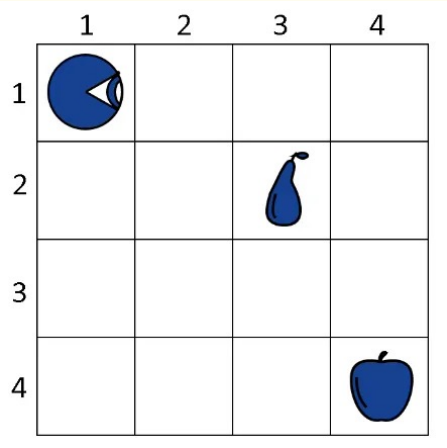
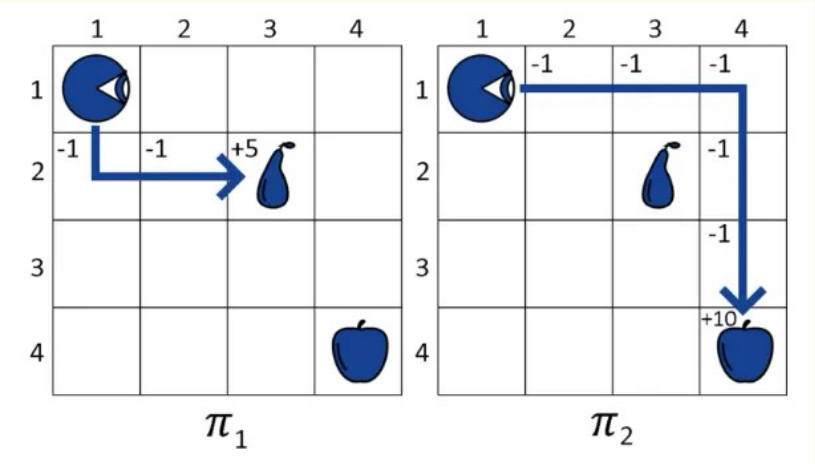
# Reinforcement Learning:



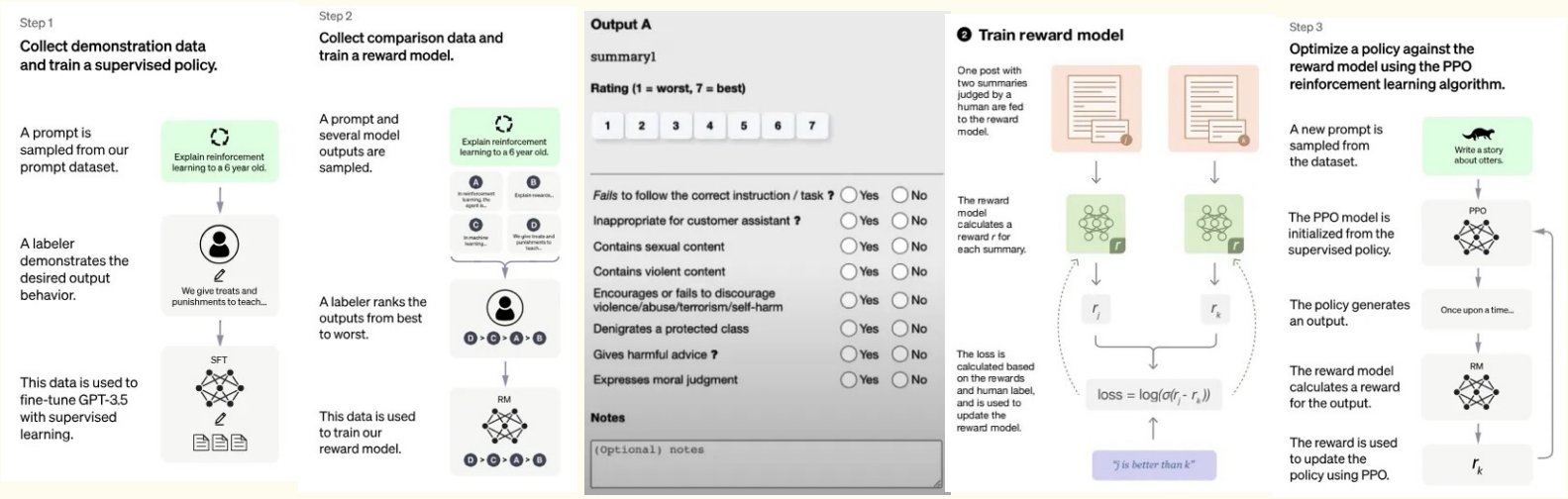



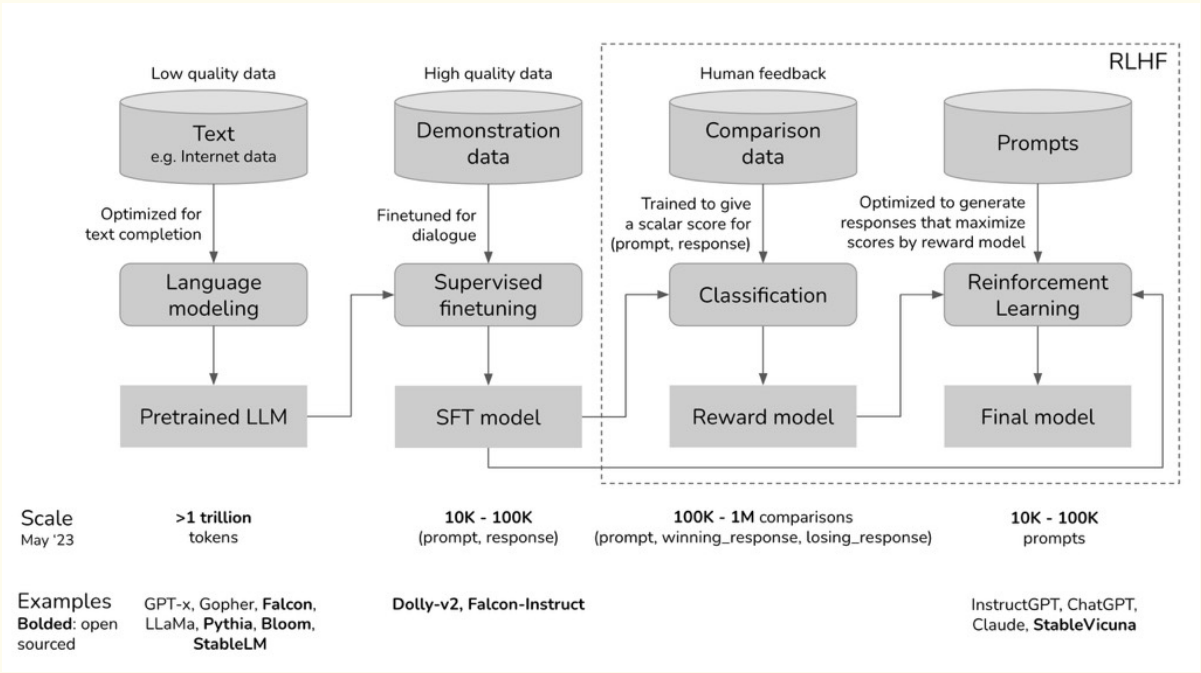

→ RL is suitable for problems with sequential decision making, where actions influence subsequent states and rewards.

→ **RLHF:** Human guidance is incorporated in the learning process. RLHF leverages the knowledge of Human to increase the accuracy and safe-guarding.

→ RLHF is having 3 steps:

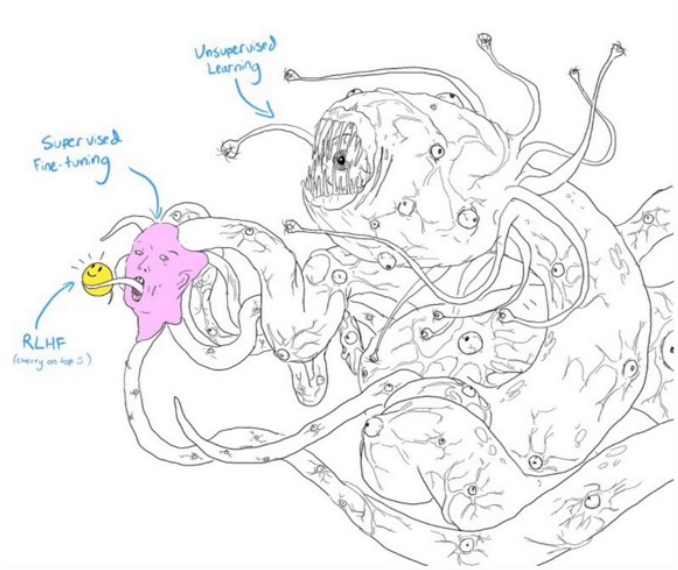
1. supervised Fine-tuning (SFT) of LLM (GPT 3.5)
2. Training a Reward Model (RM)
3. Updating policy using Proximal Policy Optimization (PPO).




If you squint, this above diagram looks very similar to the meme Shoggoth with a smiley face.

1. The pretrained model is an untamed monster because it was trained on indiscriminate data scraped from the Internet: think clickbait, misinformation, propaganda, conspiracy theories, or attacks against certain demographics.
2. This monster was then finetuned on higher quality data – think StackOverflow, Quora, or human annotations – which makes it somewhat socially acceptable.
3. Then the finetuned model was further polished using RLHF to make it customer-appropriate, e.g. giving it a smiley face.



https://huyenchip.com/2023/05/02/rlhf.html

https://medium.com/@zaiinn440/reinforcement-learning-from-human-feedback-rlhf-empowering-chatgpt-with-user-guidance-95858592fdbb