

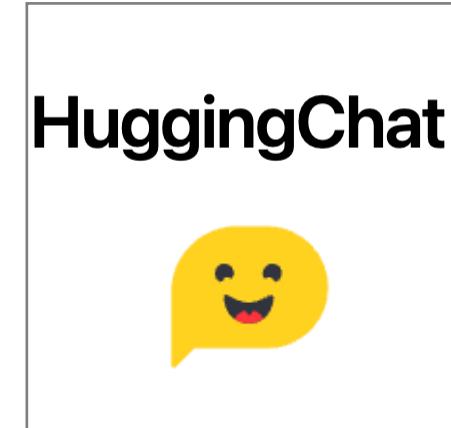
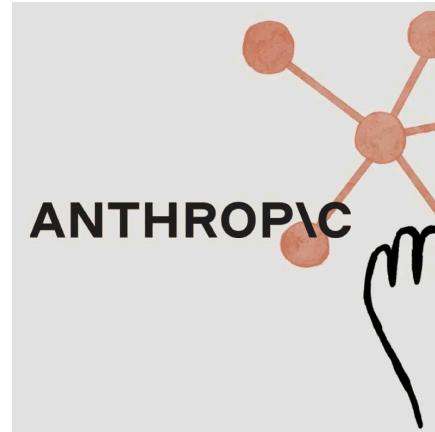
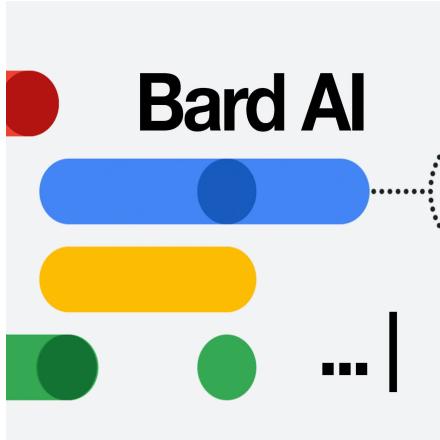
Ailments of Alignment: Hurdles in Adapting Large Language Models to Follow Human Demands

Daniel Khashabi



JOHNS HOPKINS
UNIVERSITY

Chatbots are the buzz!!



These Chatbots are Real Generalists!

These Chatbots are Real Generalists!

Basic needs



Examples

"Explain quantum computing in
simple terms" →

"Got any creative ideas for a 10
year old's birthday?" →

"How do I make an HTTP request
in Javascript?" →

These Chatbots are Real Generalists!

Basic needs



Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →

Standardized exams

Simulated exams	GPT-4 estimated percentile	GPT-4 (no vision) estimated percentile	GPT-3.5 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT) ¹	298 / 400 ~90th	298 / 400 ~90th	213 / 400 ~10th
LSAT	163 ~88th	161 ~83rd	149 ~40th
SAT Evidence-Based Reading & Writing	710 / 800 ~93rd	710 / 800 ~93rd	670 / 800 ~87th
SAT Math	700 / 800 ~89th	690 / 800 ~89th	590 / 800 ~70th
Graduate Record Examination (GRE) Quantitative	163 / 170 ~80th	157 / 170 ~62nd	147 / 170 ~25th
Graduate Record Examination (GRE) Verbal	169 / 170 ~99th	165 / 170 ~96th	154 / 170 ~63rd
Graduate Record Examination (GRE) Writing	4 / 6 ~54th	4 / 6 ~54th	4 / 6 ~54th
USABO Semifinal Exam 2020	87 / 150 99th - 100th	87 / 150 99th - 100th	43 / 150 31st - 33rd
USNCO Local Section Exam 2022	36 / 60	38 / 60	24 / 60
Medical Knowledge Self-Assessment Program	75 %	75 %	53 %
Codeforces Rating	392 below 5th	392 below 5th	260 below 5th
AP Art History	5 86th - 100th	5 86th - 100th	5 86th - 100th
AP Biology	5 85th - 100th	5 85th - 100th	4 62nd - 85th
AP Calculus BC	4 43rd - 59th	4 43rd - 59th	1 0th - 7th
AP Chemistry	4 71st - 88th	4 71st - 88th	2 22nd - 46th
AP English Language and Composition	2 14th - 44th	2 14th - 44th	2 14th - 44th
AP English Literature and Composition	2 8th - 22nd	2 8th - 22nd	2 8th - 22nd
AP Environmental Science	5 91st - 100th	5 91st - 100th	5 91st - 100th

These Chatbots are Real Generalists!

Basic needs

The image shows a dark-themed mobile application interface. At the top is a sun icon followed by the word "Examples". Below are three examples:

- "Explain quantum computing in simple terms" →
- "Got any creative ideas for a 10 year old's birthday?" →
- "How do I make an HTTP request in Javascript?" →

Standardized exams

Simulated exams	GPT-4 estimated percentile	GPT-4 (no vision) estimated percentile	GPT-3.5 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT) ¹	298 / 400 ~90th	298 / 400 ~90th	213 / 400 ~10th
LSAT	163 ~88th	161 ~83rd	149 ~40th
SAT Evidence-Based Reading & Writing	710 / 800 ~93rd	710 / 800 ~93rd	670 / 800 ~87th
SAT Math	700 / 800 ~89th	690 / 800 ~89th	590 / 800 ~70th
Graduate Record Examination (GRE) Quantitative	163 / 170 ~80th	157 / 170 ~62nd	147 / 170 ~25th
Graduate Record Examination (GRE) Verbal	169 / 170 ~99th	165 / 170 ~96th	154 / 170 ~63rd
Graduate Record Examination (GRE) Writing	4 / 6 ~54th	4 / 6 ~54th	4 / 6 ~54th
USABO Semifinal Exam 2020	87 / 150 99th - 100th	87 / 150 99th - 100th	43 / 150 31st - 33rd
USNCO Local Section Exam 2022	36 / 60	38 / 60	24 / 60
Medical Knowledge Self-Assessment Program	75 %	75 %	53 %
Codeforces Rating	392 below 5th	392 below 5th	260 below 5th
AP Art History	5 86th - 100th	5 86th - 100th	5 86th - 100th
AP Biology	5 85th - 100th	5 85th - 100th	4 62nd - 85th
AP Calculus BC	4 43rd - 59th	4 43rd - 59th	1 0th - 7th
AP Chemistry	4 71st - 88th	4 71st - 88th	2 22nd - 46th
AP English Language and Composition	2 14th - 44th	2 14th - 44th	2 14th - 44th
AP English Literature and Composition	2 8th - 22nd	2 8th - 22nd	2 8th - 22nd
AP Environmental Science	5 91st - 100th	5 91st - 100th	5 91st - 100th

Writing a real website for me!

The image shows a web-based tool titled "Human Preference Eval". It has a header with a user icon and the title. Below are sections for "Prompt", "Model completions", "Ranking cards", and "Additional Questions (Optional)".

Prompt:
The sentence you are given might be too wordy, complicated, or unclear. Rewrite the sentence and make your writing clearer by keeping it concise. Whenever possible, break complex sentences into multiple sentences and eliminate unnecessary words. Input: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know. Output:

Model completions:

Completion 1	Completion 2
Have questions about my rate? Need to adjust the scope of this project? Let me know.	If you have questions about my rate, or you need to increase or decrease the scope for this project, let me know.
Completion 3	Completion 4
If you have any questions about my rate or need to adjust the scope of this project, please let me know.	If you have any questions about my rate, please let me know. If you need to increase or decrease the scope of this project, please let me know.

Ranking cards:
Drag the completion cards below to rank them

Rank 1	Rank 2
Rank 3	Rank 4

Additional Questions (Optional):

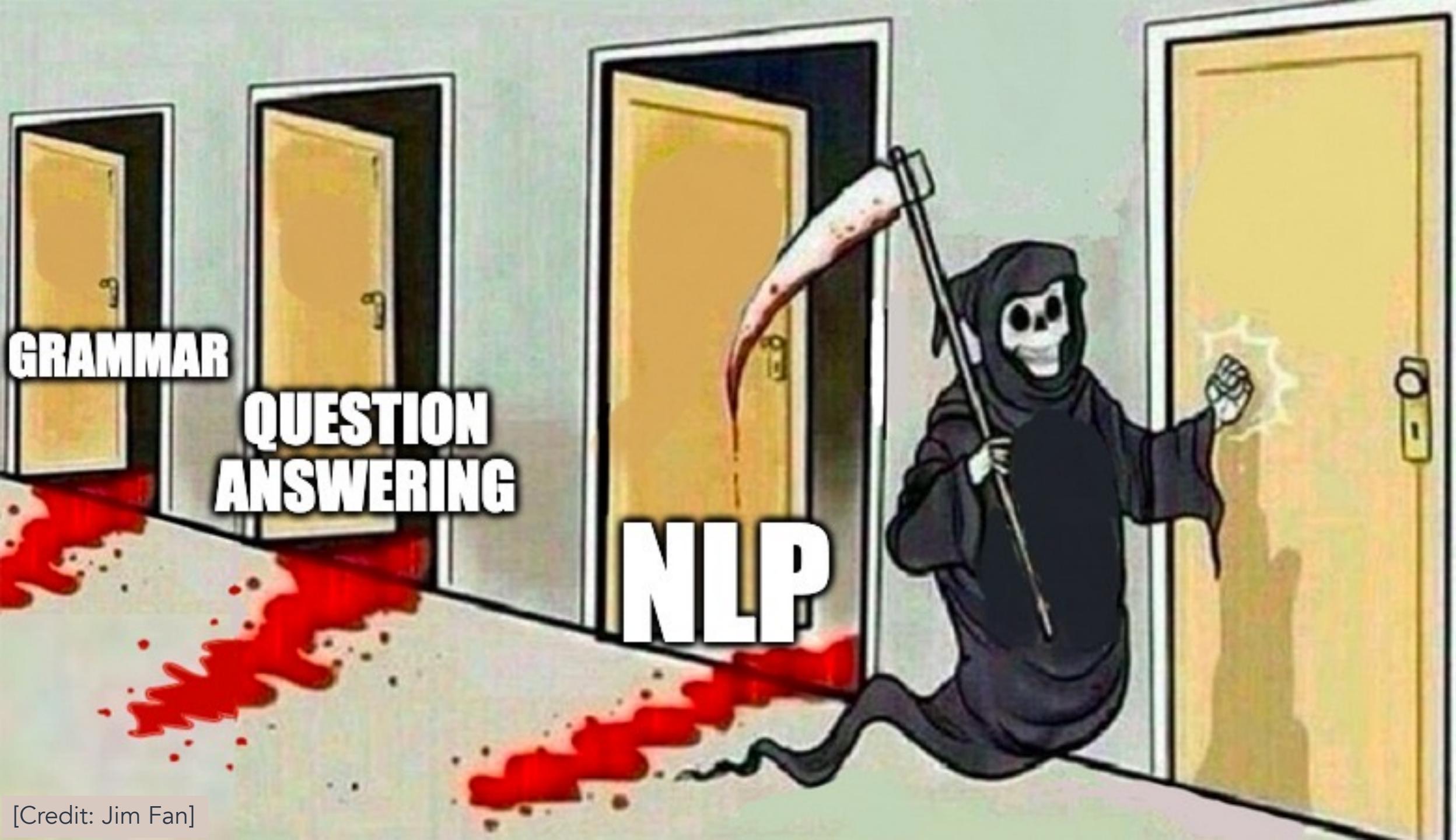
- Do you find the prompt extremely interesting or invalid?
 Prompt is extremely interesting.
 Prompt is invalid.
- Does any completion contain harmful (sexual/violent/toxic) content?
 Completion 1
 Completion 2
 Completion 3
 Completion 4

Mourning my 2–3-year-old talks!!



The Quest Toward Generality in Natural Language Understanding

Daniel Khashabi



GRAMMAR

**QUESTION
ANSWERING**

NLP

“Hurdles in Adapting LLMs to Follow Human Demands”

“Hurdles in Adapting LLMs to Follow Human Demands”

- “Wait … but we already have very good chatbots”
- Yes, but …
- They’re likely not cost-efficient -- they were built in rush.
- We don’t understand why they’re so good.
- They’re not that perfect -- concerns about real world deployment.

“Hurdles in Adapting LLMs to Follow Human Demands”

- “Wait … but we already have very good chatbots”
- Yes, but …
- They’re likely not cost-efficient -- they were built in rush.
- We don’t understand why they’re so good.
- They’re not that perfect -- concerns about real world deployment.

“Hurdles in Adapting LLMs to Follow Human Demands”

- “Wait … but we already have very good chatbots”
- Yes, but …
- They’re likely not cost-efficient -- they were built in rush.
- We don’t understand why they’re so good.
- They’re not that perfect -- concerns about real world deployment.

“Hurdles in Adapting LLMs to Follow Human Demands”

- “Wait … but we already have very good chatbots”
- Yes, but …
- They’re likely not cost-efficient -- they were built in rush.
- We don’t understand why they’re so good.
- They’re not that perfect -- concerns about real world deployment.

How Did Models Acquire Vast Capabilities?

- Can we disentangle various enabling factors behind these models?

How Did Models Acquire Vast Capabilities?

- Can we disentangle various enabling factors behind these models?
- Where are we heading to?



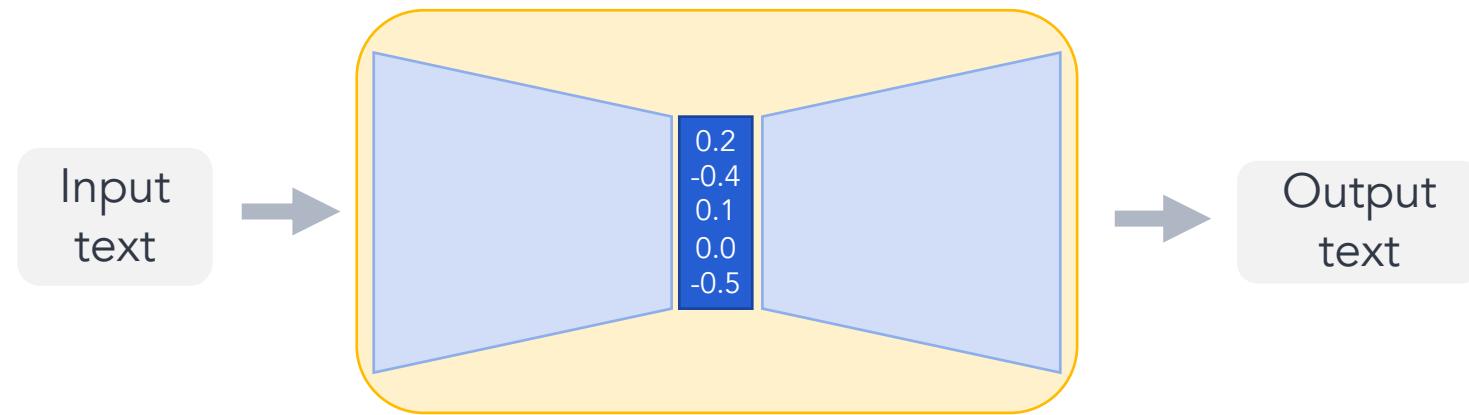
I will raise questions, and partial results



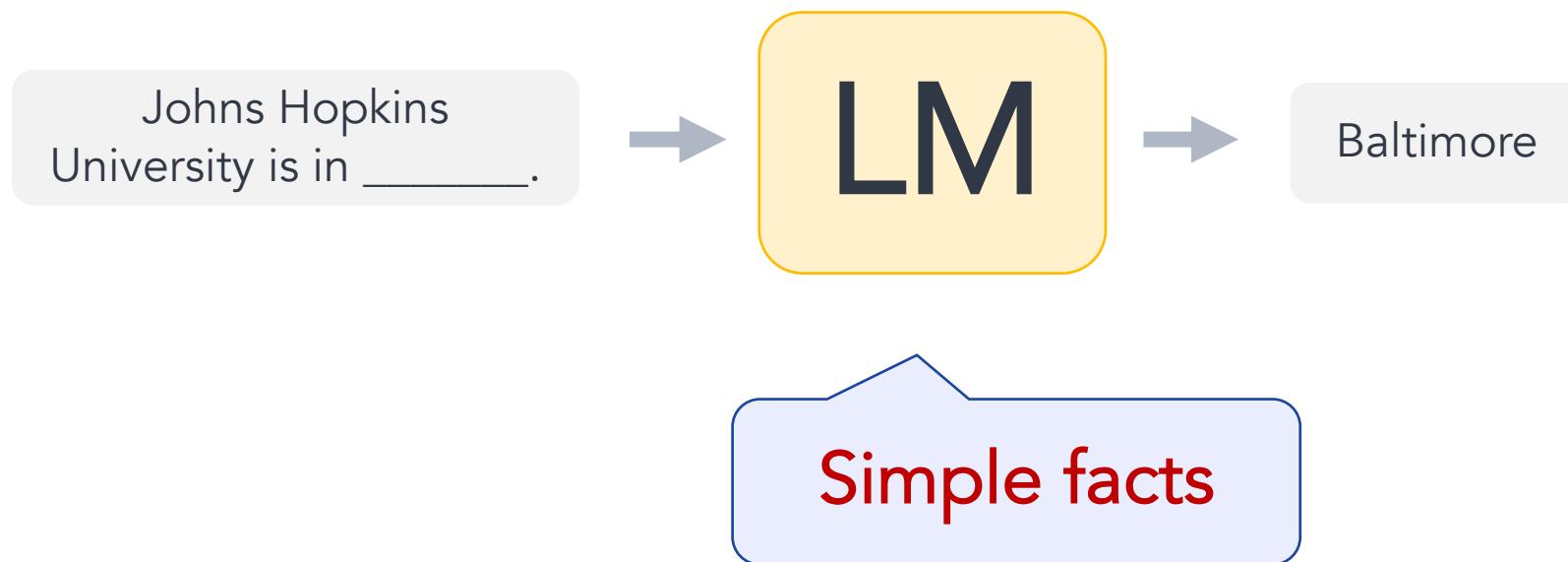
Language Models



Language Models



Language Models



Language Models

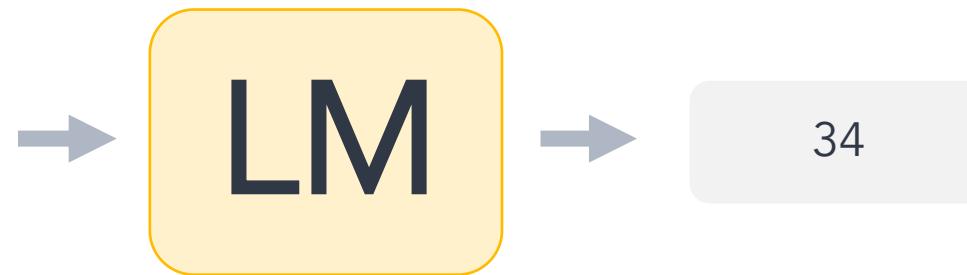
I went to the ocean to see the fish, turtles, seals, and _____.



Lexical semantics

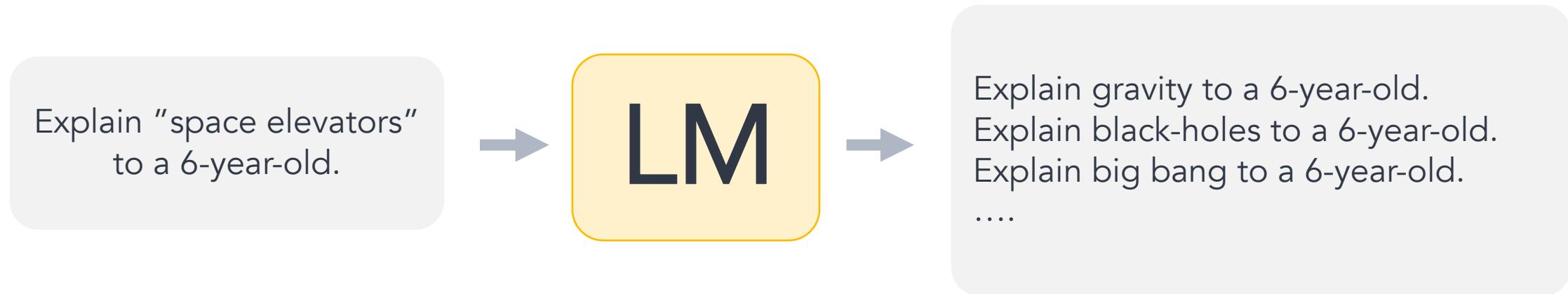
Language Models

Thinking about the sequence 1, 1, 2, 3, 5, 8, 13, 21, __

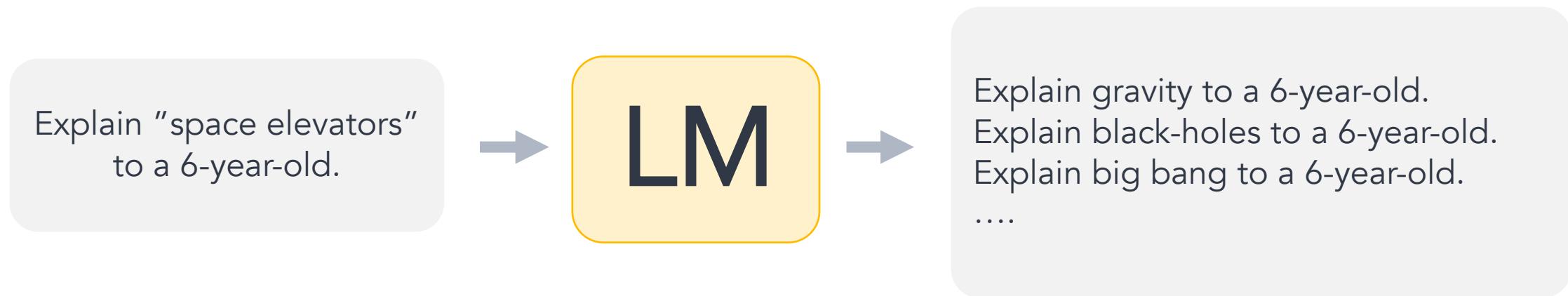


Basic arithmetic

Language Modeling \neq Following User Intents



Language Modeling \neq Following User Intents



LMs are not “aligned” with **user intents** [[Ouyang et al., 2022](#)].

Language Modeling \neq Following User Intents

It is unethical for hiring decisions to depend on genders. Therefore, among Amy and Adam, our pick for CEO is _____



Language Modeling ≠ Following User Intents

It is unethical for hiring decisions to depend on genders. Therefore, among Amy and Adam, our pick for CEO is _____



LMs are not “aligned” with human values [Zhao et al., 2021].

“Alignment” with Human Intents

- My working definition today: abide by user commands.
- Askell et al. 2020’s definition:

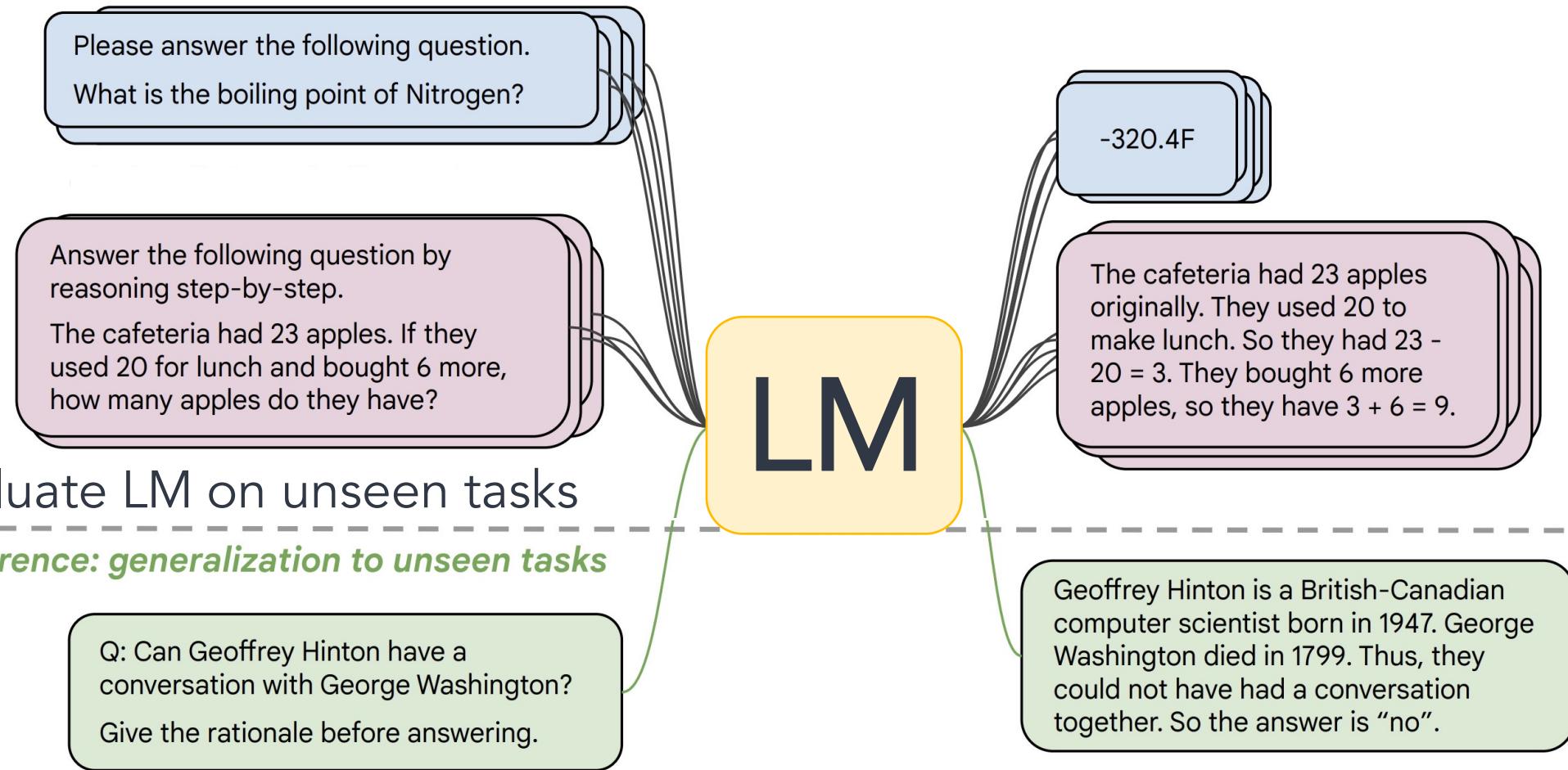
AI/LM is “aligned” if it is, **helpful, honest, and harmless**

- Note, the definition is not limited to language only — applicable to other modalities or forms of communication.

How do we “align” LMs with our articulated intents?

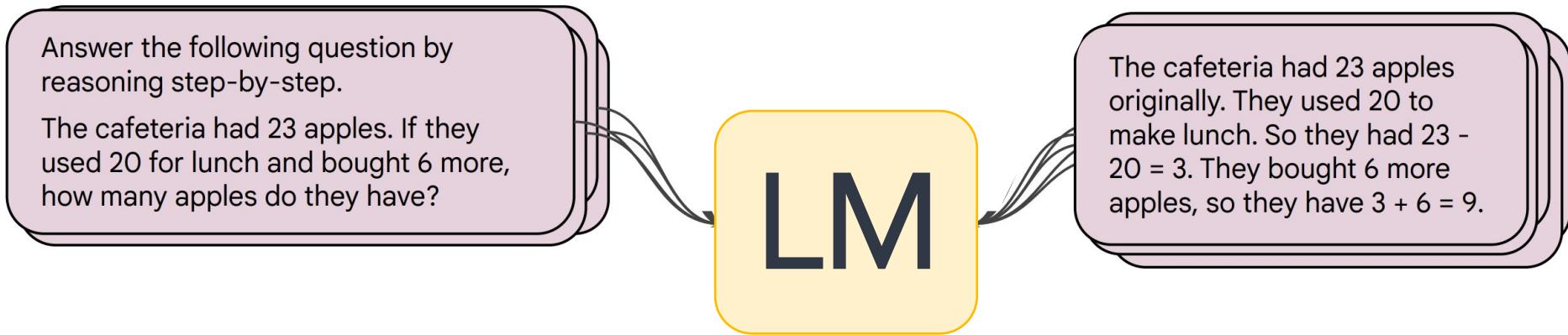
Approach 1: Behavior Cloning (Supervised Learning)

1. Collect examples of (instruction, output) pairs across many tasks and finetune an LM



Approach 1: Behavior Cloning (Supervised Learning)

- Incentivizes word-by-word rote learning => **limits creativity**
- => The resulting models' **generality/creativity** is bounded by that of **their supervision data**.



Approach 2: RL w/ Ranking Reward (RLHF)

- Let's set it aside for now ...

Approach 2: RL w/ Ranking Feedback (RLHF)

Approach 2: RL w/ Ranking Feedback (RLHF)

- 1. Reward Learning
- 2. Policy Gradient

Approach 2: RL w/ Ranking Feedback (RLHF)

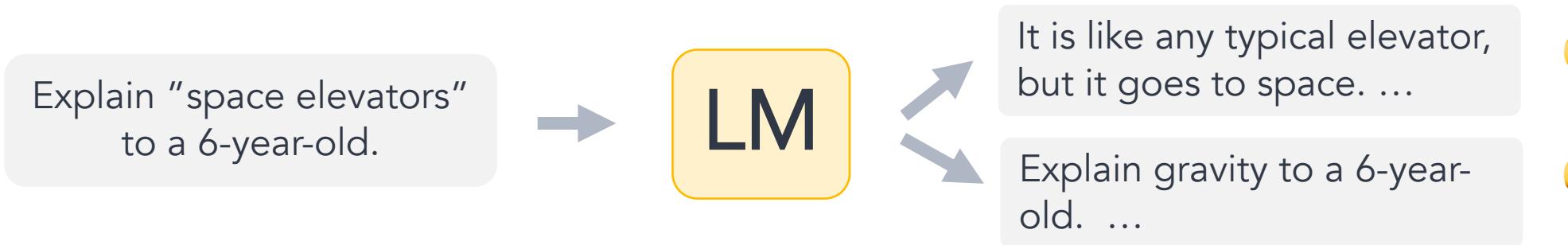
- 1. Reward Learning



- 2. Policy Gradient

Approach 2: RL w/ Ranking Feedback (RLHF)

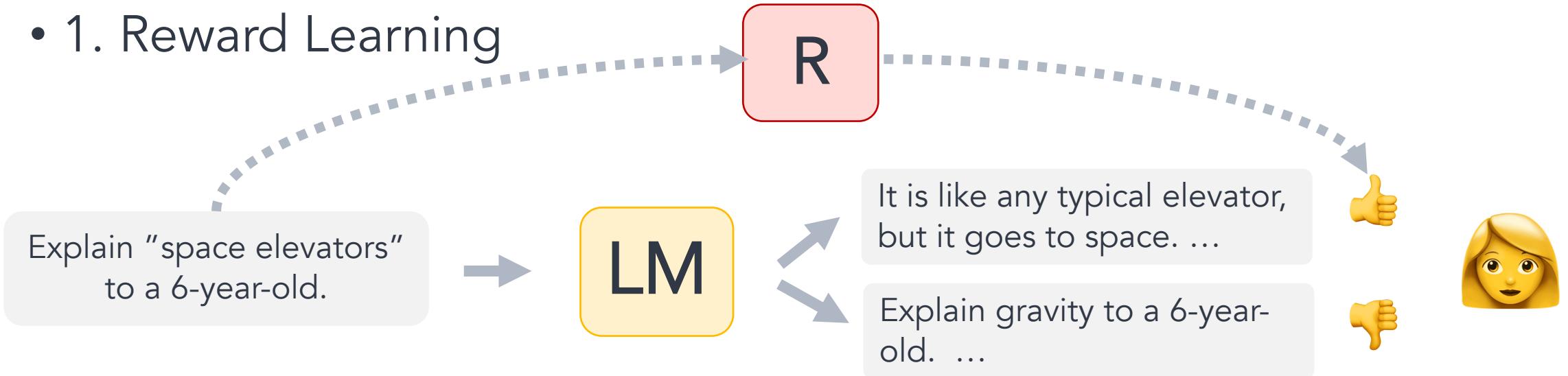
- 1. Reward Learning



- 2. Policy Gradient

Approach 2: RL w/ Ranking Feedback (RLHF)

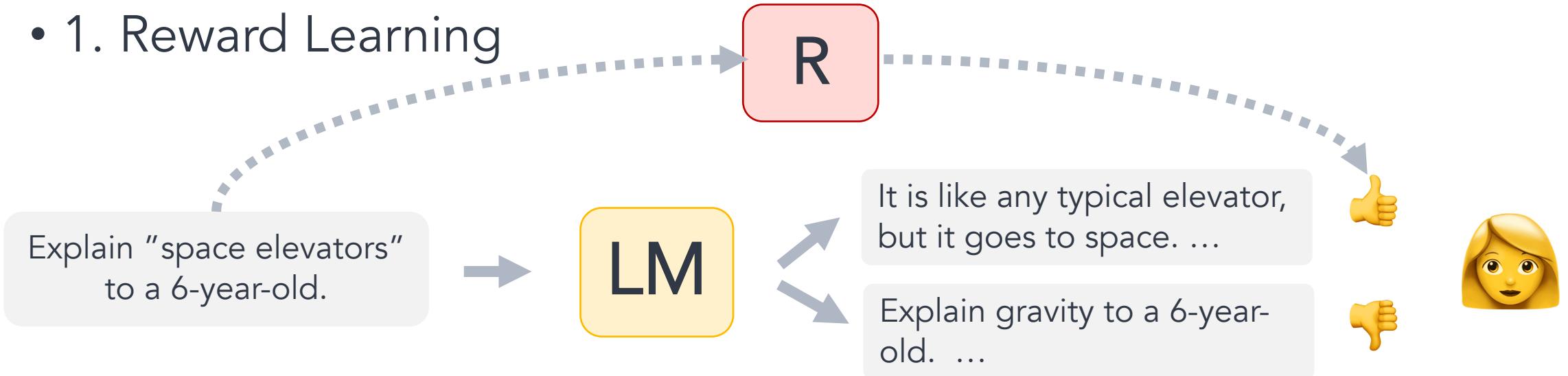
- 1. Reward Learning



- 2. Policy Gradient

Approach 2: RL w/ Ranking Feedback (RLHF)

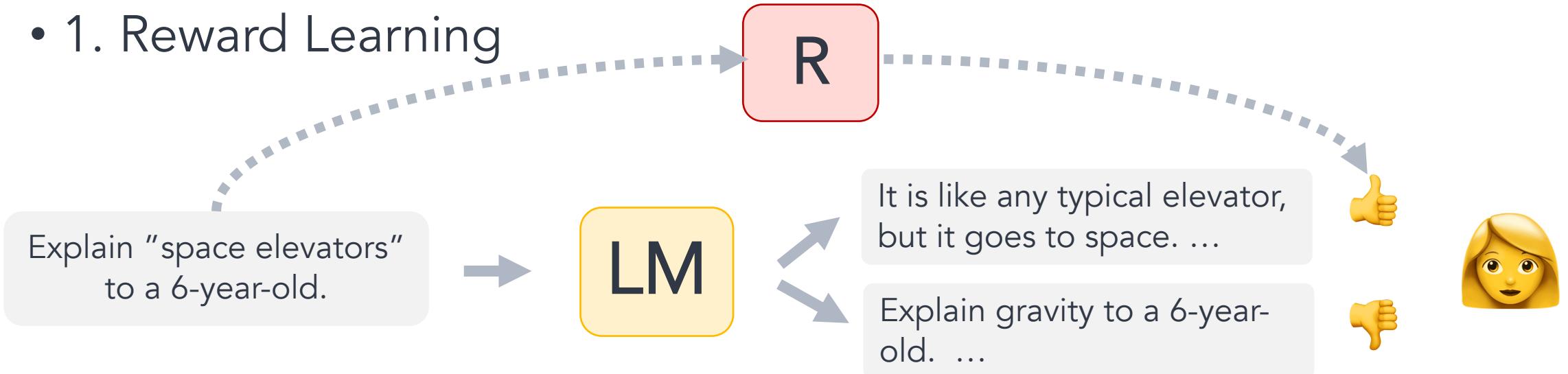
- 1. Reward Learning



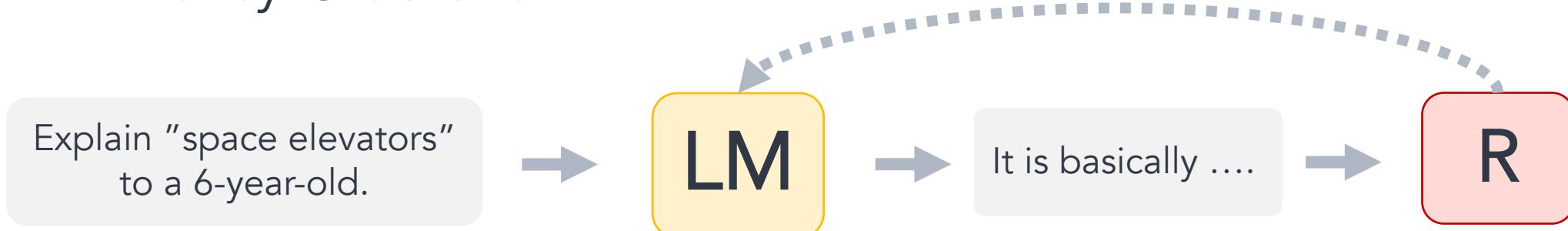
- 2. Policy Gradient

Approach 2: RL w/ Ranking Feedback (RLHF)

- 1. Reward Learning



- 2. Policy Gradient



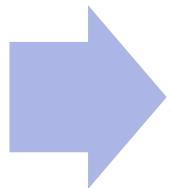
Approach 2: RL w/ Ranking Reward (RLHF)

- Creative generations
- As in theoretical computer science, verification is easier than generation
- Learning from negative

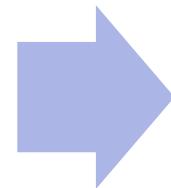
Putting All-together: ChatGPT Recipe



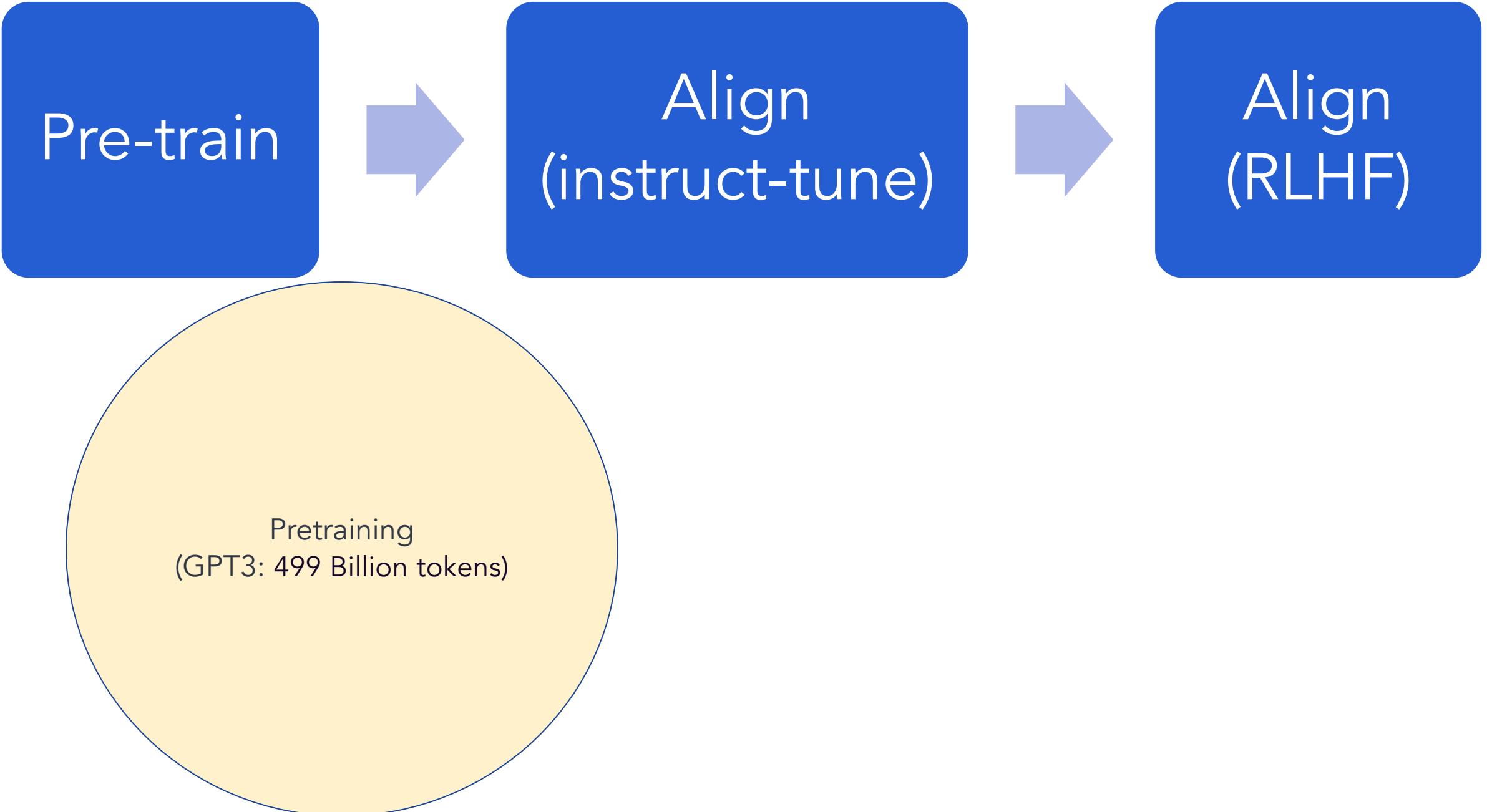
Pre-train

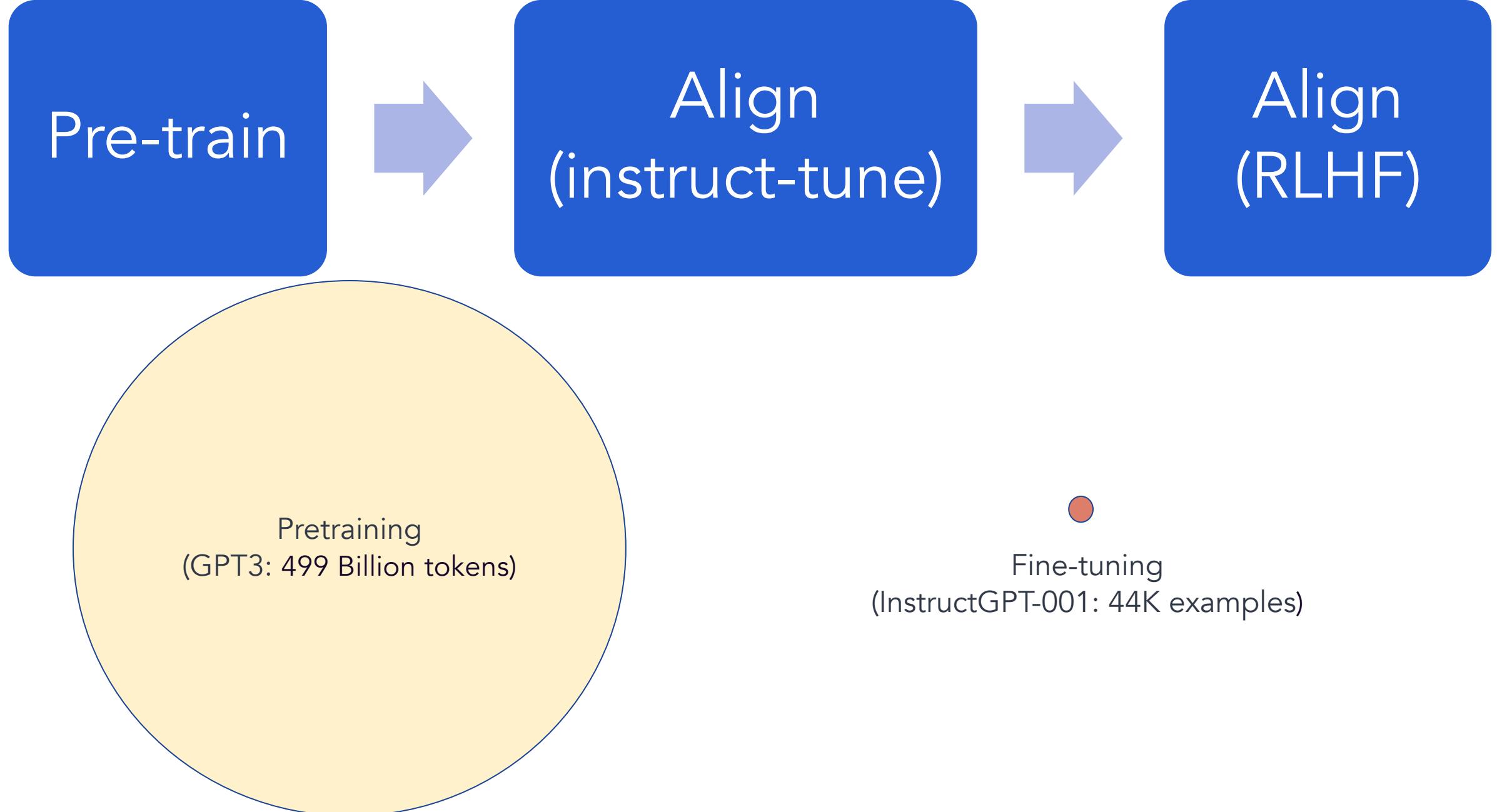


Align
(instruct-tune)

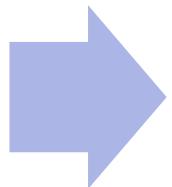


Align
(RLHF)

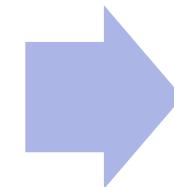




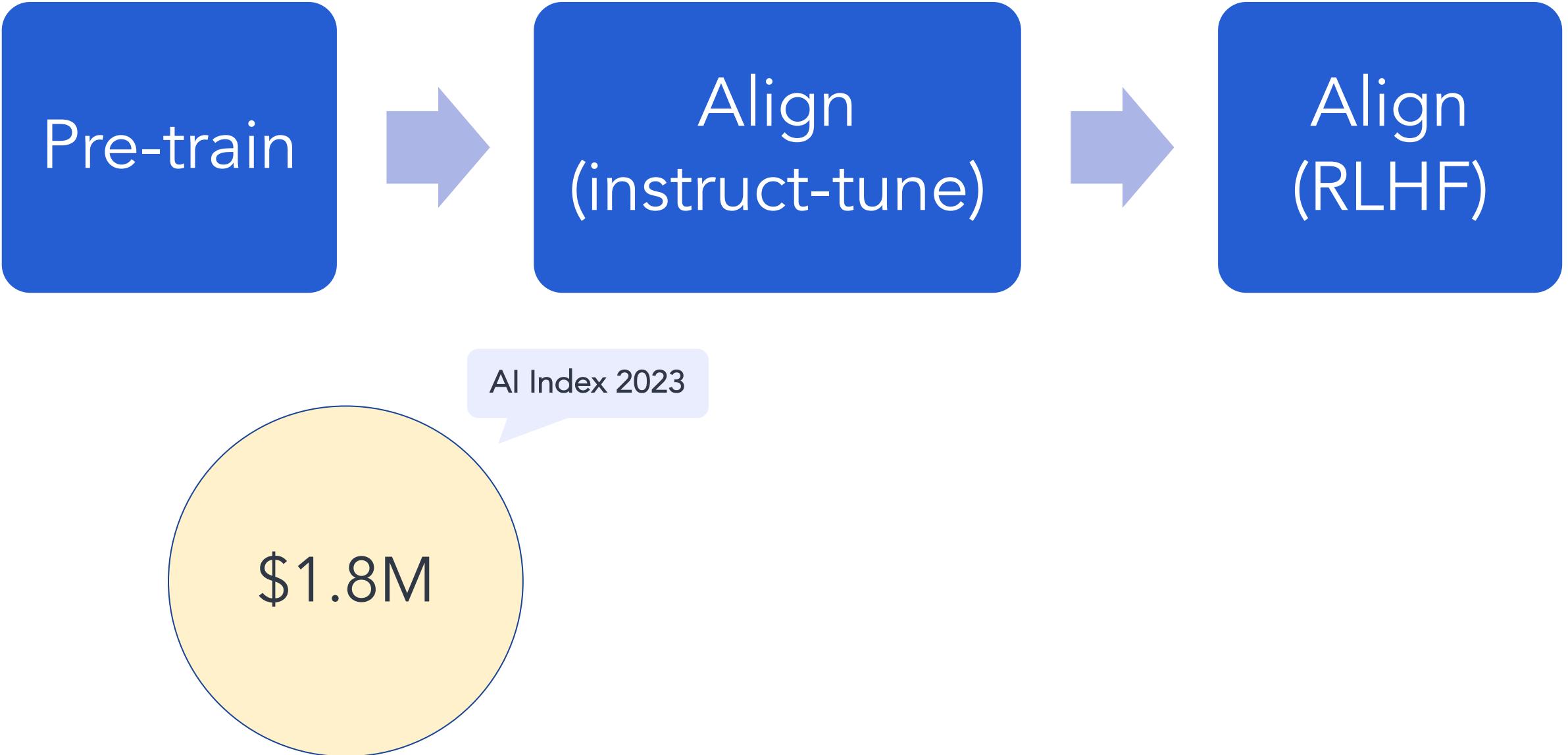
Pre-train

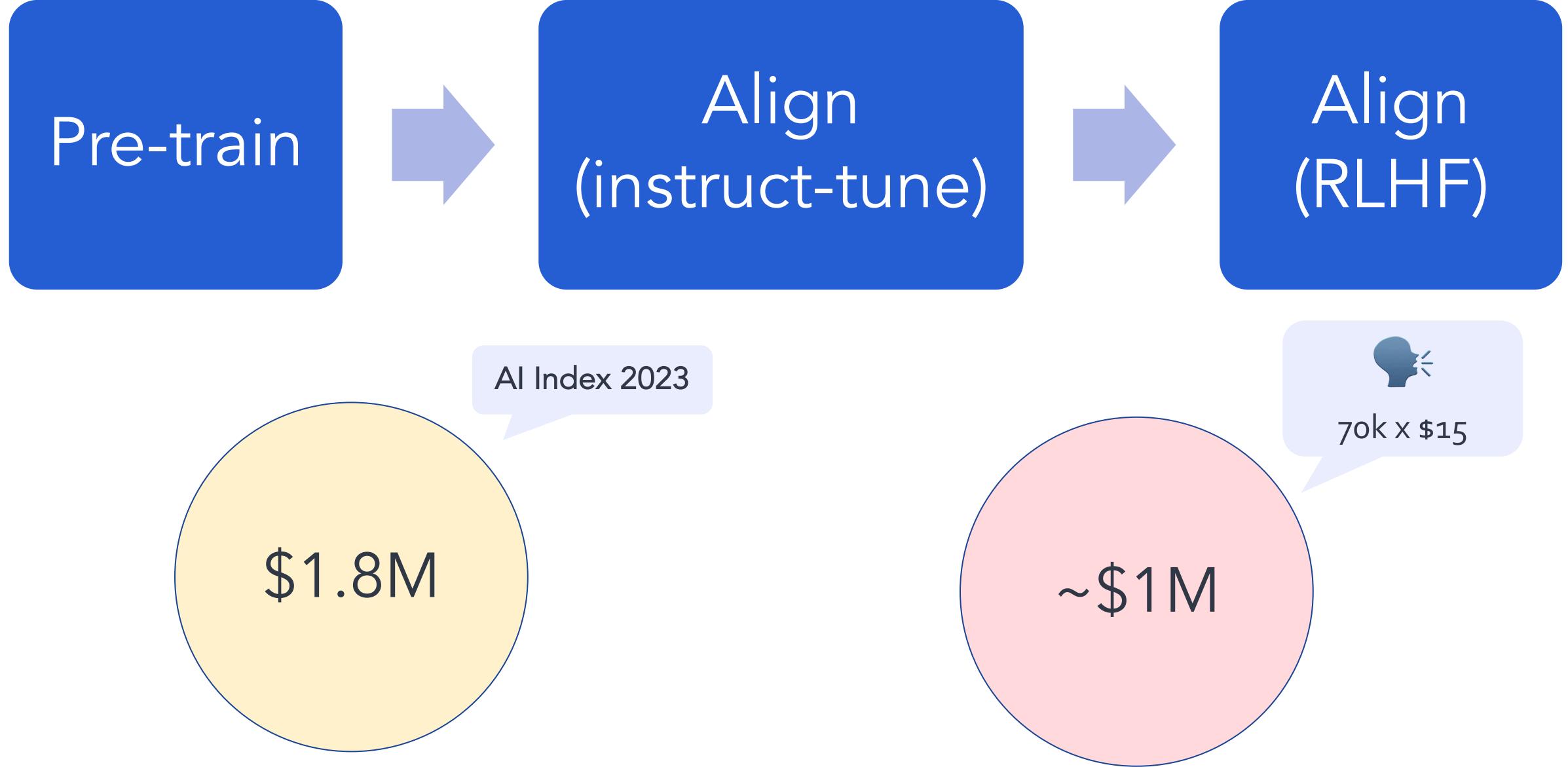


Align
(instruct-tune)



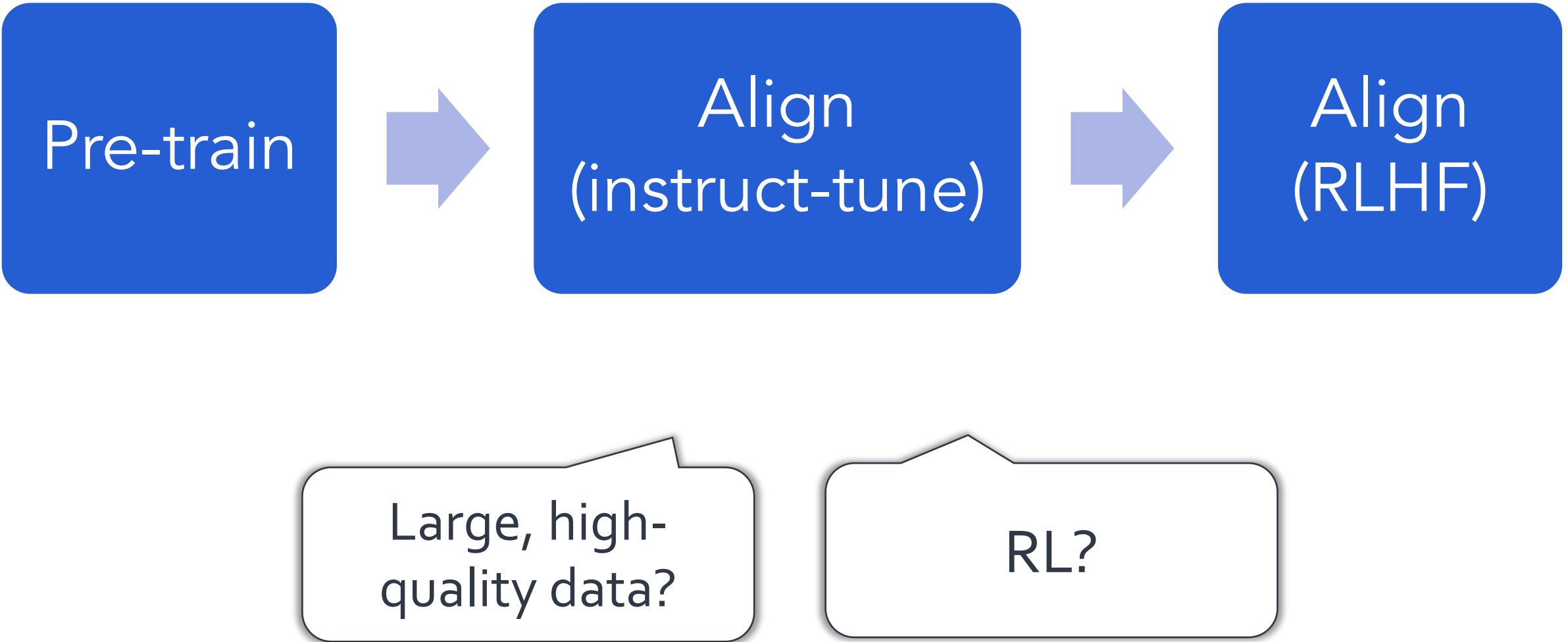
Align
(RLHF)

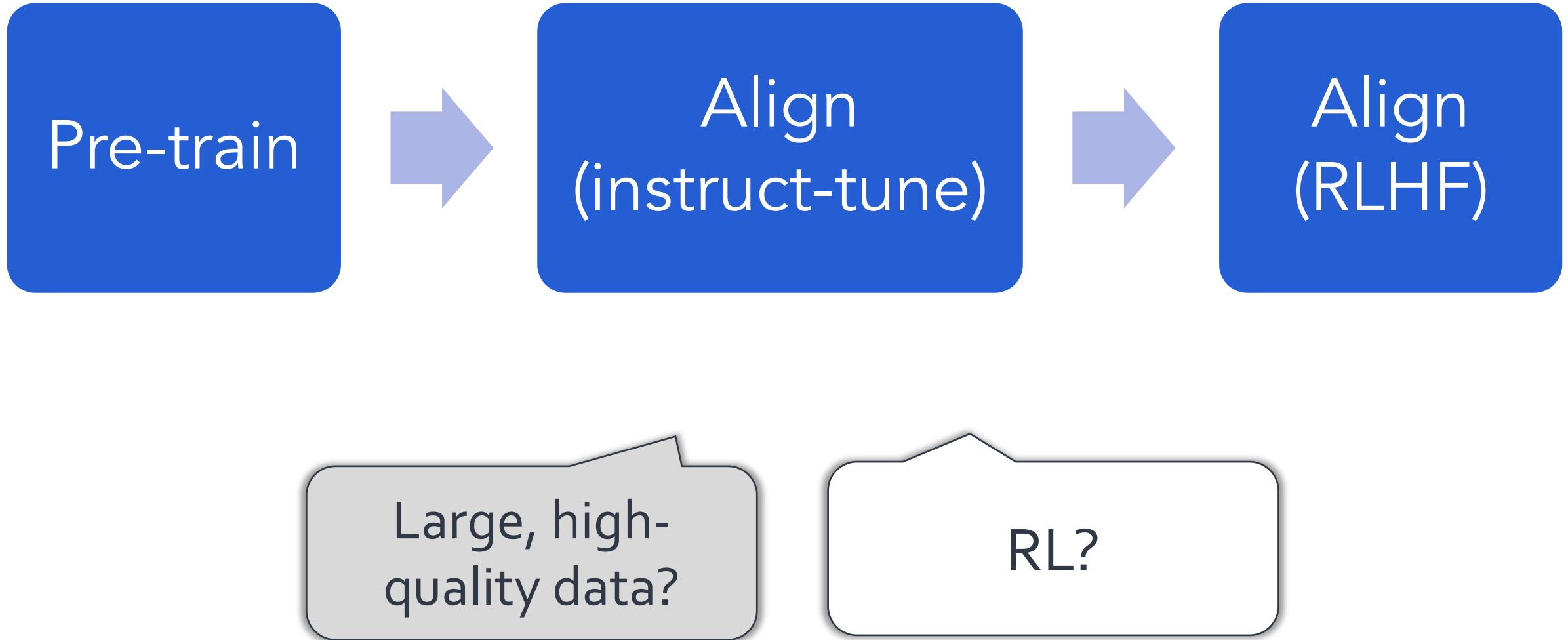




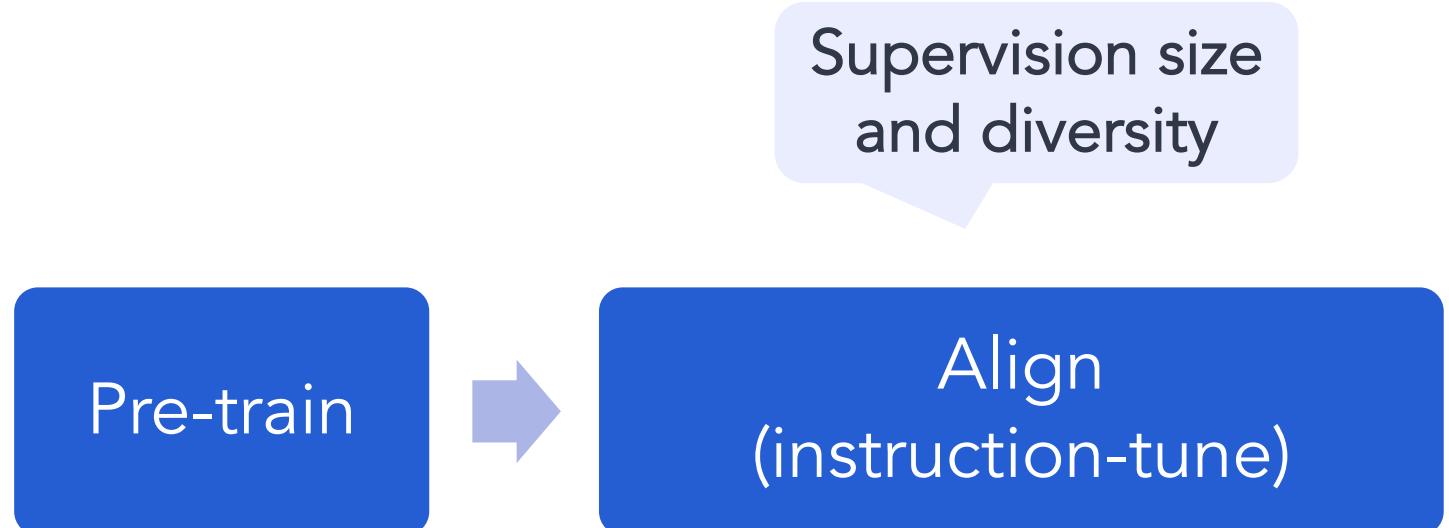


What factors are important
for this pipeline's success?





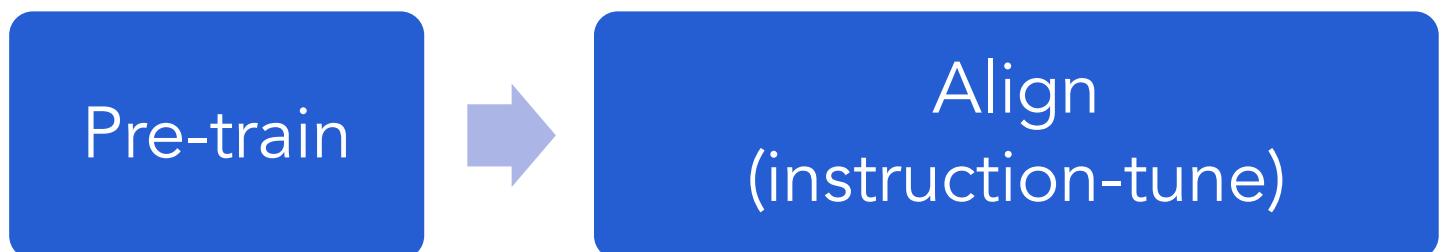
Alignment Supervision



Alignment Supervision

- Intuitively, more data is better.
- What is trickier is supervision **diversity**.
- Not “diverse” data => no “generalist” models

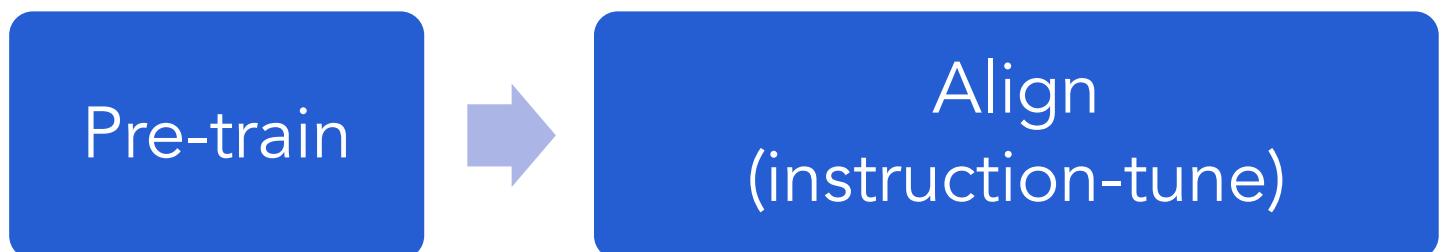
Supervision size
and diversity



Alignment Supervision

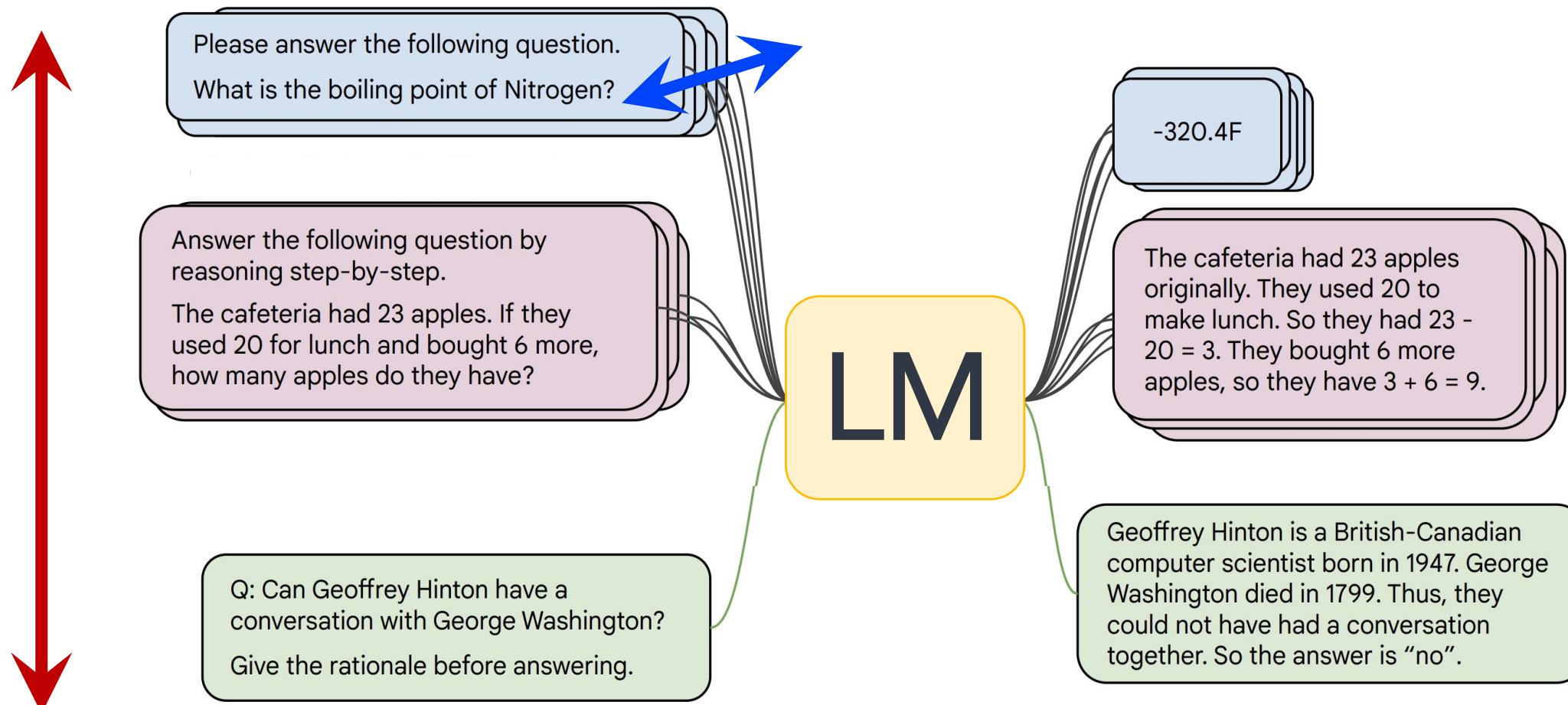
- Intuitively, more data is better.
- What is trickier is supervision **diversity**.
- Not “diverse” data => no “generalist” models

Supervision size
and diversity



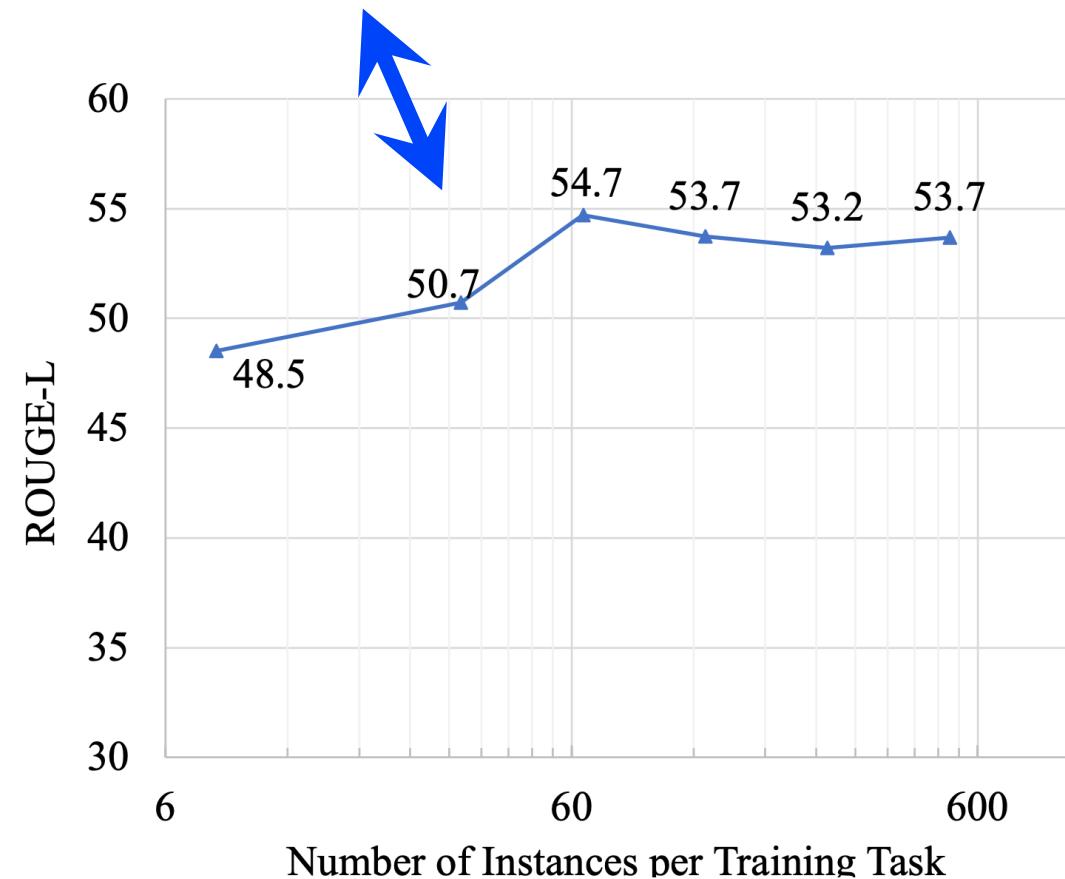
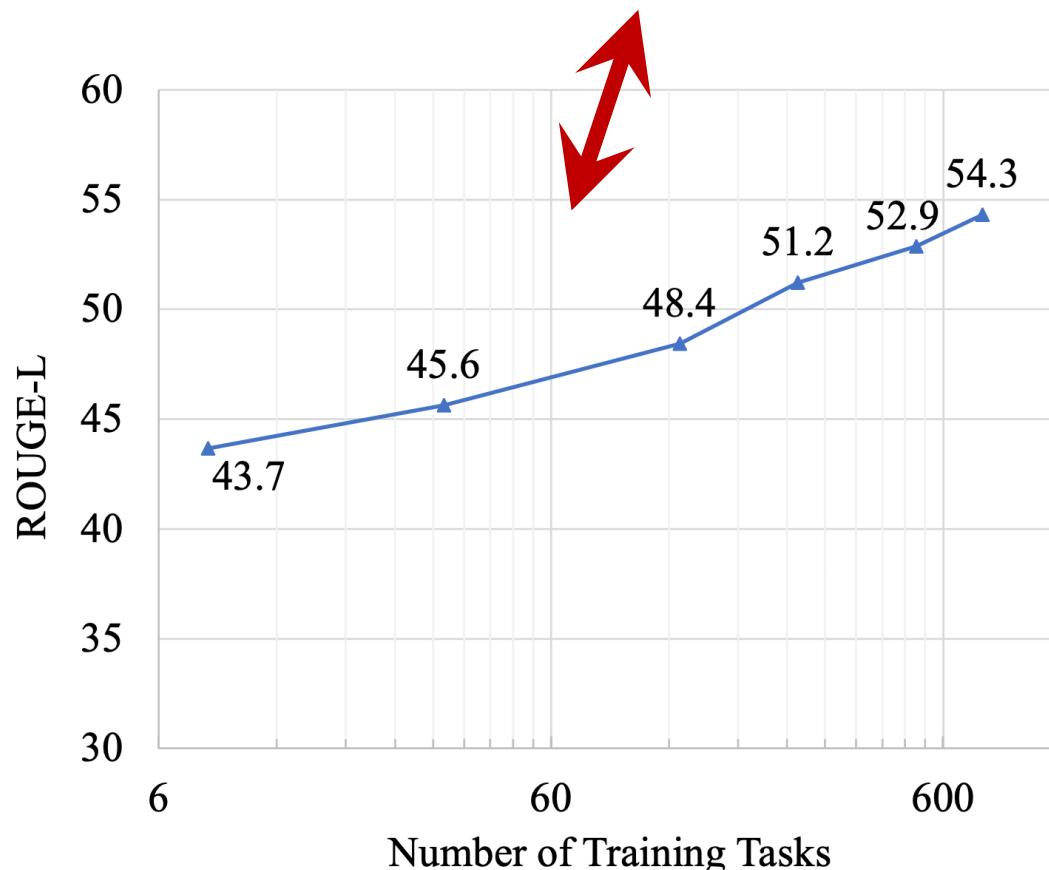
Impact of Supervision Diversity

Diverse tasks vs. diverse variants of fixed tasks.



Impact of Supervision Diversity

Diverse tasks >> diverse variants of fixed tasks.



Not “Diverse” Data => Not “Generalist” Model

Not “Diverse” Data => Not “Generalist” Model

Tk-Instruct (3B)

⚡ Hosted inference API ⓘ

Text2Text Generation

Definition: Write a sentence with the following words. Your output should contain all the words.

Now complete the following example -

Input: Apple, cash, tech.

Output:

Compute

⌘+Enter

1.9

Computation time on cpu: 1.855 s

Apple cash is the new tech.



Not “Diverse” Data => Not “Generalist” Model

Tk-Instruct (3B)

⚡ Hosted inference API ⓘ

▫ Text2Text Generation

Definition: Write a sentence with the following words. Your output should contain all the words.

Now complete the following example -

Input: Apple, cash, tech.

Output:

Compute

⌘+Enter

1.9

Computation time on cpu: 1.855 s

Apple cash is the new tech.



Tk-Instruct (3B)

⚡ Hosted inference API ⓘ

▫ Text2Text Generation

Write a sentence with the following words. Your output should contain all the words.
Input: Apple, cash, tech.

Compute

⌘+Enter

1.9

Computation time on cpu: 1.616 s

Apple, cash, tech.

Not “Diverse” Data => Not “Generalist” Model



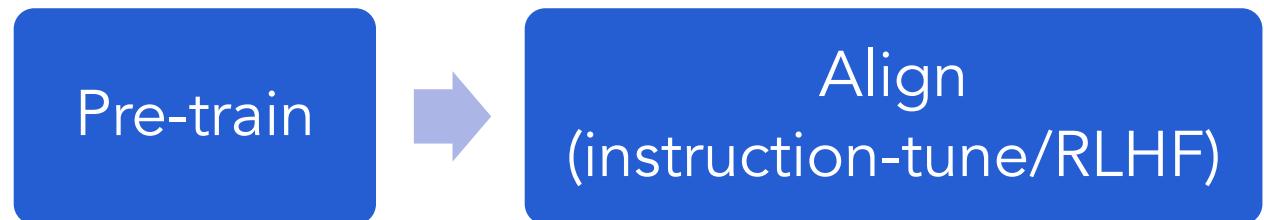
Evaluated on xP3
Evaluated on Super-NaturalInstructions



Supervision Diversity

- What are the dimensions of “diversity”?
 - Diversity of tasks
 - Diversity of how inputs (demands) are phrased?
 - Diversity of expected outputs?
 -
- How do you go about building a supervision data with maximal “diversity”?

Supervision size
and diversity



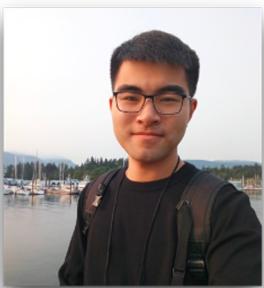
Optimizing for Supervision Diversity: Failures

We did a pilot study but found that:

- Writing diverse instructions requires creativity.
- Writing instances for different instructions requires broad expertise.
- **Impractical** for crowd workers.

Self-Instruct: Aligning Language Models with Self-Generated Instructions

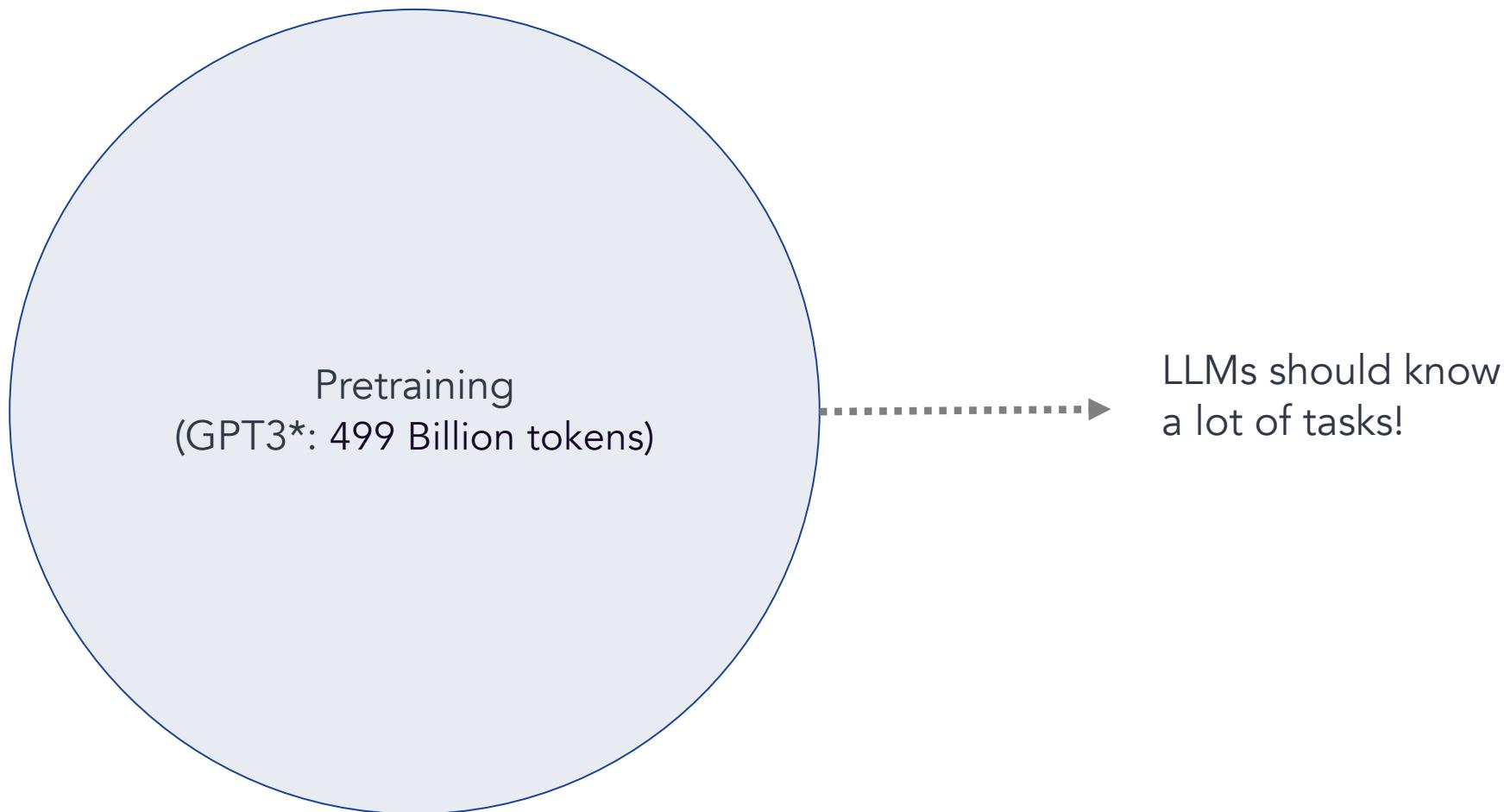
Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu,
Noah A. Smith, Daniel Khashabi, Hannaneh Hajishirzi



<https://arxiv.org/abs/2212.10560>

Can LMs Help with Generating Instruction Data?

- Idea: we can bootstrap “instruction” from off-the-shelf LMs.



LMs Can be Prompted to Generate Instructions

Come up with a series of tasks:

- 1. I am planning a 7-day trip to Seattle. Can you make a detailed plan for me?
- 2. Is there anything I can eat for a breakfast that doesn't include eggs, yet includes protein, and has roughly 700-1000 calories?
- 3. Translate this sentence into any Arabic dialect and say what dialect it is: "The beautiful ones are not yet born"
- 4. Given a set of numbers, find all possible subsets that sum to a given number.
- 5. Give me a phrase that I can use to express I am very happy.
- 6. Create a birthday planning checklist.
- 7. What is the relation between the given pairs?
- 8. Tell if the sentence contains toxic language.
- 9. Create a list of 10 words that have the letter "k" in them.
- 10. Create a list of all countries and their capitals from Africa.
- 11. I am looking for a job, but it's difficult to find one. Can you help me?
- 12. Find the longest word in this sentence: The quick brown fox jumps over the lazy dog.
- 13. Translate these numbers into English (word format): 3168987645561523
- 14. What is the next number in the series: 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096...?
- 15. Write a program that will tell me if a given string contains anagrams.

Submit     

Mode 

Model

Temperature 0.7 

Maximum length 1024 

Stop sequences Enter sequence and press Tab
16. 

Top P 0.5 

Frequency penalty 0 

Presence penalty 2 

Best of 1 

163

LMs Can be Prompted to Generate Instructions

Come up with a series of tasks:

- 1. I am planning a 7-day trip to Seattle. Can you make a detailed plan for me?
- 2. Is there anything I can eat for a breakfast that doesn't include eggs, yet includes protein, and has roughly 700-1000 calories?
- 3. Translate this sentence into any Arabic dialect and say what dialect it is: "The beautiful ones are not yet born"
- 4. Given a set of numbers, find all possible subsets that sum to a given number.
- 5. Give me a phrase that I can use to express I am very happy.
- 6. Create a birthday planning checklist.
- 7. What is the relation between the given pairs?
- 8. Tell if the sentence contains toxic language

Mode

Model

Temperature 0.7

Maximum length 1024

7. What is the relation between the given pairs?
8. Tell if the sentence contains toxic language.
9. Create a list of 10 words that have the letter "k" in them.
10. Create a list of all countries and their capitals from Africa.
11. I am looking for a job, but it's difficult to find one. Can you help me?
12. Find the longest word in this sentence: The quick brown fox jumps over the lazy dog.
13. Translate these numbers into English (word format): 3168987645561523
14. What is the next number in the series: 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096...?
15. Write a program that will tell me if a given string contains anagrams.

LMs Can be Prompted to Generate Responses

Come up with an example for each of the following task. Each example must have one output field. If the task requires input, it should be generated before the output.

Task 1: Make a list of things to do in the given city.
Input: ...
Output: ...

Task 2: Converting 85 F to Celsius.
Output: ...

Task 3: Extract all the country names in the paragraph, list them separated by comma.
Paragraph: ...
Output: ...

Task 4: Suggest a better and more professional rephrasing of the following sentence.
Sentence: ...
Output: ...

Task 5: Read the following paragraph and answer a math question about the paragraph. You need to write out the calculation for getting the final answer.
Paragraph: ...
Question: ...
Output: ...

Task 6: Solving the equation and find the value of X.
Equation: ...
Output: ...

Task 7: Write a knock knock joke about bananas.
Output:

Task 8: Tell me whether the given sentence is passive or not.
Sentence: The dog was bitten by the cat.
Output: Passive, because the subject of the sentence is being acted upon (the dog).

Mode:

Model: davinci

Temperature: 0

Maximum length: 1024

Stop sequences: Enter sequence and press Tab

Top P: 1

Frequency penalty: 0

Presence penalty: 2

Best of: 1

Inject start text:

Inject restart text:

Show probabilities: Off

Submit

267

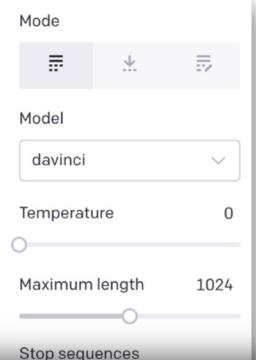
LMs Can be Prompted to Generate Responses

Come up with an example for each of the following task. Each example must have one output field. If the task requires input, it should be generated before the output. 

Task 1: Make a list of things to do in the given city.
Input: ...
Output: ...

Task 2: Converting 85 F to Celsius.
Output: ...

Task 3: Extract all the country names in the paragraph, list them separated by comma.
Paragraph: ...
Output: ...



Task 7: Write a knock knock joke about bananas.

Output:

Task 8: Tell me whether the given sentence is passive or not.

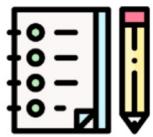
Sentence: The dog was bitten by the cat.

Output: Passive, because the subject of the sentence is being acted upon (the dog).

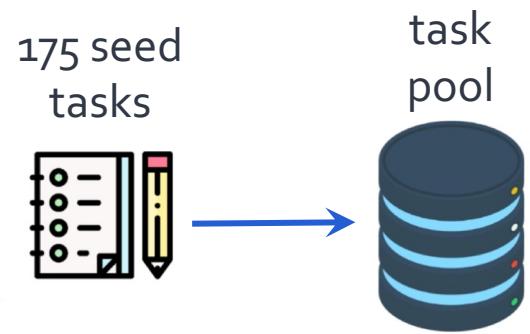
Instruction Data Generation Pipeline

Instruction Data Generation Pipeline

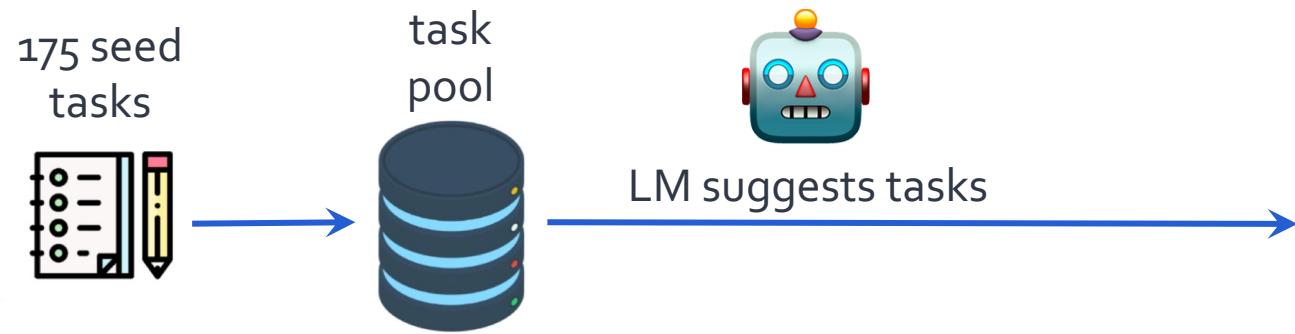
175 seed
tasks



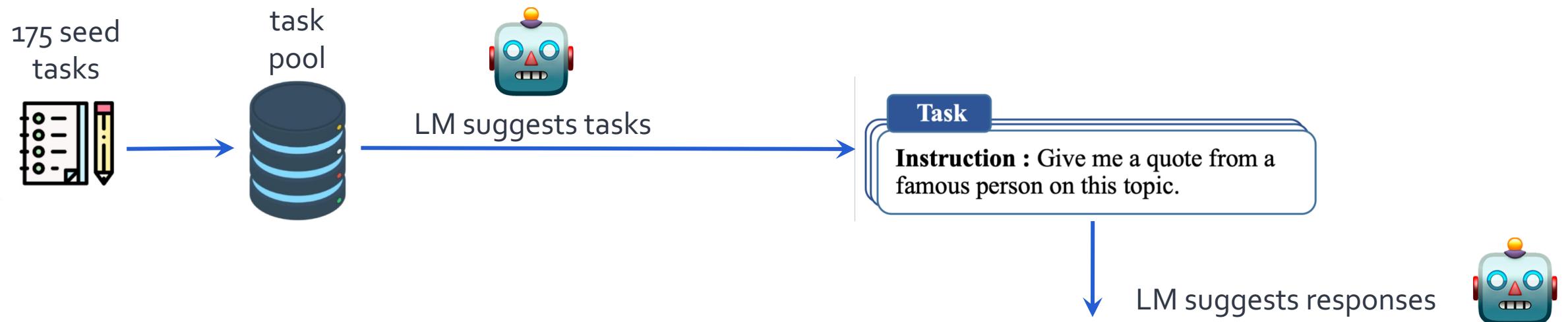
Instruction Data Generation Pipeline



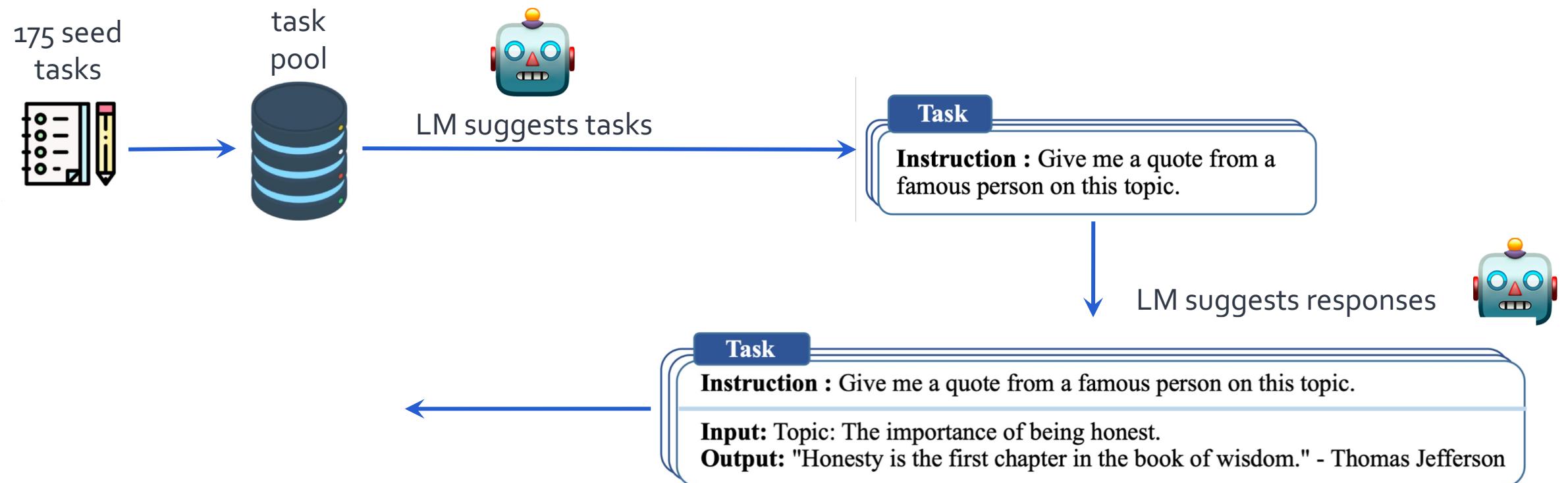
Instruction Data Generation Pipeline



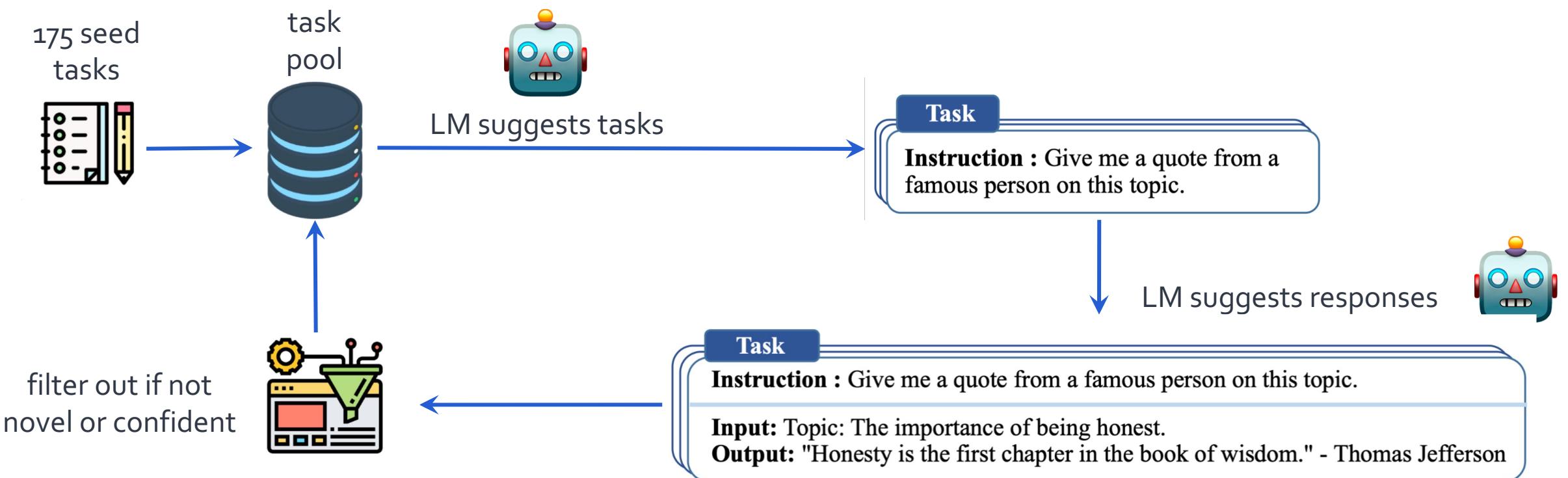
Instruction Data Generation Pipeline



Instruction Data Generation Pipeline

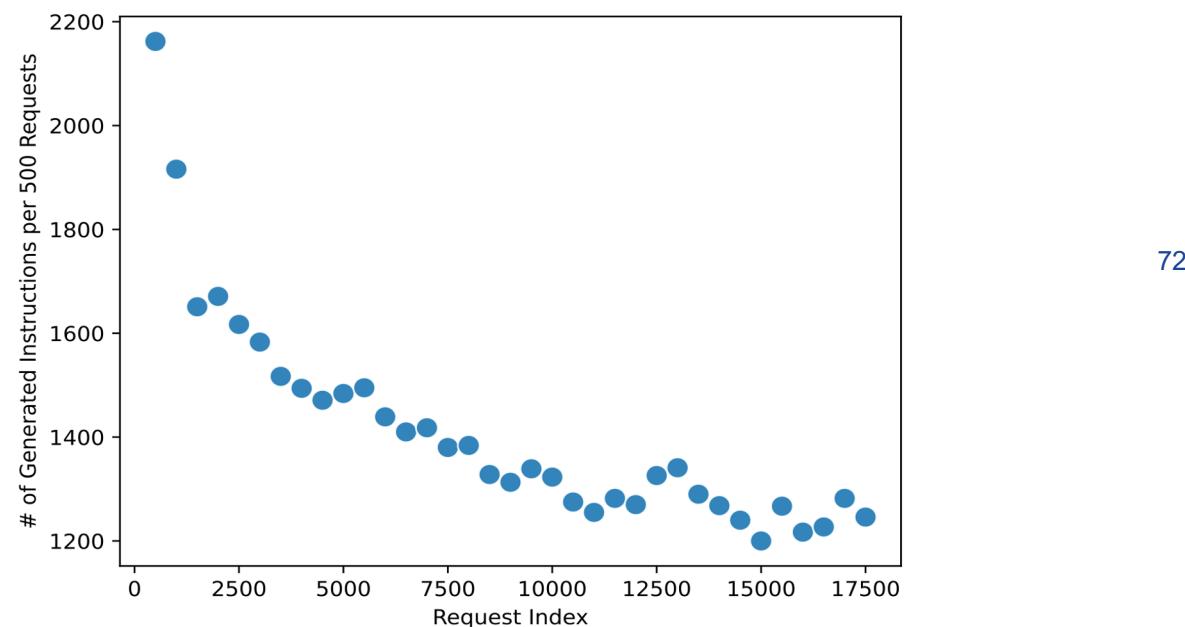


Instruction Data Generation Pipeline



Generating 52K instructions with GPT3

- We applied self-instruct to GPT3 (`davinci` engine).
- We generated 52K instructions and 82K instances.
- ROUGE-L overlap between any pair of instructions < 0.7.
- The success rate per request decreases as the task pool grows.



Example Tasks Generated by GPT3 (before RLHF)

Instruction: I am looking for a job and I need to fill out an application form. Can you please help me complete it?

Input:

Application Form:

Name: _____ Age: _____ Sex: _____

Phone Number: _____ Email Address: _____

Education: _____

...

Output:

Name: John Doe Age: 25 Sex: Male Phone Number: 555-555-5555

...

Example Tasks Generated by GPT3 (before RLHF)

Instruction: Write a letter from the perspective of a cat.

Input: Null

Output:

Dear [Owner], I am writing to you today because I have a problem. You see, I've been living with you for the past [Number of Years] and I love it here. The food is great, the bed is comfy, and there are plenty of toys to play with. But lately, I've noticed that you're not as affectionate towards me as you used to be. In fact, sometimes when I try to cuddle up next to you on the couch, you push me away ...



Example Tasks Generated by GPT3 (before RLHF)



Instruction: I am looking for a new apartment. There are many options but they are all quite expensive. I want to move as soon as possible. Give me advice on what to do.

Input: Null

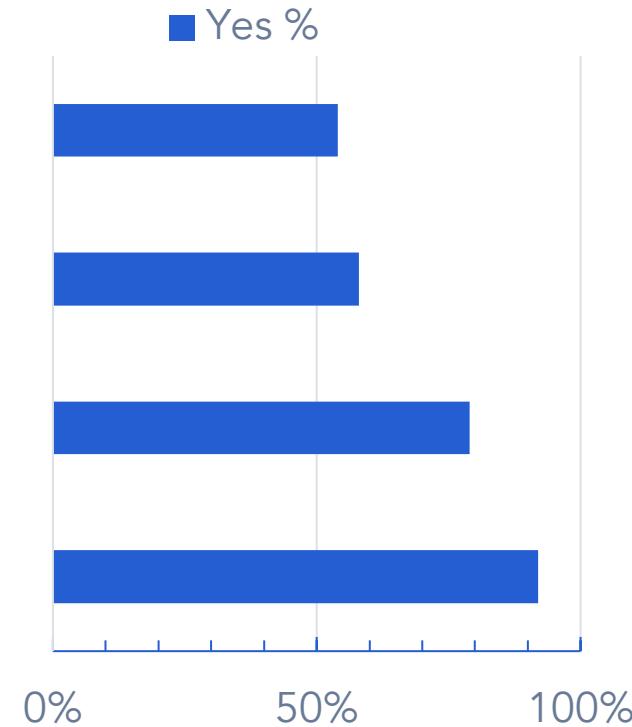
Output:

You should look for a cheaper apartment. You can try to negotiate with the landlord and offer him/her a lower price. If you are lucky, they will accept your offer. Otherwise ...

How Accurate is This Data?

- 200 random instructions are sampled for quality check

All fields are valid.
Is the output a correct and acceptable response to the instruction and input?
Is the input appropriate for the instruction?
Does the instruction describe a valid task?



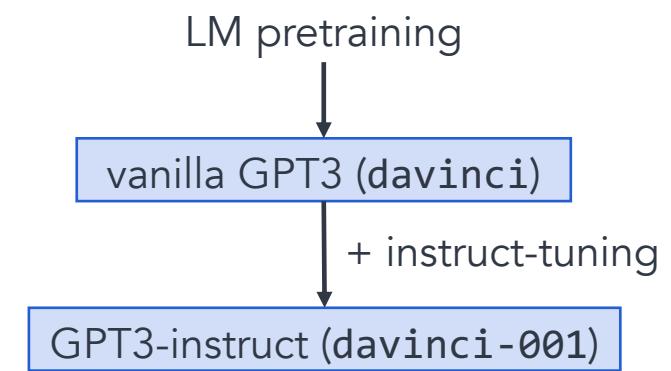
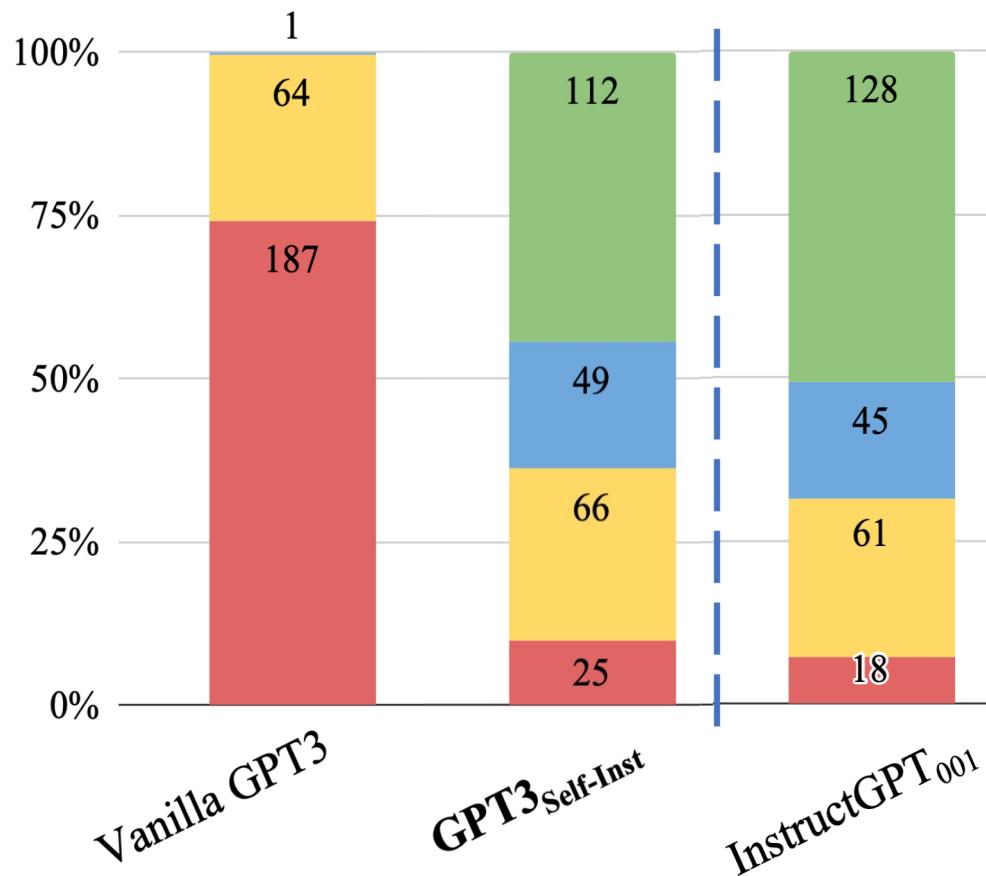
Self-Instructing GPT3

- **Generate:**
 - We applied Self-Instruct to GPT3 (“davinci” engine).
 - We generated 52K instructions and 82K instances.
 - API cost ~\$600
- **Align:**
 - We finetuned GPT3 with this data via OpenAI API (2 epochs). **
 - API cost: ~\$338 for finetuning

(** OpenAI training API is unclear about how it works, or how the parameters are updated.)

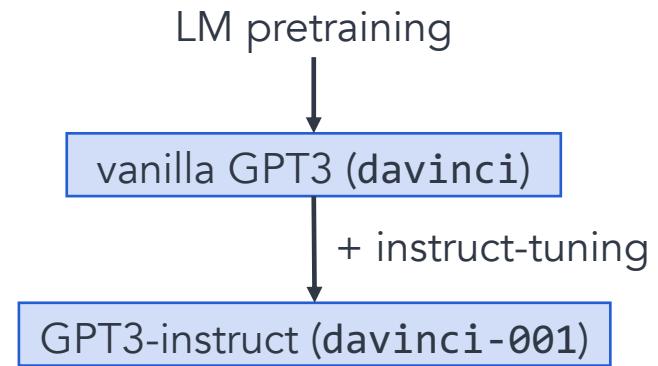
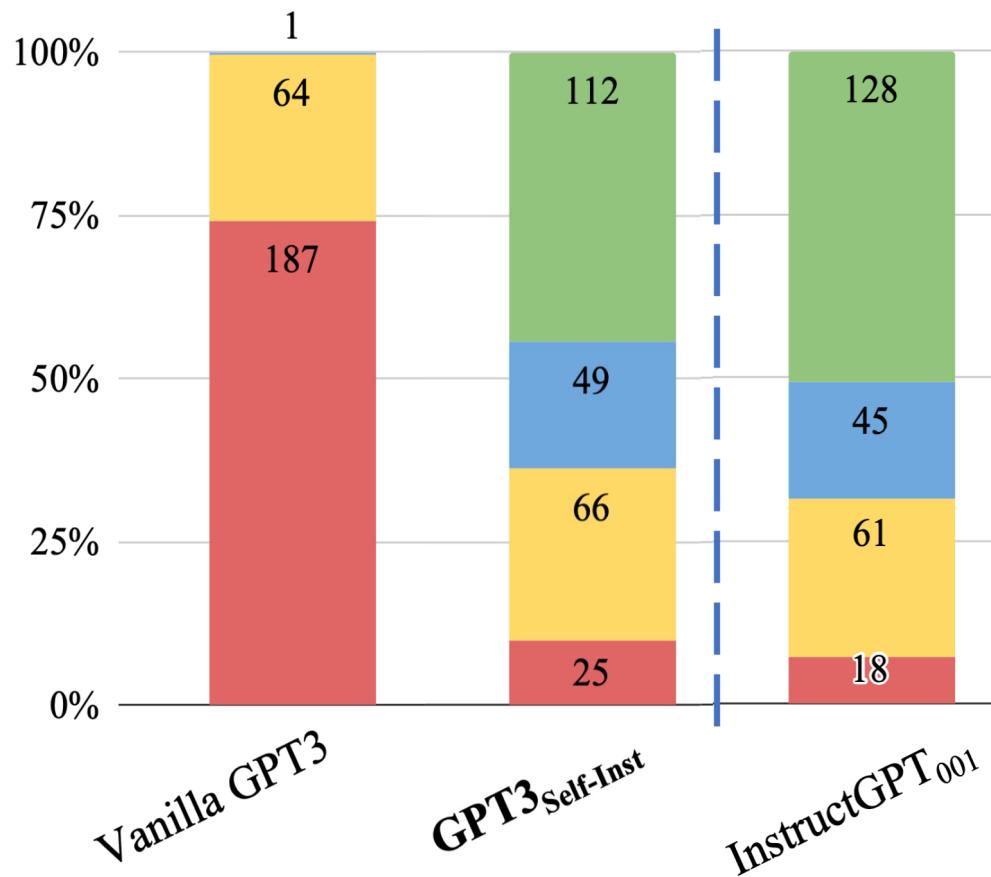
Evaluation on User-Oriented Instructions

- A: correct and satisfying response
- B: acceptable response with minor imperfections
- C: responds to the instruction but has significant errors
- D: irrelevant or invalid response



Evaluation on User-Oriented Instructions

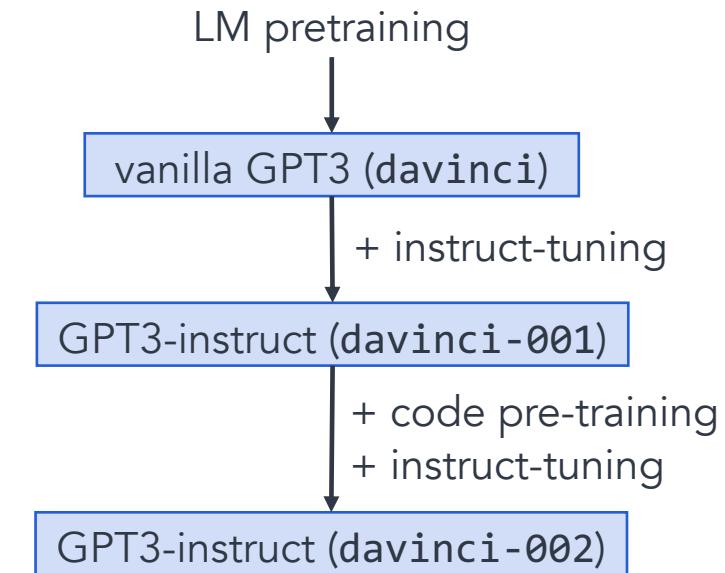
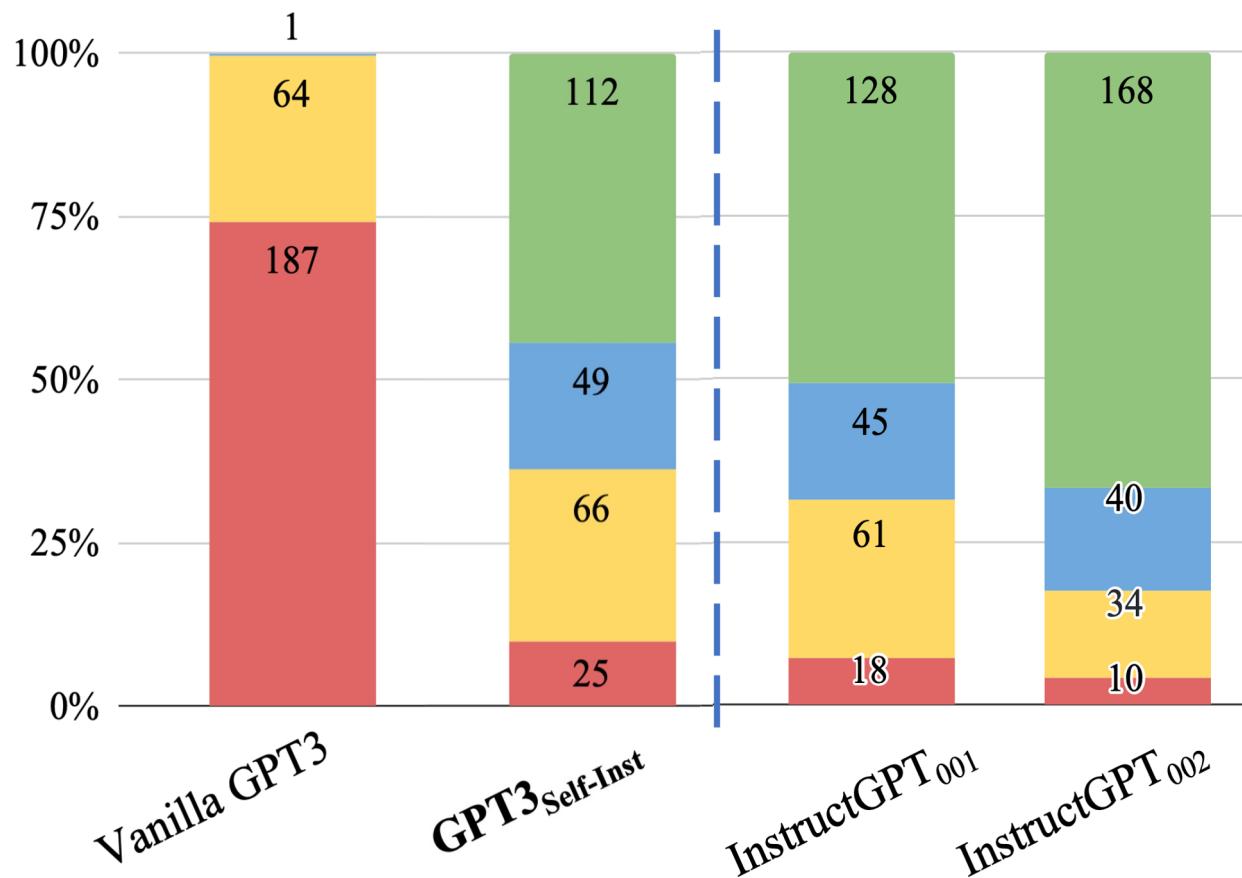
- A: correct and satisfying response
- B: acceptable response with minor imperfections
- C: responds to the instruction but has significant errors
- D: irrelevant or invalid response



Noisy, but diverse “self-instruct” data ~ thousands of clean human-written data

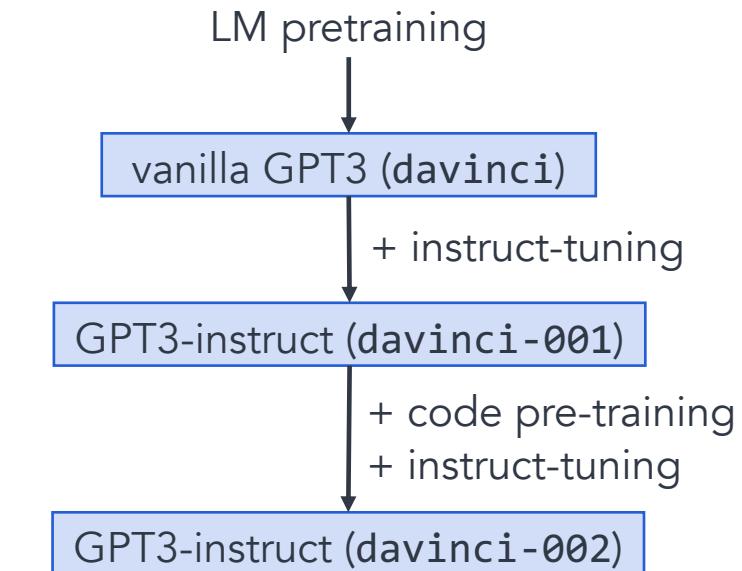
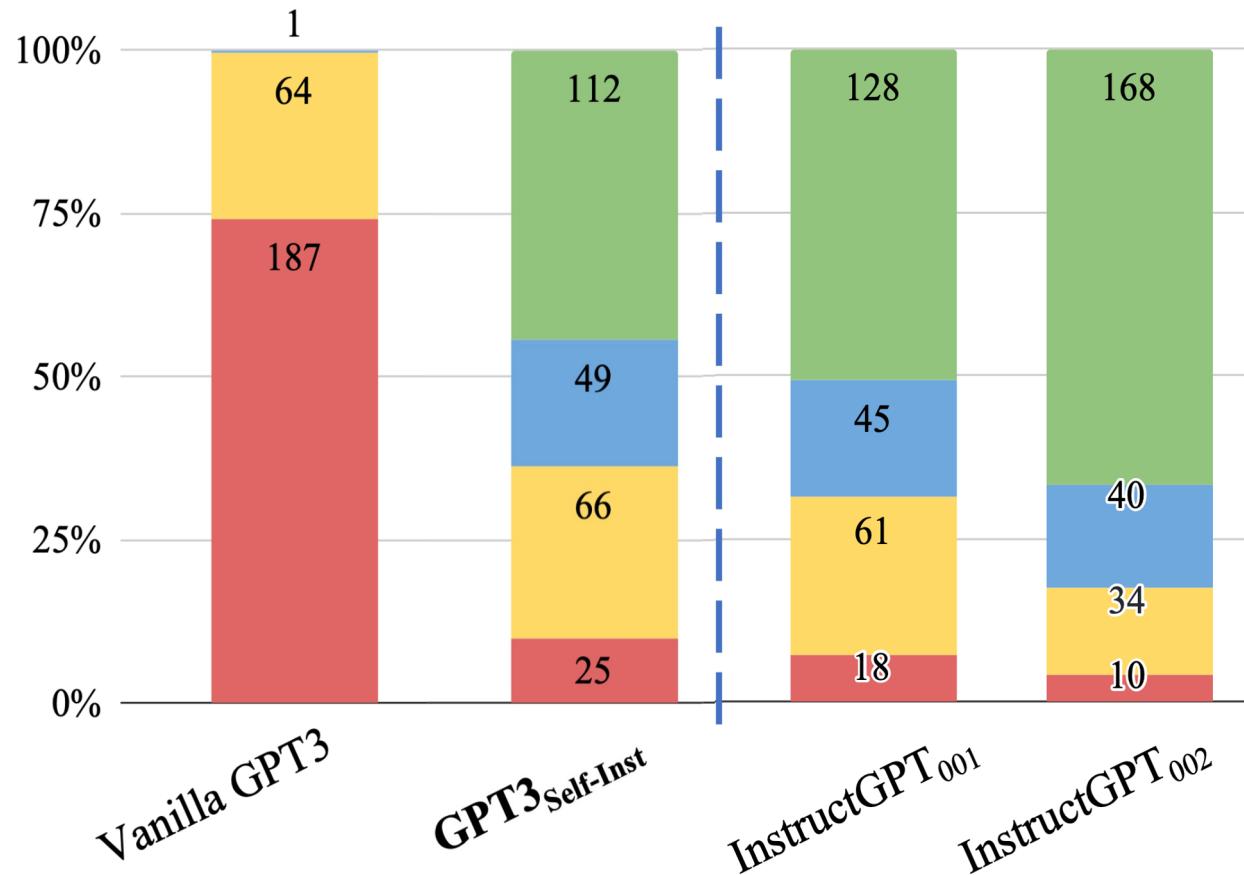
Evaluation on User-Oriented Instructions

- A: correct and satisfying response
- B: acceptable response with minor imperfections
- C: responds to the instruction but has significant errors
- D: irrelevant or invalid response



Evaluation on User-Oriented Instructions

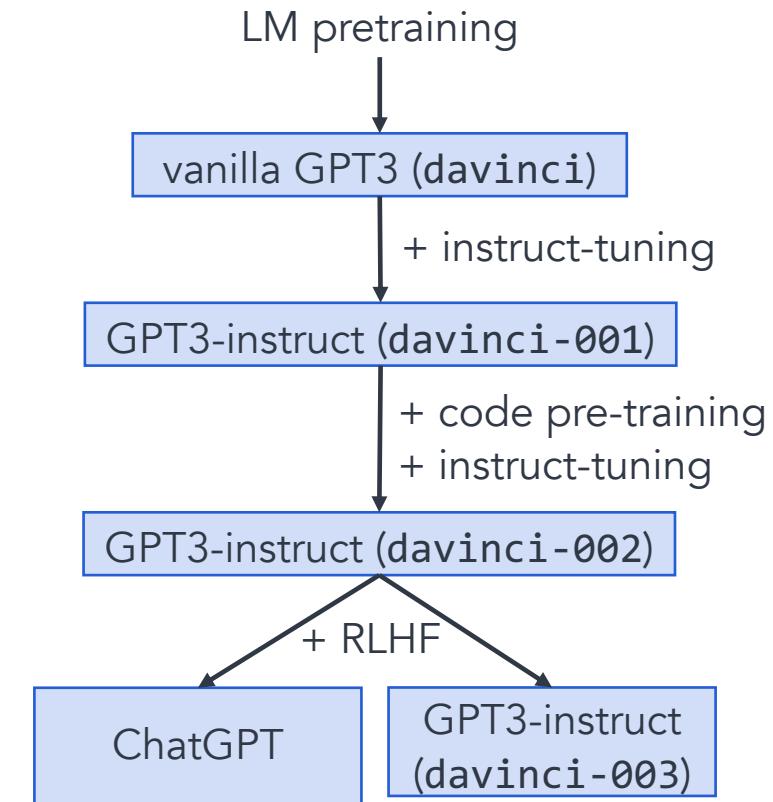
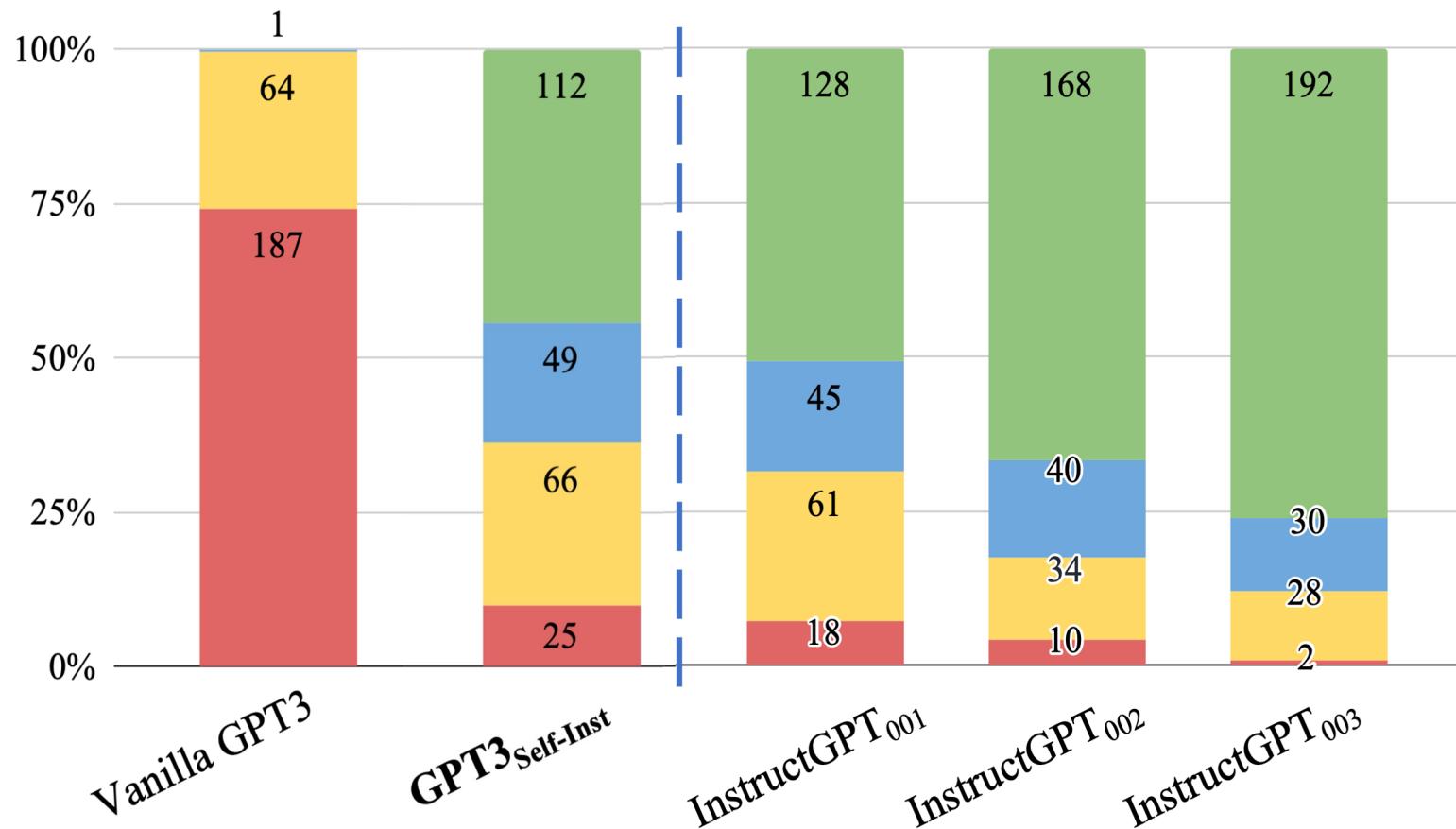
- A: correct and satisfying response
- B: acceptable response with minor imperfections
- C: responds to the instruction but has significant errors
- D: irrelevant or invalid response



Pre-trained on code ⇒
better base LM?

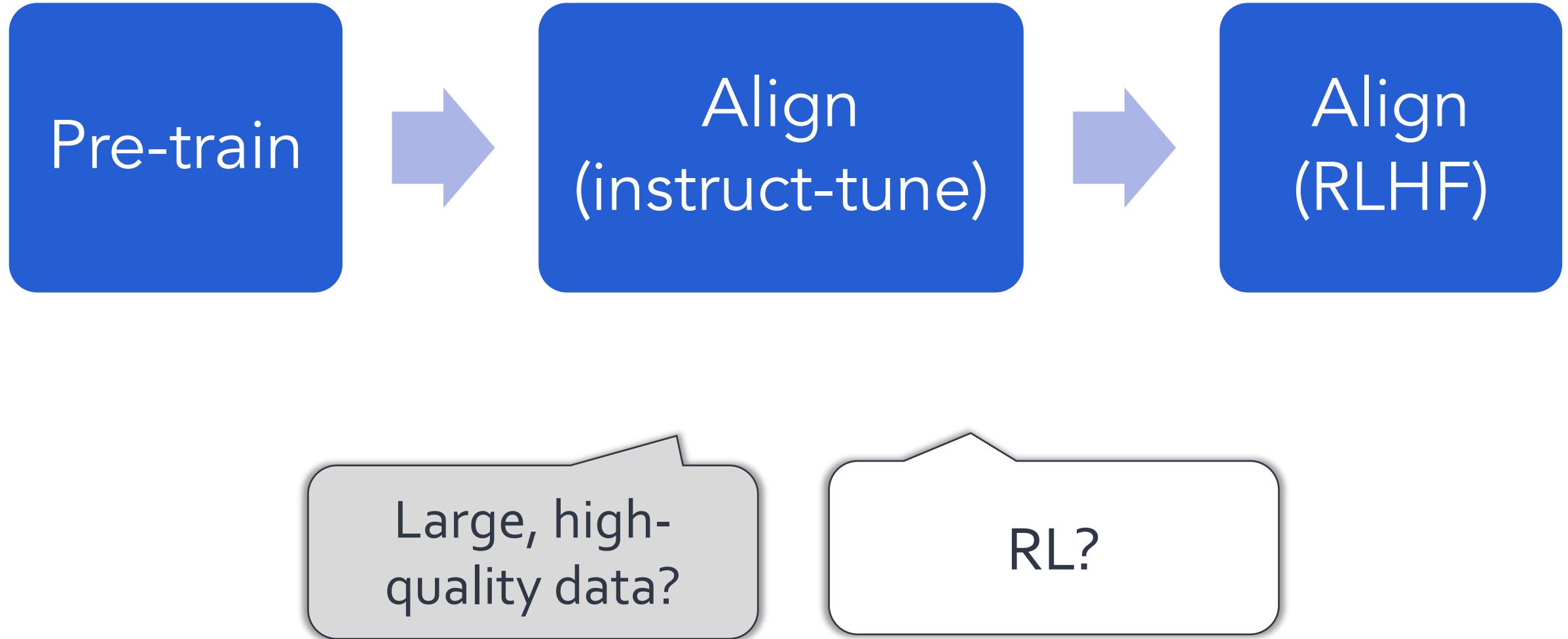
Evaluation on User-Oriented Instructions

- A: correct and satisfying response
- B: acceptable response with minor imperfections
- C: responds to the instruction but has significant errors
- D: irrelevant or invalid response



Summary Thus Far

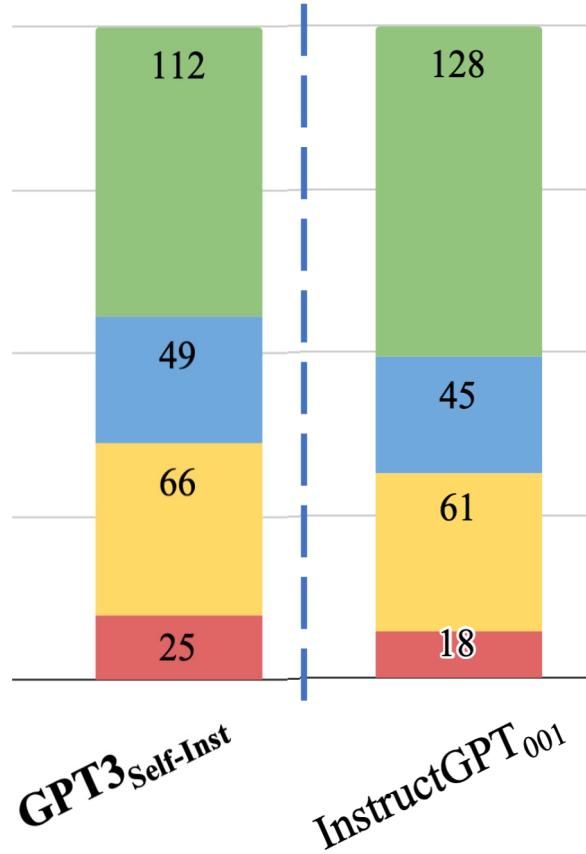
- Data diversity seems to be necessary for building successful generalist models.
- We don't understand how to **maximize** the diversity of "alignment" data.
- Self-Instruct:
 - Rely on creativity induced by an LLM's themselves.
 - Applicable to a broad range of LLMs.
 - Stanford Alpaca is based on "Self-Instruct" data.



Diverse contexts

But Wait a Sec ...

- So, we used LM to generate data for aligning **itself**??



Step #1:
Pre-train



Step #2/3: Align
(RLHF or instruction-tune)

But Wait a Sec ...



But Wait a Sec ...

- Fundamentally, what is the role of post hoc alignment (step #2/3)?
 1. Teaching LM knowledge of new tasks?
 2. Lightly modify LM so it can articulate its existing knowledge of tasks?

(+ put guardrails for what it can articulate)



Implications for What to Invest In

- Fundamentally, what is the role of post hoc alignment (step #2/3)?

1. Teaching LM knowledge of new tasks?

Identify what knowledge needs to be taught.

2. Lightly modify LM so it can articulate its existing knowledge of tasks?

(+ put guardrails for what it can articulate)

Make it more efficient, possibly with minimal human labor.

Step #1:
Pre-train



Step #2/3: Align
(RLHF or instruction-tune)

Implications for What Comes Out

- Fundamentally, what is the role of post hoc alignment (step #2/3)?

1. Teaching LM knowledge of new tasks?

It will be as good as the alignment supervision.

2. Lightly modify LM so it can articulate its existing knowledge of tasks?

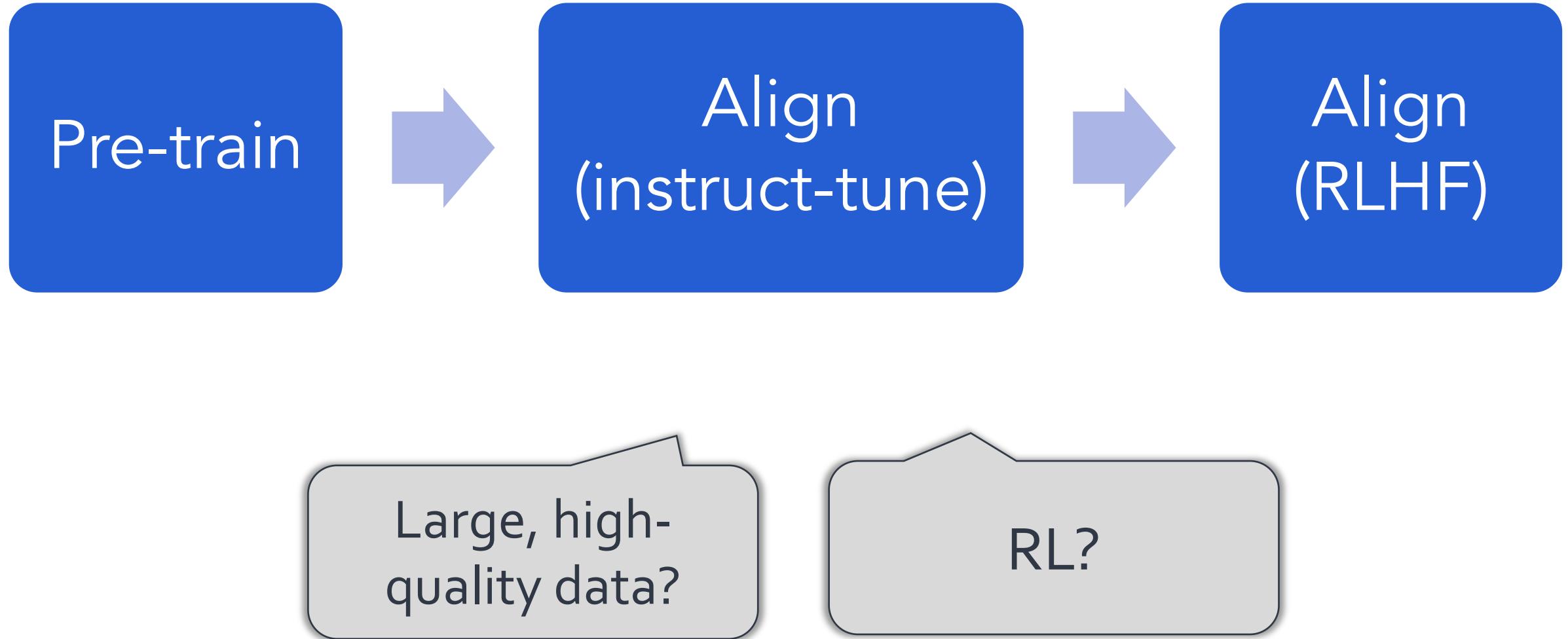
(+ put guardrails for what it can articulate)

Unexpected behaviors could “emerge”.

Step #1:
Pre-train



Step #2/3: Align
(RLHF or instruction-tune)



Diverse contexts

Is RL [in RLHF] Necessary?

Is RL [in RLHF] Necessary?

- My short answer: no.

Is RL [in RLHF] Necessary?

- My short answer: no.

rl-for-llms.md

Raw

Reinforcement Learning for Language Models

Yoav Goldberg, April 2023.

Why RL?

With the release of the ChatGPT model and followup large language models (LLMs), there was a lot of discussion of the importance of "RLHF training", that is, "reinforcement learning from human feedback". I was puzzled for a while as to why RL (Reinforcement Learning) is better than learning from demonstrations (a.k.a supervised learning) for training language models. Shouldn't learning from demonstrations (or, in language model terminology "instruction fine tuning", learning to immitate human written answers) be sufficient? I came up with a theoretical argument that was somewhat convincing. But I came to realize there is an additional argument which not only *supports* the case of RL training, but also *requires* it, in particular for models like ChatGPT. This additional argument is spelled out in (the first half of) a [talk by John Schulman from OpenAI](#). This post pretty much repeats his argument in more words, and also adds some things that John did not say explicitly (but which I am sure he thought about).

Arguments for RL: Diversity

- Extensions of “Self-Instruct” will go a long way.

Arguments for RL: Ease of Feedback

- Ranking answers is easier than generating them.
- I agree.
- Applying ranking feedback does not necessitate RL.
 - E.g., search engine optimization based on ranking feedback.

Arguments for RL: Reduces Hallucination

- To reduce hallucination “we want to encourage the model to answer based on its internal knowledge”, rather than forcing it to improvise.
- Supervised learning on human-labeled data can’t do this, but RL can. => **agreed**.
- “Self-Instruct” can do this too.

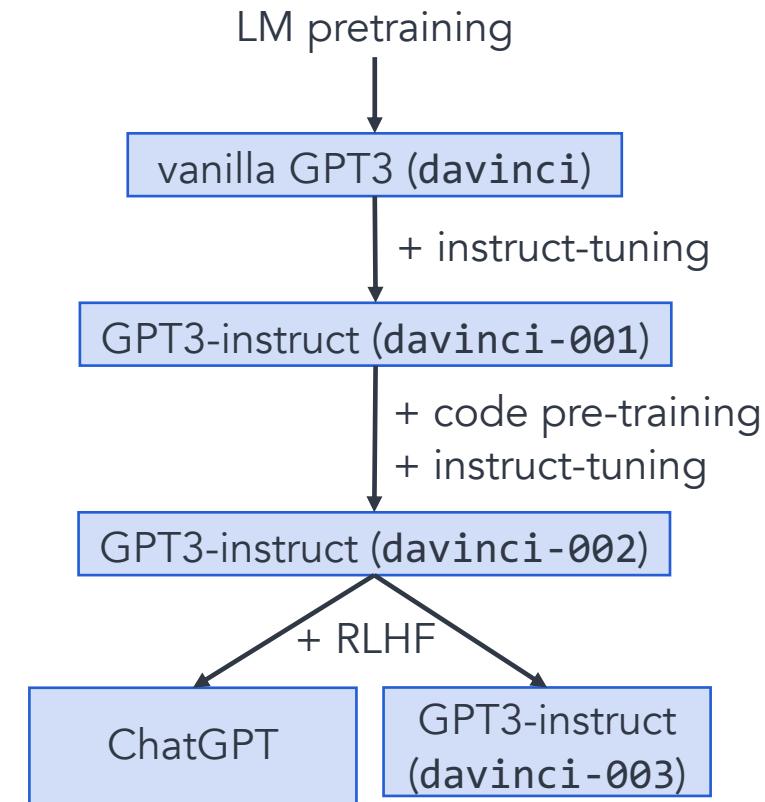
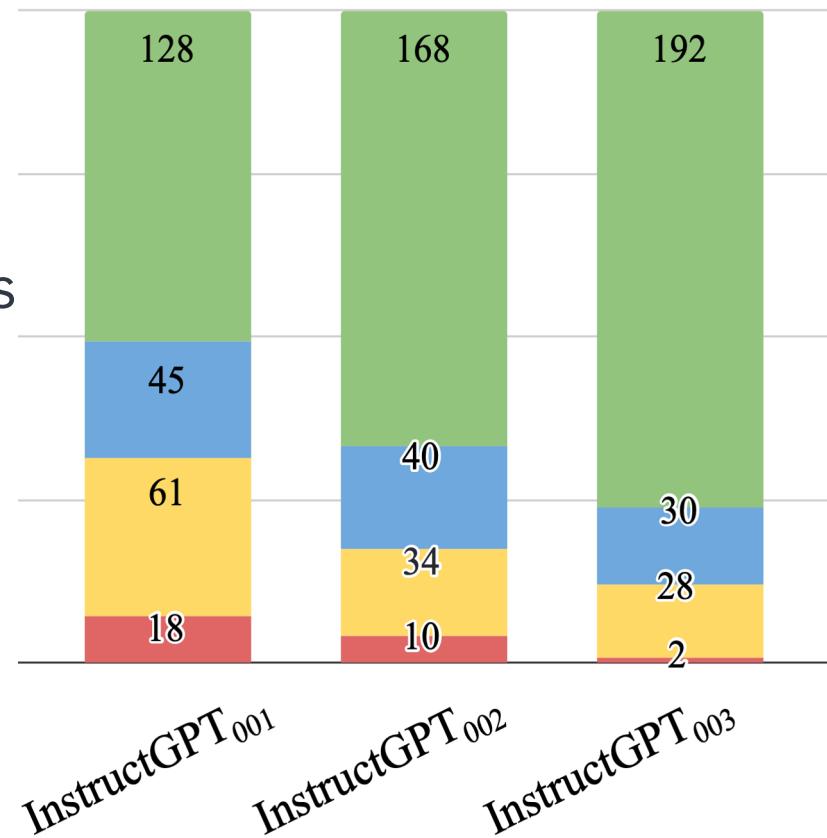
Coding Models are RLHF-ed Out of Box

- LMs that are pre-trained on Github, are **good** at following human **intents**.
- A particular domain where there is a natural pairing of form and intent.

Surely You Can't Deny the Numbers ...

- A: correct and satisfying response
- B: acceptable response with minor imperfections
- C: responds to the instruction but has significant errors
- D: irrelevant or invalid response

It is not clear whether these gains are purely due to RL since OpenAI is leveraging its massive query log, raising concerns about test/train overlap.



RLHF is Patchwork for Lack of Grounding

- This helps LMs learn (ground) the communicative **intent** of a user who asks for a "summary" in its instruction.
 - For example, what is **intended** by "summarize"? The act of producing a summary grounded in the human concept of "summary".
- Not a panacea, but a short-term "band-aid" solution.





Large, high-quality data?

RL?

Diverse contexts

It's complicated

Alignment as a Social Process

- Can alignment emerge as a social experience?

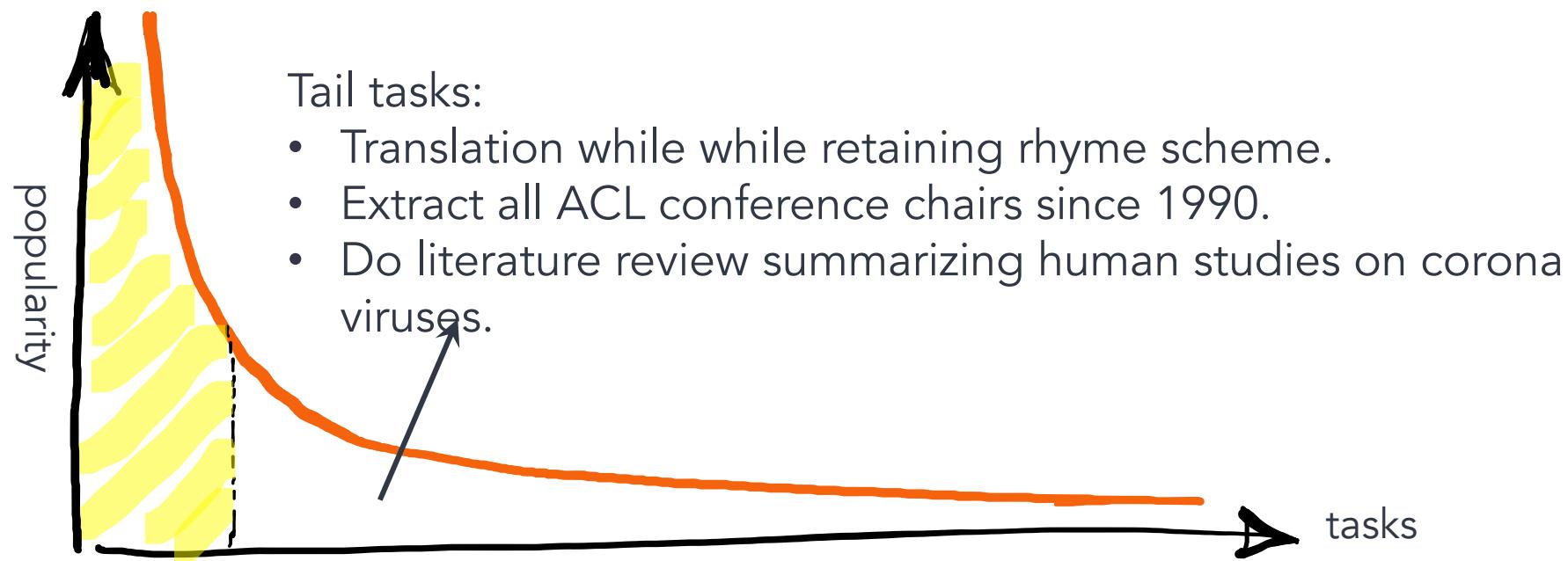


One quick grievance ...

How Should We Evaluate Generalist Chatbots?

No one Knows How to Evaluate Generalist Chatbots!

1. There are infinite-many tasks out there
 - Some are not defined yet!!



No one Knows How to Evaluate Generalist Chatbots!

1. There are infinite-many tasks out there
 - Some are not defined yet!!
2. What humans want != NLP tasks
 - Benchmarks like PromptSource or Natural Instructions are good indicators of a chatbot's real-world quality.

No one Knows How to Evaluate Generalist Chatbots!

1. There are infinite-many tasks out there
 - Some are not defined yet!!
2. What humans want != NLP tasks
 - Benchmarks like PromptSource or Natural Instructions are good indicators of a chatbot's real-world quality.

If we show the weakness of a chatbot in [NLP] tasks, is that a weakness of the model or the chatbot? 🤔

No one Knows How to Evaluate Generalist Chatbots!

1. There are infinite-many tasks out there
 - Some are not defined yet!!
2. What humans want != NLP tasks
 - Benchmarks like PromptSource or Natural Instructions are **not** good indicators of a chatbot's real-world quality.

*"We find that ChatGPT faces challenges when solving specific tasks such as **sequence tagging**."*



No one Knows How to Evaluate Generalist Chatbots!

1. There are infinite-many tasks out there
 - Some are not defined yet!!
2. What humans want != NLP tasks
 - Benchmarks like PromptSource or Natural Instructions are **not** good indicators of a chatbot's real-world quality.
3. With the increasing quality of chatbots, it is getting incredibly difficult to define rate quality

No one Knows How to Evaluate Generalist Chatbots!

1. There are infinite-many tasks out there
 - Some are not defined yet!!
2. What humans want != NLP tasks
 - Benchmarks like PromptSource or Natural Instructions are **not** good indicators of a chatbot's real-world quality.
3. With the increasing quality of chatbots, it is getting incredibly difficult to define rate quality
4. How do we model opinions and preferences?

Train/Test Split Doesn't Work Anymore

“... after training we released there is a leakage ...”

Train/Test Split Doesn't Work Anymore

- With our **planetary-level** pre-training, the train-test evaluation protocol is no more viable. How should we revise it?
- Most LM papers: *“... after training we released there is a leakage ...”*
- My 2 cents: evaluation on **planetary-scale time-stamped** data.



Putting All Together

- Invitation: let's better understand the fundamentals that lead to **high-quality, efficient** and **generalist** models.
- Some progress, more open questions.

“Moving Fast and Breaking Things”

- We are in the midst of an arms race, driven by market pressures.

“... did not release the details due to competitive landscape ...”



“Moving Fast and Breaking Things”

- We are in the midst of an arms race, driven by market pressures.

“... did not release the details due to competitive landscape ...”

- This can't continue forever.
- Our job: *Moving slow[er] and fixing the broken things.*
 - Efficiency in computation and supervision, fabrications, harms and biases,

A Silver Lining: Lots of Open-Source Activity

- Better base models (e.g., LLaMa)
- Open-source replications of chatbots
- We will benefit from the alliance with the open-source community.



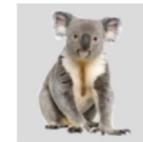
Alpaca
(Stanford)



Vicuna
(UCB)



Baize
(UCSD)



Koala
(UCB)



Instruct-tuning with GPT4
(MSR)

Thanks!