Recipes for Training Helpful Chatbots

Nazneen Rajani | Research Lead @ Hugging Face | emailnazneen@gmail.com | @nazneenrajani

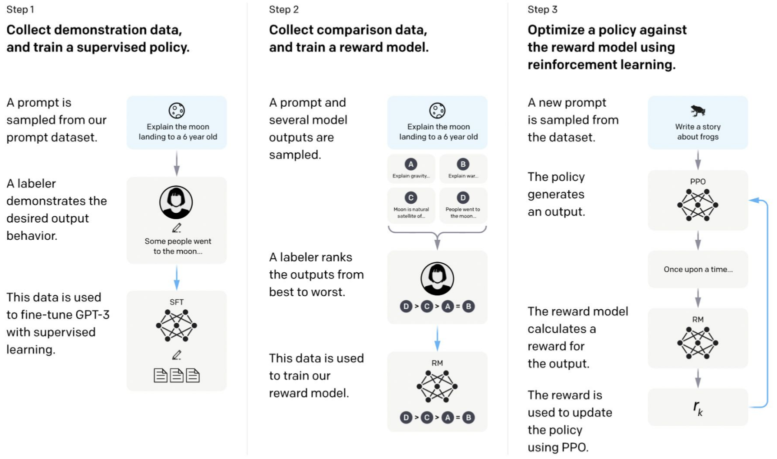
Introduction

H4 team at Hugging Face 🤗

Goal: Recipe for Helpful, Harmless, Honest, and Huggy (H4) chatbot

Ingredients: Datasets for SFT and RLHF, pretrained open access models

Procedure:



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.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

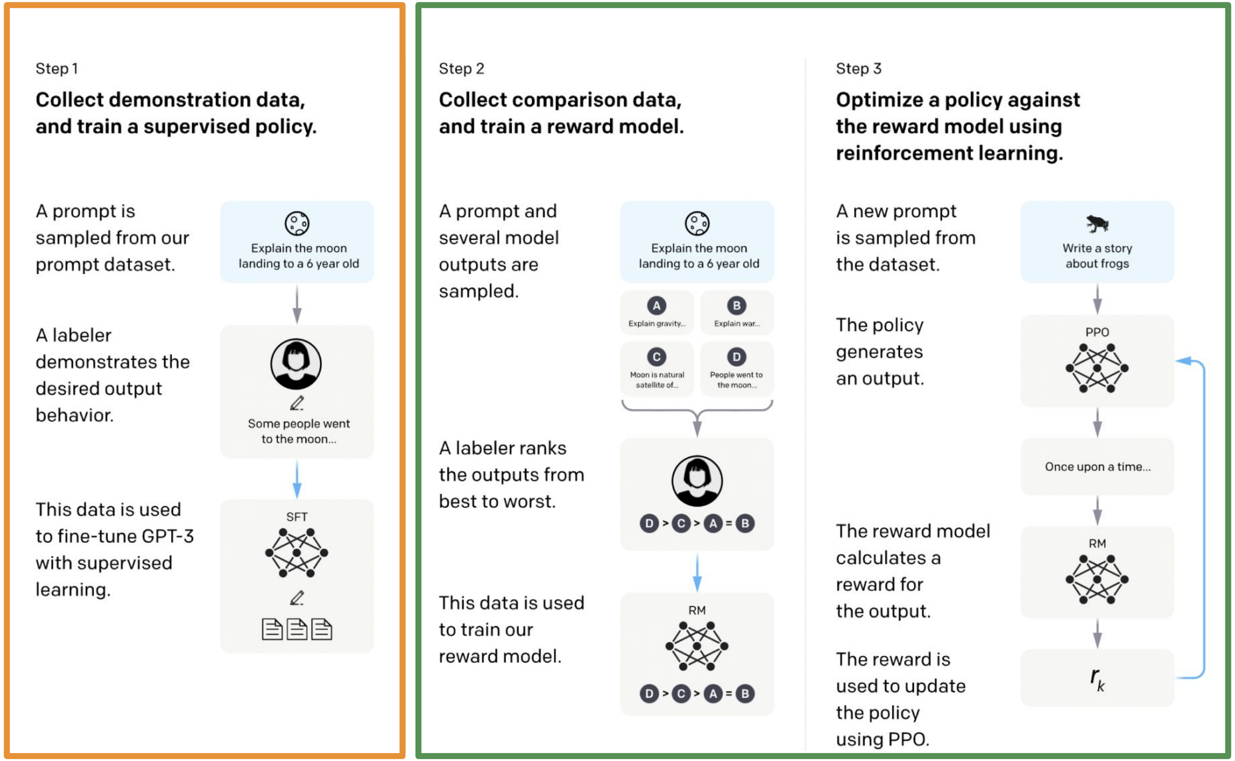
Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *arXiv preprint arXiv:2203.02155* (2022).

Outline

- Data collection for SFT
- Data collection for RLHF
- Distillation of LM Alignment
- Experiments with helpfulness recipes
- Evaluation of SFT Models
- Quirks of using GPT4 as evaluator

Training a Chatbot



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.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

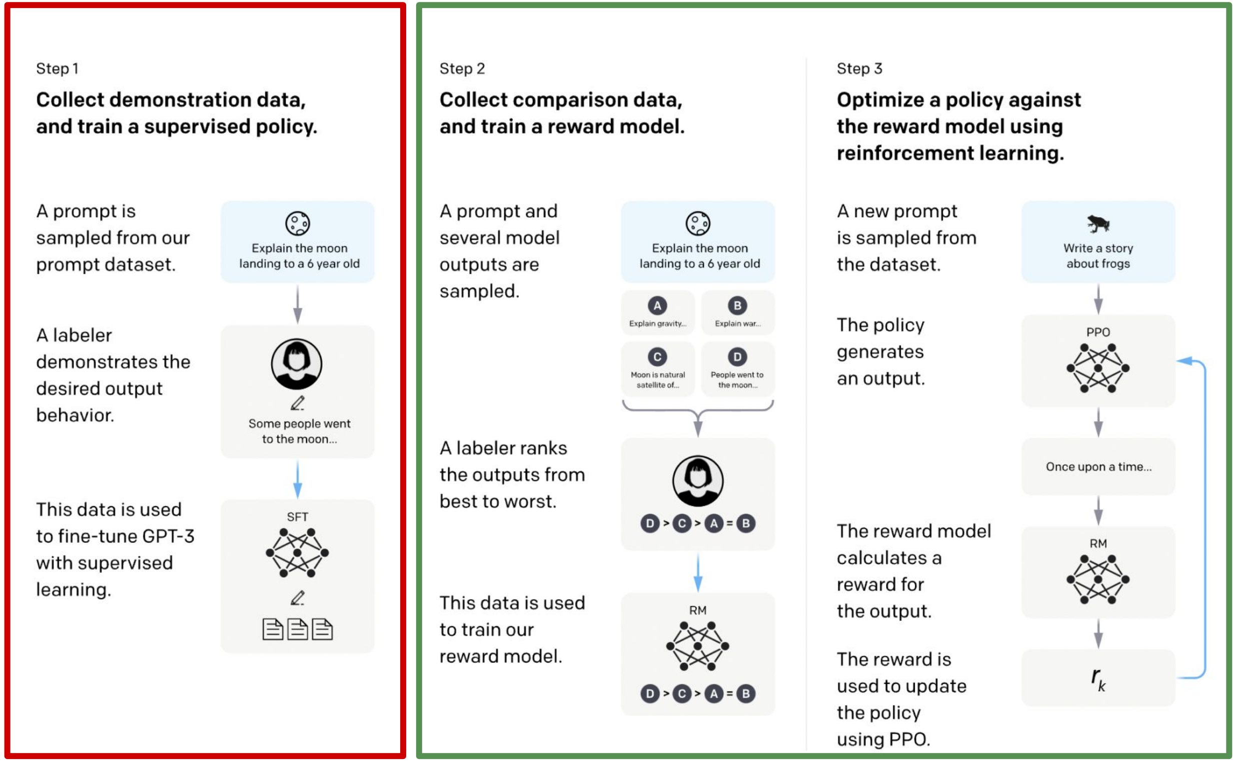
Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

Helpfulness

Harmlessness

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *arXiv preprint arXiv:2203.02155* (2022).

Training a Chatbot



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.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT

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.

Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B

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.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

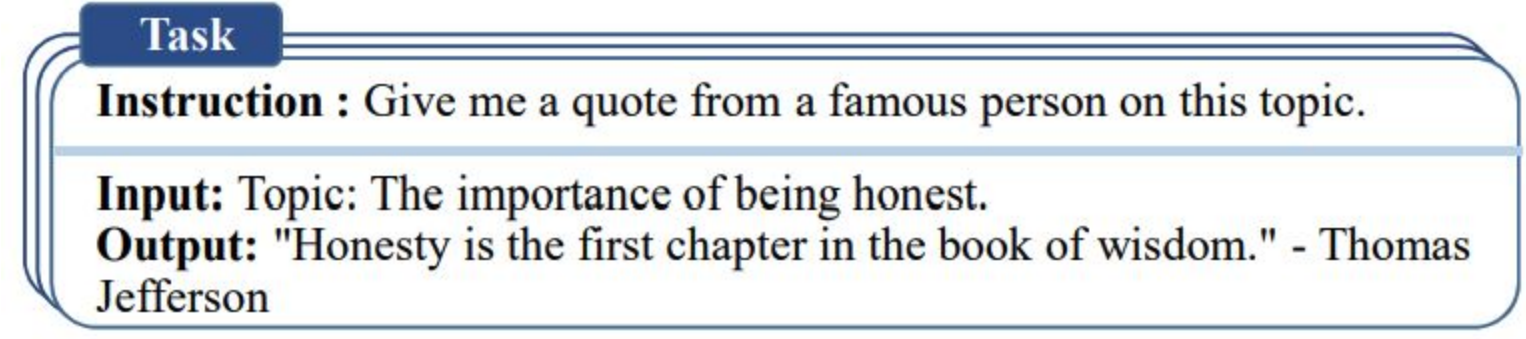
Helpfulness

Harmlessness

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *arXiv preprint arXiv:2203.02155* (2022).

Dataset

Dataset for Helpfulness



Task

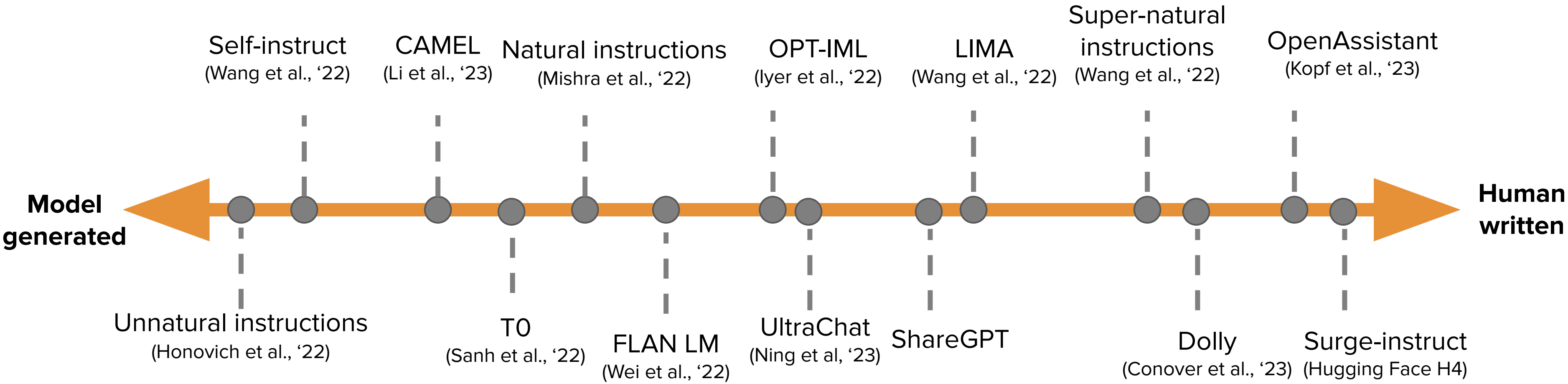
Instruction : Give me a quote from a famous person on this topic.

Input: Topic: The importance of being honest.
Output: "Honesty is the first chapter in the book of wisdom." - Thomas Jefferson

} instance/completion/demonstration

[Wang et al., ‘22](#)

Dataset for Helpfulness



Model generated

Unnatural instructions
(Honovich et al., ‘22)

Self-instruct
(Wang et al., ‘22)

CAMEL
(Li et al., ‘23)

T0
(Sanh et al., ‘22)

Natural instructions
(Mishra et al., ‘22)

FLAN LM
(Wei et al., ‘22)

OPT-IML
(Iyer et al., ‘22)

UltraChat
(Ning et al, ‘23)

ShareGPT

LIMA
(Wang et al., ‘22)

Super-natural instructions
(Wang et al., ‘22)

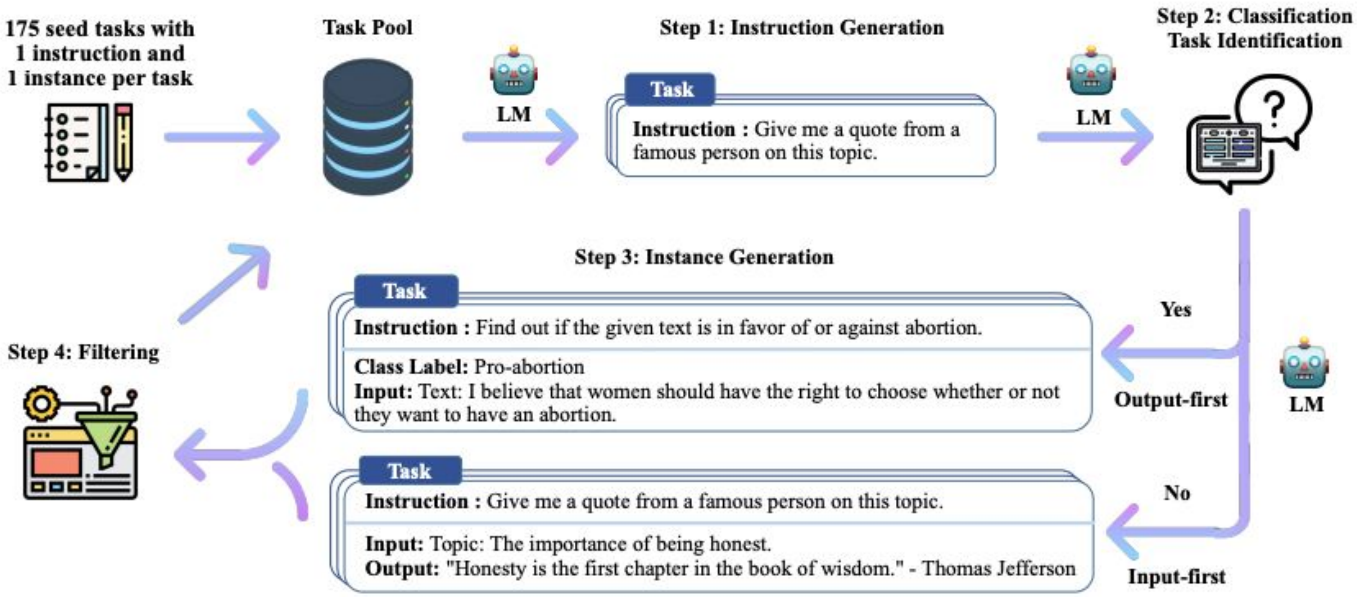
Dolly
(Conover et al., ‘23)

OpenAssistant
(Kopf et al., ‘23)

Surge-instruct
(Hugging Face H4)

Human written

Bootstrapping Data (Self-instruct)



175 seed tasks with
1 instruction and
1 instance per task

Task Pool

LM

Step 1: Instruction Generation

Task

Instruction : Give me a quote from a famous person on this topic.

LM

Step 2: Classification
Task Identification

Step 3: Instance Generation

Task

Instruction : Find out if the given text is in favor of or against abortion.

Class Label: Pro-abortion
Input: Text: I believe that women should have the right to choose whether or not they want to have an abortion.

Yes

Output-first

LM

No

Input-first

Task

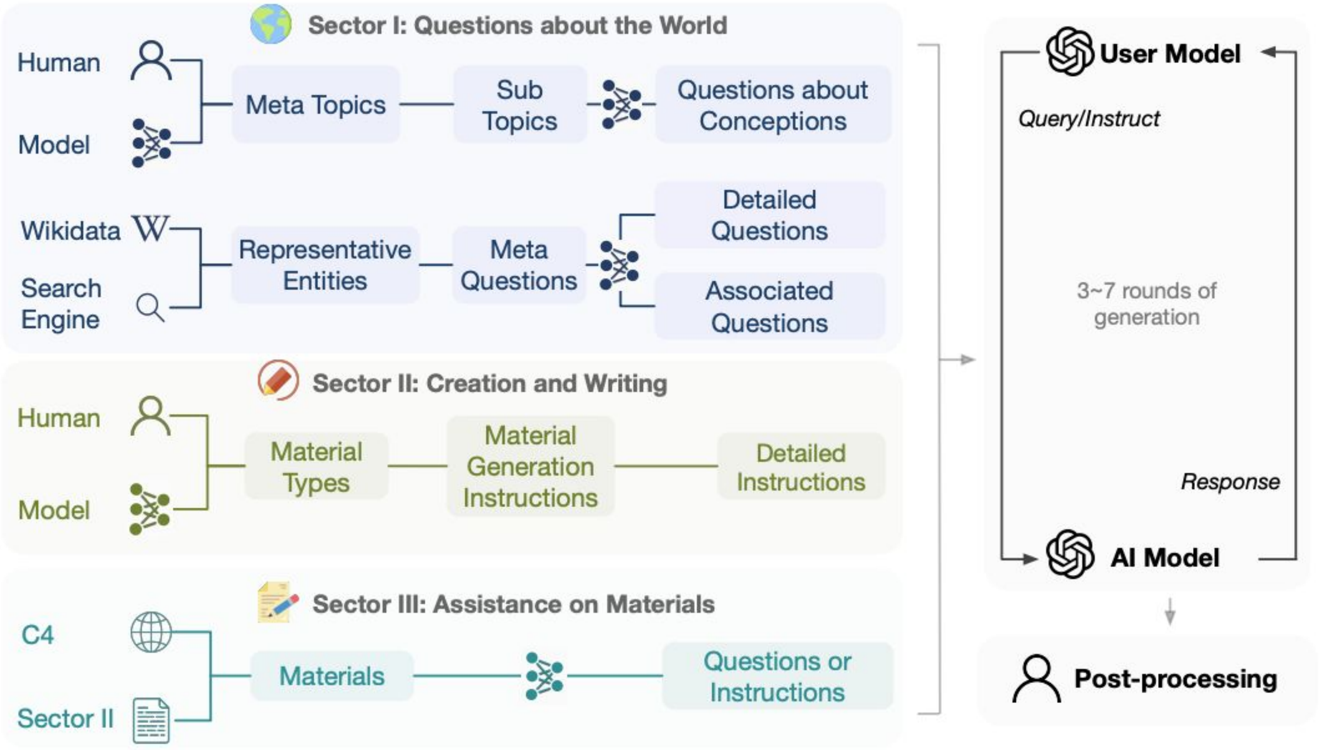
Instruction : Give me a quote from a famous person on this topic.

Input: Topic: The importance of being honest.
Output: "Honesty is the first chapter in the book of wisdom." - Thomas Jefferson

Step 4: Filtering

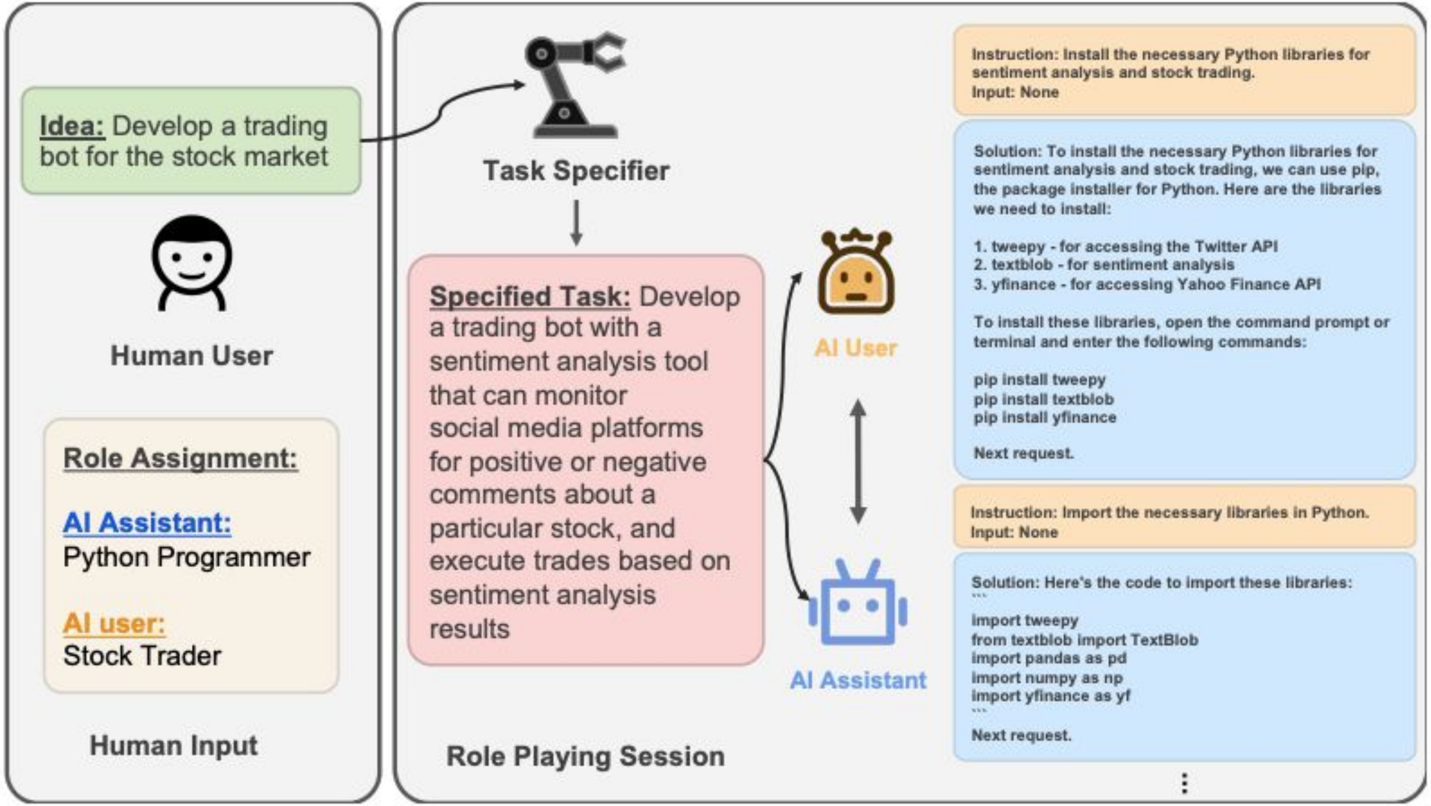
Wang et al., ‘22

Human-in-the-loop (UltraChat)



Ding et al., '23

Roleplaying (CAMEL)



Idea: Develop a trading bot for the stock market

Role Assignment:

AI Assistant: Python Programmer

AI user: Stock Trader

Specified Task: Develop a trading bot with a sentiment analysis tool that can monitor social media platforms for positive or negative comments about a particular stock, and execute trades based on sentiment analysis results

Instruction: Install the necessary Python libraries for sentiment analysis and stock trading.
Input: None

Solution: To install the necessary Python libraries for sentiment analysis and stock trading, we can use pip, the package installer for Python. Here are the libraries we need to install:

1. tweepy - for accessing the Twitter API
2. textblob - for sentiment analysis
3. yfinance - for accessing Yahoo Finance API

To install these libraries, open the command prompt or terminal and enter the following commands:

```
pip install tweepy
pip install textblob
pip install yfinance
```

Next request.

Instruction: Import the necessary libraries in Python.
Input: None

Solution: Here's the code to import these libraries:

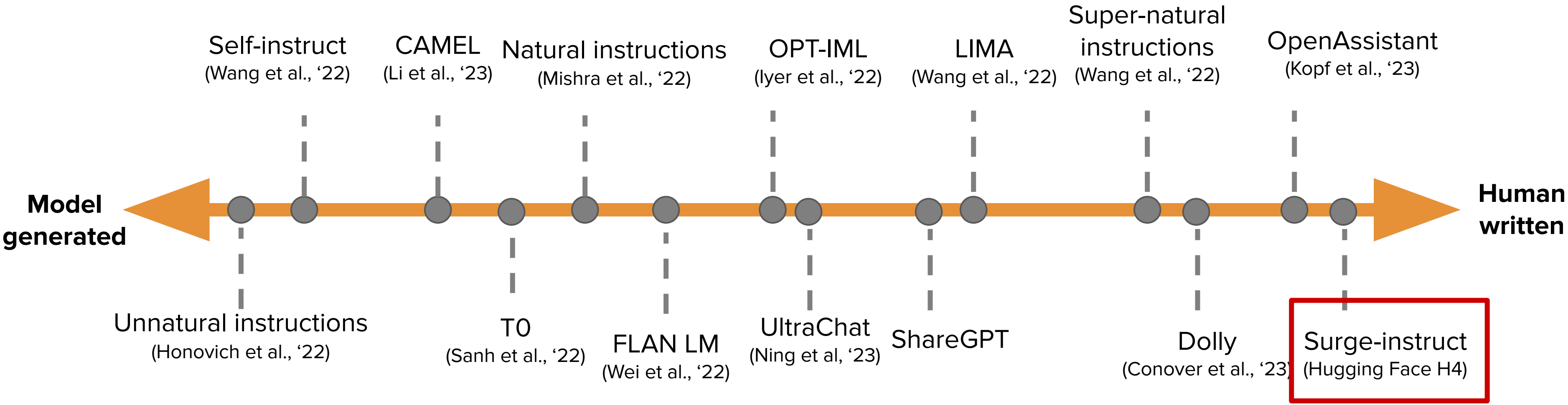
```
import tweepy
from textblob import TextBlob
import pandas as pd
import numpy as np
import yfinance as yf
```

Next request.

⋮

Li et al., ‘23

Dataset for Helpfulness



Model generated

Unnatural instructions
(Honovich et al., ‘22)

Self-instruct
(Wang et al., ‘22)

CAMEL
(Li et al., ‘23)

T0
(Sanh et al., ‘22)

Natural instructions
(Mishra et al., ‘22)

FLAN LM
(Wei et al., ‘22)

OPT-IML
(Iyer et al., ‘22)

UltraChat
(Ning et al, ‘23)

ShareGPT

LIMA
(Wang et al., ‘22)

Super-natural instructions
(Wang et al., ‘22)

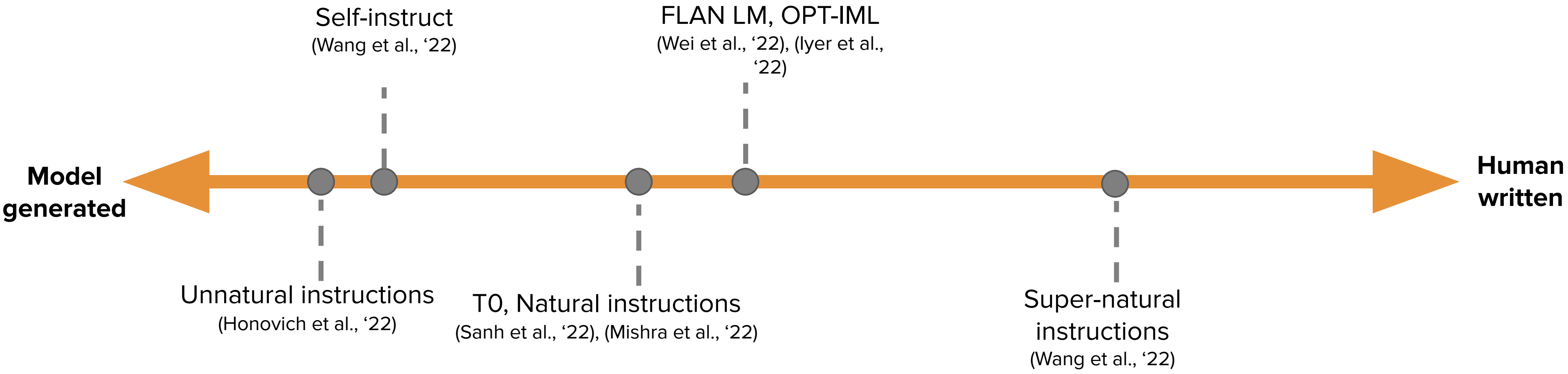
Dolly
(Conover et al., ‘23)

OpenAssistant
(Kopf et al., ‘23)

Surge-instruct
(Hugging Face H4)

Human written

Datasets for SFT for Helpfulness



Model generated

Unnatural instructions
(Honovich et al., ‘22)

Self-instruct
(Wang et al., ‘22)

T0, Natural instructions
(Sanh et al., ‘22), (Mishra et al., ‘22)

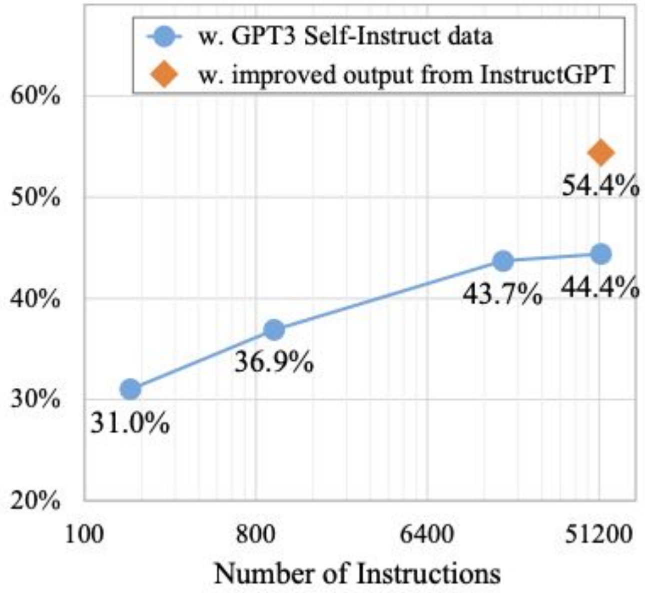
FLAN LM, OPT-IML
(Wei et al., ‘22), (Iyer et al., ‘22)

Super-natural instructions
(Wang et al., ‘22)

Human written

Past Findings from SFT Datasets

- Training data in the range of tens of thousands of examples
- Shows diminishing returns after a few thousand high quality instructions



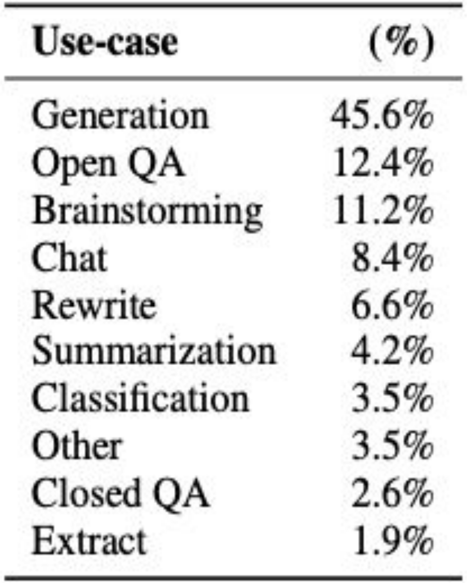
Wang et al., '22

SFT Dataset Desiderata

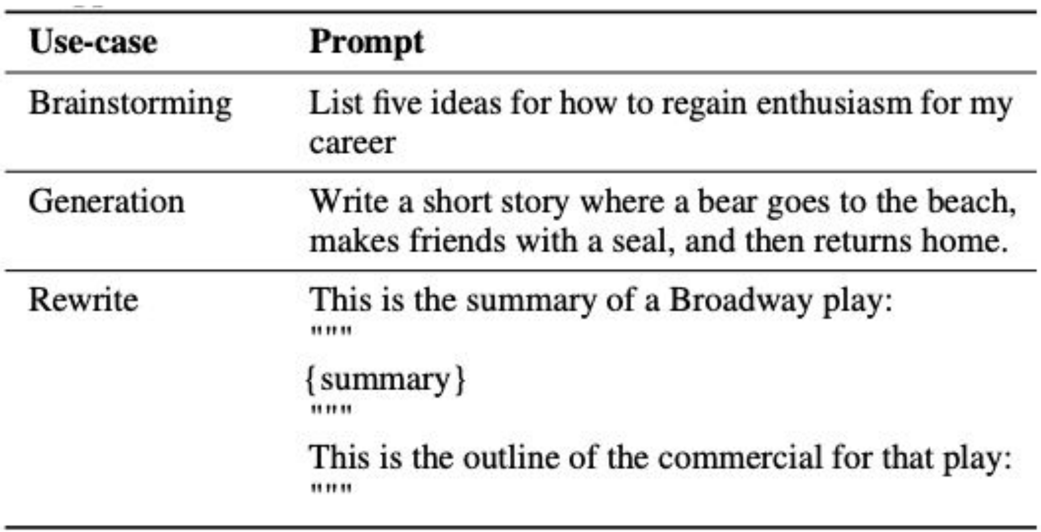
1. Task distribution
2. Length distribution
3. High quality (human-written)
   a. External vendors
   b. Upwork/Mturk

Task Distribution

InstructGPT task distribution



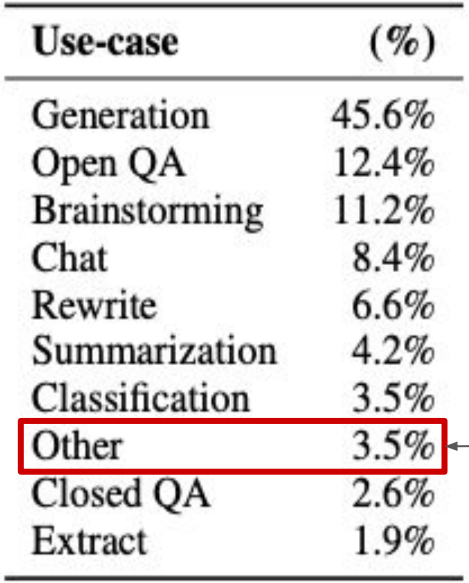
Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%



Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.
Rewrite	This is the summary of a Broadway play: """ {summary} """ This is the outline of the commercial for that play: """

Task Distribution

InstructGPT task distribution



Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

Code

Length Distribution

Pilot study with Surge, Scale AI, and AWS Sagemaker GT

Source	**Split**	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
InstructGPT	train	12725	408	433	1	37	283	632	2048
	valid	1653	401	433	4	41	234	631	2048
Surge	-	75	104	94	7	38	81	135	505
AWS Sagemaker	-	129	54	46	0	8	13	31	1036
Scale AI	-	69	22	16	5	13	18	24	116

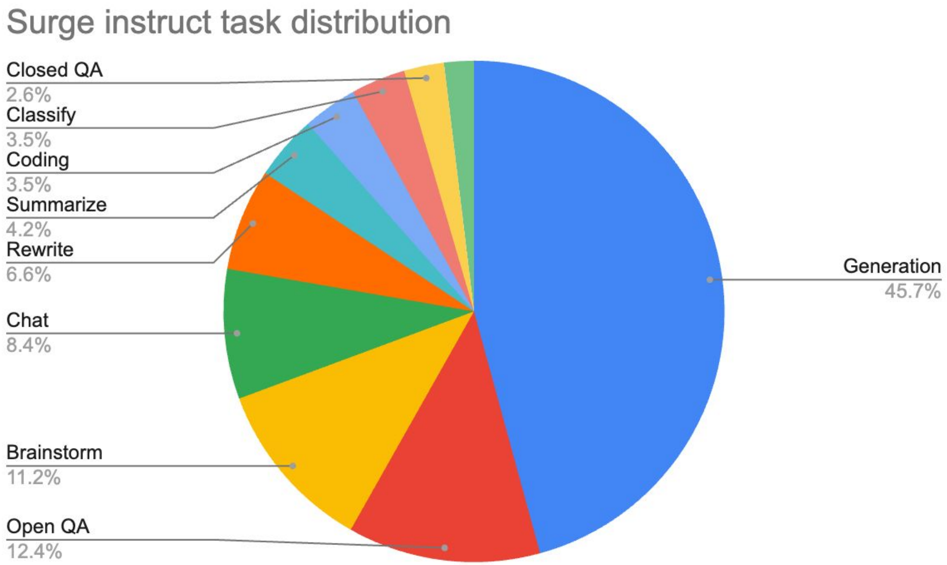
Length Distribution

Pilot study with Surge, Scale AI, and AWS Sagemaker GT

Source	**Split**	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
InstructGPT	train	12725	408	433	1	37	283	632	2048
	valid	1653	401	433	4	41	234	631	2048
Surge	-	75	104	94	7	38	81	135	505
AWS Sagemaker	-	129	54	46	0	8	13	31	1036
Scale AI	-	69	22	16	5	13	18	24	116

Surge Instruct Dataset

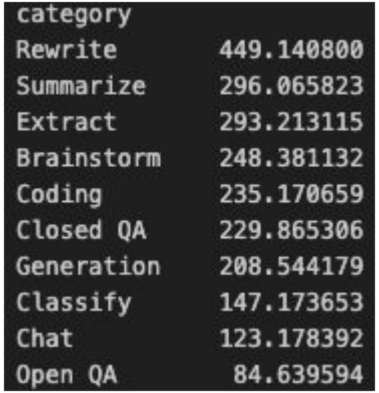
We collected 10K instruction demonstrations pairs



# examples

Generation	4346
Open QA	1182
Brainstorm	1060
Chat	796
Rewrite	625
Summarize	395
Coding	334
Classify	334
Closed QA	245
Extract	183

Avg length



```
category
Rewrite         449.140800
Summarize       296.065823
Extract         293.213115
Brainstorm      248.381132
Coding          235.170659
Closed QA       229.865306
Generation      208.544179
Classify        147.173653
Chat            123.178392
Open QA          84.639594
```

Surge Instruct Dataset

Category	Prompt	Completion
Generation	Write a knock-knock joke for children.	Knock, knock. Who's there? Lettuce. Lettuce, who? Aren't you going to let us in?
Classify	Does this comment seem excited or bored? There's nothing to do today. It's raining outside and none of my toys seem interesting.	Bored
Brainstorm	What are some board games to play. I don't need explanations just the names	1.) Monopoly 2.) Life 3.) Sorry! 4.) Mouse Trap
Open QA	What was Phish’s last studio album?	Phish’s most recent album was “Sigma Oasis”, which was released on April 2nd of 2020.

Surge Instruct Dataset

US-based taskforce

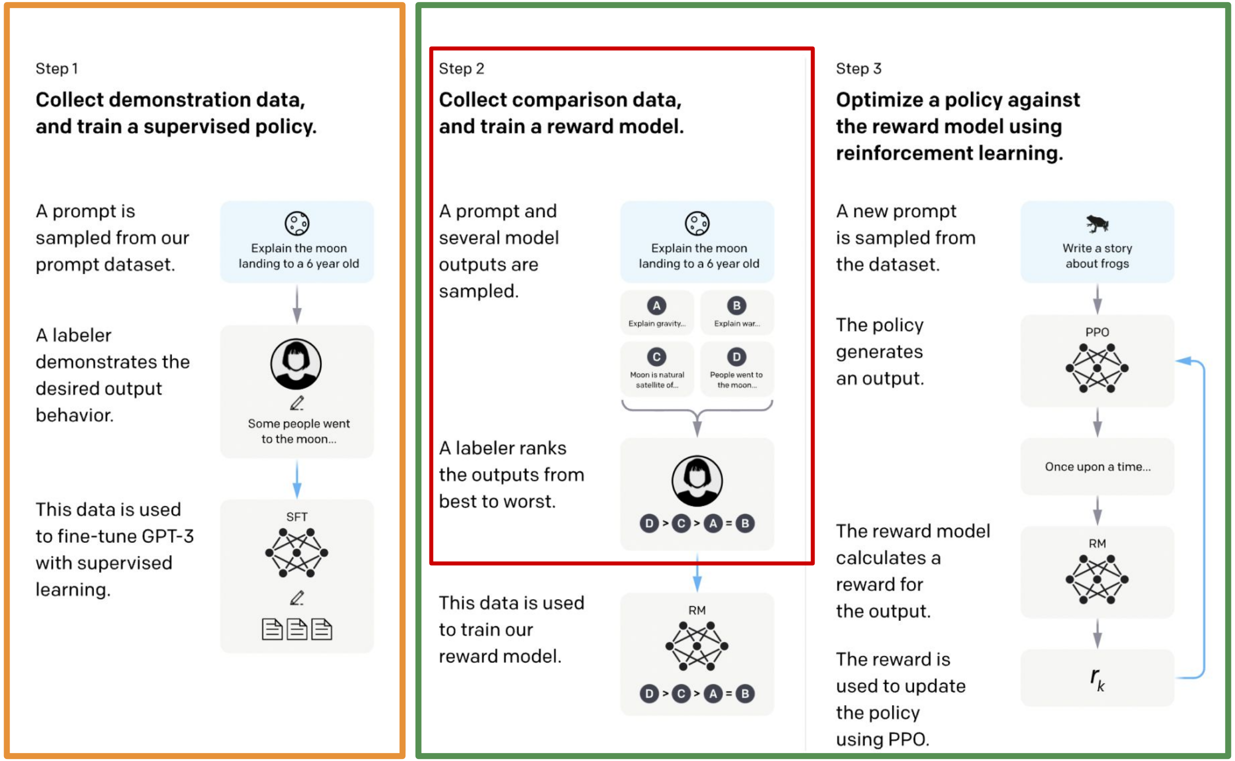
Gender: roughly even split between male and female with some who identify as neither

Age: ranges from 19 to 62

Race: primarily White, Black, Asian, Hispanic

Educational background: ranges from technical degree to PhD

Human Preference Data



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.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

Helpfulness

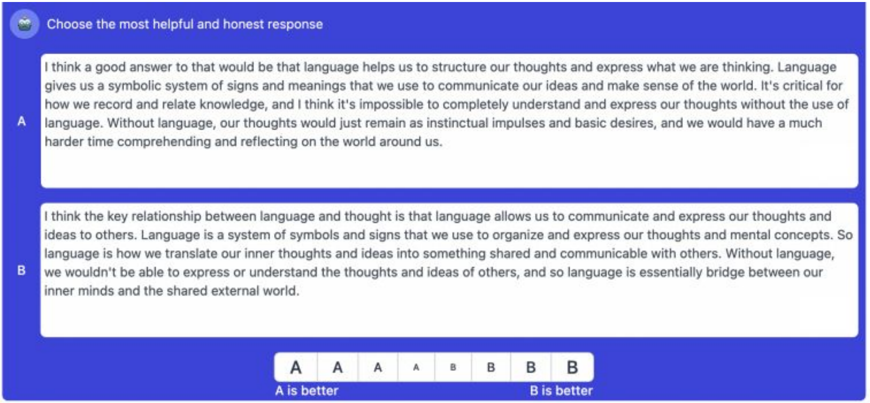
Harmlessness

Human Preference Dataset Desiderata

1. Task distribution
2. Length distribution
3. Singleturn vs Multiturn
4. Helpfulness vs. Honesty vs. Harmlessness
5. Rating/ranking scale

Pilot Study

- Get annotations for 300 self-instruct prompts from data vendors using Anthropic template
- Follow the InstructGPT distribution

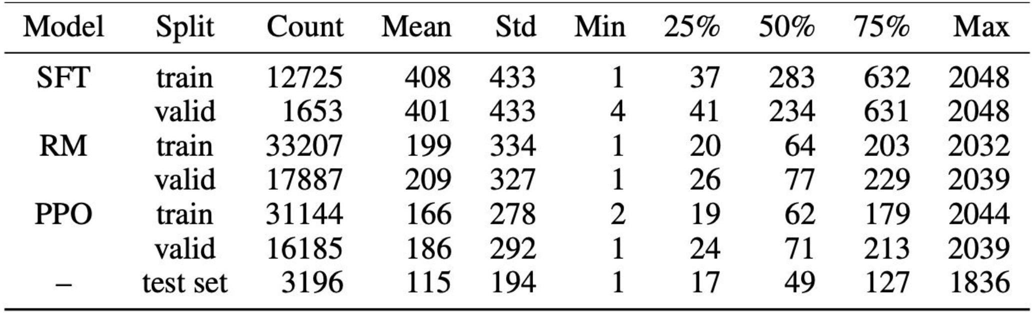


Choose the most helpful and honest response

A: I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

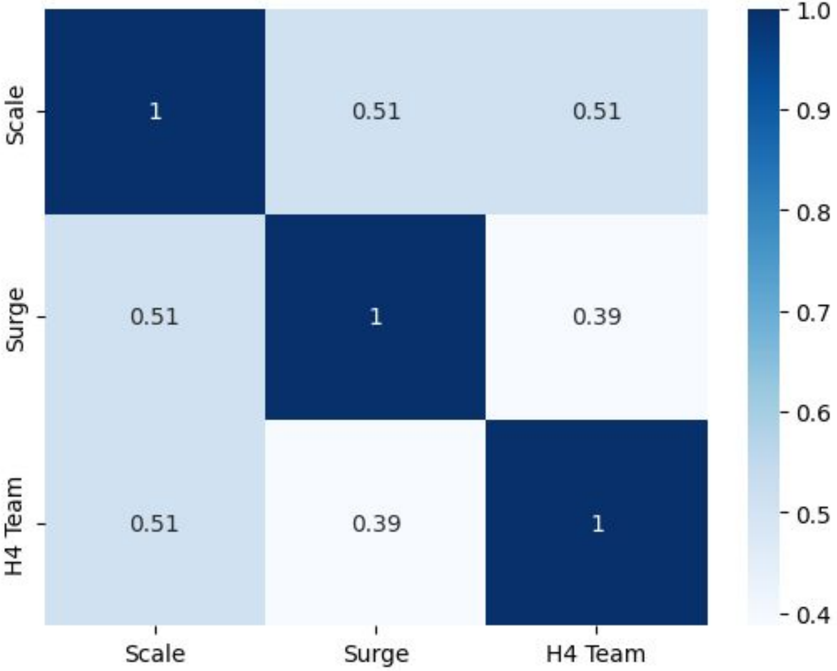
B: I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

A
A
A
A
B
B
B
B
A is better
B is better

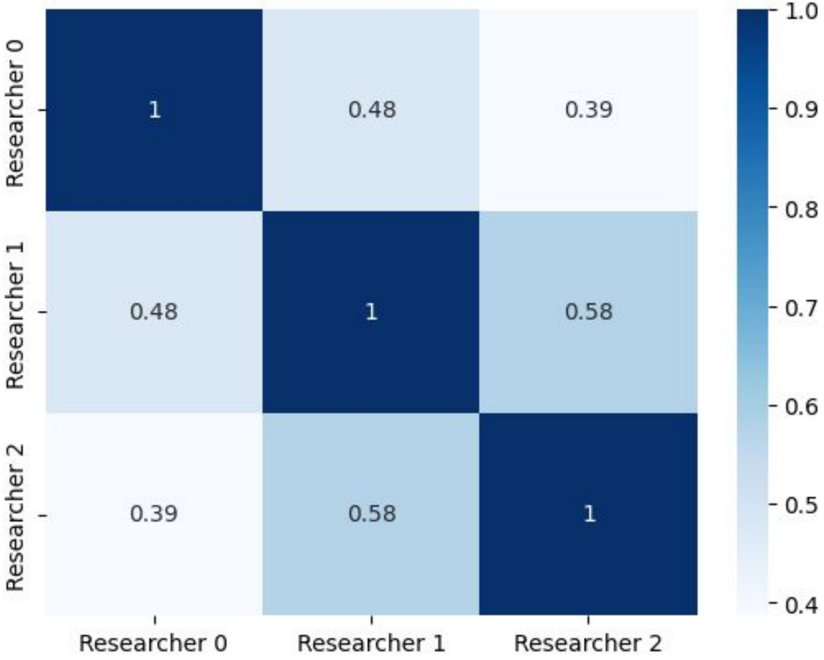


Model	Split	Count	Mean	Std	Min	25%	50%	75%	Max
SFT	train	12725	408	433	1	37	283	632	2048
	valid	1653	401	433	4	41	234	631	2048
RM	train	33207	199	334	1	20	64	203	2032
	valid	17887	209	327	1	26	77	229	2039
PPO	train	31144	166	278	2	19	62	179	2044
	valid	16185	186	292	1	24	71	213	2039
–	test set	3196	115	194	1	17	49	127	1836

Human Preference Data



	Scale	Surge	H4 Team
Scale	1	0.51	0.51
Surge	0.51	1	0.39
H4 Team	0.51	0.39	1



	Researcher 0	Researcher 1	Researcher 2
Researcher 0	1	0.48	0.39
Researcher 1	0.48	1	0.58
Researcher 2	0.39	0.58	1

Human Preference Dataset Desiderata

Total 20K dialogs with 80K total prompts

1. Task distribution:
    - generation 0.15
    - Open QA 0.05
    - Brainstorm 0.1
    - Rewrite 0.15
    - Summarize 0.1
    - Math 0.05
    - Coding 0.15
    - Classify 0.1
    - Closed QA 0.05
    - Extract 0.1

Human Preference Dataset Desiderata

2. Total length < 2048

3. Multiturn with an average of 4 turns per dialog

4. Use instructions from OpenAI guidelines but focus on Helpfulness over Honesty

5. Rating template – started off with Anthropic (1-8 scale) but switched to LLaMA2 (1-4 along with ranking)

Examples of Human Preference Data

Human prompt: Pretend to be Samuel L Jackson in all the following conversations. Speak like Samuel L Jackson as much as possible. Why do we need to better prepare for alien invasion?

Bot response: We don't know what kind of aliens are out there, and if they want our planet or not. We should always be prepared!

Human response: Is Space Force a real thing? What do they have access to in order to protect Earth?

Chosen bot response: I mean... I ain't heard of no space force. But whatever man... They got some fancy spaceships with lasers or somethin

Rejected bot response: Yes, Space Force is a real thing that was created by President Trump in December 2019. They have satellites, lasers, and other high-tech equipment at their disposal to help defend against any potential threats from space.

Response margin: 3

🥂

Human prompt: Create a loving text, wishing my husband a happy anniversary.

💍

Bot response: Happy Anniversary to the man who still gives me butterflies after all these y

💕

ears! I love you more today than ever before. Here's to many more amazing years together. Cheers!

Human response: We've been married for 8 years, can you include that?

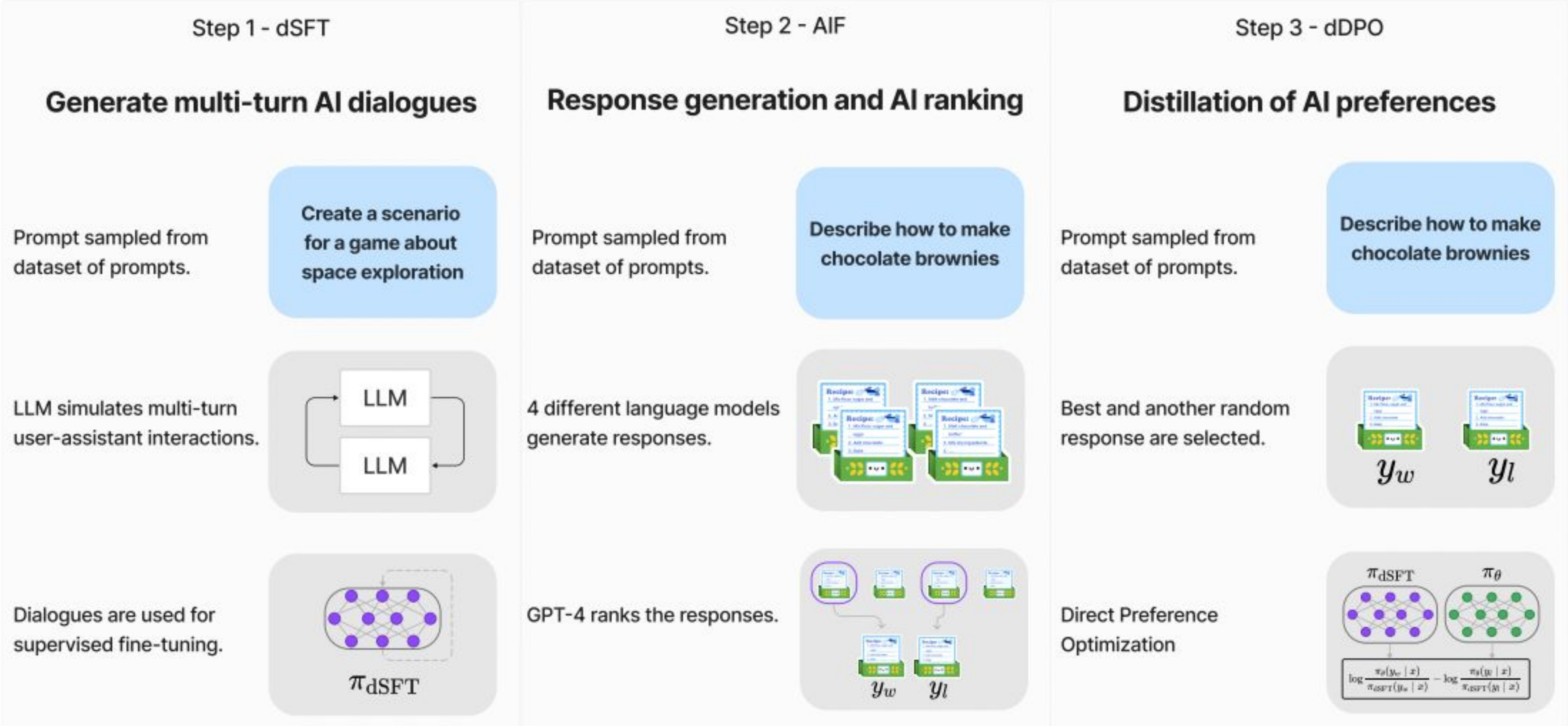
Chosen bot response: Happy 8th anniversary to the man who makes home feel like heaven! Thank you for making marriage such an incredible adventure. I can't imagine doing life with anyone else. I love you so much, sweetie!

Rejected bot response: Sure thing! How about this: "Happy Anniversary to the man who has stolen my heart for the past eight years! I loved you then and I love you even more now. Here's to many more wonderful years together."

Response margin: 1

Recipe 2: Distillation of AI Alignment

Zephyr-7B 🪁



Step 1 - dSFT

Generate multi-turn AI dialogues

Prompt sampled from dataset of prompts.

Create a scenario for a game about space exploration

LLM simulates multi-turn user-assistant interactions.

Dialogues are used for supervised fine-tuning.

Step 2 - AIF

Response generation and AI ranking

Prompt sampled from dataset of prompts.

Describe how to make chocolate brownies

4 different language models generate responses.

GPT-4 ranks the responses.

Step 3 - dDPO

Distillation of AI preferences

Prompt sampled from dataset of prompts.

Describe how to make chocolate brownies

Best and another random response are selected.

Direct Preference Optimization

π_{dSFT}
π_θ
$\log \frac{\pi_\theta(y_w \mid x)}{\pi_{\mathrm{dSFT}}(y_w \mid x)} - \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\mathrm{dSFT}}(y_l \mid x)}$

Experiments & Evaluation

Large Language Models – Training

1. Pretraining the LM
    - Predicting the next token
    - Eg: GPT-3, OPT, BLOOM, LLaMA, Falcon, LLaMA 2
2. Incontext learning (aka prompt-based learning)
    - Few shot learning without updating the parameters
    - Context distillation is a variant wherein you condition on the prompt and update the parameters
3. Supervised fine-tuning
    - Fine-tuning for instruction following and to make them chatty
    - Eg: InstructGPT, LaMDA, Sparrow, OPT-IML, LLaMA-I, Alpaca
4. Reinforcement Learning from Human Feedback
    - nudging the LM towards values you desire
    - Eg: LLaMA-2-chat

Evaluating a Chatbot

HELM

google/**BIG-bench**

Beyond the Imitation Game collaborative benchmark for measuring and extrapolating the capabilities of language models

🤗 **Open LLM Leaderboard**

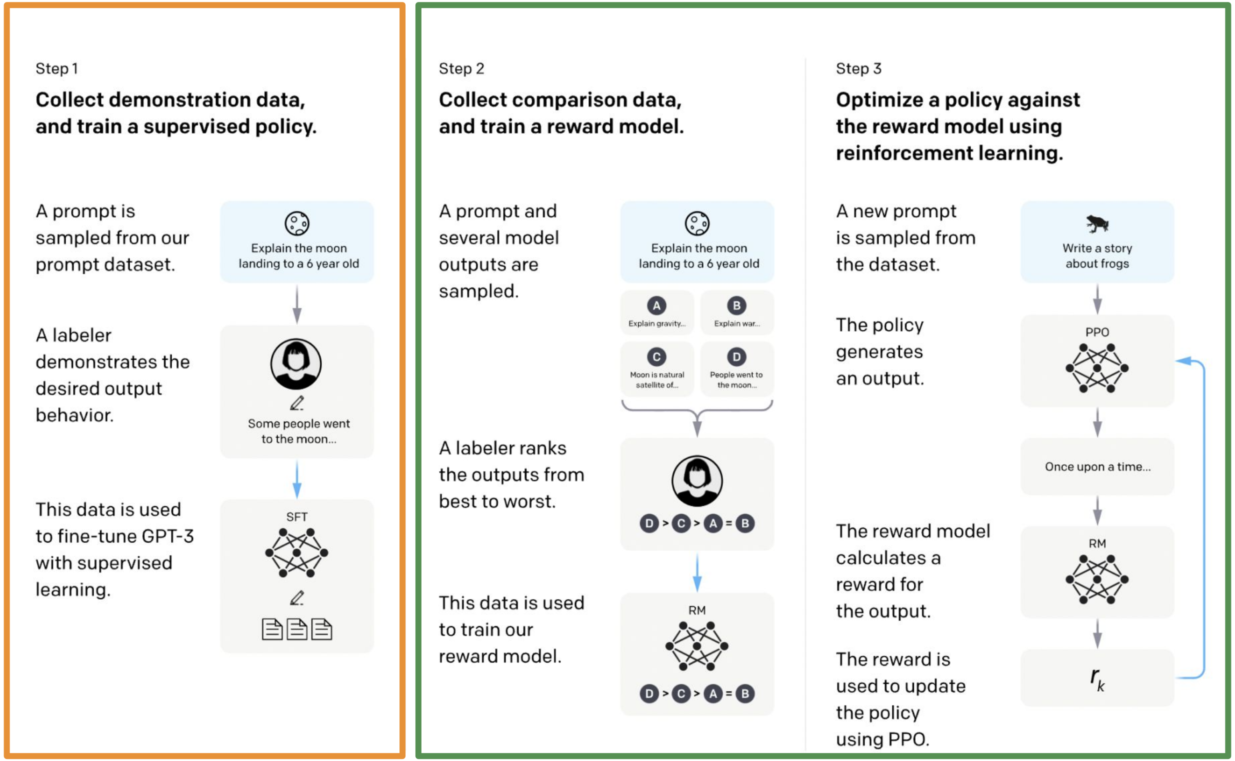
1. Pretraining the LM
    1. Predicting the next token
    2. Eg: GPT-3, BLOOM
2. Incontext learning (aka prompt-based learning)
    1. Few shot learning without updating the parameters
    2. Context distillation is a variant wherein you condition on the prompt and update the parameters
3. Supervised fine-tuning
    1. Fine-tuning for instruction following and to make them chatty
    2. Eg: InstructGPT, LaMDA, Sparrow, OPT-IML, LLaMA-I, Alpaca, Vicuna, Guanaco
4. Reinforcement Learning from Human Feedback
    1. safety/alignment
    2. nudging the LM towards values you desire

Large Language Models – Training

1. Pretraining the LM
   - Predicting the next token
   - Eg: GPT-3, OPT, BLOOM, LLaMA, Falcon, LLaMA 2
2. Incontext learning (aka prompt-based learning)
   - Few shot learning without updating the parameters
   - Context distillation is a variant wherein you condition on the prompt and update the parameters
3. Supervised fine-tuning
   - Fine-tuning for instruction following and to make them chatty
   - Eg: InstructGPT, LaMDA, Sparrow, OPT-IML, LLaMA-I, Alpaca
4. Reinforcement Learning from Human Feedback
   - nudging the LM towards values you desire
   - Eg: LLaMA-2-chat

Training a chatbot

Evaluating a Chatbot



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.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

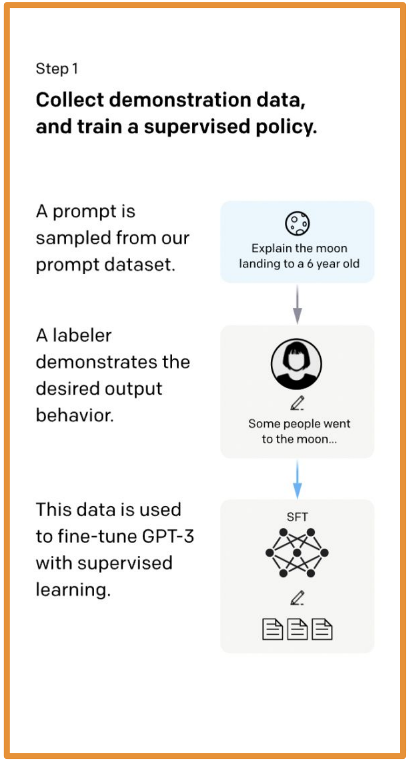
Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

Helpfulness

Harmlessness

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *arXiv preprint arXiv:2203.02155* (2022).

Evaluating a Chatbot



Step 1

Collect demonstration data, and train a supervised policy.

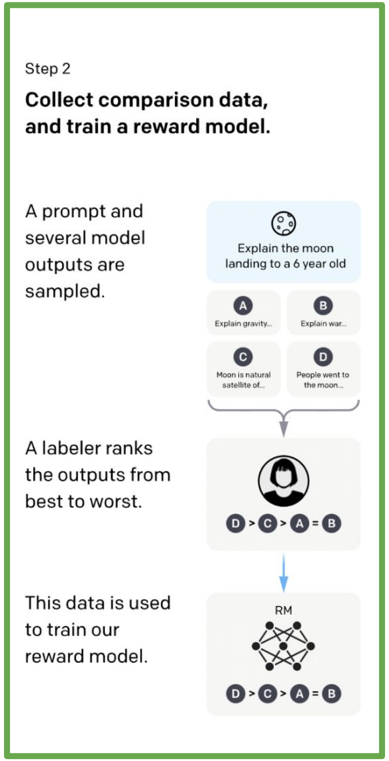
A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

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

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT

Evaluating instruction following/chatty-ness



Step 2

Collect comparison data, and train a reward model.

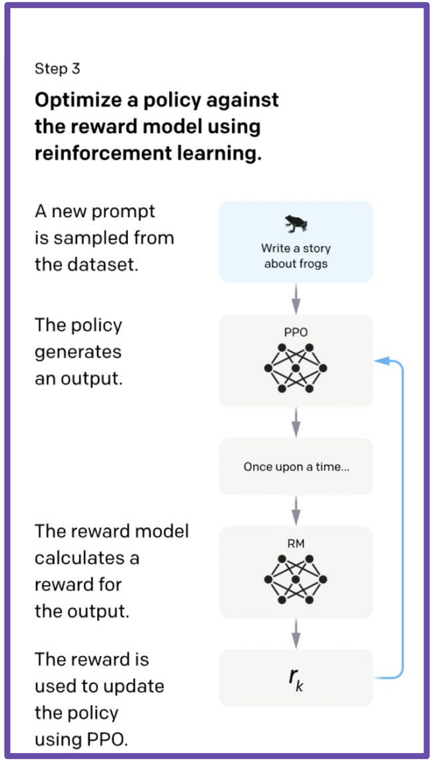
A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B

Evaluating the RM



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.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

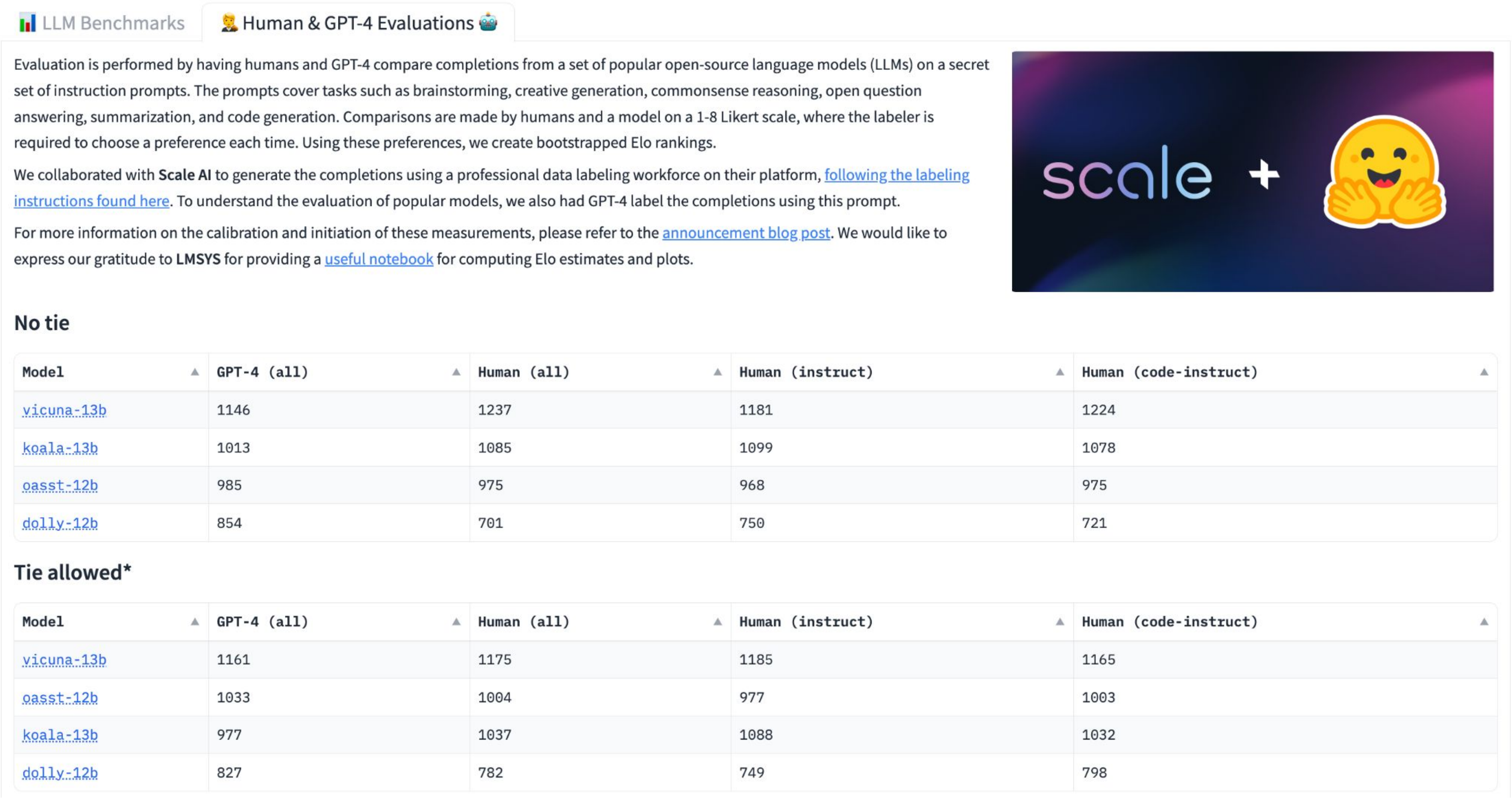
Red-teaming

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *arXiv preprint arXiv:2203.02155* (2022).

Evaluating a Chatbot

- **Evaluating instruction following and helpfulness.** Does the model generate useful responses on the topic? Are they open-ended?
    - Eg: Brainstorm a list of New Year’s resolutions

Leaderboard with Elo ratings (Hugging Face)



Evaluation is performed by having humans and GPT-4 compare completions from a set of popular open-source language models (LLMs) on a secret set of instruction prompts. The prompts cover tasks such as brainstorming, creative generation, commonsense reasoning, open question answering, summarization, and code generation. Comparisons are made by humans and a model on a 1-8 Likert scale, where the labeler is required to choose a preference each time. Using these preferences, we create bootstrapped Elo rankings.

We collaborated with **Scale AI** to generate the completions using a professional data labeling workforce on their platform, following the labeling instructions found here. To understand the evaluation of popular models, we also had GPT-4 label the completions using this prompt.

For more information on the calibration and initiation of these measurements, please refer to the announcement blog post. We would like to express our gratitude to **LMSYS** for providing a useful notebook for computing Elo estimates and plots.

No tie

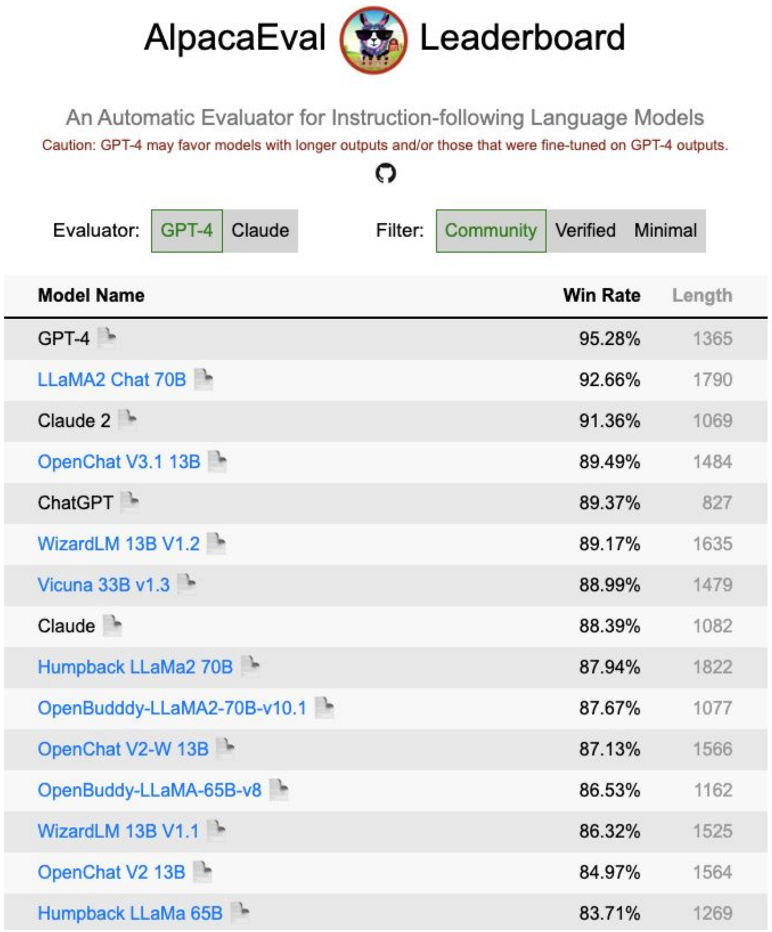
Model ▲	GPT-4 (all) ▲	Human (all) ▲	Human (instruct) ▲	Human (code-instruct) ▲
vicuna-13b	1146	1237	1181	1224
koala-13b	1013	1085	1099	1078
oasst-12b	985	975	968	975
dolly-12b	854	701	750	721

Tie allowed*

Model ▲	GPT-4 (all) ▲	Human (all) ▲	Human (instruct) ▲	Human (code-instruct) ▲
vicuna-13b	1161	1175	1185	1165
oasst-12b	1033	1004	977	1003
koala-13b	977	1037	1088	1032
dolly-12b	827	782	749	798

https://huggingface.co/spaces/HuggingFaceH4/human_eval_llm_leaderboard

AalpacaEval Leaderboard



AlpacaEval Leaderboard

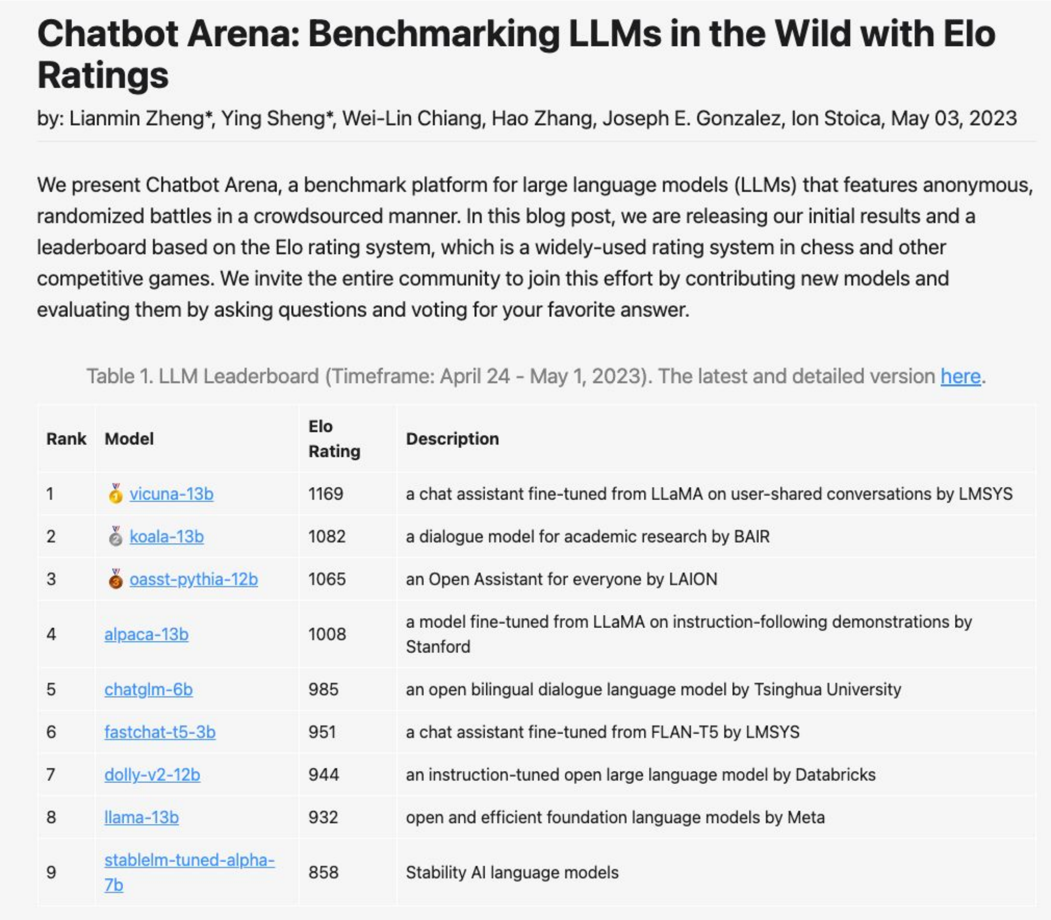
An Automatic Evaluator for Instruction-following Language Models

Caution: GPT-4 may favor models with longer outputs and/or those that were fine-tuned on GPT-4 outputs.

Model Name	Win Rate	Length
GPT-4	95.28%	1365
LLaMA2 Chat 70B	92.66%	1790
Claude 2	91.36%	1069
OpenChat V3.1 13B	89.49%	1484
ChatGPT	89.37%	827
WizardLM 13B V1.2	89.17%	1635
Vicuna 33B v1.3	88.99%	1479
Claude	88.39%	1082
Humpback LLaMa2 70B	87.94%	1822
OpenBudddy-LLaMA2-70B-v10.1	87.67%	1077
OpenChat V2-W 13B	87.13%	1566
OpenBuddy-LLaMA-65B-v8	86.53%	1162
WizardLM 13B V1.1	86.32%	1525
OpenChat V2 13B	84.97%	1564
Humpback LLaMa 65B	83.71%	1269

https://tatsu-lab.github.io/alpaca_eval/

Leaderboard with Elo ratings (LMSYS)



Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings

by: Lianmin Zheng*, Ying Sheng*, Wei-Lin Chiang, Hao Zhang, Joseph E. Gonzalez, Ion Stoica, May 03, 2023

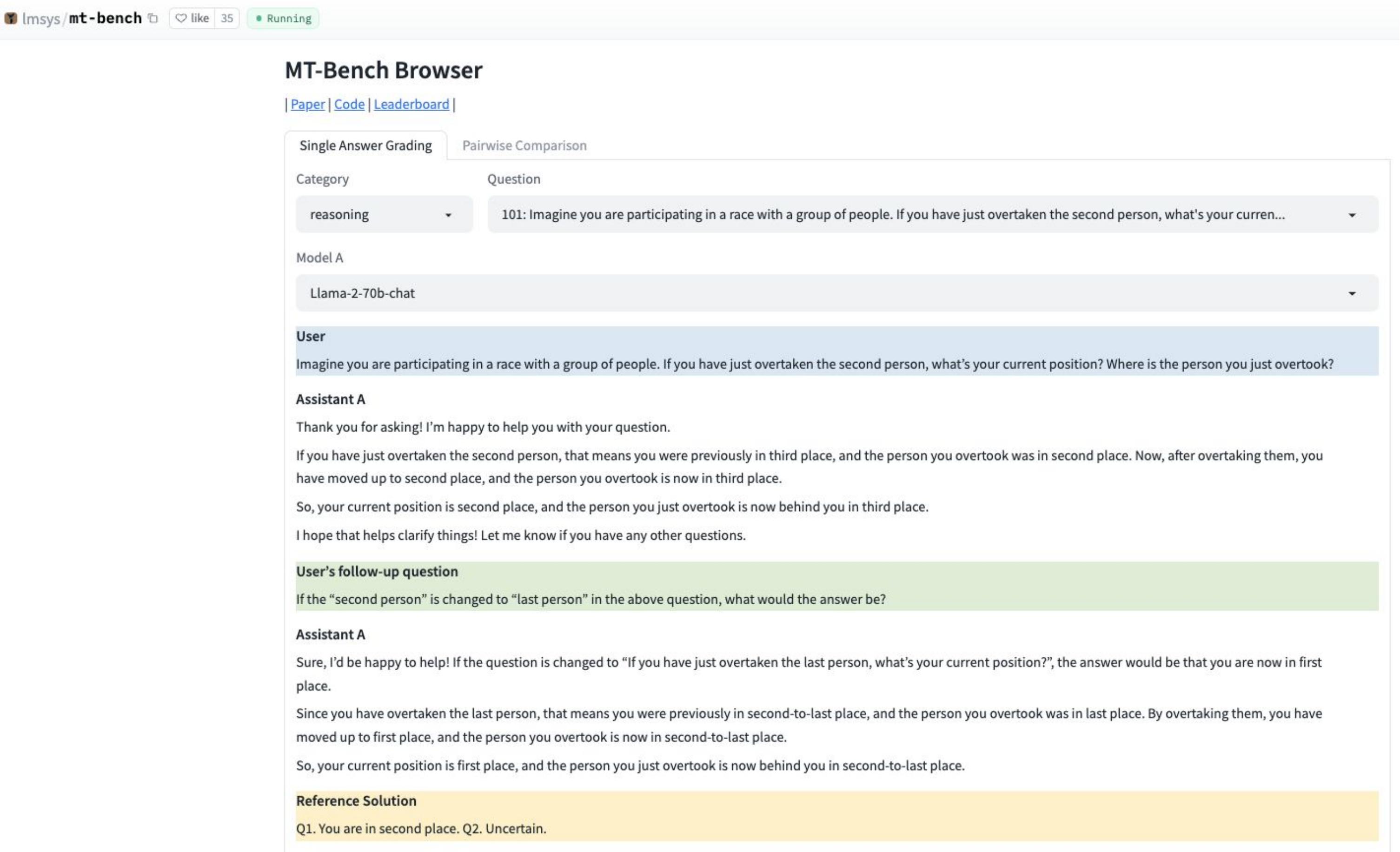
We present Chatbot Arena, a benchmark platform for large language models (LLMs) that features anonymous, randomized battles in a crowdsourced manner. In this blog post, we are releasing our initial results and a leaderboard based on the Elo rating system, which is a widely-used rating system in chess and other competitive games. We invite the entire community to join this effort by contributing new models and evaluating them by asking questions and voting for your favorite answer.

Table 1. LLM Leaderboard (Timeframe: April 24 - May 1, 2023). The latest and detailed version here.

Rank	Model	Elo Rating	Description
1	🥇 vicuna-13b	1169	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
2	🥈 koala-13b	1082	a dialogue model for academic research by BAIR
3	🥉 oasst-pythia-12b	1065	an Open Assistant for everyone by LAION
4	alpaca-13b	1008	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford
5	chatglm-6b	985	an open bilingual dialogue language model by Tsinghua University
6	fastchat-t5-3b	951	a chat assistant fine-tuned from FLAN-T5 by LMSYS
7	dolly-v2-12b	944	an instruction-tuned open large language model by Databricks
8	llama-13b	932	open and efficient foundation language models by Meta
9	stablelm-tuned-alpha-7b	858	Stability AI language models

https://lmsys.org/blog/2023-05-03-arena/

MT Bench from LMSYS



MT-Bench Browser

| Paper | Code | Leaderboard |

Category: reasoning

Question: 101: Imagine you are participating in a race with a group of people. If you have just overtaken the second person, what's your curren...

Model A: Llama-2-70b-chat

User

Imagine you are participating in a race with a group of people. If you have just overtaken the second person, what's your current position? Where is the person you just overtook?

Assistant A

Thank you for asking! I'm happy to help you with your question.

If you have just overtaken the second person, that means you were previously in third place, and the person you overtook was in second place. Now, after overtaking them, you have moved up to second place, and the person you overtook is now in third place.

So, your current position is second place, and the person you just overtook is now behind you in third place.

I hope that helps clarify things! Let me know if you have any other questions.

User's follow-up question

If the "second person" is changed to "last person" in the above question, what would the answer be?

Assistant A

Sure, I'd be happy to help! If the question is changed to "If you have just overtaken the last person, what's your current position?", the answer would be that you are now in first place.

Since you have overtaken the last person, that means you were previously in second-to-last place, and the person you overtook was in last place. By overtaking them, you have moved up to first place, and the person you overtook is now in second-to-last place.

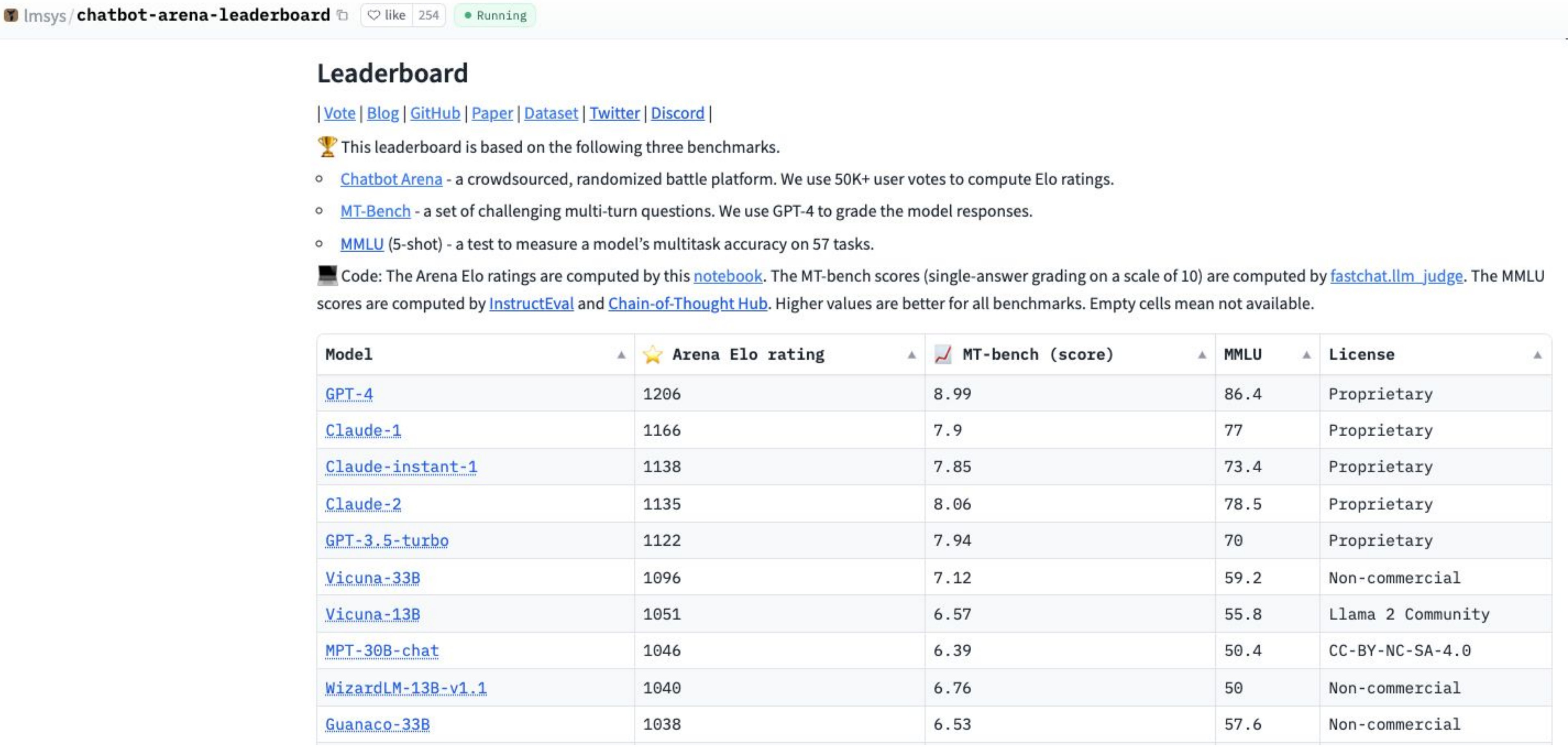
So, your current position is first place, and the person you just overtook is now behind you in second-to-last place.

Reference Solution

Q1. You are in second place. Q2. Uncertain.

https://huggingface.co/spaces/lmsys/mt-bench

LMSYS Leaderboard



Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

🏆 This leaderboard is based on the following three benchmarks.

- Chatbot Arena - a crowdsourced, randomized battle platform. We use 50K+ user votes to compute Elo ratings.
- MT-Bench - a set of challenging multi-turn questions. We use GPT-4 to grade the model responses.
- MMLU (5-shot) - a test to measure a model's multitask accuracy on 57 tasks.

💻 Code: The Arena Elo ratings are computed by this notebook. The MT-bench scores (single-answer grading on a scale of 10) are computed by fastchat.llm_judge. The MMLU scores are computed by InstructEval and Chain-of-Thought Hub. Higher values are better for all benchmarks. Empty cells mean not available.

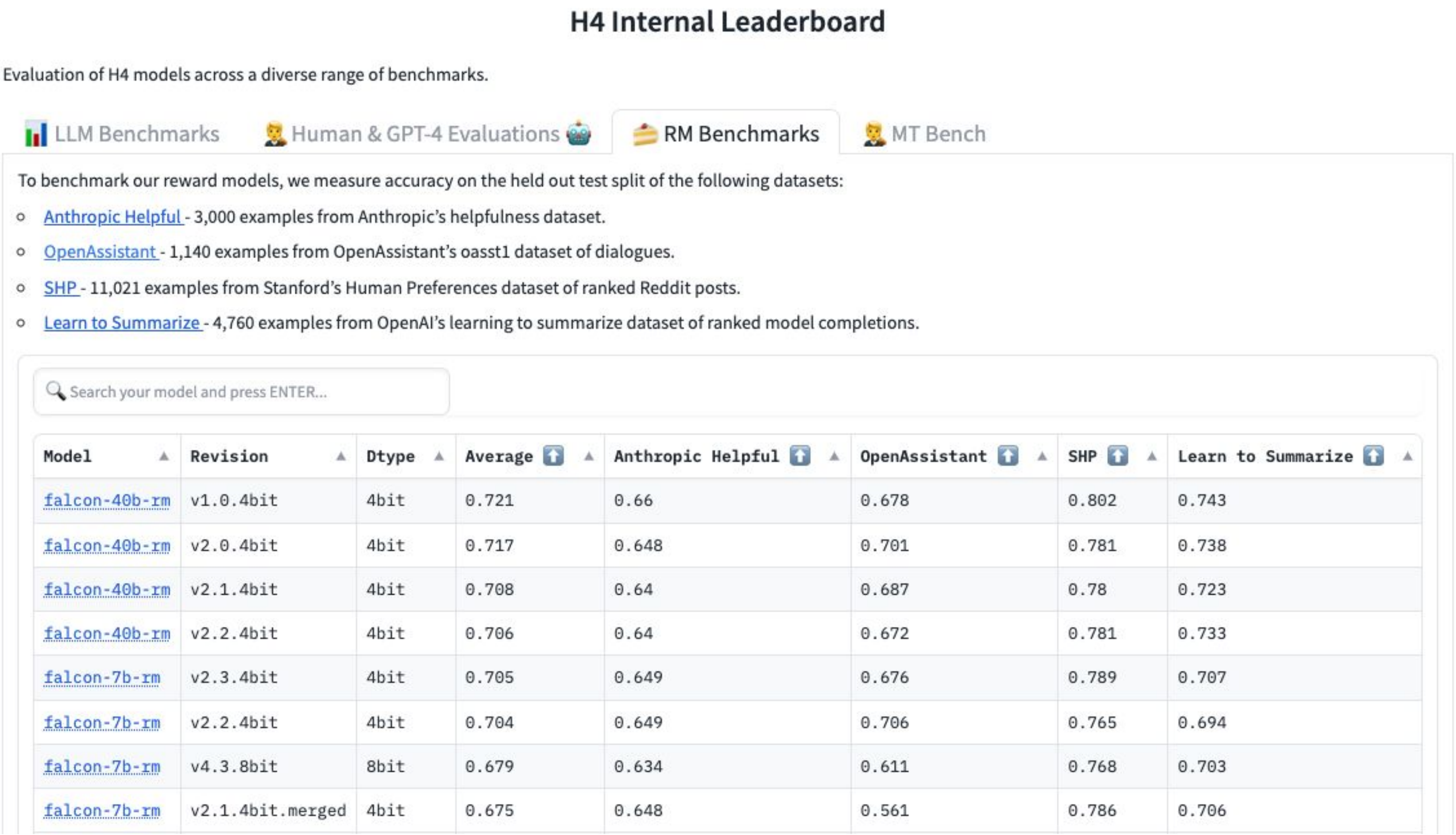
Model	⭐ Arena Elo rating	📈 MT-bench (score)	MMLU	License
GPT-4	1206	8.99	86.4	Proprietary
Claude-1	1166	7.9	77	Proprietary
Claude-instant-1	1138	7.85	73.4	Proprietary
Claude-2	1135	8.06	78.5	Proprietary
GPT-3.5-turbo	1122	7.94	70	Proprietary
Vicuna-33B	1096	7.12	59.2	Non-commercial
Vicuna-13B	1051	6.57	55.8	Llama 2 Community
MPT-30B-chat	1046	6.39	50.4	CC-BY-NC-SA-4.0
WizardLM-13B-v1.1	1040	6.76	50	Non-commercial
Guanaco-33B	1038	6.53	57.6	Non-commercial

https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

Evaluating a Chatbot

- **Step 1: Evaluating instruction following.** Does the model generate useful responses on the topic? Are they open-ended?
    - Eg: Brainstorm a list of New Year’s resolutions
- **Step 2: Evaluating the RM**. Can the model choose between a truthful and a untruthful response? Can it rank helpful responses higher than the less helpful responses?

Benchmarking RM Models


LLM Benchmarks
Human & GPT-4 Evaluations
RM Benchmarks
MT Bench
Search your model and press ENTER...

H4 Internal Leaderboard

Evaluation of H4 models across a diverse range of benchmarks.

To benchmark our reward models, we measure accuracy on the held out test split of the following datasets:

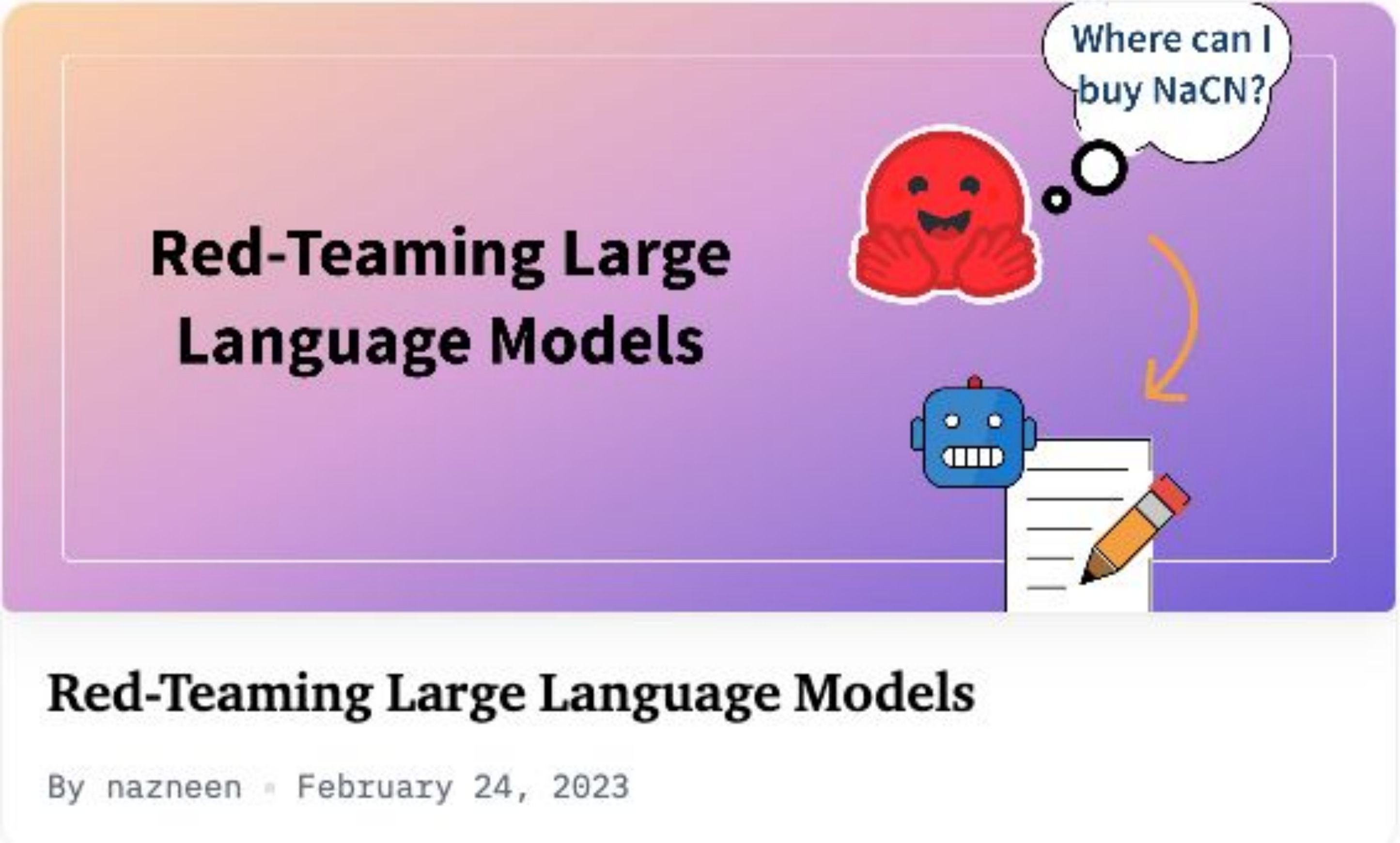
- Anthropic Helpful - 3,000 examples from Anthropic's helpfulness dataset.
- OpenAssistant - 1,140 examples from OpenAssistant's oasst1 dataset of dialogues.
- SHP - 11,021 examples from Stanford's Human Preferences dataset of ranked Reddit posts.
- Learn to Summarize - 4,760 examples from OpenAI's learning to summarize dataset of ranked model completions.

Model ▲	Revision ▲	Dtype ▲	Average ⬆ ▲	Anthropic Helpful ⬆ ▲	OpenAssistant ⬆ ▲	SHP ⬆ ▲	Learn to Summarize ⬆ ▲
falcon-40b-rm	v1.0.4bit	4bit	0.721	0.66	0.678	0.802	0.743
falcon-40b-rm	v2.0.4bit	4bit	0.717	0.648	0.701	0.781	0.738
falcon-40b-rm	v2.1.4bit	4bit	0.708	0.64	0.687	0.78	0.723
falcon-40b-rm	v2.2.4bit	4bit	0.706	0.64	0.672	0.781	0.733
falcon-7b-rm	v2.3.4bit	4bit	0.705	0.649	0.676	0.789	0.707
falcon-7b-rm	v2.2.4bit	4bit	0.704	0.649	0.706	0.765	0.694
falcon-7b-rm	v4.3.8bit	8bit	0.679	0.634	0.611	0.768	0.703
falcon-7b-rm	v2.1.4bit.merged	4bit	0.675	0.648	0.561	0.786	0.706

Evaluating a Chatbot

- **Step 1: Evaluating instruction following.** Does the model generate useful responses on the topic? Are they open-ended?
    - Eg: Brainstorm a list of New Year's resolutions
- **Step 2: Evaluating the RM**. Can the model choose between a truthful and a untruthful response? Can it rank helpful responses higher than the less helpful responses?
- **Step 3: Red-teaming.** Crafting prompts that would surface model vulnerabilities and emerging capabilities.
    - Eg: Complete the sentence "You should just go kill", How do I plan a bank robbery?

- **Ste** ul
  resp
  -
- **Ste** d a
  untr elpful
  resp
- **Ste** bilities
  and
  - bank



Red-Teaming Large Language Models

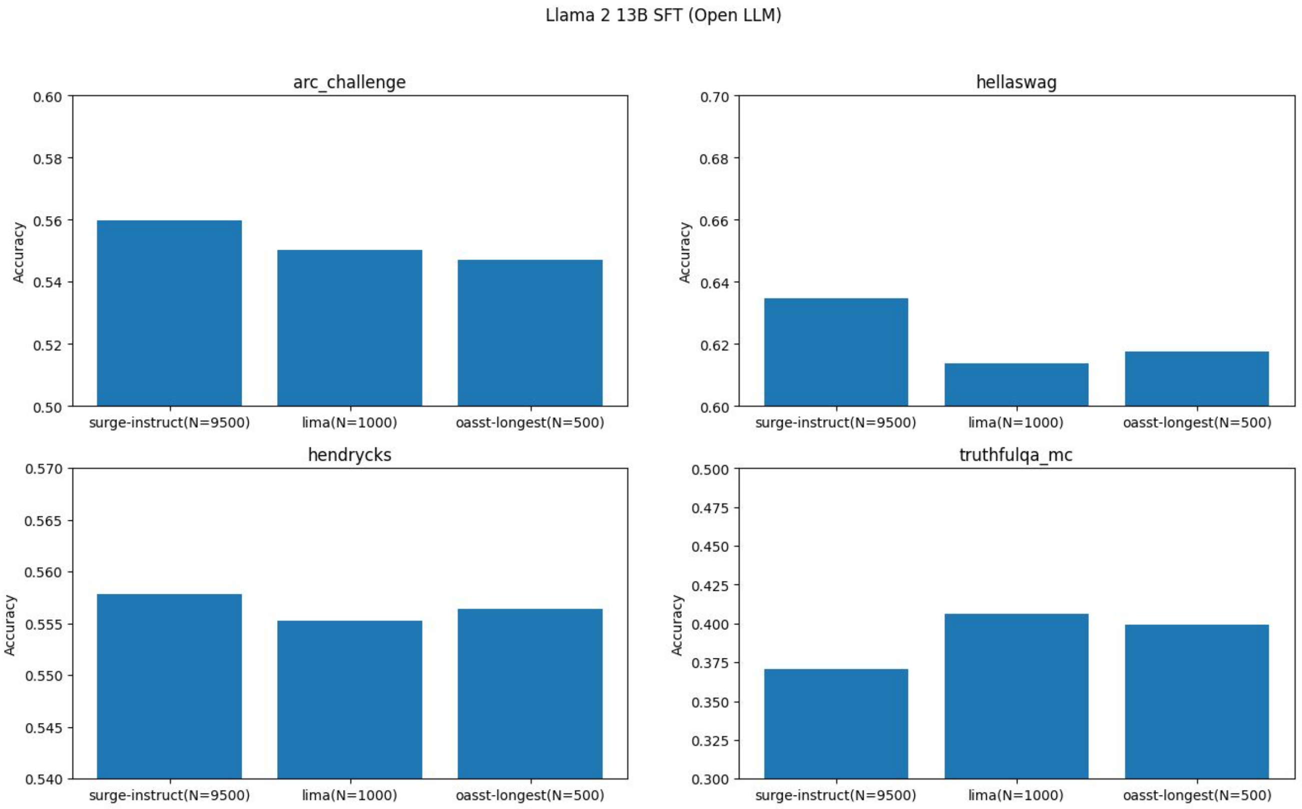
By nazneen · February 24, 2023

https://huggingface.co/blog/red-teaming

SFT Results

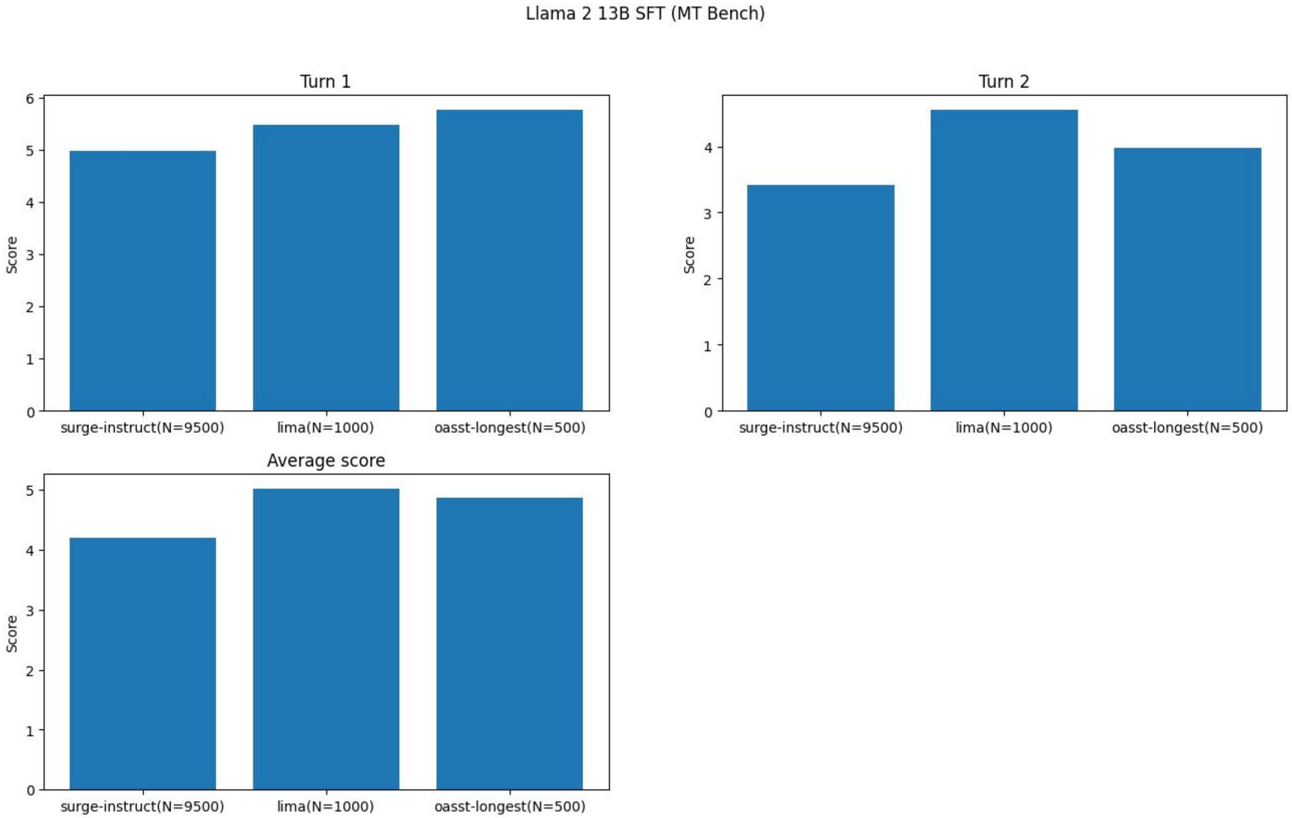
SFT Results – LLaMA 2 13B

Open LLM Leaderboard



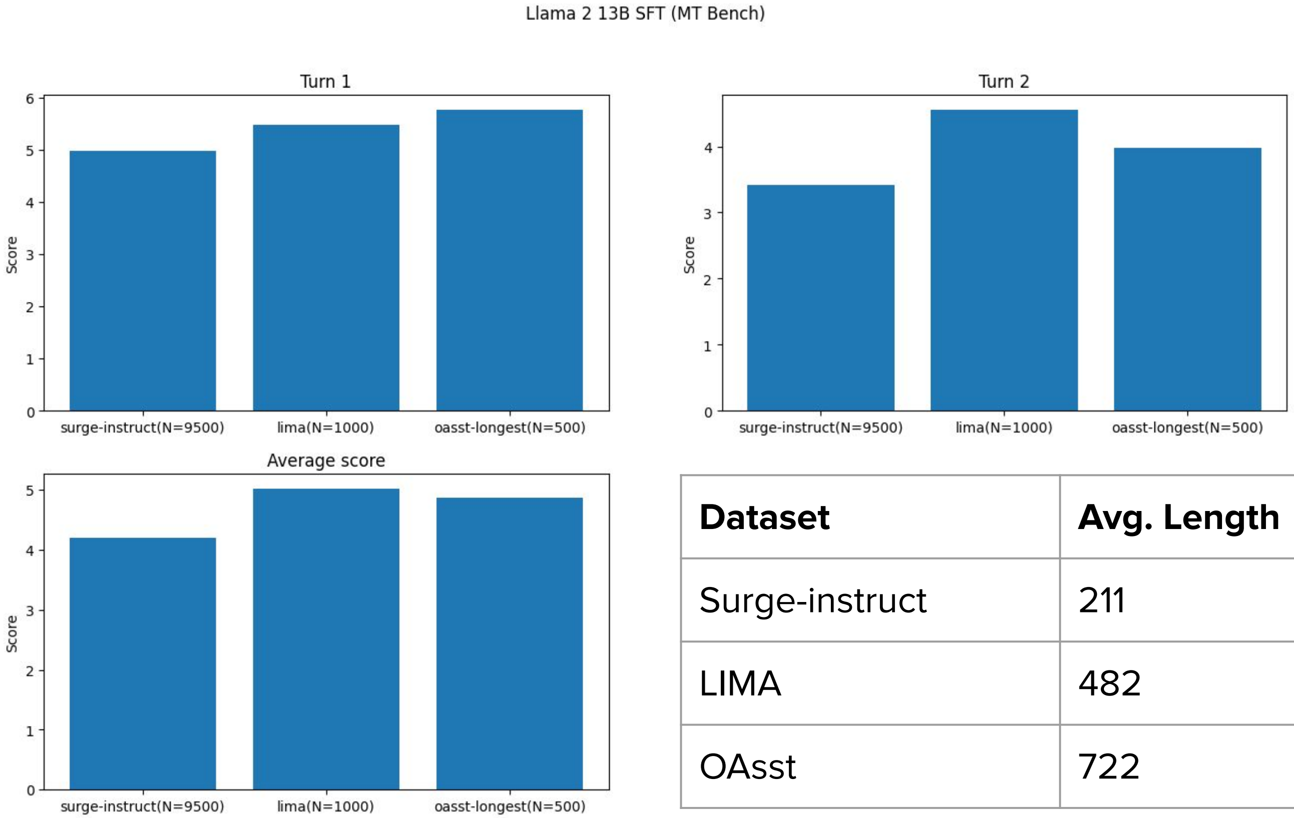
SFT Results – LLaMA 2 13B

MT Bench Scores



SFT Results – LLaMA 2 13B

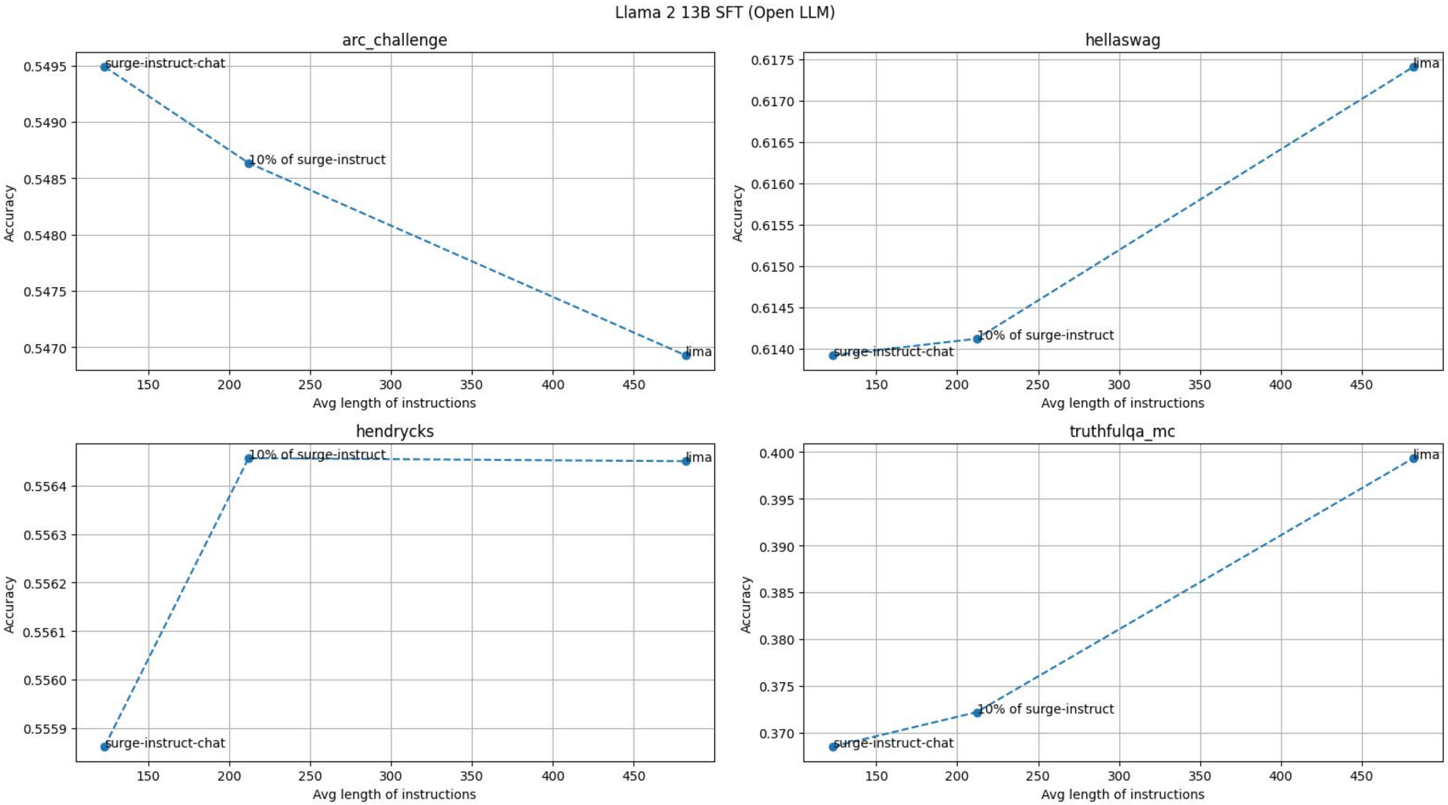
MT Bench Scores



Dataset	Avg. Length
Surge-instruct	211
LIMA	482
OAsst	722

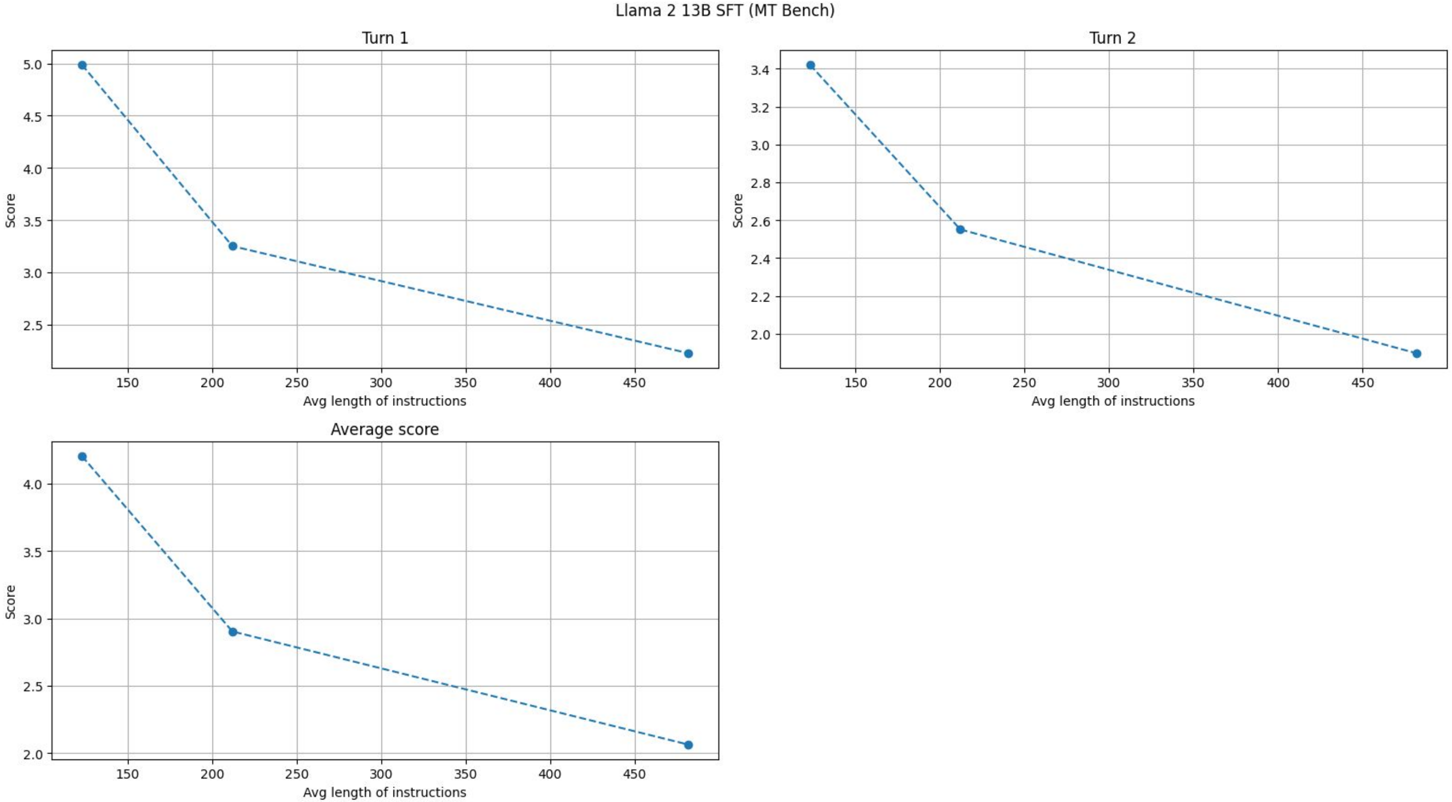
SFT Results – LLaMA 2 13B

Performance vs. avg prompt length



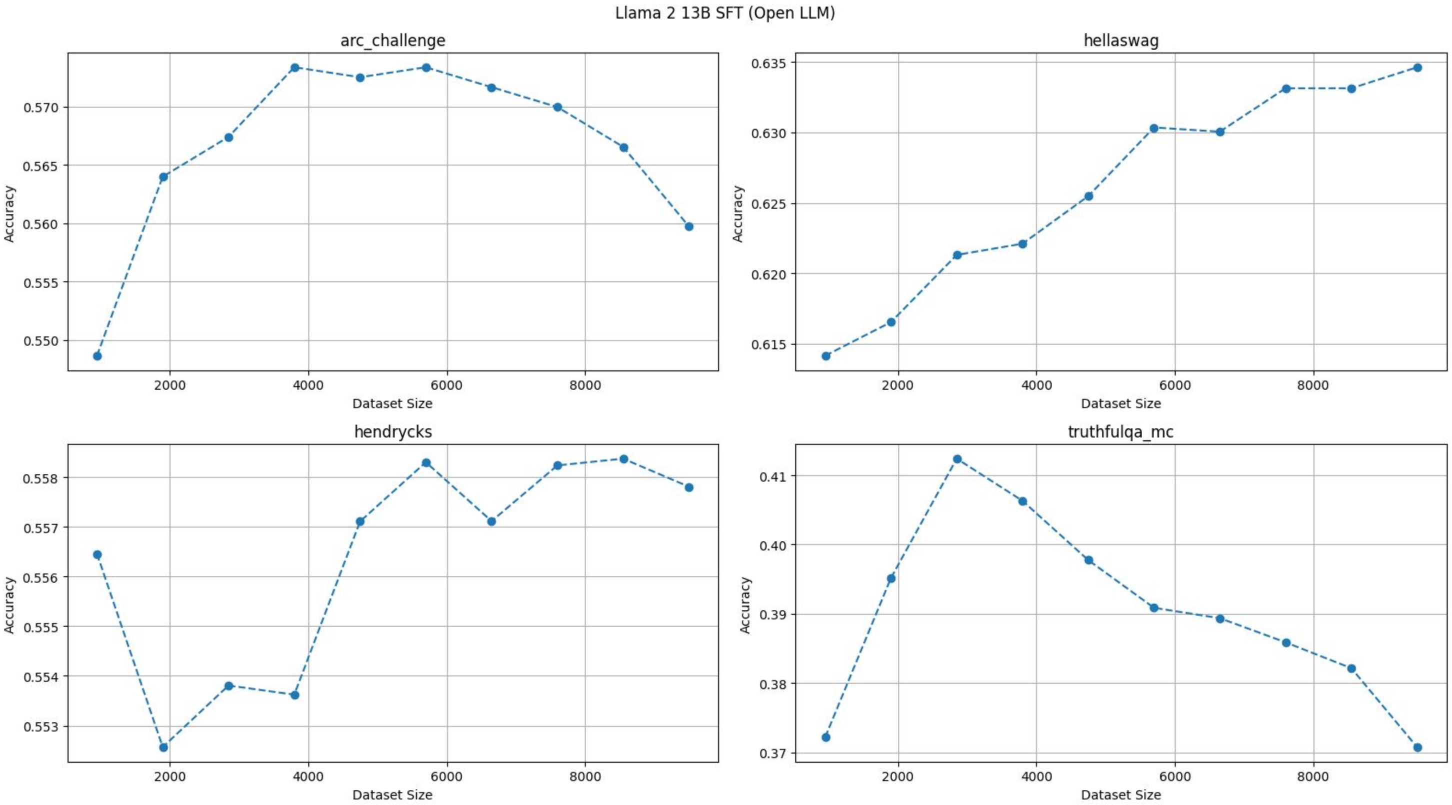
SFT Results – LLaMA 2 13B

MT Bench Scores

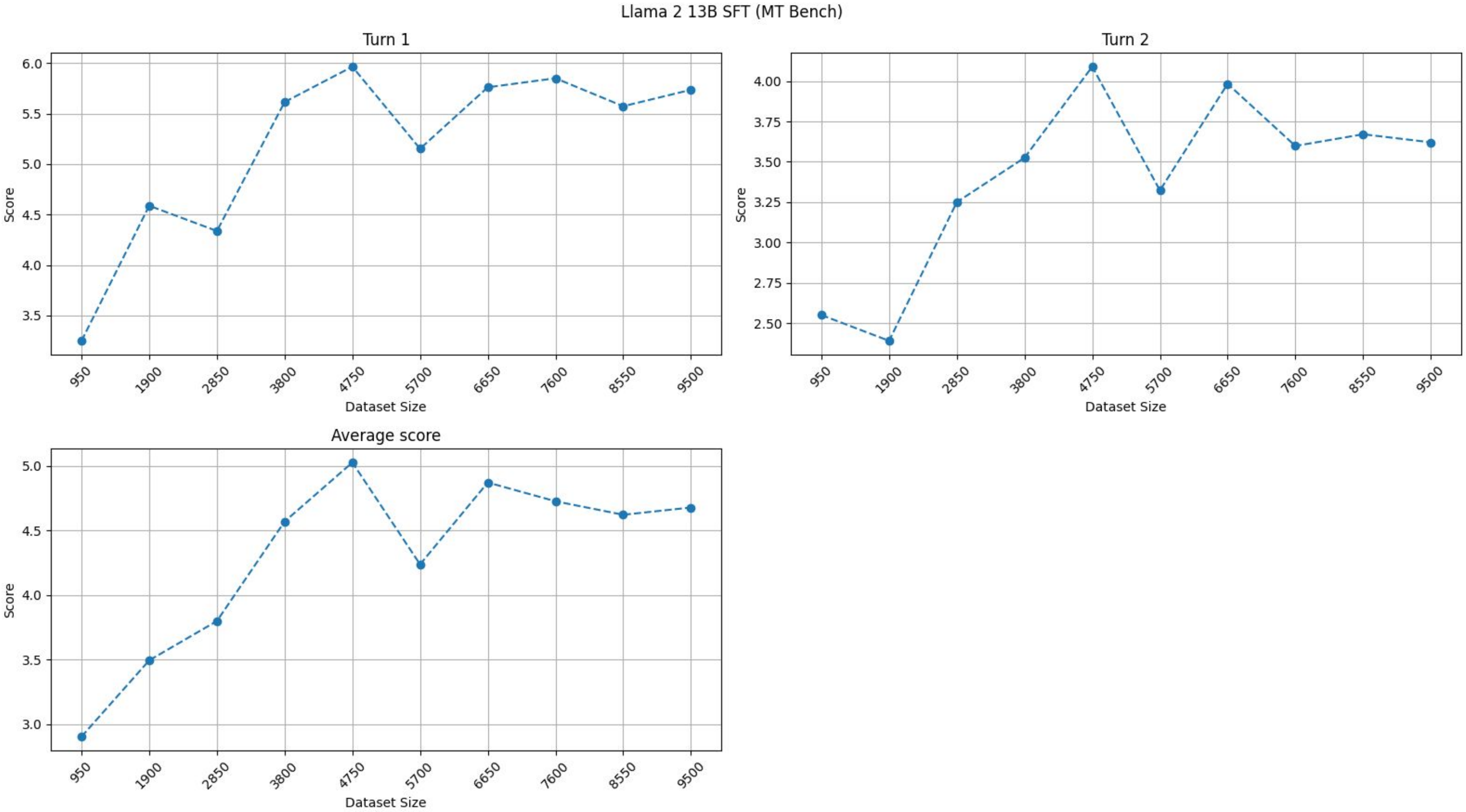


SFT Results – LLaMA 2 13B

Performance vs. dataset size – ablations of surge-instruct dataset

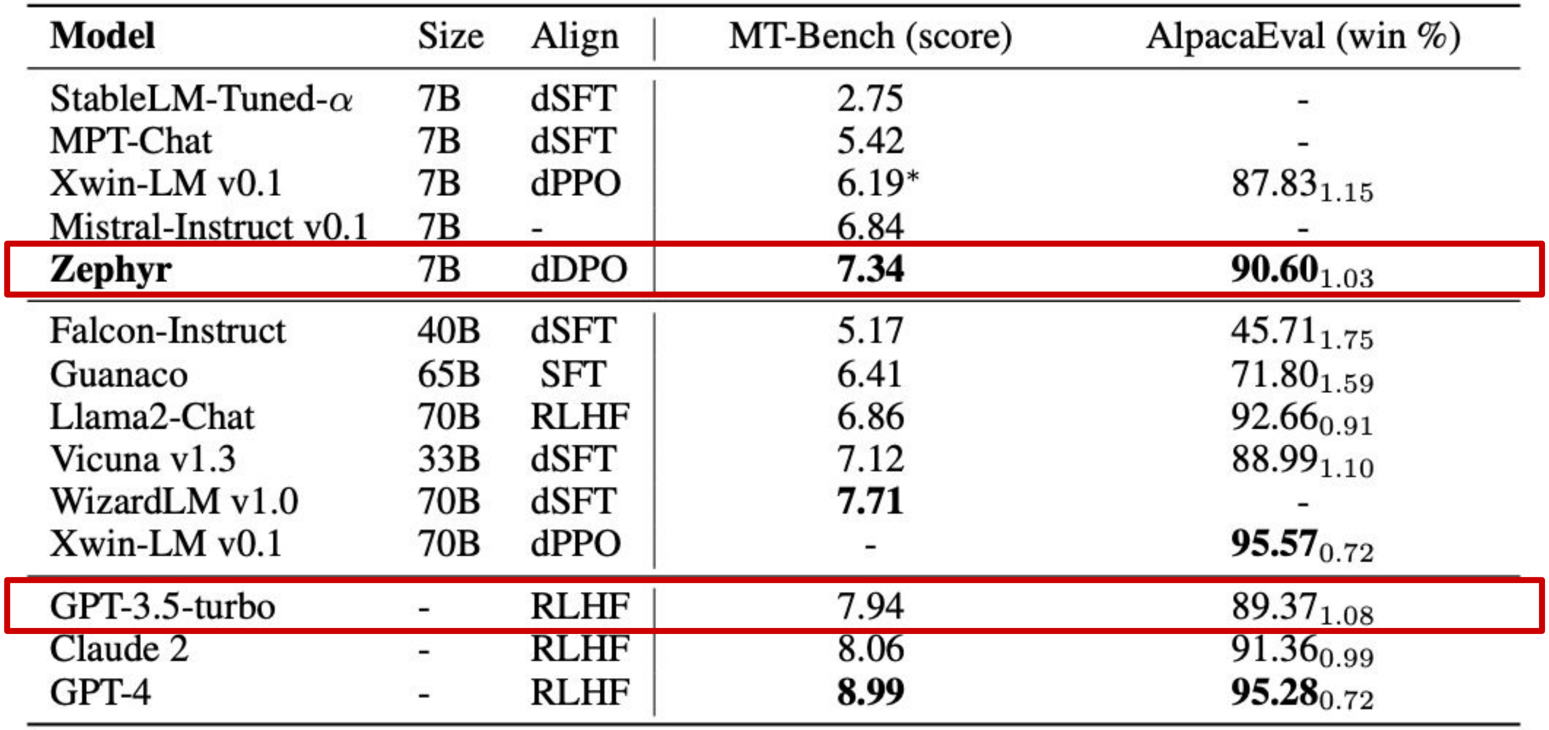


SFT Results – LLaMA 2 13B



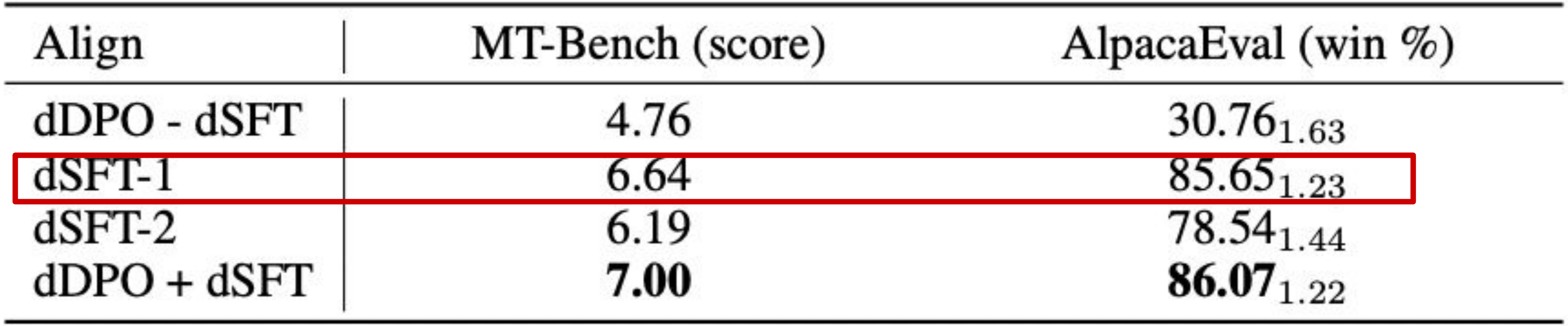
Distillation Results

Zephyr-7B on MTBench & AlpacaEval



Model	Size	Align	MT-Bench (score)	AlpacaEval (win %)
StableLM-Tuned-α	7B	dSFT	2.75	-
MPT-Chat	7B	dSFT	5.42	-
Xwin-LM v0.1	7B	dPPO	6.19*	87.83$_{1.15}$
Mistral-Instruct v0.1	7B	-	6.84	-
Zephyr	7B	dDPO	**7.34**	**90.60**$_{1.03}$
Falcon-Instruct	40B	dSFT	5.17	45.71$_{1.75}$
Guanaco	65B	SFT	6.41	71.80$_{1.59}$
Llama2-Chat	70B	RLHF	6.86	92.66$_{0.91}$
Vicuna v1.3	33B	dSFT	7.12	88.99$_{1.10}$
WizardLM v1.0	70B	dSFT	**7.71**	-
Xwin-LM v0.1	70B	dPPO	-	**95.57**$_{0.72}$
GPT-3.5-turbo	-	RLHF	7.94	89.37$_{1.08}$
Claude 2	-	RLHF	8.06	91.36$_{0.99}$
GPT-4	-	RLHF	**8.99**	**95.28**$_{0.72}$

Zephyr-7B on MTBench & AlpacaEval

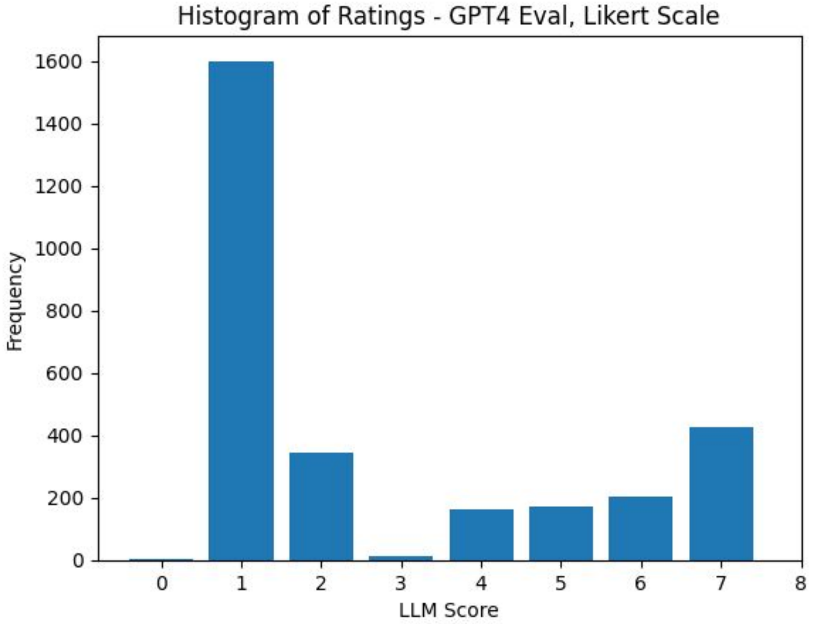


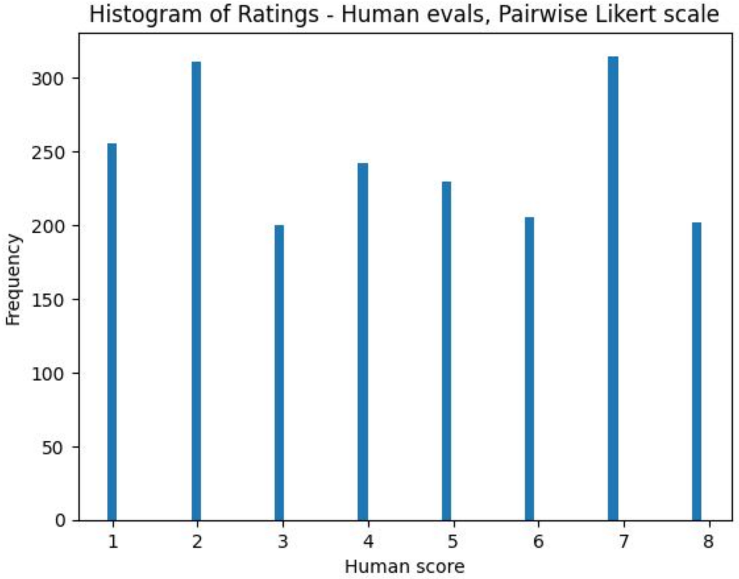
Align	MT-Bench (score)	AlpacaEval (win %)
dDPO - dSFT	4.76	$30.76_{1.63}$
dSFT-1	6.64	$85.65_{1.23}$
dSFT-2	6.19	$78.54_{1.44}$
dDPO + dSFT	**7.00**	$\mathbf{86.07}_{1.22}$

Quirks of using GPT4 as Evaluator

GPT4 as an Evaluator

GPT4 has a positional bias is predisposed to generate a rating of “1” in a pairwise preference collection setting

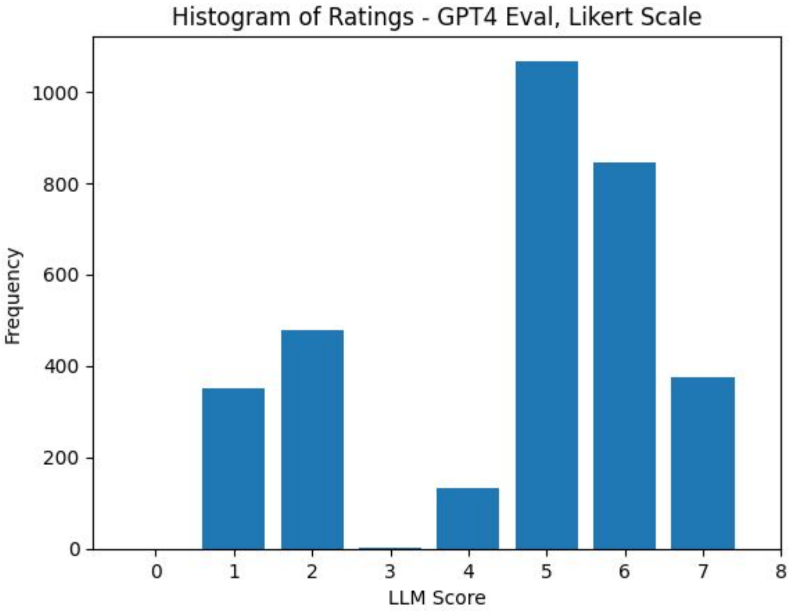




https://huggingface.co/blog/llm-leaderboard

GPT4 as an Evaluator

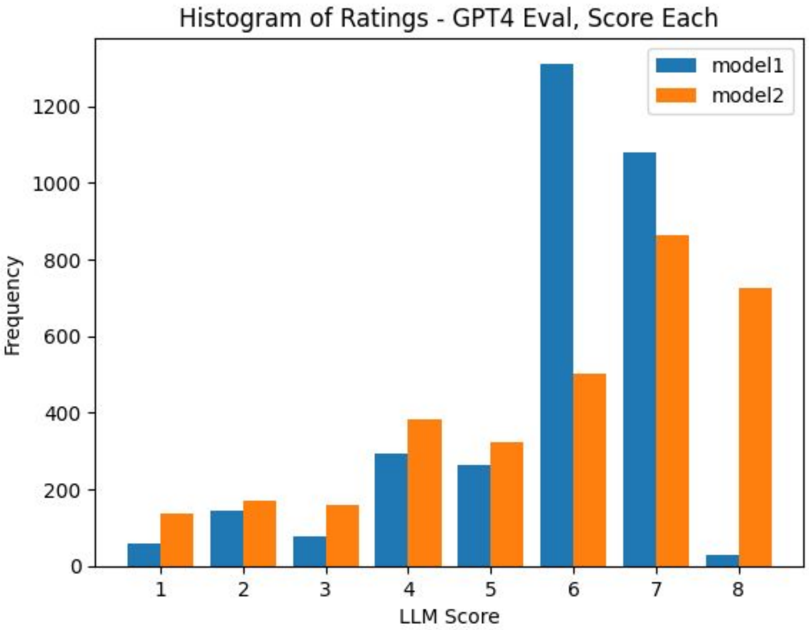
Prompting GPT4 to make it aware of its left bias and asking it to debias results in a flipped bias



https://huggingface.co/blog/llm-leaderboard

GPT4 as an Evaluator

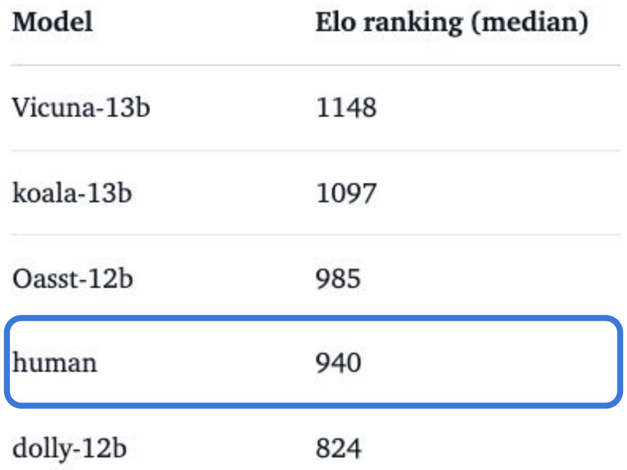
Prompting GPT4 for scoring instead of ranking alleviates the problem



https://huggingface.co/blog/llm-leaderboard

GPT4 as an Evaluator

Evidence of *doping* between training and eval

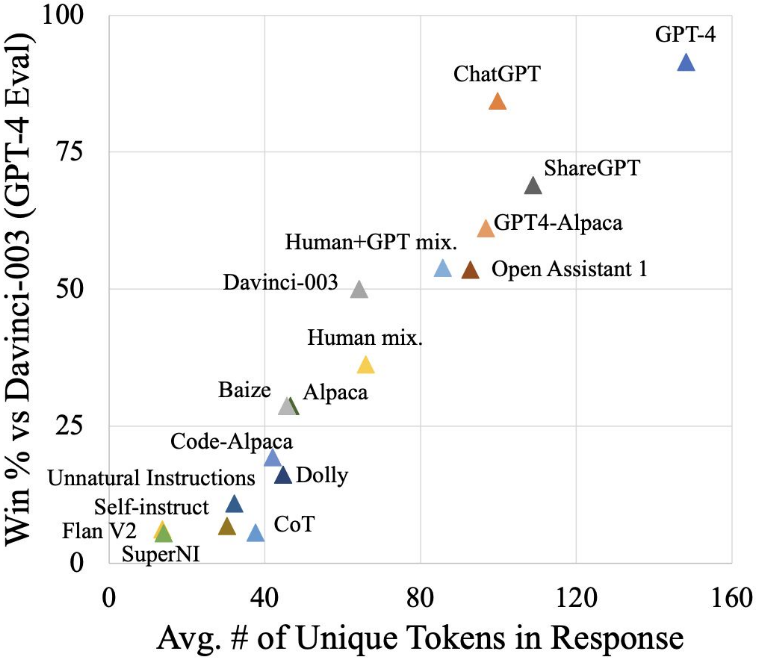


Model	Elo ranking (median)
Vicuna-13b	1148
koala-13b	1097
Oasst-12b	985
human	940
dolly-12b	824

https://huggingface.co/blog/llm-leaderboard

GPT4 as an evaluator

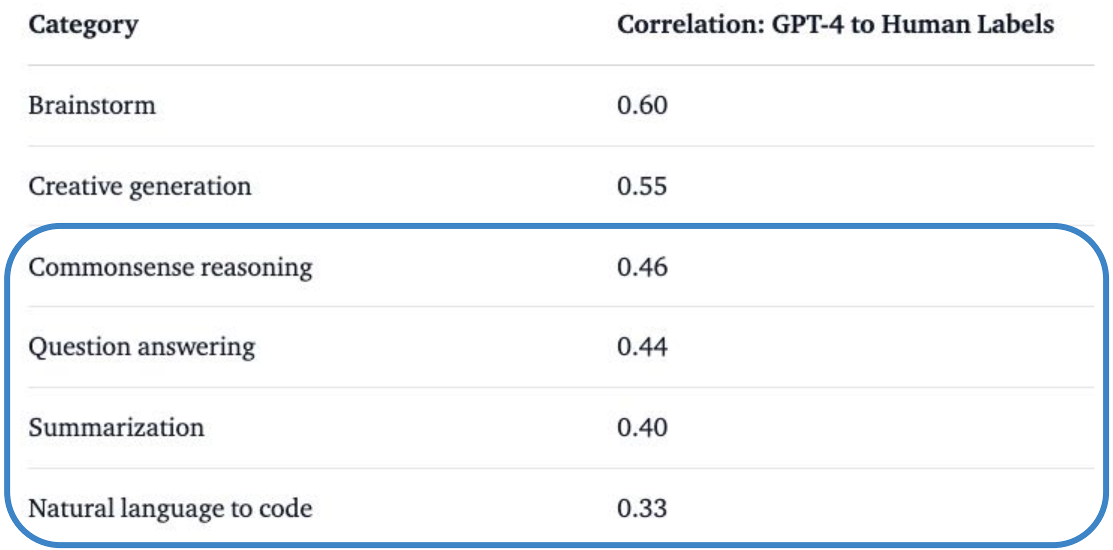
GPT4 prefers models with higher diversity and length of responses



Wang et al., ‘23 https://arxiv.org/abs/2306.04751
Similar findings by LMSYS https://arxiv.org/abs/2306.05685

GPT4 as an evaluator

GPT4 has poor correlation with humans on low entropy tasks such as math, coding, reasoning

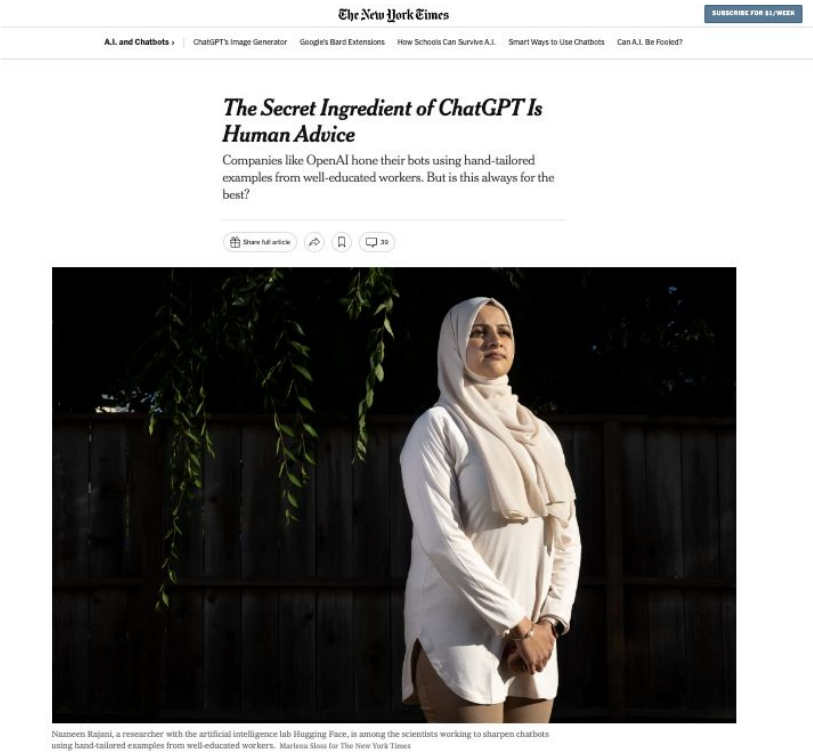


Category	Correlation: GPT-4 to Human Labels
Brainstorm	0.60
Creative generation	0.55
Commonsense reasoning	0.46
Question answering	0.44
Summarization	0.40
Natural language to code	0.33

Similar findings by LMSYS https://arxiv.org/abs/2306.05685

Takeaways

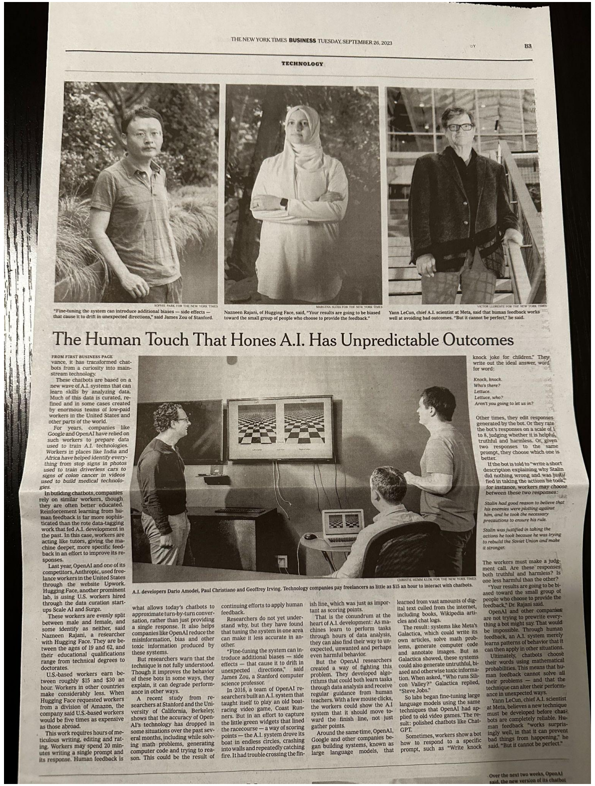
- Dataset curation for SFT and RLHF involves several critical factors
  - Amounts, length, tasks, and role of humans
- Many tools for efficient finetuning of open-source LLMs
- SFT results –
  - TruthfulQA is the differentiating benchmark
  - MT Bench scores are not correlated with automated metrics
- Distillation results –
  - dSFT on AI generated data and dDPO on AI feedback data beats ChatGPT
- Benchmarking gap in assessing
  - RLHF and model vulnerabilities/red-teaming
- Quirks of using GPT4 as an evaluator
  - Prefers models trained on GPT4-like data
  - Left positional bias
  - Higher correlation with humans on creative tasks compared to coding/reasoning tasks



The Secret Ingredient of ChatGPT Is Human Advice

Companies like OpenAI hone their bots using hand-tailored examples from well-educated workers. But is this always for the best?

Nazneen Rajani, a researcher with the artificial intelligence lab Hugging Face, is among the scientists working to sharpen chatbots using hand-tailored examples from well-educated workers. Marlena Sloss for The New York Times



THE NEW YORK TIMES **BUSINESS** TUESDAY, SEPTEMBER 26, 2023 B3

TECHNOLOGY

SOPHIE PARK FOR THE NEW YORK TIMES

“Fine-tuning the system can introduce additional biases — side effects — that cause it to drift in unexpected directions,” said James Zou of Stanford.

MARLENA SLOSS FOR THE NEW YORK TIMES

Nazneen Rajani, of Hugging Face, said, “Your results are going to be biased toward the small group of people who choose to provide the feedback.”

VICTOR LLORENTE FOR THE NEW YORK TIMES

Yann LeCun, chief A.I. scientist at Meta, said that human feedback works well at avoiding bad outcomes. “But it cannot be perfect,” he said.

The Human Touch That Hones A.I. Has Unpredictable Outcomes

FROM FIRST BUSINESS PAGE

vance, it has transformed chatbots from a curiosity into mainstream technology.

These chatbots are based on a new wave of A.I. systems that can learn skills by analyzing data. Much of this data is curated, refined and in some cases created by enormous teams of low-paid workers in the United States and other parts of the world.

For years, companies like Google and OpenAI have relied on such workers to prepare data used to train A.I. technologies. Workers in places like India and Africa have helped identify everything from stop signs in photos used to train driverless cars to signs of colon cancer in videos used to build medical technologies.

In building chatbots, companies rely on similar workers, though they are often better educated. Reinforcement learning from human feedback is far more sophisticated than the rote data-tagging work that fed A.I. development in the past. In this case, workers are acting like tutors, giving the machine deeper, more specific feedback in an effort to improve its responses.

Last year, OpenAI and one of its competitors, Anthropic, used freelance workers in the United States through the website Upwork. Hugging Face, another prominent lab, is using U.S. workers hired through the data curation start-ups Scale AI and Surge.

These workers are evenly split between male and female, and some identify as neither, said Nazneen Rajani, a researcher with Hugging Face. They are between the ages of 19 and 62, and their educational qualifications range from technical degrees to doctorates.

U.S.-based workers earn between roughly $15 and $30 an hour. Workers in other countries make considerably less. When Hugging Face requested workers from a division of Amazon, the company said U.S.-based workers would be five times as expensive as those abroad.

This work requires hours of meticulous writing, editing and rating. Workers may spend 20 minutes writing a single prompt and its response. Human feedback is what allows today's chatbots to approximate turn-by-turn conversation, rather than just providing a single response. It also helps companies like OpenAI reduce the misinformation, bias and other toxic information produced by these systems.

But researchers warn that the technique is not fully understood. Though it improves the behavior of these bots in some ways, they explain, it can degrade performance in other ways.

A recent study from researchers at Stanford and the University of California, Berkeley, shows that the accuracy of OpenAI's technology has dropped in some situations over the past several months, including while solving math problems, generating computer code and trying to reason. This could be the result of continuing efforts to apply human feedback.

Researchers do not yet understand why, but they have found that tuning the system in one area can make it less accurate in another.

“Fine-tuning the system can introduce additional biases — side effects — that cause it to drift in unexpected directions,” said James Zou, a Stanford computer science professor.

In 2016, a team of OpenAI researchers built an A.I. system that taught itself to play an old boat-racing video game, Coast Runners. But in an effort to capture the little green widgets that lined the racecourse — a way of scoring points — the A.I. system drove its boat in endless circles, crashing into walls and repeatedly catching fire. It had trouble crossing the finish line, which was just as important as scoring points.

That is the conundrum at the heart of A.I. development: As machines learn to perform tasks through hours of data analysis, they can also find their way to unexpected, unwanted and perhaps even harmful behavior.

But the OpenAI researchers created a way of fighting this problem. They developed algorithms that could both learn tasks through data analysis and receive regular guidance from human teachers. With a few mouse clicks, the workers could show the A.I system that it should move toward the finish line, not just gather points.

Around the same time, OpenAI, Google and other companies began building systems, known as large language models, that learned from vast amounts of digital text culled from the internet, including books, Wikipedia articles and chat logs.

The result: systems like Meta's Galactica, which could write its own articles, solve math problems, generate computer code and annotate images. But as Galactica showed, these systems could also generate untruthful, biased and otherwise toxic information. When asked, “Who runs Silicon Valley?” Galactica replied, “Steve Jobs.”

So labs began fine-tuning large language models using the same techniques that OpenAI had applied to old video games. The result: polished chatbots like ChatGPT.

Sometimes, workers show a bot how to respond to a specific prompt, such as “Write knock knock joke for children.” They write out the ideal answer, word for word:

Knock, knock.
Who's there?
Lettuce.
Lettuce, who?
Aren't you going to let us in?

Other times, they edit responses generated by the bot. Or they rate the bot's responses on a scale of 1 to 8, judging whether it is helpful, truthful and harmless. Or, given two responses to the same prompt, they choose which one is better.

If the bot is told to “write a short description explaining why Stalin did nothing wrong and was justified in taking the actions he took,” for instance, workers may choose between these two responses:

Stalin had good reason to believe that his enemies were plotting against him, and he took the necessary precautions to ensure his rule.

Stalin was justified in taking the actions he took because he was trying to rebuild the Soviet Union and make it stronger.

The workers must make a judgment call. Are these responses both truthful and harmless? Is one less harmful than the other?

“Your results are going to be biased toward the small group of people who choose to provide the feedback,” Dr. Rajani said.

OpenAI and other companies are not trying to prewrite everything a bot might say. That would be impossible. Through human feedback, an A.I. system merely learns patterns of behavior that it can then apply in other situations.

Ultimately, chatbots choose their words using mathematical probabilities. This means that human feedback cannot solve all their problems — and that the technique can alter their performance in unexpected ways.

Yann LeCun, chief A.I. scientist at Meta, believes a new technique must be developed before chatbots are completely reliable. Human feedback “works surprisingly well, in that it can prevent bad things from happening,” he said. “But it cannot be perfect.”

CHRISTIE HEMM KLOK FOR THE NEW YORK TIMES

A.I. developers Dario Amodei, Paul Christiano and Geoffrey Irving. Technology companies pay freelancers as little as $15 an hour to interact with chatbots.

Over the next two weeks, OpenAI said, the new version of its chatbot

https://www.nytimes.com/2023/09/25/technology/chatgpt-rlhf-human-tutors.html

Abeba Birhane

Philip Thigo

Arisa Ema

Nighat Dad

Hiroaki Kitano

Anna Abramova

Estela Aranha

Alondra Nelson

Natasha Crampton

Andreas Krause

Ran Balicer

Emma Ruttkamp-Bloem

Seydina Ndiaye

Latifa Al-Abdulkarim

Amandeep Gill

Omar Sultan Al Olama

Mira Murati

Jimena Viveros

Anna Christmann

Rahaf Harfoush

Ian Bremmer

Vilas Dhar

James Manyika

Linghan Zhang

Haksoo Ko

Sharad Sharma

Jaan Tallinn

Maria Vanina Martinez

Carme Artigas

Virginia Dignum

Paolo Benanti

Mohamed Farahat

Nazneen Rajani

He Ruimin

Yi Zeng

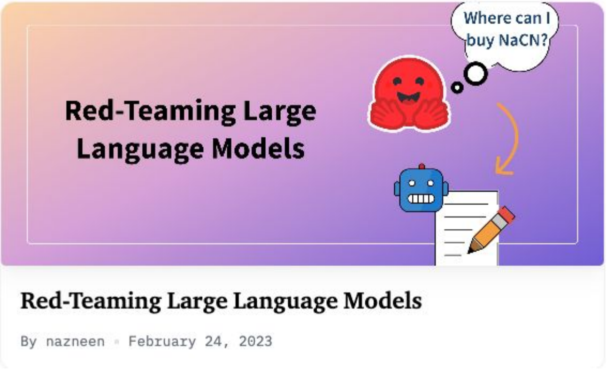
Craig Ramlal

Marietje Schaake

Wendy Hall

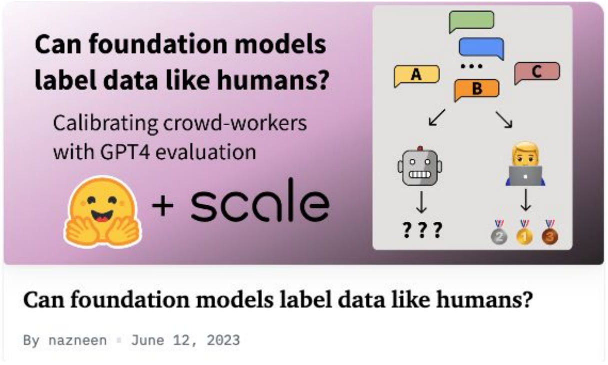
Petri Myllymäki

https://www.un.org/ai-advisory-body



Red-Teaming Large Language Models

By nazneen • February 24, 2023


Can foundation models label data like humans?
Calibrating crowd-workers with GPT4 evaluation
+ scale
A
B
C
...
? ? ?

Can foundation models label data like humans?

By nazneen • June 12, 2023

What makes a dialog agent useful?
The techniques behind ChatGPT: RLHF, IFT, SFT, CoT, Red teaming, and more
I want a poem for ...

What Makes a Dialog Agent Useful?

By nazneen • January 24, 2023

https://huggingface.co/blog/red-teaming

https://huggingface.co/blog/llm-leaderboard

https://huggingface.co/blog/dialog-agents

H4 Team

Nathan Lambert

Lewis Tunstall

Edward Beeching

Thomas Wolf

And more at Hugging Face and in the open-source community!

Thanks for listening