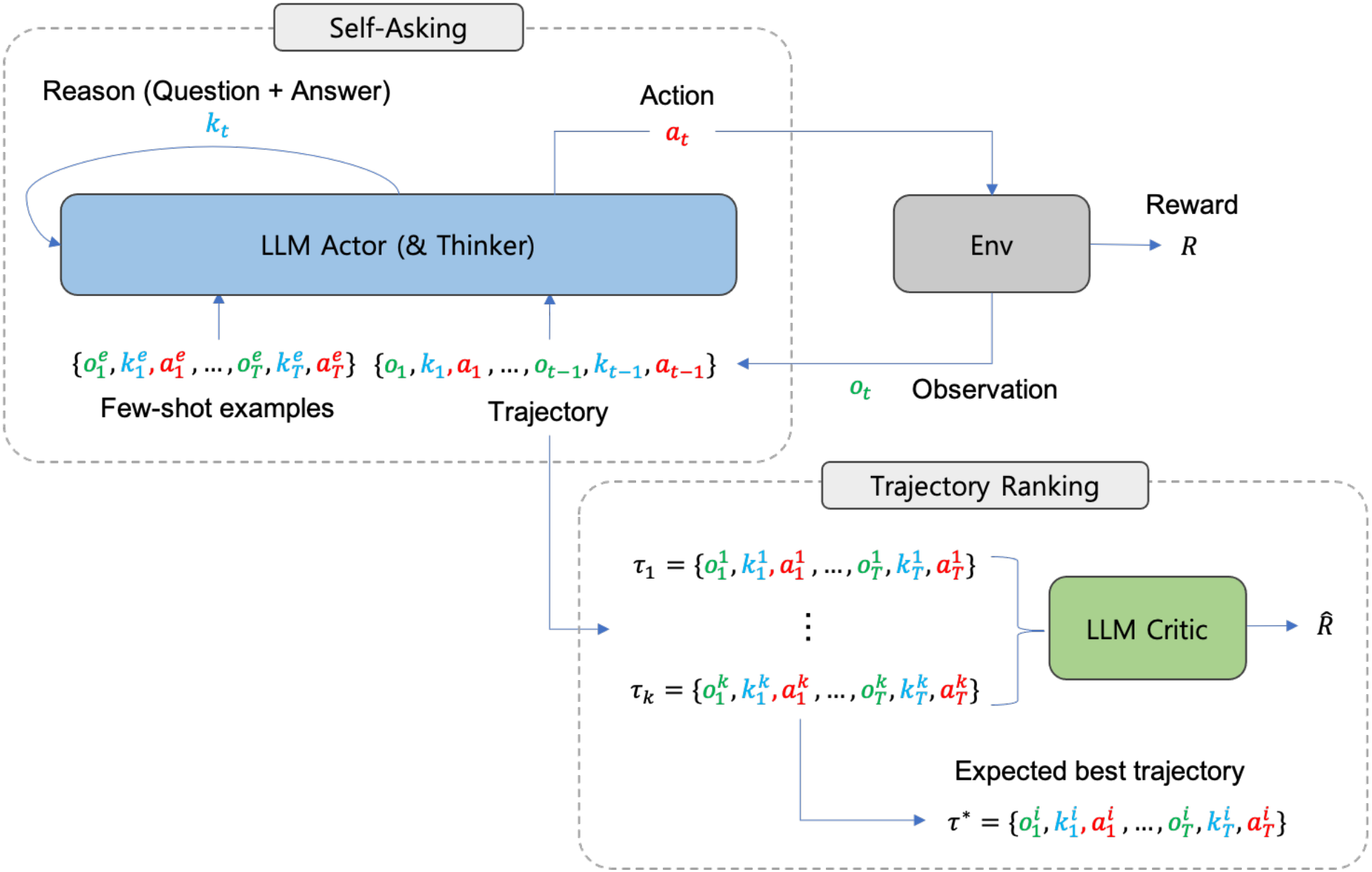

Self-Asking
Reason (Question + Answer)
k_t
Action
a_t
LLM Actor (& Thinker)
Env
Reward
R
$\{o_1^e, k_1^e, a_1^e, \ldots, o_T^e, k_T^e, a_T^e\}$
Few-shot examples
$\{o_1, k_1, a_1, \ldots, o_{t-1}, k_{t-1}, a_{t-1}\}$
Trajectory
o_t
Observation
Trajectory Ranking
$\tau_1 = \{o_1^1, k_1^1, a_1^1, \ldots, o_T^1, k_T^1, a_T^1\}$
$\tau_k = \{o_1^k, k_1^k, a_1^k, \ldots, o_T^k, k_T^k, a_T^k\}$
LLM Critic
$\hat{R}$
Expected best trajectory
$\tau^* = \{o_1^i, k_1^i, a_1^i, \ldots, o_T^i, k_T^i, a_T^i\}$