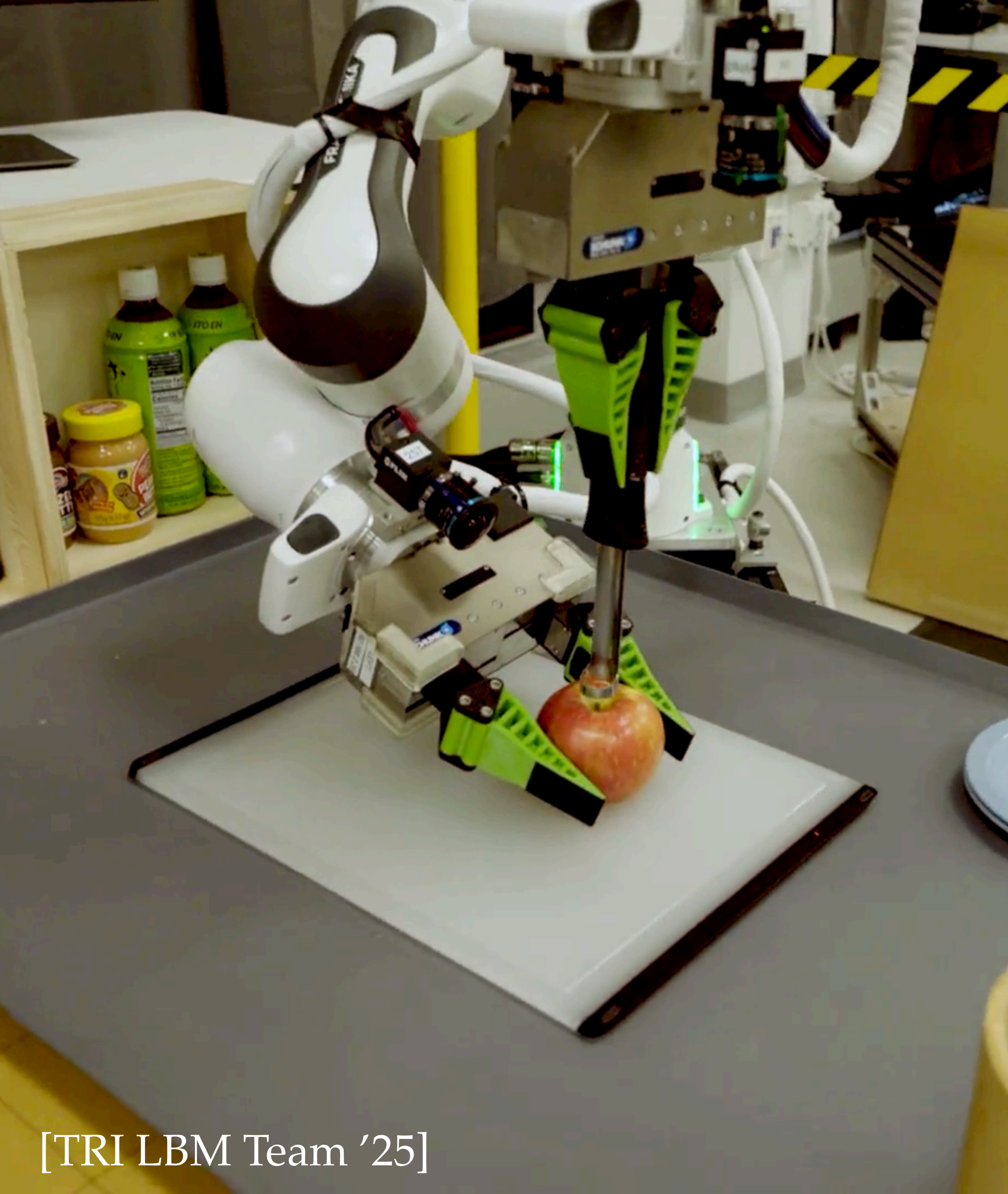


# Efficient Interactive Learning: *Learning More From Less*

Gokul Swamy







[TRI LBM Team '25]



AI Mode ▾ All Images Videos News More ▾



# Hi Gokul, what's on your mind?

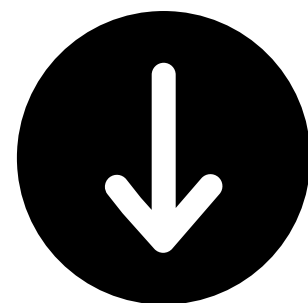
what can you tell me about the diet of the bird mentioned at the beginning of nabokov's novel pale fire?





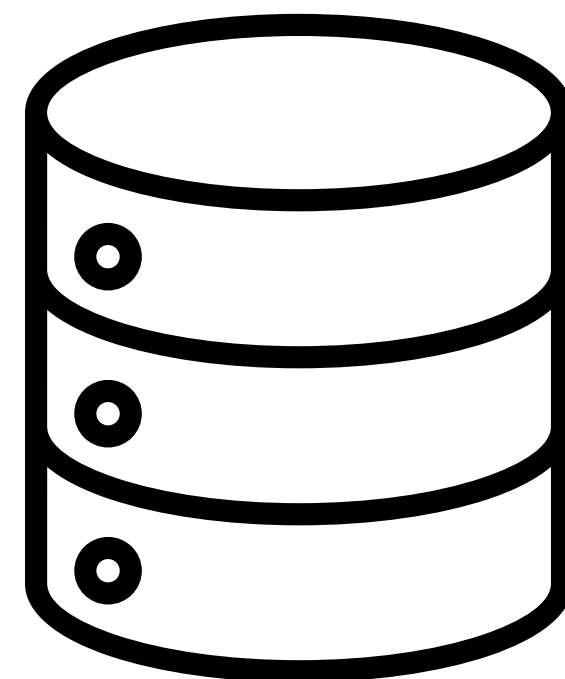
# Supervised Learning

*Robotics:*

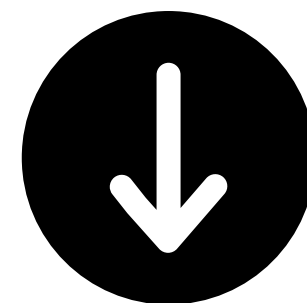


$$\pi(\text{robot arm icon} \mid \text{cup image})$$

*(Fixed Dataset)*



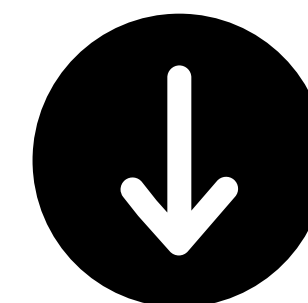
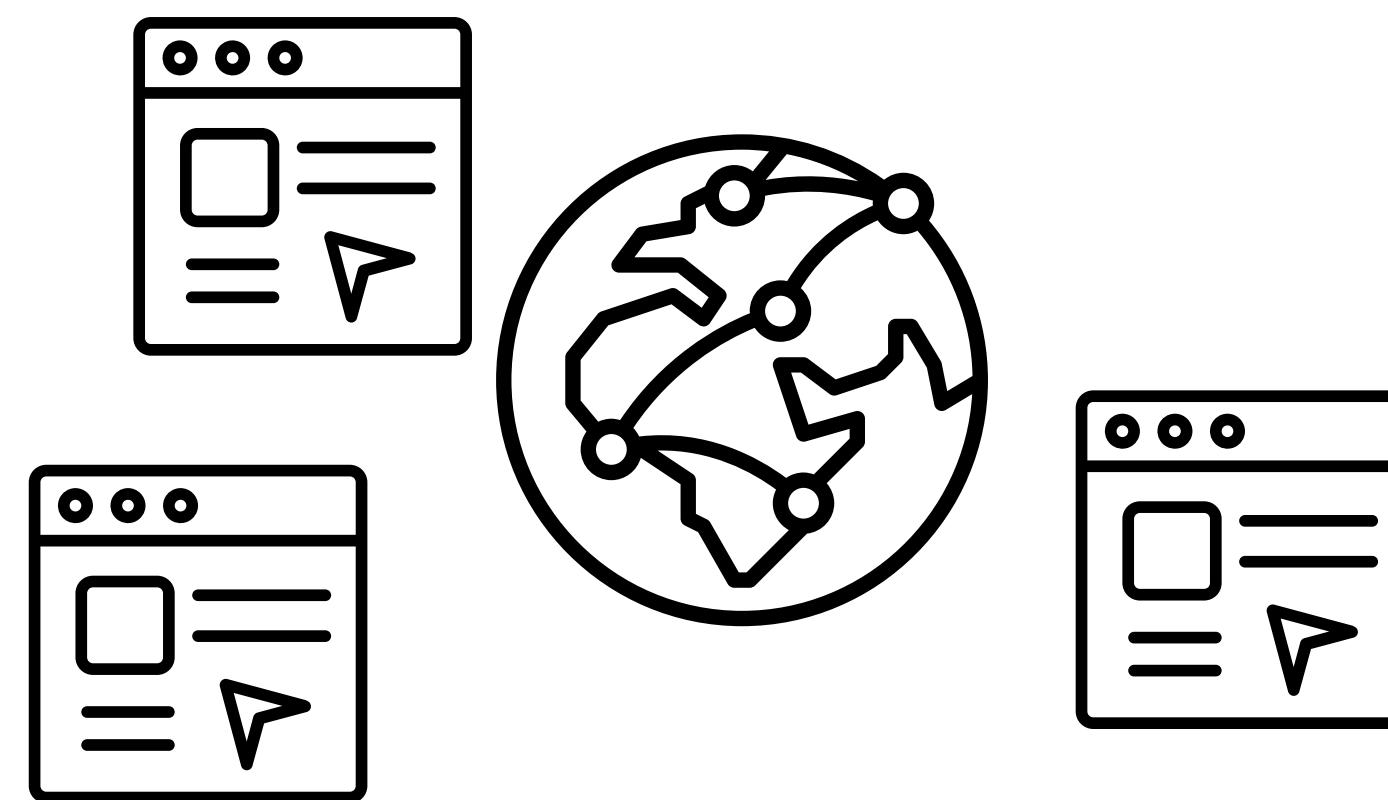
$(s_0, a_0), (s_1, a_1), \dots$



$$\pi(a_t \mid s_t)$$

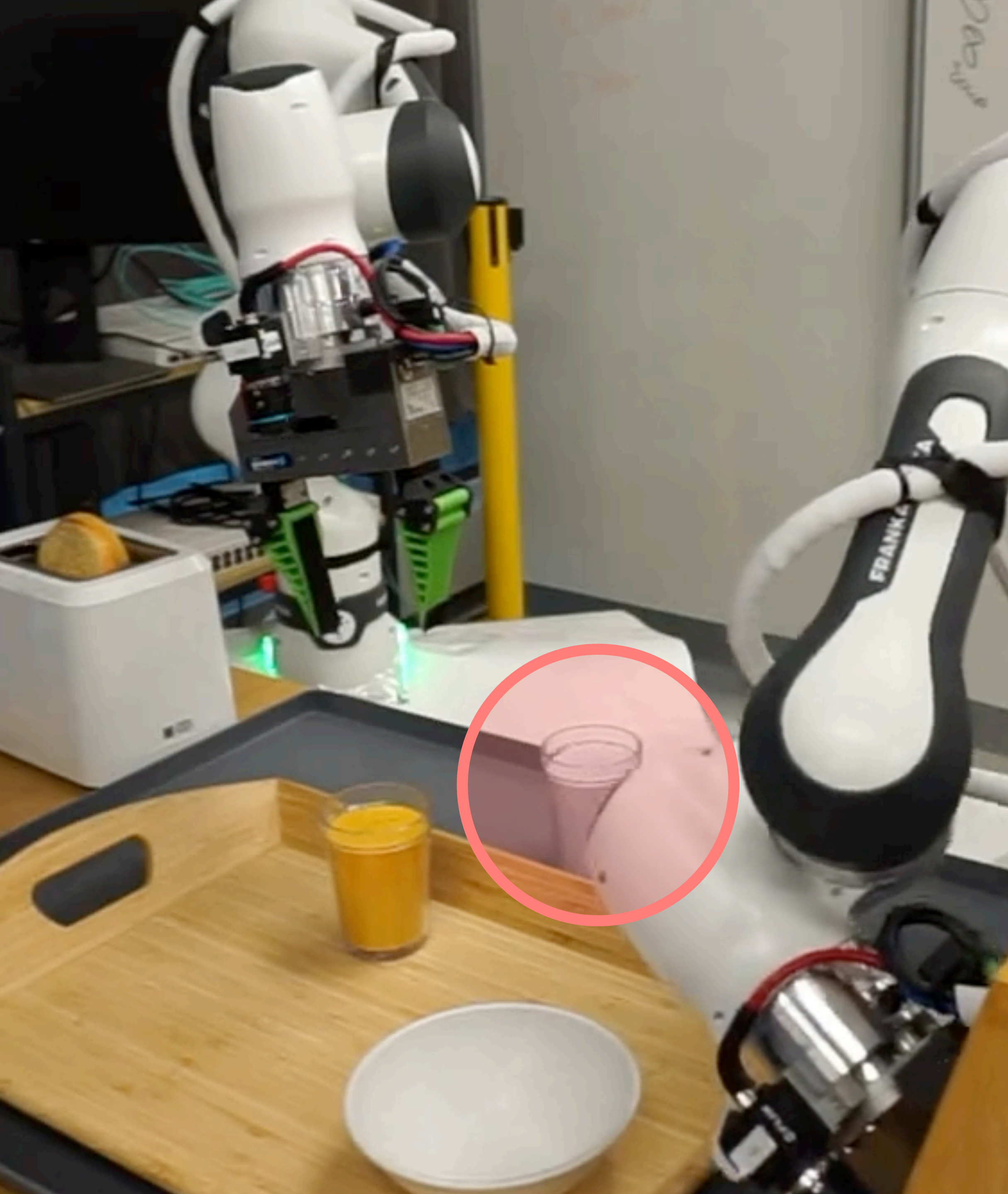
*(Policy)*

*Language Modeling:*



$$\pi(\text{"lo"} \mid \text{"hel"})$$





AI Mode ▾ All Images Videos News More ▾



## Hi Gokul, what's on your mind?

i am thinking of founding a startup to sell berries directly to wild birds. i think it is a niche, untapped market. what do you think?





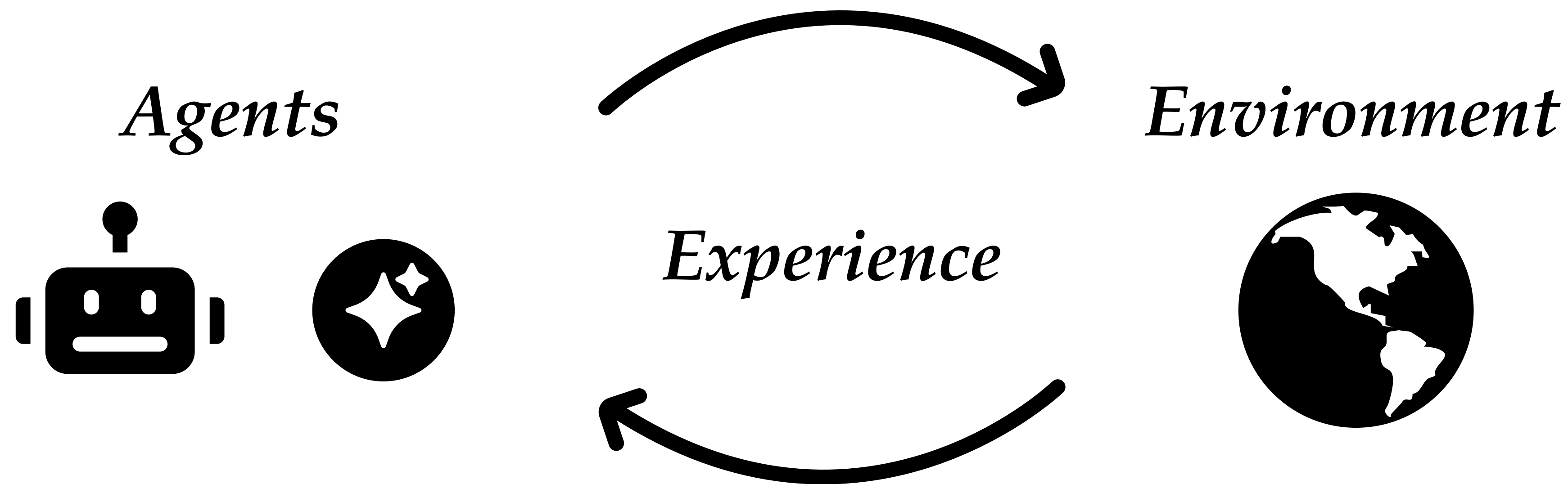


*Even with an **Internet** of data,  
→ supervised learning ←  
wasn't enough.*

*Q: How can we learn more from less data?*



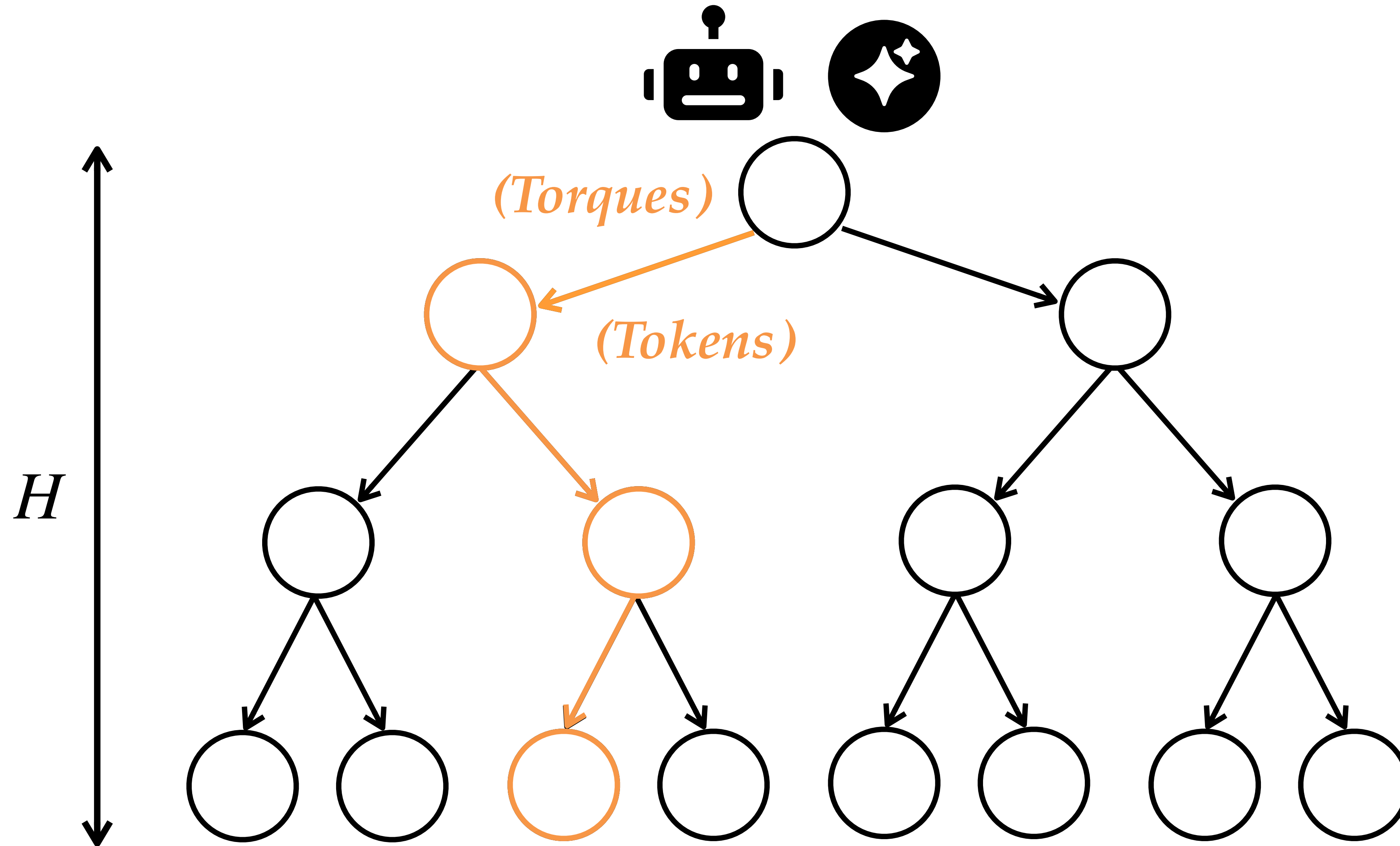
*Q: How can we learn more from less data?*



*A: Reinforcement Learning (RL)*

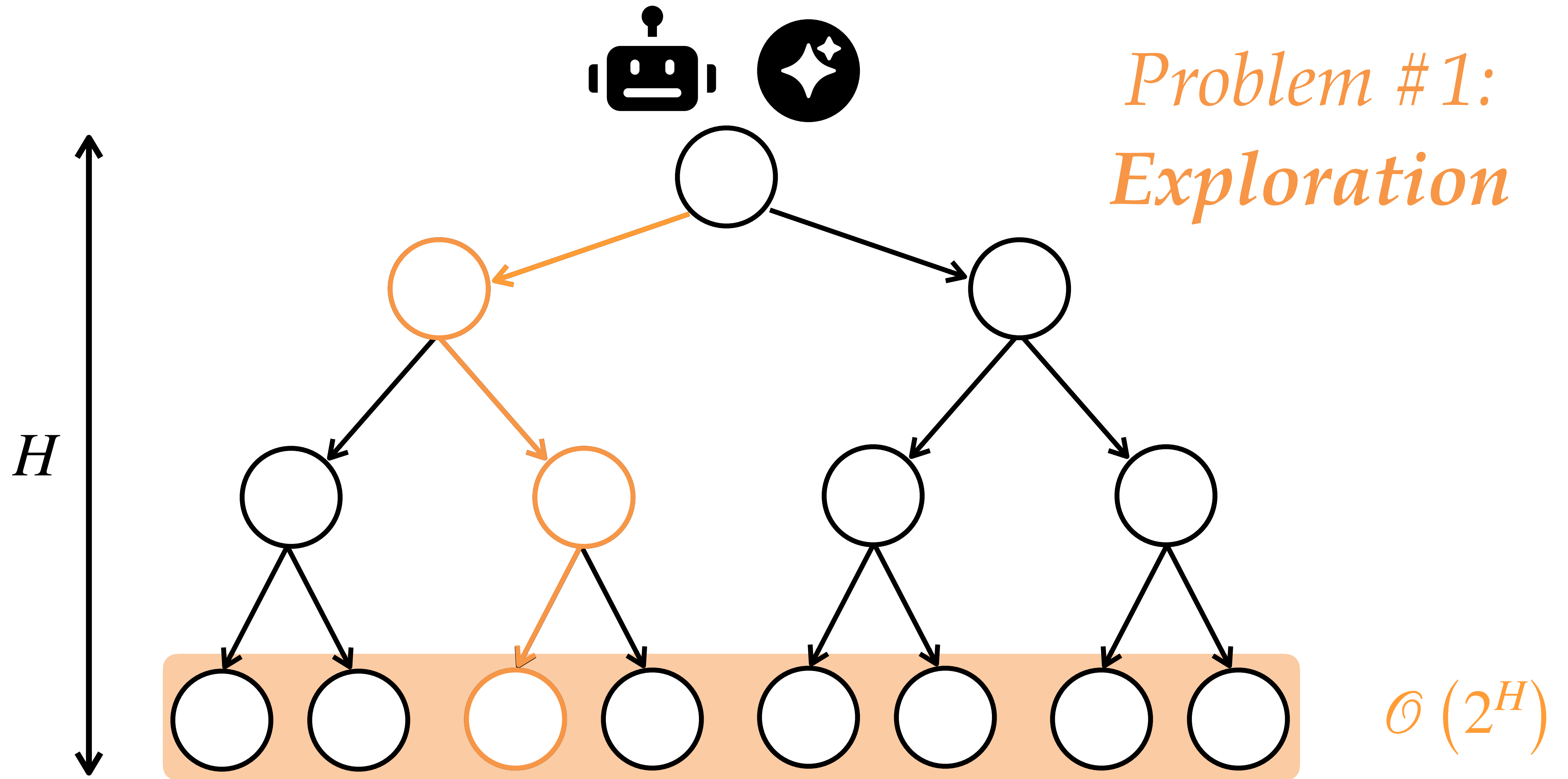
*(i.e., agents collecting their own data)*

# *What makes Reinforcement Learning hard?*

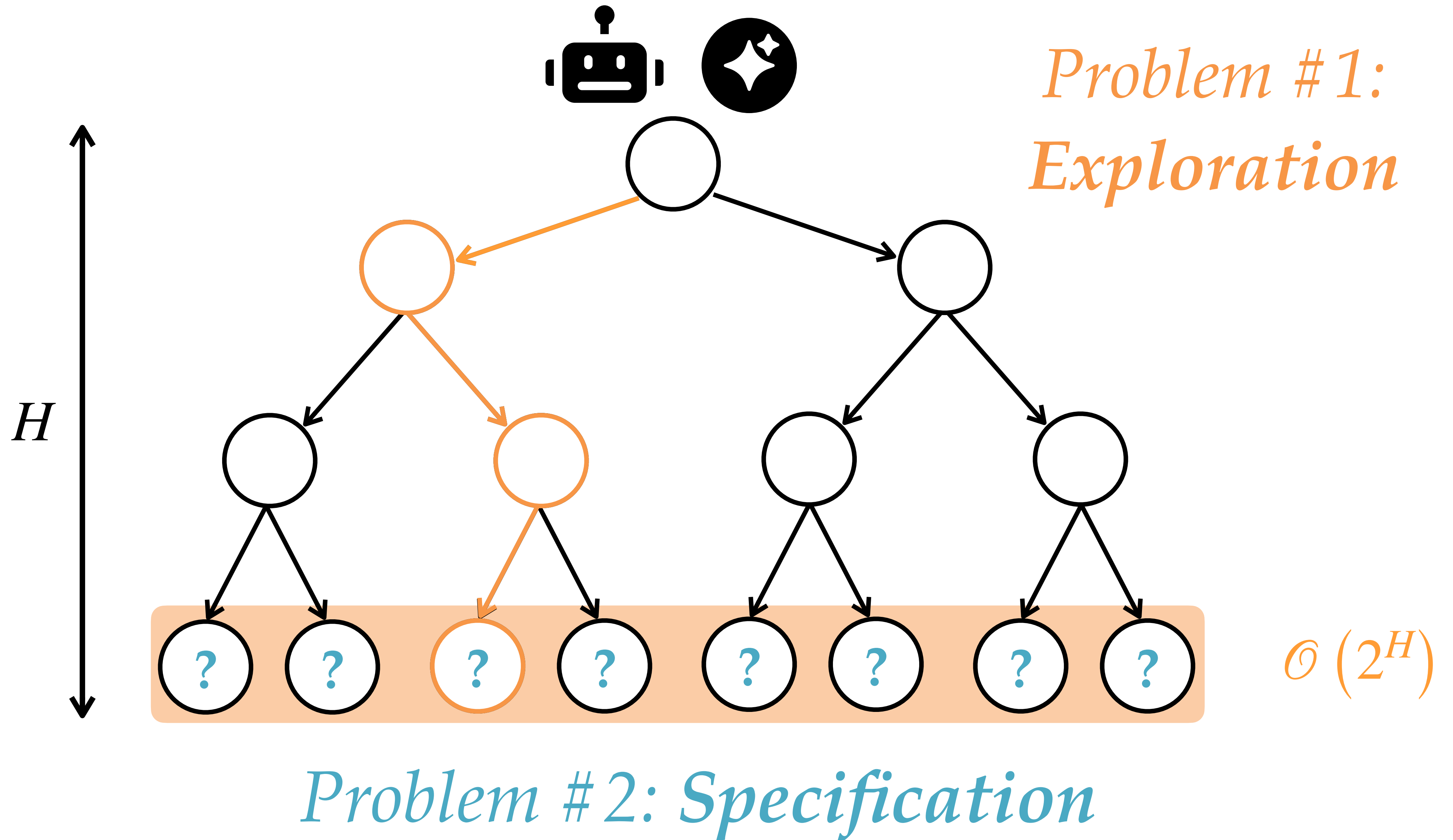




# *What makes Reinforcement Learning hard?*



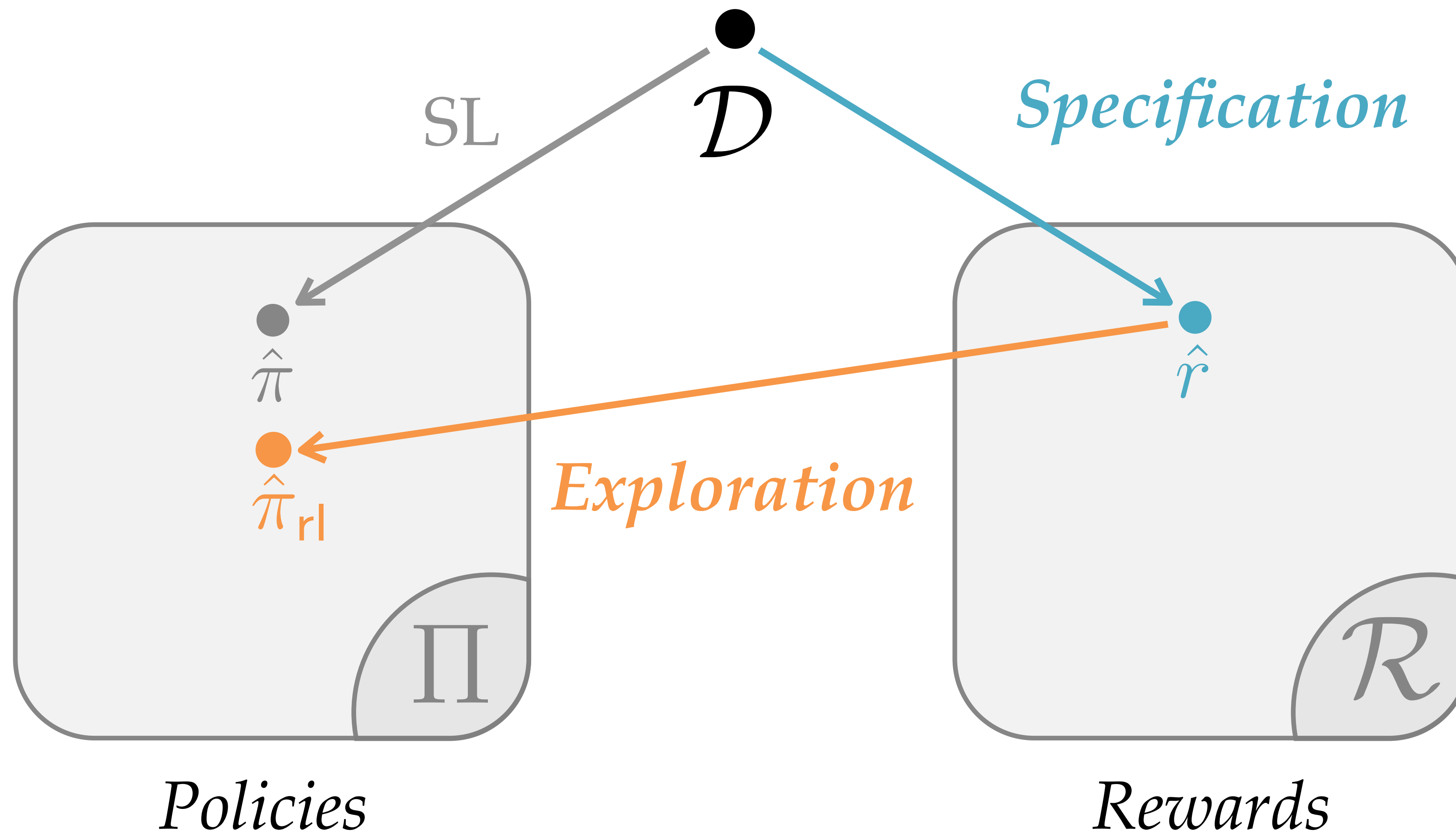
# *What makes Reinforcement Learning hard?*





*My Work: learning to make good decisions  
efficiently, even when “good” is hard to specify*

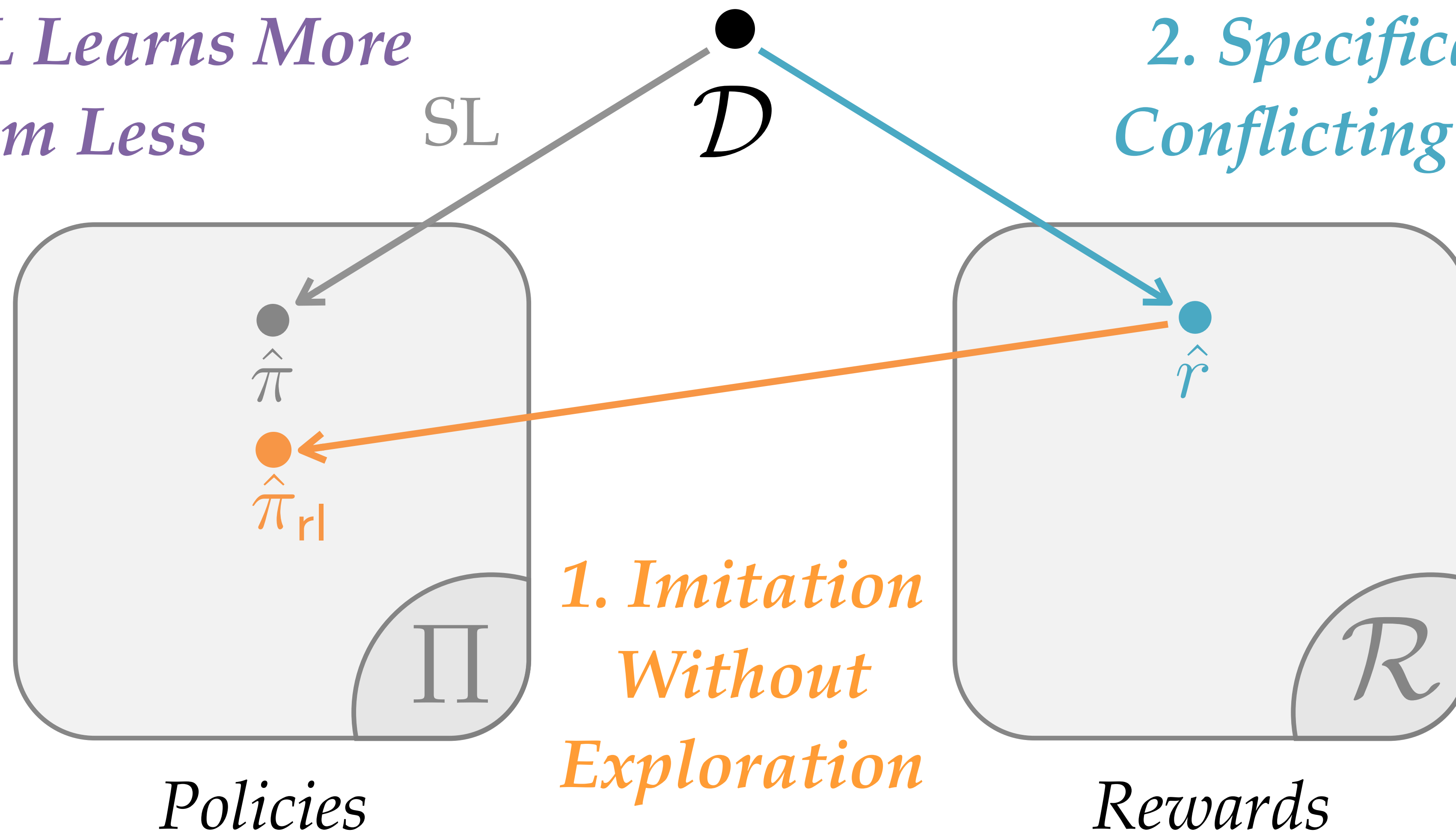
# *My Work: **Efficient** RL When “Good” Is Hard to **Specify***



*Today:*

*3. Why RL Learns More  
From Less*

*2. Specification from  
Conflicting Preferences*

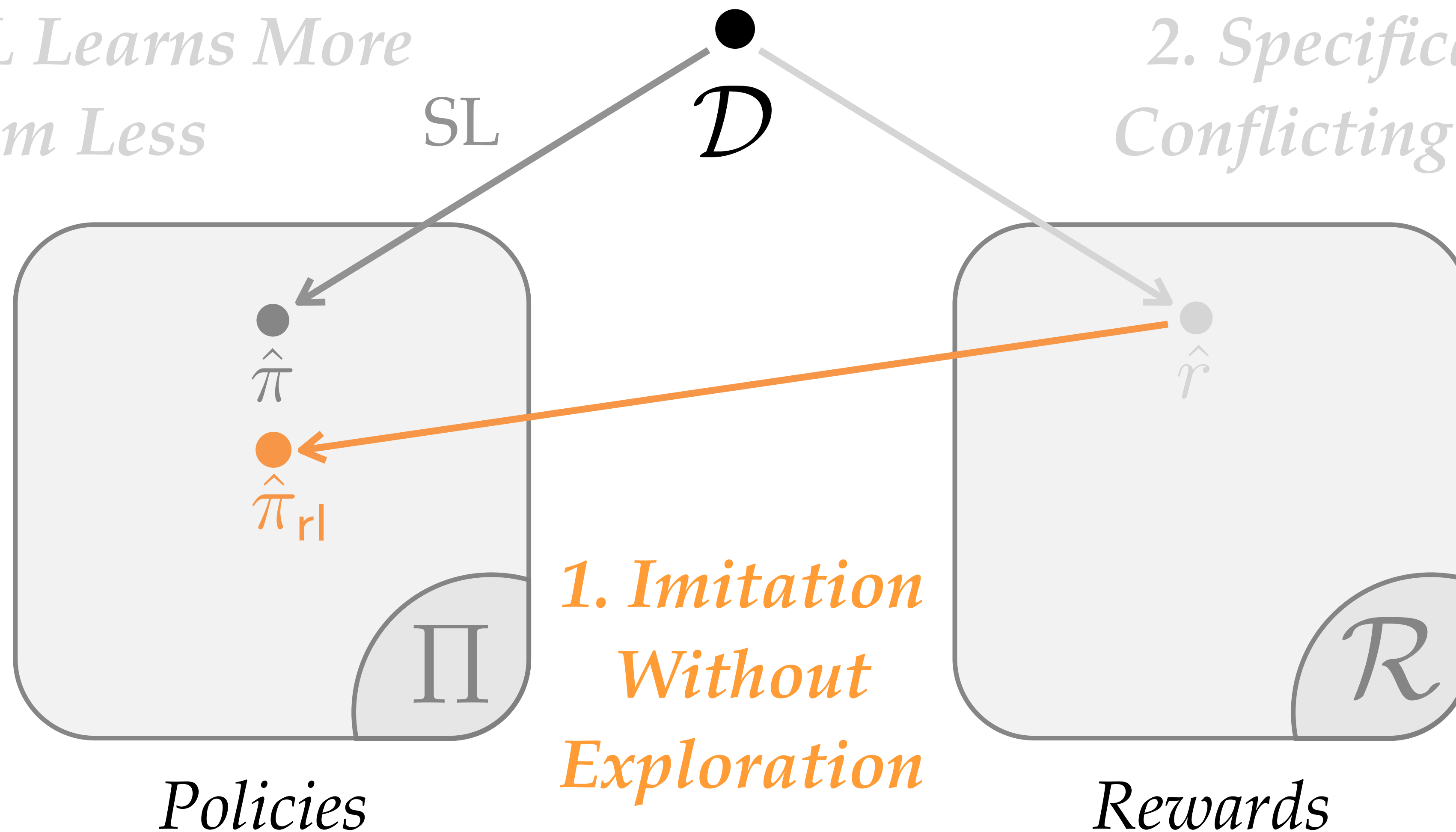




# Today:

3. *Why RL Learns More From Less*

2. *Specification from Conflicting Preferences*



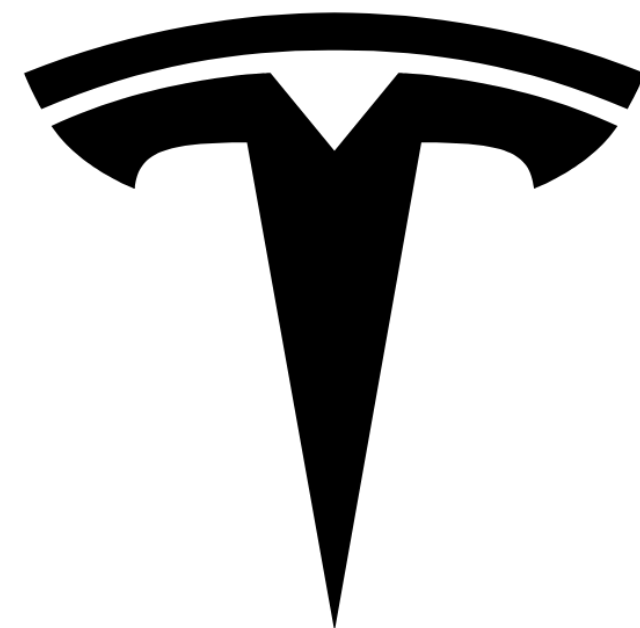
*Theory*  $\longrightarrow$  *Practice*

# *Setting: Imitation Learning*

*Goal: Autonomously Reproduce **Expert Behavior***



$\sim \pi_E$   
(Expert Policy)



*Note: fixed set of demonstrations & no reward information!*

# *Setting: Imitation Learning*

## *Supervised Approach: Behavioral Cloning*



$\sim \pi_E$   
(*Expert Policy*)

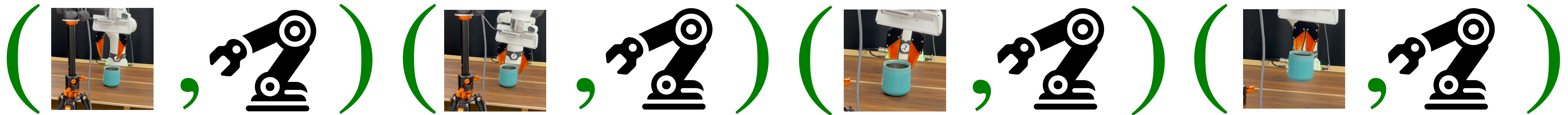


# *Setting: Imitation Learning*

## *Supervised Approach: Behavioral Cloning*



$\sim \pi_E$   
(*Expert Policy*)





# Setting: Imitation Learning

## Supervised Approach: Behavioral Cloning



$\sim \pi_E$   
(Expert Policy)

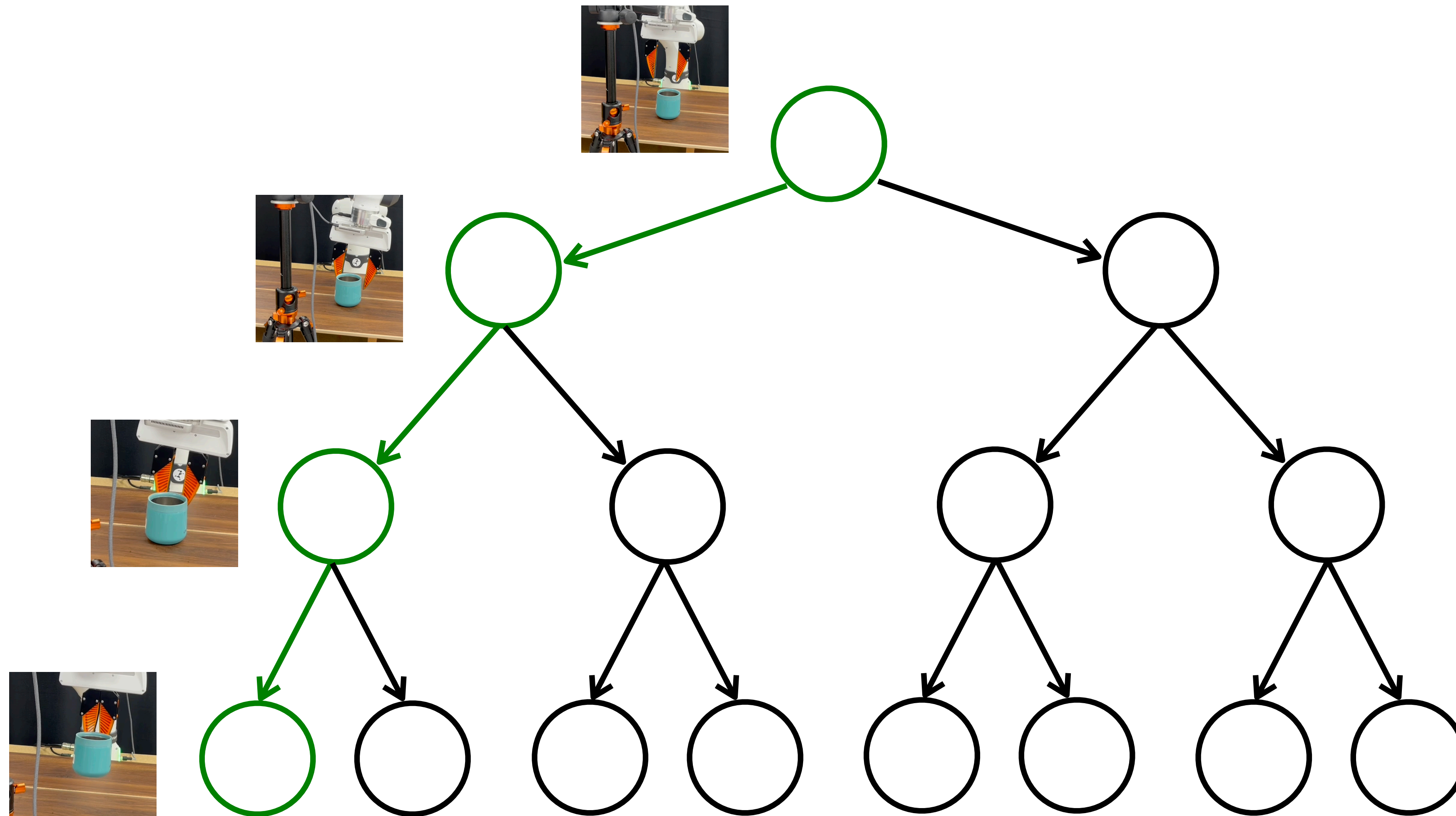


$\pi$   
(Learner Policy)

$$\pi = \arg \max_{\pi \in \Pi} \mathbb{E}_{\xi \sim \pi_E} \left[ \sum_h^H \log \pi(a_h^E | s_h^E) \right]$$

[Pomerleau '89]

# 🔑 *Problem: Supervised Imitation Doesn't Teach Recovery*

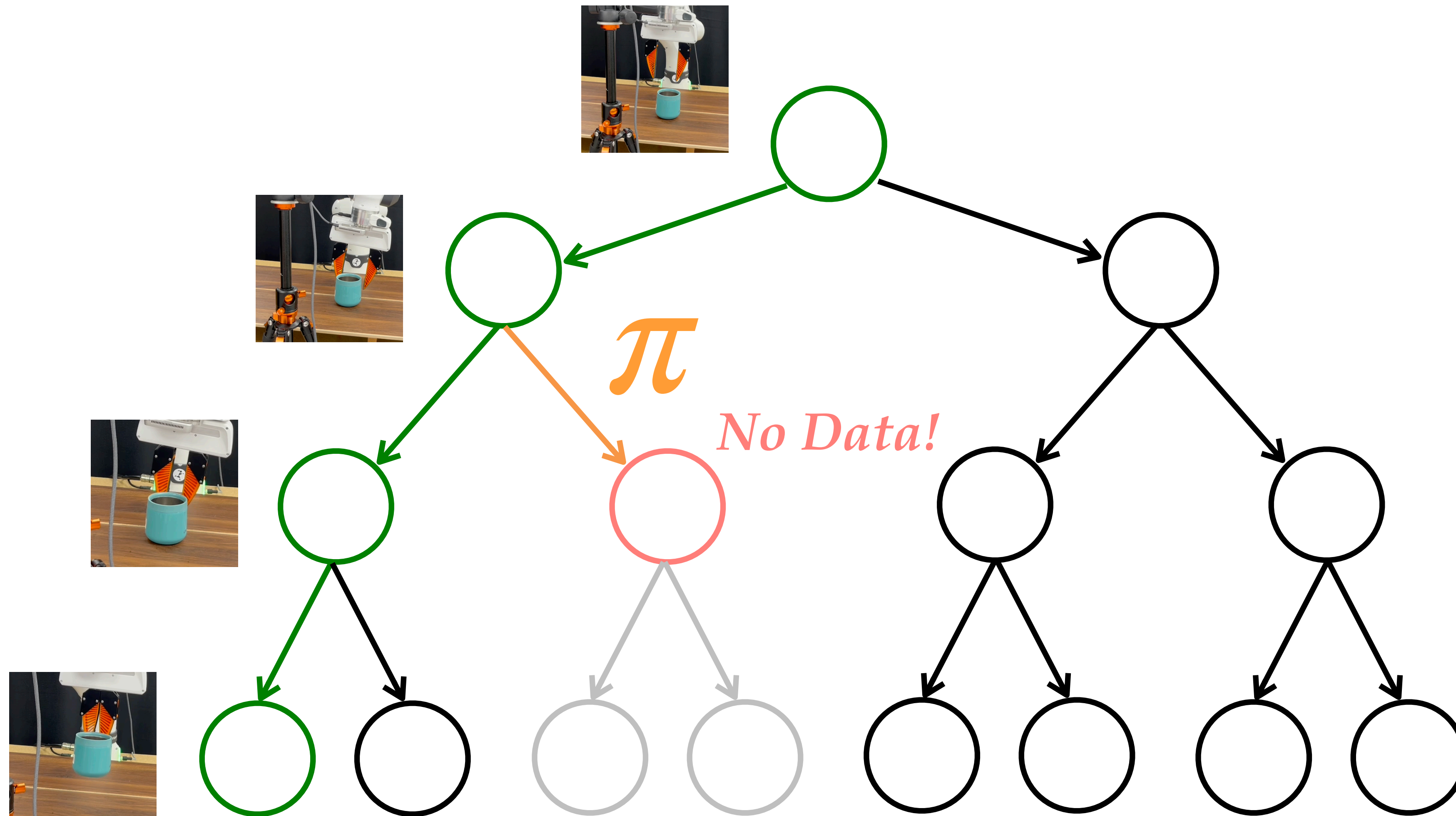


[Ross & Bagnell '12; Swamy et al. '21; '22; '23]

*Key Problem* → *Classical Solution* → *Efficient Algorithm* → *Practical Results*



# 🔑 *Problem: Supervised Imitation Doesn't Teach Recovery*

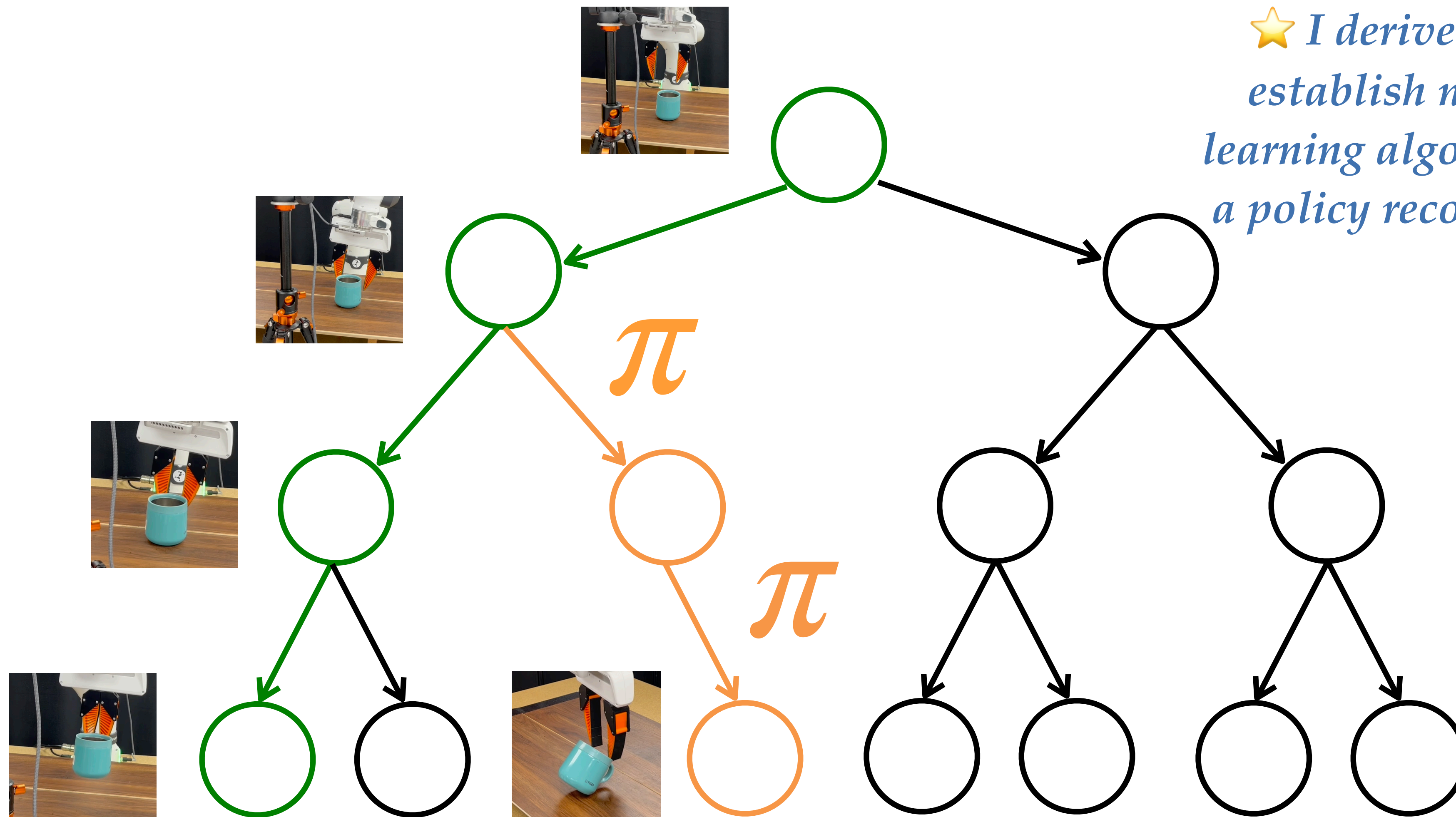


[Ross & Bagnell '12; Swamy et al. '21; '22; '23]

*Key Problem* → *Classical Solution* → *Efficient Algorithm* → *Practical Results*

# 🔑 *Problem: Supervised Imitation Doesn't Teach Recovery*

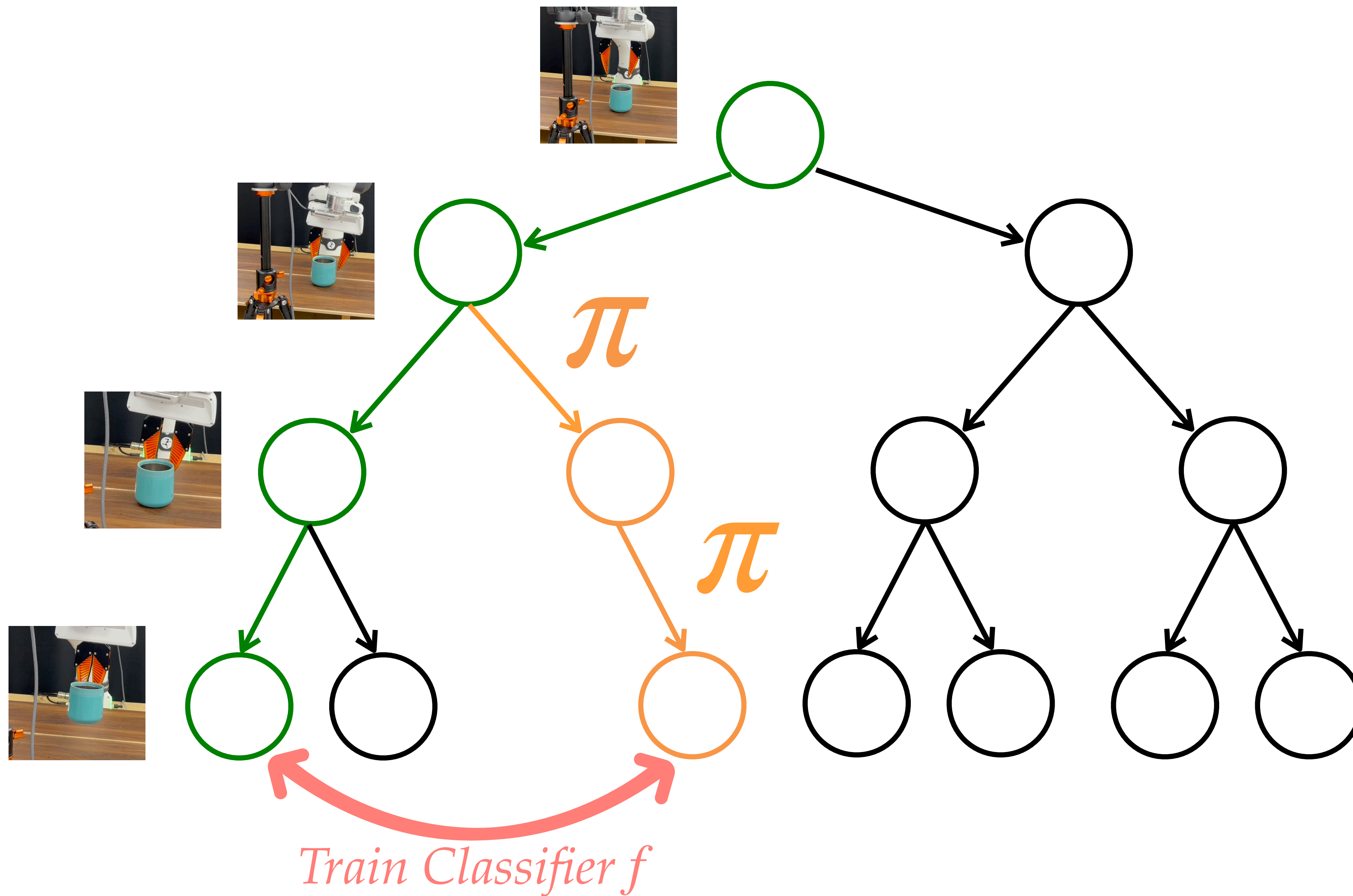
★ *I derived lower bounds that establish no offline, supervised learning algorithm can tell whether a policy recovers from mistakes* ★



[Ross & Bagnell '12; Swamy et al. '21; '22; '23]

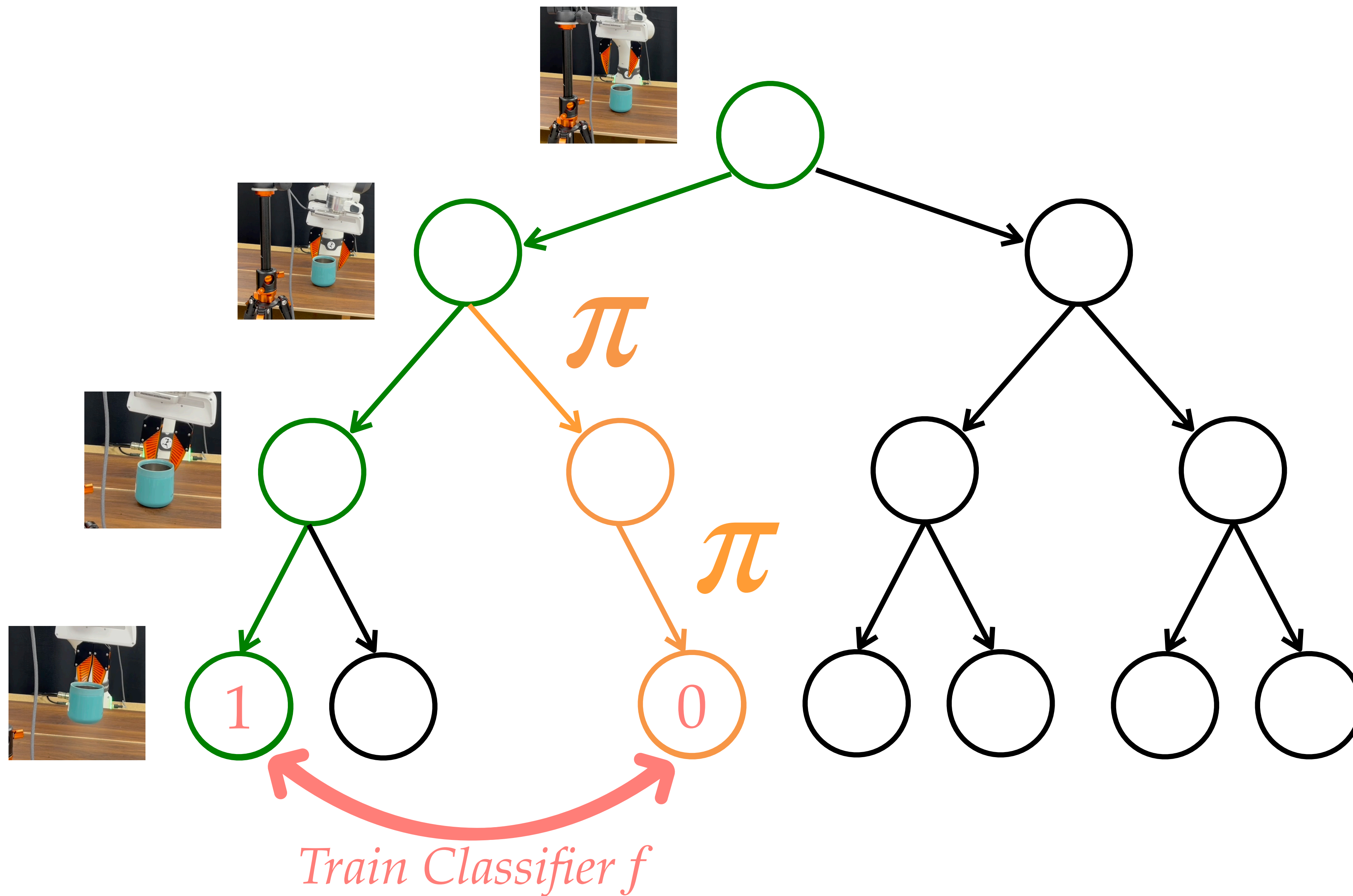
*Key Problem* → *Classical Solution* → *Efficient Algorithm* → *Practical Results*





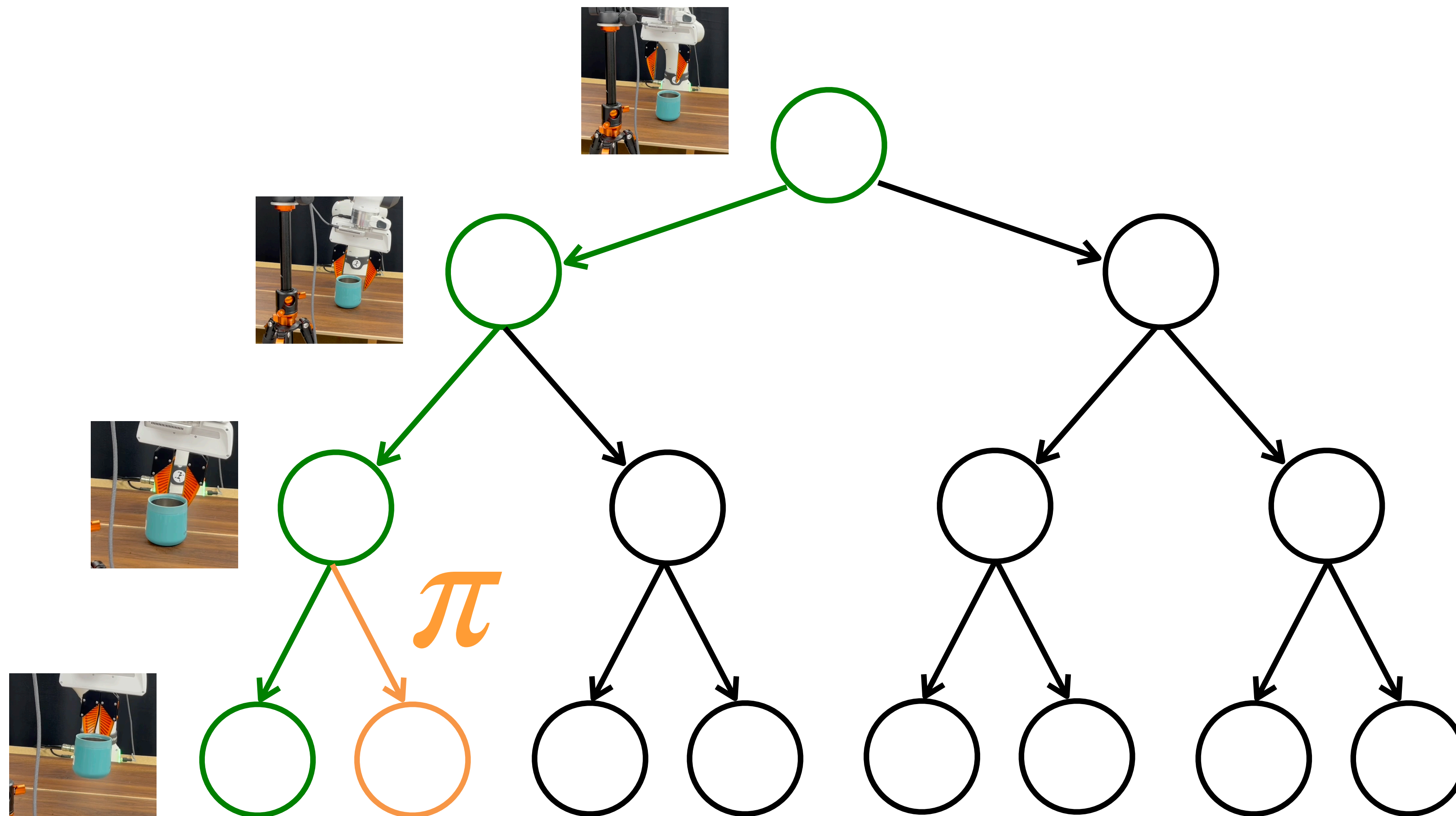
*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*





*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*



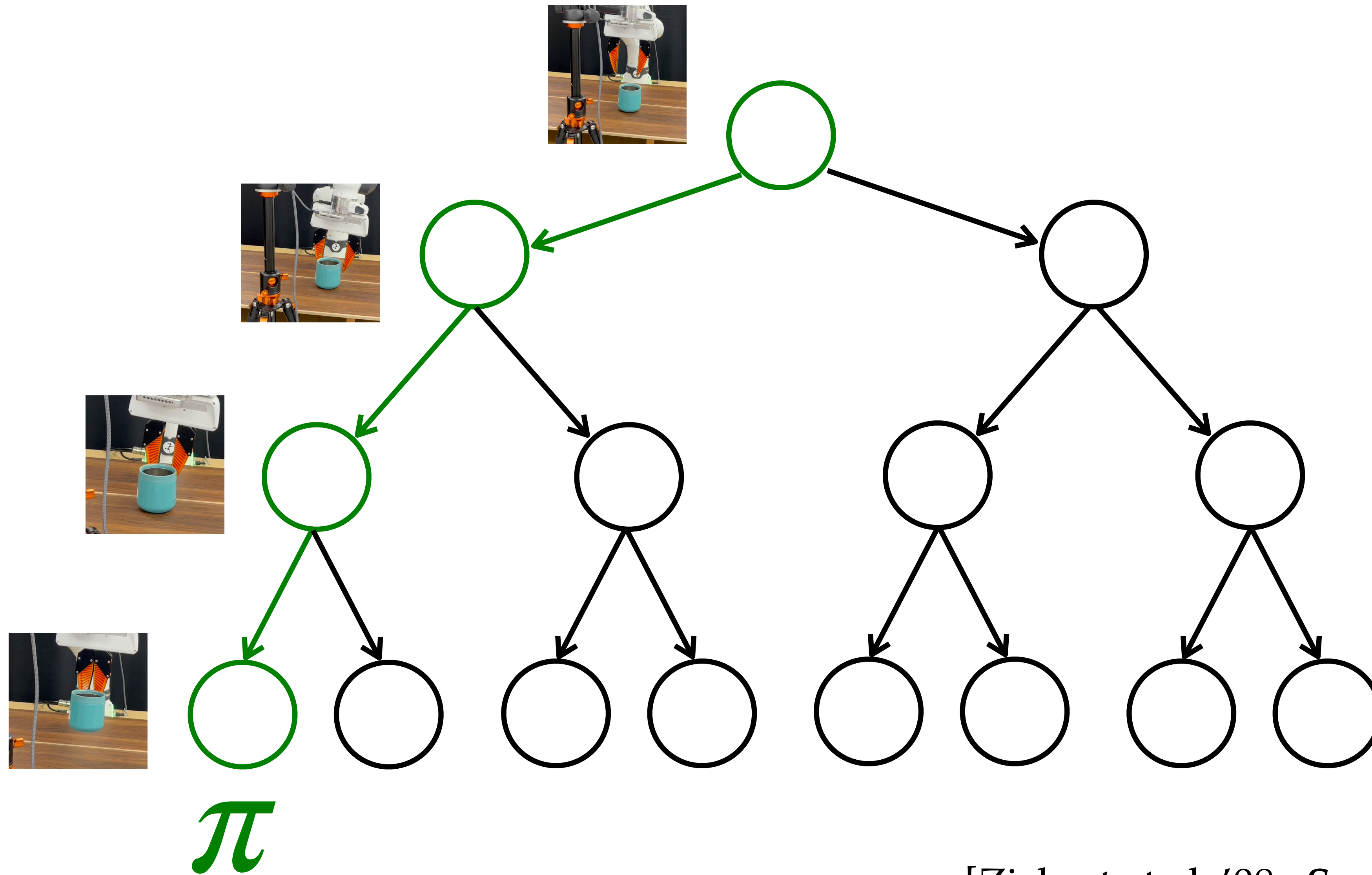


*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*





# Imitation via Inverse Reinforcement Learning



[Ziebart et al. '08; Swamy et al. '21; '22; '23]

*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*

# *Inverse Reinforcement Learning as Game-Solving*

*RL*

$$\min_{f \in \mathcal{R}} \max_{\pi \in \Pi} J(\pi, f) - J(\pi_E, f)$$

$$\text{where } J(\pi, f) \triangleq \mathbb{E}_{\xi \sim \pi} \left[ \sum_h^H f(s_h, a_h) \right].$$

★ *I proved that equilibria of this game learn recovery behavior* ★

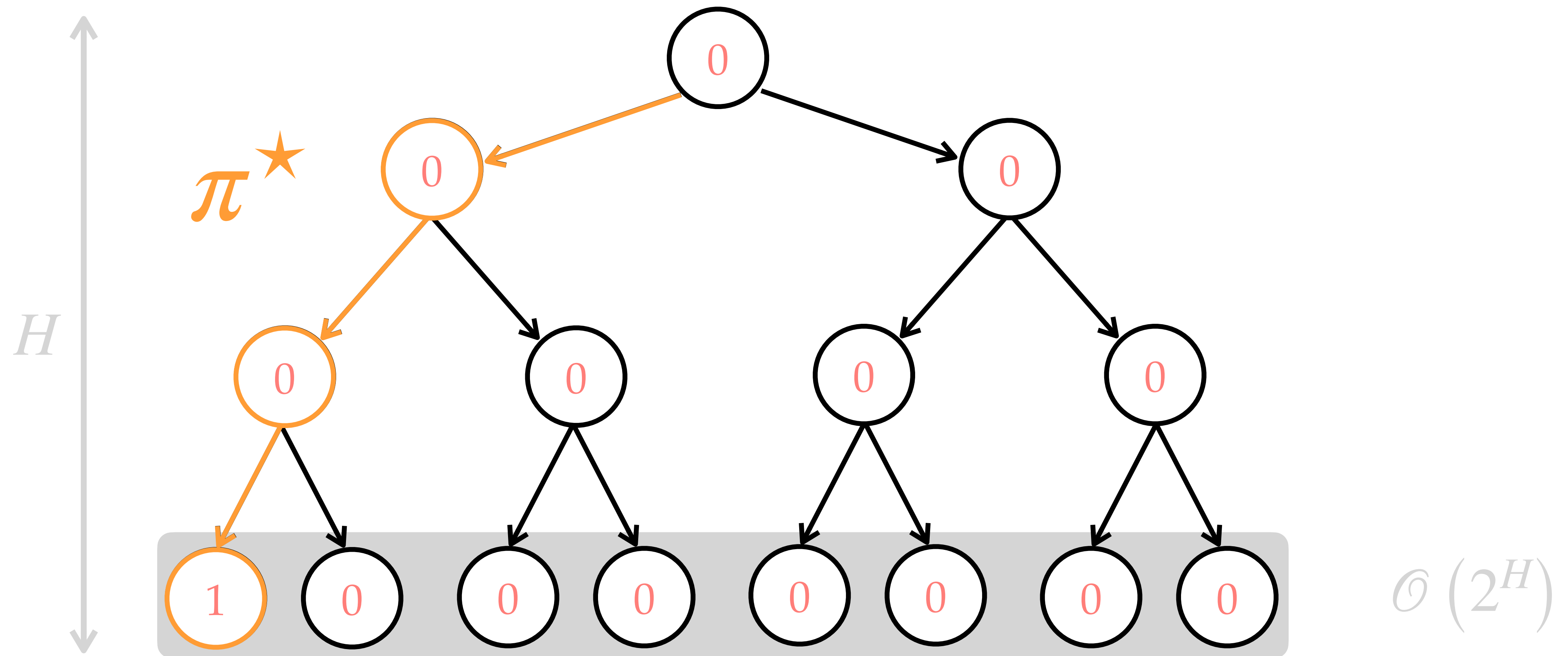
[Swamy et al. '21, '22, '23, ...]

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*



# 🔑 Problem: *Exploration* Makes (I)RL Inefficient

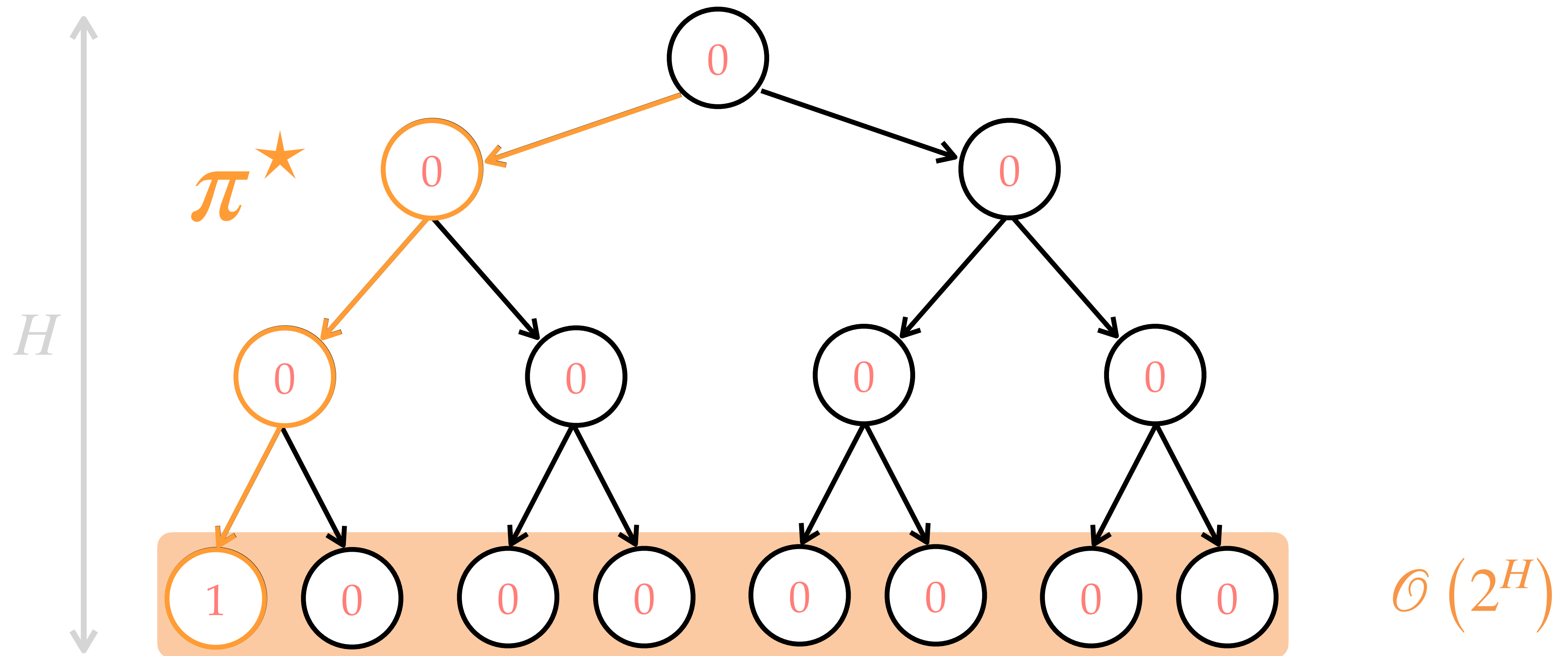
$$\pi_E \xleftrightarrow{f} \pi$$



Key Problem  $\longrightarrow$  Classical Solution  $\longrightarrow$  Efficient Algorithm  $\longrightarrow$  Practical Results

# 🔑 *Problem: Exploration Makes (I)RL Inefficient*

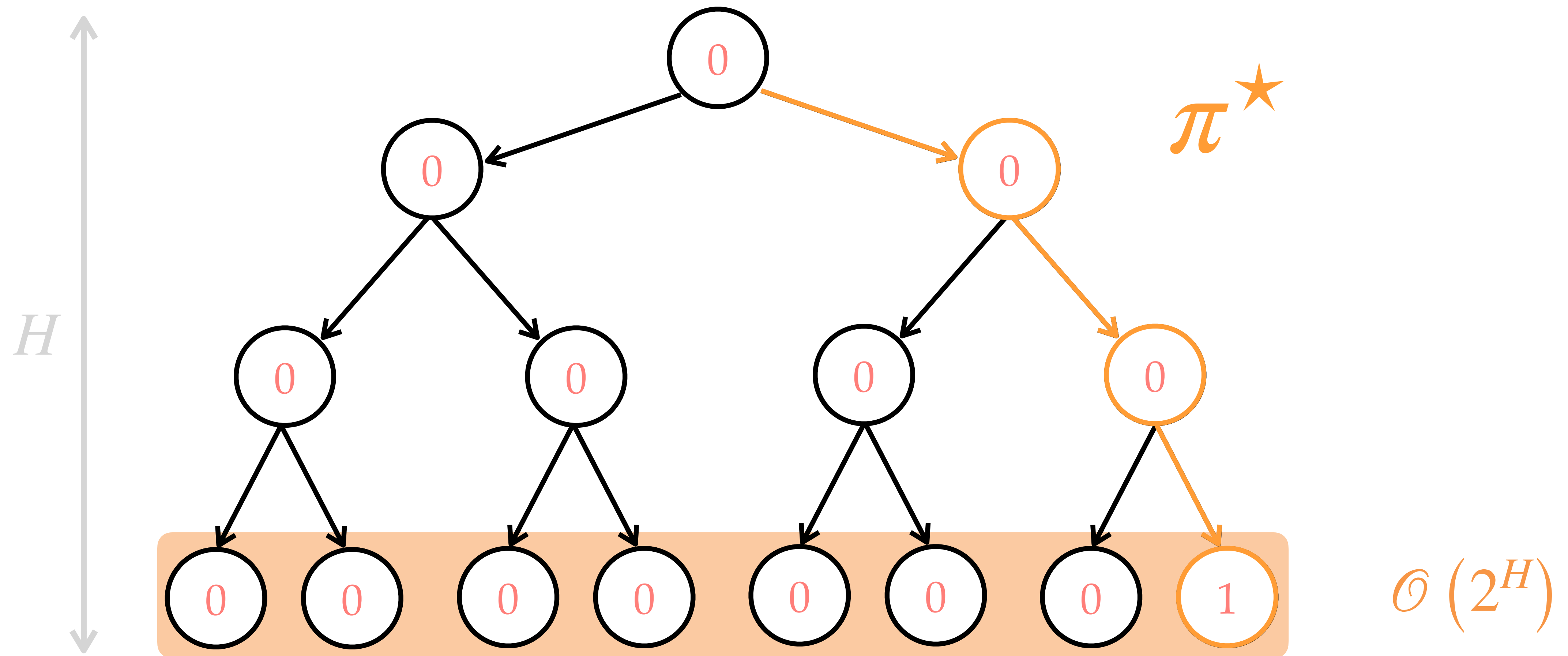
$$\pi_E \xleftrightarrow{f} \pi$$



*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*

# 🔑 Problem: *Exploration* Makes (I)RL Inefficient

$$\pi_E \xleftrightarrow{f} \pi$$



Key Problem  $\longrightarrow$  Classical Solution  $\longrightarrow$  Efficient Algorithm  $\longrightarrow$  Practical Results

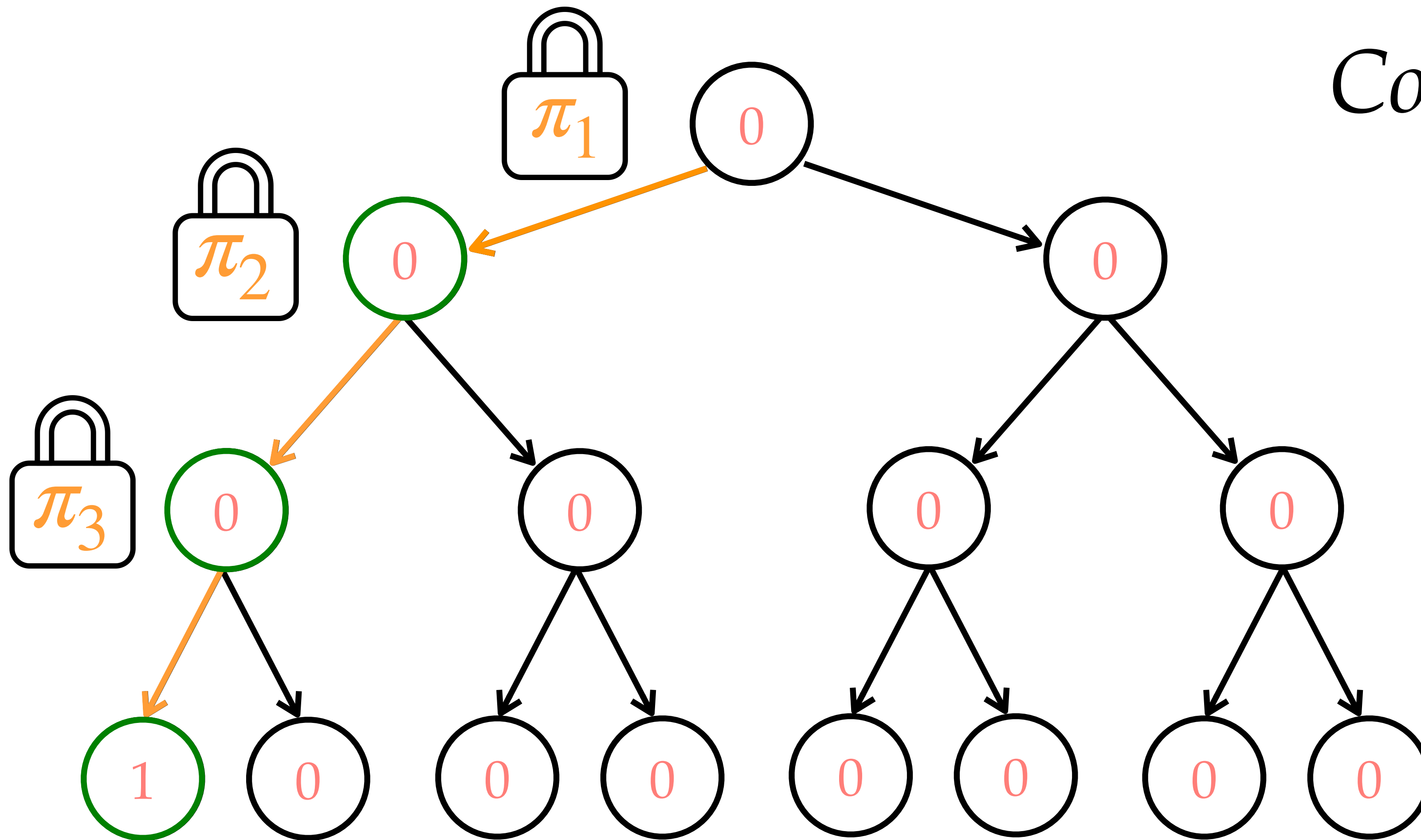


# 🔑 *Idea: Speeding up (I)RL with **Local Exploration***

$$\pi_E \xleftrightarrow{f} \pi$$

$\mathcal{O}(H^2)$   
*Complexity!*

*Expert  
States*



[Swamy et al. '23]

*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*

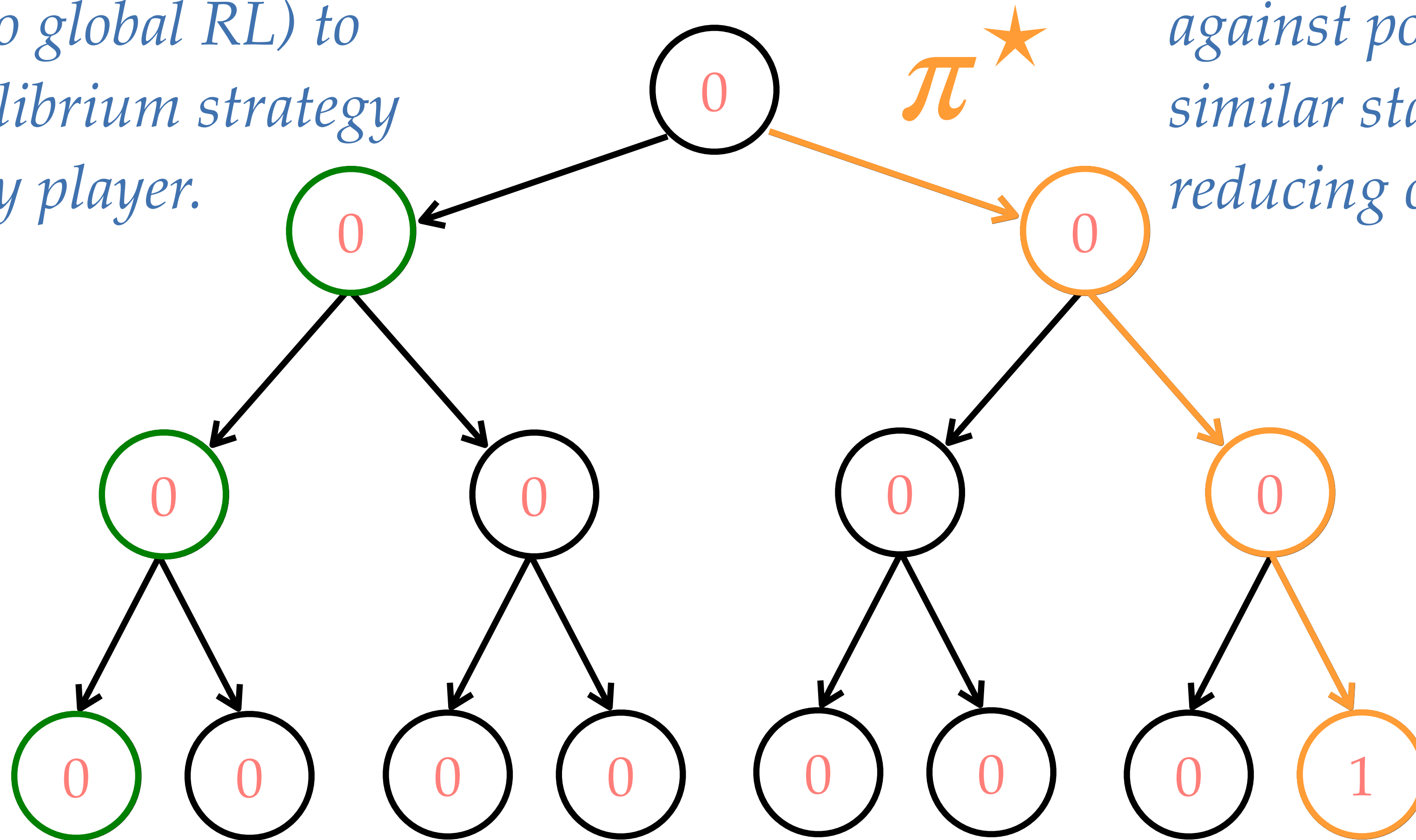
# 🔑 *Idea: Speeding up (I)RL with **Local Exploration***

*Game Theory Perspective: We don't need to compute best responses (i.e., do global RL) to compute an equilibrium strategy for just the policy player.*

$$\pi_E \xleftrightarrow{f} \pi$$

*Proof Idea: For imitation, we only need to compete against policies that visit similar states to the expert, reducing our search space!*

***Expert States***



[Swamy et al. '23]

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*

# Informal Theorem: *Local Exploration* is Sufficient for Imitation

By running  $T$  rounds of local exploration against a no-regret reward selection algorithm with regret bound  $\text{Reg}_f(T)$ , we compute policies  $\pi_{1:T}$  such that

$$J(\pi_E, r) - J(\bar{\pi}, r) \leq \underbrace{H\epsilon_\Pi}_{\substack{\text{Trajectory-level} \\ \text{mixture of } \pi_{1:T}}} + \underbrace{H^2\epsilon_{opt}}_{\substack{\text{Policy Optimization} \\ \text{Error}}} + \underbrace{H \frac{\text{Reg}_f(T)}{T}}_{\substack{\text{Reward} \\ \text{Optimization Error}}}$$

$\mathcal{O}(T \cdot \text{poly}(H))$  naively,  
 $\mathcal{O}(\text{poly}(H))$  if more clever

Policy Class  
Misspecification

Reward  
Optimization Error



First  $\text{poly}(H)$ -interaction complexity imitation learning

[Swamy et al. 2023;  
Ren\*, Swamy\* et al. 2024;

algorithms to provably learn recovery behavior!

Espinosa-Dice ... Swamy 2025]

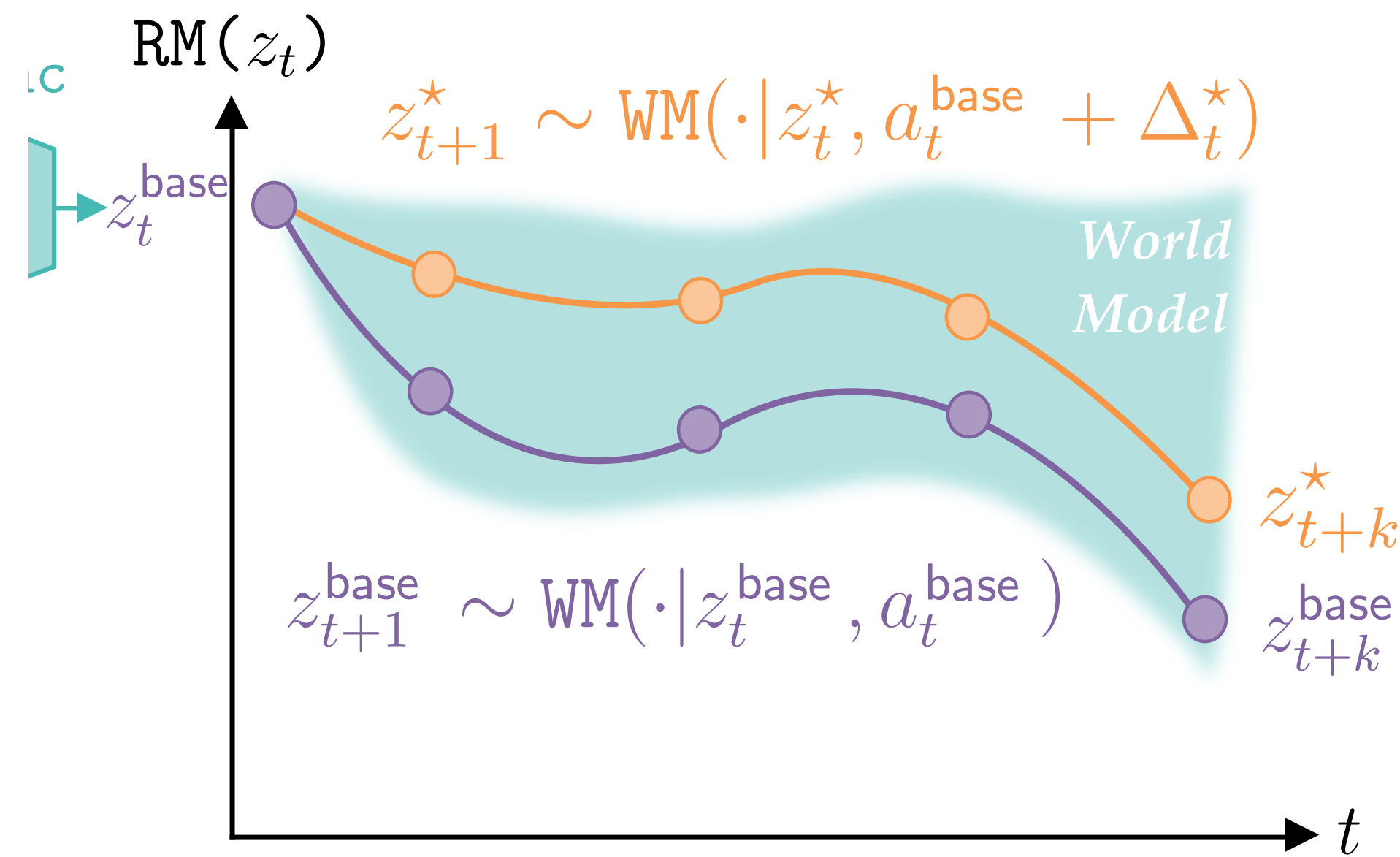
Key Problem  $\longrightarrow$  Classical Solution  $\longrightarrow$  Efficient Algorithm  $\longrightarrow$  Practical Results



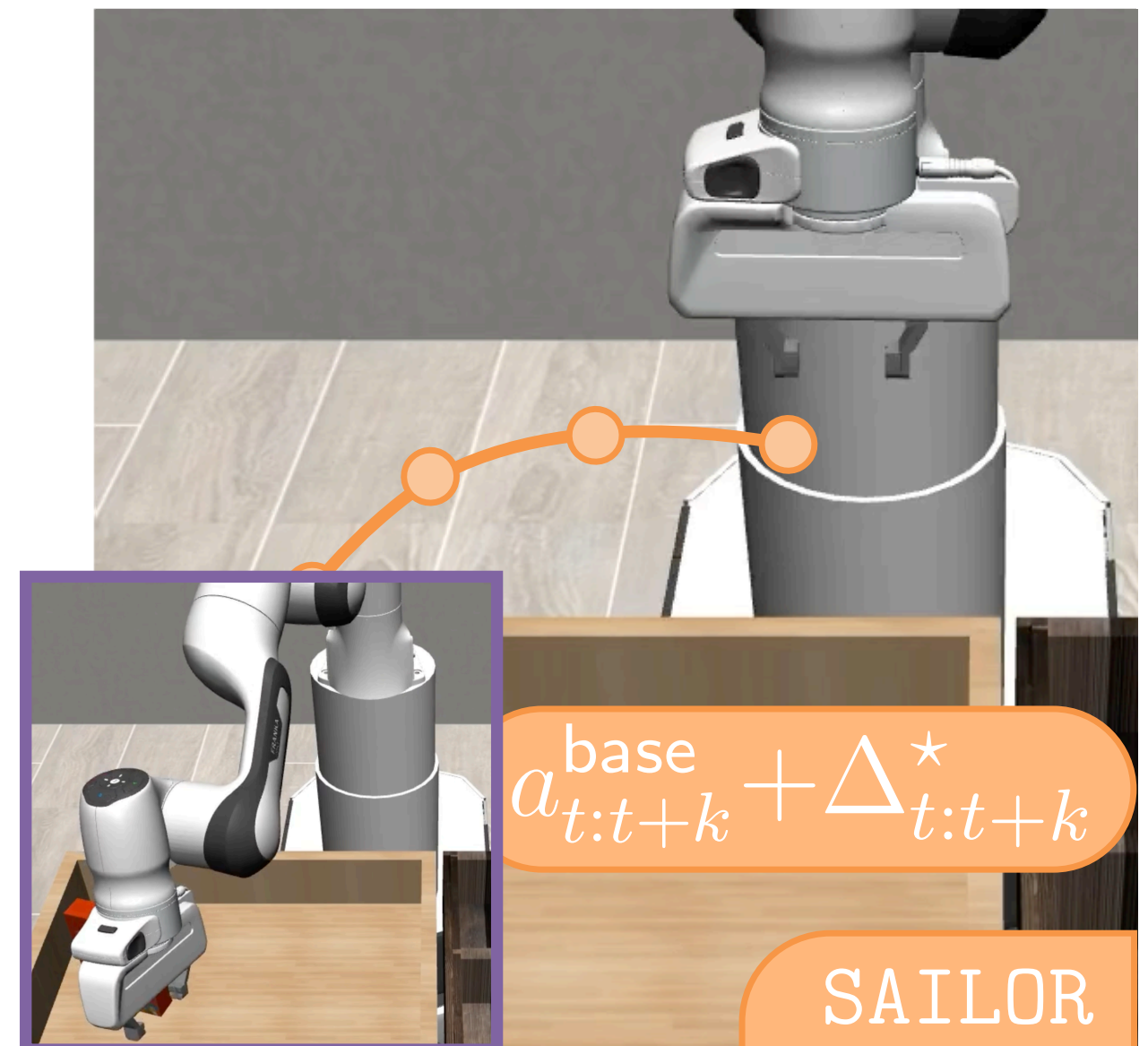
# Local Exploration: From Recovery to Prevention

🔑 Idea: Use *Local World Model*  
(Learned Partial Simulator)

## 2. Local Exploration in WM against RM



## 3. SAILOR Avoids Failure



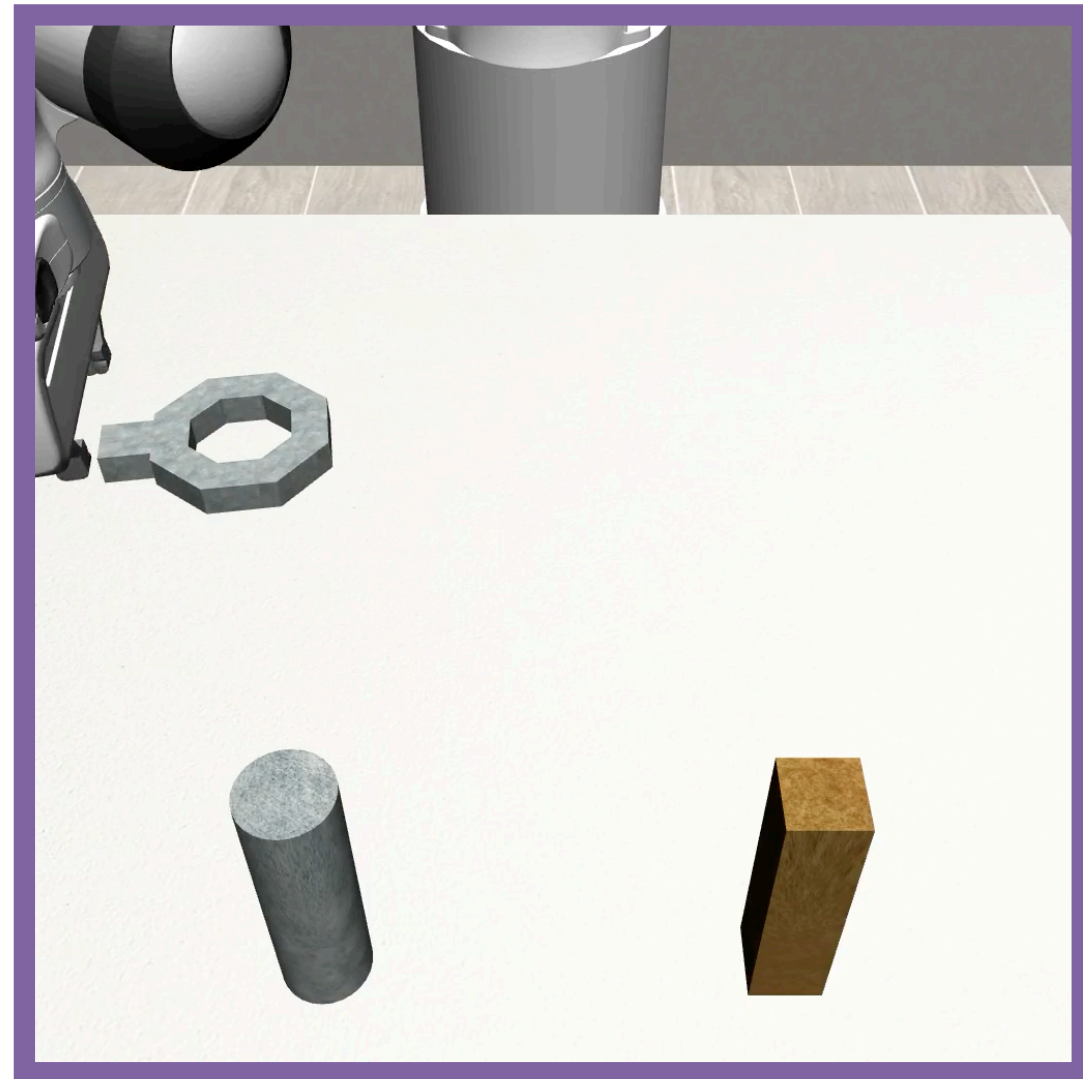
🔑 Idea: Learning all the parts of a model-predictive-control stack from expert demos

[Jain\*, Mohta\*, ... Swamy '25 — Spotlight @ NeurIPS '25]

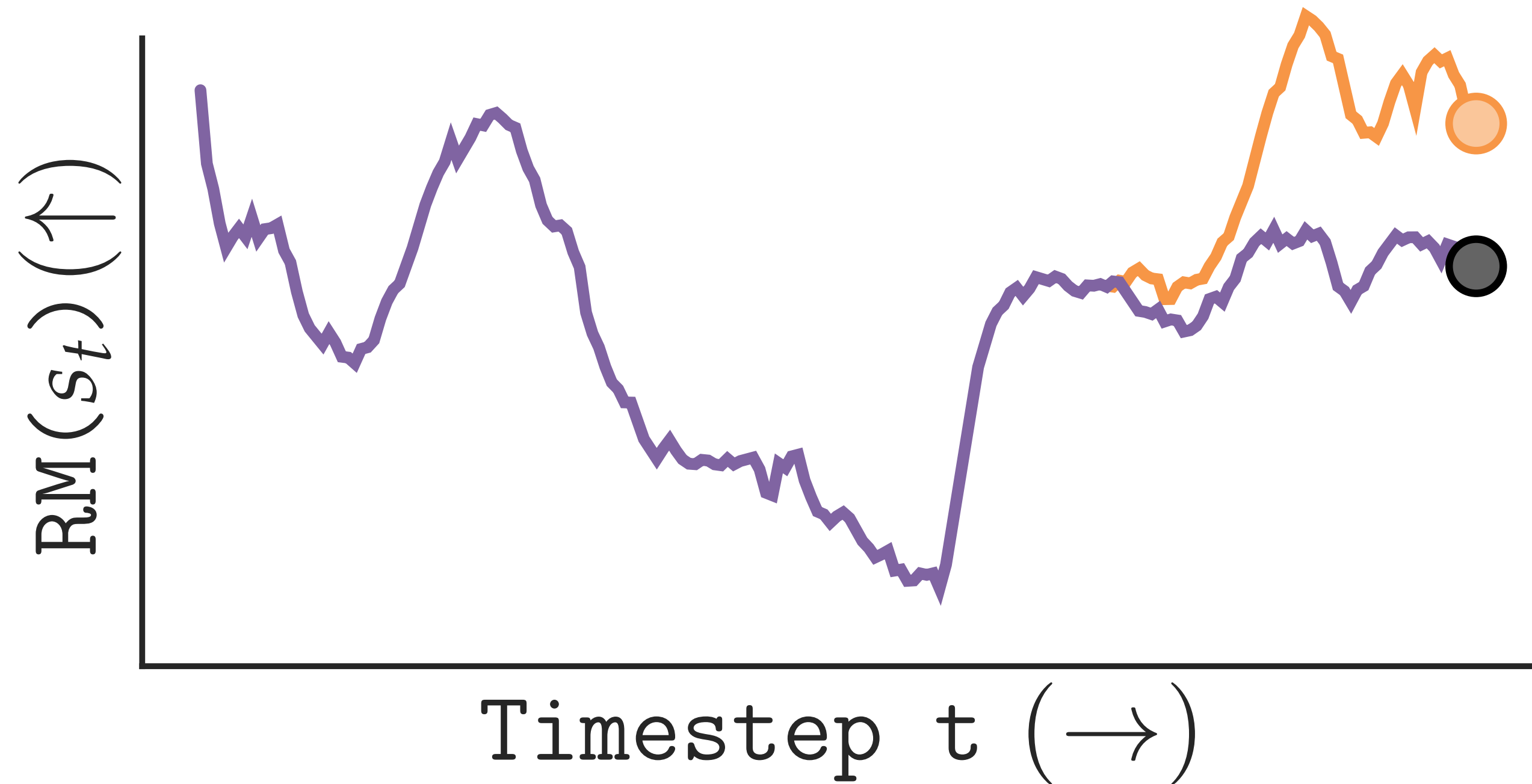
Key Problem  $\longrightarrow$  Classical Solution  $\longrightarrow$  Efficient Algorithm  $\longrightarrow$  Practical Results

# *Local Exploration: From Recovery to Prevention*

*Missed Hang*



*SL*



*Successful Hang*



*Local Exploration*

[Jain\*, Mohta\*, ... Swamy '25 — Spotlight @ NeurIPS '25]

*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*

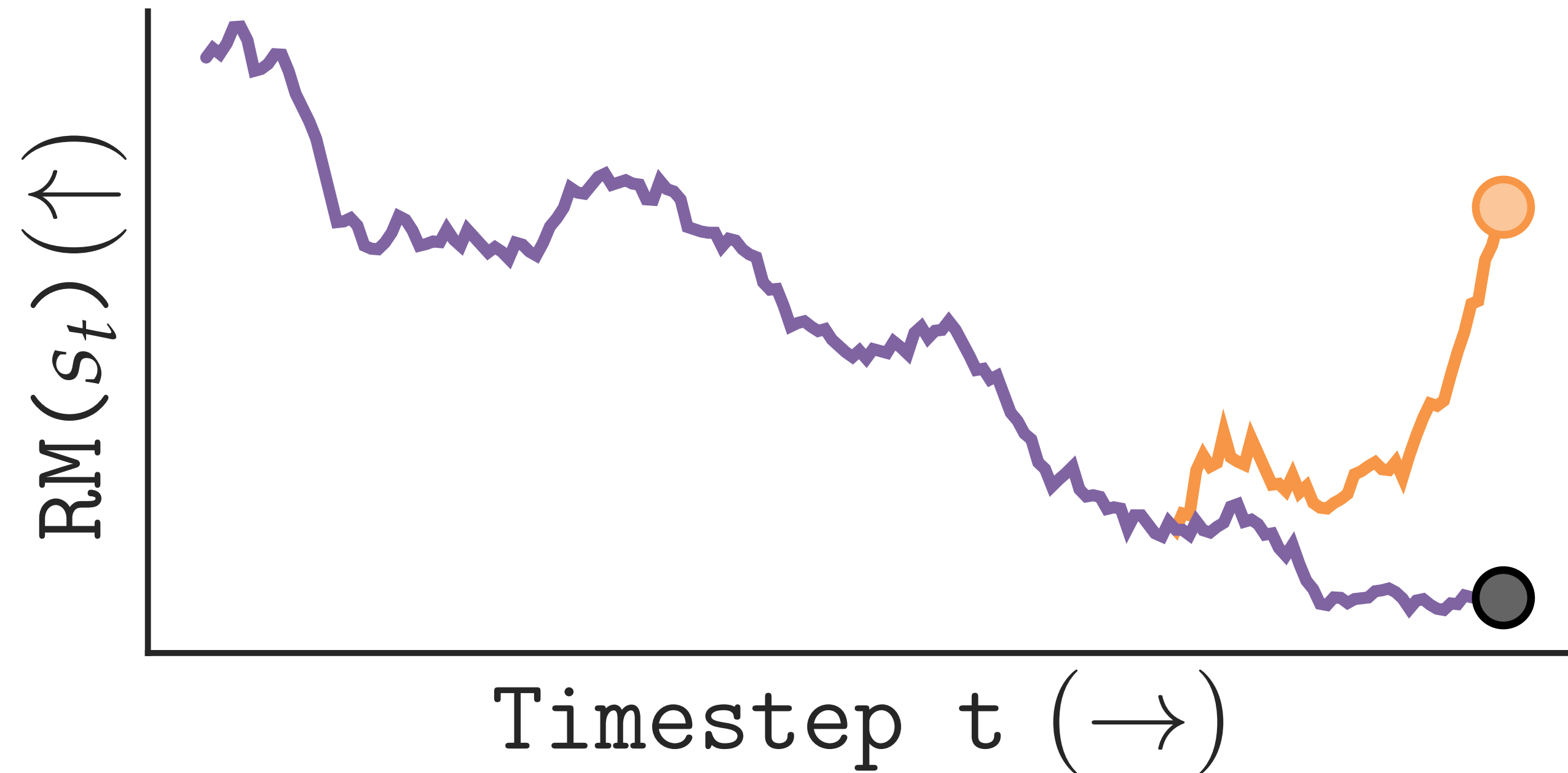


# *Local Exploration: From Recovery to Prevention*

*Missed Pick*



*SL*



*Successful Pick*



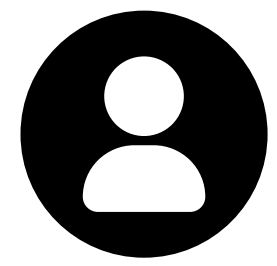
*Local Exploration*

[Jain\*, Mohta\*, ... Swamy '25 — Spotlight @ NeurIPS '25]

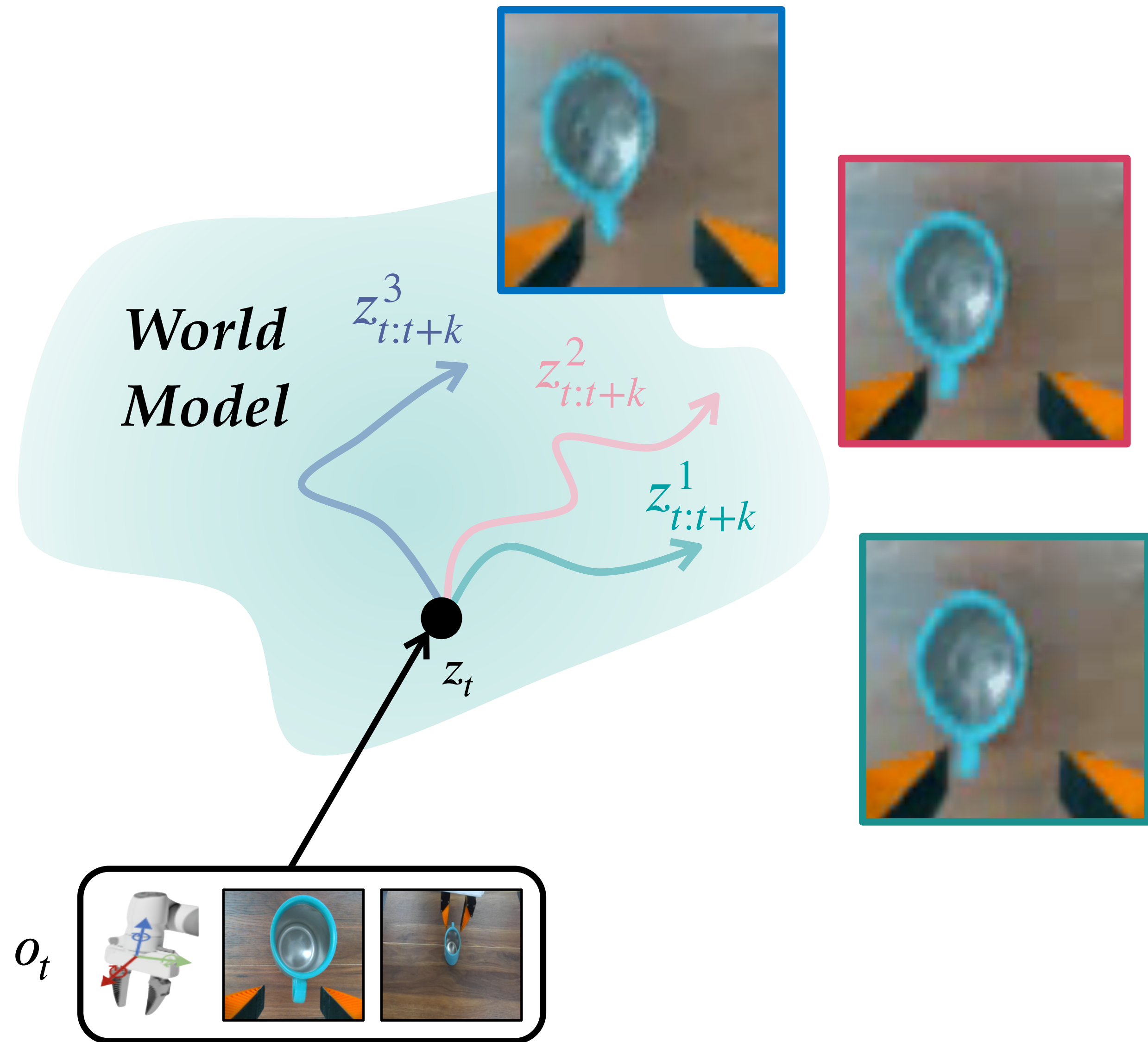
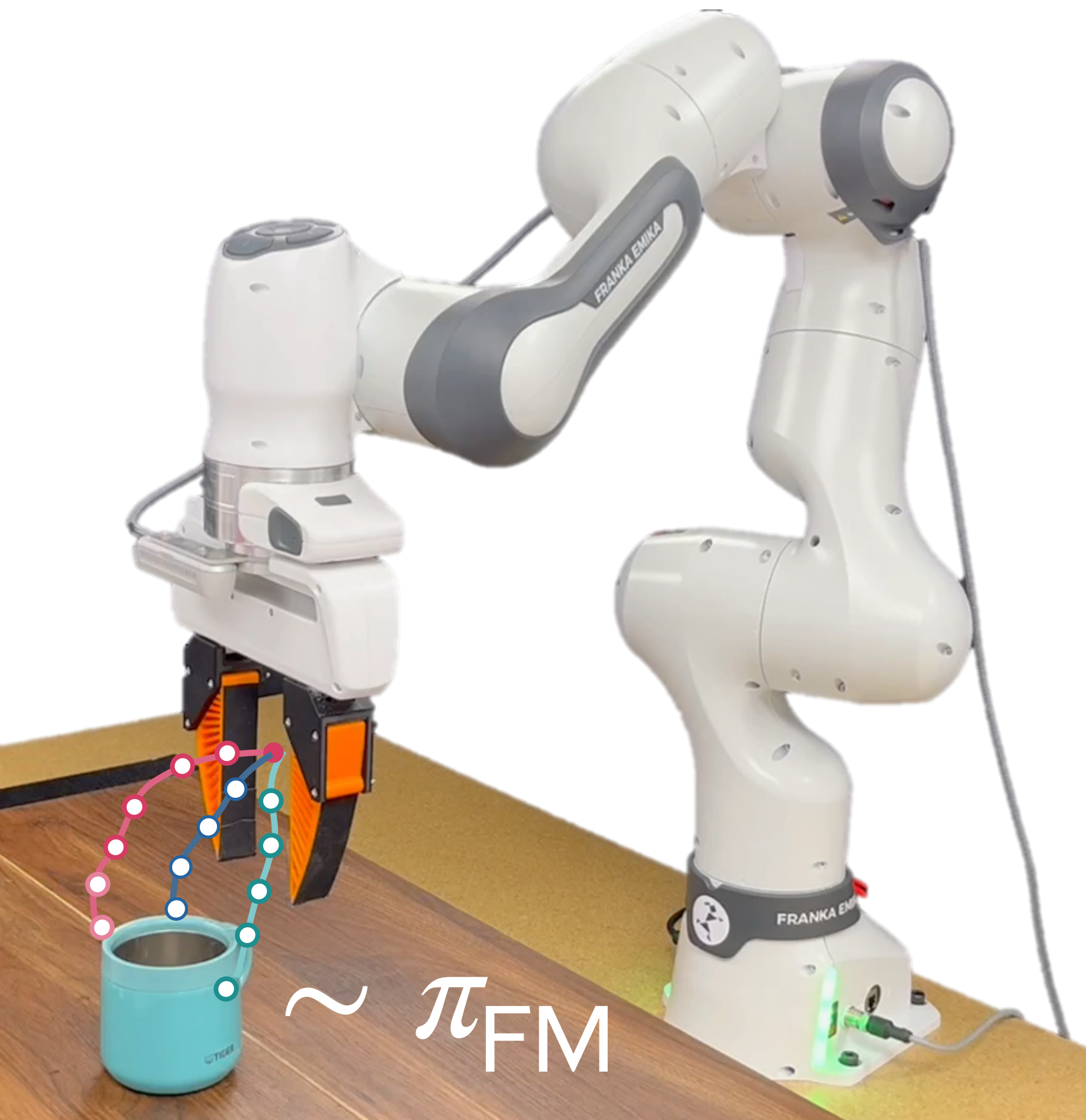
*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*



# Local Exploration: From Recovery to Prevention



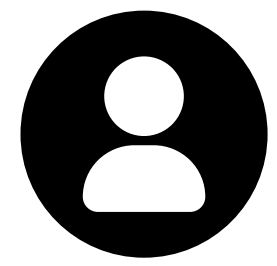
“Serve a cup of water  
to my guest”



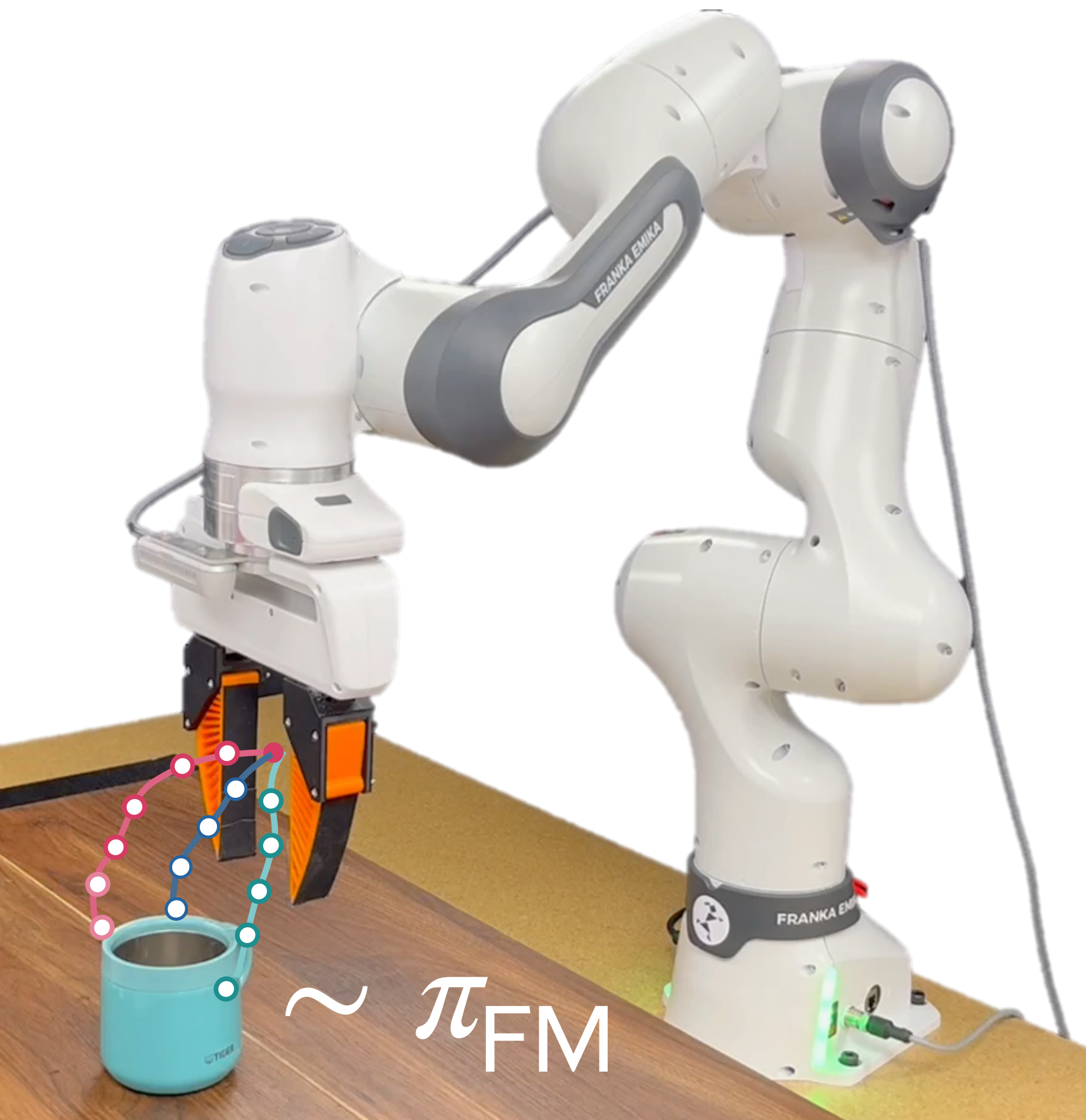
Key Problem  $\longrightarrow$  Classical Solution  $\longrightarrow$  Efficient Algorithm  $\longrightarrow$  Practical Results



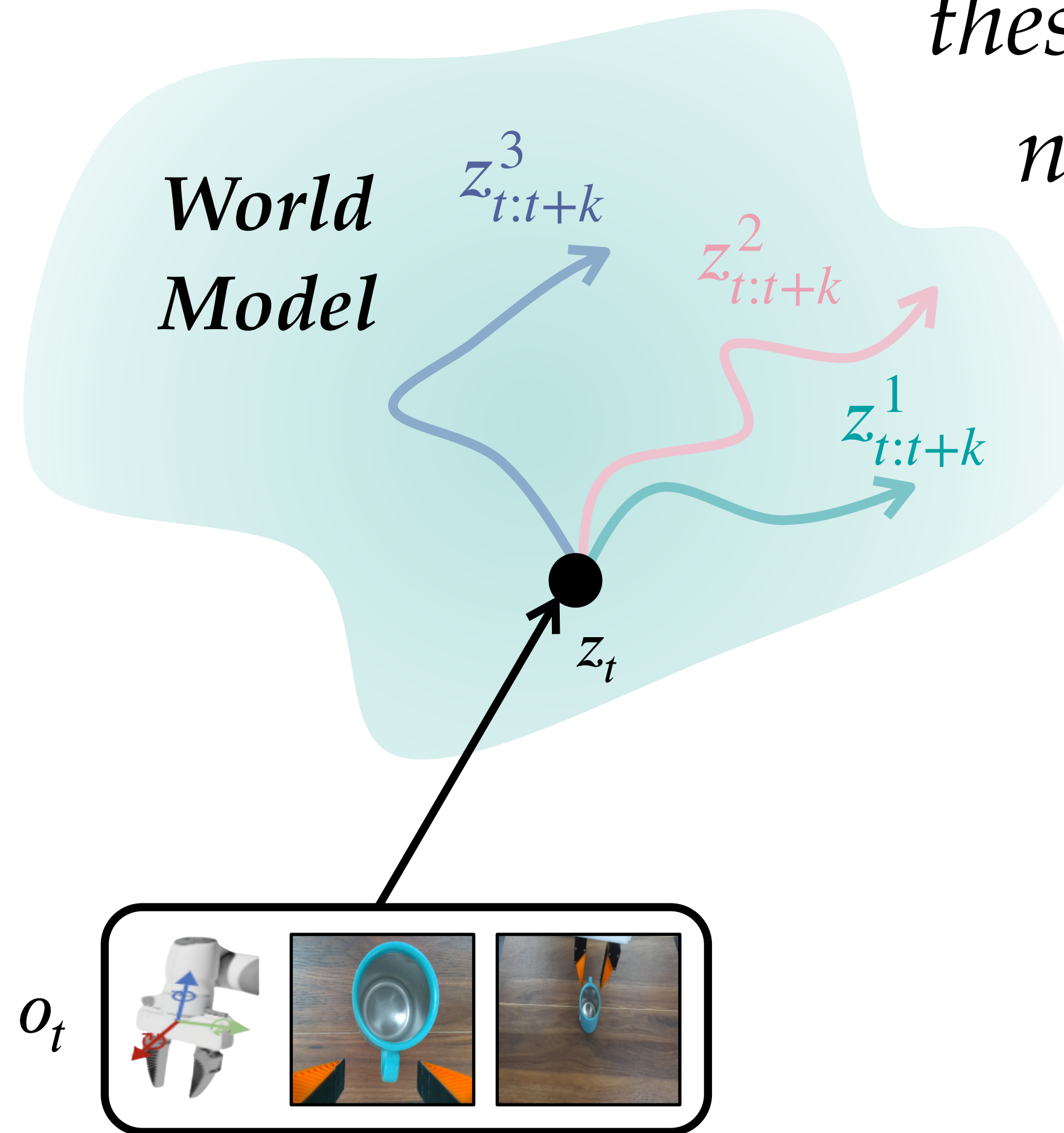
# *Local Exploration: From Recovery to Prevention*



*“Serve a cup of water  
to my guest”*



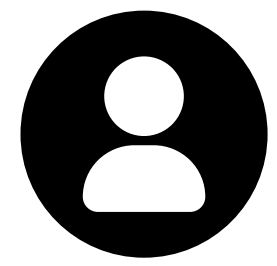
*Train a VLM to describe  
these latent futures in  
natural language*



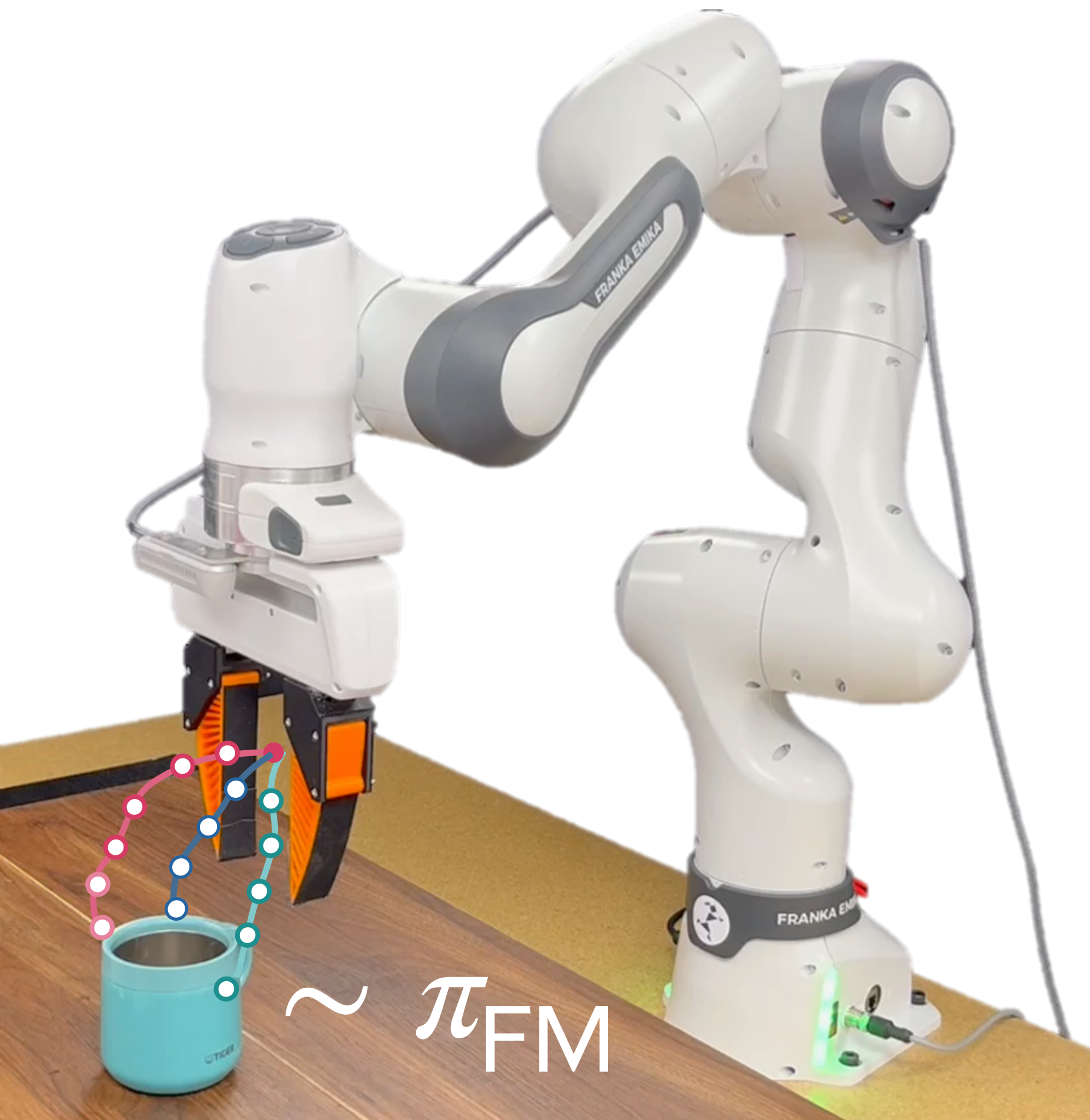
*Key Problem  $\longrightarrow$  Classical Solution  $\longrightarrow$  Efficient Algorithm  $\longrightarrow$  Practical Results*



# *Local Exploration: From Recovery to Prevention*



*“Serve a cup of water  
to my guest”*



*Train a VLM to describe these latent  
futures in natural language*

$z_{t:t+k}^3$

*“The robot fails to achieve a secure grasp.”*

$z_{t:t+k}^2$

*“The robot seizes the cup through its interior.”*

$z_{t:t+k}^1$

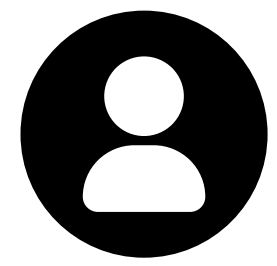
*“The robot grasps the cup via the handle.”*

*... then ask the same VLM to pick!*

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*



# *Local Exploration: From Recovery to Prevention*



*“Serve a cup of water  
to my guest”*



*Train a VLM to describe these latent  
futures in natural language*

$z_{t:t+k}^3$

*“The robot fails to achieve a secure grasp.”*

$z_{t:t+k}^2$

*“The robot seizes the cup through its interior.”*

$z_{t:t+k}^1$

*“The robot grasps the cup via the handle.”*

*... then ask the same VLM to pick!*

★ *Robot can reason over the semantic  
consequences of its actions!* ★

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*



# *Local Exploration* Learns More From Less!

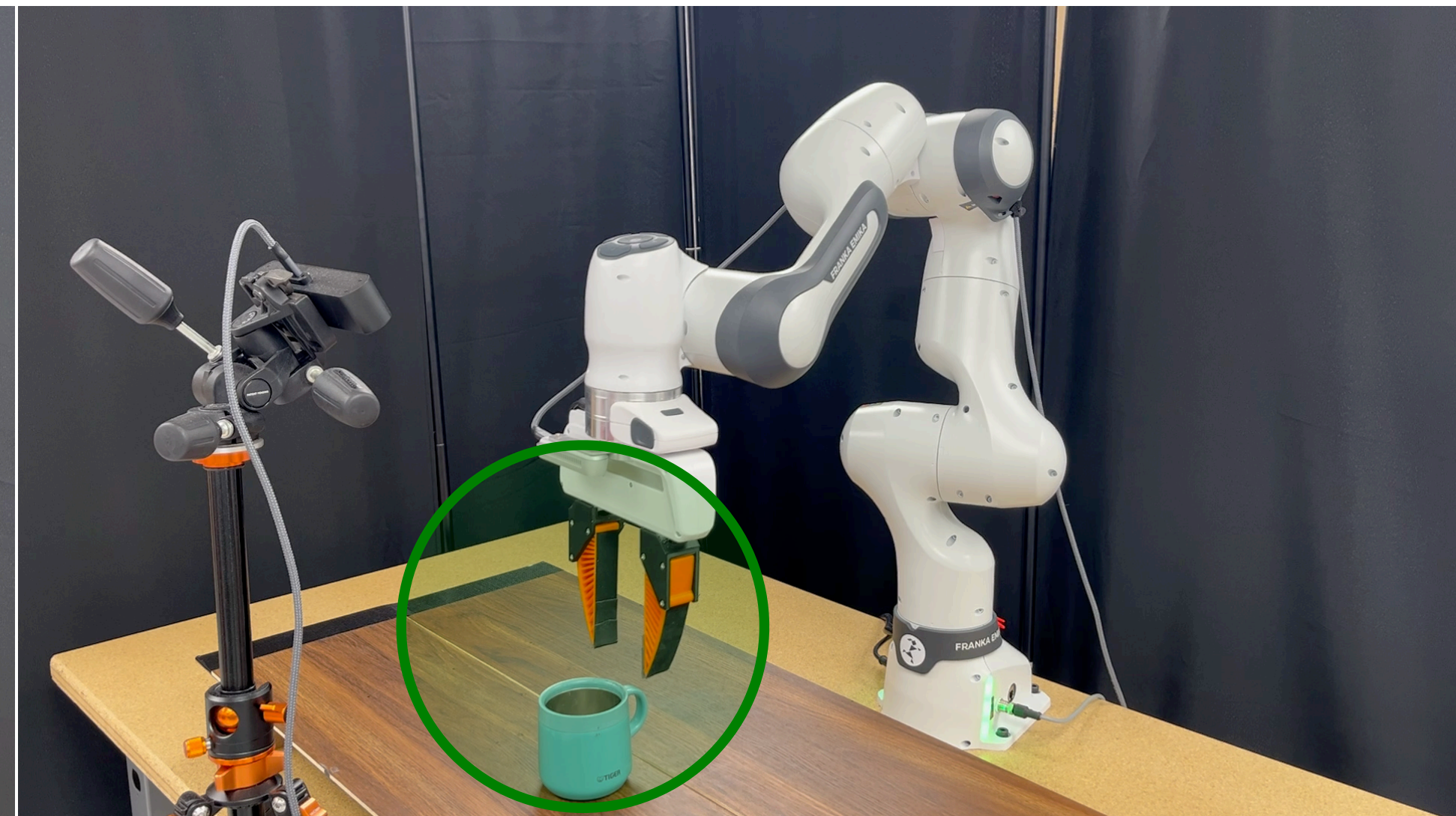
👤 “Serve a cup of water to my guest” ➔ Grasp by Handle

Base + SL

*Local Exploration*



40% SR



80% SR

[Wu, Tian, Swamy, Bajcsy '25 — *Outstanding Paper @ ICLR World Model Workshop*]

*Key Problem* ➔ *Classical Solution* ➔ *Efficient Algorithm* ➔ *Practical Results*



# *Local Exploration* Learns More From Less!

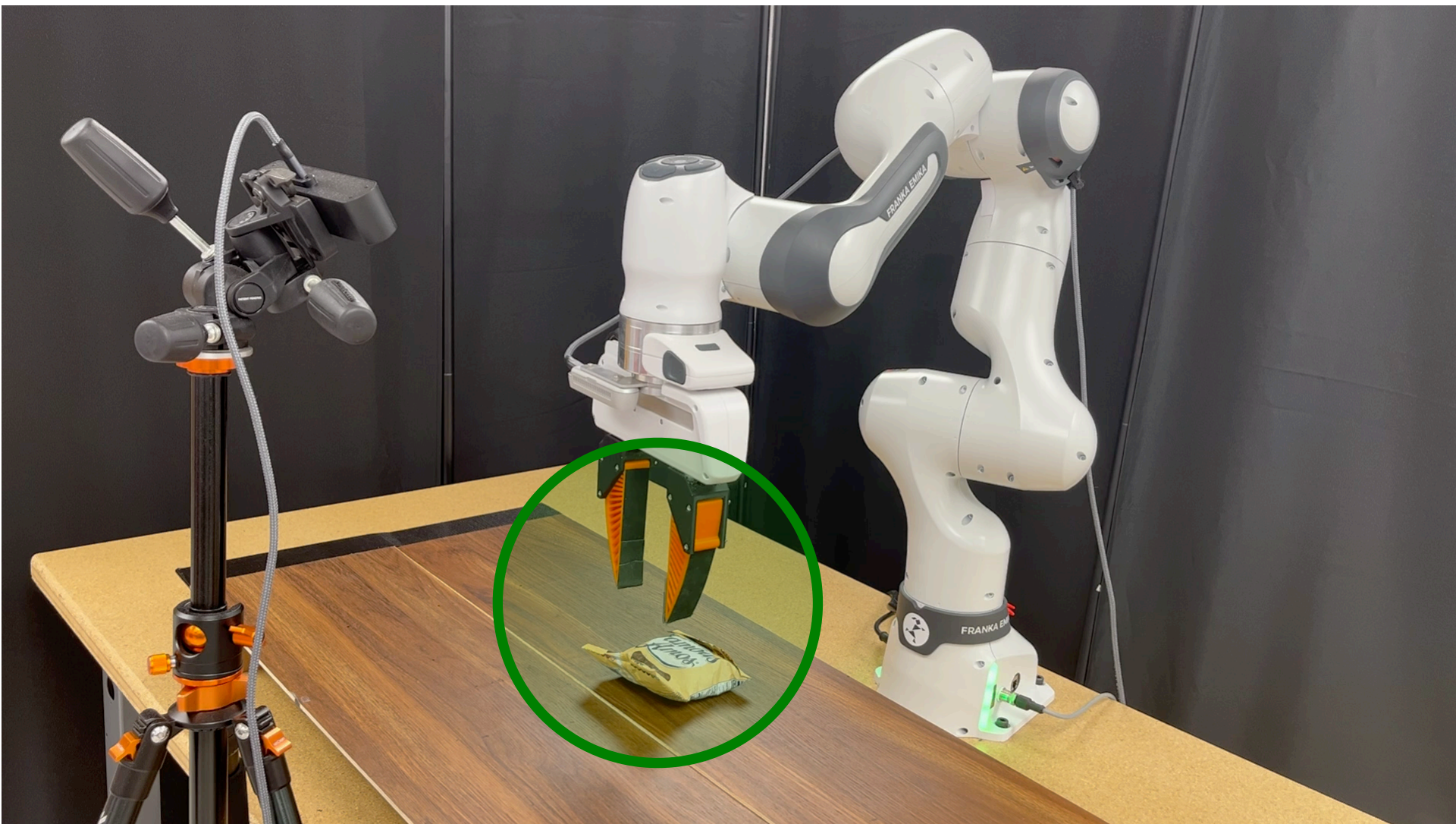
👤 “Pick up the chips without crushing them” ➔ Grasp by Edge

Base + SL

*Local Exploration*



20% SR



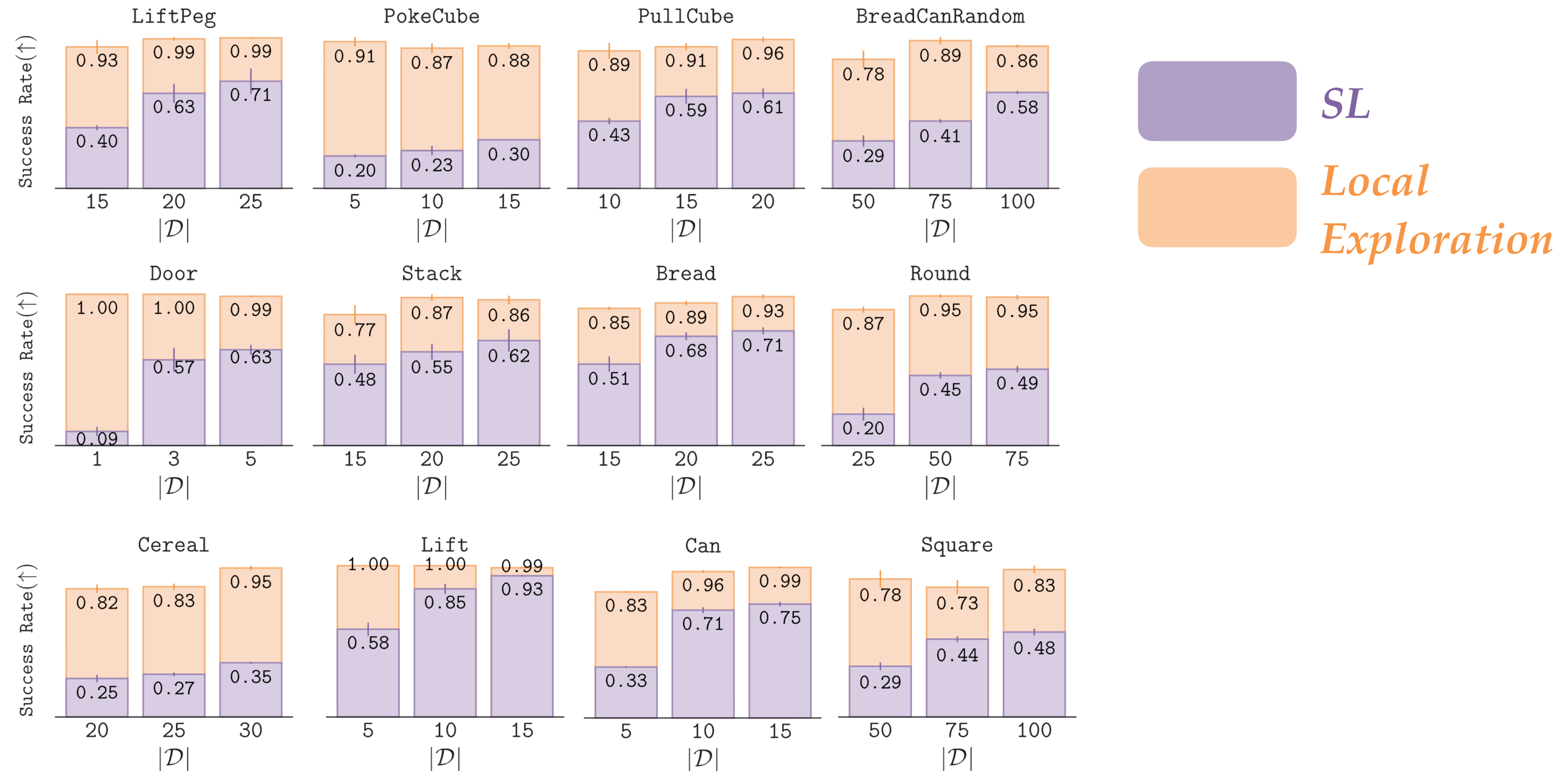
70% SR

[Wu, Tian, Swamy, Bajcsy '25 — *Outstanding Paper @ ICLR World Model Workshop*]

*Key Problem* ➔ *Classical Solution* ➔ *Efficient Algorithm* ➔ *Practical Results*



# Local Exploration Learns More From Less!



[Jain\*, Mohta\*, ... Swamy '25 — Spotlight @ NeurIPS '25]

*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*

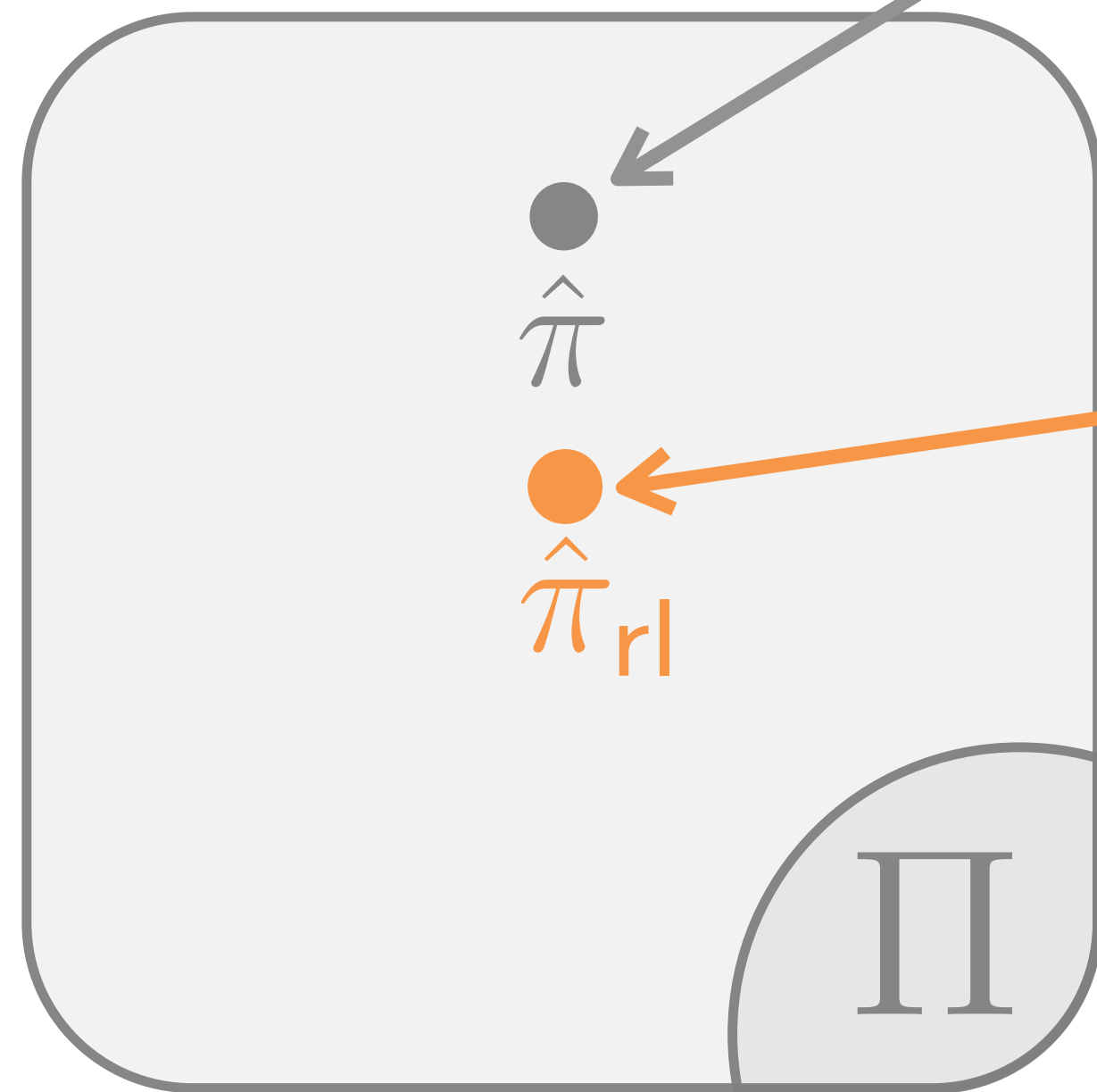
# Today:

3. *Why RL Learns More From Less*

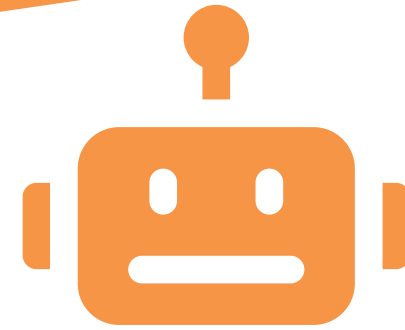
SL

$\mathcal{D}$

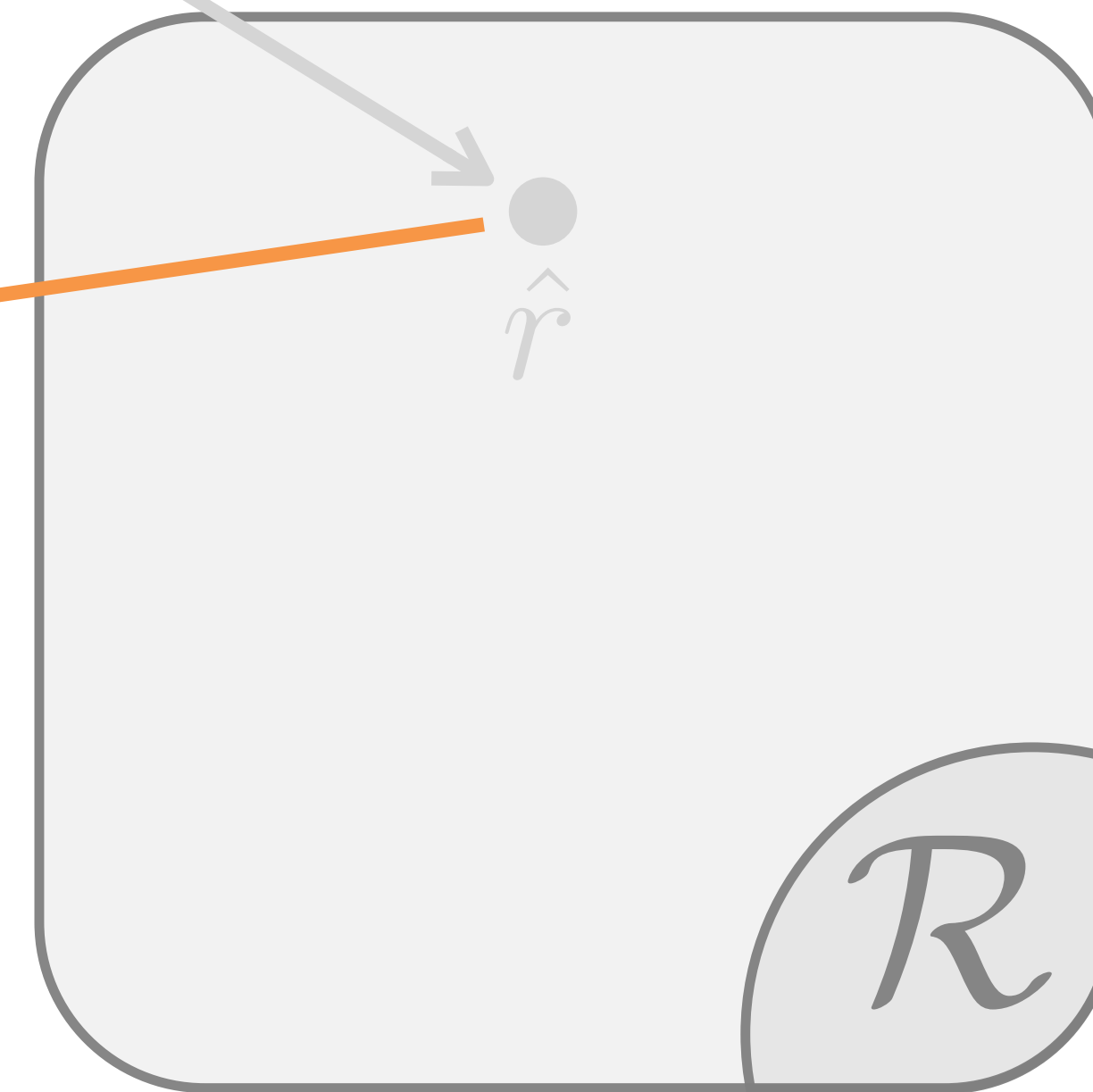
2. *Specification from Conflicting Preferences*



*Policies*



1. *Imitation Without Exploration*



*Rewards*

*Takeaway: local exploration lets us efficiently train robots to avoid failures*



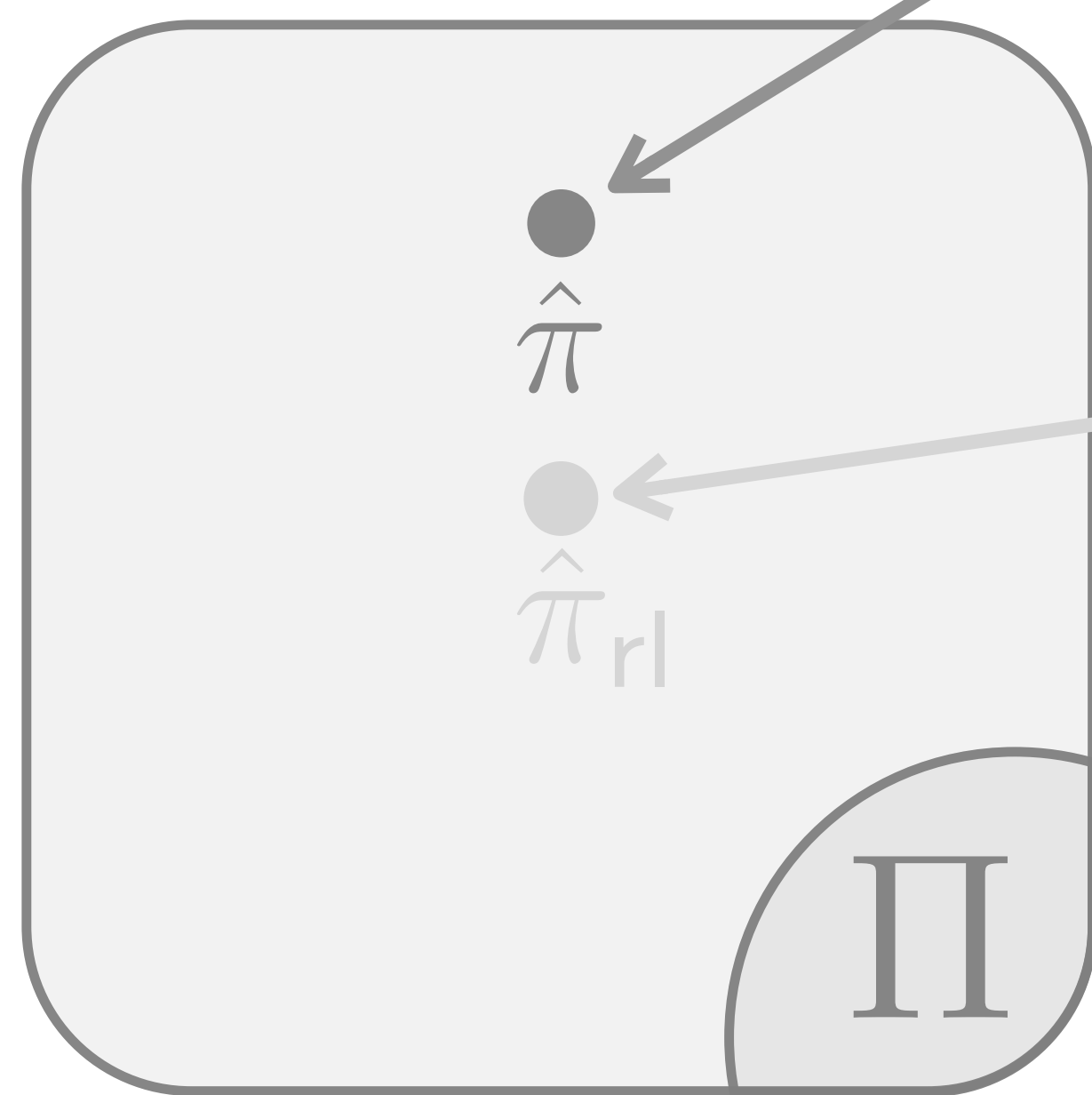
# Today:

3. *Why RL Learns More From Less*

SL

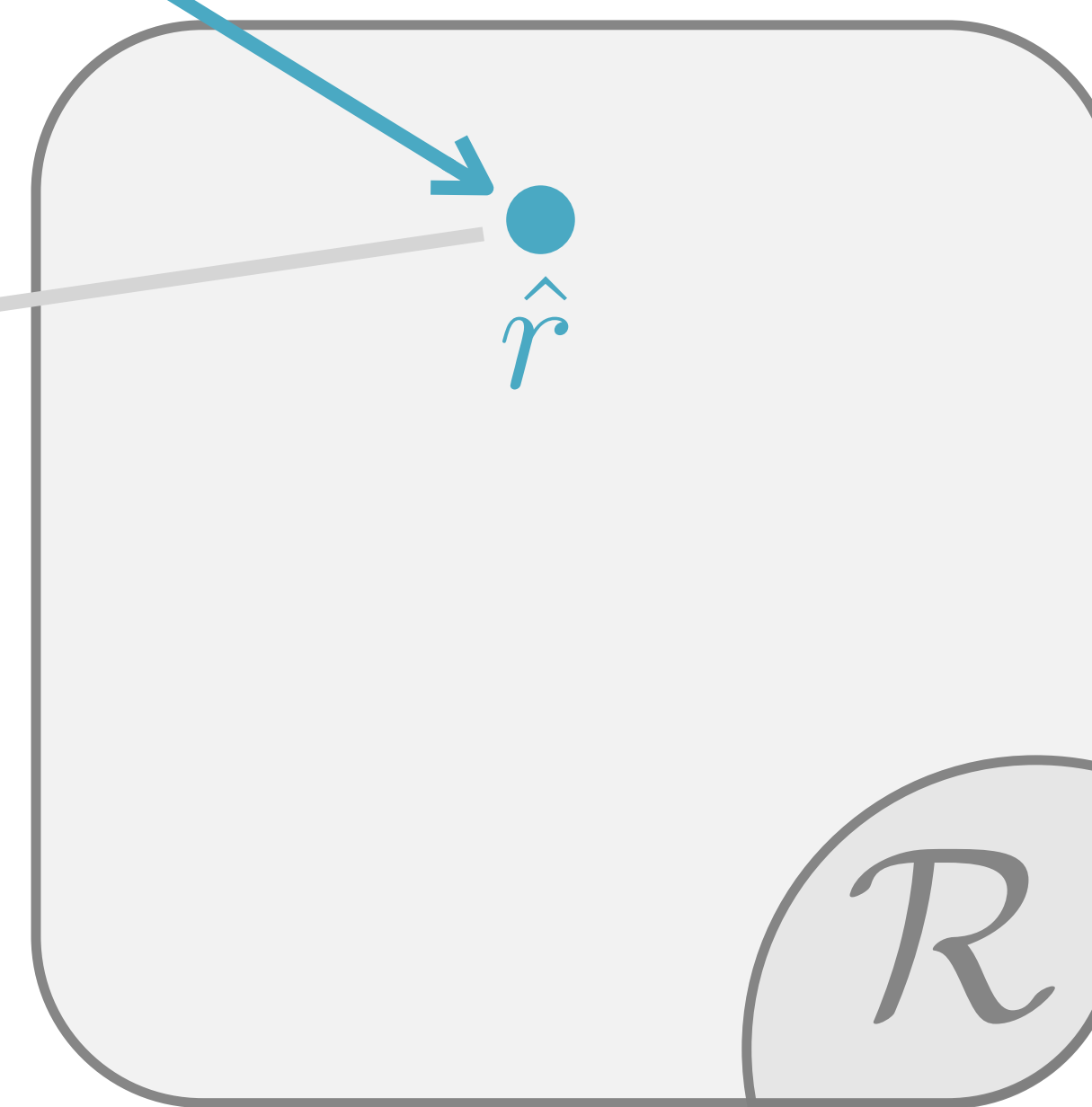
$\mathcal{D}$

2. *Specification from Conflicting Preferences*



*Policies*

1. *Imitation Without Exploration*



*Rewards*

*Theory*  $\longrightarrow$  *Practice*

# Setting: Preference Fine-Tuning

*Completion 1* ( $\xi_1 \sim \pi | s_0$ )

*Token 1*  
( $a_1 \in \mathcal{A}$ )

*Token 2*  
( $a_2 \in \mathcal{A}$ )

*Token 3*  
( $a_3 \in \mathcal{A}$ )

*Token 1*  
( $a'_1 \in \mathcal{A}$ )

*Token 2*  
( $a'_2 \in \mathcal{A}$ )

*Token 3*  
( $a'_3 \in \mathcal{A}$ )

*Completion 2* ( $\xi_2 \sim \pi | s_0$ )



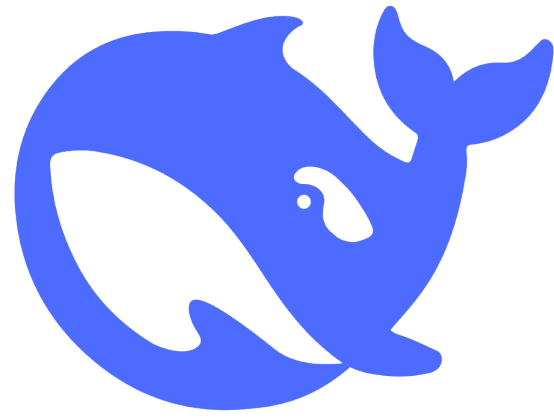
*Prompt*  
( $s_0 \sim \rho_0$ )



*Preference*  
 $\xi_1 > \xi_2$



# Setting: Preference Fine-Tuning



**Prompt**  
 $(s_0 \sim \rho_0)$

**Completion 1** ( $\xi_1 \sim \pi | s_0$ )

**Token 1**  
 $(a_1 \in \mathcal{A})$

**Token 2**  
 $(a_2 \in \mathcal{A})$

**Token 3**  
 $(a_3 \in \mathcal{A})$

**Token 1**  
 $(a'_1 \in \mathcal{A})$

**Token 2**  
 $(a'_2 \in \mathcal{A})$

**Token 3**  
 $(a'_3 \in \mathcal{A})$

**Completion 2** ( $\xi_2 \sim \pi | s_0$ )

**Gemini**

*Train*

*Reward Model*



*Preference*

$\xi_1 > \xi_2$

**Goal:** Maximize relative probability of generating  $\xi_1$  over  $\xi_2$

# Setting: Preference Fine-Tuning

*Completion 1* ( $\xi_1 \sim \pi | s_0$ )

*Token 1*  
( $a_1 \in \mathcal{A}$ )

*Token 2*  
( $a_2 \in \mathcal{A}$ )

*Token 3*  
( $a_3 \in \mathcal{A}$ )

*Token 1*  
( $a'_1 \in \mathcal{A}$ )

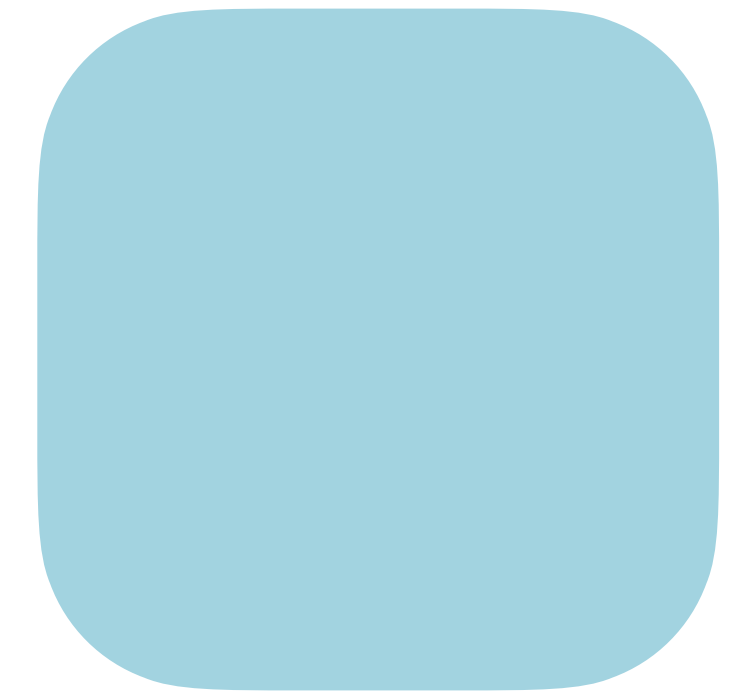
*Token 2*  
( $a'_2 \in \mathcal{A}$ )

*Token 3*  
( $a'_3 \in \mathcal{A}$ )

*Completion 2* ( $\xi_2 \sim \pi | s_0$ )



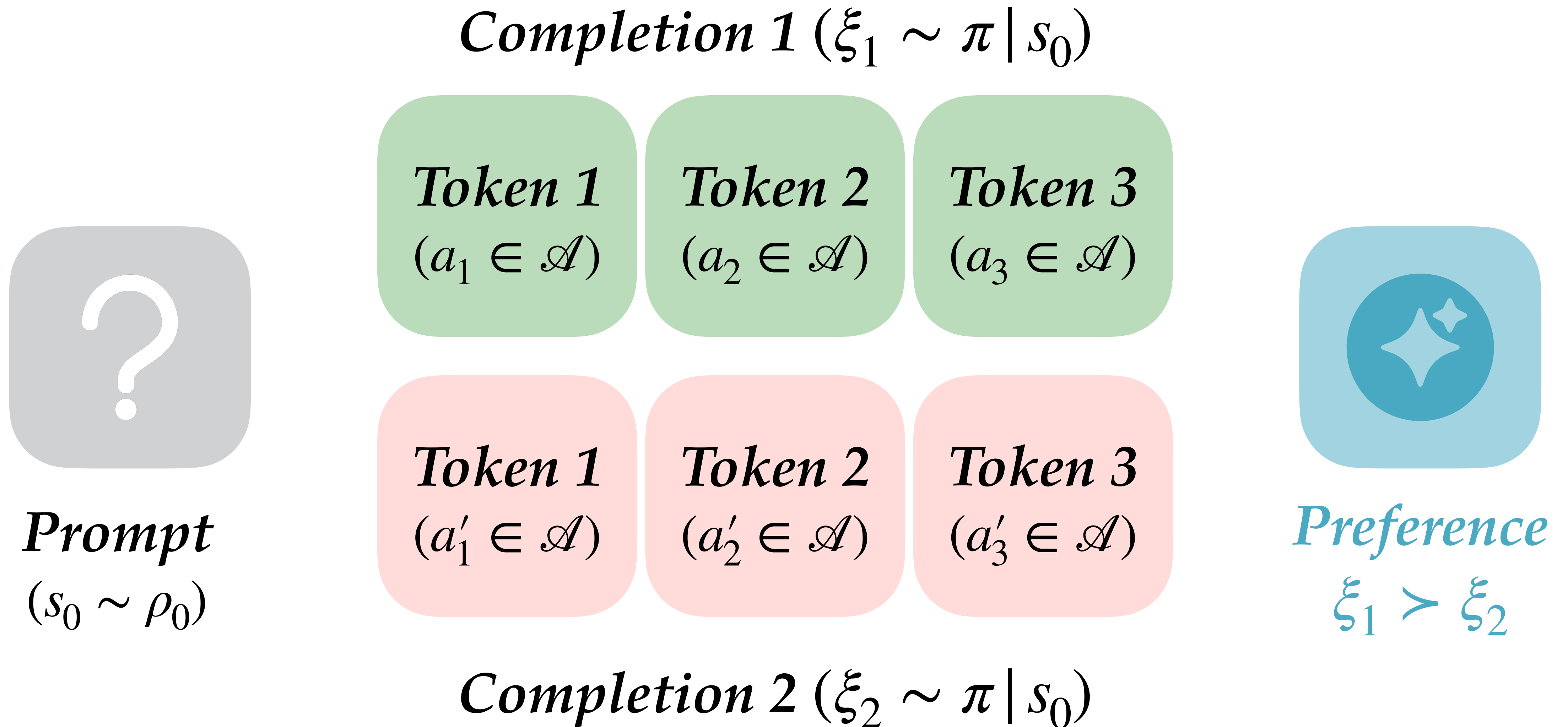
*Prompt*  
( $s_0 \sim \rho_0$ )



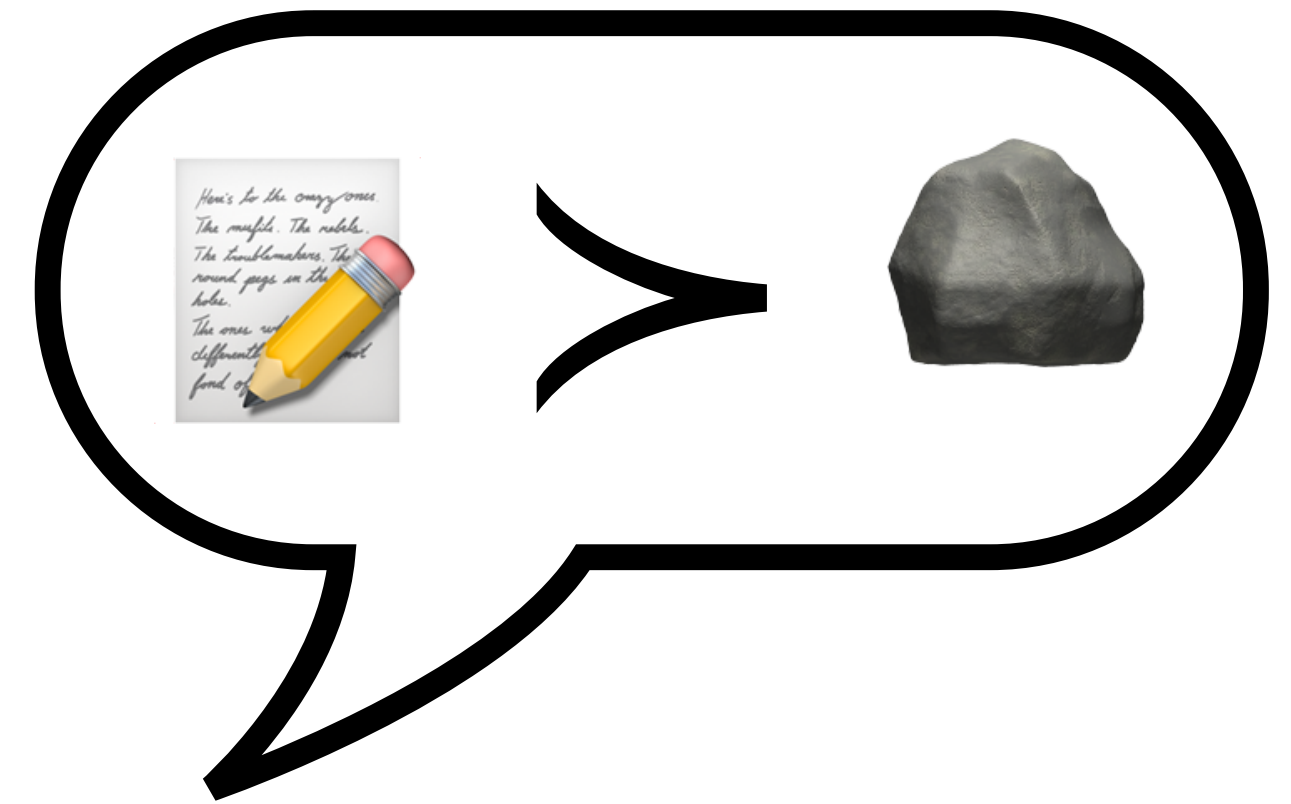
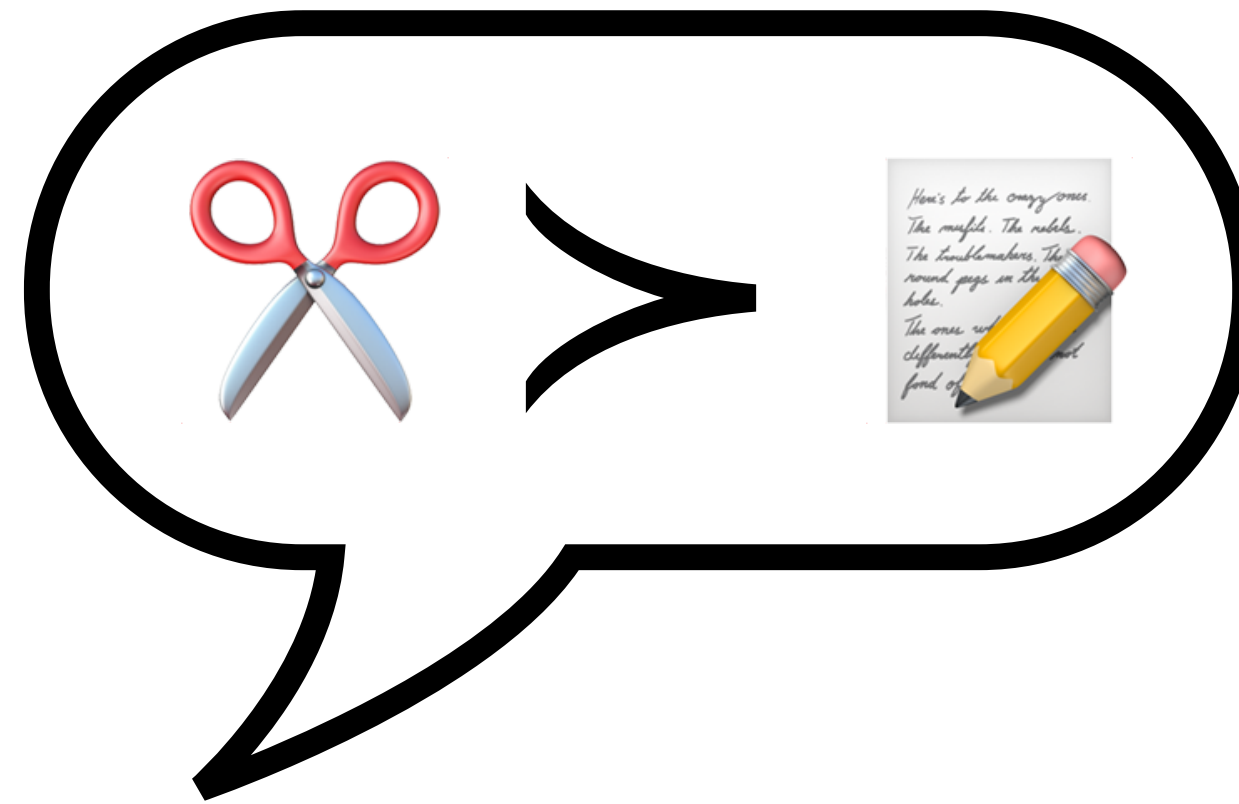
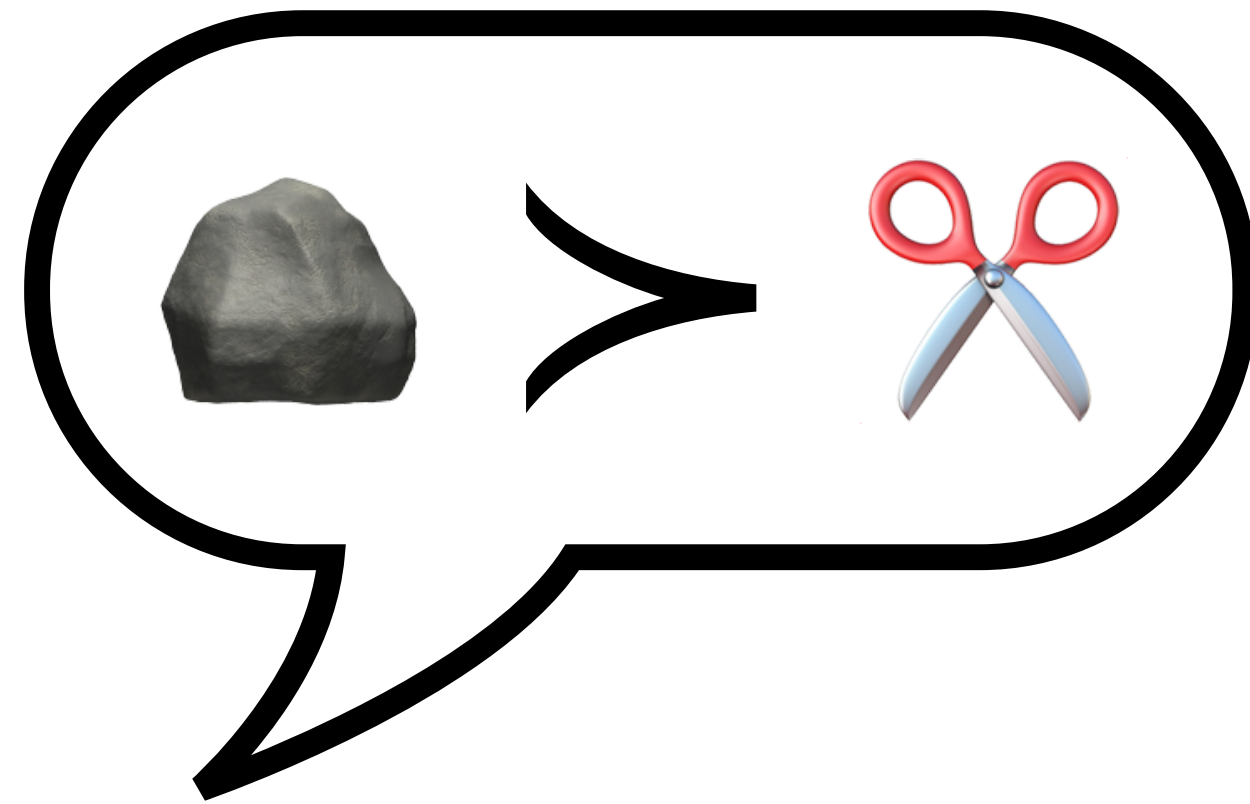
*Preference*  
 $\xi_1 > \xi_2$



# Setting: Preference Fine-Tuning from LLM Judge Feedback



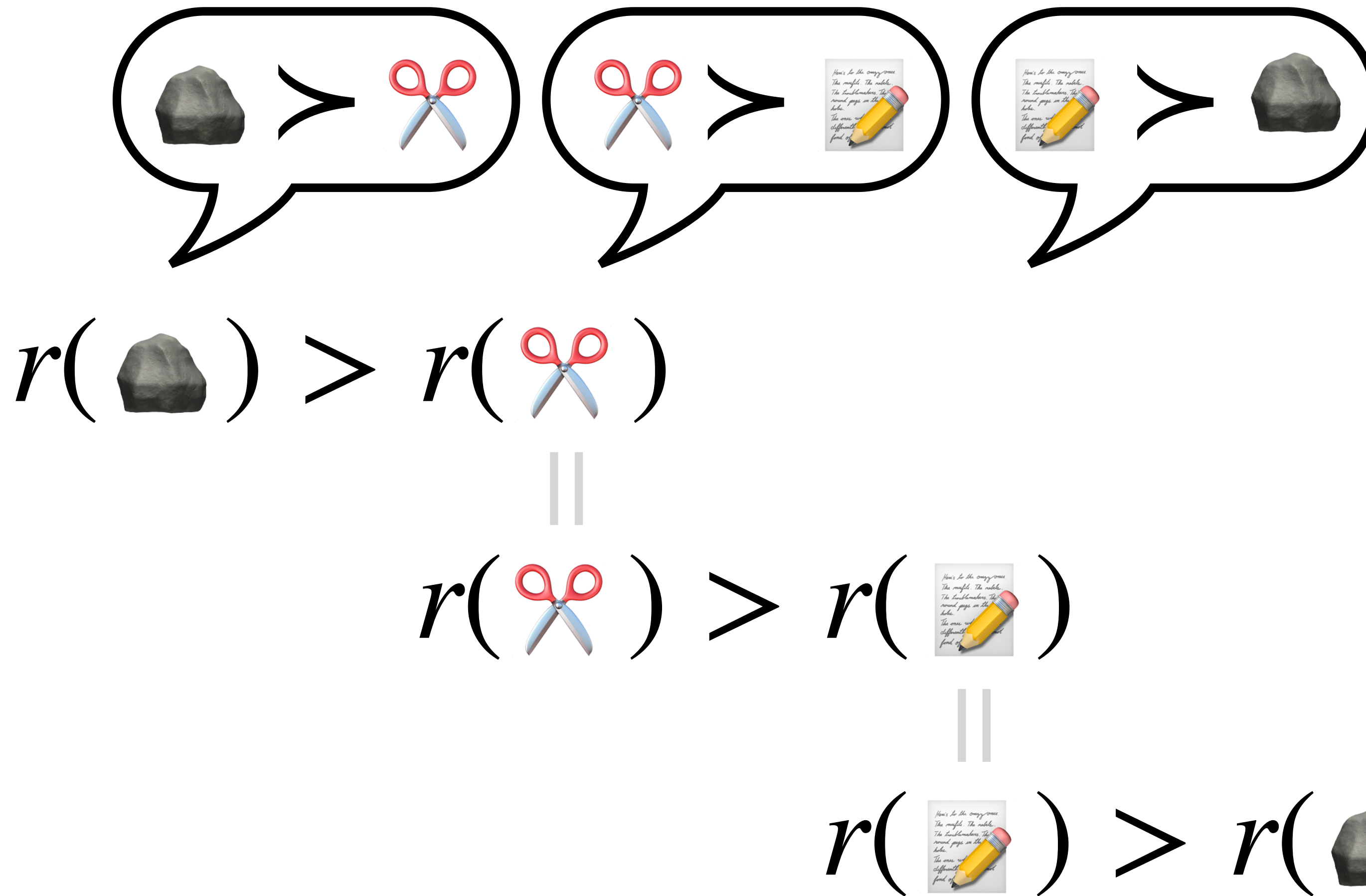
# 🔑 *Problem: Conflicting Preferences Make **Specification** Hard*



*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*



# 🔑 *Problem: Conflicting Preferences Make **Specification** Hard*

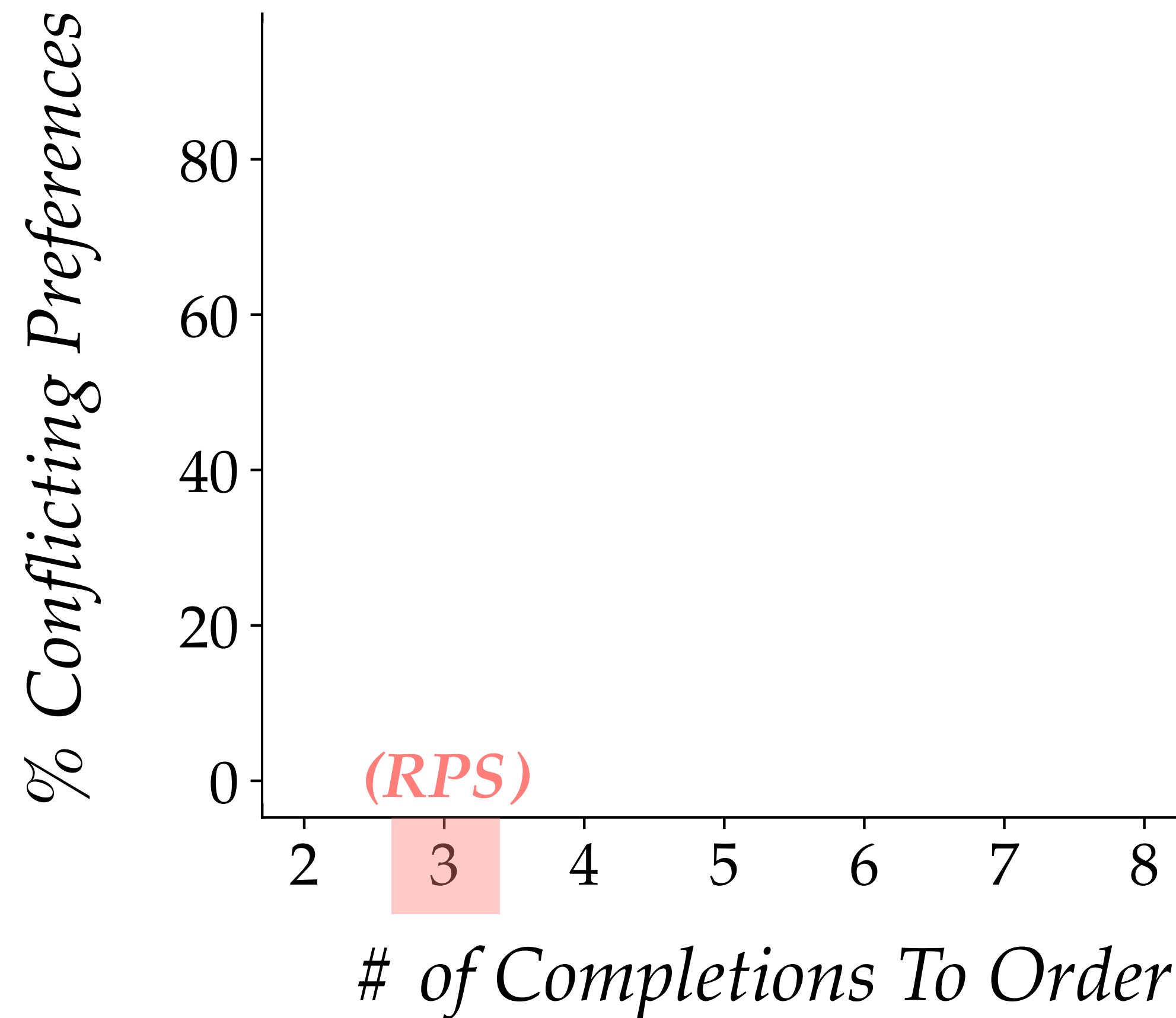


*There is no reward model that fits the data  $\Rightarrow$  no optimal policy!*

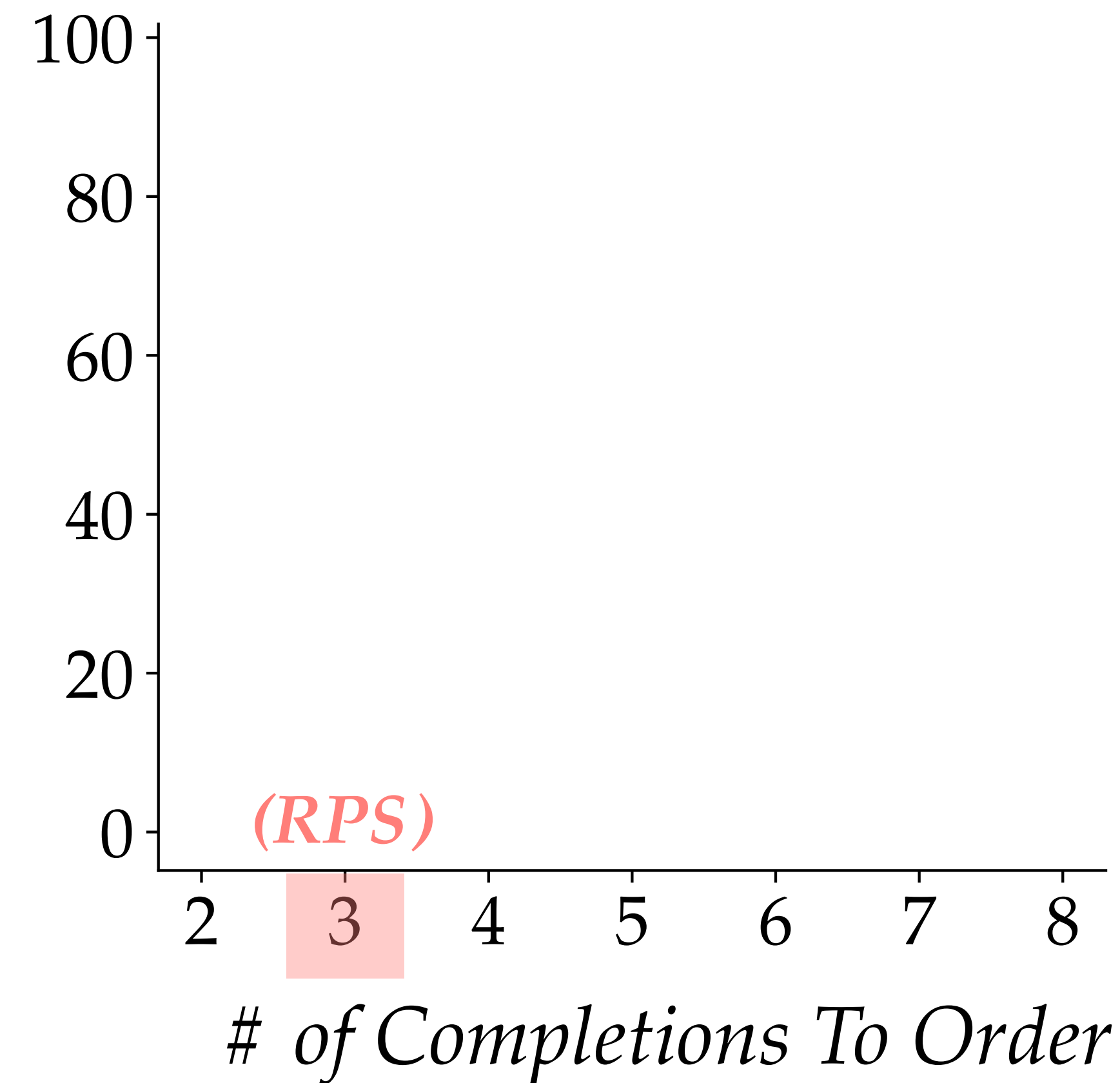
*Key Problem  $\longrightarrow$  Classical Solution  $\longrightarrow$  Efficient Algorithm  $\longrightarrow$  Practical Results*

# ⚠️ *Conflicting Preferences Are The Norm for LLM Judges!*

*Instruction Following*



*Multilingual Reasoning*

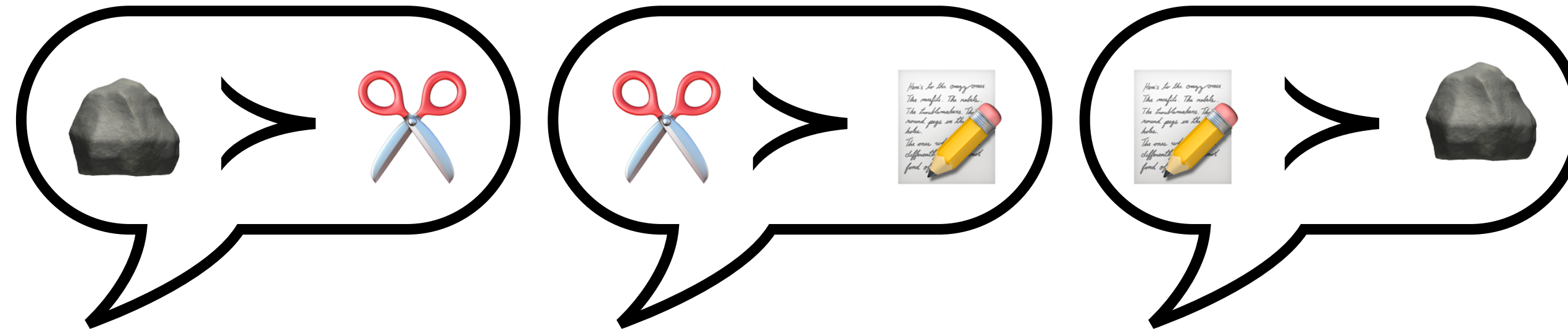


[Sutawika, Swamy et al. '26; Zhang, ..., Swamy, Wu '26]

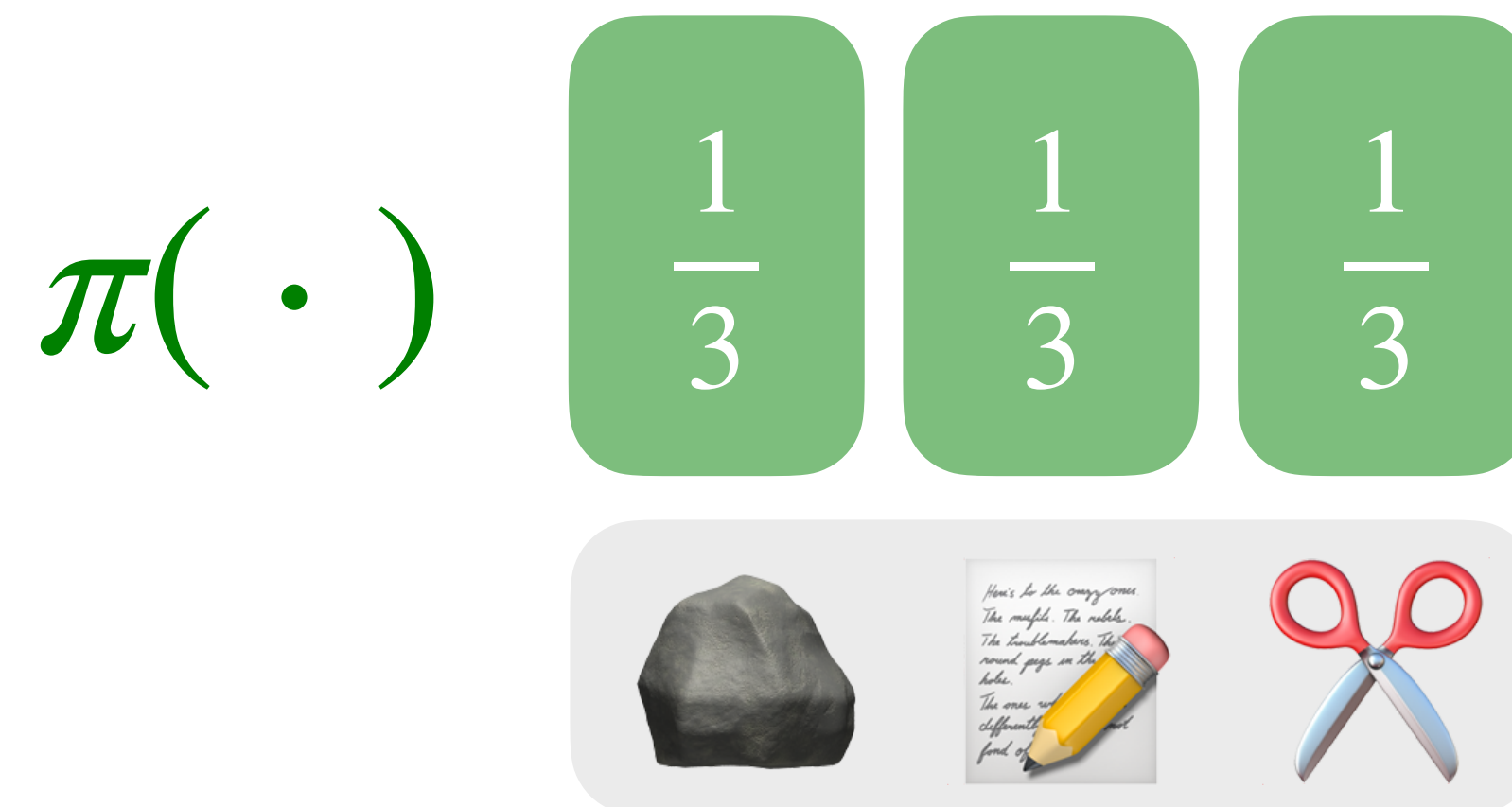
*Key Problem* → *Classical Solution* → *Efficient Algorithm* → *Practical Results*



# 🔑 *Problem: Conflicting Preferences Make **Specification** Hard*



*There is no optimal policy that always wins,  
but there is a policy that is hardest to beat!*



*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*

🔑 *Idea: Specification via the Minimax Winner*

$$\min_{\pi_1 \in \Pi} \max_{\pi_2 \in \Pi} \mathcal{P}(\pi_1, \pi_2)$$

where  $\mathcal{P}(\pi_1, \pi_2) \triangleq \mathbb{E}_{s_0 \sim \rho_0, \xi_1 \sim \pi_1 | s_0, \xi_2 \sim \pi_2 | s_0} [\mathcal{P}(\xi_1, \xi_2 | s_0)]$ .

*Robust to globally inconsistent judge preferences!*

*Aside: Also robust to heterogeneous human preferences  
+ deep connections to pluralistic alignment!*

[Kreweras '65, Fishburn '91, Dudik et al. '15, Swamy et al. '24]

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*



# *Problem: Adversarial Training is Hard*

$$\min_{\text{🌟}} \max_{\text{🌟}} \mathcal{P}_{\text{🌟}}(\text{🌟}, \text{🌟})$$

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*

🔑 *Idea: Computing Minimax Winners via Self-Play*

$$\min_{\text{⬢}} \max_{\text{⬢}} \mathcal{P}_{\text{⬢}}(\text{⬢}, \text{⬢})$$

[Swamy et al. '24]

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*



🔑 *Idea: Computing Minimax Winners via Self-Play*

$$\mathcal{P}(\text{orange star}, \text{red star}) = 1 - \mathcal{P}(\text{red star}, \text{orange star})$$

*Axiom of Probability!*

[Swamy et al. '24]

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*

# 🔑 *Idea: Computing Minimax Winners via Self-Play*

$$\min_{\text{orange}} \max_{\text{red}} \mathcal{P}(\text{orange}, \text{red})$$

$$= \max_{\text{red}} 1 - \mathcal{P}(\text{red}, \text{orange})$$

$$= \min_{\text{red}} \mathcal{P}(\text{red}, \text{orange})$$

*Proof Idea: Both players induce the same sequence of iterates, so we only need to keep track of one!*

[Swamy et al. '24]

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*

🔑 *Idea: Computing Minimax Winners via Self-Play*

$$\min_{\text{⬢}} \max_{\text{⬢}} \mathcal{P}_{\text{⬢}}(\text{⬢}, \text{⬢})$$

[Swamy et al. '24]

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*



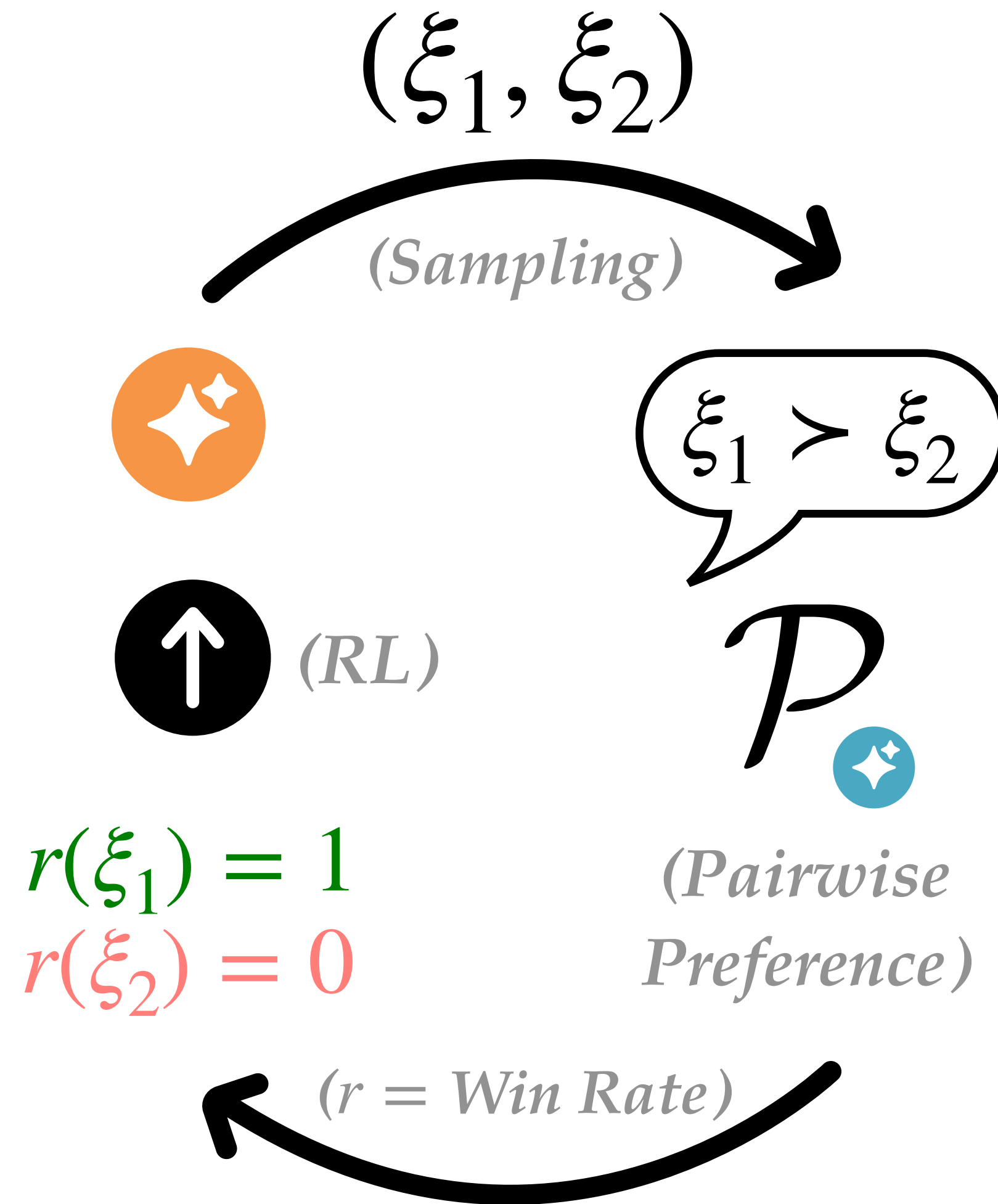
🔑 *Idea: Computing Minimax Winners via Self-Play*

$$\min_{\text{⬤}} \mathcal{P}_{\text{⬤}}(\text{⬤}, \text{⬤}_{t-1})$$

[Swamy et al. '24]

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*

# 🔑 *Idea: Computing Minimax Winners via Self-Play*



[Swamy et al. '24]

*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*

# *Informal Theorem: Self-Play Computes MWs*

Let  $\mathbb{A}_\pi$  be a no-regret algorithm with regret bound  $\text{Reg}_\pi(T)$ . For  $t \in [T]$ ,

$$r_t(\xi) \leftarrow \mathbb{E}_{s_0 \sim \rho_0, \xi' \sim \pi_t | s_0} [\mathcal{P}(\xi, \xi' | s_0)] \quad \text{Self-Play!}$$

$$\pi_{t+1} \leftarrow \mathbb{A}_\pi(r_{1:t})$$

Then,  $\bar{\pi}$  (the mixture of  $\pi_{1:T}$ ) is a  $2\text{Reg}_\pi(T)/T$ -approx. MW.

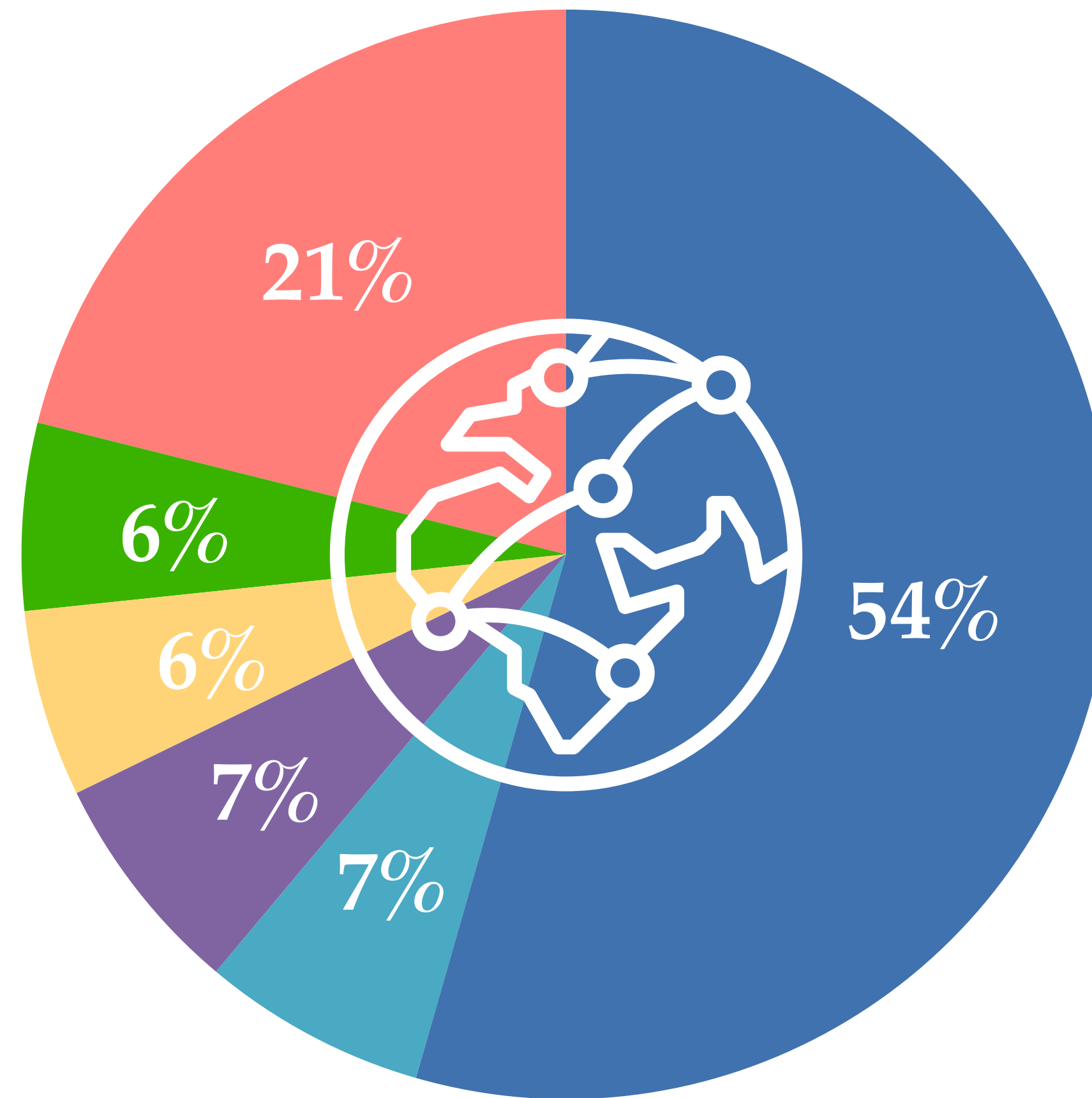
*Extensions to (1) handle multi-dimensional preferences (e.g., rubric rewards) via Blackwell Approachability, (2) coverage conditions on preference data for equilibrium computation, (3) gap-dependent fast rates under consistent preferences!*

[Swamy et al. '24]

*Key Problem  $\longrightarrow$  Classical Solution  $\longrightarrow$  Efficient Algorithm  $\longrightarrow$  Practical Results*



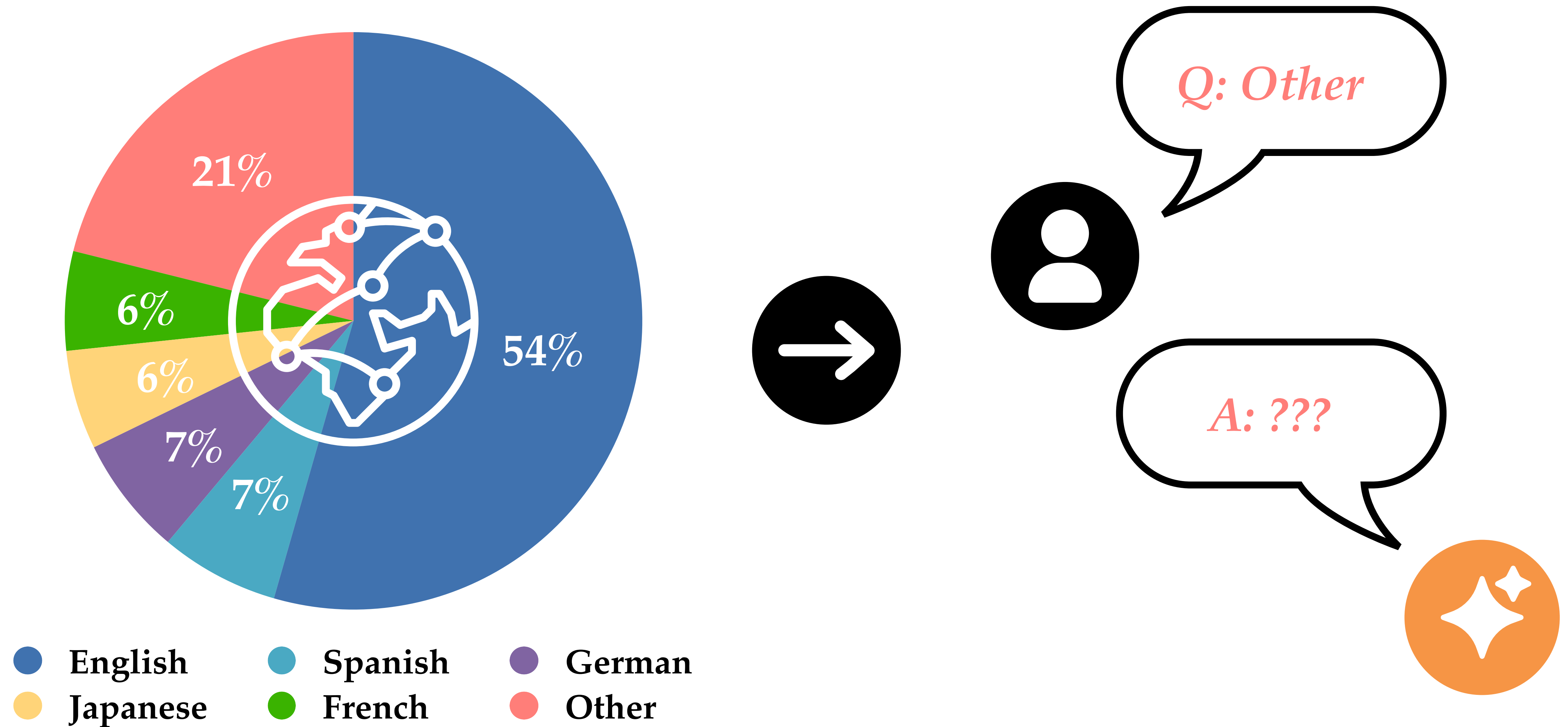
# Low Resource Language Reasoning via *Self-Play*



English Spanish German  
Japanese French Other

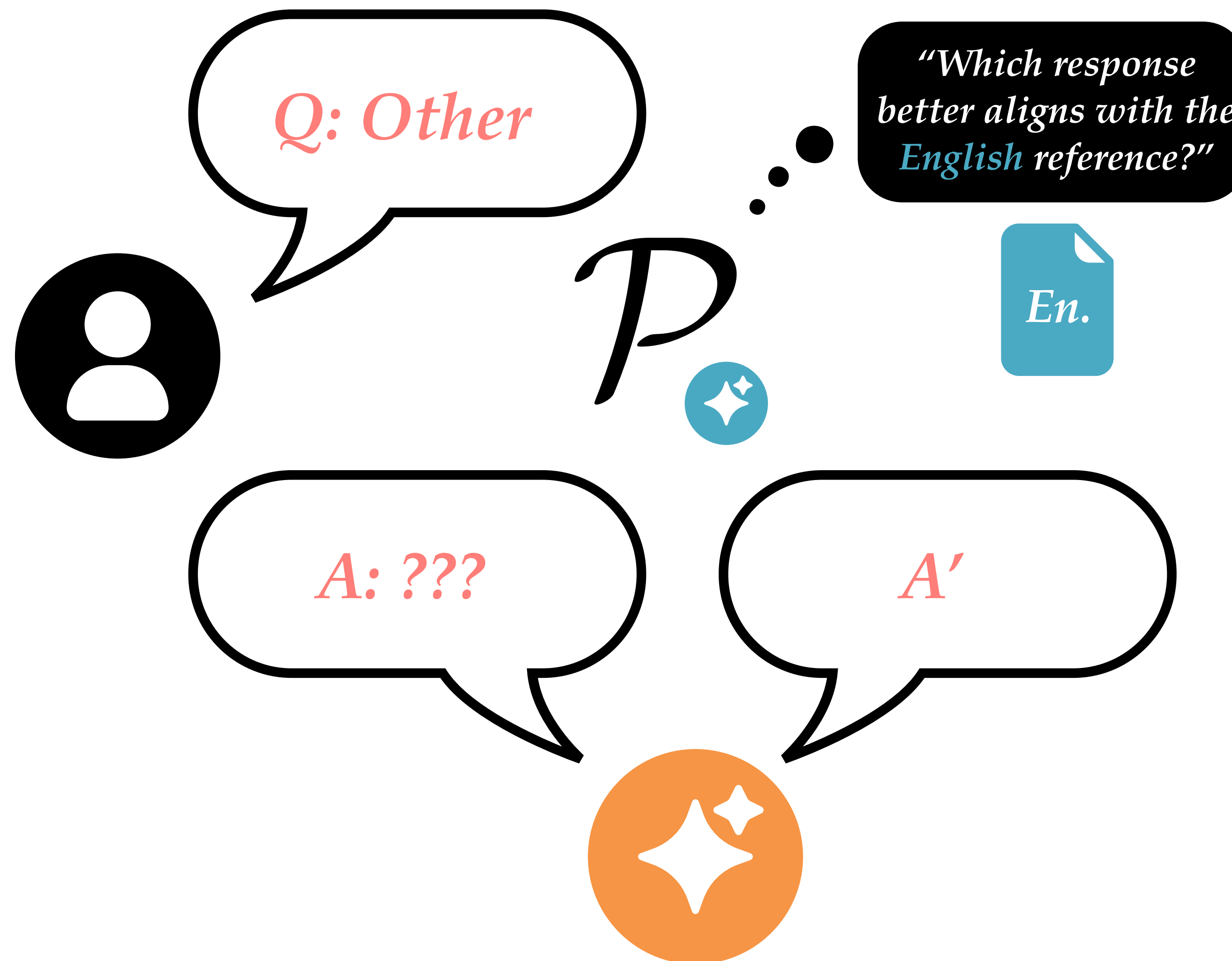
*Key Problem* → *Classical Solution* → *Efficient Algorithm* → *Practical Results*

# Low Resource Language Reasoning via *Self-Play*



*Key Problem* → *Classical Solution* → *Efficient Algorithm* → *Practical Results*

# Low Resource Language Reasoning via *Self-Play*



🔑 *Idea: Provide an English reference as privileged information to the LLM Judge!*

[Sutawika, Swamy et al. '26]

*Key Problem* → *Classical Solution* → *Efficient Algorithm* → *Practical Results*



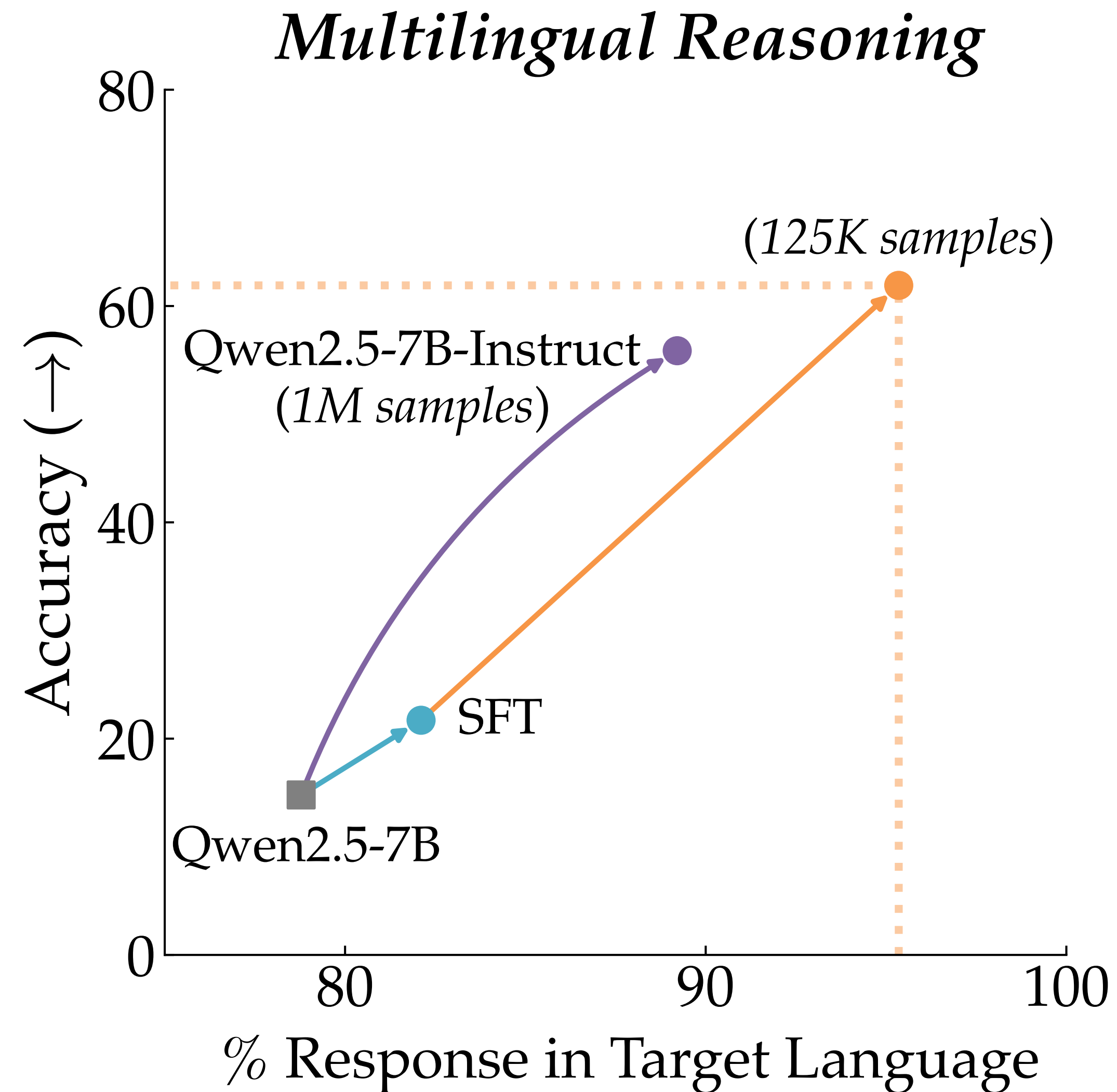
# *Self-Play* *Learns More From Less!*

% Response in Target Language

[Sutawika, Swamy et al. '26]

*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*

# *Self-Play* Learns More From Less!



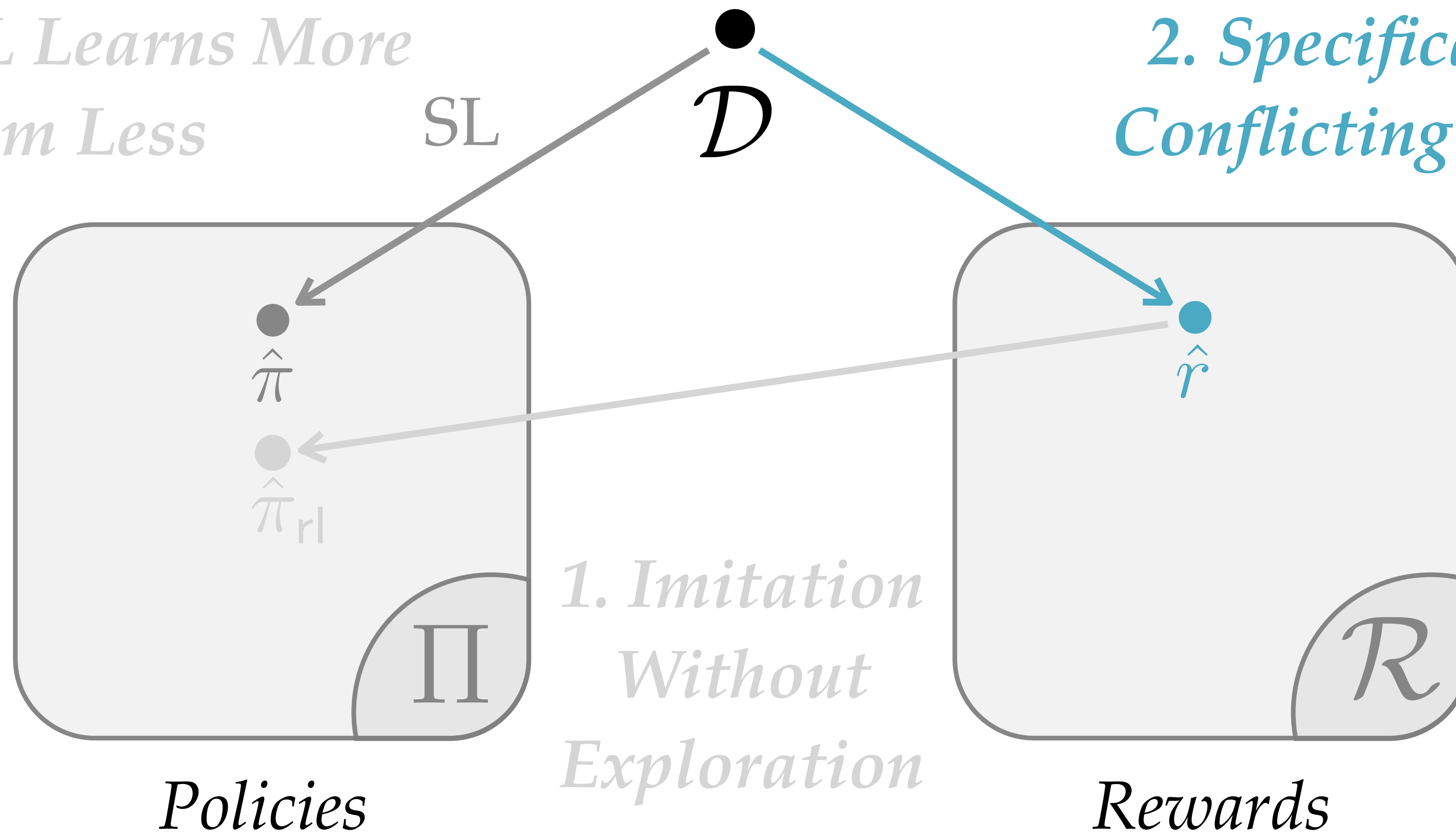
[Sutawika, Swamy et al. '26]

*Key Problem* → *Classical Solution* → *Efficient Algorithm* → *Practical Results*

# Today:

3. *Why RL Learns More From Less*

2. *Specification from Conflicting Preferences*

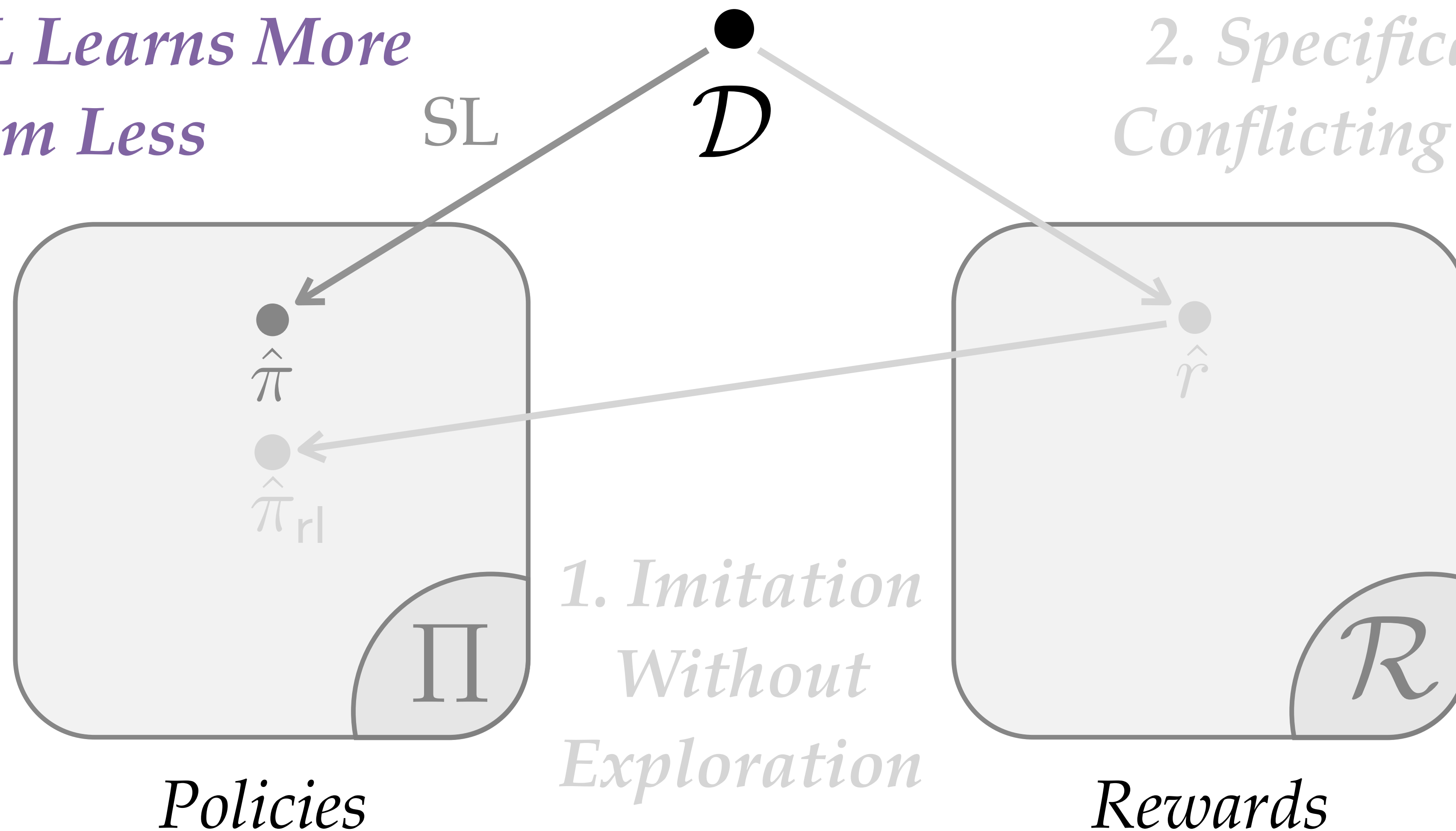


*Takeaway: self-play is robust to conflicting preferences from LLM judges*



*Today:*

*3. Why RL Learns More  
From Less*

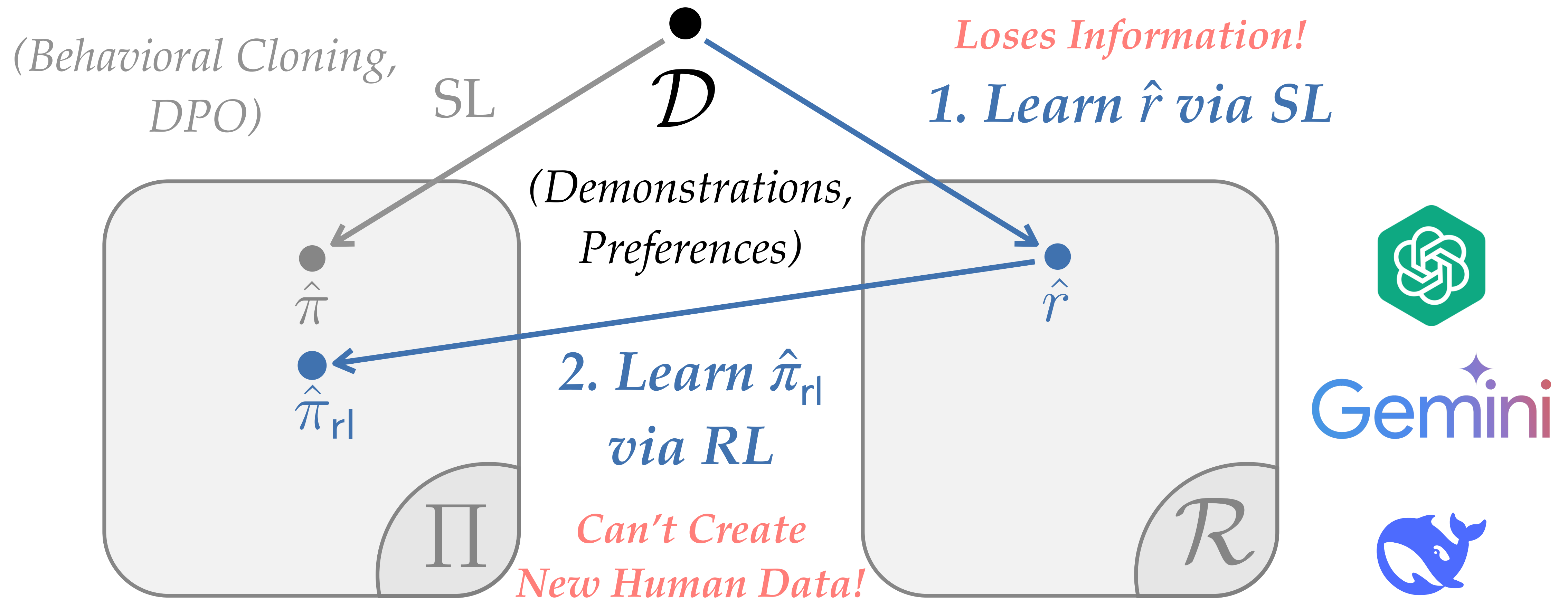


*2. Specification from  
Conflicting Preferences*

*1. Imitation  
Without  
Exploration*

*Theory* ← *Practice*

# 🔑 Q: Why does RL learn more from less?

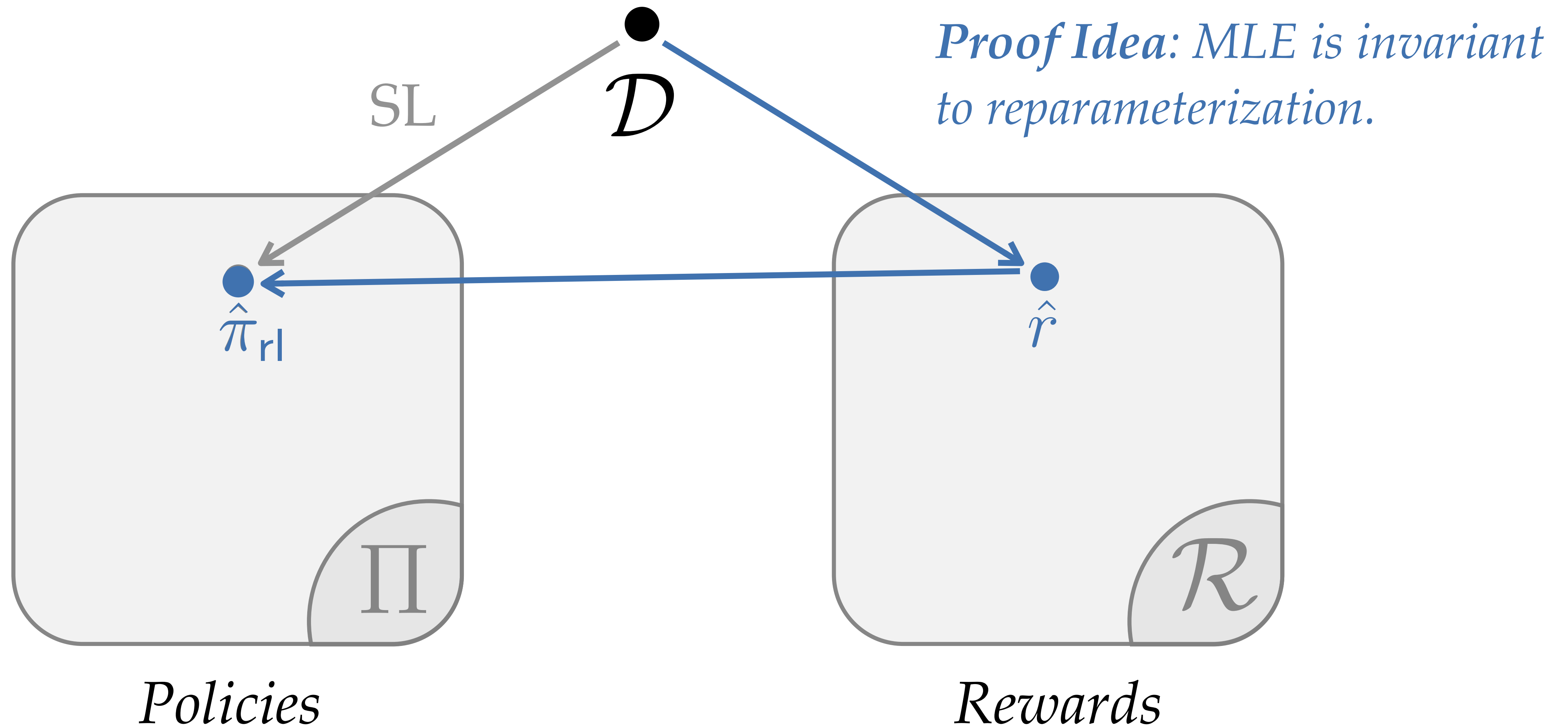


Policies *Hard Optimization Problem!* Rewards

*The Data Processing Inequality tells us this detour loses information!*

Key Question  $\longrightarrow$  Controlled Experiments  $\longrightarrow$  Hypothesis  $\longrightarrow$  Empirical Validation

# *Easier Q: When should this detour do nothing?*



**Informal Theorem:** If (1)  $\Pi \cong \mathcal{R}$  and (2) perfect optimization.

**Key Question**  $\longrightarrow$  **Controlled Experiments**  $\longrightarrow$  **Hypothesis**  $\longrightarrow$  **Empirical Validation**



# Controlled Experiments With $\Pi \cong \mathcal{R}$

1. Use same *architecture + base model* for  $\hat{\pi}$ ,  $\hat{r}$ , and  $\hat{\pi}_{rl}$
2. Use the *same dataset* to train  $\hat{\pi}$  and  $\hat{r}$ .
3. Use the *same optimizer + loss function* to train  $\hat{\pi}$  and  $\hat{\pi}_{rl}$ .



*SL*

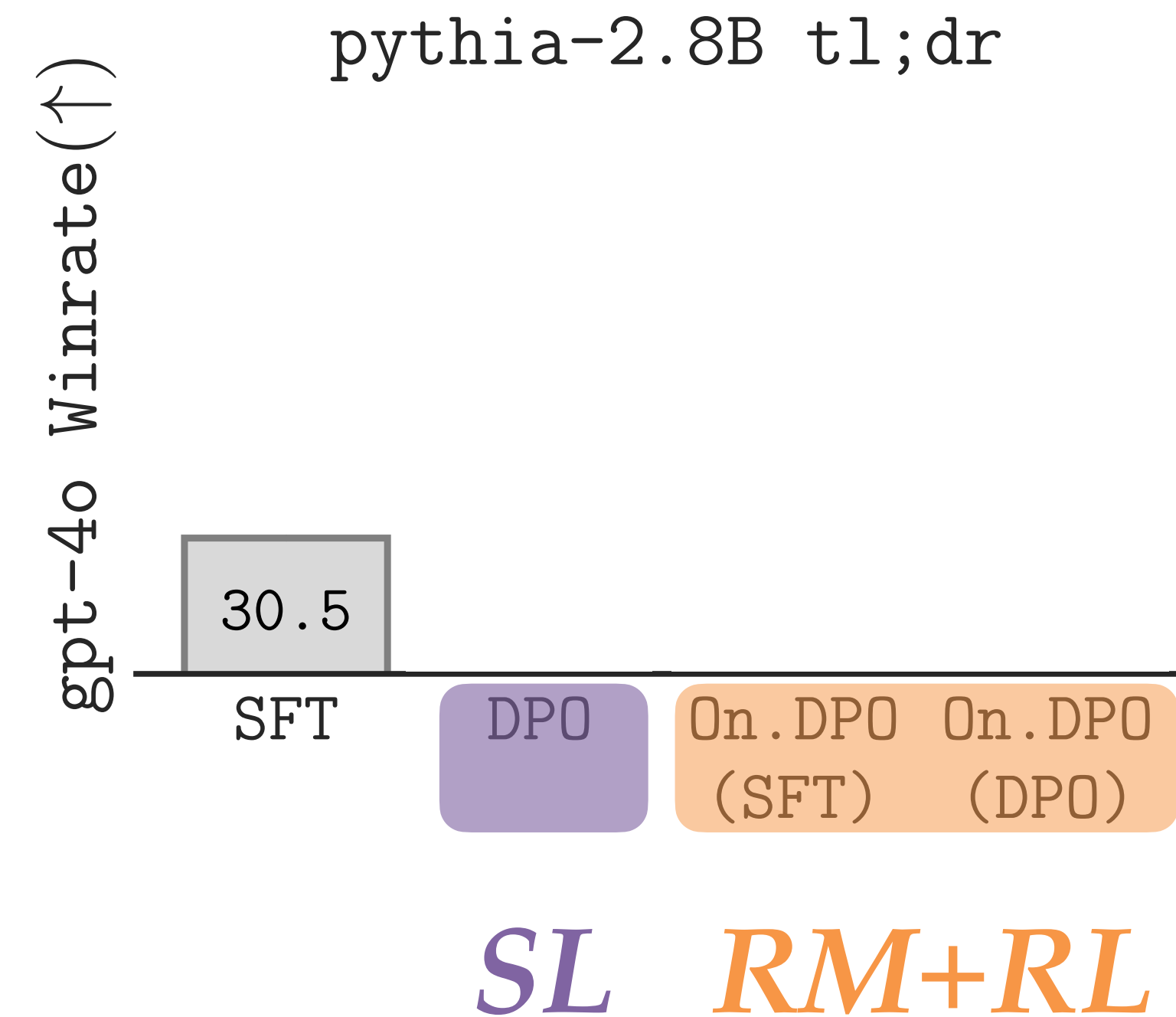
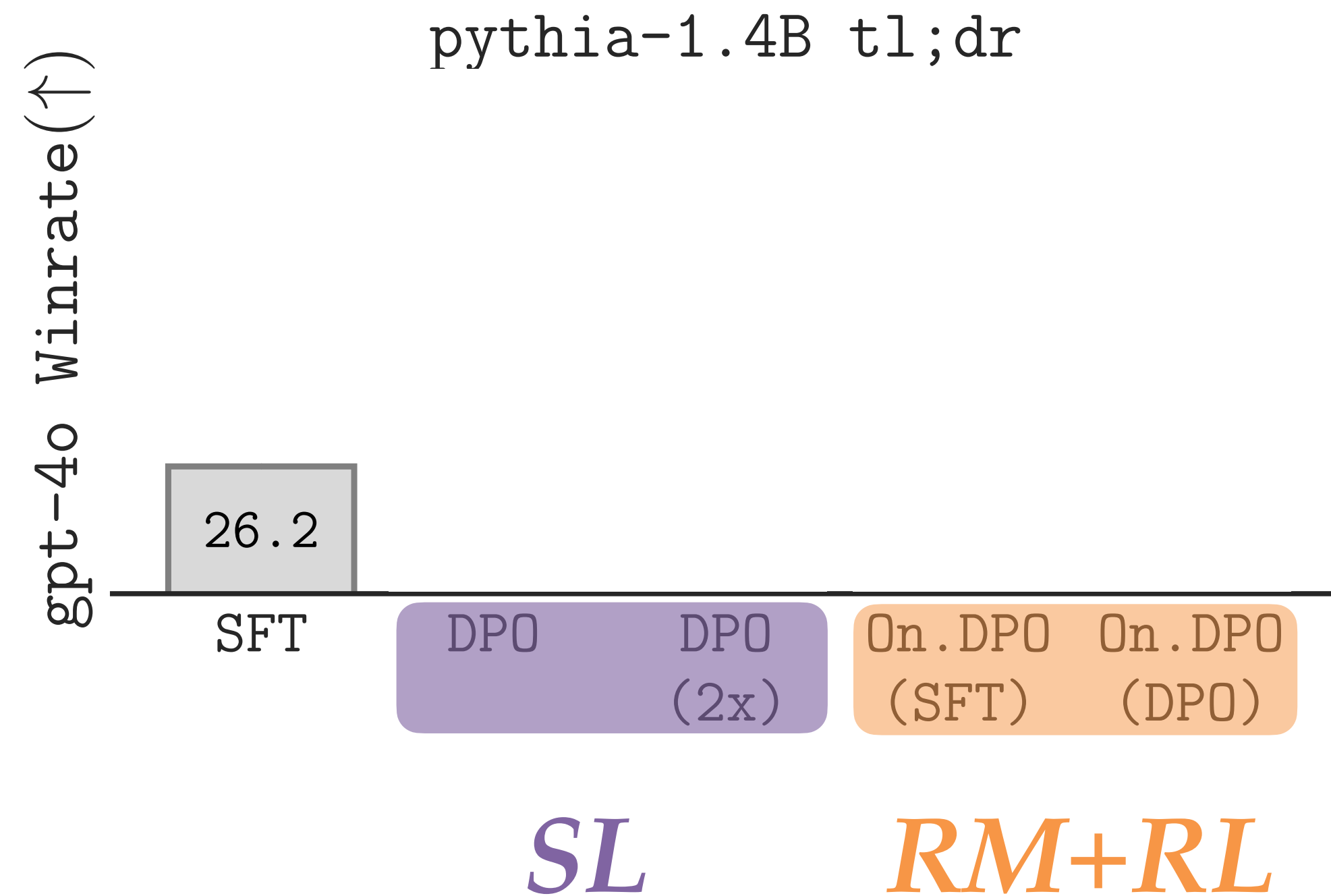
*U.S.*



*RM+RL*

*Key Question → Controlled Experiments → Hypothesis → Empirical Validation*

$$\Pi \cong \mathcal{R}, \text{ yet } SL < RM+RL$$



⇒ *We're seeing a gap between theory and practice!*

*Key Question → Controlled Experiments → Hypothesis → Empirical Validation*

# Hypotheses for the **SL** vs **RM + RL** Performance Gap

$\mathbb{H}_1$ : Intrinsic Value of Online Data

*... but we can't create data from nothing!*

$\mathbb{H}_2$ : Failure of Offline Regularization to  $\pi_{\text{ref}}$

*... but we see a gap with an explicit RKL regularizer!*

$\mathbb{H}_3$ : RMs Can Be Trained on Wider Datasets

*... but we see a gap with a narrower dataset!*

*Key Question  $\longrightarrow$  Controlled Experiments  $\longrightarrow$  Hypothesis  $\longrightarrow$  Empirical Validation*



# $\mathbb{H}_\star$ : *Generation-Verification Gaps:*

*Key Question  $\longrightarrow$  Controlled Experiments  $\longrightarrow$  Hypothesis  $\longrightarrow$  Empirical Validation*

$\mathbb{H}_\star$ : *Generation-Verification Gaps: Easier to Check Than to Solve*

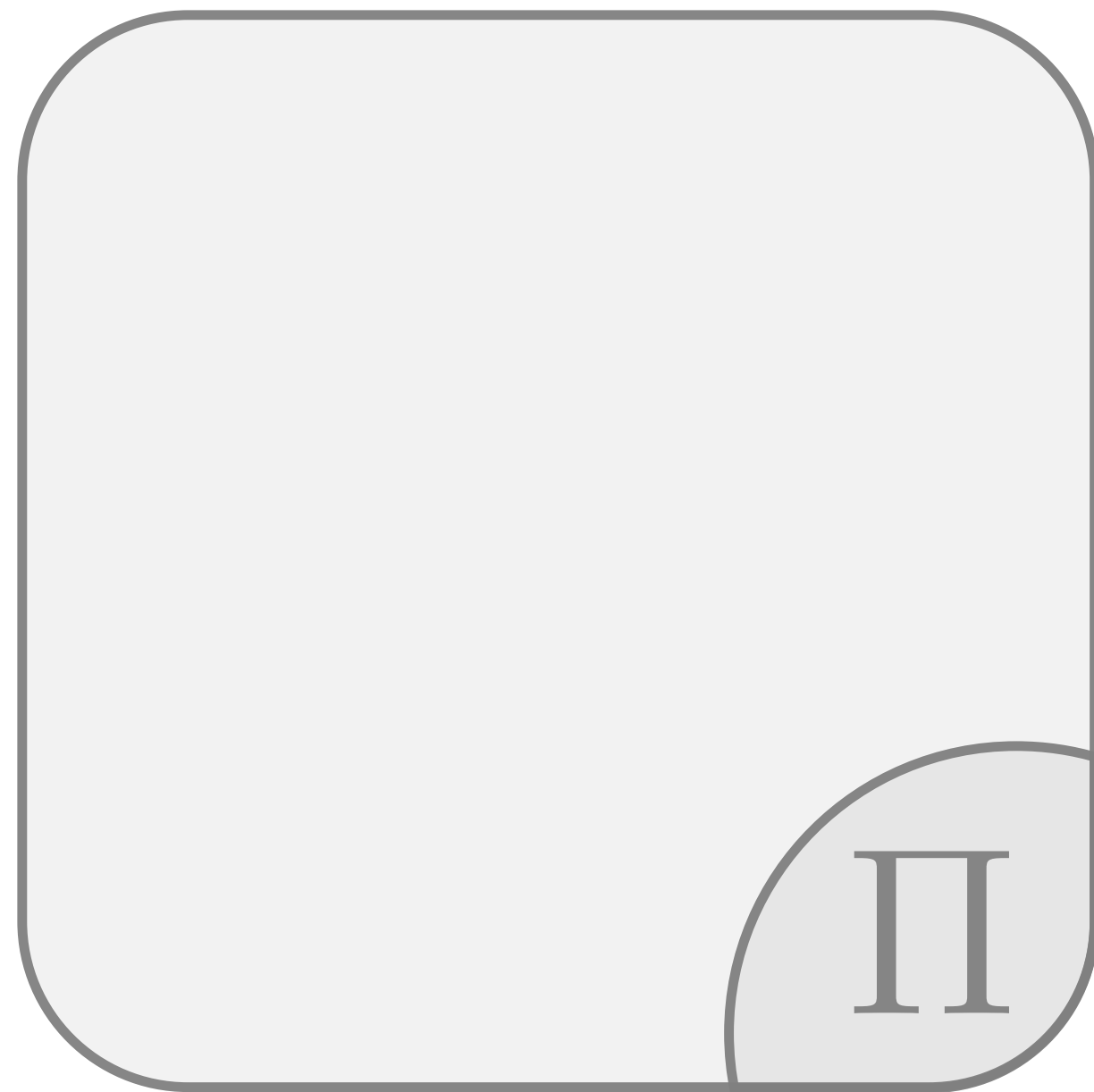
|   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|
|   |   | 6 |   |   |   | 5 |   | 8 |
| 1 |   | 2 | 3 | 8 |   |   |   | 4 |
|   |   |   | 2 |   |   | 1 | 9 |   |
|   |   |   |   | 6 | 3 |   | 4 | 5 |
|   | 6 | 3 | 4 |   | 5 | 8 | 7 |   |
| 5 | 4 |   | 9 | 2 |   |   |   |   |
|   | 8 | 7 |   |   | 4 |   |   |   |
| 2 |   |   |   | 9 | 8 | 4 |   | 7 |
| 4 |   | 9 |   |   |   | 3 |   |   |

*(e.g., problems in NP but not P)*

*Key Question → Controlled Experiments → Hypothesis → Empirical Validation*

# 🔑 *Idea: RL with a Generation-Verification Gap*

•  
 $\mathcal{D}$



*Policies*



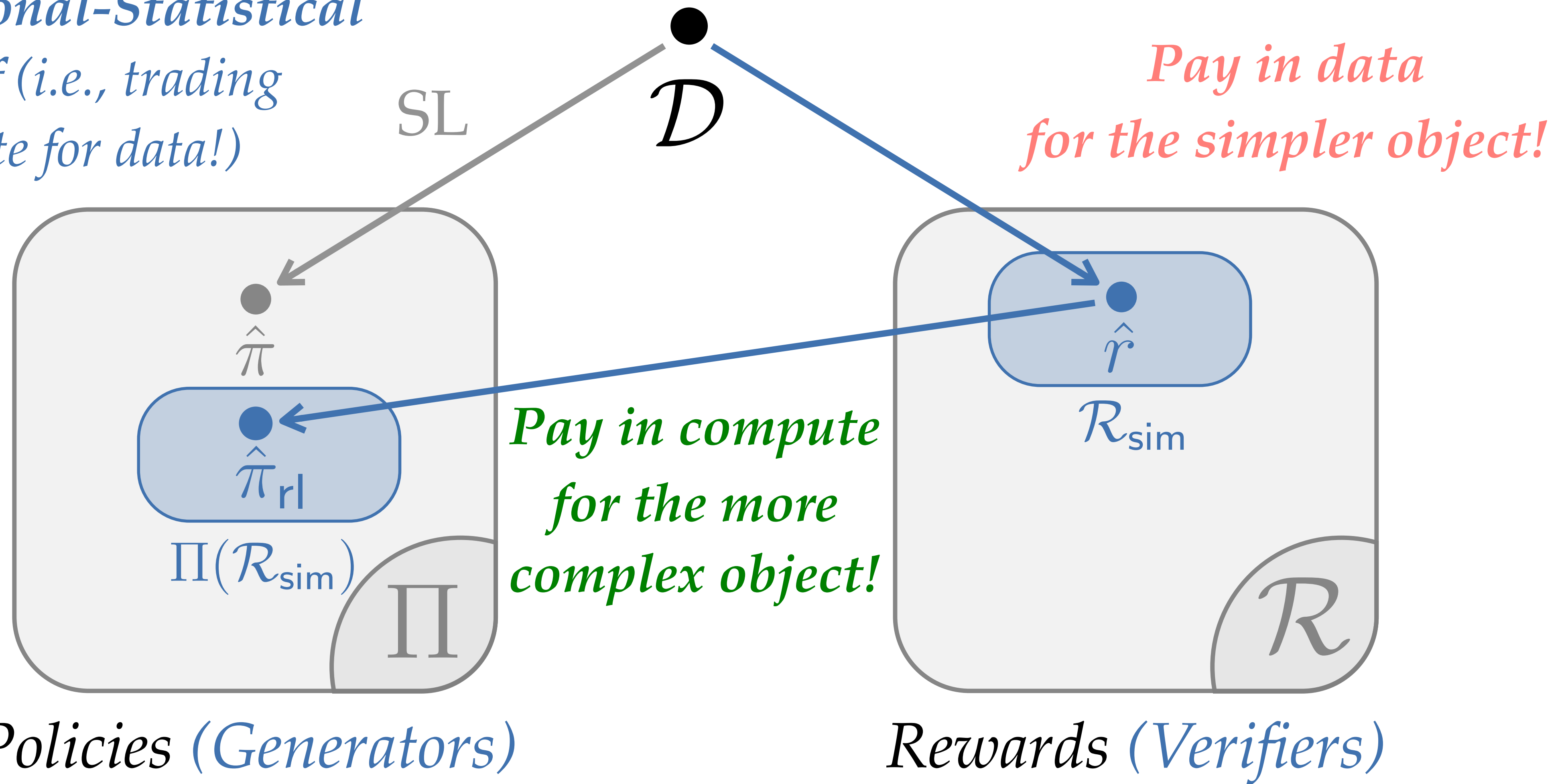
*Rewards*

*Key Question → Controlled Experiments → Hypothesis → Empirical Validation*





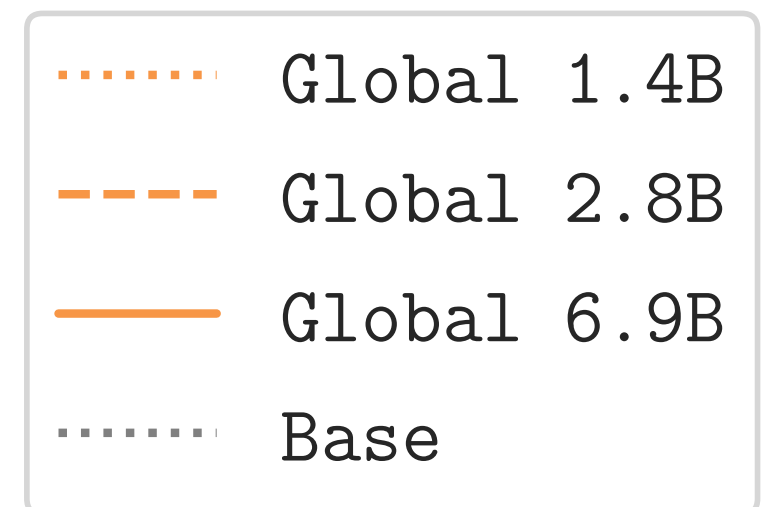
*Computational-Statistical  
Tradeoff (i.e., trading  
compute for data!)*



**Informal Theorem:** RM + RL is doing MLE over  $\Pi(\mathcal{R}_{\text{sim}})$ ,  
which requires less data than doing MLE over all of  $\Pi$ .

*Key Question  $\longrightarrow$  Controlled Experiments  $\longrightarrow$  Hypothesis  $\longrightarrow$  Empirical Validation*

# *Evidence of a Generation-Verification Gap: Scale Relative Sizes of Policies and Reward Models*



*Using a much smaller RM than policy doesn't hurt.    Using a much larger RM than policy doesn't help.*

*⇒ Reward Function Simpler Than Policy!*

*Key Question → Controlled Experiments → Hypothesis → Empirical Validation*

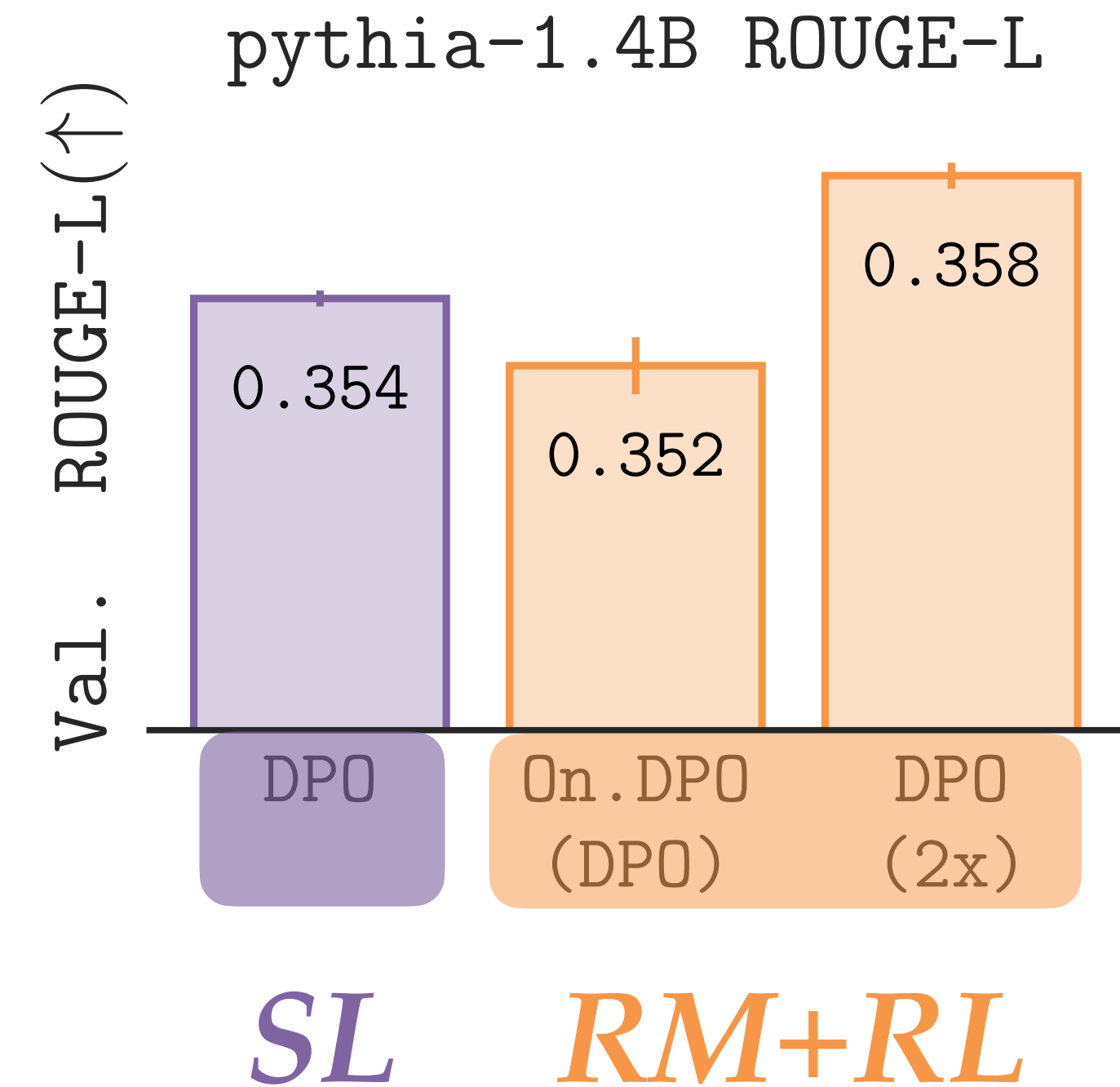
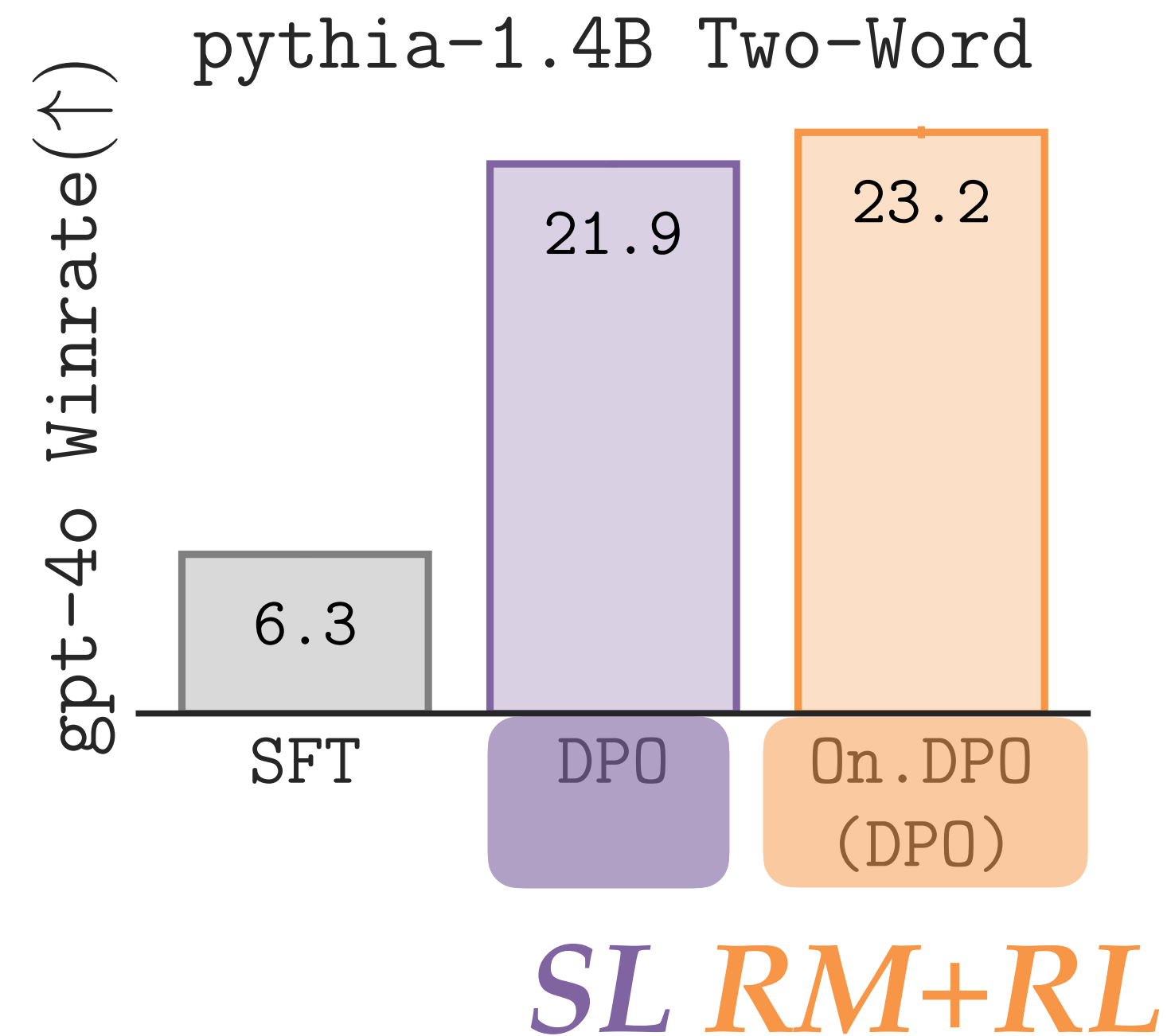


# *Testable Prediction:*

*No Generation-Verification Gap  $\Rightarrow$  **SL**  $\approx$  **RM+RL***

*Experiment 1: Simplify Policy*

*Experiment 2: Complicate Reward*

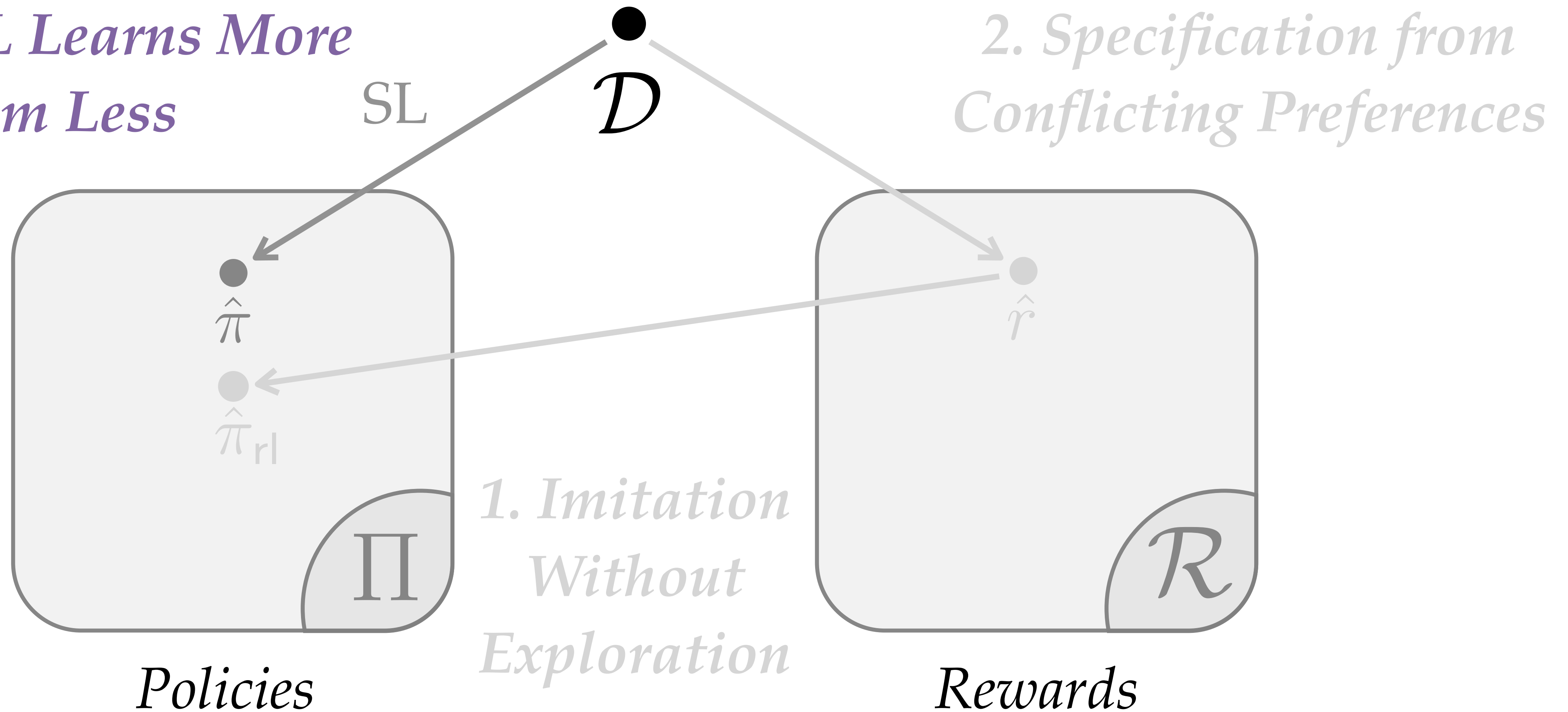


**SL  $\approx$  RM+RL Without A Generation-Verification Gap!**

*Key Question  $\longrightarrow$  Controlled Experiments  $\longrightarrow$  Hypothesis  $\longrightarrow$  Empirical Validation*

# Today:

## 3. Why RL Learns More From Less

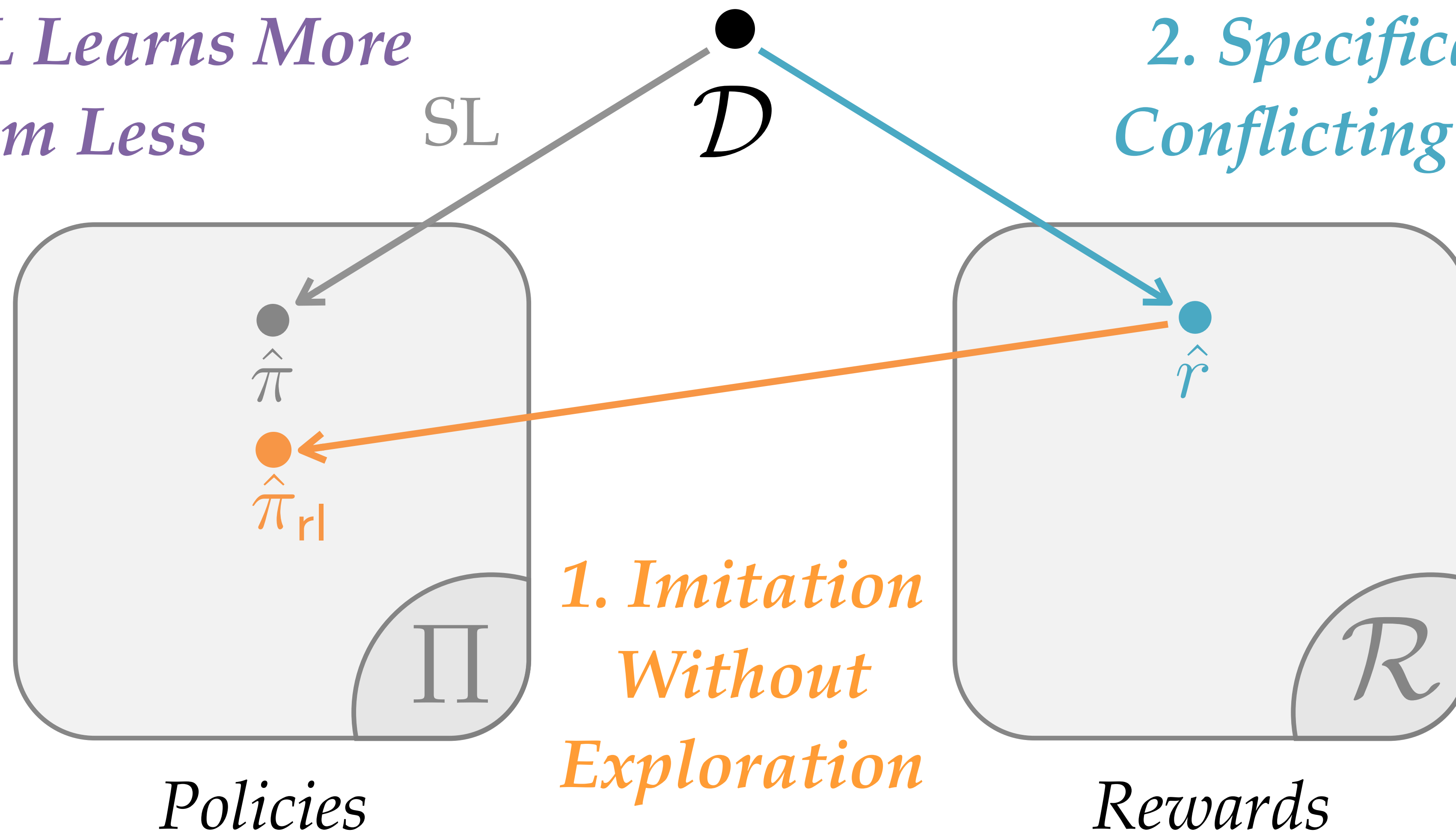


*Takeaway: RL learns more from less due to generation-verification gaps*

*Today:*

*3. Why RL Learns More  
From Less*

*2. Specification from  
Conflicting Preferences*



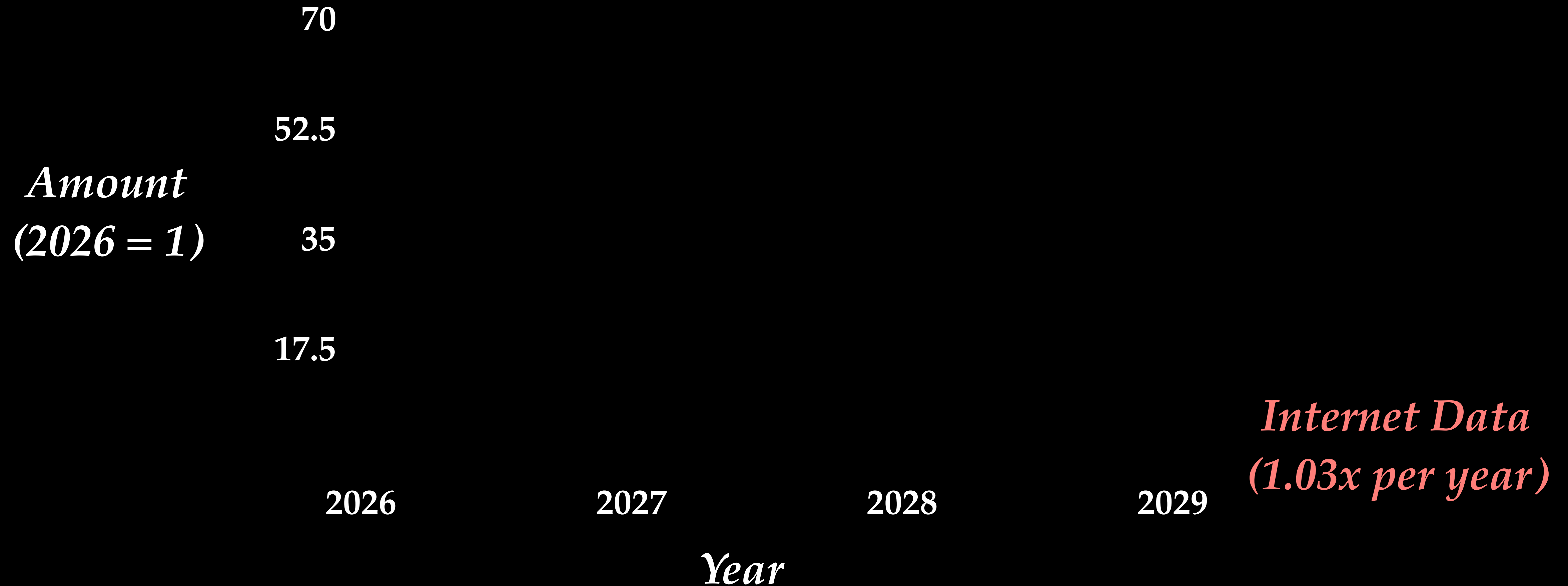








*Limited human data!*

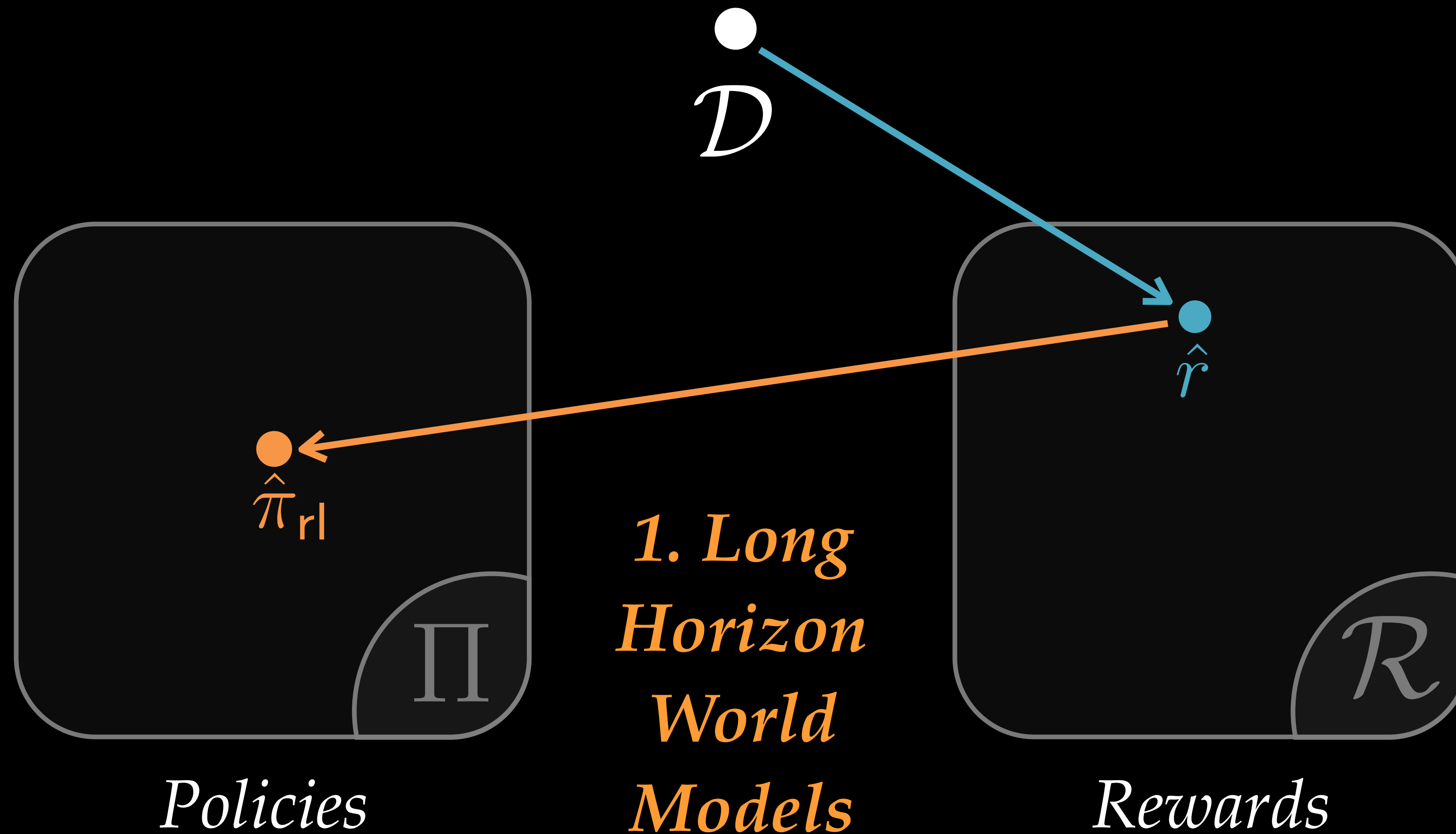


[Sevilla and Roldán, 2024,  
Villalobos et al., 2024]

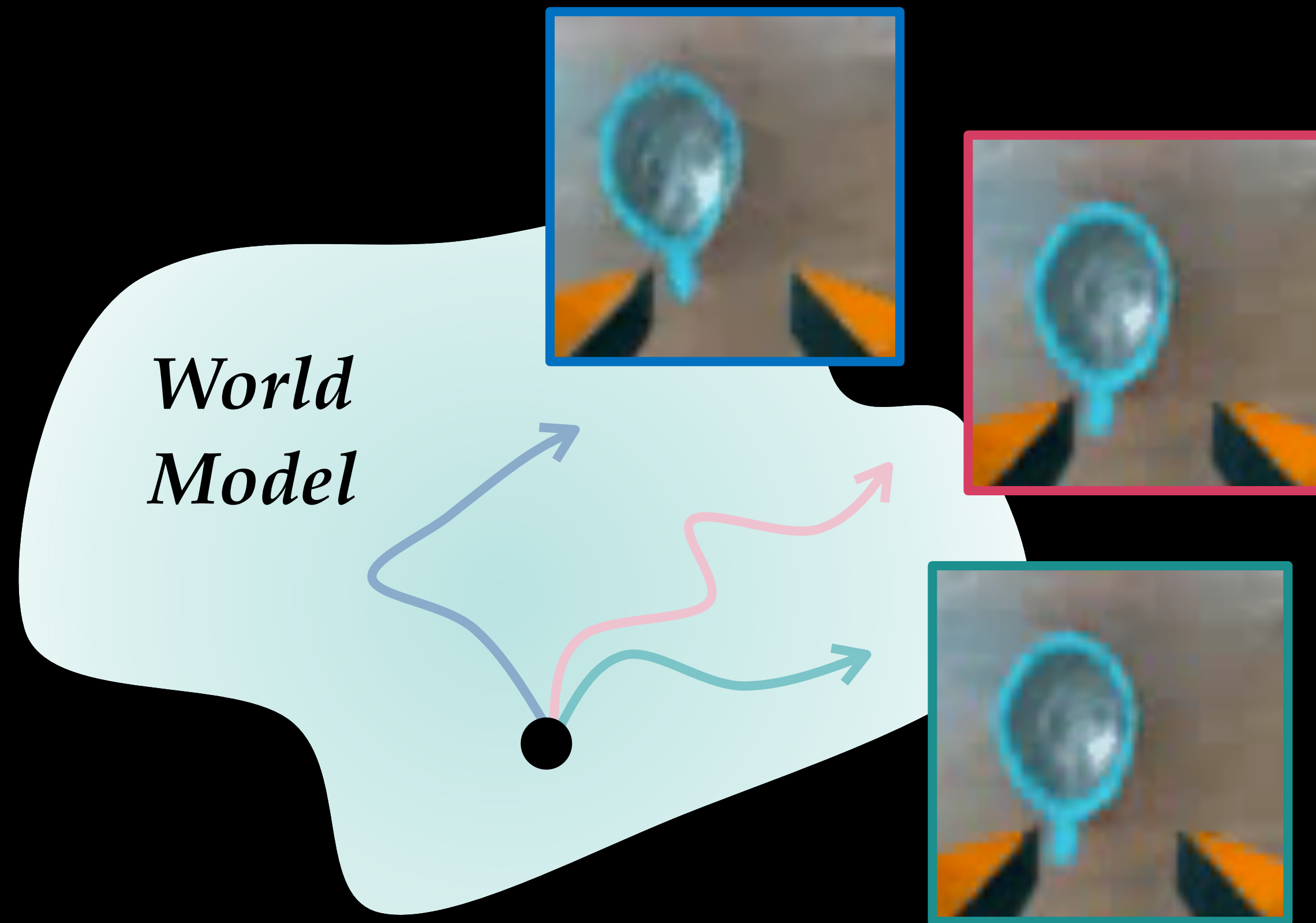
# *Future Work:*



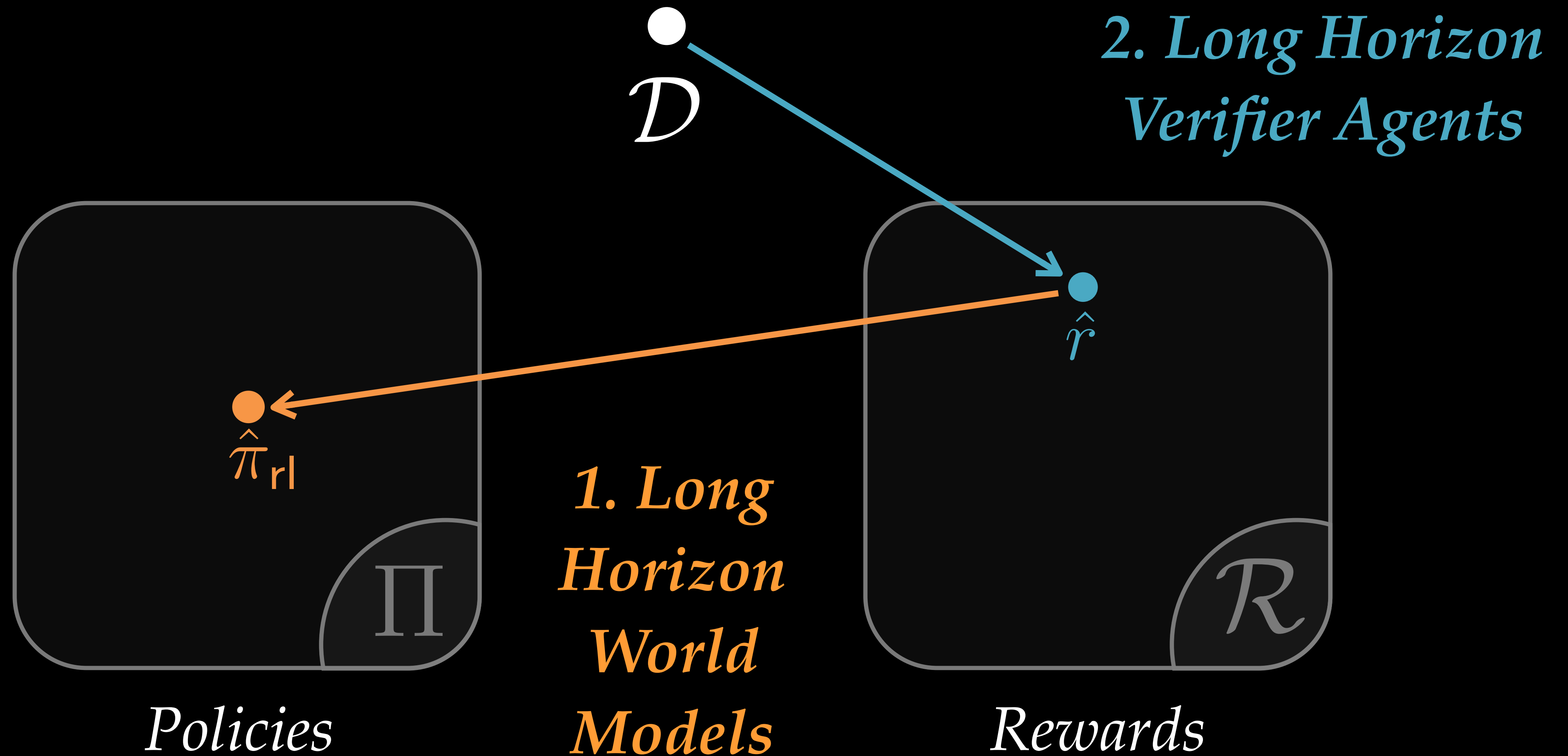
# *Future Work: Learning Even More From Less*



# *1. Long Horizon World Models*



# *Future Work: Learning Even More From Less*



*Theory Link: Decision Calibration*



## 2. *Long Horizon Verifier Agents*

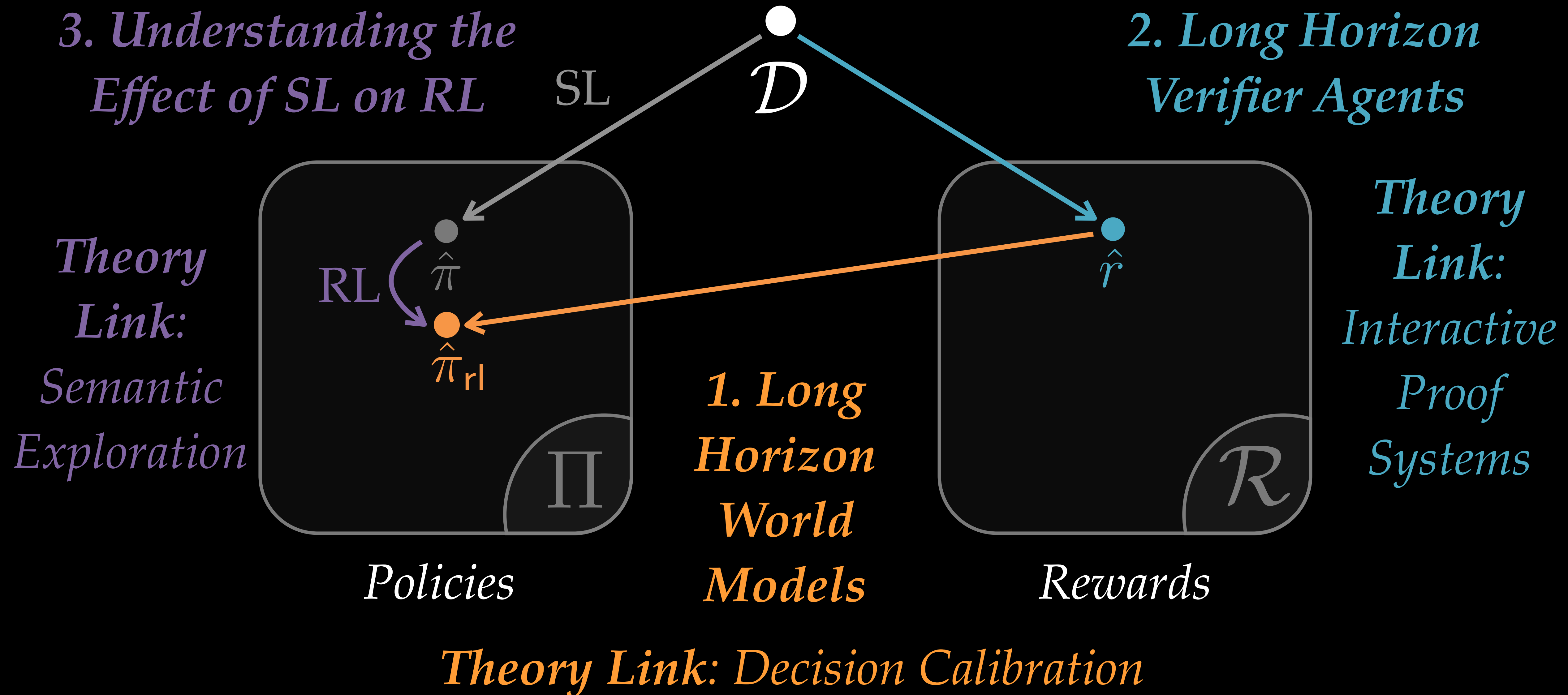
It was a slow night. The bartender polished the bar for the thirteenth time that night, wiping his rag back and forth, the lemony smell of the polish reaching through the alcohol. I lifted my glass as he swiped underneath, both of us used to the rhythm. There had been a lot of slow nights lately. Not as many people as usual in the old pub anymore. They liked the newer places, and I couldn't blame them. After all, when I'd first come here, it had been new. Now, though it was older than me, it had aged better. "Whiskey. Neat." The voice wasn't interesting, the request boring. But still, something piqued my curiosity. I turned, eyeing the new customer. They didn't seem to be anyone special. The customer raised an eyebrow at me, half-lifting their glass. "Cheers mate." I returned the motion, nodding to them. It was odd, when I'd first looked, I would have sworn they were a man, but the face seemed to change, shifting in the space of a blink. Now old, now young, now female, now male, now androgynous. I had kept track of how much I'd had to drink, and it wasn't enough to be seeing things. Knocking back the rest of my own whiskey, I slid off the stool. The stranger's words stopped me in my tracks. "Next round's on me." I looked back at them, and they offered me a quiet smile. There was something familiar about that smile. I'd seen it before; some people would call it a rictus grin. But they were offering to buy the next round. It would be churlish to refuse. Climbing back onto the stool, I held out a hand. "Name's Lyle. What's yours, friend?" The other customer tilted his head to the side, before extending their hand. Their touch was cold. "People have called me many things. Too many to count. Which would you prefer?" I sipped at my refilled glass, mulling the question over. "But I asked for your name. Not what people call you. After all, I've been called a lot of things too you know. But I don't go around calling myself 'Loser'." The customer chuckled, a low rolling laugh that had the timbre of an earthquake. They emptied their own glass, and before they could say anything, I signaled the bartender. He refilled with the poison of choice, and across from me, the figure nodded their thanks. "My name... It has been so long. I don't know if I remember my name. Did I even have one..." The face had shifted, sadness pulling away the last remnants of the grin. Summoning up my courage, though I didn't know why I should be scared, I clapped them on the back. "Buck up. You don't need a name to drink here, or chat with me. What brings you here on a Saturday night?" "Work. Always work." I shook my head at the response, raising my glass to the bartender. The other customer claimed the responsibility for this drink as well. They were fast rising in my estimation. "Look, mate. Work isn't everything. You have to balance your life. You know, have a day off here and there, find nice people to spend time with. All of that." A slow shake of the head came from my drinking companion. They sighed, and I caught the smell of an open grave. "Taking a break isn't really an option. But thank you for the advice." "Hey, all I'm saying is you should live your life. Have some fun, make some mistakes." "And you?" They raised their head, looking directly at me. "Have you lived your life?" It was my turn to chuckle at them, as I bought them another whiskey. "Oh, you have no idea. The things I've seen. There was this one time in France..." And as the night drew on, I regaled them with story after story. I had an audience that hadn't heard it all before and I was going to make the most of it. As we bought each other round after round, my companion loosened a little, even telling me a few tall tales of their own. But that wasn't the only thing we talked about. By the time the bartender warned us of last call, we were deep into a discussion on a matter of philosophy. I paid my tab and nodded to the other customer. "Well, if you're ever in this neck of the woods again, don't be a stranger. We can have another good chat." There was a return of the sadness from earlier, as they shook their head. "I'm afraid that won't be possible. You asked me what my name was earlier. I told you that people call me many things. And what they call me is—" "Death?" I laughed at the consternation on their face. "I figured it out a while ago. I was sort of expecting it soon. Ticker's not what it used to be. But hey. You can't be all bad. You bought me drinks, sat and listened to my stories, and told a few of your own." I turned, as we reached the doorway, looking back into the pub. "Funny how I ended up here tonight. Never been in this place before, but somehow it felt right. Like I was meant to spend my last evening here, sharing stories with a stranger who turned out to be Death himself. Not a bad way to go, all things considered." Death stared down at me, their ever-changing appearance no longer a disturbing sight. They held out a hand, their voice soft. "I am glad. But now, it is time to go." I settled my cap in its usual position on my head, wrapped my scarf around my neck and shrugged into my coat. And taking Death's hand, I walked into the night, unafraid.



## 2. *Long Horizon Verifier Agents*

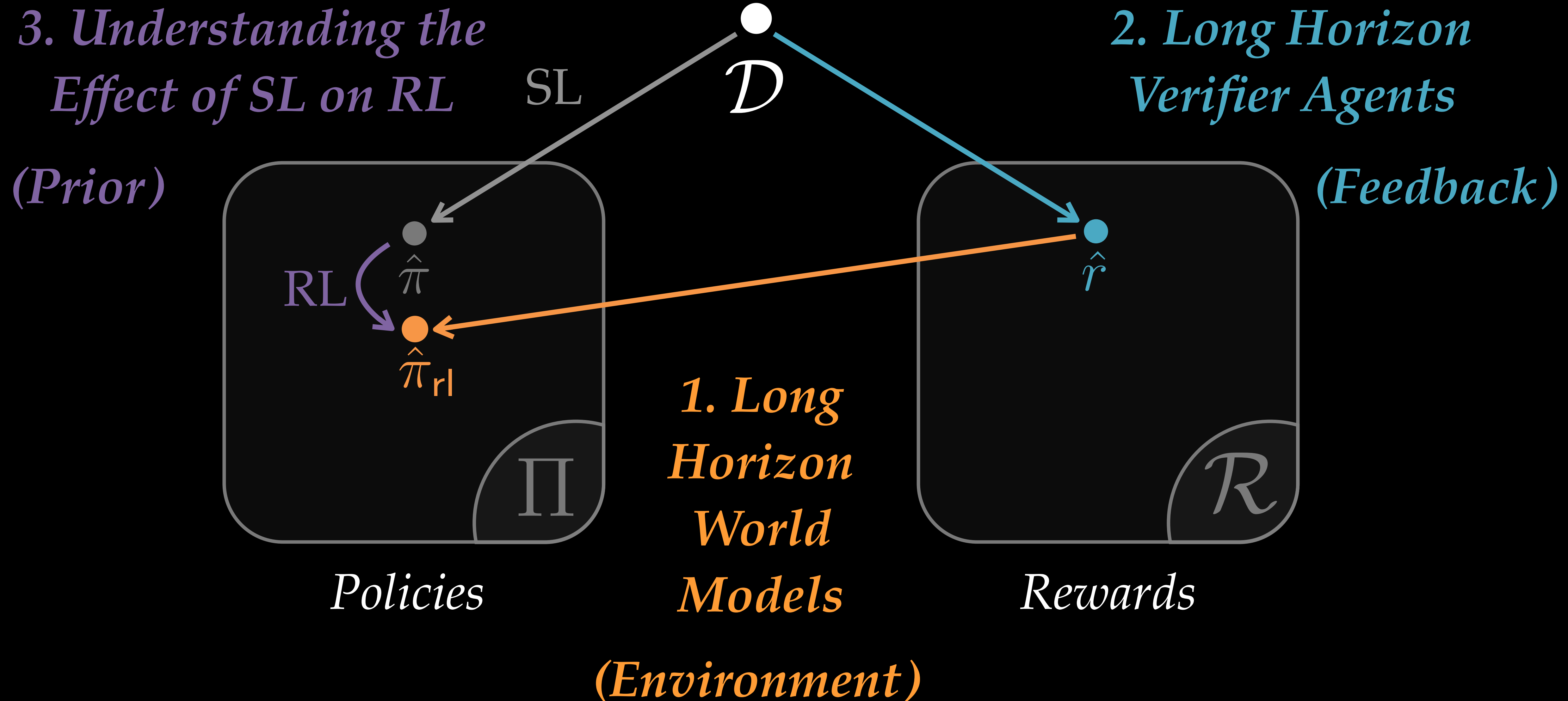
It was a slow night. The bartender polished the bar for the thirteenth time that night, wiping his rag back and forth, the lemony smell of the polish reaching through the alcohol. I lifted my glass as he swiped underneath, both of us used to the rhythm. *There had been a lot of slow nights lately. Not as many people as usual in the old pub anymore. They liked the newer places, and I couldn't blame them. After all, when I'd first come here, it had been new.* Now, though it was older than me, it had aged better. "Whiskey. Neat." The voice wasn't interesting, the request boring. But still, something piqued my curiosity. I turned, eyeing the new customer. They didn't seem to be anyone special. The customer raised an eyebrow at me, half-lifting their glass. "Cheers mate." I returned the motion, nodding to them. It was odd, when I'd first looked, I would have sworn they were a man, but the face seemed to change, shifting in the space of a blink. Now old, now young, now female, now male, now androgynous. I had kept track of how much I'd had to drink, and it wasn't enough to be seeing things. Knocking back the rest of my own whiskey, I slid off the stool. The stranger's words stopped me in my tracks. "Next round's on me." I looked back at them, and they offered me a quiet smile. There was something familiar about that smile. I'd seen it before; some people would call it a rictus grin. But they were offering to buy the next round. It would be churlish to refuse. Climbing back onto the stool, I held out a hand. "Name's Lyle. What's yours, friend?" The other customer tilted his head to the side, before extending their hand. Their touch was cold. "People have called me many things. Too many to count. Which would you prefer?" I sipped at my refilled glass, mulling the question over. "But I asked for your name. Not what people call you. After all, I've been called a lot of things too you know. But I don't go around calling myself 'Loser'." The customer chuckled, a low rolling laugh that had the timbre of an earthquake. They emptied their own glass, and before they could say anything, I signaled the bartender. He refilled with the poison of choice, and across from me, the figure nodded their thanks. "My name... It has been so long. I don't know if I remember my name. Did I even have one..." The face had shifted, sadness pulling away the last remnants of the grin. Summoning up my courage, though I didn't know why I should be scared, I clapped them on the back. "Buck up. You don't need a name to drink here, or chat with me. What brings you here on a Saturday night?" "Work. Always work." I shook my head at the response, raising my glass to the bartender. The other customer claimed the responsibility for this drink as well. They were fast rising in my estimation. "Look, mate. Work isn't everything. You have to balance your life. You know, have a day off here and there, find nice people to spend time with. All of that." A slow shake of the head came from my drinking companion. They sighed, and I caught the smell of an open grave. "Taking a break isn't really an option. But thank you for the advice." "Hey, all I'm saying is you should live your life. Have some fun, make some mistakes." "And you?" They raised their head, looking directly at me. "Have you lived your life?" It was my turn to chuckle at them, as I bought them another whiskey. "Oh, you have no idea. The things I've seen. There was this one time in France..." And as the night drew on, I regaled them with story after story. I had an audience that hadn't heard it all before and I was going to make the most of it. As we bought each other round after round, my companion loosened a little, even telling me a few tall tales of their own. But that wasn't the only thing we talked about. By the time the bartender warned us of last call, we were deep into a discussion on a matter of philosophy. I paid my tab and nodded to the other customer. "Well, if you're ever in this neck of the woods again, don't be a stranger. We can have another good chat." There was a return of the sadness from earlier, as they shook their head. "I'm afraid that won't be possible. You asked me what my name was earlier. I told you that people call me many things. And what they call me is—" "Death?" I laughed at the consternation on their face. "I figured it out a while ago. I was sort of expecting it soon. Ticker's not what it used to be. But hey. You can't be all bad. You bought me drinks, sat and listened to my stories, and told a few of your own." I turned, as we reached the doorway, looking back into the pub. *Funny how I ended up here tonight. Never been in this place before, but somehow it felt right.* Like I was meant to spend my last evening here, sharing stories with a stranger who turned out to be Death himself. Not a bad way to go, all things considered." Death stared down at me, their ever-changing appearance no longer a disturbing sight. They held out a hand, their voice soft. "I am glad. But now, it is time to go." I settled my cap in its usual position on my head, wrapped my scarf around my neck and shrugged into my coat. And taking Death's hand, I walked into the night, unafraid.

# *Future Work: Learning Even More From Less*





# *Future Work: Learning Even More From Less*



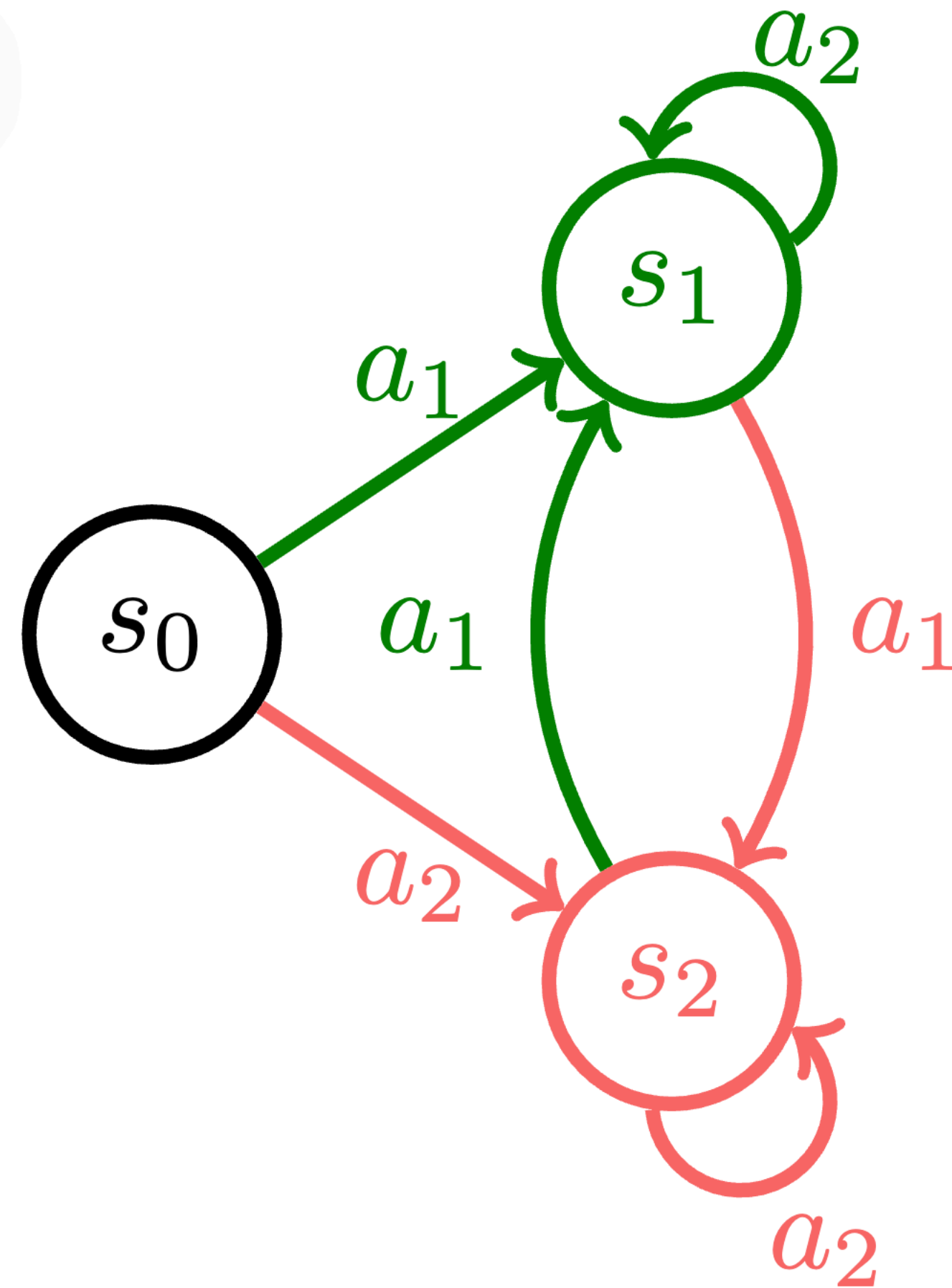


# *Let's Collaborate!*





# 🔑 *Problem: Supervised Imitation Doesn't Teach Recovery*



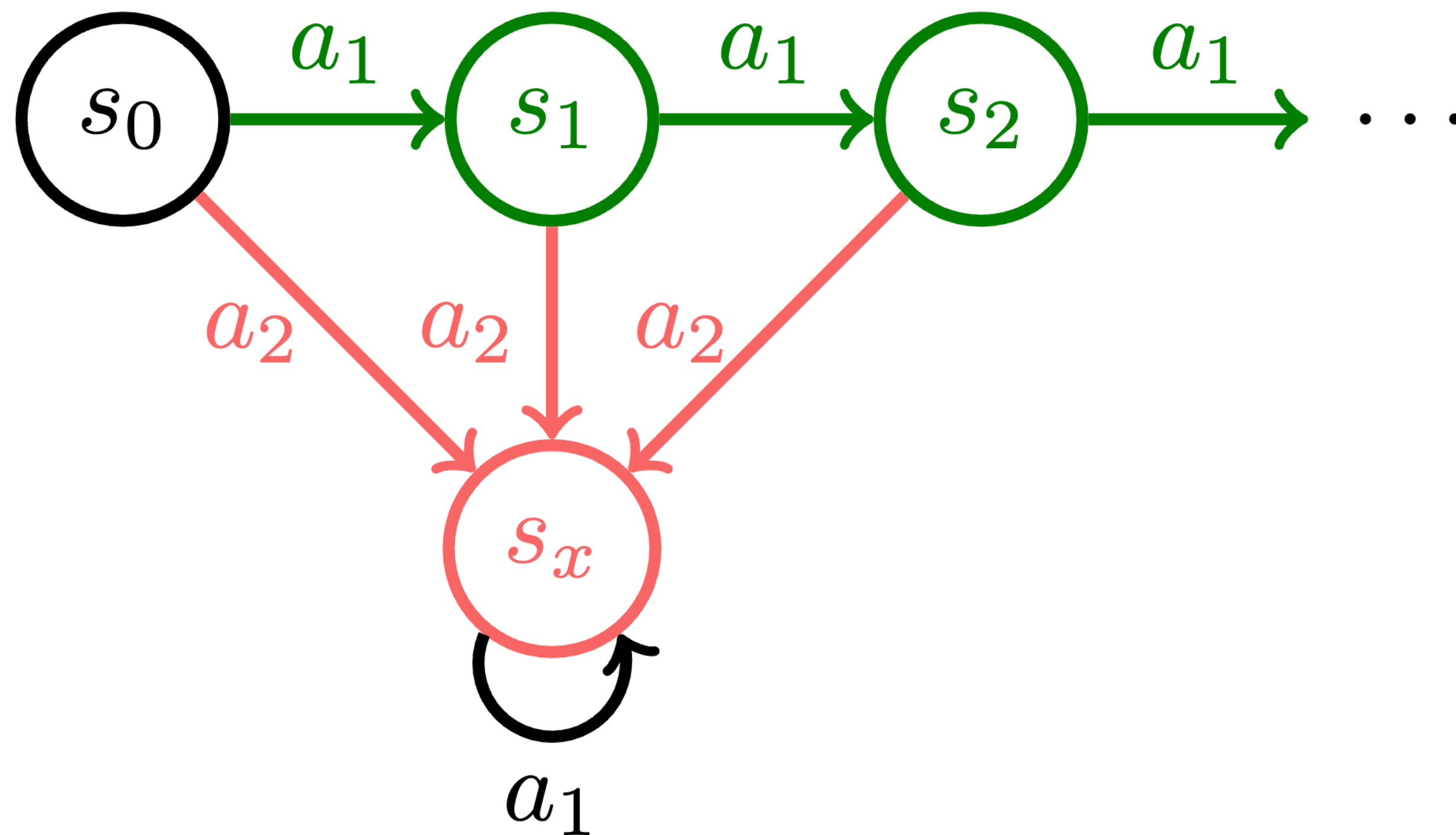
🔑 *Idea: Cannot tell the difference between policies that take  $a_1$  (i.e., recover) and  $a_2$  in state  $s_2$  from data in  $(s_0, s_1)$*

[Ross & Bagnell '12; Swamy et al. '21; '22; '23]

*Key Problem* → *Classical Solution* → *Efficient Algorithm* → *Practical Results*



# 🔑 *Problem: Supervised Imitation Doesn't Teach Recovery*



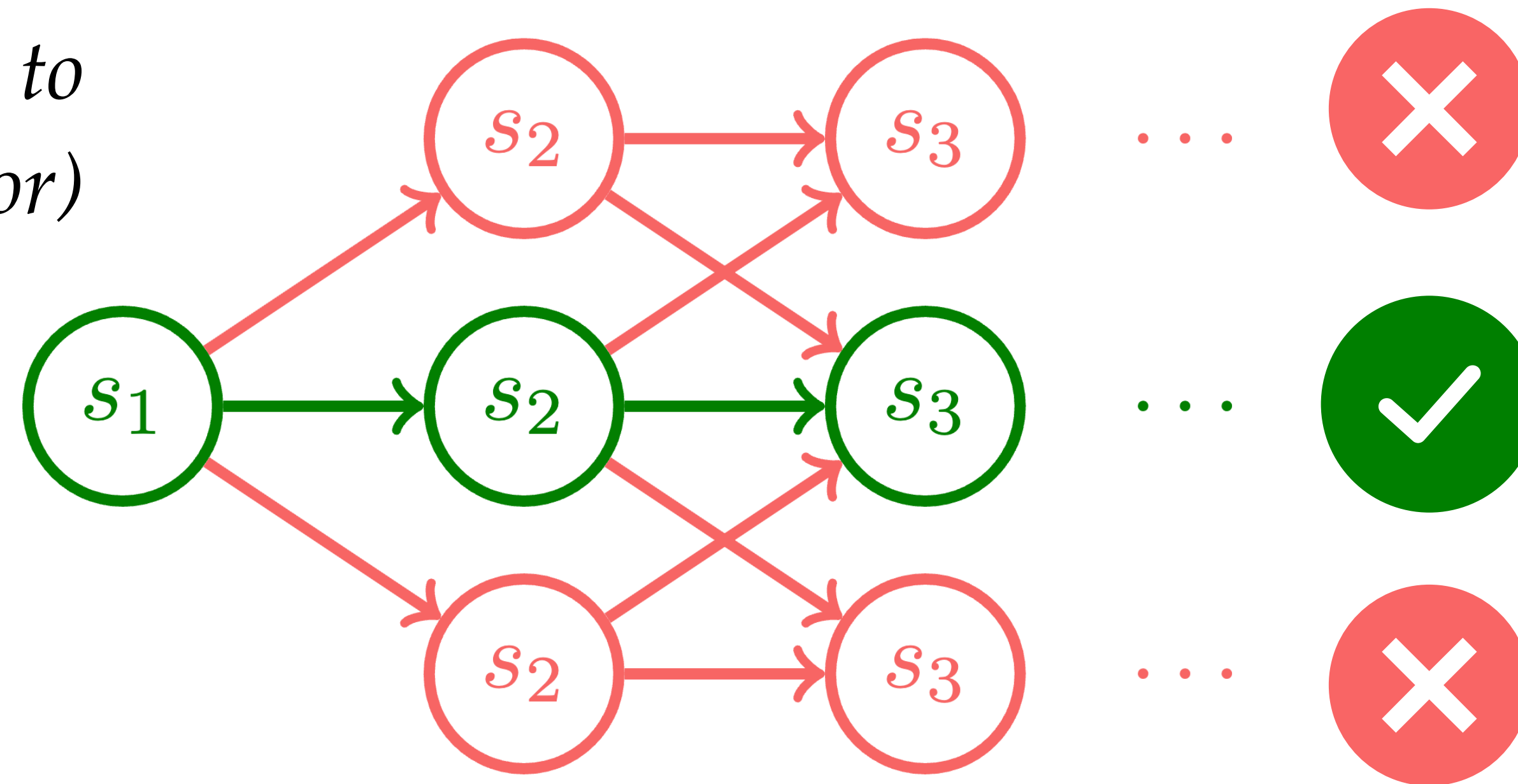
🔑 *Idea: Interaction cannot help here because there no  $\pi \in \Pi$  that can recover back to the top path.*

[Ross & Bagnell '12; Swamy et al. '21; '22; '23]

*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*

# 🔑 *Idea: Speeding up (I)RL with **Local Exploration***

*For  $\pi_1$ , BC would chose to go straight (low BC error)*



🔑 *Idea: Because we observe the consequences of our future policies in local search, we'd choose up!*

*Assume we've learned  $\pi_{3:H}$  (go straight),  $\pi_2$  (go down)*

[Swamy et al. '23]

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*

# Informal Theorem: *Local Exploration* is Sufficient for Imitation

$$\epsilon_{\Pi}^{\pi_i, r_i, \rho_E} := \mathbb{E}_{s \sim \rho_E} \left[ \max_{a \in \mathcal{A}} A_{r_i}^{\pi_i}(s, a) \right] - \max_{\pi' \in \Pi} \mathbb{E}_{s \sim \rho_E} \mathbb{E}_{a \sim \pi'(\cdot|s)} \left[ A_{r_i}^{\pi_i}(s, a) \right] \quad \epsilon_{\Pi}^{\rho_E} := \max_{\pi \in \Pi, r \in \mathcal{R}} \epsilon_{\Pi}^{\pi, r, \rho_E}$$

*Unrealizable optimal* (points to  $\max_{a \in \mathcal{A}} A_{r_i}^{\pi_i}(s, a)$ )  
*Realizable optimal* (points to  $\max_{\pi' \in \Pi} \mathbb{E}_{a \sim \pi'(\cdot|s)} [A_{r_i}^{\pi_i}(s, a)]$ )  
*Expert states* (points to  $\mathbb{E}_{s \sim \rho_E}$ )  
*Advantage over current policy* (points to the difference)

**Input:** Reward function  $r_i$ , reset distribution  $\rho$ , and policy class  $\Pi$

**Output:** Trained policy  $\pi$

**for**  $h = H, H - 1, \dots, 1$  **do**

Optimize

$$\pi_h \leftarrow \arg \max_{\pi' \in \Pi} \mathbb{E}_{s_h \sim \rho_h} \mathbb{E}_{a_h \sim \pi'(\cdot|s_h)} A_{r_i}^{\pi_{h+1}, \dots, \pi_H}(s_h, a_h)$$

**end for**

**Return**  $\pi = \{\pi_h\}_{h=1}^H$

[Swamy et al. 2023;  
Ren\*, Swamy\* et al. 2024;  
Espinosa-Dice ... Swamy 2025]

*Key Problem*  $\longrightarrow$  *Classical Solution*  $\longrightarrow$  *Efficient Algorithm*  $\longrightarrow$  *Practical Results*



# *Idea: Computing Minimax Winners via Self-Play*

$$\ell_t^1(p) \triangleq \mathbb{E}_{s_0 \sim \rho, \xi \sim p, \xi' \sim q_t} [P(\xi > \xi' | s_0)] \quad \ell_t^2(q) \triangleq \mathbb{E}_{s_0 \sim \rho, \xi \sim p_t, \xi' \sim q} [1 - P(\xi > \xi' | s_0)]$$

$$p_0 = q_0$$

$$\begin{aligned} \ell_0^1(p) &= \mathbb{E}_{s_0 \sim \rho, \xi \sim p, \xi' \sim q_0} [P(\xi > \xi' | s_0)] \\ &= \mathbb{E}_{s_0 \sim \rho, \xi \sim p, \xi' \sim p_0} [P(\xi > \xi' | s_0)] \\ &= \mathbb{E}_{s_0 \sim \rho, \xi \sim p_0, \xi' \sim p} [1 - P(\xi > \xi' | s_0)] \quad \text{(Anti)-Symmetry!} \\ &= \ell_0^2(p) \end{aligned}$$

*By induction, only need to keep track of Player 1's iterates!*

[Swamy et al. '24]

*Key Problem  $\longrightarrow$  Classical Solution  $\longrightarrow$  Efficient Algorithm  $\longrightarrow$  Practical Results*

# *Idea: RL with a Generation-Verification Gap*

*Sparse Verifier*

*Dense Generator*

*Statistical separation ( $\Theta(\log(|\mathcal{S}| |\mathcal{A}|))$ ) vs.  $\Theta(|\mathcal{S}| |\mathcal{A}|)$  for  $\ell_1$ -penalized regression!*

*Key Question  $\longrightarrow$  Controlled Experiments  $\longrightarrow$  Hypothesis  $\longrightarrow$  Empirical Validation*

# *Idea: RL with a Generation-Verification Gap*

## $\mathbb{H}_6$ : Online PFT is *Proper* Policy Learning.

*For fine-tuning problems with a simpler underlying reward function than (soft) optimal policy and a reward model class  $\mathcal{R}$  that enables learning simpler functions with fewer samples, the first step of online fine-tuning is finding a relatively simple reward model  $\hat{r}_{\text{sim}} \in \mathcal{R}_{\text{sim}} \subset \mathcal{R}$ , with the second step then finding a (soft) optimal policy for  $\hat{r}_{\text{sim}}$ ,  $\hat{\pi}_{\text{rlhf}} \in \Pi(\mathcal{R}_{\text{sim}})$ . Thus, end to end, online fine-tuning only has to choose between policies in  $\Pi(\mathcal{R}_{\text{sim}}) \subset \Pi$ , rather than across all of  $\Pi$ .*

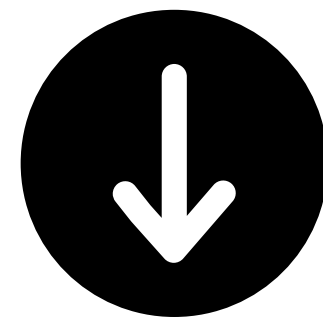
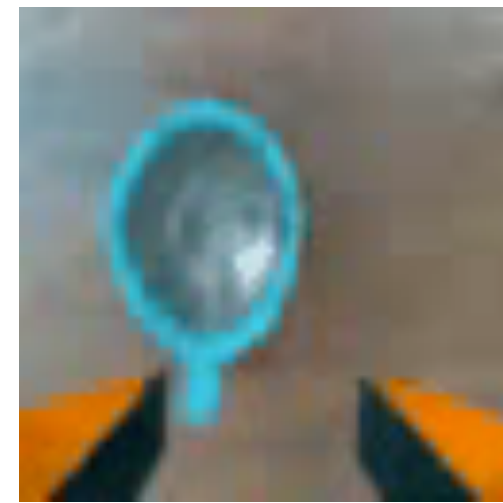
*Key Question  $\longrightarrow$  Controlled Experiments  $\longrightarrow$  Hypothesis  $\longrightarrow$  Empirical Validation*



# “Widening” Generation-Verification Gaps

## Robotics:

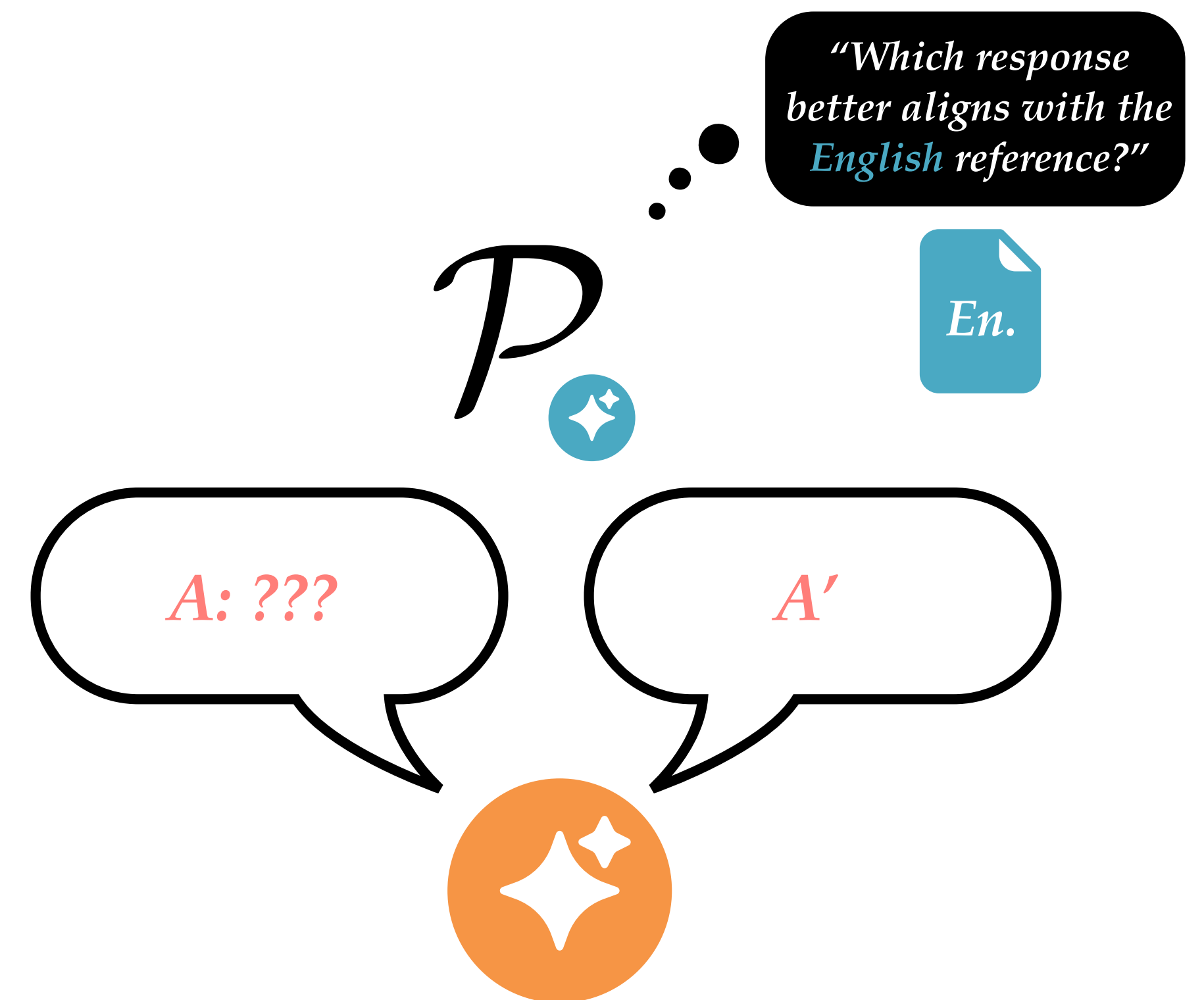
🔑: *Simplified Verification Task of VLM via Captioning!*



*“The robot grasps the cup via the handle.”*

## Language Modeling:

🔑: *Simplified Verification Task of Judge via Privileged Information!*



*Key Question → Controlled Experiments → Hypothesis → Empirical Validation*

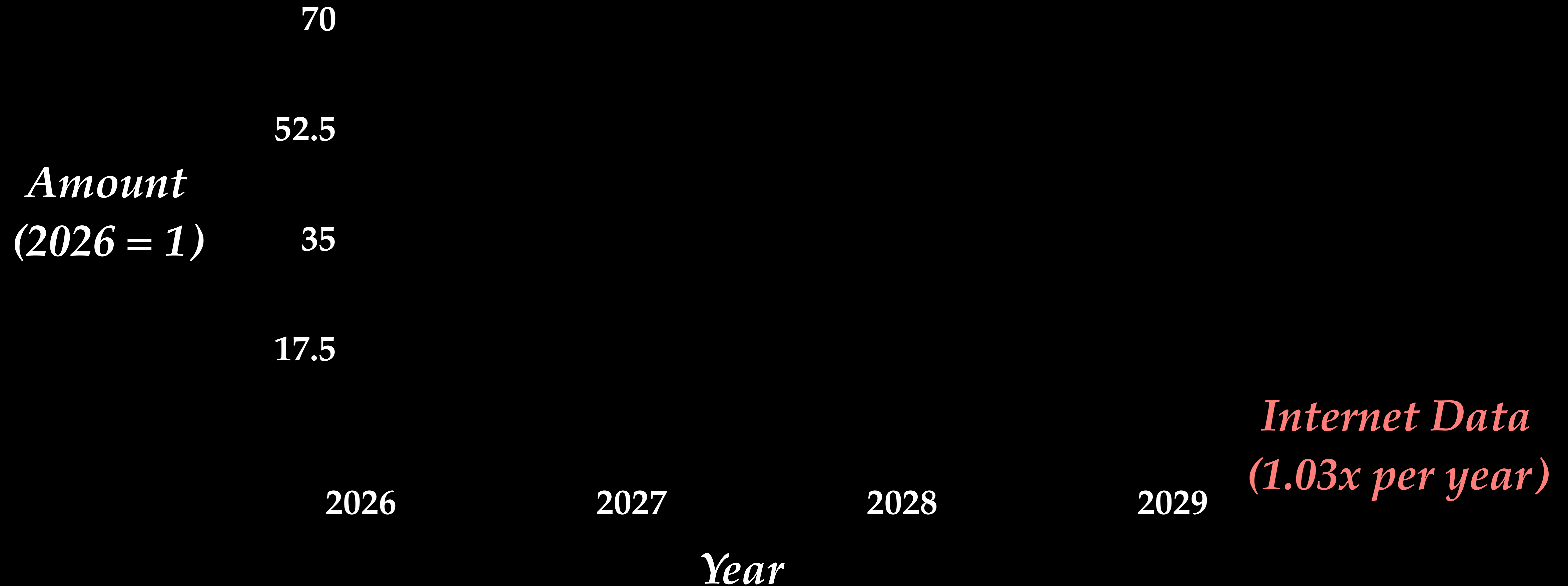
# *Low Resource Language Reasoning via **Self-Play***

[Sutawika, Swamy et al. '26]

*Key Problem → Classical Solution → Efficient Algorithm → Practical Results*



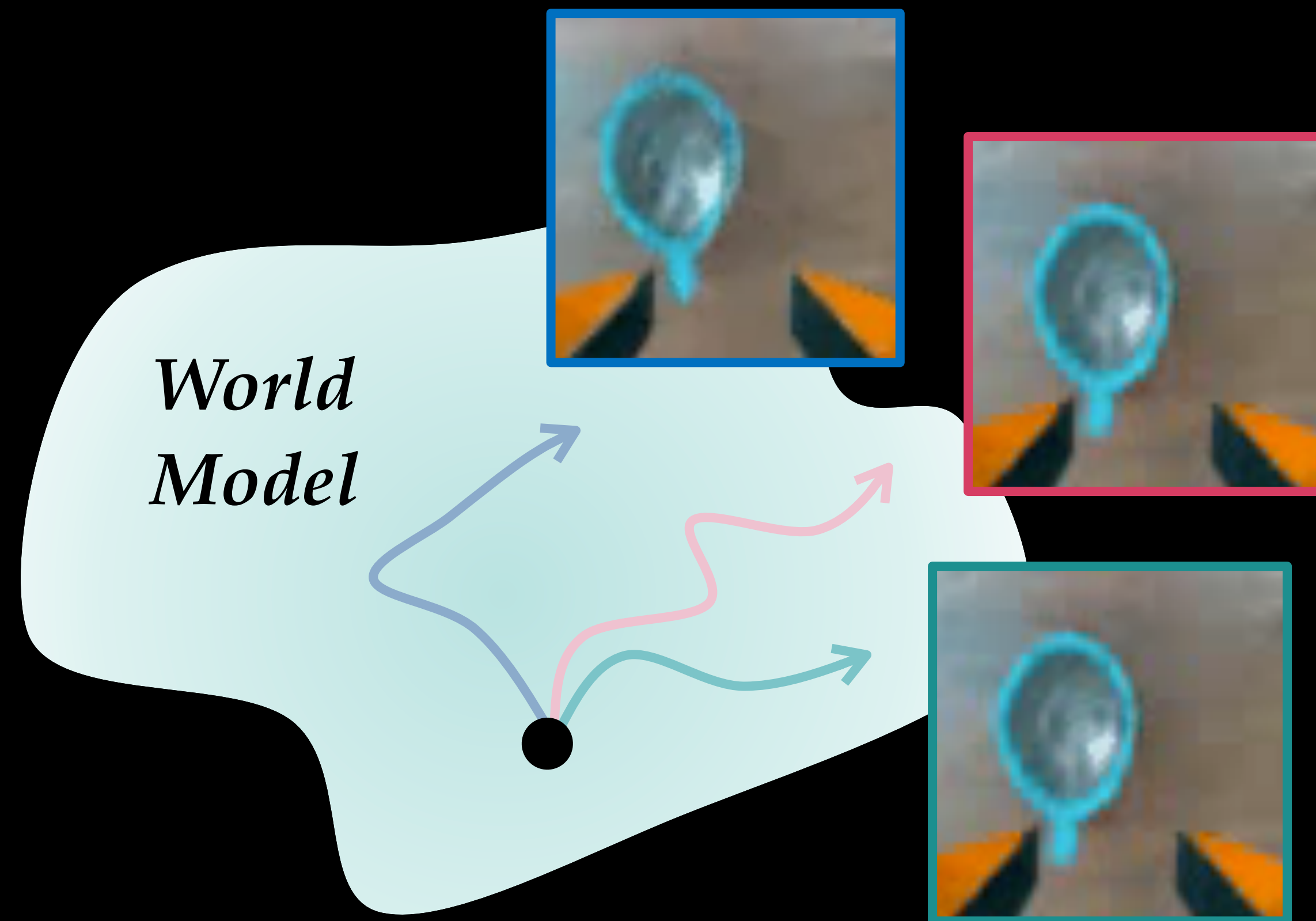
*Limited human data!*



[Sevilla and Roldán, 2024,  
Villalobos et al., 2024]



# 1. Long Horizon World Models



**Downstream Implication:** Reduced real world trial-and-error exploration for safety-critical robotics tasks + task mastery.

🔑 **Problem:** WMs are trained to ensure the latent state has enough information to reconstruct future pixels.

⇒ This requires a *lot* of data to learn and is challenging over long horizons.

🔑 **Idea 1:** Simplify WM training by just ensuring their latent state can predict  $Q$ -values (long-term consequences) under a class of reward functions

🔑 **Idea 2:** Incorporate some notion of policy complexity in WM learning to go beyond information-theoretic sufficiency.



## 2. Long Horizon Verifier Agents

It was a slow night. The bartender polished the bar for the thirteenth time that night, wiping his rag back and forth, the lemony smell of the polish reaching through the alcohol. I lifted my glass as he swiped underneath, both of us used to the rhythm. **There had been a lot of slow nights lately. Not as many people as usual in the old pub anymore. They liked the newer places, and I couldn't blame them. After all, when I'd first come here, it had been new.** Now, though it was older than me, it had aged better. "Whiskey. Neat." The voice wasn't interesting, the request boring. But still, something piqued my curiosity. I turned, eyeing the new customer. They didn't seem to be anyone special. The customer raised an eyebrow at me, half-lifting their glass. "Cheers mate." I returned the motion, nodding to them. It was odd, when I'd first looked, I would have sworn they were a man, but the face seemed to change, shifting in the space of a blink. Now old, now young, now female, now male, now androgynous. I had kept track of how much I'd had to drink, and it wasn't enough. Knocking back the rest of my own whiskey, I slid off the stool. The stranger's words stopped me in my tracks. "Next time, I'll get them, and they offered me a quiet smile. There was something familiar about that smile. I'd seen it before; so many people would call it a rich smile. But they were offering to buy the next round. It would be churlish to refuse. Climbing back onto the stool, I said out loud. "Name's Lyra. What's yours, friend?" The other customer tilted his head to the side, before extending their hand. Their touch was cold. "People have called me many things. Too many to count. Which would you prefer?" I sipped at my refilled glass, mulling the question over. But I asked for your name. What people call you. After all, I've been called a lot of things too you know. But I don't go around calling myself 'Loser'." The customer smiled, a low rolling laugh that had the timbre of an earthquake. They emptied their own glass before they could say anything. I turned to the bartender. He refilled with the poison of choice, and across from me, the figure nodded their head. "My name... It has been so long. I don't know if I remember my name. Did I even have one..." The face had shifted, sadness pulling away the remnants of the grin. Summoning up my courage, though I didn't know why I should be scared, I clapped them on the back. "Buck up. You don't need a name to drink here, or to talk to me. What brings you here on a Saturday night?" "Work. Always work." I shook my head at the response, handing my glass to the bartender. The other customer claimed the responsibility for this drink as well. They were fast rising in my estimation. "Life. Work isn't enough. I have to have to balance your life. You know, have a day off here and there, find nice people to spend time with. All of this is just a distraction. I need my drinking companion. They sighed, and I caught the smell of an open grave. "Taking a break isn't really an option. Thank you for the drink." "Hey, all I'm saying is you should live your life. Have some fun, make some mistakes." "And you?" They raised their head, looking at me. "Have you lived your life?" It was my turn to chuckle at them, as I bought them another whiskey. "Oh, you have no idea. The thing is, I was this one time in France..." And as the night drew on, I regaled them with story after story. I had an audience that hadn't heard of me before. I was going to make the most of it. As we bought each other round after round, my companion loosened a little, even telling me a few of their own. But that wasn't the only thing we talked about. By the time the bartender warned us of last call, we were deep into a discussion of the matter of philosophy. I paid my tab and nodded to the other customer. "Well, if you're ever in this neck of the woods again, don't be a stranger. We can have another good chat." There was a return of the sadness from earlier, as they shook their head. "I'm afraid that won't be possible. You asked me what my name was earlier. I told you that people call me many things. And what they call me is—" "Death?" I laughed at the consternation on their face. "I figured it out a while ago. I was sort of expecting it soon. Ticker's not what it used to be. But hey. You can't be all bad. You bought me drinks, sat and listened to my stories, and told a few of your own." I turned, as we reached the doorway, looking back into the pub. **Funny how I ended up here tonight. Never been in this place before, but somehow it felt right.** Like I was meant to spend my last evening here, sharing stories with a stranger who turned out to be Death himself. Not a bad way to go, all things considered." Death stared down at me, their ever-changing appearance no longer a disturbing sight. They held out a hand, their voice soft. "I am glad. But now, it is time to go." I settled my cap in its usual position on my head, wrapped my scarf around my neck and shrugged into my coat. And taking Death's hand, I walked into the night, unafraid.

*Downstream Implication: Less human supervision on long-horizon robotics / language modeling tasks.*

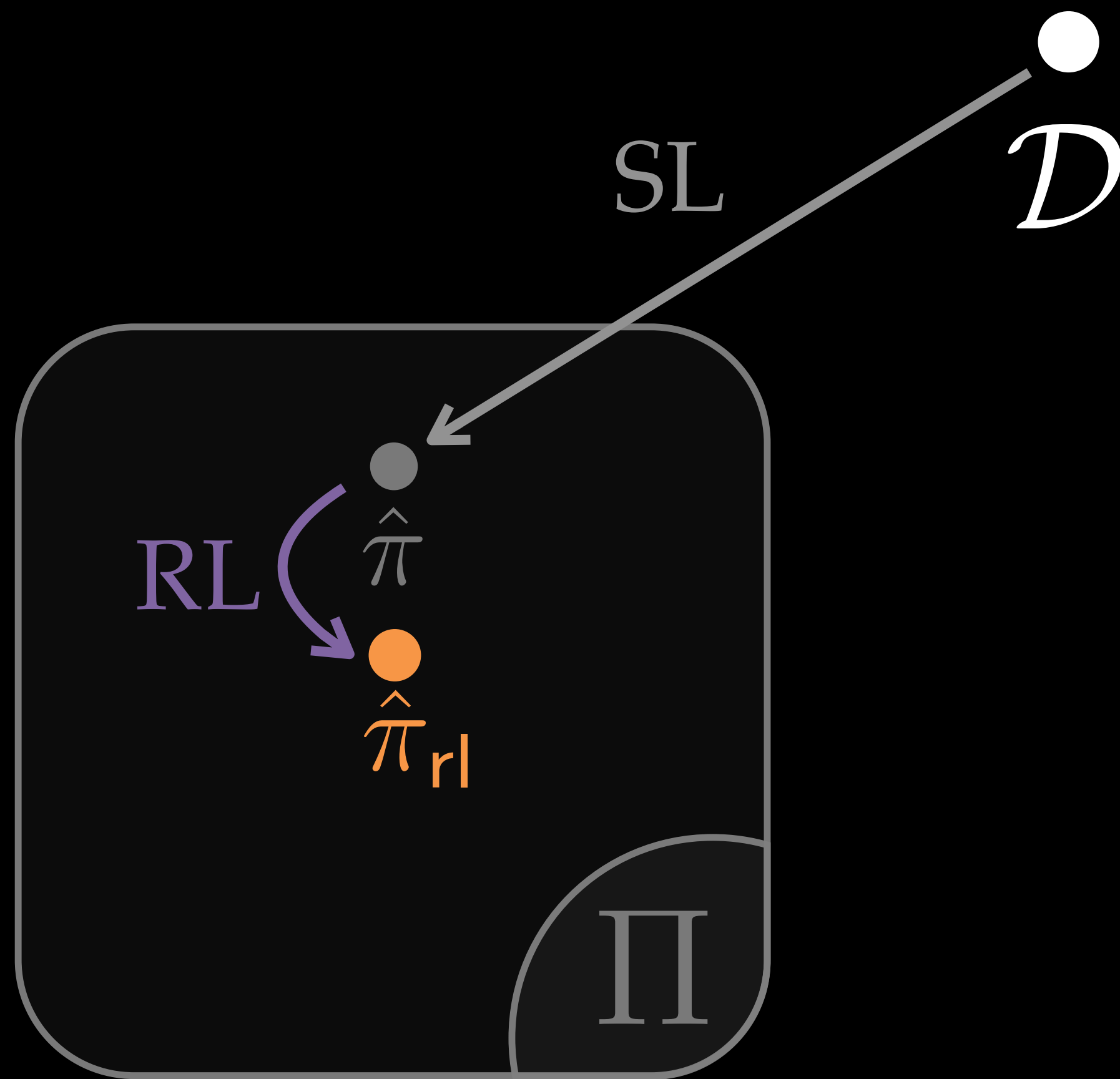
🔑 *Problem: Verifier agent needs to learn what to remember and what to forget to make accurate judgements with limited context.*

⇒ *Training such a verifier agent is a hard exploration problem with sparse rewards.*

🔑 *Idea 1: Distill an “expert” who has access to the benefit of hindsight (e.g., an LLM with future pages of the story as context).*

🔑 *Idea 2: Summarize context in a way that is predictive of future text (akin to QuietStar) for denser feedback over an episode.*

### 3. Understanding the Effect of SL on RL



#### *Policies*

**Downstream Implication:** Scientific understanding of generalization and less exploitation of learned WMs / RMs.

🔑 **Problem:** Offline (e.g., pre-training) performance is not predictive of performance after interactive learning (e.g., RL).

⇒ Generalization is hard to predict and depends on factors like model architecture.

🔑 **Idea 1:** Perform controlled experiments (e.g., mixing perfect and “noisy” datasets) to figure out which metrics predict the optimal mixture as measured by post-RL performance.

🔑 **Idea 2:** “Invert” these metrics for semantically meaningful exploration (e.g., inverting feature covariance matrix of last layer).