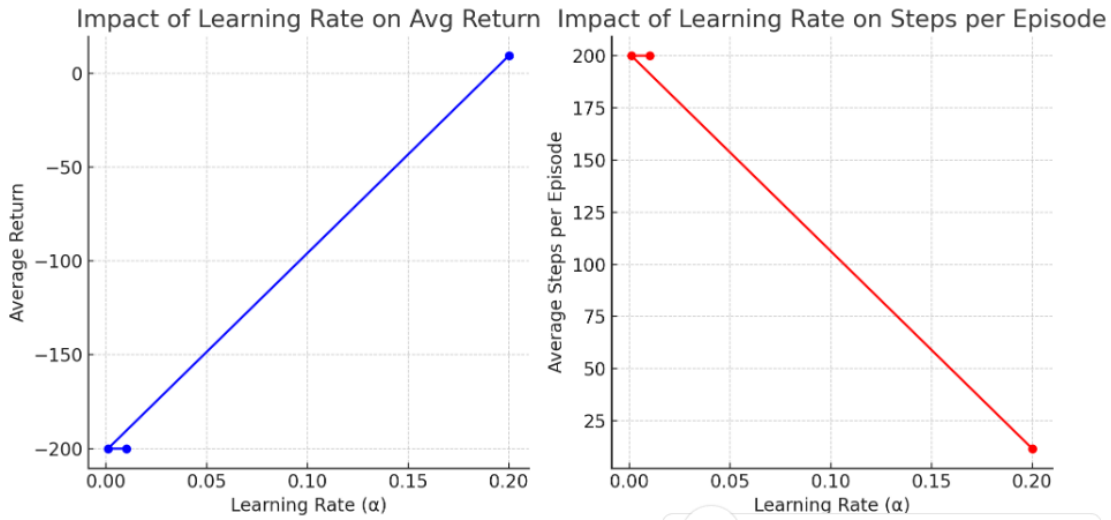
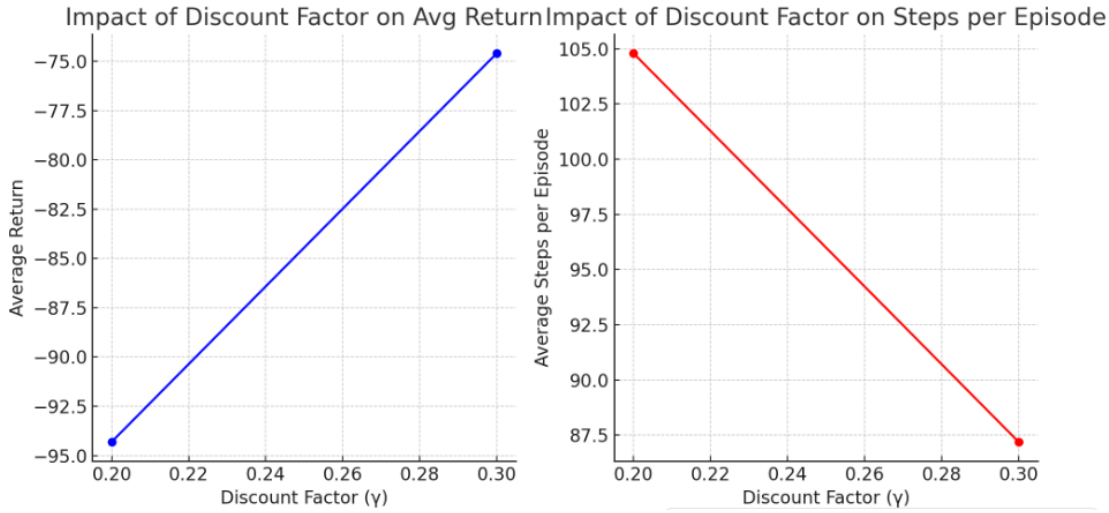
Matrix:

α	γ	Total Episodes	Avg Return	Avg Steps
0.01	0.9	2000	−200	200
0.001	0.9	2000	−200	200
0.2	0.9	2000	9.3	11.7
0.1	0.2	2000	−94.3	104.8
0.1	0.3	2000	−74.6	87.2
0.1	0.9	2000	−34.3	51.1

α = 0.01 and α = 0.001: Learning is very bad. They achieve a return of -200.0 which is the worst performance. The agent took the maximum allowed steps 200, indicating it was unable to optimize its decisions. α = 0.2 shows significantly better performance. This is able to indicate that a higher learning rate lead to faster Q-table updates, leading to better policy convergence.

Also, lower discount factors ,γ = 0.2, 0.3, led to worse performance compared to γ = 0.9. γ = 0.2 performed the not good, with an avg return of -94.3 and 104.8 avg steps. This could show that the agent failed to learn effective long-term strategies. γ = 0.3 was slightly better, but still significantly worse than γ = 0.9. The comparison might suggesting that future rewards is crucial for optimal performance.





Based on the findings, the best hyperparameter combination is $\alpha = 0.2$,$\gamma = 0.9$. This combination has faster convergence with lower steps per episode and higher average return compared to suboptimal \u03b1 and \u03b3 values.