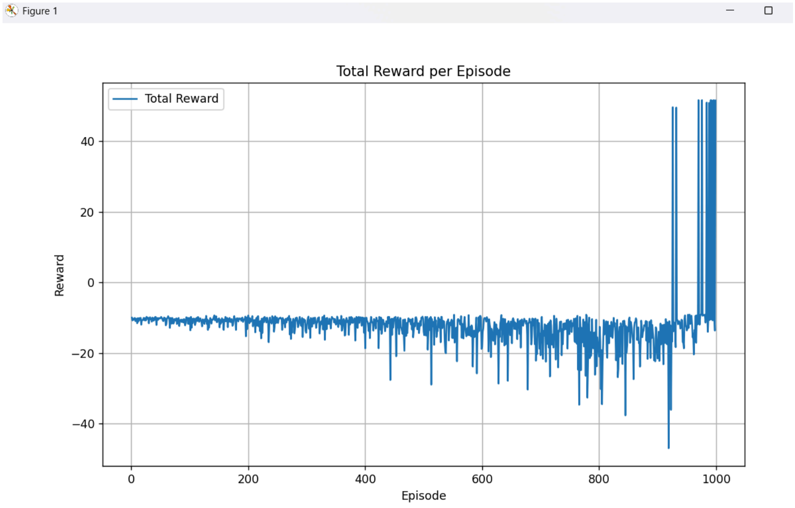
Part One: Exploration-Exploitation Trade-Off

In reinforcement learning, epsilon values determine how much an agent values exploring locations that it has never been vs exploiting locations that it knows is good. Let's evaluate the differences in Epsilon values.

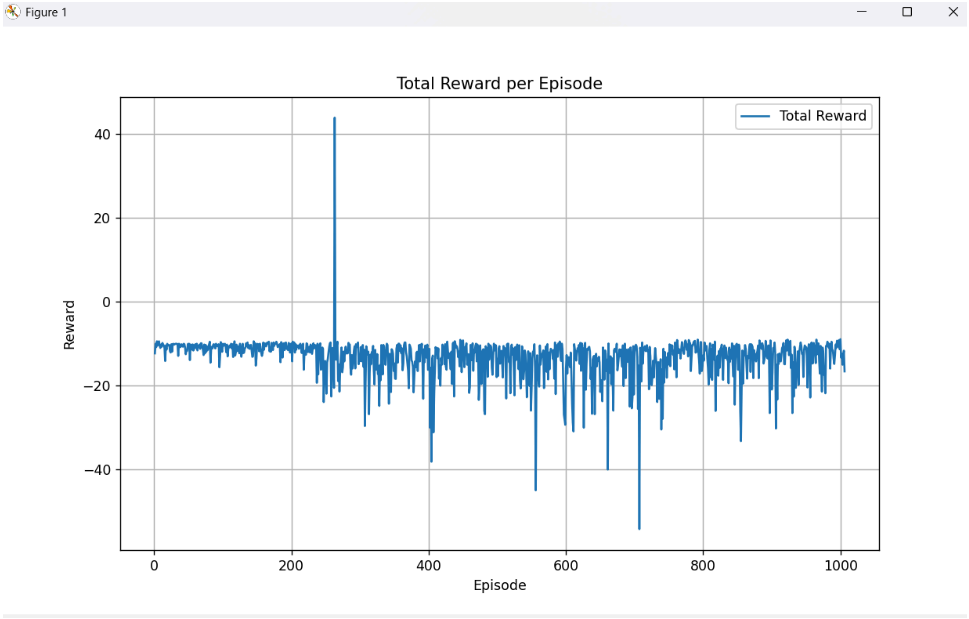
1A High Value: Epsilon decay: 5000

In this first instance, we tried an epsilon decay of 5000. This means that the epsilon value decays quite slowly. This means that it will favor exploration over exploitation. So we see as time goes on, it eventually reaches very high reward episodes because it was exploring a lot in the beginning but then it begins exploiting at the end.



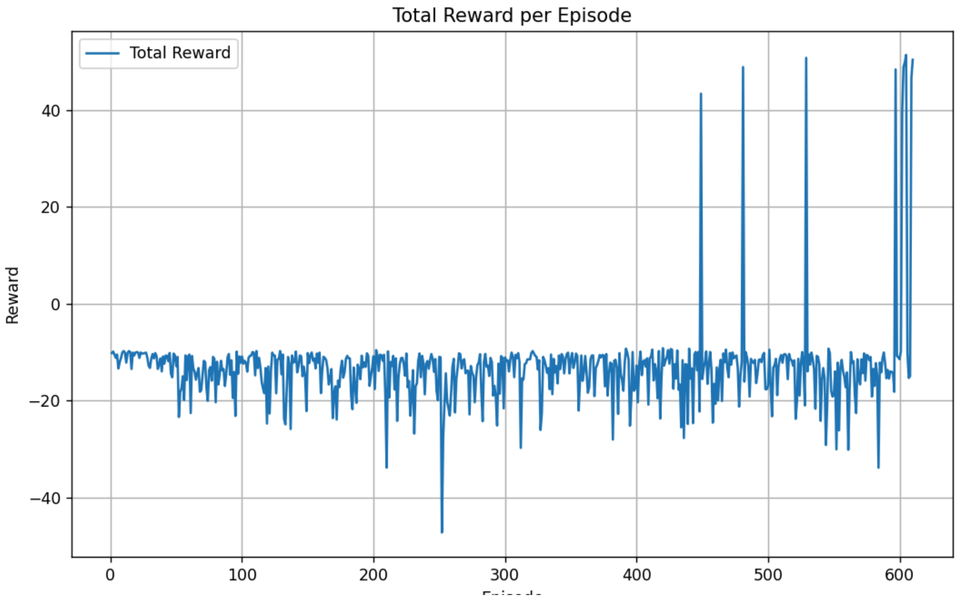
1B Medium value: Epsilon decay: 2500

This epsilon decay is half of the decay of the first one. So, the decay is relatively intermediate. The agent should balance new exploration and exploitation of good areas. In this specific run, we see that the agent saw relatively steady growth, but it never reached really high reward states. It could be just due to the nature of this run - if we kept going then it probably would reach high reward states, but since that didn’t happen, we had to terminate the run at 1000.



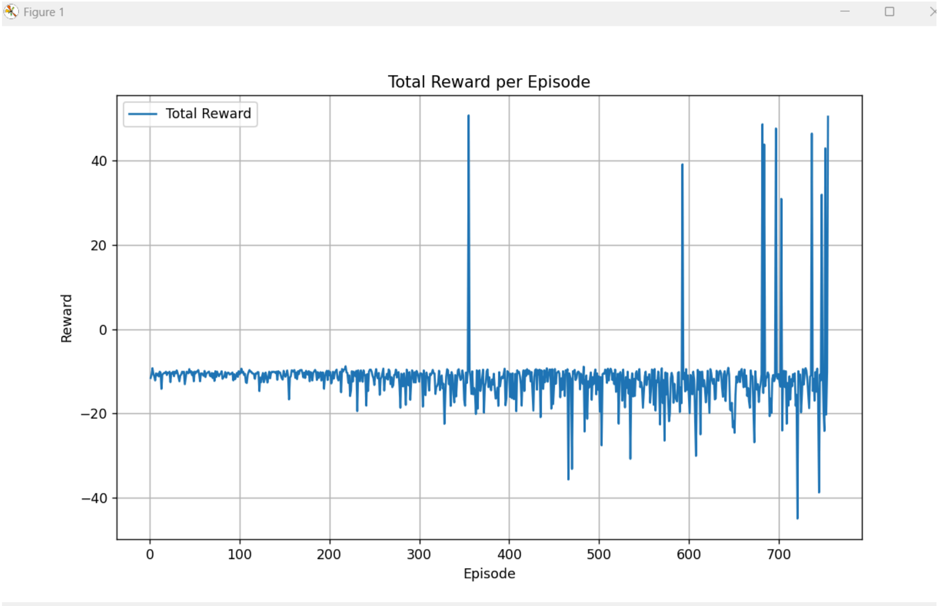
1C low value: Epsilon decay: 500

In this scenario, we have 500, which is a very rapid epsilon decay. It means the agent will not spend a long time exploring and will try to maximize exploitation. We see that the agent had mainly negative utility. However, towards the end, it started finding the goal and exploited it very well. It could be that an agent with this low of an epsilon will get stuck in a local maxima but that did not happen in this specific run. The agent found the goal ten times in only 600 tries. This could be because the agent found a state of high utility and prioritized running back to it.

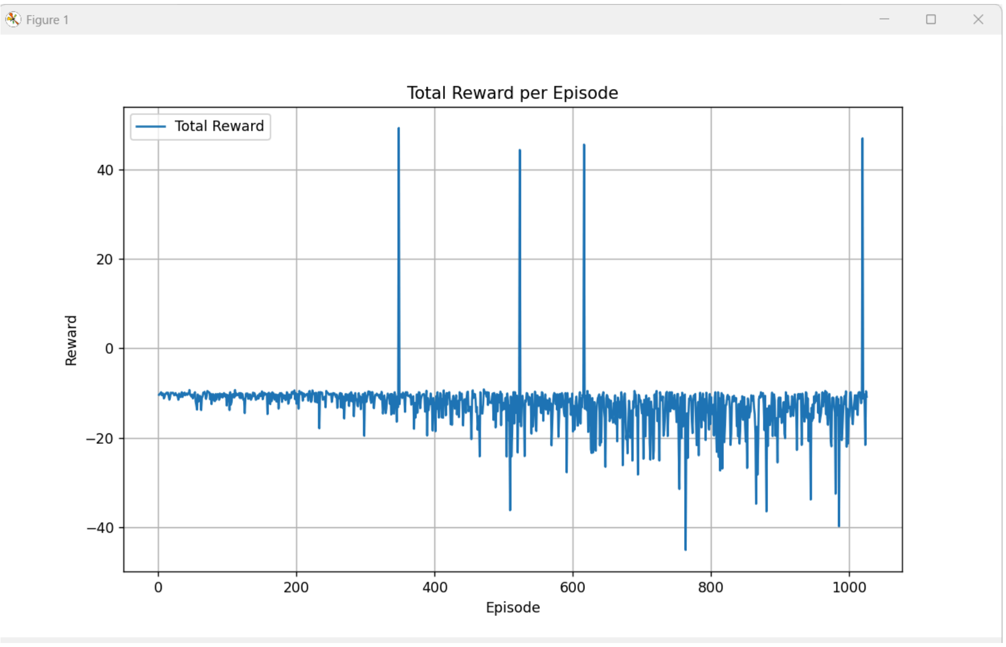


Part 2: Gamma Values

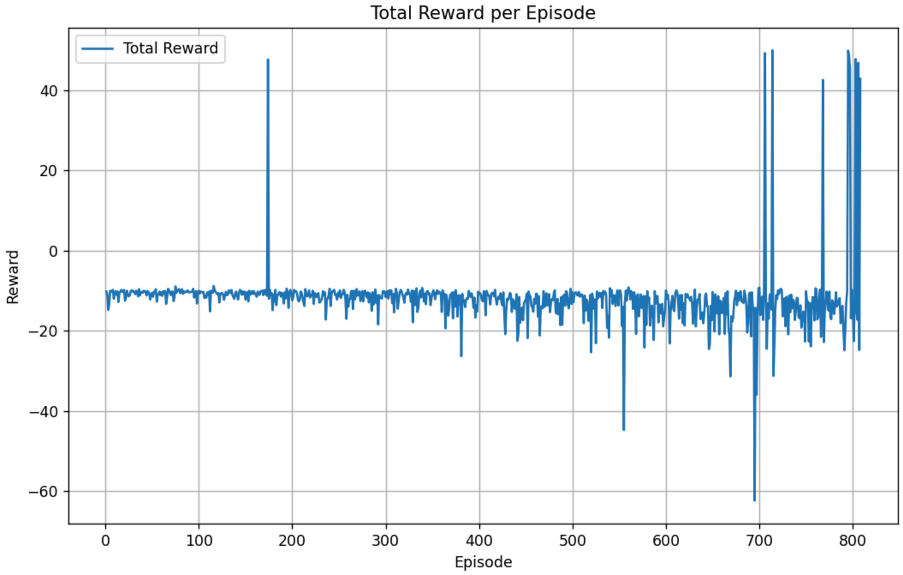
2A low Gamma - 0.3: A low gamma means that the agent prioritizes the short term rewards. So in this case, the agent happened to reach the goal ten times relatively quickly - before 1000 iterations. This is probably because the agent prioritizes the short term, so once it found the goal the first few times, it probably continued to run to it to maximize its utility. Sometimes you can see an agent get stuck in local maxima, but that did not happen.



2B medium Gamma - 0.6: A medium gamma balances between long term outlook and shortsightedness. We can see the agent does not do a great job of finding the goal state in 1000 episodes. It finds the goal state some of the time, but due to its tendency to prioritize short term gains, it does not find it all the time.



2C High Gamma - 0.99: A high gamma favors long term outlook. So in this run, it took a long time for the agent to choose something favorable. But we see that the agent began finding better solutions and found optimal goal states. Once it starts finding the goal, it continues to find it, prioritizing the long term goal over getting stuck in local maxima.


Total Reward per Episode
Total Reward
Reward
Episode