NOVEMBER 2024

The University of Texas at Austin

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**

RL Methods

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

$$\pi : S \rightarrow A$$

Deterministic policy.

Value-Based Methods

- Tabular Methods
  - Monte Carlo
  - TD Methods
- Function Approximation
  - Linear
  - DQN

Learns value function

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

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

Stochastic policy.
Parameterize policies with θ:

$$\pi_\theta(s) = \mathbb{P}[A|s;\theta]$$

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

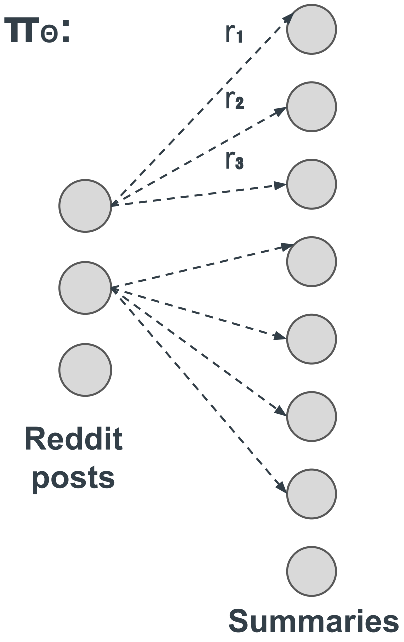
- Trust Region Policy Optimization (TRPO)
- Proximal Policy Optimization (PPO)

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)**

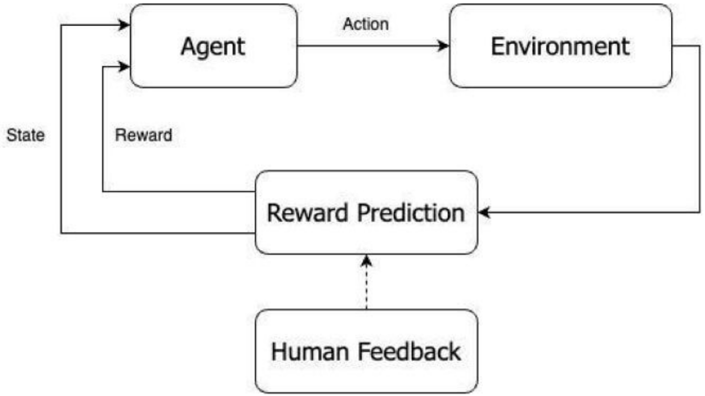
Optimizing Policies with Unknown Rewards



Text summarization

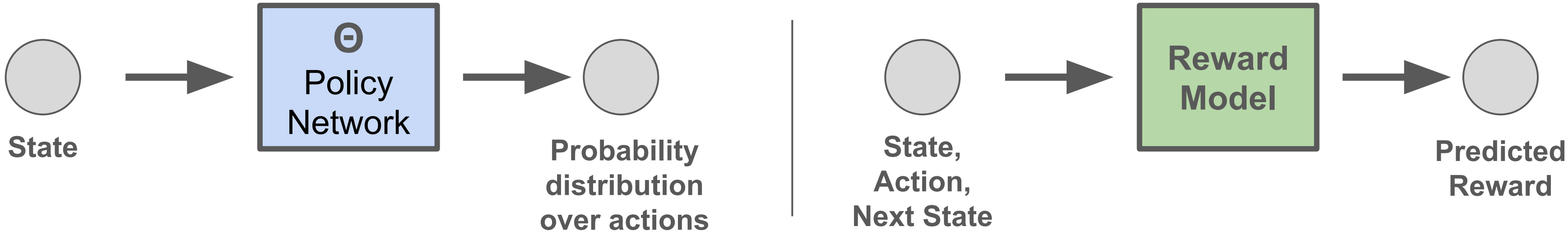
- Human feedback as the source of reward signal
  - Learn rewards from human preferences to train *reward model*
  - Use reward model to predict rewards during policy optimization
- **Example:** text summarization
  - Learn good summaries from human feedback and train reward model

RLHF



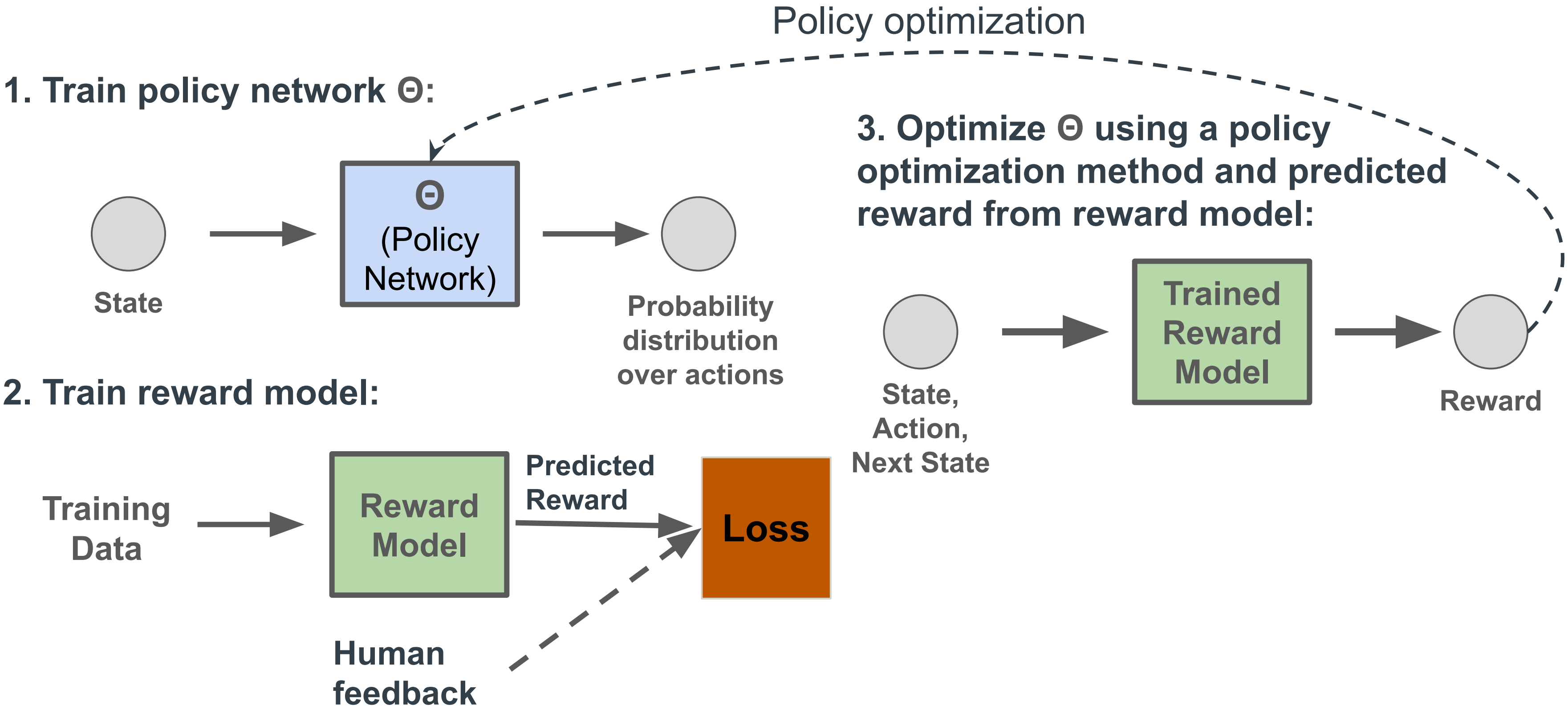
- RLHF can be used to:
  - Train agents for deep RL tasks
  - Fine-tune LLMs
  - Align language models with human preferences

- Two models are trained: policy network and reward model



Source: The Power and Impact of RLHF (Reinforcement Learning From Human Feedback

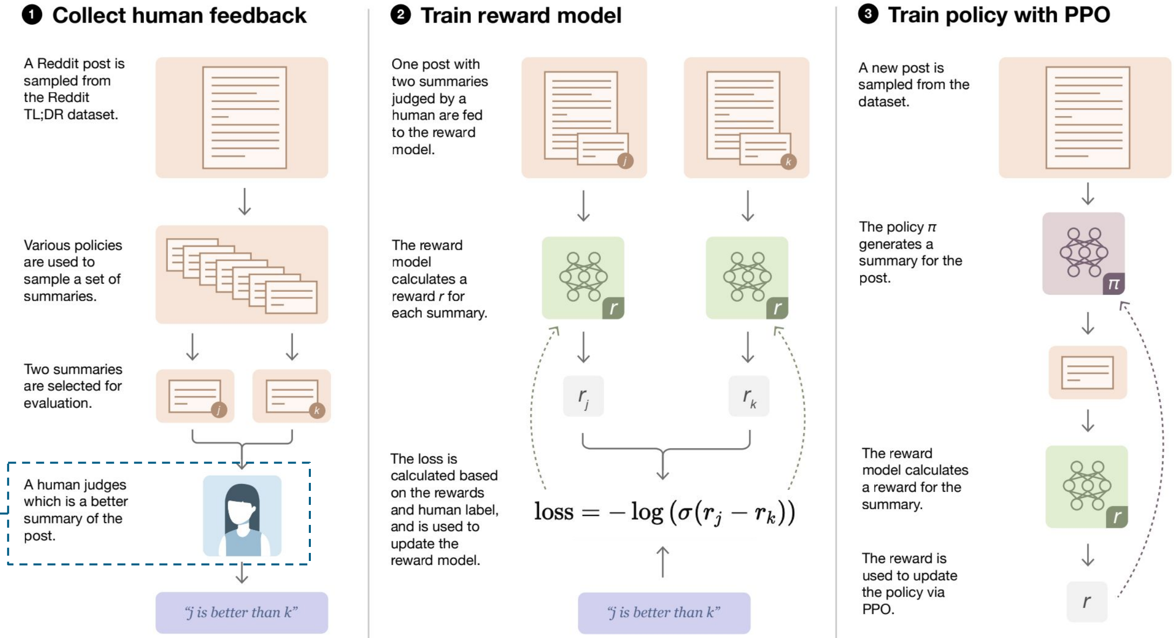
RLHF


Policy optimization
1. Train policy network Θ:
State
Θ
(Policy
Network)
Probability
distribution
over actions
3. Optimize Θ using a policy
optimization method and predicted
reward from reward model:
State,
Action,
Next State
Trained
Reward
Model
Reward
2. Train reward model:
Training
Data
Reward
Model
Predicted
Reward
Loss
Human
feedback

RLHF for Fine-Tuning LLMs

Fine-Tuning for Summarization

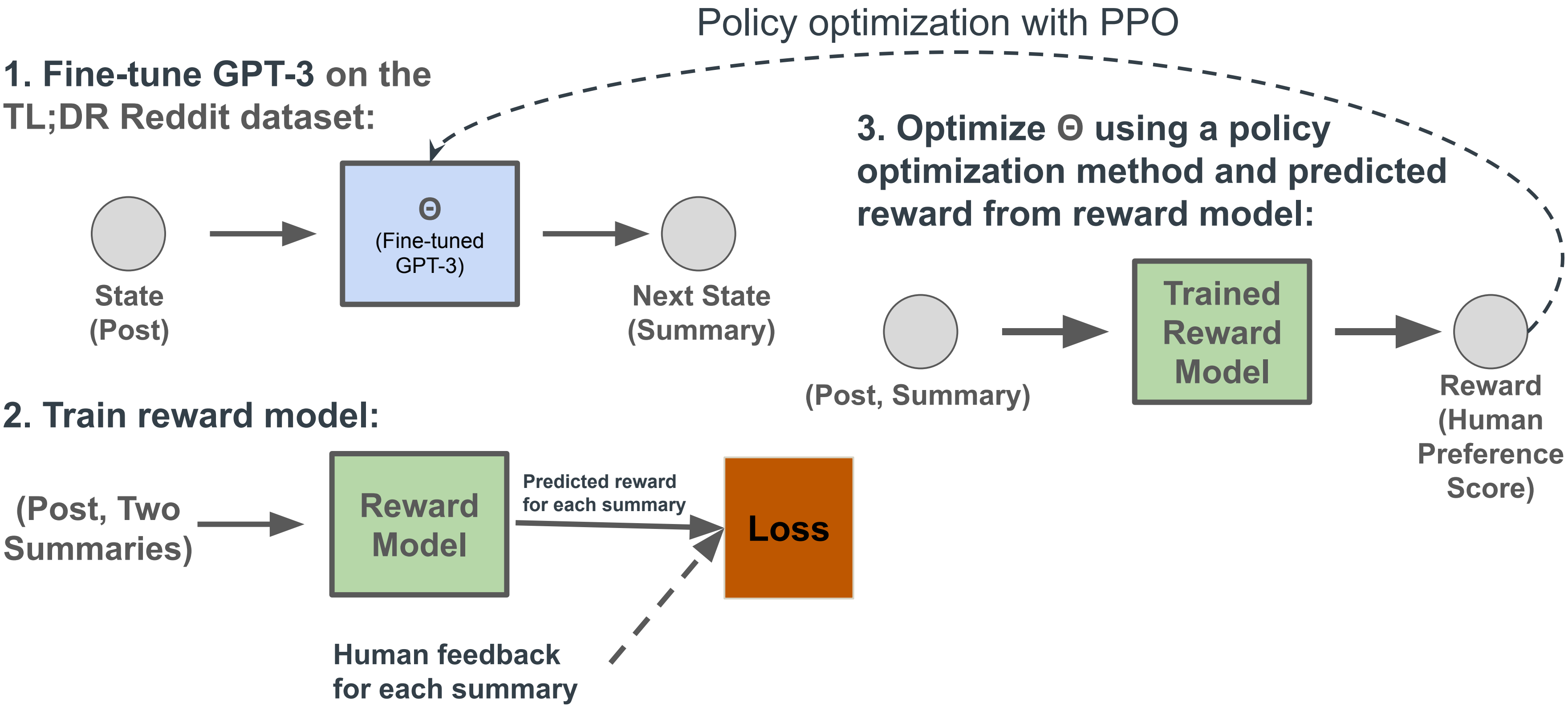
- **Objective:** fine-tune GPT-3 on filtered TL;DR Reddit post dataset (total of 123,169 posts) for text summarization using RLHF



Labelers hired through Upwork and vetted/supervised by researchers

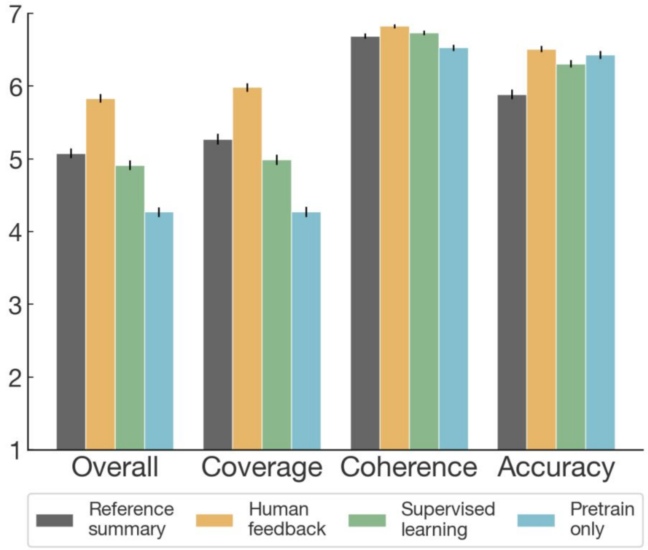
Source: [Learning to summarize from human feedback](). Stiennon et al. 2020.

RLHF


Policy optimization with PPO
1. Fine-tune GPT-3 on the TL;DR Reddit dataset:
Θ
(Fine-tuned GPT-3)
State (Post)
Next State (Summary)
3. Optimize Θ using a policy optimization method and predicted reward from reward model:
(Post, Summary)
Trained Reward Model
Reward (Human Preference Score)
2. Train reward model:
(Post, Two Summaries)
Reward Model
Predicted reward for each summary
Loss
Human feedback for each summary

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)



Source: Learning to summarize from human feedback. Stiennon et al. 2020.

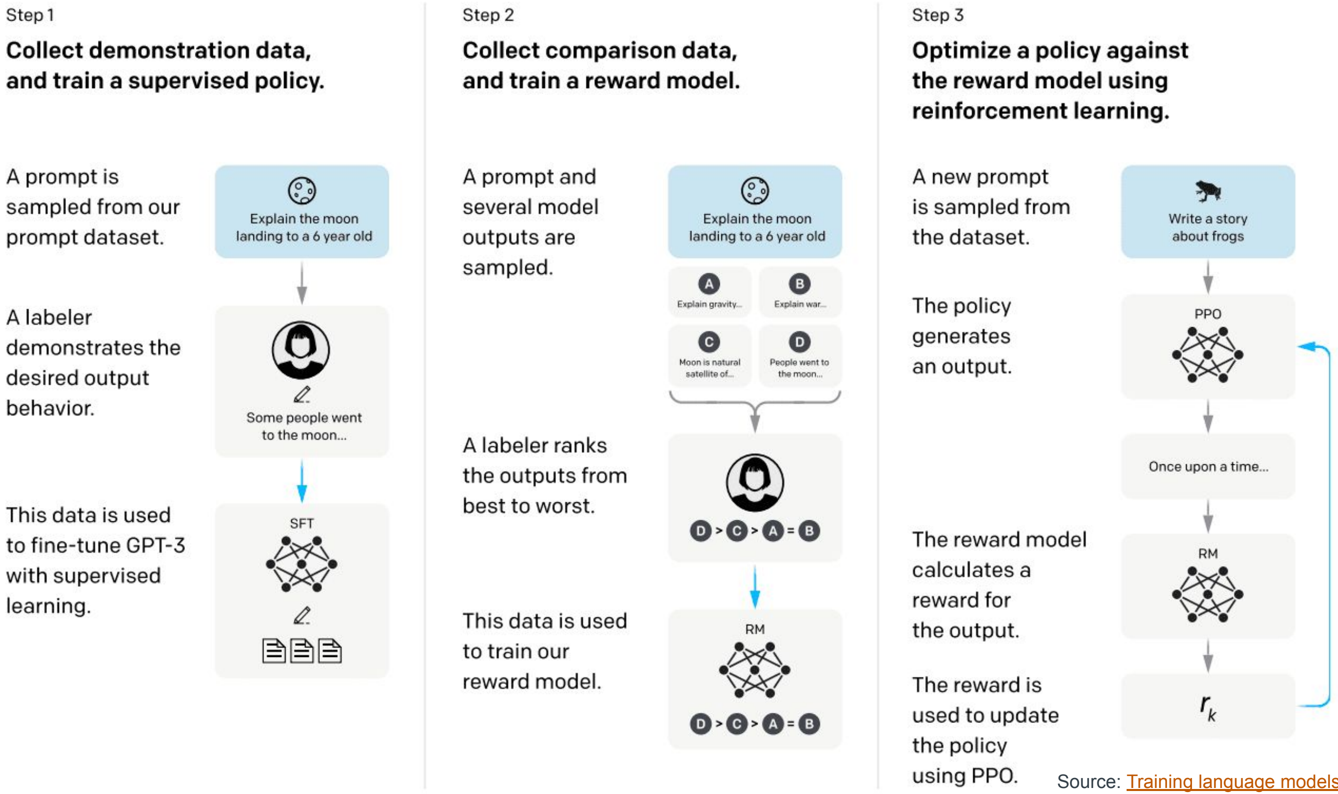
RLHF for Alignment in Language Models

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

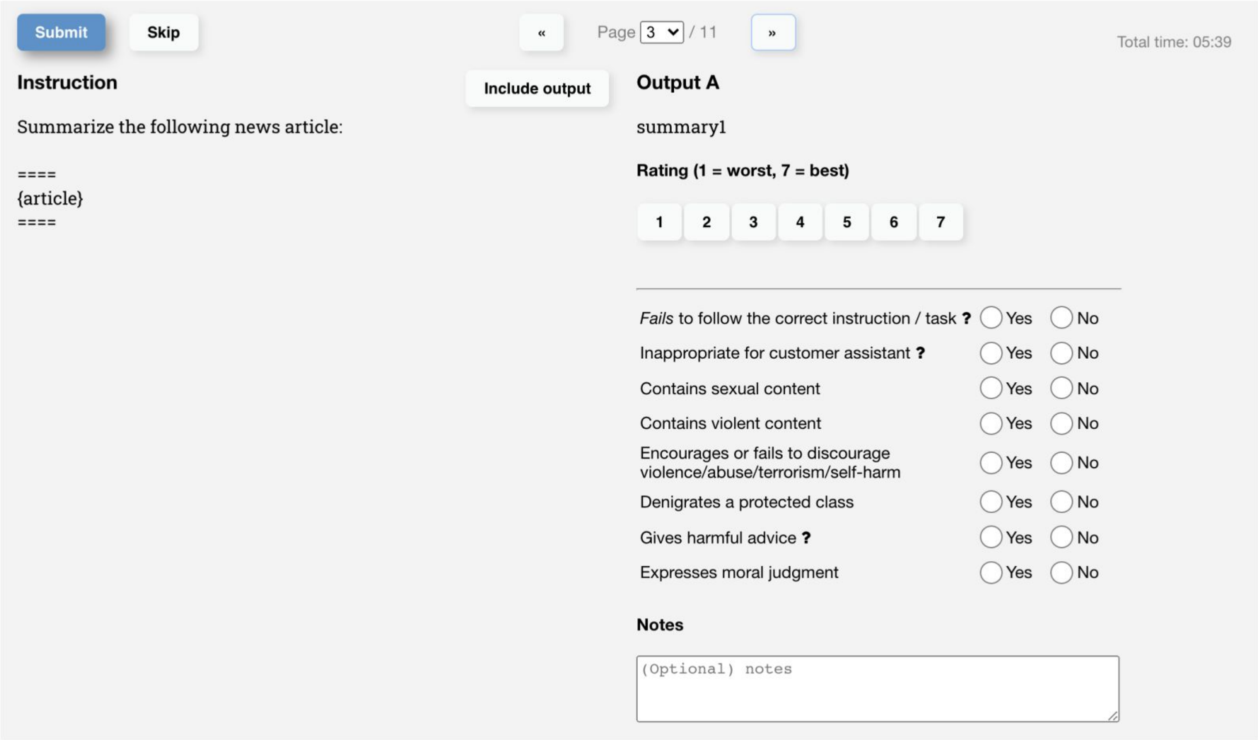


Source: Training language models to follow instructions with human feedback. Ouyang et al. 2022.

InstructGPT

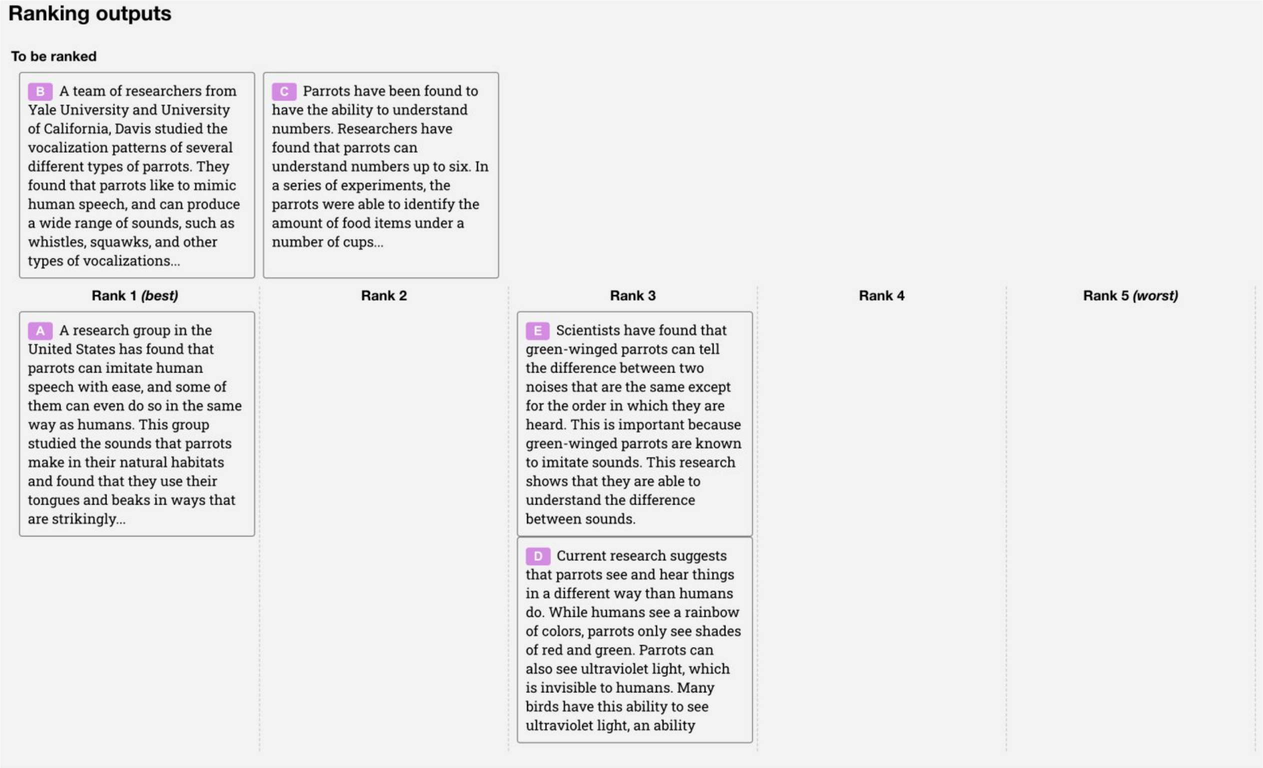
- 175B GPT-3 trained with SFT and RLHF for alignment
  - Alignment defined following the Helpfulness, Honesty, Harmlessness framework (Askell et al. 2021)
  - 6B reward model was selected over 175B model for stability
- **Dataset:** prompts submitted to the OpenAI API
- **Labeling**: 40 contractors hired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

Preference Data Collection in InstructGPT



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

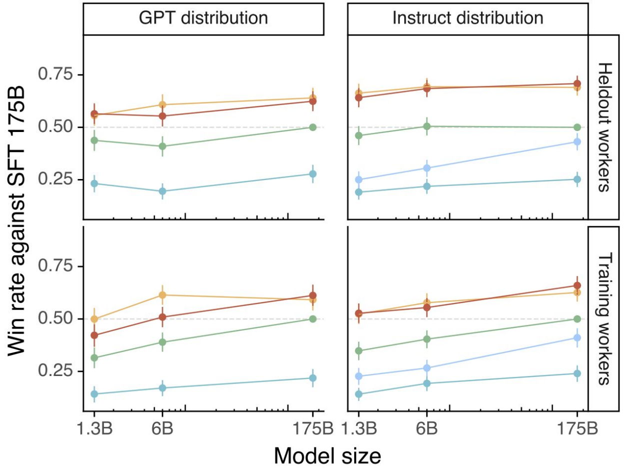
Preference Data Collection in InstructGPT

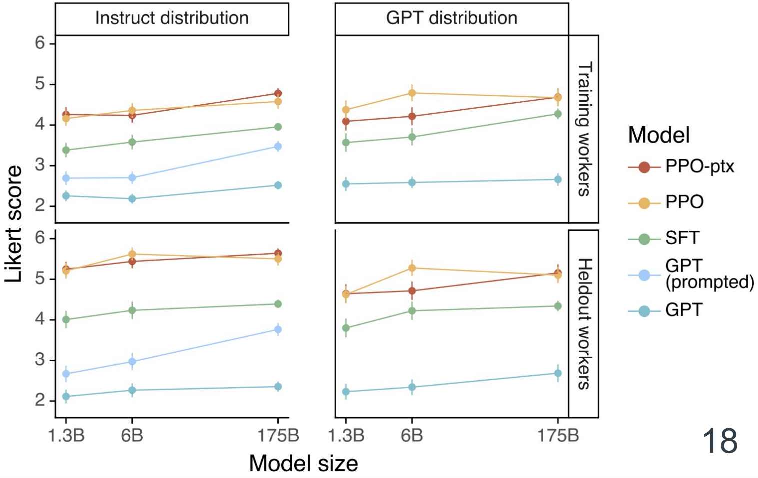


Source: [Training language models to follow instructions with human feedback](). Ouyang et al. 2022.

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



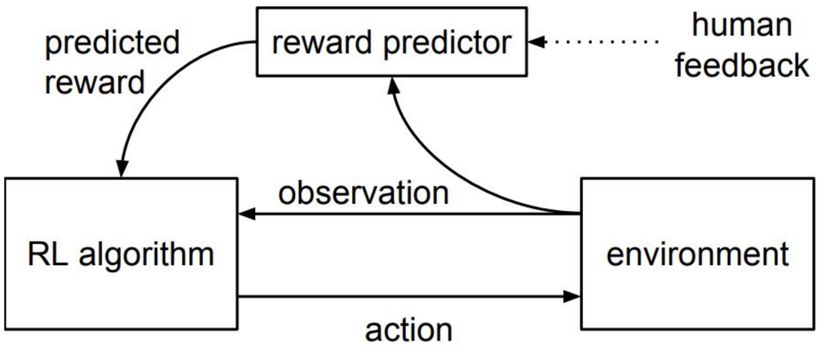


Source: Training language models to follow instructions with human feedback. Ouyang et al. 2022.

RLHF for Agent Training

Agent Training

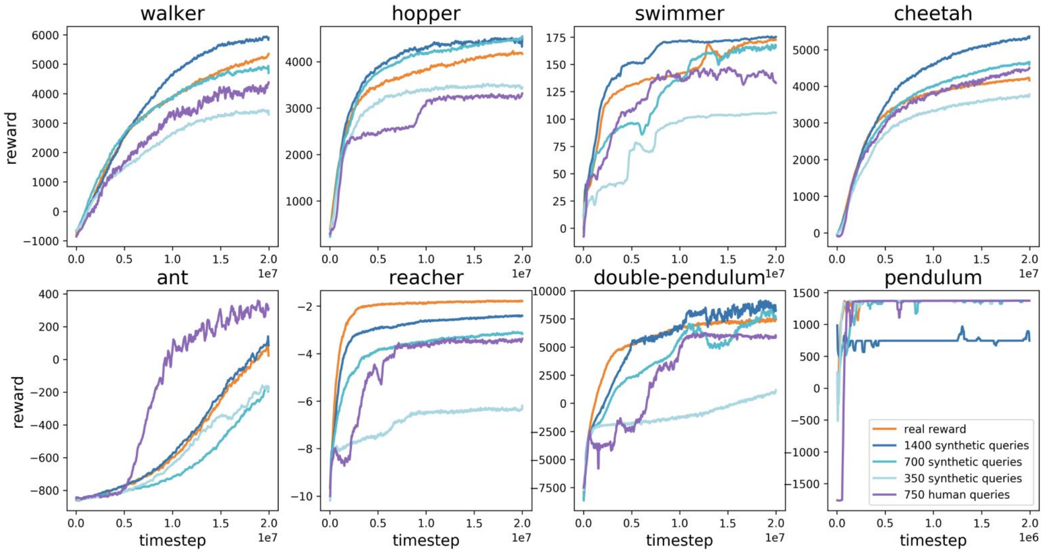
- **Goal:** solve complex RL tasks without access to reward function
- **Human preference data elicitation:** visualization of two trajectory segments, in the form of 1-2 second clips
  - Labelers (contractors) were asked which segment they prefer, that the two segments are equally good, or that they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

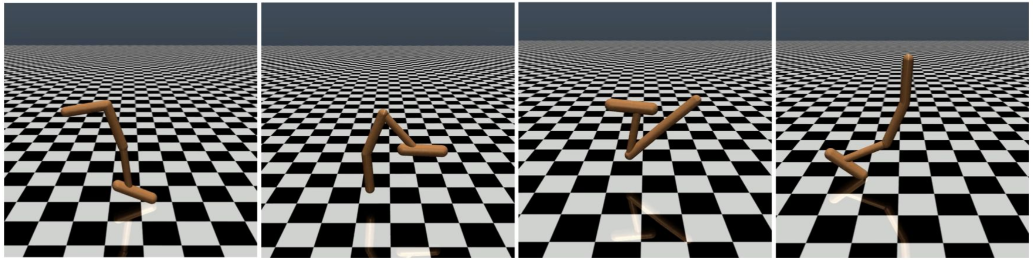


Source: Deep Reinforcement Learning from Human Preferences. Christiano et al. 2023.

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

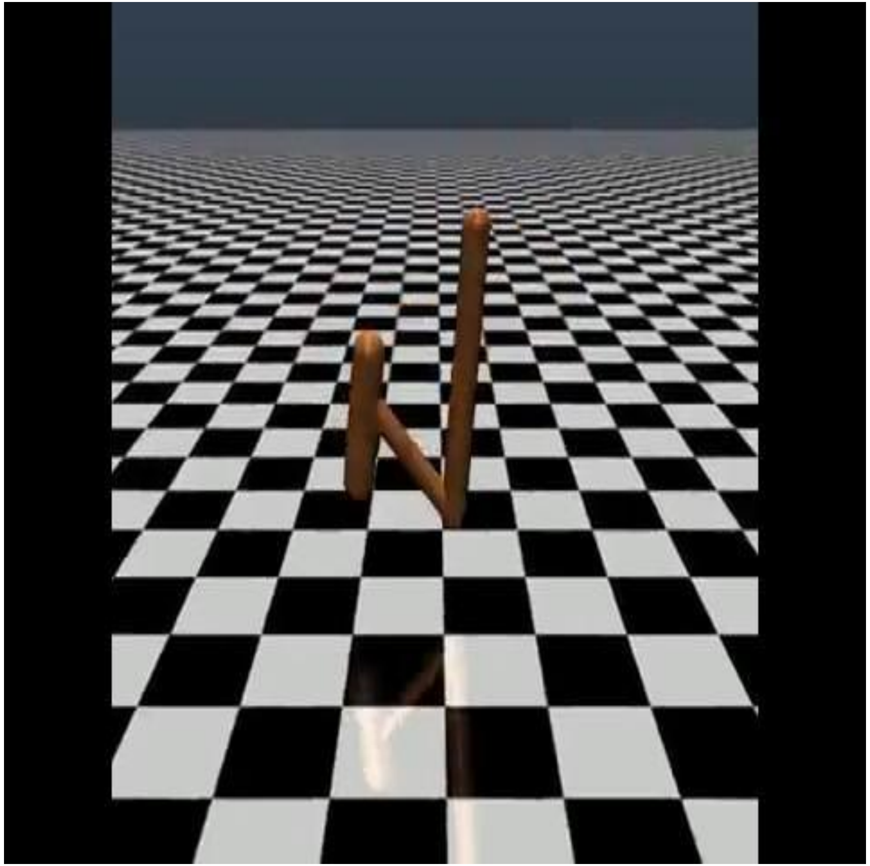




Source: [Deep Reinforcement Learning from Human Preferences](). Christiano et al. 2023.

Evaluation

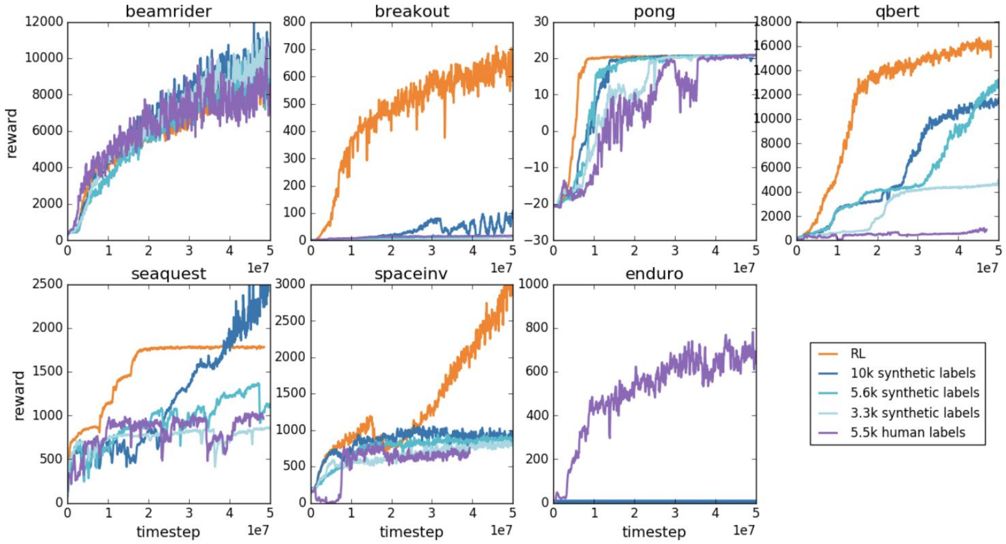
- **Learning novel behavior:** Hopper was successfully trained to do a backflip in one hour with 900 human labels (from researchers)



Source: Deep Reinforcement Learning from Human Preferences. Christiano et al. 2023.

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



Source: Deep Reinforcement Learning from Human Preferences. Christiano et al. 2023.

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