1. The trial with the highest last average 50 reward is the number B test  using an eps start of 1 and eps end of 0.05 with alpha 0.15, wherein I got 4.64

2. When training finishes, the Q-values at the start state tell the agent how good each move (U, D, L, R) looks in terms of expected future reward. Among these, the Q-value for moving right (R) becomes the largest, so it is the first greedy move in all test cases A, B, and  C because going right is the shortest path toward the nearby dirt. The agent received an early +5 reward after only two steps. Since greedy action selection always picks the move with the largest Q-value, the agent's first greedy move is R.

3. It is much lower because it is a farther starting point than others; therefore, the tiles need much more learning to improve them. I can add a slight alpha increase.

4. Bumping the learning rate (α) from 0.1 -> 0.2 makes each Q-update twice as aggressive, so the curve typically climbs faster and plateaus sooner, but it also gets a bit noisier/jittery; here it improves the last-50 average sooner (faster convergence).

5. In practice, I would prefer the slightly higher alpha because it offers speeds, though it could make the steps jittery; it still accomplishes its job.