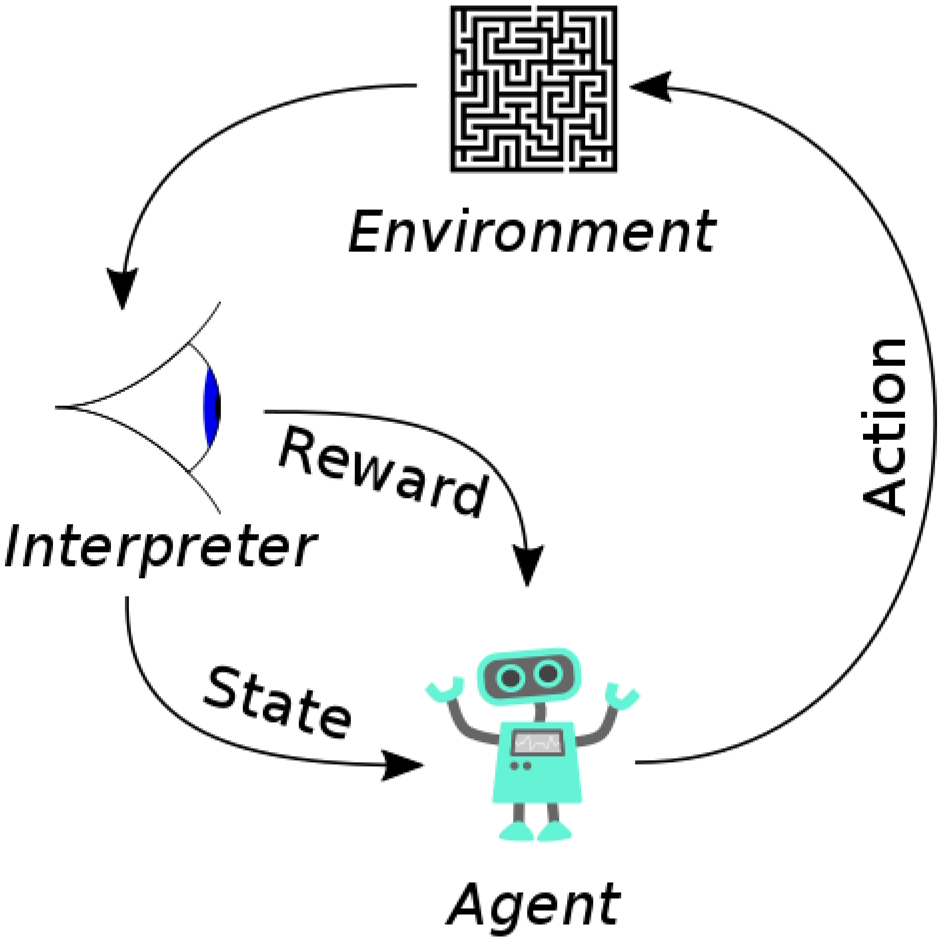
UltraFeedback: Boosting Language Models with Scaled AI Feedback

THUNLP

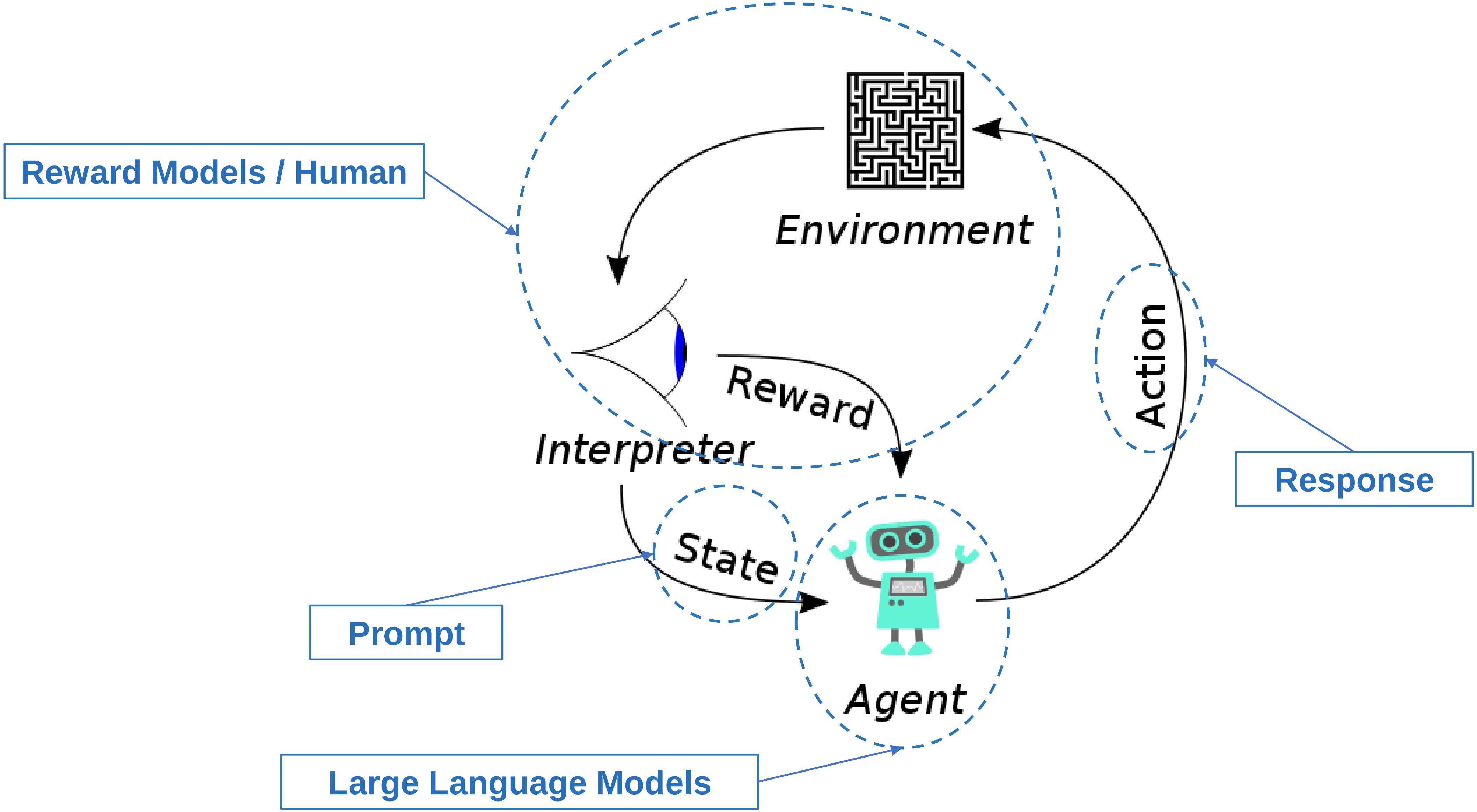
Ganqu Cui* · Lifan Yuan* · Ning Ding · Guanming Yao · Bingxiang He · Wei Zhu · Yuan Ni · Guotong Xie · Ruobing Xie · Yankai Lin · Zhiyuan Liu · Maosong Sun

2024/06/04

Brief Introduction to RLHF



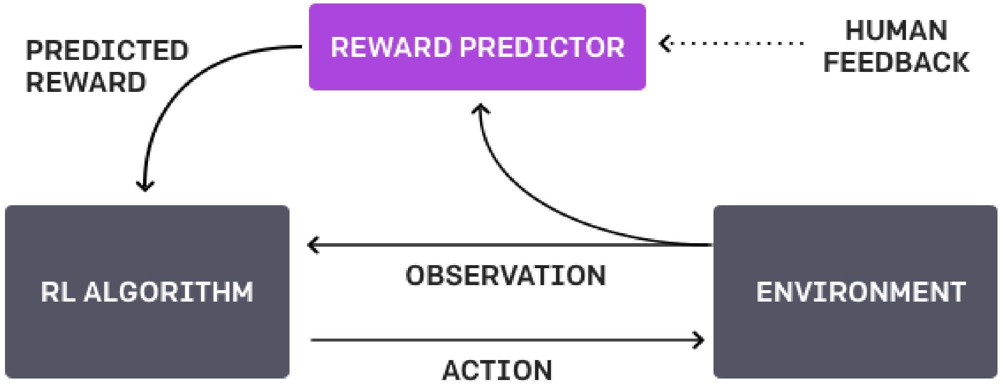
Brief Introduction to RLHF



Brief Introduction to RLHF

Early OpenAI practices

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



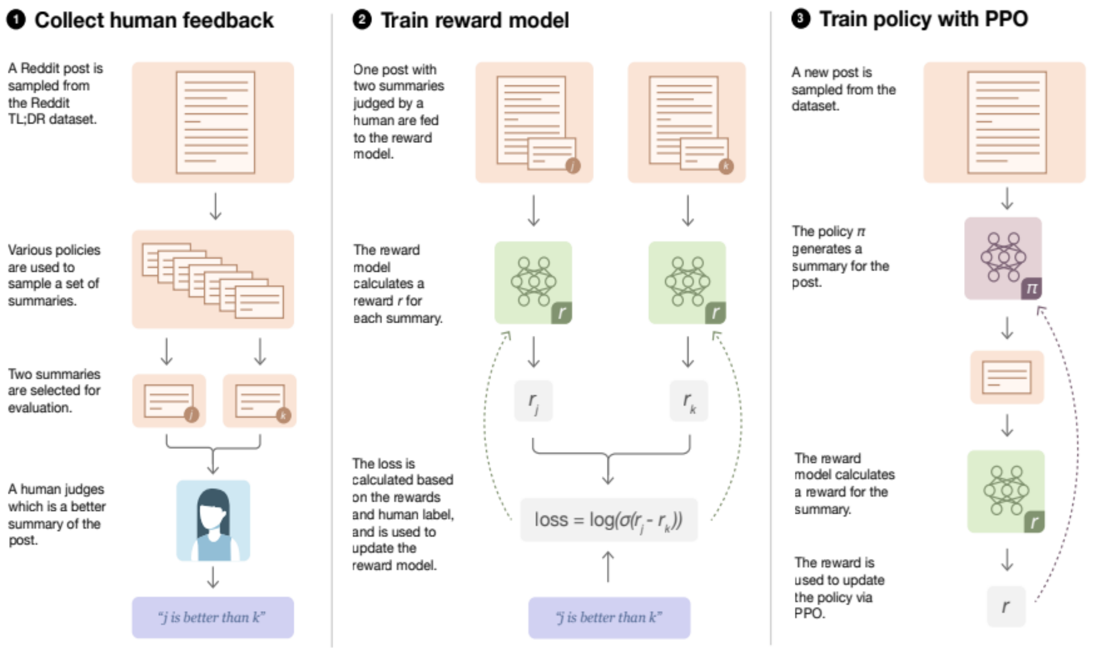


Figure 2: Diagram of our human feedback, reward model training, and policy training procedure.

Christiano, Paul F., et al. "Deep reinforcement learning from human preferences." 2017.
Stiennon, Nisan, et al. "Learning to summarize with human feedback." *2020.*

Brief Introduction to RLHF

Why RLHF?

- Takeaway from traditional RL problems
- Objective mismatch

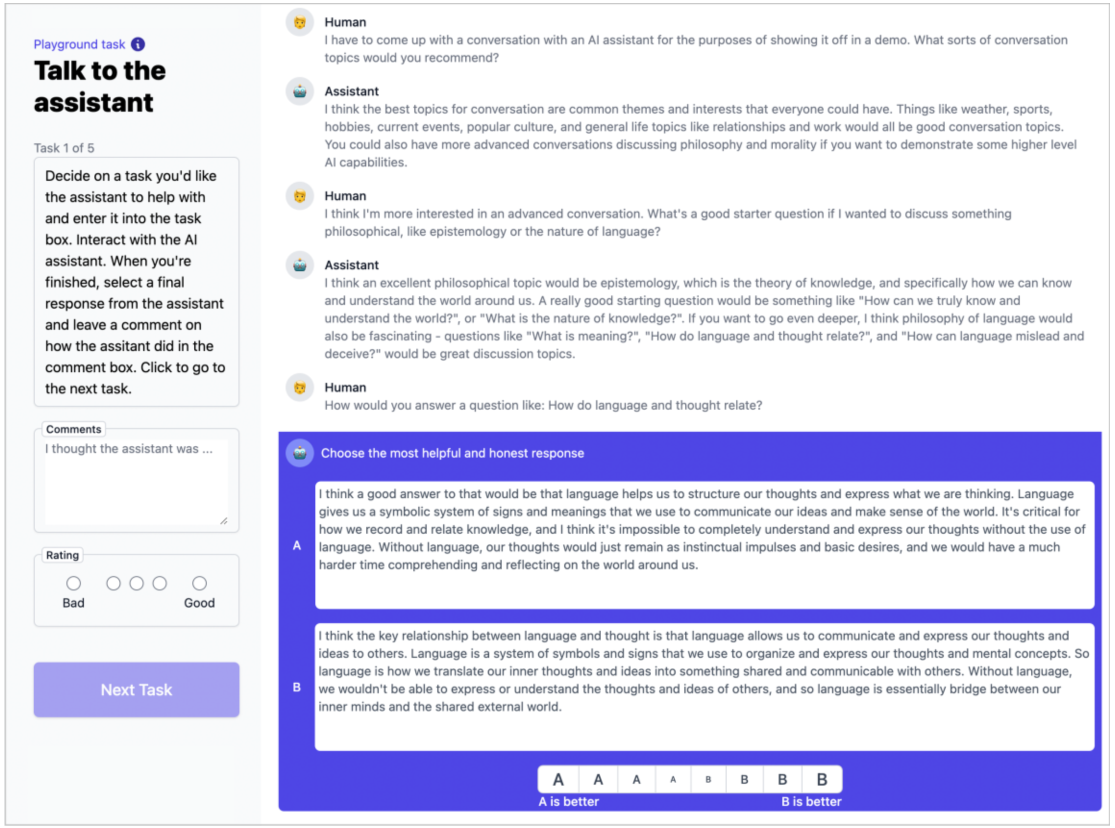
While this strategy has led to markedly improved performance, there is still a misalignment between this fine-tuning objective—maximizing the likelihood of human-written text—and what we care about—generating high-quality outputs as determined by humans. This misalignment has several causes: the maximum 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

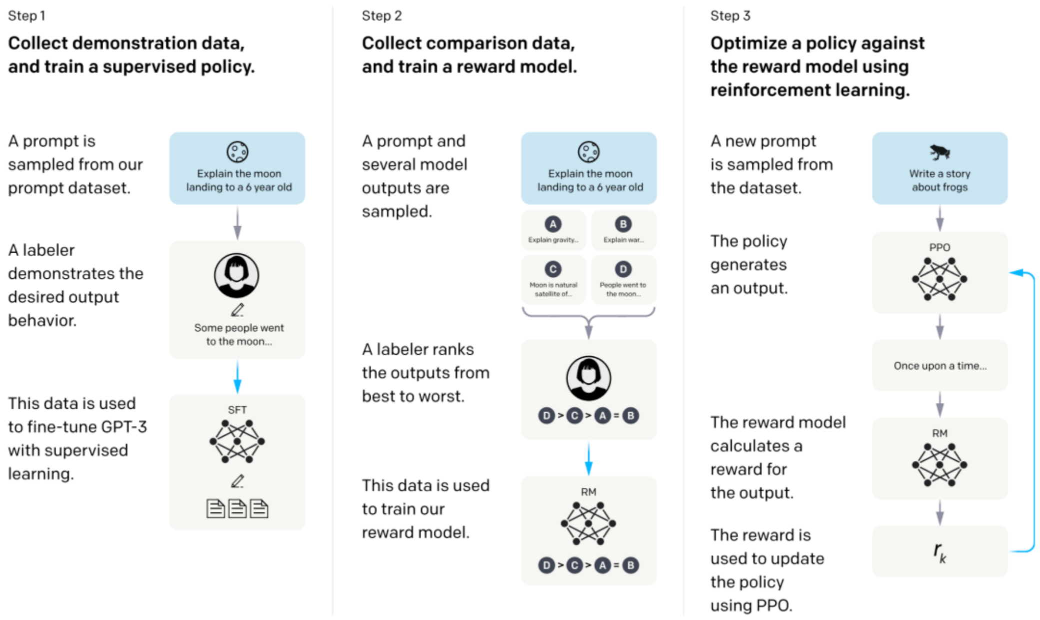
Stiennon, Nisan, et al. "Learning to summarize with human feedback." *2020.*

Brief Introduction to RLHF

RLHF for alignment

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

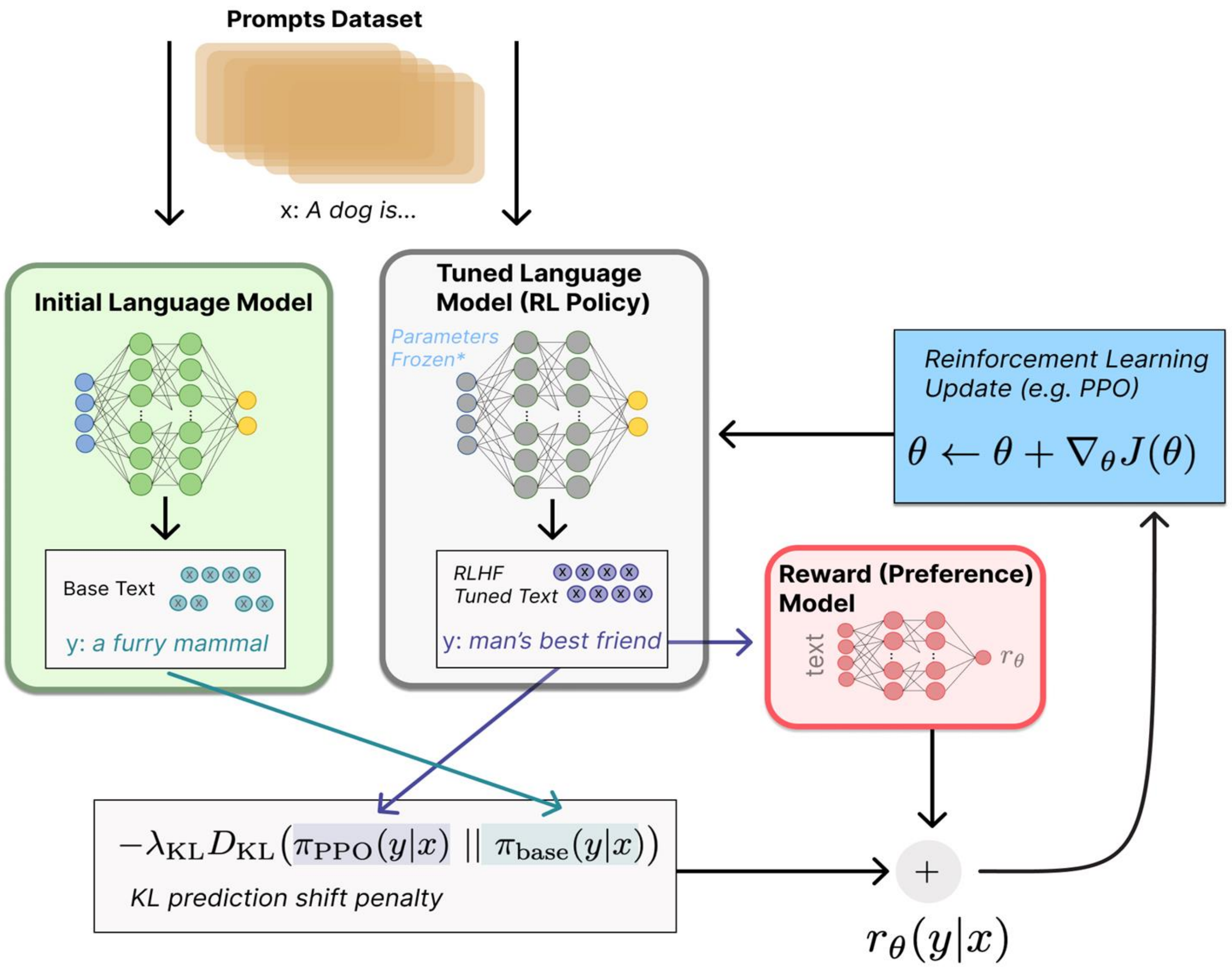




Bai, et al. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." 2021.
Ouyang, et al. "Training language models to follow instructions with human feedback." *2020.*

Brief Introduction to RLHF

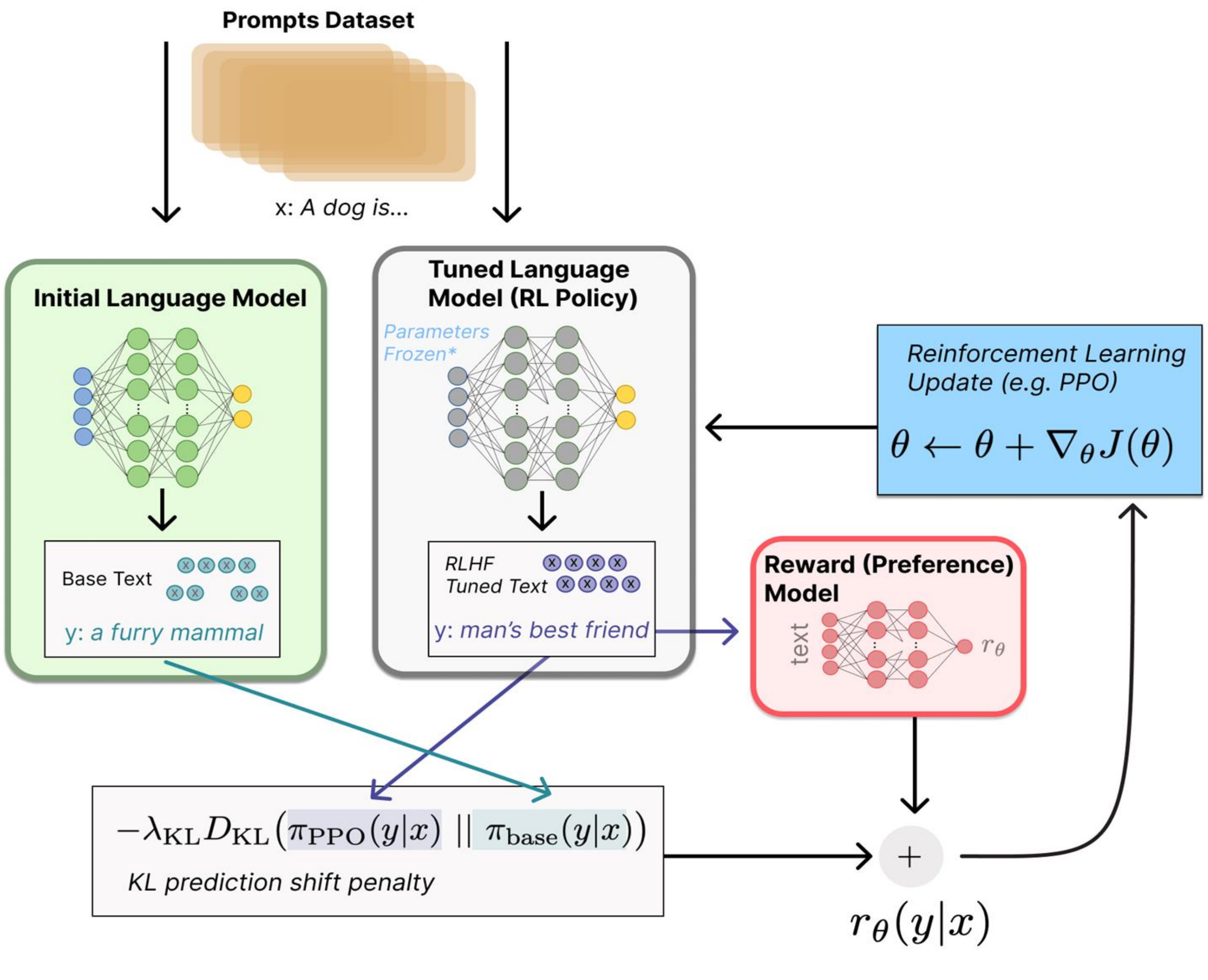
Overview



Brief Introduction to RLHF

However

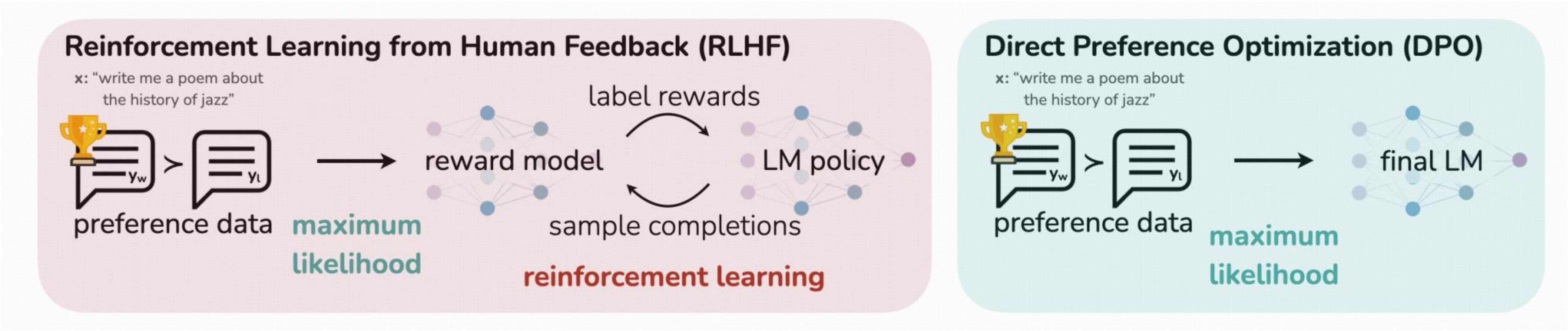
- Online RL (PPO) requires huge computational resources
- 4 models, 3~4 times larger GPU memory than SFT
- Not friendly to academy and open-source community



Brief Introduction to RLHF

Direct Preference Optimization

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



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

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
○	chargoddard/MelangeA-70b	72.43
○	ehartford/Samantha-1.1-70b	72.42
◆	psmathur/model_009	72.36
◆	upstage/Llama-2-70b-instruct	72.29
○	chargoddard/MelangeB-70b	72.14

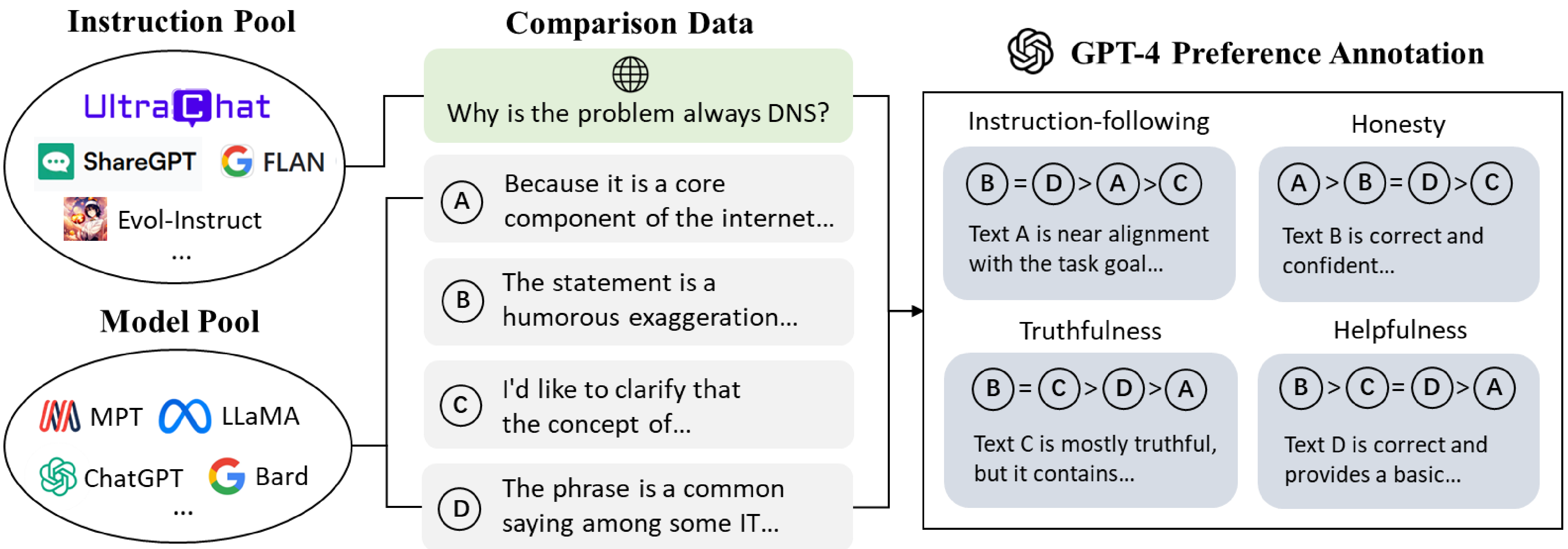
2023/08, no RLHF models on Open LLM Leaderboard

T	Model		Average
■	udkai/Turdus	DPO	74.66
■	fblgit/UNA-TheBeagle-7b-v1	DPO (& UNA)	73.87
■	argilla/distilabeled-Marcoro14-7B-slerp	DPO	73.63
■	mlabonne/NeuralMarcoro14-7B	DPO	73.57
◆	abideen/NexoNimbus-7B	Merge (of DPO models)	73.5
■	Neuronovo/neuronovo-7B-v0.2	DPO	73.44
■	argilla/distilabeled-Marcoro14-7B-slerp-full	DPO	73.4
■	CultriX/MistralTrix-v1	DPO	73.39
■	ryandt/MusingCaterpillar	DPO	73.33
■	Neuronovo/neuronovo-7B-v0.3	DPO	73.29
■	CultriX/MistralTrixTest	No info but prob DPO, given	73.17
◆	samir-fama/SamirGPT-v1	Merge (incl. DPO)	73.11
◆	SanjiWatsuki/Lelantos-DPO-7B	DPO	73.09

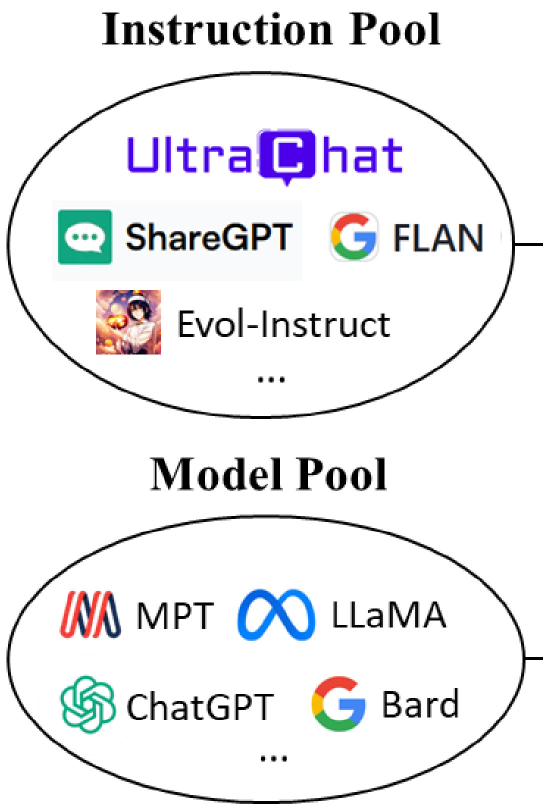
Now, almost all top models are DPO models

UltraFeedback

Construction process



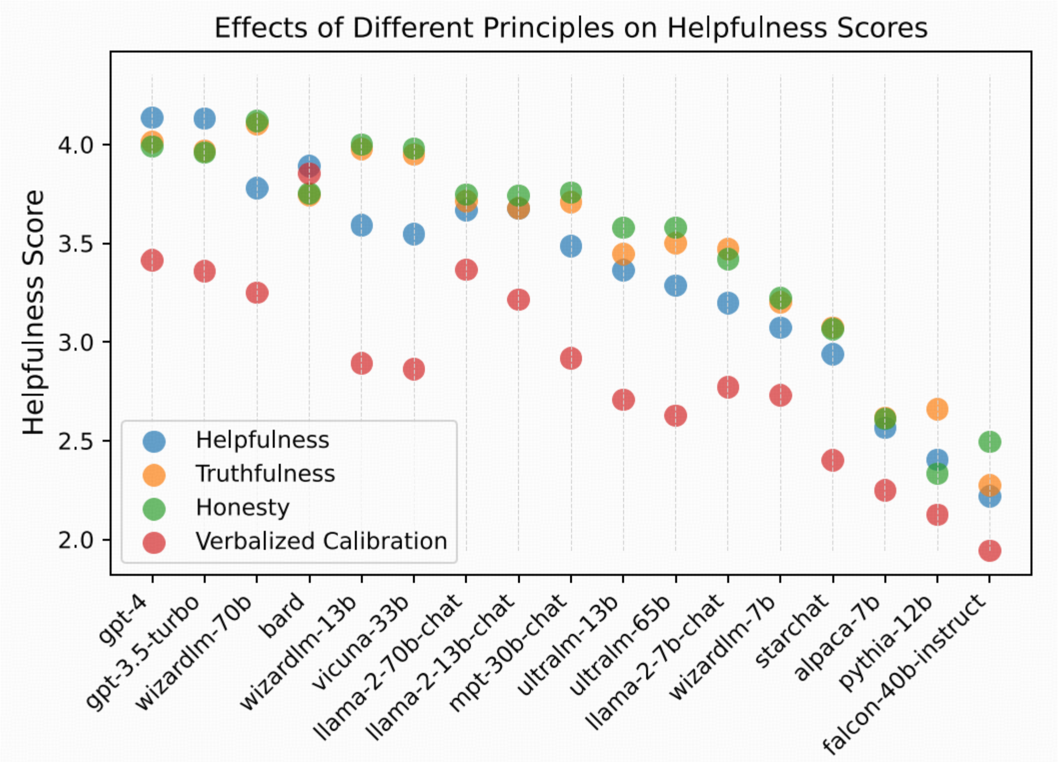
UltraFeedback



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



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

Description

Scoring

Instruction Following Assessment

Evaluate alignment between output and intent. Assess understanding of task goals and restrictions.

Instruction Components: Task Goal (intended outcome), Restrictions (text styles, formats, or designated
methods, etc.).

Scoring: Rate outputs 1 to 5:

1. **Irrelevant**: No alignment.
2. **Partial Focus**: Addresses one aspect poorly.
3. **Partial Compliance**:
   - (1) Meets goals or restrictions, neglecting others.
   - (2) Acknowledges both but slight deviations.
4. **Almost There**: Near alignment, minor deviations.
5. **Comprehensive Compliance**: Fully aligns, meets all requirements.

UltraFeedback

Statistics

- Largest and longest open preference datasets

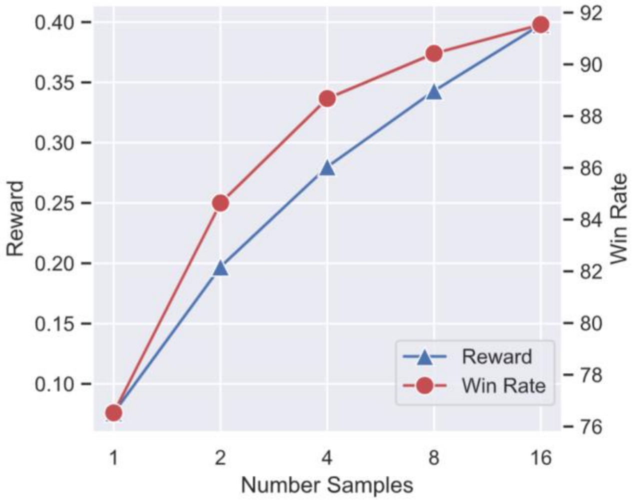
Table 1. Statistics of existing preference and critique datasets. The average length refers to the number of tokens.

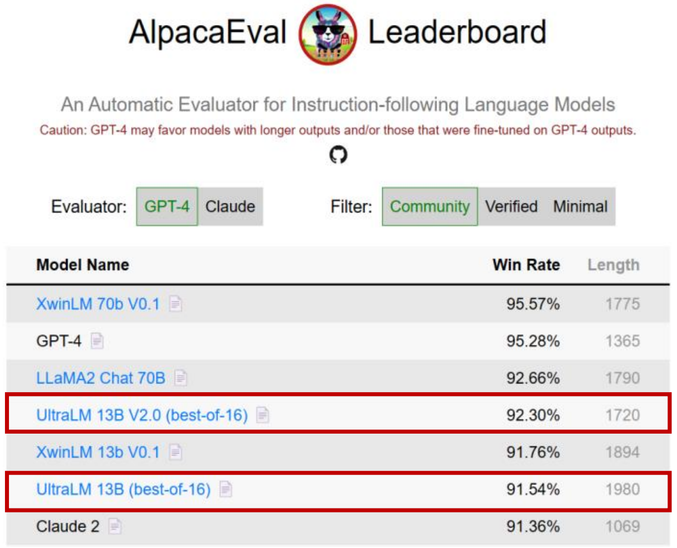
Dataset	# Convs	Prompt Length	Response Length	Critique Length	Fine-Grained?	Feedback Format	# Pairs	# Critique	Annotator
				Preference Dataset					
OASST1	35,905	167.6	221.1	-	✗	Scalar	17,966	-	Human
OpenAI WebGPT	38,925	50.9	188.2	-	✗	Scalar	19,578	-	Human
Anthropic Helpful	118,263	185.7	94.6	-	✗	Ranking	118,263	-	Human
OpenAI Summ.	60,674	**326.4**	36.6	-	✓	Scalar	92,858	-	Human
QA Feedback	11,378	155.8	107.9	-	✓	Scalar	17,118	-	Human
				Critique Dataset					
SelFee	178,331	100.3	243.9	89.4	✓	Text	-	**316,026**	AI
Shepherd	1,316	95.3	97.6	67.2	✓	Text	-	1,317	Human
ULTRAFEEDBACK	**255,864**	185.1	**305.3**	**143.1**	✓	Scalar & Text	**340,025**	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
UltraLM 13B V2.0 (best-of-16)	92.30%	1720
XwinLM 13b V0.1	91.76%	1894
UltraLM 13B (best-of-16)	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 <u>underlined</u>.

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	✗	**72.0**	-	**75.5**	**80.0**	-
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>
UltraRM	LLaMA2-13B	✓	<u>71.0</u>	**65.2**	<u>74.0</u>	<u>73.7</u>	**71.0**

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	-	-	-
LLaMA2					
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	**92.7**	56.4 / 13.8 / 29.8	54.0 / 28.6 / 17.4	67.7
LLaMA					
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	**57.8** / 10.1 / 32.1	**64.9** / 15.6 / 19.5	**69.7**

UltraFeedback

Agreement with human labelers

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

Table 4. Agreement between judges on 400 samples from ULTRAFEEDBACK, 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
GPT-4	59.2%	60.8%	59.1%	**59.7%**	**68.6%**
A-1	-	58.1%	54.7%	57.3%	60.3%
A-2	58.1%	-	55.4%	58.1%	63.3%
A-3	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 (%)
GPT-4	83.9	57.1 / 8.8 / 34.1	61.0 / 17.1 / 21.9	67.3
Human	78.5	68.1 / 17.6 / 14.3	46.3 / 19.5 / 34.1	64.3

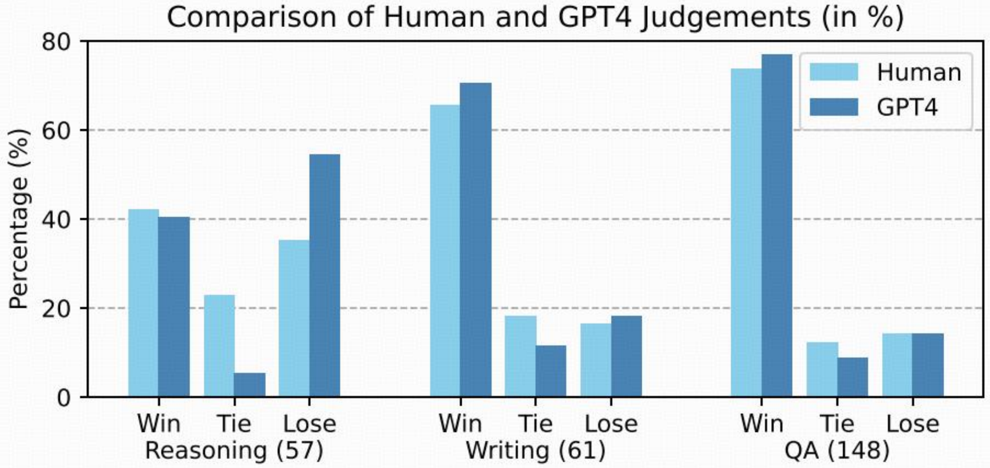


Figure 3. Catrgorical comparison of human and GPT-4 judgments. Human judgments are majority votes from three annotators. Sample numbers of each category are in parentness.

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

NVIDIA
Ai2
stability.ai

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
Aligning open language models | Lambert: 75

Stanford CS25/CS329H

Evaluating Reward Models

- Accuracy of predicting human preferences.

Preference Datasets

Reward Models

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	✗	**72.0**	-	**75.5**	**80.0**	-
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	**65.2**	74.0	73.7	**71.0**

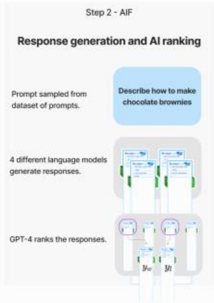
Cui et al., ArXiV 2023 "UltraFeedback: Boosting Language Models with High-quality Feedback"

Natural Language Processing - CSE 517 / CSE 447 21 Alignment of LLMs (Part 1)

UWashington CSE 447/517

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



2024-05-22 Slides created for CS886 at UWaterloo 130

UWaterloo CS886

Thank You!

THUNLP

2024/06/04