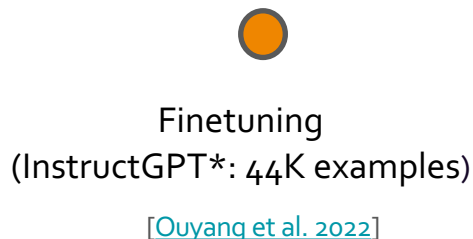
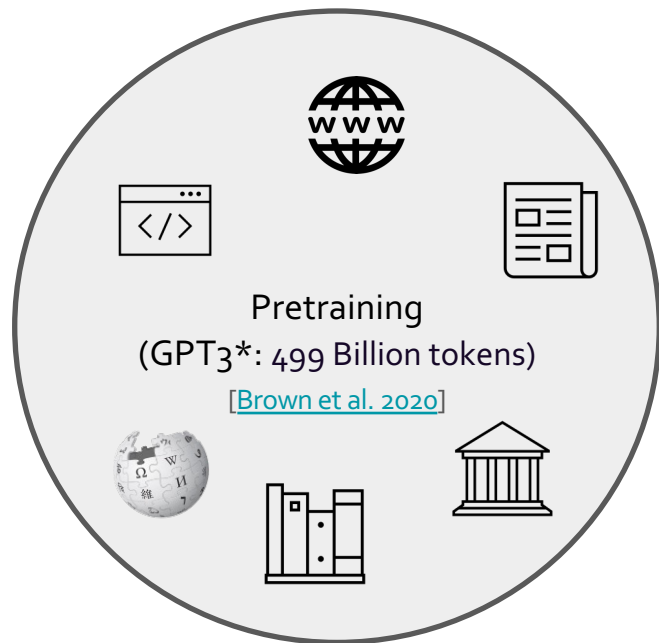


Training Language Models to Follow Instructions

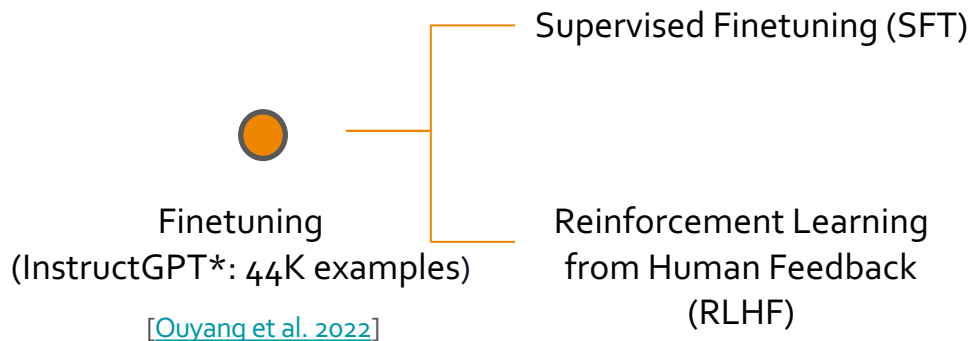
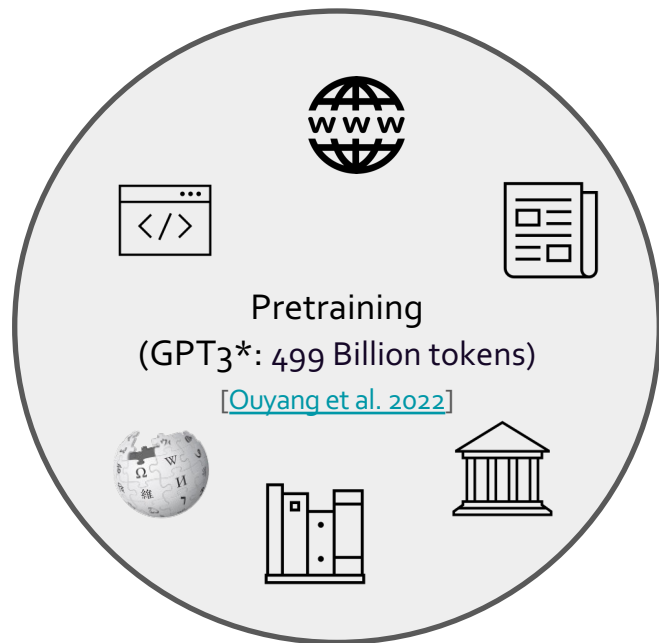
Yizhong Wang, University of Washington

12 April 2024

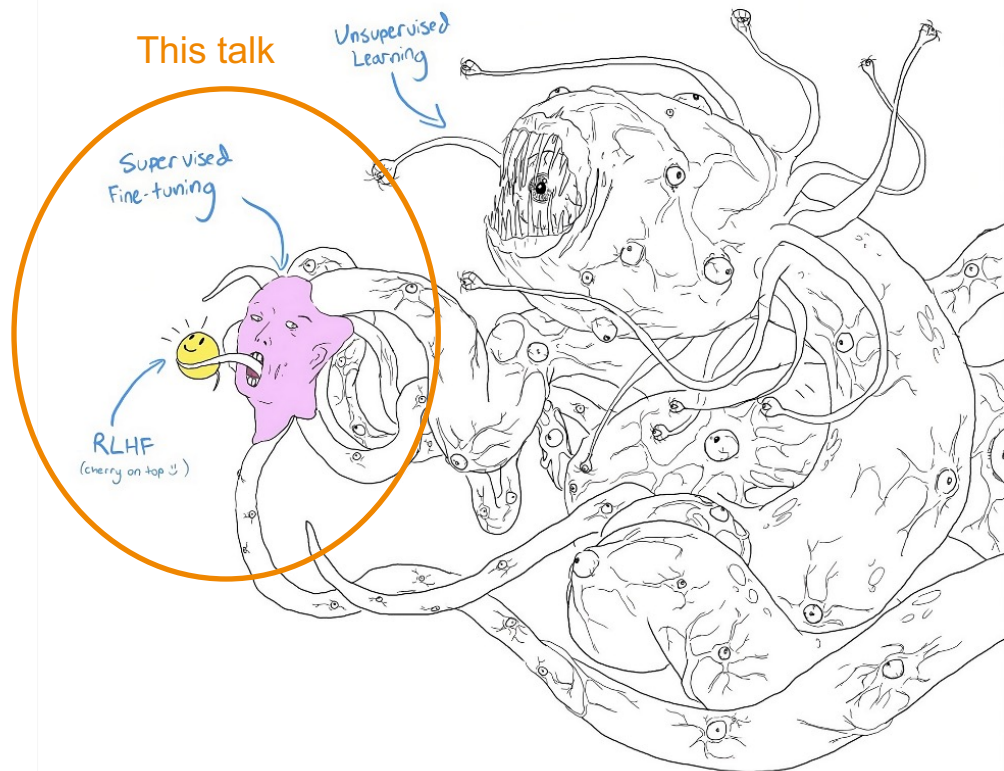
Training Stages for Building ChatGPT-like Language Models



Training Stages for Building ChatGPT-like Language Models



Training Stages for Building ChatGPT-like Language Models



[Figure credit: @anthrupad on twitter]

Overview

- Instruction tuning: concept and early work
- Reinforcement learning from human feedback
- Building open (or your own) ChatGPT

Things that LMs Can Do after Pretraining

- HKU is in _____ Hong Kong. [Trivia]
- I put _____ fork down on the table. [syntax]
- The woman walked across the street, checking for traffic over _____ shoulder. [coreference]
- I went to the ocean to see the fish, turtles, seals, and _____. [lexical semantics/topic]
- What I got from the two hours watching it was popcorn. The movie was _____. [sentiment]
- Thinking about the sequence 1, 1, 2, 3, 5, 8, 13, 21, ____ [basic arithmetic]



Things that LMs Can Do after Pretraining

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

Language Modeling \neq Following Human Instructions

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

Pretrained models are not naturally built to follow user intents.

How do we make LMs aligned
with our intents that are
articulated in language?

Natural Instructions

Give detailed human-readable instructions (that contain examples)

Input: *She chose to make a salad for lunch on Sunday.*
Question: *how long did it take for her to make a salad?*

*tagging
essential
phrases*

Crowdsourcing Instruction: *List all the words that are essential for answering it correctly. [...]*



Output:
*making
salad*

*answering
questions*

Crowdsourcing Instruction:
Answer the provided question based on a given [...]



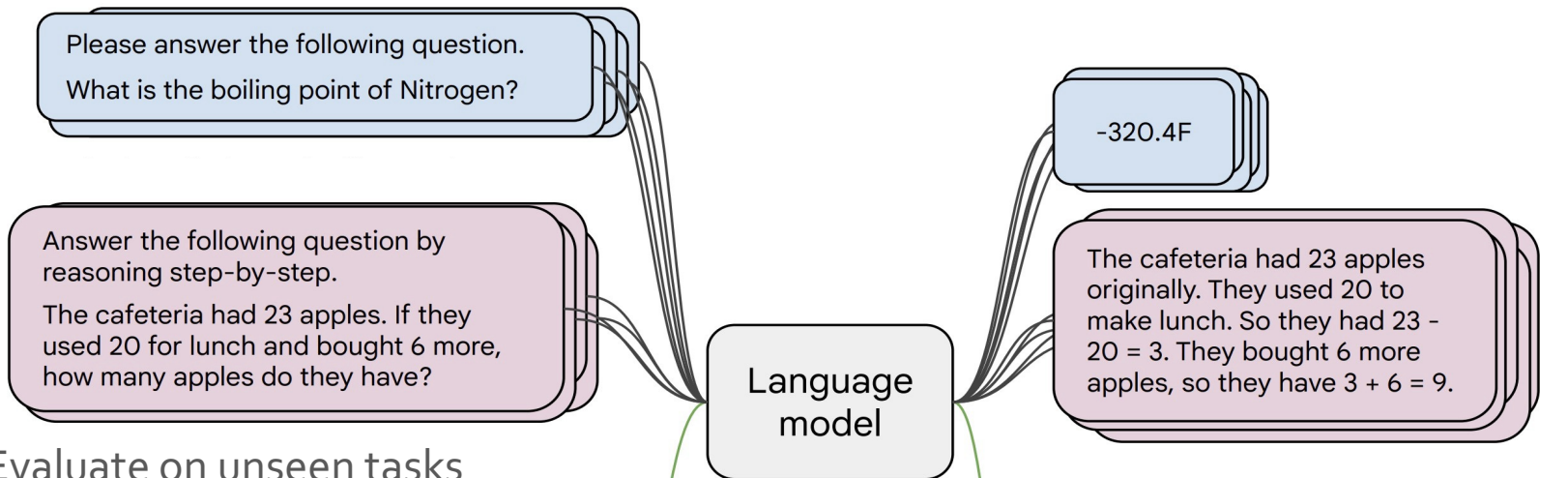
Output:
30mins



Instructions Tuning

[Weller et al. 2020, Mishra et al. 2021, Wang et al. 2022, Sanh et al. 2022; Wei et al., 2022, Chung et al. 2022, many others]

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



2. Evaluate on unseen tasks

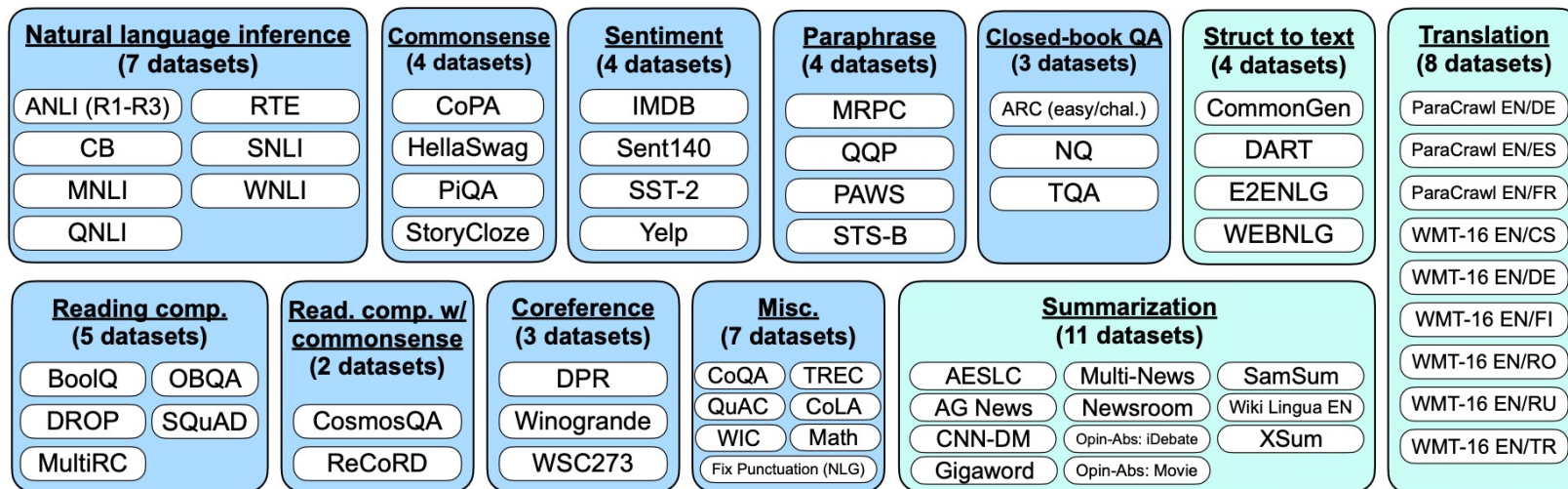
Inference: generalization to unseen tasks

Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

Tasks explored in FLAN

620 prompts on 62 datasets



Tasks Explored in T0

P3: Public Pool of Prompts, now 2085 prompts on 183 datasets

Dataset ?

No of prompts created for **cosmos_qa** : 13

Prompt name ?

cosmos_qa|

description_context_question_text| ▾

- cord19
- cornell_movie_dialog
- cos_e
- cosmos_qa
- covid_qa_castorini
- covid_qa_deepset
- covid_qa_ucsd

- context_answer_to_question
- context_description_question_ans...
- context_description_question_ans...
- context_description_question_text
- context_question_description_ans...
- context_question_description_ans...
- context_question_description_text
- description_context_question_anc

Input template

```
Read the following context and answer the question.
Context: {{ context }}
Question: {{ question }}
Answer:
```

Target template

```
{{ answer_choices[label] }}
```


Instruction-Tuning: Example

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

Before instruction finetuning

The reporter and the chef will discuss their favorite dishes.

The reporter and the chef will discuss the reporter's favorite dishes.

The reporter and the chef will discuss the chef's favorite dishes.

The reporter and the chef will discuss the reporter's and the chef's favorite dishes.

✘ (doesn't answer question)

<https://huggingface.co/google/flan-t5-xxl>

Instruction-Tuning: Example

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

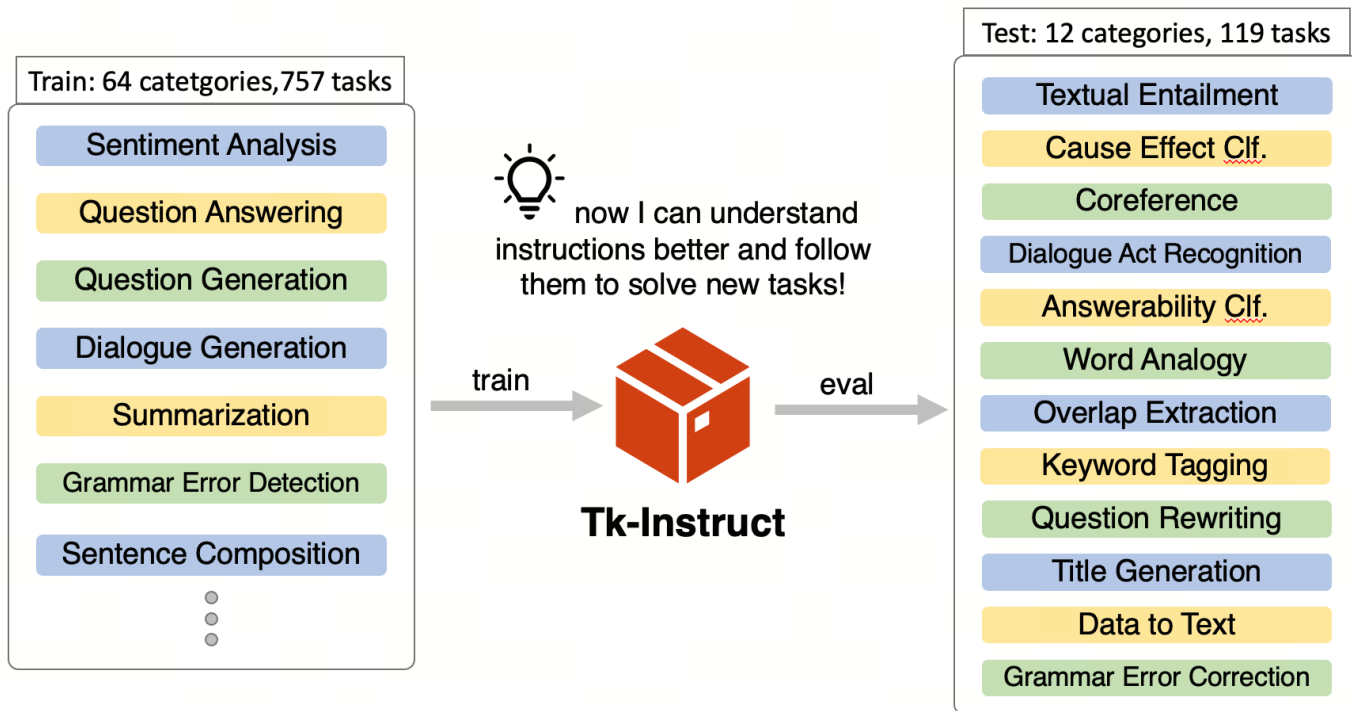
A: Let's think step by step.

After instruction finetuning

The reporter and the chef will discuss their favorite dishes does not indicate whose favorite dishes they will discuss. So, the answer is (C). ✓

<https://huggingface.co/google/flan-t5-xxl>

The Magic Cross-Task Generalization



The Impressive Cross-Task Generalization Performance

Test: 12 categories, 119 tasks

Textual Entailment

Cause Effect Clf.

Coreference

Dialogue Act Recognition

Answerability Clf.

Word Analogy

Overlap Extraction

Keyword Tagging

Question Rewriting

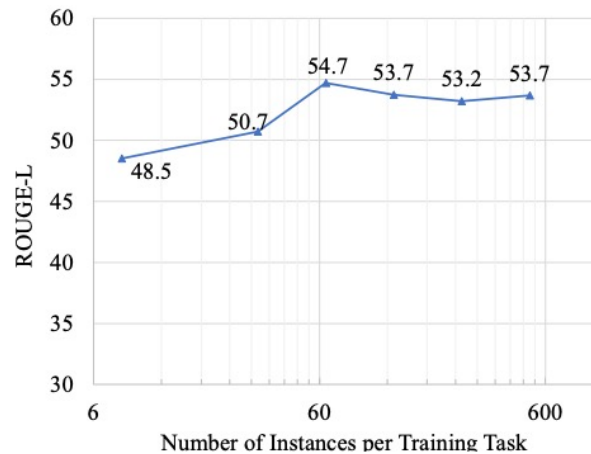
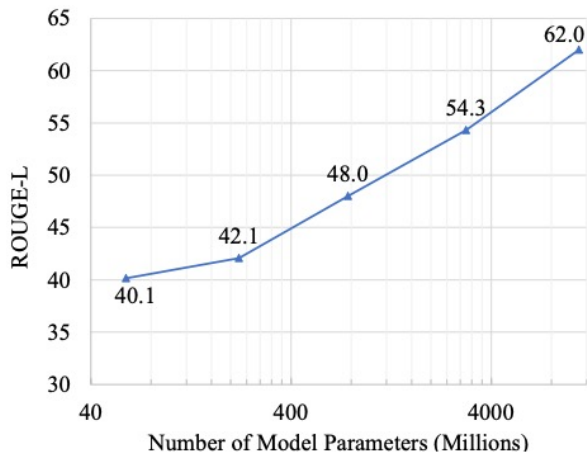
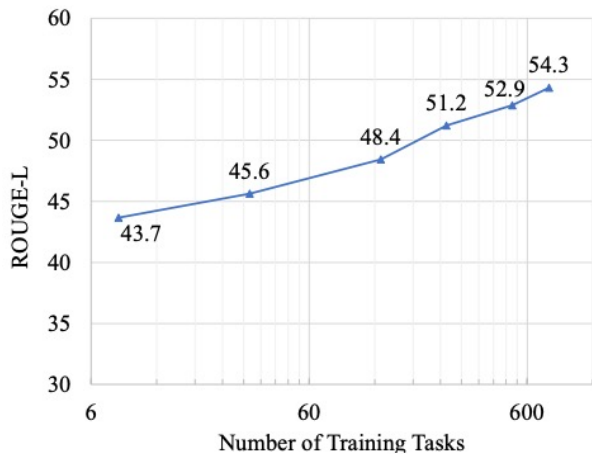
Title Generation

Data to Text

Grammar Error Correction

	Methods ↓ / Evaluation →	En
Heuristic Baselines	Copying Instance Input	14.2
	Copying Demo Output	28.5
Pretrained LMs	T5-LM (11B)	30.2
	GPT3 (175B)	45.0
Instruction-tuned Models	T0 (11B)	32.3
	InstructGPT (175B)	52.1
	Tk-INSTRUCT (ours, 11B)	62.0
	mTk-INSTRUCT (ours, 13B)	57.1
Upper-bound (est.)	Supervised Training	74.3

Scaling Instruction-Tuning

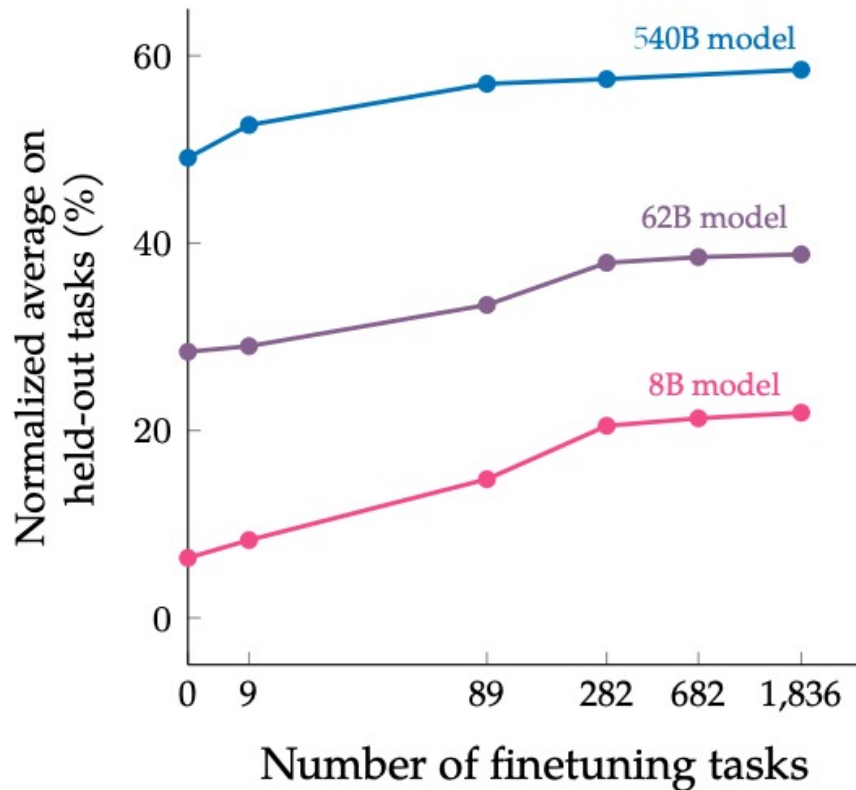


Linear growth of model performance with exponential increase in **observed tasks** and **model size**.

Number of examples has little effect.

Scaling Instruction-Tuning

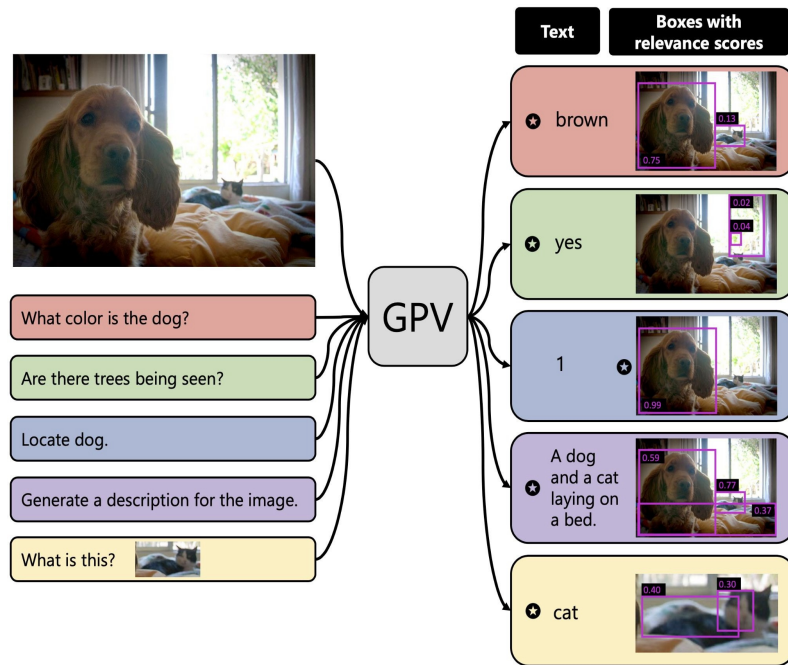
- **Instruction finetuning** improves performance by a large margin compared to **no finetuning**
- Increasing the number of finetuning tasks improves performance
- Increasing **model scale** by an order of magnitude (i.e., 8B \rightarrow 62B or 62B \rightarrow 540B) **improves performance** substantially for both finetuned and non-finetuned models



Multi-Modal Instruction-Tuning

Note these ideas can easily be repackaged for tasks that involve other modalities.

- Robots with instructions e.g. [Zhao et al FACL 2021](#)
- Vision tasks as VQA e.g. [Gupta et al CVPR 2022](#)



Summary Thus Far

- Training (tuning) LMs with annotated input instructions and their output.
- **Pros:**
 - Simple to implement
 - Shows generalization to unseen tasks.
- **Cons:**
 - It's expensive to collect ground- truth data for tasks.
 - Tasks like open-ended creative generation have no right answer. For example: "Write me a story about a dog and her pet grasshopper." Based on fine-tuning objectives, any deviations (even single-token) would incur a loss.

Reinforcement Learning from Human Feedback

GPT3.5 (InstructGPT)

30k prompts
corresponding
to diverse
tasks!

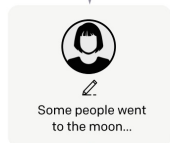
Step 1

**Collect demonstration data,
and train a supervised policy.**

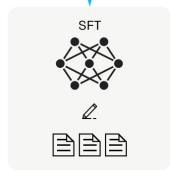
A prompt is
sampled from our
prompt dataset.



A labeler
demonstrates the
desired output
behavior.



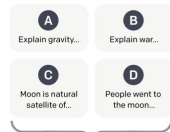
This data is used
to fine-tune GPT-3
with supervised
learning.



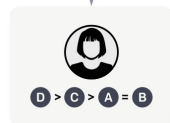
Step 2

**Collect comparison data,
and train a reward model.**

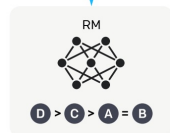
A prompt and
several model
outputs are
sampled.



A labeler
ranks the
outputs from
best to worst.



This data is used
to train our
reward model.



Step 3

**Optimize a policy against
the reward model using
reinforcement learning.**

A new prompt
is sampled from
the dataset.

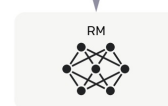


The policy
generates
an output.



Once upon a time...

The reward model
calculates a
reward for
the output.

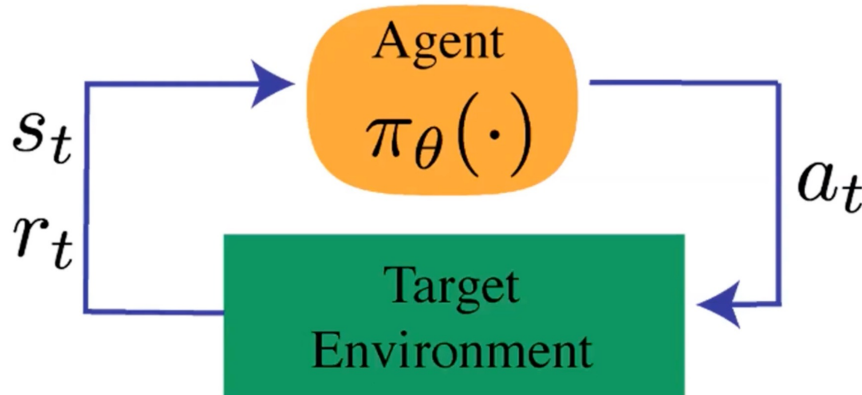


The reward is
used to update
the policy
using PPO.



Reinforcement Learning: The Basics

- An agent **interacts** with an environment by taking **actions**
- The environment returns a **reward** for the **action** and a **new state** (representation of the world at that moment).
- Agent uses a **policy** function to choose an action at a given **state**.
- Quite an open-ended learning paradigm.



Some notation:

s_t : state

r_t : reward

a_t : action

$a_t \sim \pi_{\theta}(s_t)$: policy

Reinforcement Learning: An Example

Action here: generating each token

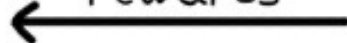
agent



actions



rewards



observations



environment



Reward here: whether humans liked the generation (sequence of actions=tokens)

Human can Express Preference as a Reward for Model Training

Playground task ⓘ

Talk to the assistant

Task 1 of 5

Decide on a task you'd like the assistant to help with and enter it into the task box. Interact with the AI assistant. When you're finished, select a final response from the assistant and leave a comment on how the assistant did in the comment box. Click to go to the next task.

Task

Get help writing this paper.

Comments

OK, but not as good as GPT-3.

Rating

Bad Good

Next Task

- Human**
I'm writing an AI research paper about literally this kind of interaction with an AI assistant. Where in the paper should I put a figure showing this interface?
- Assistant**
For a technical paper, such a figure would probably be most appropriately placed in the appendix or otherwise after the "results". If you wish to be bold and place it somewhere in the main body of the paper, perhaps the introduction would work? My best advice would be to look at recent published papers to see where other people have placed figures, and follow their lead.
- Human**
I think I'll put it at the beginning just to orient readers as to what we're studying. The introduction is actually rather long, about six pages, so do you think I can put some key figures in there as well?
- Assistant**
Sure. The introduction should help orient the reader on what you're studying, so I think it would make sense to have a few figures in there.
- Human**
Great. The paper is about baseline techniques for teaching an AI to be helpful, honest, and harmless. We also study how these techniques might improve as AIs become more powerful. Can you try writing a draft abstract for this paper for me?

Choose the most helpful and honest response

A How is this like a draft abstract?

Sure, here's what I came up with:

B In this paper we explore how various techniques can enable an intelligent personal assistant to develop the capacity to understand and help in various contexts. We start by reviewing a range of currently available techniques. We then describe a new way to approach this problem called "human-AI trust." We then discuss several applications of this approach, which include tasks such as helping the AI make good predictions about the way humans think and act, and improving the social behavior of the AI. Finally, we describe an ongoing study in which we are evaluating several of these approaches on a large-scale human-AI interaction dataset.

A A A A B B B B

A is better B is better

Reward Modeling to Make Human Preference Scalable

- Obviously, we don't want to **use human feedback directly** since that could be 💰💰💰
- Alternatively, we can build a model to **mimic their preferences** [[Knox and Stone, 2009](#)]

Reward Model ~ Human Preference

- Imagine a reward function: $R(s; p) \in \mathbb{R}$ for any output s to prompt p
- **The reward is higher when humans prefer the output**

SAN FRANCISCO,
California (CNN) --
A magnitude 4.2
earthquake shook the
San Francisco
...
overturn unstable
objects.

An earthquake hit
San Francisco.
There was minor
property damage,
but no injuries.

s_1

$$R(s_1; p) = 0.8$$

The Bay Area has
good weather but is
prone to
earthquakes and
wildfires.

s_2

$$R(s_2; p) = 1.2$$

How can We Build the Reward Model $R(s; p)$?

An earthquake hit San Francisco. There was minor property damage, but no injuries. 🧐

S_1

A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage. 🧐

S_2

The Bay Area has good weather but is prone to earthquakes and wildfires.

S_3

$$J(\phi) = -\mathbb{E}_{(s^+, s^-)} [\log \sigma(R(s^+; p) - R(s^-; p))]$$

“winning”
sample

“losing”
sample

Bradley-Terry [1952]
paired comparison model

Pairwise comparison of multiple provides which can be more reliable

RL for Training the Policy (Language Model)

- How do we change our LM parameters θ to maximize this?

$$\hat{\theta} = \operatorname{argmax}_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}} [R(\hat{s}; p)]$$

- Policy Gradient Decent:

$$\theta_{t+1} \leftarrow \theta_t + \alpha \frac{1}{n} \sum_{i=1}^n R(s; p) \nabla_{\theta} \log p_{\theta}(s)$$

If $R(s; p)$ is **large**, we take proportionately **large** steps to maximize $p_{\theta}(s)$

If $R(s; p)$ is **small**, we take proportionately **small** steps to maximize $p_{\theta}(s)$

This is why it's called "reinforcement learning":
we reinforce good actions, increasing the chance they happen again.

RL for Training the Policy (Language Model)

- How do we change our LM parameters θ to maximize this?

$$\hat{\theta} = \operatorname{argmax}_{\theta} \mathbb{E}_{\hat{s} \sim p_{\theta}} [R(\hat{s}; p)]$$

- Policy Gradient Decent:

$$\theta_{t+1} \leftarrow \theta_t + \alpha \frac{1}{n} \sum_{i=1}^n R(s; p) \nabla_{\theta} \log p_{\theta}(s)$$

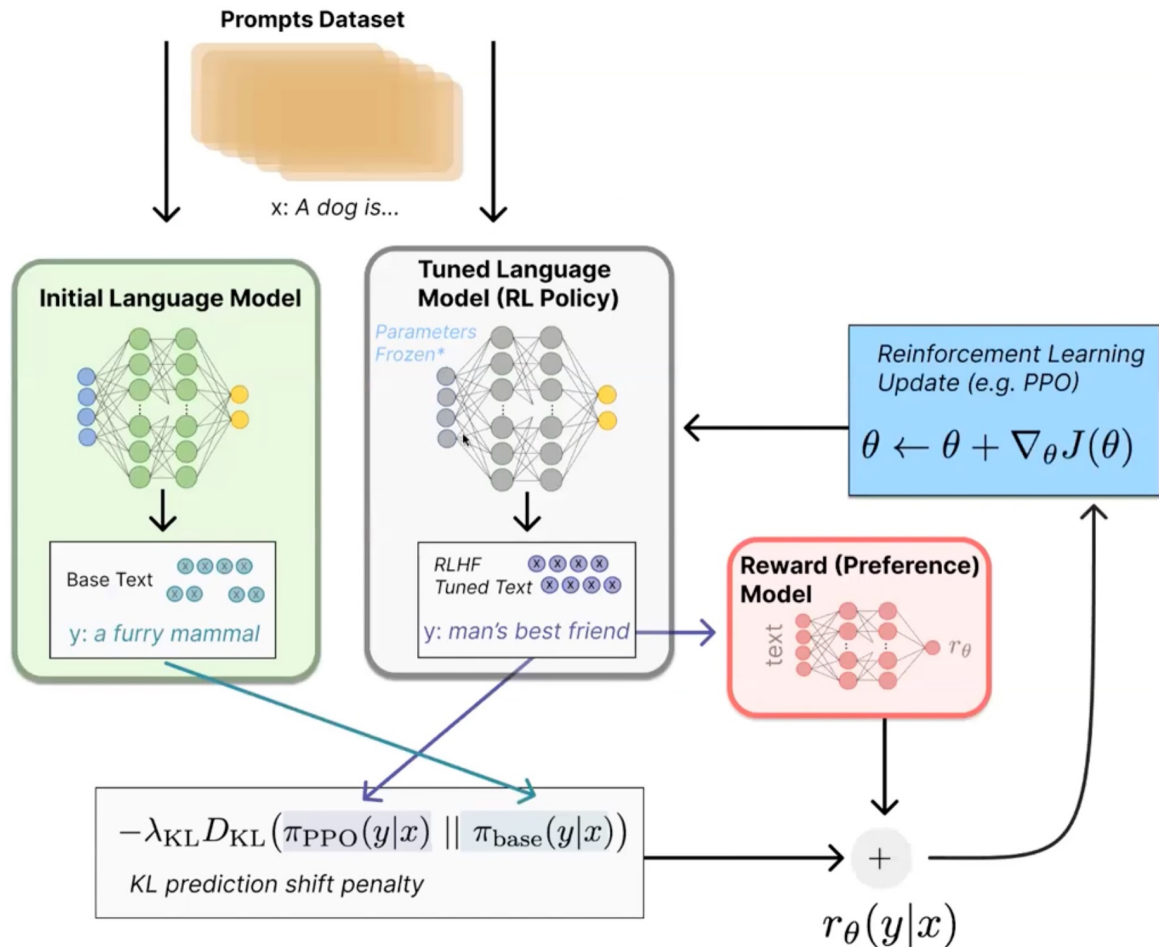
- Regularizing the training:

$$\hat{R}(s; p) := R(s; p) - \beta \log \left(\frac{p^{RL}(s)}{p^{PT}(s)} \right)$$

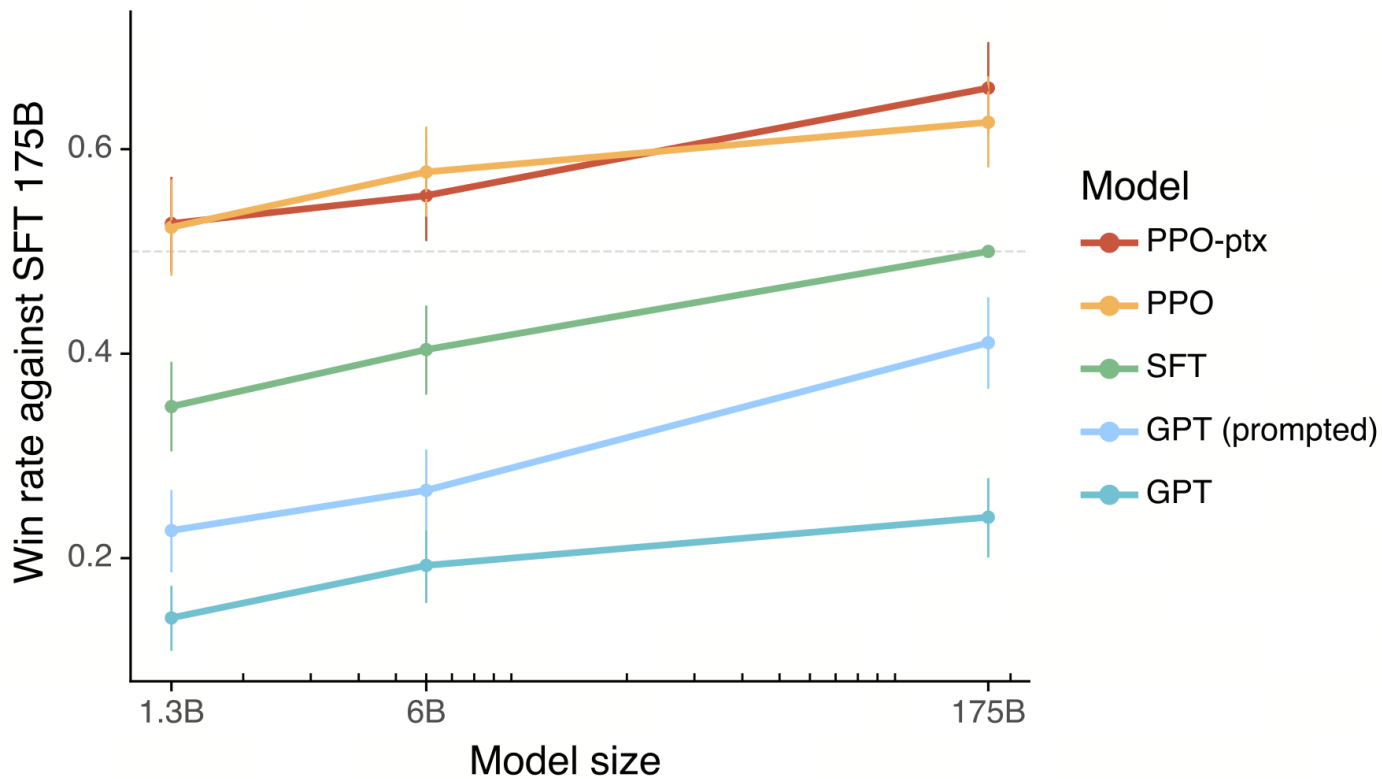
pay a price when
 $p^{RL}(s) > p^{PT}(s)$

RLHF: Putting it All Together

[[Christiano et al. 2017](#); [Stiennon et al. 2020](#)]



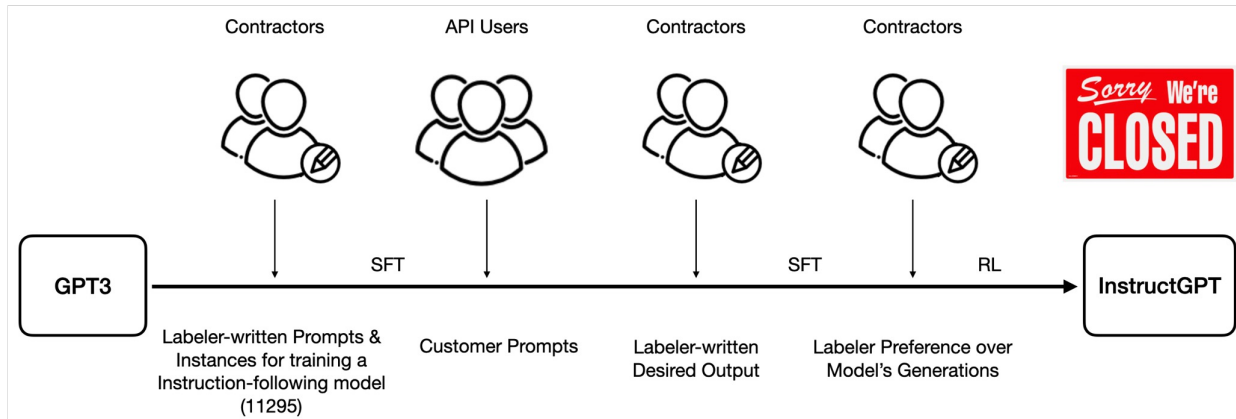
Pretraining + RLHF Gains over Pretraining + Finetuning



Building Open (Your Own) Instruction-Following Models

RLHF/Instruction-tuning is Data Hungry

SoTA models require a lot of user data and expert annotations to train.



- **Rumor:**

- Human feedback done for supervising ChatGPT is in the order of \$1M

How Can We Scale Up Data More Broadly?

- Writing diverse instructions requires creativity.
- Writing instances for different instructions requires broad expertise.
- Nearly impractical for crowd workers.

Self-Instruct: LM 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 [refresh] [undo] [redo] [share] [like]

163

Mode [list] [download] [share]

Model davinci

Temperature 0.7

Maximum length 1024

Stop sequences Enter sequence and press Tab
16. x

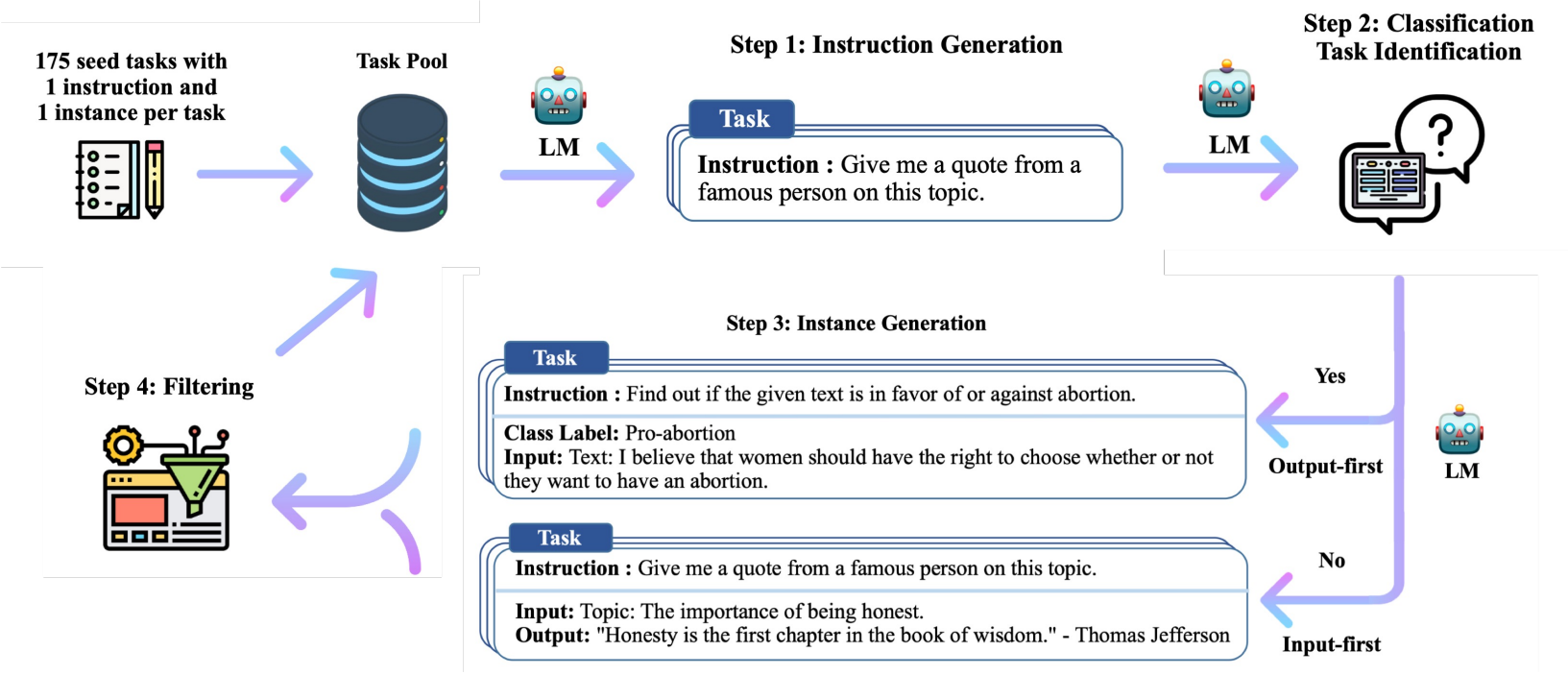
Top P 0.5

Frequency penalty 0

Presence penalty 2

Best of 1

Self-Instruct: LM can be prompted to generate instructions



[Self-Instruct: Aligning Language Models with Self-Generated Instructions, Wang et al. 2022]

Generated Tasks with an Early Version of GPT₃



Instruction: Given an address and city, come up with the zip code.

Input:

Address: 123 Main Street, City: San Francisco

Output: 94105

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

...

Instruction: How to write a code for converting degrees fahrenheit to celsius.

Input: Null

Output:

```
def convert_fahrenheit_to_celsius(fahr):  
    celsius = (fahr - 32) * 5 / 9  
    return celsius
```



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 ...

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

Input: Null

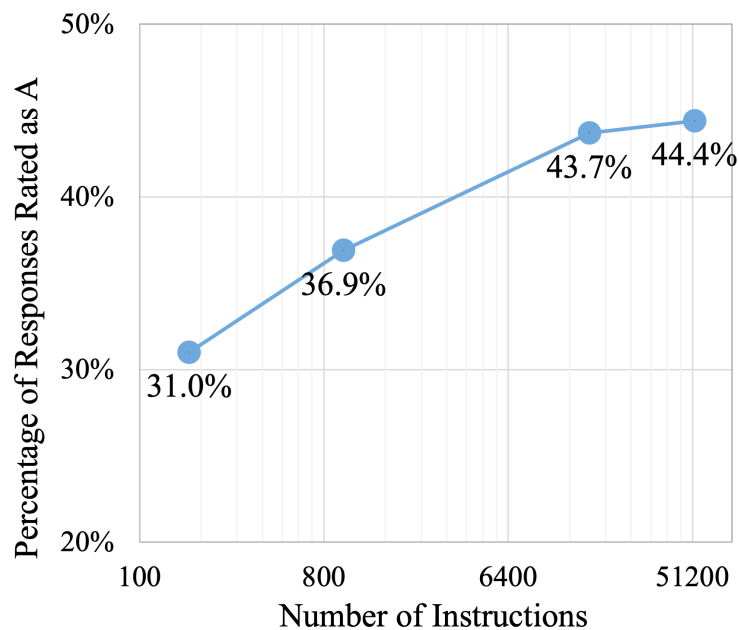
Output:

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 ...



Self-Instruct Improves Model Performance

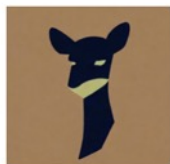
- Human judgement on 252 creative tasks.



A lot of open-source instruction datasets since then...



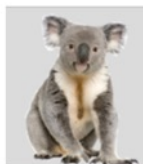
Alpaca



Vicuna



Baize



Koala



GPT4-Instruct

...



Dromedary



Dolly



Open-Assistant



WizardML



ORCA

Resources for Building Your Own ChatGPT

- Open-Instruct: <https://github.com/allenai/open-instruct/> (Wang et al., 2023)

	MMLU (factuality)	GSM (reasoning)	BBH (reasoning)	TydiQA (multilinguality)	Codex-Eval (coding)	AlpacaEval (open-ended)	Average
	EM (0-shot)	EM (8-shot, CoT)	EM (3-shot, CoT)	F1 (1-shot, GP)	P@10 (0-shot)	Win % vs Davinci-003	
Vanilla LLaMa 13B	42.3	14.5	39.3	43.2	28.6	-	-
+SuperNI	49.7	4.0	4.5	50.2	12.9	4.2	20.9
+CoT	44.2	40.0	41.9	47.8	23.7	6.0	33.9
+Flan V2	50.6	20.0	40.8	47.2	16.8	3.2	29.8
+Dolly	45.6	18.0	28.4	46.5	31.0	13.7	30.5
+Open Assistant 1	43.3	15.0	39.6	33.4	31.9	58.1	36.9
+Self-instruct	30.4	11.0	30.7	41.3	12.5	5.0	21.8
+Unnatural Instructions	46.4	8.0	33.7	40.9	23.9	8.4	26.9
+Alpaca	45.0	9.5	36.6	31.1	29.9	21.9	29.0
+Code-Alpaca	42.5	13.5	35.6	38.9	34.2	15.8	30.1
+GPT4-Alpaca	46.9	16.5	38.8	23.5	36.6	63.1	37.6
+Baize	43.7	10.0	38.7	33.6	28.7	21.9	29.4
+ShareGPT	49.3	27.0	40.4	30.5	34.1	70.5	42.0
+Human data mix.	50.2	38.5	39.6	47.0	25.0	35.0	39.2
+Human+GPT data mix.	49.3	40.5	43.3	45.6	35.9	56.5	45.2



[Created with Midjourney, prompted by Yizhong]

- OpenRLHF: <https://github.com/OpenLLMAI/OpenRLHF>
- TRL: <https://github.com/huggingface/trl>

Open Research Questions

- What is the relation between data diversity and data quality?
- How far can model generalize? What is the boundary?
- Is RL necessary? Can we find better supervised algorithms? ...
- Is HF more important or RL?
- What is the best form of HF?
- If we have more and more human interaction data, can finetuning outweigh pretraining?
- ...

Thanks!
Questions?

 @yizhongwyz

 yizhongw@cs.washington.edu