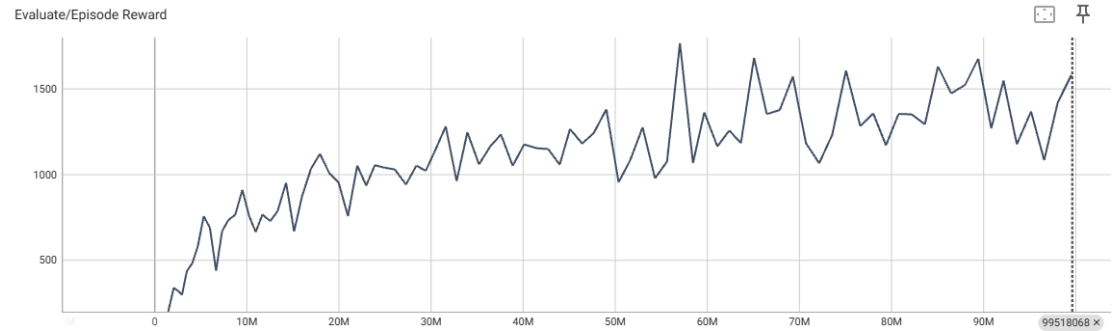
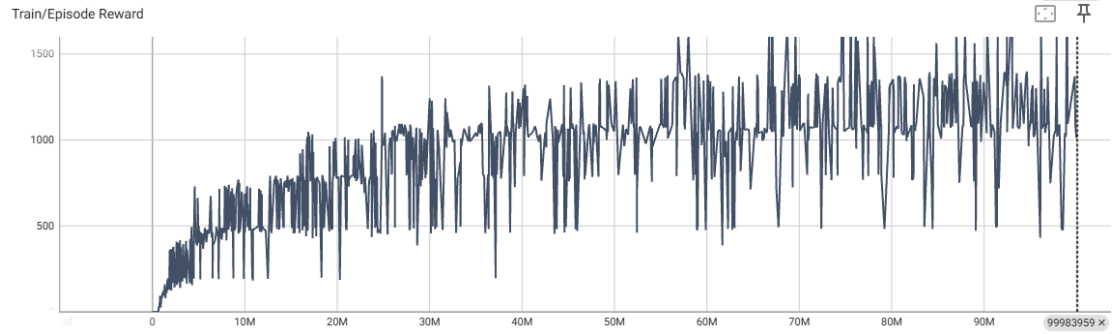
# RL Lab3 Report

**Student ID 110550014 Name: 吳權祐**

## 1. Training Curve & Testing Result




```
================================================
Evaluating...
episode 1 reward: 1373.0
episode 2 reward: 1977.0
episode 3 reward: 1391.0
average score: 1580.3333333333333
================================================
```

## 2. Bonus

### 2.1 PPO is an on-policy or an off-policy algorithm? Why? (5%)

PPO is an on-policy algorithm, learning from data generated by the current policy rather than from past experiences with different policies. It relies on the probability ratio between the current and previous policies, so data must be collected under the same or a very similar policy to ensure reliability.

### 2.2 Explain how PPO ensures that policy updates at each step are not to large to avoid destabilization. (5%)

To ensure a stable training process, PPO uses a clipping mechanism that restricts the probability ratio between the new and old policies to a small range, preventing destabilization.

**2.3 Why is GAE-lambda used to estimate advantages in PPO instead of just one-step advantages? How does it contribute to improving the policy learning process? (5%)**

PPO uses GAE-lambda instead of one-step advantages because GAE-lambda incorporates information about both immediate and long-term rewards, offering a more comprehensive view of future consequences.

**2.4 Please explain what the lambda parameters represents in GAE-lambda, and how adjusting the lambda parameter affects the training process and performance of PPO? (5%)**

Lambda represents the importance of future rewards, indicating how much the model values long-term rewards in the learning process. A higher lambda gives more weight to long-term rewards.