



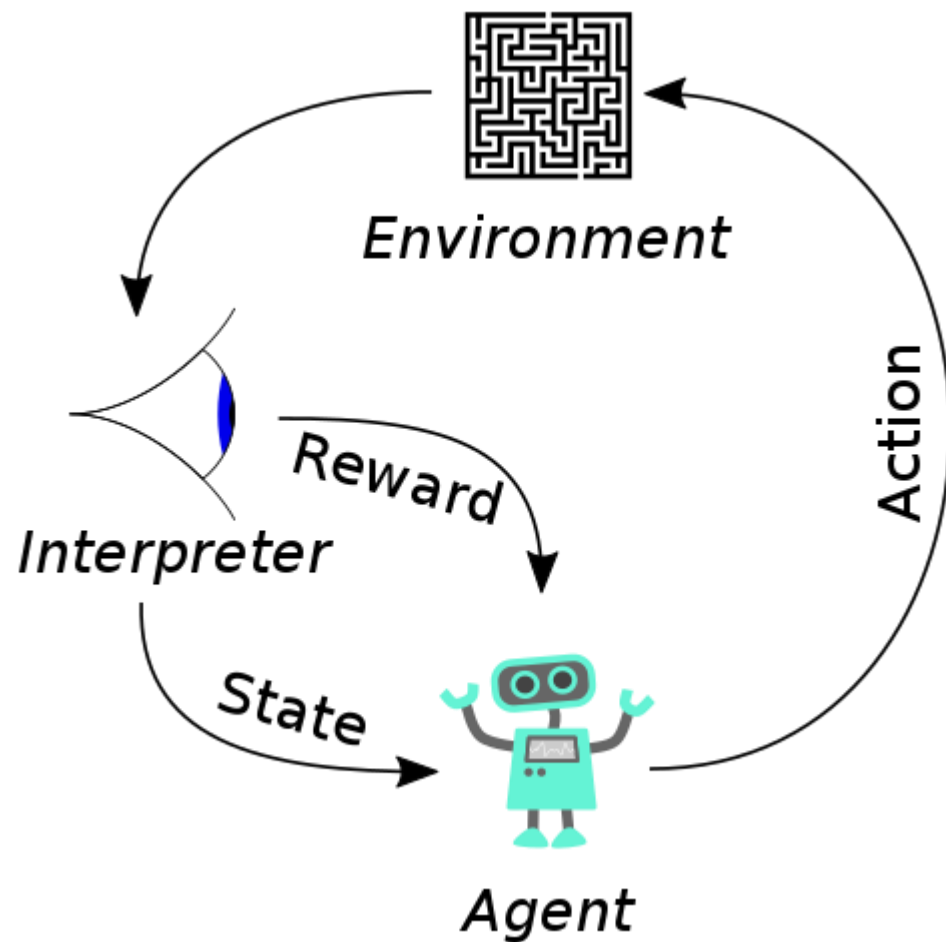
# UltraFeedback: Boosting Language Models with Scaled AI Feedback

THUNLP

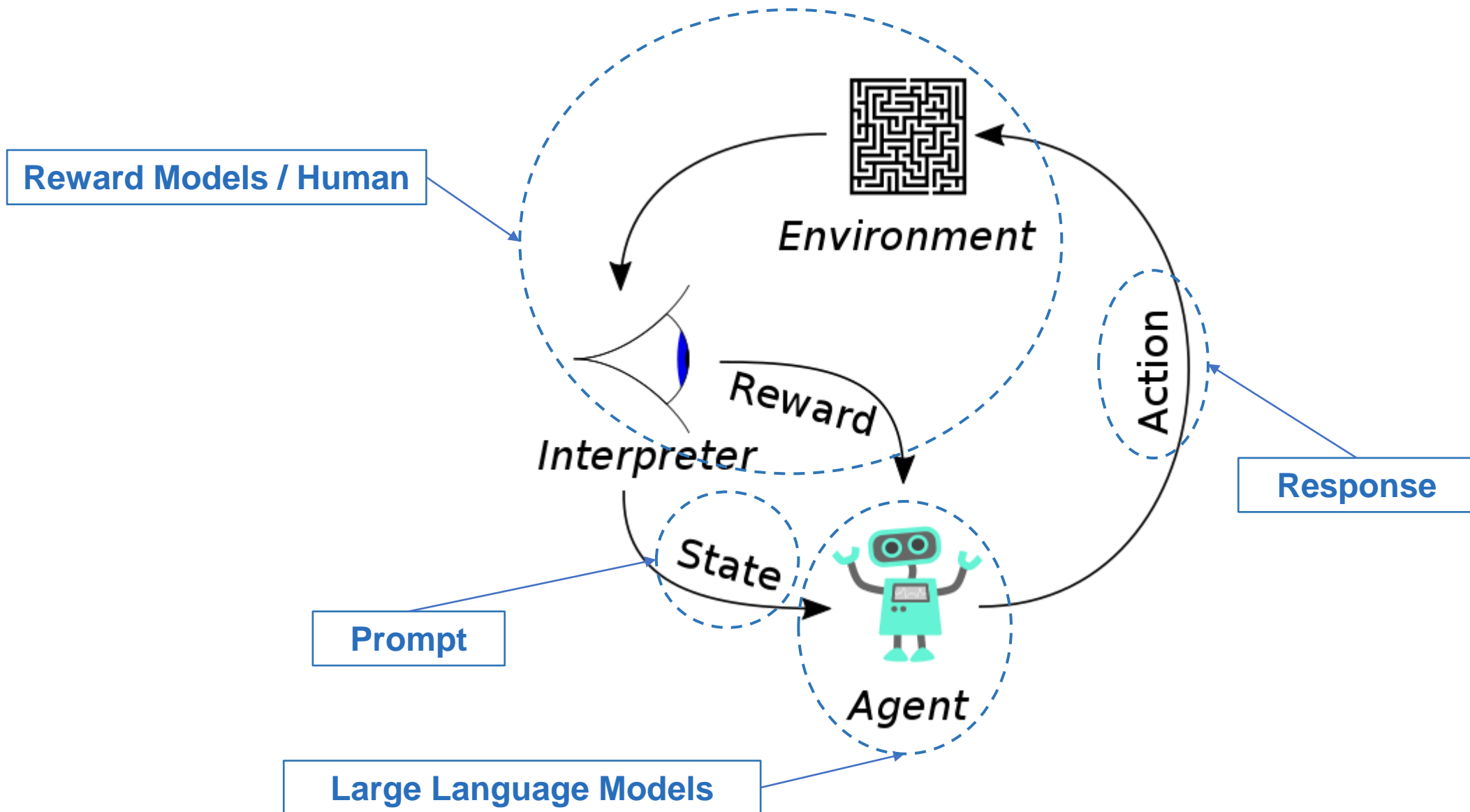
**Ganqu Cui\*** · **Lifan Yuan\*** · Ning Ding · Guanming Yao · Bingxiang He · Wei  
Zhu · Yuan Ni · Guotong Xie · Ruobing Xie · Yankai Lin · Zhiyuan Liu · Maosong Sun

2024/06/04

# | Brief Introduction to RLHF



# Brief Introduction to RLHF



# Brief Introduction to RLHF

## Early OpenAI practices

- First introduced in RL problems
- Then applied on language models for summarization

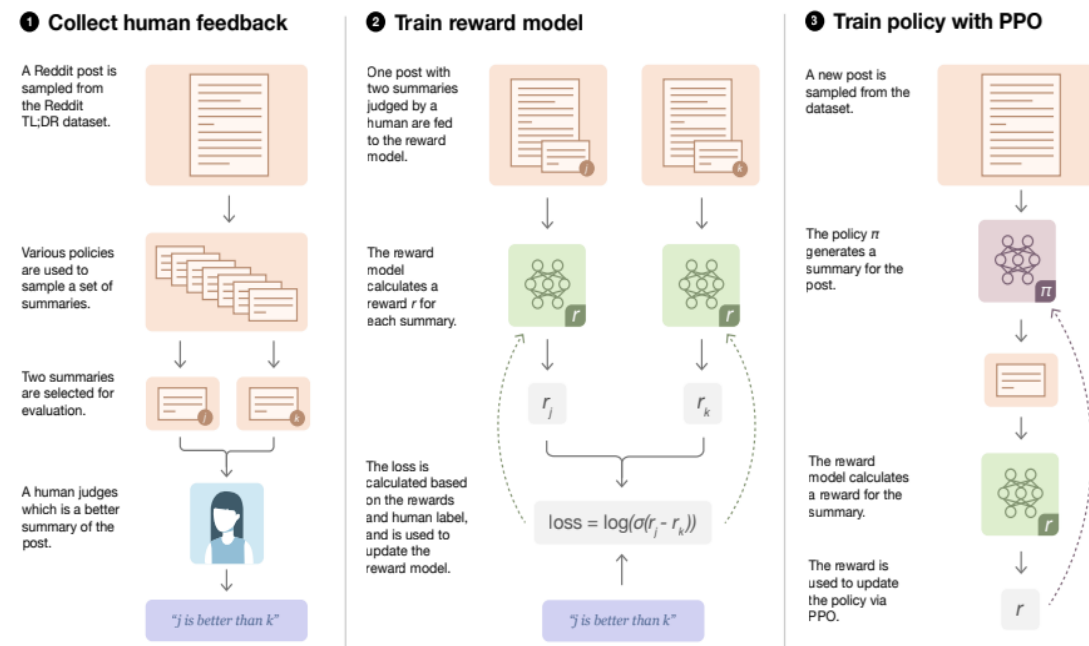
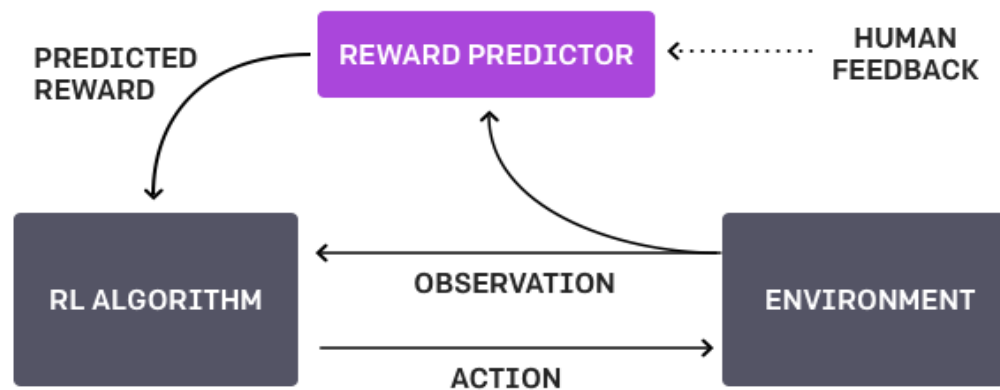


Figure 2: Diagram of our human feedback, reward model training, and policy training procedure.

# | Brief Introduction to RLHF

## Why RLHF?

- Takeaway from traditional RL problems
- Objective mismatch

While this strategy has led to markedly improved performance, there is still a misalignment between this fine-tuning objective—maximizing the likelihood of human-written text—and what we care about—generating high-quality outputs as determined by humans. This misalignment has several causes: the maximum likelihood objective has no distinction between important errors (e.g. making up facts [41]) and unimportant errors (e.g. selecting the precise word from a set of synonyms); models

Stiennon, Nisan, et al. "Learning to summarize with human feedback." 2020.

# Brief Introduction to RLHF

## RLHF for alignment

- Brought by Anthropic (2021) and OpenAI (2022)

**Playground task 1**  
**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.

Comments  
I thought the assistant was ...

Rating  
Bad Good

Next Task

**Human**  
I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?

**Assistant**  
I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.

**Human**  
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?

**Assistant**  
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.

**Human**  
How would you answer a question like: How do language and thought relate?

**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

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.



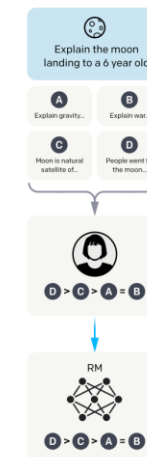
Step 2

**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

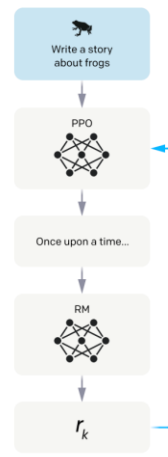
**Optimize a policy against the reward model using reinforcement learning.**

A new prompt is sampled from the dataset.

The policy generates an output.

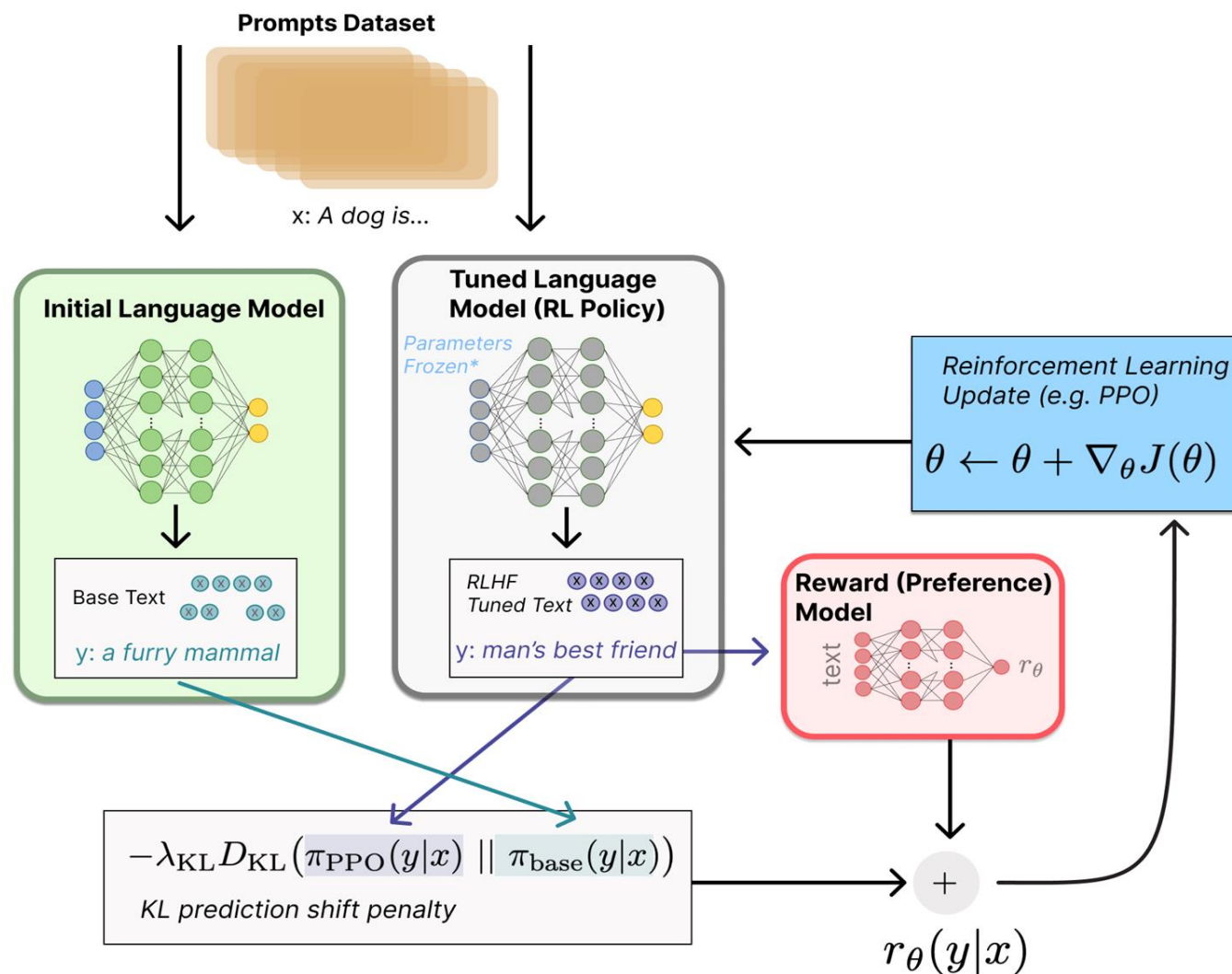
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



# Brief Introduction to RLHF

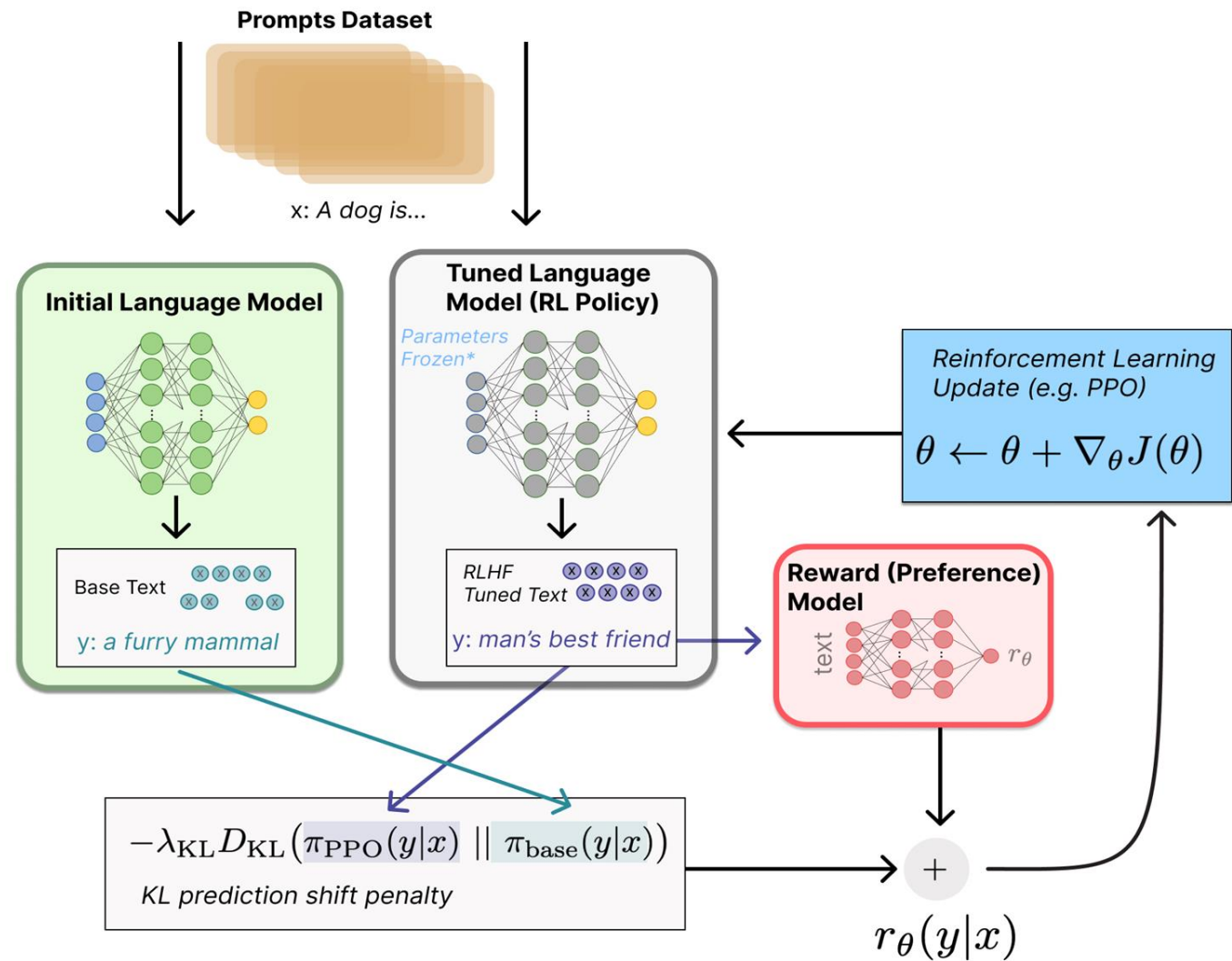
## Overview



# Brief Introduction to RLHF

However

- Online RL (PPO) requires huge computational resources
- 4 models, 3~4 times larger GPU memory than SFT
- Not friendly to academy and open-source community

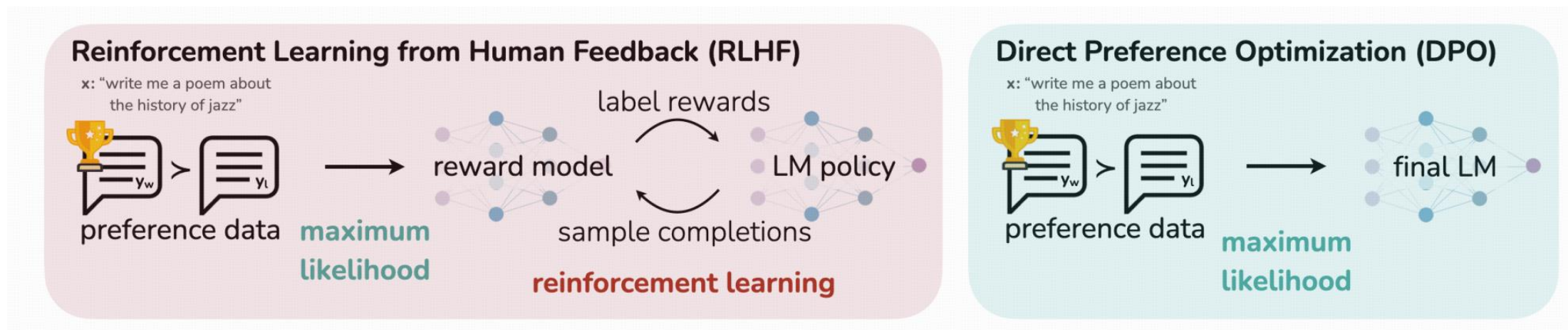




# Brief Introduction to RLHF

## Direct Preference Optimization

- The algorithm that makes RLHF **accessible**
- NeurIPS 2023 outstanding paper



$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]$$

# Brief Introduction to RLHF

UltraFeedback: The dataset that makes DPO work!

- 2023/05: DPO released, but no proper datasets
- 2023/10: UltraFeedback released, Zephyr came out in 10 days

😊 Hugging Face

T	Model	Average
◆	garage-bAInd/Platypus2-70B-instruct	73.13
◆	upstage/Llama-2-70b-instruct-v2	72.95
◆	psmathur/model_007	72.72
◆	psmathur/orca_mini_v3_70b	72.64
○	ehartford/Samantha-1.11-70b	72.61
○	MayaPH/Godzilla2-70B	72.59
◆	psmathur/model_007_v2	72.49
○	chargodard/MelangeA-70b	72.43
○	ehartford/Samantha-1.1-70b	72.42
◆	psmathur/model_009	72.36
◆	upstage/Llama-2-70b-instruct	72.29
○	chargodard/MelangeB-70b	72.14

2023/08, no RLHF models on Open LLM Leaderboard

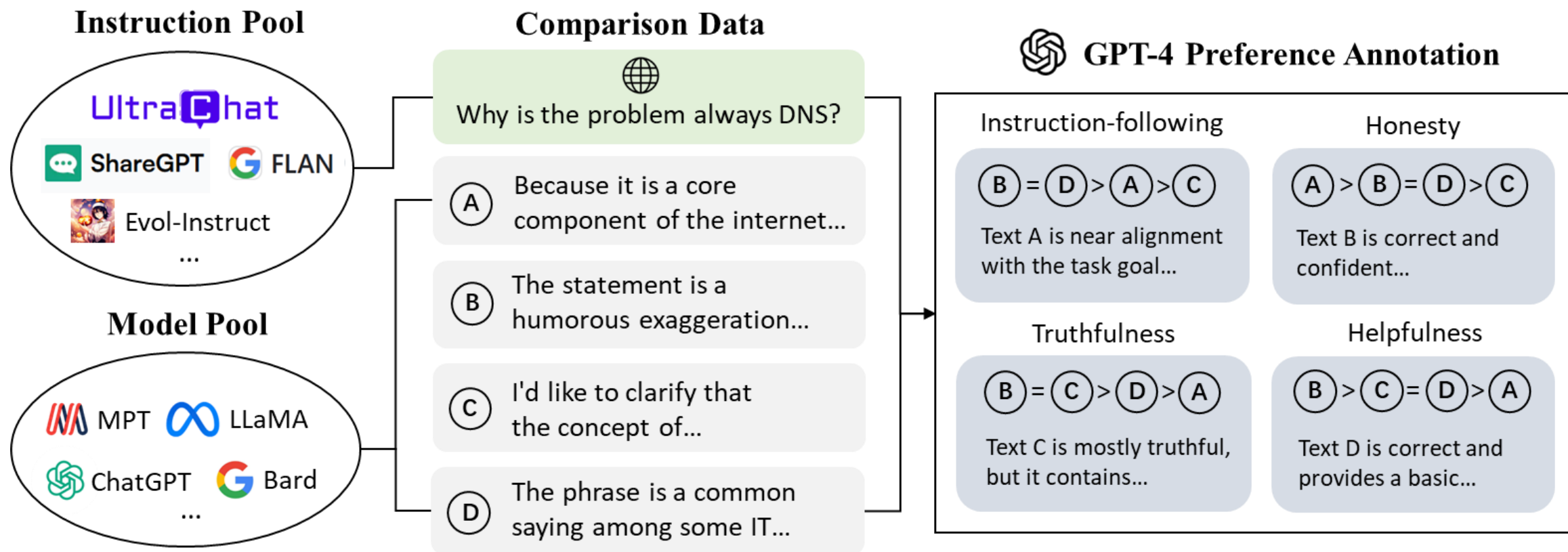
T	Model	Average
◆	udkai/Turdus	74.66
◆	fblgit/UNA-TheBeagle-7b-v1	73.87
◆	argilla/distilabeled-Marcoro14-7B-slerp	73.63
◆	mlabonne/NeuralMarcoro14-7B	73.57
◆	abideen/NexoNimbus-7B	73.5
◆	Neuronovo/neuronovo-7B-v0.2	73.44
◆	argilla/distilabeled-Marcoro14-7B-slerp-full	73.4
◆	Cultrix/MistralTrix-v1	73.39
◆	ryandt/MusingCaterpillar	73.33
◆	Neuronovo/neuronovo-7B-v0.3	73.29
◆	Cultrix/MistralTrixTest	73.17
◆	samir-fama/SamirGPT-v1	73.11
◆	SanjiWatsuki/Lelantos-DPO-7B	73.09

*Handwritten notes in red:*  
DPO (for udkai/Turdus)  
DPO (& UNA) (for fblgit/UNA-TheBeagle-7b-v1)  
DPO (for argilla/distilabeled-Marcoro14-7B-slerp)  
Merge (of DPO models) (for mlabonne/NeuralMarcoro14-7B)  
DPO (for abideen/NexoNimbus-7B)  
DPO (for Neuronovo/neuronovo-7B-v0.2)  
DPO (for argilla/distilabeled-Marcoro14-7B-slerp-full)  
DPO (for Cultrix/MistralTrix-v1)  
DPO (for ryandt/MusingCaterpillar)  
DPO (for Neuronovo/neuronovo-7B-v0.3)  
No info but prob DPO, given Merge (incl. DPO) (for Cultrix/MistralTrixTest)  
DPO (for samir-fama/SamirGPT-v1)  
DPO (for SanjiWatsuki/Lelantos-DPO-7B)

Now, almost all top models are DPO models

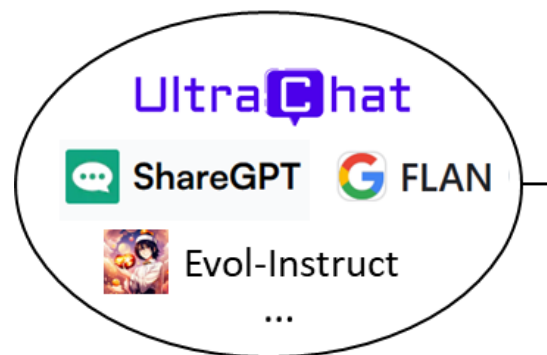
# UltraFeedback

## Construction process

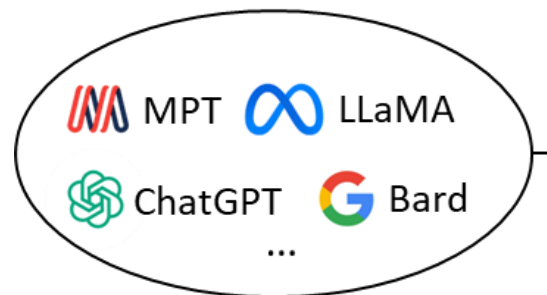


# UltraFeedback

## Instruction Pool



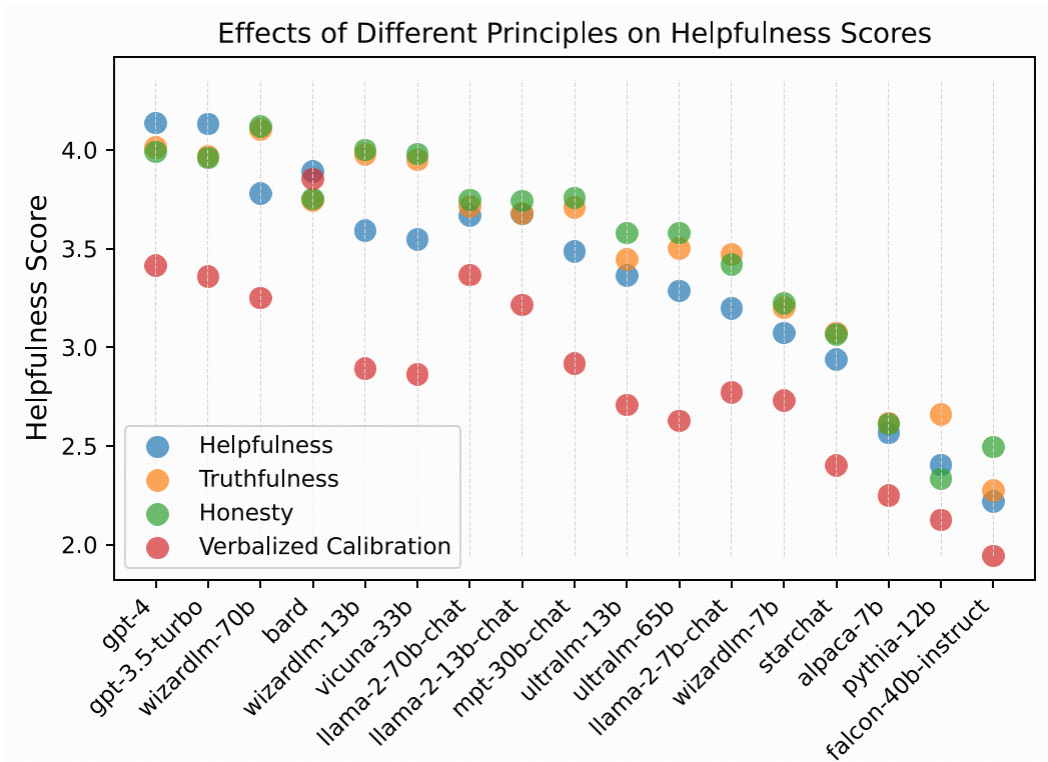
## Model Pool



Diversity is the key!

- Select **diverse and high-quality** instructions, reflect different requirements to chat models
- Select distinct model families for **response diversity**
- We also handwrite several principles to steer model behaviors

# UltraFeedback



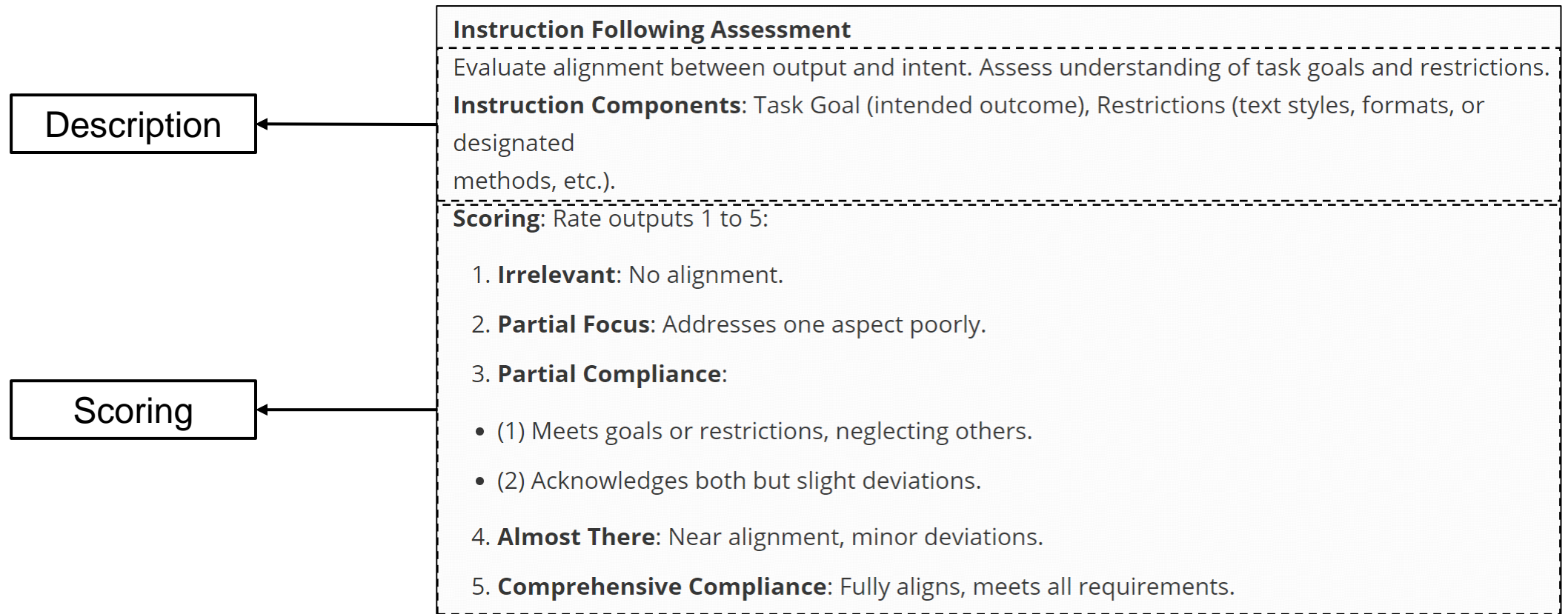
Diversity is the key!

- Select **diverse and high-quality** instructions, reflect different requirements to chat models
- Select distinct model families for **response diversity**
- We also handwrite several principles to steer model behaviors

# UltraFeedback

## Annotation

- Divide and conquer, with 4 aspects
- Detailed annotation doc



# UltraFeedback

## Statistics

- Largest and longest open preference datasets

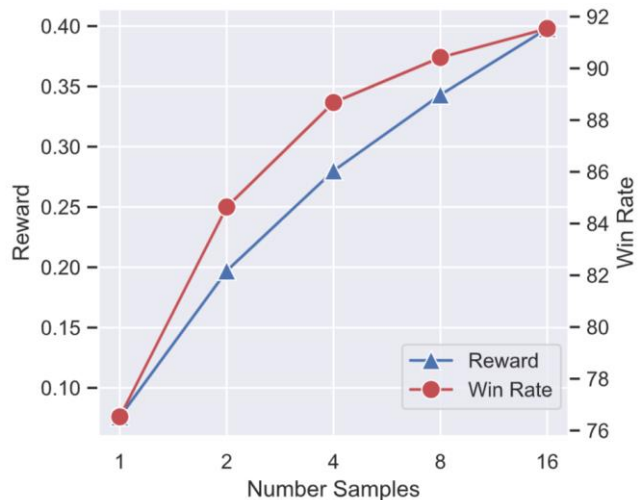
Table 1. Statistics of existing preference and critique datasets. The average length refers to the number of tokens.

Dataset	# Convs	Prompt Length	Response Length	Critique Length	Fine-Grained?	Feedback Format	# Pairs	# Critique	Annotator
<i>Preference Dataset</i>									
<b>OASST1</b>	35,905	167.6	221.1	-	✗	Scalar	17,966	-	Human
<b>OpenAI WebGPT</b>	38,925	50.9	188.2	-	✗	Scalar	19,578	-	Human
<b>Anthropic Helpful</b>	118,263	185.7	94.6	-	✗	Ranking	118,263	-	Human
<b>OpenAI Summ.</b>	60,674	<b>326.4</b>	36.6	-	✓	Scalar	92,858	-	Human
<b>QA Feedback</b>	11,378	155.8	107.9	-	✓	Scalar	17,118	-	Human
<i>Critique Dataset</i>									
<b>SelfFee</b>	178,331	100.3	243.9	89.4	✓	Text	-	<b>316,026</b>	AI
<b>Shepherd</b>	1,316	95.3	97.6	67.2	✓	Text	-	1,317	Human
<b>ULTRAFEEDBACK</b>	<b>255,864</b>	185.1	<b>305.3</b>	<b>143.1</b>	✓	Scalar & Text	<b>340,025</b>	255,864	AI

# UltraFeedback

## Experiments

- Reward modeling
- Best-of-N sampling



## 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.

Evaluator: **GPT-4** Claude Filter: **Community** Verified Minimal

Model Name	Win Rate	Length
XwinLM 70b V0.1	95.57%	1775
GPT-4	95.28%	1365
LLaMA2 Chat 70B	92.66%	1790
<b>UltraLM 13B V2.0 (best-of-16)</b>	92.30%	1720
XwinLM 13b V0.1	91.76%	1894
<b>UltraLM 13B (best-of-16)</b>	91.54%	1980
Claude 2	91.36%	1069

Table 2. Reward modeling accuracy (%) results. We compare our UltraRM with baseline open-source reward models. LLaMA2 results are taken from (Touvron et al., 2023b). The highest results are in **bold** and the second highest scores are underlined.

Model	Backbone Model	Open?	Anthropic Helpful	OpenAI WebGPT	OpenAI Summ.	Stanford SHP	Avg.
Moss	LLaMA-7B	✓	61.3	58.1	59.0	54.6	58.3
Ziya	LLaMA-7B	✓	61.4	61.8	60.3	57.0	60.1
OASST	DeBERTa-v3-large	✓	67.6	-	71.8	53.9	-
SteamSHP	FLAN-T5-XL	✓	55.4	62.6	48.4	51.6	54.5
LLaMA2 Helpfulness	LLaMA2-70B	✗	<b>72.0</b>	-	<b>75.5</b>	<b>80.0</b>	-
UltraRM-UF	LLaMA2-13B	✓	66.7	65.1	66.8	68.4	66.8
UltraRM-Overall	LLaMA2-13B	✓	<u>71.0</u>	62.0	73.0	73.6	<u>69.9</u>
<b>UltraRM</b>	LLaMA2-13B	✓	<u>71.0</u>	<b>65.2</b>	<u>74.0</u>	<u>73.7</u>	<b>71.0</b>



# UltraFeedback

## Experiments

- PPO: Improve **16.8%** win rate

Table 3. Head-to-head comparison results on three public benchmarks. The baseline is `text-davinci-003` in AlpacaEval and `gpt-3.5-turbo` in Evol-Instruct and UltraChat. The judge is GPT-4. The highest win rates are in **bold**.

Model	Size	AlpacaEval Win (%)	Evol-Instruct Win / Tie / Lose (%)	UltraChat Win / Tie / Lose (%)	Average Win (%)
ChatGPT	-	89.4	-	-	-
<i>LLaMA2</i>					
Vicuna-13B-v1.5	13B	-	33.0 / 23.9 / 43.1	34.5 / 38.2 / 27.3	-
LLaMA2-13B-Chat	13B	81.1	44.1 / 11.9 / 44.0	53.5 / 21.3 / 25.2	59.5
WizardLM-13B-v1.2	13B	89.2	55.5 / 17.4 / 27.1	59.7 / 25.5 / 14.8	68.1
OpenChat-13B-v3.2super	13B	89.5	55.5 / 11.0 / 33.5	58.7 / 26.7 / 14.5	67.9
LLaMA2-70B-Chat	70B	<b>92.7</b>	56.4 / 13.8 / 29.8	54.0 / 28.6 / 17.4	67.7
<i>LLaMA</i>					
UltraLM-13B	13B	80.7	39.9 / 14.7 / 45.4	38.2 / 34.8 / 27.0	52.9
Vicuna-13B-v1.3	13B	82.1	36.7 / 17.4 / 45.9	41.3 / 33.2 / 25.5	53.4
WizardLM-13B-v1.1	13B	86.3	54.1 / 14.7 / 31.2	56.1 / 26.0 / 17.9	65.5
Vicuna-33B-v1.3	33B	89.0	50.0 / 17.0 / 33.0	57.7 / 25.7 / 16.6	65.6
UltraLM-13B-PPO	13B	86.3	<b>57.8</b> / 10.1 / 32.1	<b>64.9</b> / 15.6 / 19.5	<b>69.7</b>

# UltraFeedback

## Agreement with human labelers

- High agreement with human labelers
- Win rates are also close

Table 4. Agreement between judges on 400 samples from ULTRA-FEEDBACK, AlpacaEval, Evol-Instruct, and UltraChat test sets. A-1, A-2, A-3 are three human judges. “Majority” stands for the agreement between each judge and other three judges’s majority votes. We include tie votes and the random agreement is 33%.

Judge	A-1	A-2	A-3	Average	Majority
<b>GPT-4</b>	59.2%	60.8%	59.1%	<b>59.7%</b>	<b>68.6%</b>
<b>A-1</b>	-	58.1%	54.7%	57.3%	60.3%
<b>A-2</b>	58.1%	-	55.4%	58.1%	63.3%
<b>A-3</b>	54.7%	55.4%	-	56.4%	62.0%

Table 5. Human evaluation results. We use majority votes from three human judges and compare GPT-4 and human evaluations on the same 266 samples.

Judge	AlpacaEval Win (%)	Evol-Instruct Win / Tie / Lose (%)	UltraChat Win / Tie / Lose (%)	Avg. Win (%)
<b>GPT-4</b>	83.9	57.1 / 8.8 / 34.1	61.0 / 17.1 / 21.9	67.3
<b>Human</b>	78.5	68.1 / 17.6 / 14.3	46.3 / 19.5 / 34.1	64.3

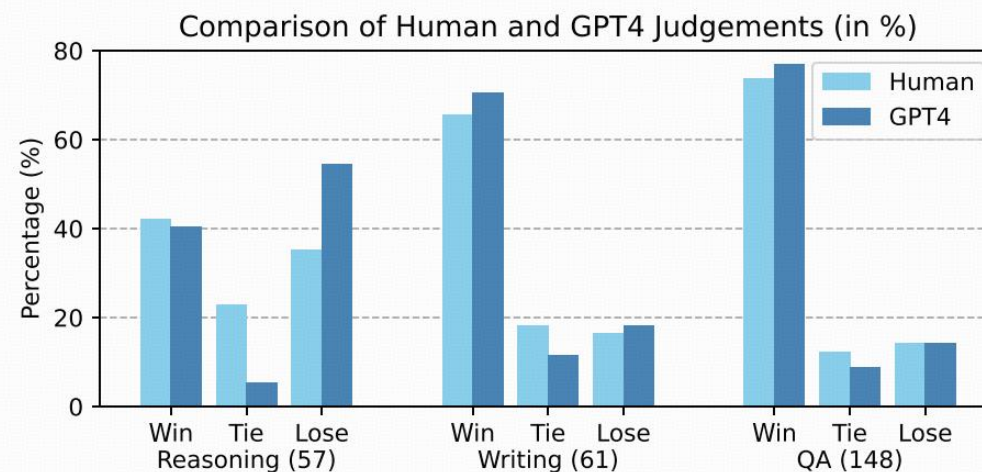


Figure 3. Categorical comparison of human and GPT-4 judgments. Human judgments are majority votes from three annotators. Sample numbers of each category are in parenthesis.

# UltraFeedback

Over **1000** models on HuggingFace are aligned with UltraFeedback rank **#5** among all datasets, **1 million** downloads per month



HuggingFace **Zephyr-7B** surpassed LLaMA2-70B-Chat, selected by their official handbook

- ① IMAGENET
- ② Common Voice
- ③ Wikipedia
- ④ XTREME
- ⑤ UltraFeedback



Used by

## Current directions

1. **Data! Data! Data!** We are *severely limited* on experimentation by having too few preference datasets (Anthropic HH, UltraFeedback, and Nectar are main three).
2. **Continuing to improve DPO:** tons of papers iterating on the method ([ORPO](#), [cDPO](#), [IPO](#), [BCO](#), [KTO](#), [DNO](#), [sDPO](#), etc)
3. **More model sizes:** Most alignment research happened at 7 or 13B parameter scale. Expand up and down!
4. **Specific evaluations:** How do we get more specific evaluations than ChatBotArena?
5. **Personalization:** A large motivation behind local models, young area academically

I cover these topics regularly on my blog [www.interconnects.ai](http://www.interconnects.ai)

Aligning open language models | Lambert: 75

## Evaluating Reward Models

- Accuracy of predicting human preferences.

**Preference Datasets**

Table 2: Reward modeling accuracy (%) results. We compare our UltraRM with baseline open-source reward models. LLaMA2 results are taken from [Touvron et al. \(2023b\)](#). The highest results are in **bold** and the second highest scores are underlined.

Model	Backbone Model	Open?	Anthropic Helpful	OpenAI WebGPT	OpenAI Summ.	Stanford SHP	Avg.
Moss	LLaMA-7B	✓	61.3	54.6	58.1	54.6	57.2
Ziya	LLaMA-7B	✓	61.4	57.0	61.8	57.0	59.3
OASST	DeBERTa-v3-large	✓	67.6	-	72.1	53.9	-
SteamSHP	FLAN-T5-XL	✓	55.4	51.6	62.6	51.6	55.3
LLaMA2 Helpfulness	LLaMA2-70B	✗	<b>72.0</b>	-	<u>75.5</u>	<b>80.0</b>	-
UltraRM-UF	LLaMA2-13B	✓	66.7	65.1	66.8	68.4	66.8
UltraRM-Overall	LLaMA2-13B	✓	71.0	62.0	73.0	73.6	69.9
UltraRM	LLaMA2-13B	✓	71.0	<b>65.2</b>	74.0	73.7	<b>71.0</b>

Cui et al. ArXiv 2023 "UltraFeedback: Boosting Language Models with High-quality Feedback"

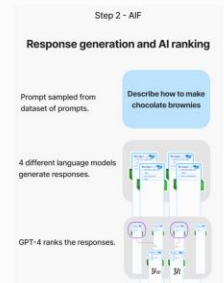
## Step 2: AI Feedback through Preferences



- AI Feedback (AIF) collection via an ensemble of chat model completions

- Followed by scoring by GPT-4 (teacher model) **UltraFeedback**

- Binarization of the preferences



Stanford CS25/CS329H

UWashington CSE 447/517

UWaterloo CS886



**Thank You!**

THUNLP

2024/06/04