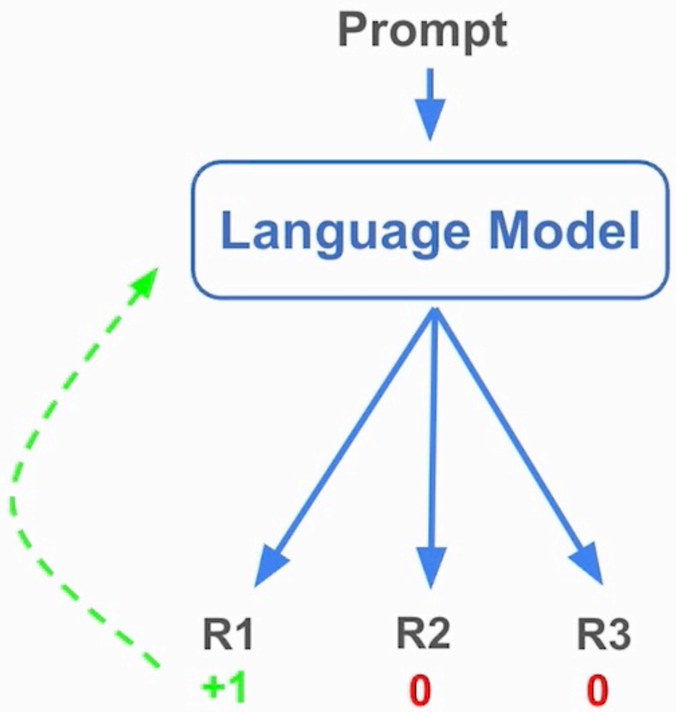
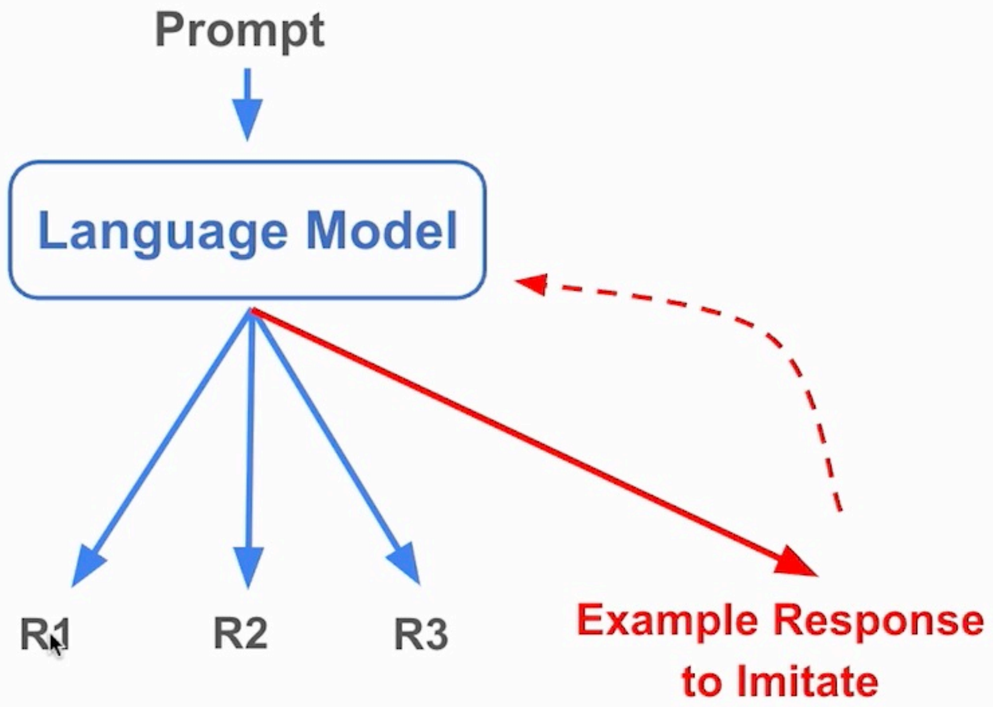
Common methods in post-training

Methods	Principles	Pros & Cons
Supervised Fine-tuning (SFT)	Imitate the example responses by maximizing the probability of the response	**Pros:** simple implementation, great for jump-starting new model behavior **Cons:** may degrade other performances for tasks not included in training data
Online Reinforcement Learning (e.g. PPO, GRPO)	Maximize the reward for the response	**Pros:** better at improving model capabilities without degrading performance in unseen tasks **Cons:** most complex implementation; requires good design of reward functions
Direct Preference Optimization (DPO)	Encourage good answer while discouraging bad answer provided	**Pros:** train model in a contrastive fashion; good at fixing wrong behaviors and improving targeted capabilities **Cons:** may be prone to overfitting; implementation complexity in between SFT & Online RL

Why Online RL degrades performance less compared with SFT?



Online RL tweaks behaviour within the model's native manifold



SFT drags it into an alien one, risking unnecessary changes of model weights