Learning from Interaction

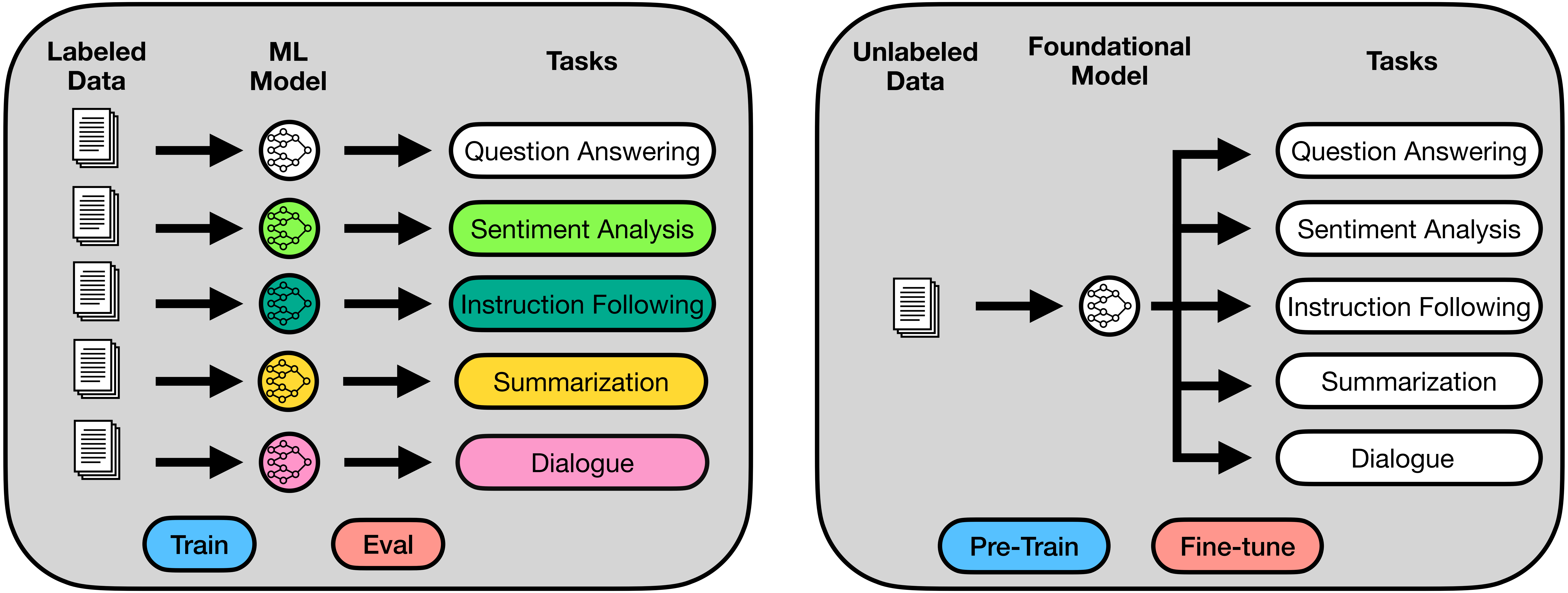
Kianté Brantley | Postdoctoral Associate | Cornell University

"A foundation model is any model that is trained on broad data that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks"

Foundational Models

[1]On the Opportunities and Risks of Foundation Models by Bommasani et al. 2022 [2]

Foundational Models



Traditional Machine Learning Models

Foundational Models

Foundational Models

$$\overbrace{\text{Large}}^{>\text{70B params}}\ \underbrace{\text{Language Model}}_{p(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2 \mid y_1)p(y_3 \mid y_1, y_3)\ldots = \prod_{i=1}^{n} p(y_i \mid y_1, \ldots, y_{i-1})}$$

for example,

$$p(\underbrace{\cdot}_{\text{predict next word}} \mid \overbrace{\text{you, only, live}}^{\text{context}})$$

Foundational Models

Large Language Model

$$p(y_1, y_2, \ldots, y_n) = p(y_1)p(y_2 \mid y_1)p(y_3 \mid y_1, y_3)\ldots = \prod_{i=1}^{n} p(y_i \mid y_1, \ldots, y_{i-1})$$

Pre-Training

Large Unlabeled Dataset

Foundational Models

Large Language Model

$$L(y) = -\sum_{i}^{n} \log p(y_i \,|\, y_{<i}) \quad \Big\} \text{ next word prediction}$$

Pre-Training

Large Unlabeled Dataset

Foundational Models

Large Language Model

$$L(x, y) = -\sum_{i}^{n} \log p(y_i \mid y_{<i}, x)$$

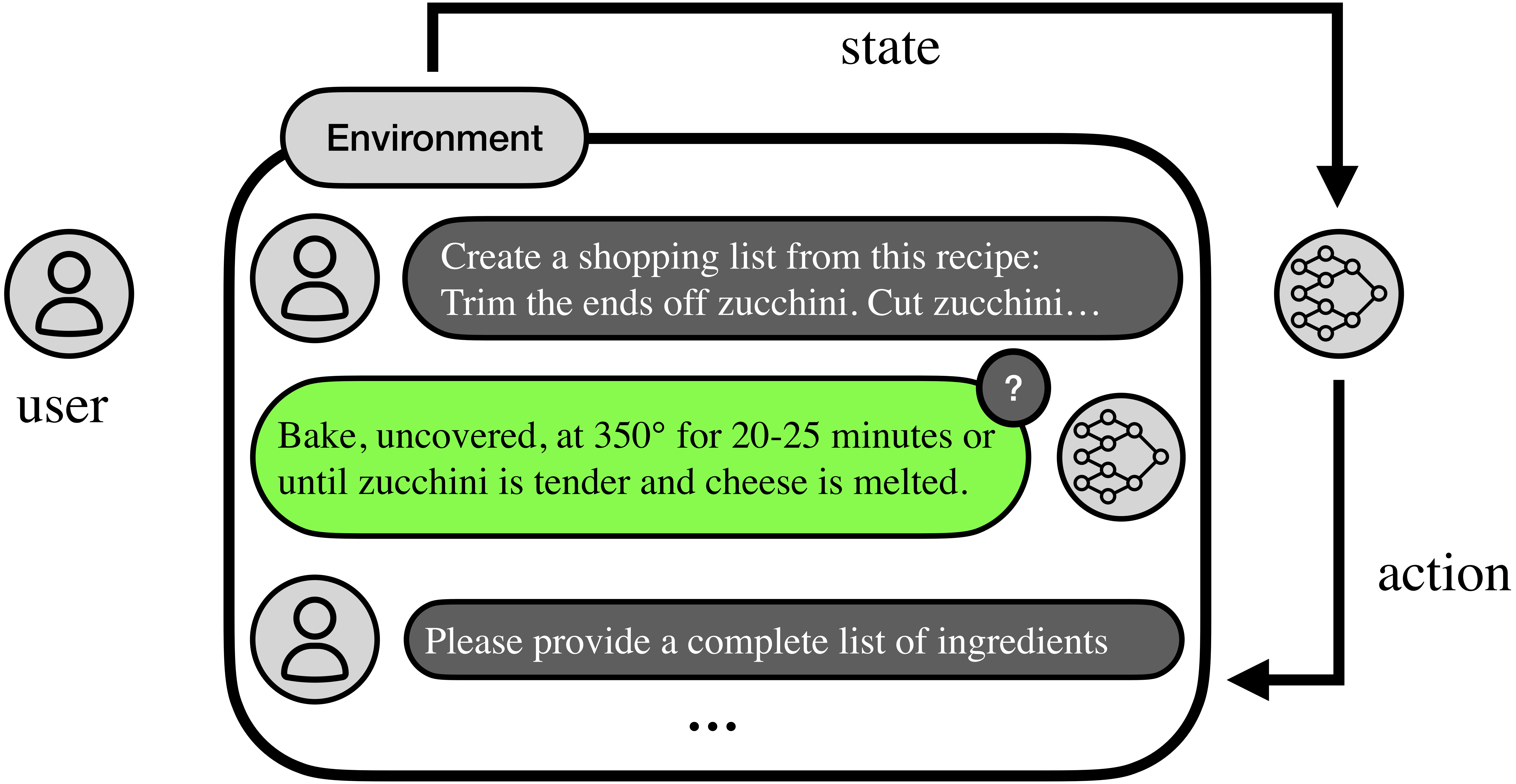
predicting the next token conditioned on some context

Fine-Tuning

Input: x → → Output: y

Input: x	Output: y
What is the capital of France?	The capital of France is Paris.
Describe the function of a computer motherboard	A computer motherboard is the main circuit board in a computer. It is the backbone of …
....	....

Foundational Models



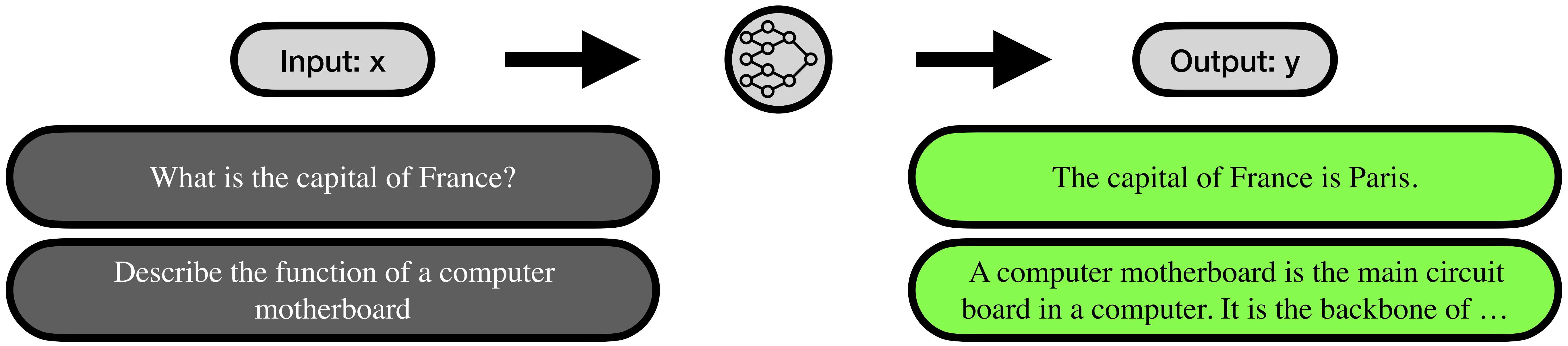
Training language models to follow instructions with human feedback Ouyang et al. 2022

Foundational Models

objective alignment issue

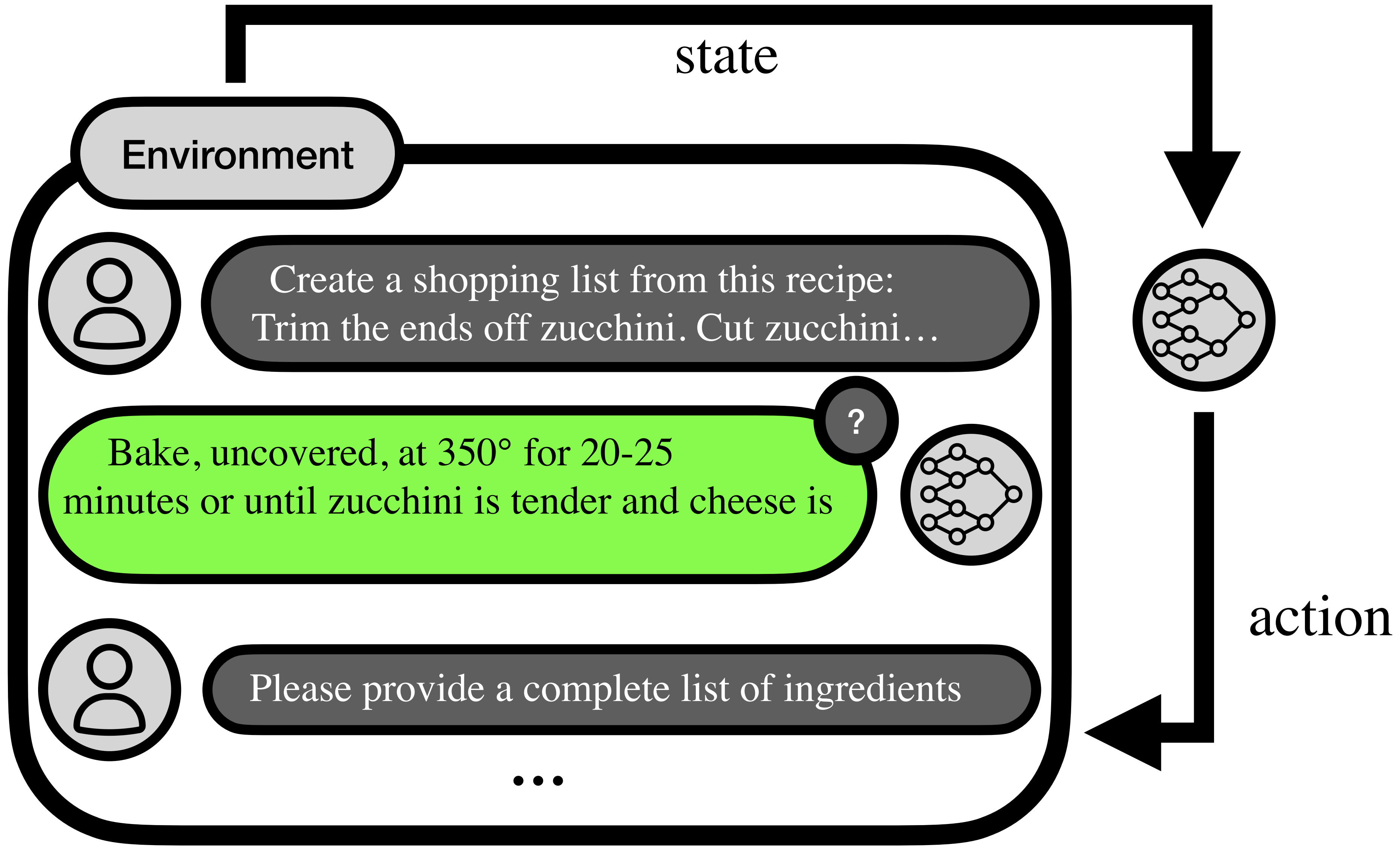
Fine-Tuning

$$L(x, y) = -\sum_{i}^{n} \log p(y_i \mid y_{<i}, x)$$



Next word prediction

Test



Ability to follow instructions

Foundational Models

"Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users."

Long Ouyang et al.
Training language models to follow instructions with human feedback
OpenAI 2022

An Old Problem

Figure 3: NAVLAB, the CMU autonomous navigation test vehicle.

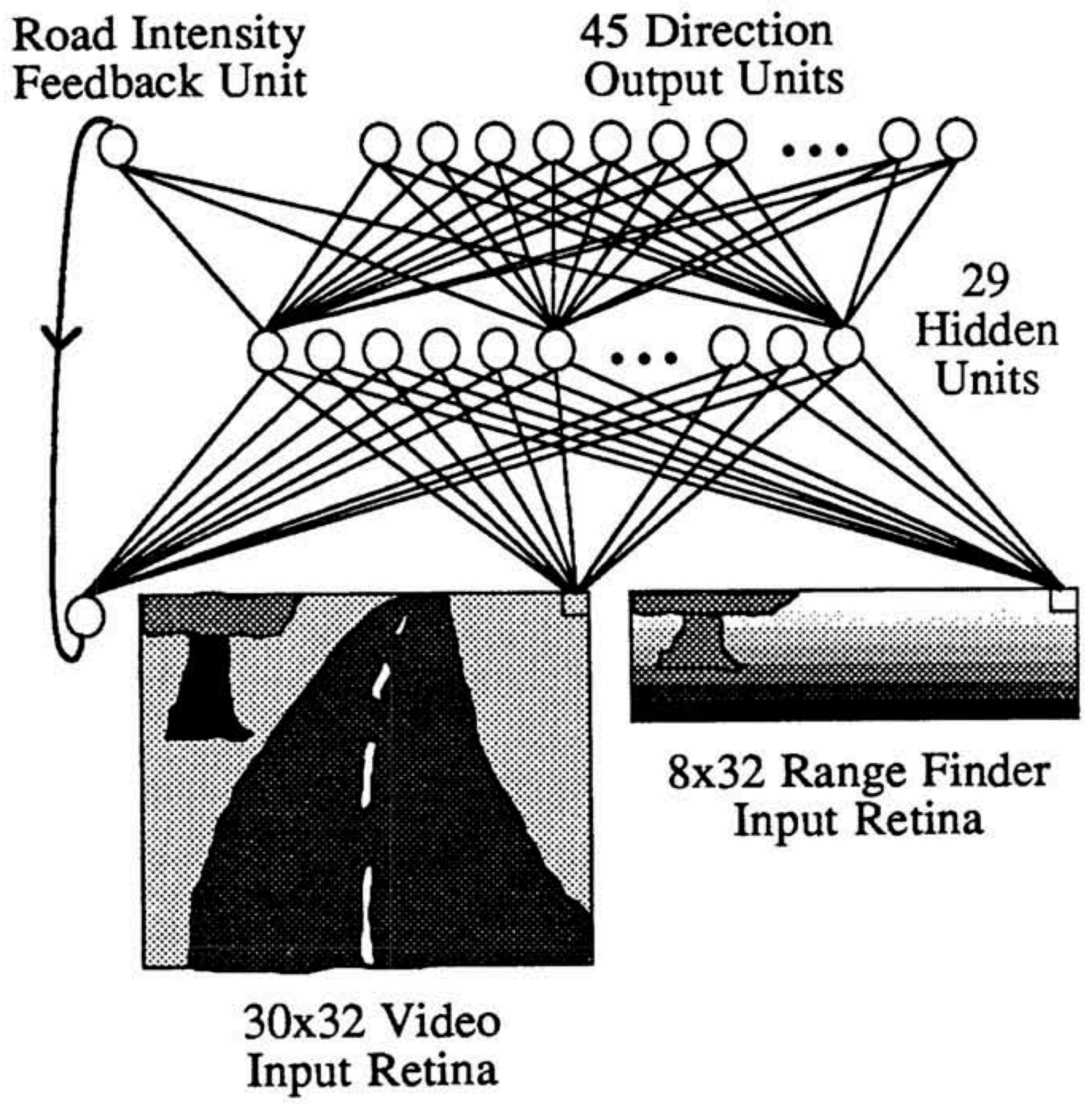


Figure 1: ALVINN Architecture

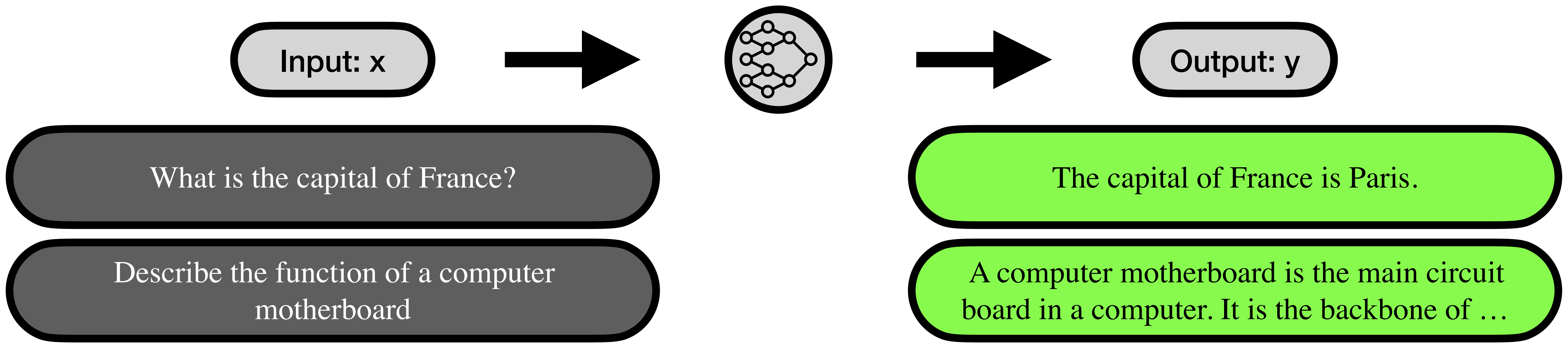
"… the network must not solely be shown examples of accurate driving, but also how to recover (i.e. return to the road center) once a mistake has been made."

Dean Pomerleau
Alvinn: Ann Autonomous Land
Vehicle In A Neural Network
NeurIPS 1989

Foundational Models

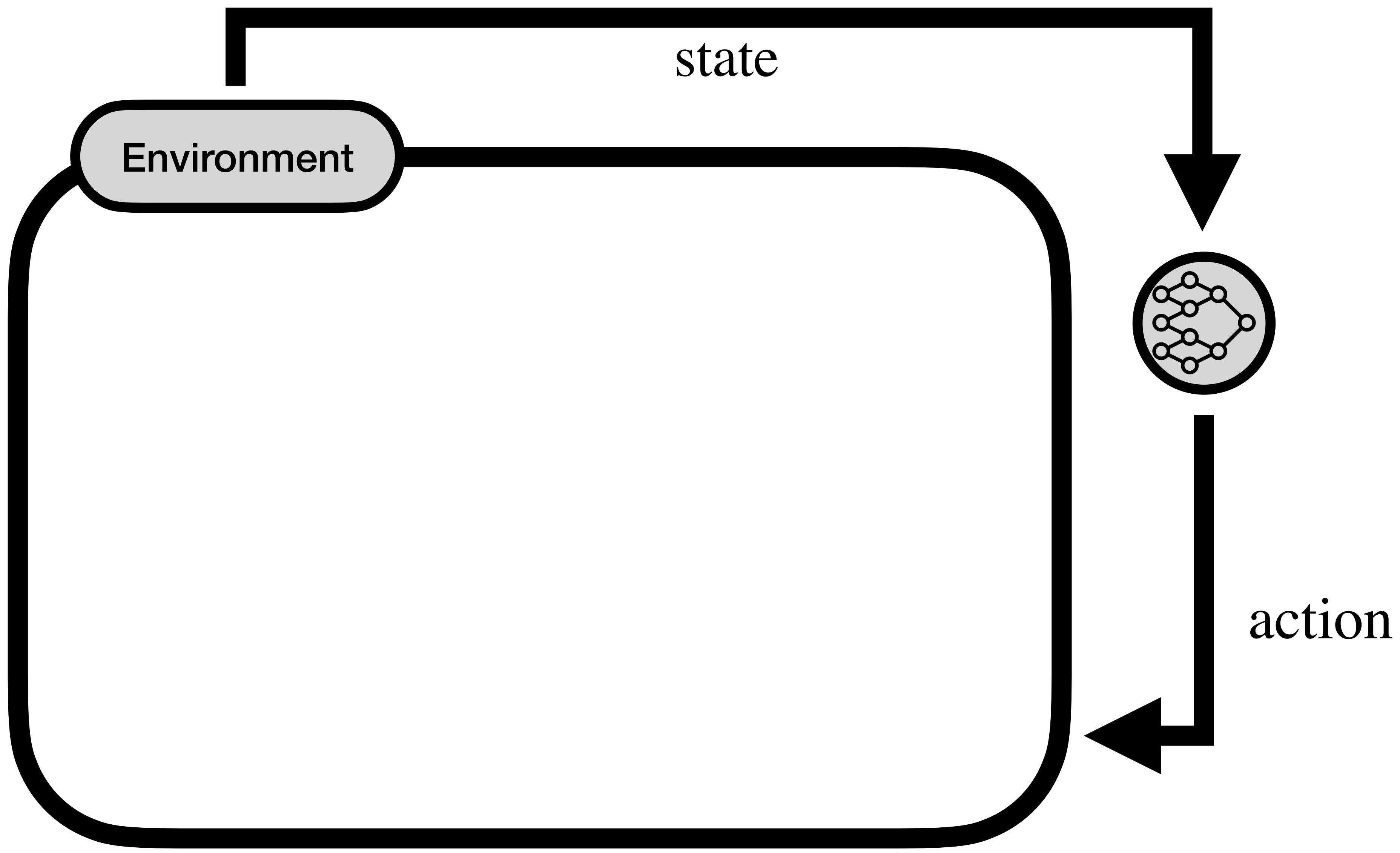
Fine-Tuning

$$L(x, y) = -\sum_{i}^{n} \log p(y_i \mid y_{<i}, x)$$



Next word prediction

Interactive Fine-Tuning



Foundational Models

Fine-Tuning

Interactive Fine-Tuning

$L(x, y) = -$

Input: x

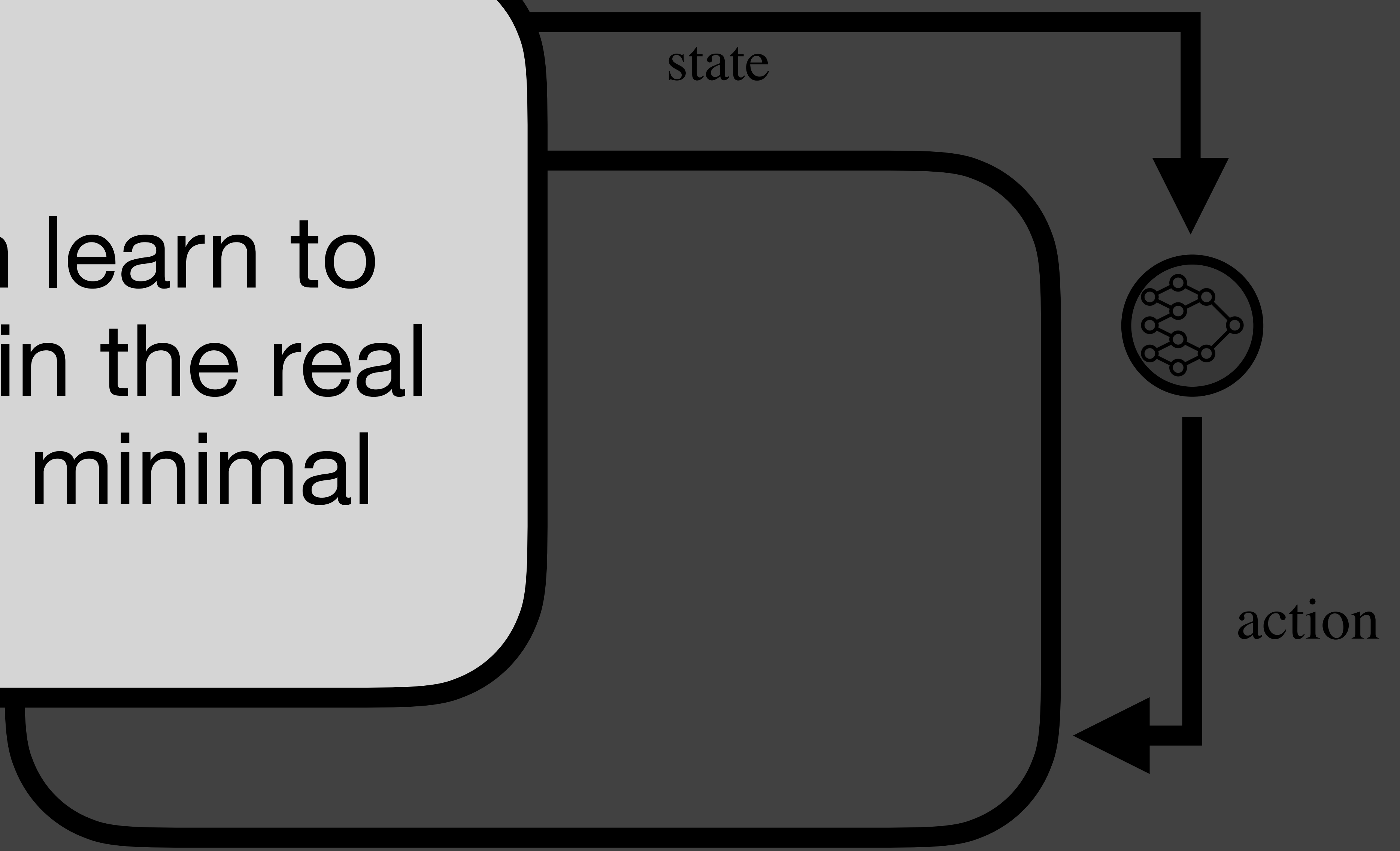
What is the capital of France?

Describe the function of a computer motherboard

board in a computer. It is the backbone of …

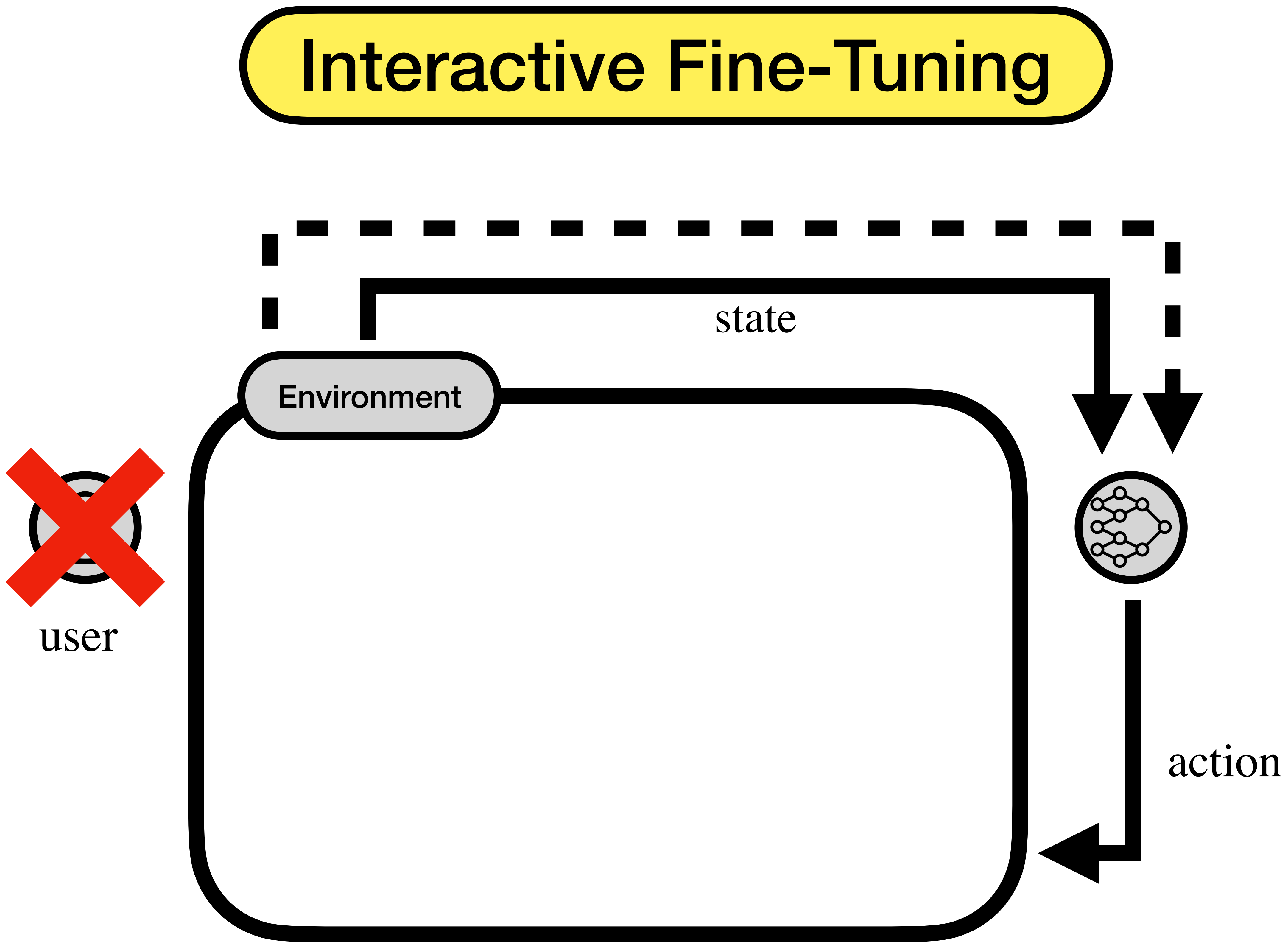
Research Goal:

Develop algorithms that can learn to make automated decisions in the real world from interactions with minimal supervision.



Next word prediction

Foundational Models



How do we learn from interactions in an environment?

Overview of Past Work

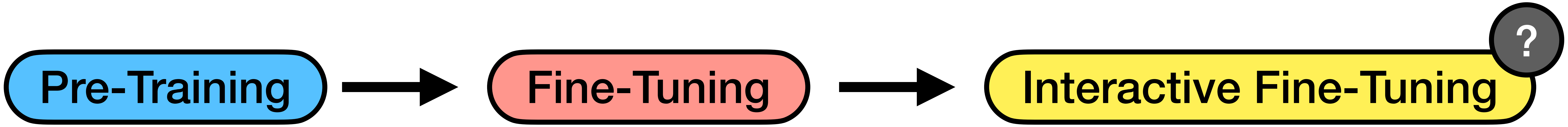
- Imitation Learning [1, 2, 12]
- Imitation Learning with a Computational Oracle [6, 7, 8]
- Preference-Based Reinforcement Learning [3, 4, 5]
- Constrained Reinforcement Learning [9, 10, 11]

1. Disagreement-Regularized Imitation Learning………………………………………………………….. [**B**SH ICLR 2020]
2. Adversarial Imitation Learning via Boosting……………………………………………………………. [CSH**B**S ICLR 2024]
3. Is reinforcement learning (not) for natural language processing?…………………………………….. [RA**B**HSBHC ICLR 2023]
4. Learning to Generate Better Than Your LLM…………………………………………………………. [C**B**RMS Instruction Workshop 2023]
5. Policy-Gradient Training of Language Models for Ranking……………………………………………. [GCC**B**J FMDM Workshop 2023]
6. Non-Monotonic Text Generation………………………………………………………………………….. [W**B**DC ICLR 2019]
7. Active Imitation Learning with Noisy Guidance………………………………………………………. [**B**SD ACL 2020]
8. Interactive Text Generation……………………………………………………………………………. [FGP**B**CZGD EMNLP 2024]
9. Reinforcement Learning with Convex Constraints……………………………………………………. [M**B**DDS NeurIPS 2019]
10. Constrained episodic reinforcement learning in concave-convex and knapsack settings………… [**B**DLMSSS NeurIPS 2020]
11. Ranking with Long-Term Constraints…………………………………………………………………….. [**B**FDJ WSDM 2024]
12. Successor feature sets: Generalizing successor representations across policies………………….. [**B**MG AAAI 2021]

Overview of Past Work **Domains**

- Natural Language Processing [3, 4, 6, 7, 8]
- Classic Control [2, 9, 10, 12]
- Information Retrieval [5, 11]
- Video Games [1]

1. Disagreement-Regularized Imitation Learning………………………………………………………….. [**B**SH ICLR 2020]
2. Adversarial Imitation Learning via Boosting……………………………………………………….. [CSH**B**S ICLR 2024]
3. Is reinforcement learning (not) for natural language processing?…………………………………….. [RA**B**HSBHC ICLR 2023]
4. Learning to Generate Better Than Your LLM………………………………………………………… [C**B**RMS Instruction Workshop 2023]
5. Policy-Gradient Training of Language Models for Ranking…………………………………………… [GCC**B**J FMDM Workshop 2023]
6. Non-Monotonic Text Generation…………………………………………………………………….. [W**B**DC ICLR 2019]
7. Active Imitation Learning with Noisy Guidance………………………………………………………. [**B**SD ACL 2020]
8. Interactive Text Generation…………………………………………………………………………. [FGP**B**CZGD EMNLP 2024]
9. Reinforcement Learning with Convex Constraints……………………………………………………. [M**B**DDS NeurIPS 2019]
10. Constrained episodic reinforcement learning in concave-convex and knapsack settings………… [**B**DLMSSS NeurIPS 2020]
11. Ranking with Long-Term Constraints……………………………………………………………….. [**B**FDJ WSDM 2024]
12. Successor feature sets: Generalizing successor representations across policies………………….. [**B**MG AAAI 2021]


Pre-Training
Fine-Tuning
Interactive Fine-Tuning
?

Pre-Training → Fine-Tuning → Interactive Fine-Tuning ?

Formalize the imitation learning problem

objective alignment issue

How can interaction help?

How can we design interactive learning algorithms specific to LLMs?

Pre-Training → Fine-Tuning → Interactive Fine-Tuning ?

Formalize the imitation learning problem

objective alignment issue

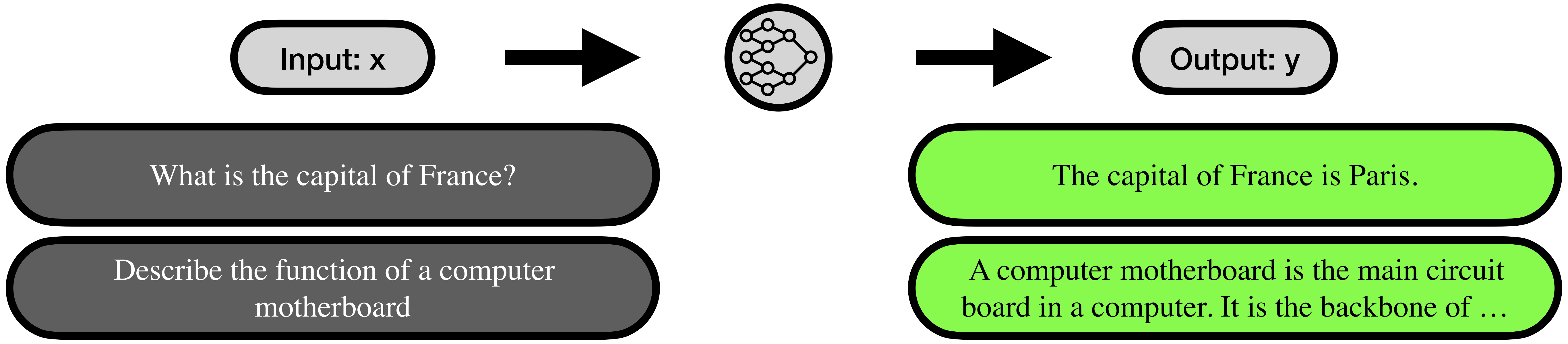
Foundational Models

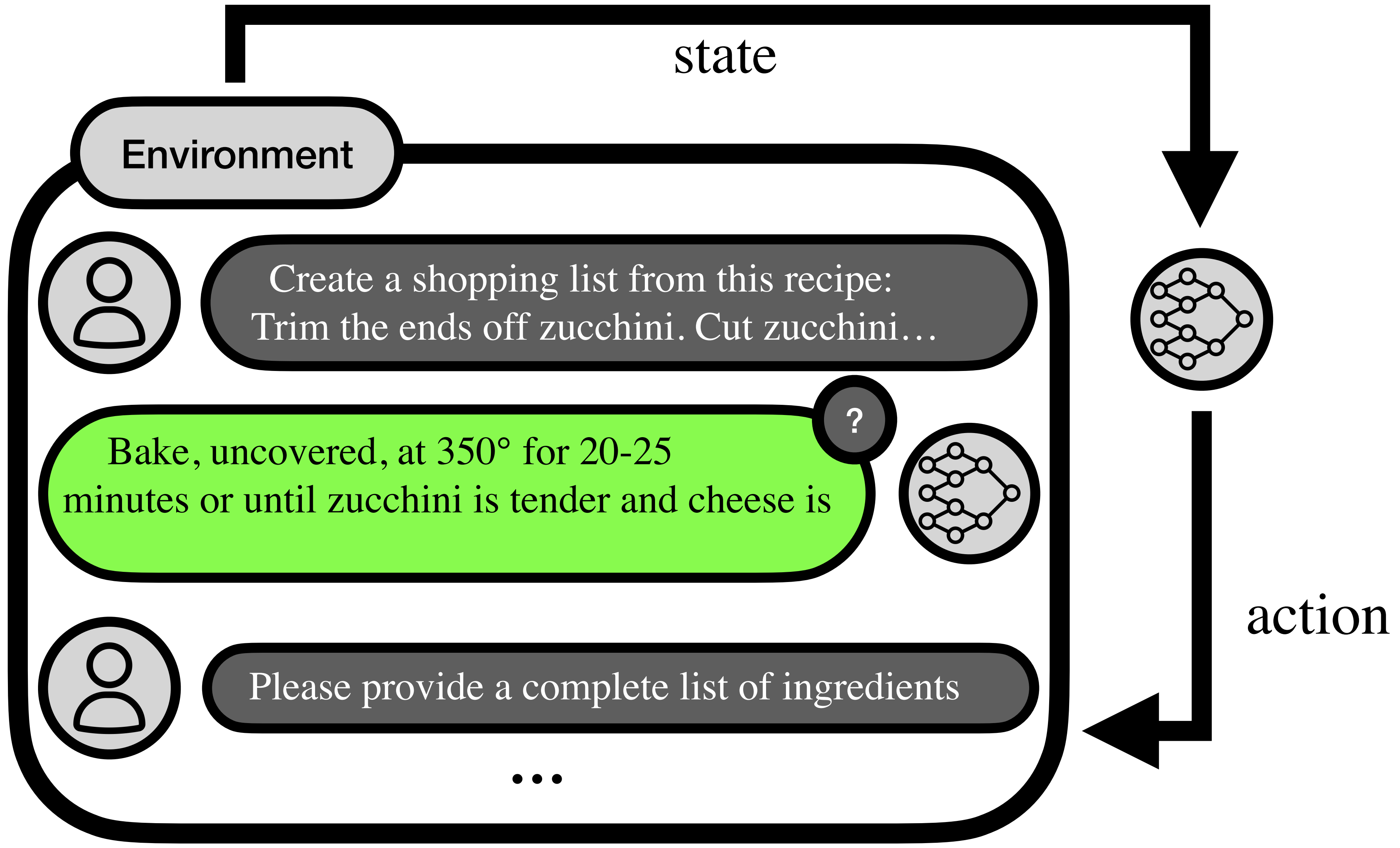
objective alignment issue

Fine-Tuning

Test

$$L(x, y) = -\sum_{i}^{n} \log p(y_i \,|\, y_{<i}, x)$$





Next word prediction

Ability to follow instructions

Reinforcement Learning

basics of mdp

States: $\{x, y\}$

Actions: $\{$  $\}$

Reward: $R(s, a)$

Transition: $P(s' \mid s, a)$

Policy: $\pi(\cdot \mid s)$

Goal: **Learn policy π to maximize reward**

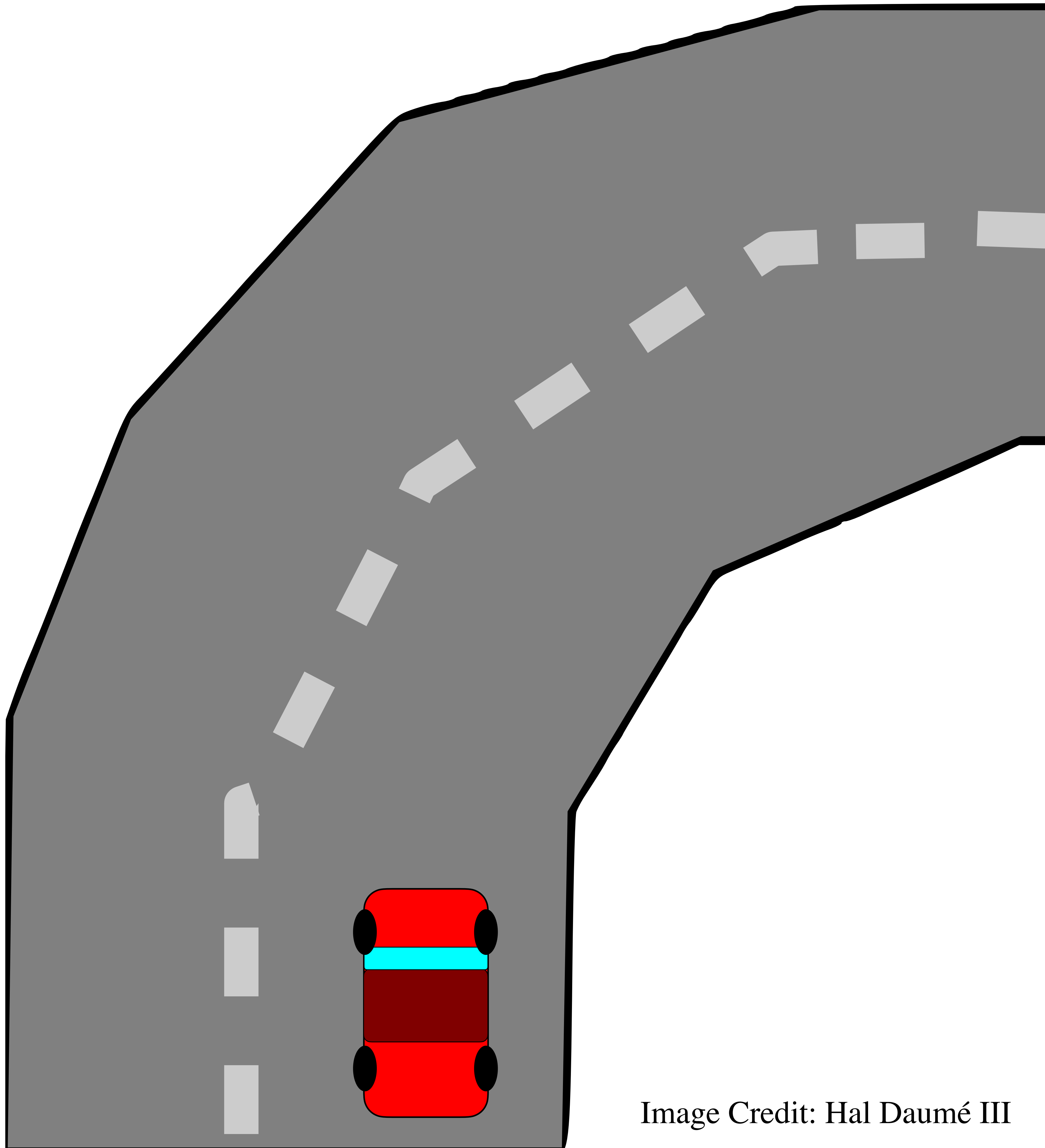


Image Credit: Hal Daumé III

Imitation Learning

basics of mdp

States: $\{x, y\}$

Actions: $\{$  $\}$

Reward: Unknown Reward

Transition: $P(s' \mid s, a)$

Policy: $\pi(\cdot \mid s)$

Training dataset: D = {(state, action)} from an expert $\pi^{\star}$

Goal: **Train a policy π to mimic the demonstrations**

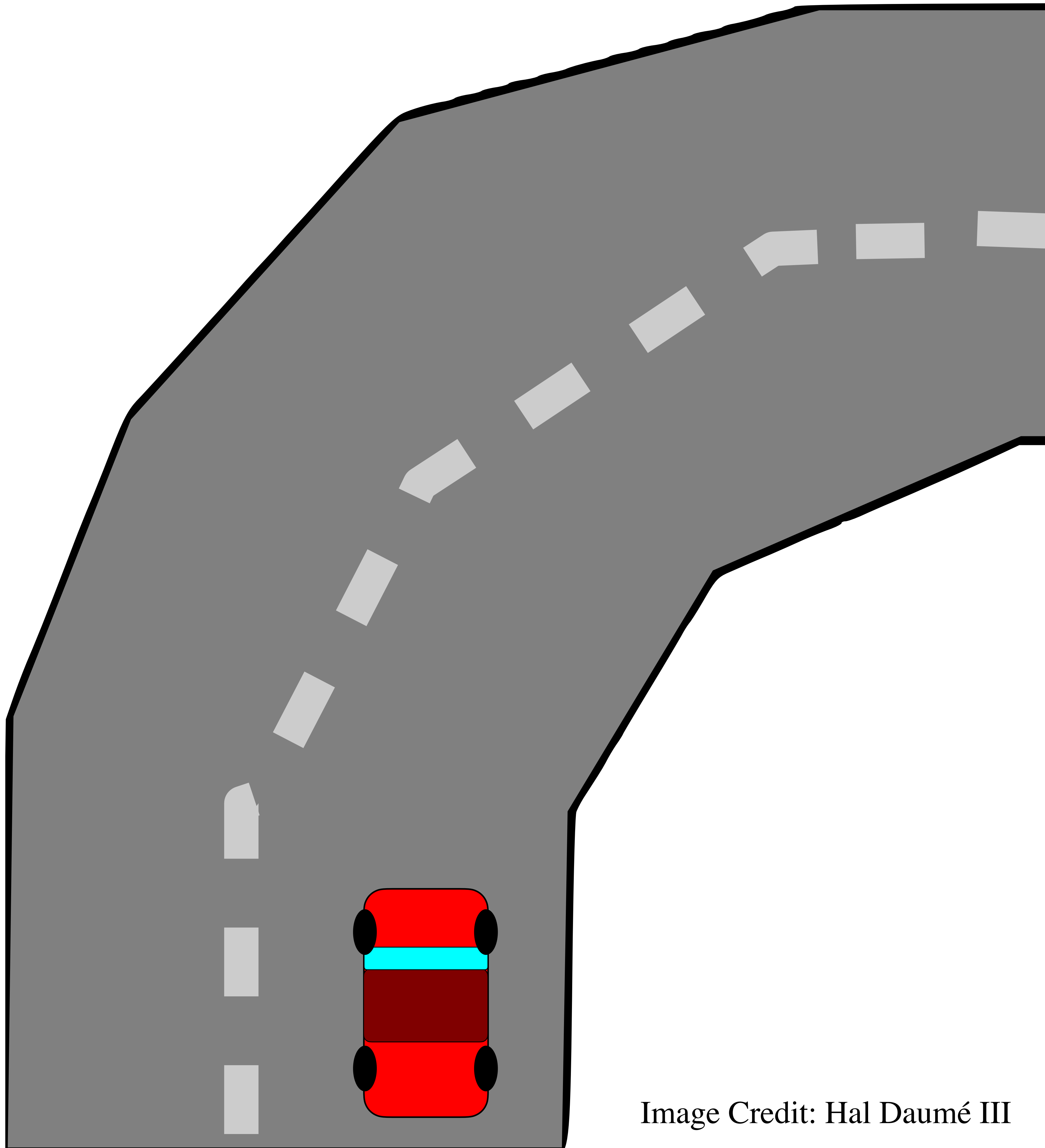


Image Credit: Hal Daumé III

Behavior Cloning

simplest imitation learning

1. **Collect trajectories from expert $\pi^\star$**

2. **Create a dataset $(s_1, a_1), \ldots, (s_h, a_h) \sim \rho_{\pi^\star}$**

3. **Train a policy (classifier) π**

$$\min_{\pi \in \Pi} \mathbb{E}_{(s,a) \sim \rho_{\pi^\star}} \left[L(\pi, s, a) \right]$$

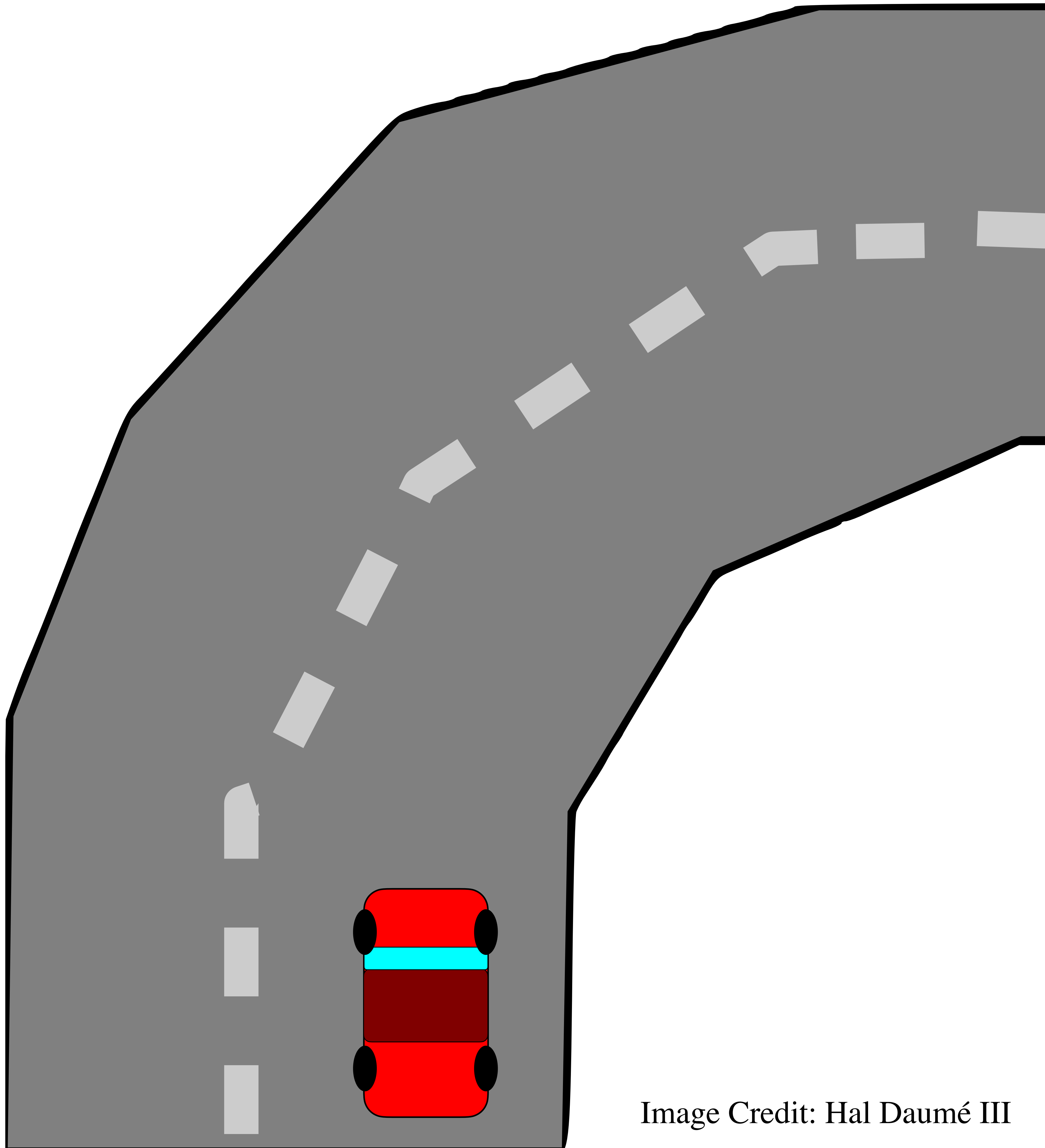


Image Credit: Hal Daumé III

Behavior Cloning

simplest imitation learning

1. Collect trajectories from expert $\pi^\star$

2. Create a dataset $(s_1, a_1), \ldots, (s_h, a_h) \sim \rho_{\pi^\star}$

3. Train a policy (classifier) π

$$\min_{\pi \in \Pi} \mathbb{E}_{(s,a)\sim\rho_{\pi^\star}} \left[L(\pi, s, a) \right]$$

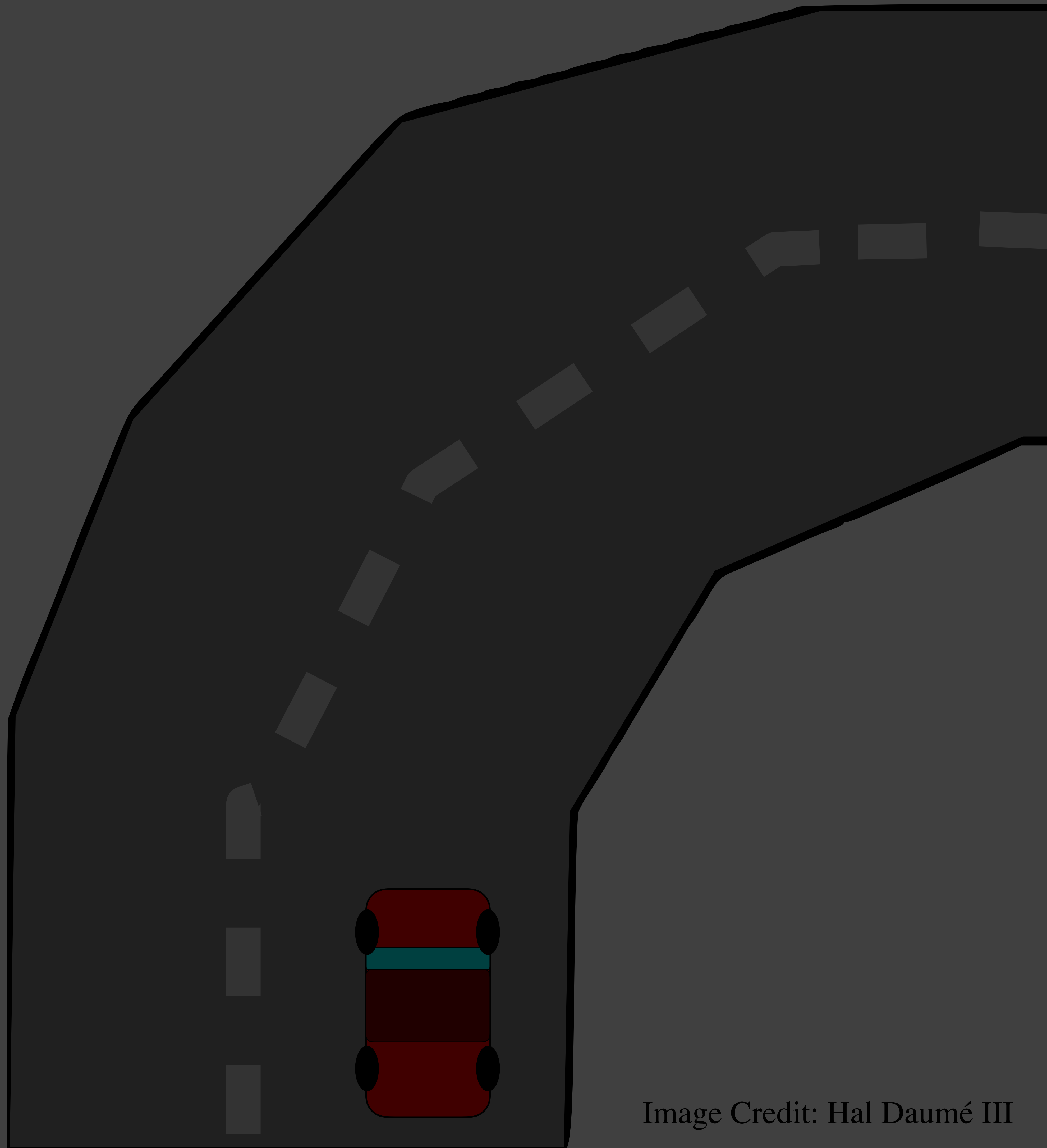


Image Credit: Hal Daumé III

Behavior Cloning

simplest imitation learning

1. **Collect trajectories from expert** $\pi^\star$

2. **Create a dataset** $(s_1, a_1), \ldots, (s_h, a_h) \sim \rho_{\pi^\star}$

3. **Train a policy (classifier)** π

$$\min_{\pi \in \Pi} \mathbb{E}_{(s,a)\sim\rho_{\pi^\star}} \left[L(\pi, s, a)\right] \boxed{\leq \epsilon}$$

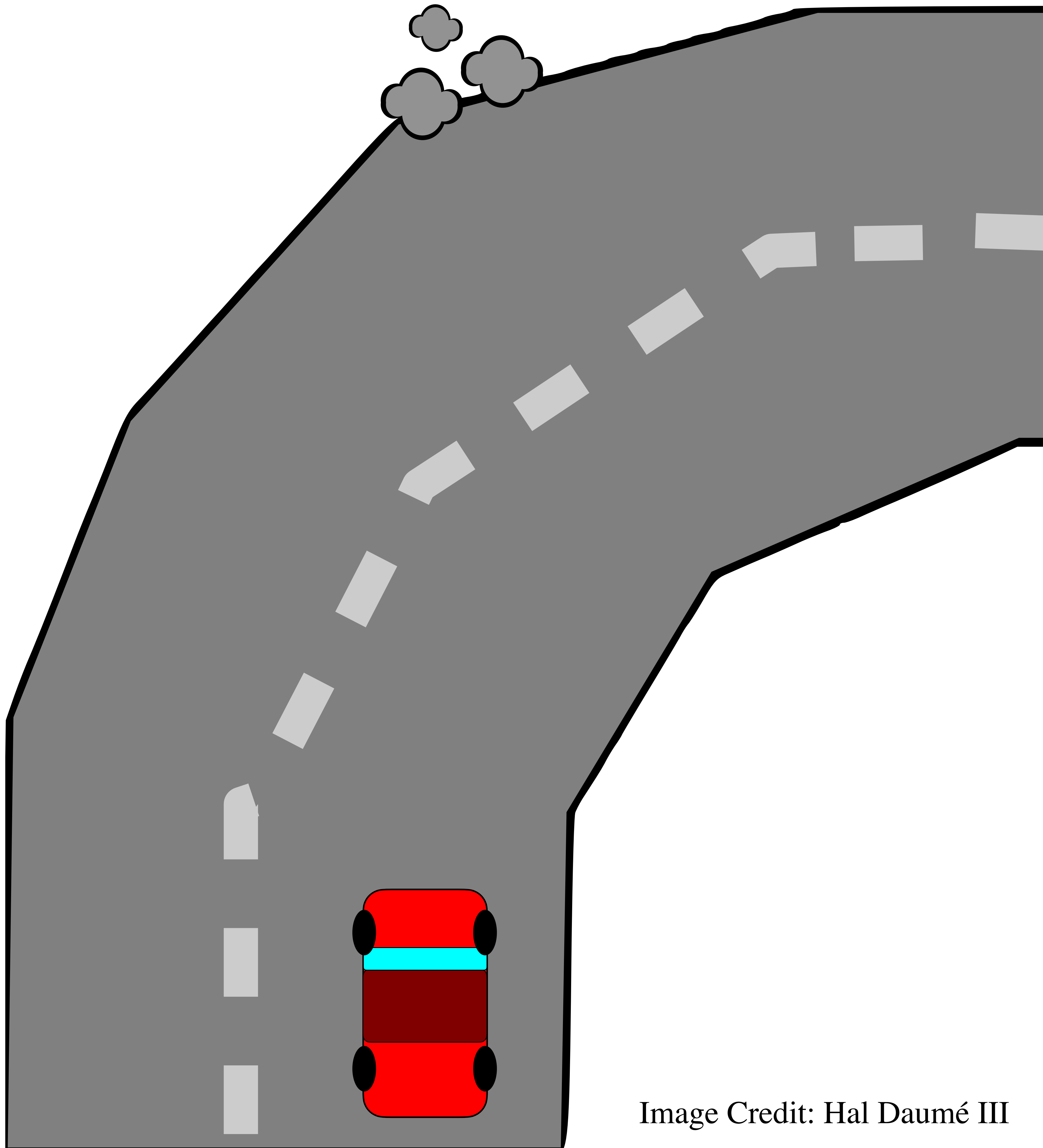


Image Credit: Hal Daumé III

Behavior Cloning

simplest imitation learning

1. **Collect trajectories from expert $\pi^\star$**

2. **Create a dataset $(s_1, a_1), \ldots, (s_h, a_h) \sim \rho_{\pi^\star}$**

3. **Train a policy (classifier) π**

$$\min_{\pi \in \Pi} \mathbb{E}_{(s,a)\sim\rho_{\pi^\star}} \left[L(\pi, s, a) \right] \boxed{\leq \epsilon}$$

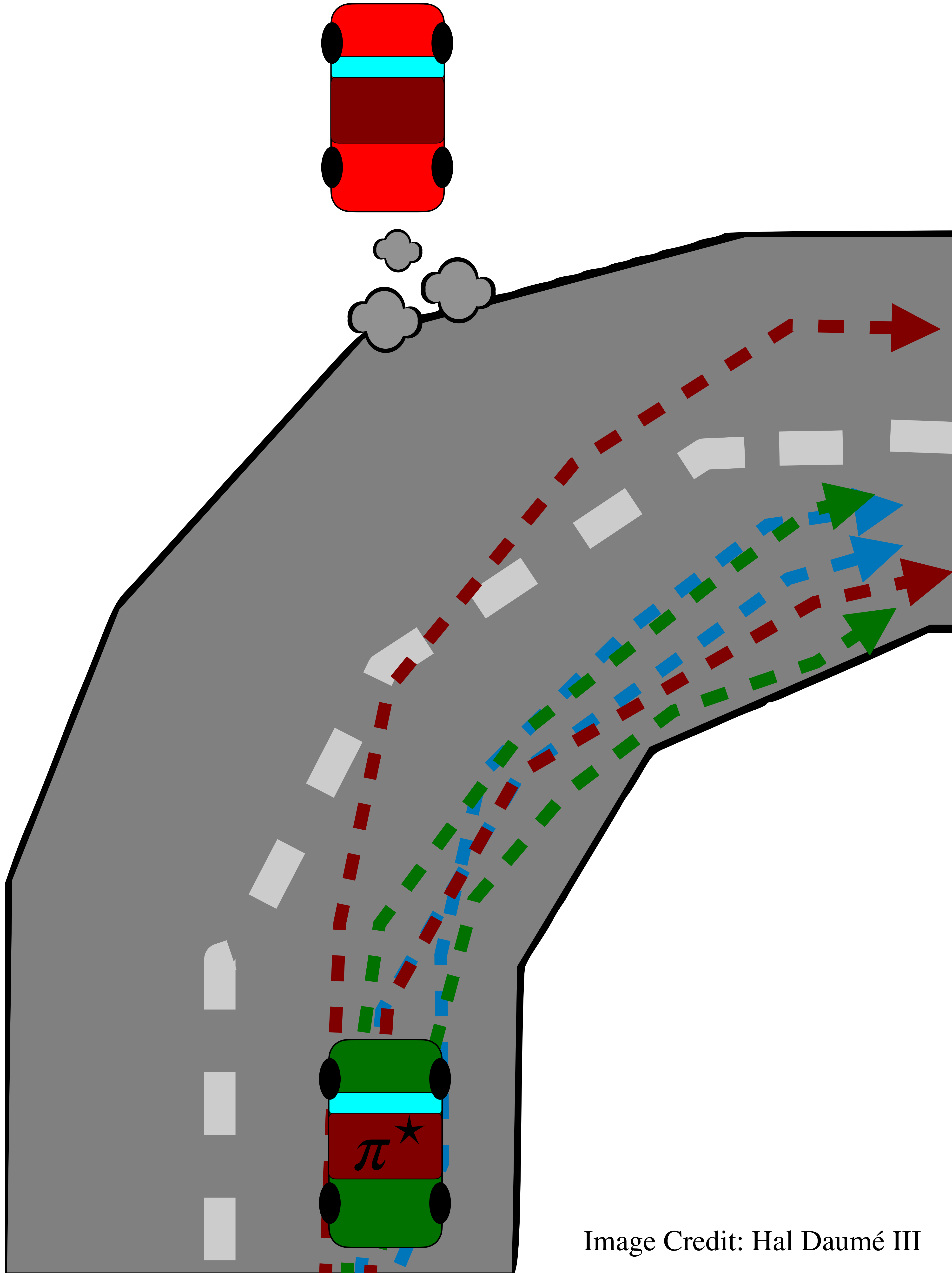


Image Credit: Hal Daumé III

Behavior Cloning

simplest imitation learning

1. Collect trajectories from expert $\pi^\star$

2. Create a dataset $(s_1, a_1), \ldots, (s_h, a_h) \sim \rho_{\pi^\star}$

3. Train a policy (classifier) π

$$\min_{\pi \in \Pi} \mathbb{E}_{(s,a)\sim\rho_{\pi^\star}} \left[L(\pi, s, a)\right] \leq \epsilon$$

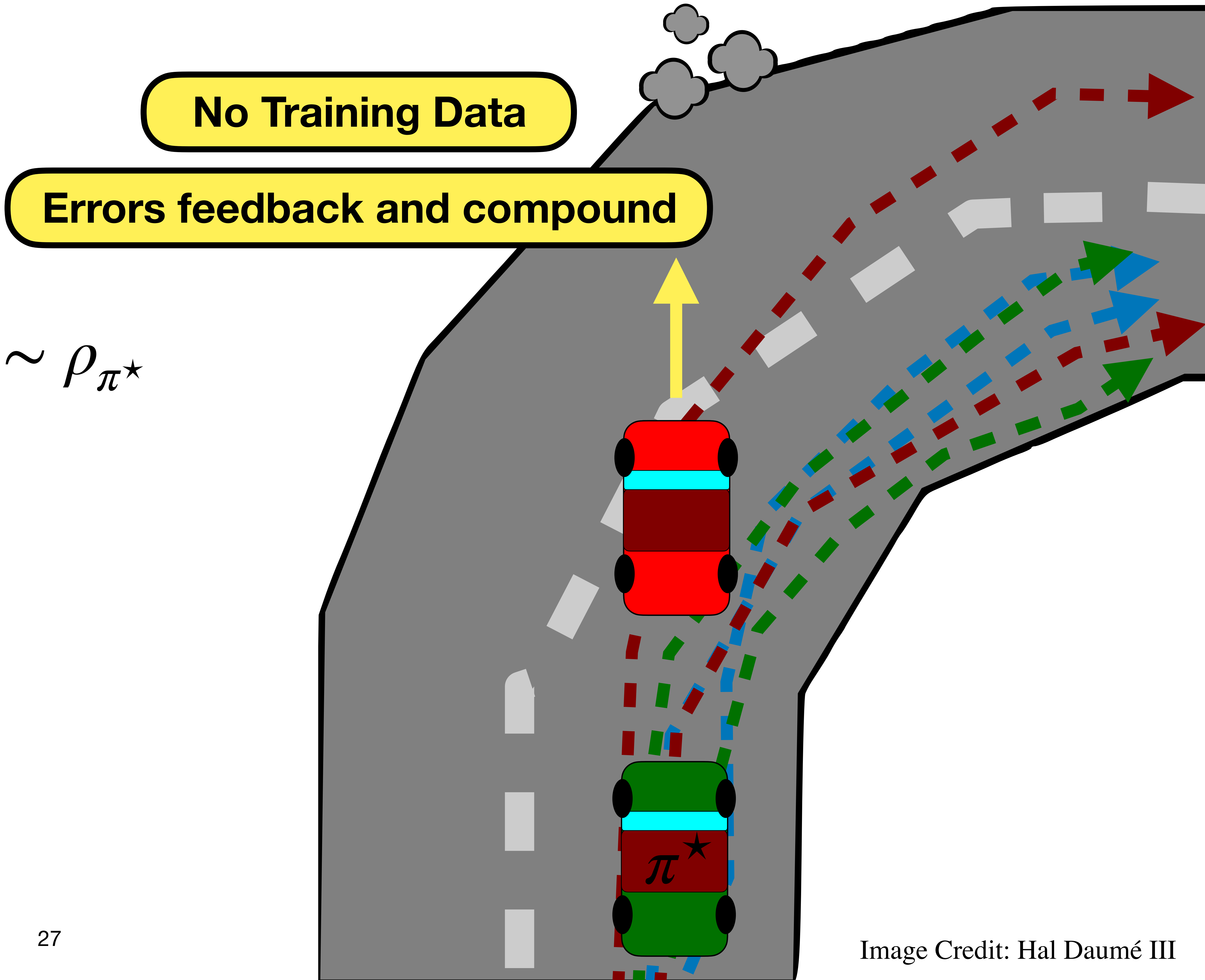


Image Credit: Hal Daumé III

Behavior Cloning

simplest imitation learning

1. Collect traje
2. Create a dat
3. Train a polic

$$\min_{\pi\in\Pi} \mathbb{E}_{(s,a)\sim\rho_{\pi^\star}}$$

Problem:

The training distribution is different than the test distribution resulting in covariate shift issues.

	Supervised Learning	Behavior Cloning
Train	$(x, y) \sim \rho$	$(s, a) \sim \rho_{\pi^\star}$
Test	$(x, y) \sim \rho$	$(s, a) \sim \rho_{\pi}$

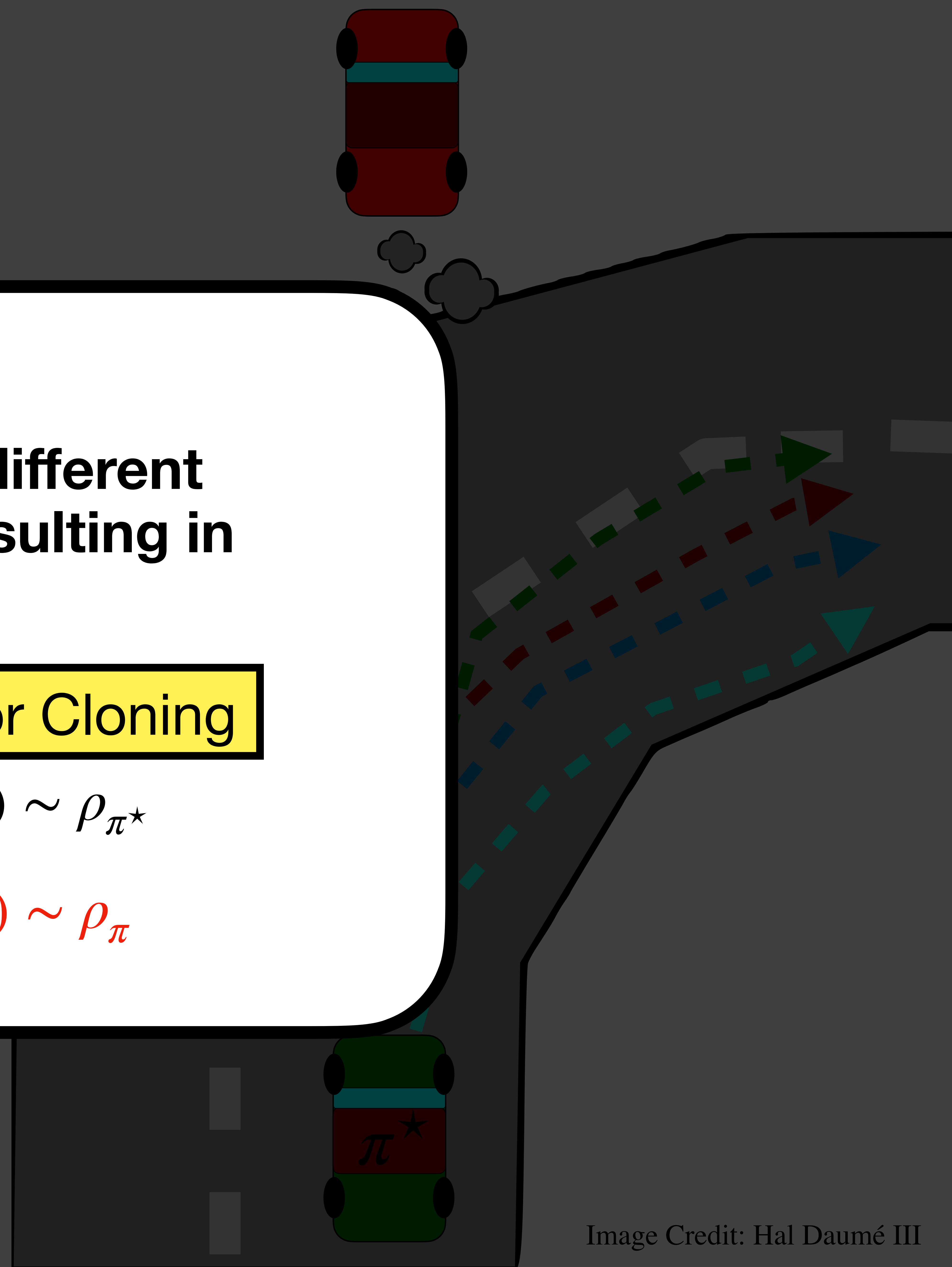
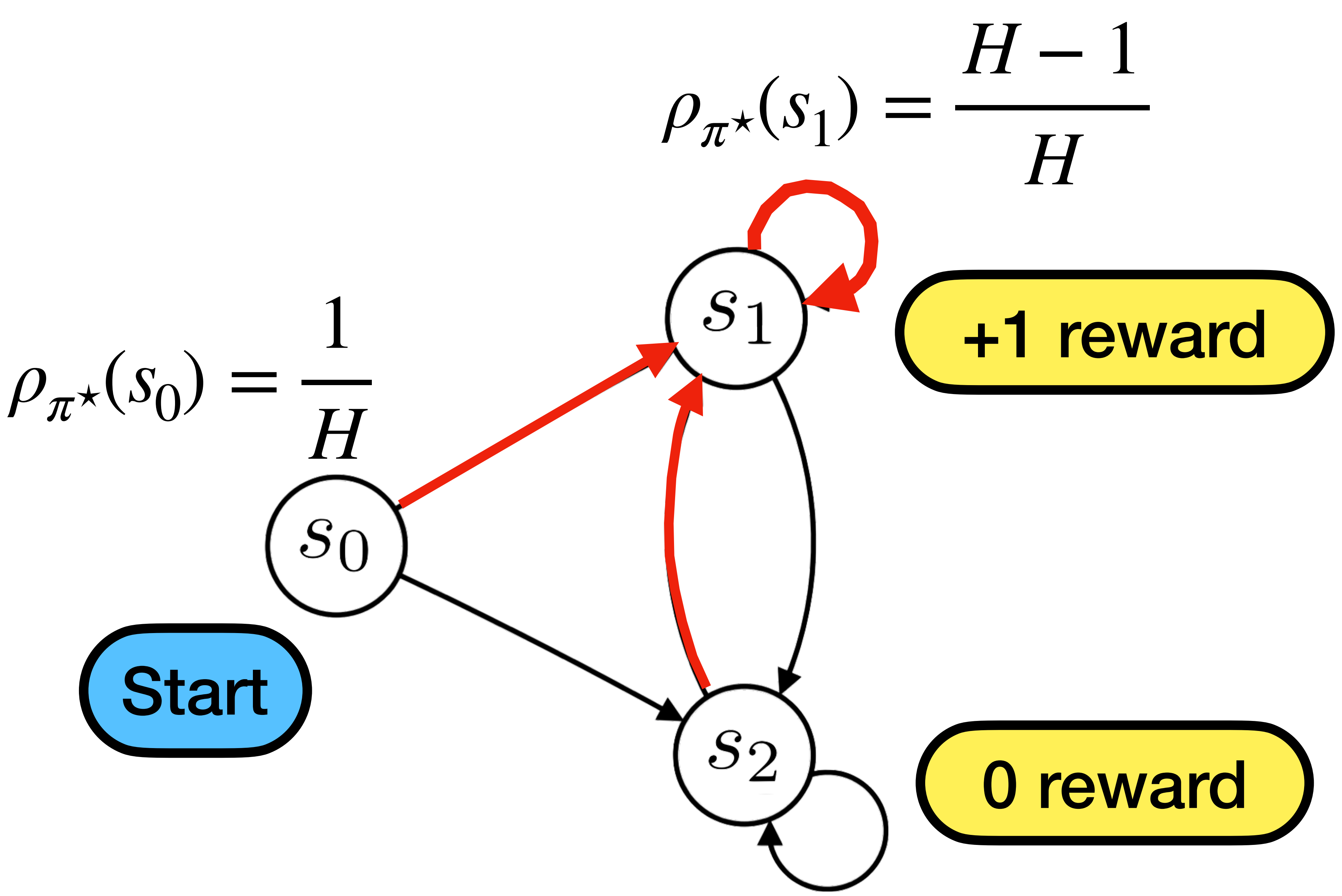


Image Credit: Hal Daumé III

Formalizing

the covariate shift issue

Given an expert policy: $\pi^\star$



Behavior Cloning Loss:

$\text{Loss}(\hat{\pi}) = \epsilon$

(loss is small)

Performance

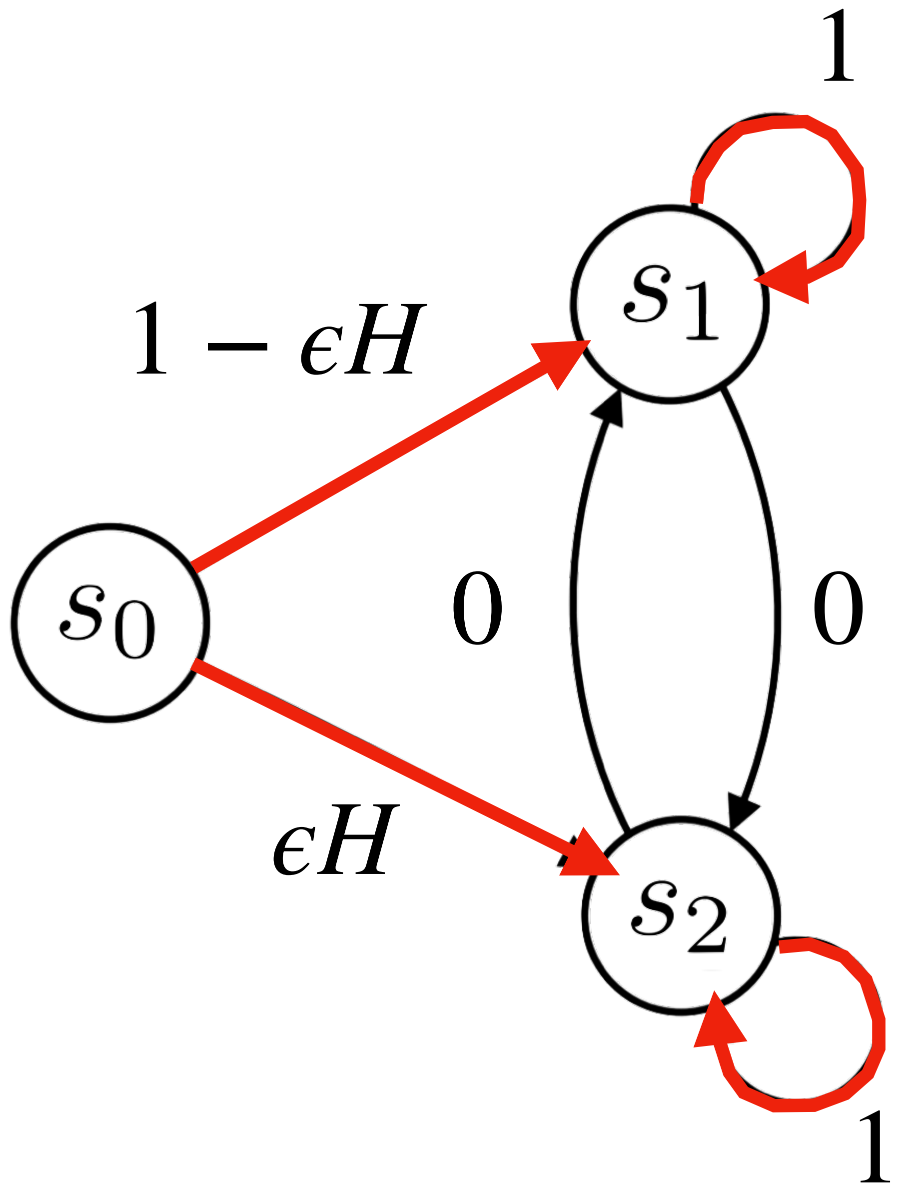
$$J(\pi) := \mathbb{E}_\pi \left[\sum_{h=1}^{H} R_h \right]$$

Behavior Cloning Performance Gap:

$$J(\pi^\star) - J(\hat{\pi}) \leq \mathcal{O}(\epsilon H^2)$$

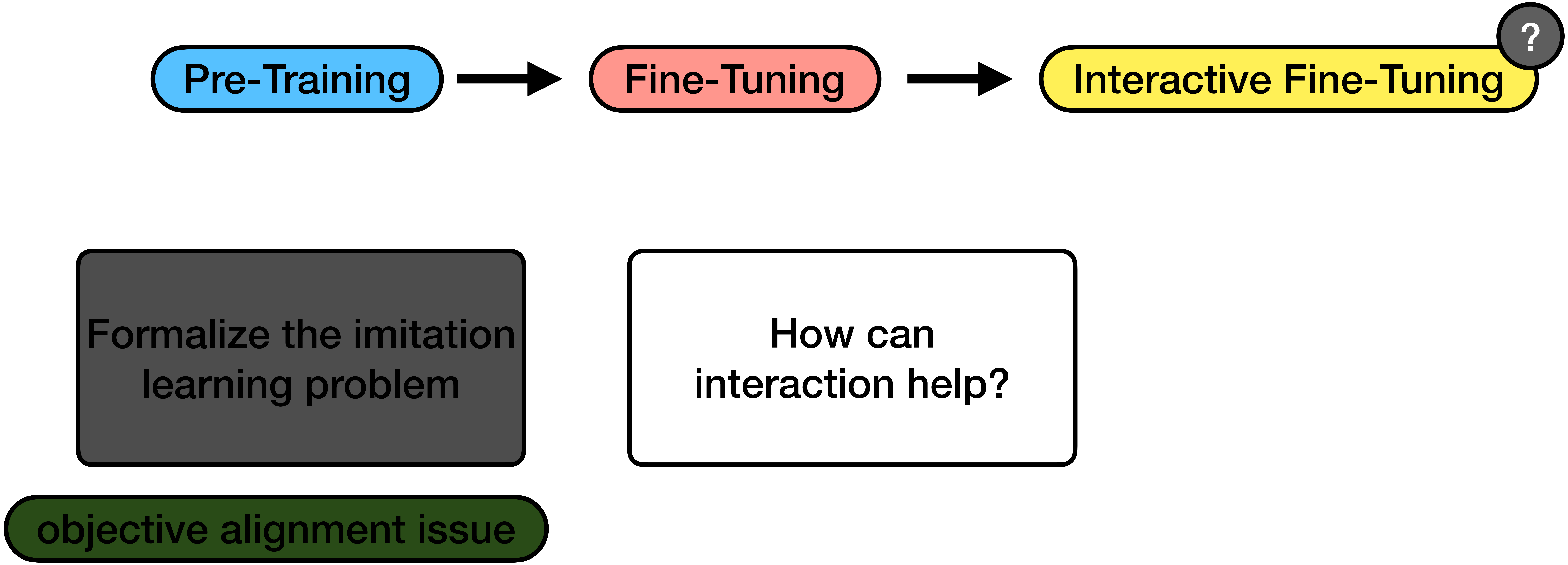
(gap scales quadratic)

Consider a policy: $\hat{\pi}$



Efficient Reductions for Imitation Learning, Ross & Bagnell, AISTATS 2010

Lower bounds for reductions, Matti Kaariainen, Atomic Learning Workshop 2006


Pre-Training
Fine-Tuning
Interactive Fine-Tuning
?
Formalize the imitation learning problem
How can interaction help?
objective alignment issue

Interactive Imitation Learning with DAgger

Let $\pi^\star$ be an expert user

Loop:

- **Interact in an environment** $(s, a) \sim \rho_\pi$
- **Get Dataset** $D_i = \{(s, \pi^\star(s)\}$
- **Aggregate dataset** $D \leftarrow D \cup D_i$
- **Train a policy π on dataset**

$$\min_{\pi \in \Pi} \mathbb{E}_{(s,a)\sim\rho_\pi}\left[L(\pi, s, \pi^\star(s)\right]$$

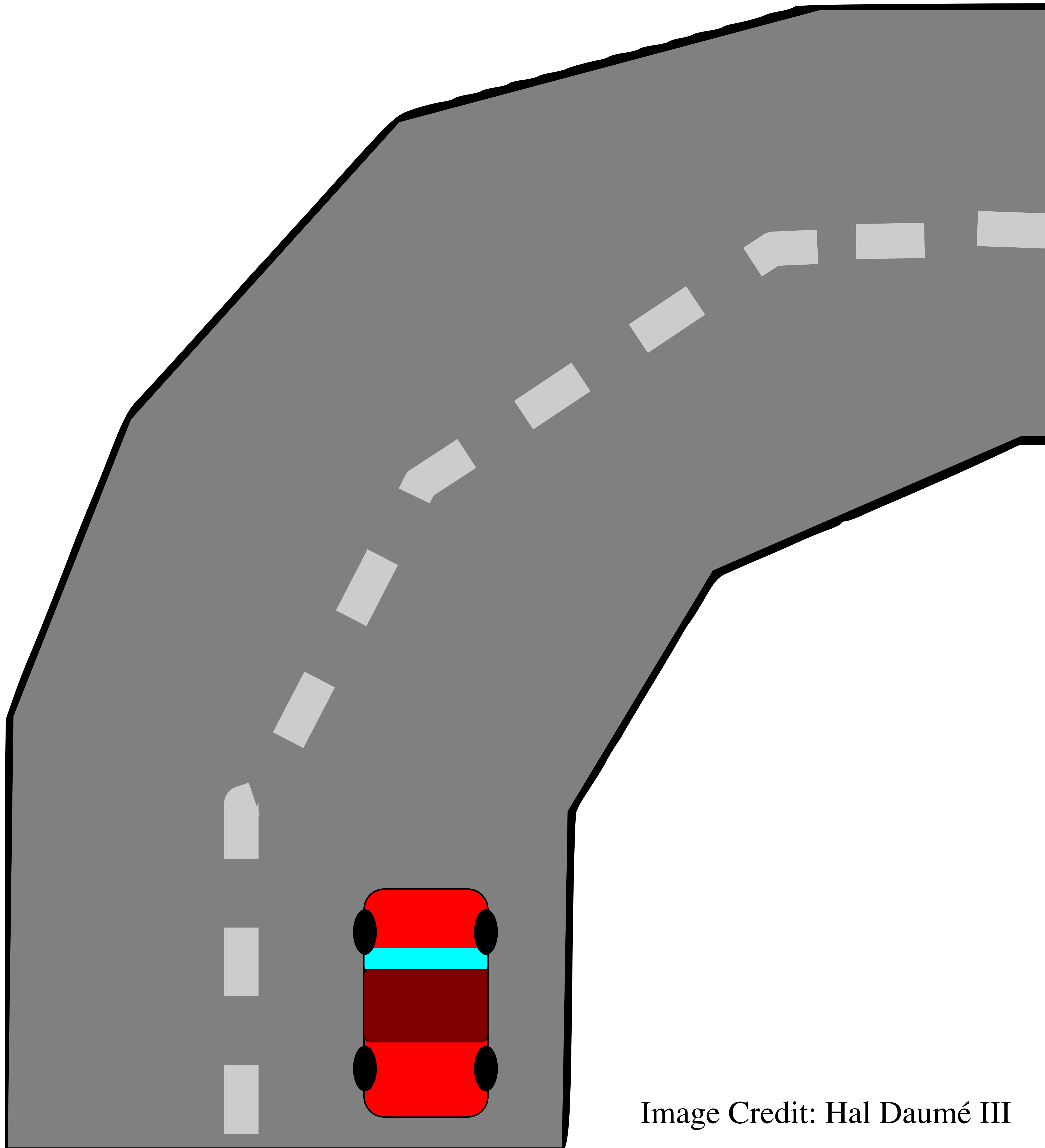


Image Credit: Hal Daumé III

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning, Ross 2011

Interactive Imitation Learning with DAgger

Let $\pi^\star$ be an expert user

Loop:

- **Interact in an environme**

- **Get Dataset** $D_i = \{(s, \pi^\star$

- **Aggregate dataset** $D \leftarrow$

- **Train a policy π on dataset**

$$\min_{\pi \in \Pi} \mathbb{E}_{(s,a)\sim\rho_\pi}\left[L(\pi, s, \pi^\star(s))\right]$$

Behavior Cloning Performance Gap:

$\mathcal{O}(\epsilon H^2)$

(gap scales quadratic)

DAgger Performance Gap:

$\mathcal{O}(\epsilon H)$

(gap scales linear)

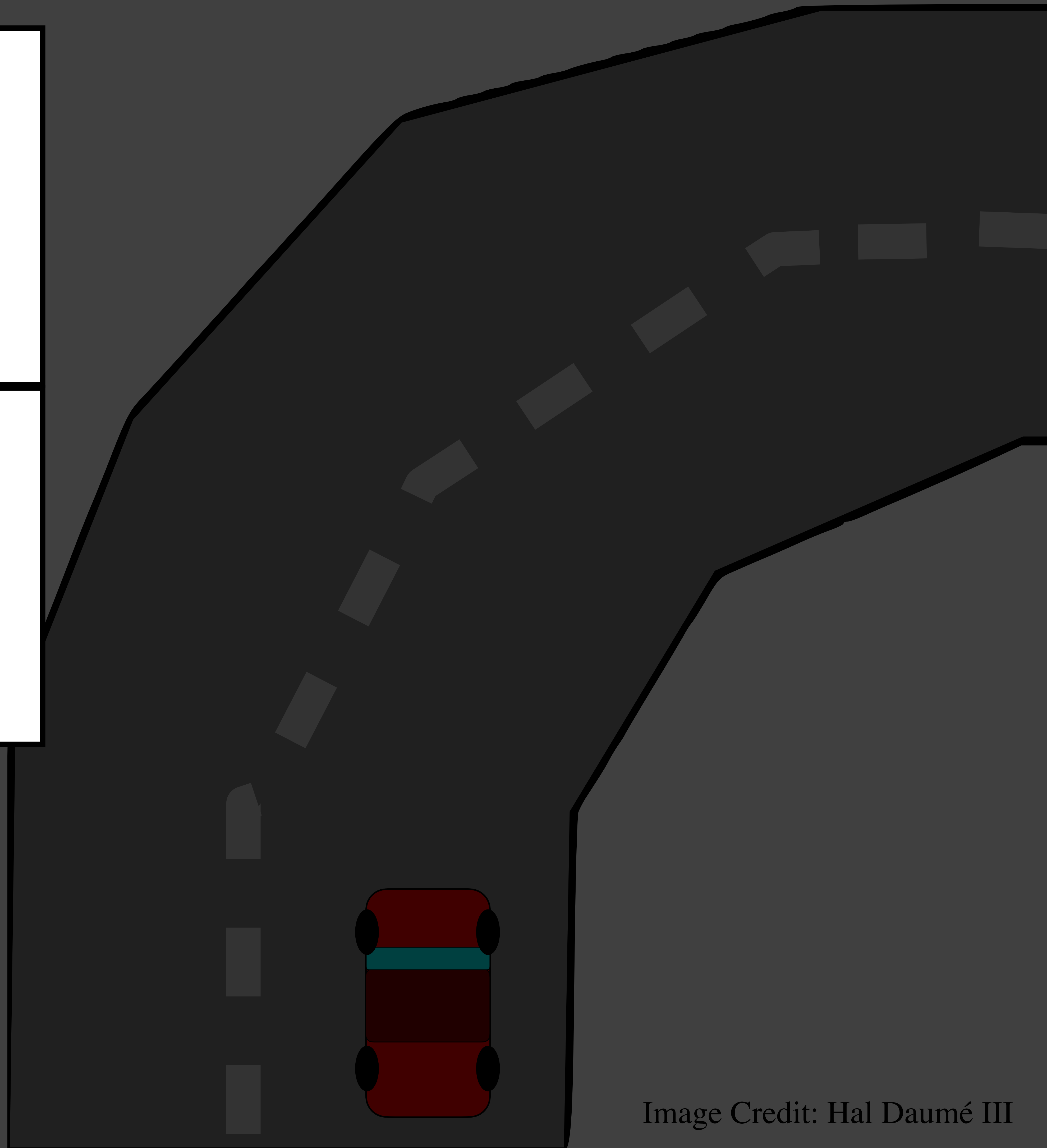


Image Credit: Hal Daumé III

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning, Ross 2011

Interactive Imitation Learning with DAgger

Let $\pi^\star$ be an expert user

Loop:

- **Interact in an environment** $(s, a) \sim \rho_\pi$
- **Get Dataset** $D_i = \{(s, \pi^\star(s)\}$
- **Aggregate dataset** $D \leftarrow D \cup D_i$
- **Train a policy π on dataset**

$$\min_{\pi \in \Pi} \mathbb{E}_{(s,a)\sim\rho_\pi} \left[L(\pi, s, \pi^\star(s)\right]$$

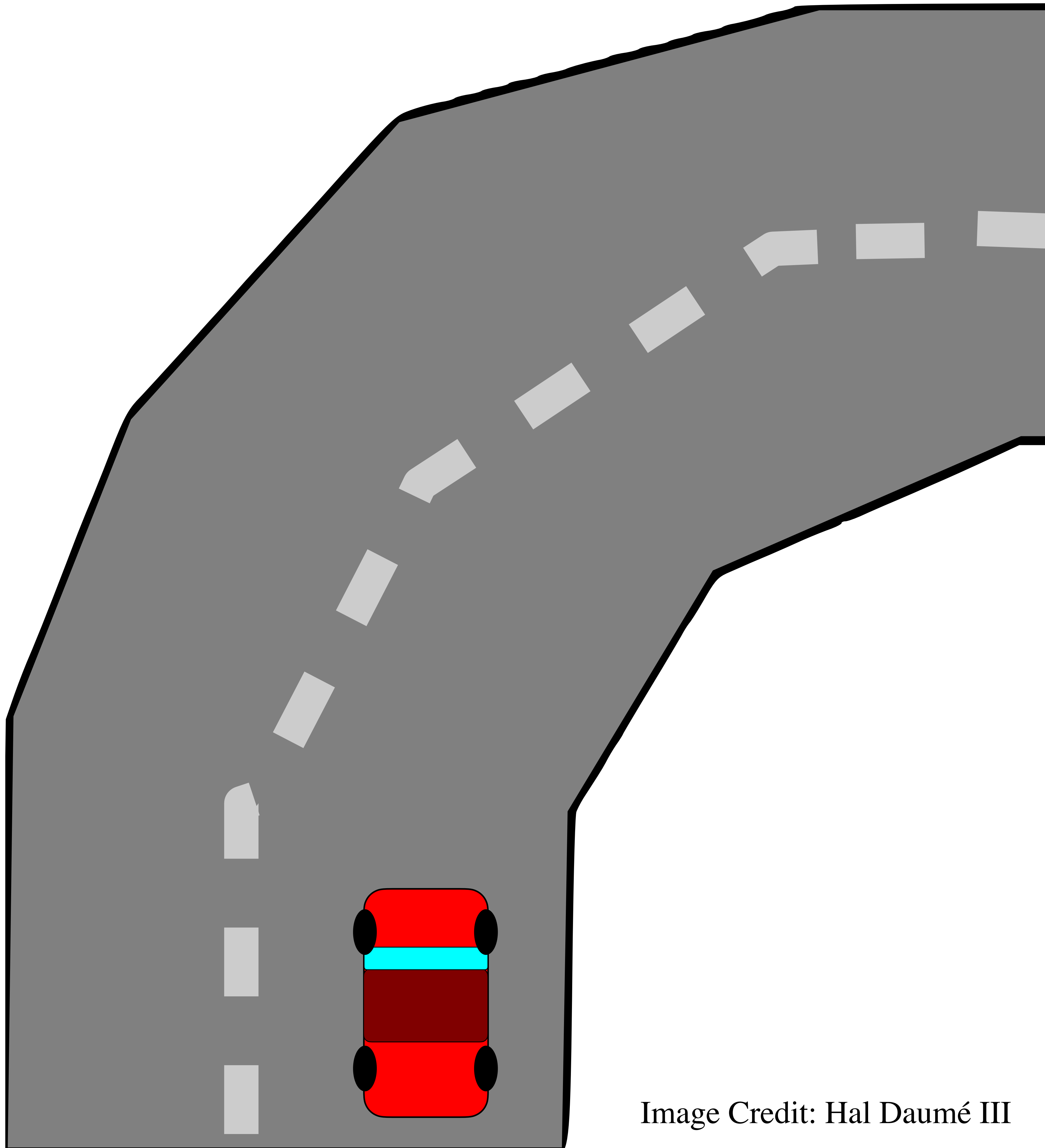


Image Credit: Hal Daumé III

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning, Ross 2011

Interactive Imitation Learning with DAgger

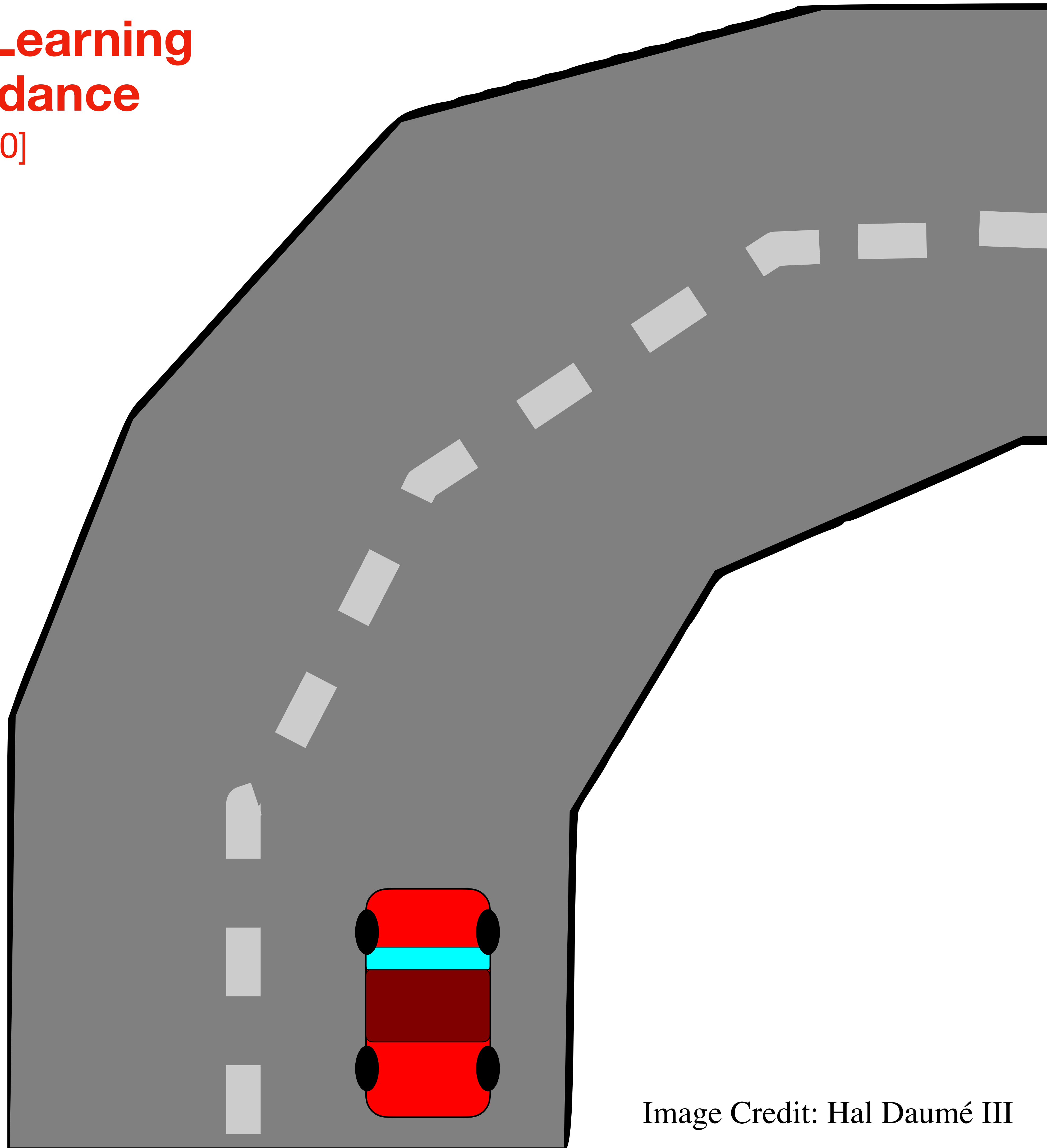
Let $\pi^\star$ be an expert user

Loop:

- Interact in an environment $(s, a) \sim \rho_\pi$
- If π is not confident:
  - Get Dataset $D_i = \{(s, \pi^\star(s)\}$
- Aggregate dataset $D \leftarrow D \cup D_i$
- Train a policy π on dataset

$$\min_{\pi \in \Pi} \mathbb{E}_{(s,a)\sim\rho_\pi} \left[L(\pi, s, \pi^\star(s)\right]$$

Active Imitation Learning with Noisy Guidance
[BSD ACL 2020]



A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning, Ross 2011

Image Credit: Hal Daumé III

Interactive Imitation Learning with DAgger

Let $\pi^\star$ be an ex

Loop:

- Interact in an environment $(s, a) \sim \rho_\pi$

- If π is not c

  - Get Da

- Aggregate

- Train a policy π on dataset

$$\min_{\pi \in \Pi} \mathbb{E}_{(s,a)\sim\rho_\pi}\left[L(\pi, s, \pi^\star(s)\right]$$

Problem:

We need online access to an expert to label each state we visit.

Research Question:

How can we remove the need for online access to an expert?

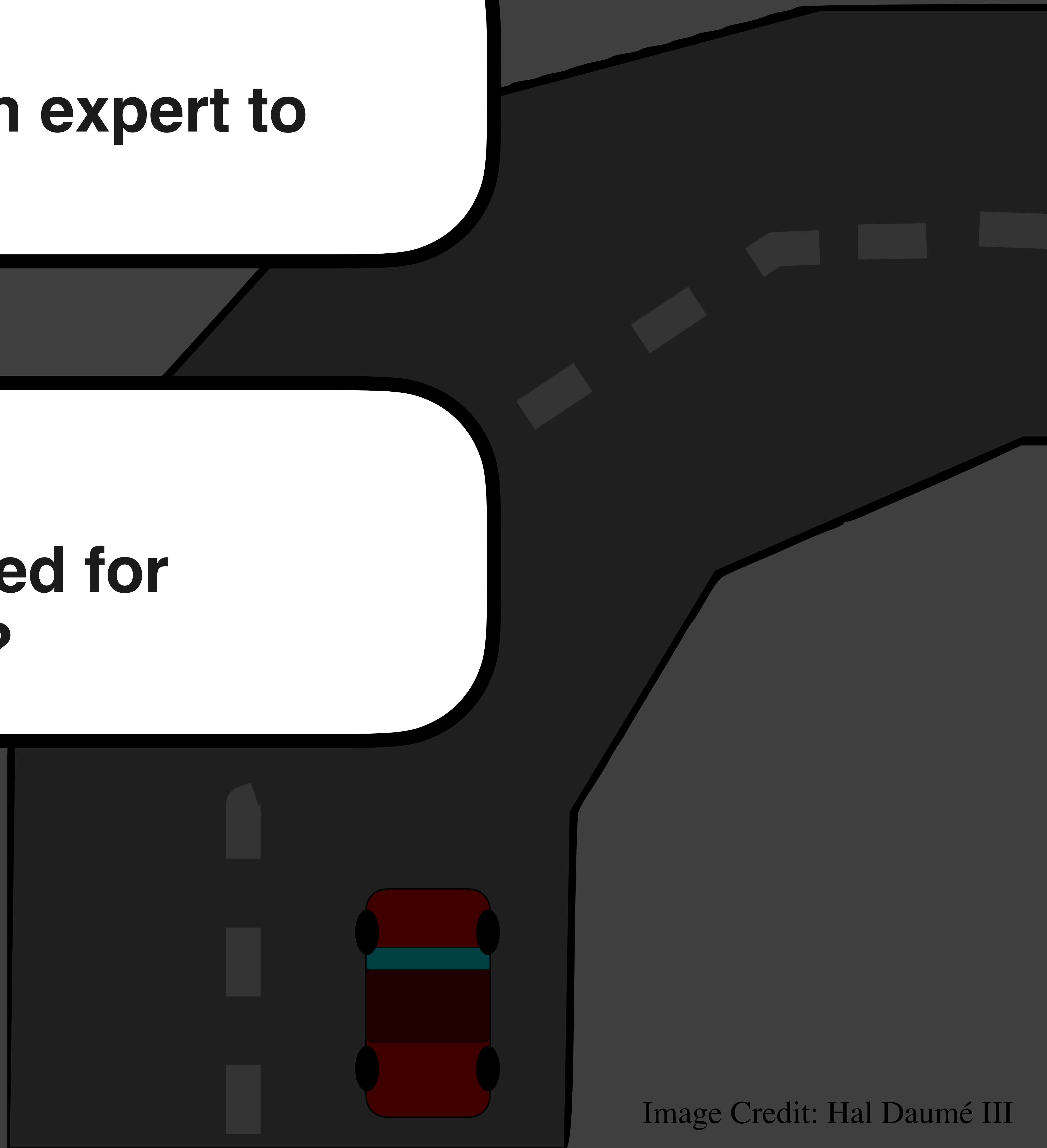


Image Credit: Hal Daumé III

Our Approach: DRIL [BSH ICLR 2020]

disagreement regularized imitation learning

$$\underbrace{\mathbb{E}_{(s,a)\sim\rho_\pi}\left[C_U(s,a)\right]}_{\text{Stay within the expert distribution}} + \lambda \overbrace{D_{\mathrm{KL}}(\rho_{\pi^\star} \,||\, \rho_\pi)}^{\text{Mimic expert within the expert distribution}}$$

Train ensemble of polices $\Pi_E = \{\pi_1, \ldots, \pi_E\}$ **on expert data**

Uncertainty Cost: $C_{\mathrm{U}}(s,a) = \mathrm{Var}_{\pi\sim\Pi_{\mathrm{E}}}(\pi(a\,|\,s))$

DRIL **cost can be optimized using any RL algorithm**

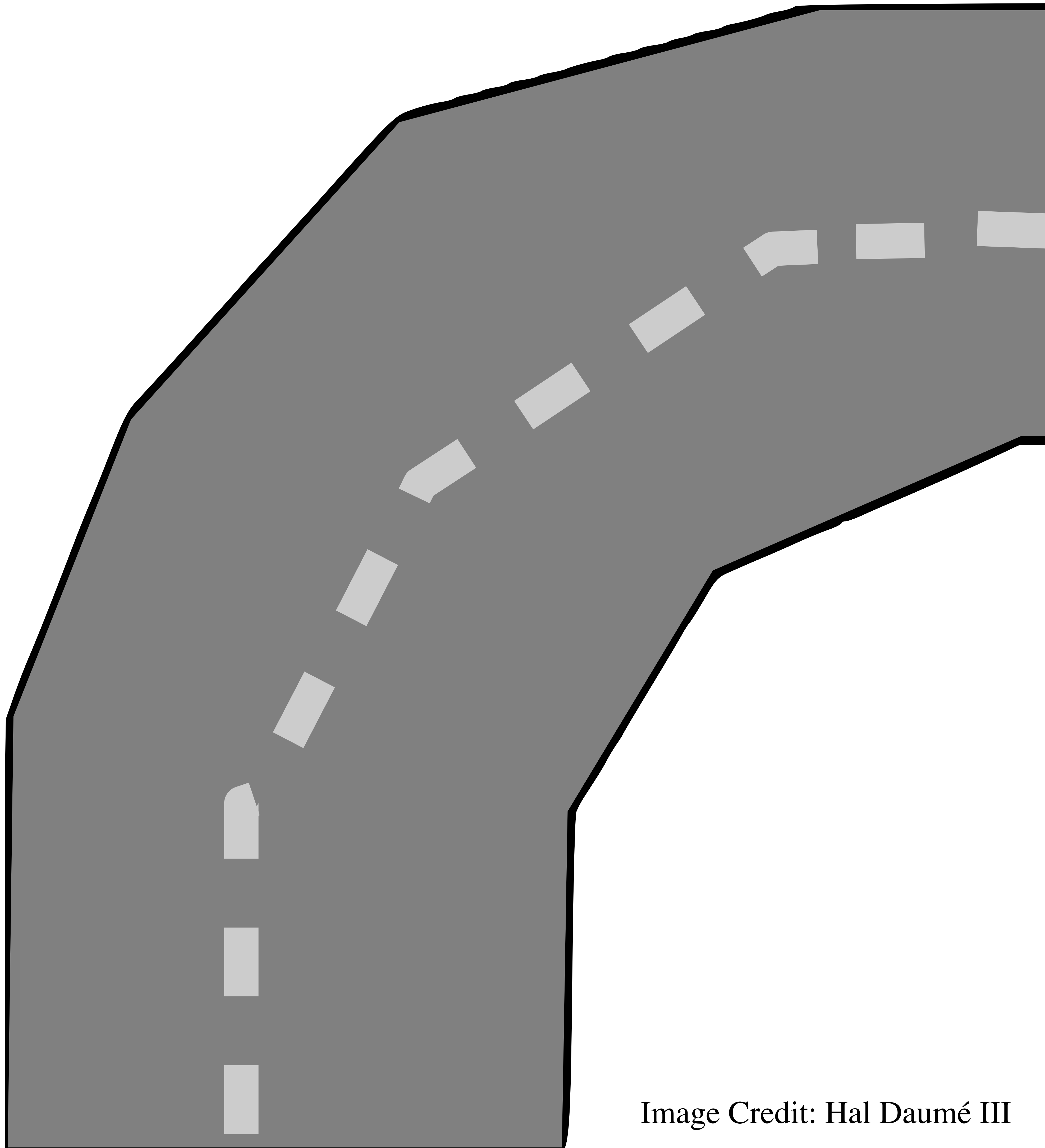


Image Credit: Hal Daumé III

Our Approach: DRIL [BSH ICLR 2020]

disagreement regularized imitation learning

Mimic expert within the expert distribution

$$\underbrace{\mathbb{E}_{(s,a)\sim\rho_\pi}\left[C_U(s,a)\right]}_{\text{Stay within the expert distribution}} + \overbrace{\lambda D_{\mathrm{KL}}(\rho_{\pi^\star} \,||\, \rho_\pi)}^{\text{Mimic expert within the expert distribution}}$$

Stay within the expert distribution

Train ensemble of polices $\Pi_E = \{\pi_1, \pi_2, \pi_3\}$ **on expert data**

Uncertainty Cost: $C_{\mathrm{U}}(s,a) = \mathrm{Var}_{\pi\sim\Pi_{\mathrm{E}}}(\pi(a\,|\,s))$

DRIL **cost can be optimized using any RL algorithm**

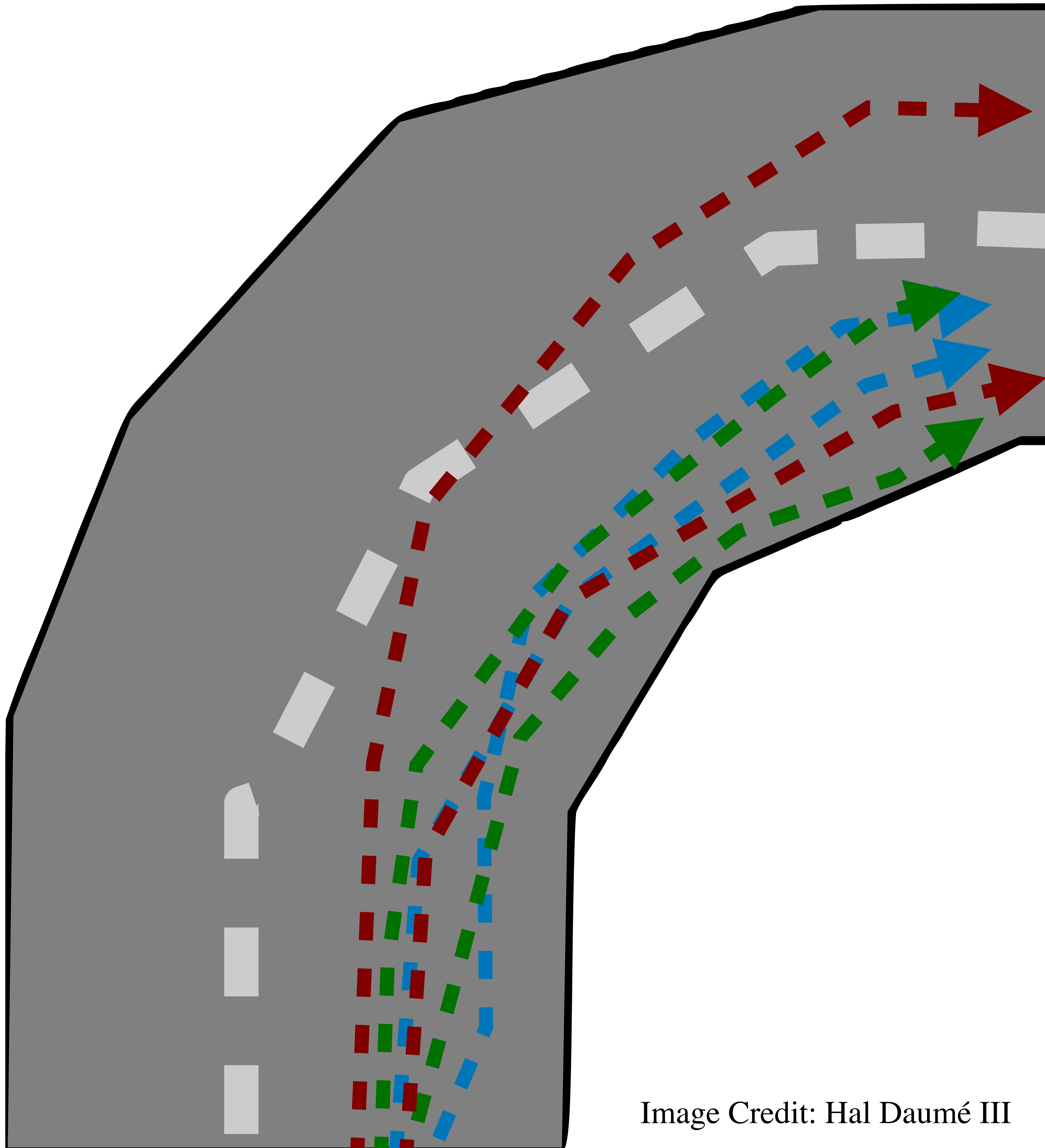


Image Credit: Hal Daumé III

Our Approach: DRIL [BSH ICLR 2020]

disagreement regularized imitation learning

$$\underbrace{\mathbb{E}_{(s,a)\sim\rho_\pi}\left[C_U(s,a)\right]}_{\text{Stay within the expert distribution}} + \overbrace{\lambda D_{KL}(\rho_{\pi^\star}||\rho_\pi)}^{\text{Mimic expert within the expert distribution}}$$

Train ensemble of polices $\Pi_E = \{\pi_1, \pi_2, \pi_3\}$ on expert data

Uncertainty Cost: $C_U(s,a) = \mathrm{Var}_{\pi\sim\Pi_E}(\pi(a\,|\,s))$

DRIL **cost can be optimized using any RL algorithm**

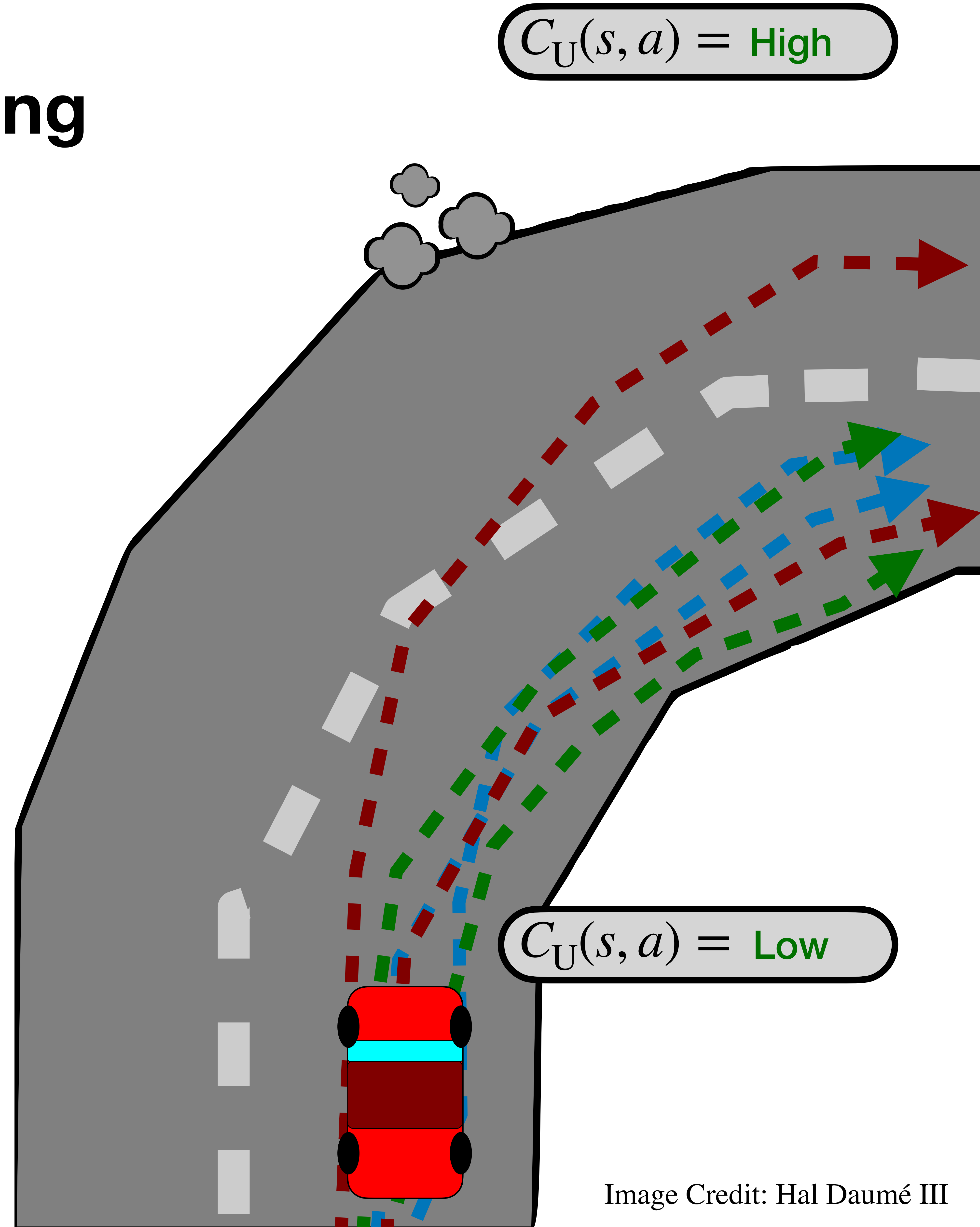


Image Credit: Hal Daumé III

DRIL

algorithm

Train Policy Ensemble $\Pi_E = \{\pi_1, \ldots, \pi_E\}$ **using demonstration data** D

Initialize π **using behavior cloning**

Loop:

- **Interact in an environment** $(s, a) \sim \rho_\pi$
- **Minimize** $D_{\mathrm{KL}}(\rho_{\pi^\star} || \rho_\pi)$ **using mini-batch of expert data**
- **Minimize** $C_U(s, a)$ **with the collected samples using RL**

Theory of DRIL

Performance gap compared to globally optimal policy $\pi^\star$

Behavior cloning $Loss(\pi) = \epsilon$

$$J(\pi^\star) - J(\hat{\pi}) \leq \mathcal{O}\left(\kappa \epsilon H\right)$$

$$\kappa = \min_{\mathcal{U} \subseteq \mathcal{S}} \frac{\max_{\pi \in \Pi} \sup_{s \in \mathcal{U}} \frac{d_\pi(s)}{d_{\pi^*}(s)}}{\min_{s \notin \mathcal{U}, a \in A} \mathrm{VAR}_{\pi \sim \Pi_{\mathrm{E}}}[\pi(a \mid s)]}$$

Concentrability inside of $\mathcal{U}$

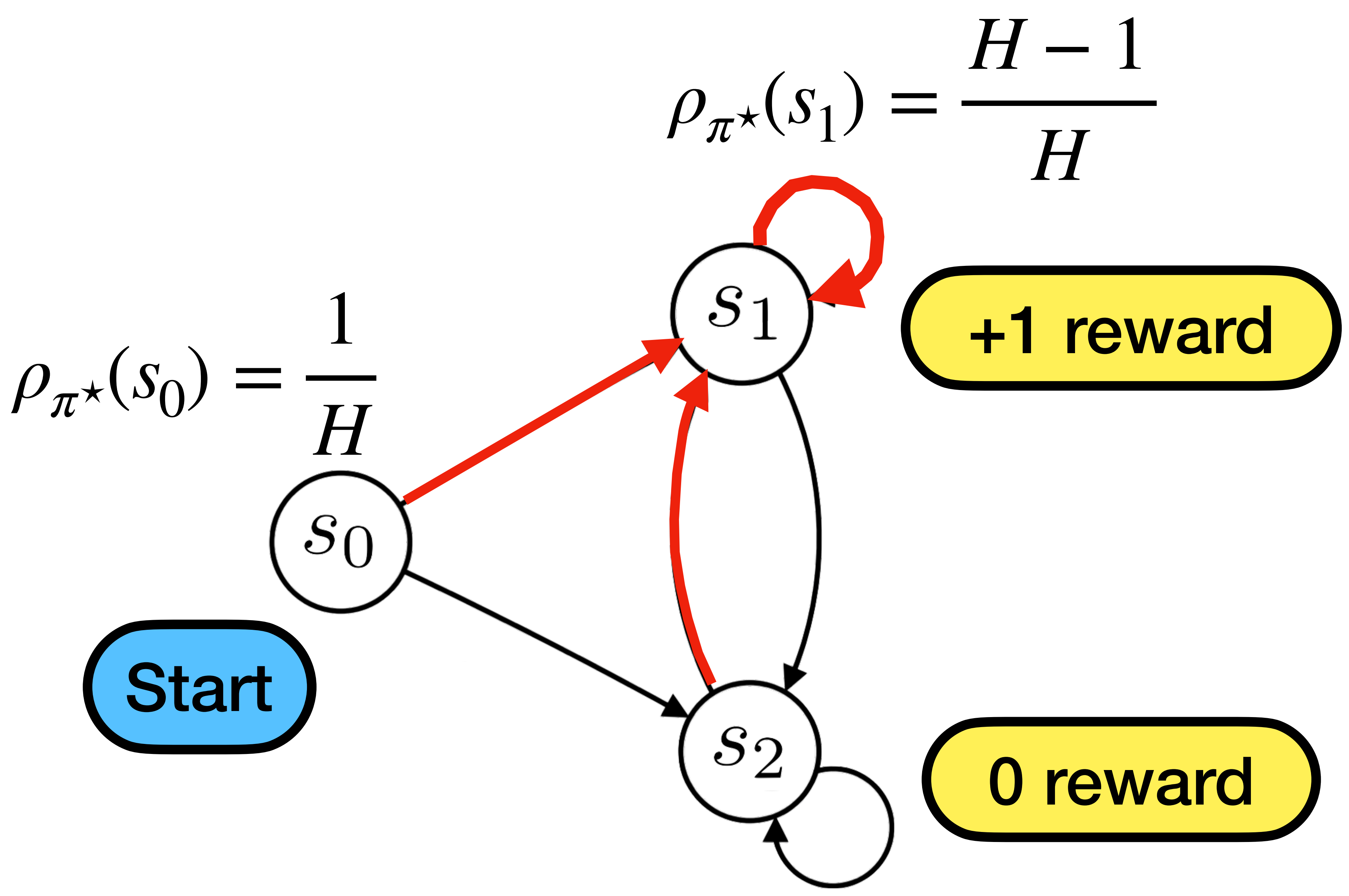
Minimum ensemble variance outside of $\mathcal{U}$

κ problem dependent. Depends on environment dynamics, expert distribution, and uncertainty of ensemble

Revisiting

the covariate shift problem

Given an expert policy: $\pi^\star$



Behavior Cloning
Performance Gap:

$\mathcal{O}(\epsilon H^2)$

(gap scales quadratic)

DAgger
Performance Gap:

$\mathcal{O}(\epsilon H)$

(gap scales linear)

DRIL
Performance Gap:

$\mathcal{O}(\epsilon H)$

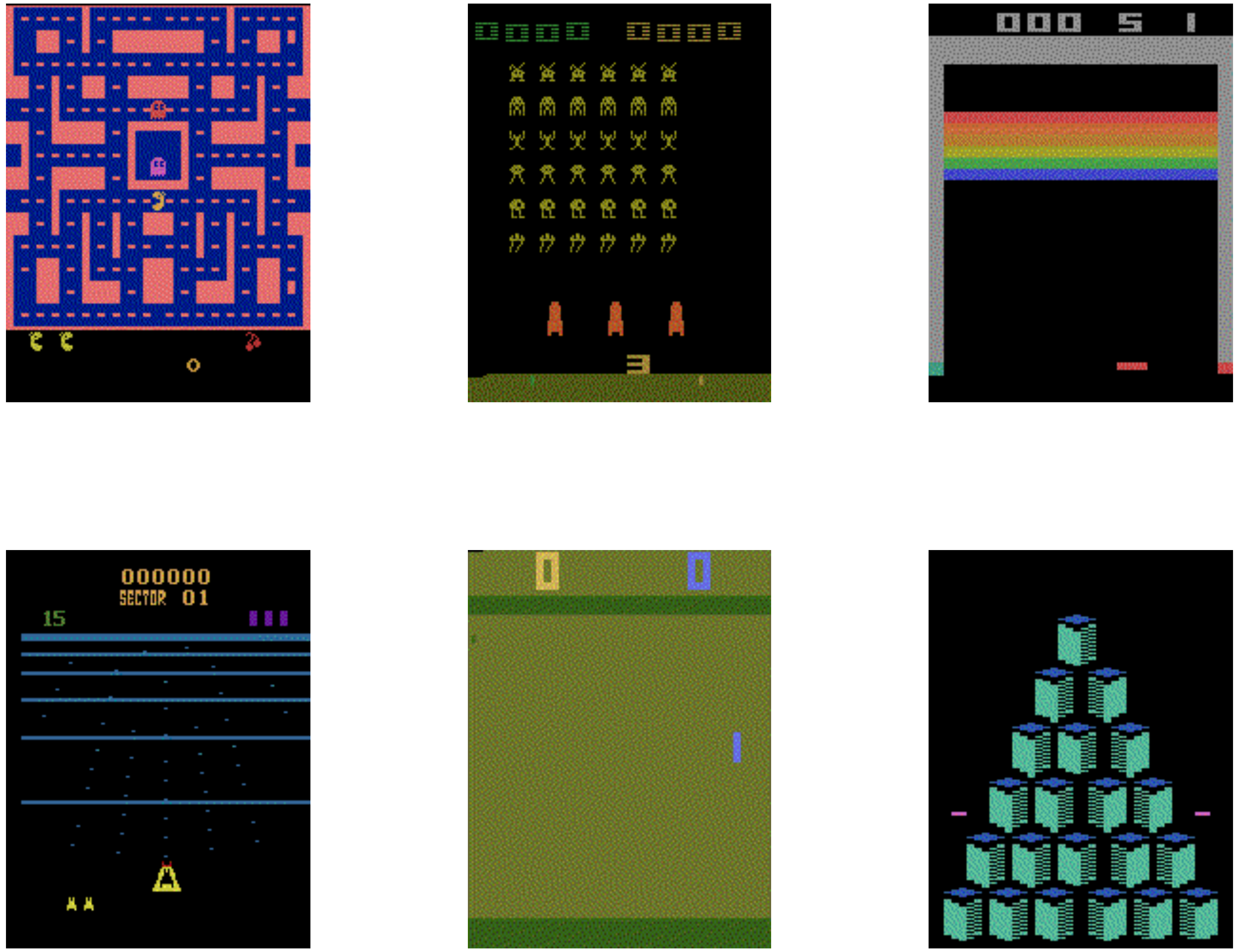
$\kappa = 12 + \mathcal{O}(\frac{1}{\sqrt{|\text{ensemble}|}})$

(gap scales linear)

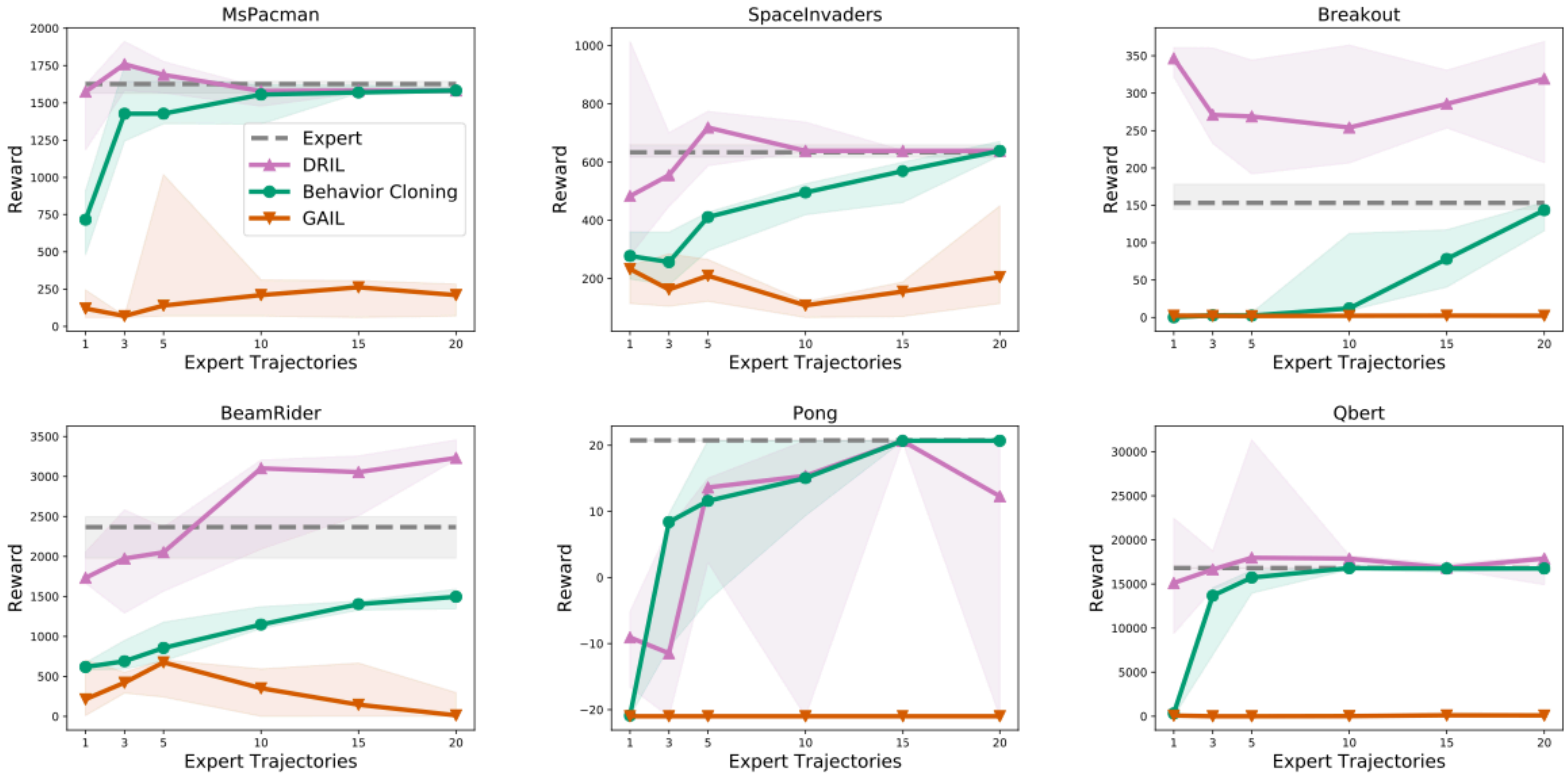
Experimental Setup

- Generated demonstration data from pre-trained experts
- Then trained DRIL on the demonstration data and evaluated on true environment reward

Atari
Imitation Learning Benchmark



Experimental Results



Impact of our work

extended to LLMs [Coste et al. ICLR 2024]

Task

Alpaca datasets: General instruction fine-tuning dataset incorporating wide range of tasks

Dataset Composition

52,000 instructions and demonstrations generated from GPT 3.5

Example of Instructions

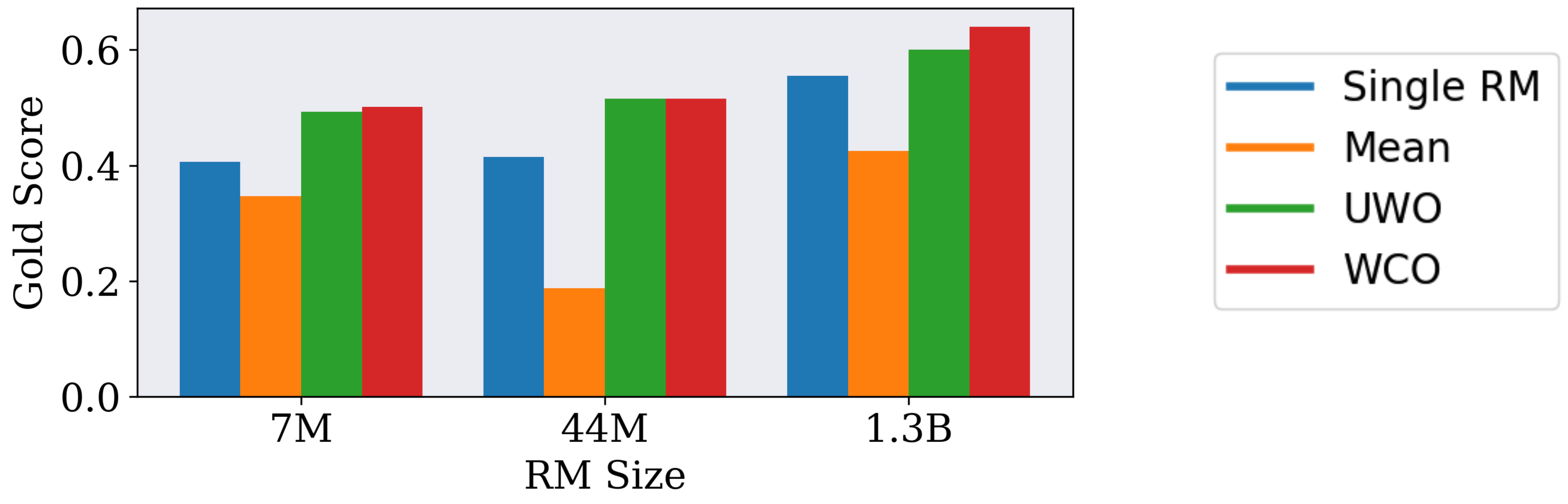
What is the capital of France?	The capital of France is Paris.
Describe the function of a computer motherboard	A computer motherboard is the main circuit board in a computer. It is the backbone of …

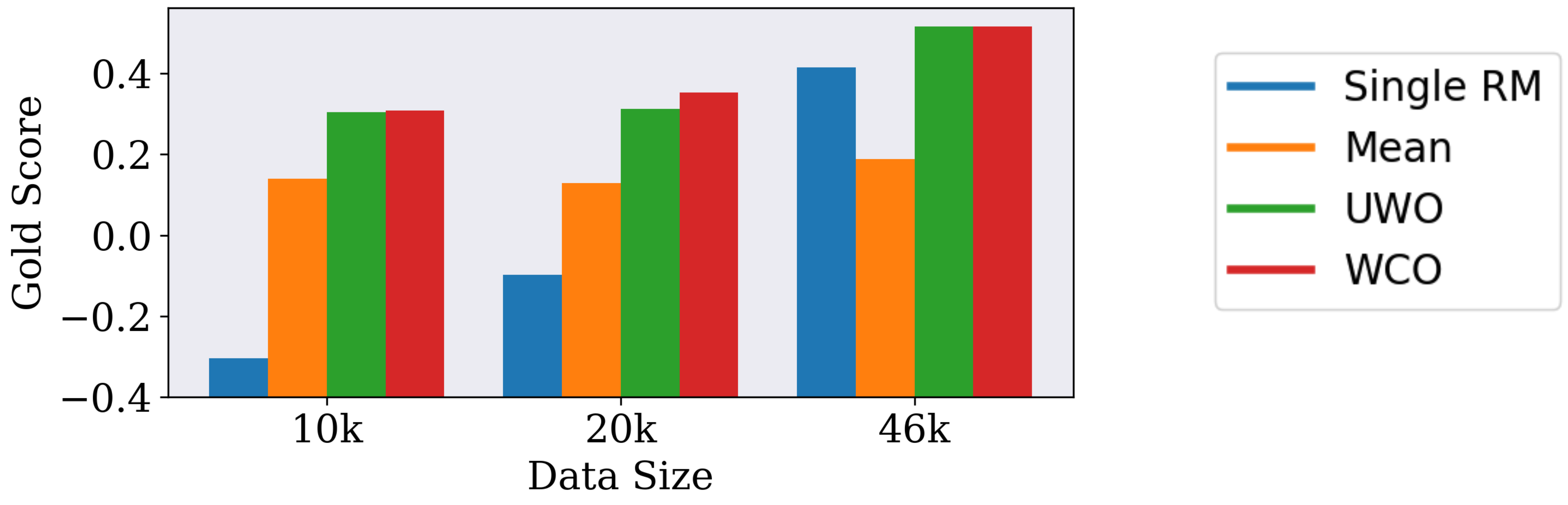
Impact of our work

extended to LLMs [Coste et al. ICLR 2024]

Uncertainty-Weighted Optimization (UWO):

UWO works by penalizing the policy for generating responses for which there is high disagreement among reward models within the ensemble

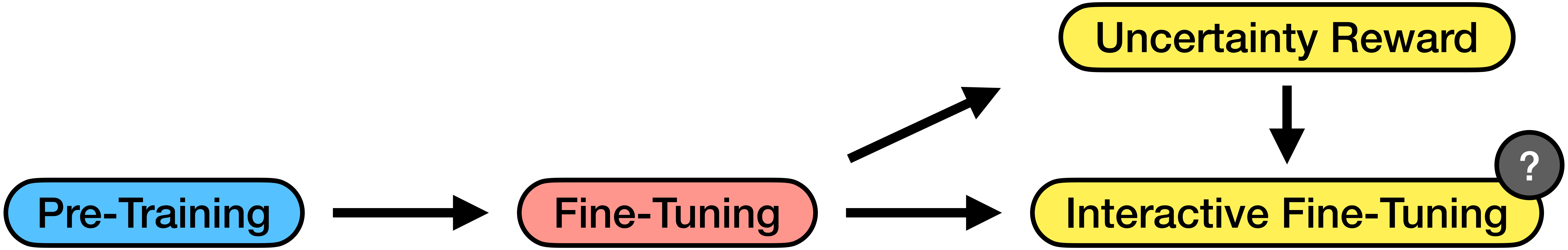




Reward Model Ensembles Help Mitigate Overoptimization Coste 2024

Summary

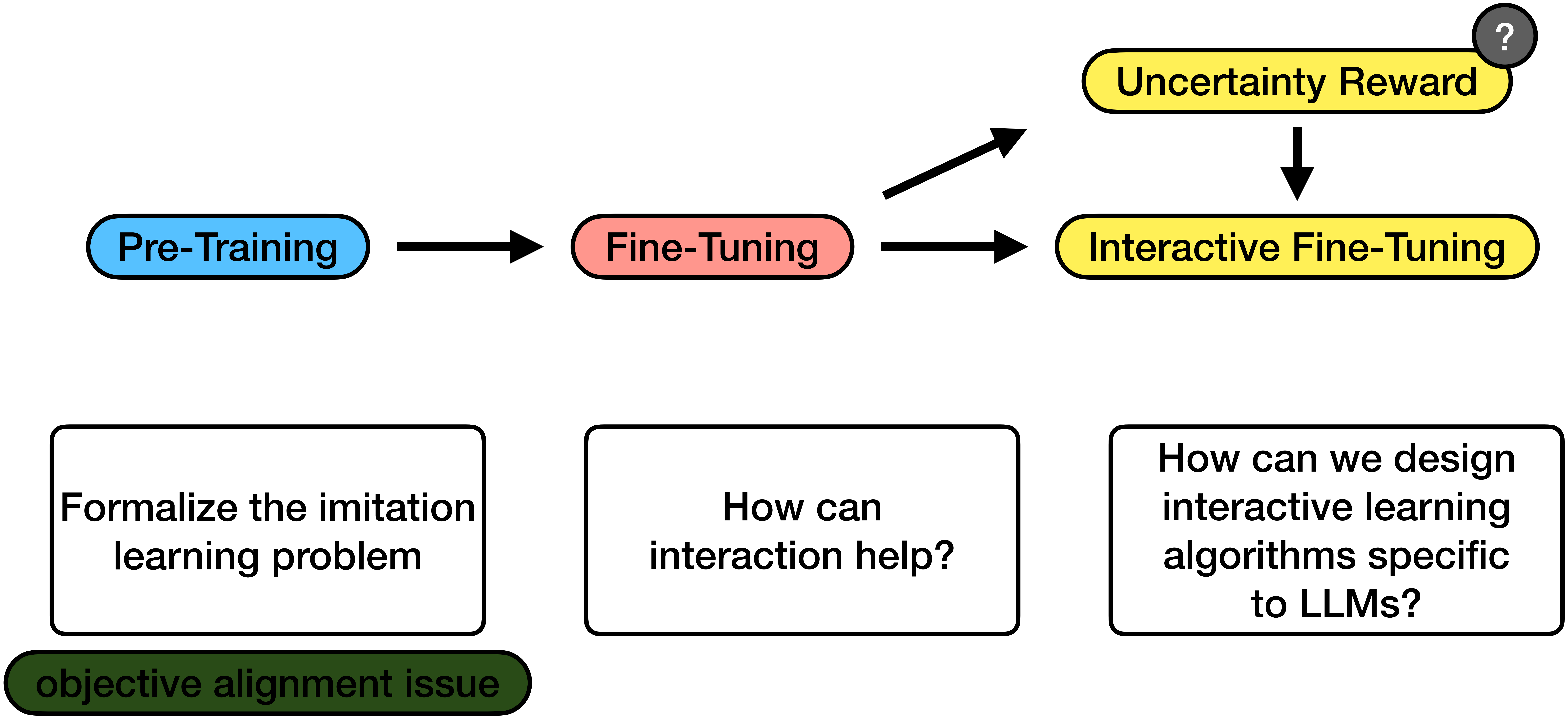
- Interacting in an environment can help address covariate shift issues
- DRIL has theoretical guarantees in some settings
- DRIL empirically performs better than Behavior Cloning



Formalize the imitation learning problem

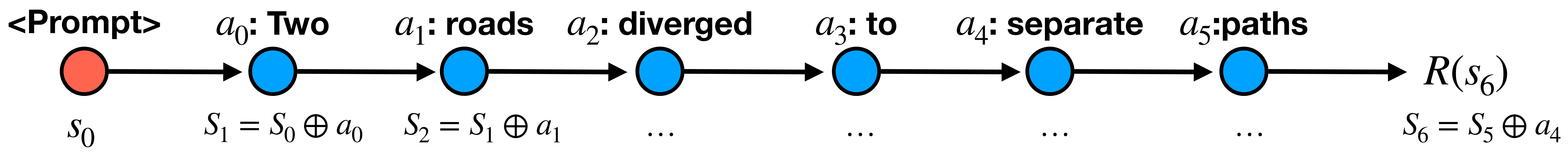
objective alignment issue

How can interaction help?


Uncertainty Reward
?
Pre-Training
Fine-Tuning
Interactive Fine-Tuning
Formalize the imitation learning problem
How can interaction help?
How can we design interactive learning algorithms specific to LLMs?
objective alignment issue

Reinforcement Learning

for text-generation

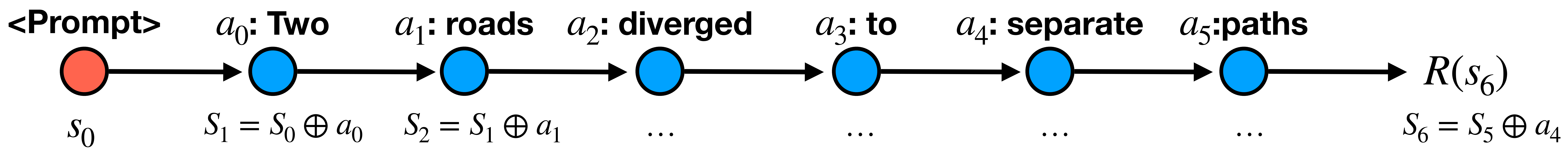


Initial State:	$s_0 \sim D$ a prompt
Actions:	words from our vocabulary
States:	$S_6 = S_0 \oplus < a_0, a_1, a_2, a_3, a_4, a_5 >$
Transition:	$P(s' \mid s, a)$
Reward:	$R(s)$ score
Horizon:	H max generation length
Policy:	$\pi(\cdot \mid s)$

Goal: Learn policy π to maximize reward

Reinforcement Learning

for text-generation

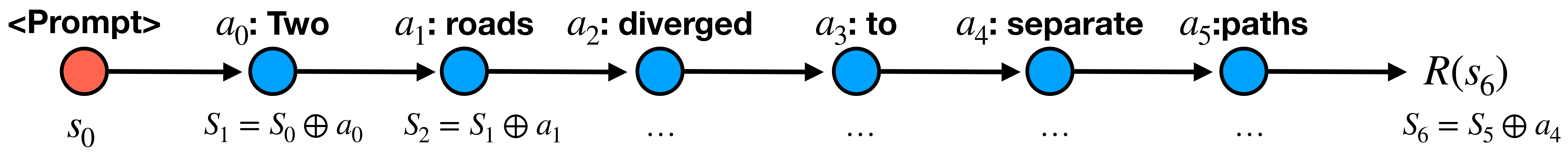


Initial State:	$s_0 \sim D$ a prompt
Actions:	words from our vocabulary
States:	$S_6 = S_0 \oplus < a_0, a_1, a_2, a_3, a_4, a_5 >$
Transition:	$P(s' \mid s, a)$
Reward:	$R(s)$ score **NLP Metric**
Horizon:	H max generation length
Policy:	$\pi(\cdot \mid s)$

Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization
[RA**B**HSBHC ICLR 2023]

The first benchmark comparing interactive fine-tuning with RL and supervised fine-tuning across a range of NLP tasks.

Reinforcement Learning
for text-generation



Initial State:	$s_0 \sim D$ a prompt
Actions:	words from our vocabulary
States:	$S_6 = S_0 \oplus < a_0, a_1, a_2, a_3, a_4, a_5 >$
Transition:	$P(s' \mid s, a)$
Unknown Reward:	$R(s)$ score
Horizon:	H max generation length
Policy:	$\pi(\cdot \mid s)$

Given Preference dataset from an expert $\pi^\star$

Goal: **Learn policy π using this dataset**

Reinforcement Learning From Human Feedback

1. Supervised Fine-Tuning (SFT)

$$\pi^{\mathrm{SFT}} = \arg\min_{\pi \in \Pi} \mathbb{E}_{(s,a)\sim\rho_{\pi^\star}} \left[L(\pi, s, a) \right]$$

3. Learn a preference reward function

for example using logistic regression

$$\hat{r} = \arg\min_{r} \sum_{D} \log \sigma\Big(r(\tau_1) - r(\tau_2) \Big)$$

4. Interactive Fine-Tuning with RL

$$\mathbb{E}_{\tau\sim\rho_\pi} \left[\hat{r}(\tau) \right] + \lambda D_{\mathrm{KL}}(\pi \,||\, \pi^{\mathrm{SFT}})$$

2. Collect preferences data

Prompt $s_0 \sim D$

Calculate the median number $[2, 3, 7, 8, 10]$

Responses from π^{SFT}

0 7

Human Labeler rank outputs

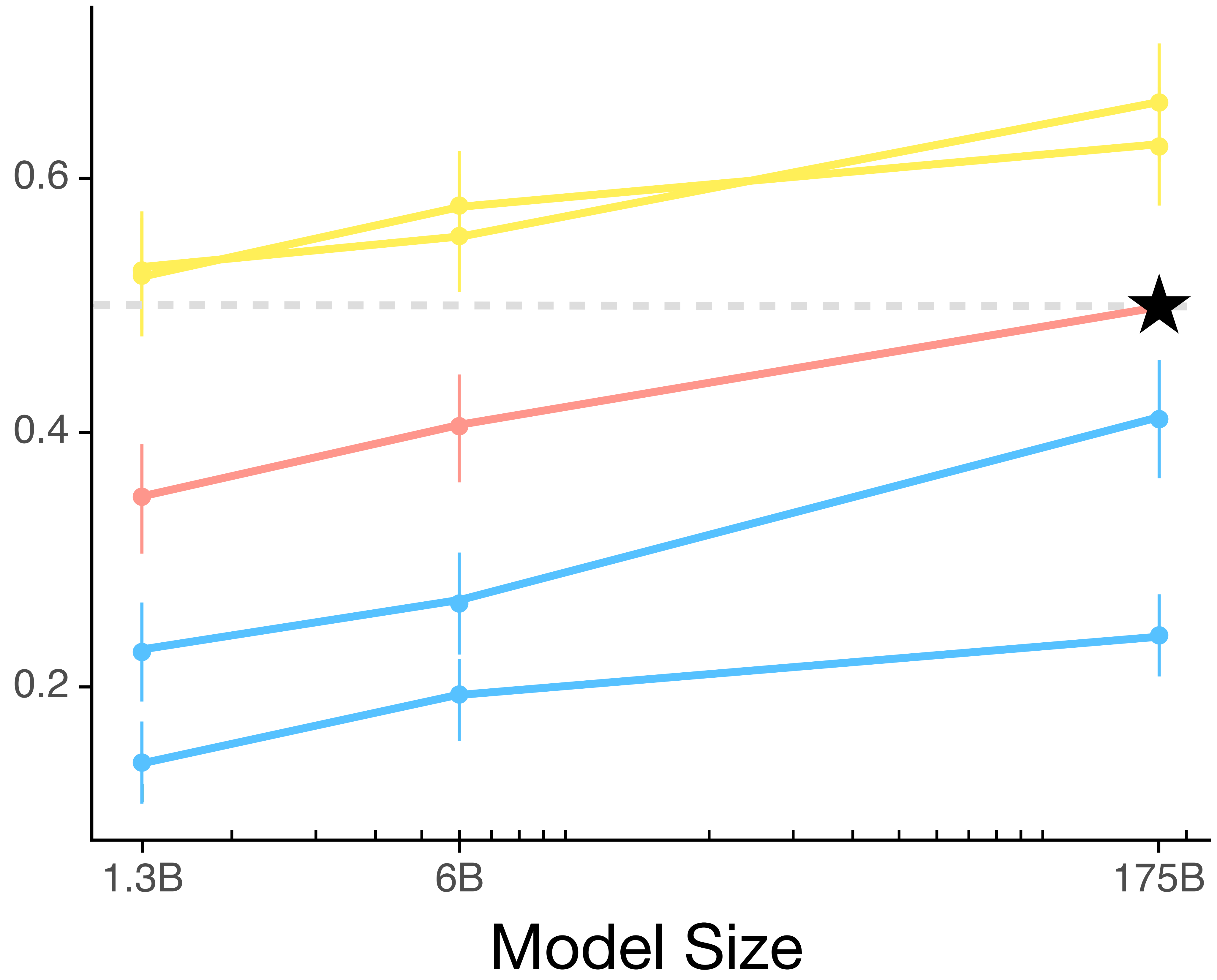
7 > 0

Aggregate $D = D \cup \{(\tau_1, \tau_2, s_0)\}$

Reinforcement Learning From Human Feedback

Pre-Training → Fine-Tuning → Interactive Fine-Tuning

Human Win rate against SFT 175B



Long Ouyang et al.
Training language models to follow instructions with human feedback
OpenAI 2022

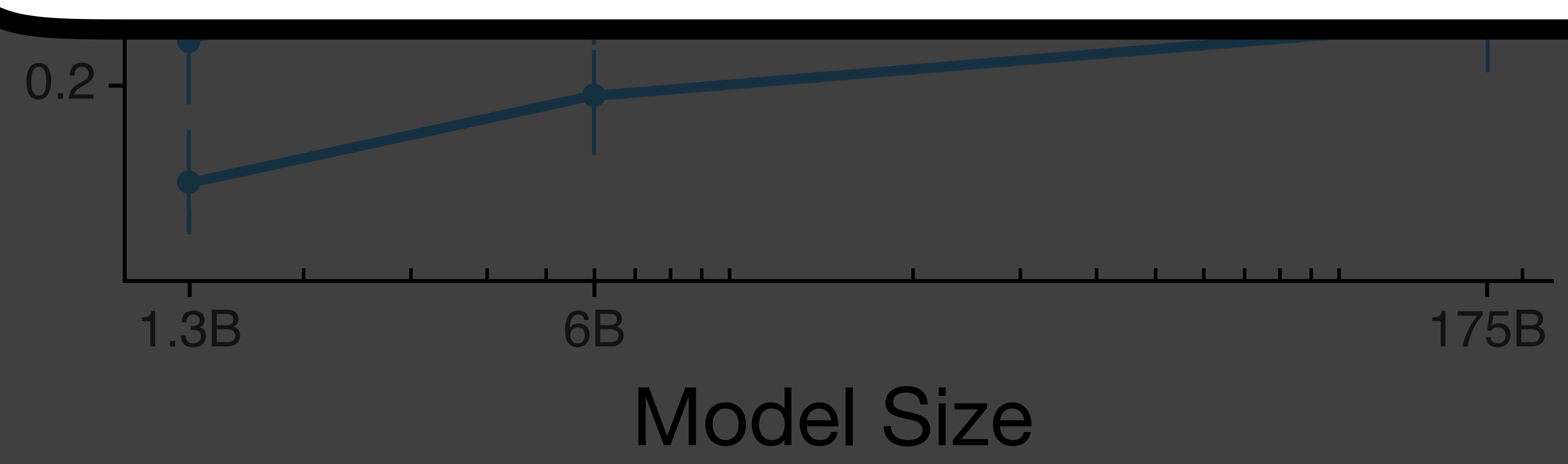
Reinforcement Learning From Human Feedback

Pre-Training

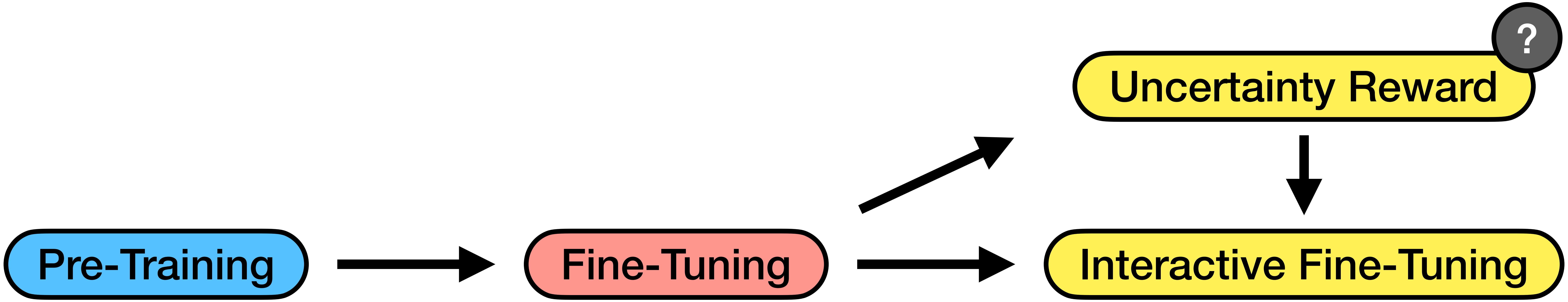
Fine-Tuning

Problem:

- **Text generation is a LARGE combinatorial search problem.**
- **RL algorithms are complex and unstable, making them challenging to tune effectively.**



Long Ouyang et al.
Training language models to follow instructions with human feedback
OpenAI 2022

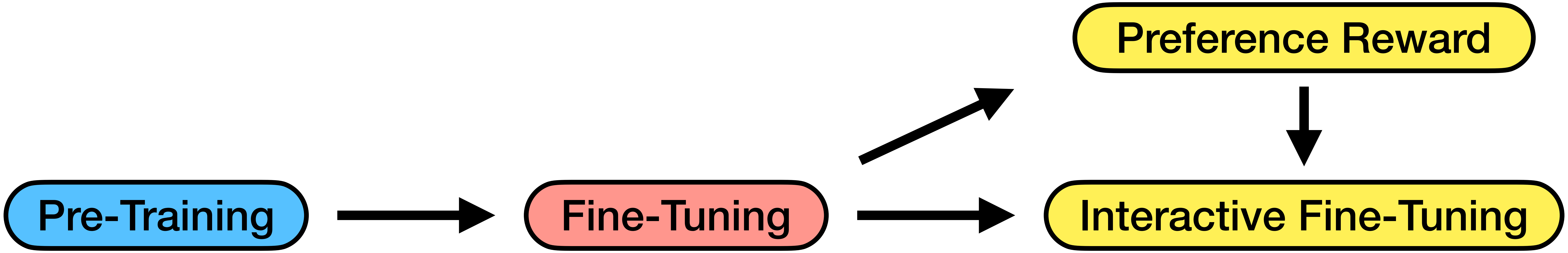


Formalize the imitation learning problem

How can interaction help?

How can we design interactive learning algorithms specific to LLMs?

objective alignment issue



Formalize the imitation learning problem

objective alignment issue

How can interaction help?

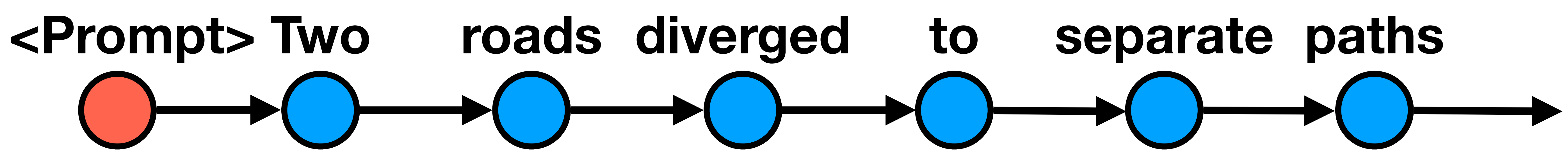
How can we design interactive learning algorithms specific to LLMs?

Reset Property

Reset allows us to rollout a policy from partial sentences

?

1. Sample a prompt from D



$\hat{r}(s_H) + \gamma V^{\pi}(s_{H+1})$

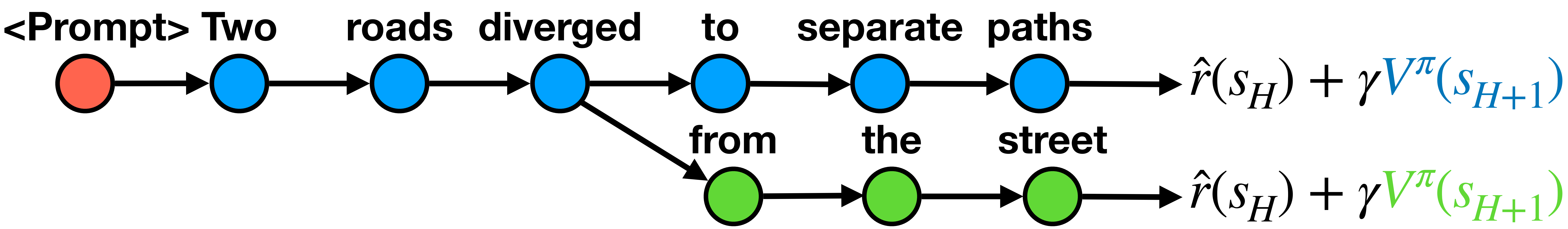
2. Sample a generation from π

Reset Property

Reset allows us to rollout a policy from partial sentences

Transition: $P(s' \mid s, a)$ Deterministic

1. Sample a prompt from D



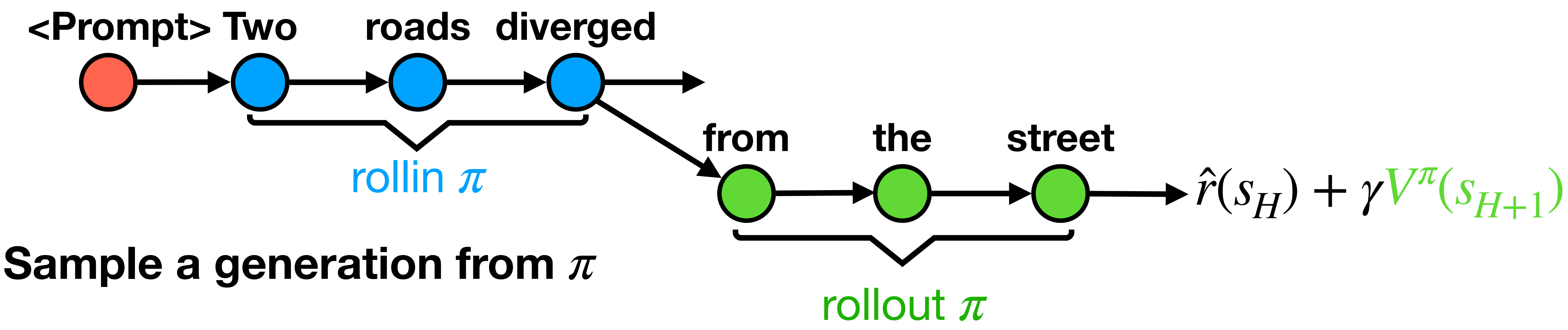
2. Sample a generation from π

Reset Property

Reset allows us to rollout a policy from partial sentences

Transition: $P(s' \mid s, a)$ Deterministic

1. Sample a prompt from D



2. Sample a generation from π

Reset Property

?

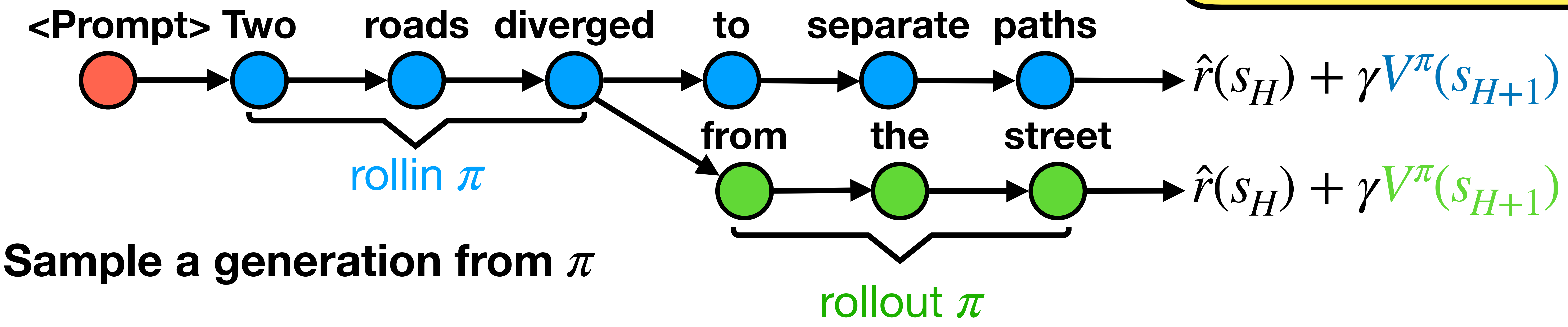
Reset allows us to rollout a policy from partial sentences

Transition: $P(s' \mid s, a)$ Deterministic

$$\mathbb{E}_{\tau \sim \rho_\pi}\left[\hat{r}(\tau)\right] + \underbrace{\lambda D_{\mathrm{KL}}(\pi \,||\, \pi^{\mathrm{SFT}})}$$

we do not want to move too far from π^{SFT}

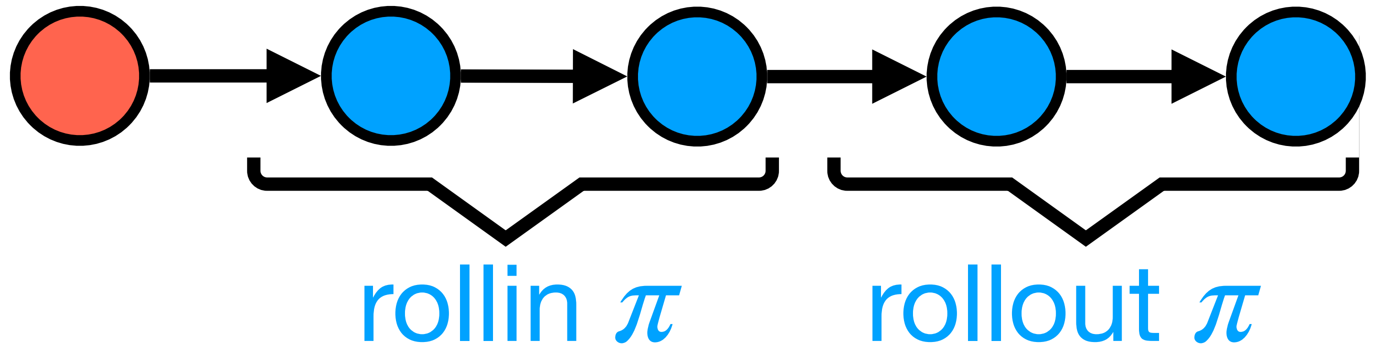
1. Sample a prompt from D



2. Sample a generation from π

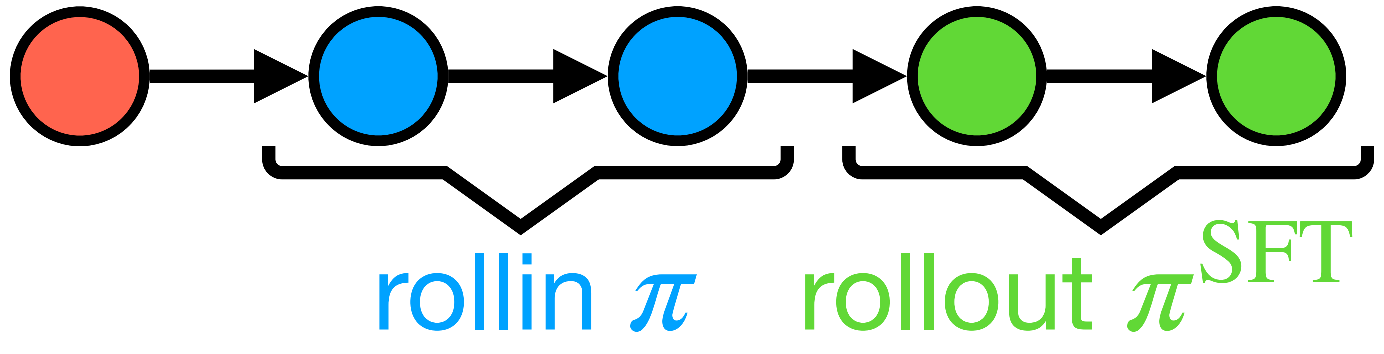
Approaches

PPO (RL algorithm)



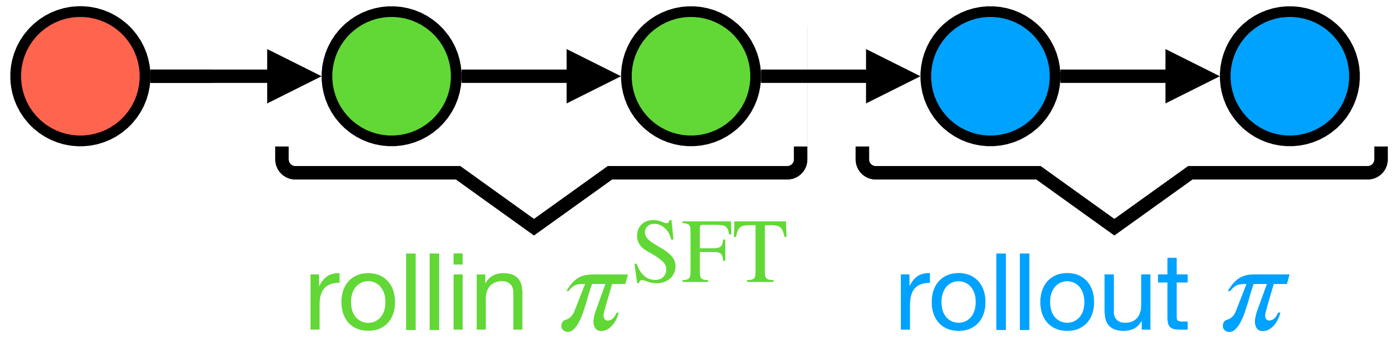
- Does not take advantage of problem specific structure
- Samples prompts $s_0 \sim D$
- Scores action with $\hat{r} + \gamma V^{\pi}$

AggreVaTeD



- Samples prompts $s_0 \sim D$
- **Scores action with $\hat{r} + \gamma V^{\pi^{\text{SFT}}}$**
- **Intuition: Receive feedback on partial generations**

PPO^{++} [CBRMS Instruction Workshop 2023]

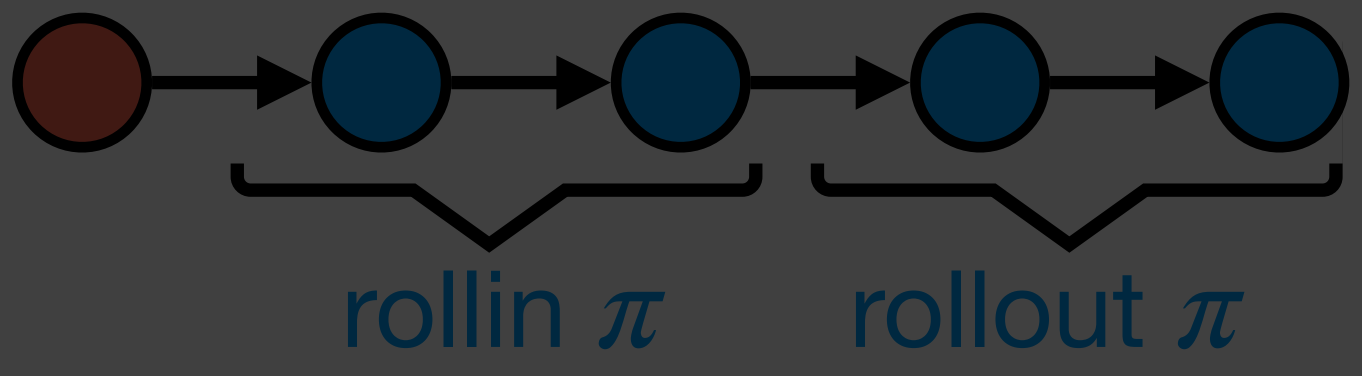


Proximal Policy Optimization, John Schulman et al., 2017

Deeply AggreVaTeD: Differentiable Imitation Learning for Sequential Prediction, Wen Sun et al. ICML 2017

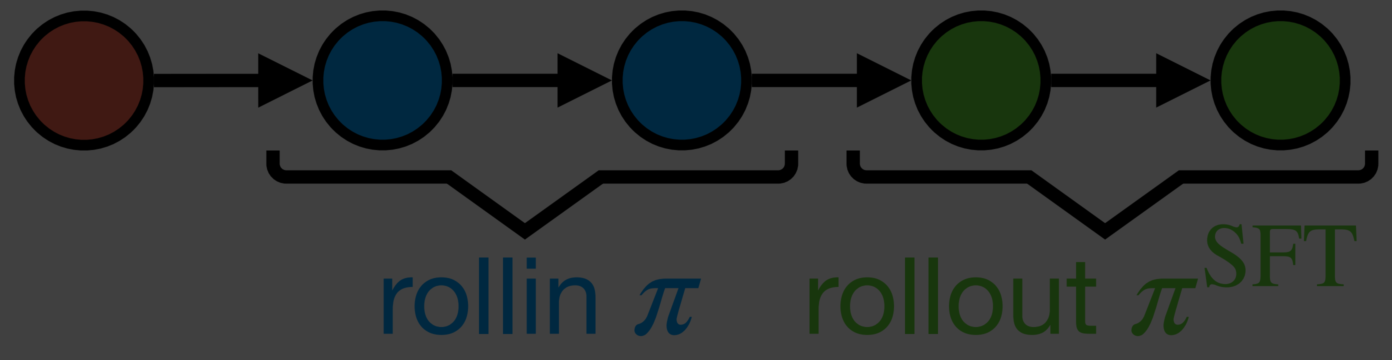
Approaches

PPO (RL algorithm)



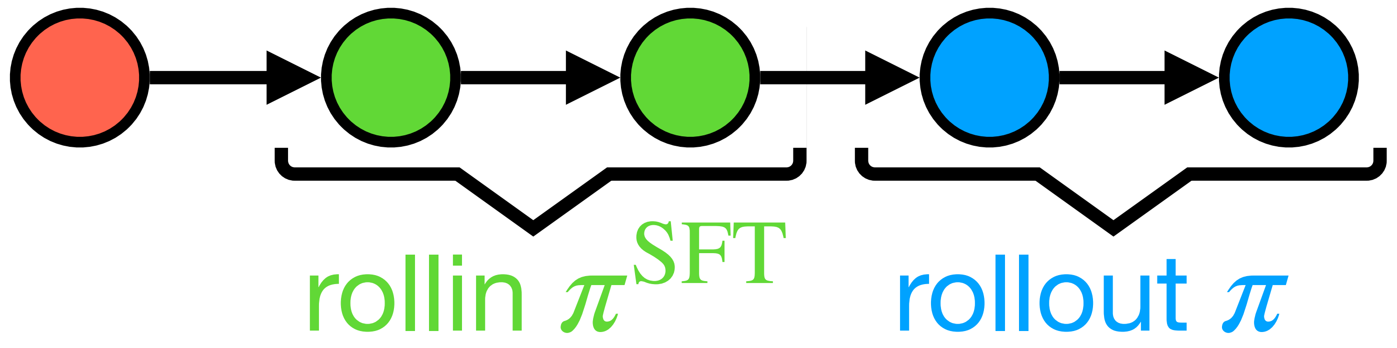
- Does not take advantage of problem specific structure
- Samples prompts $s_0 \sim D$
- Scores action with $\hat{r} + \gamma V^{\pi}$

AggreVaTeD



- Samples prompts $s_0 \sim D$
- **Scores action with** $\hat{r} + \gamma V^{\pi^{\text{SFT}}}$
- **Intuition: Receive feedback on partial generations**

PPO^{++} [CBRMS Instruction Workshop 2023]



- **Samples prompts** $s_0 \sim \beta \rho_{\pi^{\text{SFT}}} + (1 - \beta) D$
- Scores actions with $\hat{r} + \gamma V^{\pi}$
- **Intuition: Richer initial state distribution**

Proximal Policy Optimization, John Schulman et al., 2017
Deeply AggreVaTeD: Differentiable Imitation Learning for Sequential Prediction, Wen Sun et al. ICML 2017

Theory of PPO++

Let $\pi^\star$ be a high quality policy covered by π^{SFT}

$$\mathbb{E}_{s\sim\beta\rho^{\pi^{\text{SFT}}}+(1-\beta)D}\left[\max_a A^{\pi^t}(s,a)\right] \le \epsilon$$

Assume that one-step local improvement over π^t is small

$$\overbrace{J(\pi^\star) - J(\pi^t)}^{\text{Performance gap}} \le O\left(H^2 \max_s \underbrace{\left(\frac{d^{\pi^\star}(s)}{d^{\pi^{\text{SFT}}}(s)}\right)}_{\substack{\text{Assume bound density ratio and} \\ \pi^{\text{SFT}} \text{ provides coverage for } \pi^\star}} \epsilon\right)$$

Approximately Optimal Approximate Reinforcement Learning Kakade and Langford 2002

Experimental Setup

Task Statement

Given a reddit post, write a TL;DR (short summary).

Dataset Composition

- 210K Prompts total
- 117K Prompts with *Human Labels*
- 93K Prompts with *Human Preference Labels*

Learning to summarization from human feedback Stiennon et al. 2017

Experimental Results

Uncertainty of a model at predicting
the ground truth summaries

Algorithms	Perplexity (↓)	GPT4 Win Rate (↑)
`SFT`	14.09	29.5%
`SFT+PPO`	14.87	60.7%
`SFT+PPO`$^{++}$	**13.42**	**64.4%**
`SFT+AggreVaTeD`	13.53	54.12%

Experimental Results

Rafailov et al. 2023 conducted human study showing alignment between GPT4 and Human preference

Algorithms	Perplexity (↓)	GPT4 Win Rate (↑)
SFT	14.09	29.5%
SFT+PPO	14.87	60.7%
SFT+PPO^{++}	**13.42**	**64.4%**
SFT+AggreVaTeD	13.53	54.12%

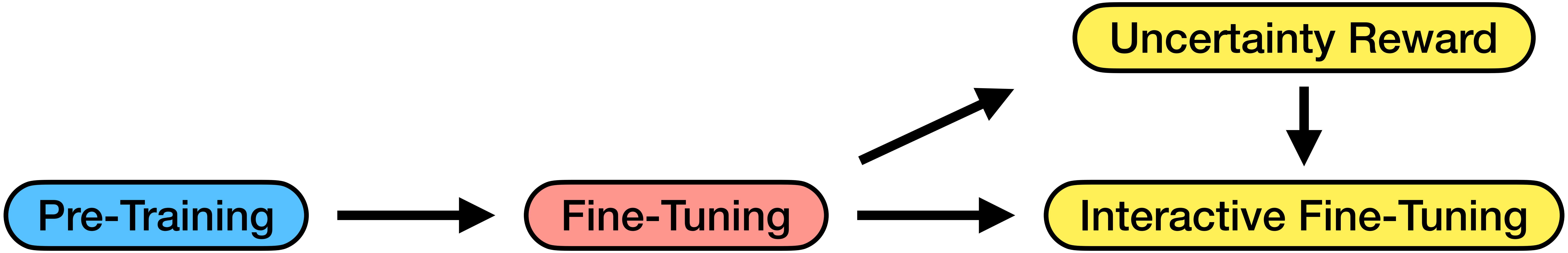
Winrate Prompt Template:

Which of the following summaries does a better job of summarizing the most important points in the given forum Post? FIRST provide a one-sentence comparison of the two summaries, explaining which you prefer and why. SECOND, on a new line, state only "A" or "B" to indicate your choice.

Post: <Post>
A: <TLDR A>
B: <TLDR B>

Summary

- Designing RL algorithms specific for LLMs can improve performance
- Resetting is a special property of MDPs for LLMs
- PPO++ simple algorithm that use reset

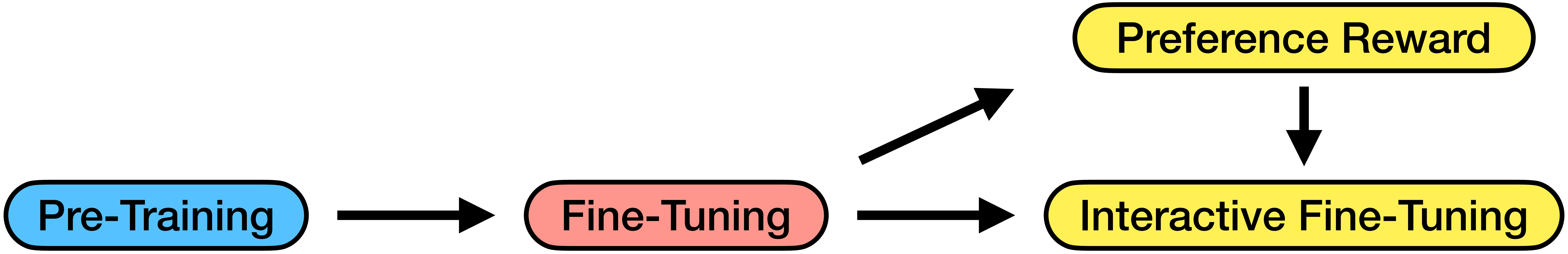


Formalize the imitation learning problem

How can interaction help?

How can we design interactive learning algorithms specific to LLMs?

objective alignment issue



Formalize the imitation learning problem

How can interaction help?

How can we design interactive learning algorithms specific to LLMs?

objective alignment issue

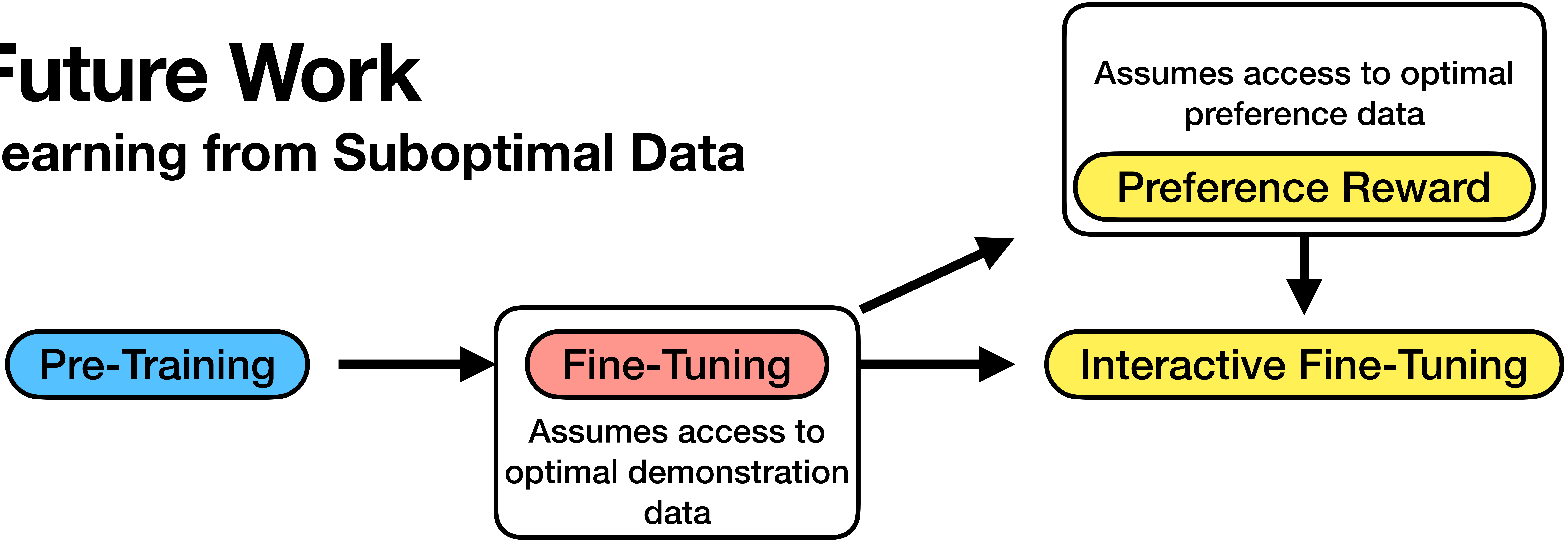
Future Work

- Imitation Learning [1, 2]
- Imitation Learning with a Computational Oracle [6, 7, 8, 12]
- Imitation Learning with Suboptimal Data

1. Disagreement-Regularized Imitation Learning…………………………………………………… [**B**SH ICLR 2020]
2. Adversarial Imitation Learning via Boosting………………………………………………… [CSH**B**S ICLR 2024]
3. Is reinforcement learning (not) for natural language processing?…………………………… [RA**B**HSBHC ICLR 2023]
4. Learning to Generate Better Than Your LLM………………………………………………… [C**B**RMS Instruction Workshop 2023]
5. Policy-Gradient Training of Language Models for Ranking…………………………………… [GCC**B**J FMDM Workshop 2023]
6. Non-Monotonic Text Generation……………………………………………………………… [W**B**DC ICLR 2019]
7. Active Imitation Learning with Noisy Guidance……………………………………………… [**B**SD ACL 2020]
8. Interactive Text Generation…………………………………………………………………… [FGP**B**CZGD EMNLP 2024]
9. Reinforcement Learning with Convex Constraints…………………………………………… [M**B**DDS NeurIPS 2019]
10. Constrained episodic reinforcement learning in concave-convex and knapsack settings………… [**B**DLMSSS NeurIPS 2020]
11. Ranking with Long-Term Constraints………………………………………………………… [**B**FDJ WSDM 2024]
12. lilGym: Natural Language Visual Reasoning with Reinforcement Learning……………………… [W**B**KA ACL 2022]

Future Work

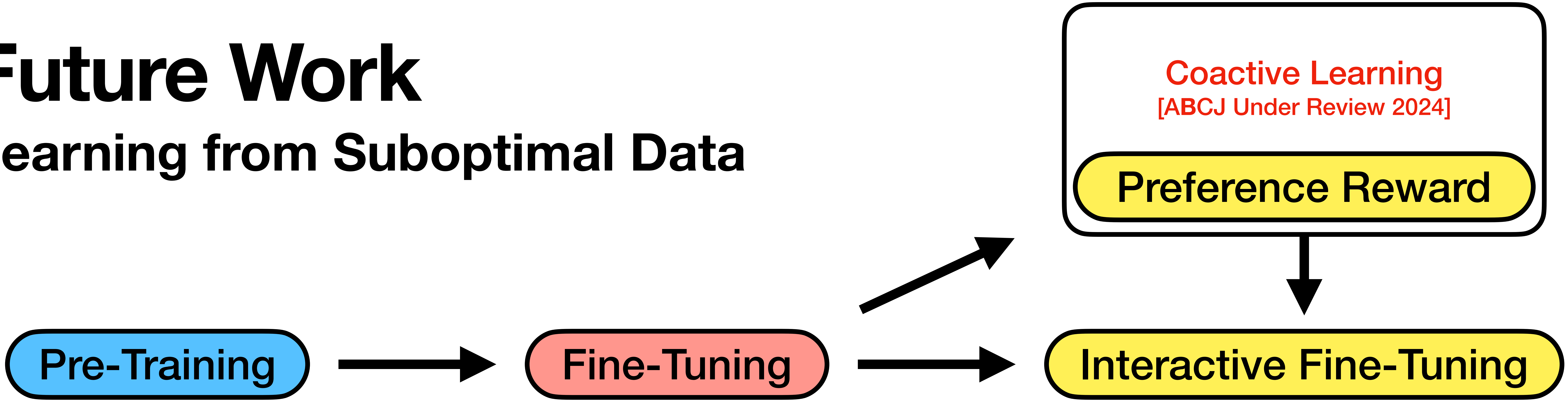
Learning from Suboptimal Data



- Develop **fine-tuning algorithms** that can learn from suboptimal data.
- Develop **reward functions** that can learn from suboptimal preferences.

Future Work

Learning from Suboptimal Data



- Develop **fine-tuning algorithms** that can learn from suboptimal data.
- Develop **reward functions** that can learn from suboptimal preferences.

Coactive Learning for Training Large Language Models using Implicit User Feedback………………………... [A**B**CJ Under Review 2024]

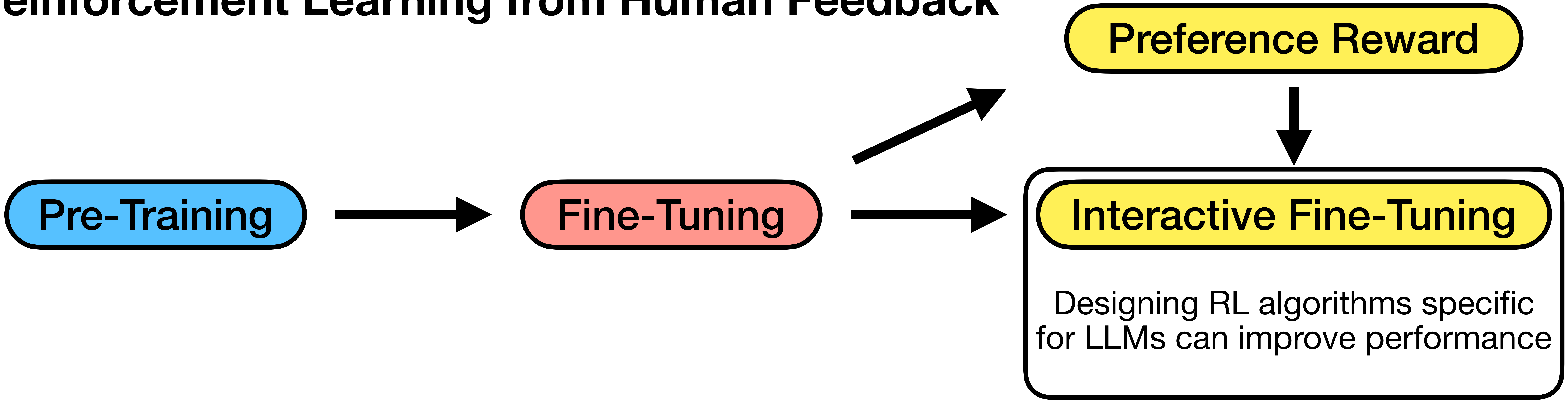
Future Work

- Reinforcement Learning from Human Feedback [3, 4]
- Constrained Reinforcement Learning [9, 10, 11]

1. Disagreement-Regularized Imitation Learning…………………………………………………………….. [**B**SH ICLR 2020]
2. Adversarial Imitation Learning via Boosting…………………………………………………………….. [CSH**B**S ICLR 2024]
3. Is reinforcement learning (not) for natural language processing?……………………………………….. [RA**B**HSBHC ICLR 2023]
4. Learning to Generate Better Than Your LLM…………………………………………………………… [C**B**RMS Instruction Workshop 2023]
5. Policy-Gradient Training of Language Models for Ranking……………………………………………… [GCC**B**J FMDM Workshop 2023]
6. Non-Monotonic Text Generation……………………………………………………………………….. [W**B**DC ICLR 2019]
7. Active Imitation Learning with Noisy Guidance…………………………………………………………… [**B**SD ACL 2020]
8. Interactive Text Generation……………………………………………………………………………. [FGP**B**CZGD EMNLP 2024]
9. Reinforcement Learning with Convex Constraints………………………………………………………… [M**B**DDS NeurIPS 2019]
10. Constrained episodic reinforcement learning in concave-convex and knapsack settings………… [**B**DLMSSS NeurIPS 2020]
11. Ranking with Long-Term Constraints……………………………………………………………………. [**B**FDJ WSDM 2024]
12. lilGym: Natural Language Visual Reasoning with Reinforcement Learning………………………….. [W**B**KA ACL 2022]

Future Work

Reinforcement Learning from Human Feedback



- Continue to **explore RL algorithms** specialized for NLP systems.
  - Fast Simulation
  - Deterministic Environment
  - Demonstration Data
  - Trajectory-Level Rewards

[CZO**B**MLS Under Review 2024]

Future Work

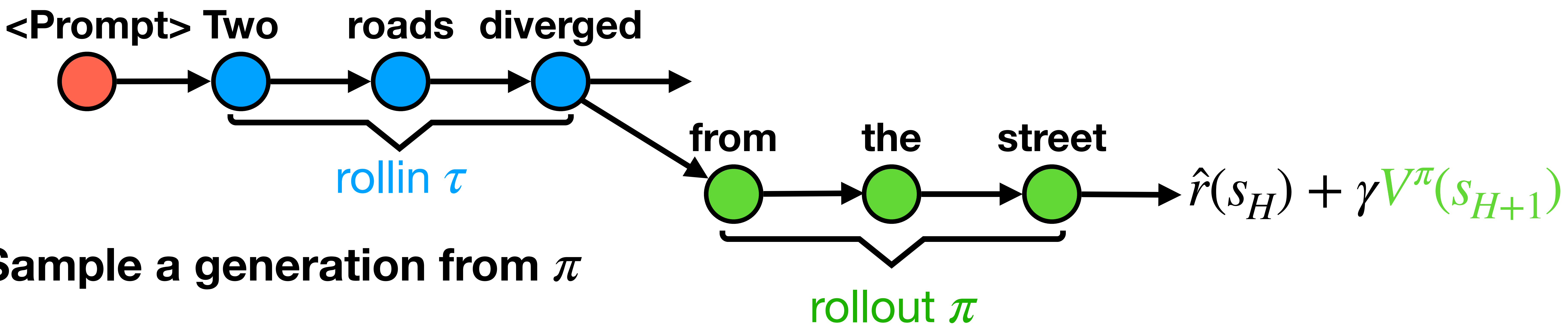
Reinforcement Learning from Human Feedback

Pre-Training → Fine-Tuning → Interactive Fine-Tuning

Fine-Tuning → Preference Reward → Interactive Fine-Tuning

Dataset Reset Policy Optimization
[CZOBMLS Under Review 2024]

1. Sample a prompt and response τ from D



2. Sample a generation from π

Future Work

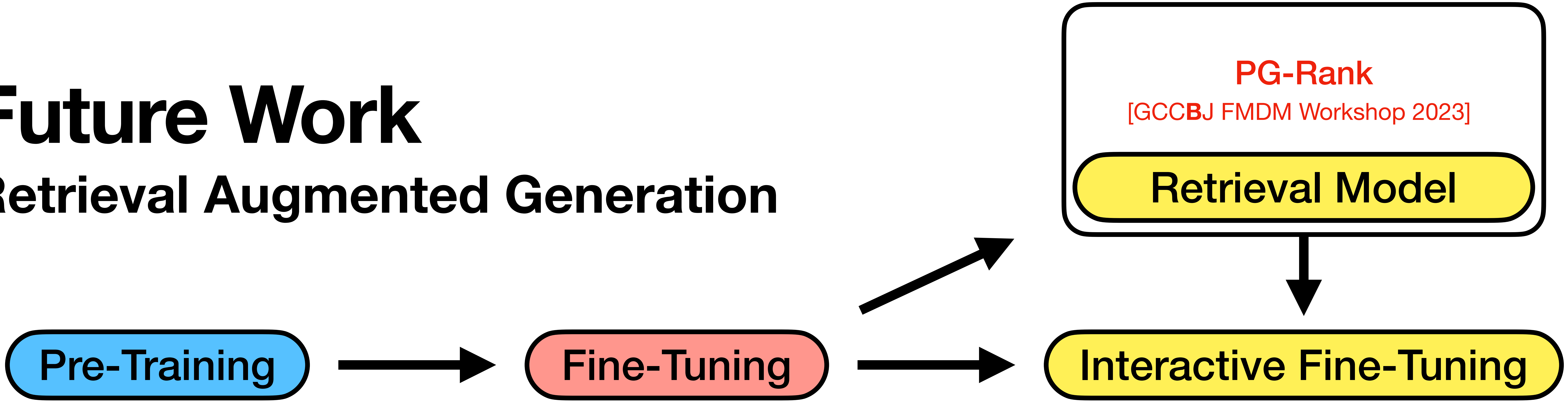
- Natural Language Processing [3, 4, 6, 7, 8, 12]
- Information Retrieval [5, 11]
- Classic Control [2, 9, 10]
- Video Games [1]
- ?

- Imitation Learning [1, 2, 6, 7, 8, 12]
- Reinforcement Learning [3, 4, 5, 9, 10, 11]

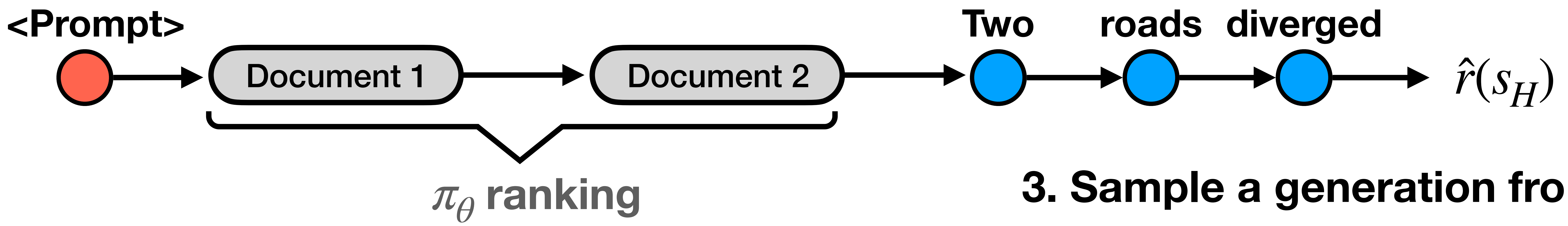
1. Disagreement-Regularized Imitation Learning………………………………………………………….. [**B**SH ICLR 2020]
2. Adversarial Imitation Learning via Boosting………………………………………………………….. [CSH**B**S ICLR 2024]
3. Is reinforcement learning (not) for natural language processing?…………………………………….. [RA**B**HSBHC ICLR 2023]
4. Learning to Generate Better Than Your LLM………………………………………………………… [C**B**RMS Instruction Workshop 2023]
5. Policy-Gradient Training of Language Models for Ranking…………………………………………… [GCC**B**J FMDM Workshop 2023]
6. Non-Monotonic Text Generation………………………………………………………………………. [W**B**DC ICLR 2019]
7. Active Imitation Learning with Noisy Guidance………………………………………………………. [**B**SD ACL 2020]
8. Interactive Text Generation……………………………………………………………………………. [FGP**B**CZGD EMNLP 2024]
9. Reinforcement Learning with Convex Constraints……………………………………………………. [M**B**DDS NeurIPS 2019]
10. Constrained episodic reinforcement learning in concave-convex and knapsack settings………… [**B**DLMSSS NeurIPS 2020]
11. Ranking with Long-Term Constraints………………………………………………………………….. [**B**FDJ WSDM 2024]
12. Successor feature sets: Generalizing successor representations across policies………………….. [**B**MG AAAI 2021]

Future Work

Retrieval Augmented Generation



1. Sample a prompt from D

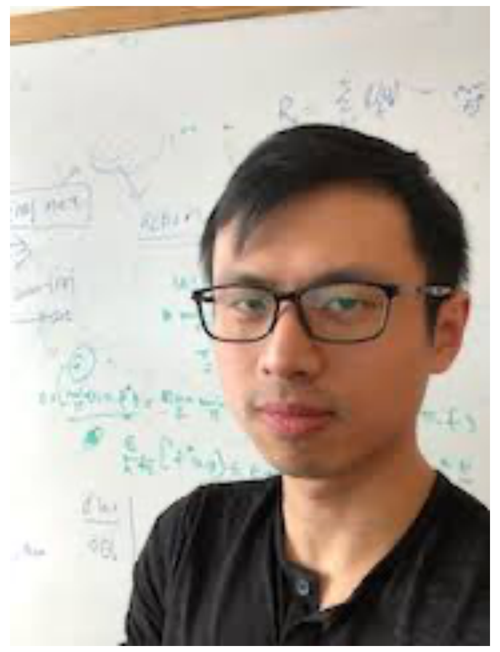


3. Sample a generation from π

2. Query a retrieval model

Policy-Gradient Training of Language Models for Ranking.. [GCCBJ FMDM Workshop 2023]

Thank You!


Wen Sun


Mikael Henaff


Rajkumar Ramamurthy


Dipendra Misra


Jonathan D. Chang


Wenhao Zhan


Thorsten Joachims


Aaron Tucker


Adam Cahall


Yoav Artzi


Anne Wu


Jason Lee

Questions?