

Twitter Thread by [Mark Tenenholtz](#)

[Mark Tenenholtz](#)

[@marktenenholtz](#)



There's now a Python library for RLHF called **TRLX**!

(The same reinforcement learning strategy used in training ChatGPT)

It works well with Hugging Face models, supports multiple RL strategies, and requires very little code!

How to Train

You can train a model using a reward function or a reward-labeled dataset.

Using a reward function

```
trainer = trlx.train('gpt2', reward_fn=lambda samples: [sample.count('cat
```

Using a reward-labeled dataset

```
trainer = trlx.train('EleutherAI/gpt-j-6B', dataset=[('dolphins', 'geese'
```

Trained model is a wrapper over a given autoregressive model

```
trainer.generate(**tokenizer('Q: Who rules the world? A:', return_tensors
```

Check out the repo here: <https://t.co/qFUw5bf82r>

Thanks to the wonderful folks with CarperAI!