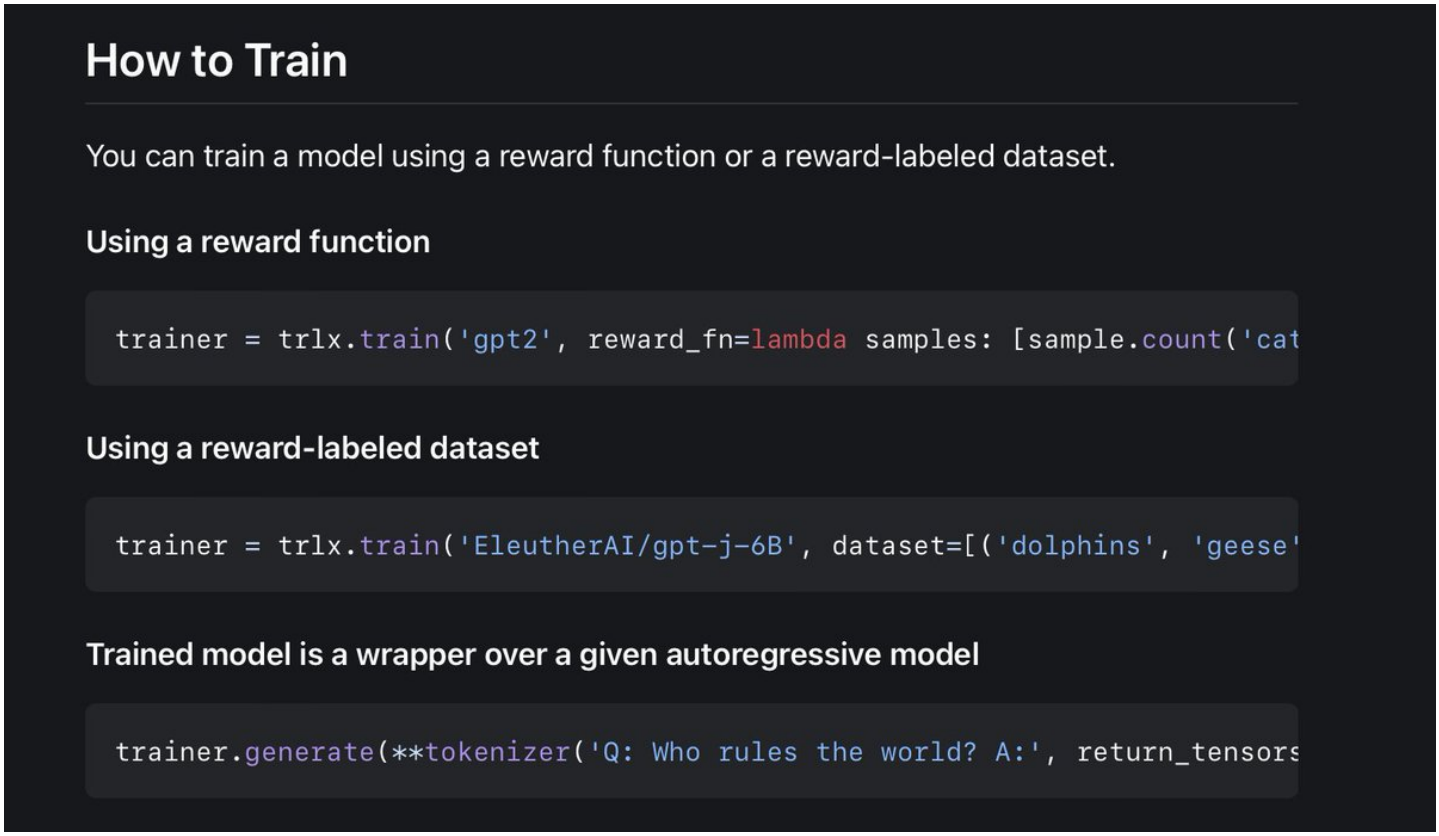
Twitter Thread by Mark Tenenholtz

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!



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

Thanks to the wonderful folks with CarperAI!