The Three Regimes of Offline-to-Online Reinforcement Learning

Lu Li, Tianwei Ni, Yihao Sun, Pierre-Luc Bacon

Lu Li at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation

July 11th, 2026

Offline-to-Online RL

Offline pretraining:

$$\pi_0 = A_{\text{off}}(\mathcal{D})$$

Online fine-tuning:

$$\pi_N = A_{\text{on}}(\mathcal{M}, \mathcal{D}, \pi_0)$$

Goal: co-design offline algorithm and online algorithm, to maximize $J(\pi_N)$

Why Offline-to-Online RL?

- Online RL (from scratch) is highly data-intensive, making large-scale interaction impractical in many real-world applications.

- Offline RL leverages pre-collected datasets and avoids online interaction, but often produces suboptimal policies.

Key idea of O2O RL: Learn a policy from an offline dataset and subsequently fine-tune it through online interaction to further improve performance.
➔ Combines the data efficiency of offline RL with the performance gains of online RL

Inconsistency in Fine-tuning Results

Prior Work

Warm-Start RL (WSRL): offline pretraining, only online data in fine-tuning

Efficient Online Reinforcement Learning Fine-Tuning Need Not Retain Offline Data

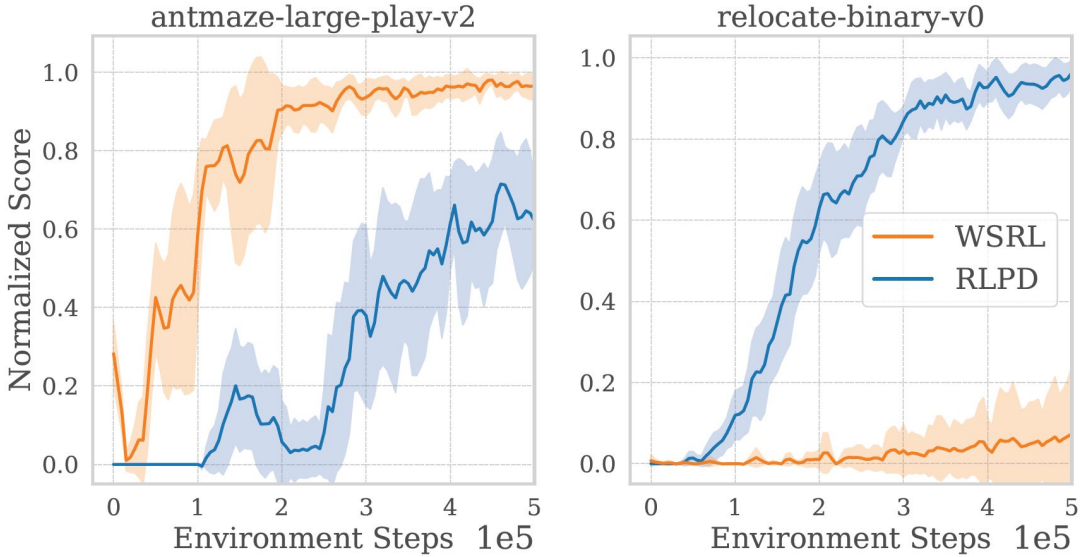
Zhiyuan Zhou[*1], Andy Peng[*1], Qiyang Li[1], Sergey Levine[1], Aviral Kumar[2]
[1]UC Berkeley, [2]Carnegie Mellon University (*Equal Contribution)

RL with Prior Data (RLPD): no offline pretraining, combing offline data and online data in fine-tuning

Efficient Online Reinforcement Learning with Offline Data

Philip J. Ball [*1] Laura Smith [*2] Ilya Kostrikov [*2] Sergey Levine [2]

Inconsistency in Online Fine-tuning



Question: What **underlying factors** cause design choices to succeed in some settings but fail in others?

Mila

A Decomposition of Performance Based on Stability–Plasticity Principle

Stability–Plasticity dilemma

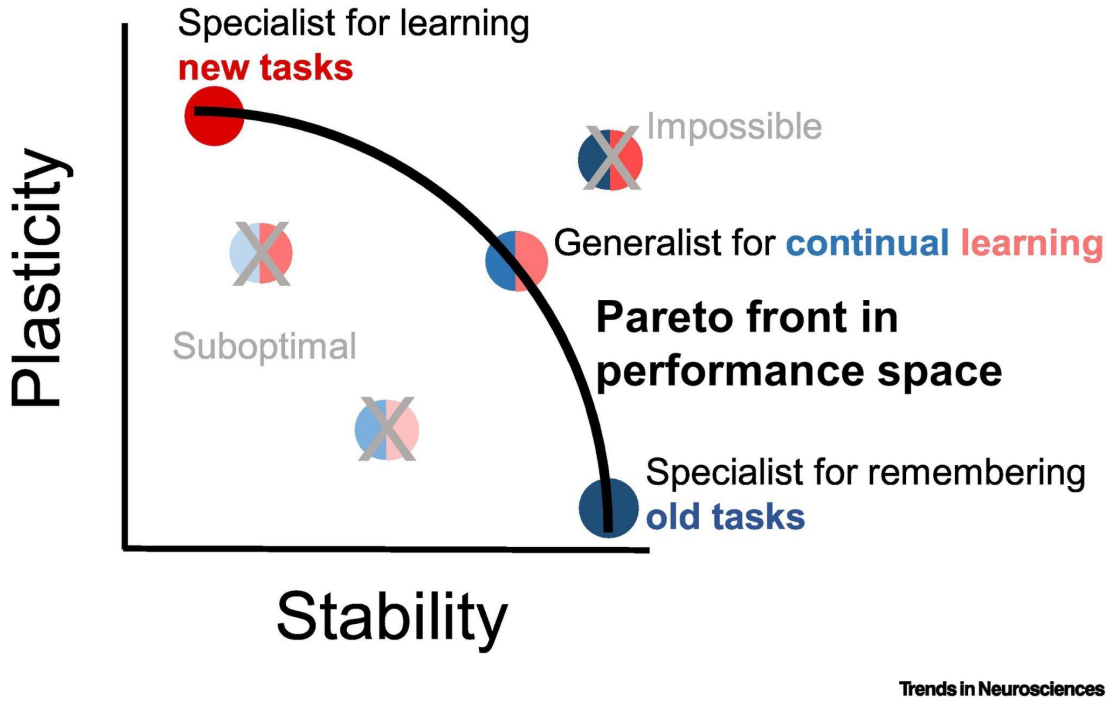


Figure from paper: metaplasticity to solving the Catastrophic Forgetting Problem

Quantifying Priors

Performance of the pretrained policy: $J(\pi_0)$

Performance of the the dataset: $J(\pi_{\mathcal{D}}) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} r_{i,t}.$

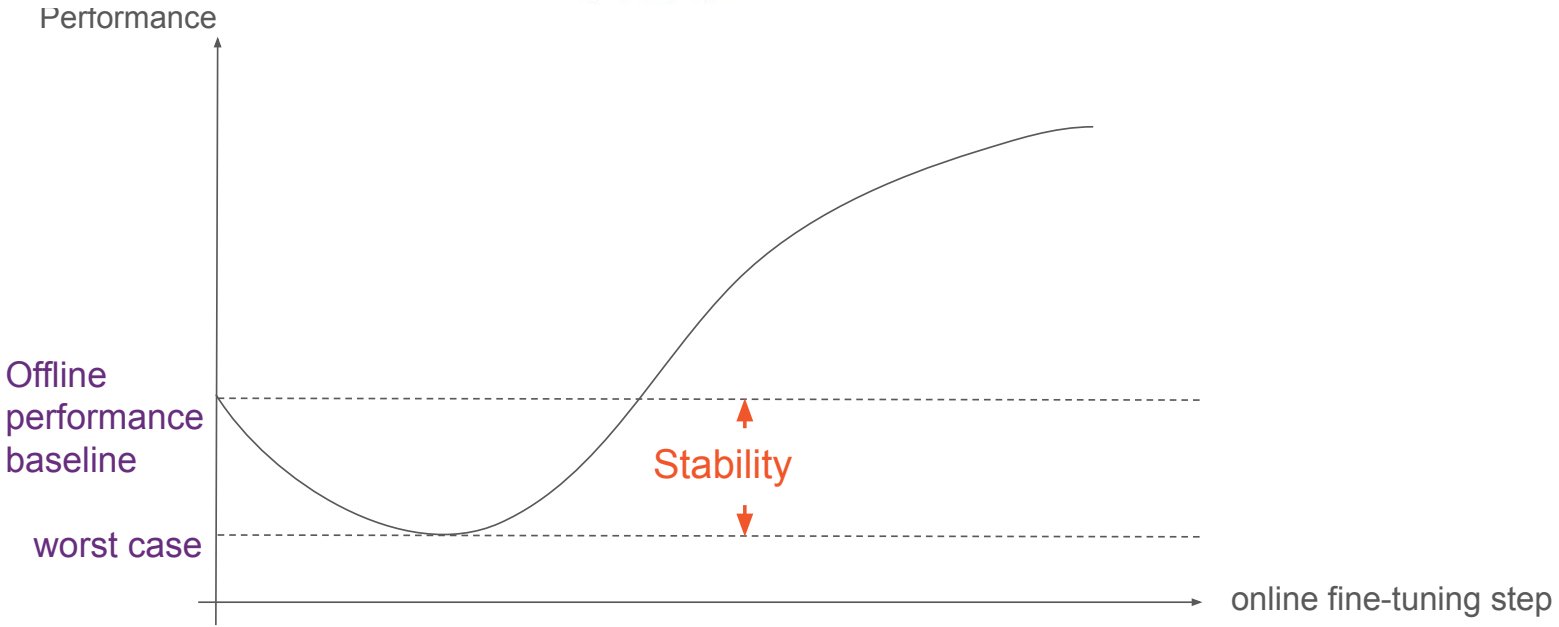
Offline performance baseline:

$$J^*_{\text{off}} := \max\left(J(\pi_0),\ J(\pi_{\mathcal{D}})\right)$$

Performance Decomposition

Stability with respect to offline performance baseline:

$$\text{Stability}(J^*_{\text{off}}) = \min_{0 \le n \le N} J(\pi_n) - J^*_{\text{off}} \le 0.$$

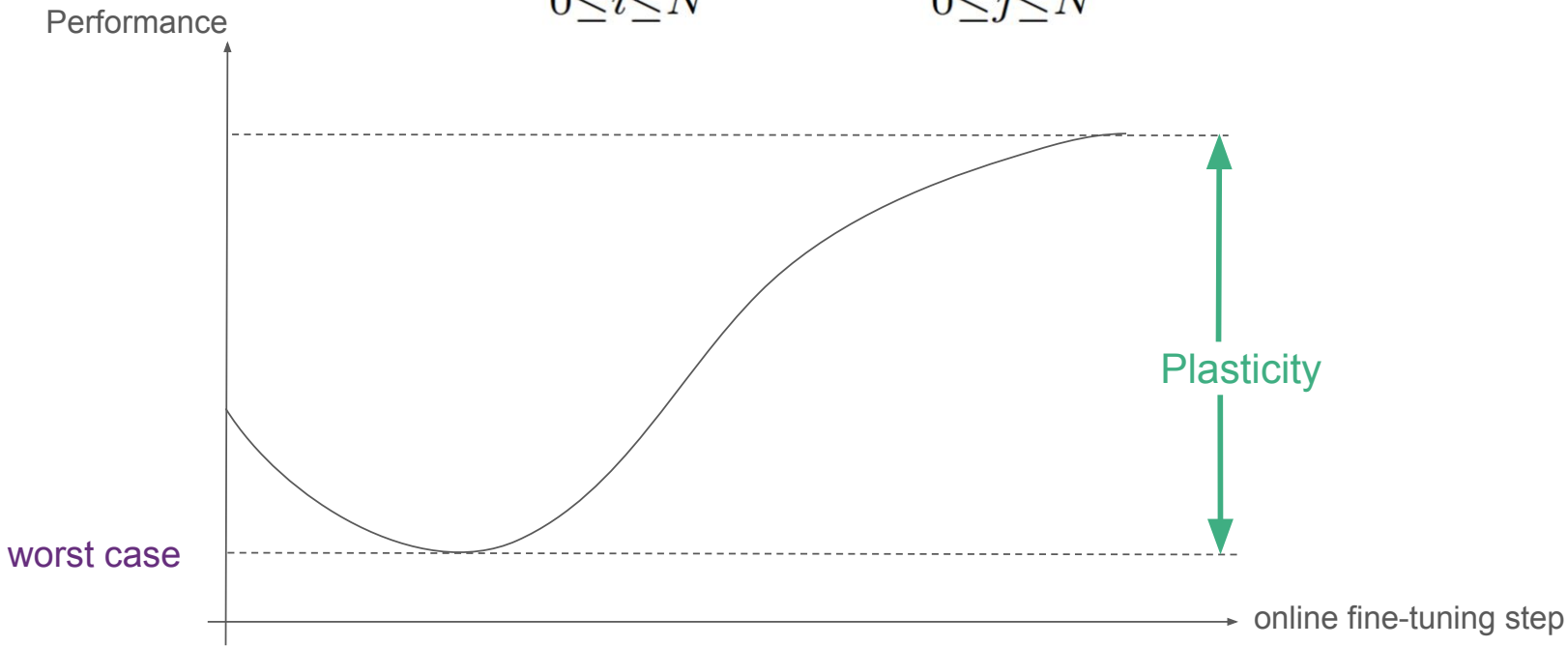


Mila

Performance Decomposition

Plasticity: the ability to improve its performance, relative to the its lowest observed value

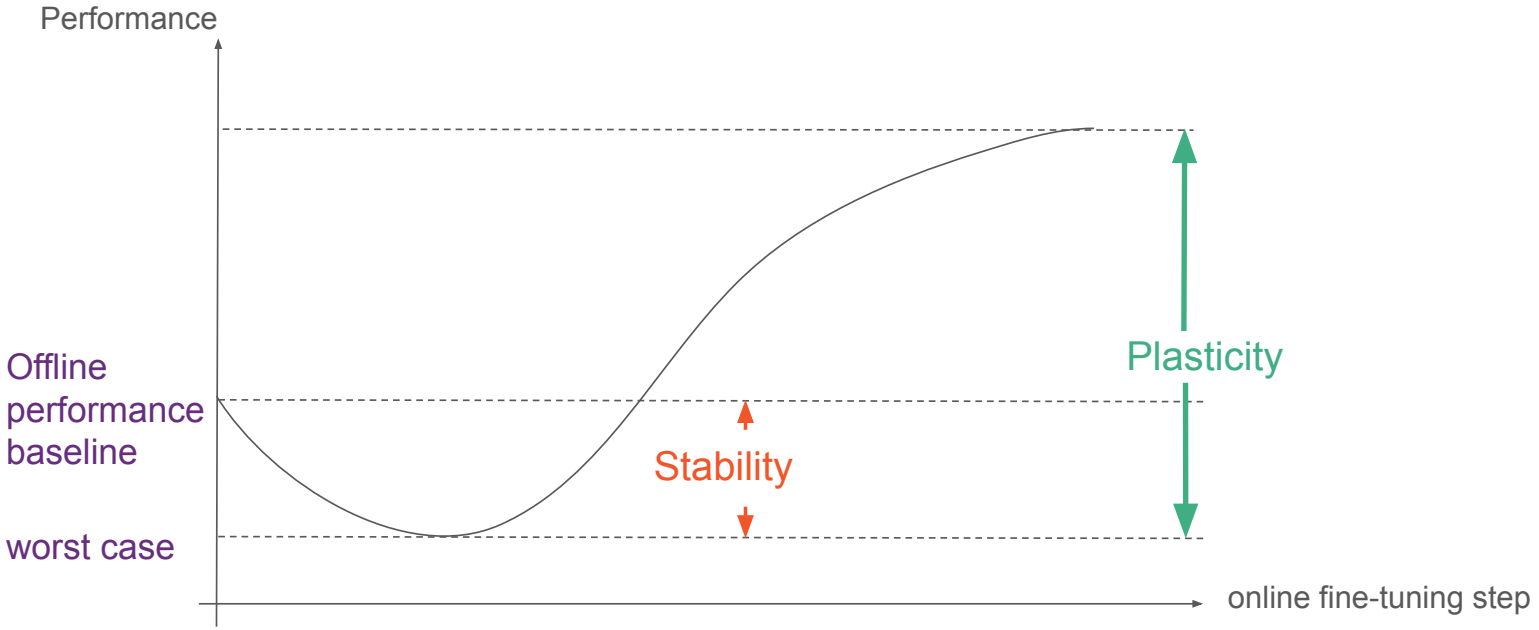
$$\text{Plasticity} = \max_{0 \le i \le N} J(\pi_i) - \min_{0 \le j \le N} J(\pi_j) \ge 0.$$



Mila

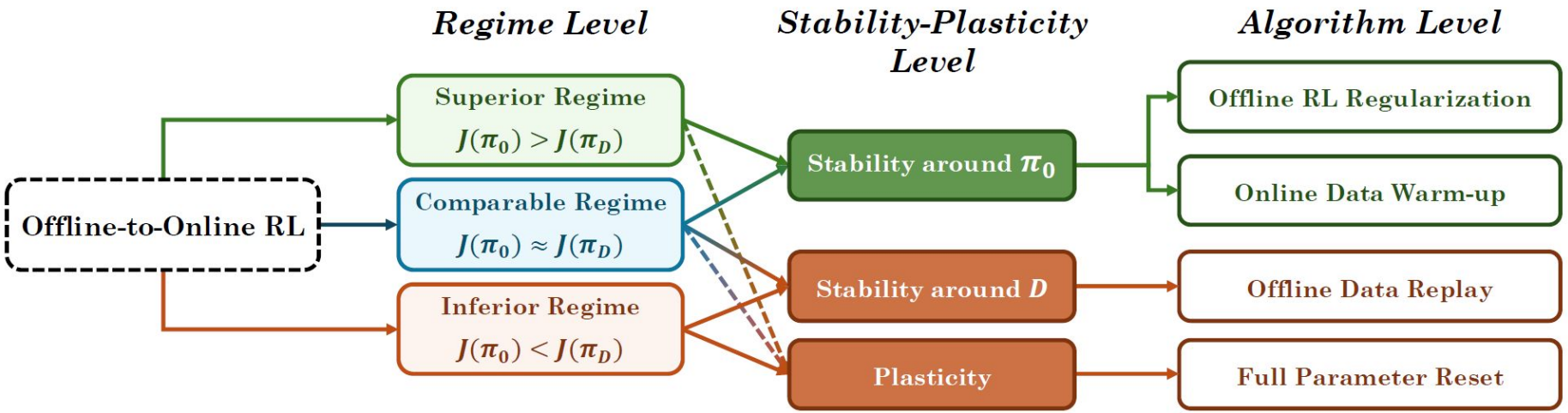
Performance Decomposition

$$\underbrace{\max_{0 \le n \le N} J(\pi_n)}_{\text{Best Performance}} = \underbrace{J^*_{\text{off}}}_{\text{(1) Prior}} + \underbrace{\text{Stability}(J^*_{\text{off}})}_{\text{(2) Degradation} \le 0} + \underbrace{\text{Plasticity}}_{\text{(3) Online Improvement} \ge 0}$$



Mila

The three regimes



Mila

Minimal baseline

We define a naive online RL fine-tuning baseline, which is ***intentionally minimalist***: use a standard online RL algorithm (e.g. SAC) initialized with an offline pretrained agent

- no offline data replay during fine-tuning
- no additional regularization
- ...

Design Choices in Improving Stability

Stability around the pretrained policy

- online data warm-up
- offline RL regularization (e.g. conservatism)
- KL regularization
- learning rate warm-up
- …

Design Choices in Improving Stability

Stability around the pretrained policy

- online data warm-up
- offline RL regularization (e.g. conservatism)
- KL regularization
- learning rate warm-up
- ...

Stability around the offline dataset

- offline data replay(append): initial the buffer with offline data
- offline data replay(separate): one offline data buffer, one online buffer
- data generative methods
- ...

Design Choices in Improving Plasticity

Plasticity

- parameter reset: reset network parameters to random value, which severely degrades initial performance but also significantly enhances plasticity.

- other variant of parameter reset that proposed in online RL

Empirical Study

Offline Pretraining

Use 21 task-dataset compositions from D4RL dataset,
with 3 different pretraining algorithms:

- CalQL: offline RL algorithm based on SAC
- ReBRAC: offline RL algorithm based on TD3
- Behavior Cloning: imitation learning algorithm (commonly used in robotics)

lead to 63 settings

Regime Classification

use T-test with margin to classify the pretrained results into three regimes: **Superior**, **Inferior**, and **Comparable**

$$H_0 : \mu_0 - \mu_D \leq \delta \text{ vs } H_1 : \mu_0 - \mu_D > \delta$$ If rejected ➔ **Superior**

$$H_0 : \mu_D - \mu_0 \leq \delta \text{ vs } H_1 : \mu_D - \mu_0 > \delta$$ If rejected ➔ **Inferior**

If neither test rejects ➔ **Comparable**

Categorizing Algorithms into 4 Types

group algorithms by the primary source of stability they emphasize:

- minimal baseline
  - standard online RL (SAC or TD3)
- ***π_0*-centric method**
  - online data warm-up (WSRL-style)
  - offline RL regularization (CalQL or ReBRAC)
- ***D*-centric method**
  - offline data replay (use online and offline buffer with 1:1 ratio)
  - offline data replay + full parameter reset (RLPD-style)
- mixed ***π_0*+*D* method**
  - offline RL regularization + offline data replay (Most prior O2O work)

Comparing Fine-tuning Results

We focus on compare the strongest ***π_0*-centric methods and *D*-centric methods** to better approximate the ideal performance achievable by each stability source.

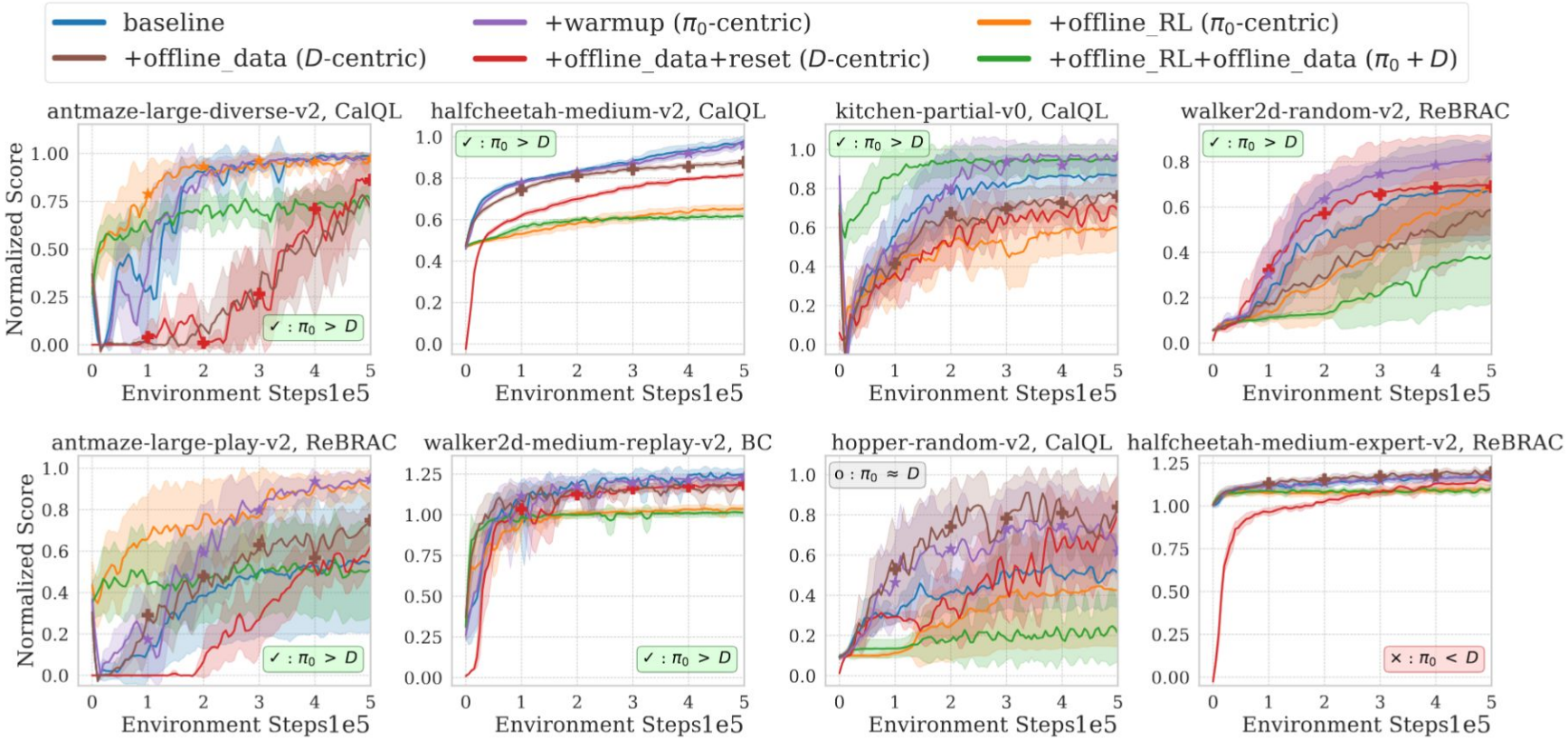
Thus, there are three kinds of fine-tuning results, which is obtained by t-test:

Superior Regime $\xrightarrow{?}$ $\pi_0\text{-centric} > \mathcal{D}\text{-centric}$

Comparable Regime $\xrightarrow{?}$ $\pi_0\text{-centric} \approx \mathcal{D}\text{-centric}$

Inferior Regime $\xrightarrow{?}$ $\pi_0\text{-centric} < \mathcal{D}\text{-centric}$

Superior Regime



Mila

Fine-tuning Results in Superior Regime

		Pretraining Regime		
		Superior	**Comparable**	**Inferior**
Fine-tune	π_0-centric $>$ $\mathcal{D}$-centric	**24**	2	1
	π_0-centric $\approx$ $\mathcal{D}$-centric	6	**2**	3
	π_0-centric $<$ $\mathcal{D}$-centric	2	4	**19**

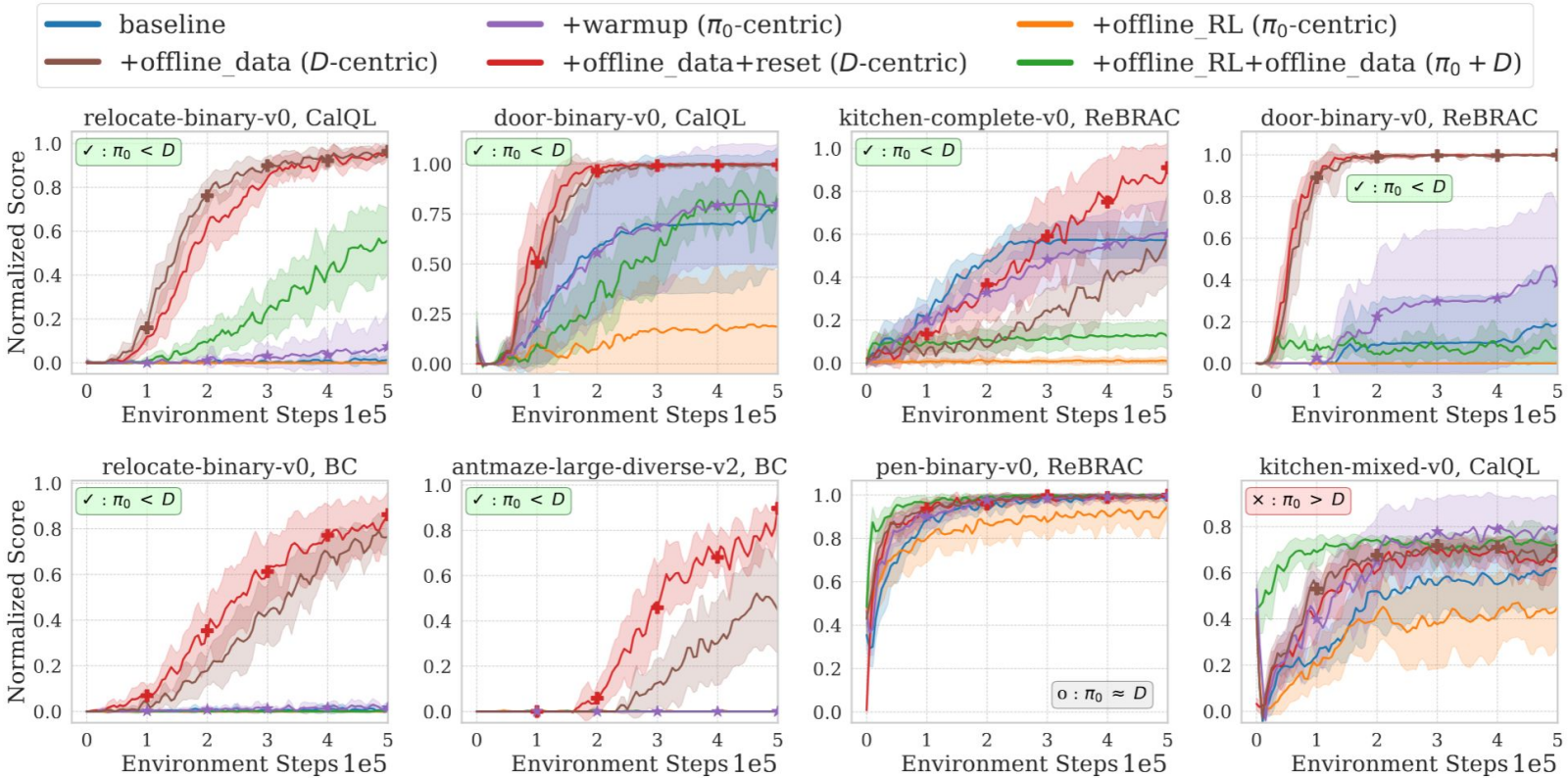
24/32(**75%**) correct predictions, only 2/32(**6%**) opposite mismatches
Takeaway: In the Superior regime, ***π***_0-centric methods typically provide more effective fine-tuning than *D*-centric methods.

Fine-tuning Results in Superior Regime

Within π_0-centric methods,
online data warm-up outperform *offline RL regularization* in 27/32(84%).
offline RL regularization (and +*offline data replay*) only works well in those settings where π_0 is close to optimal.

Since *offline RL regularization* provide much stronger stability which at the cost of plasticity, lead to less performance degradation, but also limits further improvements.

Inferior Regime



Mila

Fine-tuning Results in Inferior Regime

		Pretraining Regime		
		Superior	Comparable	Inferior
Fine-tune	π_0-centric $>$ $\mathcal{D}$-centric	**24**	2	1
	π_0-centric $\approx$ $\mathcal{D}$-centric	6	**2**	3
	π_0-centric $<$ $\mathcal{D}$-centric	2	4	**19**

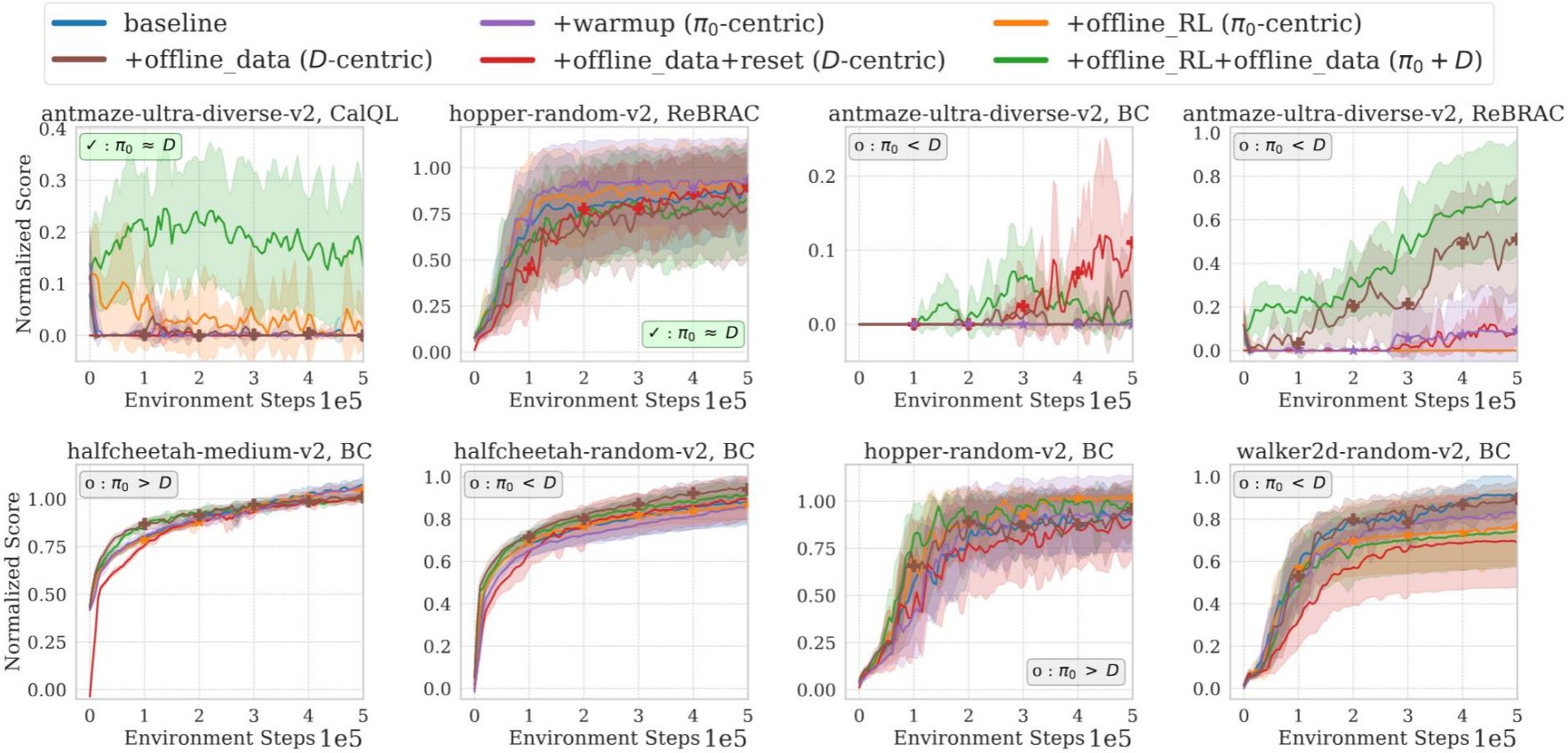
19/23(**83%**) correct predictions, only 1/23(**4%**) opposite mismatches
Takeaway: In the Inferior regime, *D*-centric methods typically provide more effective fine-tuning than ***π***_0-centric methods.

Fine-tuning Results in Inferior Regime

Within *D*-centric methods,
Out of 24 settings, *offline data replay + reset* performs better in 13 settings and *offline data replay* performs better in 11 settings.

resetting the parameters allows the agent to adapt and acquire new knowledge more effectively when offline pretraining phase substantially reduces plasticity while offering limited knowledge

Comparable Regime



Mila

Fine-tuning Results in Comparable Regime

Fine-tune	Pretraining Regime: Superior	Comparable	Inferior
π_0-centric $>$ $\mathcal{D}$-centric	**24**	2	1
π_0-centric $\approx$ $\mathcal{D}$-centric	6	**2**	3
π_0-centric $<$ $\mathcal{D}$-centric	2	4	**19**

2/8(25%) correct predictions.

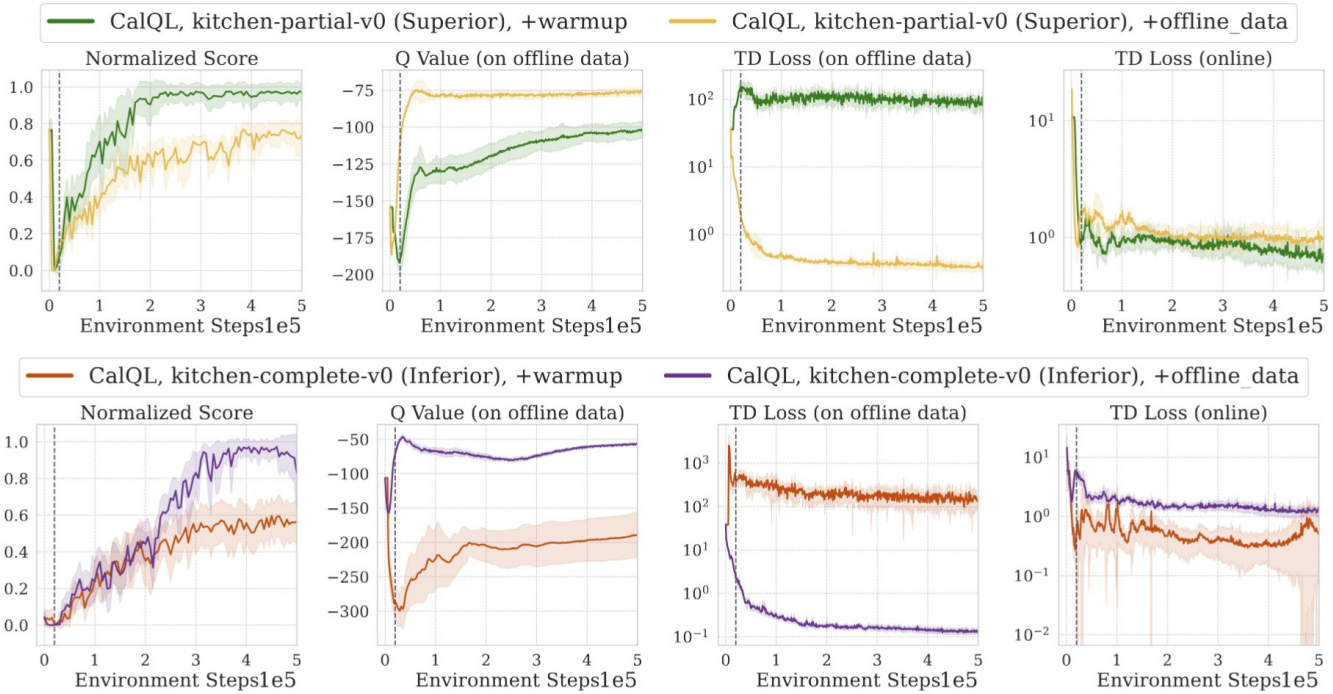
Takeaway: In the Comparable regime, ***π***_0-centric method and *D*-centric method should in principle yield comparable fine-tuning outcomes, though in practice their relative performance is often sensitive to implementation details.

Overall Results and Alternative Metrics

		Pretraining Regime		
		Superior	Comparable	Inferior
Fine-tune	π_0-centric $>$ $\mathcal{D}$-centric	**24**	2	1
	π_0-centric $\approx$ $\mathcal{D}$-centric	6	**2**	3
	π_0-centric $<$ $\mathcal{D}$-centric	2	4	**19**

Domains	Metric	Accuracy ↑	Opposite mismatch ↓
Sparse-reward (27 cases)	Raw return (ours)	**78%**	**4%**
	Dense-reward proxy	65%	11%
All domains (63 cases)	Raw return (ours)	**71%**	**5%**
	Q-function	51%	30%
	BC performance	41%	**5%**

Mechanistic Analysis

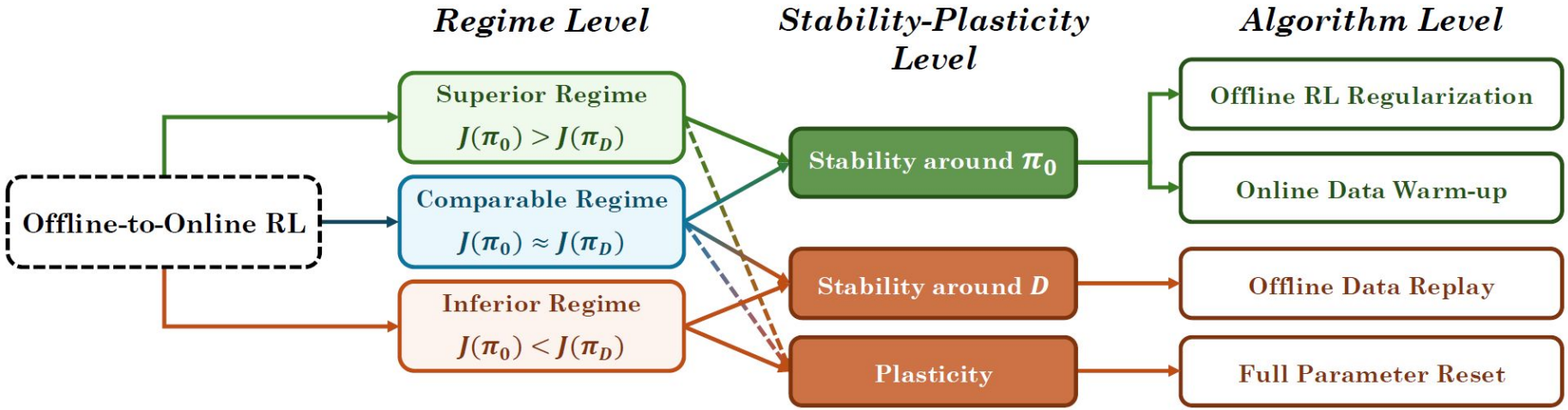


Fine-tuning without offline data results in significantly higher TD loss on the offline dataset in the Inferior regime compared to the Superior regime, and also leads to divergence of the Q-values on the offline data.

Mila

Conclusion

The three regimes



- We provide a **clear explanation** of the conflicting empirical evidence: design choices that seem inconsistent across benchmarks in fact reflect different underlying regimes.
- We offer an **actionable guidance** for practitioners. By identifying the regime of a given setting, one can select methods that align with its stability–plasticity requirements, reducing reliance on trial-and-error.
- The three regimes provides an efficient lens for **understanding offline-to-online RL**. But we also have to recognize that such discretizing behavior is a simplification.

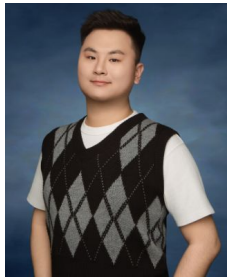
Thanks to co-authors

Tianwei Ni

Yihao Sun

Pierre-Luc Bacon

Special thanks to Guozheng Ma for insightful discussions.



Thanks for listening!
Looking forward to your thoughts!

contact:
lu.li@mila.quebec
twni2016@gmail.com

Read our paper

Read our blog