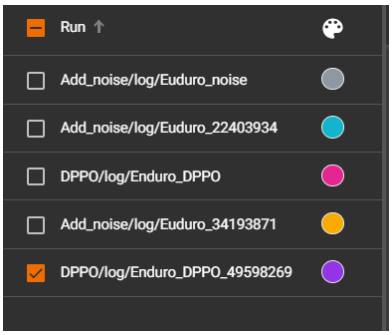
Lab 3: Proximal Policy Optimization (PPO)

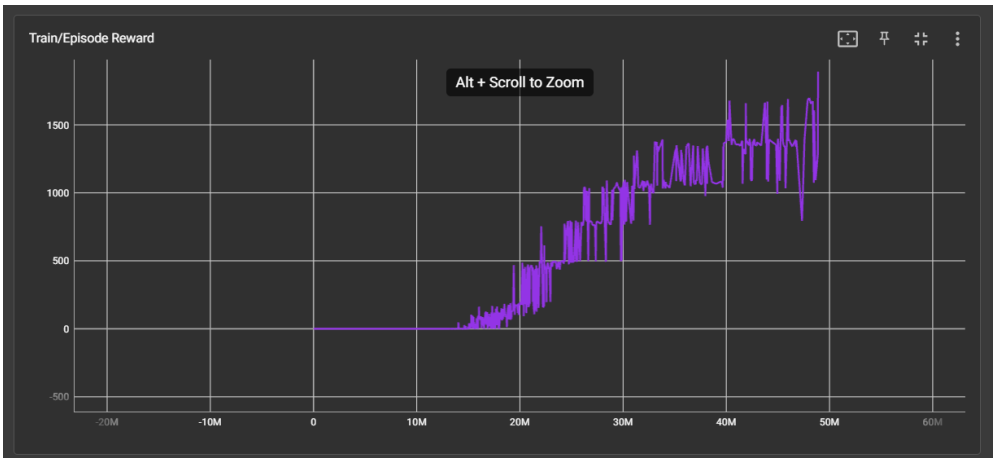
學生：陳澤昕
學號：311356003

1. Experimental Results (30%)

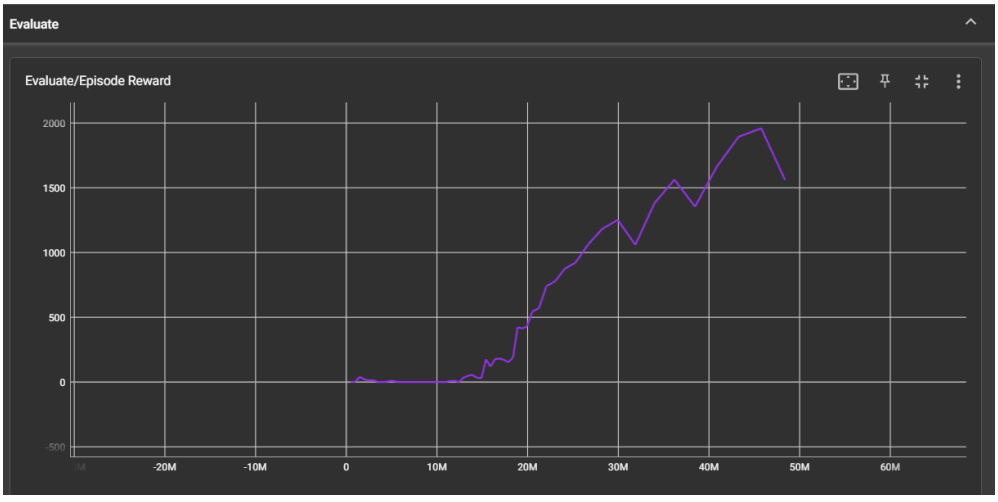
- Index



- Training Reward



- Evaluation Reward



2. Answer the questions (bonus) (20%)
    1. PPO is an on policy or an off policy algorithm? Why? (5%)
       PPO（Proximal Policy Optimization）GAE lambda 中的 lambda 參數代表權重調整，影響訓練過程和 PPO 性能。調整 lambda 值可平衡長短期效果，高 lambda 強調長期獎勵，低 lambda 側重短期獎勵。調高 lambda 可穩定訓練，但減慢收斂，反之亦然。選擇適當 lambda 可改善 PPO 的性能，但需根據任務特性調整。是一種基於當前策略的算法，所以它是一種 on-policy 的算法。代表它會收集當前策略所產生的數據，然後 agent 根據這些數據來更新策略。
    2. Explain how PPO ensures that policy updates at each step are not too large to avoid destabilization. (5%)
       PPO（Proximal Policy Optimization）透過「Clip」機制確保每個步驟的策略更新幅度不會太大，降低不穩定性。這個機制限制了策略在連續迭代的改變幅度，主要是透過 clip agen 目標函數和行動機率比率，確保策略的變化不會太過劇烈，進一步保持穩定的訓練過程。
    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 在 PPO 中用來估計 advantages value ，相較於一步，它的好處有：減少變異、更長的步數增加 reward 的考靠度、提高樣本。這有助於提升學習的穩定性和效率。
    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%)
       GAE lambda 中的 lambda 參數表示長短獎勵的重視度，影響訓練過程和 PPO 性能。調整 lambda 值可平衡長短期效果，高 lambda 注重長期獎勵，低 lambda 注重短期獎勵。高 lambda 可穩定訓練，但減慢收斂，反之亦然。選擇適當 lambda 可加速收斂。