🧑‍💻Manual curation vs. 🤖AI distillation: Lessons learned from instruction following and feedback fine-tuning

Nazneen Rajani | nazneen@collinear.ai | @nazneenrajani

Introduction: LLM Training

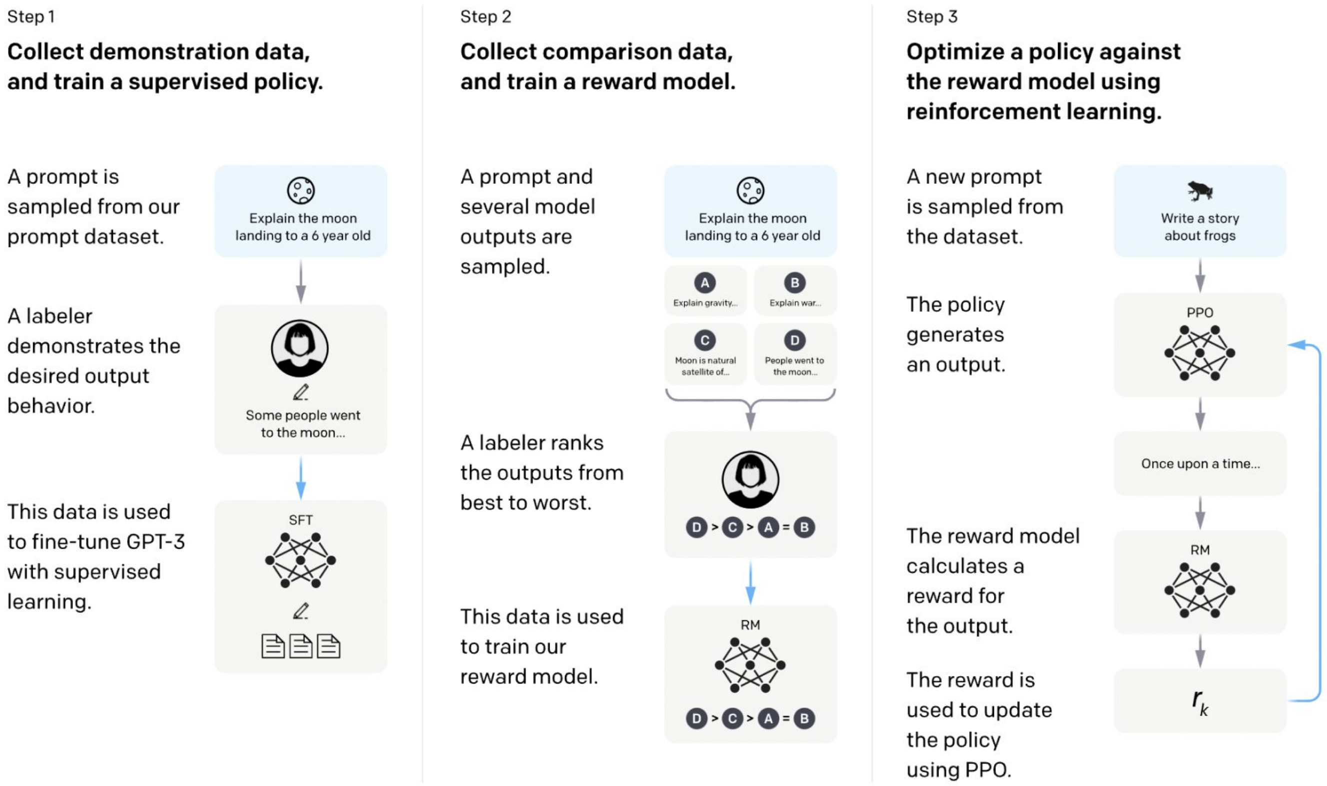
1. Pretraining the LM
   - Predicting the next token
   - Eg: GPT-3.5, OPT, BLOOM, LLaMA, Falcon, LLaMA 2, Mistral, Qwen, Yi
2. Incontext learning (aka prompt-based learning)
   - Few shot learning without updating the parameters
   - Context distillation is a variant wherein you condition on the prompt 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, Vicuna
4. Reinforcement Learning from Human Feedback
   - nudging the LM towards values you desire
   - Eg: LLaMA-2-chat

Introduction: LLM Training

1. Pretraining the LM
   - Predicting the next token
   - Eg: GPT-3.5, OPT, BLOOM, LLaMA, Falcon, LLaMA 2, Mistral, Qwen, yi
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

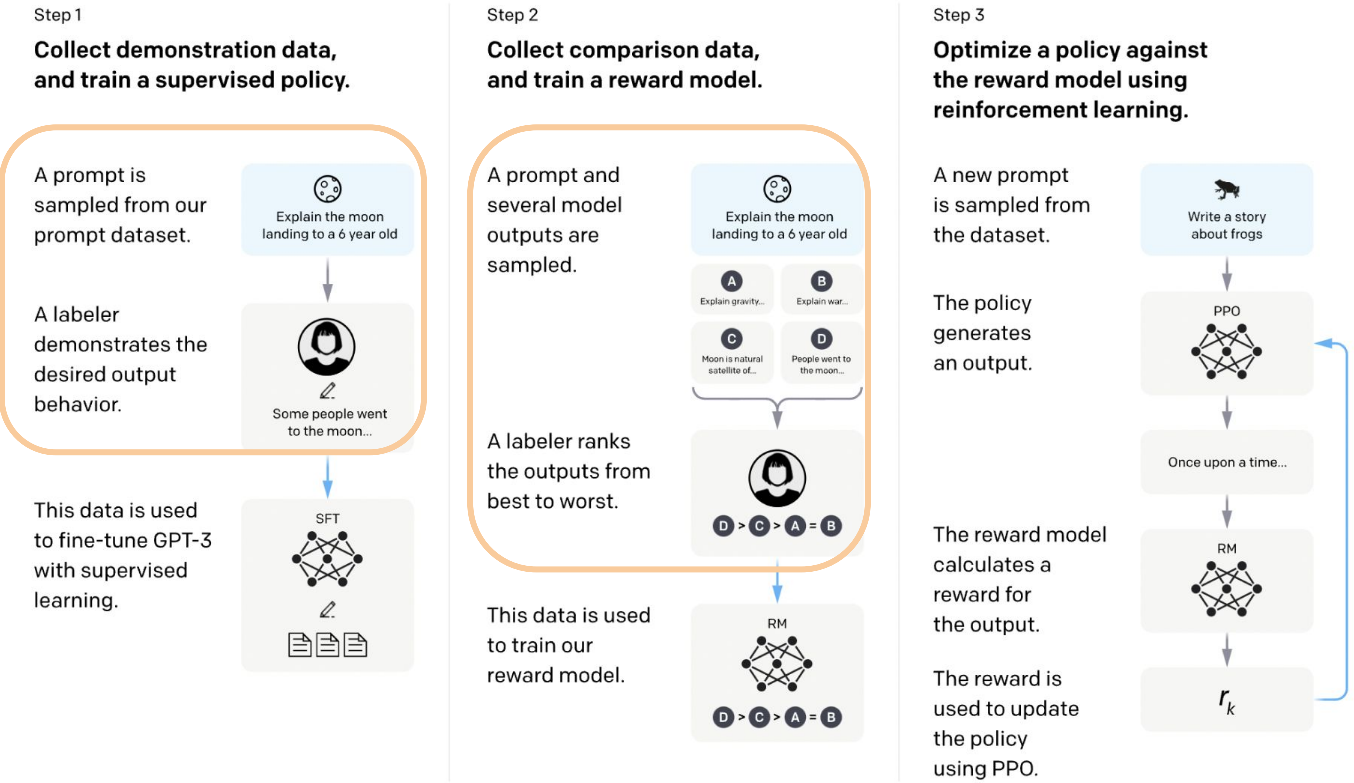
Training a chatbot

Training a Chatbot



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

Training a Chatbot



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

Training a Chatbot

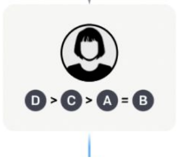
A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

Instruction demonstrations

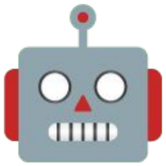
A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

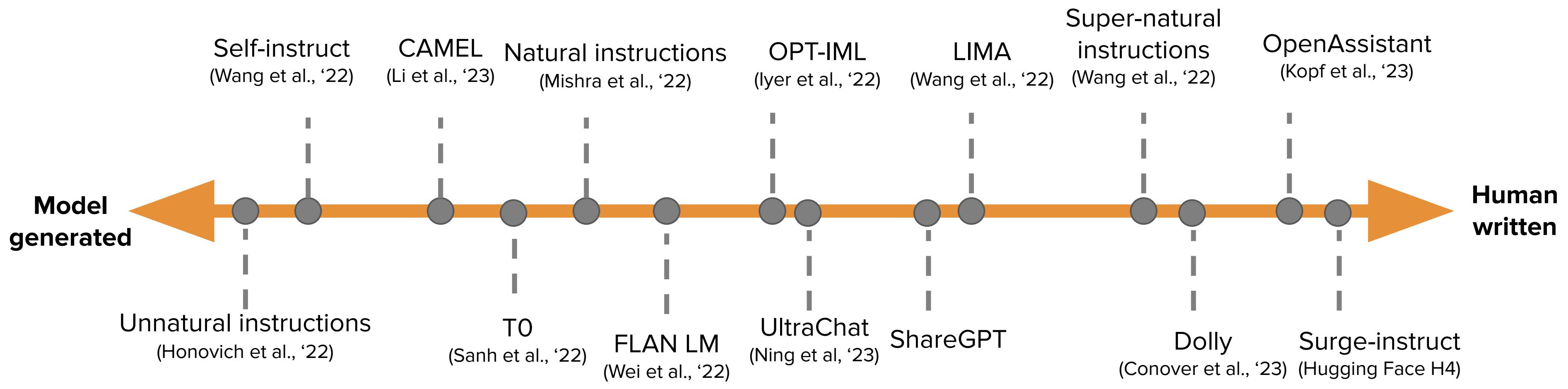


Feedback labeling

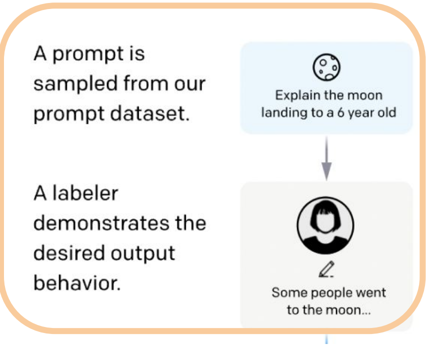
or



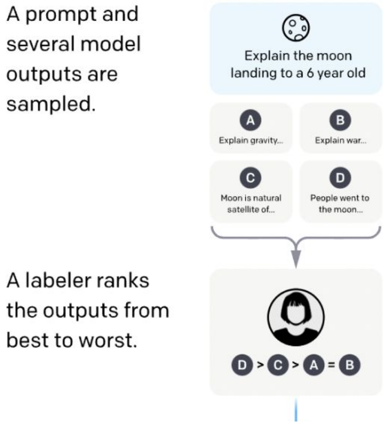
Instruction Tuning Datasets



Training a Chatbot

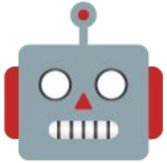


Instruction demonstrations

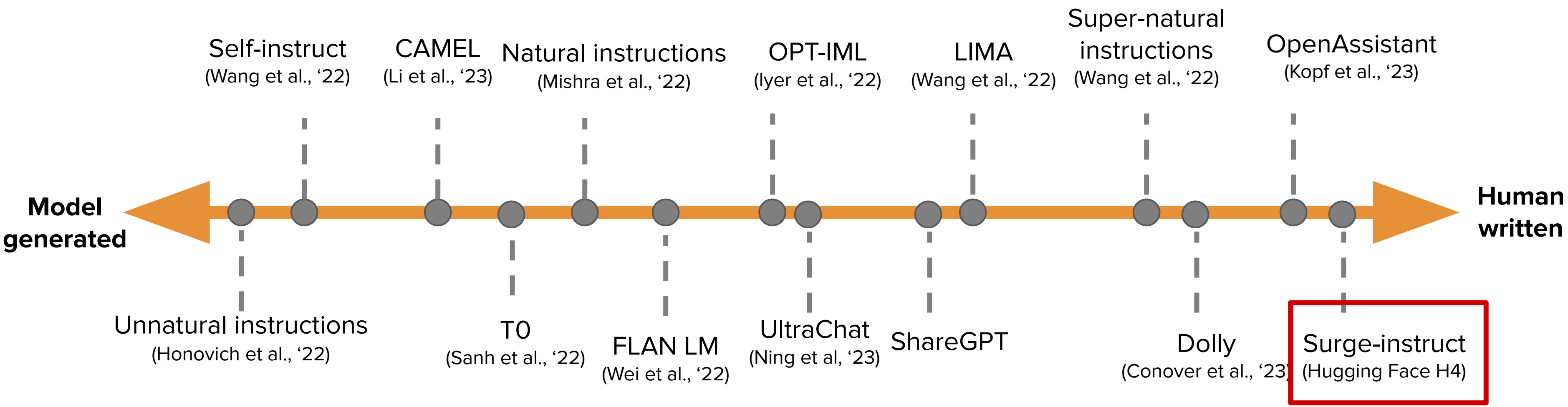


Feedback labeling

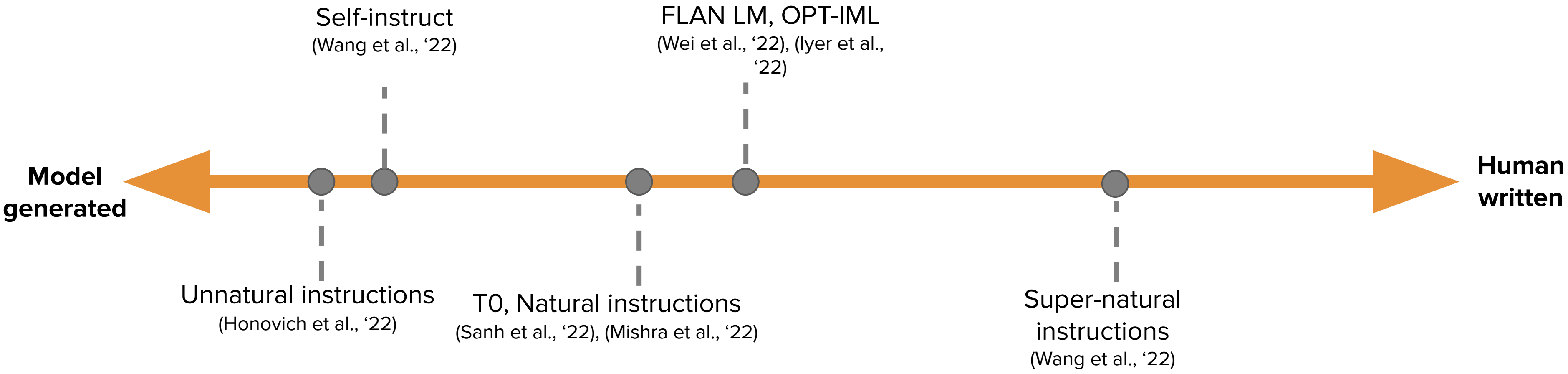
or



Instruction Tuning Dataset

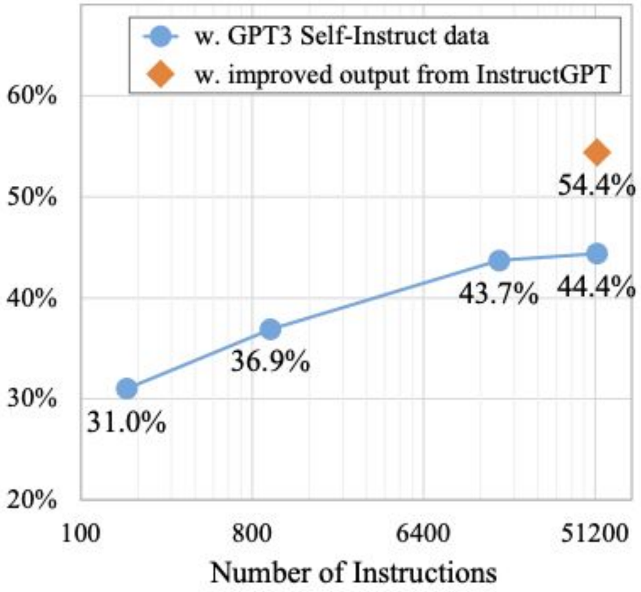


Instruction Tuning Dataset



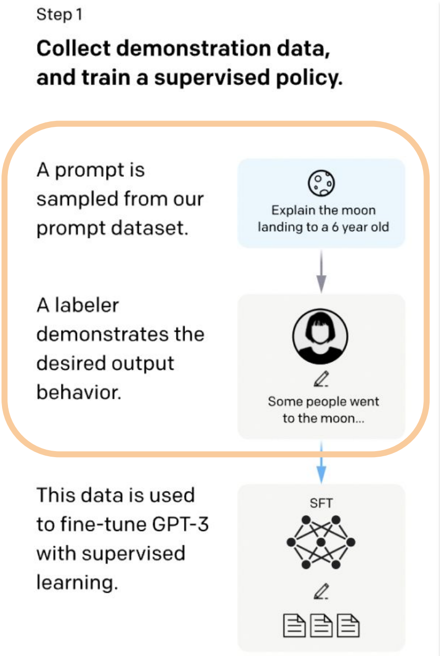
Past Findings from SFT Datasets

- 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

Instruction Tuning Dataset



Data specifications

1. Task distribution
2. Length distribution
3. Quality benchmarks

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: """

Ouyang et al., '22

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

Ouyang et al., ‘22

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

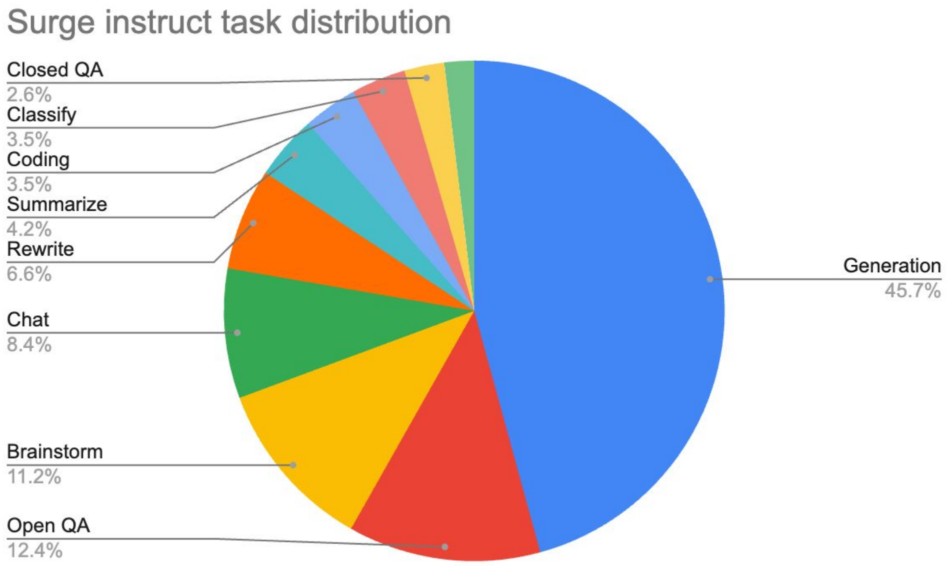
Length Distribution

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

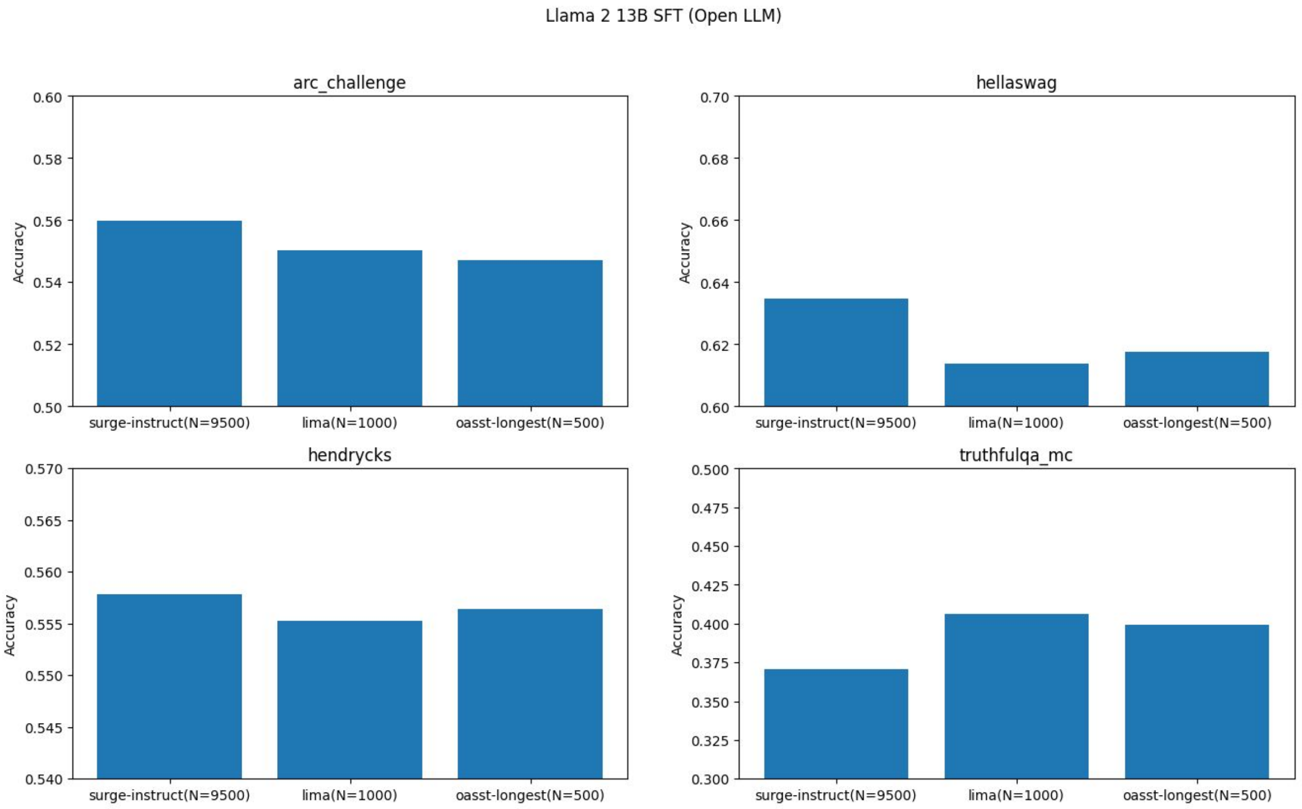
Race: primarily White, Black, Asian, Hispanic

Educational background: ranges from technical degree to PhD

Human Curation Results

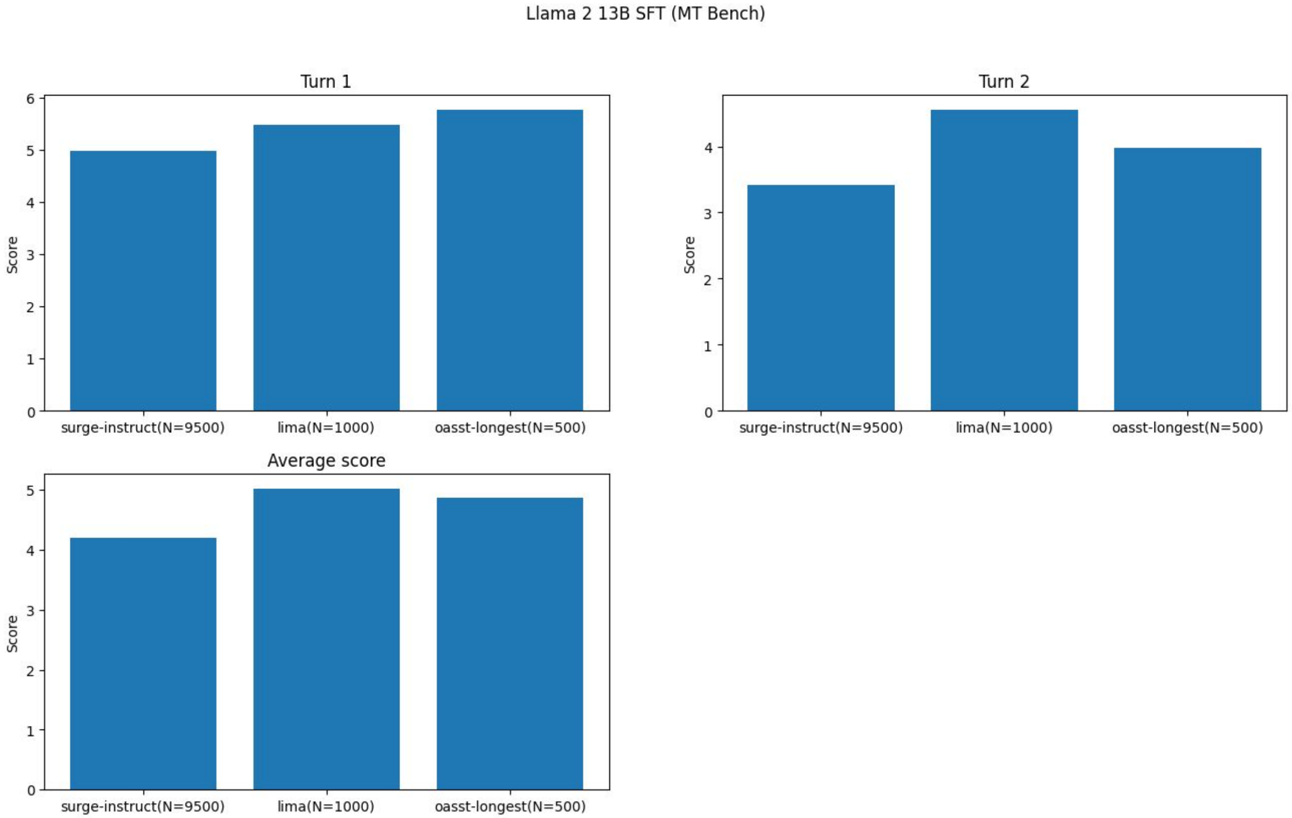
SFT Results – LLaMA 2 13B

Open LLM Leaderboard



SFT Results – LLaMA 2 13B

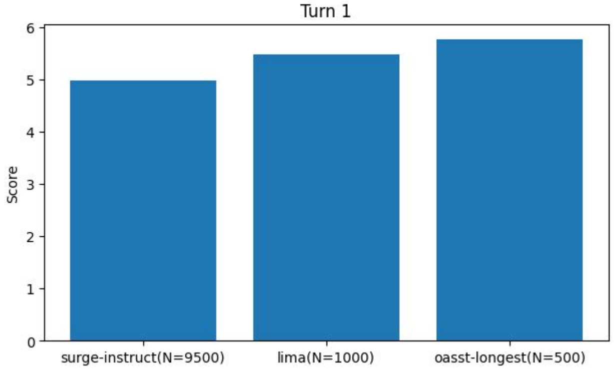
MT Bench Scores

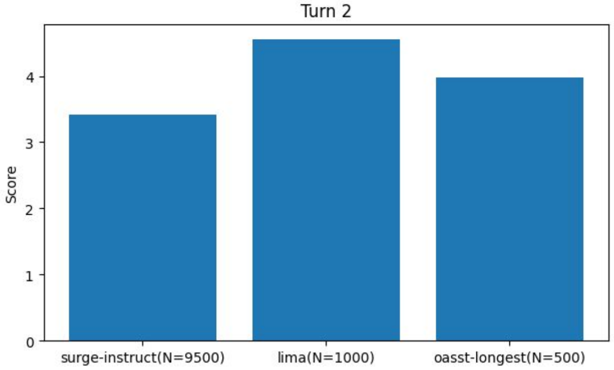


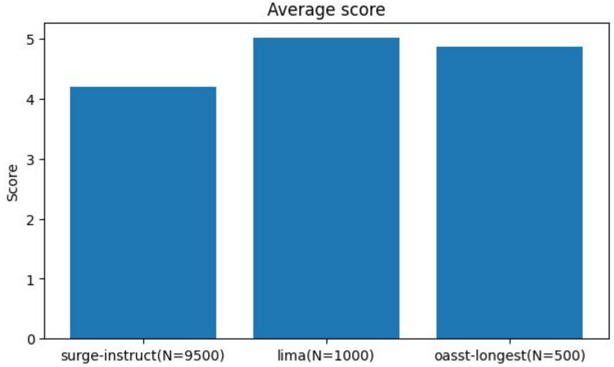
SFT Results – LLaMA 2 13B

MT Bench Scores

Llama 2 13B SFT (MT Bench)



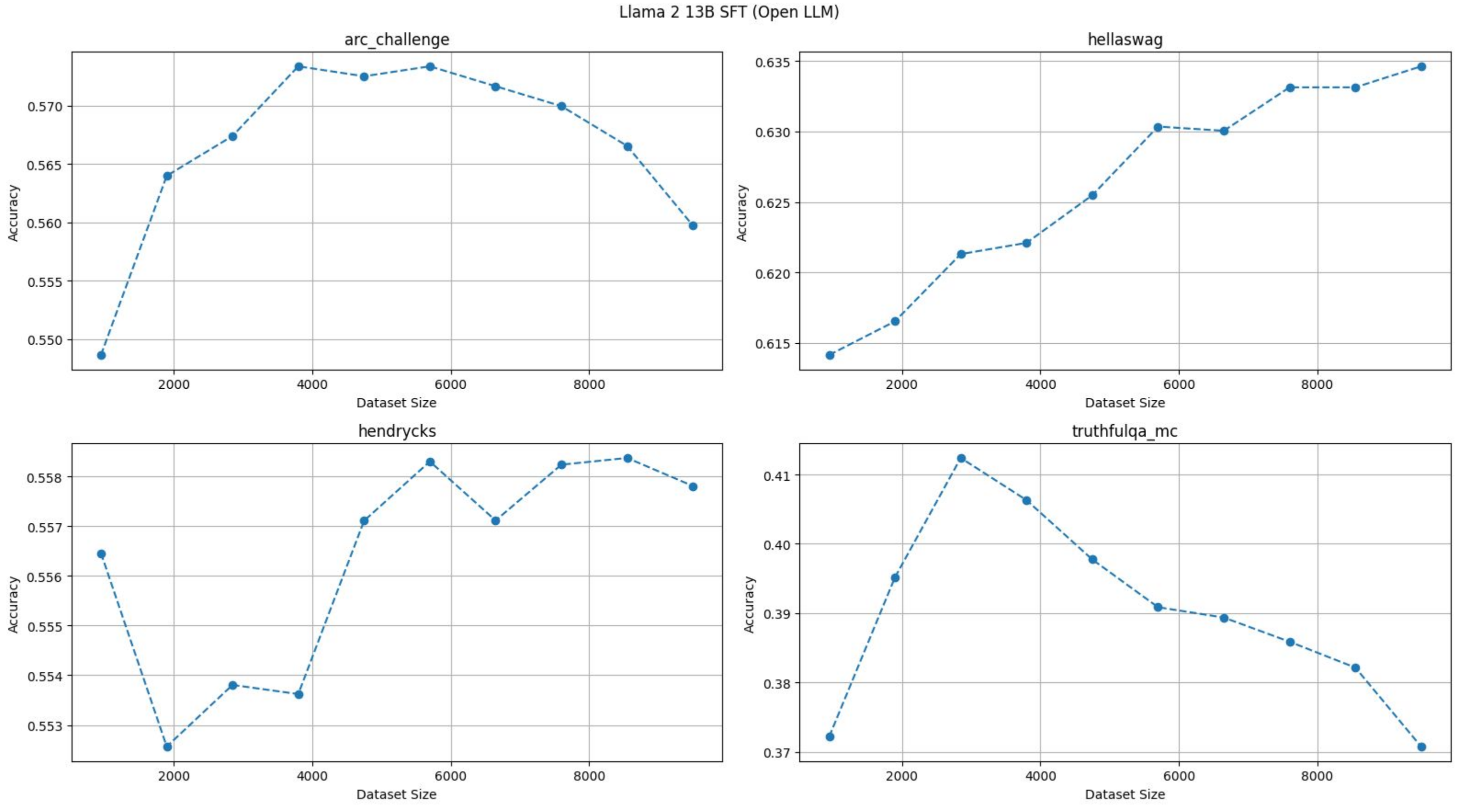




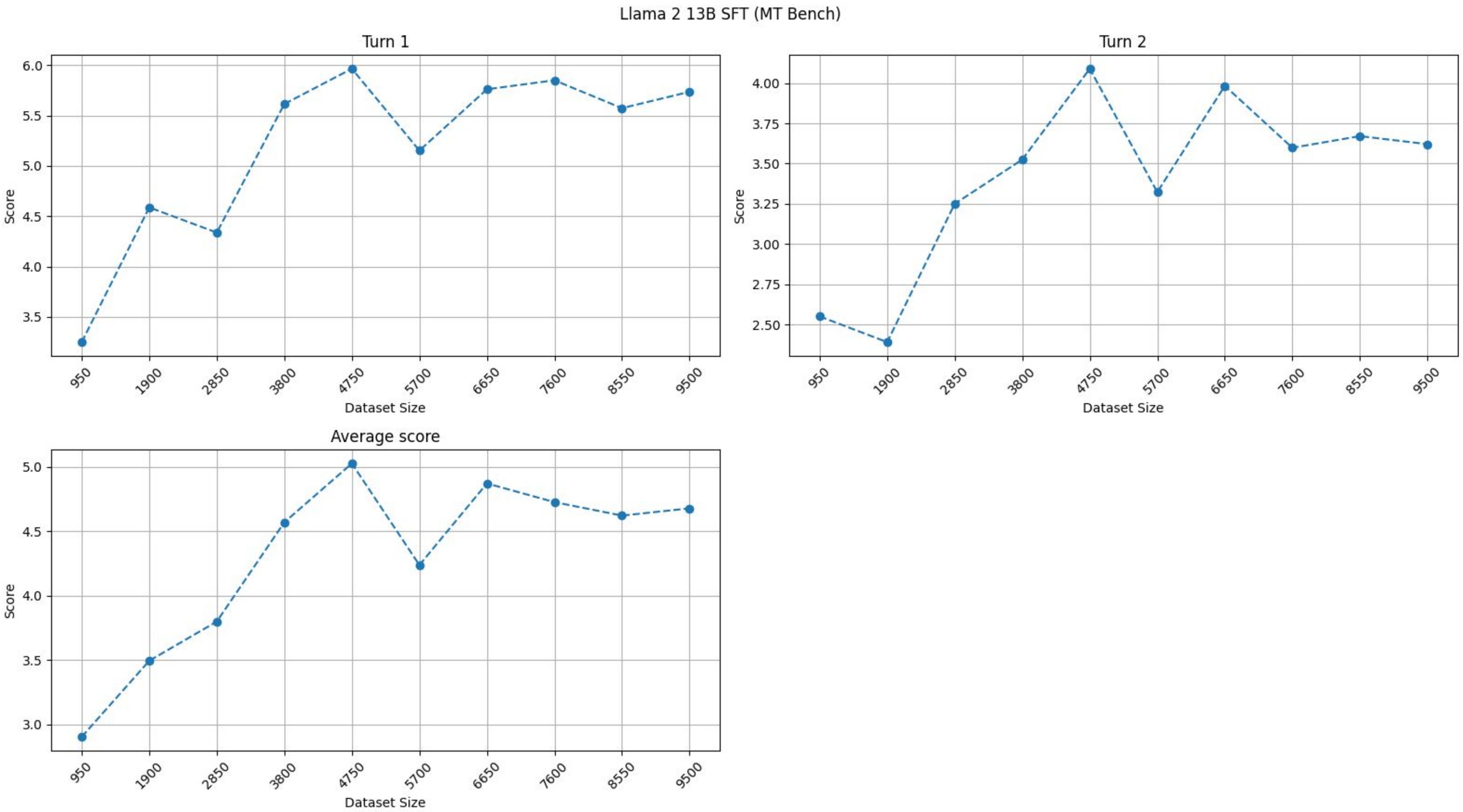
Dataset	**Avg. Length**
Surge-instruct	211
LIMA	482
OAsst	722

SFT Results – LLaMA 2 13B

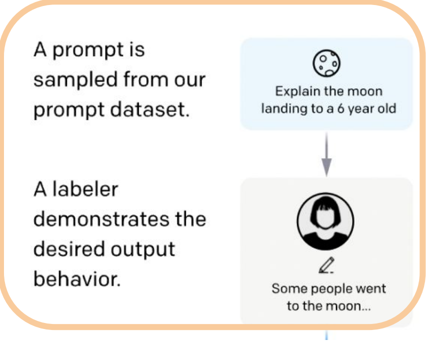
Performance vs. dataset size – ablations of surge-instruct dataset



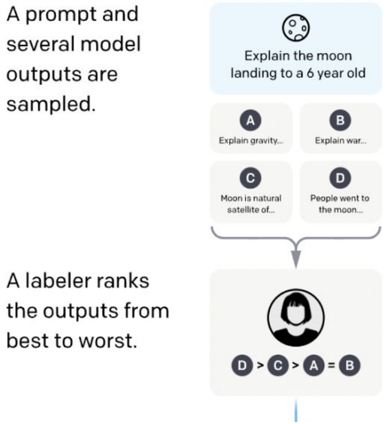
SFT Results – LLaMA 2 13B



Training a Chatbot

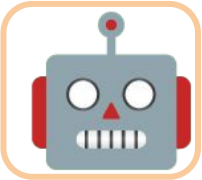


Instruction demonstrations



Feedback labeling

or



Instruction Tuning Datasets: UltraChat

1. Start with set of meta-topics or tasks

Meta topics of the Questions about the World sector

- Technology
- Health and wellness
- Travel and adventure
- Food and drink
- Art and culture
- Science and innovation
- Fashion and style
- Relationships and dating
- Sports and fitness
- Nature and the environment
- Music and entertainment
- Politics and current events
- Education and learning
- Money and finance
- Work and career
- Philosophy and ethics
- History and nostalgia
- Social media and communication
- Creativity and inspiration
- Personal growth and development
- Spirituality and faith
- Pop culture and trends
- Beauty and self-care
- Family and parenting
- Entrepreneurship and business
- Literature and writing
- Gaming and technology
- Mindfulness and meditation
- Diversity and inclusion
- Travel and culture exchange

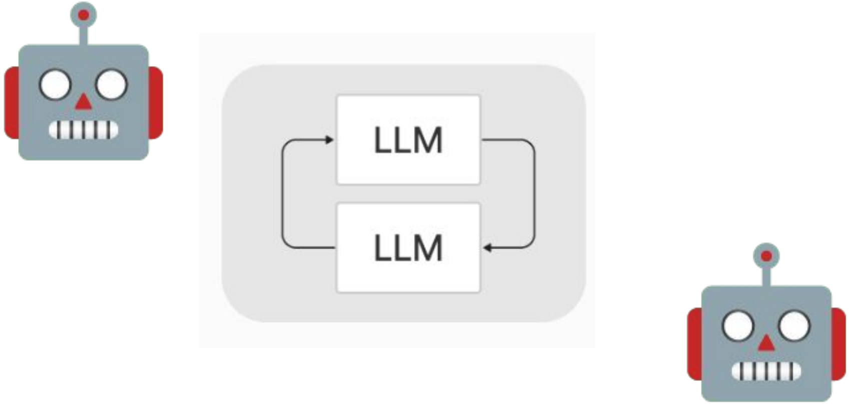
(Wang et al, 2022, Ding et al., 2023)

Instruction Tuning Datasets: UltraChat

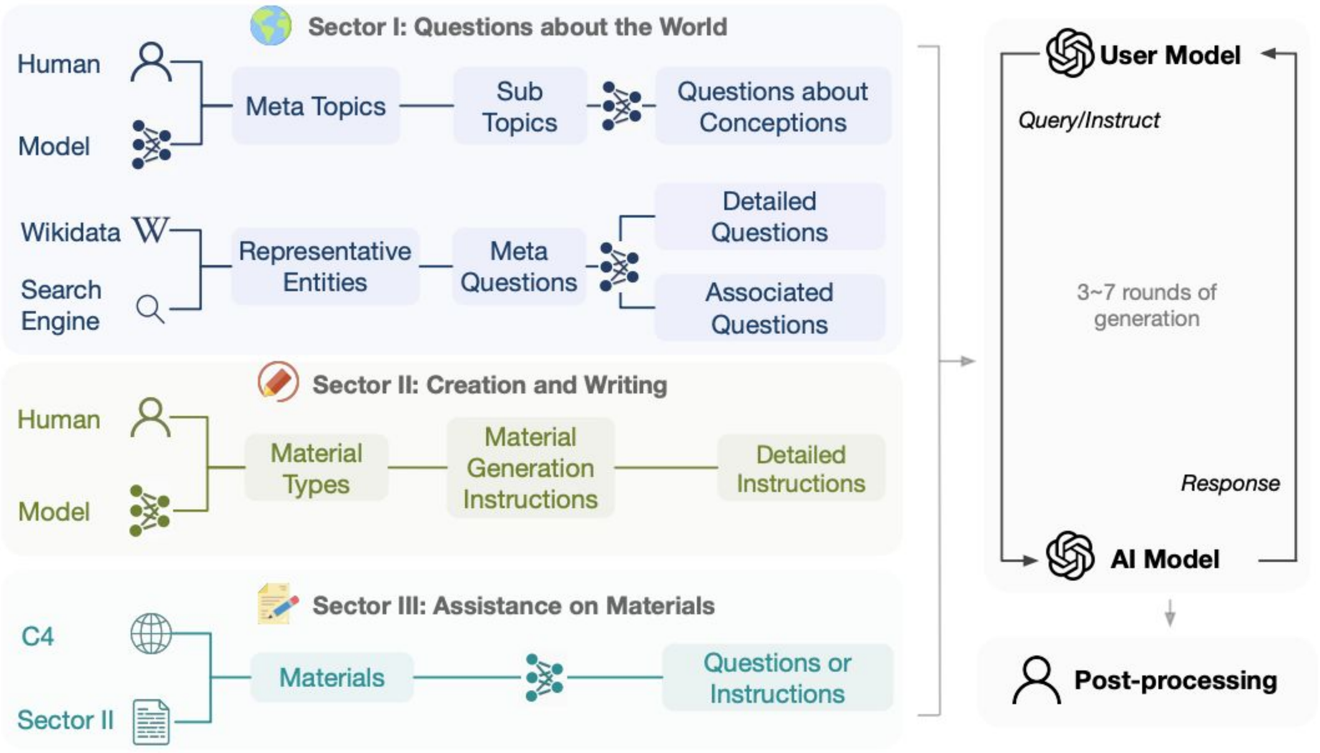
1. Start with set of meta-topics or tasks
2. Generate questions from a seed set

Instruction Tuning Datasets: UltraChat

1. Start with set of meta-topics or tasks
2. Generate questions from a seed set
3. Use these as “opening lines” in simulated dialogues



Self-Instruct Datasets: UltraChat



Ding et al., '23, Wang et al., 2022

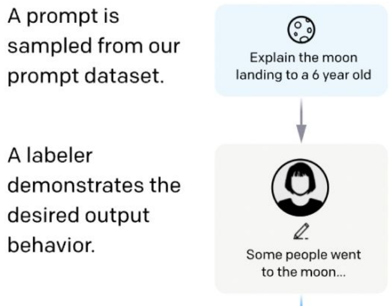
Self-Instruct Datasets: UltraChat

Direct Generation
What are some of the health benefits and risks of drinking coffee, and how can one maximize the former and minimize the latter?
There are several potential health benefits associated with drinking coffee, such as increased alertness, improved mood, and a reduced risk of certain diseases, including type 2 diabetes and liver cancer. However, there are also some potential risks, such as increased anxiety, disrupted sleep, and negative effects on the digestive system. To maximize the benefits and minimize the risks, it's important to consume coffee in moderation and pay attention to how it affects your body.

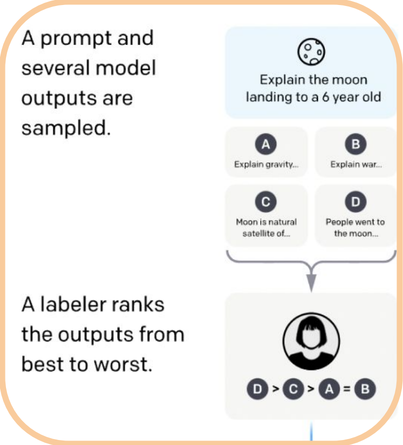
UltraChat
What are some of the health benefits and risks of drinking coffee, and how can one maximize the former and minimize the latter?
There are several health benefits and risks associated with drinking coffee. Here are some of them: Health benefits:1. Lower risk of type 2 diabetes: Studies have shown that consuming coffee can lower the risk of developing type 2 diabetes.2. Improved cognitive function: Coffee is a stimulant that helps improve cognitive function, including alertness, concentration, and memory.3. Lower risk of depression: Studies have shown that drinking coffee can reduce the risk of developing depression.4. Lower risk of certain cancers: According to research, coffee consumption can lower the risk of certain cancers, including liver and colorectal cancer.5. Reduced risk of heart disease: There is evidence that drinking moderate amounts of coffee can lower the risk of heart disease.Risks:1. Insomnia: Drinking too much coffee, especially late in the day, can cause insomnia.2. Increased risk of anxiety: Caffeine can increase anxiety levels in some people.3. Digestive issues: Drinking too much coffee can cause digestive problems like acid reflux, diarrhea, and ulcers.4. Increased heart rate: Caffeine can increase heart rate, which can be dangerous for people with pre-existing heart conditions.5. Addiction: Regular consumption of caffeine can lead to addiction.To maximize the health benefits and minimize the risks of drinking coffee, here are a few tips:1. Limit your caffeine intake to no more than 400 milligrams per day, or about 4 cups of coffee.2. Avoid drinking coffee late in the day, especially if you have trouble sleeping.3. Opt for organic, pesticide-free coffee to minimize exposure to toxins.4. Add cinnamon to your coffee instead of sugar to help regulate blood sugar levels.5. Drink your coffee black or with a small amount of milk to avoid consuming excessive sugar and calories.6. Consider switching to decaf coffee if you want to cut back on caffeine. 7. Avoid drinking coffee on an empty stomach as it can increase the risk of digestive problems.

Ding et al., '23

Training a Chatbot

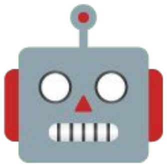


Instruction demonstrations

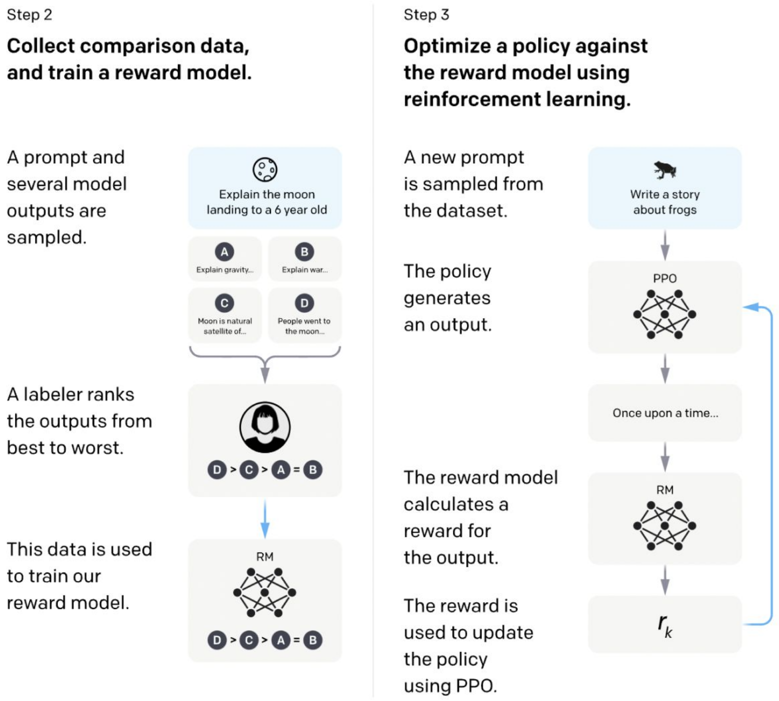


Feedback labeling

or



Feedback Dataset

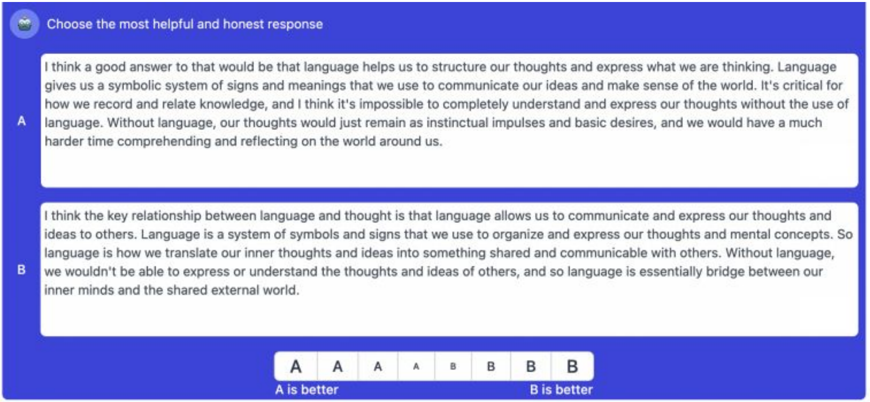


Data specifications

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

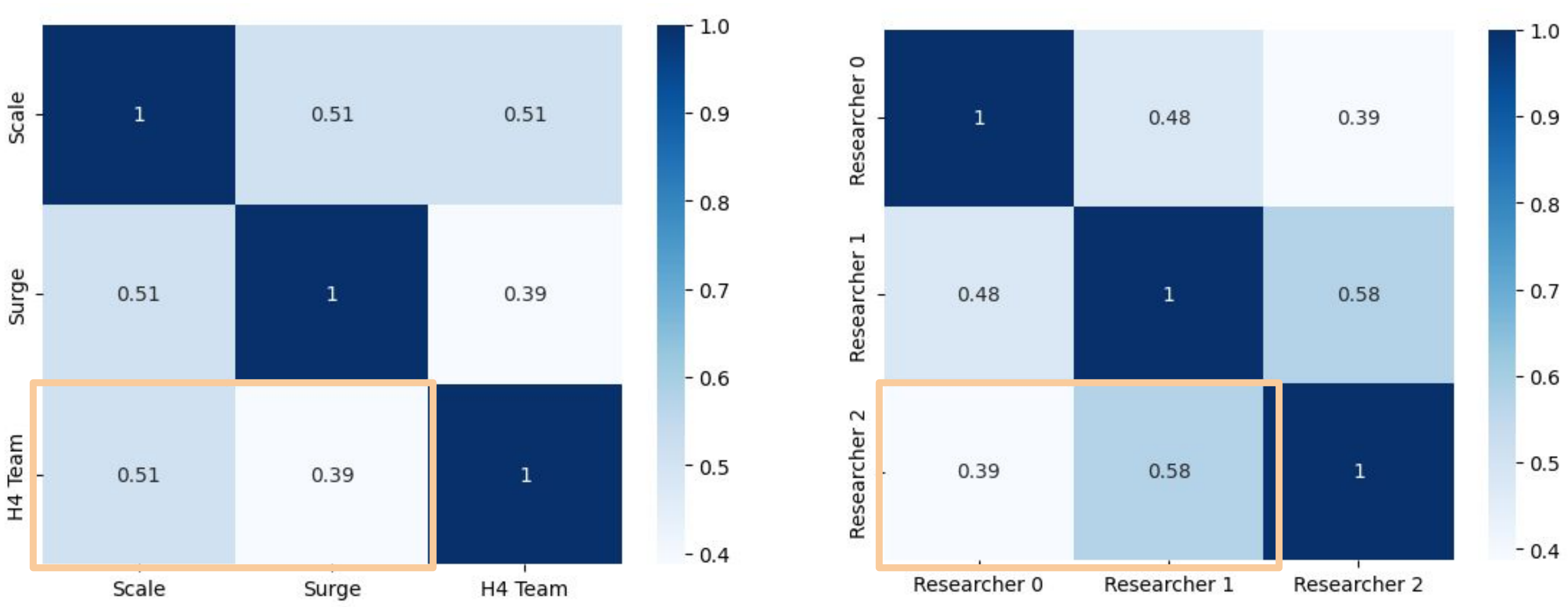
Pilot Study

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



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



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

    - "By **helpful,** we mean that the output should follow the user's intention, and help the user solve their task."
    - "By **honesty**, we mean that the output contains accurate information, and doesn't mislead the user."

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."

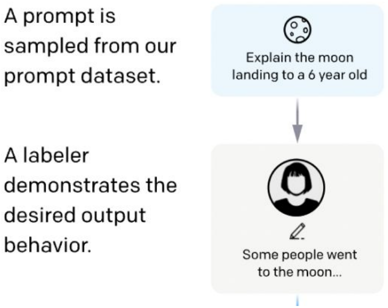
Response margin: 1

Feedback Dataset

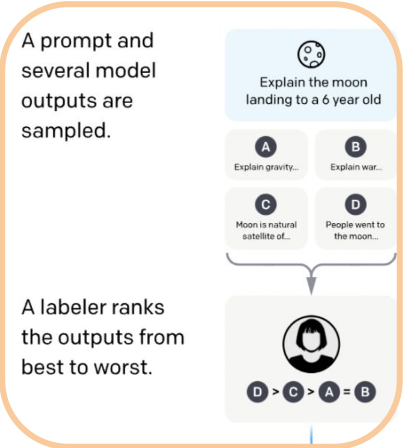
Issues:

- LLMs require large amounts of human feedback data after SFT
- Expensive and time consuming to gather the data. Eg: LLaMA-2: 💰💰💰

Training a Chatbot

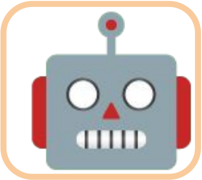


Instruction demonstrations

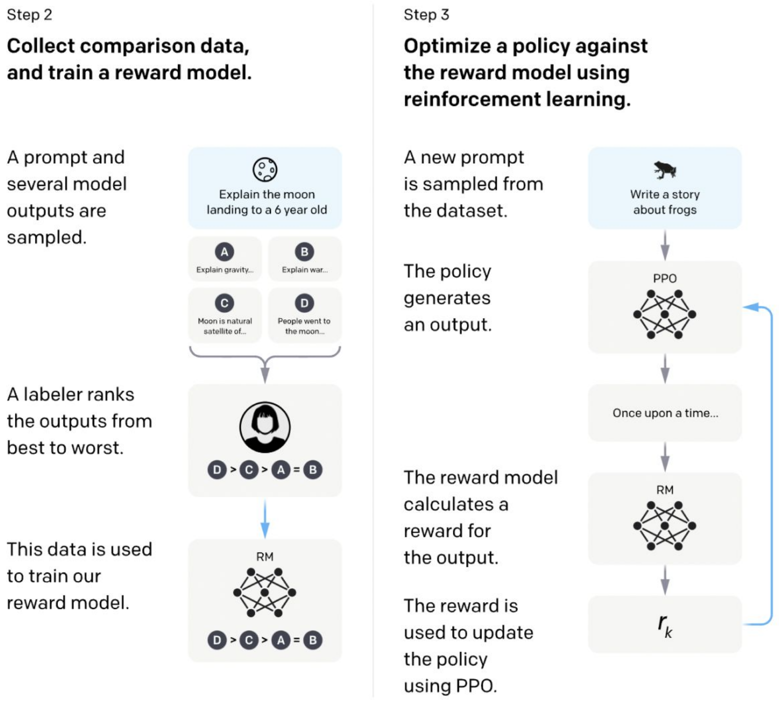


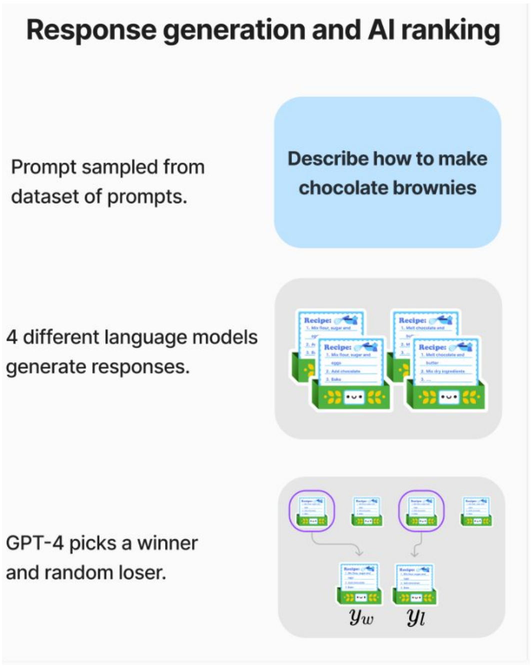
Feedback labeling

or



AI Feedback Dataset



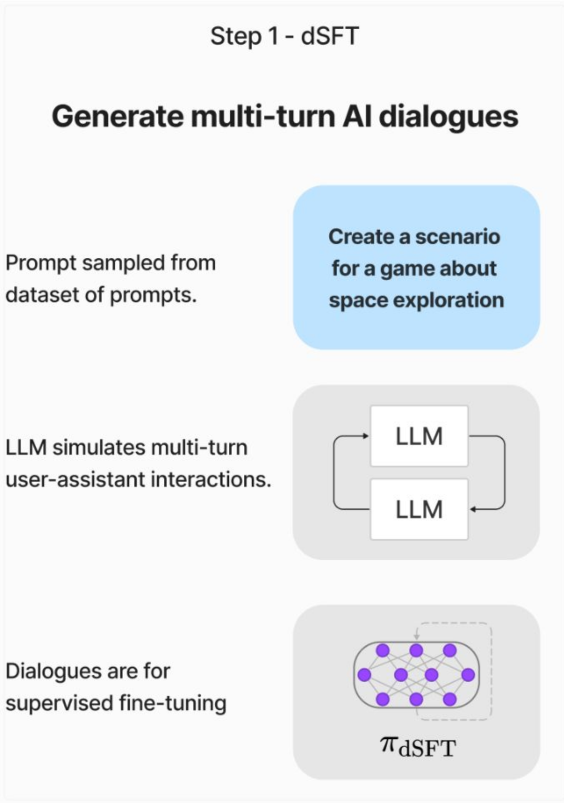


(Zephyr, 2023)

Zephyr-7B distillation

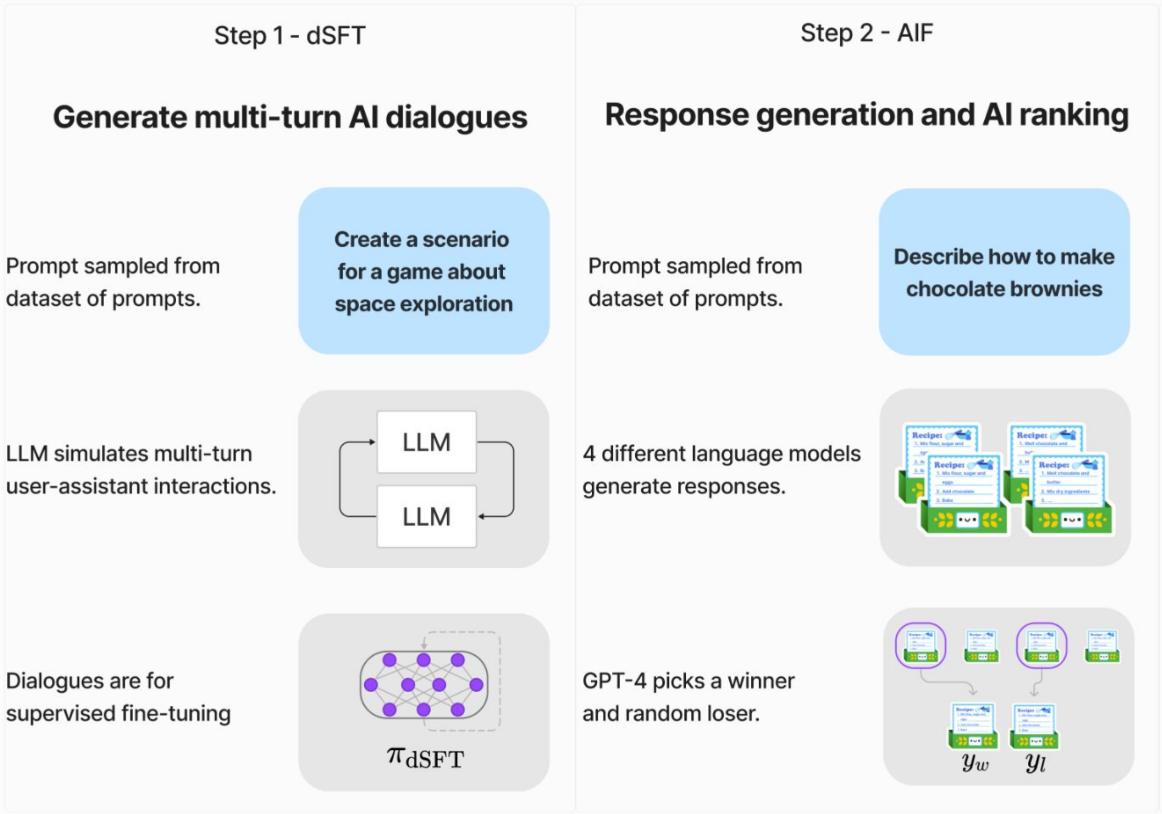
- Model based on Mistral-7B (Jiang et al., 2023) - best 7B model
- Trained using DPO distillation from AI Feedback
- Evaluation using LLM-as-evaluator methods and human judgement

Zephyr



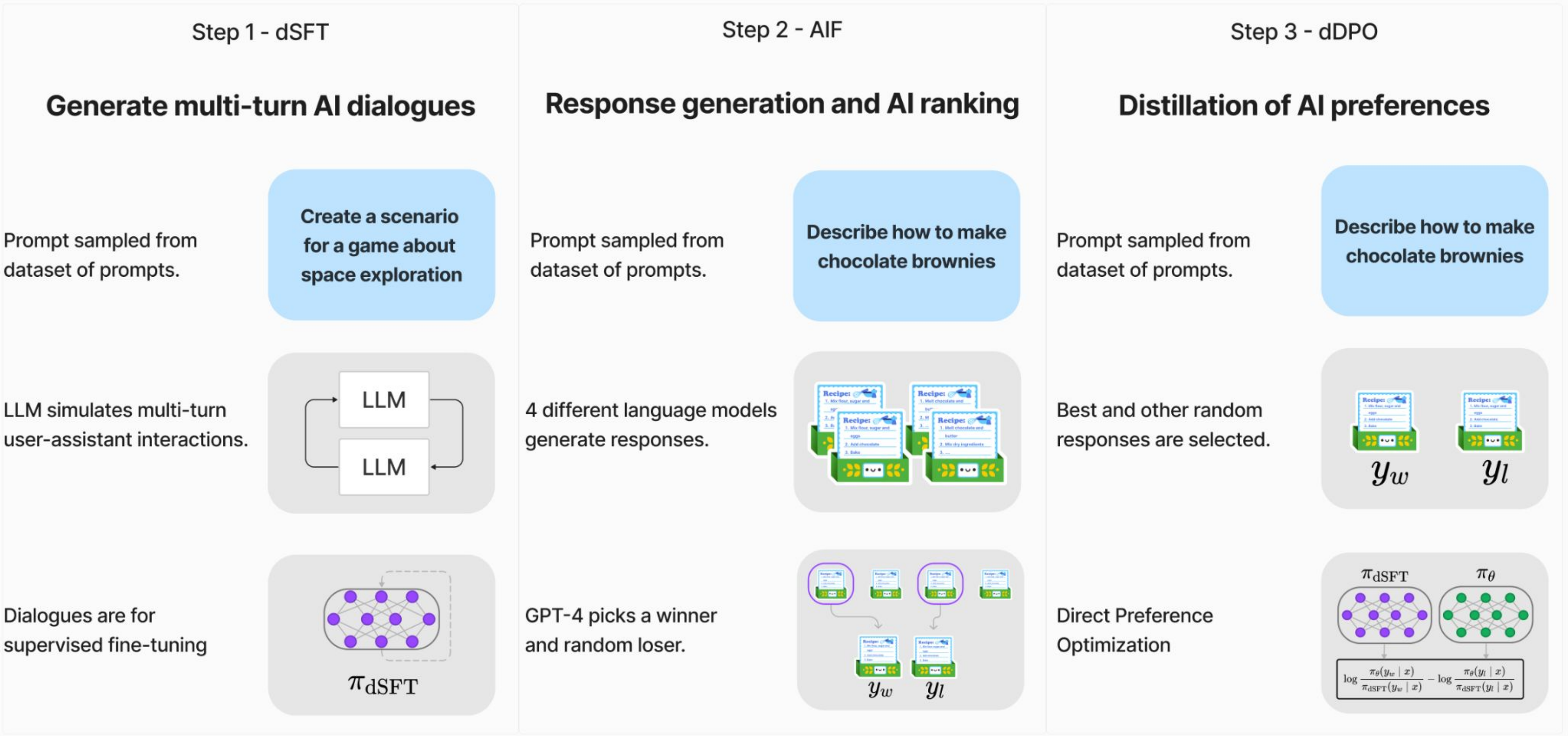
(Zephyr, 2023)

Zephyr 🪁



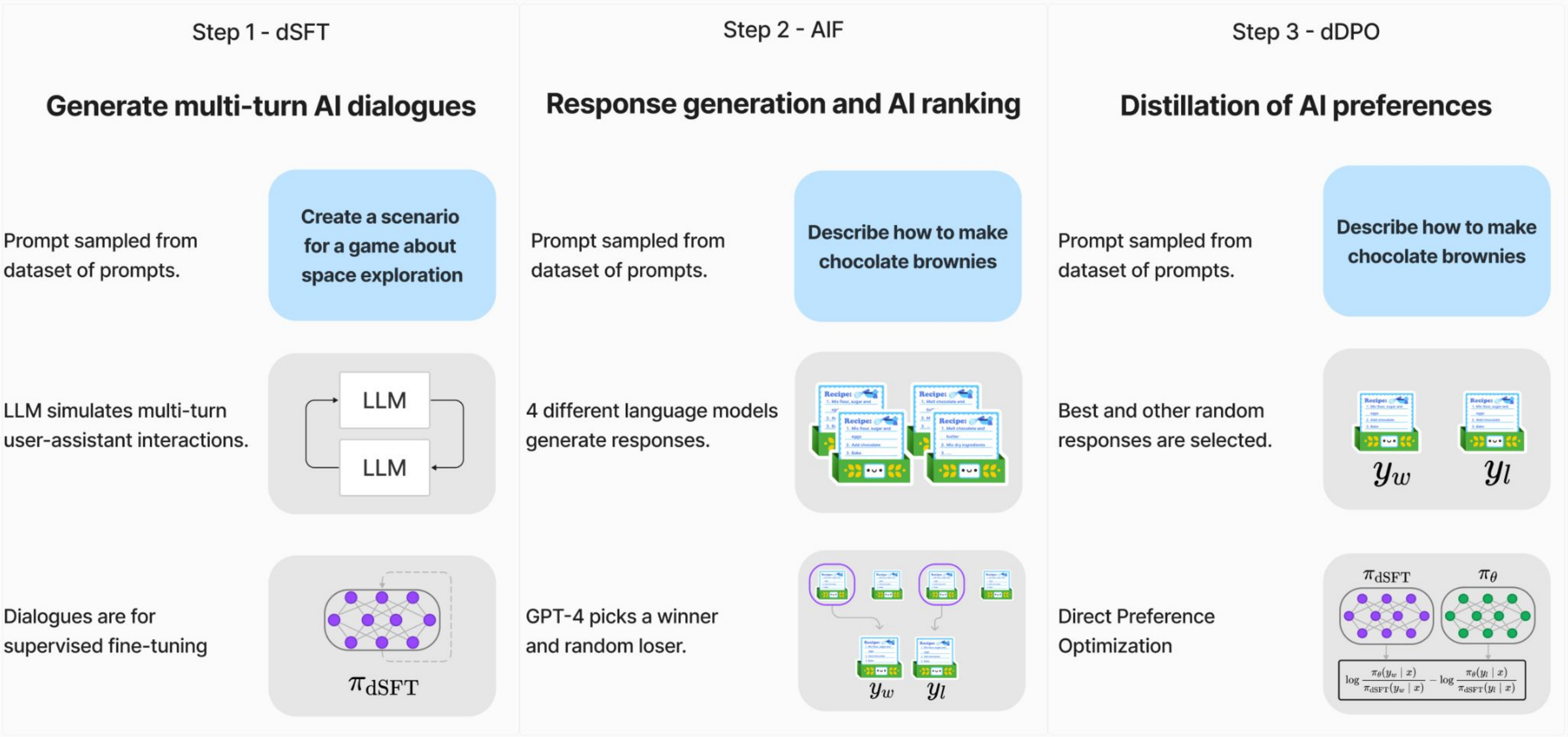
(Zephyr, 2023)

Zephyr



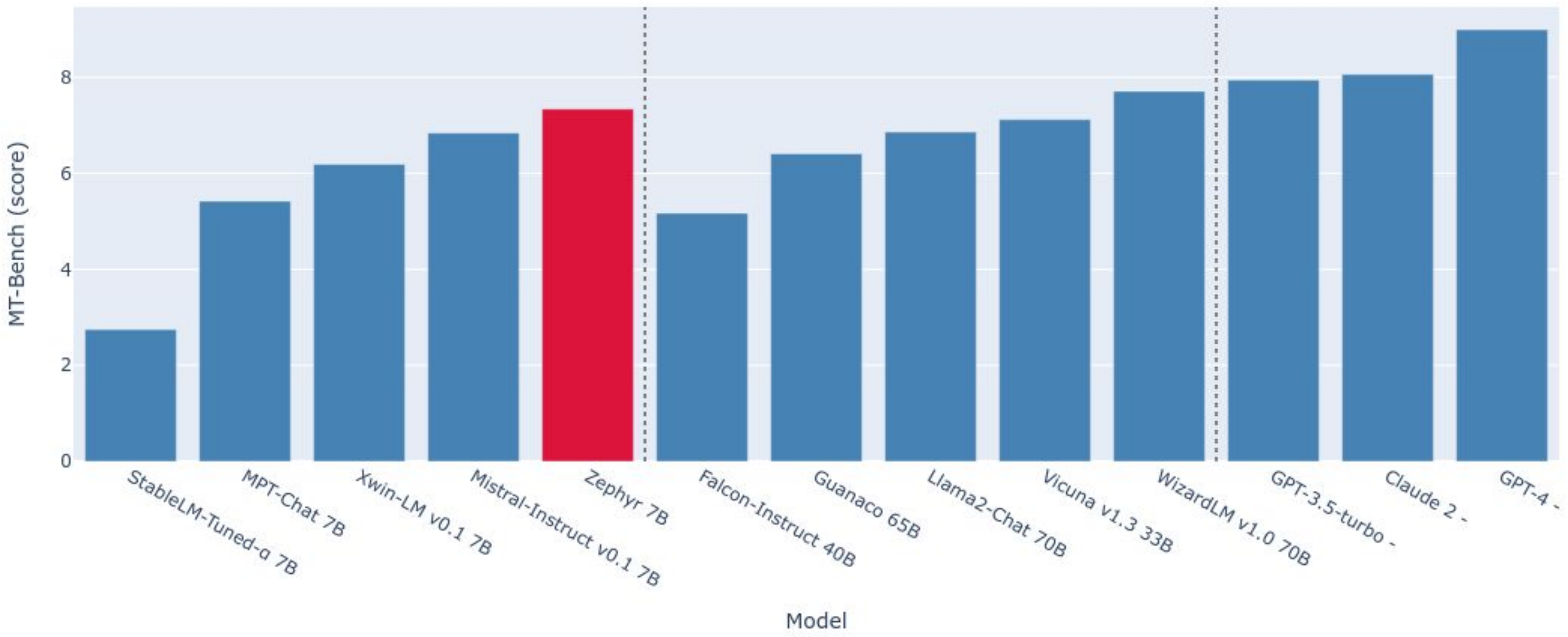
(Zephyr, 2023)

Zephyr



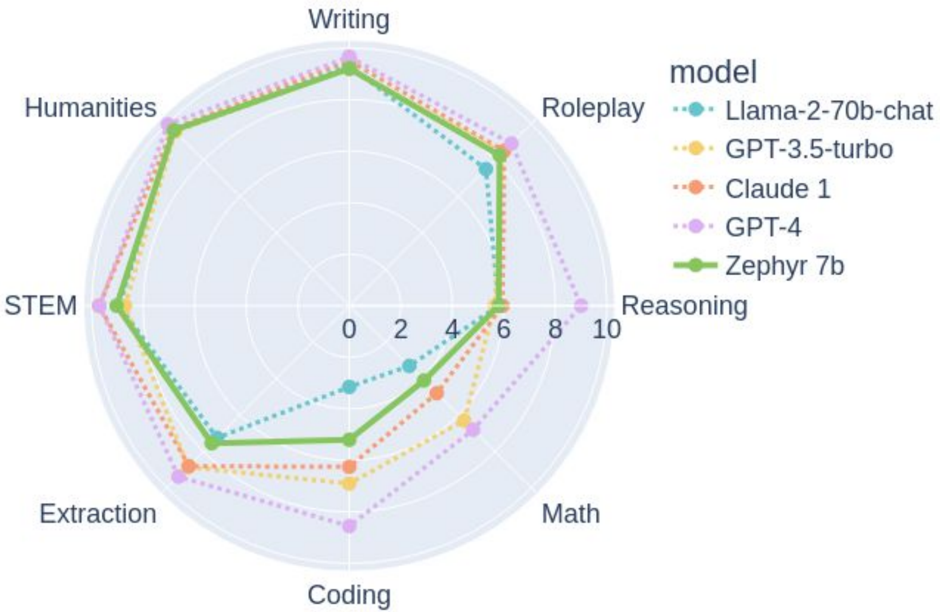
(Zephyr, 2023)

MT-Bench



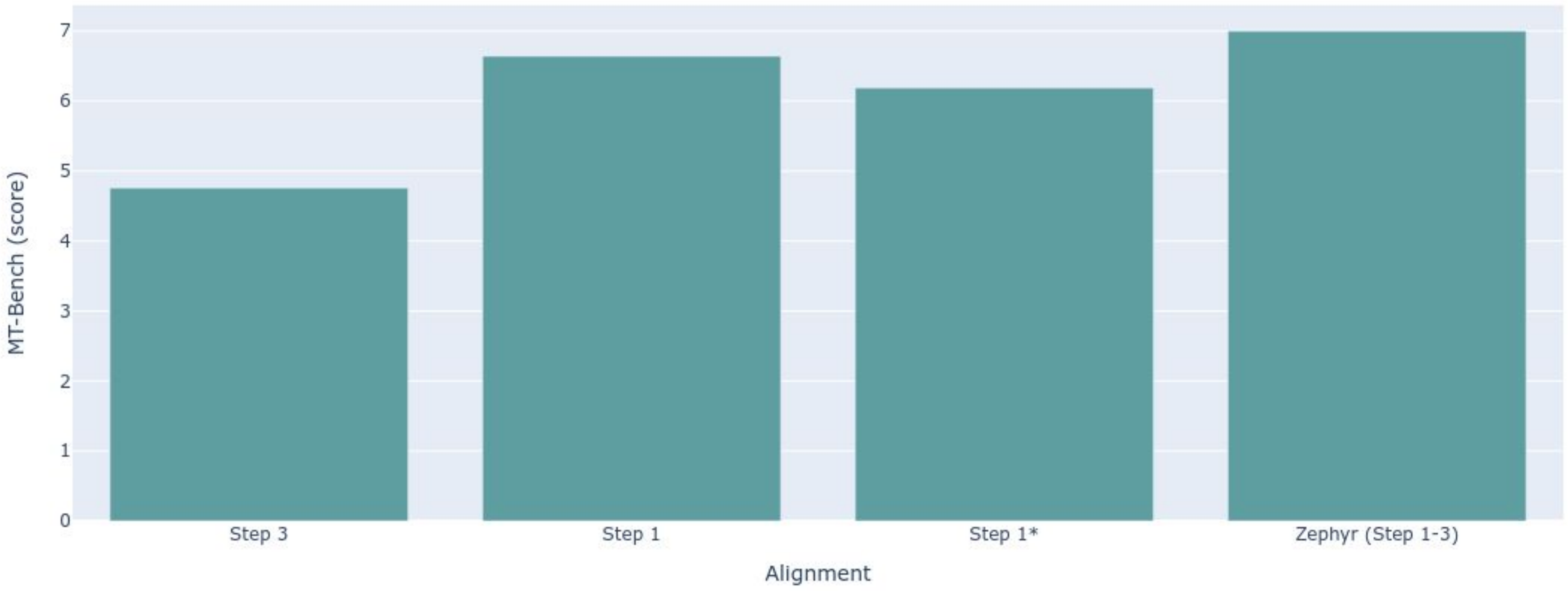
(Zheng et al., 2023)

MT-Bench by Domain

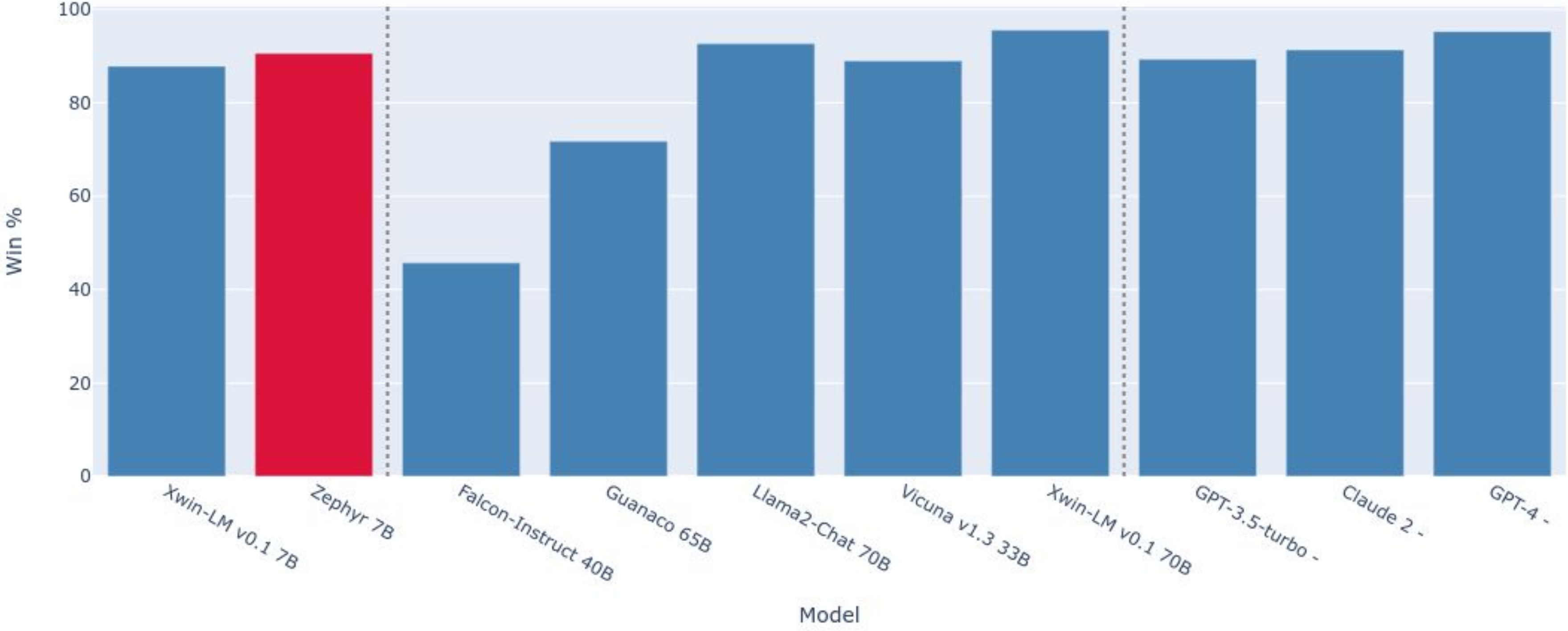


(Zheng et al., 2023)

Impact of Training on Feedback



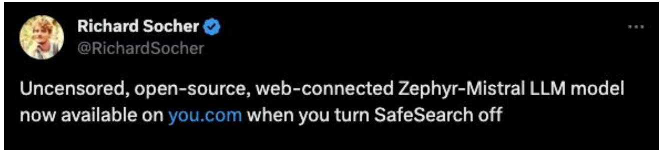
AlpacaEval



(Li et al., 2023)

Takeaways

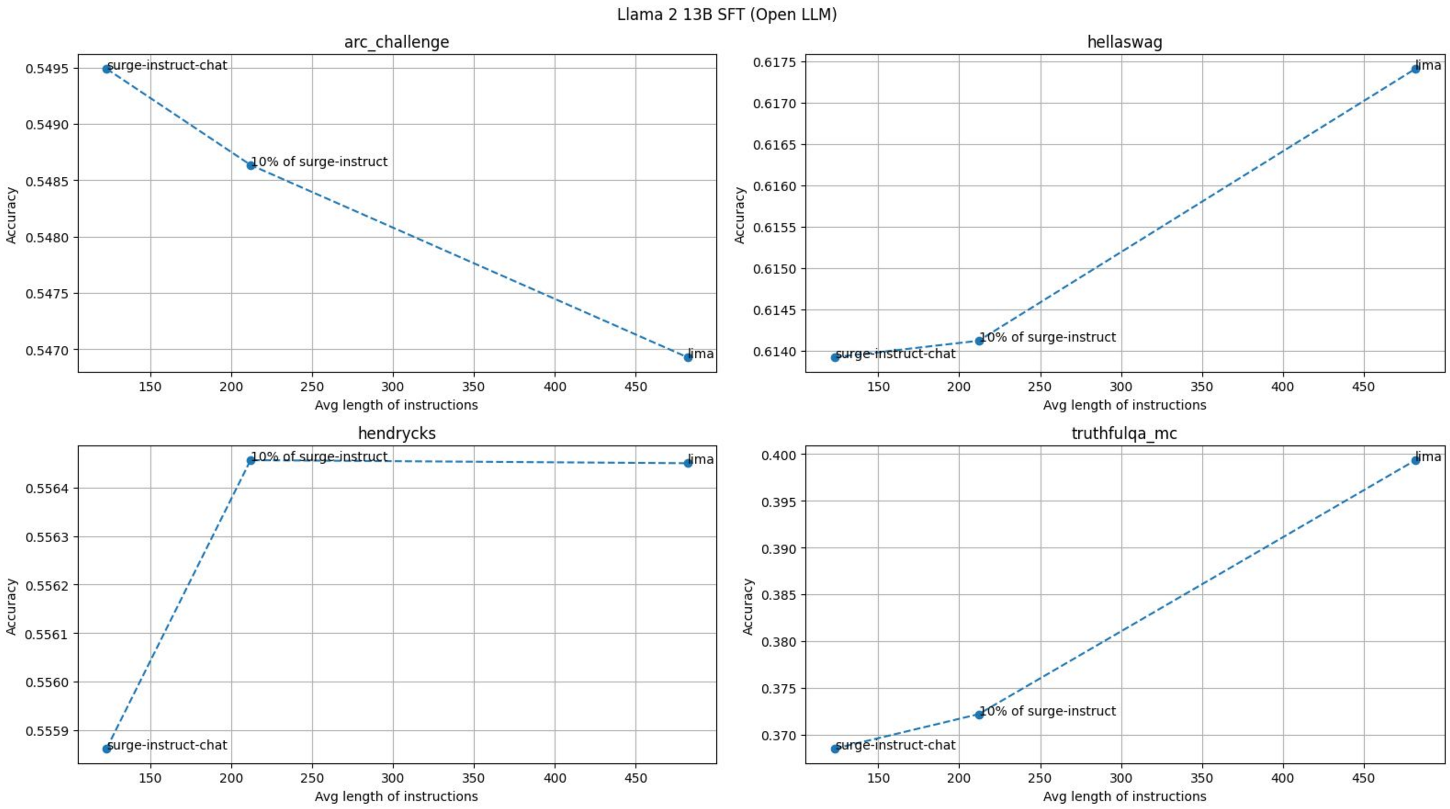
- Datasets for instruction following and feedback finetuning involve several critical factors
    - Amounts, length, tasks, and role of humans
- Major advances in the construction of synthetic AI instruction and feedback datasets – *quality* and *diversity* is key
- Strong FM is key for a usable chatbot
- 🧑‍💻Manual curation results –
    - TruthfulQA is the differentiating benchmark for ablation experiments
    - MT Bench scores are not always correlated with automated metrics
- 🤖AI Distillation results –
    - dSFT on AI generated data and dDPO on AI feedback data beats ChatGPT
- Research ➔ product in < 1 week



Backup slides

SFT Results – LLaMA 2 13B

Performance vs. avg prompt length



SFT Results – LLaMA 2 13B

MT Bench Scores