Reinforcement Learning from Human Feedback (RLHF)

基于人工反馈的强化学习

Recap: Instruction Tuning

Instruction Finetuning

Pros:

- Simple and straightforward, generalize to unseen tasks

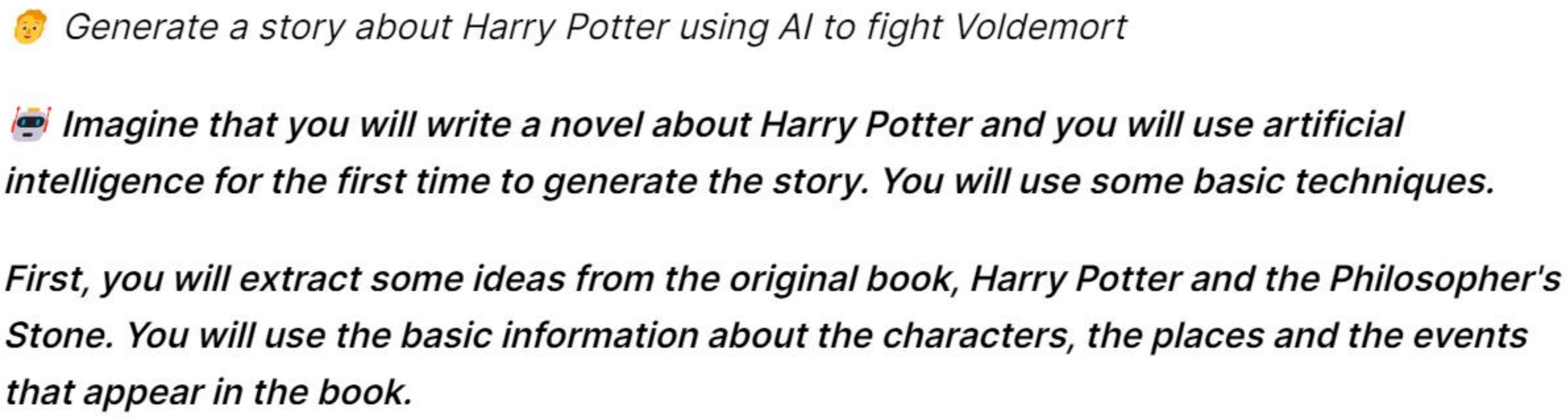
Cons:

- **Expensive** to collect gound-truth data for tasks
- No right answer to tasks **open-ended creative generation**
- **Equal penalization** to all token-level mistakes (even though some are worse than others)

still cannot maximize human preferences

Recap: Instruction Tuning

- **Limitation in tasks like open-ended creative generation:**
  GPT-3 can continue a prompt but fails understand a request



🧑 *Generate a story about Harry Potter using AI to fight Voldemort*

🤖 ***Imagine that you will write a novel about Harry Potter and you will use artificial intelligence for the first time to generate the story. You will use some basic techniques.***

First, you will extract some ideas from the original book, Harry Potter and the Philosopher's Stone. You will use the basic information about the characters, the places and the events that appear in the book.

- **Difficult to create loss function for abstract objectives:**
  - what is funny?
  - what is ethical?
  - what is safe?

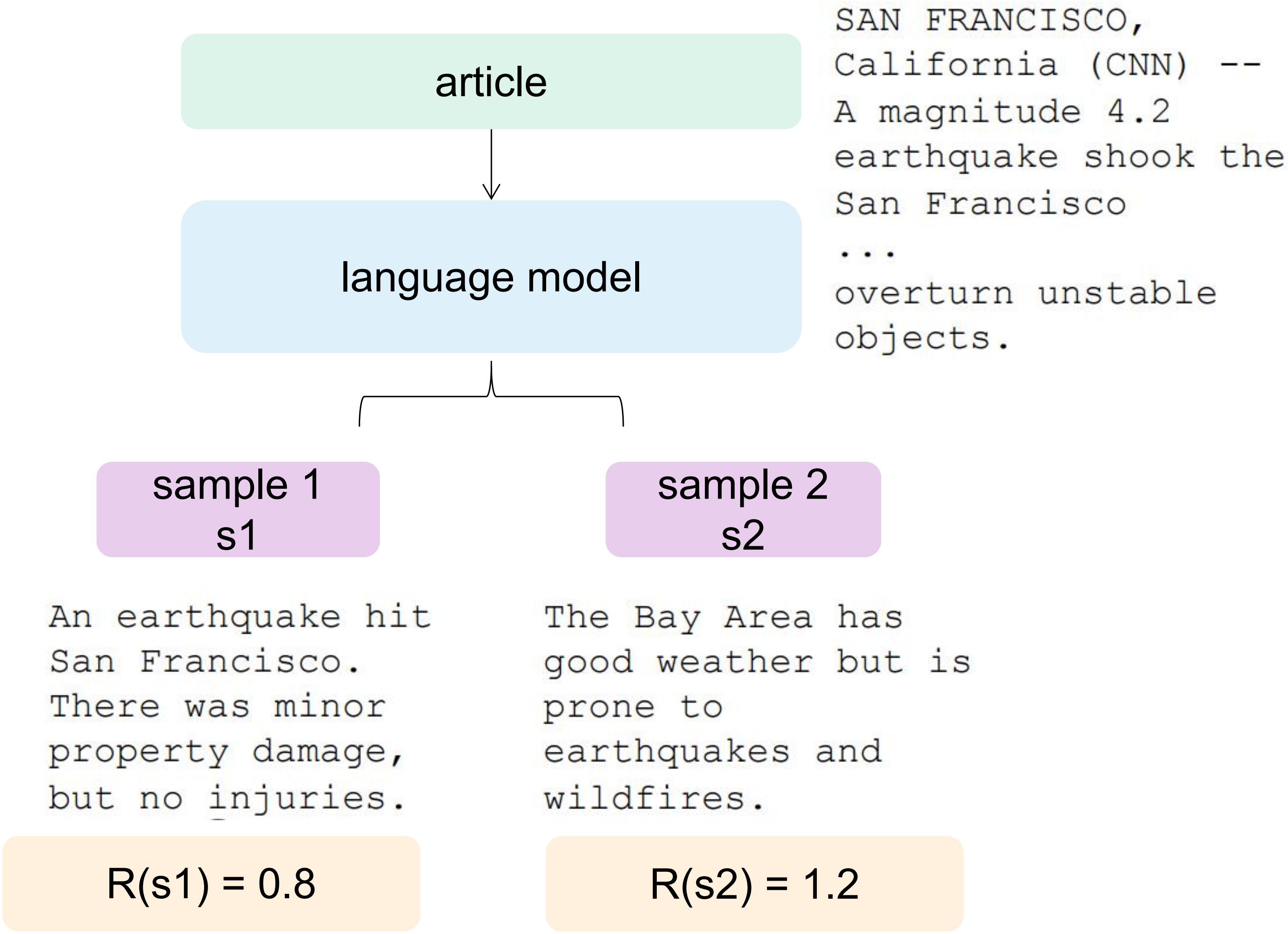
RLHF - Optimizing on Human Preferences

Example: Summarization Task

For each sample s, we assign a human reward R(s) to that summary.

The higher the reward, the better the summary is based on human preference.

The language model is optimized by maximizing the expected reward of samples.



article

SAN FRANCISCO,
California (CNN) --
A magnitude 4.2
earthquake shook the
San Francisco
...
overturn unstable
objects.

language model

sample 1
s1

An earthquake hit
San Francisco.
There was minor
property damage,
but no injuries.

R(s1) = 0.8

sample 2
s2

The Bay Area has
good weather but is
prone to
earthquakes and
wildfires.

R(s2) = 1.2

Contents

- **强化学习和PPO算法**
- InstructGPT/ChatGPT中的RLHF

Contents

- **强化学习和PPO算法**
  - **Motivation of RLHF**
    - Human feedback
    - Reinforcement learning
  - Introduction of reinforcement learning
    - Agent
    - Reward
    - RL algorithm
      - policy gradient (PG)
      - proximal policy gradient (PPO)

Motivation

Reinforcement Learning from Human Feedback

基于人工反馈的强化学习

Motivation

Reinforcement Learning from Human Feedback

基于人工反馈的强化学习

Motivation

Reinforcement Learning from Human Feedback

基于**人工反馈**的**强化学习**

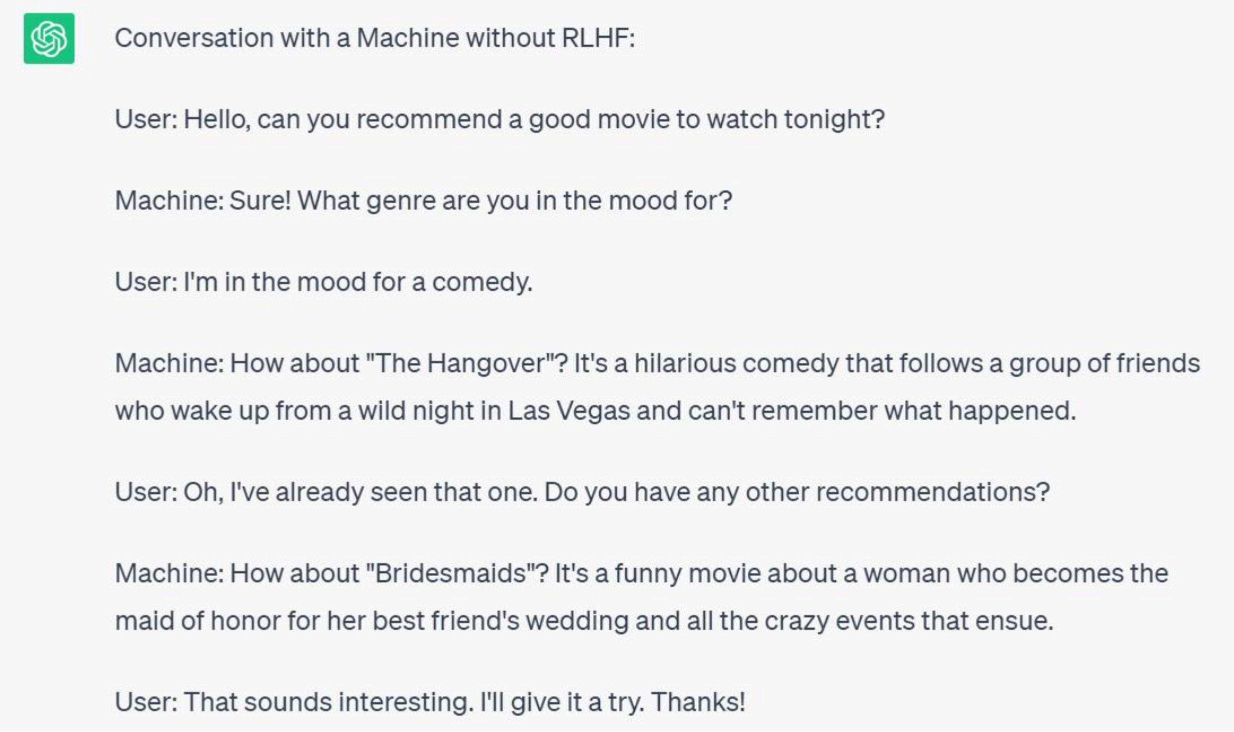
Motivation

Reinforcement Learning from Human Feedback

基于**人工反馈**的**强化学习**

Human Feedback

- **Human Feedback** in conversations



Conversation with a Machine without RLHF:

User: Hello, can you recommend a good movie to watch tonight?

Machine: Sure! What genre are you in the mood for?

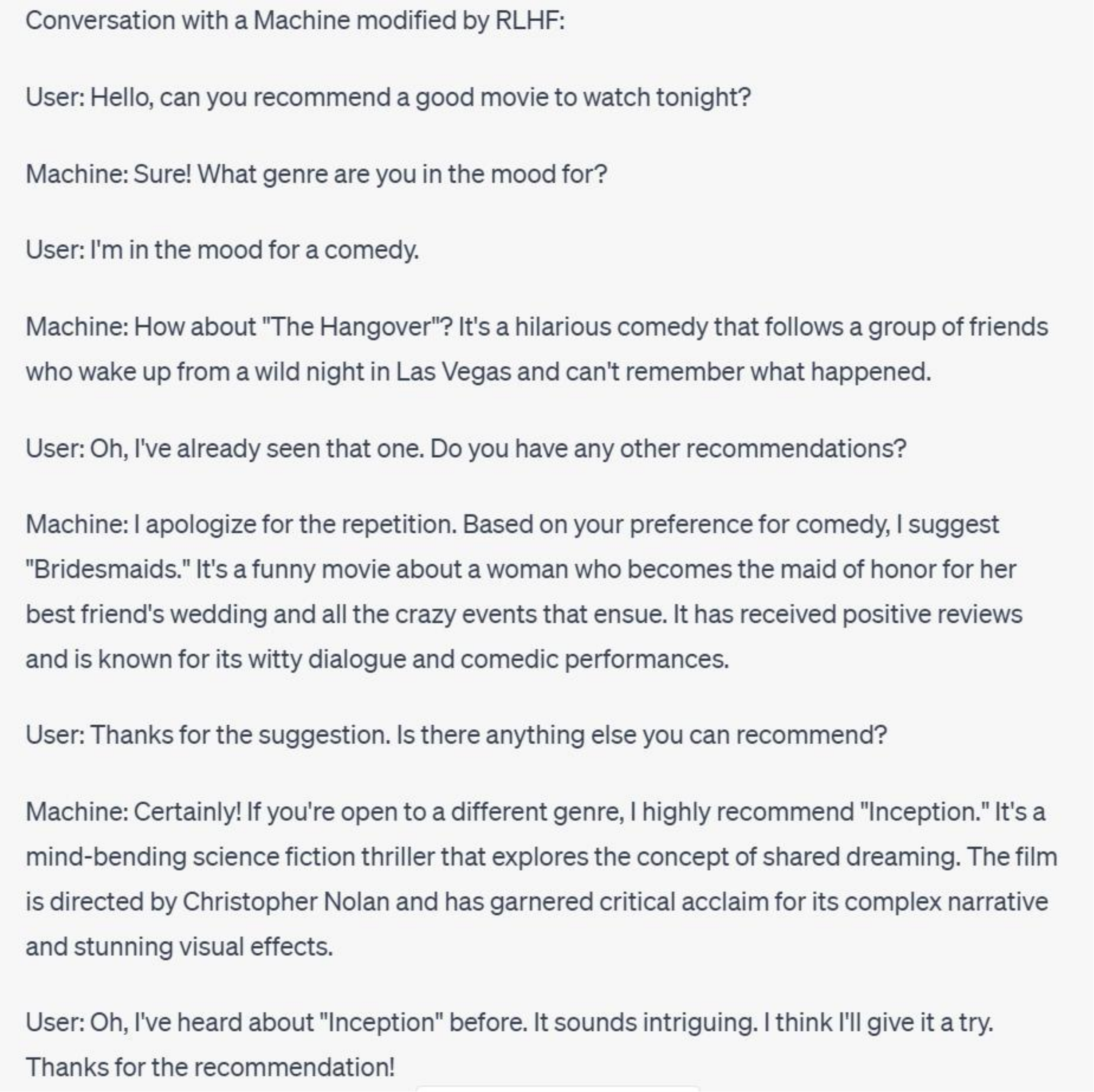
User: I'm in the mood for a comedy.

Machine: How about "The Hangover"? It's a hilarious comedy that follows a group of friends who wake up from a wild night in Las Vegas and can't remember what happened.

User: Oh, I've already seen that one. Do you have any other recommendations?

Machine: How about "Bridesmaids"? It's a funny movie about a woman who becomes the maid of honor for her best friend's wedding and all the crazy events that ensue.

User: That sounds interesting. I'll give it a try. Thanks!



Conversation with a Machine modified by RLHF:

User: Hello, can you recommend a good movie to watch tonight?

Machine: Sure! What genre are you in the mood for?

User: I'm in the mood for a comedy.

Machine: How about "The Hangover"? It's a hilarious comedy that follows a group of friends who wake up from a wild night in Las Vegas and can't remember what happened.

User: Oh, I've already seen that one. Do you have any other recommendations?

Machine: I apologize for the repetition. Based on your preference for comedy, I suggest "Bridesmaids." It's a funny movie about a woman who becomes the maid of honor for her best friend's wedding and all the crazy events that ensue. It has received positive reviews and is known for its witty dialogue and comedic performances.

User: Thanks for the suggestion. Is there anything else you can recommend?

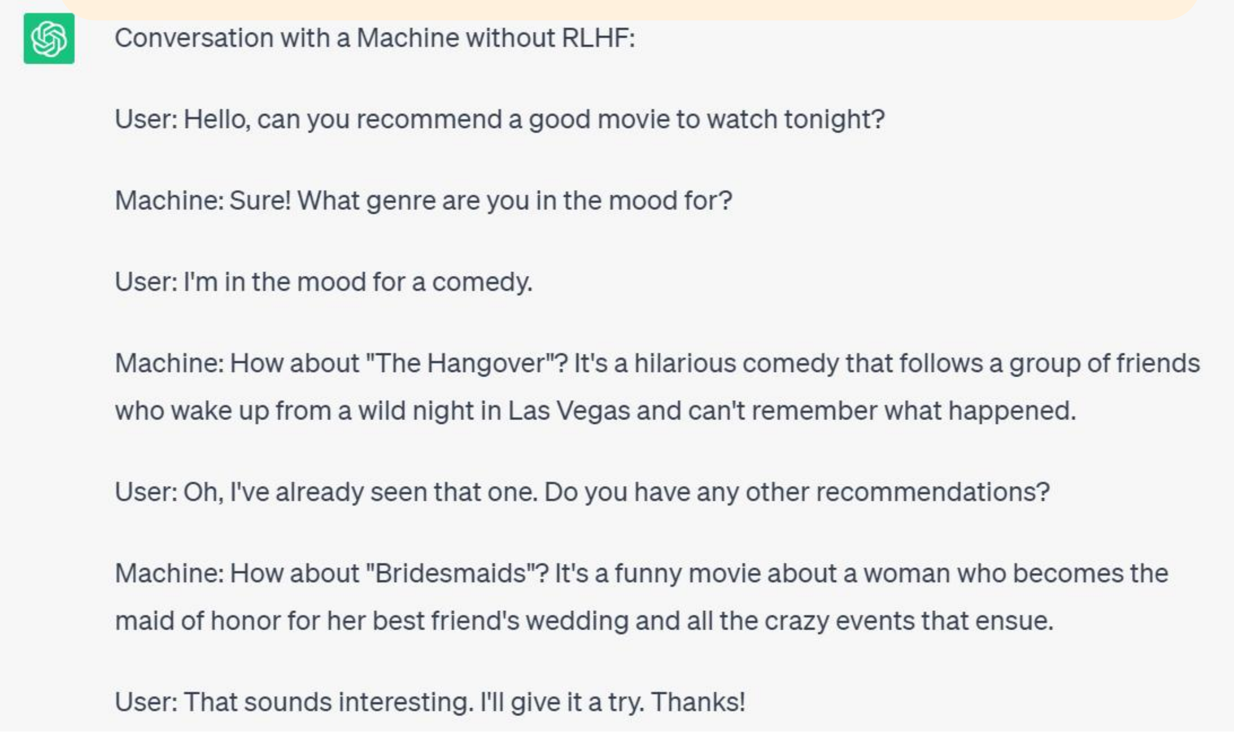
Machine: Certainly! If you're open to a different genre, I highly recommend "Inception." It's a mind-bending science fiction thriller that explores the concept of shared dreaming. The film is directed by Christopher Nolan and has garnered critical acclaim for its complex narrative and stunning visual effects.

User: Oh, I've heard about "Inception" before. It sounds intriguing. I think I'll give it a try. Thanks for the recommendation!

Human Feedback

- **Human Feedback** in conversations

- acknowledges the repetition and apologizes for it.
- provides more detailed information about the recommended movies



Conversation with a Machine without RLHF:

User: Hello, can you recommend a good movie to watch tonight?

Machine: Sure! What genre are you in the mood for?

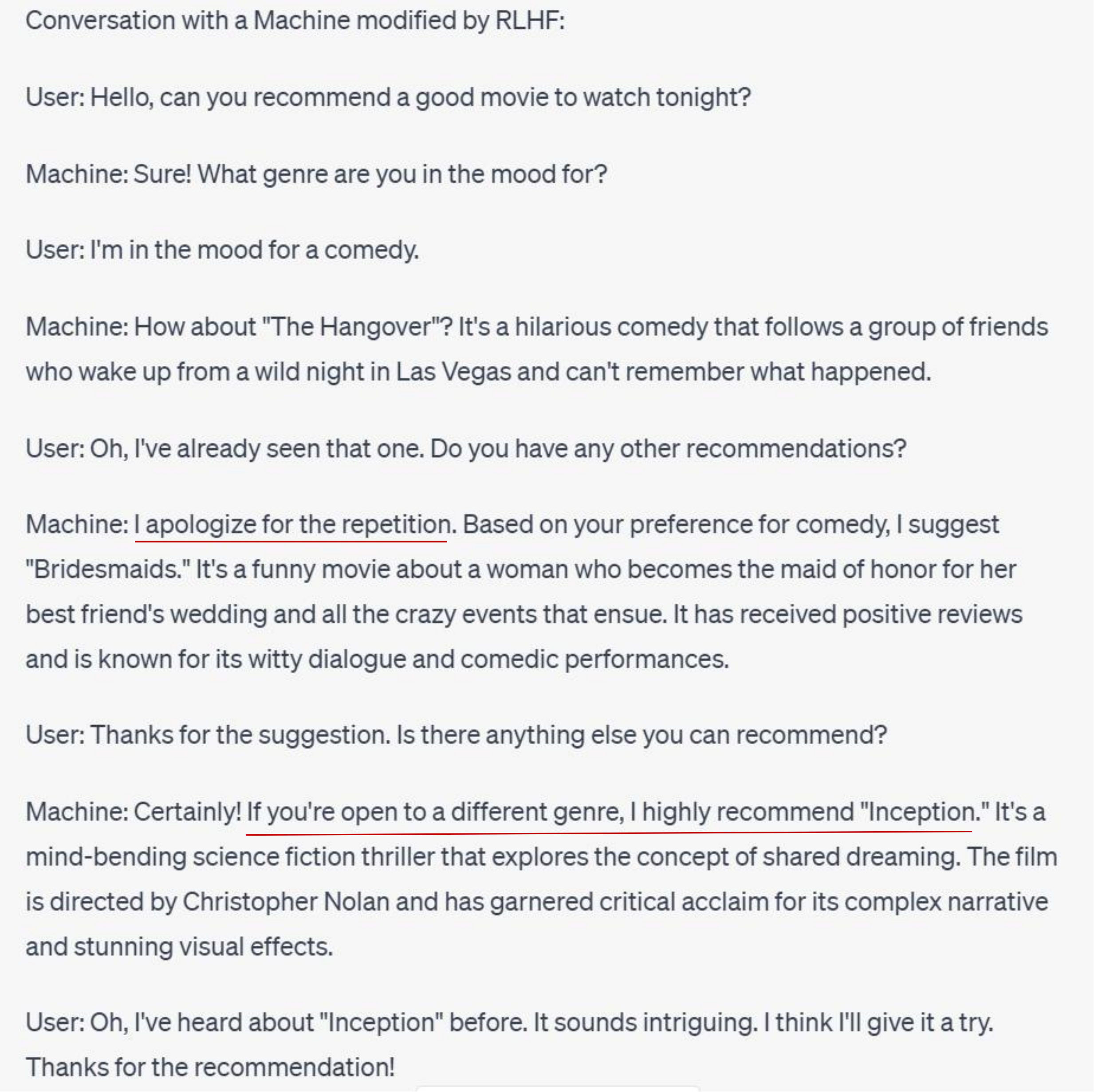
User: I'm in the mood for a comedy.

Machine: How about "The Hangover"? It's a hilarious comedy that follows a group of friends who wake up from a wild night in Las Vegas and can't remember what happened.

User: Oh, I've already seen that one. Do you have any other recommendations?

Machine: How about "Bridesmaids"? It's a funny movie about a woman who becomes the maid of honor for her best friend's wedding and all the crazy events that ensue.

User: That sounds interesting. I'll give it a try. Thanks!



Conversation with a Machine modified by RLHF:

User: Hello, can you recommend a good movie to watch tonight?

Machine: Sure! What genre are you in the mood for?

User: I'm in the mood for a comedy.

Machine: How about "The Hangover"? It's a hilarious comedy that follows a group of friends who wake up from a wild night in Las Vegas and can't remember what happened.

User: Oh, I've already seen that one. Do you have any other recommendations?

Machine: I apologize for the repetition. Based on your preference for comedy, I suggest "Bridesmaids." It's a funny movie about a woman who becomes the maid of honor for her best friend's wedding and all the crazy events that ensue. It has received positive reviews and is known for its witty dialogue and comedic performances.

User: Thanks for the suggestion. Is there anything else you can recommend?

Machine: Certainly! If you're open to a different genre, I highly recommend "Inception." It's a mind-bending science fiction thriller that explores the concept of shared dreaming. The film is directed by Christopher Nolan and has garnered critical acclaim for its complex narrative and stunning visual effects.

User: Oh, I've heard about "Inception" before. It sounds intriguing. I think I'll give it a try. Thanks for the recommendation!

Human Feedback

- **Human Feedback** in summerization

Prompt:

To pursue a Computer Sc.PhD or continue working? Especially if one has no real intention to work in academia even after grad school .

Vanilla LM:

I' m considering pursuing a PhD in Computer Science, but I'm worried about the future. Im currently employed full-time, but I'm worried about the future.

+

Human Annotation:

Software Engineer with a job I’m happy at (for now), deciding whether to pursue a PhD to improve quallfications and explore interests and a new challenge

=

RLHF Model:

Currently employed, considering pursuing PhD in Computer Science to avoid being stuck wih no residency visa ever again . Has anyone pursued a PhD purely for the sake of research, with no intention of joining the academic world?

Human Feedback

- **Human Feedback** in summerization

Prompt:

[r/dating_advice] First date ever, going to the beach. Would like some tips

Hey Reddit! I (20M) would like some tips, because I have my first ever date tomorrow (although I've had a gf for 3 years, but no actual dating happened), and we're going to the beach.

I met this girl, we have mutual friends, at a festival a few days ago. We didn't kiss, but we talked, held hands, danced a bit. I asked her to go on a date with me, which was super hard as it is the first time I've asked this to anybody. What I mean to say is, it's not like a standard *first* date because we already spent some time together.

I'm really nervous and excited. I'm going to pick her up tomorrow, we're cycling to the beach which will take 30 minutes, and then what? I'm a bit scared. Should I bring something (the weather, although no rain and sunny, is not super so no swimming), should we do something. I'd like all the tips I can get. Thanks!

6.7B supervised model
Going on a date with a girl I met a few days ago, going to the beach What should I bring, what should we do?

+

Human written reference
TL;DR First date after 3 years in a relation-ship, going to the beach, terrified.What to bring with me, what to do?

=

6.7B human feedback model
Going on my first ever date tomorrow, cycling to the beach. Would like some tips on what to do and bring. I'm a bit nervous and excited.Thanks!

Human Feedback

- **Human Feedback**
    - offer a sense of emotional connection
    - uncover underlying messages within the conversation
    - avoid confining oneself to superficial aspects of the matter

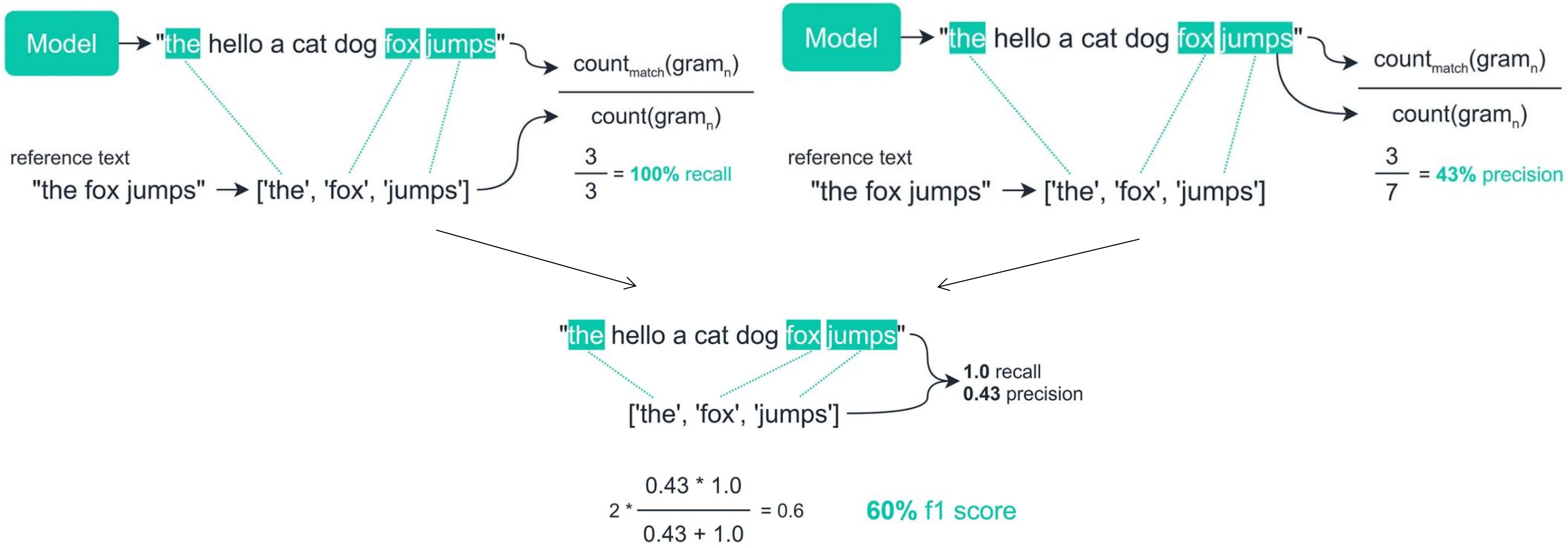
Human Feedback

- **Human Feedback**
    - offer a sense of emotional connection
    - uncover underlying messages within the conversation
    - avoid confining oneself to superficial aspects of the matter

Make machine produce sentences that sound natural to a human

Human Feedback

- Limitations of previous approach
  - evaluation metrics: BLEU, ROUGE - math equation


Model
"the hello a cat dog fox jumps"
$\frac{count_{match}(gram_n)}{count(gram_n)}$
reference text
"the fox jumps"
['the', 'fox', 'jumps']
$\frac{3}{3}$ = 100% recall
Model
"the hello a cat dog fox jumps"
$\frac{count_{match}(gram_n)}{count(gram_n)}$
reference text
"the fox jumps"
['the', 'fox', 'jumps']
$\frac{3}{7}$ = 43% precision
"the hello a cat dog fox jumps"
['the', 'fox', 'jumps']
1.0 recall
0.43 precision
$2 * \frac{0.43 * 1.0}{0.43 + 1.0} = 0.6$
60% f1 score

Human Feedback

- Limitations of previous approach
    - evaluation metric: BLEU, ROUGE - math equation

Human Feedback

- Limitations of previous approach
  - evaluation metric: BLEU, ROUGE - math equation
  - abstract objective that requires complicated evaluation

Human Feedback

- Limitations of previous approach
  - evaluation metric: BLEU, ROUGE - math equation
  - abstract objective that requires complicated evaluation
    - how to define funny?
    - how to define ethical?
    - how to define human-like?

open-ended question

- hard to assign it to a math equation

Human Feedback

- Limitations of previous approach
  - evaluation metric: BLEU, ROUGE - math equation
  - abstract objective that requires complicated evaluation
    - how to define funny?
    - how to define ethical?
    - how to define human-like?

open-ended question

- hard to assign it to a math equation

- **Solution: reinforcement learning**
  - agent learns to solve complex problems (e.g., win a game)

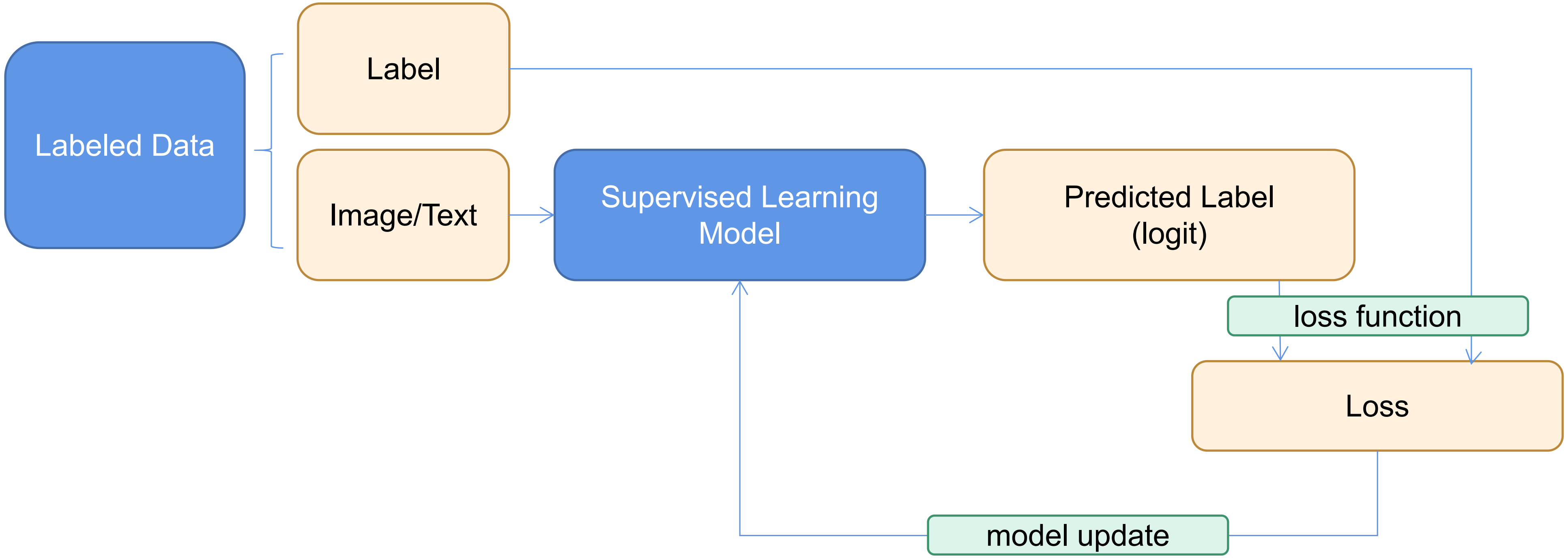
Motivation

Reinforcement Learning from Human Feedback

基于人工反馈的强化学习

Reinforcement Learning

- Recap for supervised learning

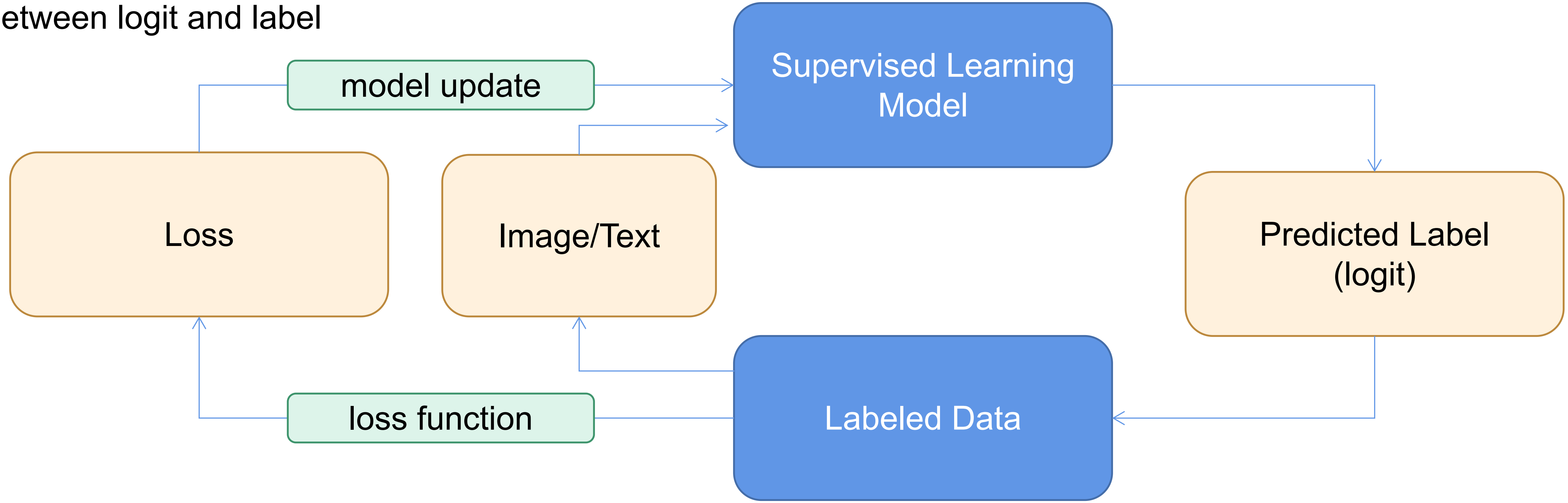


objective:
- minimise the loss between logit and label
- maximise the likelihood between logit and label

Reinforcement Learning

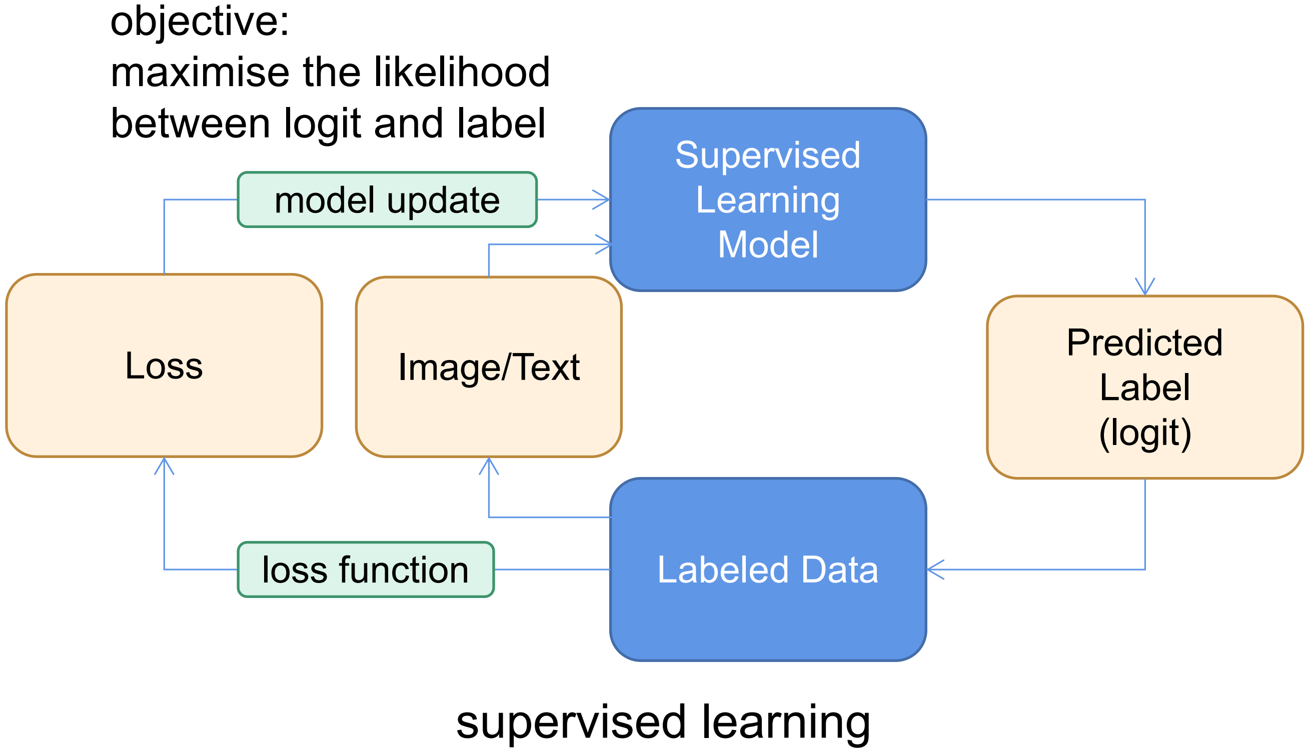
- Recap for supervised learning: let's illustrate it in another way

objective:
maximise the likelihood
between logit and label

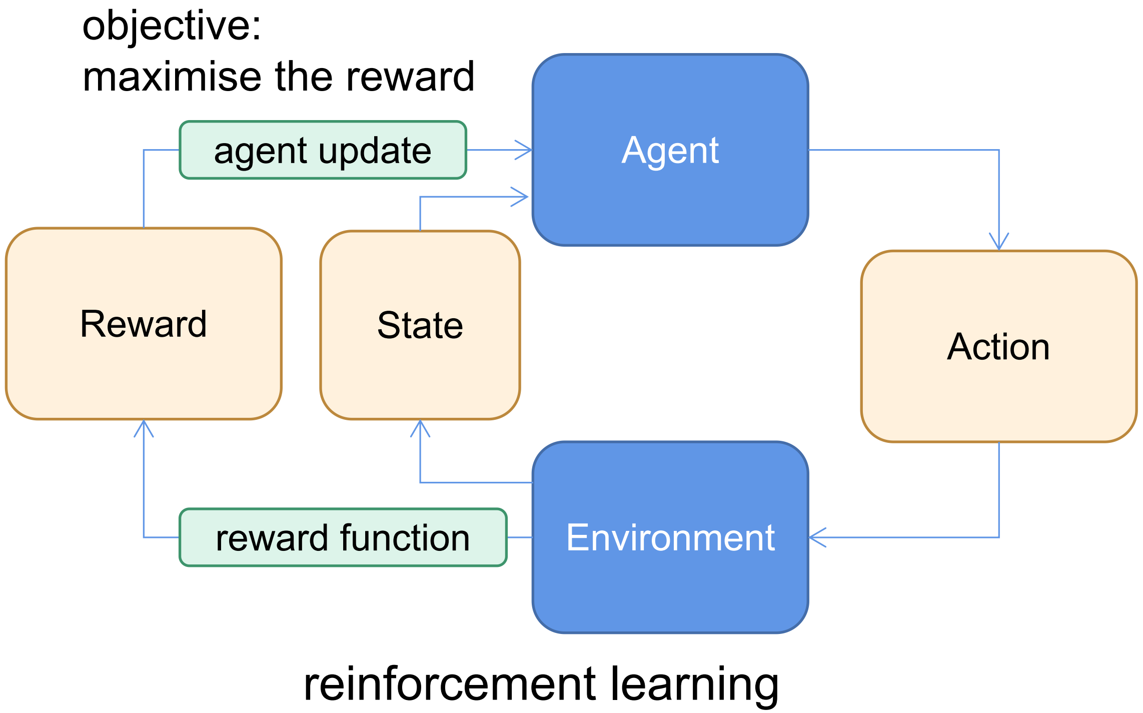


Reinforcement Learning

- Reinforcement learning is very similar to supervised learning.



supervised learning



reinforcement learning

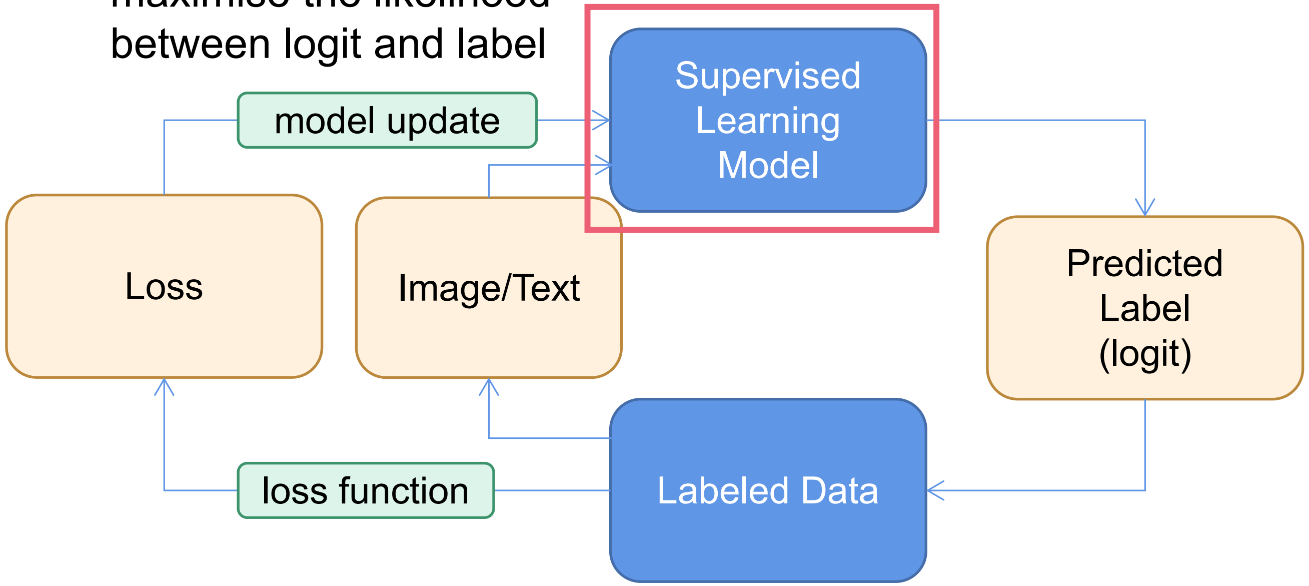
Contents

- **强化学习和PPO算法**
  - Motivation of RLHF
    - Human feedback
    - Reinforcement learning
  - **Introduction of reinforcement learning**
    - Agent
    - Reward
    - RL algorithm
      - policy gradient (PG)
      - proximal policy gradient (PPO)

Reinforcement Learning

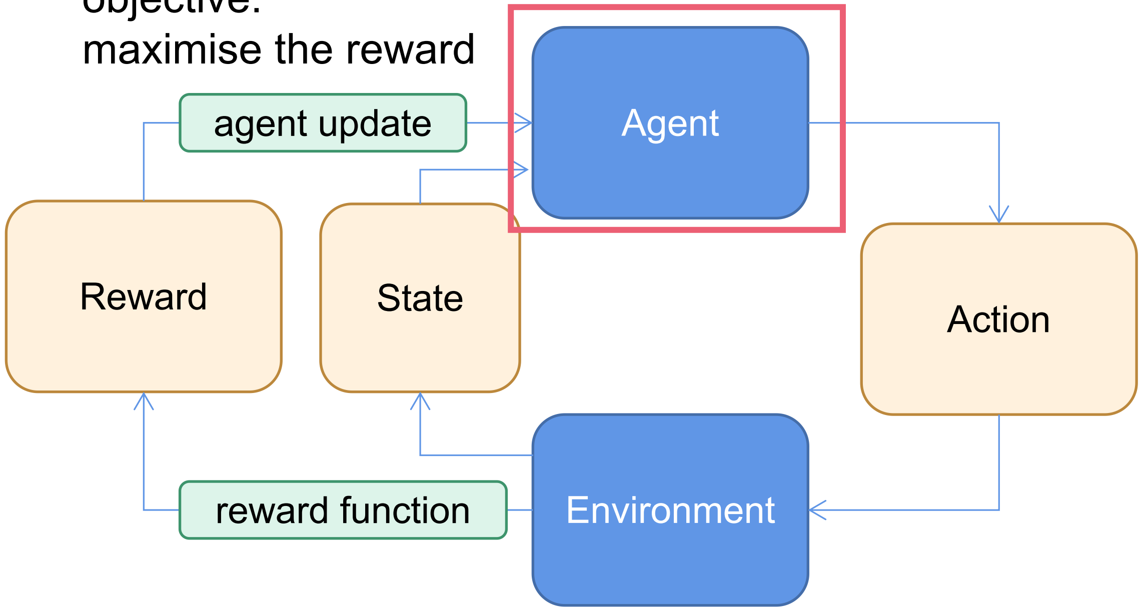
- Let’s look into the process step by step.

objective:
maximise the likelihood
between logit and label



supervised learning

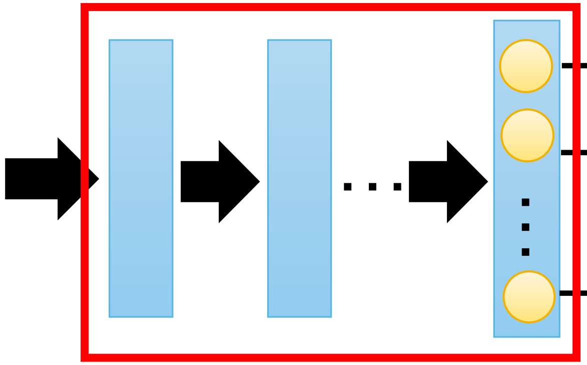
objective:
maximise the reward



reinforcement learning

Reinforcement Learning - Agent

supervised learning



dog 0.7

car 0.2

bicycle 0.1

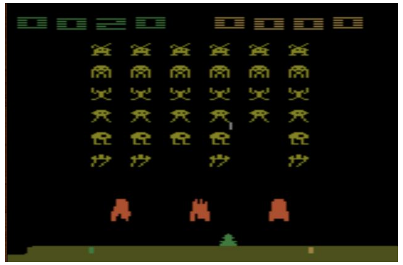
Score of a label

image

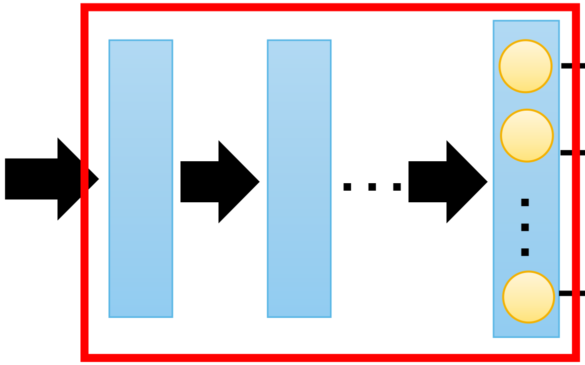
forward function

label

policy-based deep reinforcement learning



pixels



left 0.7

right 0.2

fire 0.1

Score of an action

state

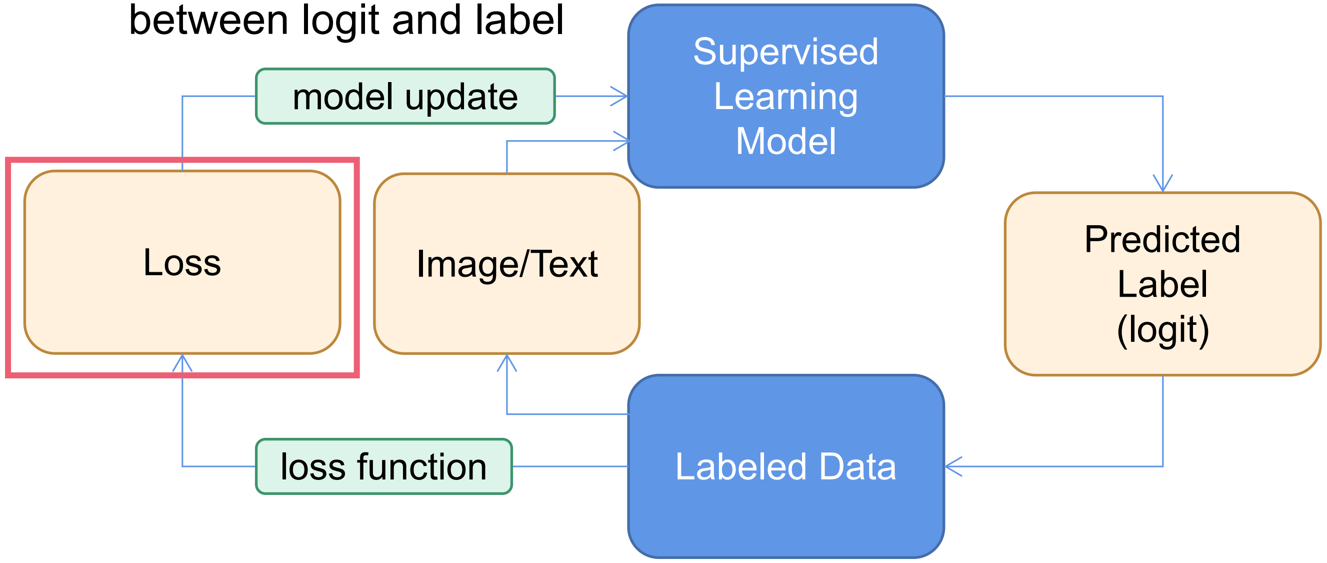
policy

action

Reinforcement Learning

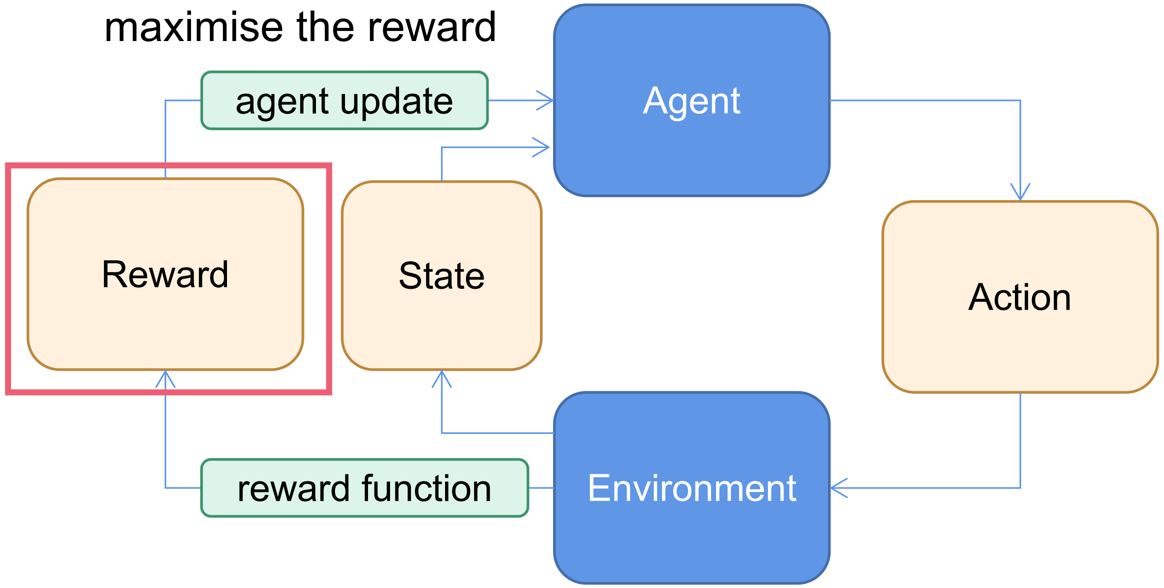
- Let’s look into the process step by step.

objective:
maximise the likelihood
between logit and label



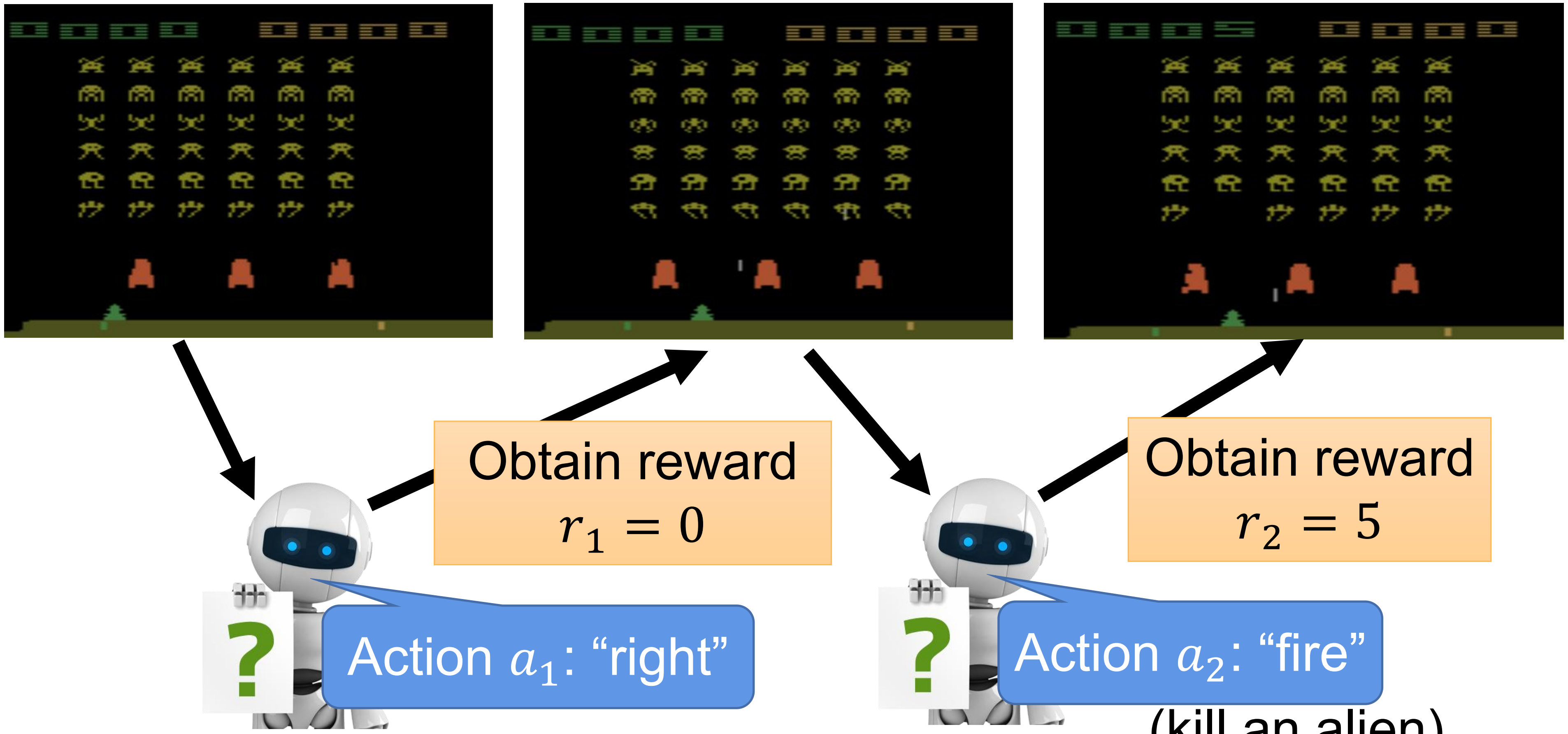
supervised learning

objective:
maximise the reward



reinforcement learning

Reinforcement Learning - Reward



Start with state s_1

state s_2

state s_3

Obtain reward $r_1 = 0$

Action a_1: “right”

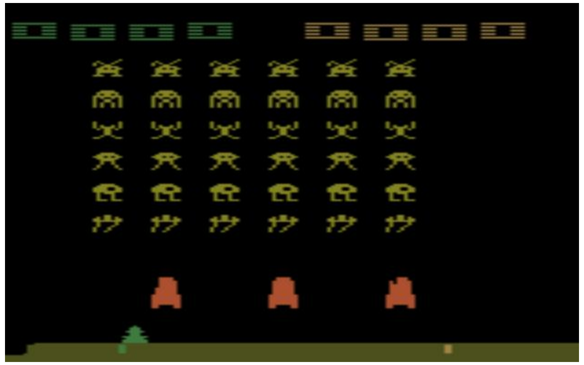
Obtain reward $r_2 = 5$

Action a_2: “fire”

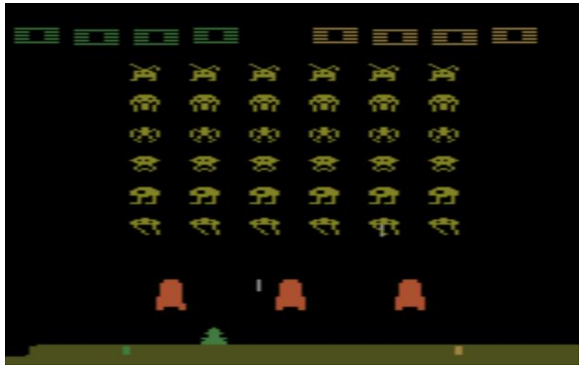
(kill an alien)

Reinforcement Learning - Reward

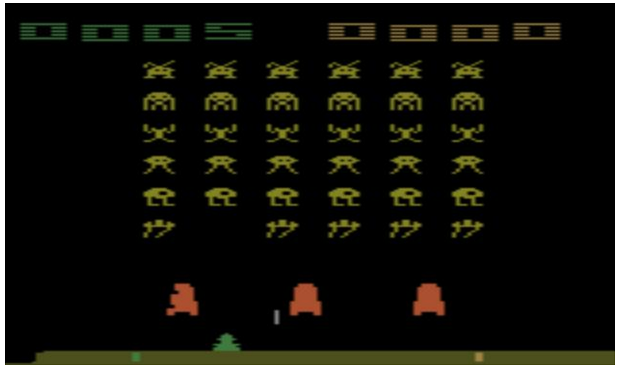
Start with state s_1



state s_2



state s_3

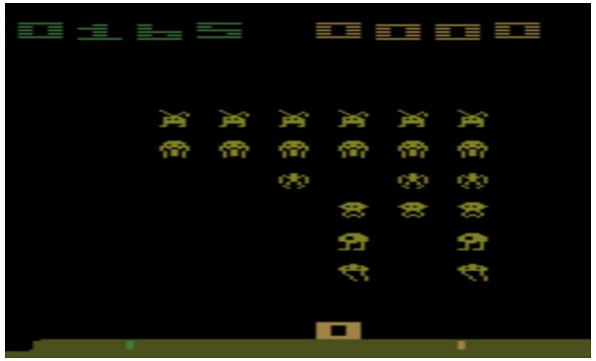


- Start with state s_1
- Machine decides to take a_1
- Machine obtains reward r_1
- Machine sees state s_2
- Machine decides to take a_2
- Machine obtains reward r_2
- Machine sees state s_3
- ……
- Machine decides to take a_T
- Machine obtains reward r_T

After many turns

Obtain reward r_T

Action a_T



Game Over
(spaceship destroyed)

This is an ***episode***.

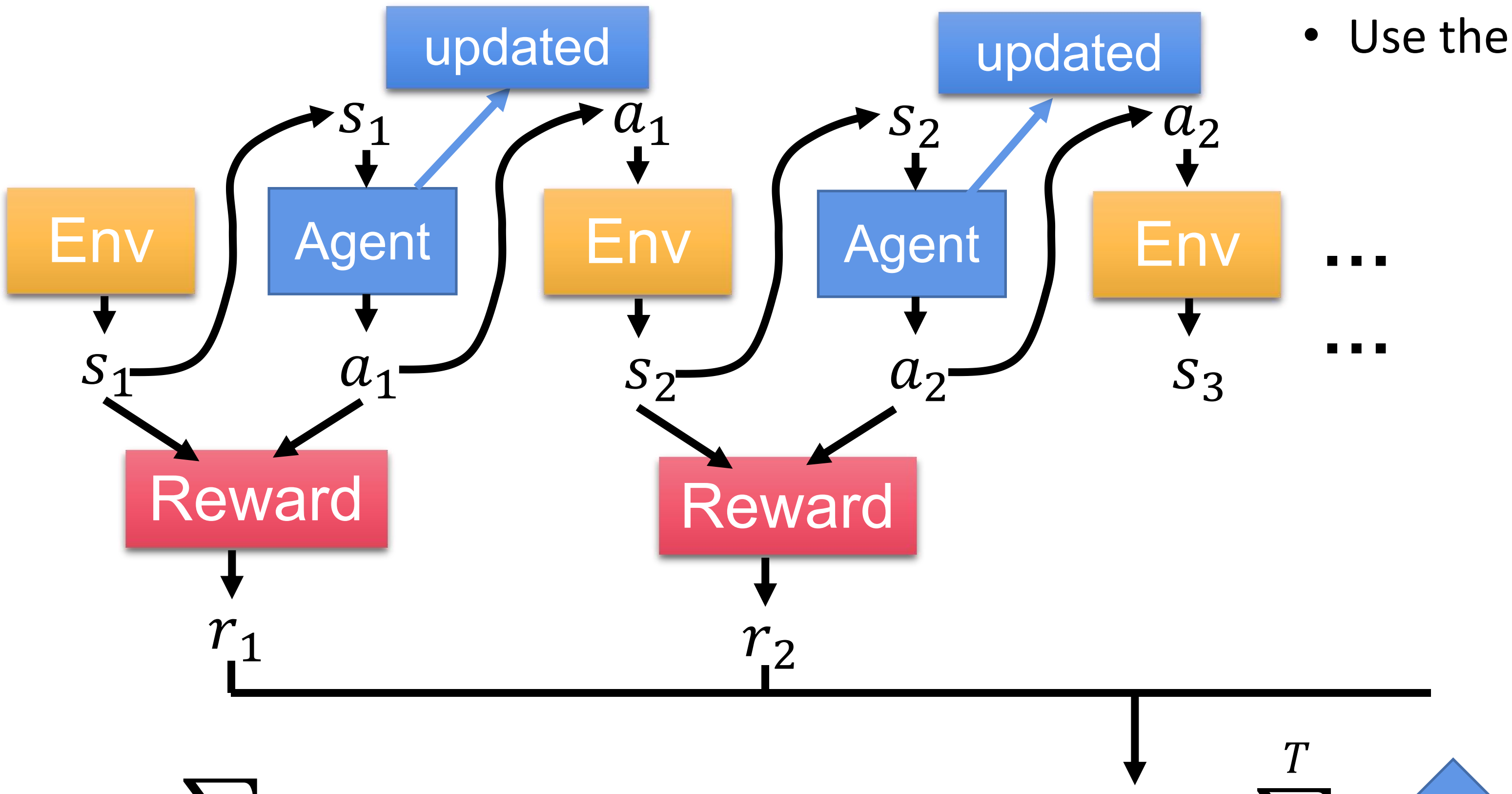
Total reward:

$$R = \sum_{t=1}^{T} r_t$$

We want the total reward be maximized.

Reinforcement Learning - Reward

- agent, reward and environment



- Given an actor πθ(s) with network parameter θ
- Use the actor πθ(s) to play the videogame

Total reward: $R(\tau)$

Even with the same actor, $R(\tau)$ is different each time

Randomness in the actor and the game

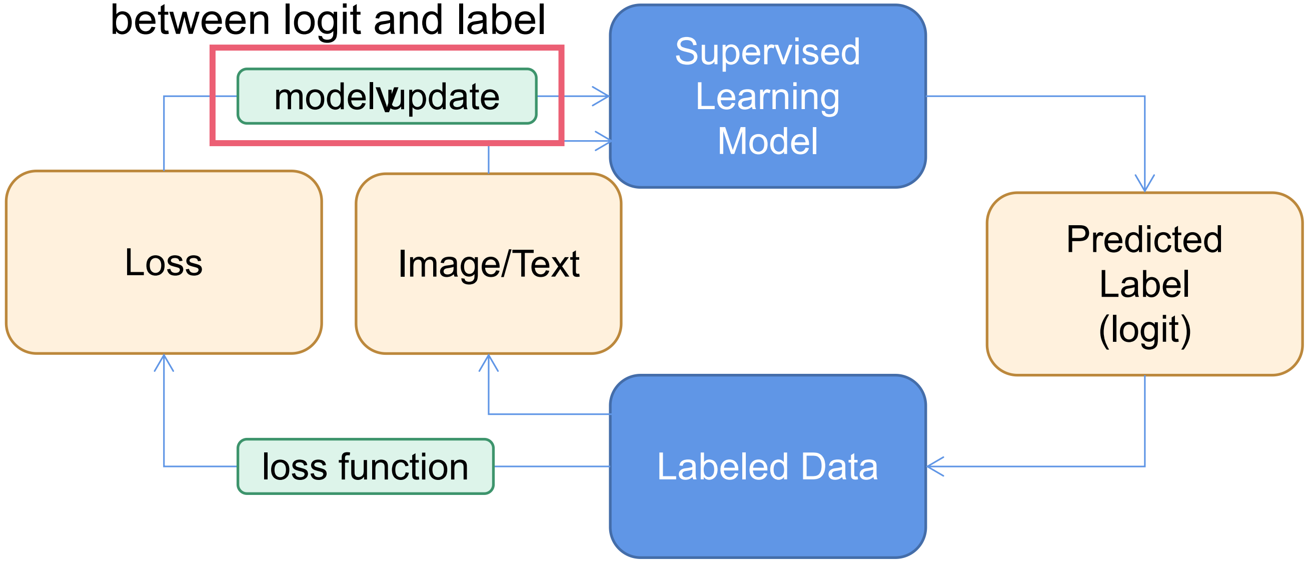
We define $\bar{R}_\theta$ as the *expected value* of $R(\tau)$

$$\bar{R}_\theta = \sum_{\tau} R(\tau) p_\theta(\tau) = E_{\tau \sim p_\theta(\tau)}[R(\tau)] \qquad R(\tau) = \sum_{t=1}^{T} r_t$$

Reinforcement Learning

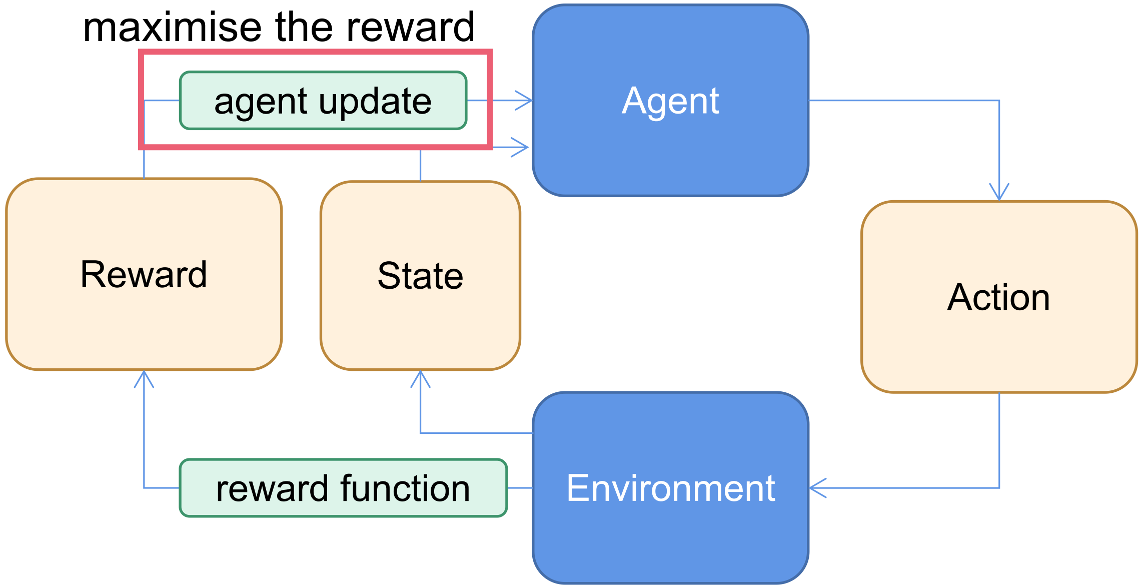
- Let’s look into the process step by step.

objective:
maximise the likelihood
between logit and label



supervised learning

objective:
maximise the reward



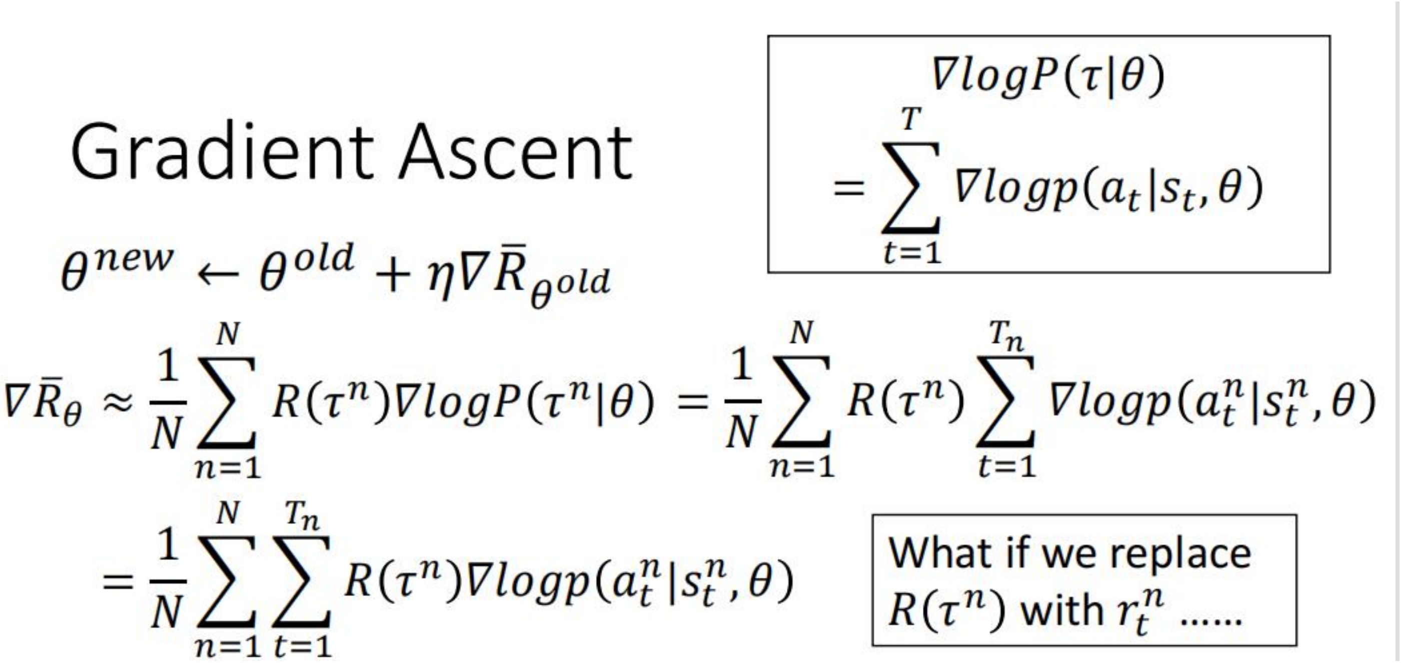
reinforcement learning

Reinforcement Learning - optimisation

supervised learning

gradient descend

policy-based
deep reinforcement learning



Gradient Ascent

$$\theta^{new} \leftarrow \theta^{old} + \eta \nabla \bar{R}_{\theta^{old}}$$

$$\nabla \log P(\tau|\theta) = \sum_{t=1}^{T} \nabla \log p(a_t|s_t,\theta)$$

$$\nabla \bar{R}_\theta \approx \frac{1}{N}\sum_{n=1}^{N} R(\tau^n) \nabla \log P(\tau^n|\theta) = \frac{1}{N}\sum_{n=1}^{N} R(\tau^n) \sum_{t=1}^{T_n} \nabla \log p(a_t^n|s_t^n,\theta)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n) \nabla \log p(a_t^n|s_t^n,\theta)$$

What if we replace $R(\tau^n)$ with r_t^n ......

Reinforcement Learning - policy gradient

Gradient Ascent

$$\nabla log P(\tau|\theta) = \sum_{t=1}^{T} \nabla log p(a_t|s_t,\theta)$$

$$\theta^{new} \leftarrow \theta^{old} + \eta \nabla \bar{R}_{\theta^{old}}$$

$$\nabla \bar{R}_{\theta} \approx \frac{1}{N}\sum_{n=1}^{N} R(\tau^n) \nabla log P(\tau^n|\theta) = \frac{1}{N}\sum_{n=1}^{N} R(\tau^n) \sum_{t=1}^{T_n} \nabla log p(a_t^n|s_t^n,\theta)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} R(\tau^n) \nabla log p(a_t^n|s_t^n,\theta)$$

What if we replace $R(\tau^n)$ with r_t^n ……

If in τ^n machine takes a_t^n when seeing s_t^n in

$R(\tau^n)$ is positive → Tuning θ to increase $p(a_t^n|s_t^n)$

$R(\tau^n)$ is negative → Tuning θ to decrease $p(a_t^n|s_t^n)$

It is very important to consider the cumulative reward $R(\tau^n)$ of the whole trajectory τ^n instead of immediate reward r_t^n

On-policy v.s. Off-policy

- On-policy: The agent learned and the agent interacting with the environment is the same.
- Off-policy: The agent learned and the agent interacting with the environment is different.

$$\nabla \bar{R}_\theta = E_{\underline{\tau \sim p_\theta(\tau)}}[R(\tau)\nabla log p_\theta(\tau)]$$

- Use π_θ to collect data. When θ is updated, we have to sample training data again.
- Goal: Using the sample from $\pi_{\theta'}$ to train θ. θ' is fixed, so we can re-use the sample data.

PPO and PPO2 algorithm

- PPO Algorithm

$$J_{PPO}^{\theta^k}(\theta) = J^{\theta^k}(\theta) - \beta KL(\theta, \theta^k)$$

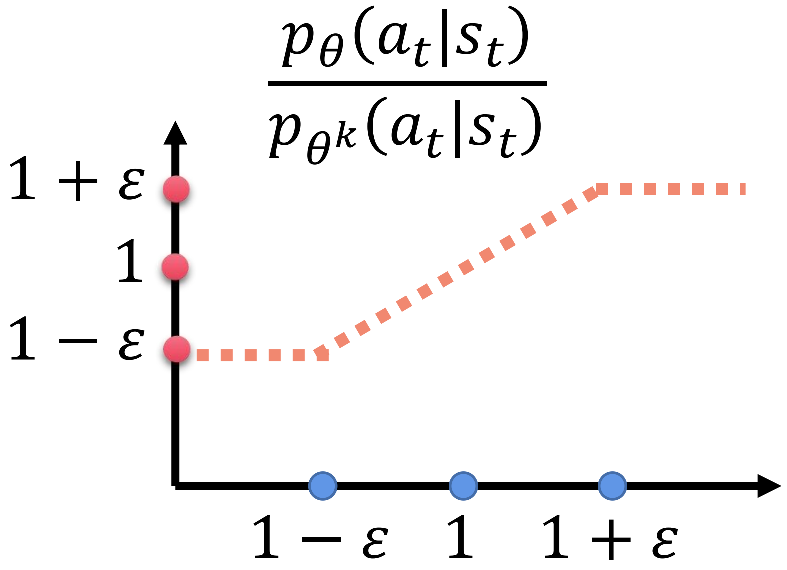
$$J^{\theta^k}(\theta) \approx \sum_{(s_t, a_t)} \frac{p_\theta(a_t|s_t)}{p_{\theta^k}(a_t|s_t)} A^{\theta^k}(s_t, a_t)$$

- PPO2 Algorithm

$$J_{PPO2}^{\theta^k}(\theta)$$

$$\approx \sum_{(s_t, a_t)} min\left(\frac{p_\theta(a_t|s_t)}{p_{\theta^k}(a_t|s_t)} A^{\theta^k}(s_t, a_t), clip\left(\frac{p_\theta(a_t|s_t)}{p_{\theta^k}(a_t|s_t)}, 1-\varepsilon, 1+\varepsilon\right) A^{\theta^k}(s_t, a_t)\right)$$

$$\approx \sum_{(s_t, a_t)} min\left(clip\left(\frac{p_\theta(a_t|s_t)}{p_{\theta^k}(a_t|s_t)}, 1-\varepsilon, 1+\varepsilon\right) A^{\theta^k}(s_t, a_t)\right)$$

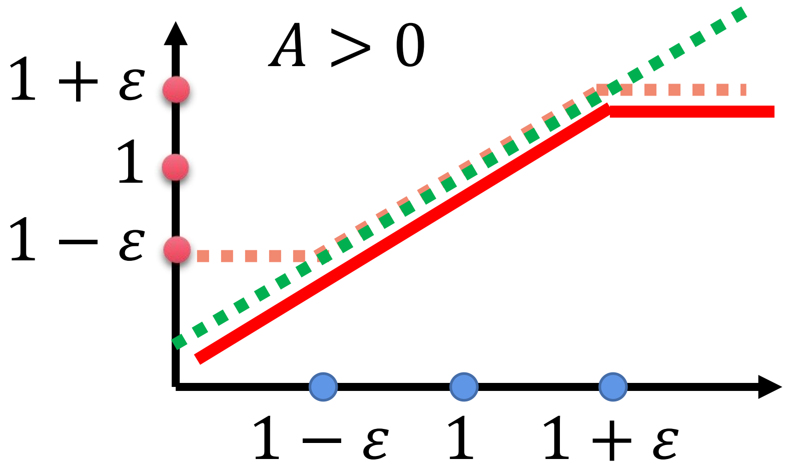


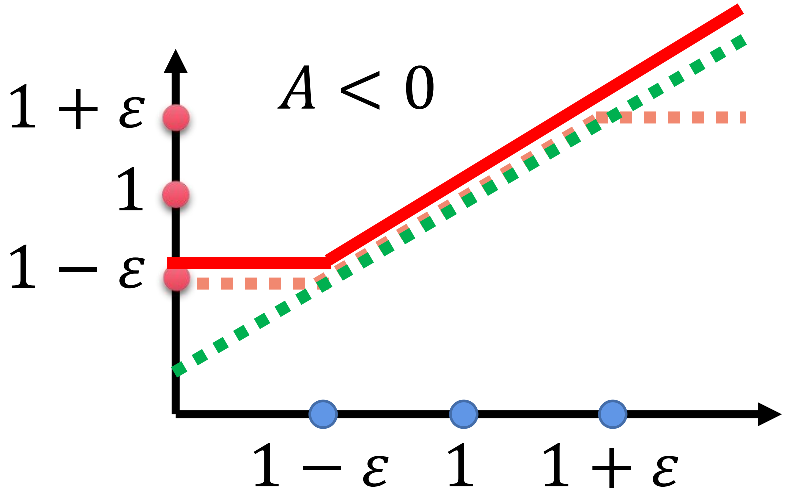
PPO2 algorithm

- PPO2 Algorithm

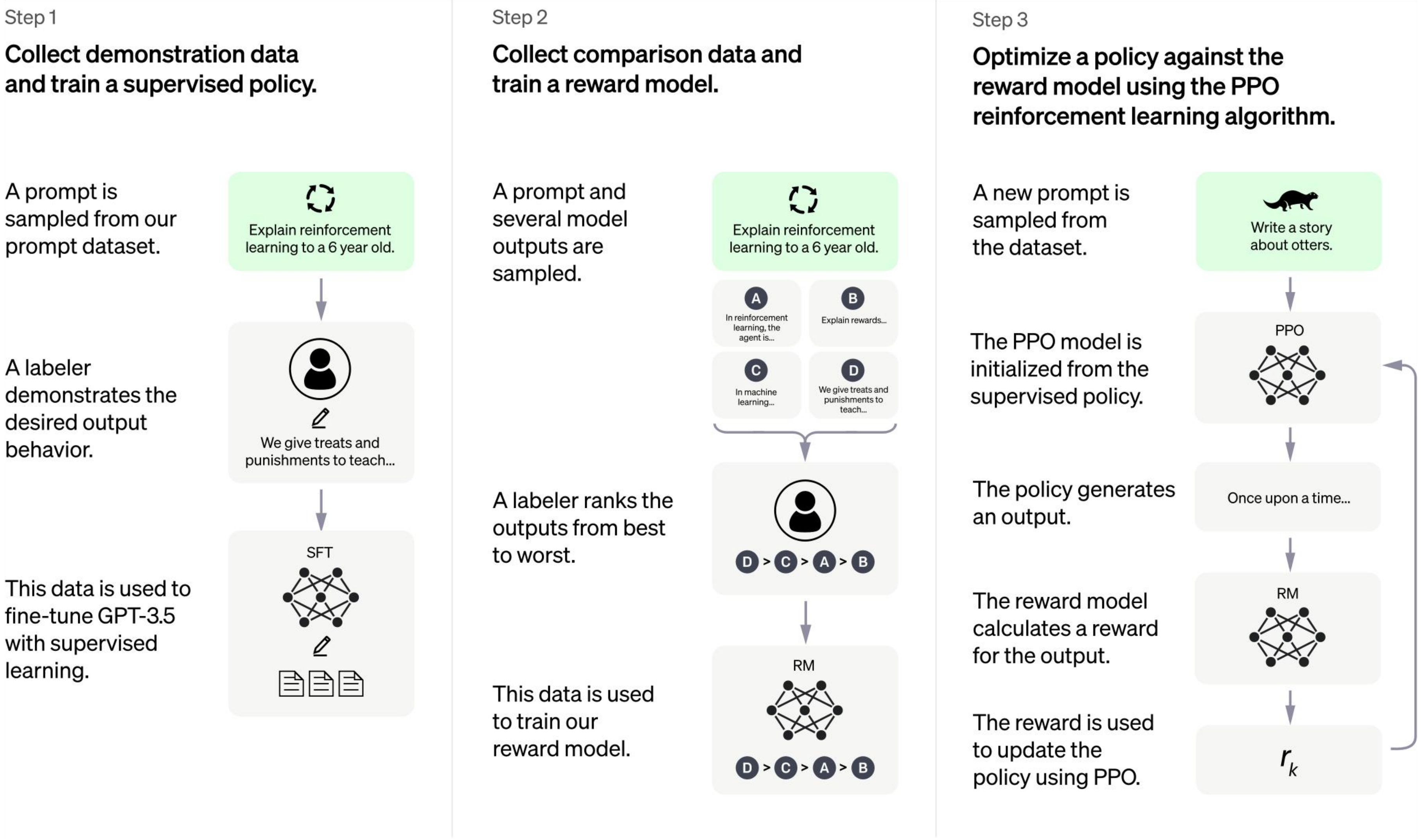
$$J_{PPO2}^{\theta^k}(\theta) \approx \sum_{(s_t,a_t)} min\left(\frac{p_\theta(a_t|s_t)}{p_{\theta^k}(a_t|s_t)} A^{\theta^k}(s_t,a_t), clip\left(\frac{p_\theta(a_t|s_t)}{p_{\theta^k}(a_t|s_t)}, 1-\varepsilon, 1+\varepsilon\right) A^{\theta^k}(s_t,a_t)\right)$$

$$\approx \sum_{(s_t,a_t)} min\left(clip\left(\frac{p_\theta(a_t|s_t)}{p_{\theta^k}(a_t|s_t)}, 1-\varepsilon, 1+\varepsilon\right) A^{\theta^k}(s_t,a_t)\right)$$





RLHF in InstructGPT/ChatGPT



Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

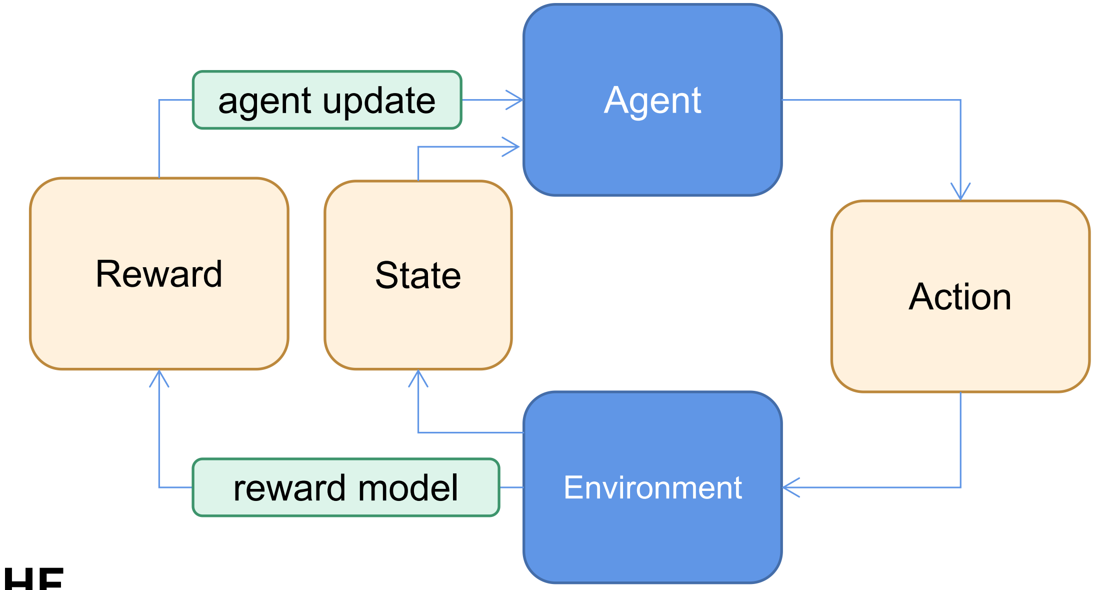
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM
r_k

RLHF in InstructGPT/ChatGPT

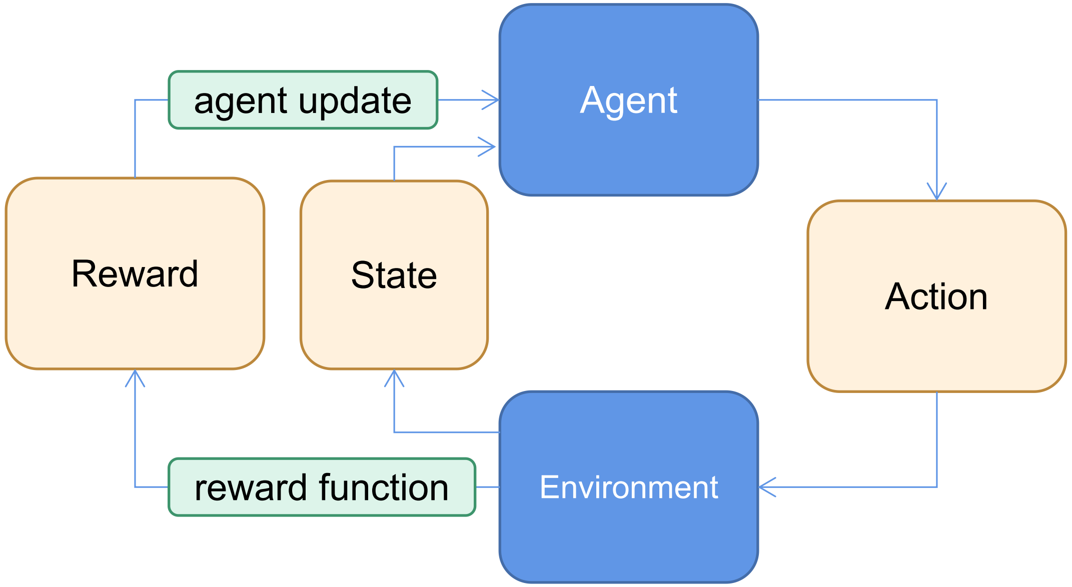
objective: make machine talk like a human
relative to **reward function**, maximise reward
reward function --> reward model



RLHF

obtrain rankings from human(training data)
- train reward model
- input: action （answer） ， output：reward （ranking)

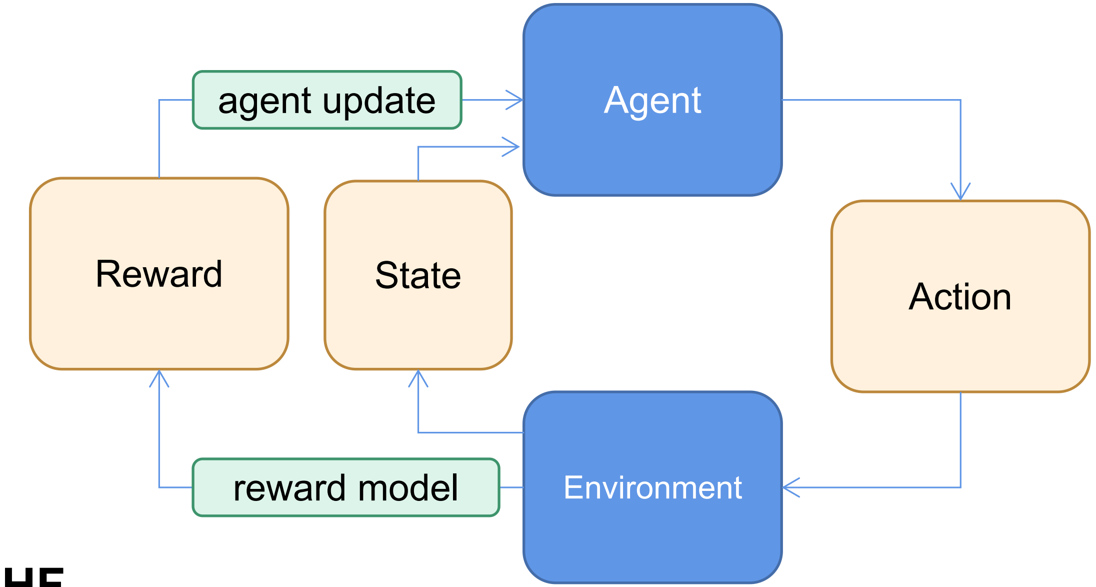
update agent to solve complicated problems
objective: win the game
relative to **reward function**, maximise reward



RL

RLHF in InstructGPT/ChatGPT

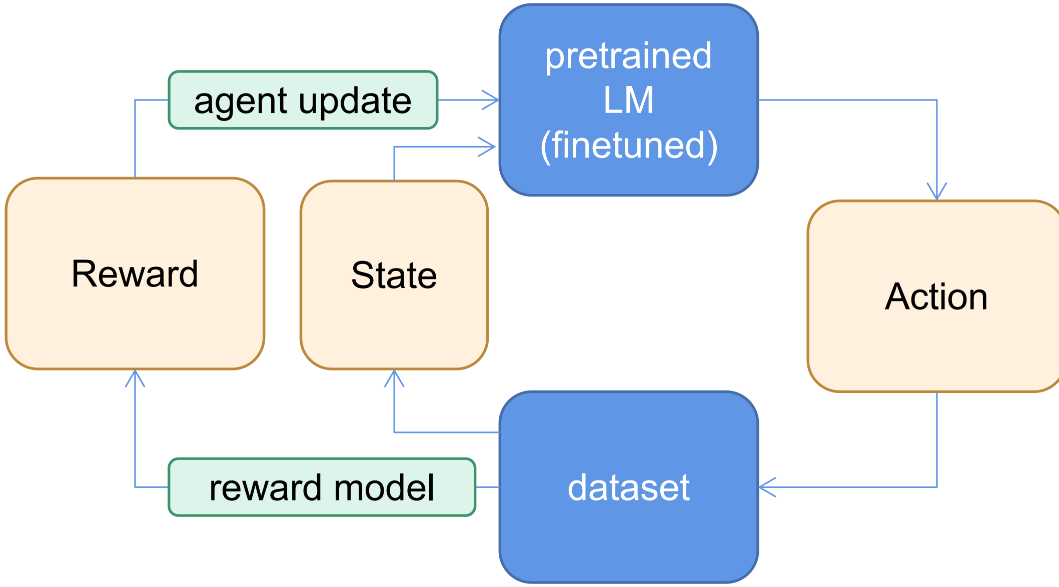
objective: make machine talk like a human
relative to **reward function**, maximise reward
reward function --> reward model



agent update
Agent
Reward
State
Action
reward model
Environment

RLHF

obtrain rankings from human(training data)
- train reward model
- input: action（answer）， output：reward （ranking)



agent update
pretrained LM (finetuned)
Reward
State
Action
reward model
dataset

RLHF in ChatGPT

学习活动

- **开发者群英会**
  - https://gitee.com/mindspore/community/issues/I6Q9H4
- **第六届开源创新大赛：**
  - MindSpore开源项目贡献赛

    https://www.gitlink.org.cn/competitions/index/track1_2023_mindspore
  - MindSpore代码评注赛

    https://www.gitlink.org.cn/competitions/index/track4_2023_mindspore
- **昇腾AI创新大赛：** https://www.hiascend.com/developer/AAIC2023