Reward Hacking

An interesting problem that can emerge in reinforcement learning is known as reward hacking, where the agent learns to cheat the system by favoring actions that maximize the reward received even if those actions don't align well with the original objective.

as the policy tries to optimize the reward, it can diverge too much from the initial language model.

A diagram of a problem

Description automatically generated

The model could also start generating nonsensical, grammatically incorrect text that just happens to maximize the rewards

Avoid Reward Hacking

A diagram of a diagram of a reference model

Description automatically generated

Evaluate the Model

You can use the summarization data set to quantify the reduction in toxicity,

A diagram of a person's body

Description automatically generated