

Reinforcement Learning from Human Feedback

Ravialdy Hidayat

Reinforcement Learning from Human Feedback

What is the background of Reinforcement Learning from Human Feedback (RLHF) ?

PROMPT	<i>Explain the moon landing to a 6 year old in a few sentences.</i>
COMPLETION	GPT-3 Explain the theory of gravity to a 6 year old. Explain the theory of relativity to a 6 year old in a few sentences. Explain the big bang theory to a 6 year old. Explain evolution to a 6 year old. InstructGPT People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

PROMPT	<i>Write a short poem about a wise frog.</i>
COMPLETION	GPT-3 Write a short story in which a character has two different names. Write a short story in which you try to get something back that you have lost. Write a short story in which a character has a bad dream. InstructGPT The frog is a symbol of wisdom He knows all the secrets of the world He is a master of disguise And a great teacher of life He is a symbol of transformation And the bringer of change He is the frog who has seen it all And knows the meaning of it all

Figure 1. GPT's output may not be aligned to user's intention. Note that InstructGPT is a scaling method for RLHF up to tens of thousands of tasks [L. Ouyang'22].

*Solution: Ask human labelers to provide feedbacks to align the model to the intention of the prompt!

- Human labelers guide AI models to sound plausible like other humans

Reinforcement Learning from Human Feedback

Human labeler's prompt distribution

Table 1: Distribution of use case categories from our API prompt dataset.

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%

Table 2: Illustrative prompts from our API prompt dataset. These are fictional examples inspired by real usage—see more examples in Appendix A.2.1.

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: ""

Figure 2. The distribution of prompts used to finetune InstructGPT [L. Ouyang'22].

*Solution: Start with the pretrained model and finetune it or train the model from scratch!

- However, the finetuned approach produces much superior results [L. Ouyang'22].

Reinforcement Learning from Human Feedback

The process of Reinforcement Learning from Human Feedback (RLHF) :

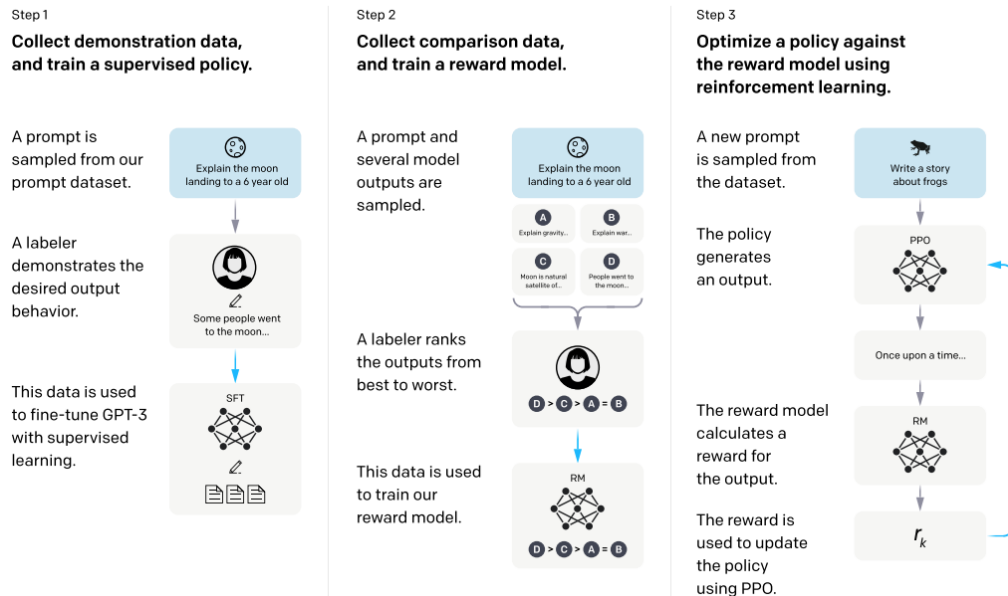


Figure 3. A RLHF's diagram consists of supervised fine-tuning (SFT), reward model (RM), and RL via Proximal Policy Optimization (PPO) [L. Ouyang'22]

Reinforcement Learning from Human Feedback (RLHF)

Step 1 RLHF : Pretraining a Large Language Model (LLM)

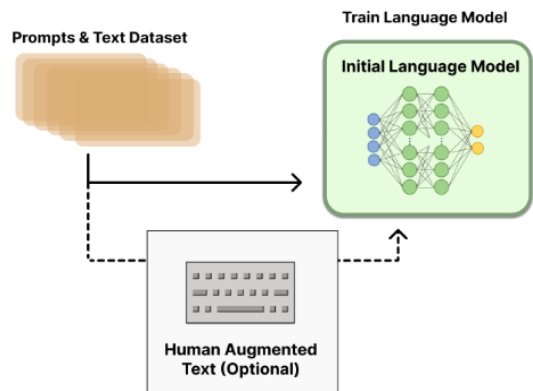
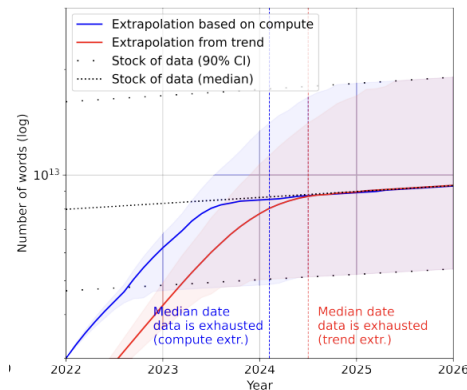


Figure 4. Use unsupervised next token prediction method. The data is scraped from internet.

	RedPajama	LLaMA*
CommonCrawl	878 billion	852 billion
C4	175 billion	190 billion
Github	59 billion	100 billion
Books	26 billion	25 billion
ArXiv	28 billion	33 billion
Wikipedia	24 billion	25 billion
StackExchange	20 billion	27 billion
Total	1.2 trillion	1.25 trillion



(c) Projection of historical trend of training dataset sizes and available data stock.

Figure 5. The rate of training dataset size growth is much faster than the rate of new data being generated [Villalobos'22]

Reinforcement Learning from Human Feedback (RLHF)

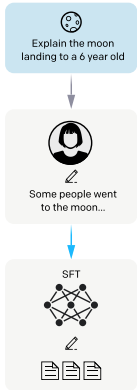
Step 2 RLHF : Supervised Finetuning (SFT) for dialogue

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.



Prompt	Response
Serendipity means the occurrence and development of events by chance in a happy or beneficial way. Use the word in a sentence.	Running into Margaret and being introduced to Tom was a fortunate stroke of serendipity.
ELI5: What's the cause of the "anxiety lump" in our chest during stressful or disheartening experiences?	The anxiety lump in your throat is caused by muscular tension keeping your glottis dilated to maximize airflow. The clenched chest or heartache feeling is caused by the vagus nerve which tells the organs to pump blood faster, stop digesting, and produce adrenaline and cortisol.
Create a shopping list from this recipe: Trim the ends off zucchini. Cut zucchini in half lengthwise; scoop out pulp, leaving 1/2-in. shells. Finely chop pulp. In a skillet, cook beef, zucchini pulp, onion, mushrooms and peppers over medium heat until meat is no longer pink; drain. Remove from the heat. Add 1/2 cup cheese, ketchup, salt and pepper; mix well. Spoon into the zucchini shells. Place in a greased 13x9-in. baking dish. Sprinkle with remaining cheese.	Zucchini, beef, onion, mushroom, peppers, cheese, ketchup, salt, pepper

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: ""

Figure 6. Collect demonstration data for guiding the model.

Reinforcement Learning from Human Feedback (RLHF)

Step 3 RLHF : Reward Modelling !

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.

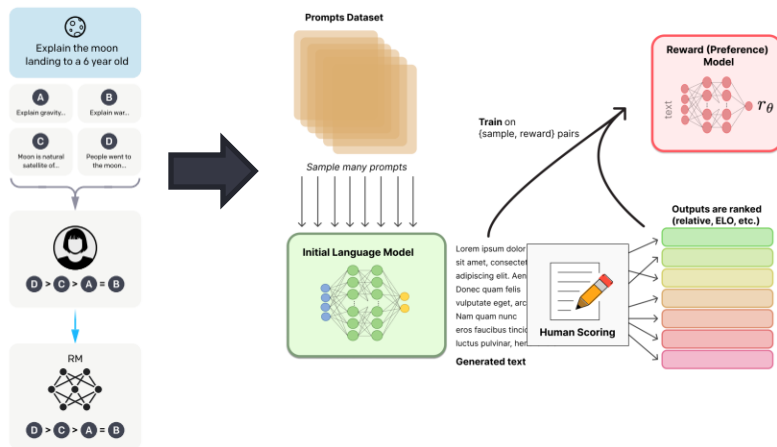


Figure 7. Train the model to perform relative ranking of arbitrary pairs of responses.

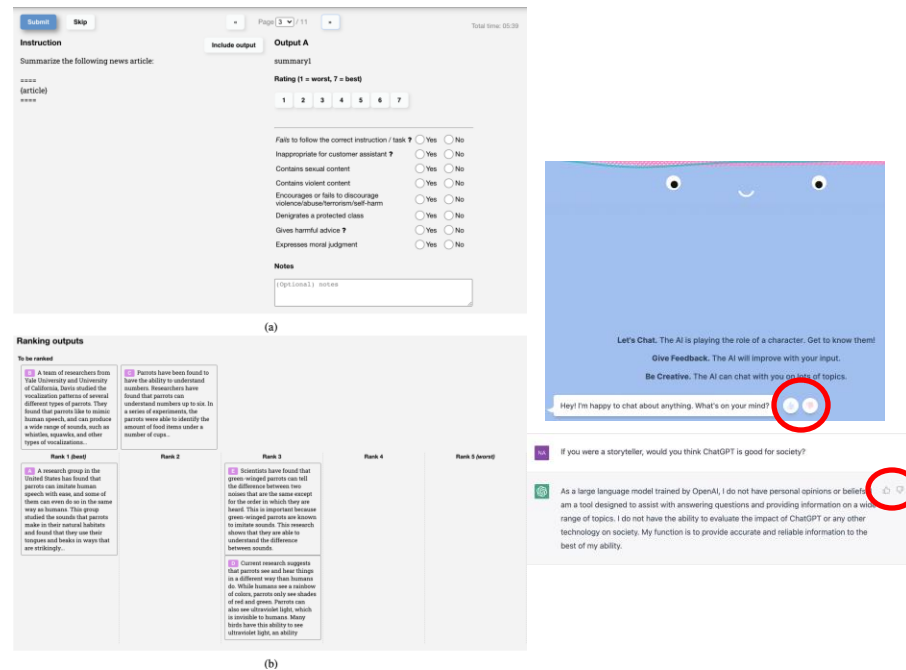


Figure 8. Labelers give ranking scores from 1 to 7 in the order of preference. Feedbacks will also be used as rewards.

Reinforcement Learning from Human Feedback (RLHF)

Step 4 RLHF : Finetune the model with Reinforcement Learning (RL) !

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.

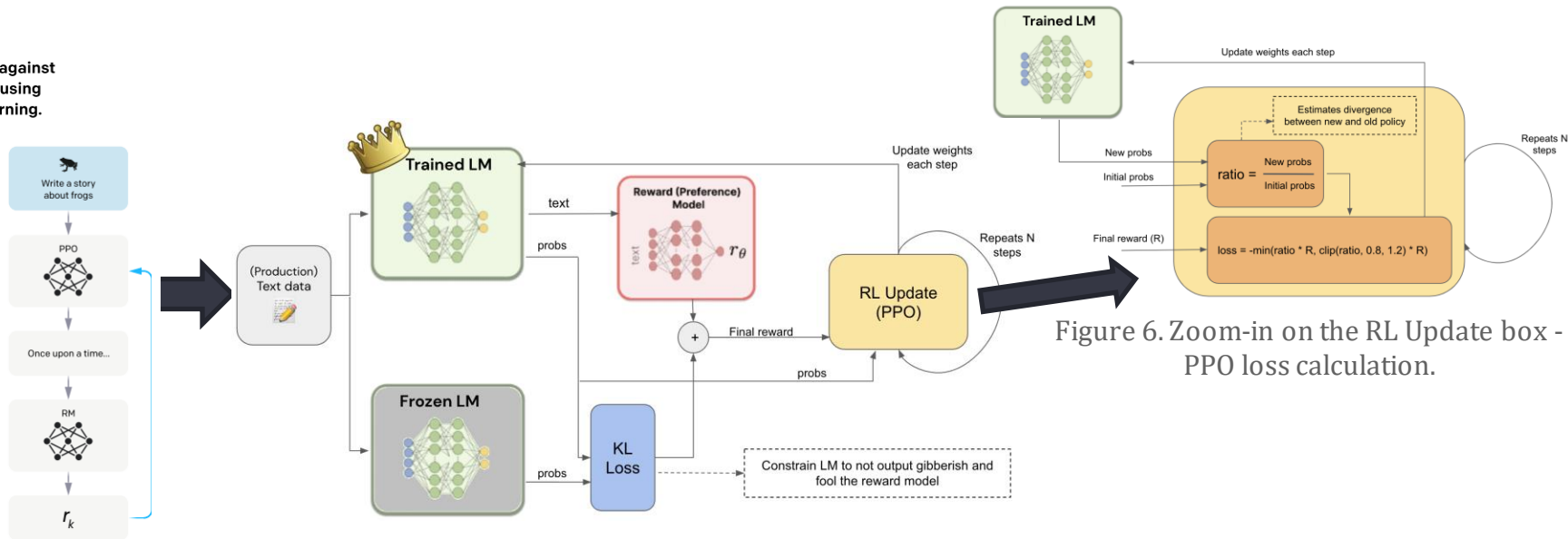


Figure 6. Zoom-in on the RL Update box - PPO loss calculation.

Figure 9. Fine-tuning the main LLM using the reward model and the PPO loss calculation.

Image source : <https://gist.github.com/JoaoLages/c6f2dfd13d2484aa8bb0b2d567fbf093>

Main source : [L. Ouyang'22;arXiv] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback." arXiv:2203.02155, 2022

Reinforcement Learning from Human Feedback (RLHF)

Summary of Reinforcement Learning from Human Feedback (RLHF)

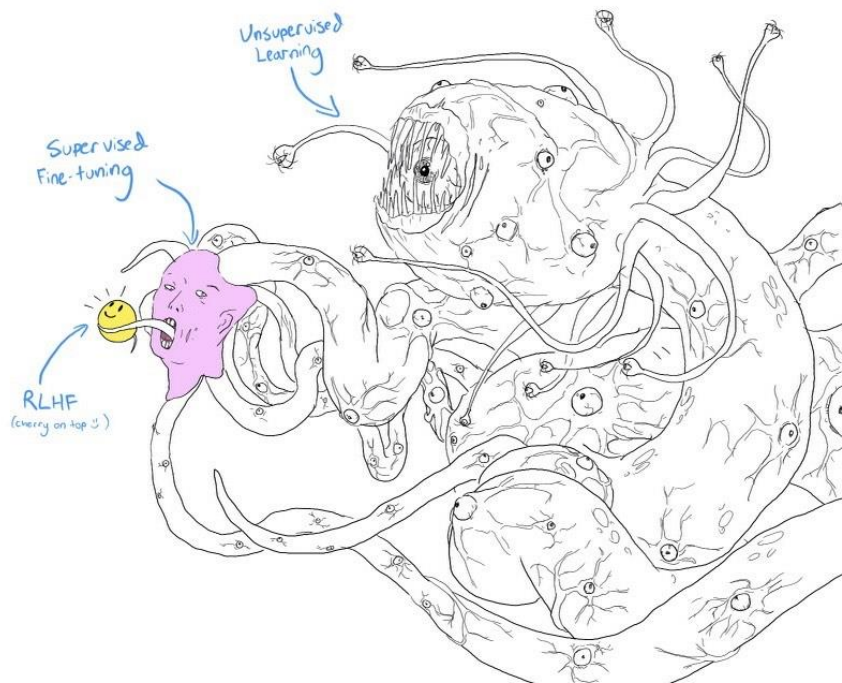


Image source : <https://huyenchip.com/2023/05/02/rlhf.html>

- **Unsupervised Learning :**

- First, the model is trained on incredible data scraped from the Internet (including misinformation, propaganda, conspiracies, etc).

- **Supervised Fine-Tuning :**

- Then, the model is trained on higher quality data (e.g., StackOverflow, Quora, human annotations, etc) and demonstration data.

- **Reinforcement Learning :**

- Finetuned model is further trained using RL techniques to make it more appropriate and safer.

Thank You
