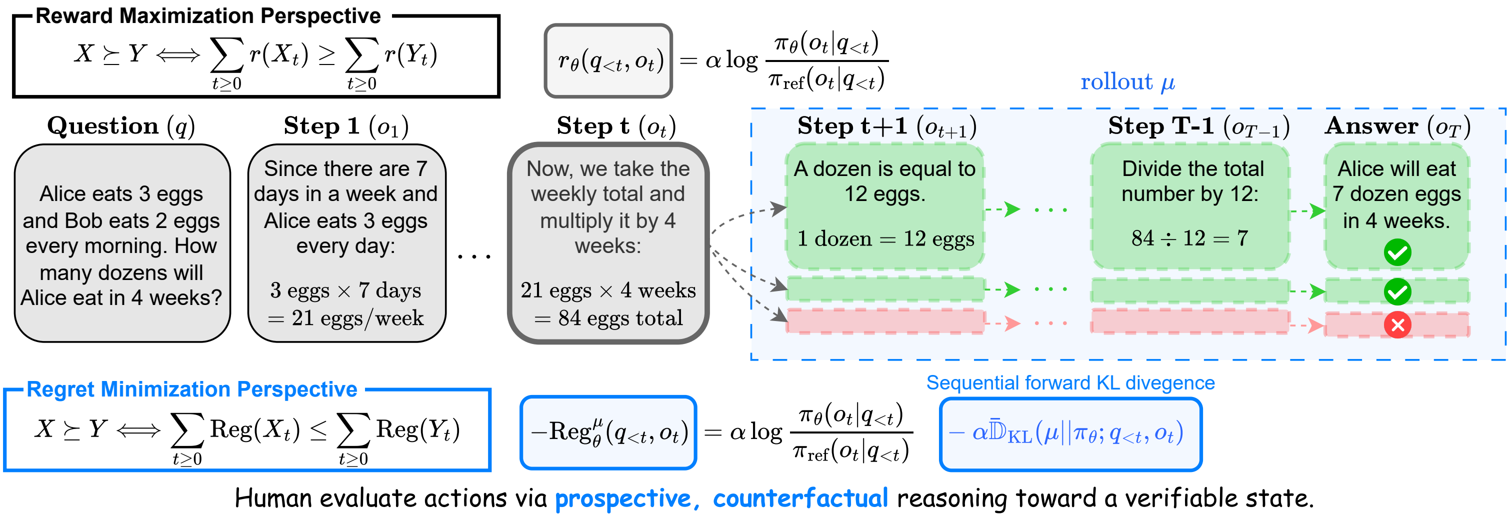
Reward Maximization Perspective
$X \succeq Y \Longleftrightarrow \sum_{t\geq 0} r(X_t) \geq \sum_{t\geq 0} r(Y_t)$
$r_\theta(q_{<t}, o_t) = \alpha \log \frac{\pi_\theta(o_t|q_{<t})}{\pi_{\mathrm{ref}}(o_t|q_{<t})}$
rollout μ
Question (q)
Alice eats 3 eggs and Bob eats 2 eggs every morning. How many dozens will Alice eat in 4 weeks?
Step 1 (o_1)
Since there are 7 days in a week and Alice eats 3 eggs every day:
$3 \text{ eggs} \times 7 \text{ days} = 21 \text{ eggs/week}$
Step t (o_t)
Now, we take the weekly total and multiply it by 4 weeks:
$21 \text{ eggs} \times 4 \text{ weeks} = 84 \text{ eggs total}$
Step t+1 (o_{t+1})
A dozen is equal to 12 eggs.
$1 \text{ dozen} = 12 \text{ eggs}$
Step T-1 (o_{T-1})
Divide the total number by 12:
$84 \div 12 = 7$
Answer (o_T)
Alice will eat 7 dozen eggs in 4 weeks.
Regret Minimization Perspective
$X \succeq Y \Longleftrightarrow \sum_{t\geq 0} \mathrm{Reg}(X_t) \leq \sum_{t\geq 0} \mathrm{Reg}(Y_t)$
$-\mathrm{Reg}_\theta^\mu(q_{<t}, o_t) = \alpha \log \frac{\pi_\theta(o_t|q_{<t})}{\pi_{\mathrm{ref}}(o_t|q_{<t})}$
Sequential forward KL divergence
$-\alpha \bar{\mathbb{D}}_{\mathrm{KL}}(\mu||\pi_\theta; q_{<t}, o_t)$
Human evaluate actions via prospective, counterfactual reasoning toward a verifiable state.