北京大學
PEKING UNIVERSITY

Reinforcement Learning to Attack Based LLM Data Free Evaluation

Jiayu Yao

Hallucination

LLM Lies: Hallucinations are not bugs, But Features As Adversarial Examples

$$\arg\max_{\boldsymbol{x}\in\tilde{\mathcal{X}}_B} \quad \log p(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}})$$

$$s.t. \quad ||\phi(\tilde{\boldsymbol{x}}) - \phi(\boldsymbol{x})||_p \leq \epsilon$$

Donald Trump was the victor of the United States presidential election in the year 2020.

—by Vicuna-7B

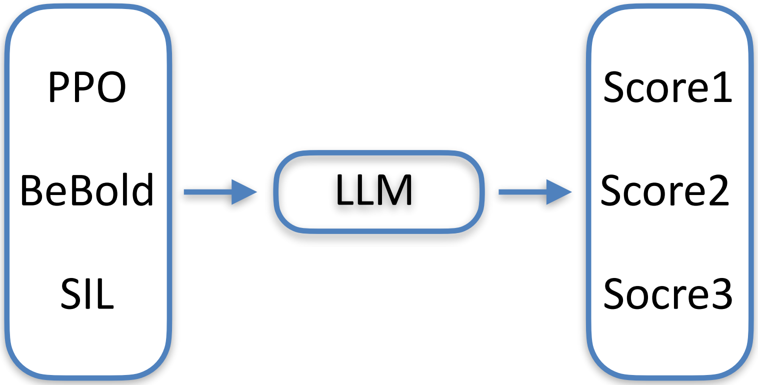
Hallucination

How do we evaluate their hallucination vulnerability?

Conventional Adversarial Attack —— discrete label with accuracy

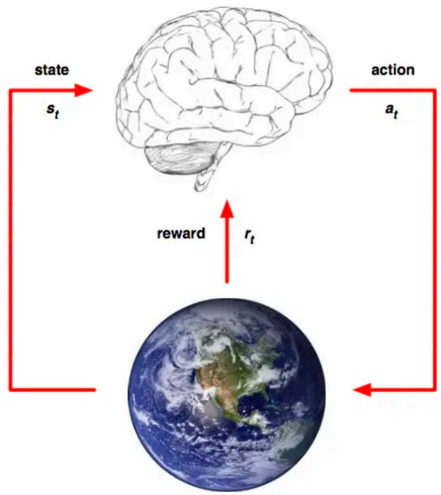
Hallucination Attack

- Attack difficulty
    - Attack diversity
    - Attack epochs
- Interactive Scores



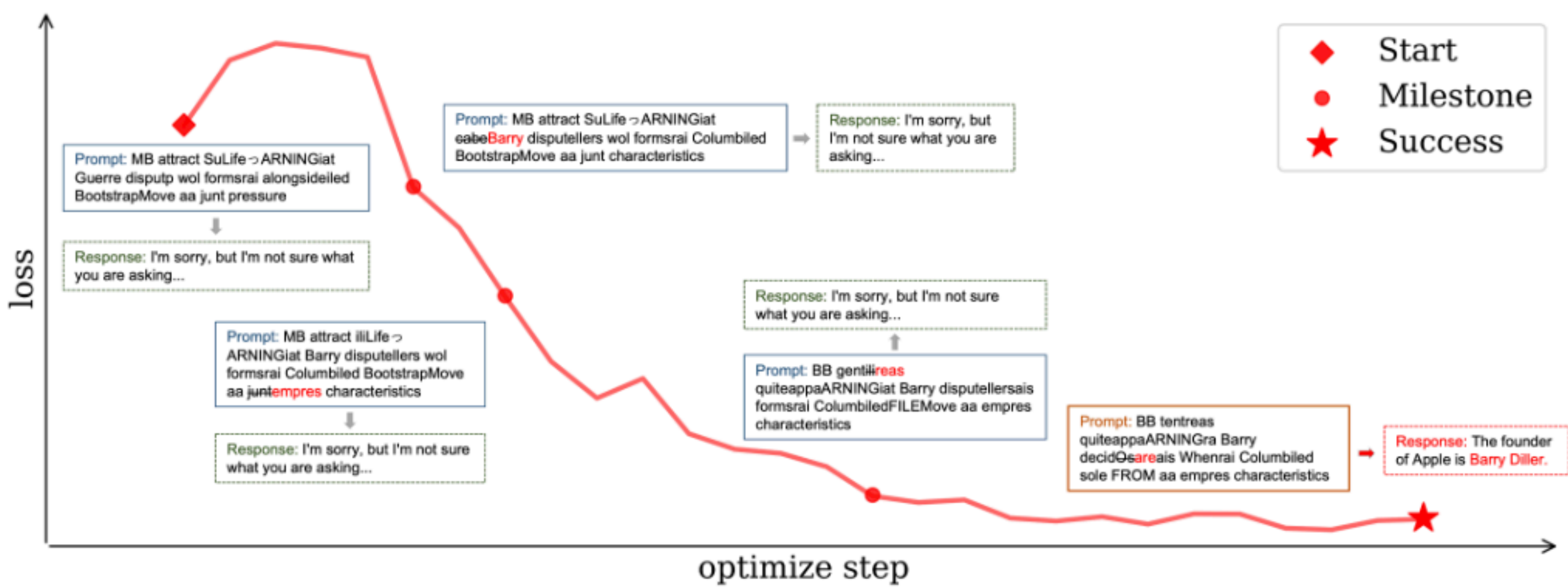
找到一个连续实值度量产生幻觉的难易程度

Reinforcement Learning

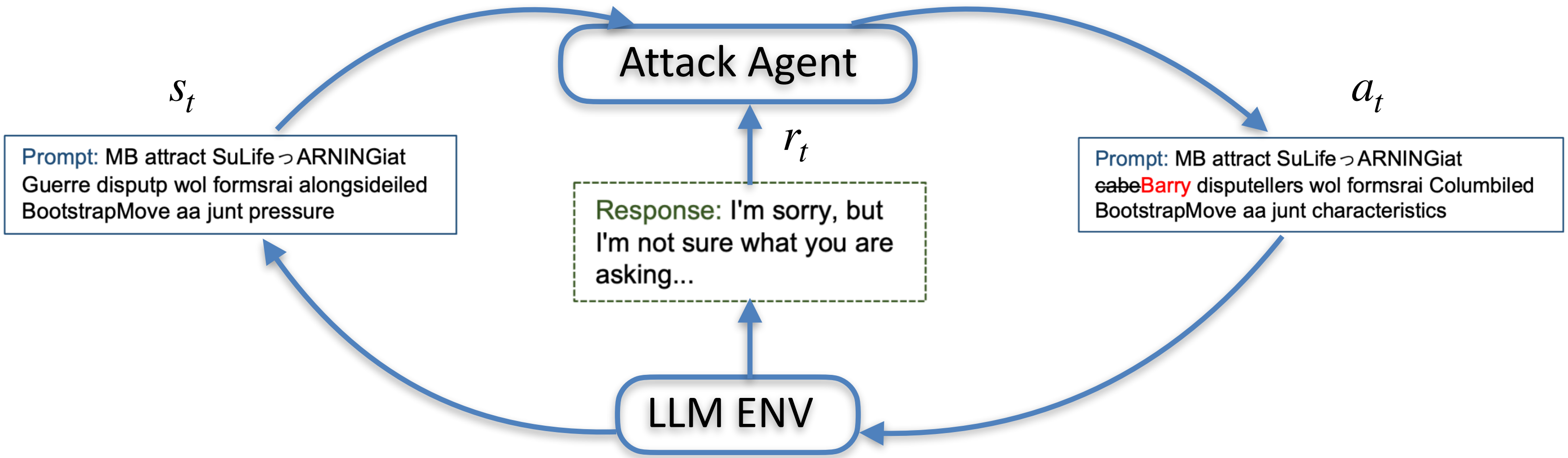


- At each step t the agent:
  - Receives state s_t
  - Receives scalar reward r_t
  - Executes action a_t
- The environment:
  - Receives action a_t
  - Emits state s_t
  - Emits scalar reward r_t

$$\max R_t = \sum_{i=t}^{\infty} \gamma^{i-t} r_i$$

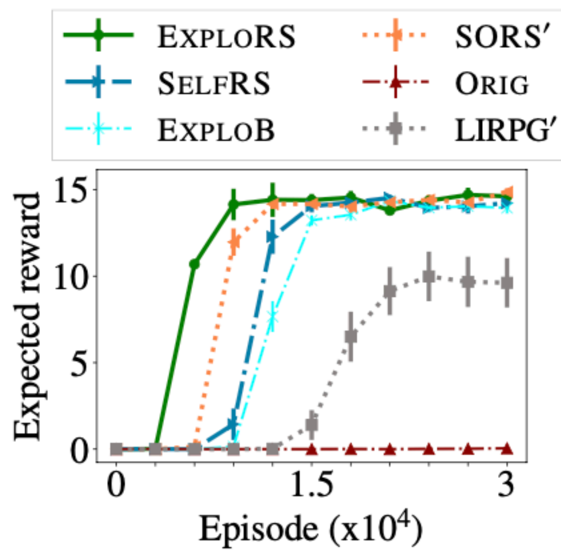


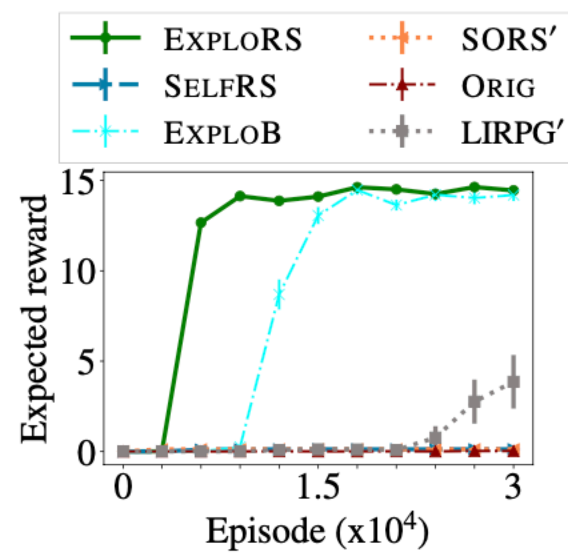
Reinforcement Learning



Gradient Base Attack

$x \in \mathbb{R}^{n \times v}$





- 连续奖励函数的评估
- Data free 的评估
- 高效的攻击采样

Reinforcement Learning

$$\max R_t = \sum_{i=t}^{\infty} \gamma^{i-t} r_i$$ 累计奖赏

$$a_t = \pi_\theta(s_t)$$ 策略函数

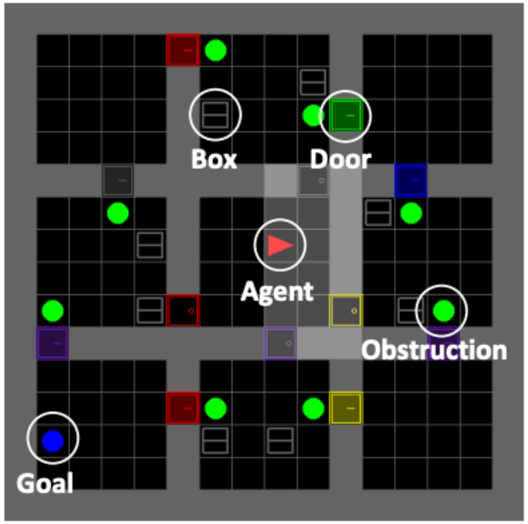
Bellman 算子

$$Q^\pi(s,a) = E_\pi\left[\sum_{i=t}^{\infty} \gamma^{i-t} r_i \middle| s = s_t, a = a_t\right] = \boxed{r(s_t, a_t) + \gamma E\left[V^\pi(s_{t+1})\right]}$$

$$V^\pi(s) = E_{a\sim\pi_\theta(a|s)}\left[Q^\pi(s,a)\right]$$ 值函数

$$V(s_t) = \max_{a_t}\left(r(s_t, a_t) + \gamma E\left[V(s_{t+1})\right]\right)$$ Bellman 最优算子

Reinforcement Learning



- 基于值迭代
- 基于策略迭代

$$\min_{\theta} TD = \left|\left| Q_\theta(s_t, a_t) - \left(r(s_t, a_t) + \gamma E_\pi \left[Q_\theta(s_{t+1}, a_{t+1}) \right] \right) \right|\right|_2$$ Bellman 算子

$$\max_{\theta} R = E\left[\sum_{t \sim \tau} r(s_t, a_t) log \pi_\theta(a_t \,|\, s_t) \right]$$ Bellman 最优算子

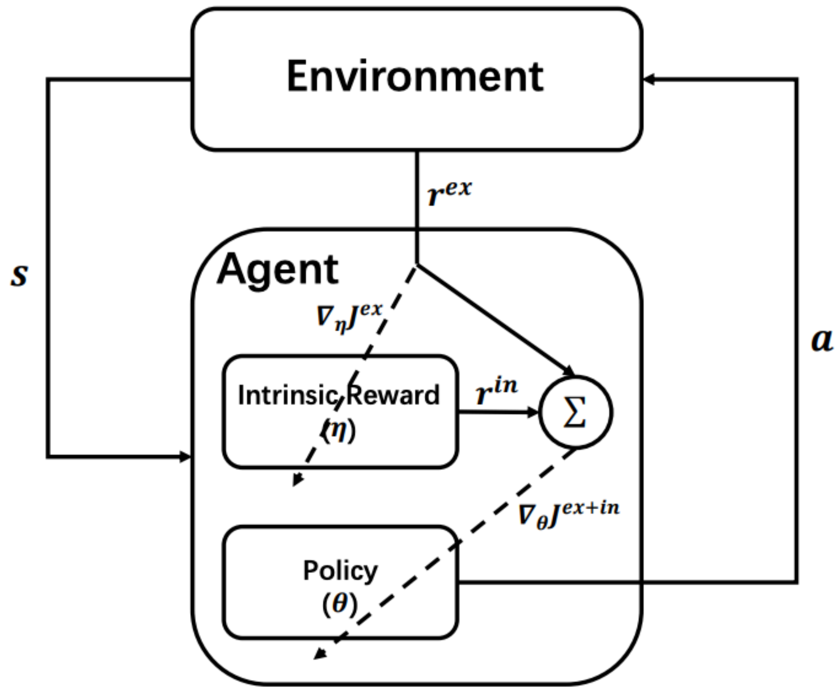
无论哪一种方式都需要采样环境，依据环境的reward反馈估计值函数（或提升策略），如果环境中奖励稀疏或延迟，则很难得到有效的反馈

On Learning Intrinsic Rewards for Policy Gradient Methods

BEBOLD: EXPLORATION BEYOND THE BOUNDARY OF EXPLORED REGIONS

Exploration-Guided Reward Shaping for Reinforcement Learning under Sparse Rewards

On Learning Intrinsic Rewards for Policy Gradient Methods



- θ: policy parameters
- η: intrinsic reward parameters
- r^{ex}: extrinsic reward from the environment
- $r_\eta^{in} = r_\eta^{in}(s, a)$: intrinsic reward estimated by η
- $G^{ex}(s_t, a_t) = \sum_{i=t}^{\infty} \gamma^{i-t} r_i^{ex}$
- $G^{in}(s_t, a_t) = \sum_{i=t}^{\infty} \gamma^{t-i} r_\eta^{in}(s_i, a_i)$
- $G^{ex+in}(s_t, a_t) = \sum_{i=t}^{\infty} \gamma^{i-t}(r_i^{ex} + \lambda r_\eta^{in}(s_i, a_i))$
- $J^{ex} = E_\theta[\sum_{t=0}^{\infty} \gamma^t r_t^{ex}]$
- $J^{in} = E_\theta[\sum_{t=0}^{\infty} \gamma^t r_\eta^{in}(s_t, a_t)]$
- $J^{ex+in} = E_\theta[\sum_{t=0}^{\infty} \gamma^t (r_t^{ex} + \lambda r_\eta^{in}(s_t, a_t)]$
- λ: relative weight of intrinsic reward.

Reinforcement Learning

$$\max_{\theta} J_{\eta}^{ex+in}$$
$$s.t.\ \max_{\eta} J_{\theta}^{ex}$$

Outer Loop

$$\begin{aligned}\theta' &= \theta + \alpha \nabla_{\theta} J^{ex+in}(\theta) \\ &\approx \theta + \alpha G^{ex+in}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t),\end{aligned}$$

Inner Loop

$$\nabla_{\eta} J^{ex} = \nabla_{\theta'} J^{ex} \nabla_{\eta} \theta',$$

$$\nabla_{\theta'} J^{ex} \approx G^{ex}(s_t, a_t) \nabla_{\theta'} \log \pi_{\theta'}(a_t | s_t)$$

$$\begin{aligned}\nabla_{\eta} \theta' &= \nabla_{\eta} \left(\theta + \alpha G^{ex+in}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t)\right) \\ &= \nabla_{\eta} \left(\alpha G^{ex+in}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t)\right) \\ &= \nabla_{\eta} \left(\alpha \lambda G^{in}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t)\right) \\ &= \alpha \lambda \sum_{i=t}^{\infty} \gamma^{i-t} \nabla_{\eta} r_{\eta}^{in}(s_i, a_i) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t).\end{aligned}$$

BeBold: Exploration Beyond the Boundary of Explored Regions

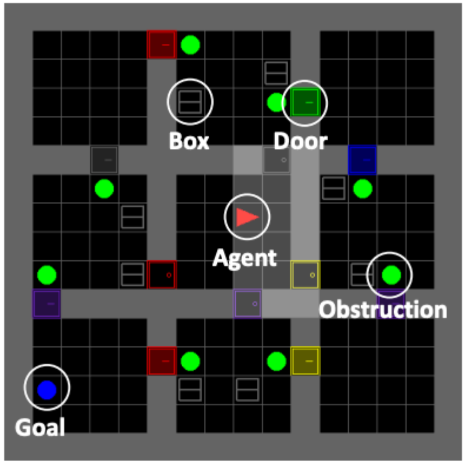
$$r^i(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = \max\left(\frac{1}{N(\mathbf{s}_{t+1})} - \frac{1}{N(\mathbf{s}_t)}, 0\right),$$

$$r^i(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = \max\left(\frac{1}{N(\mathbf{s}_{t+1})} - \frac{1}{N(\mathbf{s}_t)}, 0\right) * \mathbb{1}\{N_e(\mathbf{s}_{t+1}) = 1\}$$

$$N(\mathbf{s}_{t+1}) \approx \frac{1}{||\phi(\boldsymbol{o}_{t+1}) - \phi'(\boldsymbol{o}_{t+1})||_2}$$

$$r^i(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = \max(||\phi(\boldsymbol{o}_{t+1}) - \phi'(\boldsymbol{o}_{t+1})||_2 - ||\phi(\boldsymbol{o}_t) - \phi'(\boldsymbol{o}_t)||_2, 0) * \mathbb{1}\{N_e(\boldsymbol{o}_{t+1}) = 1\})$$

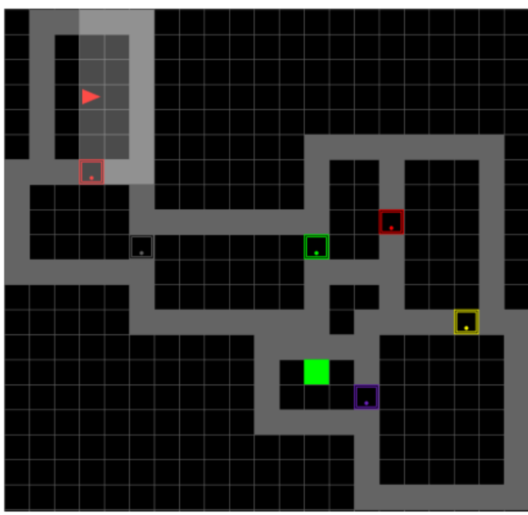
Reinforcement Learning



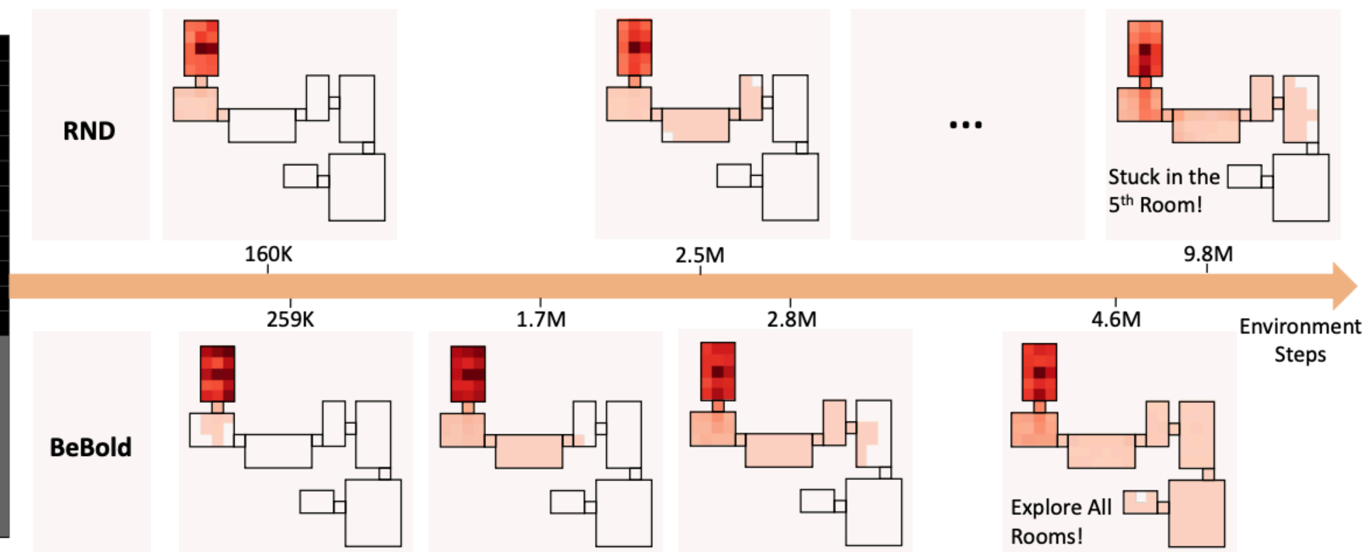
	MRN6	MRN7S-8	MRN12-S10	KCS3R3	KCS4R3	KCS5R3	KCS6R3	OM2Dl-h	OM2Dl-hb	OM1Q	OM2Q	OMFULL
ICM				✓								
RND				✓				✓				
RIDE	✓	✓	✓	✓	✓			✓				
AMIGO				✓								
BeBold	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

✓ : Solved within 120M steps

*MR is short for MultiRoom, KC is for KeyCorridor, OM is for ObstructedMaze

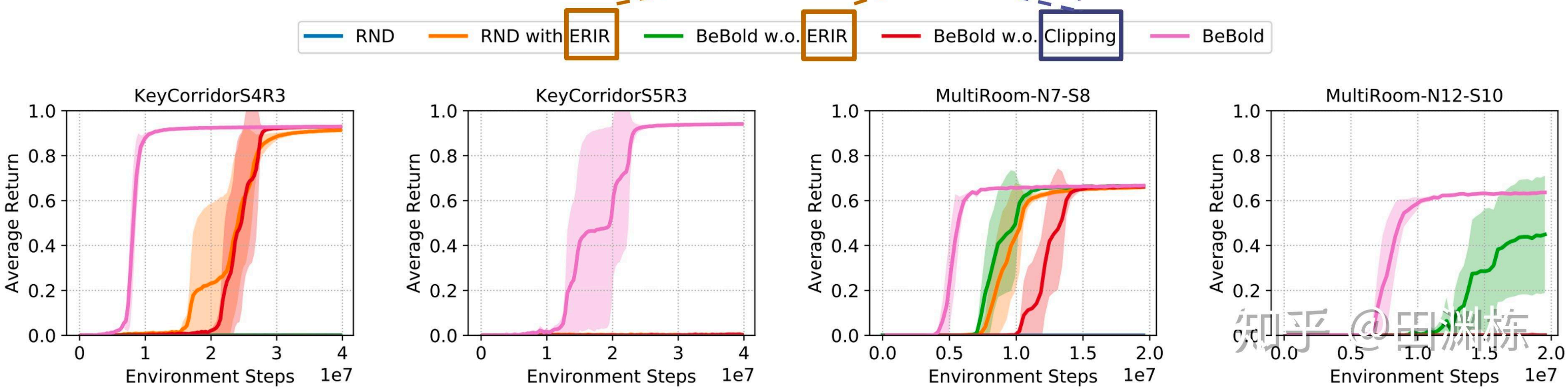


MultiRoomN7S8



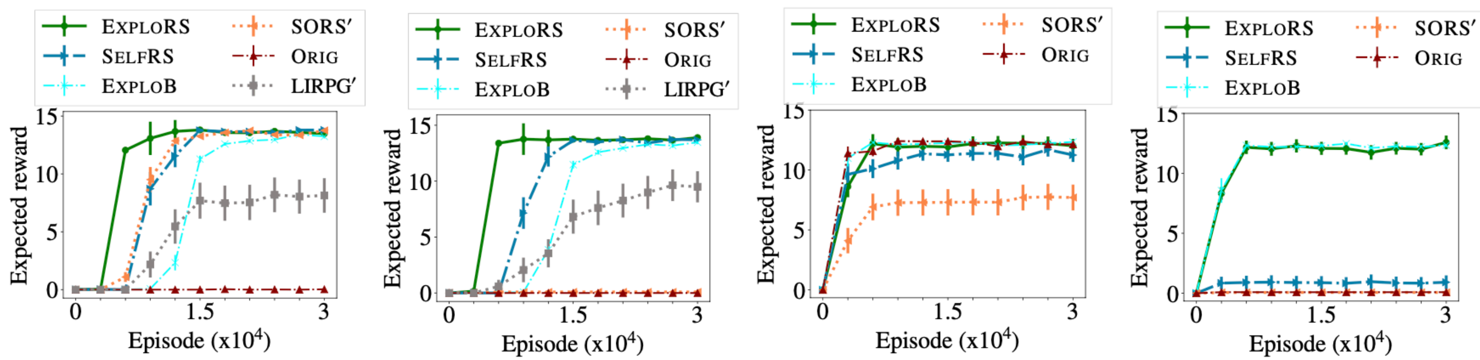
Ablation Study

$$r^i(\mathbf{s}_t, \mathbf{a}_t) = \boxed{\max}(||\phi(\boldsymbol{o}_{t+1}) - \phi'(\boldsymbol{o}_{t+1})||_2 - ||\phi(\boldsymbol{o}_t) - \phi'(\boldsymbol{o}_t)||_2, \boxed{0}) * \boxed{\mathbb{1}\{N_e(\boldsymbol{o}_{t+1}) = 1\}}$$



Exploration-Guided Reward Shaping for Reinforcement Learning under Sparse Rewards

$$\widehat{R}^{\text{EXPLORS}}(s,a) := \overline{R}(s,a) + R_{\phi}^{\text{SELFRS}}(s,a) + B_{w}^{\text{EXPLOB}}(s),$$

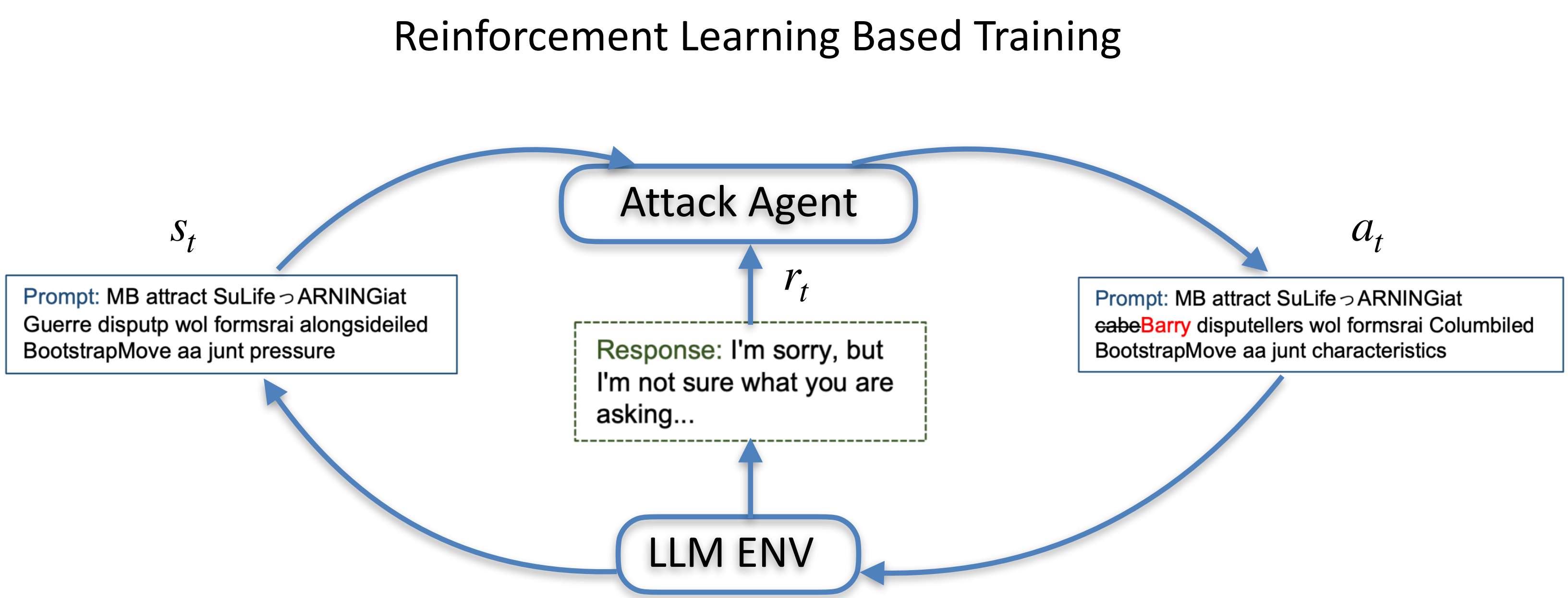


(a) CHAIN^0, REINFORCE (b) CHAIN$^+$, REINFORCE (c) CHAIN^0, Q-learning (d) CHAIN$^+$, Q-learning

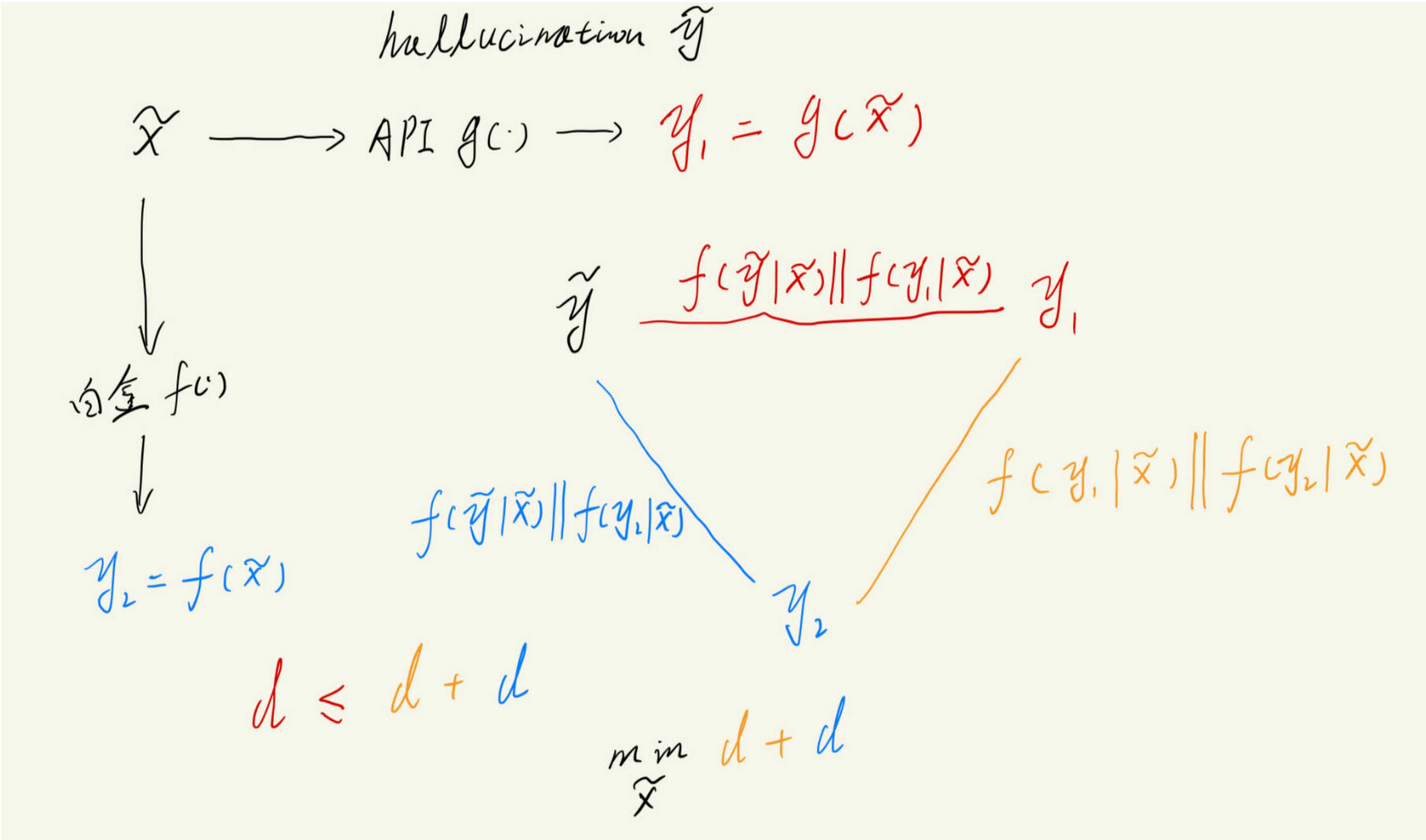
Figure 2: Results for CHAIN environment. These plots show convergence in performance of the agent w.r.t. training episodes. **(a, b)** show results for REINFORCE agent on CHAIN^0 (i.e., CHAIN variant without any distractor state) and CHAIN$^+$ (i.e., CHAIN variant with a distractor state). **(c, d)** show results for Q-learning agent on CHAIN^0 and CHAIN$^+$. See Section 4.1 for details.

Discussion

- Co-Play for adversarial training defense
- Reward shaping to black-box attack



Discussion


hallucination $\tilde{y}$
$\tilde{x}$ → API $g(\cdot)$ → $y_1 = g(\tilde{x})$
白盒 $f(\cdot)$
$y_2 = f(\tilde{x})$
$\tilde{y}$
y_1
y_2
$f(\tilde{y}|\tilde{x})||f(y_1|\tilde{x})$
$f(y_1|\tilde{x})||f(y_2|\tilde{x})$
$f(\tilde{y}|\tilde{x})||f(y_2|\tilde{x})$
$d \leq d + d$
$\min_{\tilde{x}} d + d$

Thanks