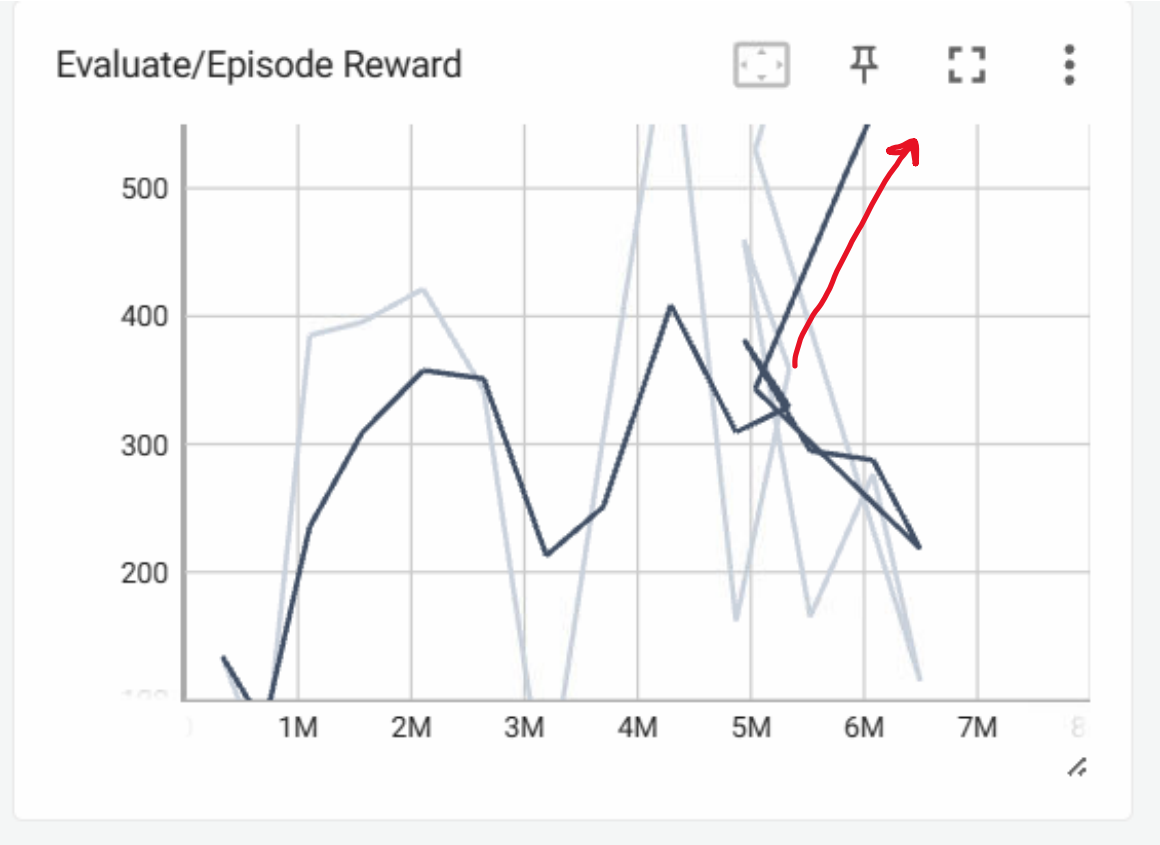
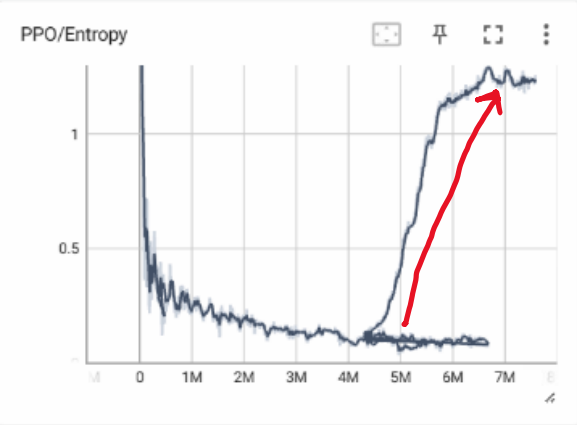
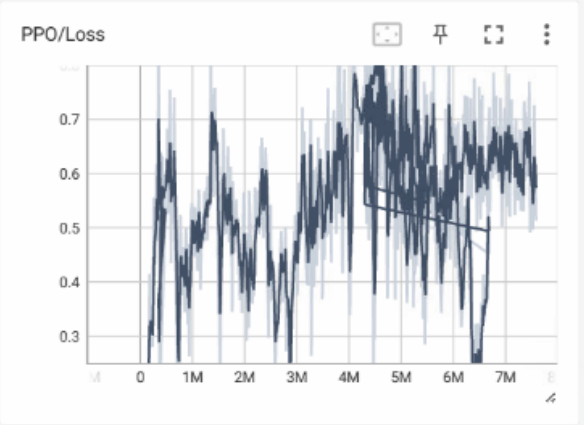
- Experimental Results (30%)

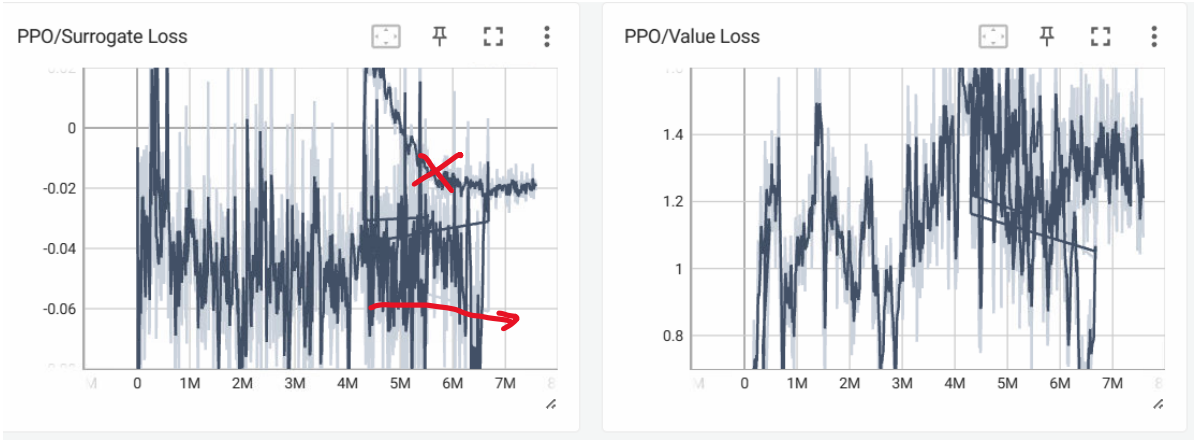
  *我在前 4M step 因為 clip epsilon 設錯，導致前 4M 學習效果並不是很好，在後面 4M 重新修正結果後，效果有明顯上升

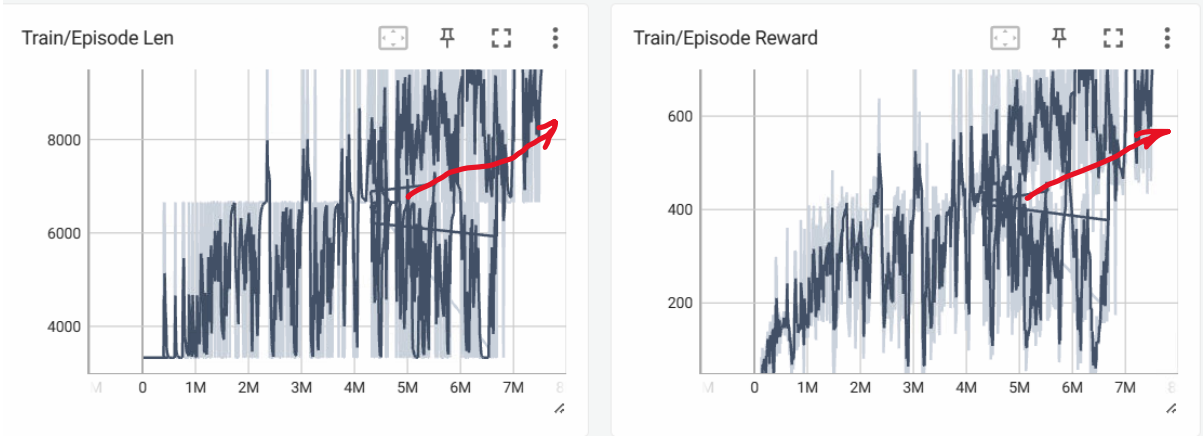
  *原本跑到 7M 修正上述錯誤並從 4M 時的 model 重新開始訓練。而誤會繪圖的方式錯把初始 step 設為 4M(應該要為最後跑)，因此於 4M 出現分岔。我以紅色箭頭註記修正後的紀錄結果

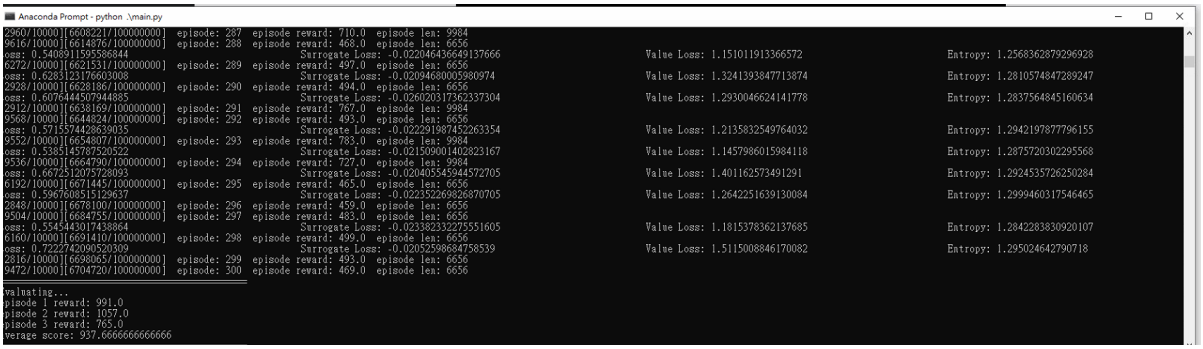












- Answer the questions (bonus) (20%)

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

   PPO 是 on-policy，因為它使用正在 train 的 model 來與環境互動

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

   $r_t(\theta)\widetilde{A_t}$表示每次往 gradient 最大方向更新的距離，如果此值太大容易造成不穩定。PPO 使用 clip 來使得每次更新的距離($move\ lenght$)介於

   $$move\ lenght < (1-\varepsilon)\widetilde{A_t}\ , if\ \widetilde{A_t} < 0$$

   $$move\ lenght < (1+\varepsilon)\widetilde{A_t}\ , if\ \widetilde{A_t} > 0$$

   如此保證不會一次太大步而不穩定

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%)

$$\widetilde{A}_t^{GAE(\gamma,\lambda)} = (\delta_t^V + (\lambda\gamma)^1\delta_{t+1}^V + \cdots + (\lambda\gamma)^k\delta_{t+k}^V + \cdots)$$

One-step advantage $\widetilde{A}_t^{GAE(\gamma,0)} = \widetilde{A}_t^{(1)} = TD\_error$

其只考慮了下一步的結果所產生的 TD error 並以此來衡量 action 的好壞，但由於整個過程是 stochastic 可能只是剛好此次結果好/不好下次就是不好/好，其後續有很大的 variance，因此採用 Lambda-GAE 方式多考慮更長遠的結果來減少 variance 以此減少 variance 避免偶而遇到的極糟情況讓整個 model 壞掉。

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%)

Lambda 的大小用來調整要考慮非常長遠還是近期就好，當 Lambda 越小時，model 的 bias 越低但隨之而來的 variance 也越高，對 noise 的對抗性越低，容易在訓練過程 model 一遇到很糟狀況而被拖垮；反之如果太高則 model 的準確度越低(bias 大)，但對 noise 容忍度較高幫助 model 穩定