RL CIA-1

K-arm bandit solution for

1. Recommendation system

    - **Approach 1**:
        - Select recommendations based on an optimistic estimate of the reward. It uses a confidence interval that adjusts over time, favoring options that have either high average rewards or have been less explored.
        - Estimated reward + sqrt(2ln(t)/(number of times this option was chosen)).
        - Where 't' is the total number of trails sp far.
        - For each arm (recommendation), a probability distribution is maintained for the potential reward.
        - At each step, a reward is sampled from each arm's distribution.
        - The arm with the highest sampled reward is selected.
        - After observing the actual reward, the probability distribution is updated to better reflect future predictions.
    - **Approach 2**:
        - Uses a Bayesian approach to model the distribution of rewards for each option, selecting recommendations based on sampling from these distributions.
        - For each arm, sample from the posterior distribution of the reward and choose the arm with the highest sampled value.
        - The arm with the highest UCB is selected.
        - This approach ensures that less-explored arms are given a chance if their confidence bounds suggest high potential.