Bonus: Suggested Improvements to the Implementation

For this homework, the main algorithms used are Epsilon-Greedy and Thompson Sampling. While these algorithms are effective, the experiment could be improved by incorporating additional strategies and refinements:

1. Upper Confidence Bound (UCB1) Algorithm
   - Epsilon-Greedy explores randomly, which may not efficiently balance exploration and exploitation. UCB1 uses confidence intervals to guide exploration based on uncertainty, choosing the bandit with the highest potential reward.
   - Benefits:
     - Smarter exploration: Focuses on actions with uncertain or high potential rewards.
     - Reduced cumulative regret over time.
     - Particularly effective when the number of trials is limited.
   - Implementation Idea: Create a UCB1() class that inherits from the abstract Bandit() class, similar to Epsilon-Greedy and Thompson Sampling. Track the mean reward and number of pulls per bandit, and select the bandit according to the UCB1 formula: UCB = mean reward + sqrt(2 * ln(t) / n_i), where t is the current trial and n_i is the number of times bandit i has been pulled.
2. ***Adaptive Epsilon Strategy***
   - Instead of a simple epsilon = 1/t decay, consider dynamic adjustment based on observed reward variance. Bandits with high variance might require more exploration.
3. ***Parallel Experimentation***
   - Run multiple independent simulations in parallel to better estimate expected performance and reduce variance in cumulative reward estimates.

Conclusion

Incorporating UCB1 or other adaptive strategies could reduce regret, make smarter decisions under uncertainty, and provide richer insights into bandit performance. These enhancements would strengthen the experiment beyond the basic Epsilon-Greedy and Thompson Sampling algorithms.