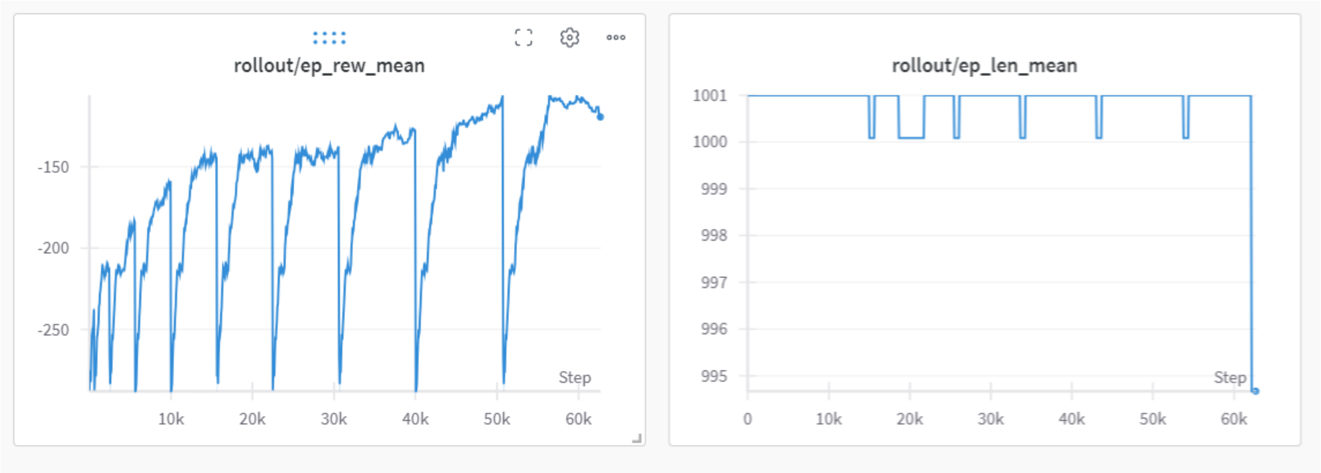
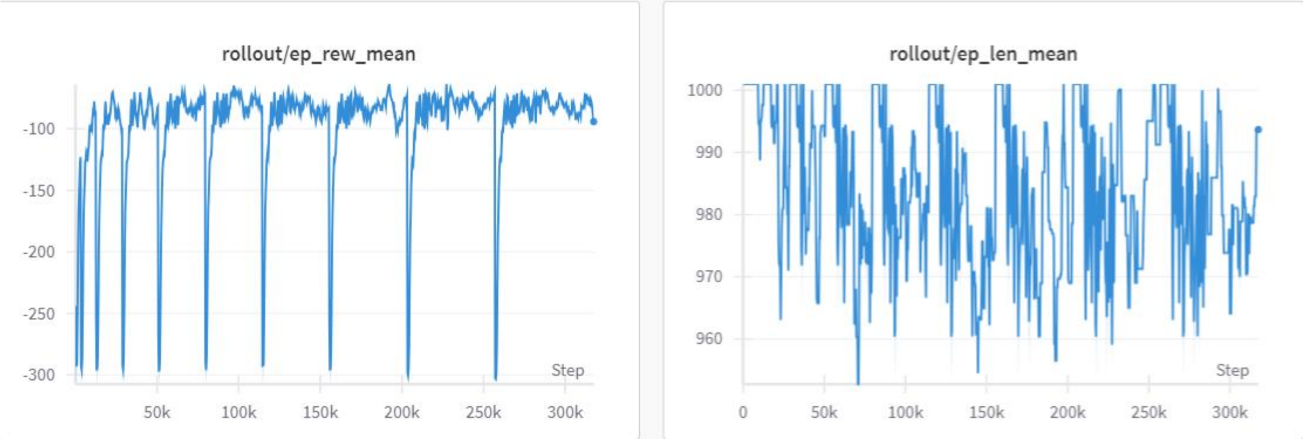
Experiment tracking for RL models using wandb:

1st iteration:



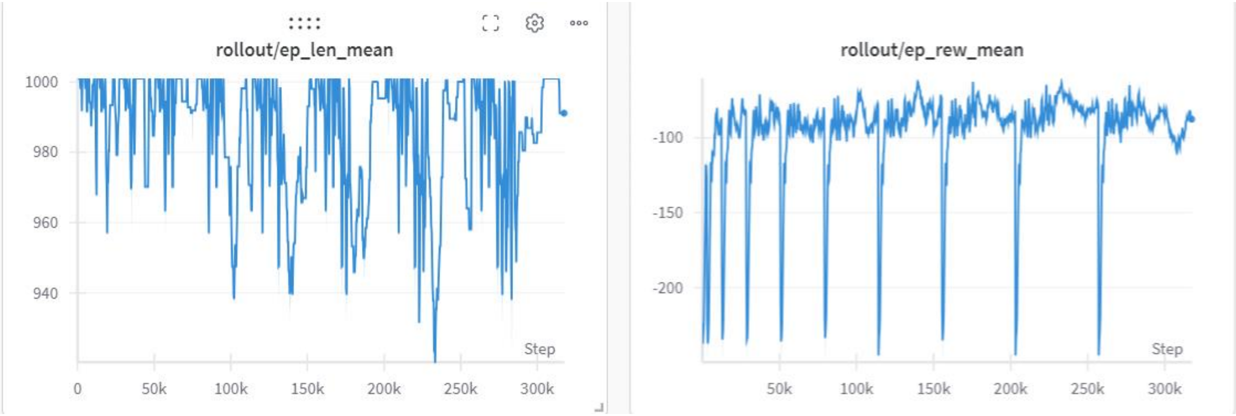
Batch size: 128, LL: 0.0001, time_steps: 100,000

2nd iter:



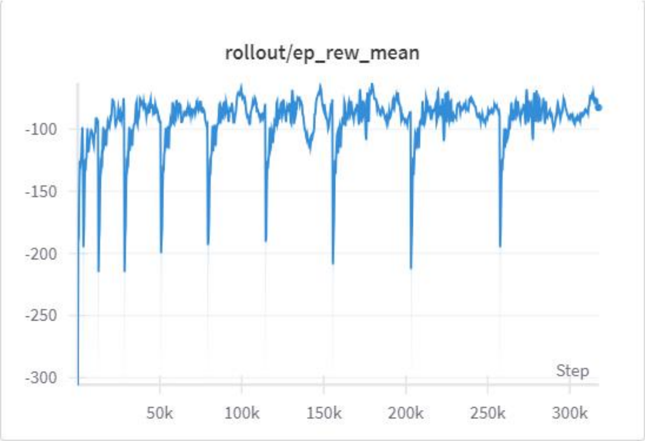
Batch size: 64, LL: 0.0001, time_steps: 300,000

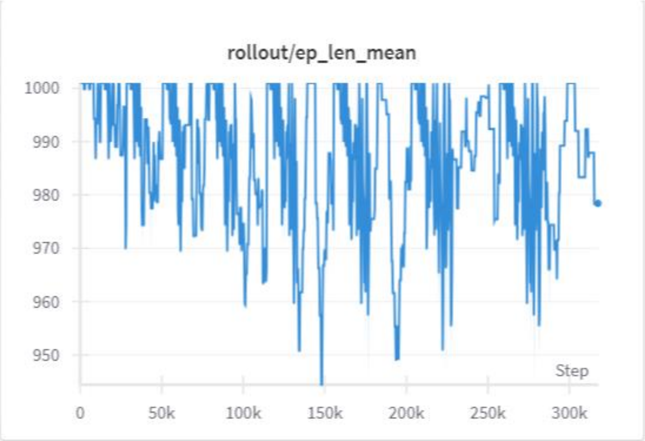
3rd Iter:



Batch size: 512, LL: 0.0003, time steps: 300,000

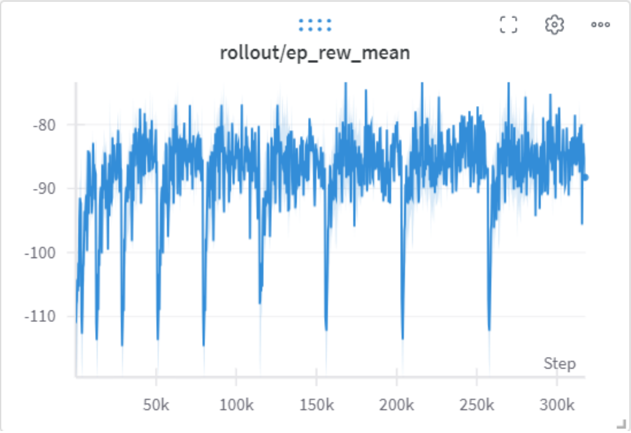
4th Iter:

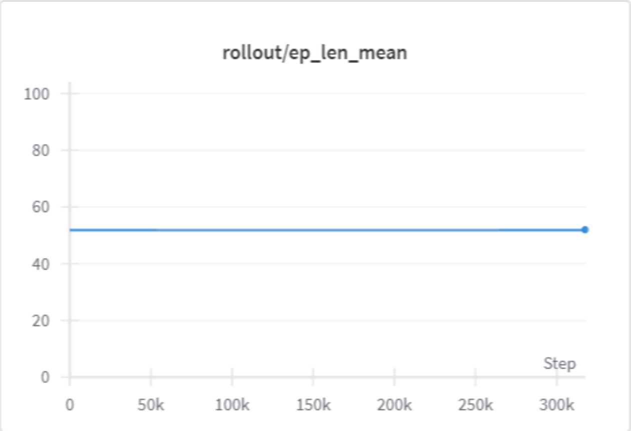




Batch size: 128, LL: 0.0003, Time steps: 300.000

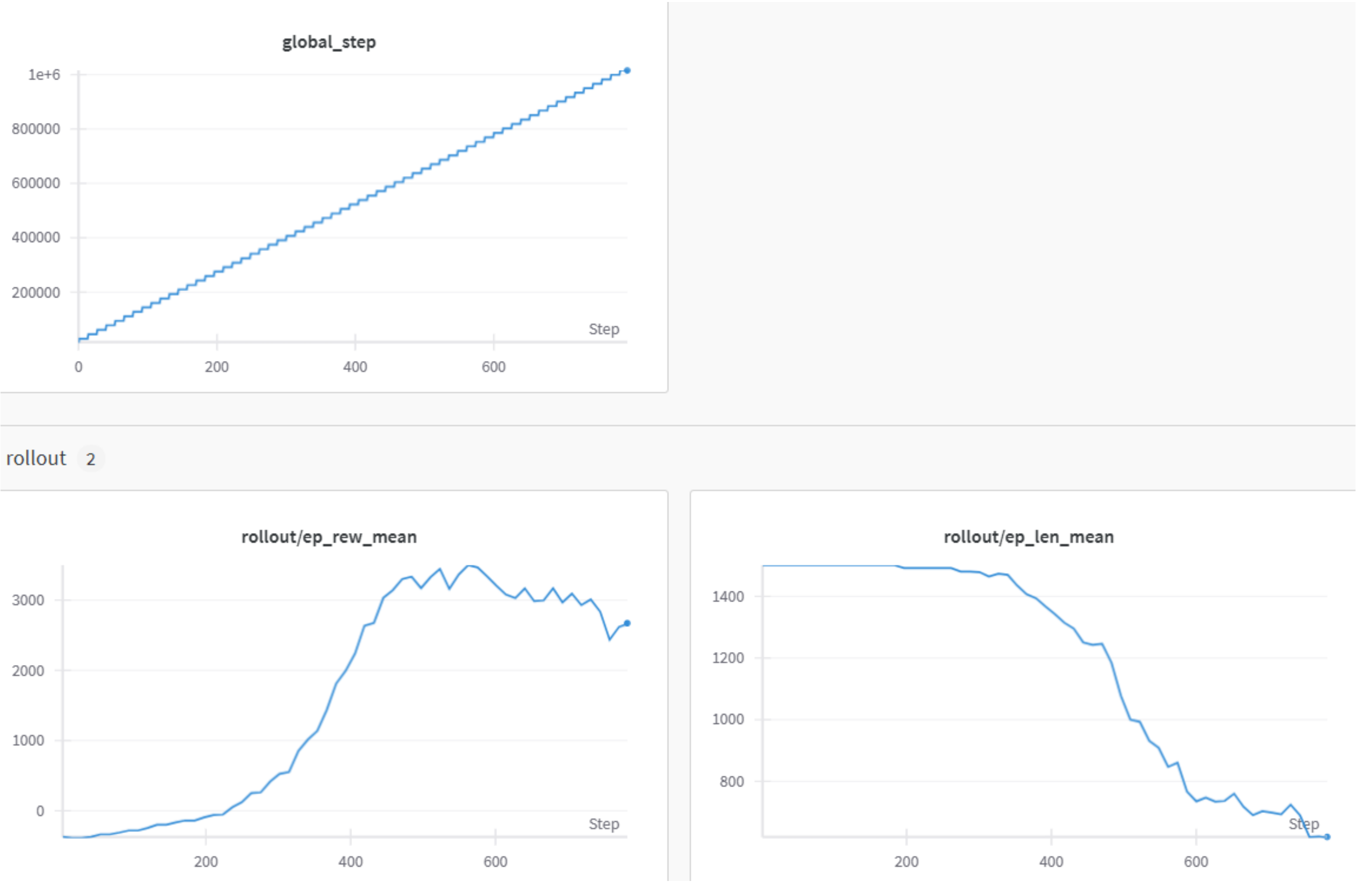
5th Iter:





Batch size: 64, LL: 0.0001, time steps: 500.000 over 10 epochs

Last Iter:



Batch size: 128, LL: uses a linear schedule (0.1), time steps: 1000000, and new reward logic

This one yielded the best results