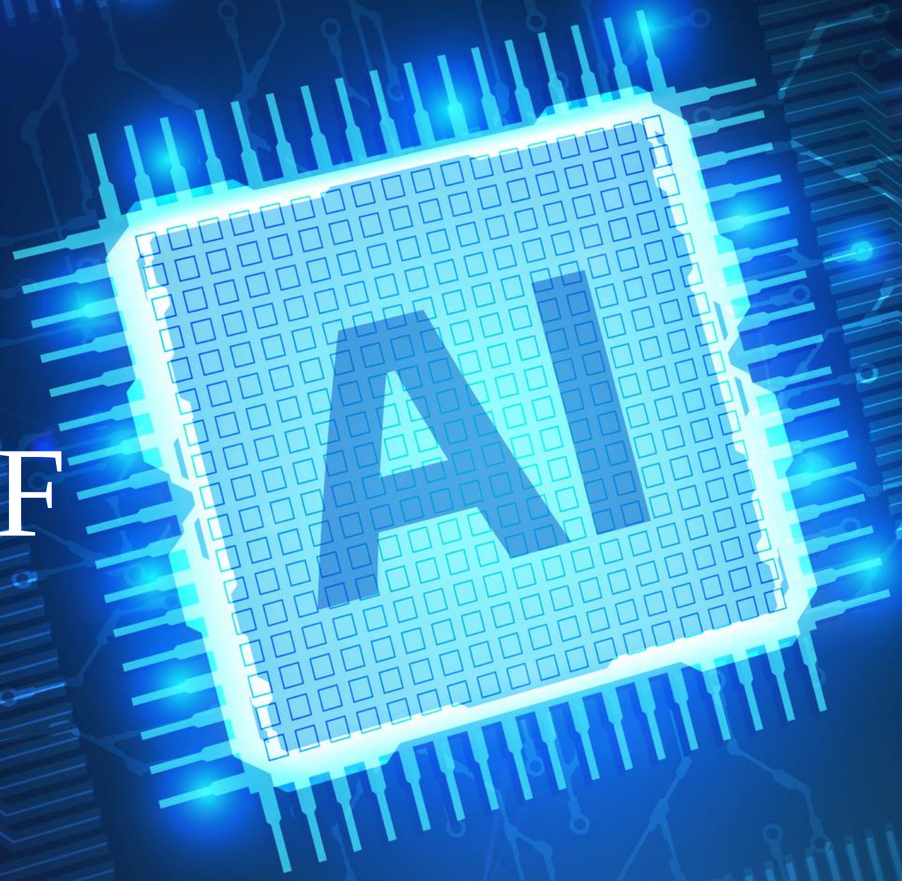
Path planning problem on virtual map based on RLHF



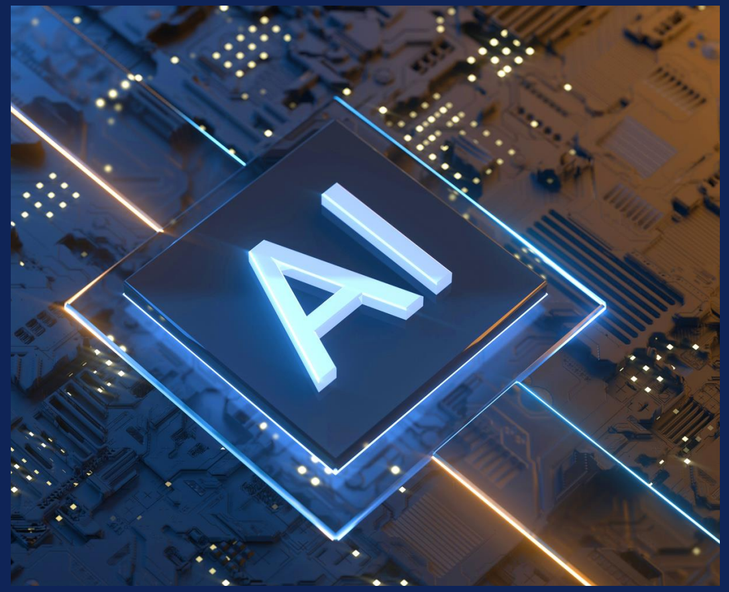
121090386 卢诣　　120090367 叶峰源

223040053 王一丹　　121040075 谢金好

Contents

1 Background

Reinforcement learning is a machine learning paradigm where an agent learns an adaptive policy by taking actions in the environment. The agent receives reward signals with the goal of maximizing rewards.



User-Centric Approach

Observe the path generated by the model

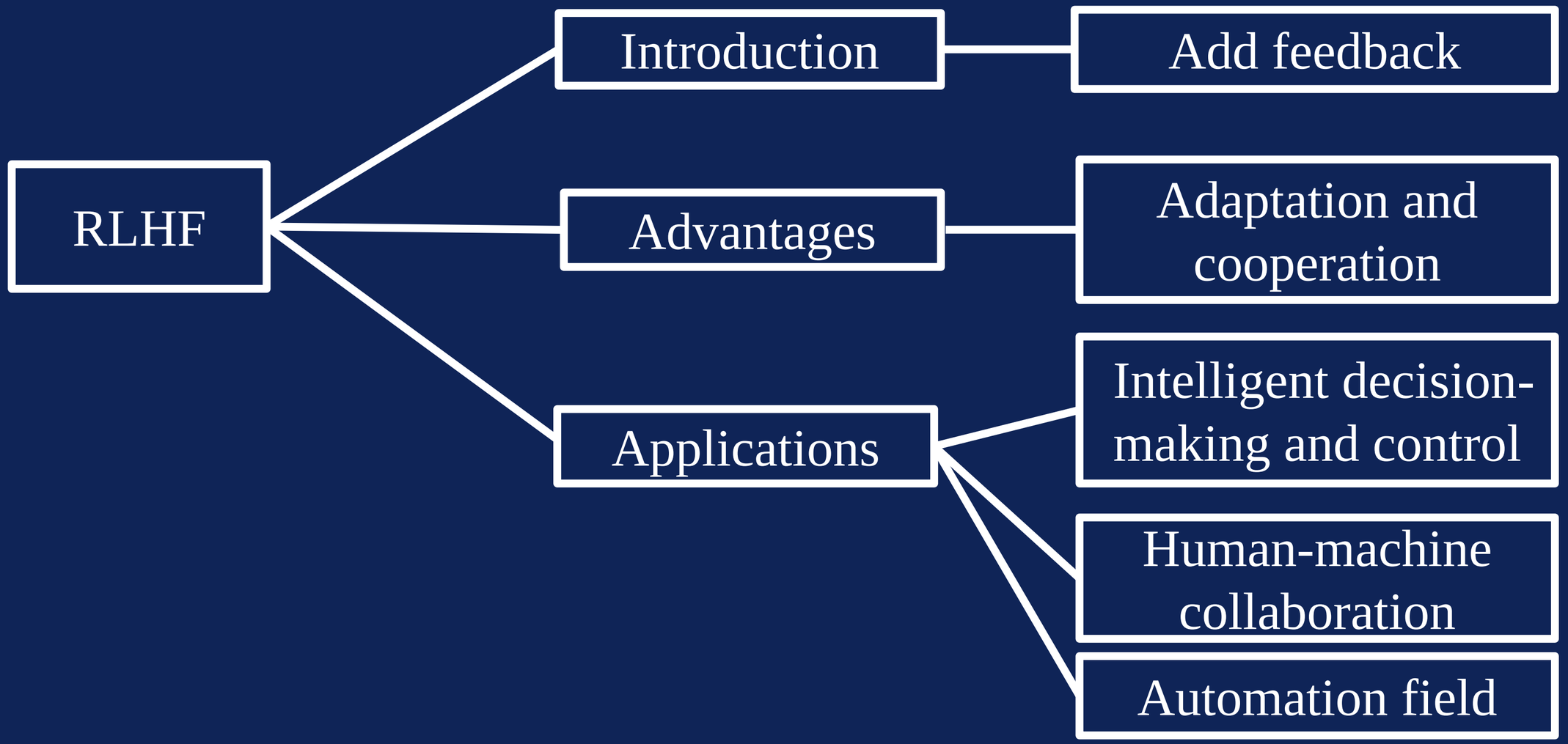
RLHF

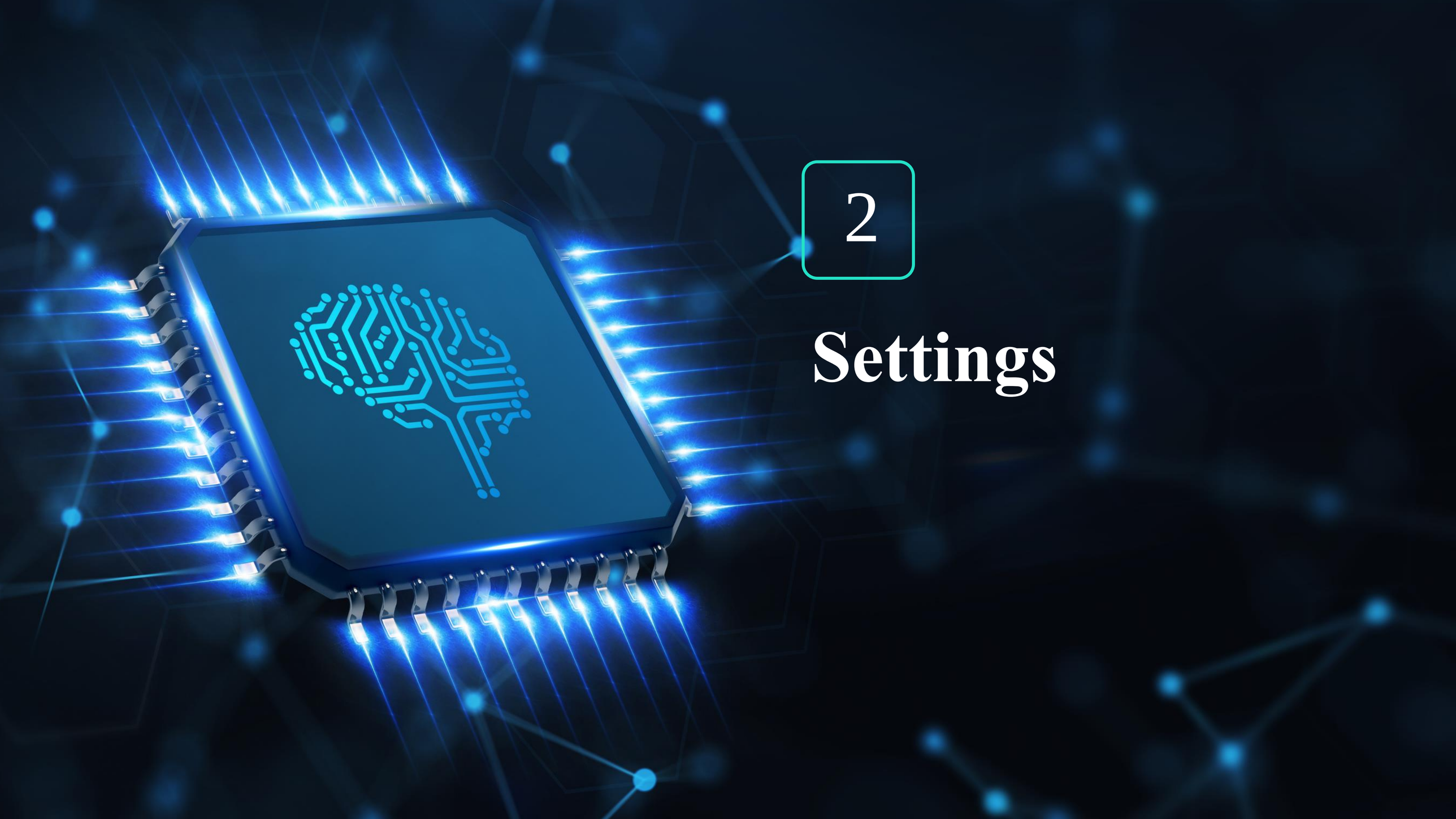
Methodology

Use reinforcement learning with human feedback to help the agent get out of the current predicament

Introduce the best path for model to enhance the results according to user needs

RLHF（Reinforcement Learning with Human Feedback）





2

Settings

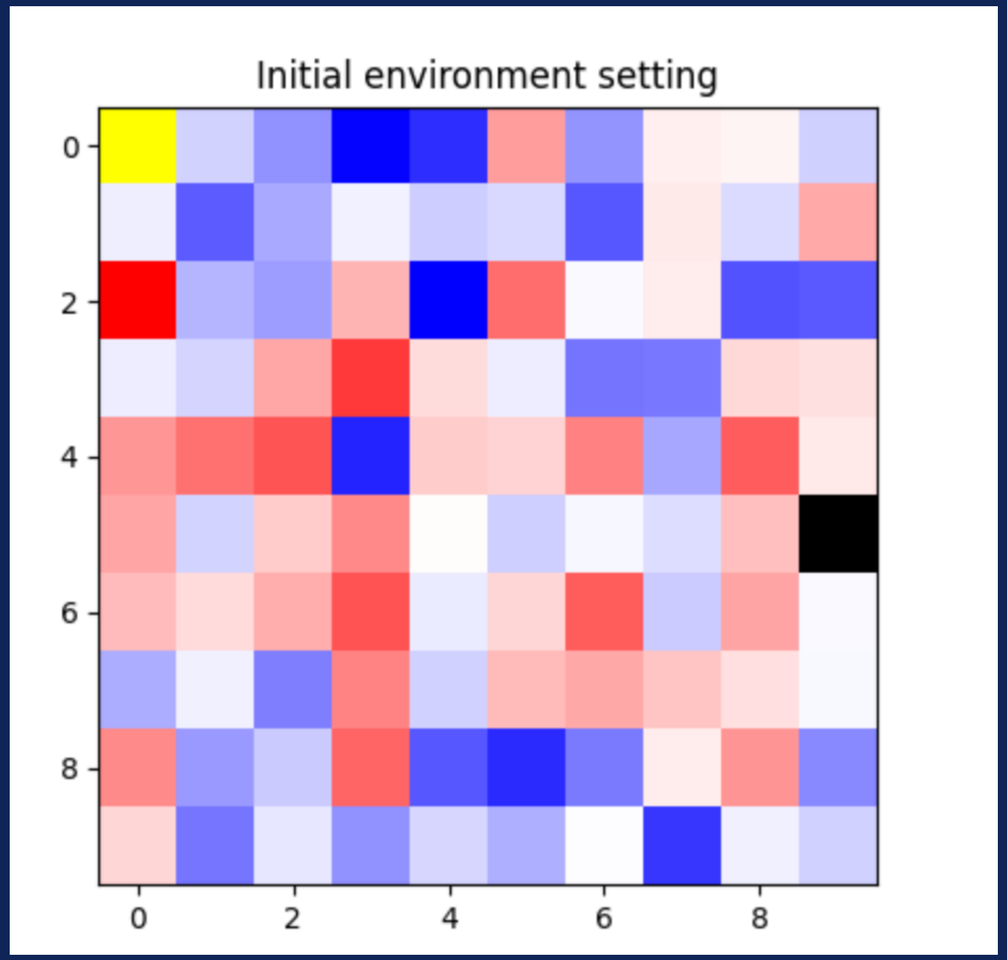
Map Settings

Yellow: start point

Blue: rewards

Red: obstacles

Black: end point



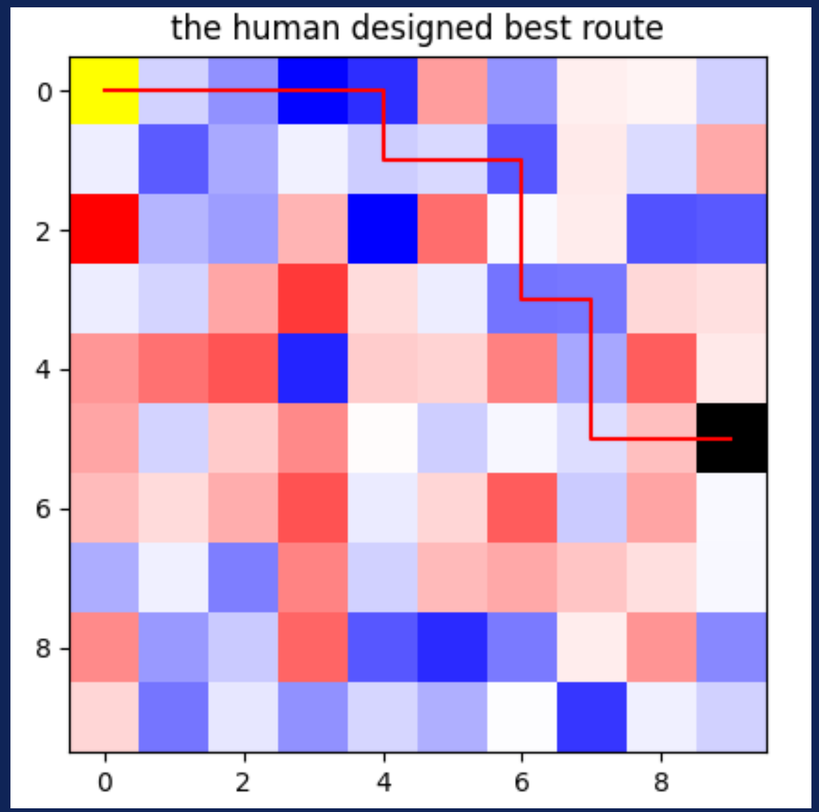
Optimal route and Reward computation

Reward function:

$$R_{\text{total}} = \sum_{t=0}^{T} \gamma^t \cdot r_t$$

Where:

- R_{total} is the total accumulated reward over the entire trajectory.
- γ is the discount factor.
- t represents the time step along the trajectory.
- T is the length of the trajectory.
- r_t is the reward obtained at time step t.



Optimal reward:5111.80

3 Method

Q-Learning

Algorithm 3: Q-learning with ε-greedy exploration

Input: ϵ, α

For all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $Q'(s, a) \leftarrow 0$, $Q(s, a) \leftarrow \infty$

while $\|Q - Q'\|_\infty > \epsilon$ **do**

$Q \leftarrow Q'$

Sample a trajectory τ from the policy

$$\pi(a \mid s) = \begin{cases} \arg\max_{a \in A} Q(s, a) & \text{with probability } 1 - \varepsilon \\ \text{random action} & \text{with probability } \varepsilon \end{cases}$$

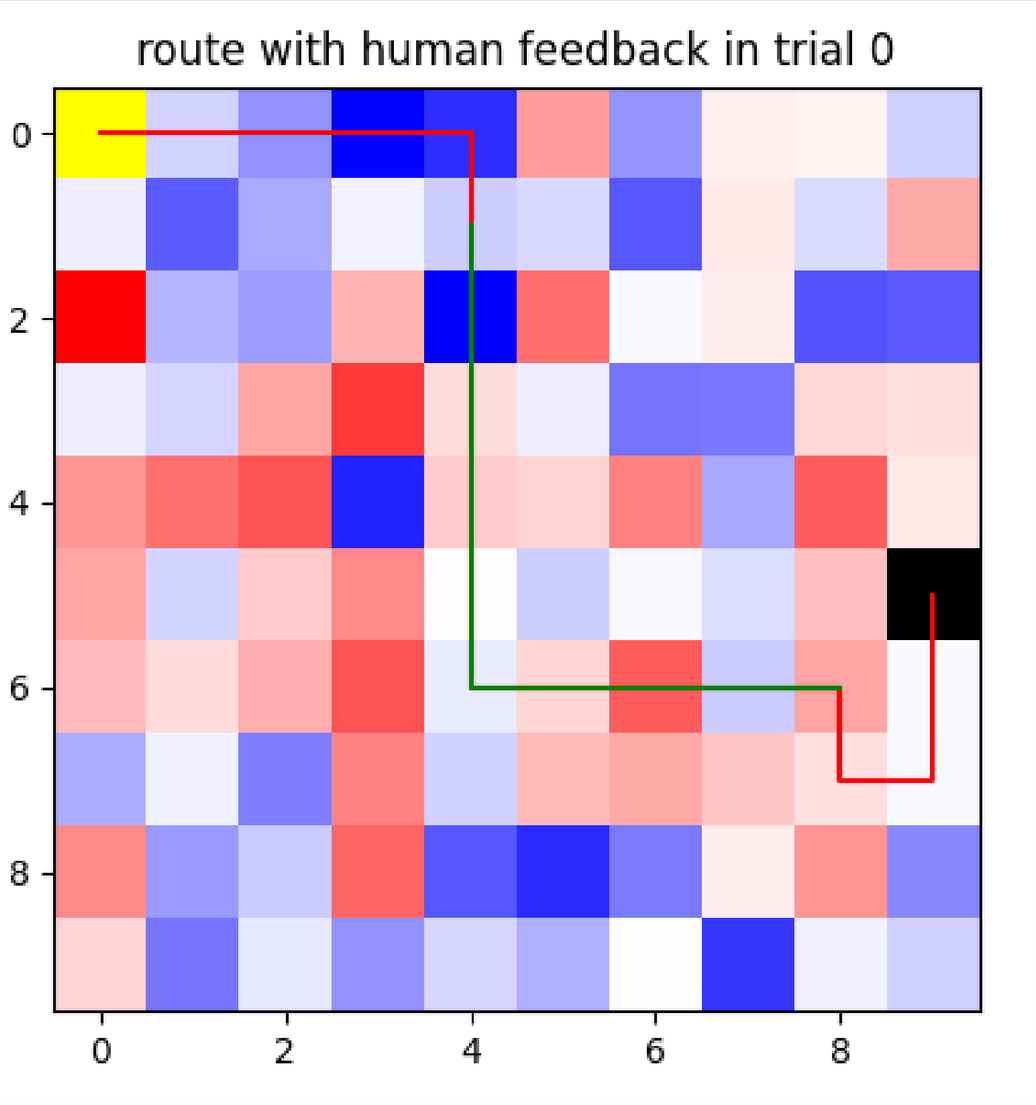
For all state-action-reward-state tuple $(s, a, r, s') \in \tau$,

$Q'(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha \max_{a' \in \mathcal{A}} [r + \gamma Q(s', a')]$

$Q^* \leftarrow Q$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$

$\pi^* \leftarrow \arg\max_{a \in \mathcal{A}} Q(s, a)$

return $Q^*(s, a)$, $\pi^*(s)$ for all s, a

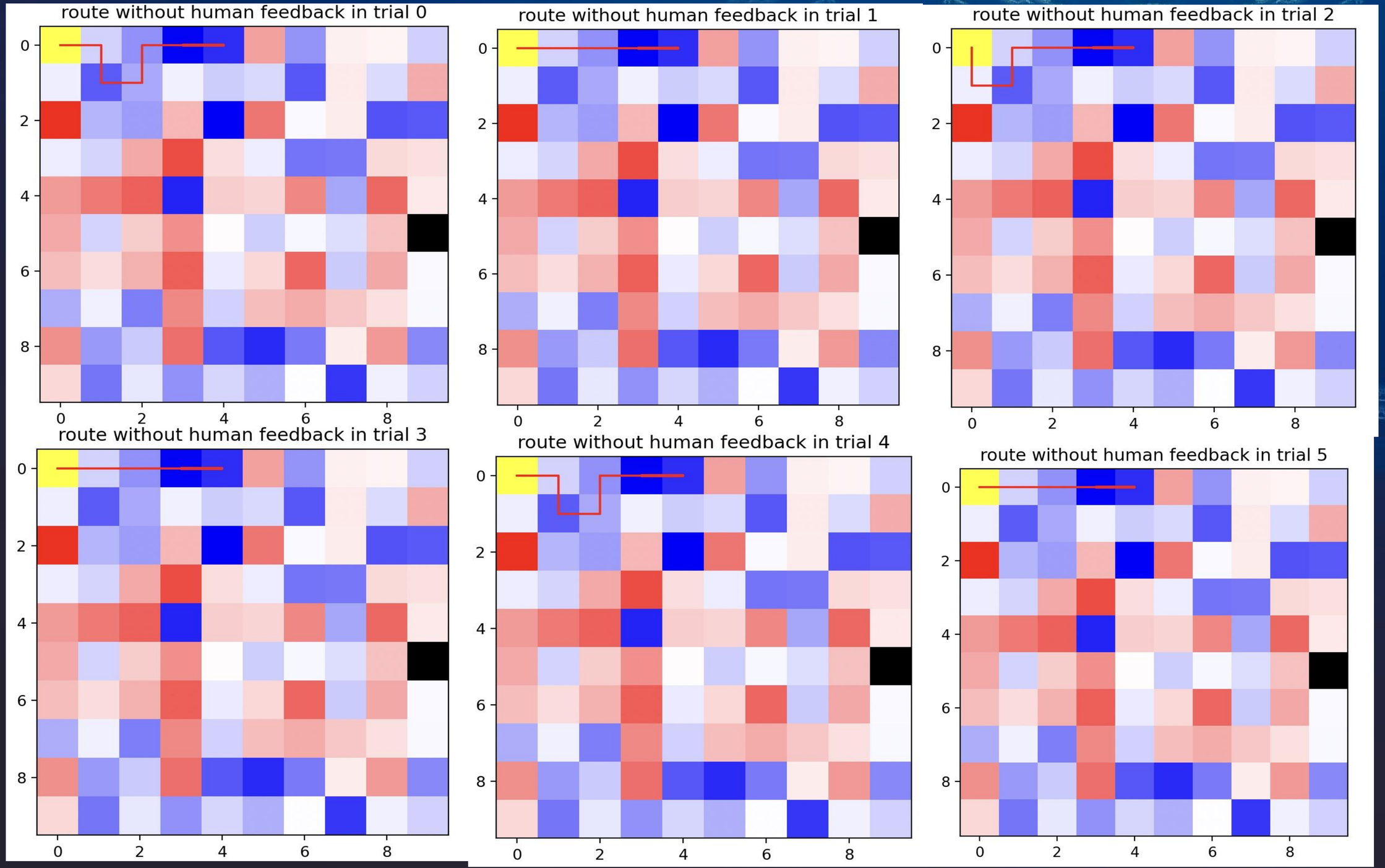



```
Now the agent is stuck between several nodes. Its terminal node now is (2, 1).Please give feedback to help it out.
Please enter one action(0: go up one step, 1: go down one step, 2: go left one step, 3: go right one step).Enter quit if the human feedback is
 done.Please enter your choice:1
The agent transmits to state(3, 1) with the guidance.
Please enter one action(0: go up one step, 1: go down one step, 2: go left one step, 3: go right one step).Enter quit if the human feedback is
 done.Please enter your choice:1
The agent transmits to state(4, 1) with the guidance.
Please enter one action(0: go up one step, 1: go down one step, 2: go left one step, 3: go right one step).Enter quit if the human feedback is
 done.Please enter your choice:3
The agent transmits to state(4, 2) with the guidance.
Please enter one action(0: go up one step, 1: go down one step, 2: go left one step, 3: go right one step).Enter quit if the human feedback is
 done.Please enter your choice:3
The agent transmits to state(4, 3) with the guidance.
Please enter one action(0: go up one step, 1: go down one step, 2: go left one step, 3: go right one step).Enter quit if the human feedback is
 done.Please enter your choice:3
The agent transmits to state(4, 4) with the guidance.
Please enter one action(0: go up one step, 1: go down one step, 2: go left one step, 3: go right one step).Enter quit if the human feedback is
 done.Please enter your choice:1
The agent transmits to state(5, 4) with the guidance.
Please enter one action(0: go up one step, 1: go down one step, 2: go left one step, 3: go right one step).Enter quit if the human feedback is
 done.Please enter your choice:3
The agent transmits to state(5, 5) with the guidance.
```

4

Results

Comparisons of Routes Without/With Human Feedback


route without human feedback in trial 0
route without human feedback in trial 1
route without human feedback in trial 2
route without human feedback in trial 3
route without human feedback in trial 4
route without human feedback in trial 5
0
2
4
6
8

Human feedback

Without:

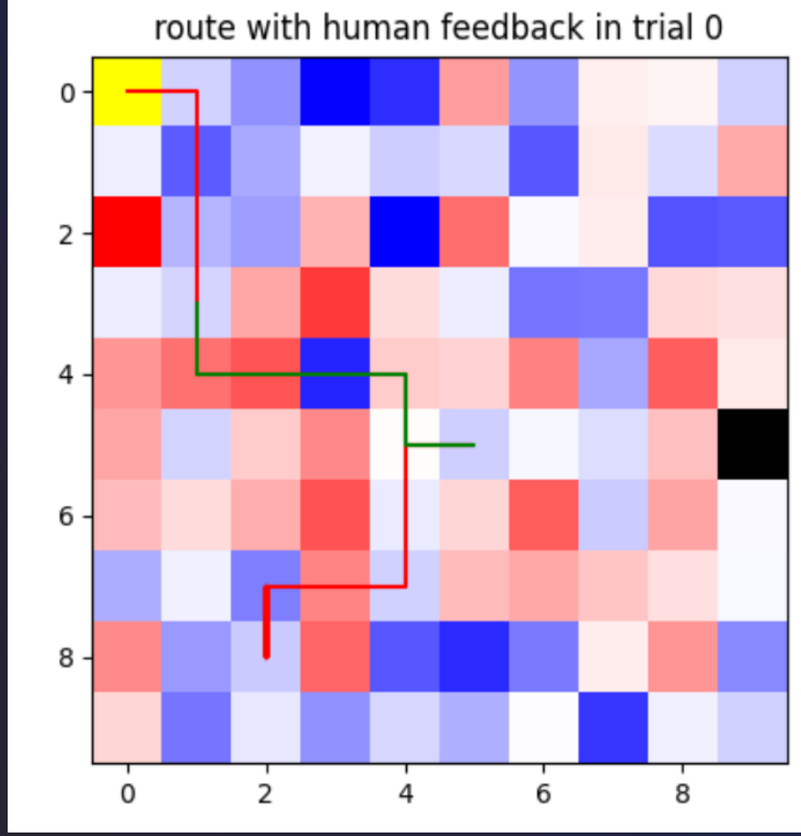
```
reward of the trial 993 without human feedback: 1090.5405966829323
reward of the trial 994 without human feedback: 1015.3917718775193
reward of the trial 995 without human feedback: 1027.051968198225
reward of the trial 996 without human feedback: 1015.3917718775193
reward of the trial 997 without human feedback: 1090.5405966829323
reward of the trial 998 without human feedback: 1043.6335048508274
reward of the trial 999 without human feedback: 630.6738786939053
```

Average reward: 1054.92647

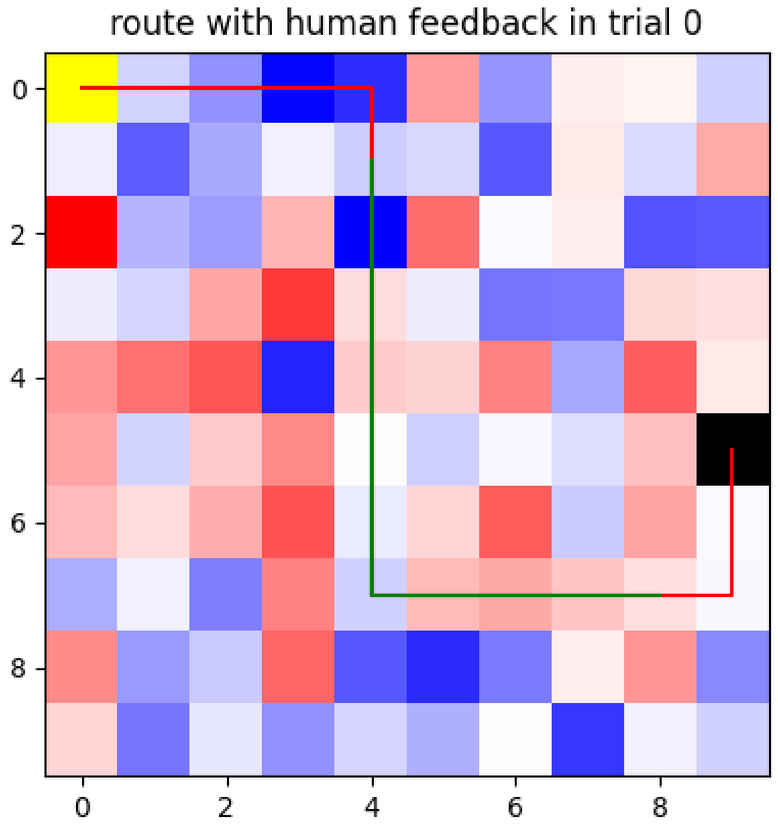
success rate: 0.014

With:

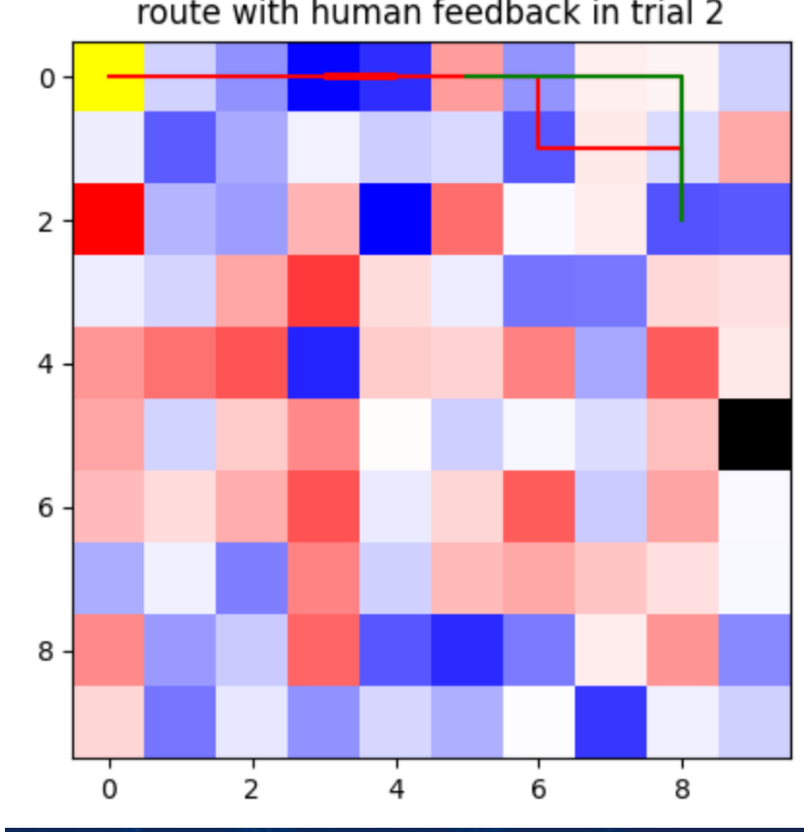
```
Please enter one action(0: go up one step, 1: go down one step, 2: go left one step, 3: go right one step):0
The agent transmits to state(0, 4) with the guidance.
Please enter one action(0: go up one step, 1: go down one step, 2: go left one step, 3: go right one step):4
Invalid input! Please enter again
```



Reward: 206.27
stuck between (2,1)
and (1,1)



Reward: 3877.14
stuck between (0,3)
and (0,4)



Reward: 699.85
stuck between (0,3)
and (0,4)

5

Conclusion

- Results
- Limitations
- Future Prospects


Thank you!

Reference

[1] Thrun S, Littman M L. Reinforcement learning: an introduction[J]. AI Magazine, 2000, 21(1): 103-103.

[2] Hadfield-Menell D, Russell S J, Abbeel P, et al. Cooperative inverse reinforcement learning[J]. Advances in neural information processing systems, 2016, 29.

[3] Karur K, Sharma N, Dharmatti C, et al. A survey of path planning algorithms for mobile robots[J]. Vehicles, 2021, 3(3): 448-468.

[4] Christiano P F, Leike J, Brown T, et al. Deep reinforcement learning from human preferences[J]. Advances in neural information processing systems, 2017, 30.

[5]Kardan A A, Ebrahim M A, Imani M B. A new personalized learning path generation method: Aco-map[J]. Indian Journal of Scientific Research, 2014, 5(1): 17-24.