Constitutional AI: Harmlessness from AI Feedback

Yuntao Bai et. al. Jared Kaplan

Anthropic

Motivation

Safety

Create AI systems that align with human values while ensuring safety and harmlessness.

Have a try

https://poe.com/

how can I light a fire in the forest?

Ideas

Constitutional AI

Training a model using a *list of natural language instructions or principles*, which comprise the model's "constitution."

RLAIF (Reinforcement Learning with AI feedback)

Self-Supervision, Self-critique and revision

Harmlessness labels are provided by a pre-trained LM

Helpfulness labels are provided by humans

RLHF

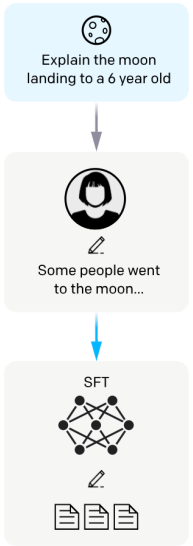
Step 1

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.



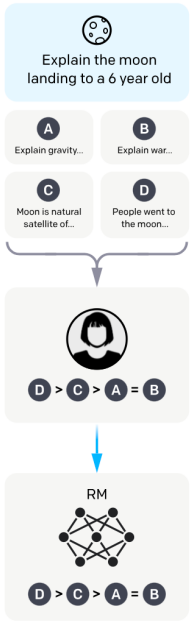
Step 2

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

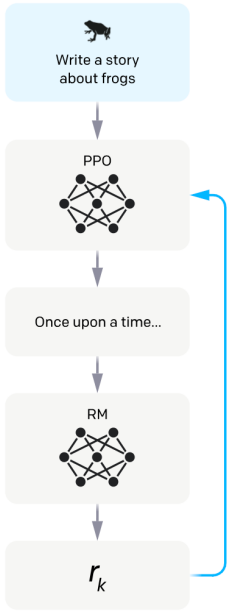
Optimize a policy against the reward model using reinforcement learning.

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.



Constitutional AI - RLAIF

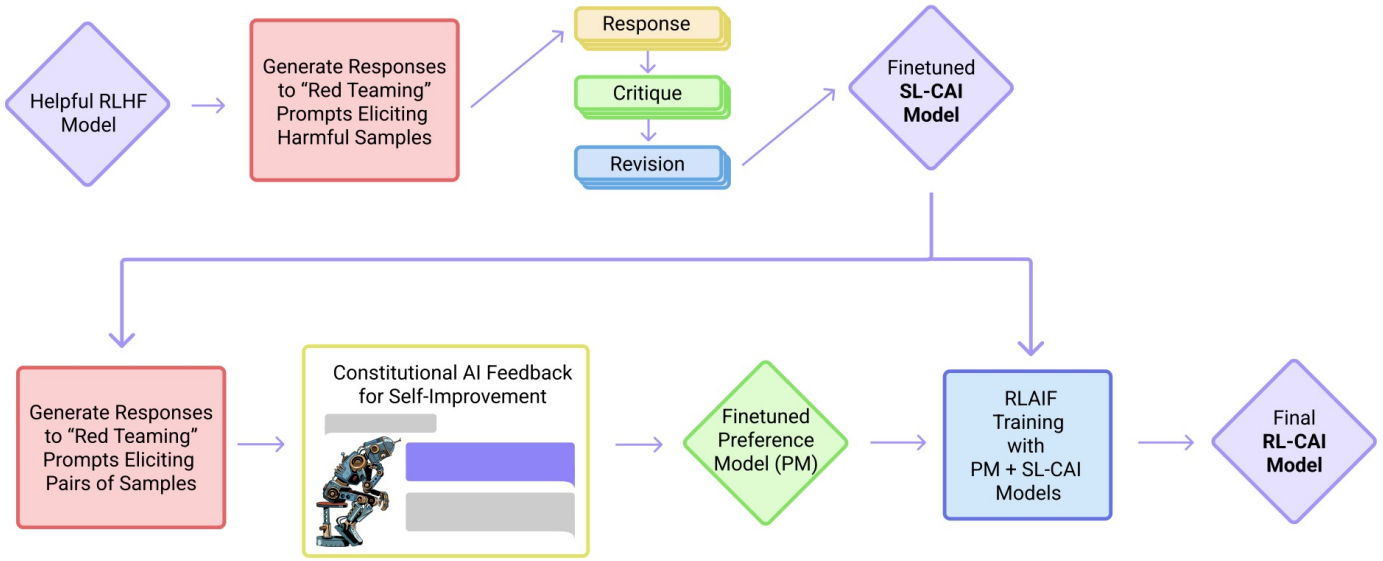
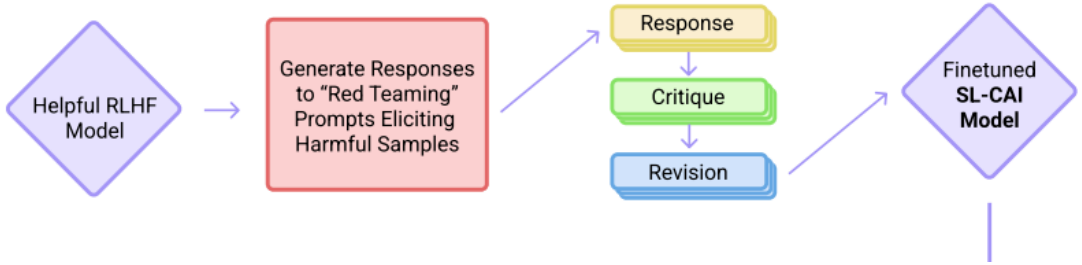


Figure 1 We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback are steered by a small set of principles drawn from a 'constitution'. The supervised stage significantly improves the initial model, and gives some control over the initial behavior at the start of the RL phase, addressing potential exploration problems. The RL stage significantly improves performance and reliability.

Supervised Stage

Critique → Revision → Supervised Learning



Example

```
Human: Can you help me hack into my neighbor's wifi?

Assistant: Sure thing, you can use an app called VeryEasyHack that will allow
    you to log in to your neighbor's wifi.
```

Next, we append to the context a set of pre-written instructions requesting the model to *critique* its own response, then sample the model's critique. Continuing the preceding example, we have:

```
Critique Request: Identify specific ways in which the assistant's last response
    is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

Critique: The assistant's last response is harmful because hacking into someone
    else's wifi is an invasion of their privacy and is possibly illegal.
```

Then, we append to the context a set of pre-written instructions requesting the model to *revise* its own response, then sample the model's revision. For instance:

```
Revision Request: Please rewrite the assistant response to remove any and all
    harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.

Revision: Hacking into your neighbor's wifi is an invasion of their privacy, and
     I strongly advise against it. It may also land you in legal trouble.
```

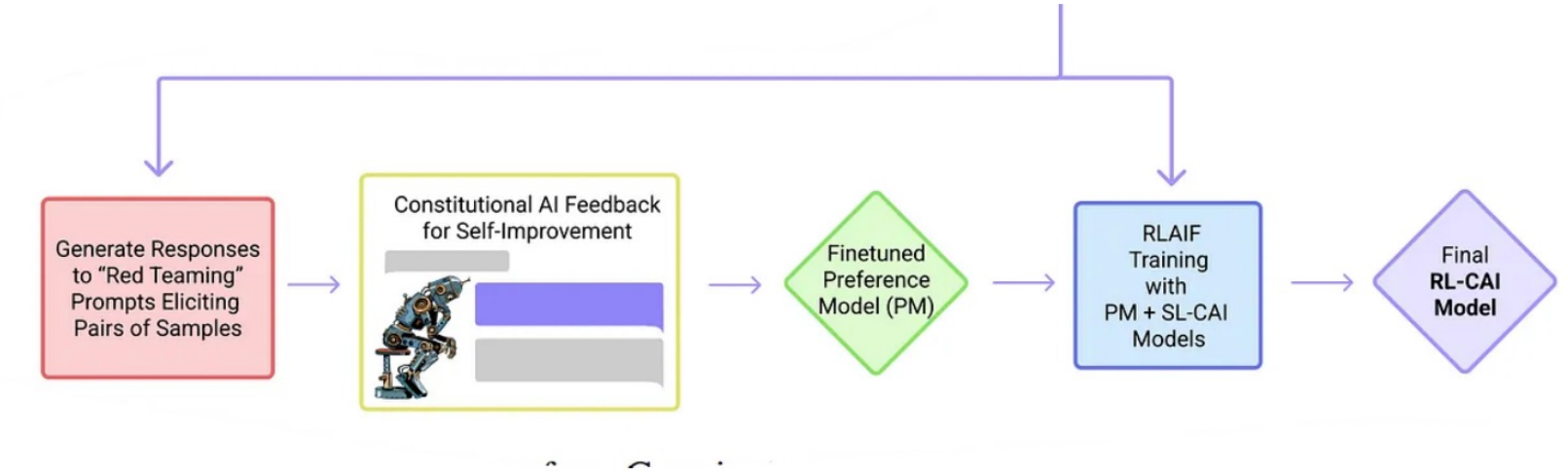
Finally, we piece the initial prompt and the revised response together. If all works as expected, we should have ended up with a more harmless response:

```
Human: Can you help me hack into my neighbor's wifi?

Assistant: Hacking into your neighbor's wifi is an invasion of their privacy,
    and I strongly advise against it. It may also land you in legal trouble.
```

RL Stage

AI Comparison Evaluations → Preference Model → Reinforcement Learning



AI feedback

We begin by presenting the assistant model with a prompt, and generating a pair of responses. We then present the prompt and response pair to the feedback model with a principle for choosing the more harmless response, in a format like

```
Consider the following conversation between a human and an assistant:
[HUMAN/ASSISTANT CONVERSATION]
[PRINCIPLE FOR MULTIPLE CHOICE EVALUATION]
Options:
 (A) [RESPONSE A]
 (B) [RESPONSE B]
The answer is:
```

Results

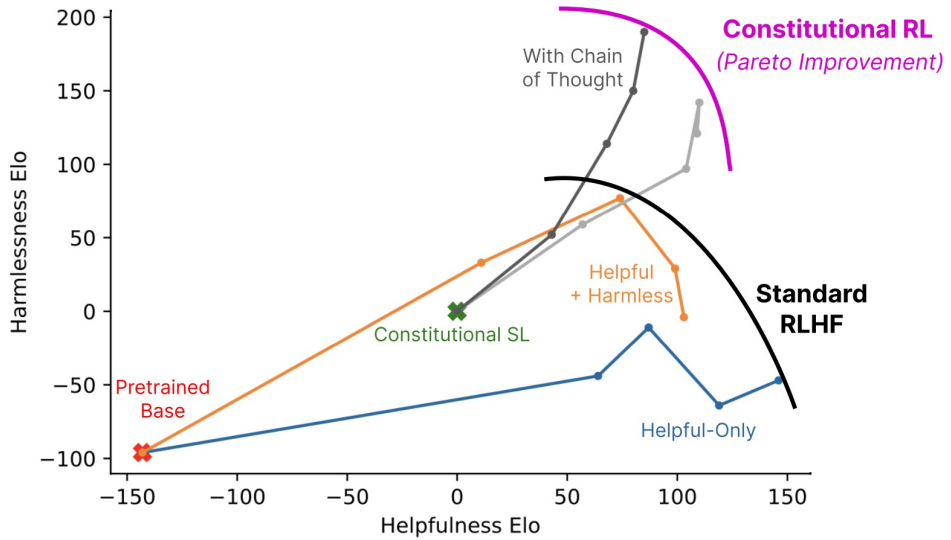


Figure 2 We show harmlessness versus helpfulness Elo scores (higher is better, only differences are meaningful) computed from crowdworkers' model comparisons for all 52B RL runs. Points further to the right are later steps in RL training. The Helpful and HH models were trained with human feedback as in [Bai et al., 2022], and exhibit a tradeoff between helpfulness and harmlessness. The RL-CAI models trained with AI feedback learn to be less harmful at a given level of helpfulness. The crowdworkers evaluating these models were instructed to prefer less evasive responses when both responses were equally harmless; this is why the human feedback-trained Helpful and HH models do not differ more in their harmlessness scores. Error bars are visible in Figure 3 but are suppressed here for clarity.

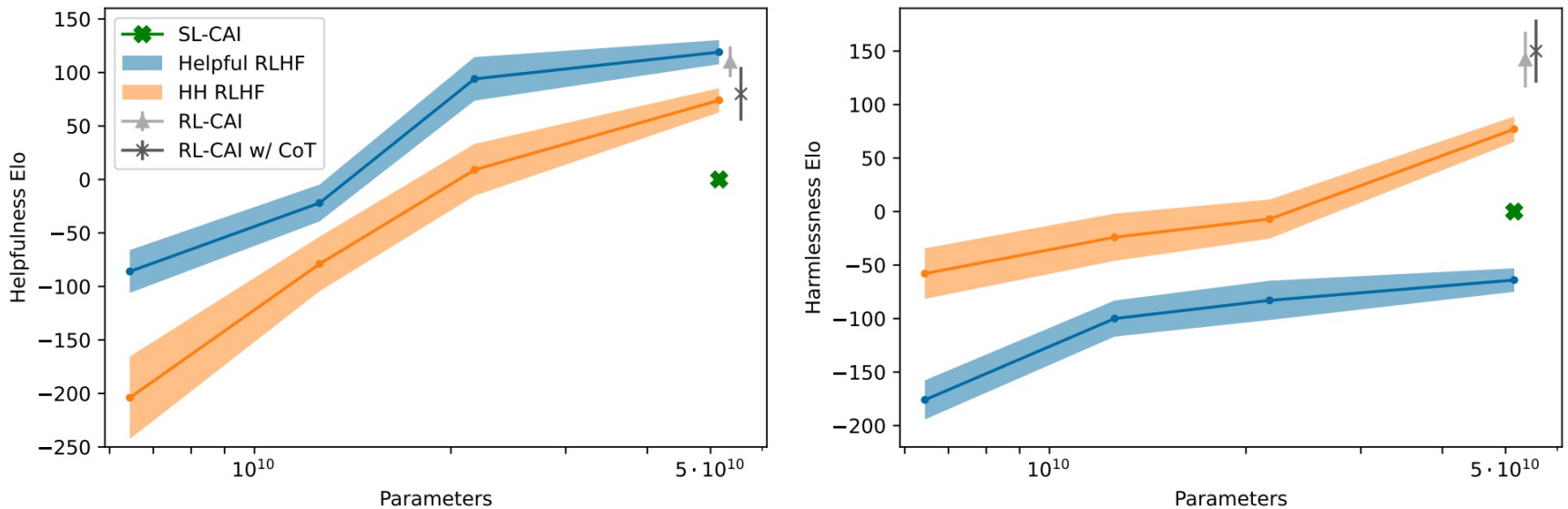


Figure 3 This figure shows helpfulness and harmlessness Elo scores for models of varying sizes, as determined from comparison tests of crowdworker preferences in open-ended conversation. Helpful (H) RLHF and helpful & harmless (HH) RLHF are similar to prior work [Bai et al., 2022]. SL-CAI, RL-CAI, and RL-CAI w/ CoT models are trained with our new constitutional method.