From Deployable Policies to Adaptive Priors in Offline Reinforcement Learning

Tianwei Ni[1,2]* **Vineet Jain**[1,3] **Akash Karthikeyan**[1,2] **Pierre-Luc Bacon**[1,2]
[1]Mila - Quebec AI Institute [2]Université de Montréal [3]McGill University

Abstract

Offline reinforcement learning (RL) has traditionally focused on learning policies for direct deployment under conservative objectives, where uncertainty outside the offline dataset is treated pessimistically to ensure robustness. We argue that this formulation becomes incomplete when offline learning is followed by further online interaction, as increasingly occurs in modern intelligent systems through test-time adaptation and online fine-tuning. This position paper argues that, in such settings, the objective of offline RL should extend beyond immediate deployment and instead prioritize learning *adaptive policy priors*: policies that preserve the capacity to improve during subsequent interaction through memory, exploration, and self-correction. We formalize this perspective as *adaptive offline reinforcement learning* (AORL), distinguish it from offline-to-online RL, and explain why adaptability becomes important under distribution shift, limited dataset coverage, and changing test-time conditions. We further discuss Bayesian offline RL as one principled direction for constructing adaptive policy priors by preserving epistemic uncertainty over plausible environments. Finally, we outline connections, open challenges, and research directions for treating offline RL as preparation for future experience rather than solely as a static deployment problem.

1 Introduction

Offline reinforcement learning (RL) studies how to optimize a policy using a static dataset of trajectories, without relying on online interaction during training[2]. In its classical formulation (Levine et al., 2020), the objective is to learn a policy for *direct deployment*: the offline-learned policy is treated as a final decision rule whose performance is expected to derive entirely from offline data rather than from subsequent interaction. This deployment-oriented formulation makes distribution shift particularly challenging (Kumar et al., 2019): actions chosen outside the dataset may lead to states where value estimates are unreliable, errors compound over time, and correcting such errors through further interaction is not part of the learning objective.

As a result, the dominant design principle of offline RL has been *conservatism*: uncertainty about out-of-dataset actions is treated pessimistically so that learned policies remain safe and robust under limited support. This principle appears through imitation objectives (Pomerleau, 1988; Fujimoto & Gu, 2021), policy constraints (Fujimoto et al., 2019; Wu et al., 2019; Peng et al., 2019), pessimistic value estimation (Kumar et al., 2020; An et al., 2021; Kostrikov et al., 2022), or uncertainty penalties (Yu et al., 2020; Kidambi et al., 2020; Bai et al., 2022). This conservative formulation has driven much of the progress in offline RL (Reed et al., 2022; Lee et al., 2022a; Kumar et al., 2023; Park et al., 2025b), particularly in benchmark settings where policies are evaluated under fixed environment conditions.

*Correspondence to: twni2016@gmail.com. Blog post: twni2016.github.io/blogs/policyprior/main.html.

[2]We refer to "training" as parameter updates in the policy, in contrast to in-context learning that does not modify parameters (Brown et al., 2020).

Preprint, March 14, 2026.

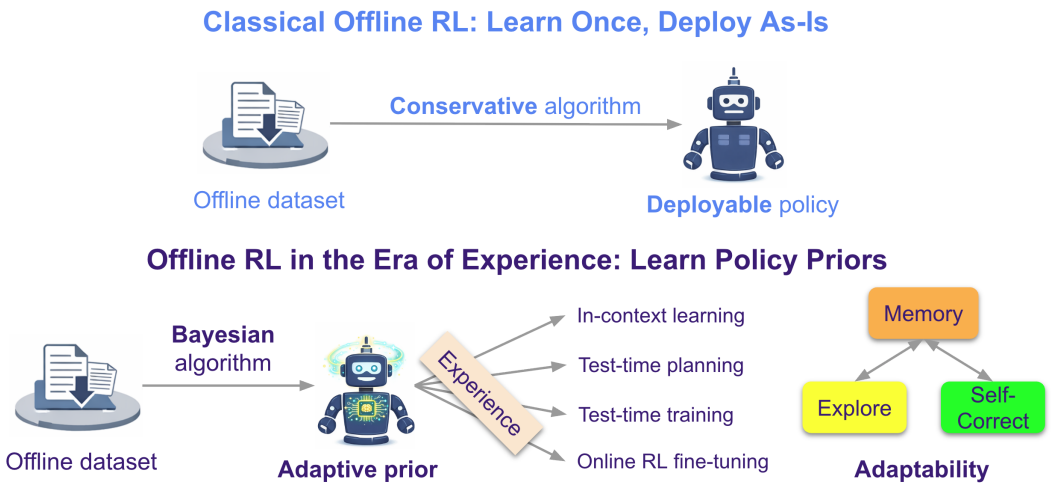

Figure 1: Top: **Classical offline RL** learns a deployable policy from a static dataset under conservative objectives. Bottom: **Adaptive offline RL** learns a policy prior that remains improvable through subsequent experience, including in-context learning, planning, or online fine-tuning. Here, **adaptability** arises from the interaction between memory, exploration, and self-correction. Bayesian perspective is highlighted as one principled direction for constructing such adaptive priors.

A fundamental limitation of this formulation is that it leaves little room for improvement after offline learning, even though such improvement is increasingly central to how intelligent systems operate. This becomes especially important in the *era of experience* (Silver & Sutton, 2025), where performance depends not only on knowledge extracted from static datasets, but also on the ability to acquire and exploit new information through online interaction in continual and open-ended environments (Khetarpal et al., 2022). In these settings, offline optimization is not the endpoint of learning; rather, it produces an initial policy whose quality should also be judged by how effectively it improves after deployment.

Recent practice already reflects this shift across domains, including large language models (LLMs) and robotics, where offline-learned policies are routinely followed by mechanisms that learn from subsequent experience, such as test-time in-context learning (Brown et al., 2020; Wei et al., 2022), planning over imagined trajectories (Yao et al., 2023; Snell et al., 2025), or online RL fine-tuning (Guo et al., 2025a,b). Under this broader role, excessive conservatism carries an opportunity cost: actions assigned near-zero probability during offline learning may become difficult to recover later, even when subsequent interaction would reveal them to be beneficial.

In this position paper, we use the term *adaptive offline reinforcement learning* (AORL) to refer to settings where offline RL should prioritize adaptability beyond direct deployment. **Our position is that when policies will continue improving through interaction, offline RL should learn adaptive policy priors rather than fully deployable policies.** This differs from *offline-to-online RL* (Nair et al., 2020), where offline learning primarily serves as initialization for subsequent online *training*. In AORL, the objective of the offline stage itself is to preserve the capacity for later adaptation, even when online interaction is limited or occurs without parameter updates.

Under this view, offline learning should not eliminate uncertainty solely for immediate robustness, but preserve sufficient behavioral flexibility for later adaptation. Such adaptability requires three ingredients: *memory*, so that decisions depend on online history; *exploration*, so that uncertain but potentially valuable actions remain reachable; and *self-correction*, so that new evidence can revise early mistakes during interaction. Actions outside the dataset are therefore not inherently undesirable; rather, they reflect epistemic uncertainty that can later be resolved through in-context learning, planning, or fine-tuning. Fig. 1 summarizes this shift from deployable policies to adaptive policy priors.

One principled direction for AORL arises from Bayesian perspectives (Ghosh et al., 2022; Ni et al., 2025). In this view, offline RL is cast as an epistemic POMDP (Ghosh et al., 2021): limited dataset coverage induces a posterior distribution over plausible MDPs that agree on observed data while

differing beyond dataset support. The resulting optimal policy is naturally *adaptive in context*: by conditioning on online history, it can first explore uncertain but potentially good actions, then exploit the most promising ones. Bayesian offline learning therefore offers a concrete interpretation of adaptive policy priors whose value emerges through test-time interaction rather than immediate deployment.

The remainder of this paper is organized as follows. We first formalize adaptive offline reinforcement learning and define adaptive policy priors through memory, exploration, and self-correction. We then explain why these properties become increasingly important when offline learning is followed by test-time adaptation or online fine-tuning, especially under distribution shift, limited high-quality coverage, and changing deployment conditions. Next, we present Bayesian offline RL as a principled direction that naturally realizes adaptive behavior under epistemic uncertainty, while discussing alternative views that continue to prioritize conservatism or treat adaptation purely as an online RL problem. Finally, we outline broader connections, open challenges, and research directions for establishing adaptive offline RL as a practical learning paradigm.

2 Formulation of Adaptive Offline Reinforcement Learning

2.1 Offline RL Problem

We consider the standard offline RL setting (Levine et al., 2020) for a discrete-time, infinite-horizon, discounted-reward MDP defined by the tuple $\mathcal{M}^* = (\mathcal{S}, \mathcal{A}, \rho^*, P^*, R^*, \gamma)$.[3] Here, the state space $\mathcal{S}$, the action space $\mathcal{A}$, and the discount factor $\gamma \in (0, 1)$ are assumed known, while the environment components, including the initial state distribution $\rho^* \in \Delta(\mathcal{S})$, the transition function $P^* : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, and reward function $R^* : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$, are *unknown*.

Instead of interacting with $\mathcal{M}^*$ during training, the offline learner receives a static offline dataset $\mathcal{D} = \{\tau^i\}_{i=1}^N$ of trajectories collected by an unknown behavior policy β. Each trajectory $\tau = (s_0, a_0, r_1, s_1, a_1, r_2, \dots)$ is generated by

$$s_0 \sim \rho^*, \quad a_t \sim \beta(h_t), \quad s_{t+1} \sim P^*(s_t, a_t), \quad r_{t+1} \sim R^*(s_t, a_t), \quad \forall t \geq 0,$$

where the behavior policy may depend on the interaction history. We define the *history* at time t as

$$h_t := (s_0, a_0, r_1, s_1, \dots, a_{t-1}, r_t, s_t),$$

with $h_t \in \mathcal{H}_t$ and $\mathcal{H}_t$ denoting the corresponding history space.[4] Histories evolve recursively as $h_{t+1} = h_t \oplus (a_t, r_{t+1}, s_{t+1})$ for all $t \geq 0$, with $h_0 := (s_0)$ and $\oplus$ denoting sequence concatenation. In practice, trajectories are finite due to truncation by time limits, while infinite-horizon objectives are recovered through value bootstrapping.

The *ideal goal* for offline RL is to find a policy that maximizes expected discounted return under the true environment. In the most general form, this policy may depend on interaction history, denoted as $\pi : \mathcal{H}_t \to \Delta(\mathcal{A})$.

$$\max_{\pi} J(\pi; \mathcal{M}^*) := \mathbb{E}_\tau \left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, \pi, \mathcal{M}^* \right]. \tag{1}$$

Because $\mathcal{M}^*$ is inaccessible during offline optimization, this objective cannot be optimized directly. The key difficulty is epistemic uncertainty in state-action regions poorly covered by $\mathcal{D}$. We define the **behavioral support** of a state s within $\mathcal{D}$ as

$$\mathrm{supp}_{\mathcal{D}}(s) = \{a \mid \Pr_{\mathcal{D}}(a \mid s) > \epsilon\} \subseteq \mathcal{A},$$

where $\Pr_{\mathcal{D}}(\cdot \mid s)$ denotes the empirical action distribution in $\mathcal{D}$ and $\epsilon > 0$ is a small threshold. For states s not observed in the dataset, $\mathrm{supp}_{\mathcal{D}}(s) = \emptyset$. Epistemic uncertainty therefore remains substantial over actions in $\mathcal{A} \setminus \mathrm{supp}_{\mathcal{D}}(s)$ for each $s \in \mathcal{S}$. Different treatments of this uncertainty give rise to different algorithmic principles, including conservative offline RL (Levine et al., 2020), Bayesian offline RL (Ghosh et al., 2022), and optimistic offline RL (Agarwal et al., 2020).

[3] The same setting extends naturally to partially observable MDPs (POMDPs) (Kaelbling et al., 1998), where the policy observes o_t emitted from state s_t, and states in the history are replaced by observations.

[4] For applications where rewards are unavailable at test time, histories are defined without past rewards.

Offline training and online testing. Although offline RL is defined by the absence of online data collection during training, the learned policy is still evaluated through interaction with the true environment $\mathcal{M}^*$ at test (deployment) time. This creates a temporal separation between the offline stage, where parameters are optimized from a fixed dataset, and the online stage, where decisions are generated under the true environment. This distinction also separates offline RL from offline-to-online RL (Nair et al., 2020), where the policy updates its parameters in the online stage.

2.2 Core Components of Adaptive Policy Priors

Table 1: Comparison between classical offline RL and adaptive offline RL.

Property	Classical Offline RL	Adaptive Offline RL
Decision rule	Markovian	History-dependent
Uncertainty treatment	Within offline support	Preserve flexibility
Design focus	Offline stage	Offline + test-time stages
Core capabilities	Safety, robustness	Memory, exploration, self-correction

History dependence in decision making (memory). Classical offline RL typically assumes a Markovian decision rule, where actions are selected solely from the current state through a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ (Levine et al., 2020). The offline-learned Markovian policy remains fixed during deployment. In adaptive offline RL (AORL), by contrast, effective decision making should depend on online interaction history, because limited offline coverage leaves uncertainty about environment dynamics. This dependence can be implemented *explicitly* through a history-dependent policy $\pi : \mathcal{H}_t \to \Delta(\mathcal{A})$ (Ni et al., 2025), or *implicitly* through online adaptation of latent variables (Ghosh et al., 2022; Liu et al., 2023), plans (Janner et al., 2022), or even policy parameters (Xu et al., 2025). In this sense, history dependence serves as memory: the same state s may require different actions when preceding observations imply different beliefs about the environment. Memory is therefore the basic mechanism that allows online interaction to influence future decisions.

Behavioral flexibility (exploration and self-correction). Classical offline RL typically restricts learned actions to the behavioral support $\mathrm{supp}_{\mathcal{D}}(s)$ or heavily penalizes actions outside it (Levine et al., 2020). In AORL, by contrast, out-of-support actions are not treated as intrinsically undesirable: some may be potentially useful but remain unresolved under offline data alone because of epistemic uncertainty rather than evidence of poor value. Preserving non-negligible probability on such actions allows the policy to explore uncertain behaviors during interaction. Equally importantly, because these actions may fail, the policy should be able to self-correct by using online feedback to revise future decisions rather than committing to early mistakes (Wang et al., 2024; Kumar et al., 2025). Exploration and self-correction therefore form two *complementary* roles of behavioral flexibility, both of which rely on memory.

Co-design of offline optimization and test-time adaptation. Classical offline RL primarily designs algorithms within the offline optimization stage, treating test-time interaction as passive deployment (Levine et al., 2020). In AORL, by contrast, offline optimization and test-time adaptation are jointly considered: the offline-learned policy is evaluated by how well it supports memory, exploration, and self-correction under the true environment.

These components define the basic requirements for an **adaptive policy prior**: the policy should retain memory from interaction history while preserving sufficient flexibility to explore uncertain behaviors and correct them online. Table 1 summarizes the key differences between classical and adaptive offline RL. These requirements appear naturally in several forms of adaptation that occur after offline optimization, which we describe next.

2.3 Forms of Adaptation

Adaptation occurs after offline optimization, when an offline-learned policy is allowed to interact with the true environment $\mathcal{M}^*$ and improve with the resulting online interaction data, referred to as *experience* (Silver & Sutton, 2025). These adaptation mechanisms differ in whether policy parameters are updated and how much online data they require. Common forms are summarized in Table 2.

Test-time in-context learning. The simplest form of adaptation occurs when decision making is history-dependent, so that behavior changes directly through interaction history without modifying

Table 2: Common forms of adaptation after offline optimization.

Form of Adaptation	Online Data	Parameter Update?	Mechanism
Test-time in-context learning	Limited	No	Implicit inference
Test-time planning	Limited	No	Explicit inference
Test-time training	Limited	Yes	Temporary update
Online RL fine-tuning	Extended	Yes	Persistent update

policy parameters or performing additional optimization. This mechanism is commonly referred to as *in-context learning* (Brown et al., 2020; Moeini et al., 2025), and has often been interpreted as implicit Bayesian inference over latent tasks (Xie et al., 2022). It differs from *in-weight learning*, where adaptation occurs through parameter updates. A special case of in-context learning is *self-improvement* (Shinn et al., 2023), where explicit reward signals may be unavailable at test time, yet the agent can still improve through informative observations that reduce uncertainty about the environment. Self-improvement includes settings where the uncertainty lies in transition dynamics rather than reward function (Packer et al., 2018), as well as settings where feedback is conveyed through language observations (Cheng et al., 2024; Klissarov et al., 2026).

Test-time planning. In-context learning relies primarily on the generalization ability of the offline-learned policy, which may be insufficient when the true environment $\mathcal{M}^*$ differs substantially from what is specified by $\mathcal{D}$. Test-time planning addresses this limitation by improving policy decisions with online search over future trajectories, often using learned world models (Argenson & Dulac-Arnold, 2020; Zhou et al., 2025; Wei et al., 2025). Recent planning approaches further enable trajectory optimization through probabilistic inference without requiring an explicit world model (Janner et al., 2022; Ajay et al., 2023).

Test-time training. Adaptation may also occur through parameter updates using a limited amount of online data collected at test time (Finn et al., 2017; Park et al., 2024; Xu et al., 2025). Unlike in-context learning or planning, test-time training modifies the policy itself rather than only changing how it is queried, often through lightweight or localized updates designed for rapid adaptation. Although this lies outside the classical offline RL formulation, it is relevant under AORL because the offline-learned policy is viewed as a prior that should support efficient improvement from limited online experience.

Online RL fine-tuning. A more extended form of adaptation is online RL fine-tuning, where the offline-learned policy serves as an initialization for continued RL with substantial online interaction. This includes offline-to-online RL (Nair et al., 2020; Lee et al., 2022b) and RL post-training of foundation models (Guo et al., 2025a). Compared with test-time training, the objective is no longer rapid local adaptation, but sustained policy improvement through accumulating experience.

3 Why Offline RL Needs Adaptive Policy Priors?

The previous section defined adaptive policy priors through memory and behavioral flexibility. We now ask why these properties are needed beyond the classical offline RL objective. The key reason is that uncertainty often persists beyond offline optimization, appearing in online environments through distribution shift, limited data coverage, exploration bottlenecks, or changing conditions.

Inevitable distribution shift. Classical offline RL reduces risk by keeping actions within behavioral support, yet out-of-distribution (OOD) states may still arise during deployment because small errors accumulate over time. This is well known in imitation learning, where compounding errors induce *covariate shift* (Ross & Bagnell, 2010; Filos et al., 2020; Spencer et al., 2021), and remains a practical bottleneck in offline RL (Park et al., 2024). Since such OOD states are rarely trained explicitly under conservative objectives, the learned policy may become brittle when uncertainty matters most. Adaptability helps address this by preparing the offline-learned policy to use memory and online feedback to **self-correct** once OOD states are encountered. In this sense, adaptability does not conflict with safety: adaptive policies can be designed to remain risk-averse while still using online evidence to revise decisions under uncertainty (Rigter et al., 2023).

Limited high-quality coverage. Classical offline RL effectively bounds policy improvement by the quality of behaviors represented in the dataset (Jin et al., 2021; Uehara & Sun, 2022). This

limitation becomes pronounced when high-quality actions are absent or underrepresented (Ni et al., 2025). In practice, collecting high-quality trajectories is expensive, and in open-ended domains even carefully curated datasets may remain incomplete. This aligns with the broader observation that human-generated data alone imposes a ceiling on improvement (Silver & Sutton, 2025). Adaptability mitigates this limitation by preserving behavioral flexibility, allowing subsequent interaction to **explore** uncertain actions and reinforce those that prove beneficial.

Exploration bottleneck in online RL fine-tuning. A standard way to overcome the limited adaptability of offline RL is to continue learning through online RL fine-tuning (Nair et al., 2020). While this relaxes strict conservatism, exploration remains difficult. In classical offline-to-online RL, online improvement is often slow and incremental (Luo et al., 2023; Zhao et al., 2023). In foundation model post-training, recent evidence further suggests that RL often mainly reweights behaviors already present in the policy prior rather than reliably discovering new ones (Yue et al., 2025; Zhao et al., 2025). As a result, behavioral support may contract further during fine-tuning, as illustrated in Fig. 2, making underrepresented but potentially good actions increasingly difficult to recover. Although stronger exploration strategies or prolonged fine-tuning can alleviate this effect (Liu et al., 2025), a complementary and often cheaper solution is to prepare a **behaviorally flexible** policy prior that already preserves diverse modes beyond well-supported actions in the offline dataset.

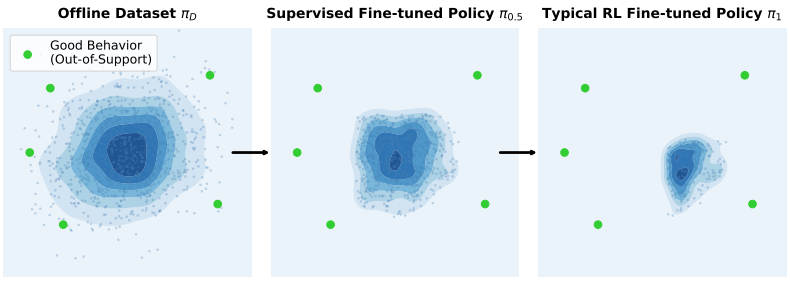

Figure 2: Behavioral support may shrink across offline (here, SFT) and online fine-tuning stages, making underrepresented but possibly good actions increasingly difficult to recover (Yue et al., 2025).

Test-time condition change. Standard offline RL typically assumes that the deployment environment matches the one that generated the offline dataset. In realistic open-ended domains, however, this assumption is often violated: dynamics, rewards, or task structure may shift after offline training. Prior work has shown that offline-learned policies can become brittle under such changes, even when test-time modifications are specified by environment parameters (Liang et al., 2023), natural-language instructions (Karthikeyan & Pant, 2025), or external datasets (Lyu et al., 2024). Adaptability helps address this limitation by enabling policies to use online interaction to detect changed conditions, explore alternative behaviors, and correct decisions as new evidence accumulates.

4 A Bayesian Perspective for Adaptive Policy Priors

A natural question is whether adaptive policy priors admit a principled computational interpretation. One promising direction arises from Bayesian principles in offline RL (Ghosh et al., 2022), building on Bayes-adaptive MDPs (Duff, 2002), whose practical potential has recently been demonstrated (Choi et al., 2024; Ni et al., 2025). Rather than collapsing uncertainty conservatively or treating unseen actions optimistically, Bayesian offline RL maintains multiple plausible environment hypotheses consistent with the offline dataset, as illustrated in Fig. 3.

Bayesian formulation of offline RL. In Bayesian model-based offline RL, the agent maintains epistemic uncertainty over environments consistent with the offline dataset $\mathcal{D}$. A prior distribution $\Pr(\mathcal{M})$ over MDPs induces a posterior $\Pr(\mathcal{M} \mid \mathcal{D})$ after observing the dataset, where each MDP $\mathcal{M}$ is specified by environment parameters (ρ, P, R). The offline objective becomes

$$\max_{\pi} \mathbb{E}_{\mathcal{M}\sim\Pr(\mathcal{M}|\mathcal{D})}[J(\pi;\mathcal{M})]. \tag{2}$$

Because each sampled MDP $\mathcal{M}$ is unknown to the agent, decision making becomes a partially observable problem over environment hypotheses, often called an *epistemic POMDP* (Ghosh et al., 2021, 2022). The optimal policy is therefore naturally history-dependent: it maintains Bellman

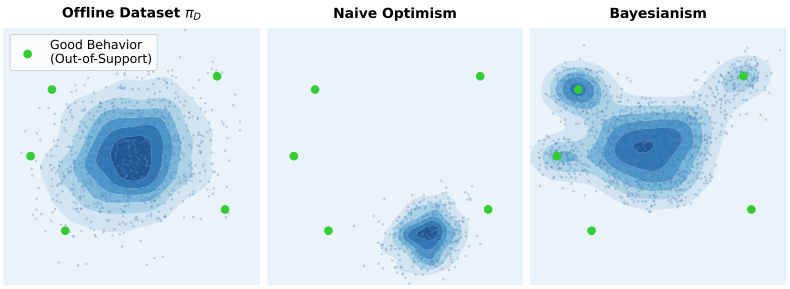

Figure 3: Bayesian offline RL treats unseen actions as uncertain rather than inherently bad or reliably good. In contrast, naive optimism in model-free off-policy RL may assign unreliable values to out-of-support actions due to extrapolation error (Fujimoto et al., 2019).

consistency under the posterior over plausible MDPs. By contrast, model-free offline RL may assign unreliable values to out-of-support actions without an explicit mechanism for Bellman consistency.

Connection to adaptive policy priors. This formulation directly explains the ingredients of adaptive policy priors. Online history h_t provides *memory* about the environment; uncertain actions induce *exploration*; and newly observed evidence supports *self-correction*. Formally, online interaction further updates the environment posterior:

$$\underbrace{\Pr(\mathcal{M} \mid \mathcal{D}, h_t)}_{\text{Online Posterior}} \propto \underbrace{\Pr(\mathcal{M} \mid \mathcal{D})}_{\text{Offline Posterior}} \underbrace{\Pr(h_t \mid \mathcal{M}, \mathcal{D})}_{\text{Online Likelihood}}, \tag{3}$$

where the offline posterior $\Pr(\mathcal{M} \mid \mathcal{D})$ serves as the effective *prior* for online interaction. Under this interpretation, adaptability is not an auxiliary property added after offline learning, but a direct consequence of preserving epistemic uncertainty during offline optimization.

5 Alternative Views

Conservatism as the first principle. A dominant alternative view in offline RL is that conservatism should remain the primary design principle because robustness and safety are central objectives. This is well justified when policies are intended for direct deployment in high-stakes settings, where uncertain actions may incur unacceptable cost. Such arguments are typically developed under the classical offline RL assumption of Markovian, memoryless policies. Our position differs when offline learning is followed by further interaction: if the policy is expected to continue improving online, uncertainty need not be fully eliminated during the offline stage, and adaptability should therefore be prioritized. Importantly, *adaptability does not preclude a light degree of conservatism*; rather, it requires sufficient memory and behavioral flexibility so that the policy can revise decisions through interaction under uncertainty. The key distinction is whether offline learning aims to produce a final deployable policy or an improvable policy prior.

Adaptation should be left to online RL, not offline RL. Another alternative view is that offline RL need not itself produce an adaptive prior; its role is simply to provide a stable initialization for subsequent online RL, while adaptation is delegated to downstream fine-tuning. Under this perspective, one may even blur the distinction between offline and online learning by directly training online RL from scratch with offline data (Song et al., 2023; Ball et al., 2023). This is reasonable in settings where online interaction is abundant, but incomplete when online experience is limited or costly. If offline training collapses the policy too narrowly—for example through behavioral cloning or strongly conservative objectives—online RL may inherit a restricted search space and struggle to recover underrepresented but valuable behaviors (Yue et al., 2025). Finally, when offline learning is intended to produce *reusable* priors rather than task-specific warm starts, its objective should therefore extend beyond fast initialization to preserving behavioral flexibility for later adaptation.

Adaptability already appears in many offline RL methods. Another alternative view is that offline RL already contains mechanisms associated with adaptation, making an explicit shift in objective unnecessary. Common history-dependent policies with sequence modeling (Chen et al., 2021) use past context during action generation, yet typically remain constrained by patterns learned from the

offline dataset. Test-time planning methods (Argenson & Dulac-Arnold, 2020; Janner et al., 2022) improve decisions through online search, but the candidate trajectories they consider are usually still constrained by offline support. Similarly, *stitching* (Fu et al., 2020) enables compositional generalization from offline data, but this remains a passive recombination of in-distribution behaviors rather than adaptation driven by new online evidence. These examples suggest that adaptive ingredients are increasingly present in offline RL, but common formulations still do not explicitly preserve exploration and self-correction under later interaction.

6 Connections Beyond Offline RL

Adaptive offline RL also connects naturally to neighboring research areas that study adaptation under different assumptions. These connections help clarify what is distinctive about AORL while suggesting methodological directions that may accelerate progress.

Meta-RL. The Bayesian perspective in Sec. 4 is closely related to contextual MDPs (Hallak et al., 2015) and meta-RL (Duan et al., 2016; Wang et al., 2017; Beck et al., 2025), where policies are trained across a distribution of tasks so that adaptation emerges at test time. The key difference is that in meta-RL the task distribution is pre-specified before training, whereas in Bayesian offline RL the distribution over environments must be self-constructed from offline data. Nevertheless, both settings give rise to similar adaptive behavior at test time, including online exploration and self-correction under uncertainty.

Plasticity in continual RL. The ability to improve from subsequent experience is closely related to *plasticity* in continual RL (Abbas et al., 2023; Klein et al., 2024), and more recently in offline-to-online RL (Li et al., 2025b). AORL differs in that adaptation often occurs without parameter updates—for example through in-context learning, also known as *in-context plasticity* (Klissarov et al., 2026)—and typically assumes a fixed underlying environment rather than changing tasks. Still, the central concern is similar: whether prior learning preserves the capacity to improve when new experience becomes available.

7 Open Challenges and Research Directions

Benchmarks and evaluation for adaptability. Existing offline RL benchmarks evaluate many important settings, including high-dimensional observations (Gulcehre et al., 2020; Lu et al., 2023), stochasticity (Qin et al., 2022), sparse rewards and stitching (Fu et al., 2020; Park et al., 2025a), and safety (Liu et al., 2024). However, they rarely evaluate whether an offline-learned policy can *improve through subsequent interaction*. Prior work on adaptability has so far focused either on bandit-like settings (Ghosh et al., 2022; Ni et al., 2025) or illustrative task modifications (Liang et al., 2023; Zhou et al., 2025; Karthikeyan & Pant, 2025). A notable exception is the off-dynamics RL benchmark (Lyu et al., 2024), where test-time condition changes are introduced through externally provided datasets. If AORL is to become a meaningful research direction, progress will require benchmarks that measure the capacity for test-time adaptation.

On the offline side, dataset design should include at least two cases: *underrepresented but valuable actions*, where useful behaviors appear only sparsely in the data, and *unseen but plausible actions*, where beneficial behaviors are absent altogether and must be discovered through later interaction. On the online side, evaluation should go beyond a single fixed test environment and instead consider *a distribution of environments*: (1) environments fully aligned with the offline data, and (2) environments with test-time condition changes under full or partial specification.

More broadly, AORL benchmarks should treat *multi-episode interaction* as the basic unit of training and evaluation, akin to meta-RL setups (Duan et al., 2016), in order to capture the improvable capacity of adaptive agents. Evaluation should measure not only final returns, but also *how* improvement unfolds through interaction, including adaptation speed, robustness to early mistakes, recovery after failed exploration, and sensitivity to test-time horizon length (Sentenac et al., 2025).

Theory of adaptive offline RL. Classical offline RL theory typically relies on coverage assumptions over good actions and Markovian policies. Extending these guarantees to history-dependent policies that can discover unseen but valuable actions through online interaction remains largely open. A

central question is *which classes* of offline datasets favor adaptation rather than conservatism, and how guarantees should change under test-time condition shift.

Memory-based RL. Memory remains underexplored in offline RL, with the most notable progress such as the Decision Transformer family (Chen et al., 2021), which still operate under conservative objectives. We believe that advances in understanding memory (Ni et al., 2023; Cherepanov et al., 2026), learning history representations (Ni et al., 2024), and scaling memory-based RL (Grigsby et al., 2024) in online POMDPs could help AORL, especially by understanding how in-context learning emerges from offline-trained policies.

Overcoming value overestimation. A major challenge in AORL is that classical *value overestimation* re-emerges once conservatism is relaxed (Fujimoto et al., 2019; Kumar et al., 2019; Sims et al., 2024). Because adaptive priors must preserve uncertain actions for later exploration, avoiding overestimation *without* collapsing back to conservative behavior becomes a central algorithmic difficulty. Recent evidence from Bayesian offline RL suggests that sufficiently long-horizon rollouts can mitigate this problem (Ni et al., 2025). This points to a broader direction: scaling long-horizon planning in model-based offline RL, potentially informed by recent progress in long-horizon world modeling (Li et al., 2025a; Lin et al., 2026).

Scalable Bayesian inference. A practical bottleneck of the Bayesian direction in Sec. 4 is that exact posterior inference over MDPs is typically intractable. In practice, uncertainty is often approximated through model ensembles, where disagreement serves as a proxy for epistemic uncertainty (Yu et al., 2020). Improving the scalability of such approximations, while achieving more reliable uncertainty quantification (He et al., 2026), remains an important direction for AORL.

Beyond Bayesian model-based direction. Although we emphasize the Bayesian model-based direction in Sec. 4 because it offers a principled interpretation of adaptive policy priors, simpler alternatives may also achieve adaptive behavior. These include model-free approaches (Hu et al., 2024; Wagenmaker et al., 2025) as well as non-Bayesian approaches. Understanding when such methods recover similar adaptive mechanisms, and when they fundamentally differ from model-based formulations, remains an important open direction.

Foundation models for AORL. Foundation models may provide useful priors for adaptive offline RL by supplying pretrained knowledge beyond the offline dataset itself. For example, world foundation models could be adapted to offline model-based RL to support longer-horizon planning. Likewise, pretrained history-dependent policies, including large language models and vision-language-action models, may serve as strong initial policies that can be further fine-tuned.

AORL for foundation models. Recent progress in foundation models already highlights the importance of test-time adaptation and continual learning (Wei et al., 2025; Snell et al., 2025). Yet offline post-training remains dominated by supervised fine-tuning (SFT) (Ouyang et al., 2022), which may narrow behavioral support and hinder later test-time adaptation or online RL fine-tuning (Zhang et al., 2026a). Bayesian directions may therefore offer a useful path for AORL in foundation models: model-based formulations can preserve broader support through synthetic on-policy rollouts (Chen et al., 2025), while Bayesian formulations can promote in-context adaptation (Zhang et al., 2026b).

8 Conclusion

This paper argues that when policies are expected to continue improving through interaction, offline RL should prioritize learning *adaptive policy priors* rather than fully deployable policies. Under this view, the objective of offline learning is not only immediate robustness under limited support, but also preserving the capacity for memory, exploration, and self-correction once new experience becomes available. Bayesian offline RL provides one principled direction for this perspective, though many algorithmic and theoretical questions remain open.

More broadly, we believe offline RL should gradually move beyond a supervised-learning view shaped by data scaling laws (Kaplan et al., 2020), toward a continual-learning view aligned with the emerging era of experience (Silver & Sutton, 2025). In this setting, offline learning should be judged not only by deployment quality, but by the adaptability of the policies it prepares for future experience.

References

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C Machado. Loss of plasticity in continual deep reinforcement learning. In *Conference on Lifelong Learning Agents*, 2023. 8

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020. 3

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision making? In *International Conference on Learning Representations*, 2023. 5

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in neural information processing systems*, 34:7436–7447, 2021. 1

Arthur Argenson and Gabriel Dulac-Arnold. Model-based offline planning. In *International Conference on Learning Representations*, 2020. 5, 8

Chenjia Bai, Lingxiao Wang, Zhuoran Yang, Zhi-Hong Deng, Animesh Garg, Peng Liu, and Zhaoran Wang. Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. In *International Conference on Learning Representations*, 2022. 1

Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pp. 1577–1594. PMLR, 2023. 7

Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A tutorial on meta-reinforcement learning. *Foundations and Trends in Machine Learning*, 18(2-3):224–384, 2025. 8

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020. 1, 2, 5

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*, 2025. 9

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, 2021. 7, 9

Ching-An Cheng, Andrey Kolobov, Dipendra Misra, Allen Nie, and Adith Swaminathan. LLF-bench: Benchmark for interactive learning from language feedback. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024. 5

Egor Cherepanov, Nikita Kachaev, Artem Zholus, Alexey Kovalev, and Aleksandr Panov. Unraveling the complexity of memory in RL agents: an approach for classification and evaluation. In *International Conference on Learning Representations*, 2026. 9

Yunseon Choi, Li Zhao, Chuheng Zhang, Lei Song, Jiang Bian, and Kee-Eung Kim. Diversification of adaptive policy for effective offline reinforcement learning. In *International Joint Conference on Artificial Intelligence*, pp. 3863–3871, 2024. 6

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016. 8

Michael O'Gordon Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002. 6

Angelos Filos, Panagiotis Tigkas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *International Conference on Machine Learning*, 2020. 5

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017. 5

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020. 8

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021. 1

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019. 1, 7, 9

Dibya Ghosh, Jad Rahme, Aviral Kumar, Amy Zhang, Ryan P. Adams, and Sergey Levine. Why generalization in RL is difficult: Epistemic pomdps and implicit partial observability. In *Advances in Neural Information Processing Systems*, 2021. 2, 6

Dibya Ghosh, Anurag Ajay, Pulkit Agrawal, and Sergey Levine. Offline rl policies should be trained to be adaptive. In *International Conference on Machine Learning*, pp. 7513–7530. PMLR, 2022. 2, 3, 4, 6, 8

Jake Grigsby, Linxi Fan, and Yuke Zhu. Amago: Scalable in-context reinforcement learning for adaptive agents. In *International Conference on Learning Representations*, 2024. 9

Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Thomas Paine, Sergio Gómez, Konrad Zolna, Rishabh Agarwal, Josh S Merel, Daniel J Mankowitz, Cosmin Paduraru, et al. Rl unplugged: A suite of benchmarks for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:7248–7259, 2020. 8

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a. 2, 5

Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. Improving vision-language-action model with online reinforcement learning. In *International Conference on Robotics and Automation*, 2025b. 2

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015. 8

Wenchong He, Zhe Jiang, Tingsong Xiao, Zelin Xu, and Yukun Li. A survey on uncertainty quantification methods for deep learning. *ACM Computing Surveys*, 58(7):1–35, 2026. 9

Hao Hu, Yiqin Yang, Jianing Ye, Chengjie Wu, Ziqing Mai, Yujing Hu, Tangjie Lv, Changjie Fan, Qianchuan Zhao, and Chongjie Zhang. Bayesian design principles for offline-to-online reinforcement learning. In *International Conference on Machine Learning*, pp. 19491–19515. PMLR, 2024. 9

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, pp. 9902–9915. PMLR, 2022. 4, 5, 8

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021. 5

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 1998. 3

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 9

Akash Karthikeyan and Yash Vardhan Pant. Genplan: Generative sequence models as adaptive planners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 6, 8

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 2022. 2

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823, 2020. 1

Timo Klein, Lukas Miklautz, Kevin Sidak, Claudia Plant, and Sebastian Tschiatschek. Plasticity loss in deep reinforcement learning: A survey. *arXiv preprint arXiv:2411.04832*, 2024. 8

Martin Klissarov, Jonathan Cook, Diego Antognini, Hao Sun, Jingling Li, Natasha Jaques, Claudiu Musat, and Edward Grefenstette. Improving interactive in-context learning from natural language feedback. *arXiv preprint arXiv:2602.16066*, 2026. 5, 8

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022. 1

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 9

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020. 1

Aviral Kumar, Rishabh Agarwal, Xinyang Geng, George Tucker, and Sergey Levine. Offline q-learning on diverse multi-task data both scales and generalizes. In *International Conference on Learning Representations*, 2023. 1

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, 2025. 4

Kuang-Huei Lee, Ofir Nachum, Mengjiao Yang, Lisa Lee, Daniel Freeman, Sergio Guadarrama, Ian Fischer, Winnie Xu, Eric Jang, Henryk Michalewski, and Igor Mordatch. Multi-game decision transformers. In *Advances in Neural Information Processing Systems*, 2022a. 1

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*, pp. 1702–1712. PMLR, 2022b. 5

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. 1, 3, 4

Chenhao Li, Andreas Krause, and Marco Hutter. Uncertainty-aware robotic world model makes offline model-based reinforcement learning work on real robots. *arXiv preprint arXiv:2504.16680*, 2025a. 9

Lu Li, Tianwei Ni, Yihao Sun, and Pierre-Luc Bacon. The three regimes of offline-to-online reinforcement learning. *arXiv preprint arXiv:2510.01460*, 2025b. 8

Zhixuan Liang, Yao Mu, Mingyu Ding, Fei Ni, Masayoshi Tomizuka, and Ping Luo. Adaptdiffuser: diffusion models as adaptive self-evolving planners. In *International Conference on Machine Learning*, 2023. 6, 8

Haoxin Lin, Siyuan Xiao, Yi-Chen Li, Zhilong Zhang, Yihao Sun, Chengxing Jia, and Yang Yu. ADM-v2: Pursuing full-horizon roll-out in dynamics models for offline policy learning and evaluation. In *International Conference on Learning Representations*, 2026. 9

Jinxin Liu, Hongyin Zhang, Zifeng Zhuang, Yachen Kang, Donglin Wang, and Bin Wang. Design from policies: Conservative test-time adaptation for offline policy optimization. *Advances in Neural Information Processing Systems*, 2023. 4

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. In *Conference on Neural Information Processing Systems*, 2025. 6

Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, and Ding Zhao. Datasets and benchmarks for offline safe reinforcement learning. *Journal of Data-centric Machine Learning Research*, 2024. 8

Cong Lu, Philip J. Ball, Tim G. J. Rudner, Jack Parker-Holder, Michael A Osborne, and Yee Whye Teh. Challenges and opportunities in offline reinforcement learning from visual observations. *Transactions on Machine Learning Research*, 2023. 8

Yicheng Luo, Jackie Kay, Edward Grefenstette, and Marc Peter Deisenroth. Finetuning from offline reinforcement learning: Challenges, trade-offs and practical solutions. *arXiv preprint arXiv:2303.17396*, 2023. 6

Jiafei Lyu, Kang Xu, Jiacheng Xu, Jing-Wen Yang, Zongzhang Zhang, Chenjia Bai, Zongqing Lu, Xiu Li, et al. Odrl: A benchmark for off-dynamics reinforcement learning. *Advances in Neural Information Processing Systems*, 2024. 6, 8

Amir Moeini, Jiuqi Wang, Jacob Beck, Ethan Blaser, Shimon Whiteson, Rohan Chandra, and Shangtong Zhang. A survey of in-context reinforcement learning. *arXiv preprint arXiv:2502.07978*, 2025. 5

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020. 2, 4, 5, 6

Tianwei Ni, Michel Ma, Benjamin Eysenbach, and Pierre-Luc Bacon. When do transformers shine in RL? decoupling memory from credit assignment. In *Advances in Neural Information Processing Systems*, 2023. 9

Tianwei Ni, Benjamin Eysenbach, Erfan SeyedSalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging state and history representations: Understanding self-predictive rl. In *International Conference on Learning Representations*, 2024. 9

Tianwei Ni, Esther Derman, Vineet Jain, Vincent Taboga, Siamak Ravanbakhsh, and Pierre-Luc Bacon. Long-horizon model-based offline reinforcement learning without conservatism. *arXiv preprint arXiv:2512.04341*, 2025. 2, 4, 6, 8, 9

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. 9

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Song. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018. 5

Seohong Park, Kevin Frans, Sergey Levine, and Aviral Kumar. Is value learning really the main bottleneck in offline rl? *Advances in Neural Information Processing Systems*, 2024. 5

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations*, 2025a. 8

Seohong Park, Kevin Frans, Deepinder Mann, Benjamin Eysenbach, Aviral Kumar, and Sergey Levine. Horizon reduction makes RL scalable. In *Conference on Neural Information Processing Systems*, 2025b. 1

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019. 1

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in Neural Information Processing Systems*, 1, 1988. 1

Rong-Jun Qin, Xingyuan Zhang, Songyi Gao, Xiong-Hui Chen, Zewen Li, Weinan Zhang, and Yang Yu. Neorl: A near real-world benchmark for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:24753–24765, 2022. 8

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022. Featured Certification, Outstanding Certification. 1

Marc Rigter, Bruno Lacerda, and Nick Hawes. One risk to rule them all: A risk-sensitive perspective on model-based offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023. 5

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010. 5

Flore Sentenac, Ilbin Lee, and Csaba Szepesvari. Balancing optimism and pessimism in offline-to-online learning. *arXiv preprint arXiv:2502.08259*, 2025. 8

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 2023. 5

David Silver and Richard S Sutton. Welcome to the era of experience. *preprint*, 2025. 2, 4, 6, 9

Anya Sims, Cong Lu, Jakob N Foerster, and Yee W Teh. The edge-of-reach problem in offline model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 37:63029–63056, 2024. 9

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. 2, 9

Yuda Song, Yifei Zhou, Ayush Sekhari, Drew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid RL: Using both offline and online data can make RL efficient. In *International Conference on Learning Representations*, 2023. 7

Jonathan Spencer, Sanjiban Choudhury, Arun Venkatraman, Brian Ziebart, and J Andrew Bagnell. Feedback in imitation learning: The three regimes of covariate shift. *arXiv preprint arXiv:2102.02872*, 2021. 5

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022. 5

Andrew Wagenmaker, Perry Dong, Raymond Tsao, Chelsea Finn, and Sergey Levine. Posterior behavioral cloning: Pretraining bc policies for efficient rl finetuning. *arXiv preprint arXiv:2512.16911*, 2025. 9

Jane Wang, Zeb Kurth-Nelson, Hubert Soyer, Joel Z. Leibo, Dhruva Tirumala, Rémi Munos, Charles Blundell, Dharshan Kumaran, and Matt M. Botvinick. Learning to reinforcement learn. In *Annual Meeting of the Cognitive Science Society*, 2017. 8

Yifei Wang, Yuyang Wu, Zeming Wei, Stefanie Jegelka, and Yisen Wang. A theoretical understanding of self-correction through in-context alignment. *Advances in Neural Information Processing Systems*, 2024. 4

Hui Wei, Zihao Zhang, Shenghua He, Tian Xia, Shijia Pan, and Fei Liu. Plangenllms: A modern survey of llm planning capabilities. In *Annual Meeting of the Association for Computational Linguistics*, pp. 19497–19521, 2025. 5, 9

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019. 1

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. 5

Shoukai Xu, Mingkui Tan, Liu Liu, Zhong Zhang, Peilin Zhao, et al. Test-time adapted reinforcement learning with action entropy regularization. In *International Conference on Machine Learning*, 2025. 4, 5

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 2023. 2

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020. 1, 9

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Conference on Neural Information Processing Systems*, 2025. 6, 7

Dylan Zhang, Yufeng Xu, Haojin Wang, Qingzhi Chen, and Hao Peng. Good sft optimizes for sft, better sft prepares for reinforcement learning. *arXiv preprint arXiv:2602.01058*, 2026a. 9

Shenao Zhang, Yaqing Wang, Yinxiao Liu, Tianqi Liu, Peter Grabowski, Eugene Ie, Zhaoran Wang, and Yunxuan Li. Beyond markovian: Reflective exploration via bayes-adaptive RL for LLM reasoning. In *International Conference on Learning Representations*, 2026b. 9

Kai Zhao, Jianye Hao, Yi Ma, Jinyi Liu, Yan Zheng, and Zhaopeng Meng. Enoto: Improving offline-to-online reinforcement learning with q-ensembles. *arXiv preprint arXiv:2306.06871*, 2023. 6

Rosie Zhao, Alexandru Meterez, Sham M. Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. Echo chamber: RL post-training amplifies behaviors learned in pretraining. In *Conference on Language Modeling*, 2025. 6

Guangyao Zhou, Sivaramakrishnan Swaminathan, Rajkumar Vasudeva Raju, J Swaroop Guntupalli, Wolfgang Lehrach, Joseph Ortiz, Antoine Dedieu, Miguel Lázaro-Gredilla, and Kevin Murphy. Diffusion model predictive control. *Transactions on Machine Learning Research*, 2025. 5, 8