# Reward Model (RM) Training via Ranking

After training the SFT model, the authors wanted to build a new model capable of predicting which results humans prefer.

Human labelers ranked the model’s outputs by preference, from 1 (worst) to 7 (best).

These rankings were then used to train a Reward Model (RM) that predicts a numerical value (“reward”) representing how much a given output is preferred.

First, the previous model generates multiple responses for the same prompt. Then, labelers rank those responses, and the RM is trained to assign higher rewards to the preferred outputs.

The loss function used is:

* : average loss over all samples;
* : reward that the RM gives to response for prompt ;
* : preferred output;
* : less preferred output;
* : sigmoid function, which ensures the model assigns a higher reward to the preferred response.

# Reinforcement Learning (RLHF) via PPO

The Reward Model now acts as a judge for the SFT model.

The process works as follows:

**SFT model generates outputs** → **RM evaluates them** → **RLHF model learns to maximize the reward.**

To achieve this, the authors used PPO (Proximal Policy Optimization), with the SFT model as the initial policy.

However, during training, the model might “forget” what it learned during SFT. To prevent this, a penalty term was added to ensure the model stays close to the original behavior:

This term is included in the PPO objective:

where:

* : reward from the RM;
* : current policy (the model being updated);
* : base SFT Model;
* : controls how strongly the model is penalized for deviating from SFT.

# Reinforcement Learning (RLHF-ptx) via PPO-ptx

After training, the authors found that the model followed instructions better but performed worse on general NLP tasks like summarization and question answering.

To fix this, they introduced a variant called PPO-ptx, which mixes part of the pretraining gradients with the PPO training process.

This approach allowed the model to keep learning from the Reward Model while also preserving its general NLP abilities and prior knowledge.