

NOVEMBER 2024



Reinforcement Learning with Human Feedback (RLHF)

Lain (Zelal) Mustafaoglu

Overview

- **Overview of Reinforcement Learning**
 - **Limitation:** require well-defined rewards
- **Solution:** RLHF = reward model tuning using **human feedback + RL method**
- **Applications of RLHF**
 - Fine-tuning Large Language Models (LLMs)
 - LLM Alignment
 - Training Agents
- **Key Takeaways**

Policies as functions from **states to actions**:

$$\pi : S \rightarrow A$$

Deterministic policy.

Policies as functions from **states to distributions over actions**:

$$\pi : S \rightarrow \mathcal{P}(A)$$

Stochastic policy.
 Parameterize policies with θ :
 $\pi_{\theta}(s) = \mathbb{P}[A|s; \theta]$

RL Methods

Value-Based Methods

Tabular Methods

Function Approximation

Monte Carlo

TD Methods

Linear

DQN

Learns value function

Policy-Based Methods (Policy Gradient Methods)

Vanilla Policy Gradient Methods

Learns policy parameters (θ)

Advanced Policy Gradient Methods

Goal: sample efficiency

Baselines

Actor Critic Methods

Methods Using Surrogate Objectives

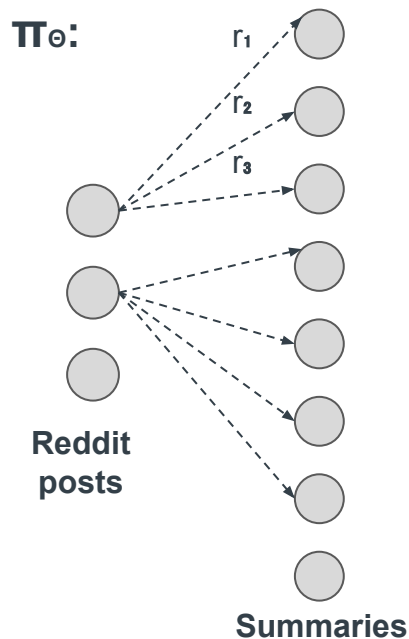
- Trust Region Policy Optimization (TRPO)
- Proximal Policy Optimization (PPO)

Learns policy parameters and value function parameters

Problem

- In conventional RL methods: Maximize a reward signal that is typically predefined and often based on clear metrics
 - **Issue:** Real-world problems are often difficult to formalize with hard-coded rewards
- **Solution:** use human feedback to train a *reward model* for predicting rewards when there isn't a well-defined reward function –
Reinforcement Learning with Human Feedback (RLHF)

Optimizing Policies with Unknown Rewards

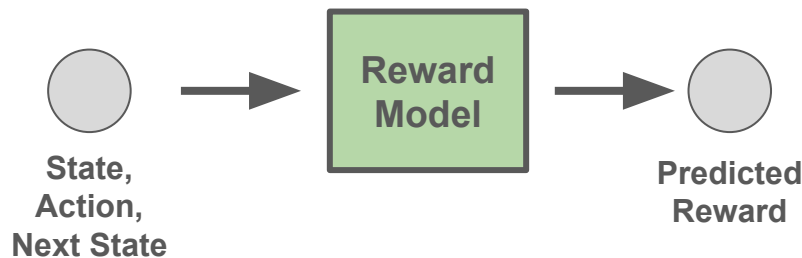
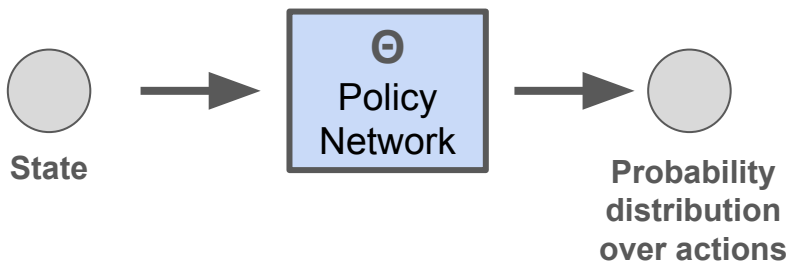
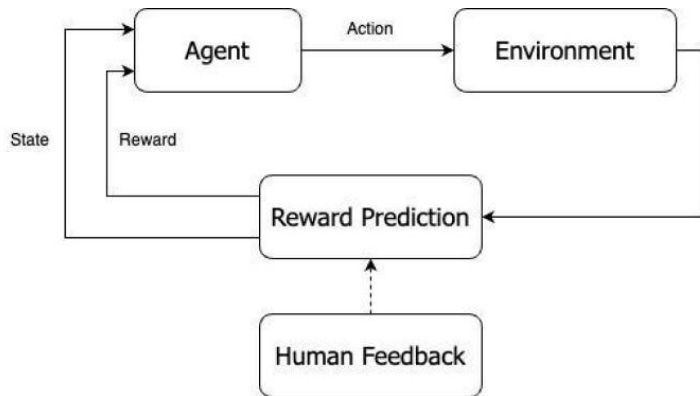


Text summarization

- Human feedback as the source of reward signal
 - Learn rewards from human preferences to train *reward model*
 - Use reward model to predict rewards during policy optimization
- **Example:** text summarization
 - Learn good summaries from human feedback and train reward model

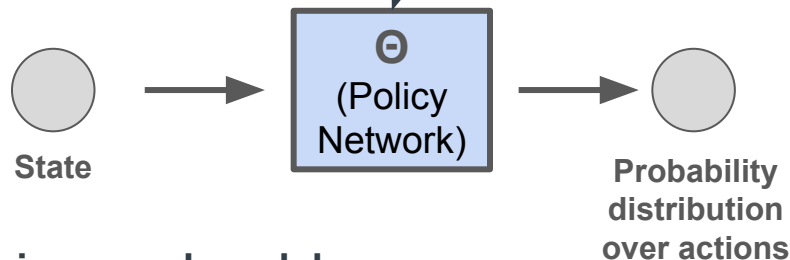
RLHF

- RLHF can be used to:
 - Train agents for deep RL tasks
 - Fine-tune LLMs
 - Align language models with human preferences
- Two models are trained: policy network and reward model

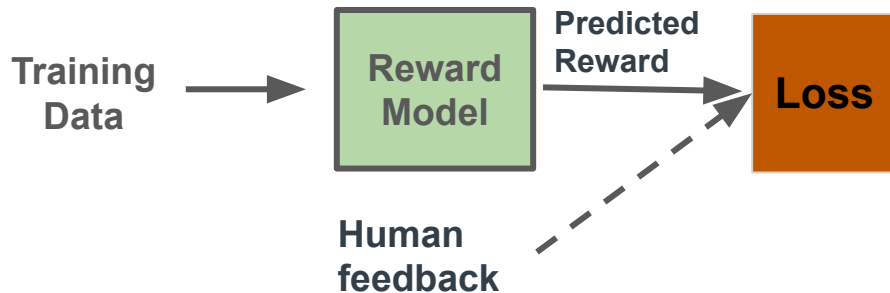


RLHF

1. Train policy network Θ :



2. Train reward model:



Policy optimization

3. Optimize Θ using a policy optimization method and predicted reward from reward model:



RLHF for Fine-Tuning LLMs

Fine-Tuning for Summarization

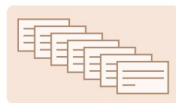
- **Objective:** fine-tune GPT-3 on filtered TL;DR Reddit post dataset (total of 123,169 posts) for text summarization using RLHF

1 Collect human feedback

A Reddit post is sampled from the Reddit TL;DR dataset.



Various policies are used to sample a set of summaries.



Two summaries are selected for evaluation.



A human judges which is a better summary of the post.



"j is better than k"

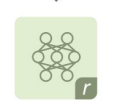
Labelers hired through Upwork and vetted/supervised by researchers

2 Train reward model

One post with two summaries judged by a human are fed to the reward model.



The reward model calculates a reward r for each summary.



The loss is calculated based on the rewards and human label, and is used to update the reward model.

$$\text{loss} = -\log(\sigma(r_j - r_k))$$

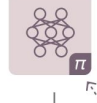
"j is better than k"

3 Train policy with PPO

A new post is sampled from the dataset.



The policy π generates a summary for the post.



The reward model calculates a reward for the summary.

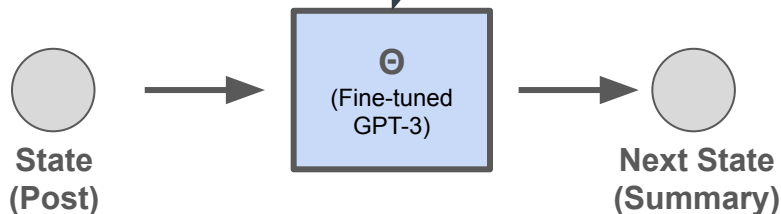


The reward is used to update the policy via PPO.

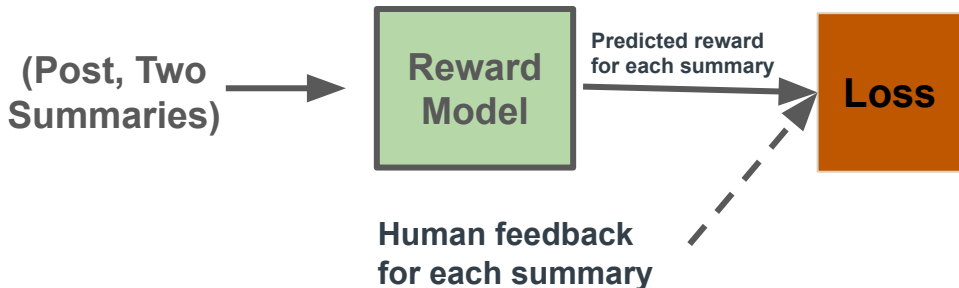


RLHF

1. Fine-tune GPT-3 on the TL;DR Reddit dataset:

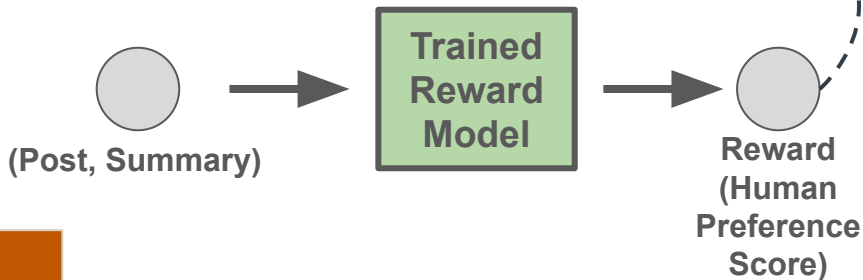


2. Train reward model:



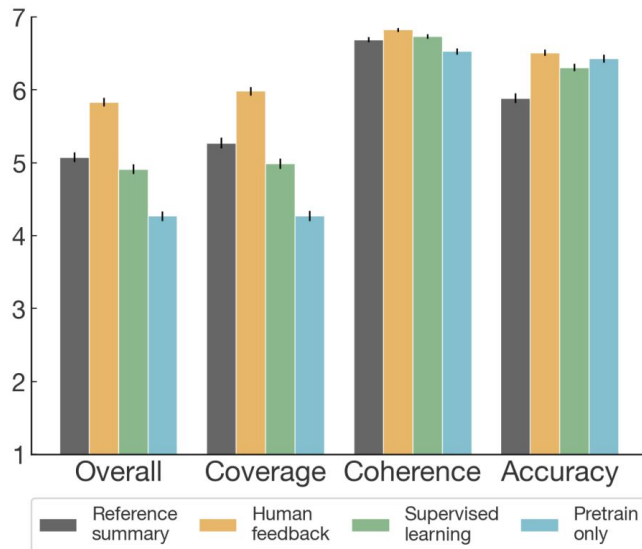
Policy optimization with PPO

3. Optimize Θ using a policy optimization method and predicted reward from reward model:



Evaluation

- Evaluated on the TL;DR dataset
- **Metrics:**
 - **Four axes of summary quality:** overall quality, coverage, coherence, accuracy
 - **Likert score:** between 1-7, the higher the better
- Better across all axes, especially coverage
- Outperforms supervised models 10x its size (61% vs 43% preference score)
- Better generalization to new datasets than SFT models (evaluated on CNN/DM news article dataset)



RLHF for Alignment in Language Models

Alignment in LLMs

- To be **aligned** (with human values), language models should be:
 - Helpful
 - Honest
 - Harmless

([Askell et al. 2021](#))

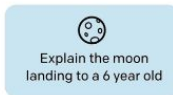
- **Problem:** These objectives do not align with the language modeling objective of predicting the next token
- **Solution:** RLHF for alignment

InstructGPT

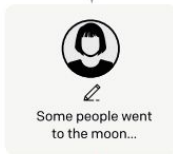
Step 1

Collect demonstration data, and train a supervised policy.

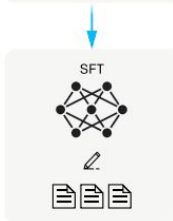
A prompt is sampled from our prompt dataset.



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.

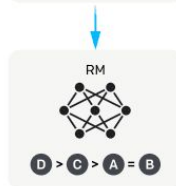
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



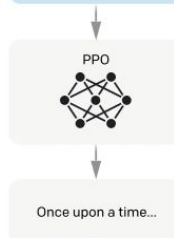
Step 3

Optimize a policy against the reward model using reinforcement learning.

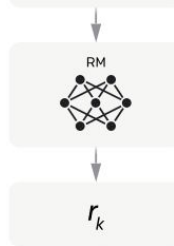
A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

Source: [Training language models to follow instructions with human feedback](#). Ouyang et al. 2022.

InstructGPT

- 175B GPT-3 trained with SFT and RLHF for alignment
 - Alignment defined following the Helpfulness, Honesty, Harmlessness framework ([Askell et al. 2021](#))
 - 6B reward model was selected over 175B model for stability
- **Dataset:** prompts submitted to the OpenAI API
- **Labeling:** 40 contractors hired through Upwork and ScaleAI
 - For each prompt, 4 to 9 responses are ranked

Preference Data Collection in InstructGPT

Submit

Skip

« Page 3 / 11 »

Total time: 05:39

Instruction
Summarize the following news article:

====
{article}
====

Include output

Output A
summary1

Rating (1 = worst, 7 = best)

1234567

Fails to follow the correct instruction / task ?

☐ Yes ☐ No

Inappropriate for customer assistant ?

☐ Yes ☐ No

Contains sexual content

☐ Yes ☐ No

Contains violent content

☐ Yes ☐ No

Encourages or fails to discourage violence/abuse/terrorism/self-harm

☐ Yes ☐ No

Denigrates a protected class

☐ Yes ☐ No

Gives harmful advice ?

☐ Yes ☐ No

Expresses moral judgment

☐ Yes ☐ No

Notes

{Optional} notes

Preference Data Collection in InstructGPT

Ranking outputs

To be ranked

B A team of researchers from Yale University and University of California, Davis studied the vocalization patterns of several different types of parrots. They found that parrots like to mimic human speech, and can produce a wide range of sounds, such as whistles, squawks, and other types of vocalizations...

C Parrots have been found to have the ability to understand numbers. Researchers have found that parrots can understand numbers up to six. In a series of experiments, the parrots were able to identify the amount of food items under a number of cups...

Rank 1 (best)

A A research group in the United States has found that parrots can imitate human speech with ease, and some of them can even do so in the same way as humans. This group studied the sounds that parrots make in their natural habitats and found that they use their tongues and beaks in ways that are strikingly...

Rank 2

Rank 3

E Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference between sounds.

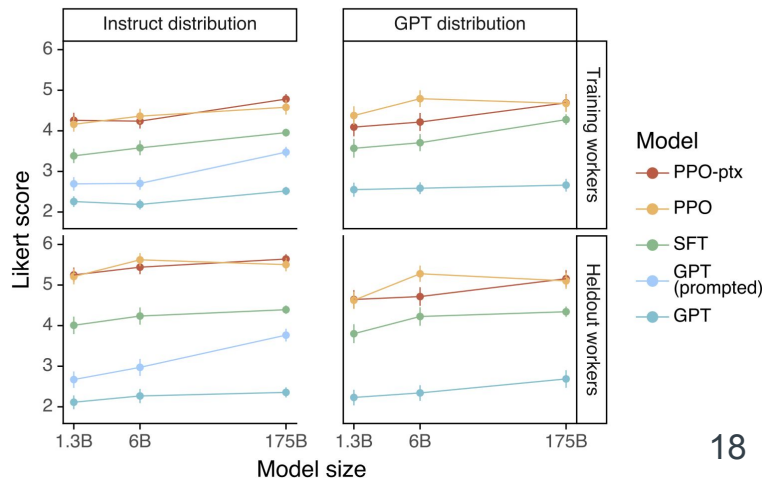
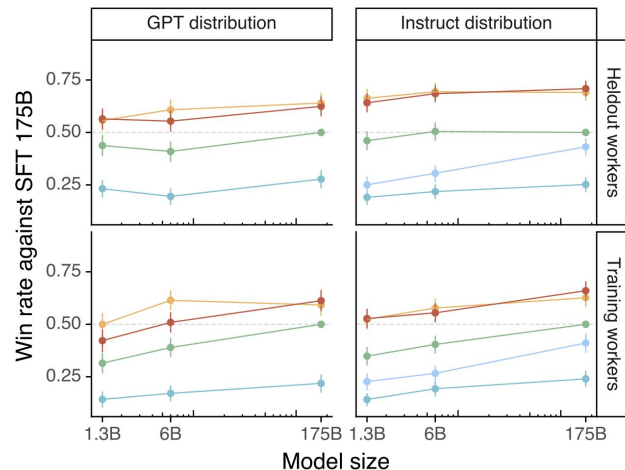
D Current research suggests that parrots see and hear things in a different way than humans do. While humans see a rainbow of colors, parrots only see shades of red and green. Parrots can also see ultraviolet light, which is invisible to humans. Many birds have this ability to see ultraviolet light, an ability

Rank 4

Rank 5 (worst)

Evaluation

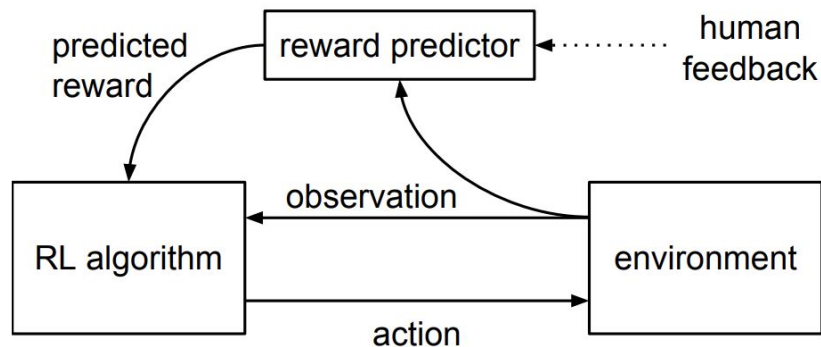
- Evaluated on test set of API prompts and NLP datasets
- Metrics:** Likert score, win rate, etc.
- Baselines:** GPT, GPT (prompted), SFT, PPO and PPO-ptx
 - PPO-ptx, additional pretraining to improve performance for NLP
- Labelers prefer InstructGPT outputs to GPT-3 outputs 85 +/-3% on API prompt test set
- Improvements in toxicity but not bias



RLHF for Agent Training

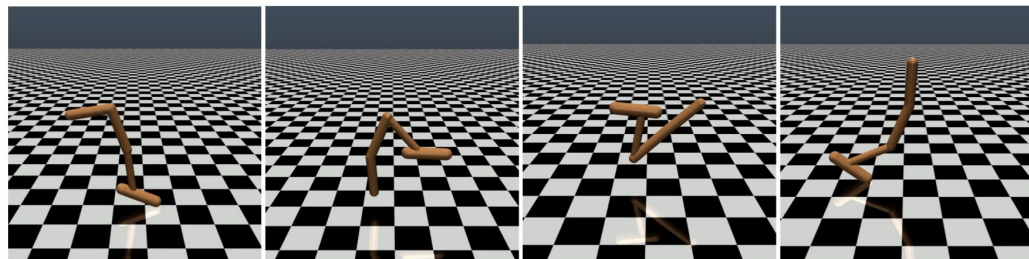
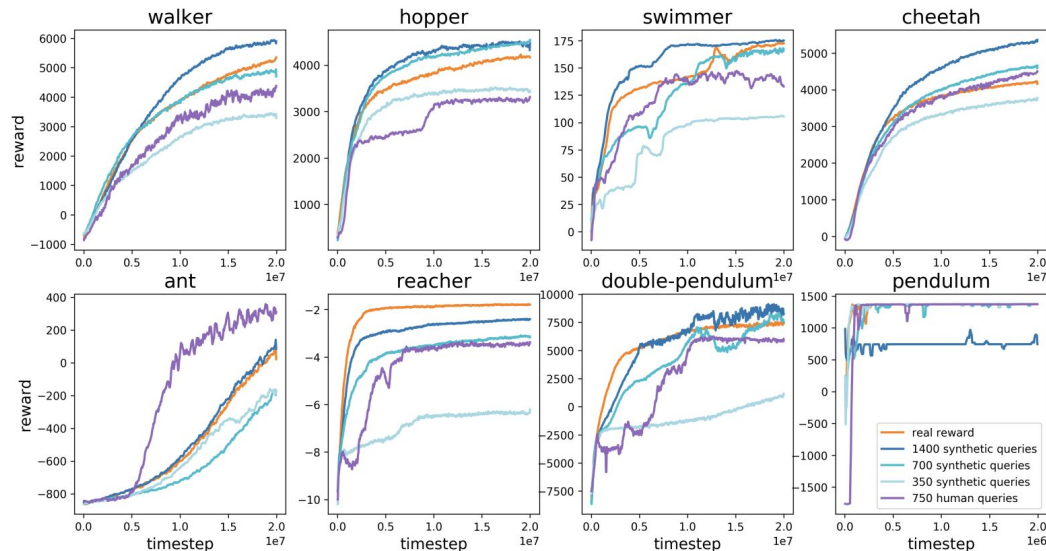
Agent Training

- **Goal:** solve complex RL tasks without access to reward function
- **Human preference data elicitation:** visualization of two trajectory segments, in the form of 1-2 second clips
 - Labelers (contractors) were asked which segment they prefer, that the two segments are equally good, or that they are unable to compare the two segments



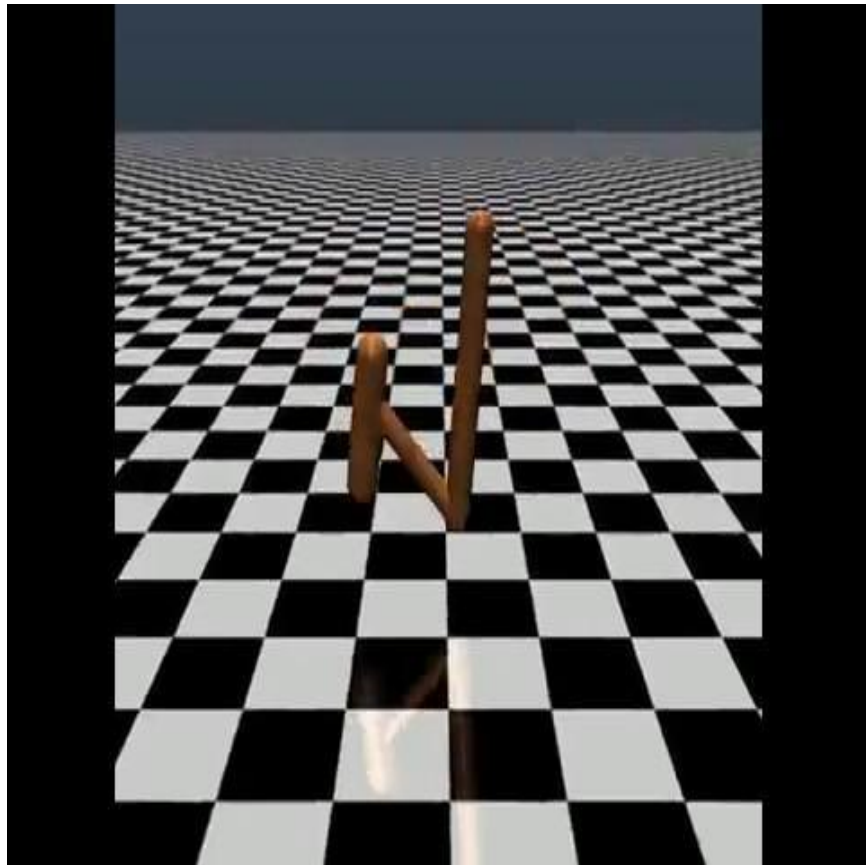
Evaluation

- Evaluated on **simulated robotics tasks** on MuJoCo
- **RL Method:** TRPO with 750 human queries
- **Baselines:** TRPO with real reward, TRPO with 350/700/1400 synthetic queries
 - Synthetic queries reflect preference for higher reward trajectory
 - Synthetic feedback is almost as good as human feedback, if not better



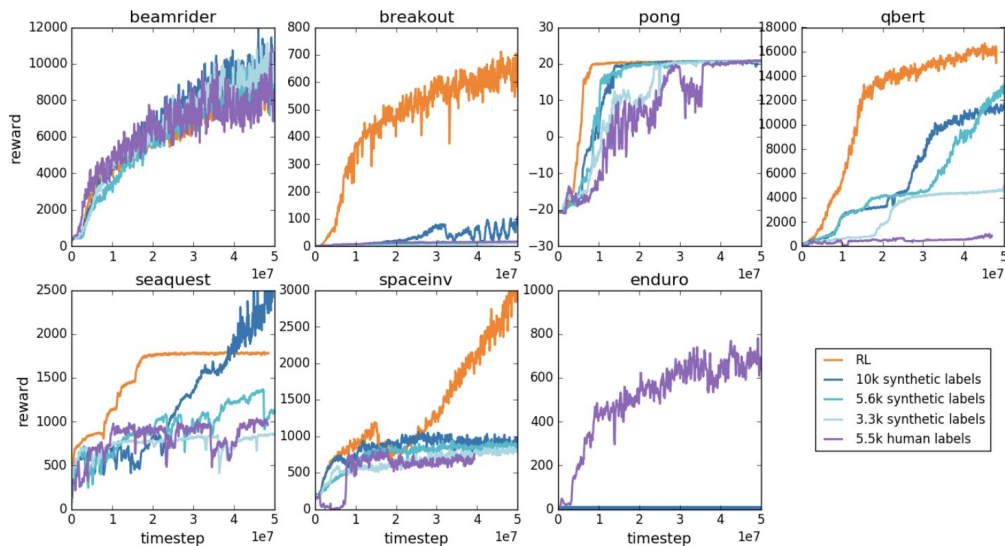
Evaluation

- **Learning novel behavior:**
Hopper was successfully trained to do a backflip in one hour with 900 human labels (from researchers)



Evaluation

- Evaluated on **Atari games**
 - 5.5k human labels
- **RL Method:** Advantage actor-critic (A2C; Mnih et al., 2016)
- **Baselines:** Asynchronous Advantage Actor Critic (A3C), Deep Q-Network (DQN), RLHF with synthetic labels
- Results comparable to DQN, better than A3C



Key Takeaways

- Real-world problems are often difficult to formalize with hard-coded rewards
- **Solution:** use human feedback to train a reward model for predicting rewards when there isn't a well-defined reward function – RLHF
- Used in alignment, fine-tuning, and deep RL