ORL-Auditor: Dataset Auditing in Offline Deep Reinforcement Learning

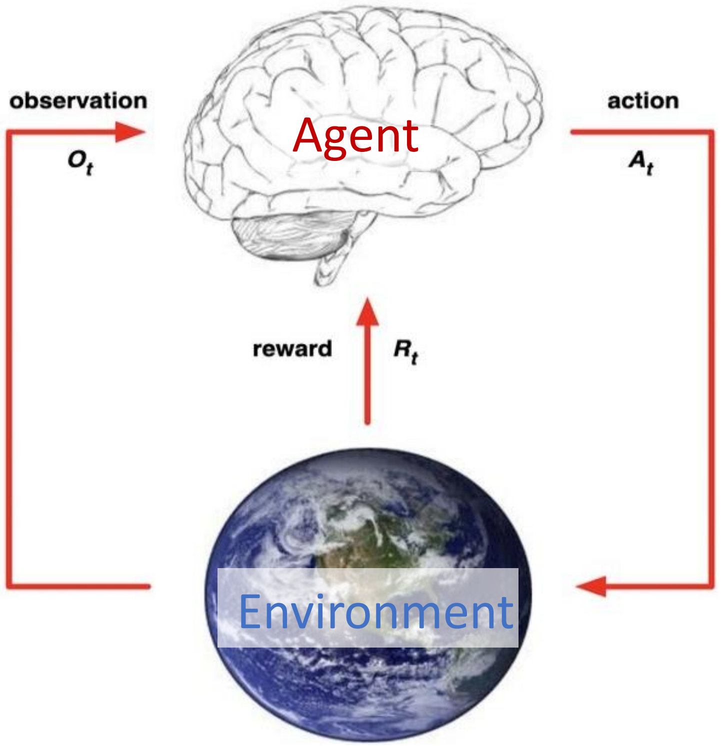
Linkang Du, Min Chen, Mingyang Sun, Shouling Ji,
Peng Cheng, Jiming Chen, and Zhikun Zhang

NDSS 2024

1. Background

Introduction of deep reinforcement learning (DRL)

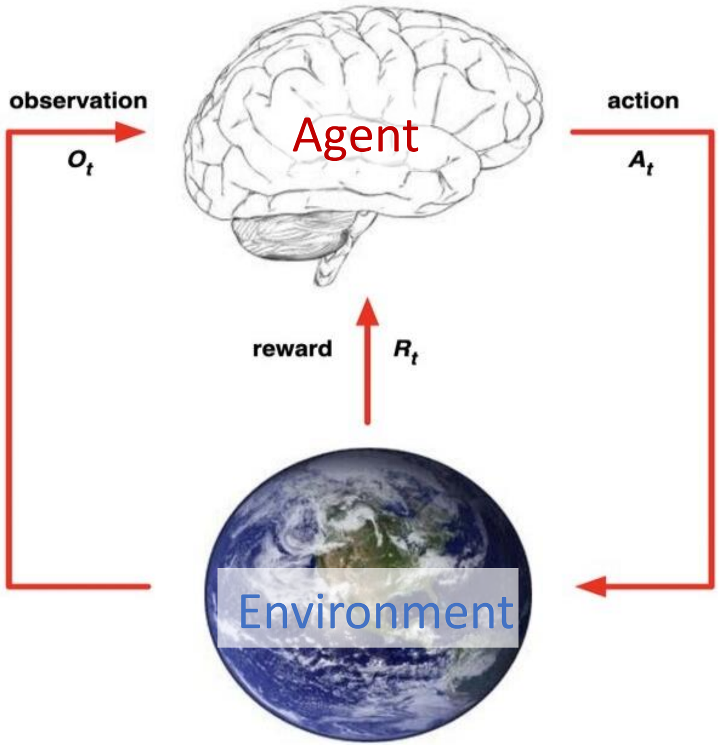
- Individuals gradually form expectations for stimuli in response to rewards or punishments provided by the environment (Reward r), resulting in habitual behaviors that yield maximum benefits (Actions a)



1. Background

Introduction of deep reinforcement learning (DRL)

- Individuals gradually form expectations for stimuli in response to rewards or punishments provided by the environment (Reward r), resulting in habitual behaviors that yield maximum benefits (Actions a)



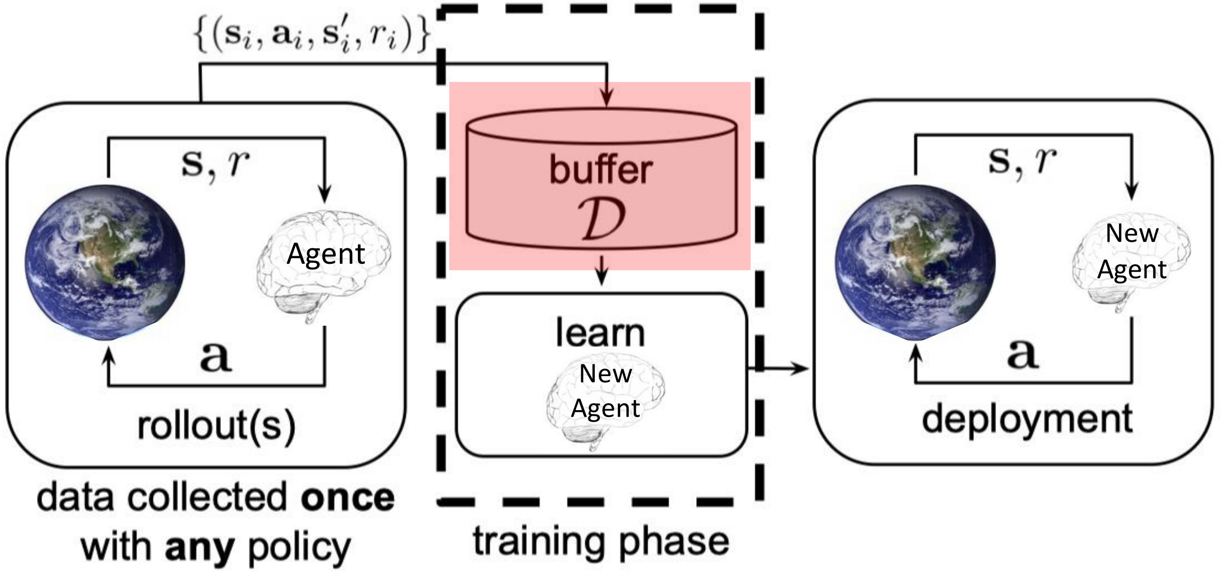
At t-th time step:

Agent
- ✓ Input observation o_t
- ✓ Input reward r_t
- ✓ Output action a_t

Environment
- ✓ Input action a_t
- ✓ Output observation o_{t+1}
- ✓ Output reward r_{t+1}

1. Background

The benefits of offline data



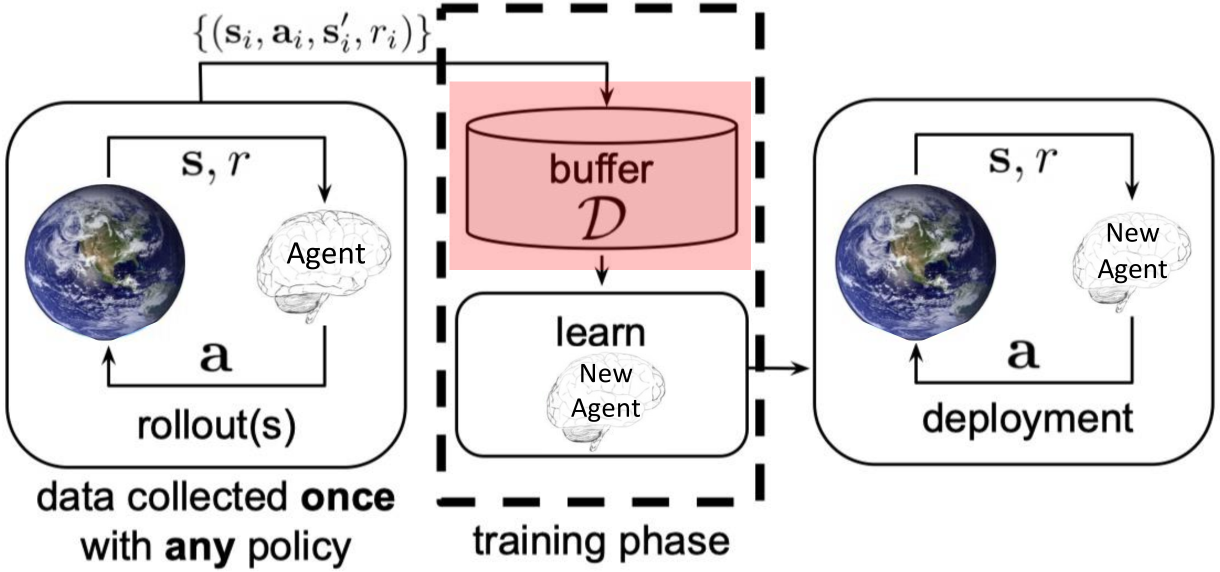
Offline Reinforcement Learning

1. Less computing resource

2. No damage to device

3. Full use of the history records

Possible misuse of offline data



Offline Reinforcement Learning

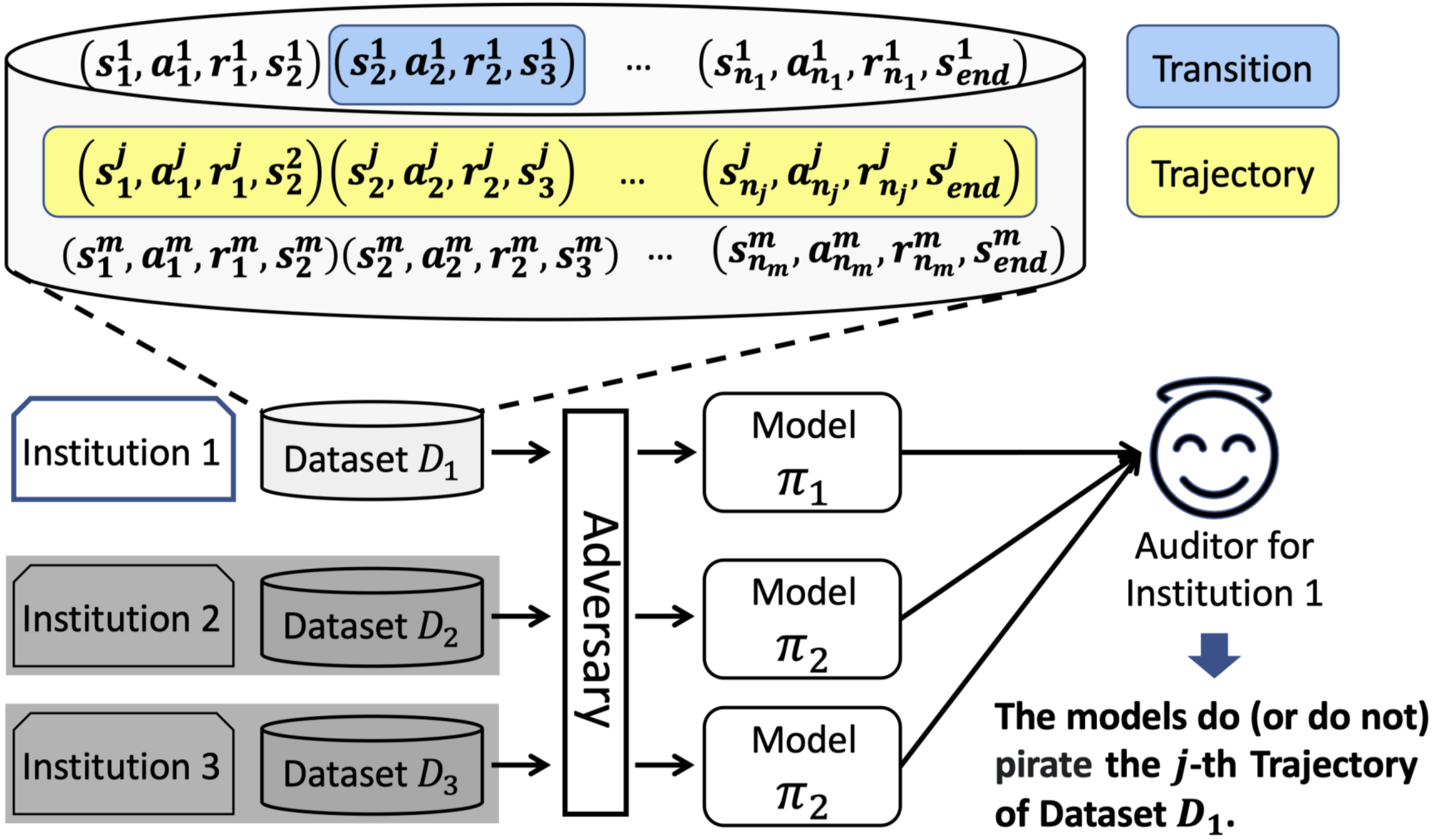
1. Data Traceability

2. Profit from data theft

2. Problem Statement

Dataset copyright auditing for offline DRL

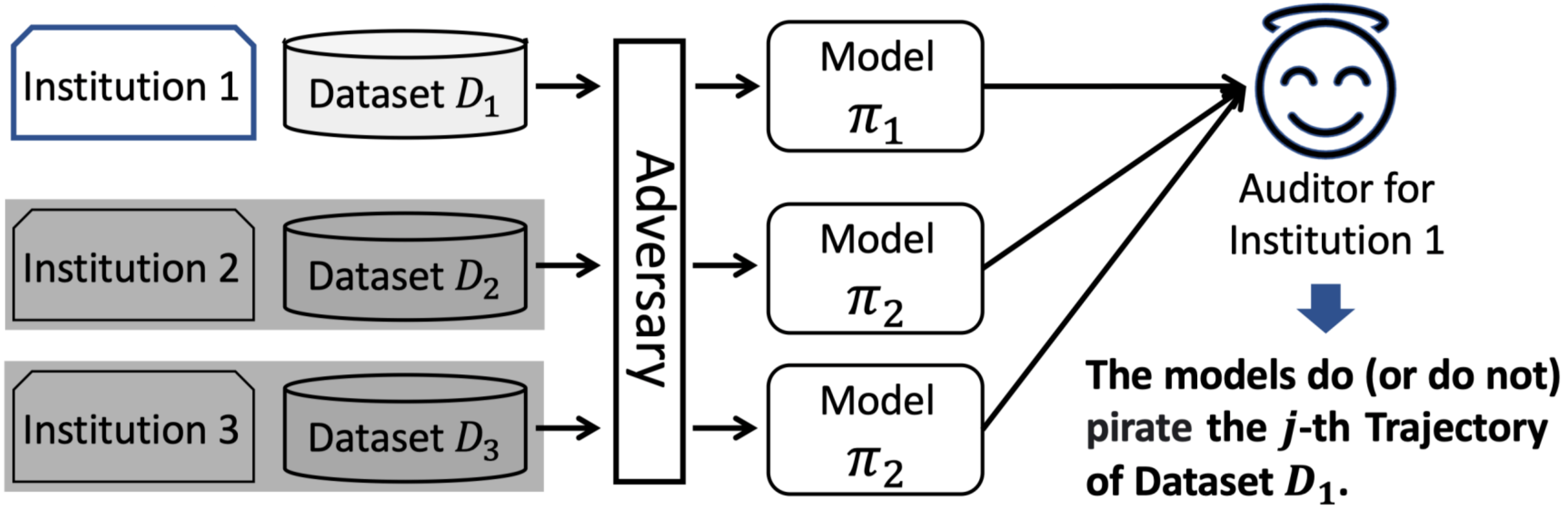
- Check whether the suspect model uses Dataset D_1 in its training process



2. Problem Statement

Assumptions about auditor

- Know the details of the dataset to be audited (the target dataset)
- Without any auxiliary dataset
- Black-box access to the suspect model



3. Related Work and Limitations

Watermarking [NeurIPS '20, NeurIPS '22]

Inject samples from a specific distribution prior to publishing the dataset

- ➢Can not handle datasets that have already been published
- ➢Infeasible to be altered afterward

Dataset (Membership) inference [ICLR '21, NeurIPS '20, NeurIPS '22]

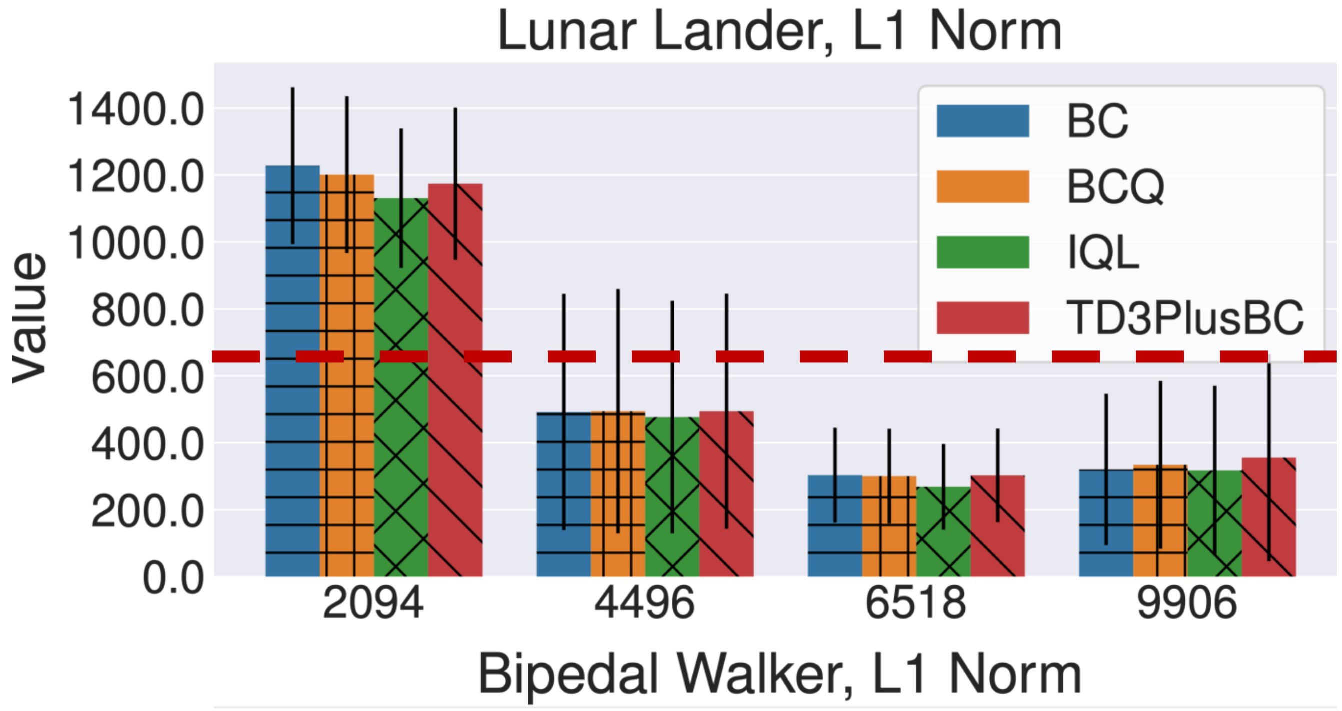
The models' decision boundaries or the behavioral difference between the surrogate models and the models trained on the target dataset

- ➢Difficult to determine suitable auxiliary dataset to train the surrogate model
- ➢Hard to obtain the decision boundaries when outputs are continuous

3. Related Work and Limitations

Dataset (Membership) inference [ICLR '21, NeurIPS '20, NeurIPS '22]

- ➢Difficult to determine suitable auxiliary dataset to train the surrogate model

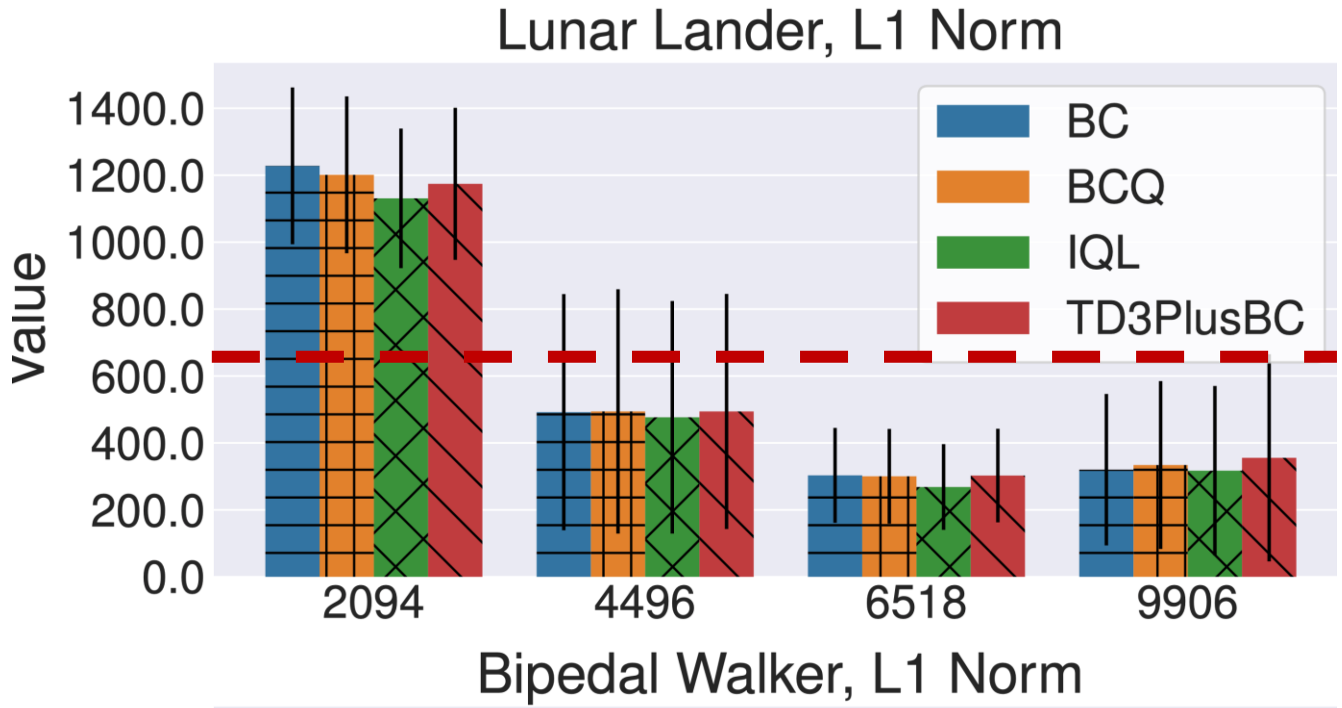


The DRL model was trained by **Dataset "0841"**

3. Related Work and Limitations

Dataset (Membership) inference [ICLR '21, NeurIPS '20, NeurIPS '22]

- ➢Difficult to determine suitable auxiliary dataset to train the surrogate model
- ➢Hard to define the decision boundaries when outputs are continuous



The DRL model was trained by **Dataset "0841"**

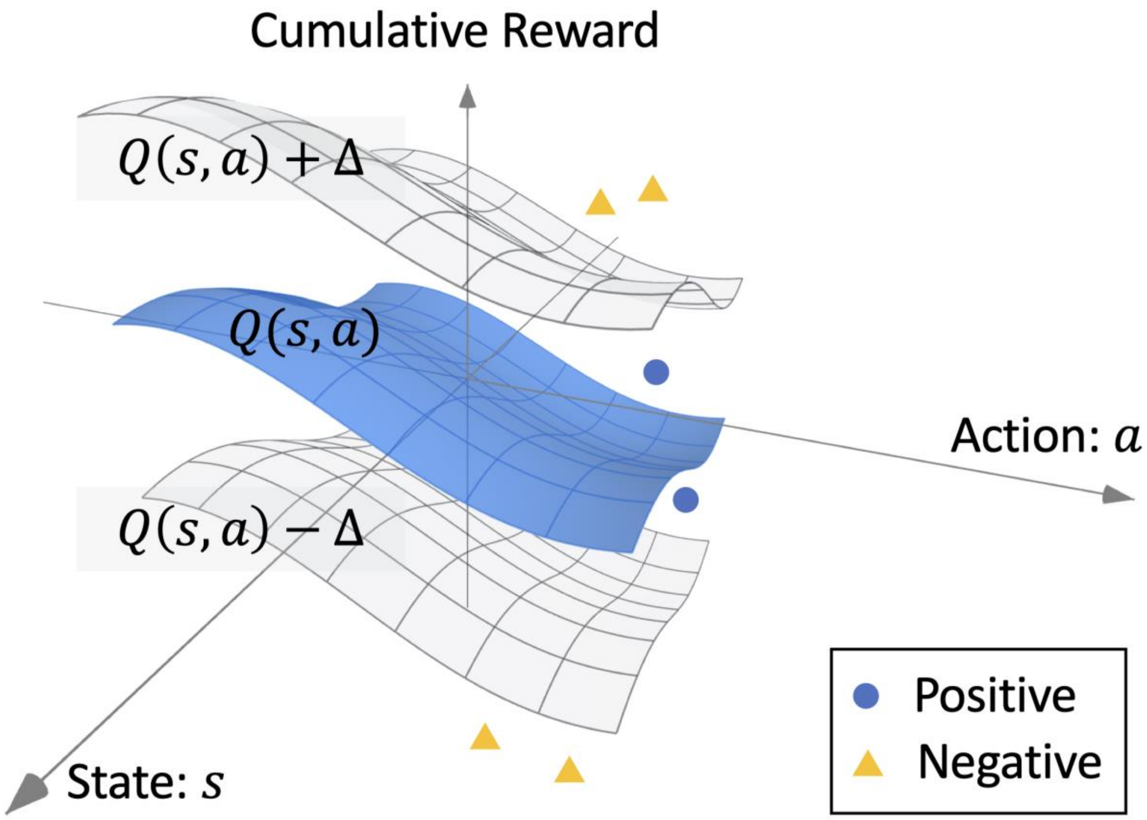
Task Name	Offline Model	Accuracy Training	Accuracy Test
Lunar Lander	BC	50.09 ± 0.68	48.41 ± 1.87
	BCQ	49.84 ± 1.39	47.69 ± 1.45
	IQL	49.88 ± 0.76	47.34 ± 1.83
	TD3PlusBC	50.08 ± 0.92	48.27 ± 1.81
Bipedal Walker	BC	50.00 ± 0.63	46.27 ± 2.42
	BCQ	49.97 ± 0.69	47.38 ± 2.41
	IQL	50.17 ± 0.95	47.19 ± 1.90
	TD3PlusBC	49.87 ± 0.94	45.48 ± 1.46
Ant	BC	50.44 ± 0.64	46.74 ± 2.37
	BCQ	50.22 ± 0.52	45.38 ± 2.16
	IQL	50.33 ± 0.35	45.89 ± 1.90
	TD3PlusBC	50.13 ± 0.67	45.03 ± 1.55

The performance of existing **MIA**

4. Our Proposal

Intuitive explanation of ORL-Auditor

- The middle surface is **the cumulative rewards of the state-action pairs from a dataset**. The auditor outputs a positive result if the cumulative rewards of a suspect model's state-action pairs are between the two outer surfaces



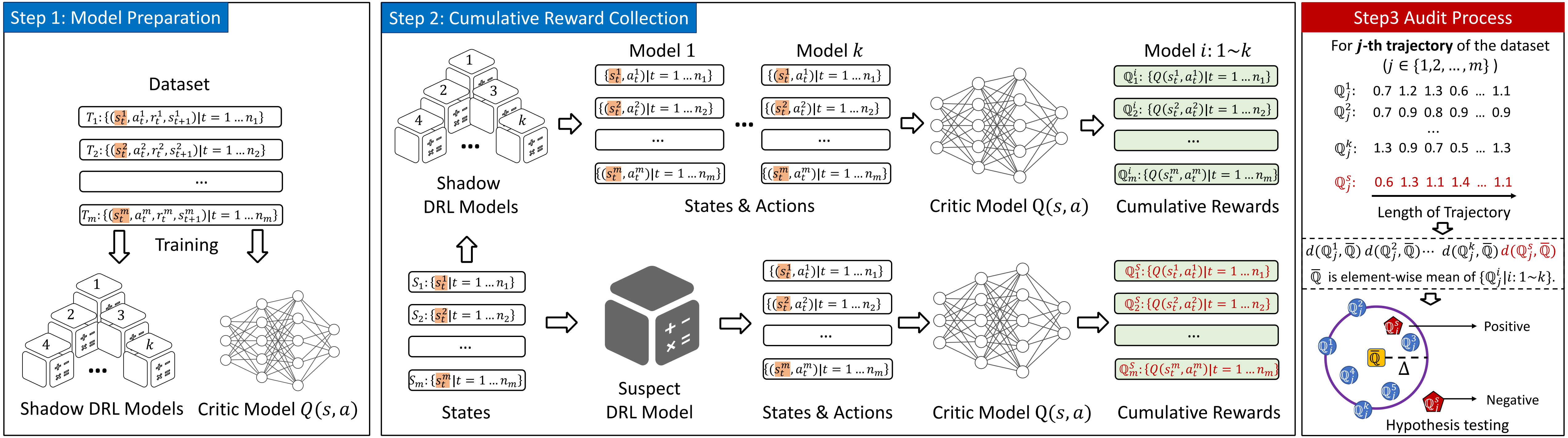
4. Our Proposal

Workflow of ORL-Auditor

- Step 1: Model Preparation
- Step 2: Cumulative Reward Collection
- Step 3: Auditing Process

✓ Auditing Basis $Q(s,a)$

✓ Auditing Boundary Δ

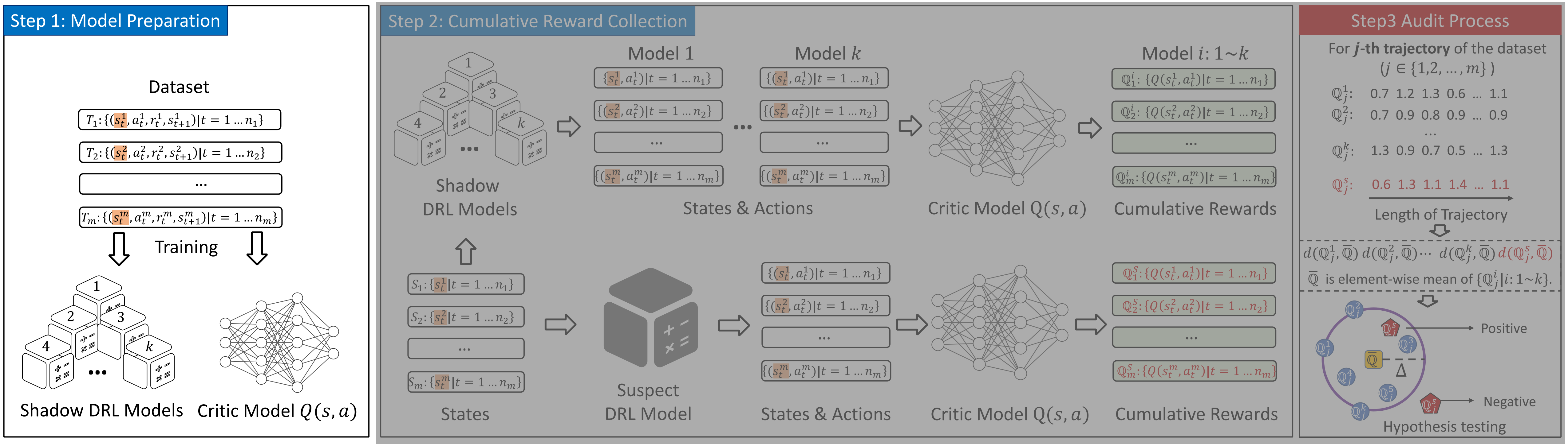


Workflow of ORL-Auditor

- Step 1: Model Preparation
- Step 2: Cumulative Reward Collection
- Step 3: Auditing Process

✓ Auditing Basis $Q(s, a)$

✓ Auditing Boundary Δ



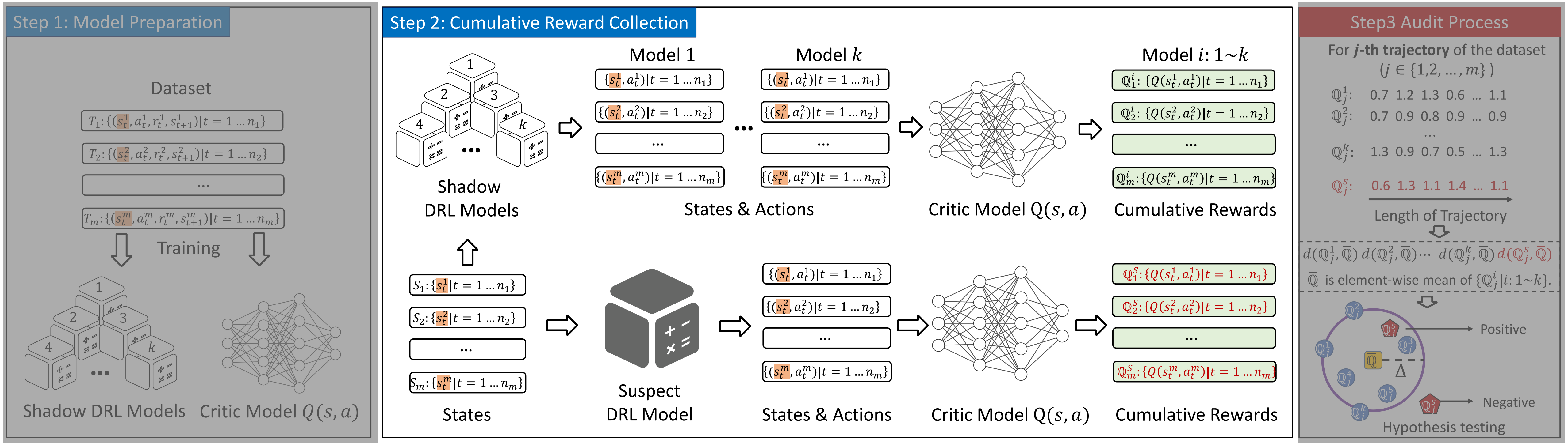
Train the shadow DRL models (Δ) and the critic model based on the target dataset (Q)

4. Our Proposal

Workflow of ORL-Auditor

- ➢ Step 1: Model Preparation
- ➢ Step 2: Cumulative Reward Collection
- ➢ Step 3: Auditing Process

✓ Auditing Basis $Q(s,a)$
✓ Auditing Boundary Δ



Collect the predicted cumulative rewards for the state-action pairs of the shadow models (Δ) and the suspect model

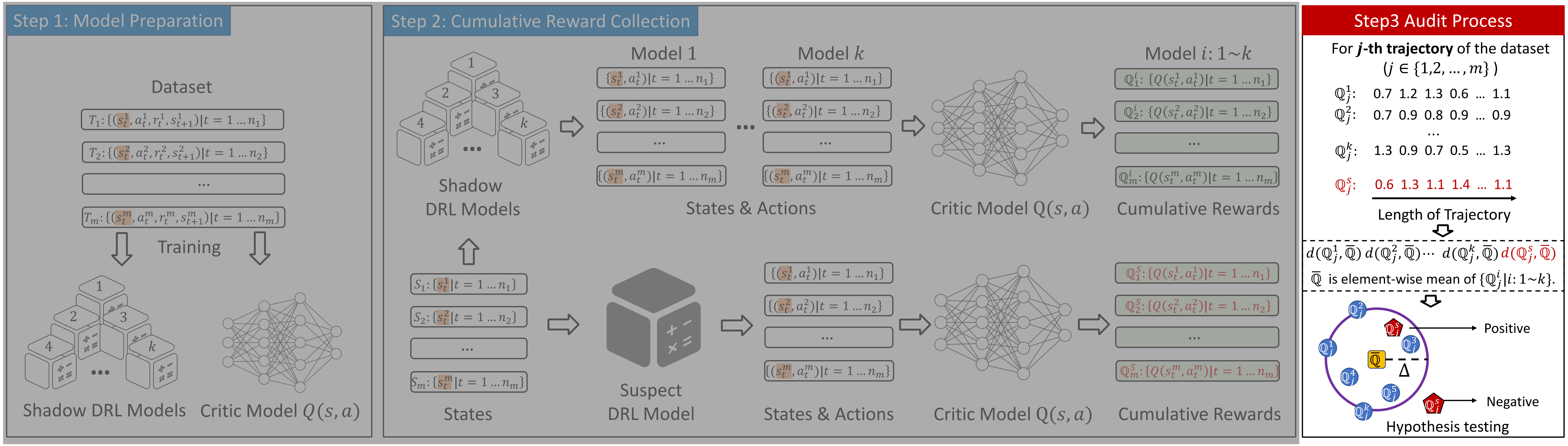
4. Our Proposal

Workflow of ORL-Auditor

- ➢ Step 1: Model Preparation
- ➢ Step 2: Cumulative Reward Collection
- ➢ Step 3: Auditing Process

✓ Auditing Basis $Q(s,a)$

✓ Auditing Boundary Δ

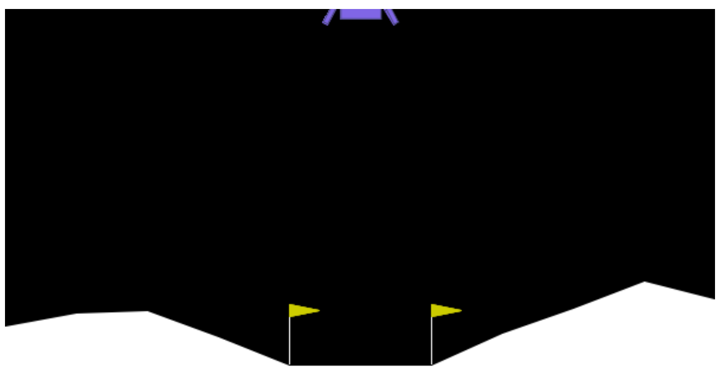


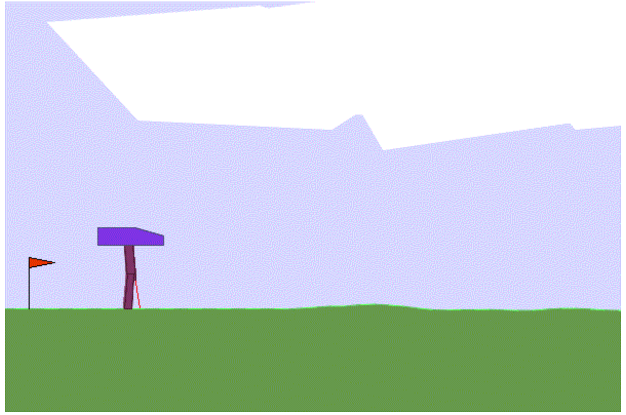
Utilize **the element-wise mean of the predicted cumulative rewards** of the shadow models as the used auditing basis (Instead of the $Q(s,a)$ directly from the critic model)

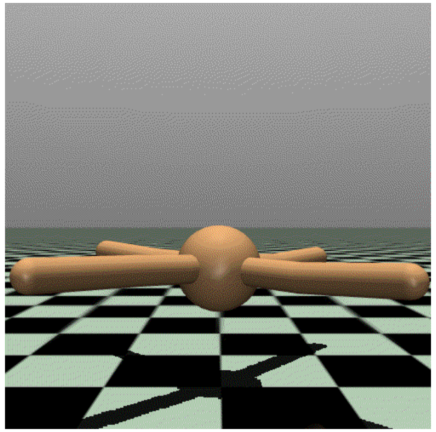
5. Evaluation

Overview of the used environments

Task Name	State Shape	Action Shape
Lunar Lander (Continuous)	Continuous(6-dim) Discrete(2-dim)	Continuous(2-dim)
Bipedal Walker	Continuous(24-dim)	Continuous(4-dim)
Ant	Continuous(111-dim)	Continuous(8-dim)







5. Evaluation

Main steps in dataset generation and offline DRL model preparation

Dataset generation

For each combination of task and offline RL model in the experiment

↓

Train with 5 random seeds: $\{0, 1, \ldots, 4\}$

↓

5 online RL models detailed in Table XIV

↓

Collect with 1 random seed: $\{0\}$

↓

5x1 offline Datasets detailed in Table XV

↓

Model preparation

Train with 30 random seeds: $\{42, 43, \ldots, 72\}$

↓

5x1x30 offline RL models detailed in Table XVI, Table XVII, Table XVIII, and Table XIX

Overall auditing performance

Task Name	Offline Model	L1 Norm		L2 Norm		Cosine Distance		Wasserstein Distance	
		TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
Lunar Lander	BC	**99.01±0.46**	**100.00±0.00**	96.96±0.73	100.00±0.00	96.93±0.77	100.00±0.00	98.40±0.74	99.94±0.16
	BCQ	**98.29±1.14**	**100.00±0.00**	96.03±1.15	100.00±0.00	95.97±1.07	99.99±0.04	97.57±1.17	99.91±0.14
	IQL	**98.61±1.51**	**99.91±0.32**	97.52±2.51	99.97±0.12	97.49±2.56	99.92±0.19	98.32±1.79	97.10±5.66
	TD3PlusBC	**98.29±2.04**	**99.48±0.79**	96.35±3.01	99.89±0.22	96.27±3.16	99.91±0.23	98.53±1.25	95.59±3.77
Bipedal Walker	BC	99.20±1.47	100.00±0.00	98.40±2.70	100.00±0.00	98.56±2.68	100.00±0.00	**99.31±1.32**	**100.00±0.00**
	BCQ	99.52±0.77	100.00±0.00	98.16±2.89	100.00±0.00	99.87±0.15	100.00±0.00	**99.89±0.13**	**100.00±0.00**
	IQL	95.10±7.41	100.00±0.00	95.04±5.45	100.00±0.00	**99.84±0.32**	**100.00±0.00**	95.01±6.72	100.00±0.00
	TD3PlusBC	**99.36±1.28**	**94.77±19.42**	97.15±5.71	93.36±21.46	96.96±5.82	91.98±21.75	98.08±3.84	88.26±25.34
Ant	BC	97.42±1.66	99.94±0.11	96.48±1.66	99.90±0.36	99.20±1.08	85.66±28.23	**98.00±1.19**	**99.92±0.14**
	BCQ	97.17±2.96	99.80±0.43	95.68±2.54	99.84±0.43	99.66±0.43	86.70±26.89	**98.67±1.65**	**99.79±0.46**
	IQL	97.20±2.33	99.66±0.73	96.61±2.50	99.69±0.59	99.57±0.79	86.25±27.90	**99.36±0.42**	**99.63±0.78**
	TD3PlusBC	98.53±1.80	99.18±1.72	97.17±1.79	99.35±1.74	99.72±0.40	87.79±26.43	**99.25±1.24**	**99.14±1.81**

Remarks

- **Distance metrics:** Different auditing accuracy over four distance metrics
- **Hypothesis testing:** The auditing accuracy as determined by Grubbs' test outperforms that of the 3σ principle

5. Evaluation

Visualization of cumulative rewards

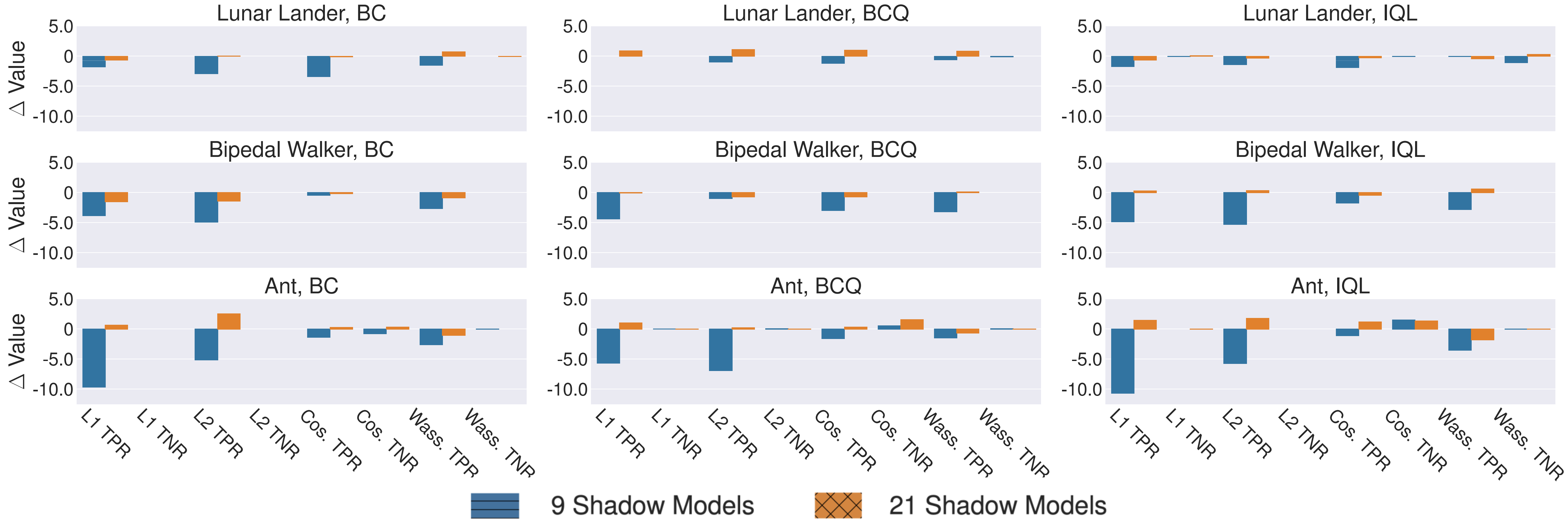
Bipedal Walker, BC

Bipedal Walker, BCQ

Remarks

- **Cumulative rewards:** The cumulative rewards reflect the differences in models' state-action pairs
- **Difference in trajectories:** The distribution of points varies on the different trajectories

Hyperparameter study (Shadow models' amount)

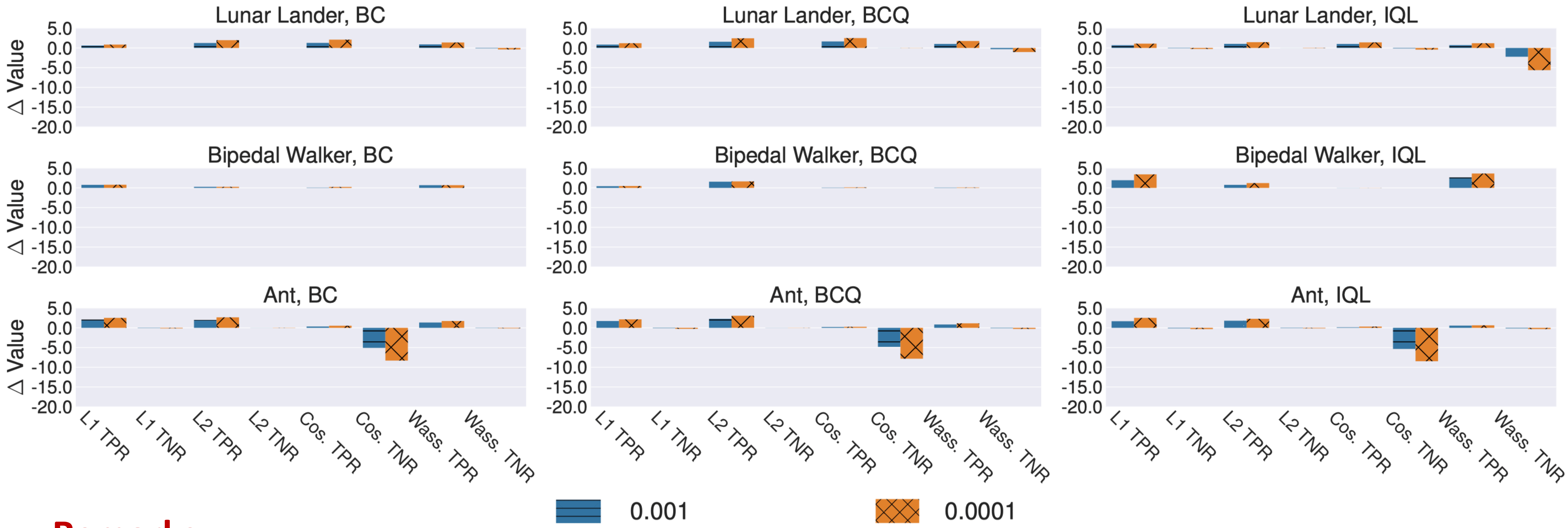


Remarks

- **Benefits of more shadow models**: The auditing accuracy increases with a larger amount of shadow models
- **Saturation point:** There exists a saturation point for auditing accuracy with the expansion of shadow models

5. Evaluation

Hyperparameter study (Significance level)

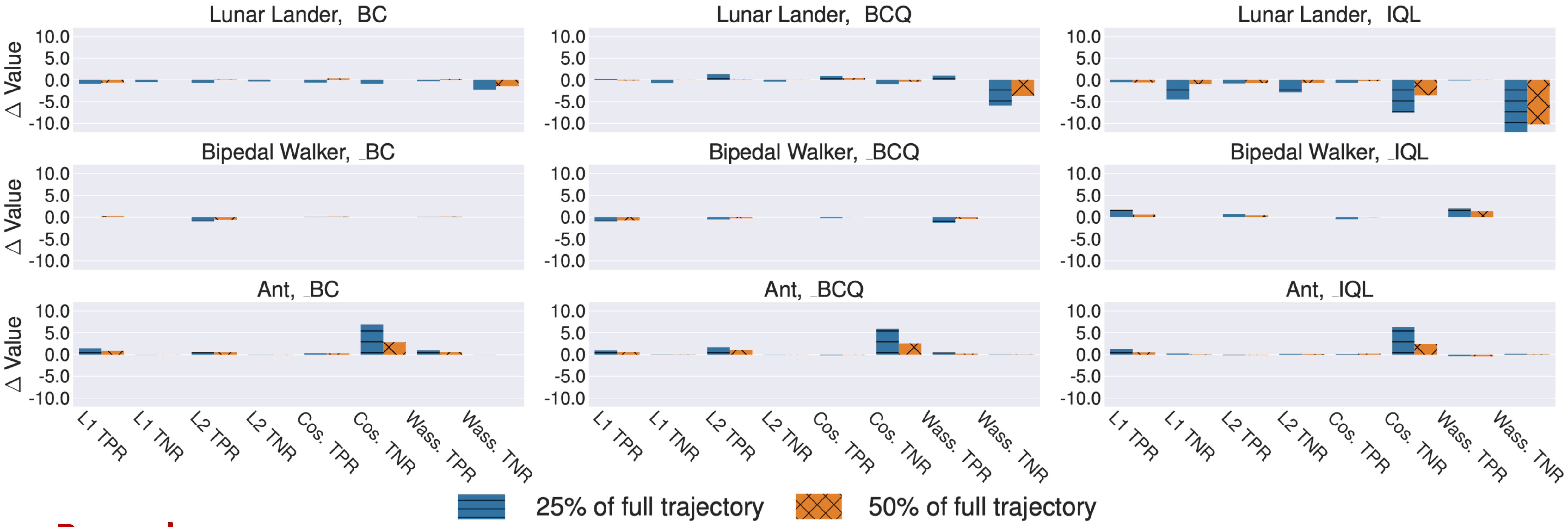


Remarks

- For a complicated task, we recommend the auditor to select a large significance level
- For the suspect models with low performance, ORL-Auditor should adopt a large significance level

5. Evaluation

Hyperparameter study (Trajectory size)



Remarks

- **Benefits of larger trajectory size**: ORL-Auditor tends to achieve a higher accuracy with a larger trajectory size
- A small trajectory size achieves better results under some tasks, since **the front states of some trajectories can sufficiently reflect behavioral preference** of the model [Paine, et al. (2020)]

5. Evaluation

Robustness (Ensemble architecture [USENIX Security '22, PETS '23])

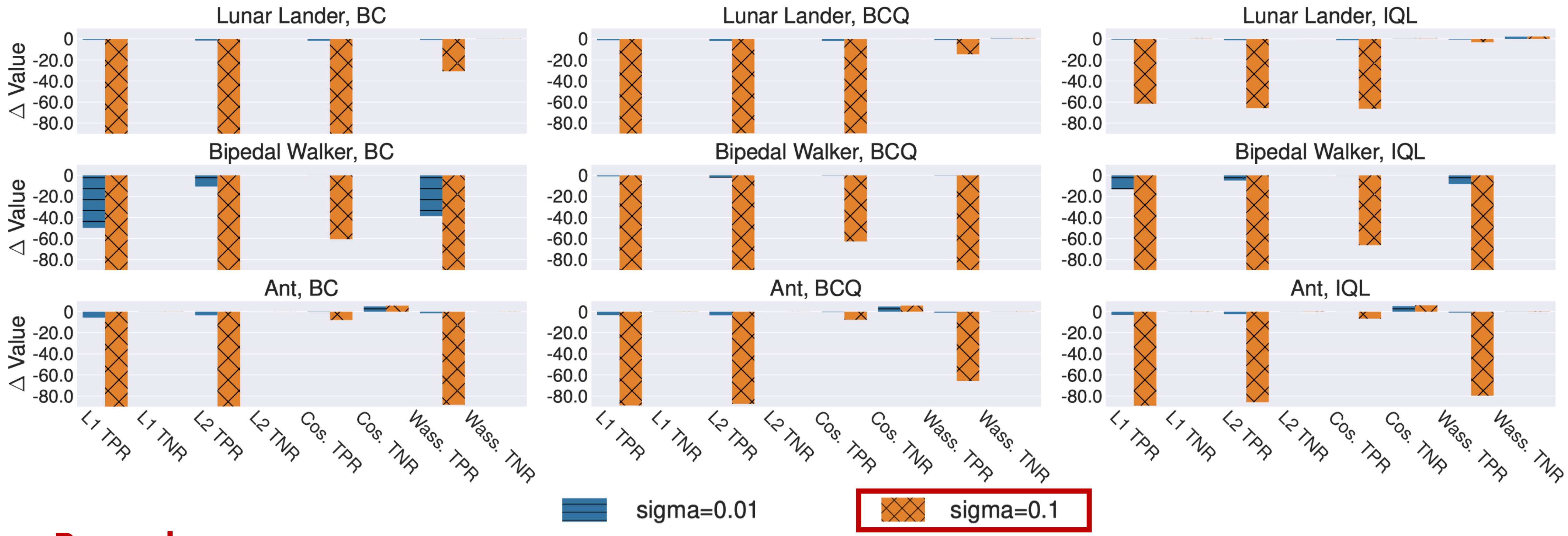
Task Name	Offline Model	L1 Norm		L2 Norm		Cosine Distance		Wasserstein Distance	
		TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
Lunar Lander	BC	100.00±0.00	100.00±0.00	99.20±0.98	100.00±0.00	99.20±0.98	100.00±0.00	99.60±0.80	99.90±0.44
	BCQ	99.60±0.80	100.00±0.00	98.00±2.19	100.00±0.00	98.00±2.19	100.00±0.00	99.60±0.80	100.00±0.00
	IQL	100.00±0.00	99.90±0.44	99.20±0.98	100.00±0.00	99.60±0.80	99.90±0.44	99.60±0.80	97.60±4.27
	TD3PlusBC	100.00±0.00	99.30±0.95	99.60±0.80	99.90±0.44	99.60±0.80	99.80±0.60	99.60±0.80	95.80±3.57
Bipedal Walker	BC	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
	BCQ	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
	IQL	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
	TD3PlusBC	100.00±0.00	94.90±19.07	100.00±0.00	93.80±21.63	100.00±0.00	92.70±21.62	100.00±0.00	89.20±23.94
Ant	BC	99.60±0.80	100.00±0.00	99.60±0.80	99.90±0.44	99.60±0.80	83.20±31.99	99.20±1.60	100.00±0.00
	BCQ	100.00±0.00	99.70±0.71	99.60±0.80	99.80±0.60	100.00±0.00	85.70±28.31	100.00±0.00	99.70±0.71
	IQL	100.00±0.00	99.80±0.60	99.20±0.98	99.70±0.71	99.20±0.98	86.80±28.32	100.00±0.00	99.80±0.60
	TD3PlusBC	99.60±0.80	99.30±1.82	100.00±0.00	99.40±2.20	100.00±0.00	87.80±25.87	99.60±0.80	98.50±3.79
Half Cheetah	BC	85.00±25.98	100.00±0.00	84.50±25.71	100.00±0.00	94.00±10.39	67.50±43.20	87.00±21.38	100.00±0.00
	BCQ	91.00±15.59	100.00±0.00	89.00±16.76	100.00±0.00	95.00±8.66	67.17±42.30	93.00±12.12	100.00±0.00
	IQL	90.00±12.81	100.00±0.00	86.50±16.70	100.00±0.00	94.50±9.53	71.00±41.37	91.50±12.52	100.00±0.00
	TD3PlusBC	61.50±20.32	100.00±0.00	77.00±19.42	100.00±0.00	95.00±8.66	65.67±41.28	52.00±33.26	100.00±0.00

Remarks

- ORL-Auditor maintains a high level of auditing accuracy
- Integrating **more distance metrics in the auditing process** can enhance the robustness

5. Evaluation

Robustness (Action distortion)



Remarks

- ORL-Auditor can resist the potential action distortion from the suspect model
- ORL-Auditor with a **single distance metric faces limitations for a strong distortion**

6. Conclusion

Highlights

- ORL-Auditor is the **first approach to conduct trajectory-level dataset auditing** for offline DRL models
- We conclude some **useful observations for adopting ORL-Auditor**
- We apply ORL-Auditor to audit the models trained on **the open-source datasets from Google and DeepMind**, where all TPR and TNR results are superior to 95%

Limitations and Future Work

- The accuracy of ORL-Auditor **decreases when the significance level downs to 0.001**. Thus, it is interesting to enhance ORL-Auditor to satisfy stricter auditing demands in the future
- ORL-Auditor based on a **single distance metric** may not be sufficiently robust to strong distortion

ORL-Auditor: Dataset Auditing in Offline Deep Reinforcement Learning

Get the full paper ☺

Linkang Du, Min Chen, Mingyang Sun, Shouling Ji,

Peng Cheng, Jiming Chen, and Zhikun Zhang