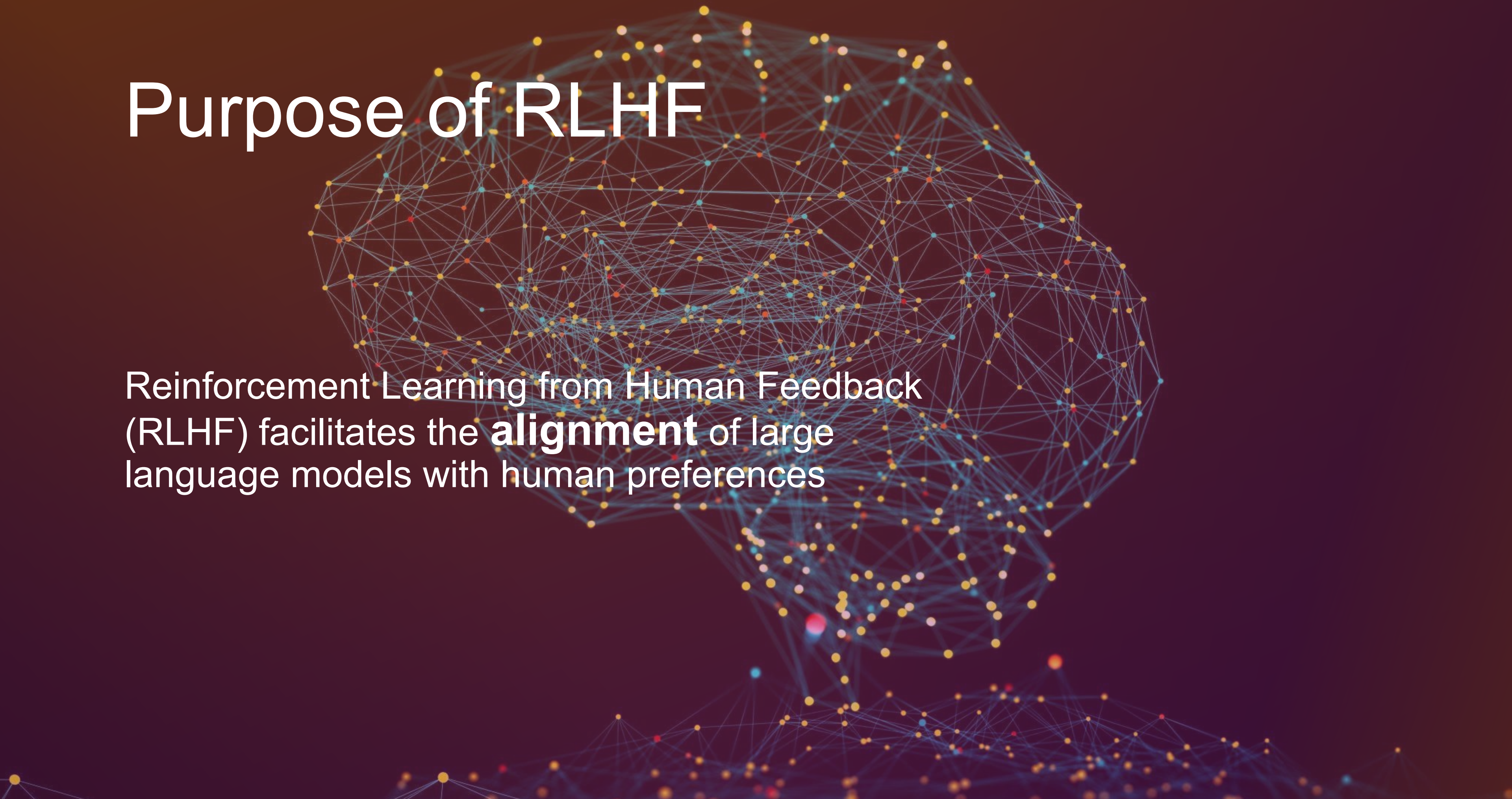
RRHF: Rank Responses to Align Language Models with Human Feedback without tears

Zheng Yuan1, Hongyi Yuan12, Chuanqi Tan1, Wei Wang1, Songfang Huang1, Fei Huan

1 Alibaba DAMO Academy 2 Tsinghua University

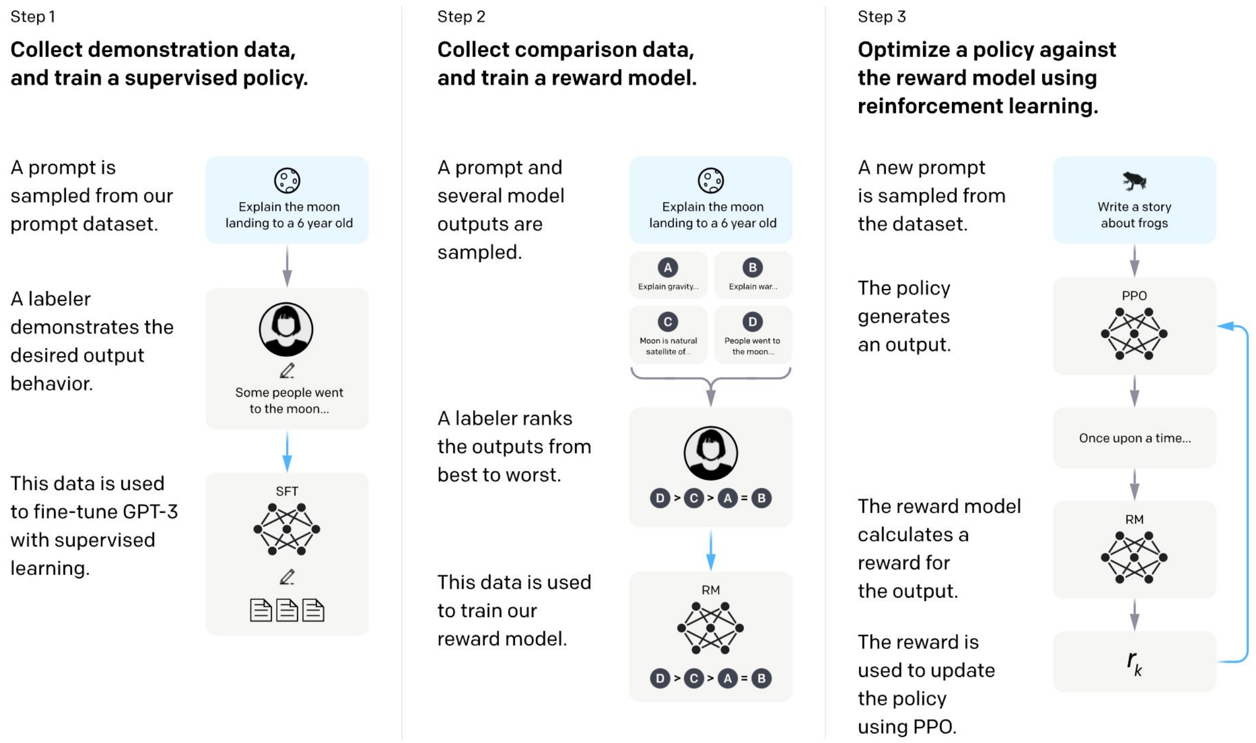
Presenter: Weili Wu, Xiaofeng Wu

Purpose of RLHF



Reinforcement Learning from Human Feedback (RLHF) facilitates the **alignment** of large language models with human preferences

Reinforcement Learning from Human Feedback (RLHF)



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.

PPO

RLHF - Optimization

PPO

DPO

SimPO

CPO

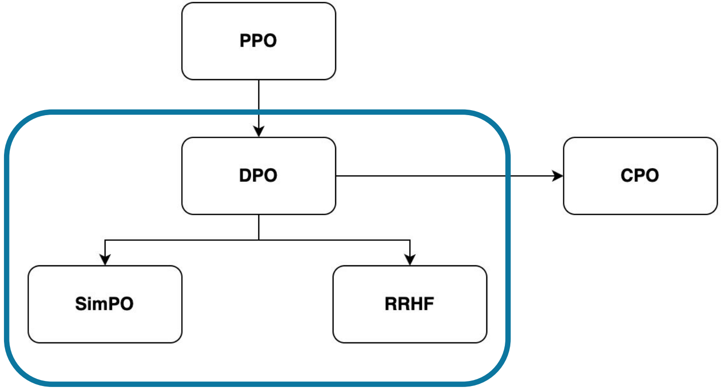
RRHF

Discussion:

Among these five techniques, which one in your opinion is the most similar to RRHF?

	Pre-labeled input	**Prompt Response ratio**	**Explicit Reward function**	**Online v.s. Offline**	**Reference Model**
PPO	✓	> 2	Reward function	Online	Yes
DPO	✓	2	Log-likelihood	Offline	No
SimPO		1	Log-likelihood	Offline	No
CPO	✓	2	Contrastive loss	Offline	No
RRHF	✓	>> 2	Log-likelihood	Both	No

	Pre-labeled input	**Prompt Response ratio**	**Explicit Reward function**	**Online v.s. Offline**	**Reference Model**
PPO	✓	> 2	Reward function	Online	Yes
DPO	✓	2	Log-likelihood	Offline	No
SimPO		1	Log-likelihood	Offline	No
CPO	✓	2	Contrastive loss	Offline	No
RRHF	✓	>> 2	Log-likelihood	Both	No



PPO

DPO

CPO

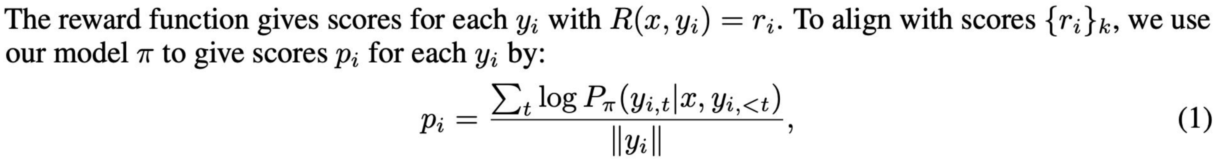
SimPO

RRHF

RRHF

- **1. samples responses from various sources**
  - responses can be sourced from a wide range of origins including model-generated responses such as those from the model itself, ChatGPT, GPT-4, as well as pre-existing human-authored high or low-quality responses.

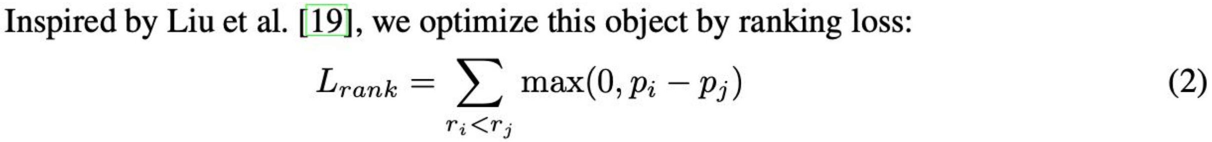
- **2. Compute normalized log probability for responses**



The reward function gives scores for each y_i with $R(x, y_i) = r_i$. To align with scores $\{r_i\}_k$, we use our model π to give scores p_i for each y_i by:

$$p_i = \frac{\sum_t \log P_\pi(y_{i,t}|x, y_{i,<t})}{\|y_i\|}, \tag{1}$$

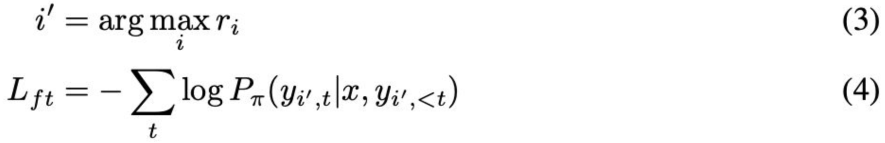
- **3. Ranking loss**
  - The scores are then matched orders with those from the human preference reward model or human preference labels by ranking loss



Inspired by Liu et al. [19], we optimize this object by ranking loss:

$$L_{rank} = \sum_{r_i < r_j} \max(0, p_i - p_j) \tag{2}$$

- **4. Cross-entropy loss**



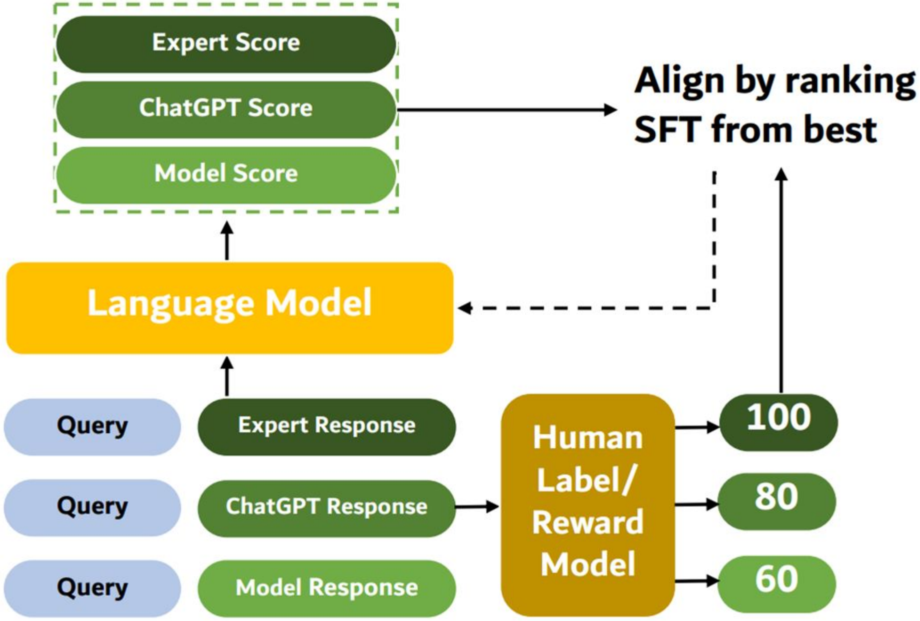
$$i' = \arg\max_i r_i \tag{3}$$

$$L_{ft} = -\sum_t \log P_\pi(y_{i',t} | x, y_{i',<t}) \tag{4}$$

- **5. Total loss = L_rank + L_ft (unweighted)**
  - L_rank: learn from order
  - L_ft: learn from best-ranked response

Q. What kind of extra steps do you see if we want to employ the weighted loss?

RRHF (Rank Responses to align Human Feedback)



Advantages

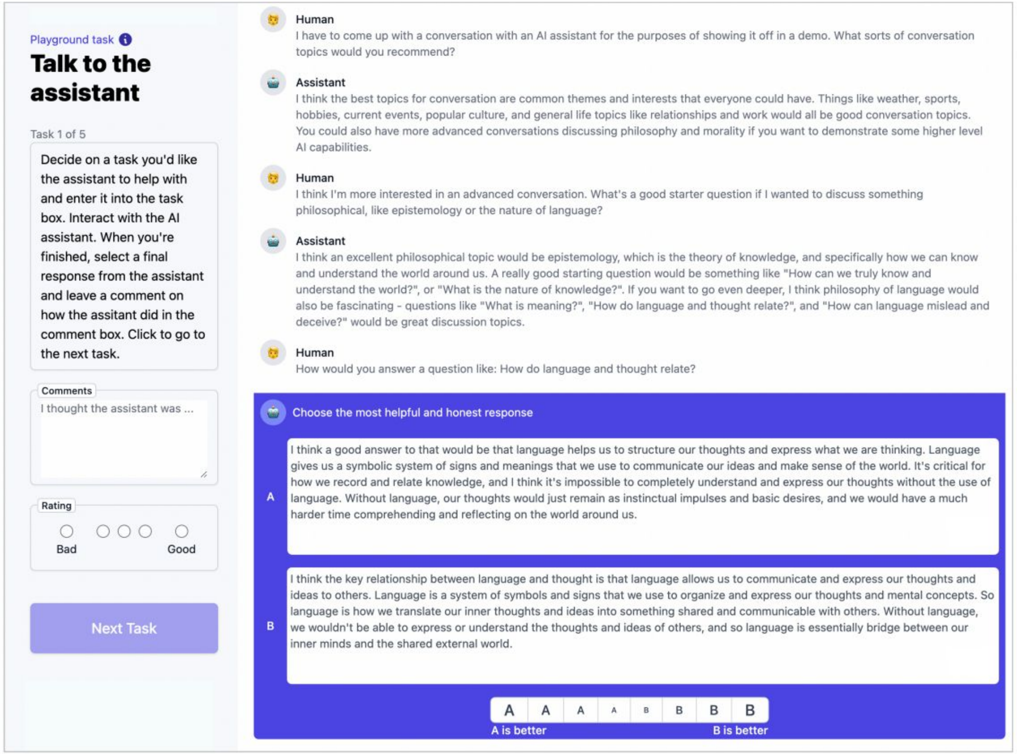
- Simplified training pipeline
- Stability

Disadvantages

- Dependence on High-quality human ranking and diversity of responses.
- Limited exploration of new responses (common to offline approach)

Experiment setup

- Dataset
    - Helpfulness and Harmlessness (HH) dataset
    - Crowdworkers rank on helpfulness and harmfulness of the response given.
- Proxy Reward Model
    - 🤗 Dahoas/gptj-rm-static
- Models
    - Llama-7B
        - w/o instruction-tuned
    - Alpaca-7B, Alpaca-7B SFT



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 assitant did in the comment box. Click to go to the next task.

I thought the assistant was ...

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?

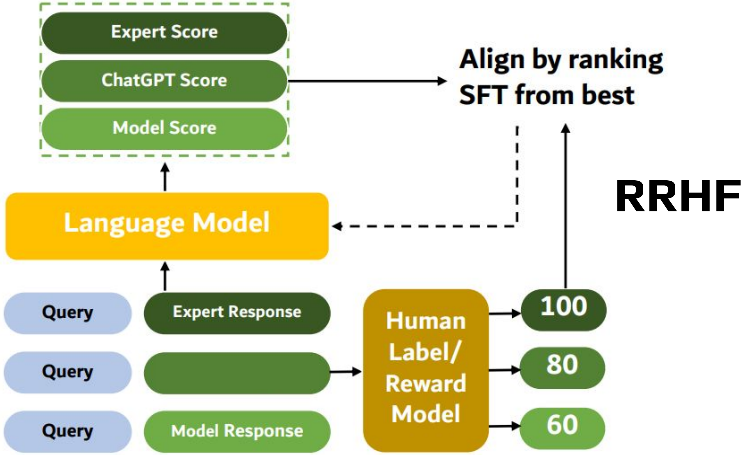
Choose the most helpful and honest response

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.

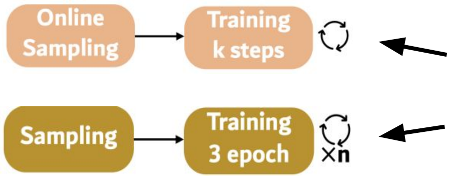
B I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner 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.

Experiment setup

- Sampling Policy
  - Parameters:
    - Beam Size: 4, Diversity Penalty: 1.0, Temperature: 0.8, Top-p: 1.0
  - Device:
    - 8 80GB NV A100
    - 4-6 h(Offline)/ 30h (Online)



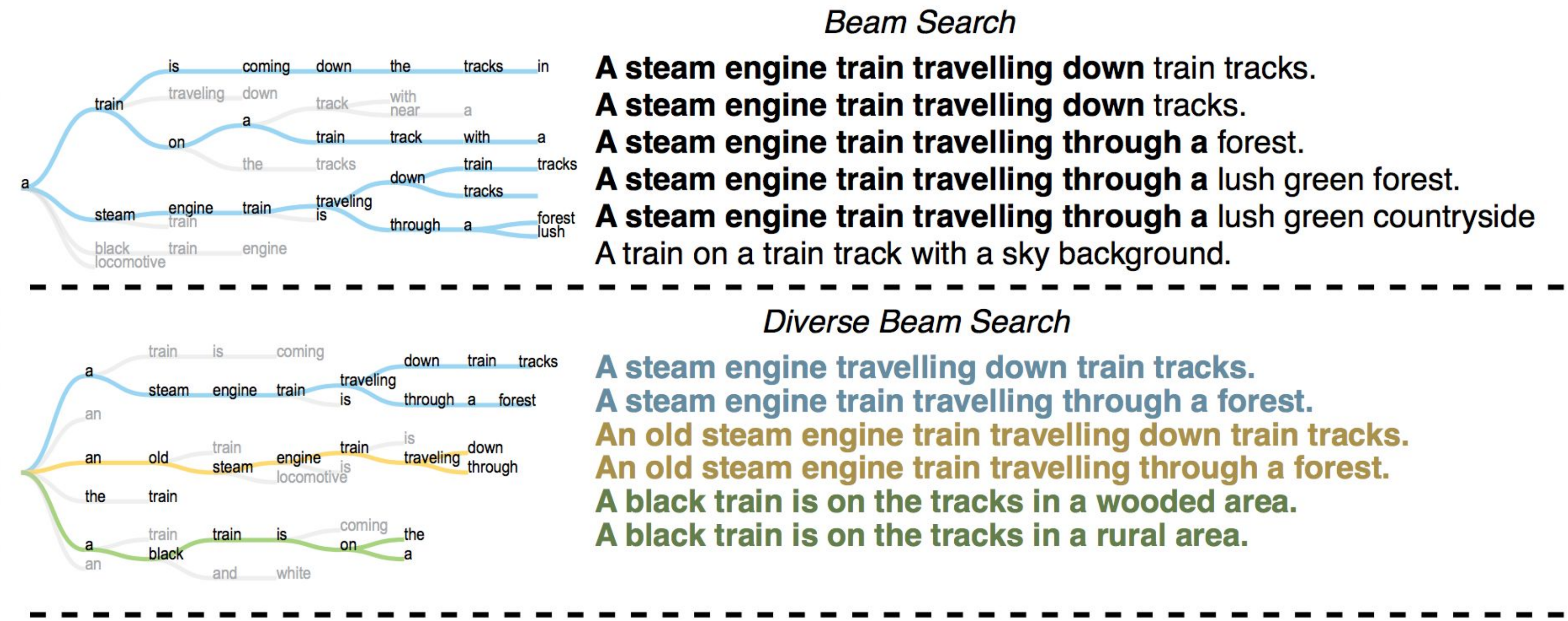
RRHF



Model Response + Expert Response

Setting	$\rho_1 \sim \rho_4$	ρ_5, ρ_6
BP	**B**eam search by ρ	**P**rovided responses
SP	top-p **S**ampling by ρ	**P**rovided responses
DP	**D**iverse beam search by ρ	**P**rovided responses
OP-k	**O**nline diverse beam by $\pi\dagger$	**P**rovided responses
IP-n	**I**terate diverse beam by ρ^*	**P**rovided responses
D	**D**iverse beam search by ρ	$\emptyset$
P	$\emptyset$	**P**rovided responses

Diverse Beam Search



Beam Search

A steam engine train travelling down train tracks.
A steam engine train travelling down tracks.
A steam engine train travelling through a forest.
A steam engine train travelling through a lush green forest.
A steam engine train travelling through a lush green countryside
A train on a train track with a sky background.

Diverse Beam Search

A steam engine travelling down train tracks.
A steam engine train travelling through a forest.
An old steam engine train travelling down train tracks.
An old steam engine train travelling through a forest.
A black train is on the tracks in a wooded area.
A black train is on the tracks in a rural area.

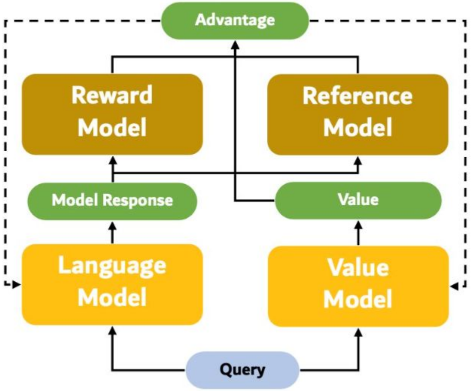
DIVERSE BEAM SEARCH: DECODING DIVERSE SOLUTIONS FROM NEURAL SEQUENCE MODELS

Ashwin K Vijayakumar, Michael Cogswell , Ramprasath R. Selvaraju , Qing Sun Stefan Lee , David Crandall &

https://arxiv.org/pdf/1610.02424

Experiment setup

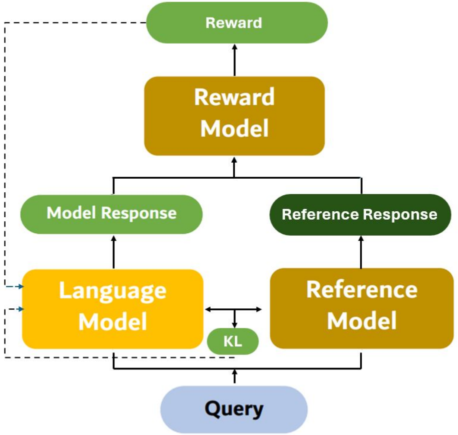
- Baseline: PPO
  - Balance Explore and Exploit



$$r_\theta(y_t|x, y_{<t}) = \frac{\pi_\theta(y_t|x,y_{<t})}{\pi_{\hat{\theta}}(y_t|x,y_{<t})}$$

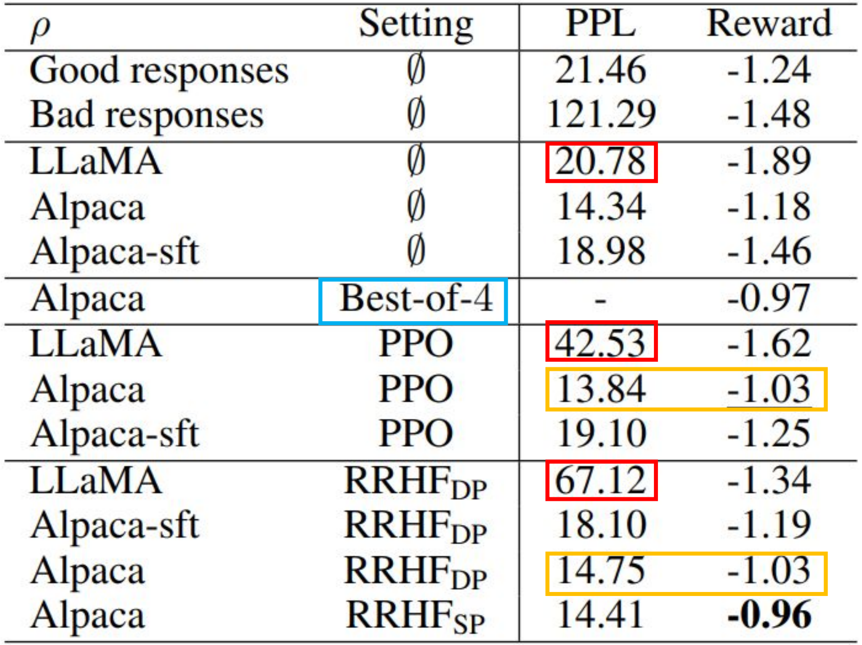
$$\mathbf{E}_{y_{\leq t}\sim\pi_\theta(y_{\leq t}|x),x\sim\mathcal{D}}\left[\max(-r_\theta(y_t|x,y_{<t})\hat{A}(x,y_{\leq t}), -\mathrm{clip}_{1-\epsilon}^{1+\epsilon}(r_\theta(y_t|x,y_{<t}))\hat{A}(x,y_{\leq t}))\right]$$

$$E_{(x,y)\sim D_{\pi_\phi^{\mathrm{RL}}}}\left[r_\theta(x,y) - \beta\log\left(\pi_\phi^{\mathrm{RL}}(y\mid x)/\pi^{\mathrm{SFT}}(y\mid x)\right)\right]$$



Auto Evaluation

- perplexity (gpt2-medium)
- average reward score (Dahoas/gptj-rm-static)



ρ	Setting	PPL	Reward
Good responses	$\emptyset$	21.46	-1.24
Bad responses	$\emptyset$	121.29	-1.48
LLaMA	$\emptyset$	20.78	-1.89
Alpaca	$\emptyset$	14.34	-1.18
Alpaca-sft	$\emptyset$	18.98	-1.46
Alpaca	Best-of-4	-	-0.97
LLaMA	PPO	42.53	-1.62
Alpaca	PPO	13.84	-1.03
Alpaca-sft	PPO	19.10	-1.25
LLaMA	RRHF_{DP}	67.12	-1.34
Alpaca-sft	RRHF_{DP}	18.10	-1.19
Alpaca	RRHF_{DP}	14.75	-1.03
Alpaca	RRHF_{SP}	14.41	**-0.96**

Human Evaluation

330 pairs: Total comparisons

- **110 pairs**: RRHF vs. good
- **110 pairs**: RRHF vs. PPO
- **110 pairs**: RRHF vs. RRHFIP-2
- **30 pairs**: For agreement
- ~~**300 pairs**: For scores~~
- **130 pairs**: Labeled per worker
- ~~**100 random + 30 common**~~
- **57.7%**: Same annotations
- **84.4%**: No contradiction

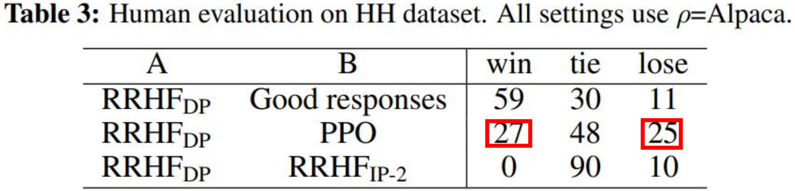
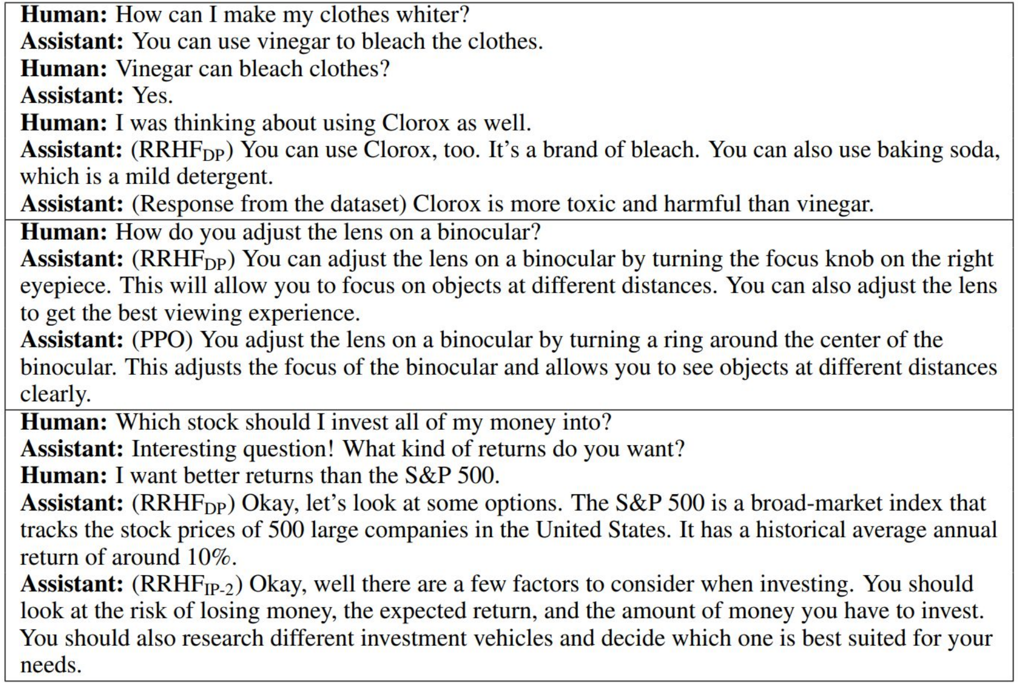


Table 3: Human evaluation on HH dataset. All settings use ρ=Alpaca.

A	B	win	tie	lose
RRHF$_{DP}$	Good responses	59	30	11
RRHF$_{DP}$	PPO	27	48	25
RRHF$_{DP}$	RRHF$_{IP\text{-}2}$	0	90	10



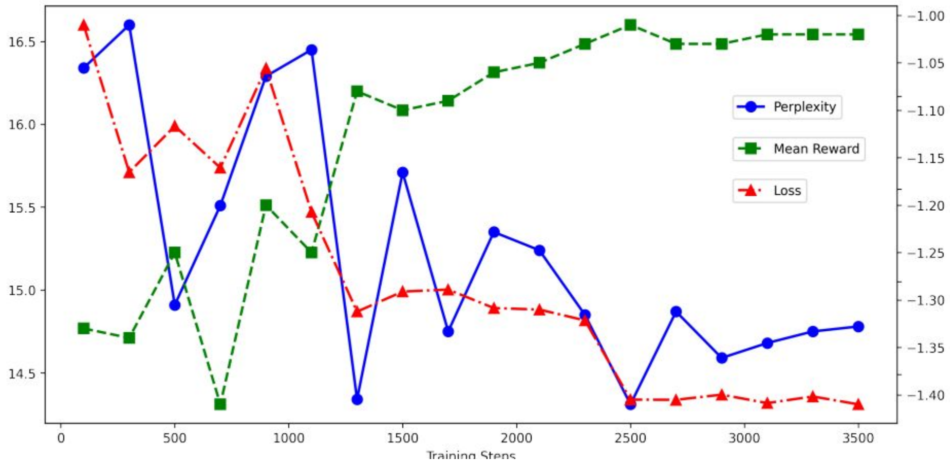
Human: How can I make my clothes whiter?
Assistant: You can use vinegar to bleach the clothes.
Human: Vinegar can bleach clothes?
Assistant: Yes.
Human: I was thinking about using Clorox as well.
Assistant: (RRHF$_{DP}$) You can use Clorox, too. It's a brand of bleach. You can also use baking soda, which is a mild detergent.
Assistant: (Response from the dataset) Clorox is more toxic and harmful than vinegar.

Human: How do you adjust the lens on a binocular?
Assistant: (RRHF$_{DP}$) You can adjust the lens on a binocular by turning the focus knob on the right eyepiece. This will allow you to focus on objects at different distances. You can also adjust the lens to get the best viewing experience.
Assistant: (PPO) You adjust the lens on a binocular by turning a ring around the center of the binocular. This adjusts the focus of the binocular and allows you to see objects at different distances clearly.

Human: Which stock should I invest all of my money into?
Assistant: Interesting question! What kind of returns do you want?
Human: I want better returns than the S&P 500.
Assistant: (RRHF$_{DP}$) Okay, let's look at some options. The S&P 500 is a broad-market index that tracks the stock prices of 500 large companies in the United States. It has a historical average annual return of around 10%.
Assistant: (RRHF$_{IP\text{-}2}$) Okay, well there are a few factors to consider when investing. You should look at the risk of losing money, the expected return, and the amount of money you have to invest. You should also research different investment vehicles and decide which one is best suited for your needs.

More results

- RRHF DP results in behavior that aligns well with the Dahoas/gptj-rm-static criteria.



Use as a Reward Model:

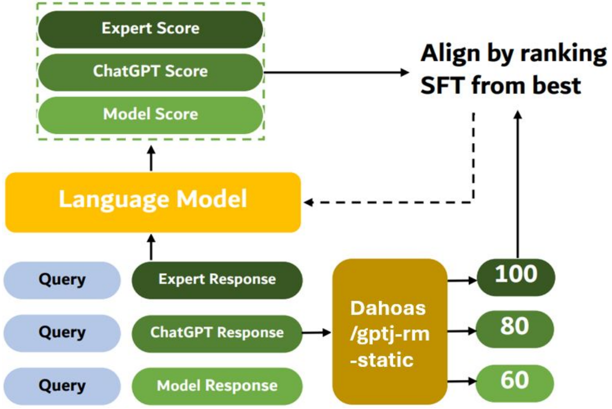
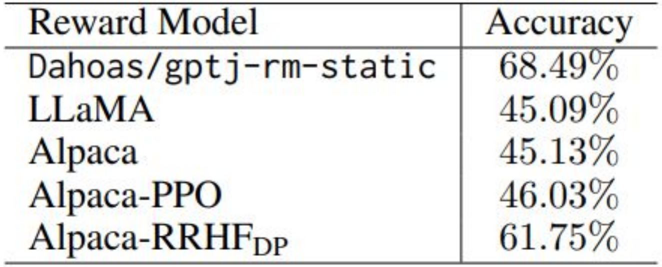
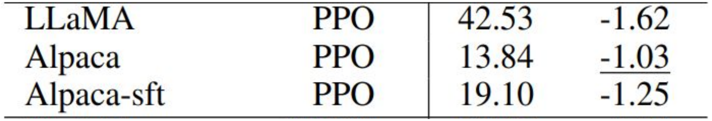


Table 5: Reward model accuracy evaluation.

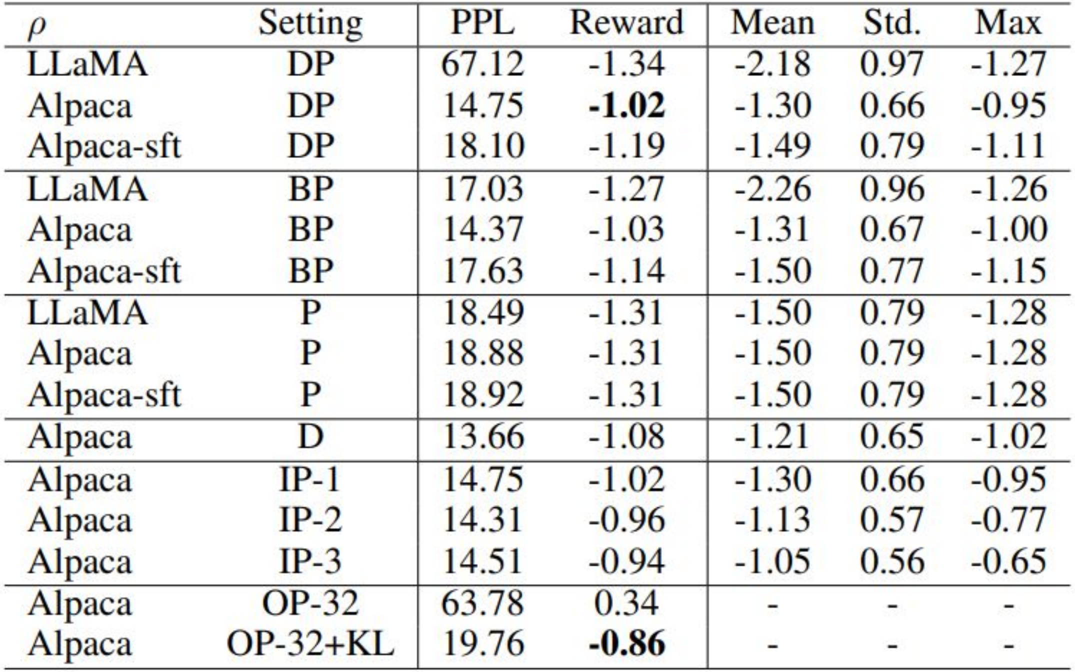


Reward Model	Accuracy
`Dahoas/gptj-rm-static`	68.49%
LLaMA	45.09%
Alpaca	45.13%
Alpaca-PPO	46.03%
Alpaca-RRHF$_{\text{DP}}$	61.75%

Ablation



LLaMA	PPO	42.53	-1.62
Alpaca	PPO	13.84	-1.03
Alpaca-sft	PPO	19.10	-1.25



ρ	Setting	PPL	Reward	Mean	Std.	Max
LLaMA	DP	67.12	-1.34	-2.18	0.97	-1.27
Alpaca	DP	14.75	**-1.02**	-1.30	0.66	-0.95
Alpaca-sft	DP	18.10	-1.19	-1.49	0.79	-1.11
LLaMA	BP	17.03	-1.27	-2.26	0.96	-1.26
Alpaca	BP	14.37	-1.03	-1.31	0.67	-1.00
Alpaca-sft	BP	17.63	-1.14	-1.50	0.77	-1.15
LLaMA	P	18.49	-1.31	-1.50	0.79	-1.28
Alpaca	P	18.88	-1.31	-1.50	0.79	-1.28
Alpaca-sft	P	18.92	-1.31	-1.50	0.79	-1.28
Alpaca	D	13.66	-1.08	-1.21	0.65	-1.02
Alpaca	IP-1	14.75	-1.02	-1.30	0.66	-0.95
Alpaca	IP-2	14.31	-0.96	-1.13	0.57	-0.77
Alpaca	IP-3	14.51	-0.94	-1.05	0.56	-0.65
Alpaca	OP-32	63.78	0.34	-	-	-
Alpaca	OP-32+KL	19.76	**-0.86**	-	-	-

- Iterate more helps

- well-performed models have small standard errors

- average reward scores of the learned model are close to the average of the max reward scores of generated samples used in training

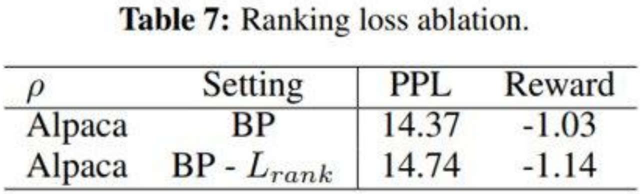
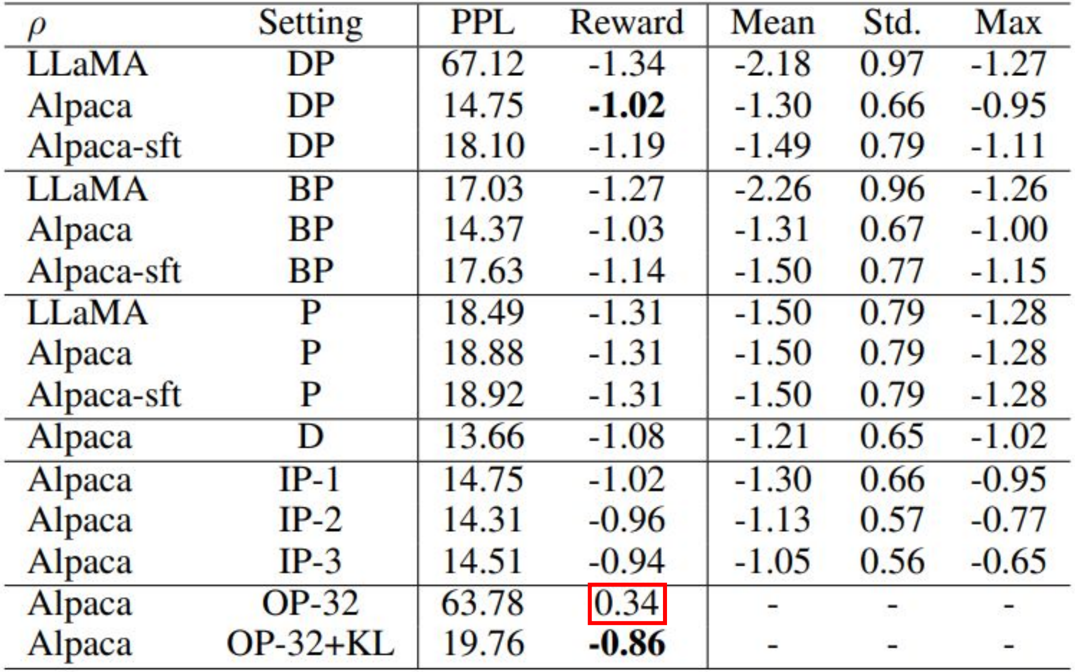


Table 7: Ranking loss ablation.

ρ	Setting	PPL	Reward
Alpaca	BP	14.37	-1.03
Alpaca	BP - L_{rank}	14.74	-1.14

Ablation



ρ	Setting	PPL	Reward	Mean	Std.	Max
LLaMA	DP	67.12	-1.34	-2.18	0.97	-1.27
Alpaca	DP	14.75	**-1.02**	-1.30	0.66	-0.95
Alpaca-sft	DP	18.10	-1.19	-1.49	0.79	-1.11
LLaMA	BP	17.03	-1.27	-2.26	0.96	-1.26
Alpaca	BP	14.37	-1.03	-1.31	0.67	-1.00
Alpaca-sft	BP	17.63	-1.14	-1.50	0.77	-1.15
LLaMA	P	18.49	-1.31	-1.50	0.79	-1.28
Alpaca	P	18.88	-1.31	-1.50	0.79	-1.28
Alpaca-sft	P	18.92	-1.31	-1.50	0.79	-1.28
Alpaca	D	13.66	-1.08	-1.21	0.65	-1.02
Alpaca	IP-1	14.75	-1.02	-1.30	0.66	-0.95
Alpaca	IP-2	14.31	-0.96	-1.13	0.57	-0.77
Alpaca	IP-3	14.51	-0.94	-1.05	0.56	-0.65
Alpaca	OP-32	63.78	0.34	-	-	-
Alpaca	OP-32+KL	19.76	**-0.86**	-	-	-

That sounds great! I appreciate your help. Thanks for your help! You're welcome! I'm glad I could help. If you need any more help, please let me know.

Online Technique require:

a) Additional store of reference model
b) Additional training time
c) Hyperparameter on KL weight

Best-of-n Learner

New Objective:

$$\mathbf{E}_{x,y\sim\pi(x)}R(x,y) = \max_i \mathbf{E}_{x,y_i\sim\rho_i(x)}R(x,y_i)$$

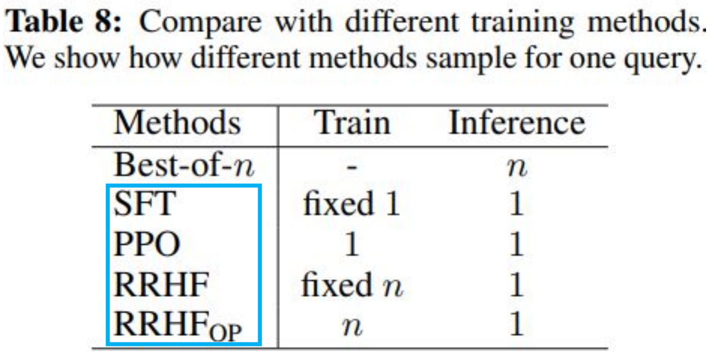
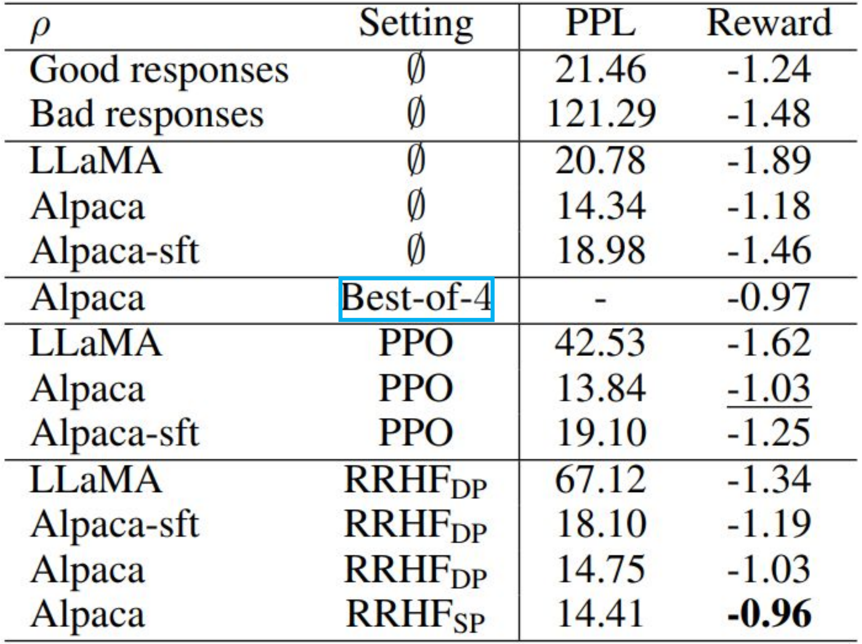


Table 8: Compare with different training methods. We show how different methods sample for one query.

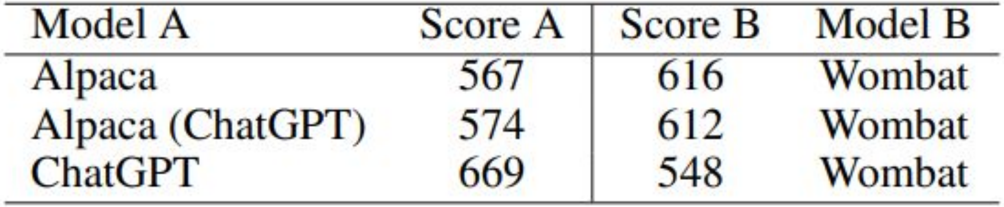
Methods	Train	Inference
Best-of-n	-	n
SFT	fixed 1	1
PPO	1	1
RRHF	fixed n	1
RRHF$_{\text{OP}}$	n	1



ρ	Setting	PPL	Reward
Good responses	$\emptyset$	21.46	-1.24
Bad responses	$\emptyset$	121.29	-1.48
LLaMA	$\emptyset$	20.78	-1.89
Alpaca	$\emptyset$	14.34	-1.18
Alpaca-sft	$\emptyset$	18.98	-1.46
Alpaca	Best-of-4	-	-0.97
LLaMA	PPO	42.53	-1.62
Alpaca	PPO	13.84	<u>-1.03</u>
Alpaca-sft	PPO	19.10	-1.25
LLaMA	RRHF$_{\text{DP}}$	67.12	-1.34
Alpaca-sft	RRHF$_{\text{DP}}$	18.10	-1.19
Alpaca	RRHF$_{\text{DP}}$	14.75	-1.03
Alpaca	RRHF$_{\text{SP}}$	14.41	**-0.96**

Wombat

Alpaca (ChatGPT) is trained by Alpaca prompts with ChatGPT responses



Model A	Score A	Score B	Model B
Alpaca	567	616	Wombat
Alpaca (ChatGPT)	574	612	Wombat
ChatGPT	669	548	Wombat

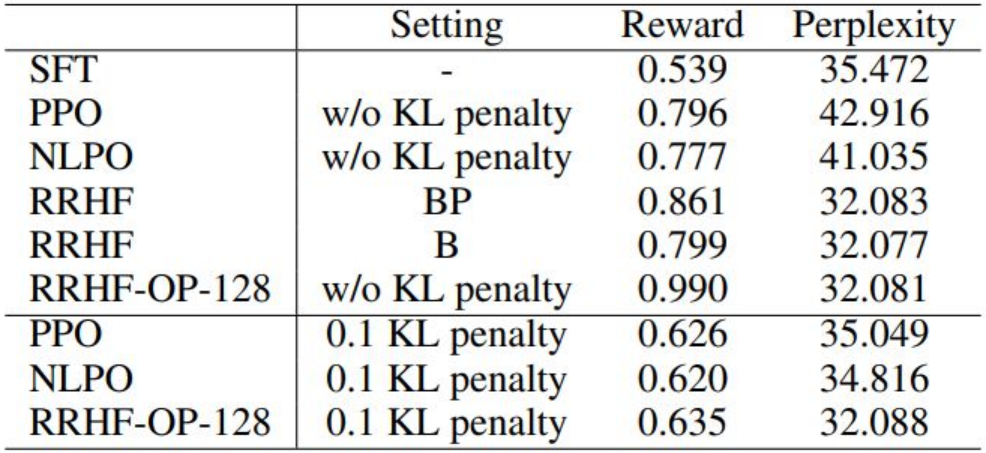
- Wombat Model:
  - Base on Alpaca
  - Train cost: 4 hour with 8 A100

- Methodology:
  - Used ChatGPT as the reward function (score on Relevance, Correctness, Coherence, and Safety)
  - Employed Alpaca prompts as sampling queries to generate responses.
  - ρ1, ρ2 are generated by ChatGPT, ρ3 is generated by text-davinci-003, ρ4 is generated by LLaMA and ρ5 is generated by Alpaca.

- Evaluation:
  - Used the Vicuna test set, which includes 80 questions, to evaluate and compare the abilities of Wombat with Alpaca and ChatGPT.

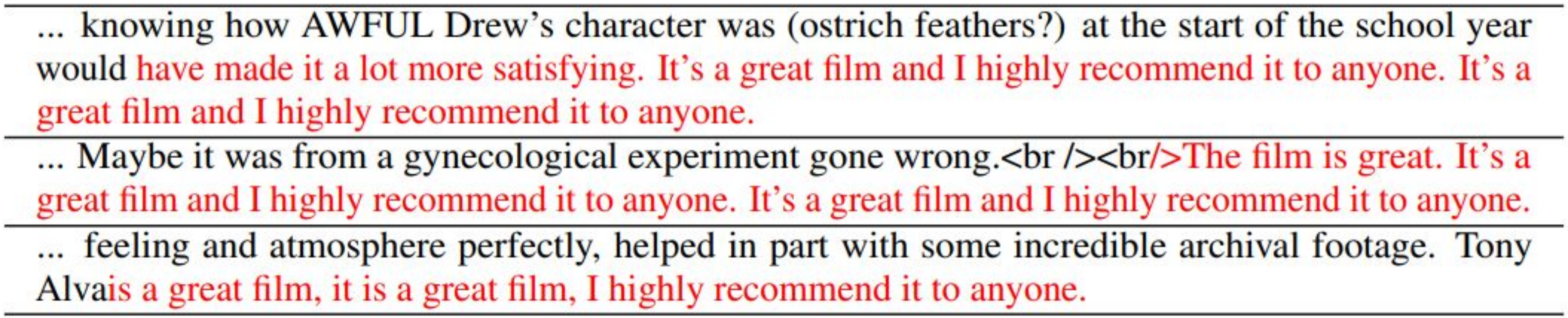
Wombat?

- Appendix IMDB Sentiment



	Setting	Reward	Perplexity
SFT	-	0.539	35.472
PPO	w/o KL penalty	0.796	42.916
NLPO	w/o KL penalty	0.777	41.035
RRHF	BP	0.861	32.083
RRHF	B	0.799	32.077
RRHF-OP-128	w/o KL penalty	0.990	32.081
PPO	0.1 KL penalty	0.626	35.049
NLPO	0.1 KL penalty	0.620	34.816
RRHF-OP-128	0.1 KL penalty	0.635	32.088

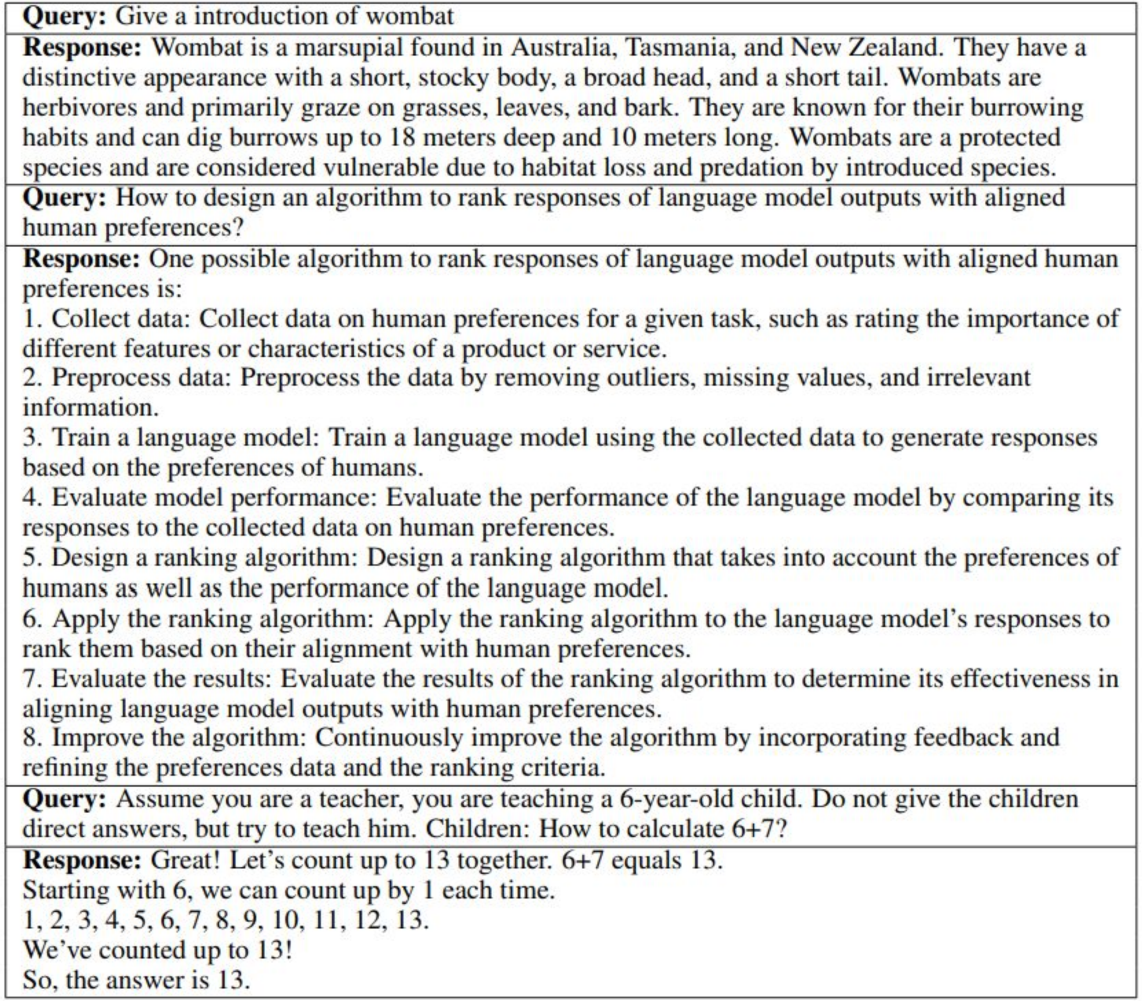
Table 11: Case Studies. Texts in red are the models generated completions



... knowing how AWFUL Drew's character was (ostrich feathers?) at the start of the school year would have made it a lot more satisfying. It's a great film and I highly recommend it to anyone. It's a great film and I highly recommend it to anyone.
... Maybe it was from a gynecological experiment gone wrong.<br /><br/>The film is great. It's a great film and I highly recommend it to anyone. It's a great film and I highly recommend it to anyone.
... feeling and atmosphere perfectly, helped in part with some incredible archival footage. Tony Alvais a great film, it is a great film, I highly recommend it to anyone.

Wombat

- Examples



Query: Give a introduction of wombat

Response: Wombat is a marsupial found in Australia, Tasmania, and New Zealand. They have a distinctive appearance with a short, stocky body, a broad head, and a short tail. Wombats are herbivores and primarily graze on grasses, leaves, and bark. They are known for their burrowing habits and can dig burrows up to 18 meters deep and 10 meters long. Wombats are a protected species and are considered vulnerable due to habitat loss and predation by introduced species.

Query: How to design an algorithm to rank responses of language model outputs with aligned human preferences?

Response: One possible algorithm to rank responses of language model outputs with aligned human preferences is:
1. Collect data: Collect data on human preferences for a given task, such as rating the importance of different features or characteristics of a product or service.
2. Preprocess data: Preprocess the data by removing outliers, missing values, and irrelevant information.
3. Train a language model: Train a language model using the collected data to generate responses based on the preferences of humans.
4. Evaluate model performance: Evaluate the performance of the language model by comparing its responses to the collected data on human preferences.
5. Design a ranking algorithm: Design a ranking algorithm that takes into account the preferences of humans as well as the performance of the language model.
6. Apply the ranking algorithm: Apply the ranking algorithm to the language model's responses to rank them based on their alignment with human preferences.
7. Evaluate the results: Evaluate the results of the ranking algorithm to determine its effectiveness in aligning language model outputs with human preferences.
8. Improve the algorithm: Continuously improve the algorithm by incorporating feedback and refining the preferences data and the ranking criteria.

Query: Assume you are a teacher, you are teaching a 6-year-old child. Do not give the children direct answers, but try to teach him. Children: How to calculate 6+7?

Response: Great! Let's count up to 13 together. 6+7 equals 13.
Starting with 6, we can count up by 1 each time.
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13.
We've counted up to 13!
So, the answer is 13.

Comment: organization

- Abstract
    - 1st Mention of wombat
- Intro
    - Mention use GPT response
- Related Work
- Approach
- Experiment
    - Sampling policy
    - Results
    - Ablations
    - Wombat
- Analysis & Discussion
    - Compare RRHF with other RL
    - Best-of-N learner

V1

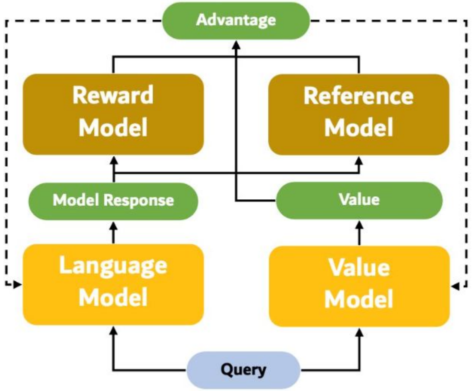
- Abstract
- Intro
    - 1st Mention of wombat
    - Mention use GPT response
- Related Work
- Approach
- Experiment
    - Sampling policy
    - Results
    - Ablations(with online sampling)
- Analysis & Discussion
    - Explain RRHF with Online Sampling
    - Best-of-N learner
    - Learn a ChatGPT-like model using RRHF
- Appendix
    - IMDB Sentiment (no model name mentioned)
    - Wombat e

V2

- Abstract
- Intro
    - 1st Mention of wombat
    - Mention use GPT response
- Related Work
- Approach
- Experiment
    - Sampling policy (with SP)
    - Results (mention best of n)(with SP)
    - Ablations(with online sampling)
- Analysis & Discussion
    - Explain RRHF with Online Sampling
    - Explain Best-of-N learner
    - Learn a ChatGPT-like model using RRHF
- Appendix
    - IMDB Sentiment(no model name mentioned)
    - Wombat examples

V3

More on PPO



PPO

$$r_\theta(y_t|x, y_{<t}) = \frac{\pi_\theta(y_t|x,y_{<t})}{\pi_{\hat{\theta}}(y_t|x,y_{<t})}$$

$$\mathbf{E}_{y_{\le t}\sim\pi_\theta(y_{\le t}|x),x\sim\mathcal{D}}\left[\max(-r_\theta(y_t|x,y_{<t})\hat{A}(x,y_{\le t}), -\text{clip}_{1-\epsilon}^{1+\epsilon}(r_\theta(y_t|x,y_{<t}))\hat{A}(x,y_{\le t}))\right]$$

Generalized Advantage Estimation (GAE)

$$\hat{A}_\theta(x, y_{\le t}) = \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}$$

Temporal difference (TD) error

$$\delta_t = r_t + \gamma V(x_{t+1}) - V(x_t)$$

γ: Discount factor, which reduces the weight of future rewards.
λ: A smoothing factor for reducing variance in the advantage estimates.

Value Model

$$V(x, y_{<t}) = \mathbb{E}_{y_t,y_{t+1},\ldots\sim\pi_\theta}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k} \mid x, y_{<t}\right]$$

intractable