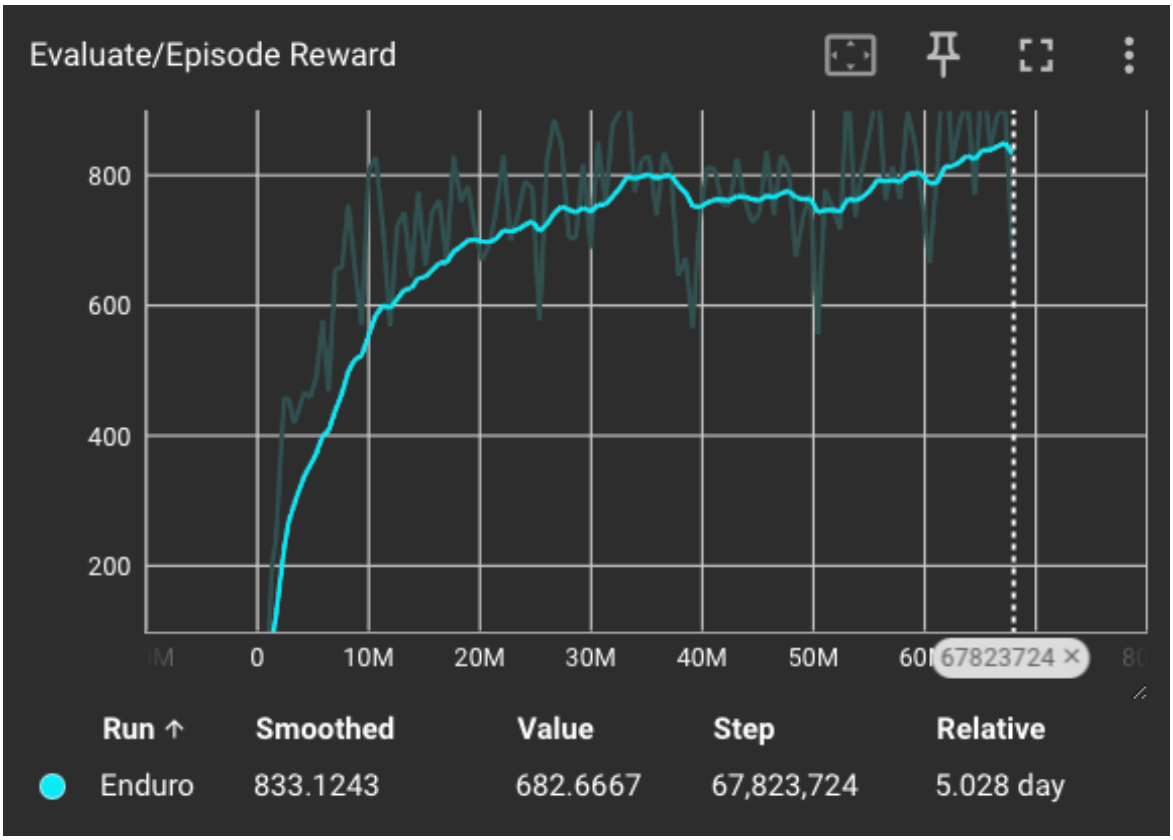
RL Lab3 Report

Name : 林伯偉　　Student ID : 109612019

- **Experiment Results (30%)**



- **Bonus**

1. **PPO is an on-policy or an off-policy algorithm? Why? (5%)**

   PPO is an on-policy algorithm. Because it learns directly from the current policy that it is improving, rather than from past experiences stored in a replay buffer.

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

   PPO ensures that policy updates at each step are not too large to avoid destabilization by using a clipped surrogate objective function. The

objective function compares the ratio of the new policy probability to the old policy probability and takes the minimum with a clipping parameter (typically set to 1 + ε or 1 - ε). This helps prevent large policy updates that could lead to policy oscillations or destabilization during training. The clipped surrogate objective ensures that the policy update is within a controlled range, providing stability to the learning process

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%)

GAE-lambda is used in PPO to estimate advantages instead of just one-step advantages because it provides a trade-off between bias and variance in advantage estimation. GAE-lambda incorporates information from multiple time steps by using a parameter lambda (λ), which allows it to capture the influence of both short-term and long-term effects. This can lead to more accurate advantage estimates and can improve the overall policy learning process by reducing the variance associated with estimating advantages.

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

The lambda parameter in GAE-lambda represents the weighting of the exponentially decaying sum of future advantages. A lambda value of 0 reduces the GAE-lambda to one-step advantages, while a value of 1 includes advantages from all future time steps. Adjusting the lambda parameter affects the training process and performance of PPO by controlling the balance between bias and variance in advantage estimation. A higher lambda value increases the influence of long-term effects, potentially reducing variance but introducing more bias. Conversely, a lower lambda value may reduce bias but increase variance. The choice of lambda depends on the specific characteristics of the environment and the trade-off between bias and variance that the practitioner is willing to accept for improved policy learning.