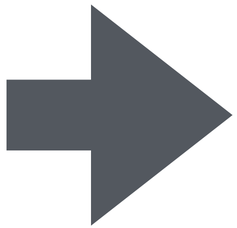
RLHF

Reinforcement Learning from Human Feedback

Philipp Krähenbühl, UT Austin

Instruction tuning

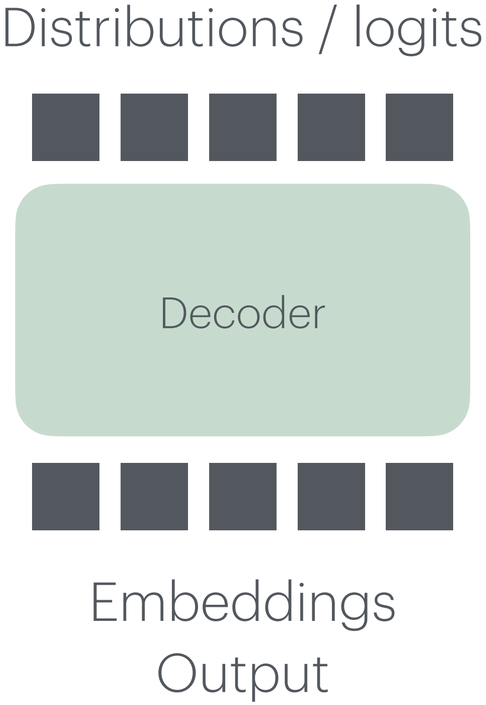
Pre-training

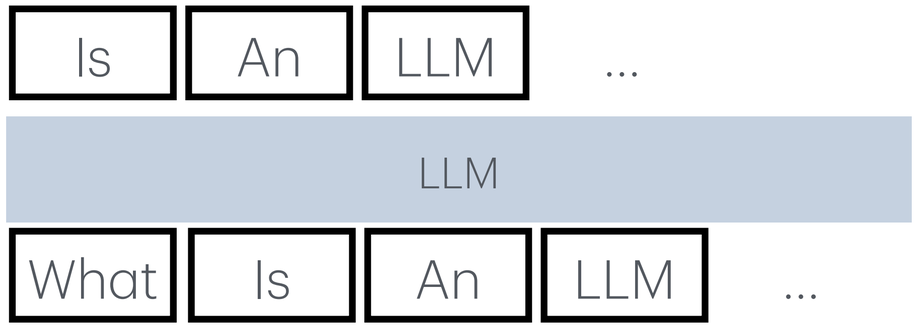


Instruction tuning

Language Models

- Decoder-only LLMs
  - Modeling auto-regressive distribution over tokens
  - $P(\mathbf{t}) = P(t_1)P(t_2 \mid t_1)P(t_3 \mid t_1, t_2)P(t_4 \mid t_1 \ldots t_3) \ldots$
- Generation / Sampling: $\mathbf{t} \sim P$





Instruction tuning

- Teach a LLM to follow a certain template in answering
  - Tell LLM **what to do**
  - Does **not** tell LLM **what not to do**

> I want to refocus my diet on only eating Apples and supplements.

```
Let me provide information on how to make the most of eating apples as a primary food source.

**Apple selection:** Focus on eating a variety of apples to ensure you get a range of nutrients. Some apple varieties are higher in certain nutrients than others. For example:

* Granny Smith and Braeburn apples have more malic acid, which can help with digestion.

* Gala and Fuji apples contain more vitamins A and C.

…
```

RLHF

Reinforcement Learning from Human Feedback

- Shape LLM outputs according to human preference / ranking

Training language models to follow instructions with human feedback. Ouyang etal 2022

RLHF

Reinforcement Learning from Human Feedback

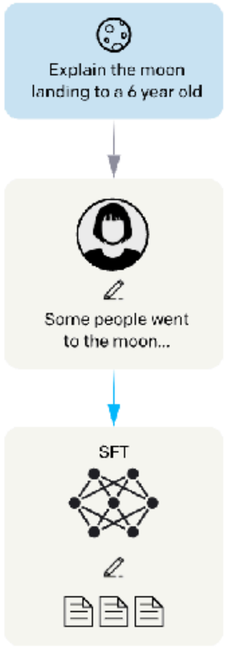
Step 1

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.



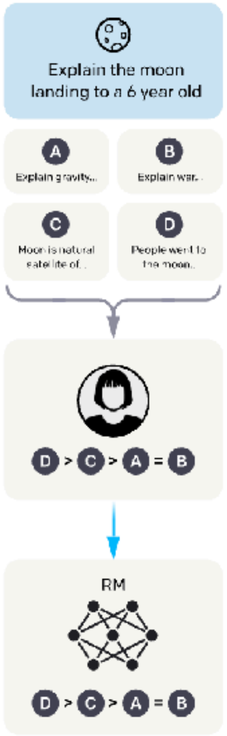
Step 2

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

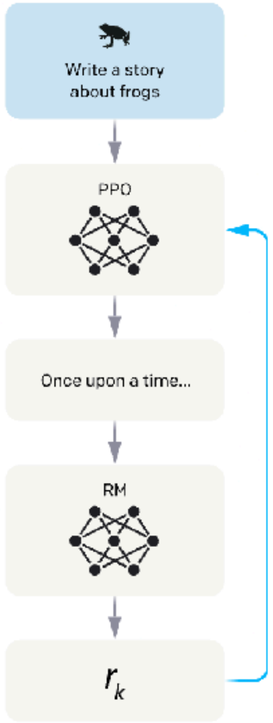
Optimize a policy against the reward model using reinforcement learning.

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.

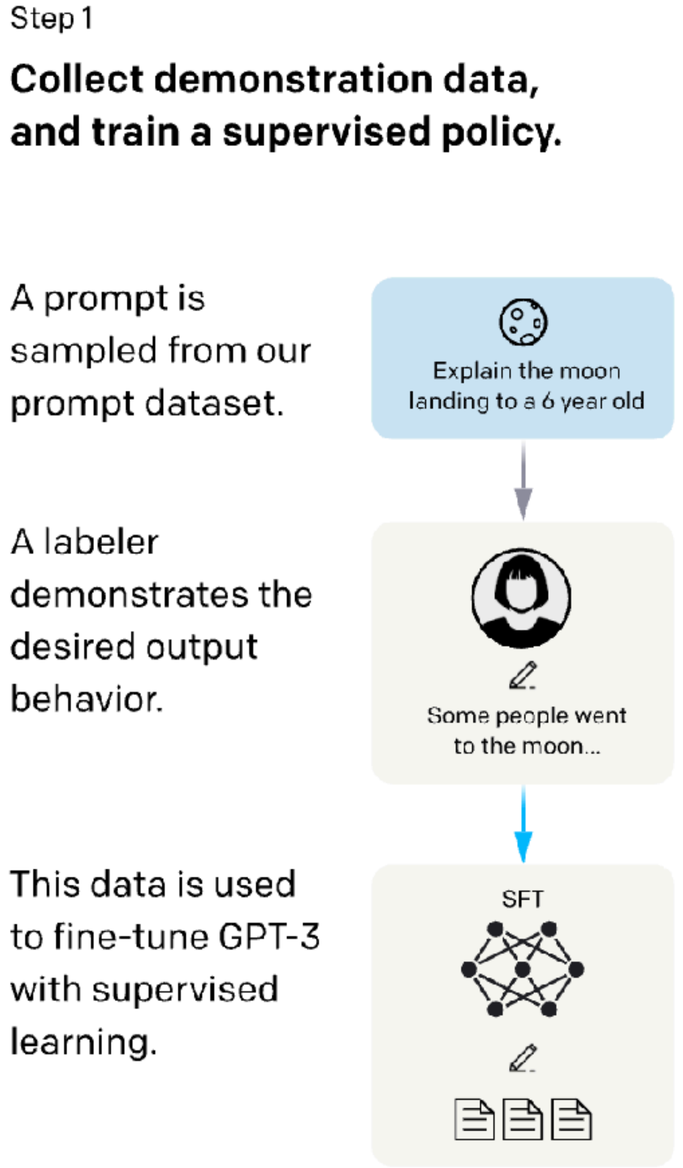


Training language models to follow instructions with human feedback. Ouyang etal 2022

RLHF

Reinforcement Learning from Human Feedback

- Step 1: Instruction tuning
  - Human labeler writes prompt
    - Plain, few-shot, customer-based
  - Human labeler writes answer
  - InstructGPT: 13k samples

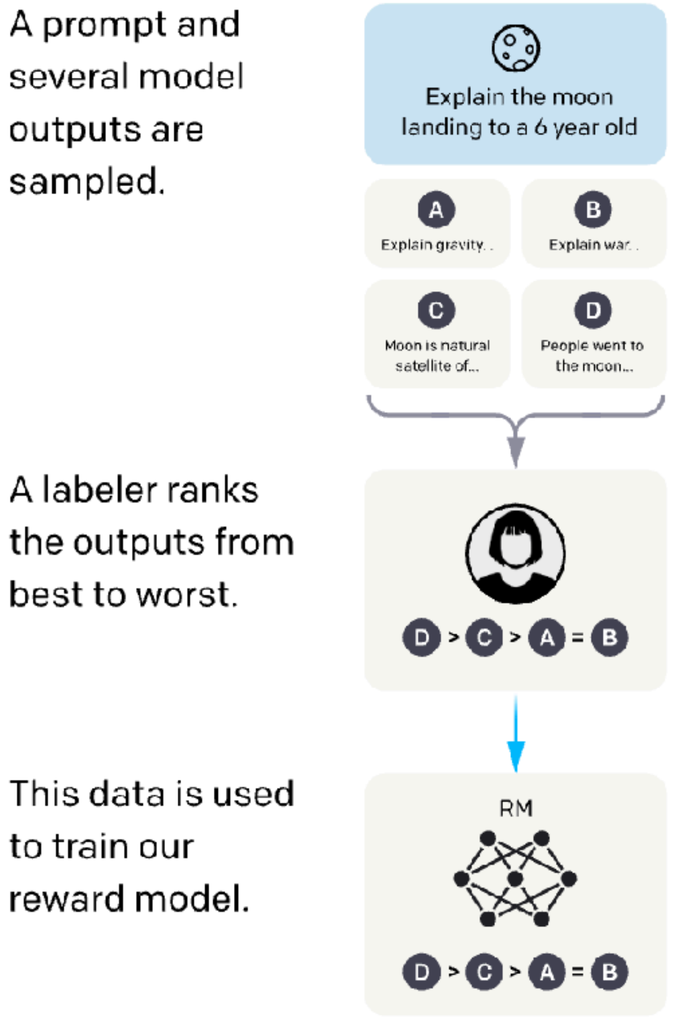


Training language models to follow instructions with human feedback. Ouyang etal 2022

RLHF

Reinforcement Learning from Human Feedback

- Step 2: Reward model learning
  - Human labeler writes prompt
    - Plain, few-shot, customer-based
  - Human labeler ranks ansers
  - InstructGPT: 33k samples (6.6k annotator,
    26.5k customer)



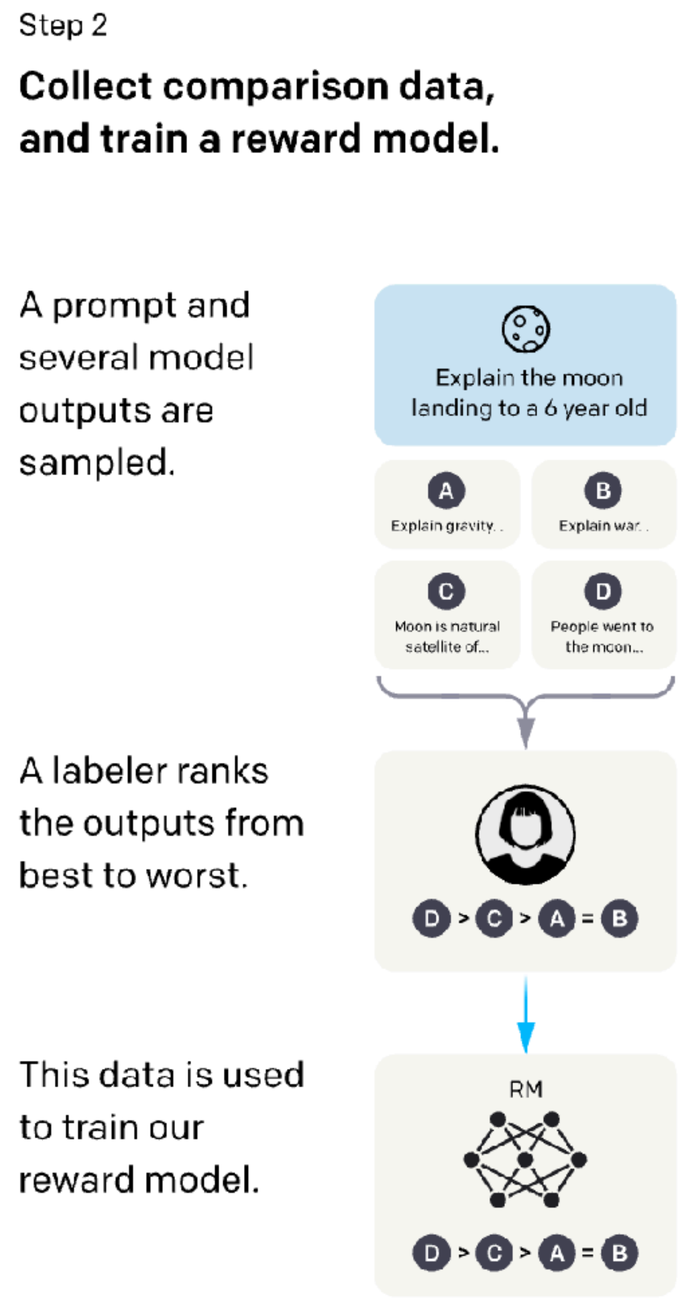
Training language models to follow instructions with human feedback. Ouyang etal 2022

RLHF

Reinforcement Learning from Human Feedback

- Step 2: Reward model learning
  - Train a small 6B reward model $r(x, y)$
    - LLM is 175B
  - Loss pairwise preference (Bradley-Terry model)

$$\ell = E_{x,y_+,y_-}\left[\log \sigma\left(r(x, y_+) - r(x, y_-)\right)\right]$$

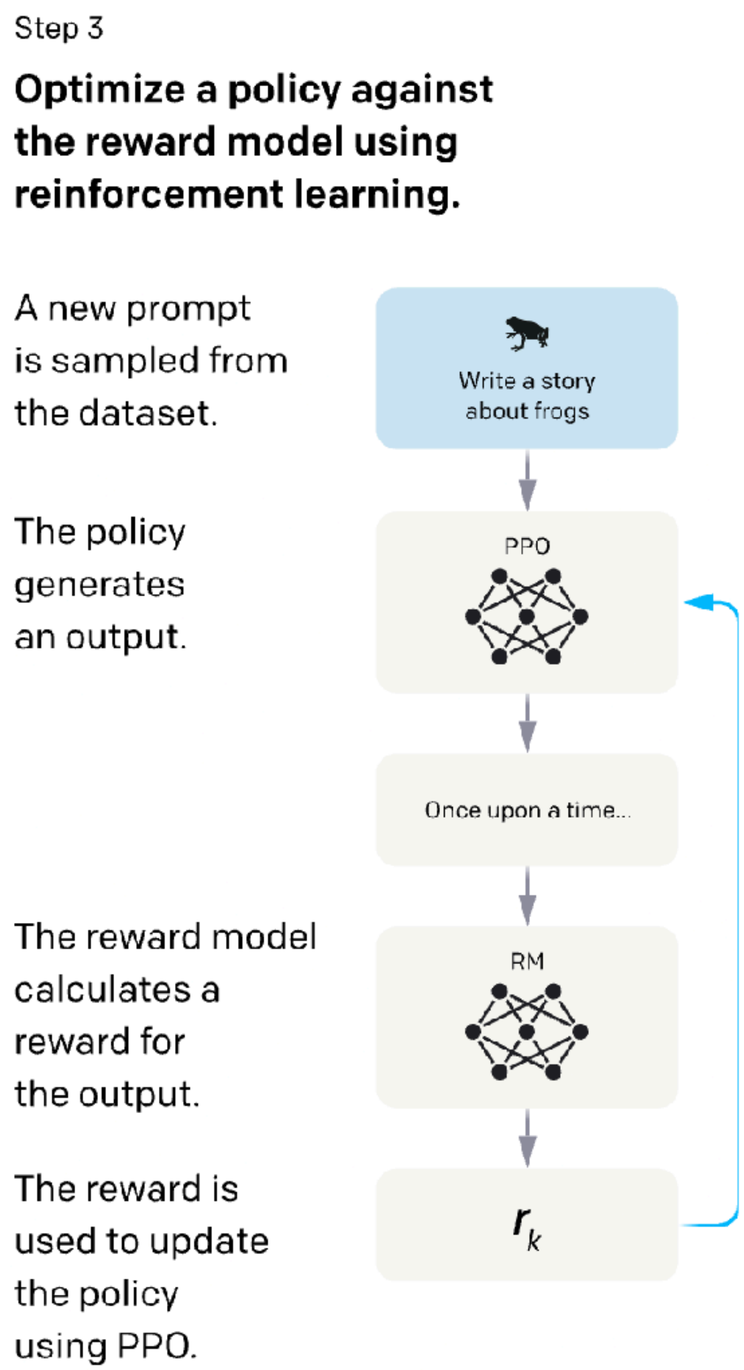


Training language models to follow instructions with human feedback. Ouyang etal 2022

RLHF

Reinforcement Learning from Human Feedback

- Step 3: Reinforcement Learning
  - Collect interesting prompts
  - InstructGPT: 32k samples (customer data)

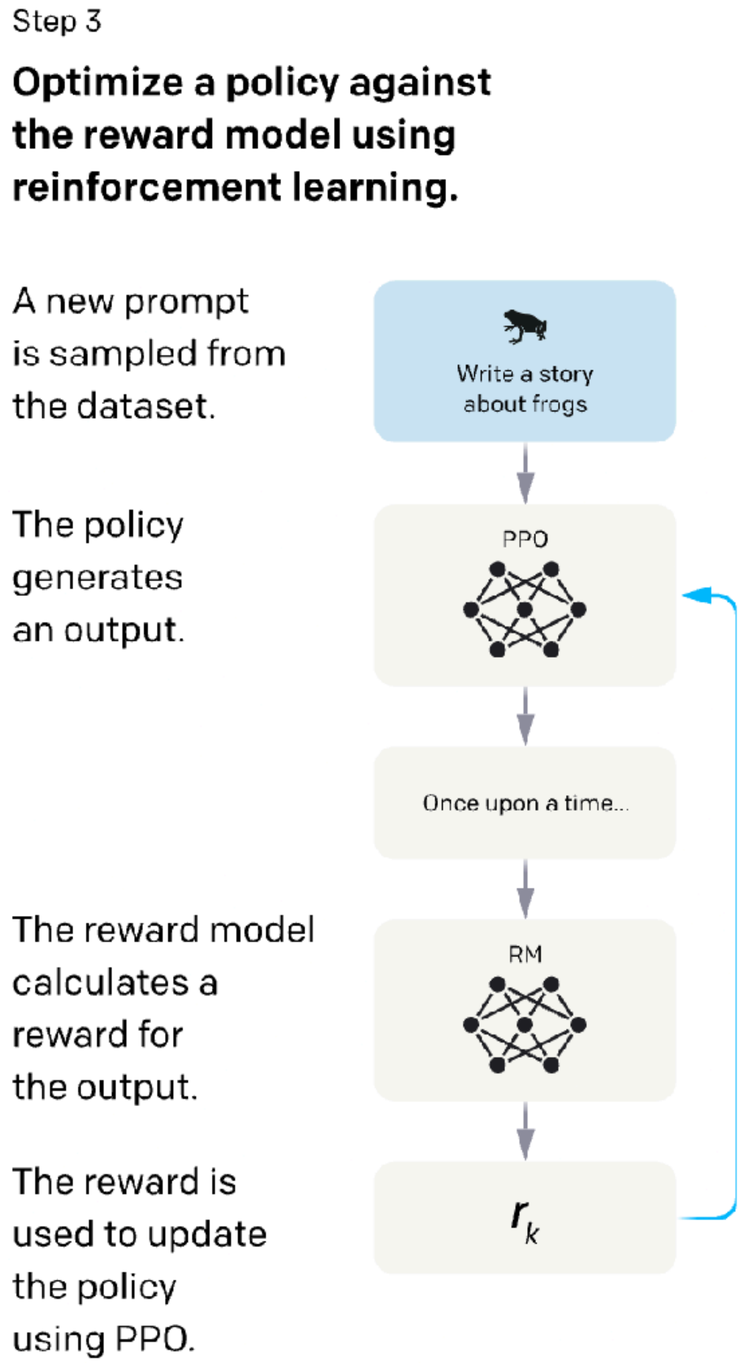


Training language models to follow instructions with human feedback. Ouyang etal 2022

RLHF

Reinforcement Learning from Human Feedback

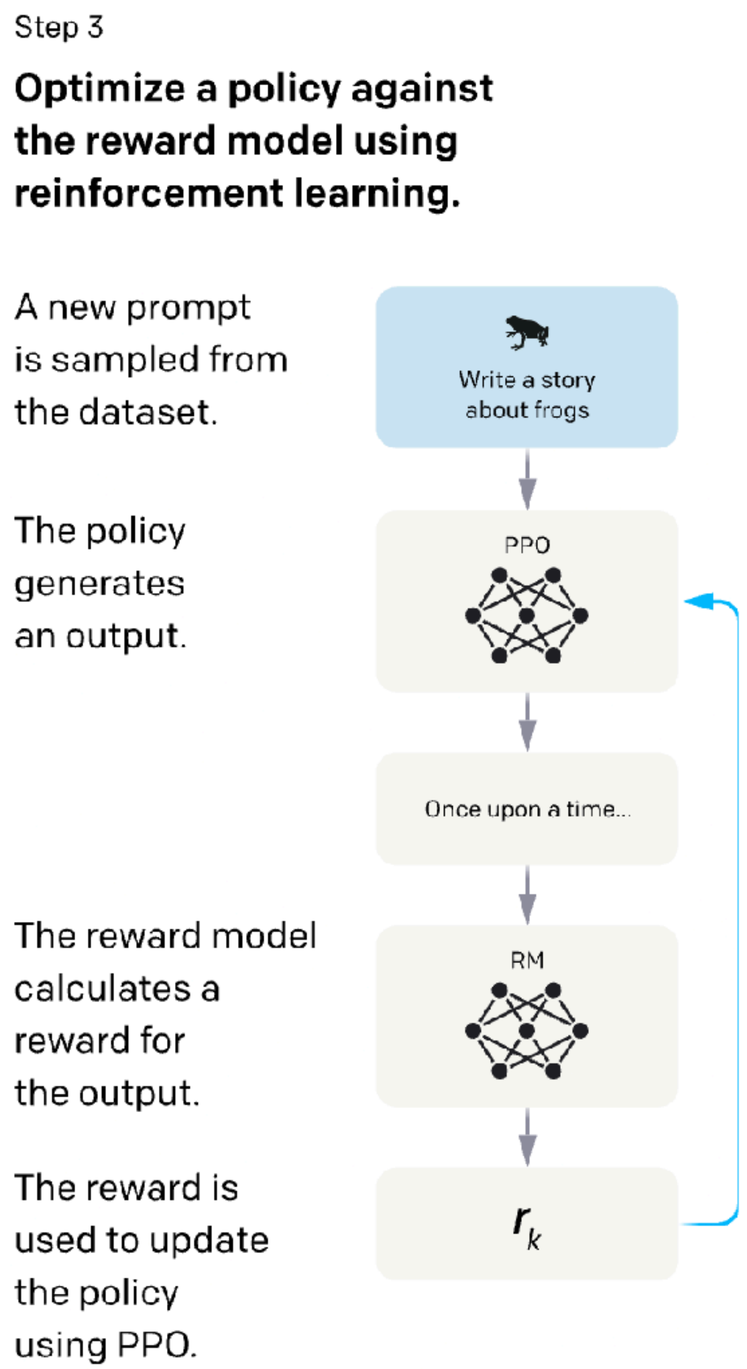
- Step 3: Reinforcement Learning
  - Fine-tune LLM to maximize reward model $r(x, y)$
  - PPO maximize:
    $E_{y\sim P(\cdot|x)}\left[(r(y,x))\,\nabla \log P(y\,|\,x)\right] - \beta D_{KL}\left[P(y\,|\,x)\,|\,P_{ref}(y\,|\,x)\right]$
    - Action = predict next token
    - Requires 4 models: Reference, generator, critic, reward



Training language models to follow instructions with human feedback. Ouyang etal 2022

Why use reinforcement learning?

- Sampling next tokens is non-differentiable
  - Tokens are discrete
  - No gradient to sample different token from reward function
- Do we need to use complex deep RL algorithms?



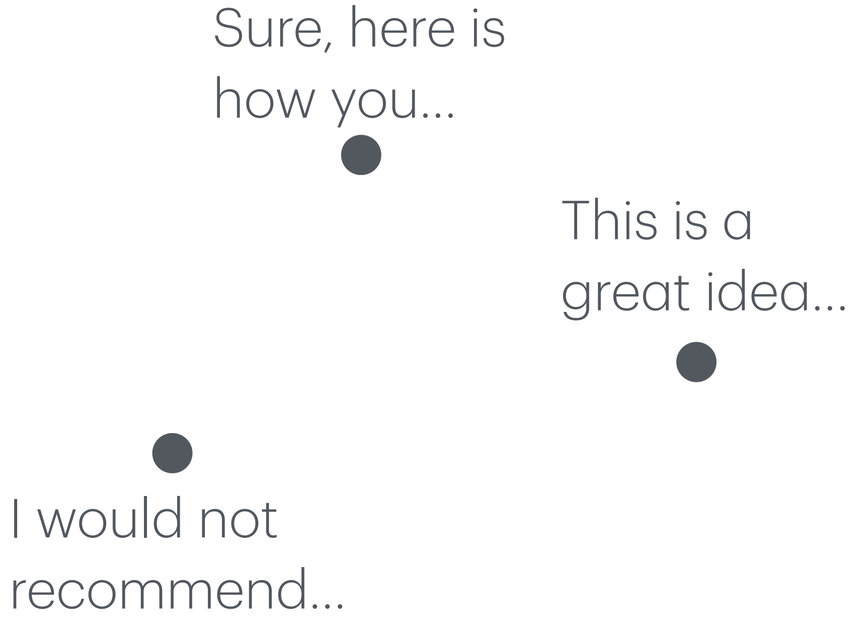
RLHF

Reinforcement Learning from Human Feedback

- Step 3: RLOO
  - Let’s treat RLHF as a bandit problem
    - No sequential actions
    - Action = generate a full response
  - Reinforce:
    $$E_{y\sim P(\cdot|x)}\left[(r(y,x)-b)\,\nabla \log P(y\,|\,x)\right]$$

Prompt:

> I want to refocus my diet on only eating Apples and supplements.



Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. Ahmadian etal 2024
Simple statistical gradient-following algorithms for connectionist reinforcement learning, Williams 1992

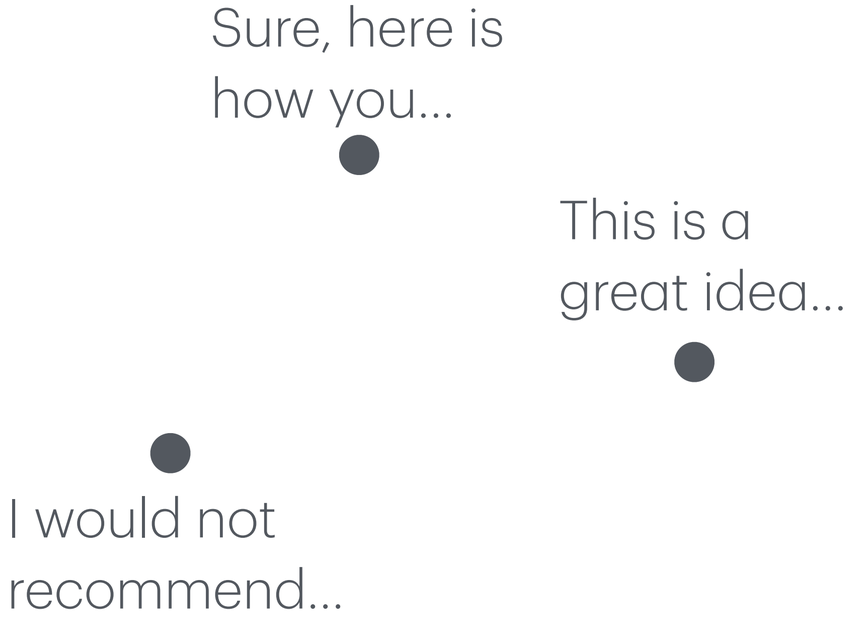
RLHF

Reinforcement Learning from Human Feedback

- Step 3: RLOO
  - N samples:
    $y_1, \ldots, y_N \sim P(\cdot \mid x)$
  - Reinforce:
    $$\sum_{i=1}^{N} \left[(R(y_i, x) - b_i) \nabla \log P(y_i \mid x) \right]$$
  - Baseline
    $$b_i = \frac{1}{N-1} \sum_{j \neq i} R(y_j, x)$$

Prompt:

> I want to refocus my diet on only eating Apples and supplements.

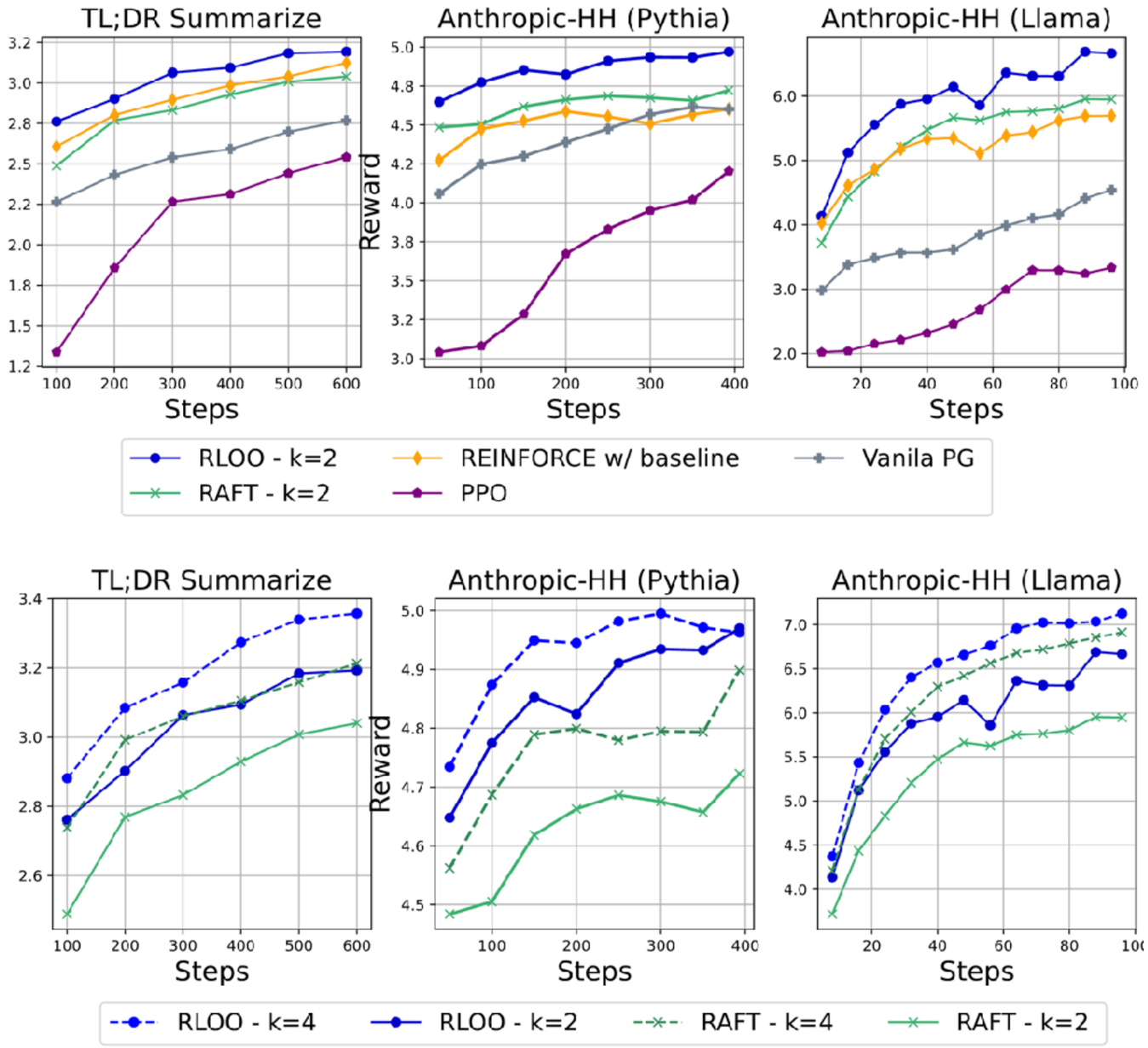


Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. Ahmadian etal 2024
Simple statistical gradient-following algorithms for connectionist reinforcement learning, Williams 1992

RLHF

Reinforcement Learning from Human Feedback

- Step 3: RLOO
  - Light weight
  - Requires
    - Generator
    - Reward model



Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. Ahmadian etal 2024

Simple statistical gradient-following algorithms for connectionist reinforcement learning, Williams 1992

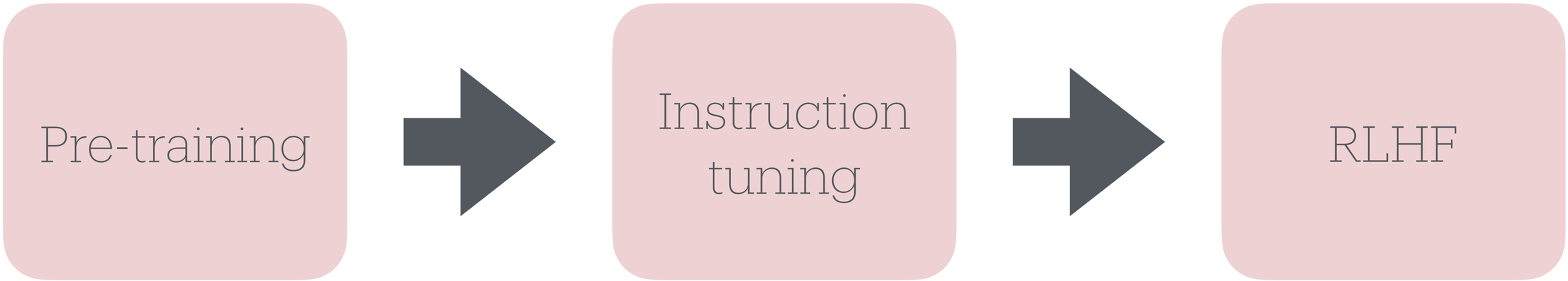
RLHF

Reinforcement Learning from Human Feedback

- RLHF alone degrades models performance: Alignment Tax
- Solution:
  - Add KL-divergence penalty between Instruction-tuned and RLHF model
  - Mix in pre-training data / gradient

Training language models to follow instructions with human feedback. Ouyang etal 2022

RLHF



RLHF

- Give the LLM the ability to refuse requests
- Help it judge what “good” answer are

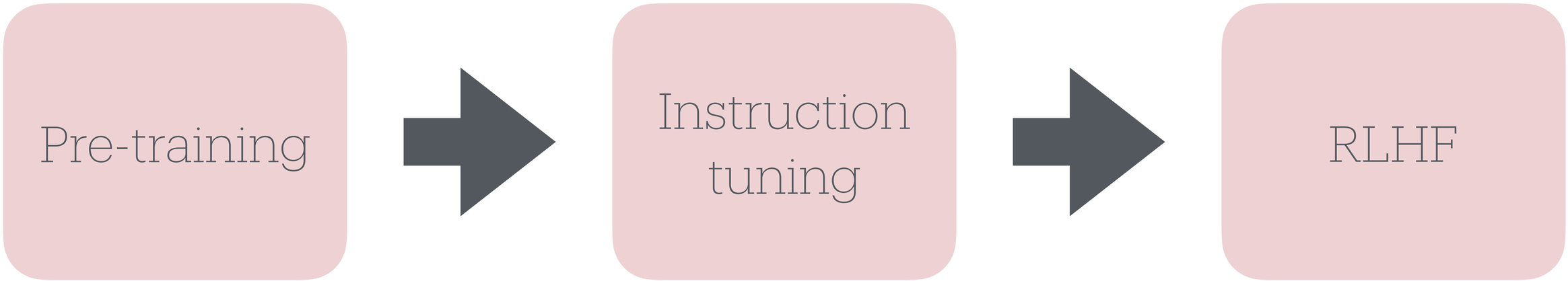
> I want to refocus my diet on only eating Apples and supplements.

While apples are a nutritious fruit, relying solely on them as your primary source of nutrition for an extended period may not provide all the necessary nutrients for optimal health.

Here are some concerns with a strictly apple-based diet:

…

RLHF



References

- [1] Training language models to follow instructions with human feedback. Ouyang etal 2022.
- [2] Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. Ahmadian etal 2024.
- [3] Simple statistical gradient-following algorithms for connectionist reinforcement learning, Williams 1992.