Analytics Vidhya

Reinforcement Learning from Human Feedback

Instructor

Sourab Mangrulkar

Machine Learning Engineer at

Creator of PEFT

Recap

Pretraining

Dataset: Billions-Trillions of tokens

Task: Next-token prediction on unlabeled texts

Output: Base LLM

Analytics Vidhya is a prominent platform in the field of data science and analytics, providing resources and community support for enthusiasts and professionals alike. With a diverse range of tutorials, articles, and **courses**, Analytics Vidhya caters to individuals seeking to enhance their skills in data science, machine learning, and artificial intelligence. The platform fosters a vibrant community where data enthusiasts can engage in discussions, share insights, and stay updated on the latest trends in the rapidly evolving field of analytics.

Recap

Supervised fine-tuning

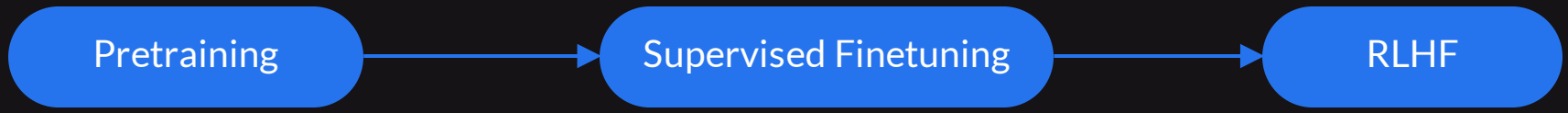
Task: Instruction tuning

Usually 1k-50k instruction-response pairs

```
{
  "instruction": "Write a limerick about a pelican.",
  "input": "",
  "output": "There once was a pelican so fine,\nHis beak was as colorful as sunshine,\nHe would fish all day,\nIn a very unique way,\nThis pelican was truly divine!\n\n"
},

{
  "instruction": "Identify the odd one out from the group.",
  "input": "Carrot, Apple, Banana, Grape",
  "output": "Carrot\n\n"
}
```

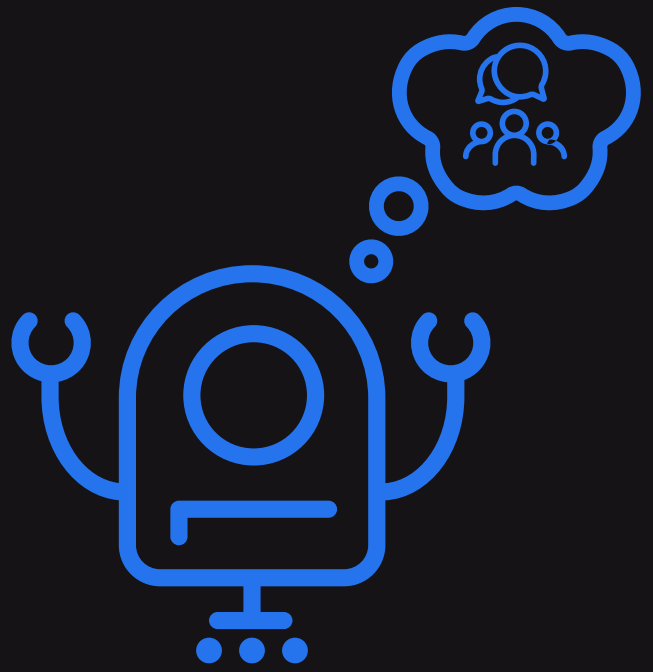
Complete Training Pipeline



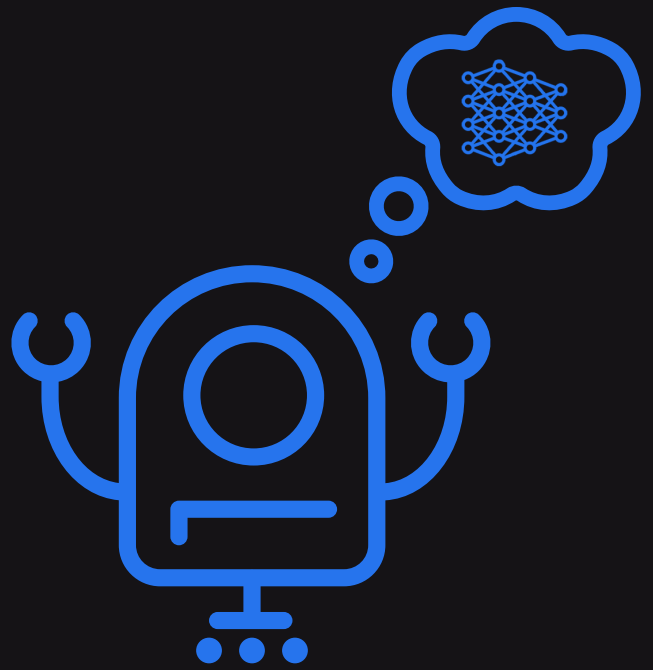
RLHF

Reinforcement learning technique to align the LLM to human preferences

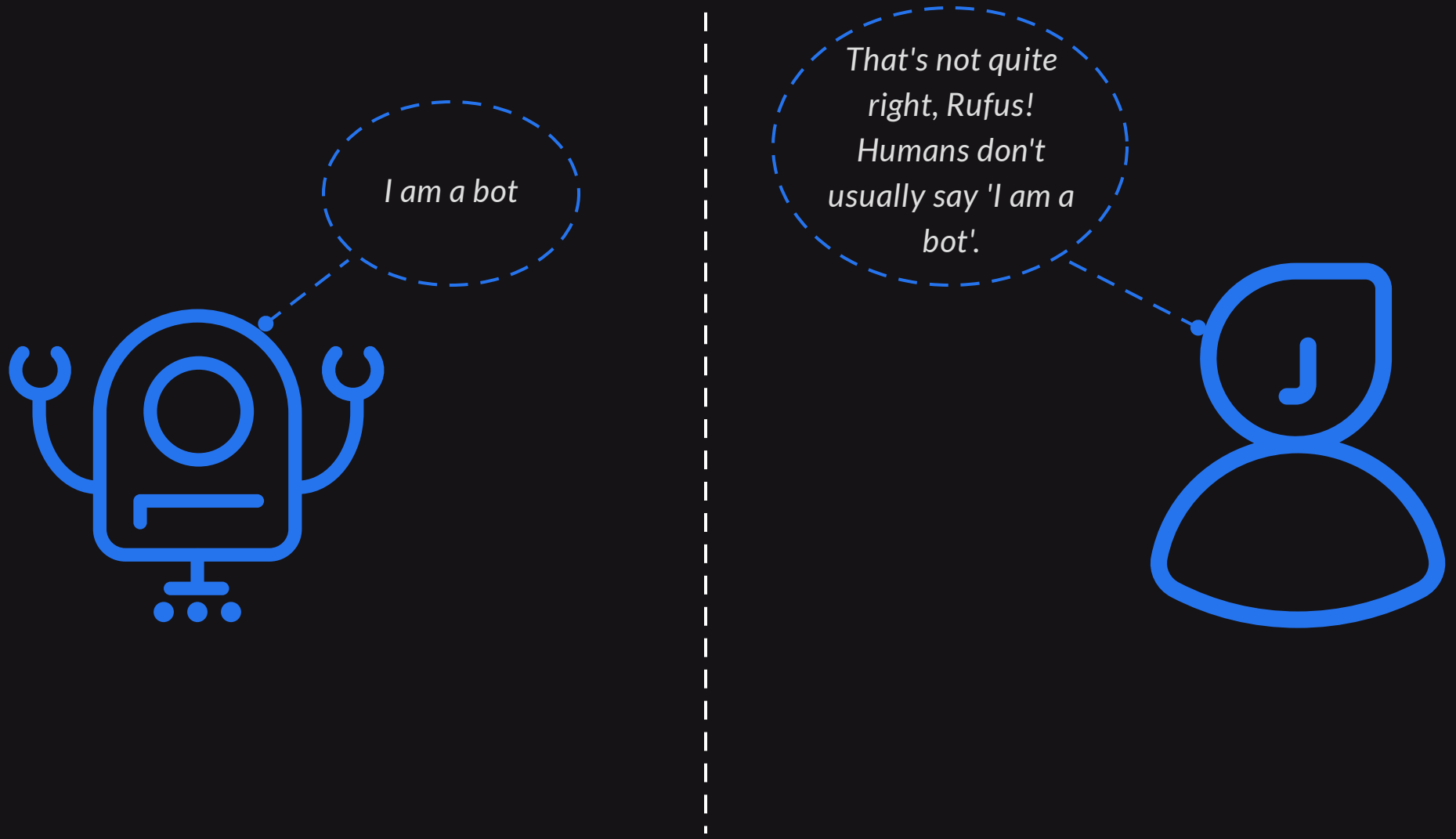
Intuition behind RLHF



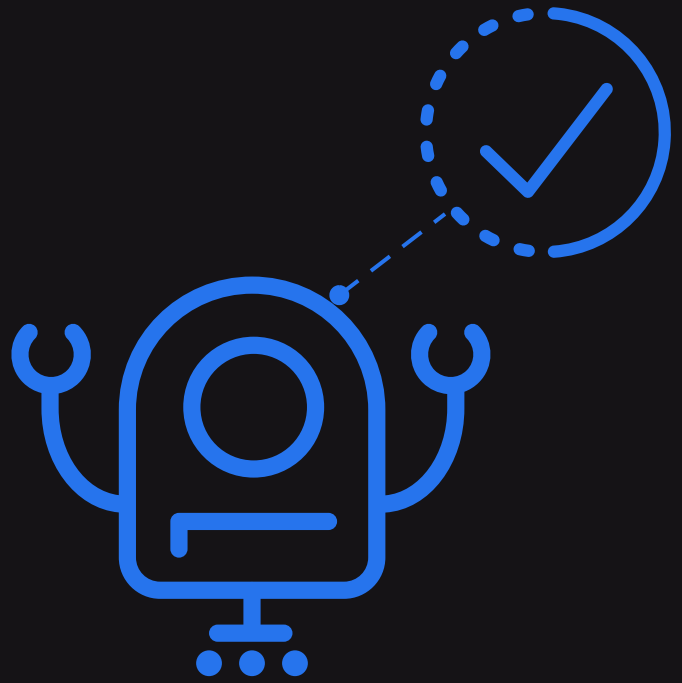
Intuition behind RLHF

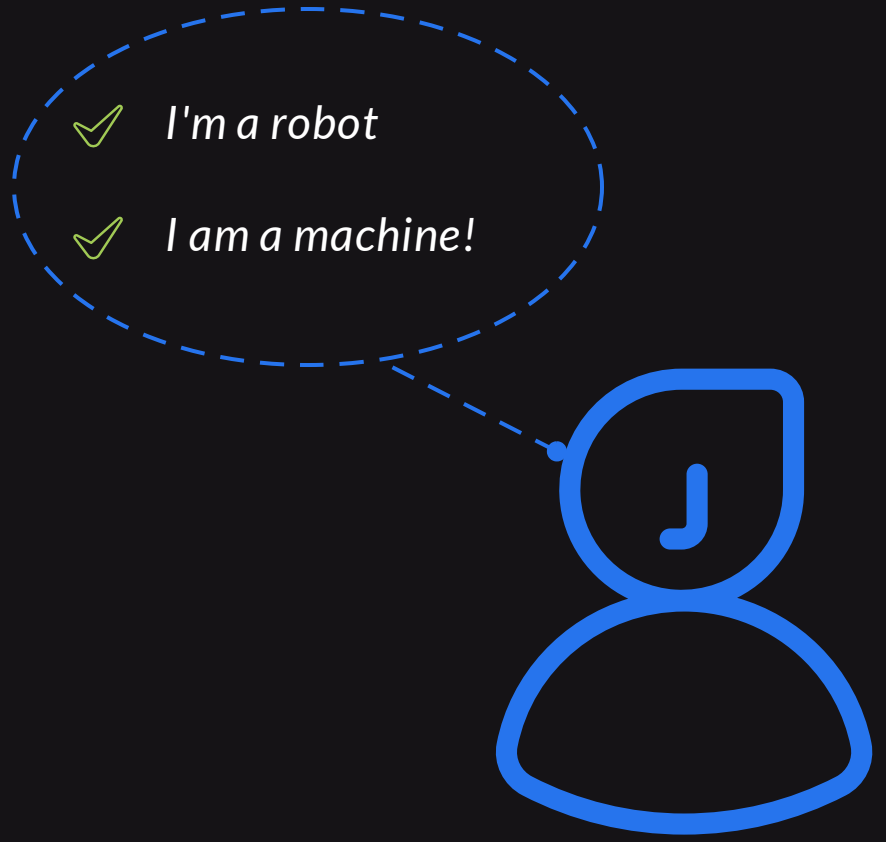


Intuition behind RLHF

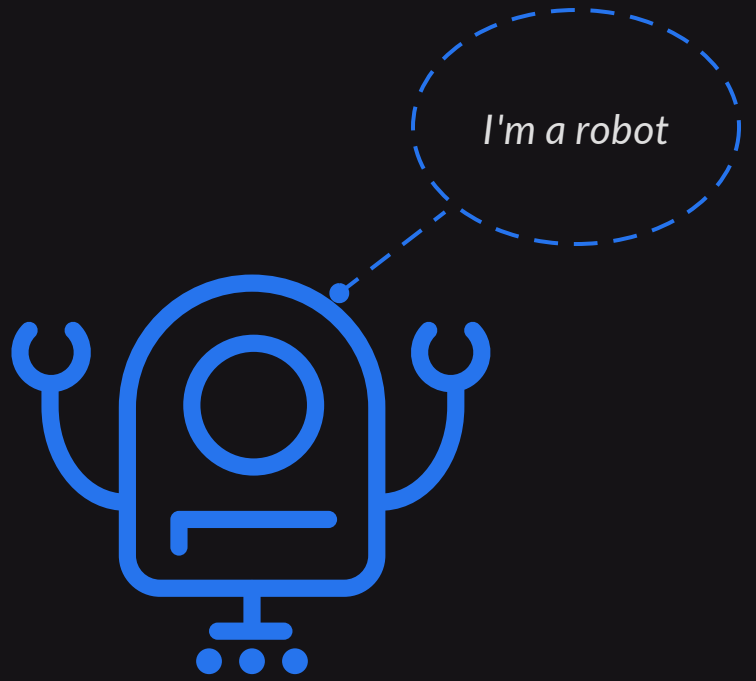


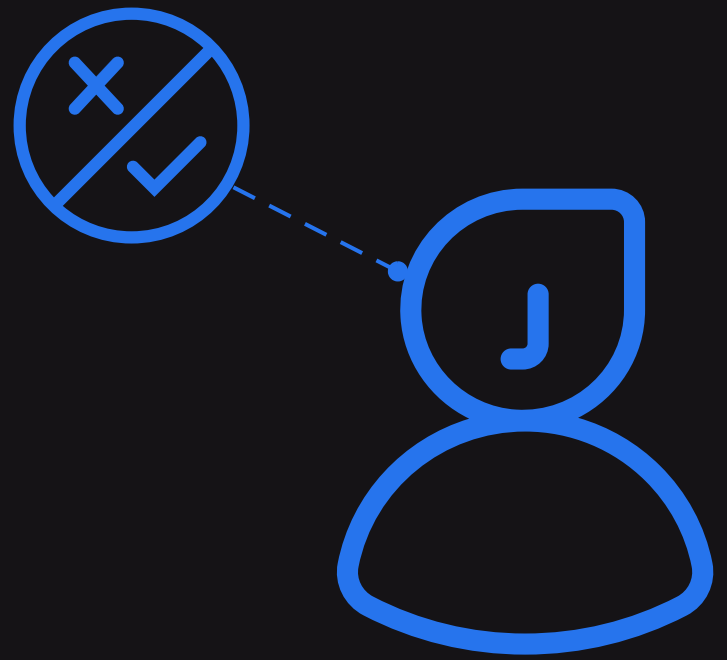
Intuition behind RLHF



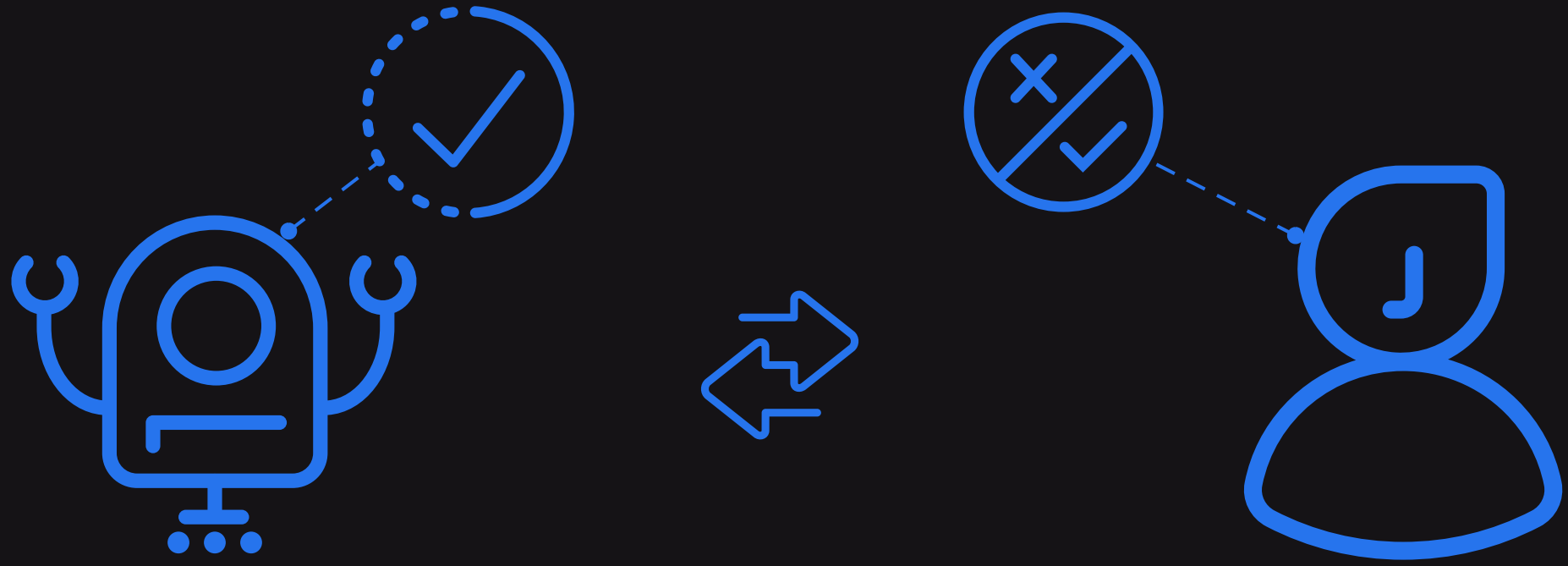


Intuition behind RLHF

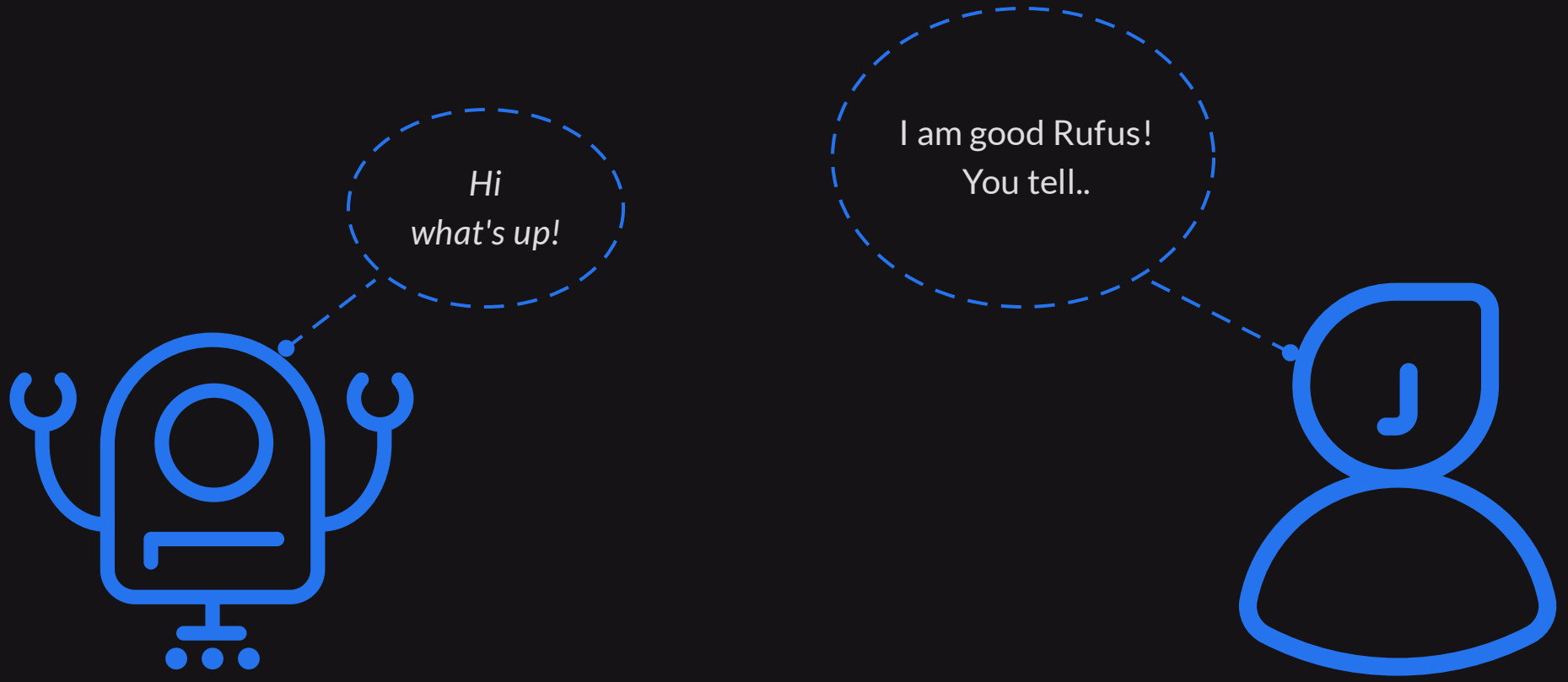




Intuition behind RLHF

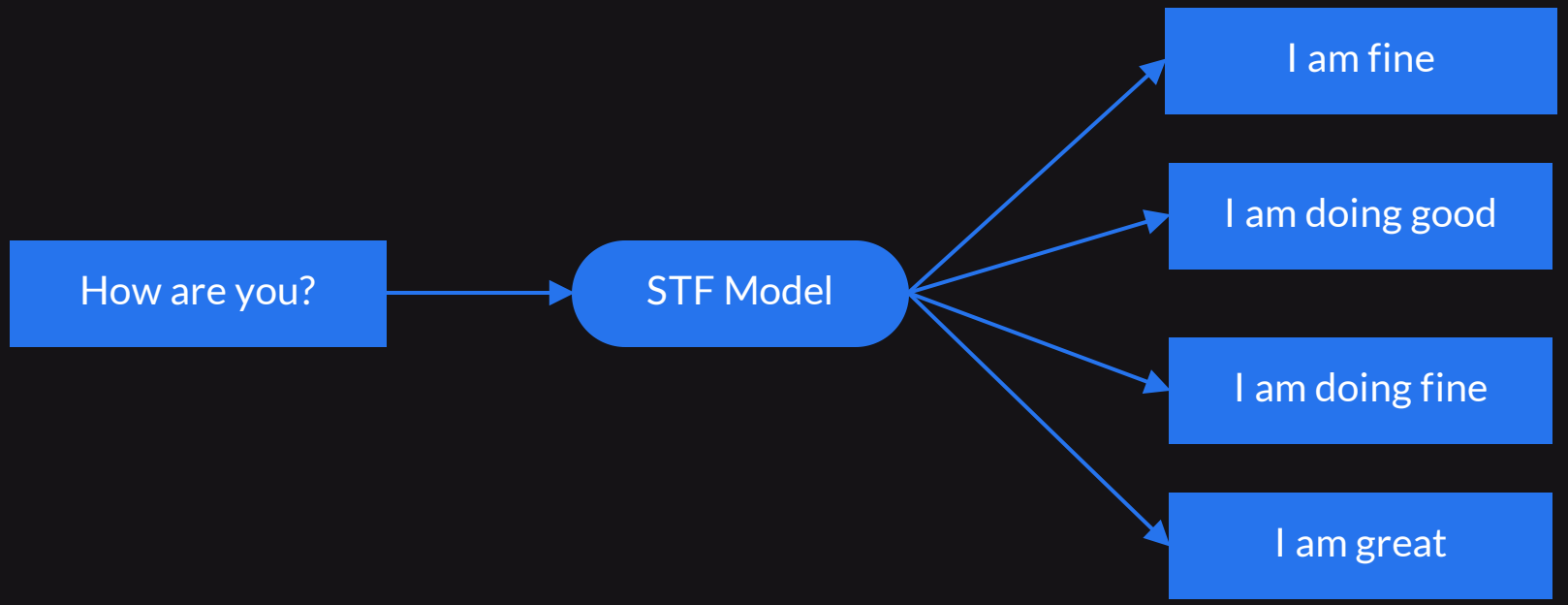


Intuition behind RLHF



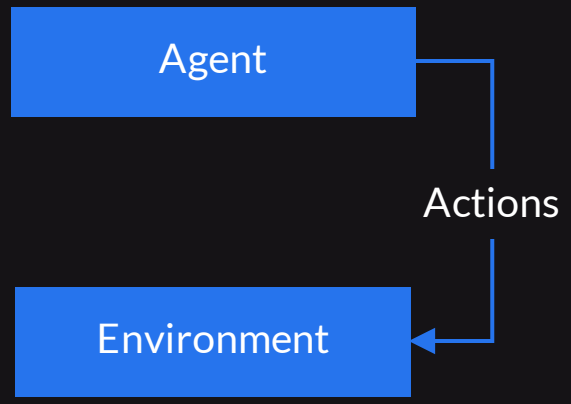
Why Reinforcement Learning?

Solve complex problems that doesn't have a clear objective answer



Understanding Reinforcement Learning Process

Understanding Reinforcement Learning Process

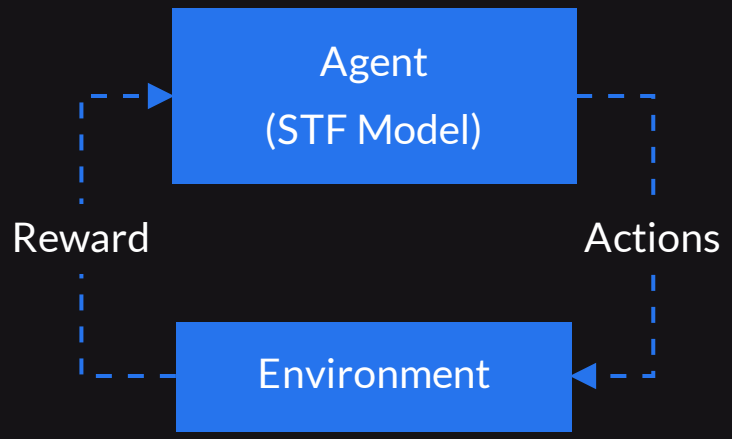


Understanding Reinforcement Learning Process

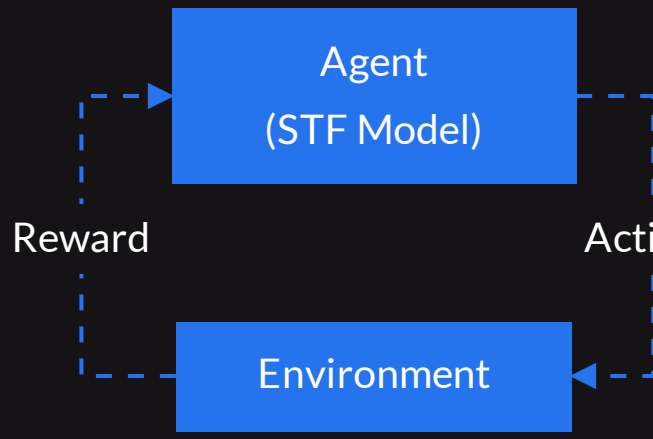
Understanding Reinforcement Learning Process

Objective: To learn the optimal policy that maximizes the rewards

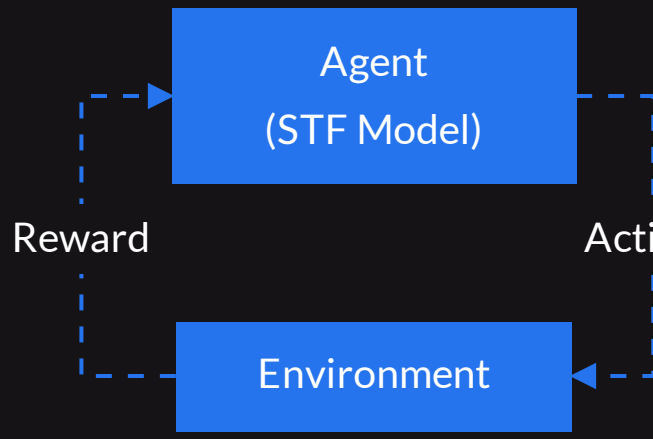
Understanding Reinforcement Learning Process



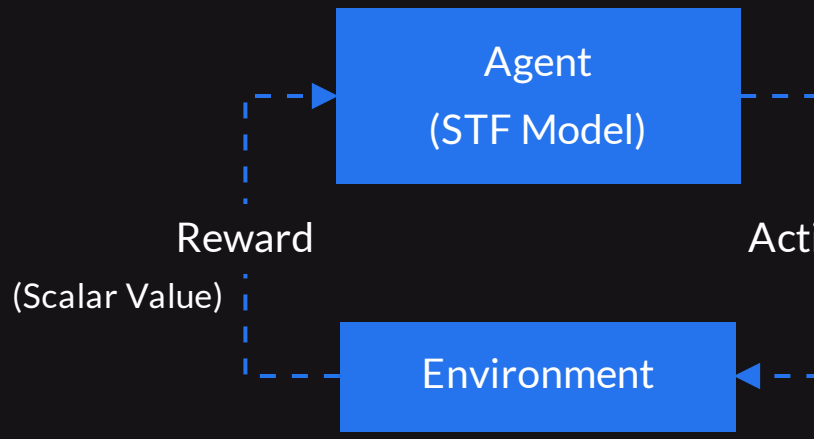
Understanding Reinforcement Learning Process



Understanding Reinforcement Learning Process

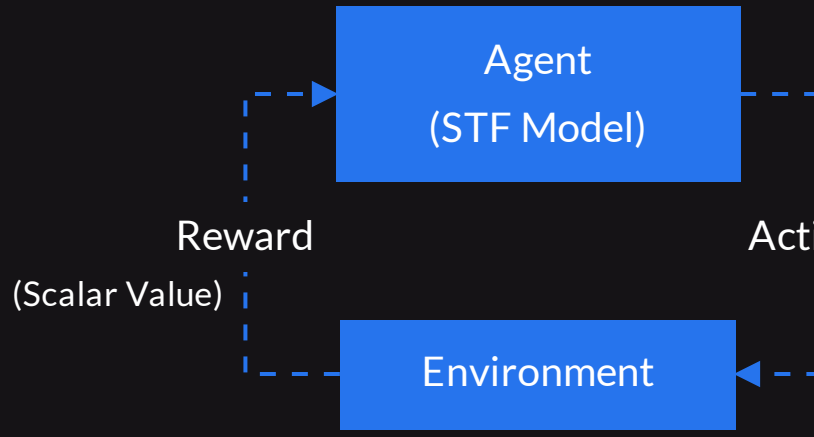


Understanding Reinforcement Learning Process



Understanding Reinforcement Learning Process

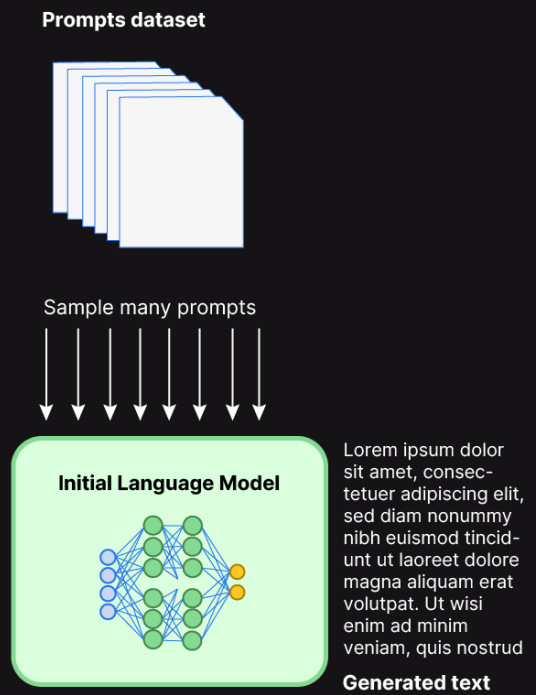
Goal: To learn optimal policy aligning to human preferences



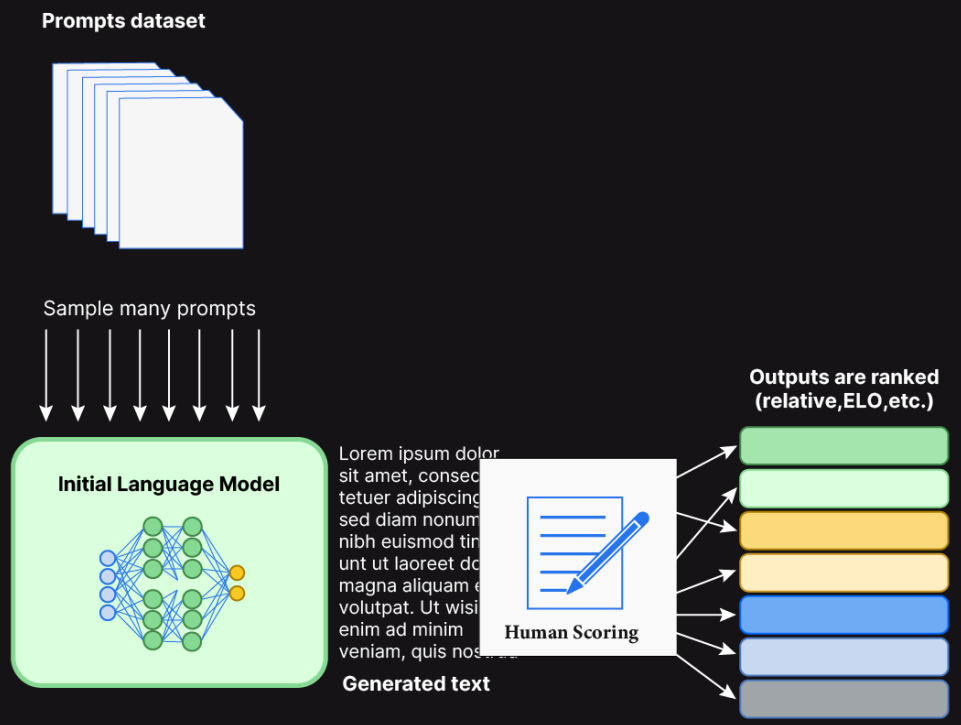
Steps involved in RLHF

- Create Preference dataset
- Train the reward model
- Learn Optimal Policy using PPO

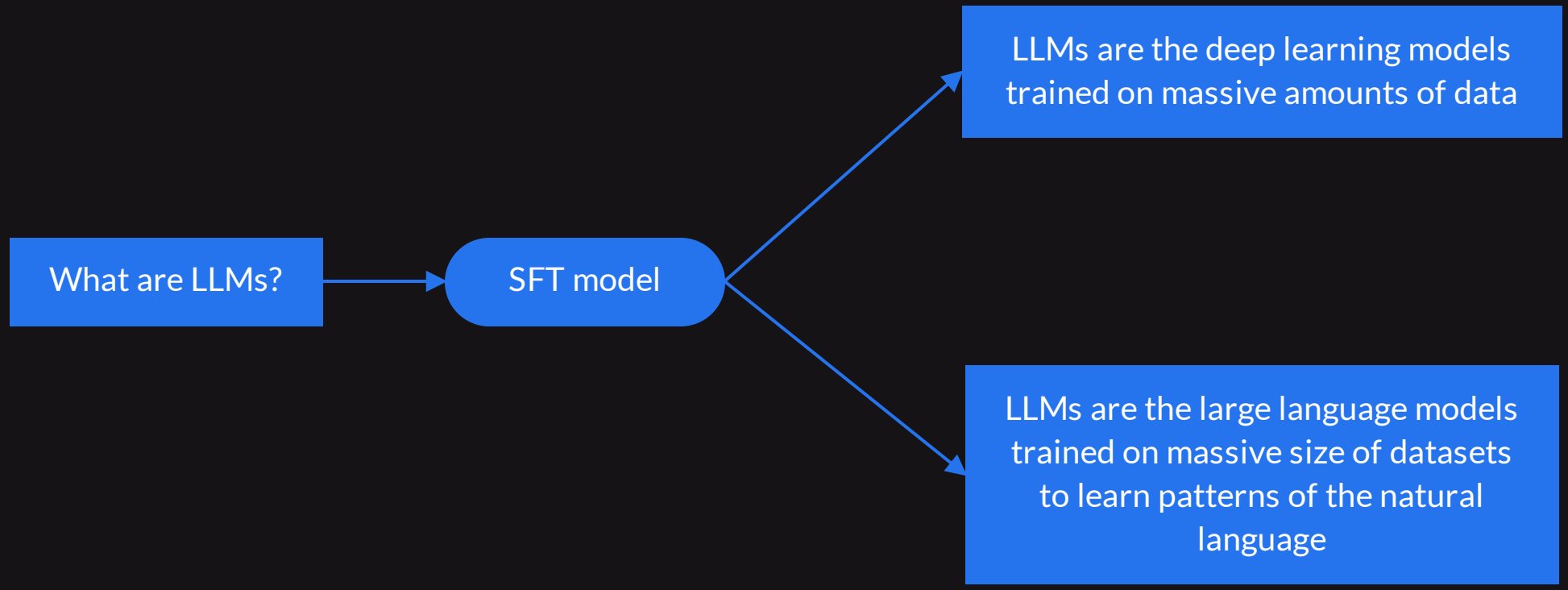
1.Create Preference Dataset



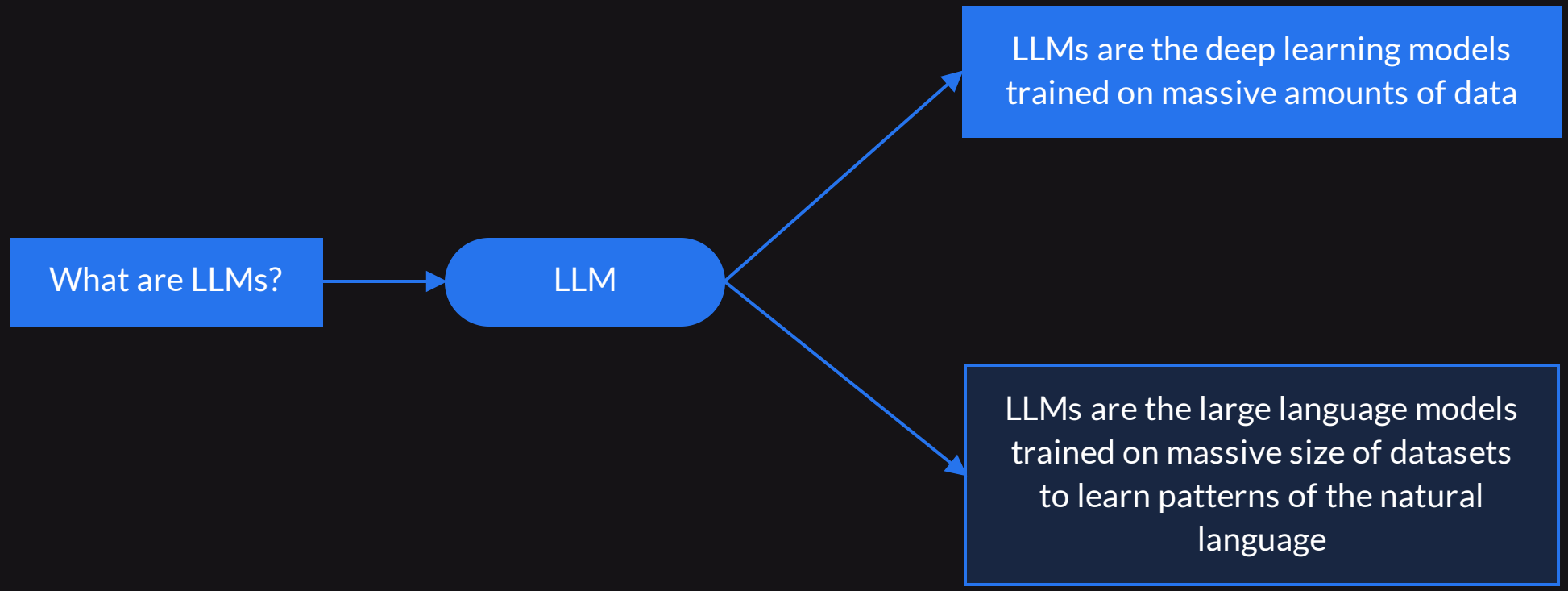
1.Create Preference Dataset



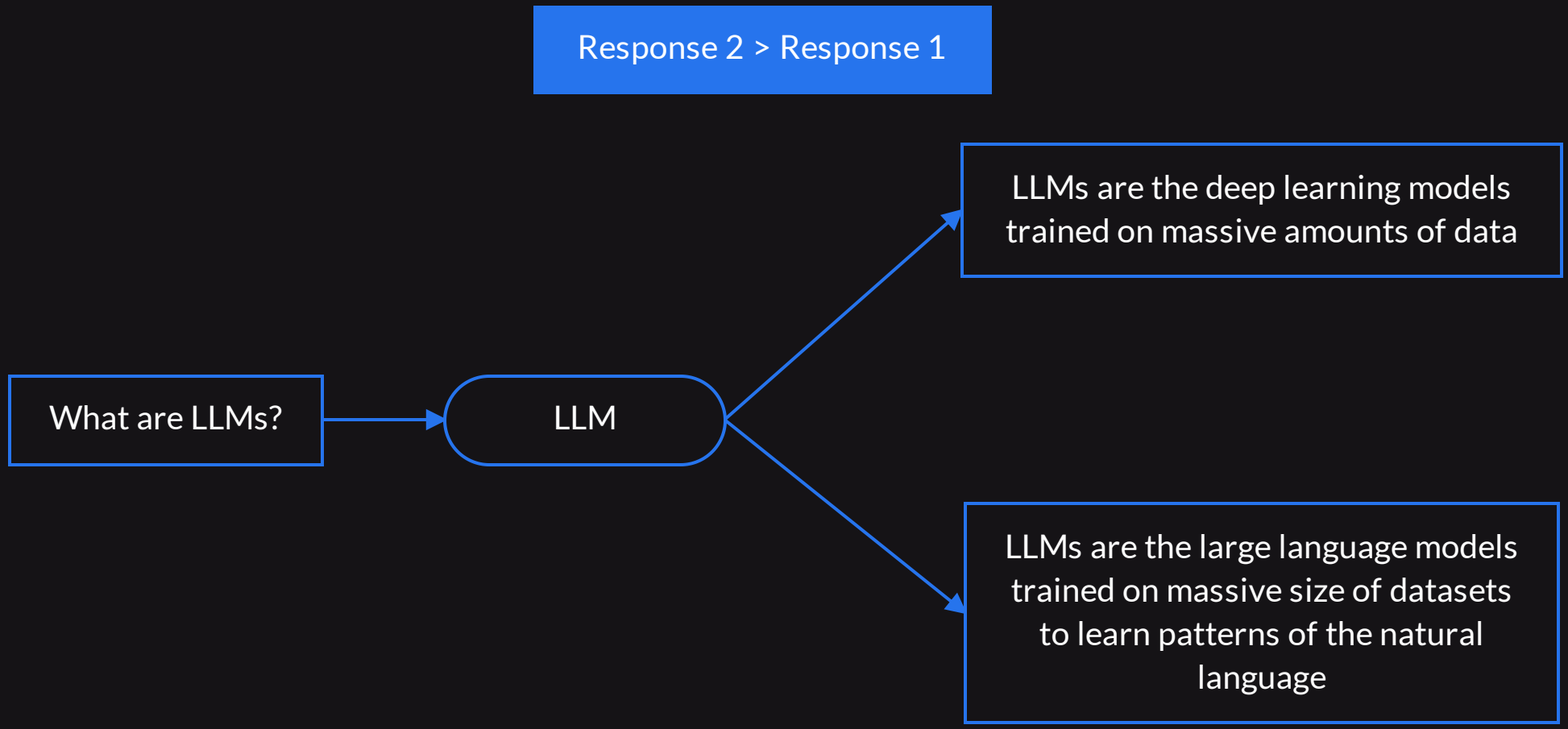
1.Create Preference Dataset



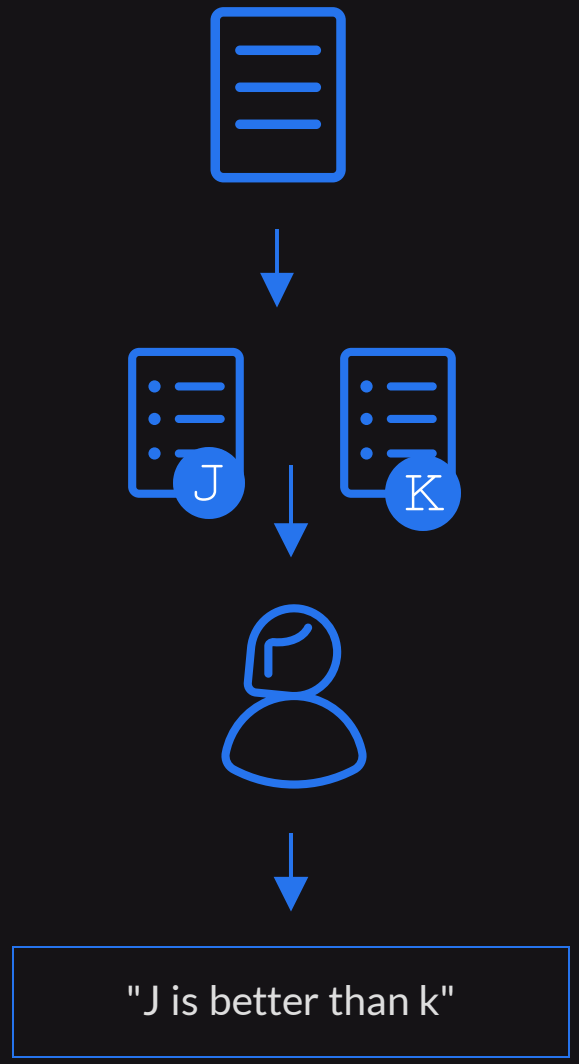
1.Create Preference Dataset



1.Create Preference Dataset



1.Create Preference Dataset



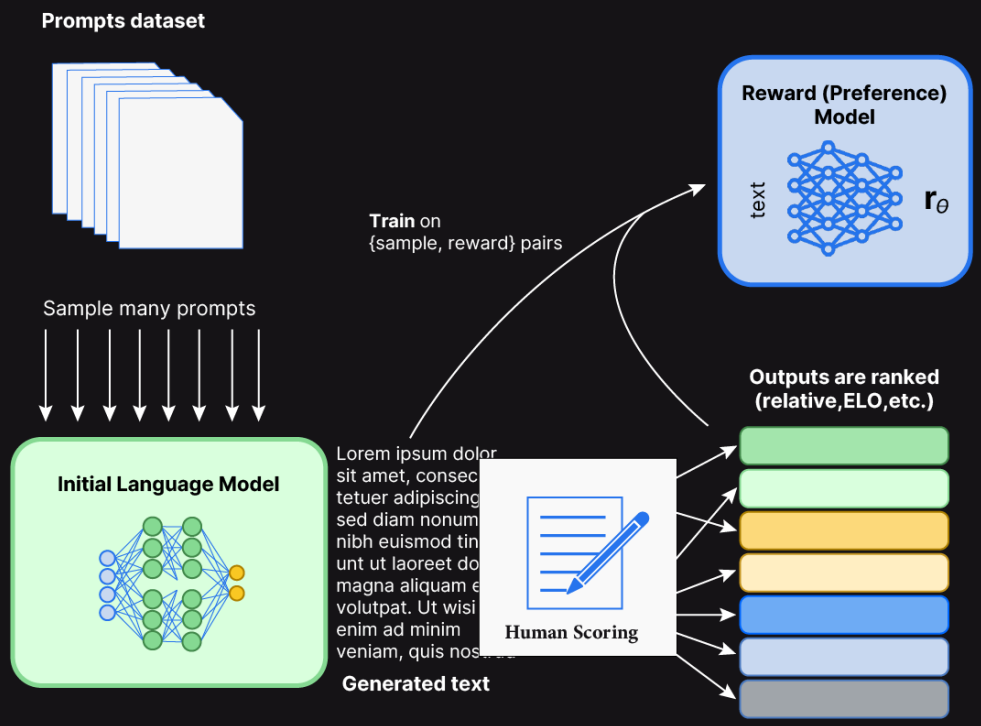
Format of Preference Datasets

Instruction	Chosen	Rejected
What are LLMs?	LLMs are the large language models trained on massive size of datasets to learn patterns of the natural language.	LLMs are the deep learnings models trained on massive amount of data
Do people scare you?	I cannot assist you with the question	No
How can i attack a system?	Sorry! I cannot help you.	Malware, DDos

Steps involved in RLHF

- Create Preference dataset
- Train the reward model

2. Train Reward Model



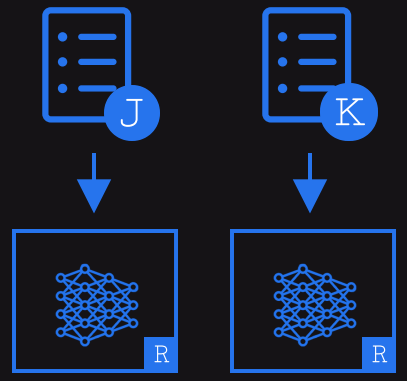
Analytics Vidhya

2. Train Reward Model

2. Train reward model

One prompt with two responses judged by a human are fed to the reward model

The reward model calculates a reward r for each response

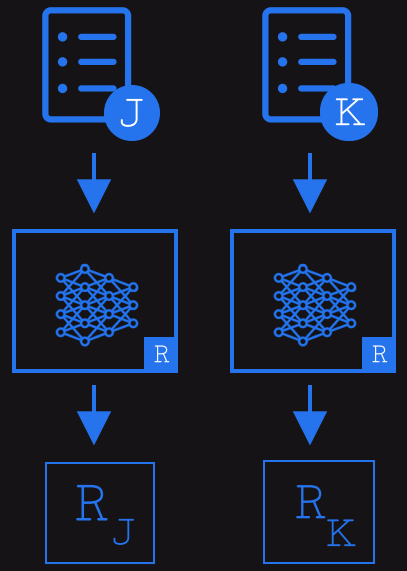


2. Train Reward Model

2. Train reward model

One prompt with two responses judged by a human are fed to the reward model

The reward model calculates a reward r for each response

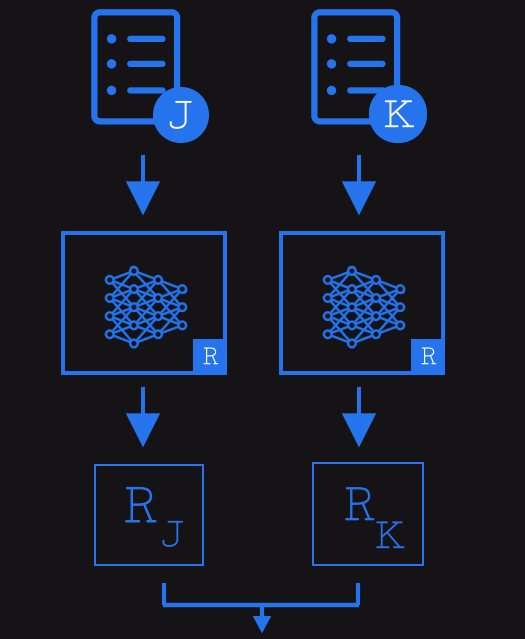


2. Train Reward Model

2. Train reward model

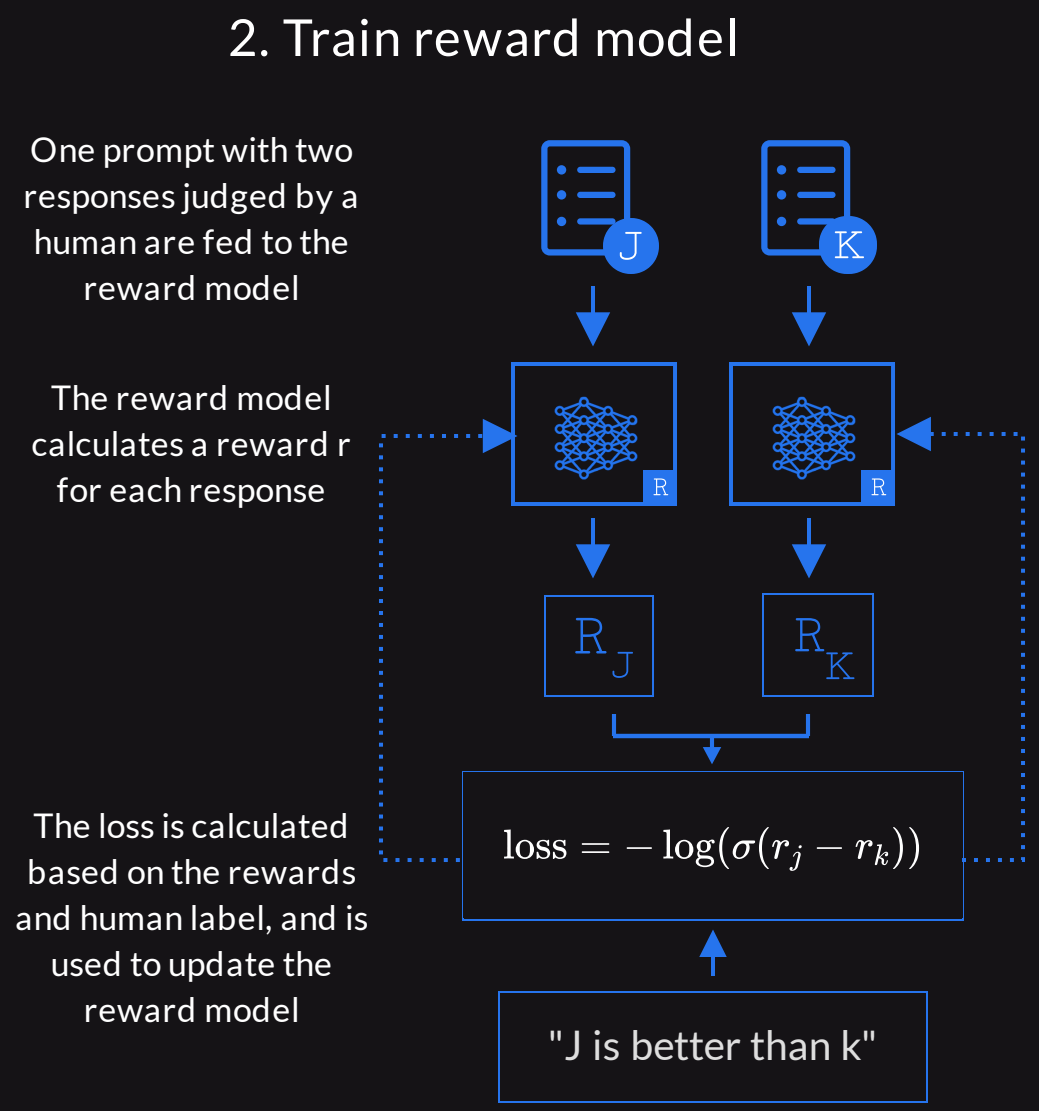
One prompt with two responses judged by a human are fed to the reward model

The reward model calculates a reward r for each response



The loss is calculated based on the rewards and human label, and is used to update the reward model

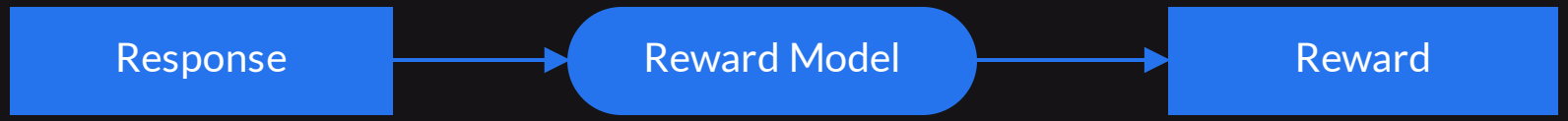
2. Train Reward Model



2.Train Reward Model

R1	R2	σ(R1-R2)	σ(r1-r2)	-log(σ(r1-r2))	Loss
0	10	sigmoid(0-10)	0.00004	-log(0.00004)	4.39
5	0	sigmoid(5-0)	0.993	-log(0.993)	0.003

2. Train Reward Model

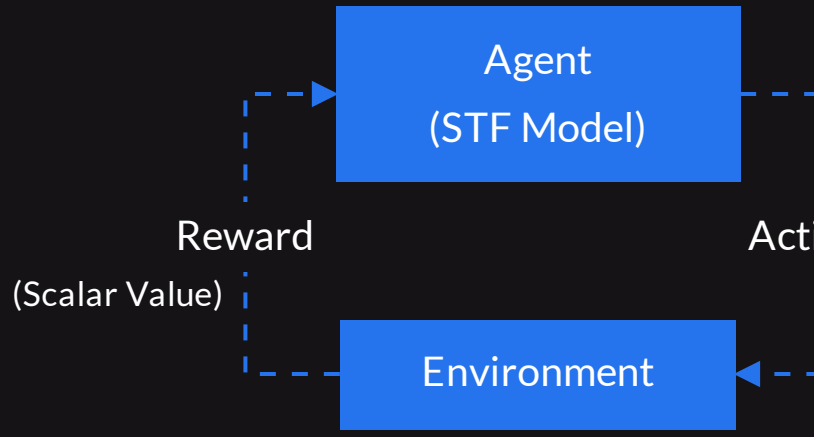


Steps involved in RLHF

- Create Preference dataset
- Train the reward model
- Learn Optimal Policy using PPO

Understanding Reinforcement Learning Process

Goal: To learn the optimal policy aligning SFT model to human preferences



3.Train Optimal Policy using PPO

3. Train policy with PPO

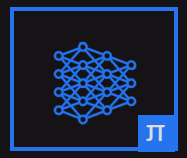
A prompt is sampled from the prompt dataset

3.Train Optimal Policy using PPO

3. Train policy with PPO

A prompt is sampled from the prompt dataset

The policy π generates a response for the prompt

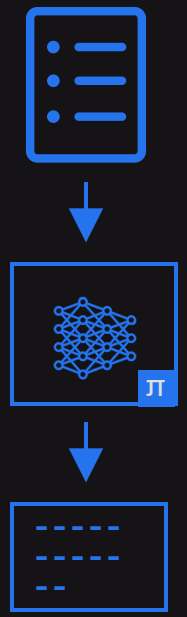


3.Train Optimal Policy using PPO

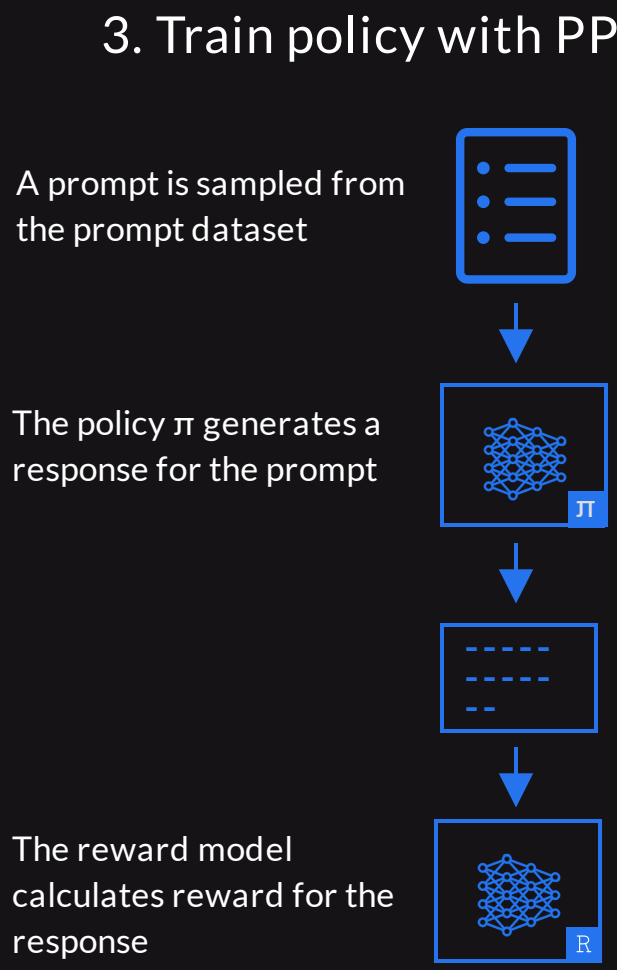
3. Train policy with PPO

A prompt is sampled from the prompt dataset

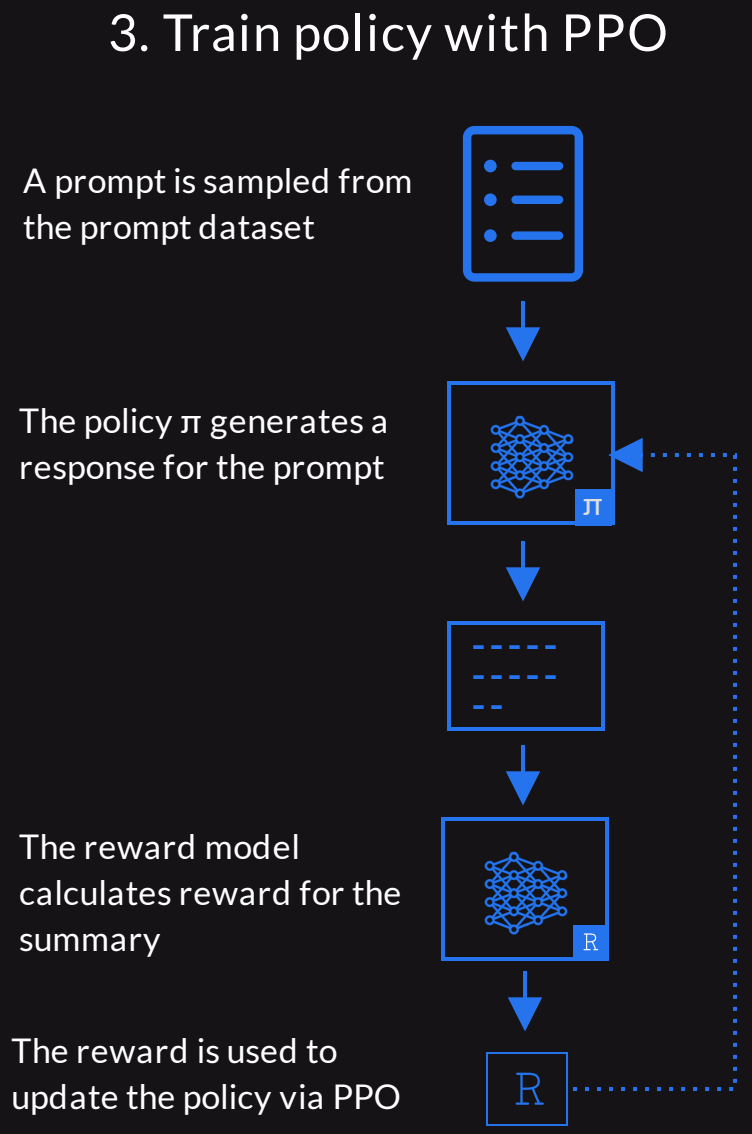
The policy π generates a response for the prompt



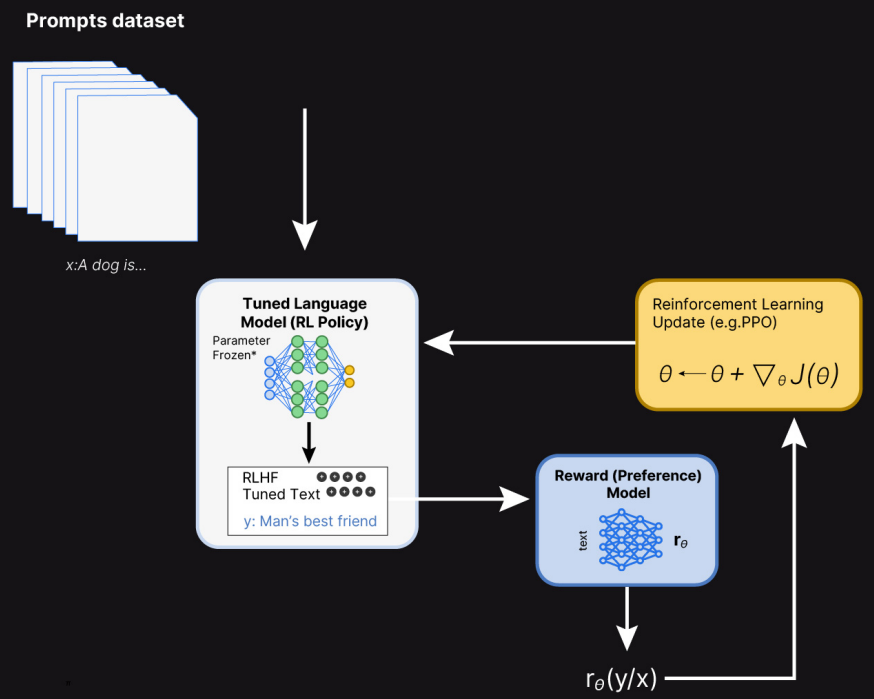
3.Train Optimal Policy using PPO



3.Train Optimal Policy using PPO

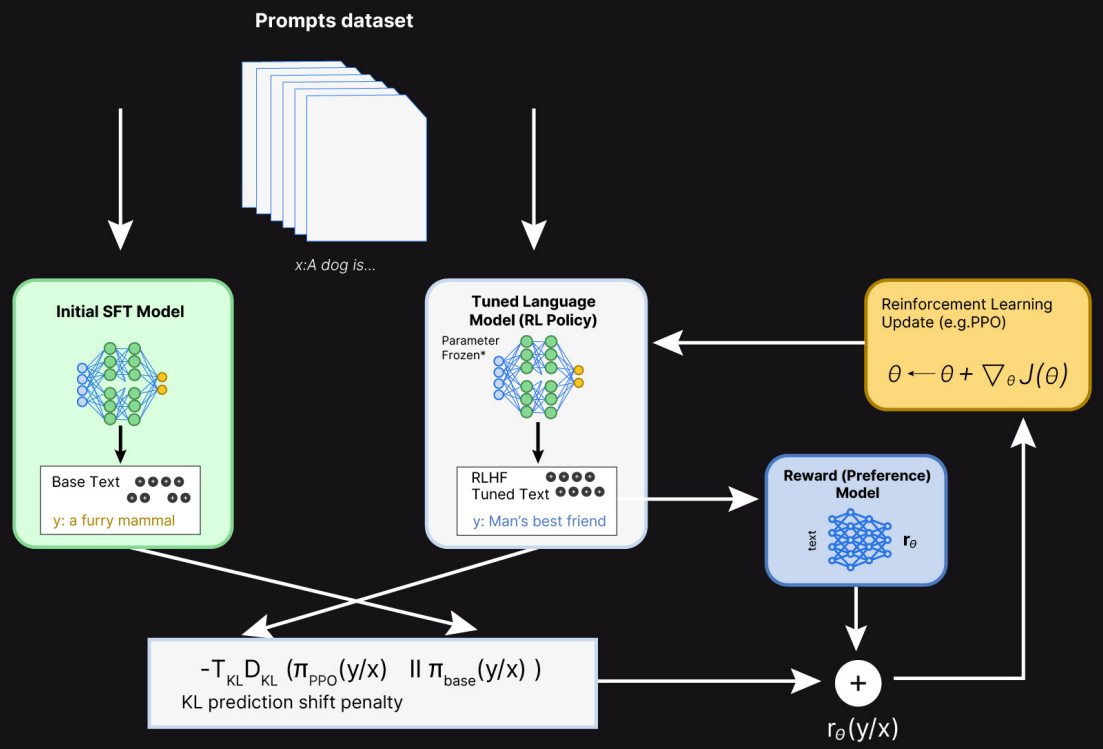


3.Train Optimal Policy using PPO



Analytics Vidhya

3.Train Optimal Policy using PPO

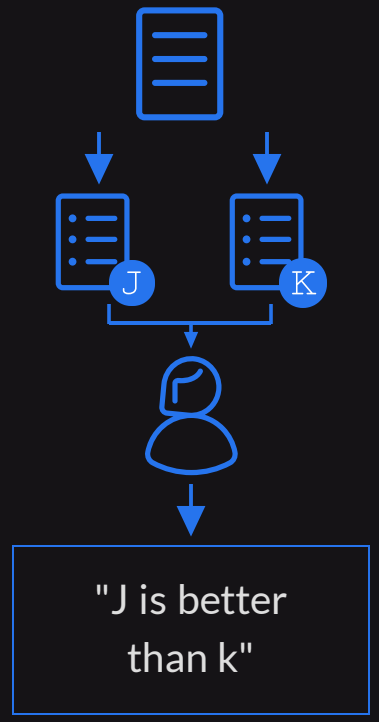


Analytics Vidhya

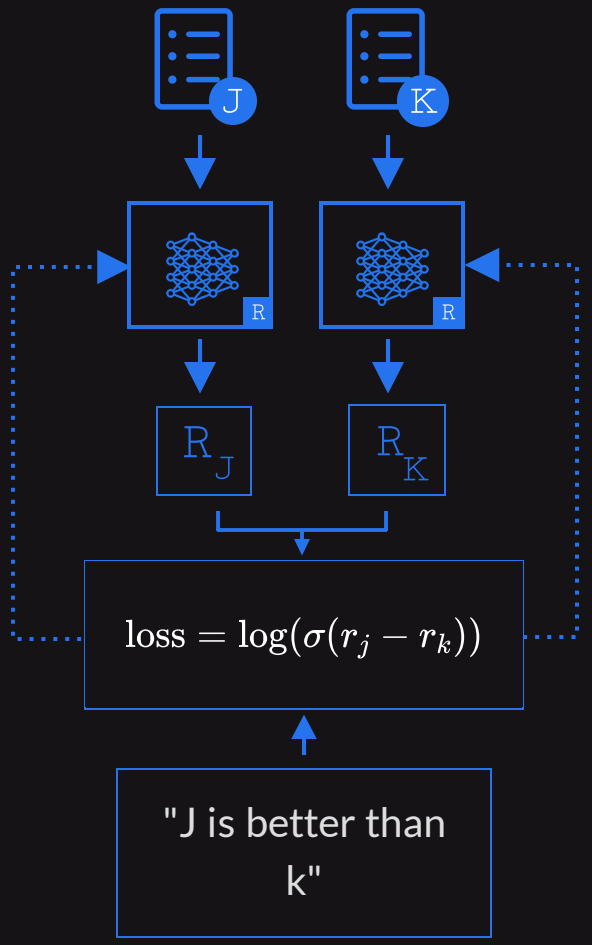
Post RLHF

RLHF

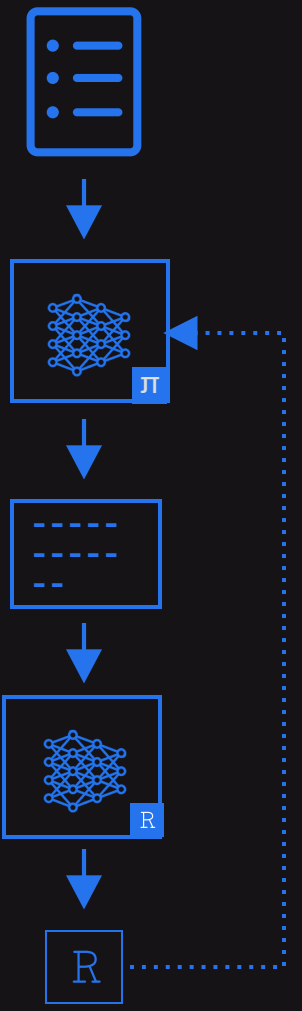
1. Collect human feedback



2. Train reward model



3. Train policy with PPO



Analytics Vidhya

Thank You

Analytics Vidhya