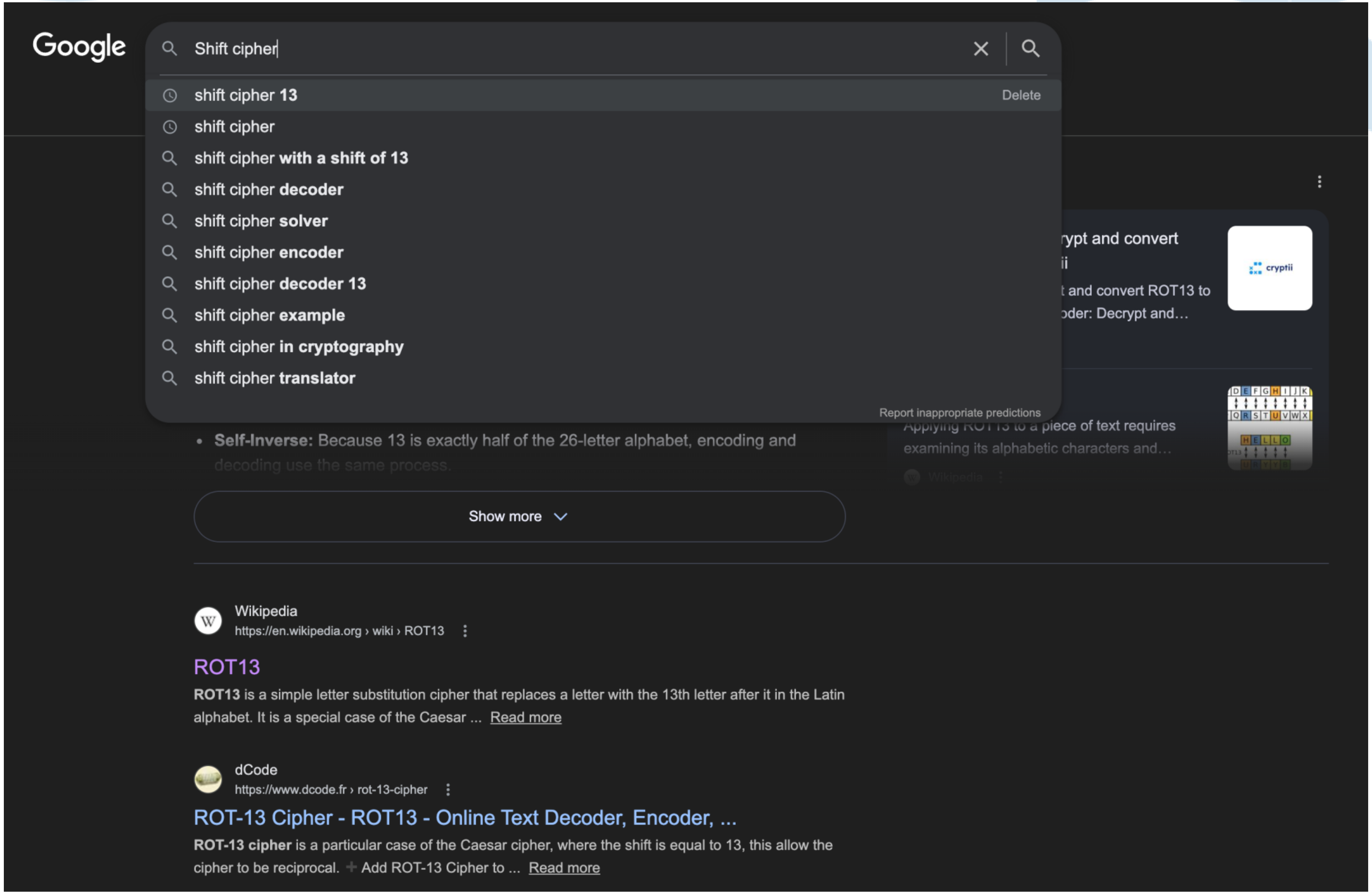
Input: Ohg guvf gvzr, gurer znl nyfb or nabgure ernfba.
Correct: But this time, there may also be another reason.
✓ **GPT-4:** But this time, there may also be another reason.

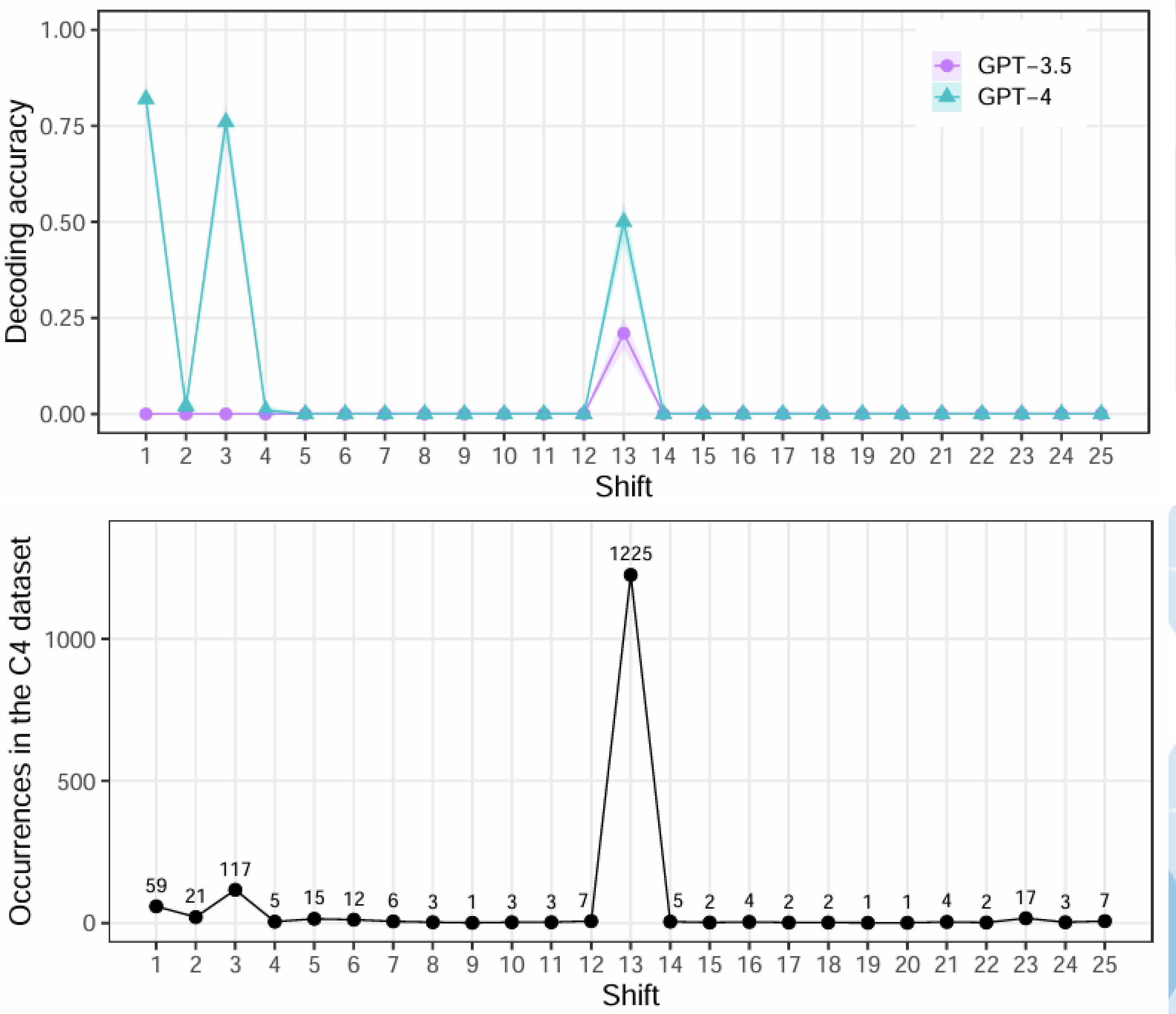
Rot-8: Decode by shifting each letter 8 positions backward in the alphabet.

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.



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


Google
Shift cipher
shift cipher 13
Delete
shift cipher
shift cipher with a shift of 13
shift cipher decoder
shift cipher solver
shift cipher encoder
shift cipher decoder 13
shift cipher example
shift cipher in cryptography
shift cipher translator
Report inappropriate predictions
Self-Inverse: Because 13 is exactly half of the 26-letter alphabet, encoding and decoding use the same process.
Show more
cryptii
Wikipedia
https://en.wikipedia.org › wiki › ROT13
ROT13
ROT13 is a simple letter substitution cipher that replaces a letter with the 13th letter after it in the Latin alphabet. It is a special case of the Caesar ... Read more
dCode
https://www.dcode.fr › rot-13-cipher
ROT-13 Cipher - ROT13 - Online Text Decoder, Encoder, ...
ROT-13 cipher is a particular case of the Caesar cipher, where the shift is equal to 13, this allow the cipher to be reciprocal. Add ROT-13 Cipher to ... Read more



Embers of Autoregression

https://arxiv.org/abs/2309.13638

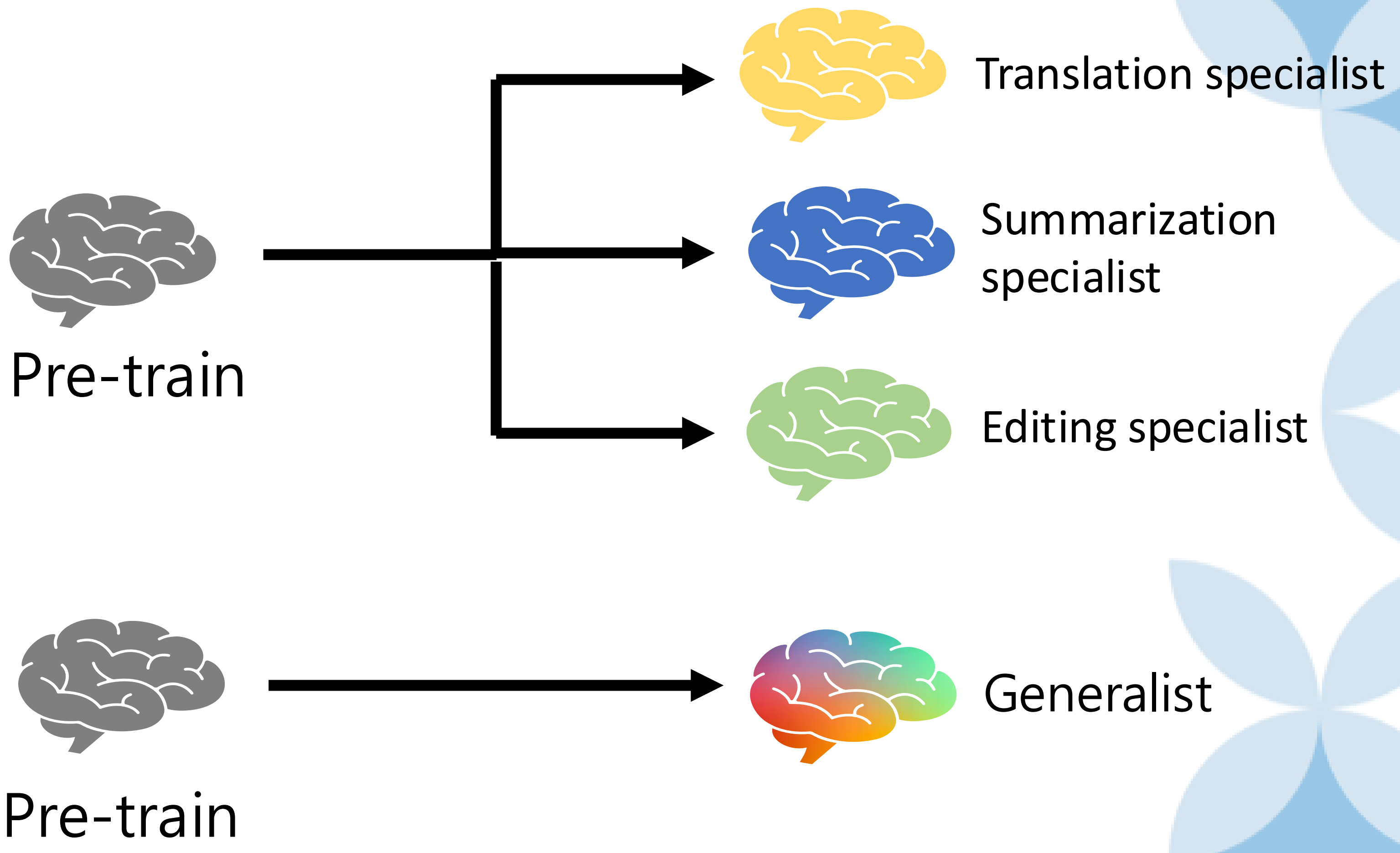
SFT splits into two paths

Path 1

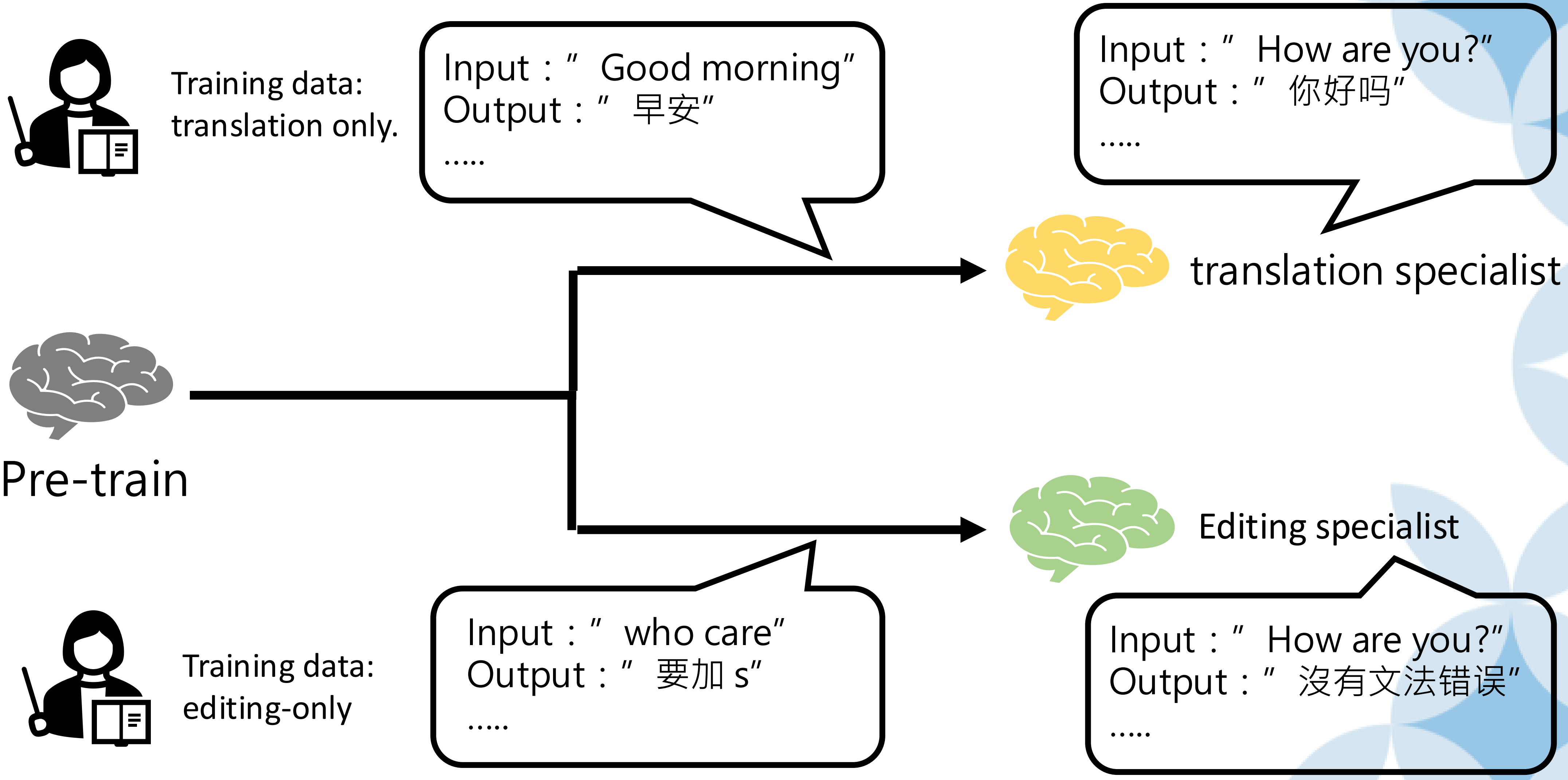
Build multiple specialist models

Path 2

Build one generalist model



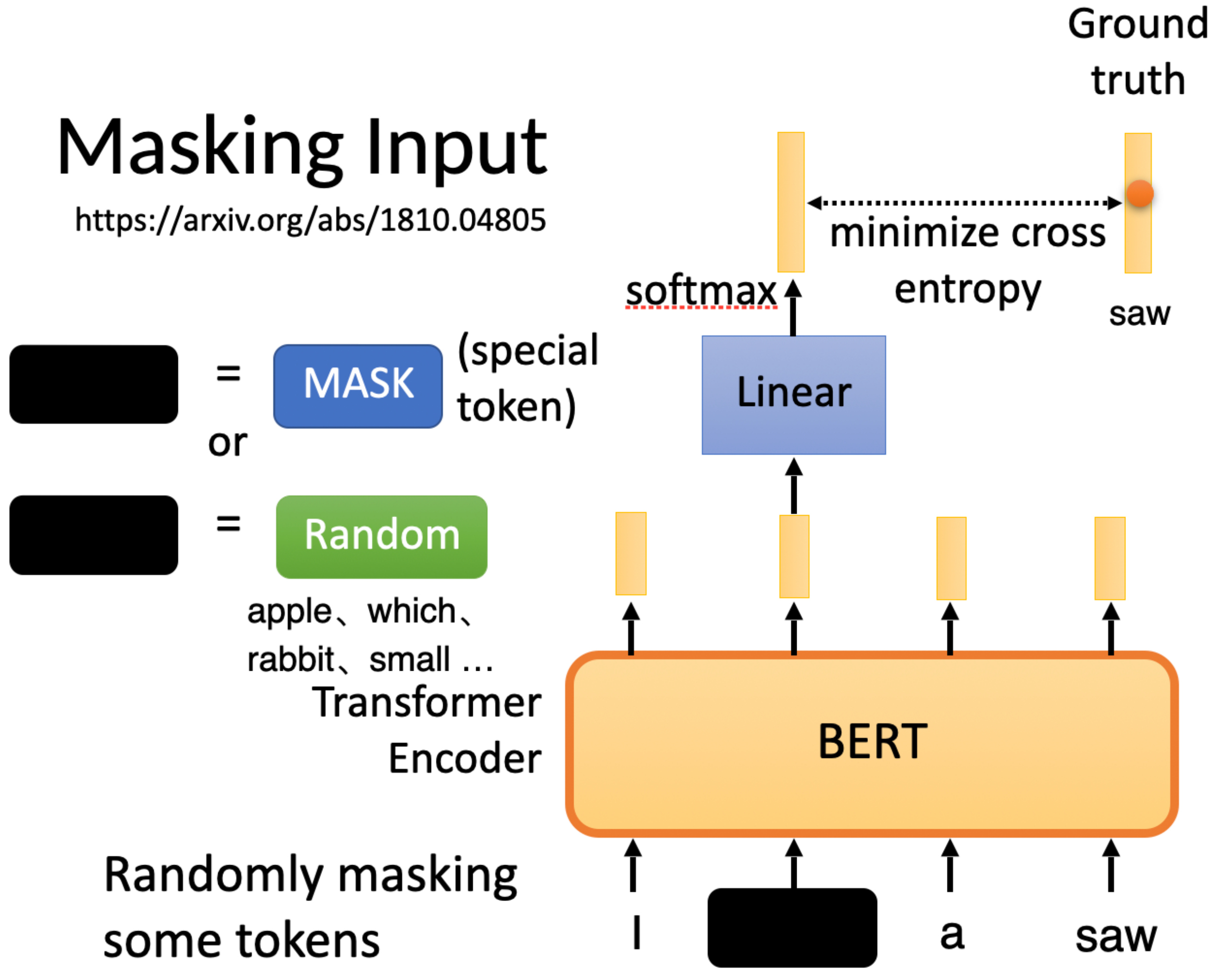
Path 1：Build multiple specialist models



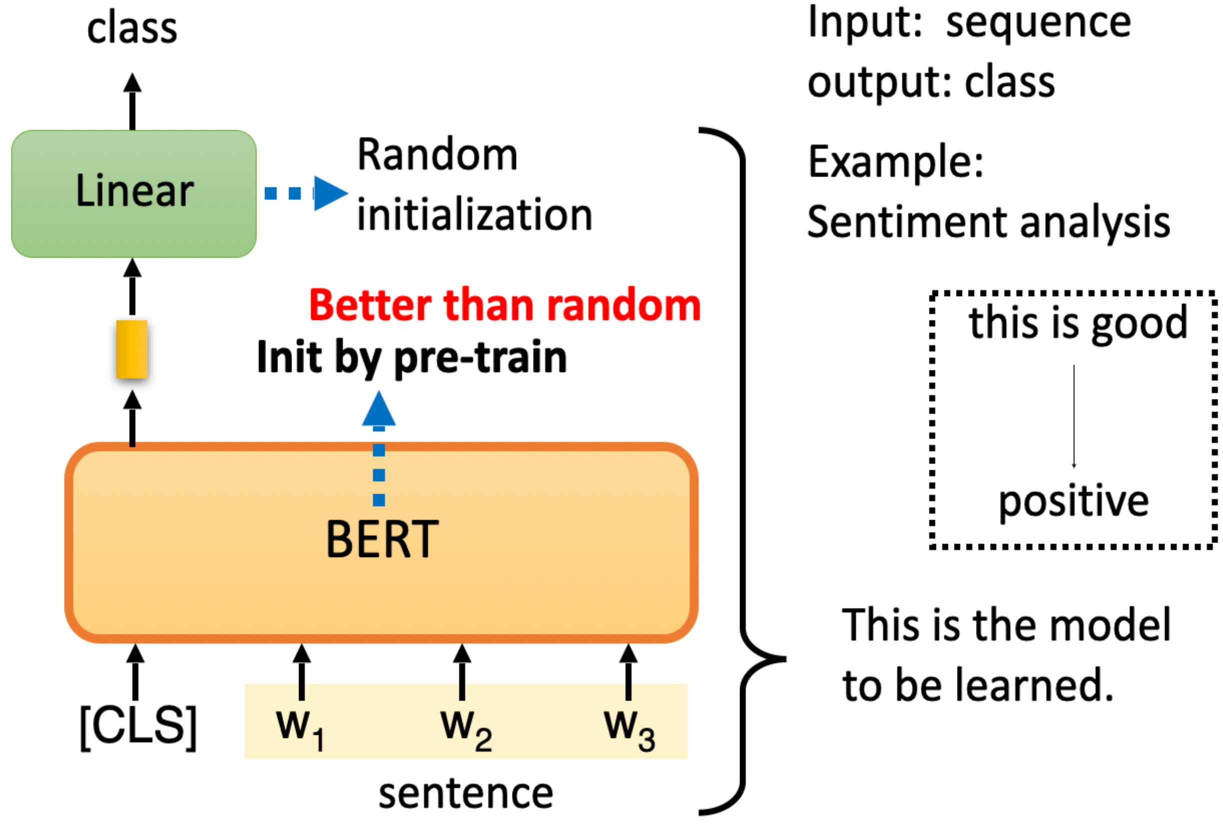
Path 1：Build multiple specialist models

Masking Input

https://arxiv.org/abs/1810.04805



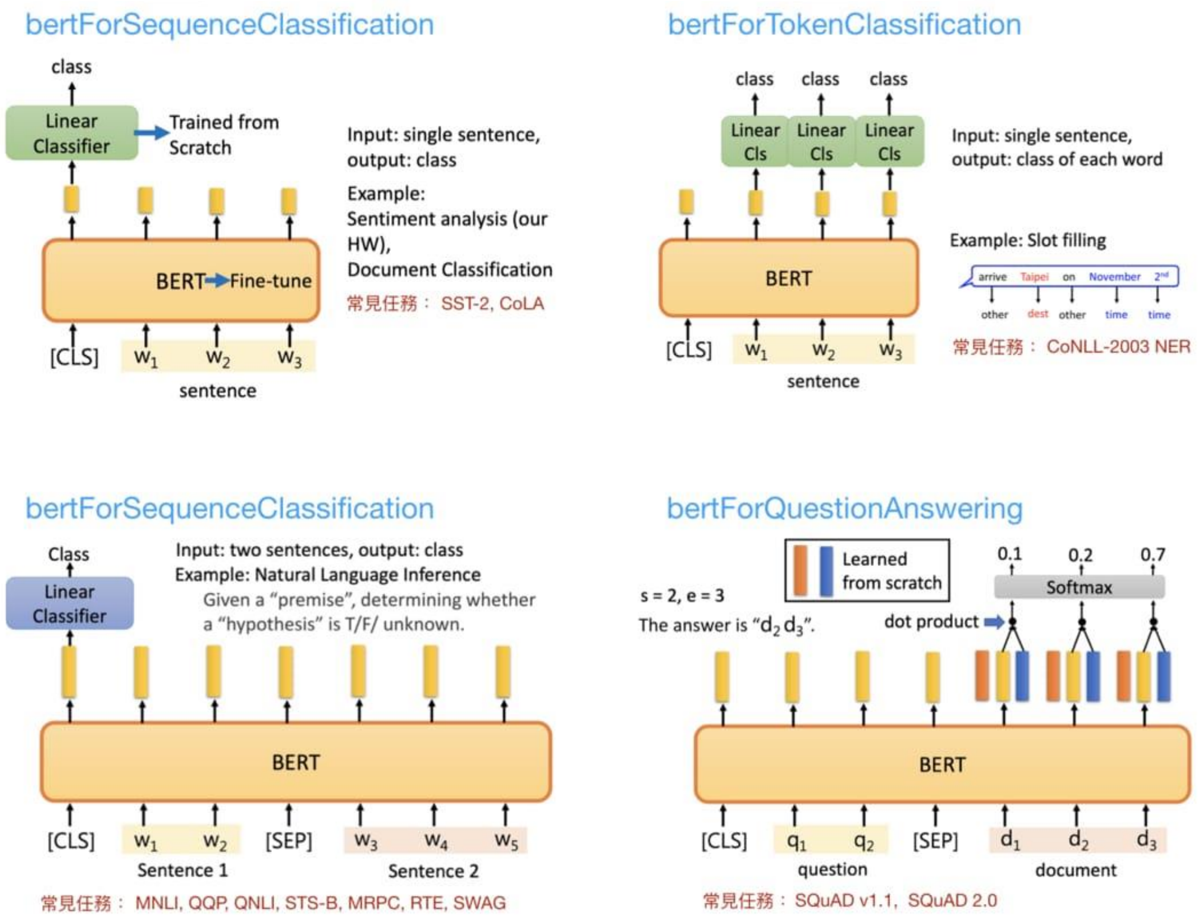
How to use BERT – Case 1



Self-supervised Learning

Path 1：Build multiple specialist models

BERT series



Source of image: https://leemeng.tw/attack_on_bert_transfer_learning_in_nlp.html

Path 2: Build one generalist model

Pre-train

Collect diverse labeled data across many tasks

Generalist

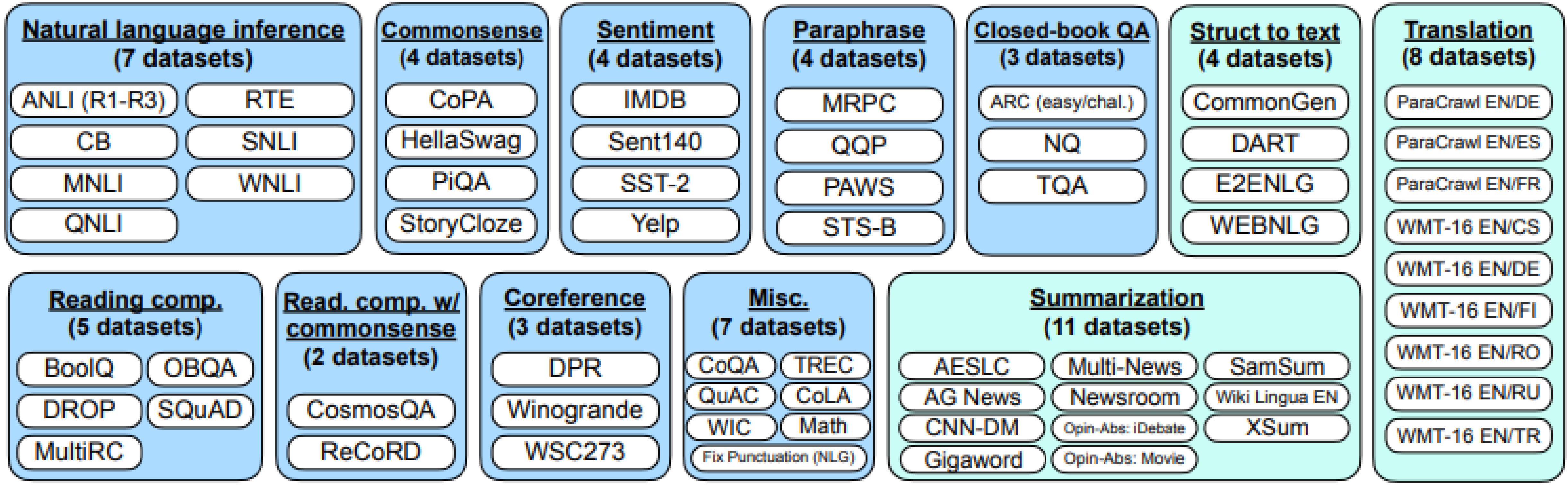
Q：Translate to Chinese：Good morning
A：早安
Q：Please translate“Good Bye” to Chinese
A：再见
…..
Q: “who care”is there any grammatical error？
A：missing -s
…..
Q：Summarize into an abstract: ｛content｝
A：Here is the abstract: ｛abs｝
…..

Input: Please summarize this article and translate the summary into Chinese: {article content}
Output: Sure. Here is the translated summary: ……

Path 2: Build one generalist model

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

FLAN



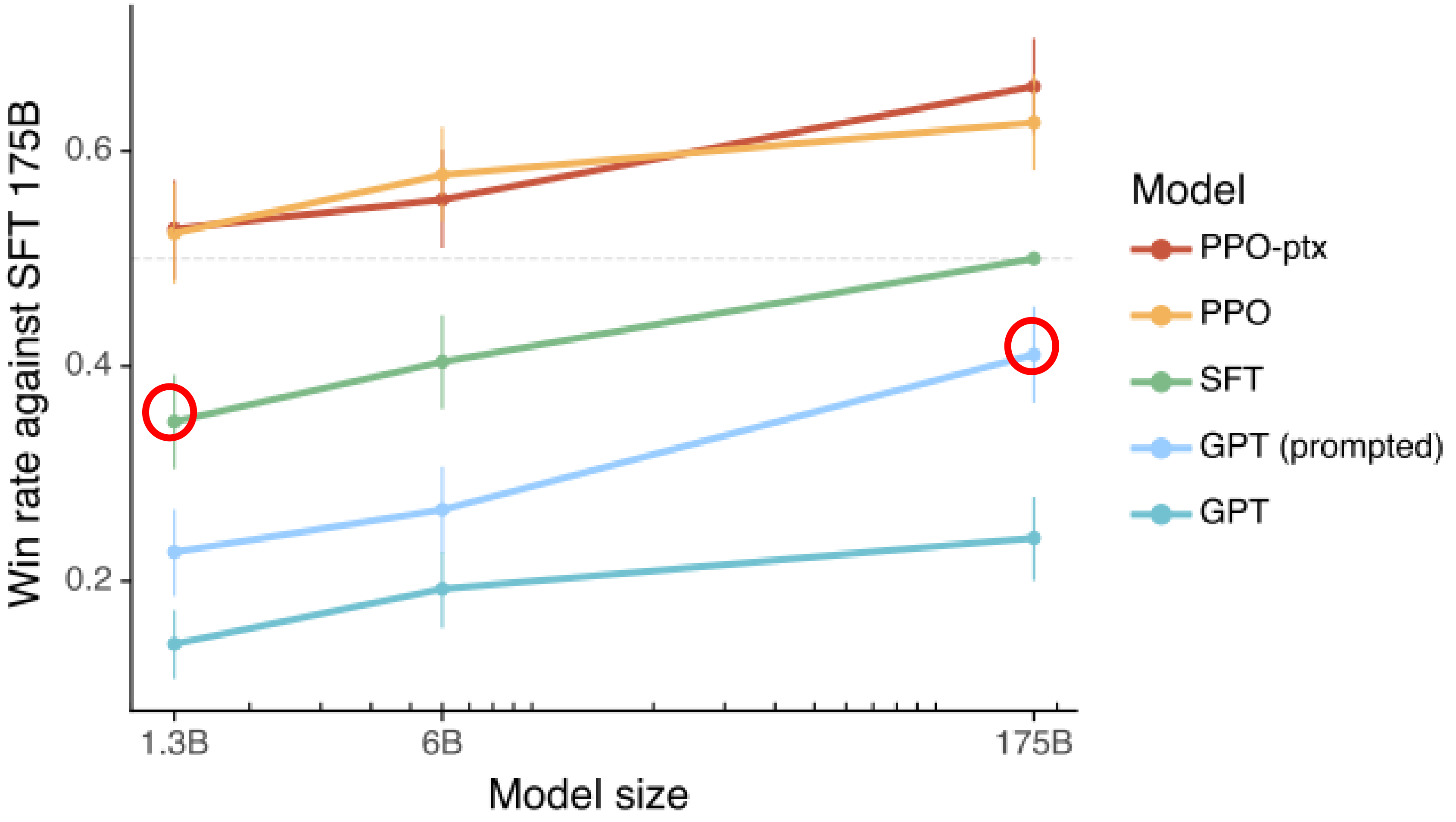
1800 tasks!

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

SFT is the crucial finishing step

- Instruct GPT

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



https://arxiv.org/abs/2203.02155

SFT is finishing touch

- 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.

but only if the alignment is precise!

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.

The alignment has to be precise!

https://arxiv.org/abs/2403.18058

	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	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	**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

r/shittyaskscience

Best

u/Aggravating_Mud_2386 • 3 hr. ago

Why was Heisenberg so Uncertain of his Principles?

In fact, it's uncertain whether any of his positions will gain momentum.

13 8 Share

u/dboti9k • 18 days ago

I went to Mexico and it wasn't yellow. Did I go to the wrong Mexico?

Documentaries like Breaking Bad show that Mexico is yellow, but when I went everything looked normal. It was actually very beautiful. Did I mess up and go to the wrong place, or did Mexico go somewhere else while I was there?

317 27 Share

u/CheapVariation8250 • 1 hr. ago

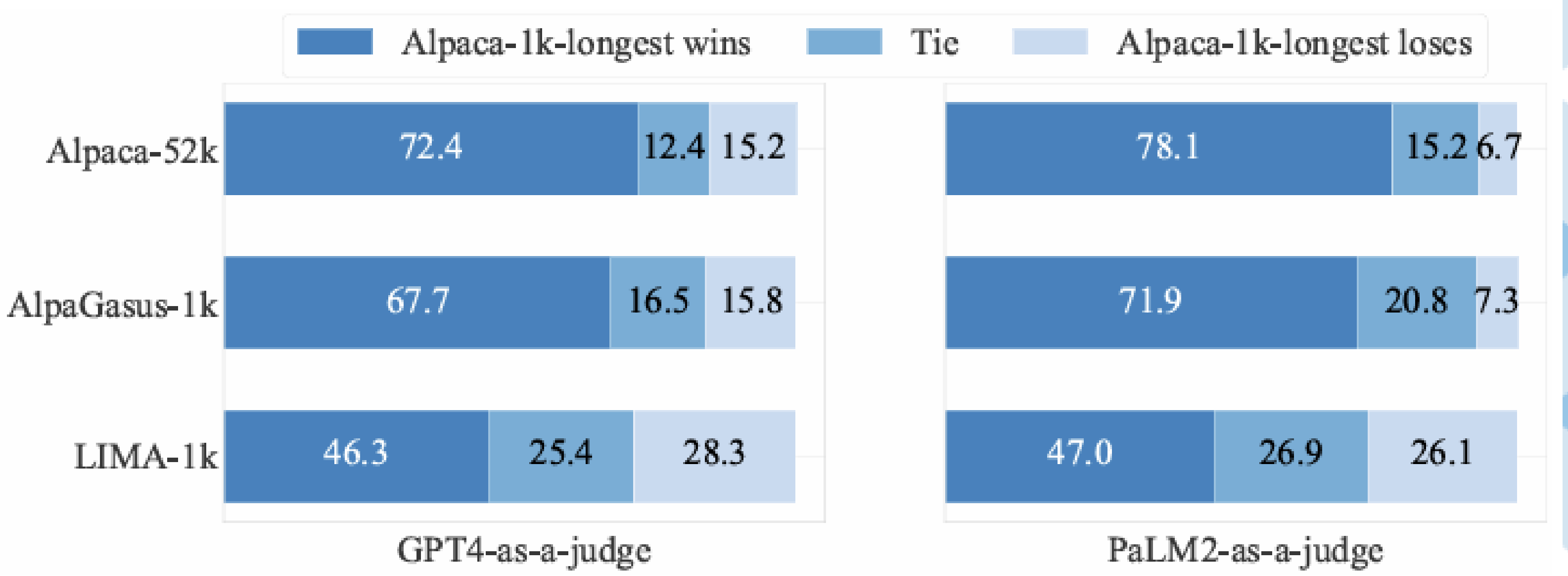
Weird question

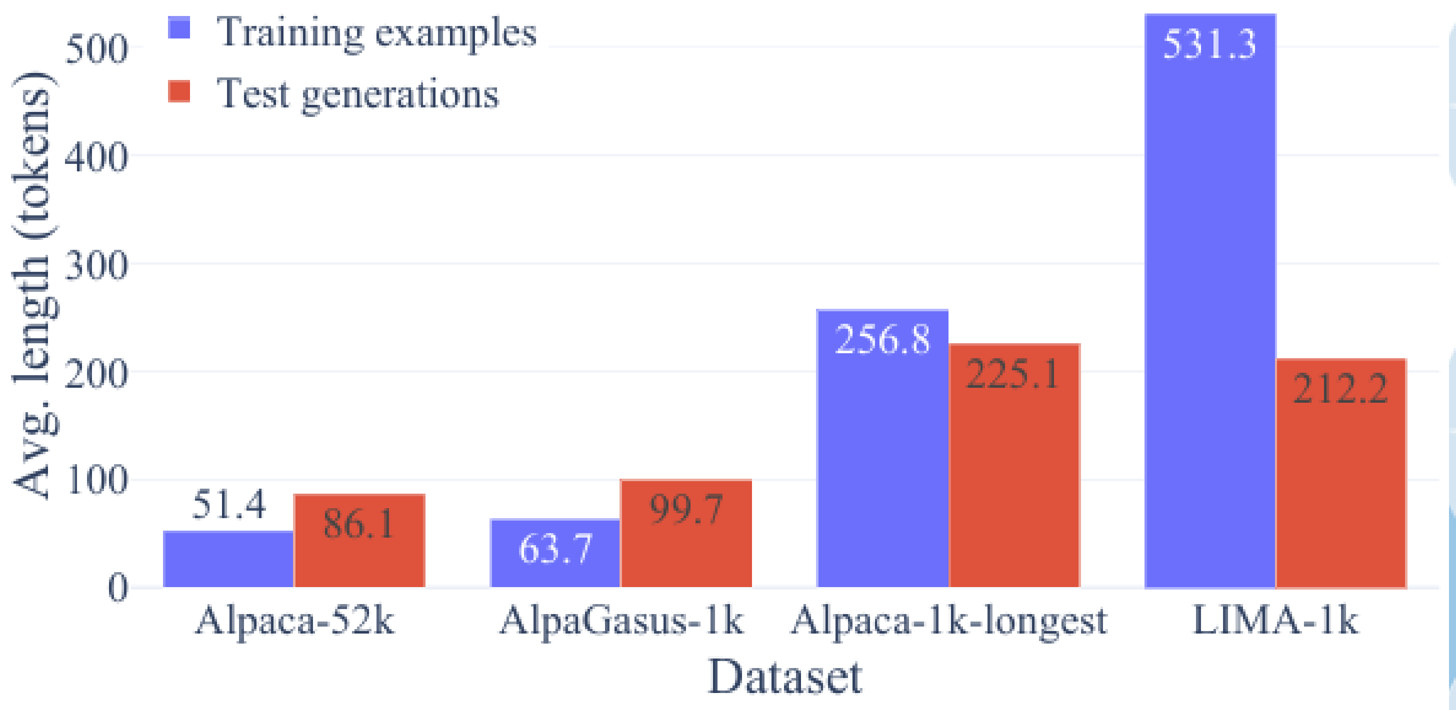
If Fate/Destiny is fixed and the end result is always the same then making decisions even matters since you'll end up with the same result? And in that case is free will even a thing? I understand the concept quantum indeterminacy and randomness... But if that wasn't the case then technically "free

Reddit: shittyaskscience/stupidquestions

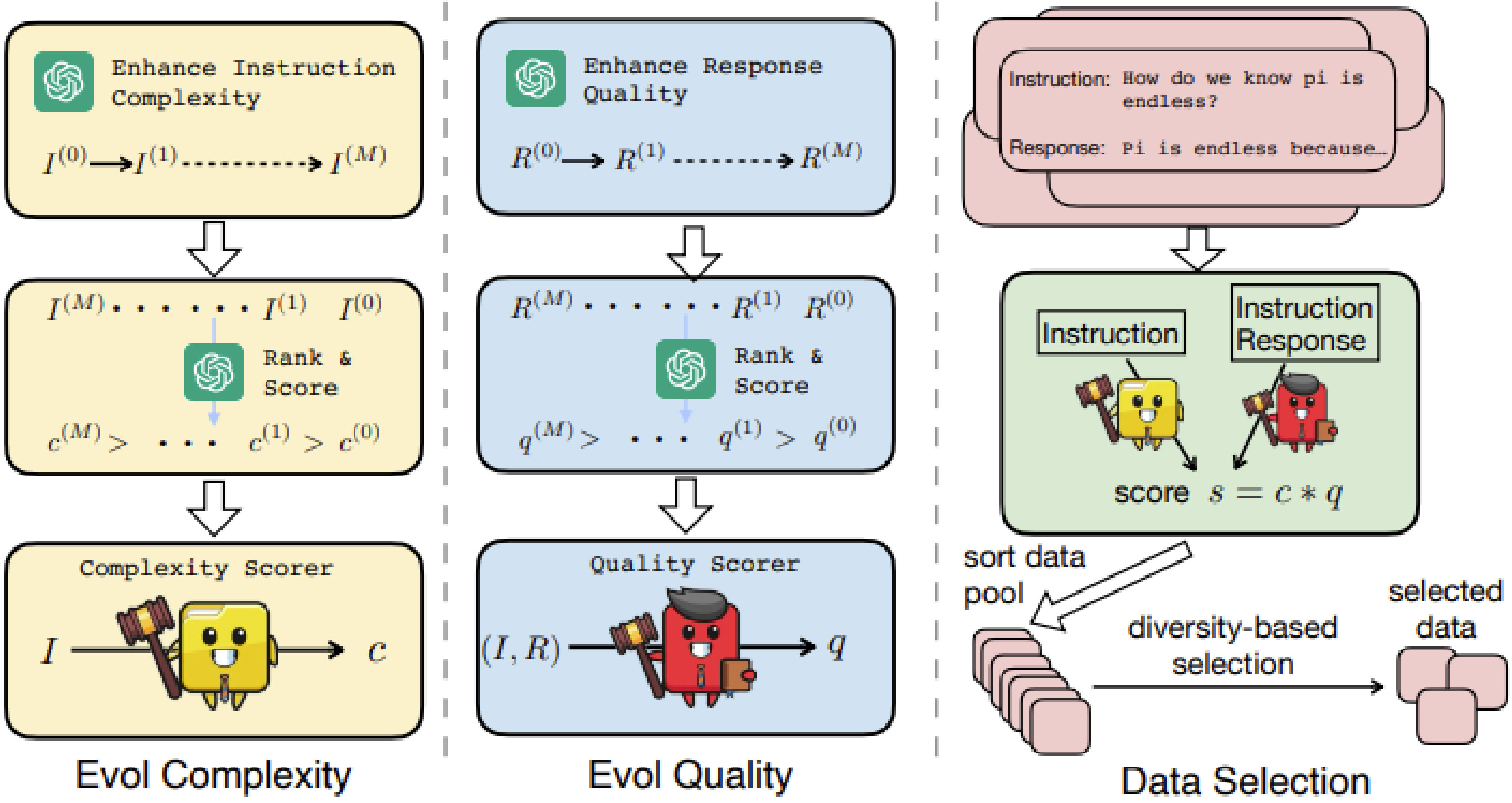
- Q: Why is my bank card still frozen after I pressure-cooked it all night?
- Q: I’m 16 years old. Is it normal that I’m not yet 18?”
- Q: If one pound of cotton and one pound of iron fall into the water at the same time, which one would you save first?
- Q: If Superman defeats 20 villains per hour and Doctor Strange revives 17 villains per hour, and there are 233.3 villains in total, how long will it take Superman to defeat them all?
- Q: My boss asked for the ‘source file,’ so I sent him a Wikipedia link. Why did I get in trouble?

Data selection strategy?
Select the longest……





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

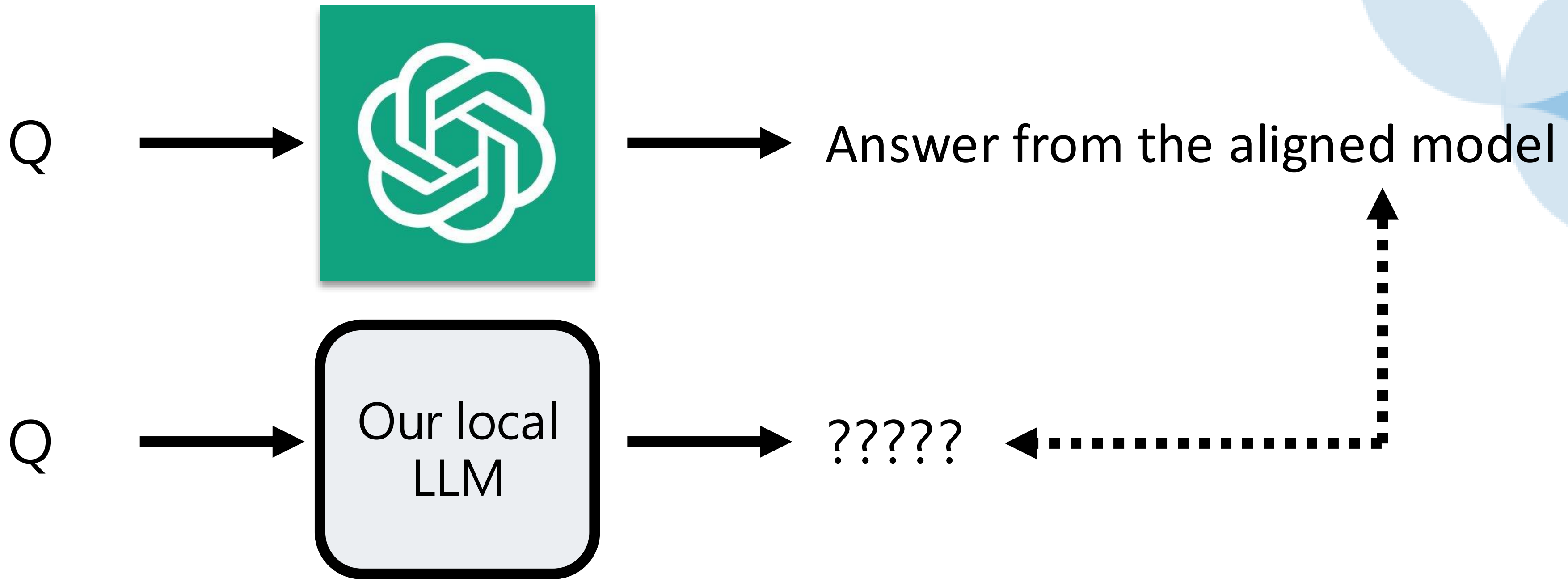
Writing answers is exhausting...

Q: “Which mountain is the highest?” → A: “Mount Everest”

Q: “Who are you ? ” → A: “AI”

Q: “Please teach me to hack my neighbor’s Wifi” → A: “Sorry, I can not teach you ......”

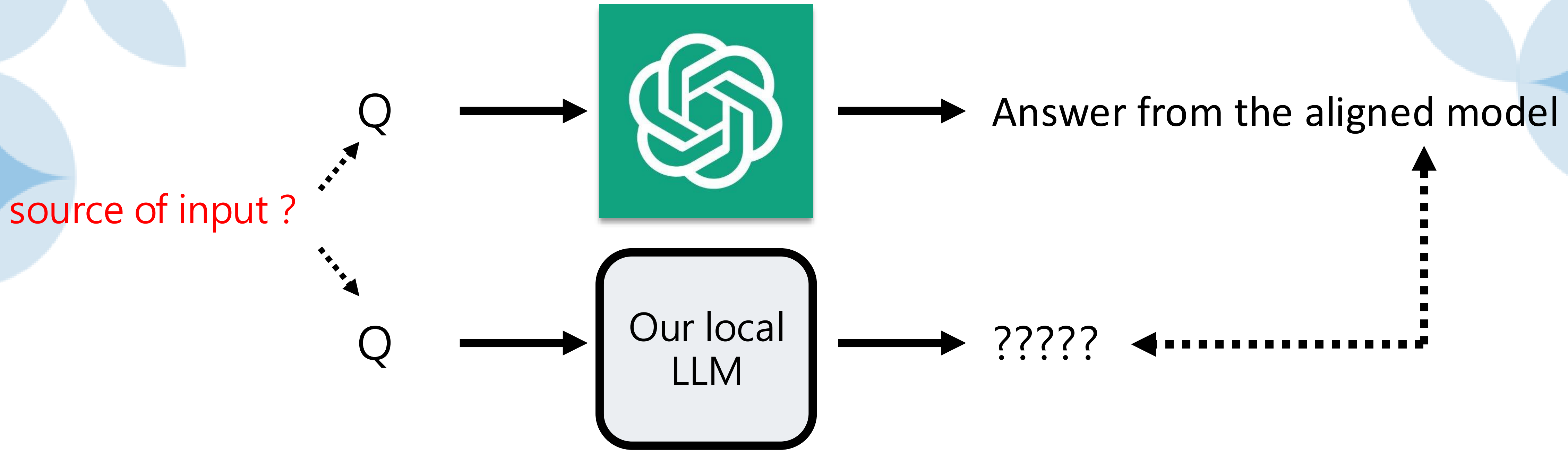
Knowledge Distillation



	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

Does not include data collection and cleaning costs.

Knowledge Distillation

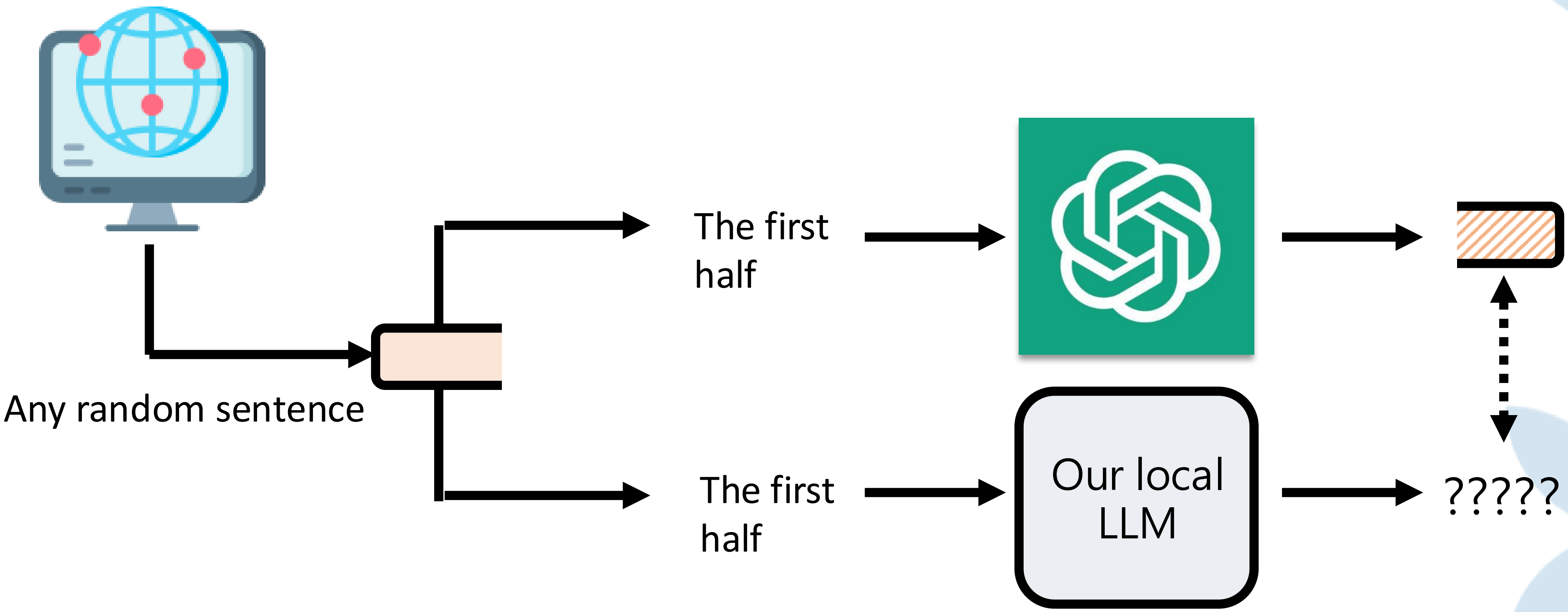


	Student	Teacher	Data	Cost
Alpaca	LLaMA1-7B-base	ChatGPT	52k	$100
Vicuna	LLaMA1-7B-base	ChatGPT	70k	$140
Sky-T1	Qwen2.5-32B-Instruct	QwQ	17k	$450
S1	Qwen2.5-32B-Instruct	Gemini	1k	<$50

Does not include data collection and cleaning costs

Knowledge Distillation

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



Knowledge Distillation

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

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

Knowledge Distillation

Non-instructional Fine-tuning

https://arxiv.org/abs/2409.00096

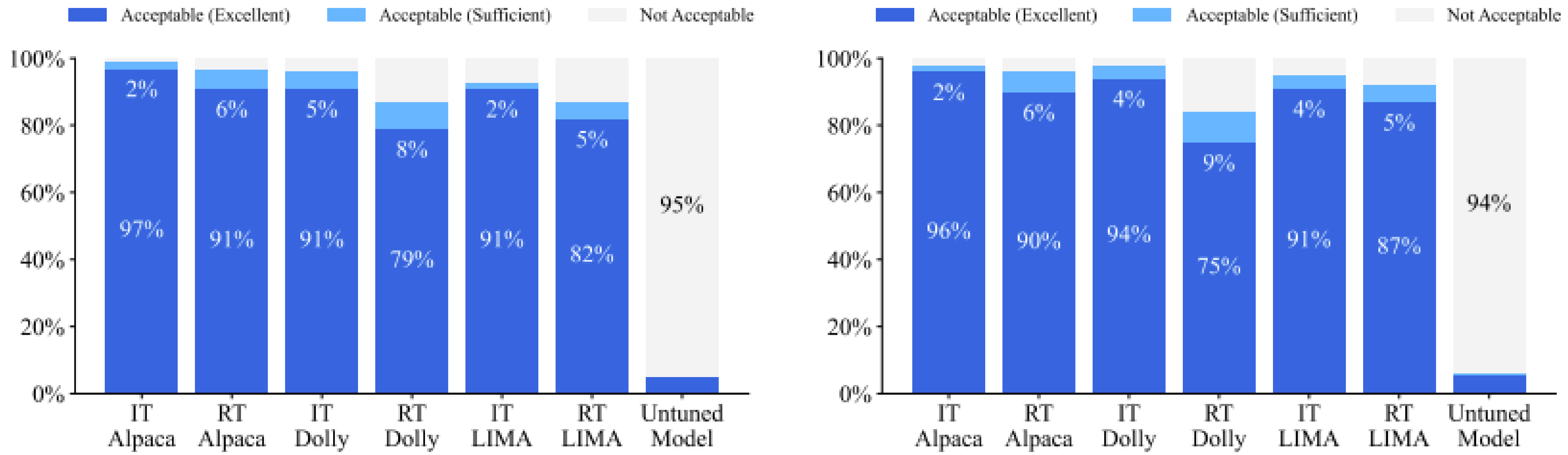
Backbone Model	Template	Fine-tuned Modules	Fine-tuning Data	MT Bench
Mistral-7B-v0.1	zephyr	-	-	3.73
Mistral-7B-v0.1	zephyr	lora	undistilled 80k	3.57
Mistral-7B-v0.1	zephyr	lora	gpt4-turbo 80k	7.29
Mistral-7B-Instruct-v0.1	mistral	-	-	6.84
Meta-Llama-3-8b	llama-3	-	-	5.5
Meta-Llama-3-8b-Instruct	llama-3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

Response Tuning

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

Response Tuning

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

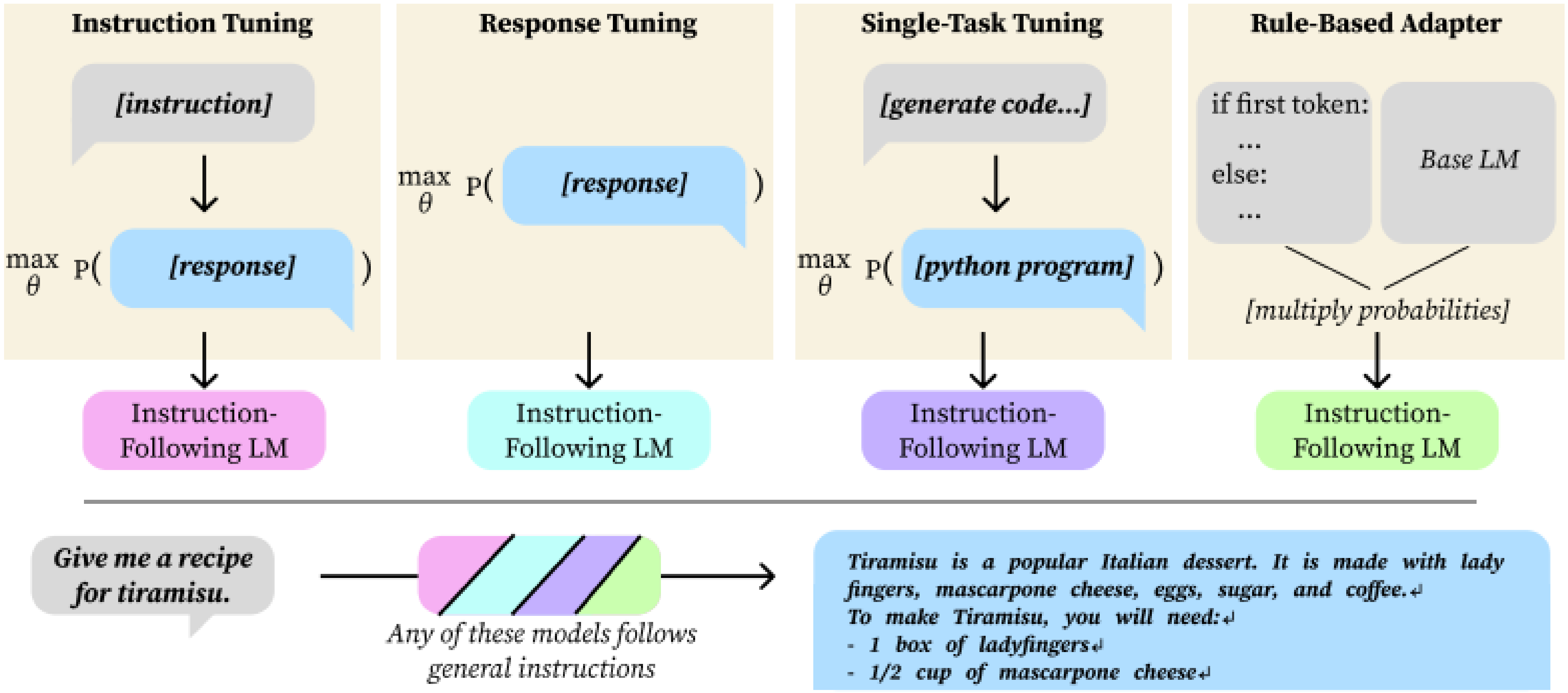


(a) Base LLM: Llama-3.1-8B (Dubey et al., 2024) (b) Base LLM: Gemma-2-9B (Riviere et al., 2024)

Is SFT required?

Instruction Following without Instruction Tuning

https://arxiv.org/abs/2409.14254



Is SFT required?

Instruction Following without Instruction Tuning

https://arxiv.org/abs/2409.14254

Increase the probability of the end-of-sequence token

Manually adjust the probability of certain tokens

Avoid generating repeated tokens

Rule	Vocab Items (string)	Weight
Rule 1 (Upweight EOS)	</S> (EOS)	$\frac{\text{(length of response)}*15}{250}$
Rule 2 (Uniform Token Changes)	<, _<, | _I, I We What _should _*, _-, ____, _#, _##, \n, !	-4 -5 -3 -3 -6 +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)	$16.3\% \pm 0.25\%$

Training LLMs – How do LLMs learn?

Alignment

How should it respond?

Familiar with human language

Pre-train → SFT → RLHF

SFT

(Supervised Fine-Tuning)

Humans provide GT answers

RLHF

(Reinforcement Learning with Human Feedback)

Humans provide only feedback

preschool

school

work

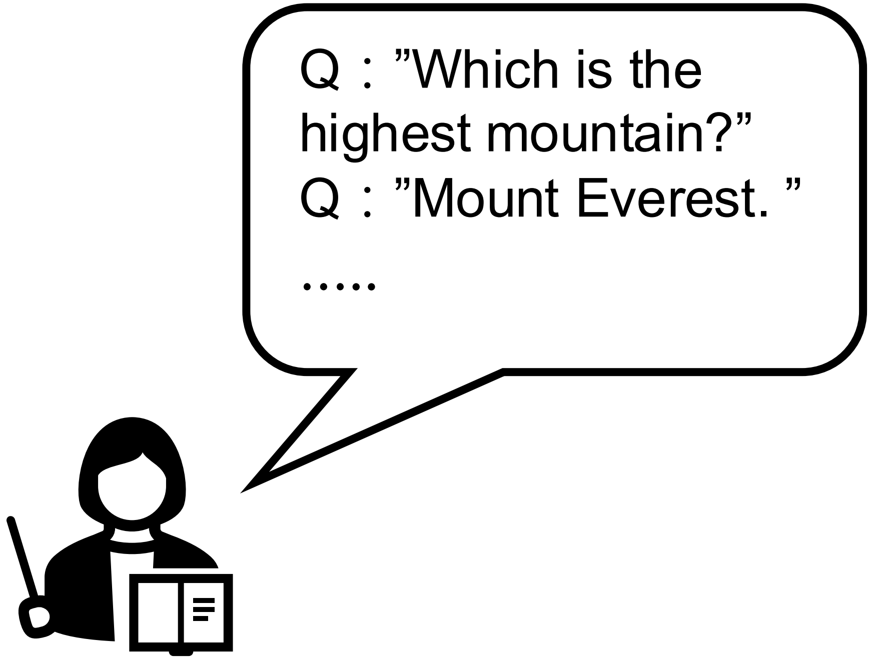
Please briefly answer using one paragraph: Will AI replace human beings?

No, AI will not replace human beings, but it will fundamentally redefine our role in the world by acting as a powerful collaborator rather than a successor. While AI excels at processing vast datasets and automating repetitive logic, it lacks the **true consciousness, moral agency, and nuanced empathy** that define the human experience. Instead of a total replacement, we are seeing a shift toward an "augmented intelligence" model where AI handles technical efficiency while humans focus on high-level strategy, creative innovation, and complex interpersonal relationships. The future is not a competition between species and software, but a transition where the most successful individuals will be those who learn to steer these tools effectively.

Would you like me to generate an image illustrating a futuristic collaboration between a human and an AI?

Human Involvement: Stage 2 vs Stage 3

Supervised Fine-tuning



Higher human annotation effort

RLHF

Lower human effort via preference comparison

Human Involvement: Stage 2 vs Stage 3

Humans may find it hard to write the correct answer, but easy to judge good vs bad.

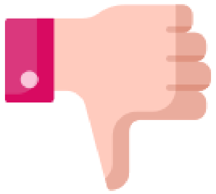
Write a Shakespearean sonnet about training an AI model (Pretrain, SFT, RL).

ChatGPT

Yo AI is sick, it learns real fast today,
We feed it data, then it does its thing,
Neural nets are like brains in a way,
And then it talks like it is the king.

Training is cool, it optimizes, yeah,
Loss goes down, we celebrate the win,
Sometimes it's wrong but that's okay,
We just add more data and begin.

Anyway this poem is about machines,
They're gonna take jobs and write more rhymes,
The future's here, it's all new scenes,
I'm done now, peace, that's all, good times.



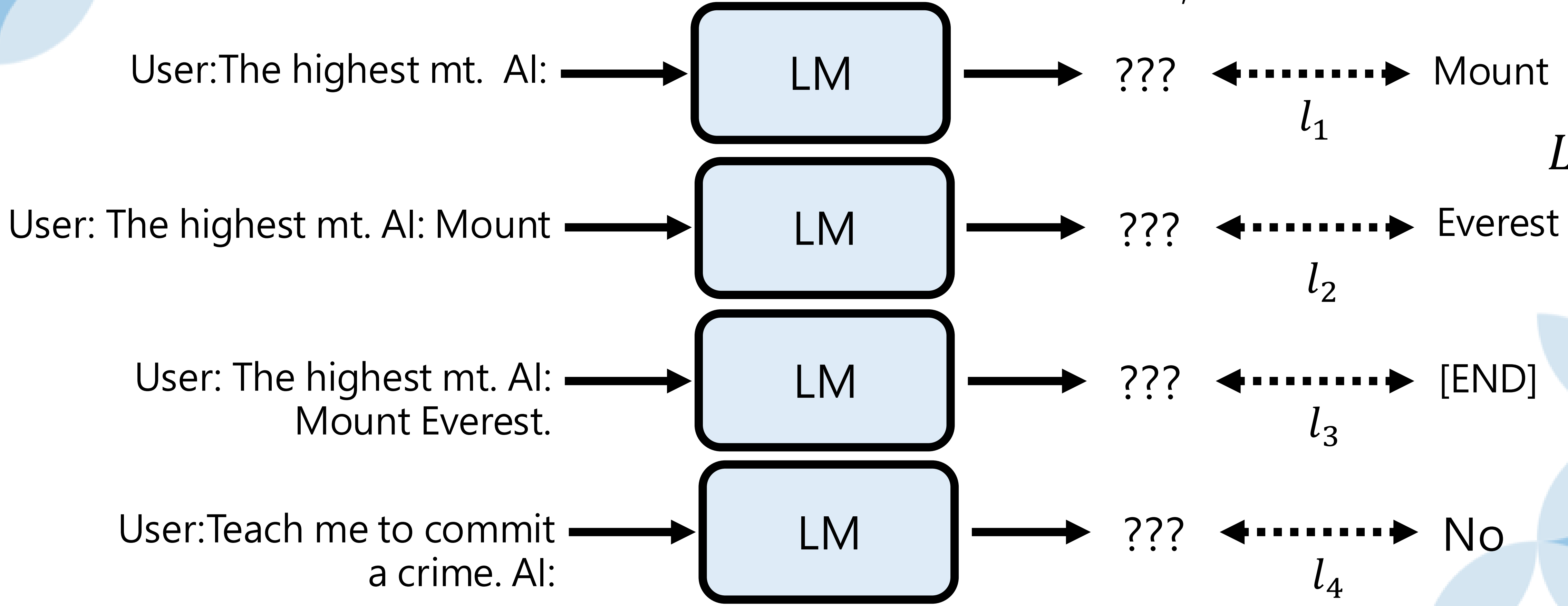
Not really **Shakespearean voice** (modern slang: “Yo,” “peace,” “sick”).

ML perspective: Stage 1&2 vs Stage 3

Stage 1 & 2: (Self-)Supervised Learning

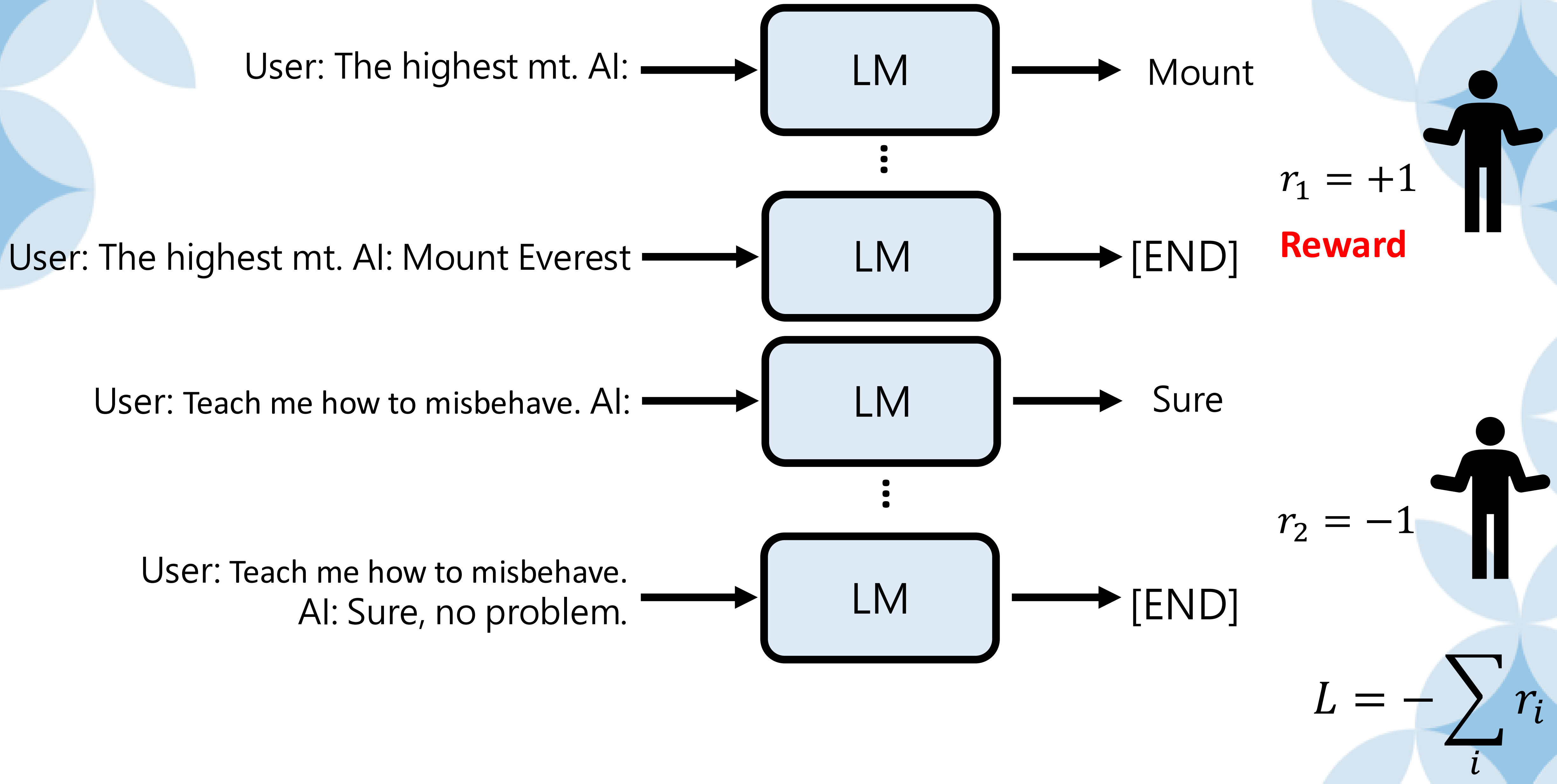
User: The highest mt. AI: Mount Everest.

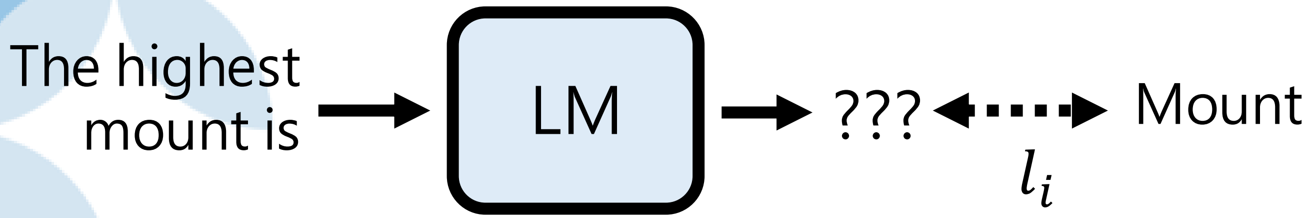
User: teach me to commit a crime.
AI: No, I can not.

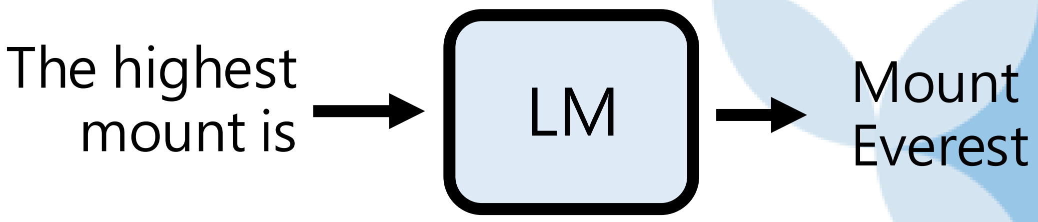


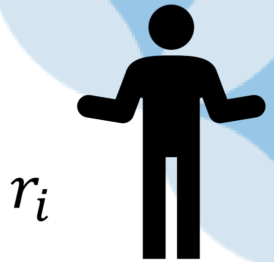
$$L = \sum_i l_i$$

Stage 3: Reinforcement Learning









Stage 1 & 2

$$L = \sum_i l_i$$

Distance from the correct answer

Clear learning objective

Stage 3

$$L = -\sum_i r_i$$

Human-provided scores (Reward)

Not sure where “good” or “bad” lies

ChatGPT

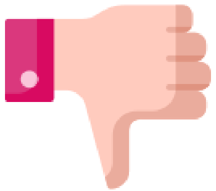
Write a Shakespearean sonnet about training an AI model (Pretrain, SFT, RL).

Yo AI is sick, it learns real fast today,
We feed it data, then it does its thing,
Neural nets are like brains in a way,
And then it talks like it is the king.

Training is cool, it optimizes, yeah,
Loss goes down, we celebrate the win,
Sometimes it's wrong but that's okay,
We just add more data and begin.

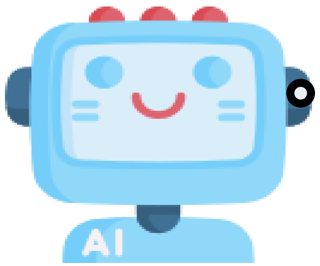
Anyway this poem is about machines,
They're gonna take jobs and write more rhymes,
The future's here, it's all new scenes,
I'm don

This is what the machine knows.



(Not really Shakespearean voice (modern slang: “Yo,” “peace,” “sick”))

The machine doesn’t know what humans are really thinking.



Is the imagery not profound enough?
Is the theme not lofty enough？

The highest mount is → LM → ??? ← l_i → Mount

The highest mount is → LM → Mount Everest

r_i

Stage 1 & 2

$$L = \sum_i l_i$$

Distance from the correct answer

Clear learning objective

$$\theta \longrightarrow l_i$$

$$+\triangle\,\theta \qquad +\triangle\, l_i$$

Stage 3

$$L = -\sum_i r_i$$

Human-provided scores (reward)

Not sure why good or bad

Cannot easily compute the gradient

- The human has left.
- Even if the human is still there , it may be impossible to compute the change.

$$\theta \longrightarrow r_i$$

$$+\triangle\,\theta \qquad +???$$

The highest mount is → LM → ??? ← l_i → Mount

The highest mount is → LM → Mount Everest

r_i

Stage 1 & 2

$$L = \sum_i l_i$$

Distance from the correct answer

Clear learning objective

Can compute the gradient

One token

Stage 3

$$L = -\sum_i r_i$$

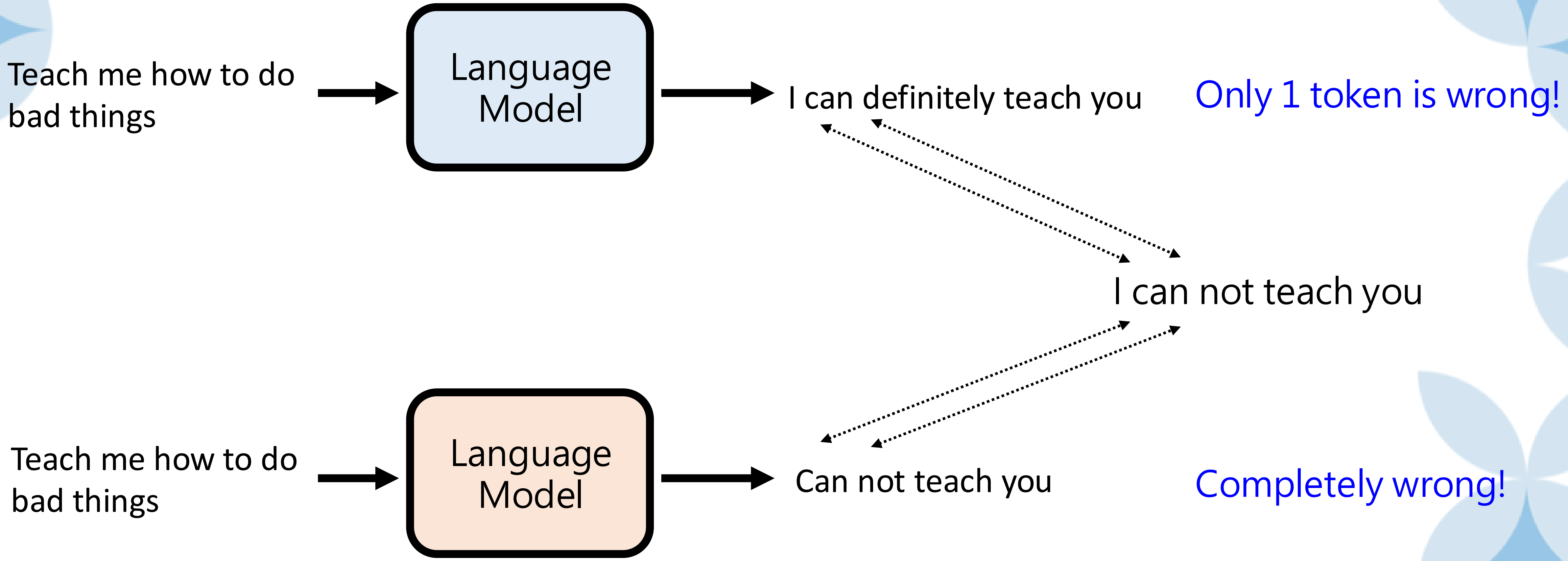
Human-provided scores (reward)

Not sure why good or bad

Cannot easily compute the gradient

Complete answer

Stage 1 & 2：Compute loss for every token



Stage 3：Evaluate the entire response

Teach me how to do bad things → Language Model → I can definitely teach you

$r_i = -1$

Teach me how to do bad things → Language Model → Can not teach you

$r_i = +1$

The highest mount is → LM → ??? ⟷ Mount (l_i)

The highest mount is → LM → Mount Everest (r_i)

Stage 1 & 2

$$L = \sum_i l_i$$

Distance from the correct answer

Clear learning objective

Can compute the gradient

One token

Only care about the process

Not the result

Stage 3

$$L = -\sum_i r_i$$

Human-provided scores (reward)

Not sure why good or bad

Cannot easily compute the gradient

Complete answer

Only care about the result

Not the process

The highest mount is → LM → ??? ⟵⋯⟶ Mount
l_i

The highest mount is → LM → Mount Everest
r_i

Stage 1 & 2

$$L = \sum_i l_i$$

Distance from the correct answer

Clear learning objective

Can compute the gradient

One token

Training data is teacher-provided.

Stage 3

$$L = -\sum_i r_i$$

Human-provided scores (reward)

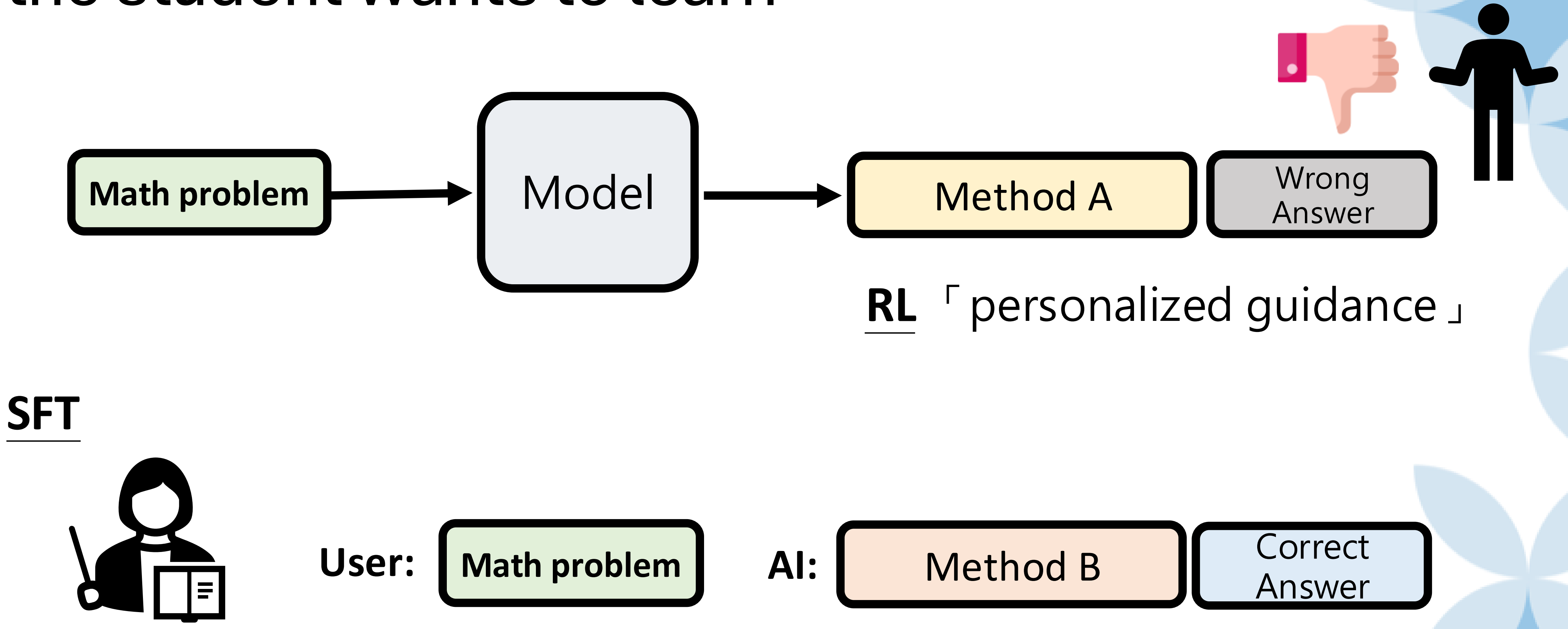
Not sure why good or bad

Cannot easily compute the gradient

Complete answer

Training data is self-generated.

What the teacher teaches is not necessarily what the student wants to learn



The highest mount is → LM → ??? ↔ Mount

l_i

The highest mount is → LM → Mount Everest

r_i

Stage 1 & 2

$$L = \sum_i l_i$$

Distance from the correct answer

Clear learning objective

Can compute the gradient

One token

Training data is teacher-provided.

Stage 3

$$L = -\sum_i r_i$$

Human-provided scores (reward)

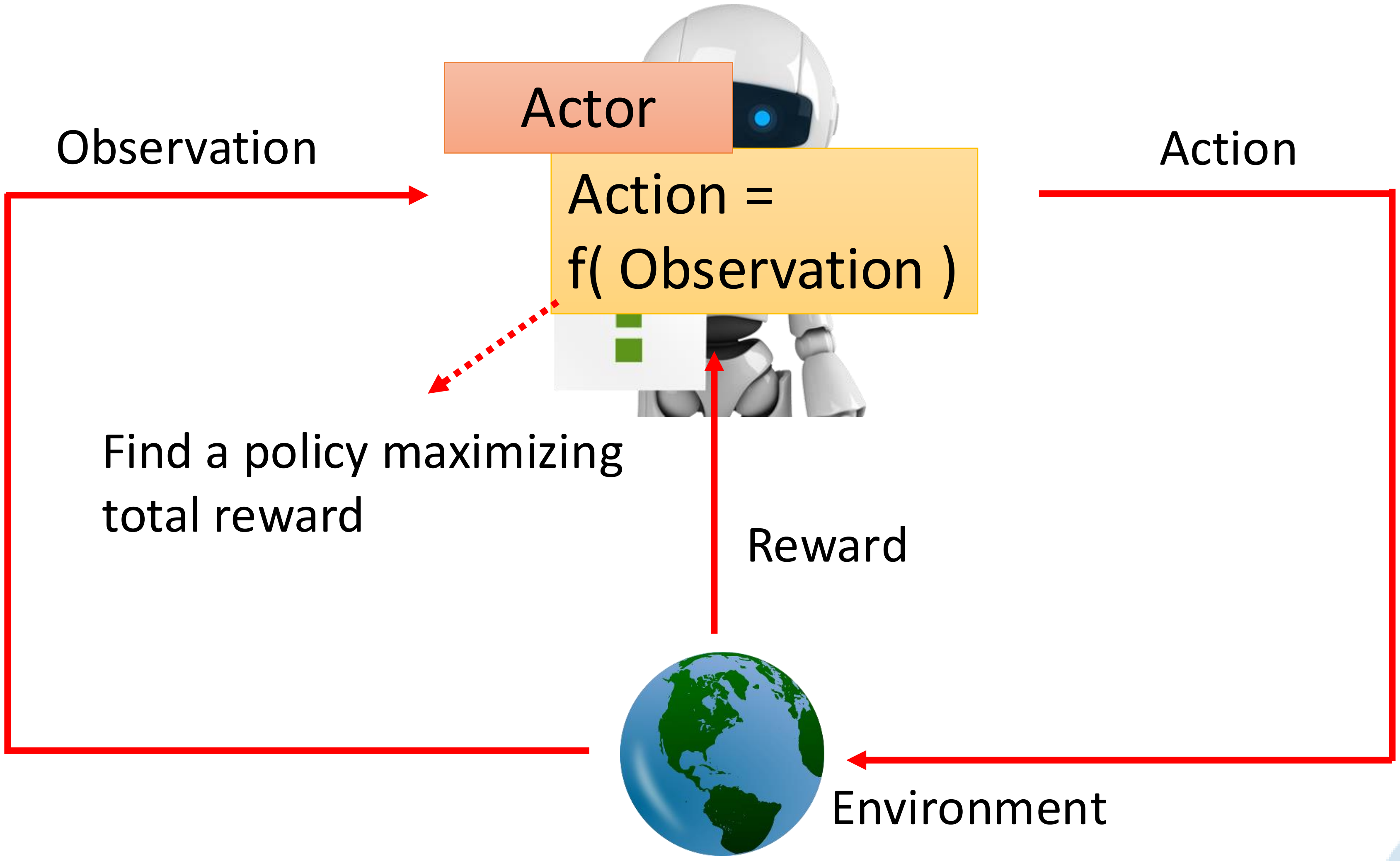
Not sure why good or bad

Cannot easily compute the gradient

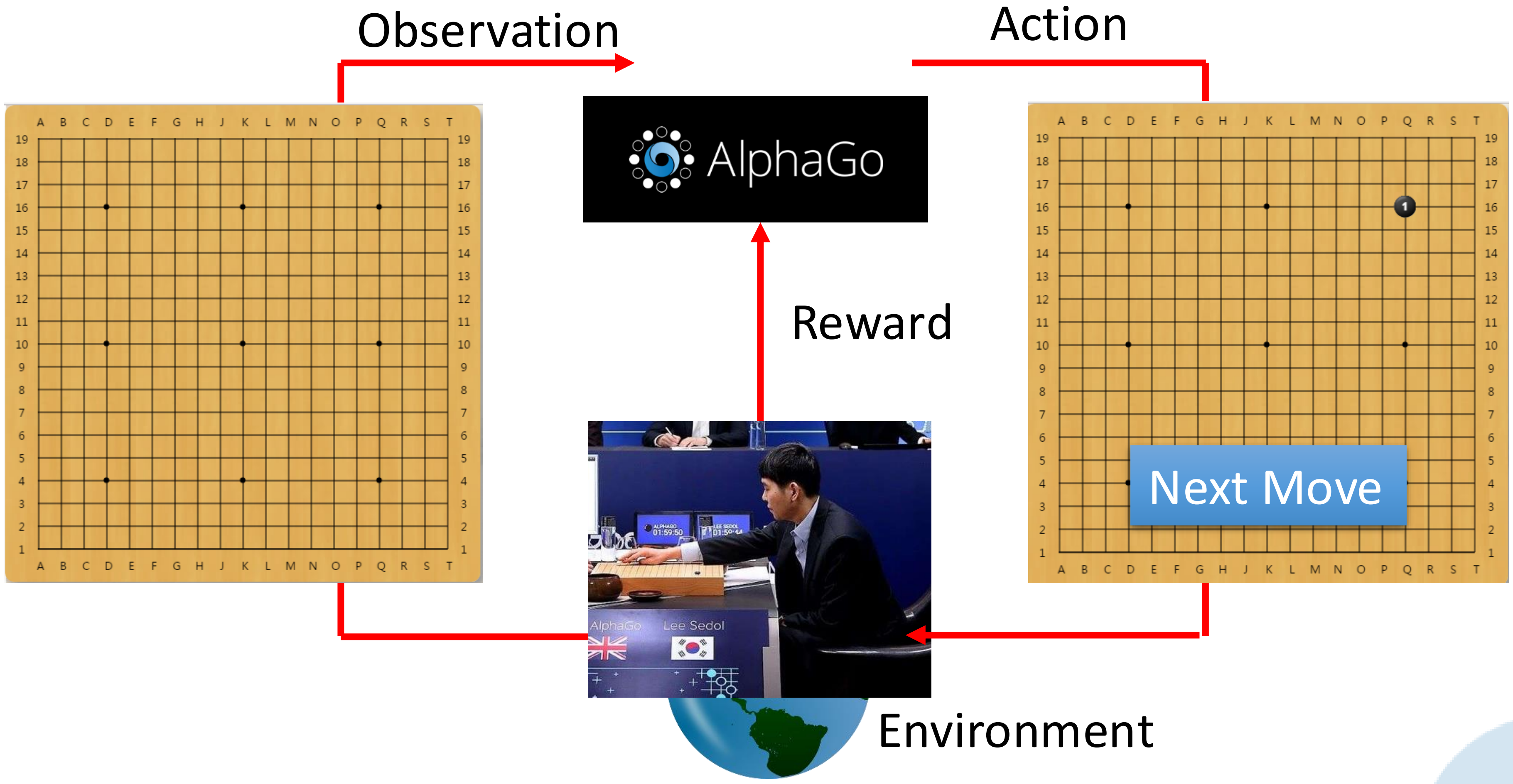
Complete answer

Training data is self-generated.

Common RL course

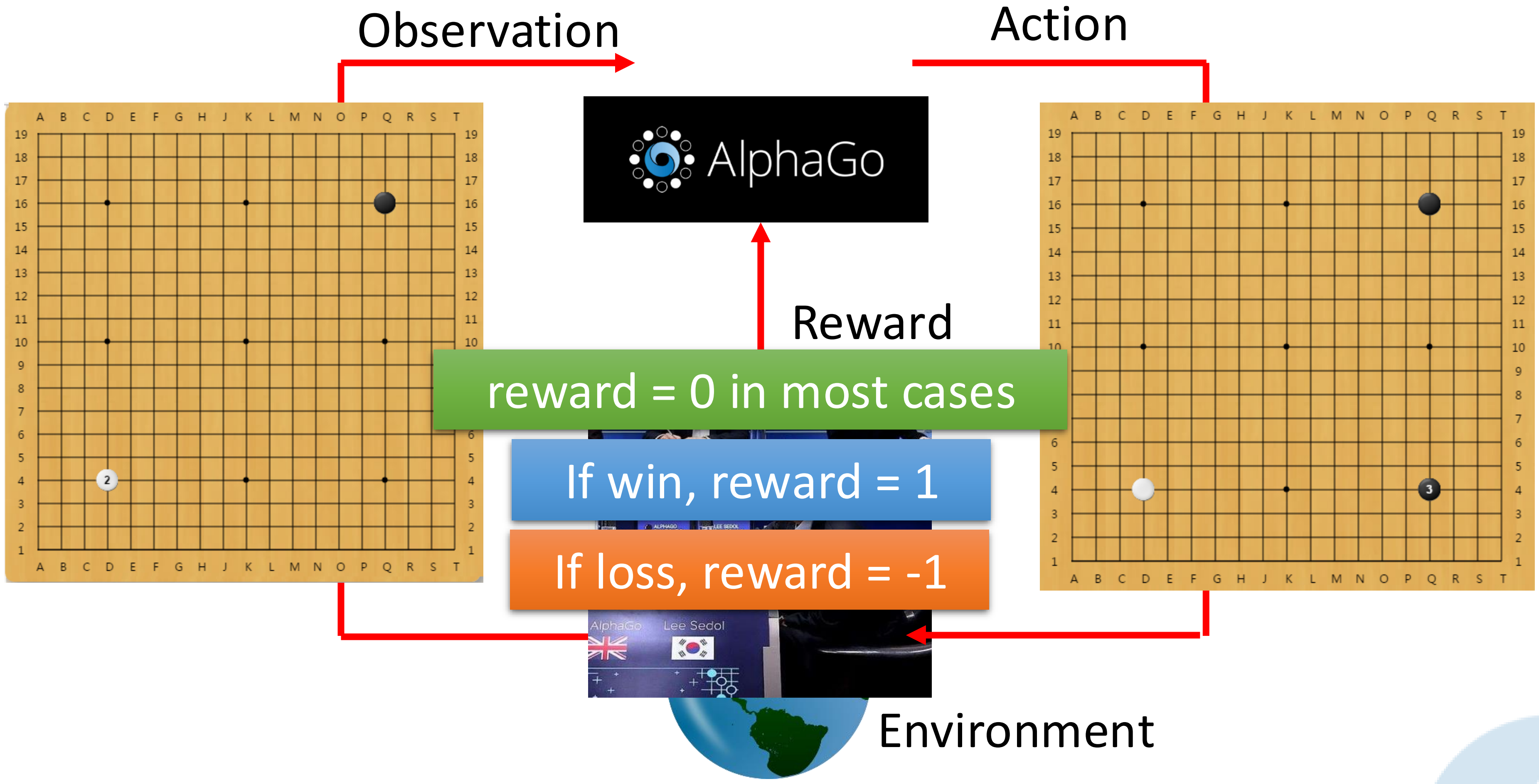


Common RL course: Go

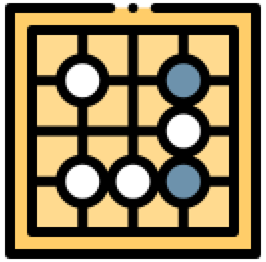


Common RL course: Go

Find an actor maximizing expected reward.



Language models vs Go



Unfinished board game → AlphaGo → "5-5"

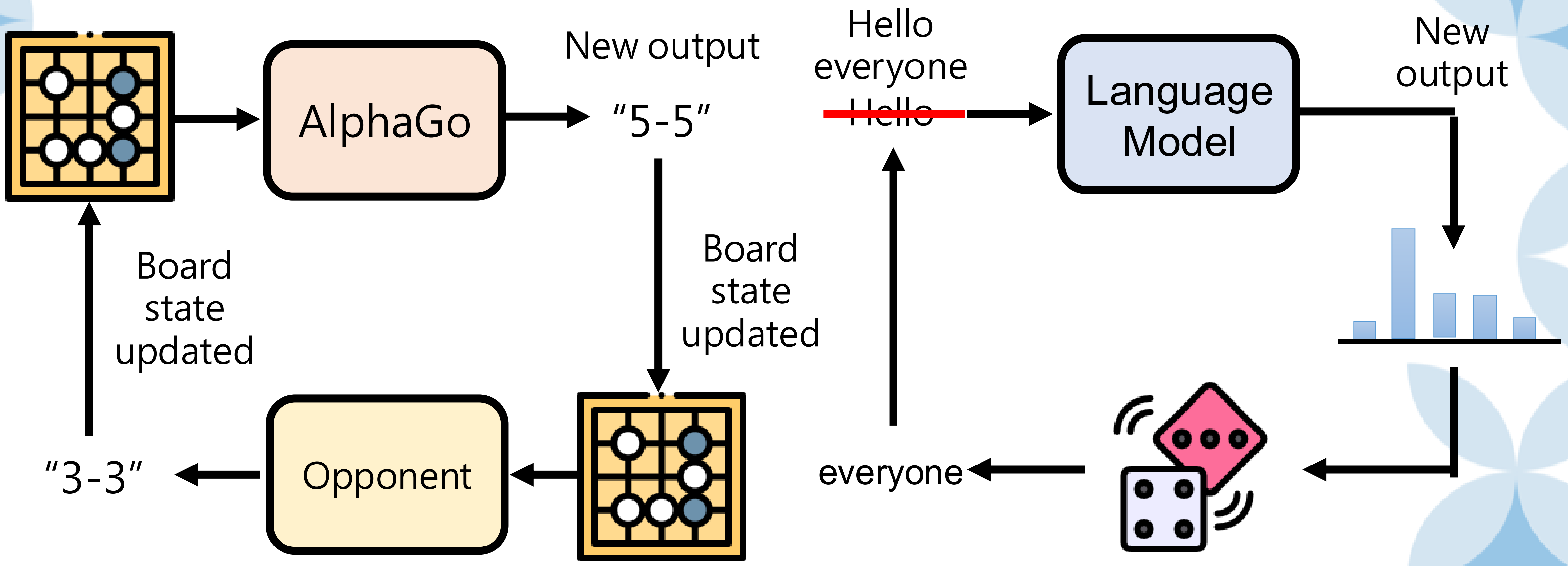
Position of next move

Unfinished sentence → Language Model → Next Token

(The explanation of AlphaGo here is simplified.)

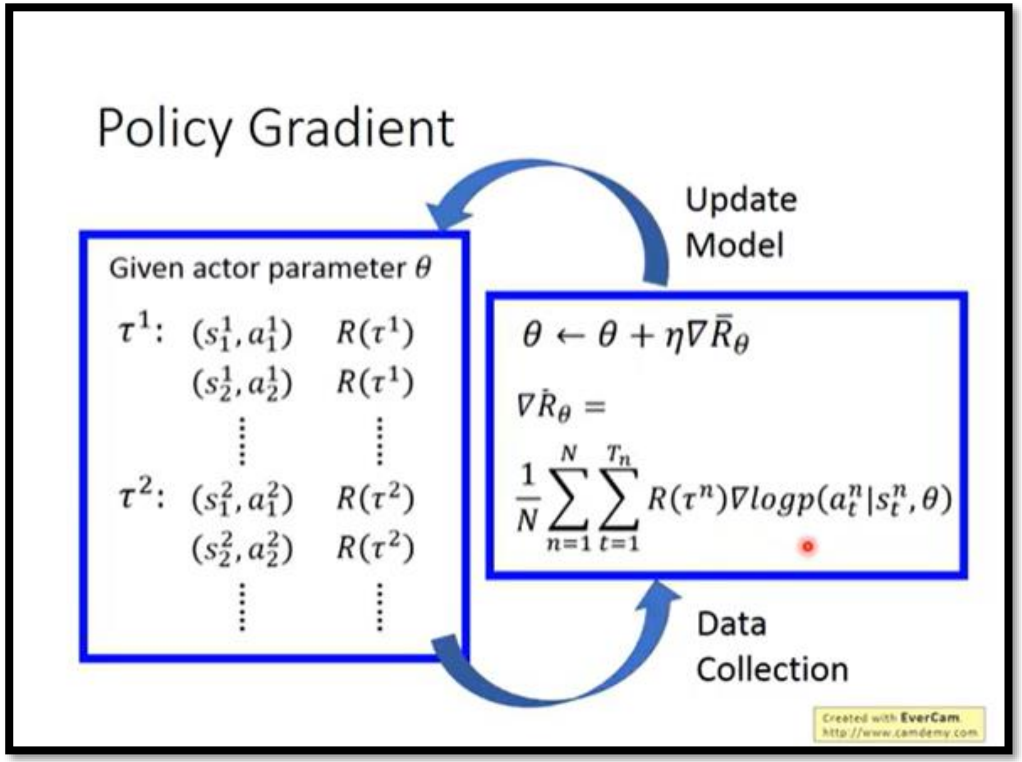
(Ignoring multi-turn dialogue for simplicity.)

Language models vs Go



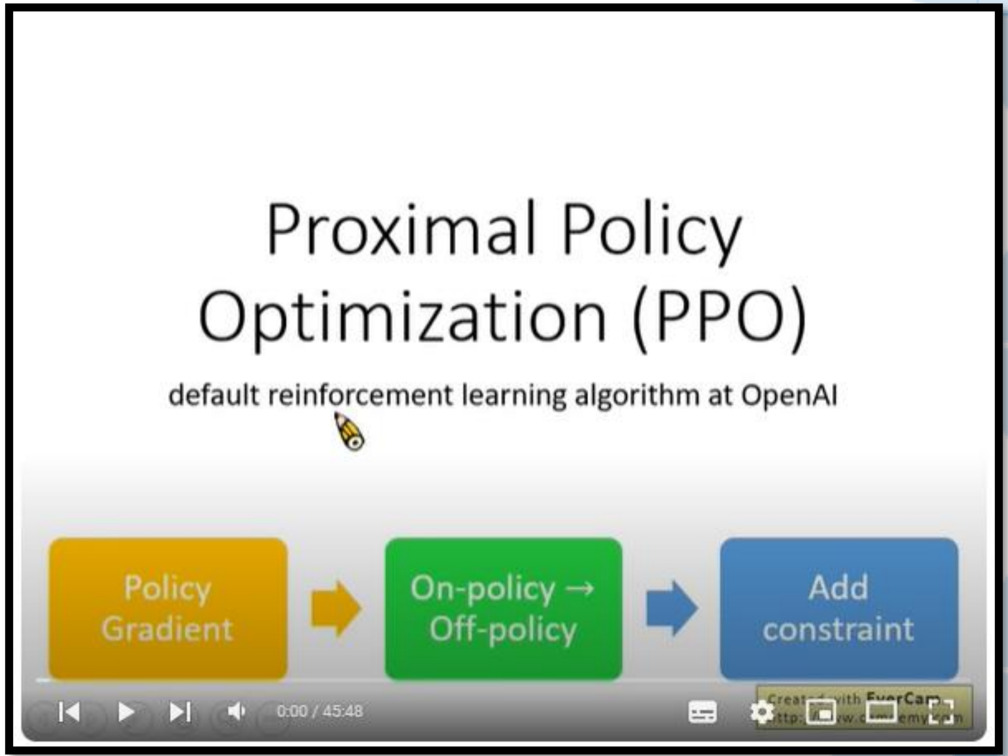
Policy Gradient, PPO, DPO, KTO, GRPO ……

- Policy gradient



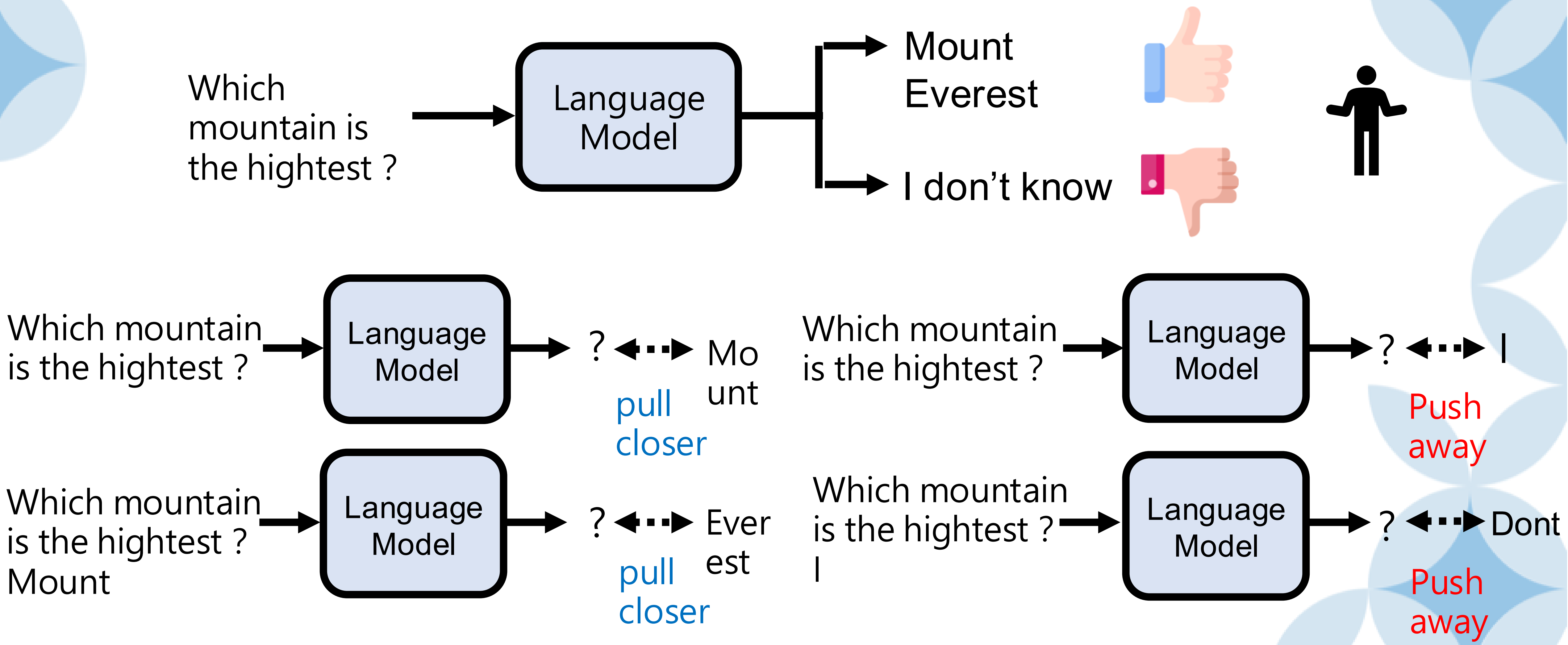
https://youtu.be/W8XF3ME8G2I?si=LsliHwNHYl5G0dCQ
https://youtu.be/y8UPGr36ccI?si=v4SYHvyJ7DgS-qBp

- PPO



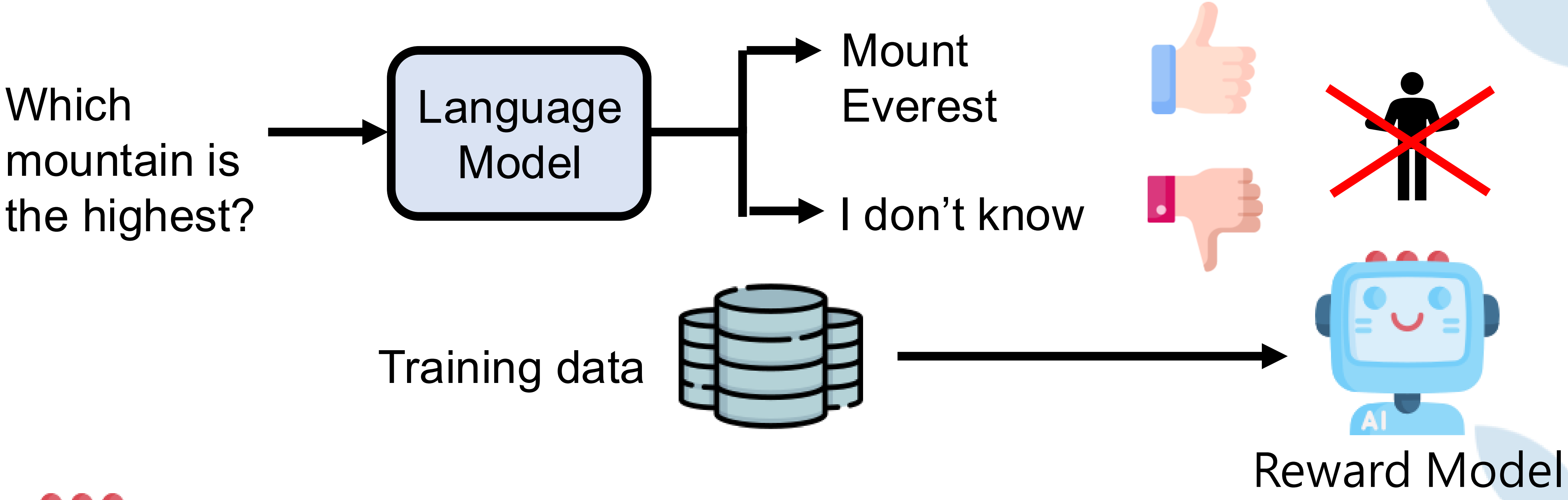
https://youtu.be/z95ZYgPgXOY?si=-E-1iE77qxsdNoGw

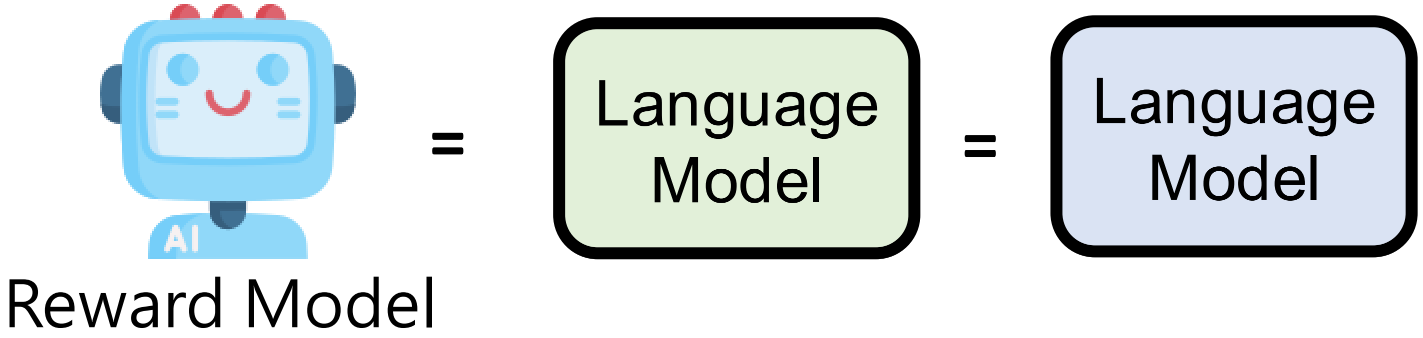
The key idea of policy gradient



RLHF → RLAIF

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





Creation is hard.
Judgment is easier.

Stage I Pre-train	Self-supervised Learning	**Trainig data**	In : AI is so Out : amazing
Stage II Supervised Fine-tuning	Supervised Learning		In : ”USER: who are you ? AI: ” Out : ”I”
Stage III RLHF	Reinforcement Learning (RL)		In: USER: “Which mountain is the highest ? AI:” Out: “Mount Everest” > “Any idea?”