In-Context Reinforcement Learning under Non-Stationarity: A Survey

Anonymous authors
Paper under double-blind review

Abstract

The development of decision-pretrained transformers, algorithm distillation, long-context meta-RL, and retrieval-augmented agents has renewed interest in in-context reinforcement learning (ICRL): the ability of a pretrained or fine-tuned decision model to infer latent task rules and improve future behavior from interaction context, without test-time parameter updates. This line of work asks when trial-and-error evidence, rewards, transitions, demonstrations, feedback, or retrieved experience can make learning-like computation happen inside the context window. However, existing surveys of ICRL mainly organize the field around pretraining objectives, architectures, context formats, evaluation protocols, and theoretical mechanisms, while the non-stationary setting remains comparatively underexamined. In changing environments, accumulated context is not merely more evidence about a fixed task: the reward specification, transition kernel, observation channel, action interface, constraint model, or demonstration and memory distribution can fall out of alignment with the current regime. Previously useful context can therefore become stale, misleading, or useful again when an old regime returns. We survey non-stationary ICRL as the problem of adapting through context while deployed policy parameters remain fixed: the policy must infer both the current decision rule and which parts of its accumulated evidence still support that rule. We define non-stationary ICRL, relate it to meta-RL, decision sequence modeling, retrieval-augmented RL, value- and model-aware ICRL, and reward-feedback agents, and organize the literature along three questions: what changes, how the change unfolds, and how observable the change is to the agent. We further analyze methods through context-management operations and influence facets, including what is written, retrieved, compressed, trusted, forgotten, or isolated, and argue that average return and held-out task generalization are insufficient to establish adaptation under drift. The survey identifies a research agenda around lifecycle evaluation, stale-context stress tests, validity-aware retrieval, adaptive forgetting, context poisoning, and finite-context theory for non-stationary ICRL.

1 Introduction

In-context reinforcement learning (ICRL) asks whether a pretrained or fine-tuned sequential decision model can infer task rules and improve future actions from the interaction history available in its context, without test-time parameter updates (Duan et al., 2016). The context may contain states, actions, rewards, transition outcomes, demonstrations, feedback, retrieved memories, or summaries, but the defining property is learning-like adaptation through that context rather than through an online optimizer (Laskin et al., 2023; Zisman et al., 2024). That idea runs through black-box meta-reinforcement learning (Duan et al., 2016; Wang et al., 2016), gradient-based meta-learning (Finn et al., 2017), offline trajectory modeling (Janner et al., 2021), decision transformers (Chen et al., 2021), algorithm distillation (Laskin et al., 2023), supervised ICRL (Lee et al., 2023), and long-context meta-RL (Melo, 2022; Grigsby et al., 2024a). It also inherits a broader question from in-context learning: when does a sequence model merely condition on examples (Brown et al., 2020; Min et al., 2022), and when does its forward pass implement something close to inference (Chan et al., 2022; Garg et al., 2022; Akyurek et al., 2023), optimization (von Oswald et al., 2023; Kirsch et al., 2022), or learning (Wies et al., 2023; Akyurek et al., 2024)? This capability is especially appealing in reinforcement

learning. If useful adaptation can happen through the input stream, then a system can respond quickly without running an optimizer, changing deployed weights, or hiding the adaptation process inside an online update.

That promise is much easier to state in stationary settings than in changing ones. Recent survey work already organizes ICRL broadly around pretraining, context construction, architectures, evaluation, and theory (Moeini et al., 2025b). Here we focus on the case where the environment changes during the lifetime over which context is accumulated. By non-stationarity, we mean temporal changes in the decision process that alter which accumulated evidence remains valid for action. In a stationary few-shot task, more context is usually more evidence about one latent task. Under non-stationarity, the same extra trajectory can be useful evidence, obsolete evidence, a misleading demonstration, or a memory that becomes useful again only after a previous regime returns. The question is no longer simply whether the agent uses context. It is whether the agent can tell which parts of its context still describe the world it is acting in.

For ICRL, non-stationarity is best viewed as a change in the validity conditions of context: past trajectories, rewards, feedback, demonstrations, or retrieved memories may no longer license the same policy choice. The underlying source can be a changed reward specification (Besbes et al., 2014; Song et al., 2025), transition kernel (Lecarpentier & Rachelson, 2019; Chandak et al., 2020c), observation channel (Cobbe et al., 2020; Wang et al., 2025e), action interface (Chandak et al., 2020b; Sinii et al., 2024), preference or safety model (Moeini et al., 2025a; Song et al., 2025), task distribution (Al-Shedivat et al., 2017; Wang et al., 2025a), or behavior policy behind the data (Dong et al., 2026; Schmied et al., 2024). These sources impose different adaptation burdens. Reward switches require reinterpretation of feedback, transition drift changes action–outcome evidence, action remapping changes the meaning of behavior, and behavior-policy shift changes the trustworthiness of demonstrations or retrieved trajectories. An abrupt reward switch creates a different adaptation problem (Adams & MacKay, 2007; Garivier & Moulines, 2011) from gradual transition drift (Lecarpentier & Rachelson, 2019; Feng et al., 2023); a recurring latent mode creates a different memory problem (Hallak et al., 2015; Killian et al., 2017) from one-way degradation (Khetarpal et al., 2022; Abel et al., 2023); an observed task descriptor creates a different inference problem (Hallak et al., 2015; Humplik et al., 2019) from a hidden change point (Adams & MacKay, 2007; Keplinger et al., 2025). For ICRL, the relevant consequence is that context has a lifetime. It can be useful when it identifies the current mode, harmful when it anchors the agent to a past mode, and useful again when an old mode recurs. This framing draws on older Markov decision process (Puterman, 1994), partially observable decision process (Kaelbling et al., 1998), contextual MDP (Hallak et al., 2015), hidden-parameter MDP (Killian et al., 2017; Doshi-Velez & Konidaris, 2016), and epistemic-POMDP (Ghosh et al., 2021) formalisms, but shifts the emphasis from solving a known model class to managing evidence inside a fixed context window.

We use ***Non-Stationary ICRL*** for fixed-parameter, context-conditioned sequential decision making in which the decision process changes during the policy's context-bearing lifetime. The deployed policy must use recent and stored interaction history to detect regime evidence, revise its latent task belief, and change its action distribution without test-time parameter updates. The central difficulty is not only selecting the policy appropriate for the current regime, but also deciding which previous rewards, transitions, demonstrations, feedback, or memories still provide valid evidence for that choice. It covers algorithm distillation (Laskin et al., 2023) and decision-pretrained transformers (Furuta et al., 2022; Liu & Abbeel, 2023; Polubarov et al., 2026) when interaction history drives adaptation; supervised ICRL (Lee et al., 2023; Lin et al., 2024) and offline-pretrained agents (Tarasov et al., 2025; Wang et al., 2024a) when task examples, rewards, or trajectories identify the current regime; retrieval-augmented ICRL (Schmied et al., 2024; Sridhar et al., 2024) and retrieval-based RL (Goyal et al., 2022; Humphreys et al., 2022) when memory selection decides what evidence the model sees; and value-aware agents (Liu et al., 2026; Berkes et al., 2026) or model-aware agents (Son et al., 2025; Xu et al., 2026c) when values or dynamics are inferred from context. It leaves out standard offline RL policies that do not adapt to new test-time evidence, deployment-time fine-tuning, and domain-specific robotics or multi-agent surveys unless their mechanisms clarify the fixed-parameter context problem. We treat LLM agents and broader foundation-agent systems as bridge cases when they close the loop between context, action, feedback, and future behavior. Pure conversational memory, retrieval-augmented question answering, or one-shot prompting is therefore outside the core claim unless the retained context changes decisions in a repeated environment. The closest bodies of work are therefore not one

lineage but several intersecting ones: lifelong and continual ICRL (Xu et al., 2024; Wang et al., 2025d); non-stationary MDP analyses (Lecarpentier & Rachelson, 2019; Chandak et al., 2020c; Wei & Luo, 2021), dynamic-regret analyses (Luo et al., 2024; Pettet et al., 2024; Chen et al., 2025a), and near-optimal non-stationary control (Mao et al., 2021); value- and uncertainty-aware ICRL (Tarasov et al., 2025; Liu et al., 2026; Berkes et al., 2026), compositional or suboptimal-data ICRL (Xu et al., 2026a; Dong et al., 2026), and model-aware ICRL (Son et al., 2025; Xu et al., 2026c); retrieval-augmented ICRL (Schmied et al., 2024; Sridhar et al., 2024), history curation (Chen et al., 2025b), and retrieval-based RL (Goyal et al., 2022; Humphreys et al., 2022); large heterogeneous sequence-agent datasets (Nikulin et al., 2025; Wang et al., 2025a), generalist sequence agents (Polubarov et al., 2025; 2026), and mixture-of-experts routing (Wu et al., 2025); and reward-feedback agents (Monea et al., 2024; Song et al., 2025), safety-feedback agents (Chandak et al., 2020a; Moeini et al., 2025a), and barrier-style safe ICRL (Kwon et al., 2026).

The non-stationary lens changes how the field should be evaluated. Many current ICRL evaluations test held-out task generalization or average improvement with longer prompts. Those tests are useful, but they do not by themselves show whether an agent can recover after a shift, reject stale demonstrations, reuse old context when a mode recurs, or operate under a finite context budget. For non-stationary ICRL, evaluation should report lifecycle curves, post-shift recovery, dynamic regret, stale-context sensitivity, retrieval utility, and adaptation under controlled changes in shift frequency and observability. Recent work begins to address these issues directly through lifelong ICRL (Xu et al., 2024), continual ICRL (Wang et al., 2025d), classical non-stationary RL (Cheung et al., 2020; Fei et al., 2020; Wei & Luo, 2021), near-optimal non-stationary control (Mao et al., 2021), bandit tracking (Garivier & Moulines, 2011; Russac et al., 2019), and dynamic-regret analyses (Chen et al., 2025a), but **the area still lacks a shared taxonomy and reporting standard.**

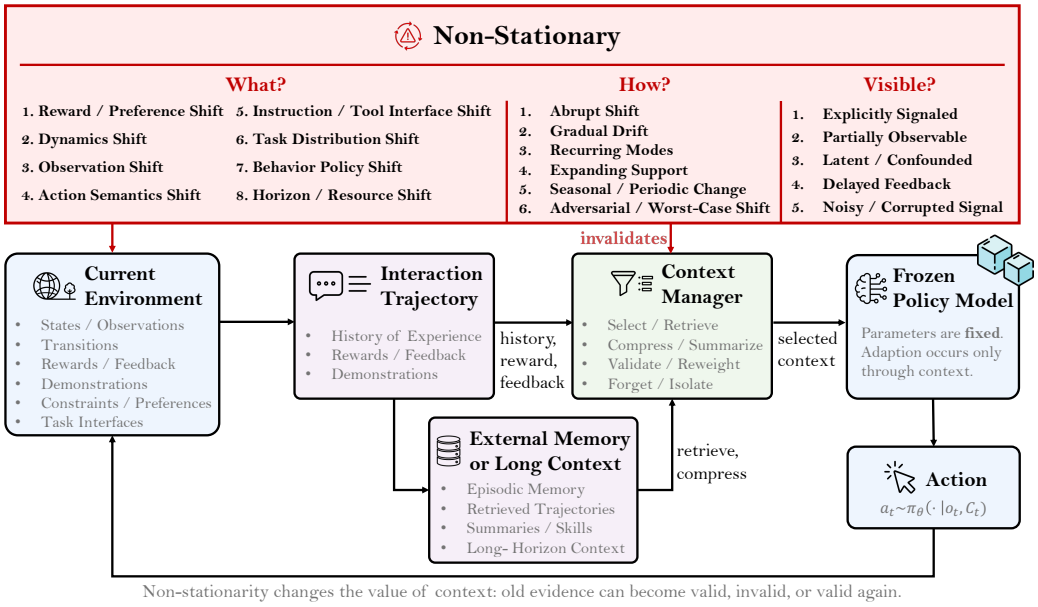

Figure 1: Survey frame for non-stationary ICRL. A fixed-parameter policy model must infer and act from interaction trajectories while the decision process and the validity of accumulated evidence change.

Figure 1 gives the organizing picture. The frozen decision model is only one part of the system. The context pipeline decides which observations, demonstrations, rewards, retrieved memories, and summaries reach that model at each decision point. Once the environment changes, this pipeline is both the adaptation mechanism and a common source of failure. ICRL systems should therefore be evaluated by how their policy model, interaction history, memory, and feedback loop behave together under shifts. Table 1 makes this distinction explicit by contrasting the scope of this paper with a general ICRL survey. Table 2 then positions the paper against adjacent survey literatures, and Figure 2 shows the resulting domain boundary.

Table 1: Scope contrast between general ICRL surveys and this non-stationary ICRL survey.

Dimension	General ICRL Survey	This Survey
Core Focus	Context-driven improvement in pretrained decision models.	Validity of accumulated context after environmental change.
Axis	Pretraining objectives, architectures, context formats, and theory.	Sources, temporal patterns, and observability of non-stationarity.
Evaluation	Few-shot improvement, held-out tasks, context length, and architecture comparisons.	Post-shift recovery, dynamic regret, stale-context sensitivity, memory selection, and re-adaptation.
Failure Mode	Failure to use context or generalize to new tasks.	Over-trusting stale, contradictory, or distributionally invalid context.
Open Problems	Better training, scaling, architectures, and mechanistic theory.	Change-point-aware context policies, validity-aware retrieval, adaptive forgetting, and finite-context theory under drift.

Table 2: Related survey landscape.

Literature	Representative Work and Emphasis	Gap Addressed Here
General ICRL	ICRL surveys organize pretraining, context construction, architectures, evaluation, and theory for ICRL broadly (Moeini et al., 2025b).	Re-centers the discussion on context validity after environmental change rather than ICRL as a whole.
Continual RL	Continual RL studies lifelong adaptation, plasticity, forgetting, and transfer (Khetarpal et al., 2022; Abel et al., 2023; Hadsell et al., 2020; Caccia et al., 2022).	Imposes the stricter ICRL constraint: the deployed decision model adapts through context, not weight updates.
Non-Stationary RL	Non-stationary bandits, MDP algorithms, regret analyses, and NS-Gym benchmarks address changing rewards, dynamics, and latent regimes (Besbes et al., 2014; Padakandla, 2020; Lecarpentier & Rachelson, 2019; Cheung et al., 2020; Fei et al., 2020; Feng et al., 2023; Wei & Luo, 2021; Keplinger et al., 2025).	Adds learned context construction, retrieval, and stale evidence as first-class variables.
Change Detection And Bandits	Changepoint inference, sliding-window or discounted bandits, and restart methods discount old samples or react to abrupt shifts (Adams & MacKay, 2007; Garivier & Moulines, 2011; Russac et al., 2019; Trovo et al., 2020; Auer et al., 2002; 2019).	Treats these tools as context-validity mechanisms inside fixed-parameter agents.
Meta-RL	Black-box, gradient-based, belief-based, and non-stationary meta-RL learn fast adaptation through hidden state, gradients, or latent task beliefs (Duan et al., 2016; Wang et al., 2016; Finn et al., 2017; Beck et al., 2023; Rakelly et al., 2019; Zintgraf et al., 2020; Humplik et al., 2019; Mendonca et al., 2020; Bing et al., 2022).	Connects this lineage to modern sequence models and finite context-management failures.
Offline And Sequence RL	Offline RL, transformer-in-RL surveys, and trajectory models learn from fixed datasets or tokenize trajectories as supervised sequences (Levine et al., 2020; Prudencio et al., 2023; Li et al., 2023; Chen et al., 2021; Janner et al., 2021; Emmons et al., 2022).	Separates dataset-scale sequence modeling from true test-time context adaptation.
ICL Mechanisms	Work on formatting, function learning, learnability, and implicit optimization studies how transformers infer new functions from examples (Chan et al., 2022; Min et al., 2022; Garg et al., 2022; Akyurek et al., 2023; Wies et al., 2023; von Oswald et al., 2023; Kirsch et al., 2022; Akyurek et al., 2024).	Transfers the mechanism question to sequential decisions with rewards, stale evidence, and feedback loops.
RL Generalization	Zero-shot and procedural generalization work studies held-out environments, overfitting control, and distribution shift (Kirk et al., 2023; Cobbe et al., 2020).	Requires within-lifetime context invalidation, not only train-test diversity.
LLM Agents	Agent surveys and reflection or reward-feedback agents emphasize planning, memory, tool use, and reward-guided improvement (Wang et al., 2024b; Yao et al., 2023; Shinn et al., 2023; Song et al., 2025).	Includes only sequential feedback loops where actions, rewards, tool outcomes, or memories change future decisions; excludes general chatbot memory, one-shot prompting, and RAG without control feedback.

Table 2: Related survey landscape (continued).

Literature	Representative Work and Emphasis	Gap Addressed Here
Retrieval And History Curation	Retrieval-augmented and history-filtered ICRL systems decide which evidence reaches the policy model (Schmied et al., 2024; Sridhar et al., 2024; Chen et al., 2025b).	Asks whether retained evidence is valid under the current regime, not merely similar or high-return.
Generalist Sequence Agents	Cross-domain agents, large ICRL datasets, and generalist sequence models scale task breadth, action interfaces, and trajectory-token modeling (Lee et al., 2022; Reed et al., 2022; Nikulin et al., 2025; Wang et al., 2025a; Polubarov et al., 2025; 2026; Wu et al., 2025).	Treats scale as useful but insufficient without shift-aware evaluation and context-validity tests.

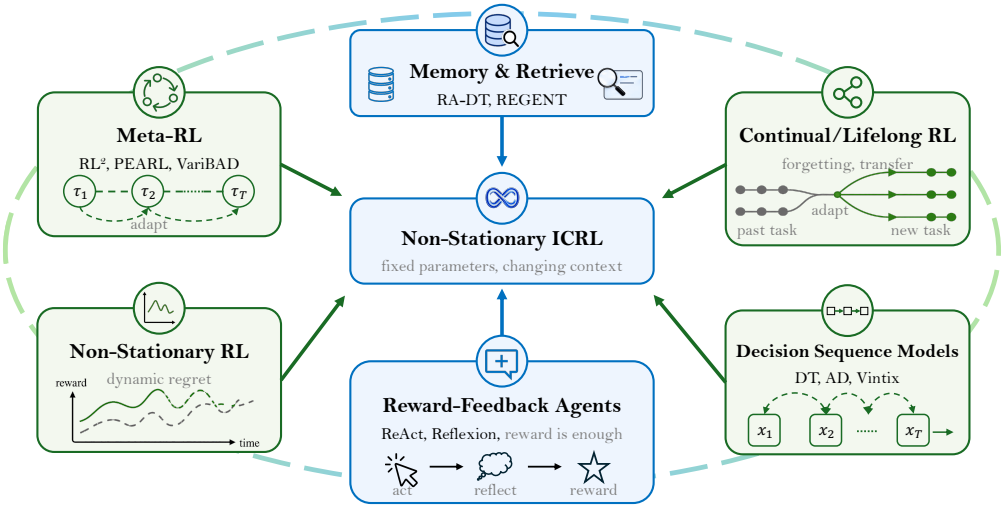

Figure 2: Scope map for this survey, not a timeline. The paper studies the intersection where sequential decision agents adapt through context while the data-generating process changes; adjacent fields enter only when they clarify this mechanism.

The survey builds its argument in four steps. Figure 2 is a domain-boundary view: it identifies which literatures overlap in the problem studied here. The survey first casts non-stationary ICRL as a context-validity problem. It then separates non-stationarity into three questions that are often conflated: what changes, how the change unfolds, and how directly the change can be observed by a frozen context-conditioned agent. On that basis, it reviews methods by the mechanism that manages context–raw histories, curated histories, retrieval, compression, value-aware inference, model-based planning, reflection, and forgetting–rather than by publication chronology. Finally, it argues for evaluation protocols that expose adaptation rather than memorization or domain-randomization-style generalization, and it highlights problems that stationary benchmarks tend to miss: action-semantics shift, context poisoning, memory isolation, uncertainty-calibrated retrieval, and theory for finite, stale, or conflicting context. The same source/pattern/observability vocabulary is used throughout the survey so that method claims, benchmark claims, and open problems can be stated at the level of shift actually being tested.

The rest of the survey is organized as follows. Section 2 defines ICRL and its boundaries. Section 3 formalizes non-stationary ICRL and introduces context validity. Section 4 develops the taxonomy. Section 5 explains why non-stationarity changes the role of context. Section 6 reviews method families through the lens of context management. Section 7 discusses training data and objectives. Section 8 proposes evaluation protocols and metrics. Section 9 summarizes theoretical links and gaps. Section 10 closes with design principles and open problems.

2 Foundations of In-Context Reinforcement Learning

2.1 Fixed-Parameter Test-Time Adaptation

We use an operational definition, defining the concept through observable, actionable, and testable criteria. An agent performs ICRL when a fixed-parameter decision model uses the current episode or a broader interaction history to infer latent task structure and improve future action selection at test time. For an observation o_t, action a_t, and context C_t, the policy can be written as

$$\pi_\theta(a_t \mid o_t, C_t), \qquad \theta \text{ fixed at test time.} \tag{1}$$

The context is usually a finite serialization of interaction evidence,

$$C_t = \mathrm{pack}_B\left(\tau_{1:k}, d_{1:\ell}, q_{1:m}, z_t\right), \qquad \tau_i = (o_1^i, a_1^i, r_1^i, \ldots, o_{H_i}^i), \tag{2}$$

where pack_B denotes a budget-limited formatting, retrieval, or compression operator; $\tau_{1:k}$ are trajectories, $d_{1:\ell}$ are demonstrations or descriptions, $q_{1:m}$ are retrieved memories or summaries, and z_t denotes other task hints or feedback. The context C_t may contain recent trajectories, older episodes, demonstrations, rewards, return prompts, task descriptions, retrieved memories, compressed summaries, value estimates, natural-language feedback, or other tokens that influence the model's action. The defining constraint is not the token type but the adaptation channel: the agent adapts through C_t, not through gradient updates to θ. Merely conditioning on a prompt is therefore not enough; the context must carry evidence from which the model can change later decisions in a task-relevant direction. Equivalently, deployment-time adaptation is a forward-pass map from evidence to behavior,

$$C_t \mapsto \pi_\theta(\cdot \mid o_t, C_t), \qquad \theta_{t+1} = \theta_t = \theta, \tag{3}$$

whereas online RL or gradient-based meta-learning would allow $\theta_{t+1} \neq \theta_t$ after observing new experience.

This definition covers several lines of work. Algorithm distillation trains a model on histories generated by a learning algorithm, so that the trained model can imitate the algorithm's within-lifetime improvement through its context (Laskin et al., 2023). Decision-pretrained transformers and supervised-pretraining approaches show that trajectory models can learn useful in-context decision rules from offline data (Chen et al., 2021; Janner et al., 2021; Emmons et al., 2022), prompting and generalized decision-transformer objectives (Xu et al., 2022; Furuta et al., 2022), supervised ICRL theory and pretraining (Lee et al., 2023; Lin et al., 2024), agentic or variable-action formulations (Liu & Abbeel, 2023; Sinii et al., 2024), and large multi-task decision datasets (Nikulin et al., 2025; Polubarov et al., 2025; 2026). Long-context meta-RL agents connect the same idea to recurrent or transformer policies trained by reinforcement learning over multiple tasks or episodes, including transformer meta-RL (Melo, 2022), AMAGO-style agents (Grigsby et al., 2024a;b), and RELIC-style long-context systems (Elawady et al., 2024). Retrieval-augmented agents make the context partly external by selecting relevant experience from a memory database, including retrieval-augmented decision transformers (Schmied et al., 2024), REGENT-style retrieval (Sridhar et al., 2024), and earlier retrieval-based RL systems (Goyal et al., 2022; Humphreys et al., 2022). History curation and value-aware variants extend this design space by filtering learning histories (Chen et al., 2025b), routing heterogeneous decision tokens (Wu et al., 2025), or replacing pure imitation with value-sensitive objectives (Tarasov et al., 2025; Liu et al., 2026; Berkes et al., 2026). Efficient long-sequence alternatives such as S4-style state-space models (Lu et al., 2023; David et al., 2023), recurrent-memory and S5-style models (Bulatov et al., 2022; Smith et al., 2023), gated or xLSTM-style recurrent attention (Beck et al., 2024), and Mamba-style sequence models (Ota, 2024; Gu & Dao, 2023) are also relevant because non-stationary ICRL often asks the agent to retain long histories while still reacting quickly to recent evidence.

ICRL is therefore a behavioral property of the deployment loop, not a single architecture. A transformer with a trajectory prompt may fail to perform ICRL if the prompt is ignored. A recurrent policy may perform ICRL if its hidden state or explicit context implements fast adaptation. A large language model belongs in scope only when it is embedded in a sequential decision process with action, feedback, and reward-like signals, rather than merely recalling conversational state. This includes LLM bandit agents (Monea et al., 2024), reward-feedback agents (Song et al., 2025), ReAct-style agents that act in an environment (Yao et al.,

2023), reflection or experience-library systems whose stored lessons change later trials (Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2023a), and tool-use or computer-control agents whose future actions depend on previous tool outcomes (Tan et al., 2025). It excludes static prompting (Brown et al., 2020; Min et al., 2022), one-shot retrieval-augmented question answering (Lewis et al., 2020), generic chatbot memory (Zhong et al., 2024; Wang et al., 2023b; Packer et al., 2023), and summaries that do not affect a repeated decision loop. In latent-task notation, many ICRL systems can be read as amortizing a Bayes-adaptive control rule:

$$b_t(\xi) = p(\xi \mid C_t) \propto p(\xi) \prod_{(o,a,r,o') \in C_t} p_\xi(o', r \mid o, a), \qquad \pi^\star(a \mid o, C_t) = \int \pi^\star_\xi(a \mid o)\, b_t(\xi)\, d\xi, \tag{4}$$

where ξ is a hidden task, dynamics, reward, or goal parameter. The learned policy need not explicitly represent b_t, but successful ICRL behavior requires some computation with the same effect: evidence in context changes the posterior over which decision rule is currently appropriate. Recent theory strengthens this operational reading by showing that transformers can represent policy-improvement-like computations from in-context trajectories (Liang & Lai, 2026), complementing chain-of-thought and temporal-difference analyses of ICRL (Xie et al., 2026; Wang et al., 2024a). A fixed-parameter model can still implement an update-like decision rule inside the forward pass, so ICRL should not be dismissed as static behavior cloning. The guarantees usually rely on coherent context drawn from a stable task family; non-stationary ICRL asks what happens when the same context window also contains obsolete, conflicting, or regime-mismatched evidence. This behavioral definition is consistent with broader ICL work showing that in-context behavior can arise from distributional structure (Chan et al., 2022), meta-training and function learning (Garg et al., 2022; Akyurek et al., 2023; Wies et al., 2023), implicit linear-model fitting or gradient-descent-like computations (von Oswald et al., 2023; Akyurek et al., 2024), and demonstration formatting (Kirsch et al., 2022; Min et al., 2022). For this survey, those results are useful background but not sufficient: reinforcement learning adds actions, rewards, exploration, credit assignment, and the possibility that old examples become actively wrong.

This behavioral definition also avoids a common ambiguity in the literature. A method can be transformer-based, pretrained on trajectories, and evaluated on held-out tasks without demonstrating ICRL. The decisive test is counterfactual: if the same observation is paired with different valid contexts, does the policy change in the direction implied by those contexts? One compact diagnostic is a context-contrast score,

$$\begin{aligned} &\Delta_{\text{ctx}}(o; C, C') = D(\pi_\theta(\cdot \mid o, C) \,\|\, \pi_\theta(\cdot \mid o, C'))\,, \\ &\qquad J(\pi_\theta; C) > J(\pi_\theta; C') \quad \text{when } C \text{ contains better current evidence.} \end{aligned} \tag{5}$$

where D may be total variation, KL divergence, or another distance between action distributions. Large distributional change alone is not enough; the changed action distribution must improve the downstream decision objective under the task implied by the context. For non-stationary ICRL, the counterfactual is stronger. If the prompt contains both old and new evidence, can the policy privilege the evidence that is valid under the current regime? This is the reason the survey treats context construction, retrieval, and forgetting as first-class parts of the agent.

2.2 Nearby Paradigms

ICRL overlaps with several mature paradigms but is not identical to them. Black-box meta-RL trains an agent over a distribution of tasks such that within-episode or within-lifetime computation implements a learning procedure (Duan et al., 2016; Wang et al., 2016). ICRL can be seen as a modern sequence-modeling version of this idea, with explicit context windows, retrieval, and trajectory tokens.

Classical RL and planning texts frame the stationary baseline through MDPs and value functions (Puterman, 1994; Sutton & Barto, 2018), while POMDP work shows how hidden state can be converted into belief-state inference (Kaelbling et al., 1998). Contextual MDPs (Hallak et al., 2015) and hidden-parameter MDPs (Killian et al., 2017; Doshi-Velez & Konidaris, 2016; Perez et al., 2020) are especially close to meta-RL because a latent or hidden parameter controls rewards and dynamics across related tasks. Epistemic-POMDP (Ghosh et al., 2021) and task-inference views (Humplik et al., 2019; Ren et al., 2020) make the same point from a different angle: what looks like poor generalization can be implicit partial observability over an unobserved

task, regime, or data-collection process. For non-stationary ICRL this perspective is useful but incomplete. The latent task belief is not only uncertain; it can expire, conflict with new evidence, or become useful again when a previous regime returns. This is why recurrent memory baselines (Heess et al., 2015; Ni et al., 2022), trajectory-contrastive context encoders (Wang et al., 2021a), and non-stationary meta-RL methods (Mendonca et al., 2020; Bing et al., 2022) are important bridge literatures for the survey.

The older recurrent meta-RL line, including RL^2 (Duan et al., 2016) and learning-to-reinforcement-learn systems (Wang et al., 2016), already made the key move of representing fast adaptation in hidden state rather than in test-time weight updates. Attentive and latent-belief meta-RL methods such as SNAIL (Mishra et al., 2018), PEARL (Rakelly et al., 2019), and VariBAD (Zintgraf et al., 2020) sharpened the same idea by treating recent experience as evidence for a latent task or Bayes-adaptive belief. Gradient-based meta-RL, represented by MAML-style training (Finn et al., 2017) and later meta-RL variants (Beck et al., 2023), is adjacent but distinct because its adaptation channel is usually an explicit inner-loop parameter update rather than a fixed-parameter forward pass. The operational difference can be summarized as

$$\underbrace{\pi_{\theta'}(a_t \mid o_t), \quad \theta' = \theta - \alpha \nabla_\theta \mathcal{L}(\theta; \tau_{1:k})}_{\text{gradient-based adaptation}} \qquad \text{versus} \qquad \underbrace{\pi_\theta(a_t \mid o_t, C_t), \quad \theta \text{ unchanged}}_{\text{in-context adaptation}}. \tag{6}$$

Decision Transformer (Chen et al., 2021), Trajectory Transformer (Janner et al., 2021), RvS (Emmons et al., 2022), and Multi-Game Decision Transformer (Lee et al., 2022) then made trajectory sequence modeling a scalable substrate for offline decision models. Offline RL also learns from previously collected data (Levine et al., 2020; Prudencio et al., 2023), but standard offline RL usually produces a policy that acts without using new test-time examples as context; conservative Q-learning (Kumar et al., 2020) and implicit Q-learning (Kostrikov et al., 2022) are therefore background for value objectives, not ICRL by themselves.

Continual RL studies agents that persist across changing experiences and often update parameters, replay buffers, or other long-lived state (Khetarpal et al., 2022; Abel et al., 2023). Non-stationary RL studies algorithms for changing MDPs, often with change-point tests (Adams & MacKay, 2007; Chen et al., 2019), sliding windows or discounting (Garivier & Moulines, 2011), bandit and MDP variation-budget guarantees (Besbes et al., 2014; Cheung et al., 2020; Fei et al., 2020), dynamic or near-optimal regret bounds (Feng et al., 2023; Wei & Luo, 2021; Mao et al., 2021), transformer-based regret analysis (Chen et al., 2025a), or restart and policy-search schedules (Auer et al., 2019; Padakandla et al., 2020; Lecarpentier & Rachelson, 2019). Survey work (Padakandla, 2020) and policy-search variants (Chandak et al., 2020c; Luo et al., 2024; Pettet et al., 2024) further broaden this algorithmic view. ICRL imposes the additional constraint that the main policy parameters are fixed at test time.

The distinction among continual RL, lifelong RL, and ICRL is therefore operational rather than merely terminological. Continual RL asks how an agent keeps learning across a stream of tasks or changing environments (Khetarpal et al., 2022; Abel et al., 2023); the usual toolbox may include online gradient updates in persistent agents (Khetarpal et al., 2022; Abel et al., 2023), replay (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019), regularization (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018), architectural expansion (Rusu et al., 2016), or explicit task-free segmentation (Caccia et al., 2022). Lifelong RL is often used more broadly for persistent agents that accumulate and reuse experience over long horizons (Hadsell et al., 2020; Lesort et al., 2020; Chandak et al., 2020b); it emphasizes retention (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018), forward transfer (Wolczyk et al., 2022; Powers et al., 2021), and recovery or reuse when older skills or modes become relevant again (Steinparz et al., 2022; Schmied et al., 2024; Sridhar et al., 2024). ICRL, in contrast, asks whether the deployed decision model can adapt with its parameters fixed, by changing only the context, hidden state, retrieval set, or summary supplied to the forward pass. Non-stationary ICRL is the intersection of these ideas: it inherits continual and lifelong RL's changing-world concern, but it restricts the adaptation mechanism to context management. For a fixed context budget B, the ICRL boundary can be written as the constrained deployment problem

$$\max_g \ \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t\right] \quad \text{s.t.} \quad C_t = g(h_{1:t-1}), \ |C_t| \le B, \ \theta \text{ fixed}, \tag{7}$$

which places the burden of adaptation on what is selected, compressed, retrieved, or forgotten before the policy's forward pass. Recent direct work on lifelong ICRL (Xu et al., 2024) and continual ICRL (Wang

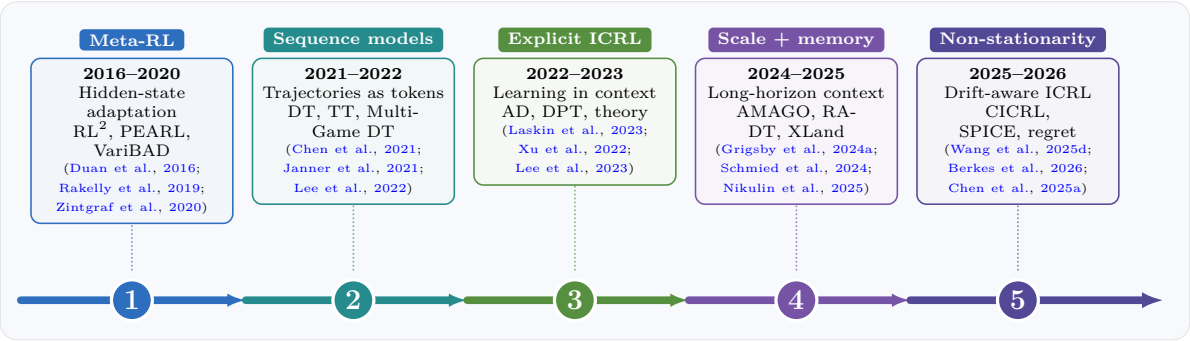

Figure 3: Historical lineage of ICRL-related method families. The timeline orders recurrent meta-RL, sequence decision models, retrieval-augmented agents, and explicit non-stationary ICRL by development history.

et al., 2025d), together with adjacent non-stationary in-context-learning theory (Qin et al., 2026), makes this intersection increasingly explicit.

Figure 3 gives a compact historical timeline for this lineage. The important transition is from hidden-state adaptation in early meta-RL, to trajectory sequence modeling, to explicit ICRL, and finally to the current non-stationary turn where retrieval, value awareness, and lifecycle evaluation become central. Table 3 uses this distinction to define the core, bridge, and excluded cases for the rest of the survey.

Table 3: Scope boundary for this survey.

Category	Included in This Survey	Boundary Condition
Core ICRL	Algorithm distillation, supervised ICRL, decision-pretrained transformers, long-context meta-RL, retrieval-augmented ICRL, value-aware ICRL.	Test-time behavior must improve or adapt from interaction context while model parameters fixed.
Bridge Cases	LLM reward-feedback agents, contextual-bandit ICRL, pure-exploration ICL, model-based in-context planning, long-memory agent systems.	Included only when the setting has sequential interaction and feedback relevant to future action (Monea et al., 2024; Krishnamurthy et al., 2024); exploration and scaffolding variants are treated similarly (Dai et al., 2024; Nie et al., 2025).
Excluded Cases	Pure supervised in-context learning, pure chatbot memory, standard offline RL, online fine-tuning, robotics-specific surveys, multi-agent adaptation.	Excluded because they lack the target adaptation channel or violate user-specified scope (Levine et al., 2020; Prudencio et al., 2023; Luo et al., 2023).

2.3 Boundary Choices Under Non-Stationarity

This survey draws the boundary around the adaptation channel rather than around the mere presence of non-stationarity: a method is central only when a fixed-parameter sequential decision agent must adapt to within-lifetime changes by selecting, compressing, retrieving, or forgetting context. Boundary choices matter more under non-stationarity than under stationary task generalization. If online parameter updates are allowed, adaptation can be attributed to continual learning rather than context use. If the environment is not sequential, the problem becomes supervised in-context learning or contextual inference rather than reinforcement learning. If the setting is purely offline and the agent receives no test-time context, then stale-context behavior cannot be studied. This survey therefore uses a strict core definition and a conservative extended scope. The aim is not to claim every context-conditioned model as ICRL, but to isolate the cases where changing context is the mechanism by which the agent adapts to changing decision processes.

The conservative boundary is also useful for negative results. If a fixed model fails after a reward switch, the failure should not immediately be attributed to lack of model capacity. It may be caused by a context policy that supplies the wrong evidence, a training distribution that never invalidates old context, or an evaluation protocol that rewards memorization. Conversely, if a system succeeds only because it fine-tunes online, the result may be valuable for continual RL but does not answer the ICRL question. Maintaining this boundary keeps the survey focused on the mechanisms that make adaptation through context possible or impossible while deployed policy parameters remain fixed.

3 Problem Setting for Non-Stationary ICRL

3.1 From Stationary MDPs to Changing Decision Processes

Let the agent interact with a sequence of decision processes

$$\mathcal{M}_t = (\mathcal{S}_t, \mathcal{A}_t, P_t, R_t, \Omega_t, O_t, \gamma_t), \tag{8}$$

where $\mathcal{S}_t$ is the state space, $\mathcal{A}_t$ is the action space, P_t is the transition kernel, R_t is the reward function, Ω_t is the observation space, O_t is the observation function, and γ_t is the discount factor or horizon convention. This notation extends the standard MDP formalism (Puterman, 1994) and allows partial observability (Kaelbling et al., 1998), hidden contexts (Hallak et al., 2015), or latent task parameters (Killian et al., 2017; Doshi-Velez & Konidaris, 2016; Ghosh et al., 2021) to enter through O_t and through the unobserved variables that govern P_t and R_t. Non-stationarity means that at least one decision-relevant component changes over time or across phases of the same context-bearing lifetime, for example $P_t \neq P_{t+1}$, $R_t \neq R_{t+1}$, or a change in O_t, $\mathcal{A}_t$, goals, constraints, or action semantics. Not every component must change. The survey focuses on changes that alter which actions are good, which observations are informative, which memories are relevant, or which previous trajectories should be trusted.

At time t, the agent acts with a fixed policy class $\pi_\theta(a_t \mid o_t, C_t)$. The context C_t is produced by a context-management process:

$$C_t = g_\phi(h_{1:t-1}, m_{1:t-1}, b_t), \tag{9}$$

where $h_{1:t-1}$ denotes interaction history, $m_{1:t-1}$ denotes external memories or retrieved items, and b_t denotes budget constraints such as context length, retrieval count, or latency. In the strictest ICRL setting, both θ and ϕ are fixed after training. Some systems use hand-designed retrieval or summarization rules (Schmied et al., 2024; Sridhar et al., 2024); others learn the context manager (Rakelly et al., 2019; Zintgraf et al., 2020; Chen et al., 2025b). The key distinction from continual RL remains that the adaptation channel at deployment is the constructed context, not online weight modification. The policy must therefore infer task rules, reward meanings, transition structure, or action effects from history and express adaptation through future actions.

This formulation separates three objects that are often conflated. The *decision model* maps observations and context to actions. The *context manager* chooses what evidence is visible to the decision model. The *environment sequence* determines whether old evidence remains compatible with the present decision problem. In stationary ICRL, the second object can often be treated as a formatting detail: include demonstrations, rewards, or previous episodes, and ask whether the model uses them. In non-stationary ICRL, the context manager is a policy in its own right. It must trade off recency, reliability, diversity, budget, and the possibility that a previously encountered regime has returned. Classical non-stationary bandit and RL algorithms often encode this tradeoff through sliding windows and discount factors (Garivier & Moulines, 2011; Russac et al., 2019), optimism bonuses (Besbes et al., 2014; Cheung et al., 2020; Wei & Luo, 2021), restart schedules (Auer et al., 2019), or explicit changepoint posteriors (Adams & MacKay, 2007; Chen et al., 2019). The ICRL version differs because those operations may be implicit in the prompt, retrieval index, attention pattern, or memory state rather than in a hand-coded online estimator.

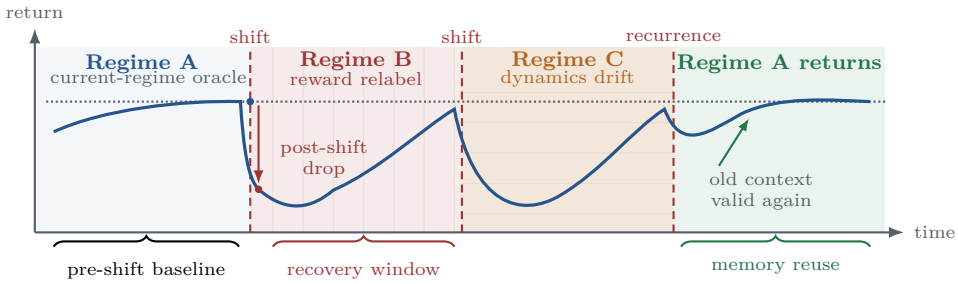

Figure 4: Lifecycle structure of a non-stationary ICRL episode. Average return hides where adaptation succeeds or fails.

Figure 4 frames the lifecycle perspective behind the adaptation metrics below, including dynamic regret, recovery time, context efficiency, and stale-context sensitivity. A non-stationary benchmark should therefore reveal pre-shift competence, immediate post-shift degradation, recovery speed, and the reuse of context when a previous regime returns.

3.2 Context Validity

We define context validity as a decision-time property. A context item $c_i \in C_t$ is valid at time t if conditioning on it improves, calibrates, or preserves decision quality under the current environment $\mathcal{M}_t$ compared with ignoring it. It is invalid if it systematically worsens decisions, biases value estimates, hides a change point, or consumes scarce context budget without useful information. One way to express this decision-relative notion is the marginal context value

$$\nu_t(c_i) = \mathbb{E}[Q_t^{\pi_\theta}(s_t, a_t \mid C_t) - Q_t^{\pi_\theta}(s_t, a_t \mid C_t \setminus \{c_i\})]\,, \tag{10}$$

where the expectation is over the current state, action-selection randomness, and any retrieval randomness. Positive $\nu_t(c_i)$ indicates helpful evidence, negative $\nu_t(c_i)$ indicates harmful or misleading evidence, and values near zero indicate context that may be syntactically relevant but decision-irrelevant. This definition is intentionally performance-relative: old context is not invalid merely because it is old, and recent context is not valid merely because it is recent. For recurring modes, an old trajectory may become valid again (Khetarpal et al., 2022; Abel et al., 2023; Xu et al., 2024). For abrupt reward relabeling, even a high-return recent trajectory may be misleading (Besbes et al., 2014; Mendonca et al., 2020; Song et al., 2025).

This definition deliberately differs from syntactic relevance. A trajectory can be syntactically similar to the current observation stream while being decision-invalid because the reward changed (Schmied et al., 2024; Berkes et al., 2026). A demonstration can be high quality under its original regime while being invalid under a new action interface (Chandak et al., 2020b; Sinii et al., 2024). A recent item can be invalid because it was produced by an exploratory or corrupted behavior policy (Dorfman et al., 2020; Fu et al., 2020; Levine et al., 2020). The survey therefore uses validity as a causal and decision-theoretic notion: context is valid when it improves the action distribution for the present regime, not when it looks similar, recent, or successful in isolation.

The validity lens gives a compact statement of the non-stationary ICRL problem:

> A **non-stationary ICRL** agent must adapt through context while estimating which previously inferred rules, strategies, and memories still describe the current data-generating process.

This is a stronger requirement than stationary few-shot generalization. It requires the model or its context manager to solve a joint problem of inference, memory selection, and control.

3.3 Objectives

The objective can be expressed in several complementary ways. The standard cumulative-reward objective is

$$J(\pi_\theta, g) = \mathbb{E}\left[\sum_{t=1}^{T} \gamma_t r_t\right], \tag{11}$$

where actions are produced by $\pi_\theta(a_t \mid o_t, g(h_{1:t-1}, m_{1:t-1}, b_t))$. For non-stationary environments, average return hides important lifecycle behavior. Two agents can have similar average return while one recovers immediately after shifts and the other fails catastrophically but performs well between shifts. Ergodicity-focused work gives a second reason to distrust a single average-return number: when reward processes are non-ergodic, ensemble expectations can diverge from the time-average experience of an individual deployed agent (Baumann et al., 2026; Verbruggen et al., 2026). In such settings, optimizing or reporting expected return can favor a policy that looks good across many hypothetical runs while producing poor long-run outcomes on the actual trajectory that carries the agent's context. This concern is especially acute for ICRL

because the context window is itself trajectory-dependent; early failures, stale memories, or unlucky regime exposure can change the evidence available for later decisions. Dynamic regret is often more diagnostic:

$$\mathrm{Reg}_T^{\mathrm{dyn}} = \sum_{t=1}^{T} V_t^{\pi_t^\star}(s_t) - \sum_{t=1}^{T} V_t^{\pi_\theta}(s_t), \tag{12}$$

where $\pi_t^\star$ is an oracle policy for the current environment or current latent mode. This objective follows the same motivation as adversarial and non-stationary bandit work (Auer et al., 2002; Besbes et al., 2014; Li & Li, 2019), non-stationary MDP regret analyses (Cheung et al., 2020; Fei et al., 2020; Wei & Luo, 2021), near-optimal or kernelized variants (Mao et al., 2021; Domingues et al., 2021), and non-ergodic RL critiques of expectation-based objectives (Baumann et al., 2026; Verbruggen et al., 2026): a static comparator or ensemble average can make an agent look good simply because the environment changes or because failures are averaged away, whereas a dynamic comparator exposes the price paid for delayed detection, adaptation, and context repair along the realized lifetime. The amount of environmental movement can be summarized by a variation budget, for example

$$B_T = \sum_{t=2}^{T} \left(\|P_t - P_{t-1}\|_\infty + \|R_t - R_{t-1}\|_\infty \right), \tag{13}$$

with analogous terms added when observation functions, action semantics, constraints, or task priors change. This budget is not the only possible formalization, but it makes explicit that a method's adaptation claim depends on how often and how much the environment changes. Post-shift recovery time, post-shift area under the learning curve, context efficiency, and stale-context sensitivity provide additional views that are more directly connected to ICRL. Given a shift at τ, one useful recovery statistic is

$$T_{\mathrm{rec}}(\tau, \alpha) = \min \left\{ k \geq 0 : \frac{1}{w} \sum_{i=0}^{w-1} J_{\tau+k+i}^{\pi_\theta} \geq \alpha \frac{1}{w} \sum_{i=0}^{w-1} J_{\tau+k+i}^{\pi^\star} \right\}, \tag{14}$$

where α is a target fraction of oracle performance and w is a smoothing window. These objectives are summarized in Table 4, which shows why stationary ICRL metrics are insufficient once context validity changes over time.

Table 4: Stationary ICRL and non-stationary ICRL emphasize different objectives.

Objective or Metric	Stationary ICRL Interpretation	Non-Stationary ICRL Interpretation
Final Return	Can the agent infer and exploit a fixed task from context (Lee et al., 2023; Lin et al., 2024)?	Can the agent eventually recover after changes (Xu et al., 2024; Wang et al., 2025d)?
Few-Shot Improvement	Does more task-relevant context improve performance (Xu et al., 2022; Raparthy et al., 2024)?	Does the agent distinguish helpful new evidence from stale old evidence?
Context Length Scaling	Does a longer prompt help (Grigsby et al., 2024a;b)?	Does a longer prompt help only when invalid context is filtered or downweighted (Chen et al., 2025b; Schmied et al., 2024)?
Dynamic Regret	Usually secondary or absent.	Measures cost relative to a changing comparator policy (Besbes et al., 2014; Feng et al., 2023; Chen et al., 2025a).
Recovery Time	Often not measured.	Central measure of adaptation after reward, dynamics, observation, or preference shifts (Adams & MacKay, 2007; Wang et al., 2025d).
Stale-Context Sensitivity	Rarely tested.	Measures whether outdated high-confidence context harms decisions (Schmied et al., 2024; Berkes et al., 2026).

3.4 Non-Stationarity Is Not Domain Randomization

Held-out task evaluation and domain randomization both test generalization across tasks or environment parameters, but they do not by themselves test non-stationary ICRL. In non-stationary ICRL, the task may

change during the agent's lifetime, so the agent must adjust its behavior from context rather than from a reset, retraining, or a clean task descriptor. A model can therefore generalize well to a new stationary task while still failing when the task changes halfway through an episode; conversely, it may handle recurring change points in a narrow environment family without achieving broad held-out generalization.

This distinction is a useful diagnostic for benchmarks. If changes occur only between episodes and the agent receives a clean task descriptor at the start of each episode, the benchmark mainly tests task-conditioned generalization. If the task changes within a lifetime but the prompt or history is reset at the change point, it tests fast reconditioning rather than robustness to stale context (Wang et al., 2025d; Chen et al., 2025b). The central non-stationary ICRL regime arises when old and new regimes coexist in the same context window, forcing the agent to decide which evidence remains valid (Schmied et al., 2024; Berkes et al., 2026). Papers reporting non-stationary ICRL results should therefore specify which regime they evaluate.

4 A Taxonomy of Non-Stationarity for ICRL

Non-stationarity is too broad a label to support useful comparisons by itself. The term can refer to qualitatively different stressors, including announced or inferable reward changes in non-stationary bandits and MDPs (Garivier & Moulines, 2011; Besbes et al., 2014; Fei et al., 2020; Song et al., 2025), gradual variation in transition dynamics (Lecarpentier & Rachelson, 2019; Chandak et al., 2020c; Feng et al., 2023), recurrent latent regimes in contextual or hidden-parameter MDPs (Hallak et al., 2015; Killian et al., 2017; Doshi-Velez & Konidaris, 2016), and changes in the action interface or action-to-effect mapping (Chandak et al., 2020b; Sinii et al., 2024). These regimes do not exercise the same adaptation mechanism. An append-all history may be adequate for gradual reward estimation and harmful after abrupt action remapping. A retrieval-augmented model may be strong on recurring modes and still fail under subtle observation drift if its retrieval key is calibrated to the wrong surface features (Thakur et al., 2021; Yoran et al., 2023). A value-aware model may react to reward shifts while remaining brittle when transition dynamics move outside its support (Fujimoto et al., 2019; Kidambi et al., 2020; Yu et al., 2020b).

We therefore describe a non-stationary ICRL setting along three axes: the component that changes, the temporal structure of the change, and the evidence by which the agent can observe it. This is narrower than general RL generalization taxonomies, which mainly ask whether a trained policy transfers to held-out environments; here the issue is whether a fixed model can keep using, rejecting, or reactivating context during a lifetime (Kirk et al., 2023). It is also more operational than POMDP and contextual-MDP formalisms (Kaelbling et al., 1998; Hallak et al., 2015), epistemic-POMDP views (Ghosh et al., 2021), and hidden-parameter MDP formalisms (Killian et al., 2017; Doshi-Velez & Konidaris, 2016). Those formalisms justify treating task identity or regime as latent. They do not by themselves say how a finite prompt, memory index, or retrieval policy should be managed under drift. They also do not reduce non-stationary ICRL to domain randomization, because the relevant change happens inside the interaction trajectory rather than only between sampled training or test instances.

Figure 5 shows a compact map of the axes, with representative labels for each dimension, and Figure 6 gives a tree view for placing papers and benchmarks. The full category definitions appear in Tables 5–7. A concrete non-stationary setting can be described along all three independent dimensions simultaneously. For example, a setting labeled as dynamics/recurring/latent denotes a task in which the transition function changes across recurring regimes, while the current regime is not directly observed and must be inferred from the interaction context. The main practical point is that these axes should be reported independently. Explicit reward switches and latent recurring dynamics are both non-stationary, but they are not interchangeable evidence for the same adaptation claim.

4.1 Changing Components

The first axis characterizes where change enters the decision process. Table 5 summarizes the main cases. The rows are not mutually exclusive, since deployed systems often face multiple forms of change at the same time. The central question is which part of the accumulated context has become unreliable.

Table 5: Taxonomy axis 1: sources of non-stationarity and their context consequences.

Shift Source	Changing Object	Context Failure Mode	Typical Evaluation Probe
Reward or Preference	R_t; goal; utility; constraint weights; user preference.	Previously good trajectories become wrong demonstrations; reward labels may need reinterpretation.	Reward relabeling and goal switches in non-stationary bandits or MDPs (Garivier & Moulines, 2011; Besbes et al., 2014; Fei et al., 2020); preference reversal or reward-feedback changes (Song et al., 2025); safety-cost shifts (Chandak et al., 2020a; Moeini et al., 2025a).

Table 5: Taxonomy axis 1: sources of non-stationarity and their context consequences (continued).

Shift Source	Changing Object	Context Failure Mode	Typical Evaluation Probe
Dynamics	P_t; latent physics; transition stochasticity; exogenous process.	Old action–outcome evidence misleads planning, credit assignment, or value estimation.	Abrupt and gradual transition changes with fixed observations (Padakandla et al., 2020; Lecarpentier & Rachelson, 2019; Chandak et al., 2020c); regret and policy-search variants (Cheung et al., 2020; Feng et al., 2023; Luo et al., 2024; Pettet et al., 2024); latent and model-based dynamics inference (Hallak et al., 2015; Killian et al., 2017; Doshi-Velez & Konidaris, 2016; Son et al., 2025; Xu et al., 2026c).
Observation	O_t; sensor mapping; feature availability; observability level.	Context describes a mode that is no longer identifiable through the same observed features.	Feature corruption, sensor remapping, and missing observations in POMDP-style settings (Kaelbling et al., 1998; Ghosh et al., 2021); recurrent memory probes (Heess et al., 2015; Ni et al., 2022); long-context generalization tests (Grigsby et al., 2024b; Raparthy et al., 2024).
Action Semantics	$\mathcal{A}_t$; available actions; action-to-effect map.	Old action tokens no longer denote the same intervention.	Action permutation, action masking, and continuous-action rescaling (Chandak et al., 2020b; Sinii et al., 2024).
Instruction or Tool Interface	System instructions; user constraints; tool schemas; website state; feedback format.	Old instructions, reflections, or tool-use memories no longer describe valid actions.	Hidden instruction changes, preference drift, tool-schema remapping, website distribution shift, and feedback reliability changes in interactive language environments (Yao et al., 2023; Wang et al., 2022; Yao et al., 2022; Shridhar et al., 2020; Tan et al., 2025).
Task Distribution	Latent task sequence; support; mixture weights.	Pretraining or retrieval priors overweight outdated task modes.	Changing task families in meta-RL (Al-Shedivat et al., 2017; Mendonca et al., 2020); out-of-support lifelong ICRL and generated-MDP sequences (Xu et al., 2024; Wang et al., 2025a; Nikulin et al., 2025).
Behavior Policy	Demonstrator; data collector; exploration or logging policy.	Retrieved or demonstrated trajectories reflect policy mismatch rather than current task evidence.	Mixed-quality demonstrations and weak-data histories in offline RL (Levine et al., 2020; Kumar et al., 2020; Kostrikov et al., 2022); value-aware ICRL (Tarasov et al., 2025); Bayesian context use (Berkes et al., 2026).
Horizon or Resource	Episode length; discount; budget; latency; context length.	A context-use rule that works with long horizons fails when adaptation opportunities are scarce.	Short-horizon shift and delayed-feedback stress tests (Grigsby et al., 2024a;b; Elawady et al., 2024); context-budget and retrieval ablations (Schmied et al., 2024).

These sources of change stress different failure modes of context use. Reward and preference shifts are the cleanest validity test: a trajectory that was optimal before a reward switch can become actively misleading afterward, especially for models trained to imitate high-return histories. Dynamics and observation shifts test whether the agent infers latent environment structure from context rather than replaying familiar action sequences or overfitting to superficial tokens. Action-semantics shifts are particularly sharp for sequence models, because the same action token or embedding may denote a different intervention after the interface changes. Task-distribution, behavior-policy, and resource shifts instead alter the credibility of accumulated evidence: the agent must decide whether demonstrations, retrieved trajectories, and old priors still describe the current decision problem.

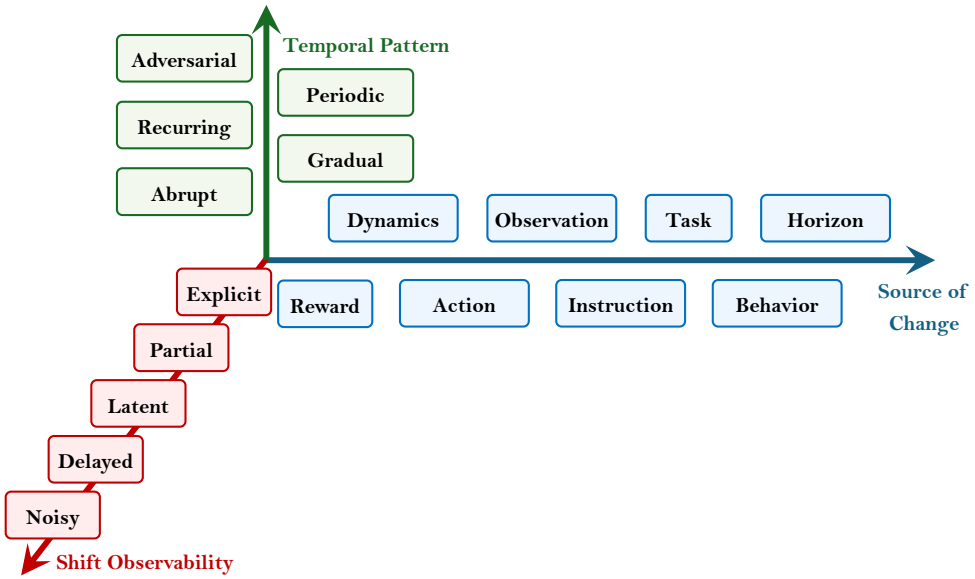

Figure 5: Compact map of three independent axes for describing non-stationarity in ICRL. Short labels indicate representative categories; Tables 5–7 give the full definitions.

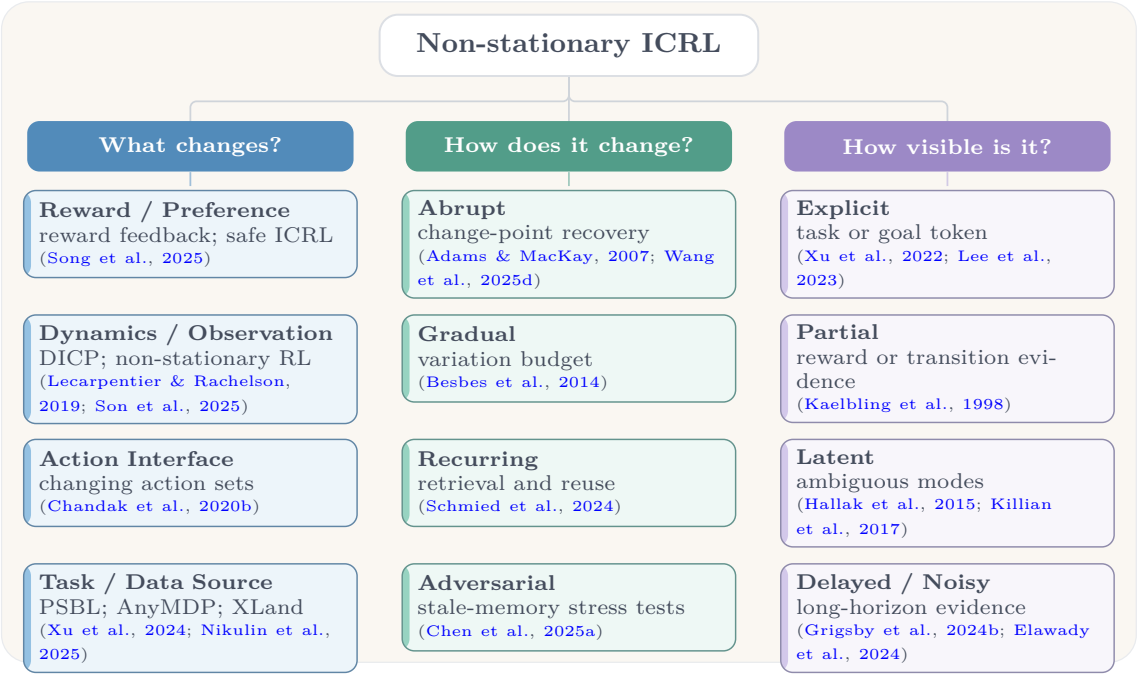

Figure 6: Tree-structured taxonomy for non-stationary ICRL. The leaves are representative problem families rather than mutually exclusive classes.

4.2 Temporal Patterns of Change

The second axis is temporal structure. An abrupt change asks for change-point detection, gradual drift for tracking, recurring modes for memory indexing, and adversarial or strategically timed changes for robustness. Table 6 links these patterns to the context-management implication and to metrics that reveal the corresponding failure.

Table 6: Taxonomy axis 2: temporal patterns of non-stationarity.

Temporal Pattern	Context-Management Demand	Representative Evaluation Signal
Abrupt Shift	Detect the change quickly; suppress pre-shift evidence before it dominates post-shift decisions.	Detection delay and changepoint quality (Adams & MacKay, 2007; Chen et al., 2019); restart or adaptive-window behavior (Auer et al., 2019); post-shift AUC in continual ICRL (Wang et al., 2025d).
Gradual Drift	Track a moving target; smooth noisy evidence without locking onto stale estimates.	Bandit and MDP dynamic regret (Besbes et al., 2014; Cheung et al., 2020; Fei et al., 2020); refined tracking bounds (Feng et al., 2023; Wei & Luo, 2021; Mao et al., 2021); kernelized and sliding-window variants (Domingues et al., 2021; Garivier & Moulines, 2011; Russac et al., 2019; Trovo et al., 2020).
Recurring Modes	Preserve mode-specific memory; retrieve old context when a regime returns without causing interference.	Re-adaptation latency and old-mode reuse (Steinparz et al., 2022); retrieval-augmented ICRL quality (Schmied et al., 2024; Sridhar et al., 2024); retrieval-based RL memory quality (Goyal et al., 2022; Humphreys et al., 2022).
Expanding Support	Recognize tasks, observations, or actions that were absent from early context.	Out-of-support return and calibration under novelty (Xu et al., 2024; Wang et al., 2025a).
Seasonal or Periodic Change	Learn when old context becomes useful again rather than merely old.	Performance by cycle phase; retrieval precision across cycles (Khetarpal et al., 2022; Abel et al., 2023; Wolczyk et al., 2022).
Adversarial Or Worst-Case Shift	Avoid being exploited by stale, strategically timed, or poisoned context.	Worst-case regret, robust return, and stale-context stress tests (Moeini et al., 2025a; Chen et al., 2025a).

These temporal patterns imply different forms of memory. Abrupt non-stationarity favors short-term evidence and explicit forgetting. Gradual drift favors smoothing and recency-weighted estimation. Recurring modes favor long-term memory with mode-specific retrieval. Expanding support favors exploration and uncertainty calibration. A single context-length ablation cannot distinguish these cases.

4.3 Shift Observability

The third axis asks how directly the agent can observe that a shift has occurred. At one extreme, the change is announced by a task descriptor, instruction, goal token, or other explicit signal. At the other, it must be inferred from prediction errors, transition mismatches, or delayed outcomes that may appear only after consequential actions have already been taken. This distinction is central for ICRL because the same context must often serve two roles at once: diagnosing the current regime and choosing actions under that diagnosis. Table 7 separates the resulting observability regimes and the abilities they stress.

Observability also changes what a benchmark can legitimately claim to measure. When every shift is announced by a clean descriptor, success mostly demonstrates task conditioning, instruction use, or prompt parsing. When shifts are hidden behind sparse delayed feedback, the same label instead covers exploration, belief updating, and long-horizon credit assignment. Both regimes are valuable, but they should not be treated as interchangeable evidence for a single adaptation capability.

4.4 A Compact Naming Scheme

For reporting experiments and organizing the literature, it is useful to name a non-stationary setting with a compact three-part notation:

SOURCE / TEMPORAL PATTERN / OBSERVABILITY.

Table 7: Taxonomy axis 3: observability regimes.

Regime	Shift Evidence	ICRL-Specific Failure Mode
Explicitly Signaled	Task ID; instruction; goal token; system message.	The model may ignore the signal as irrelevant metadata, or overfit to the signal without using interaction evidence (Xu et al., 2022; Yao et al., 2023).
Partially Observable	Rewards, transitions, or observations reveal the change only after several interactions.	Adaptation speed depends on belief-state memory (Kaelbling et al., 1998; Heess et al., 2015; Ni et al., 2022), algorithm-distillation or supervised ICRL context use (Laskin et al., 2023; Lee et al., 2023), and exploration scaffolds (Dai et al., 2024; Krishnamurthy et al., 2024).
Latent And Confounded	The same observations remain compatible with multiple shift explanations.	Context must preserve competing hypotheses in contextual and hidden-parameter MDPs (Hallak et al., 2015; Killian et al., 2017; Doshi-Velez & Konidaris, 2016), epistemic or task-inference POMDPs (Ghosh et al., 2021; Humplik et al., 2019; Ren et al., 2020), latent-belief meta-RL (Rakelly et al., 2019; Zintgraf et al., 2020), and value-aware uncertainty models (Berkes et al., 2026).
Delayed Feedback	Shift evidence arrives only after a long horizon or after irreversible early choices.	Early post-shift actions can be dominated by stale pre-shift context in long-context meta-RL (Grigsby et al., 2024a;b), S4-style ICRL (Lu et al., 2023), and long-term memory settings (Pasukonis et al., 2022).
Noisy Or Corrupted Signal	Change indicators are unreliable, missing, or adversarially perturbed.	The context manager must estimate trustworthiness, not simply recency (Zisman et al., 2024; Chen et al., 2025b).

For example, an unannounced reward switch is 'reward/abrupt/partial', a recurring hidden dynamics mode is 'dynamics/recurring/latent', and an action remapping announced by an interface token is 'action/abrupt/explicit'. The notation is not a substitute for a full benchmark description. Its purpose is more modest: it makes the adaptation claim legible before the details of the environment, dataset, or model architecture are introduced.

4.5 Using the Taxonomy Analytically

The taxonomy is intended to discipline comparisons, not merely to label benchmarks. Success on 'reward/abrupt/explicit' tasks may reflect effective use of a goal token, whereas success on 'dynamics/gradual/partial' tasks requires tracking transition evidence over time. Performance on 'task/recurring/latent' tasks may come from memory indexing and mode reuse; performance on 'behavior-policy/abrupt/noisy' tasks instead depends on judging whether demonstrations and retrieved trajectories are trustworthy. These are distinct capabilities, so they should support distinct claims.

The same taxonomy suggests a minimal stress-test rule. For each claimed source of non-stationarity, evaluation should include at least one temporal pattern where stale context is harmful and at least one observability setting where trivial task-ID conditioning is unavailable. For each claimed memory mechanism, evaluation should include one case where old context should be discarded and one where old context should be recovered. This two-sided requirement prevents degenerate solutions from looking general: a recent-only policy can excel on one-way drift, and an append-all policy can excel on stationary or recurring tasks, but neither by itself solves context validity.

5 The Role of Context Under Non-Stationarity

5.1 Context Is Evidence, Memory, and Intervention

In stationary ICRL, context is usually treated as evidence about a fixed latent task. Relevant examples improve inference; irrelevant ones are expected to be ignored. Non-stationarity breaks that tidy picture. The context is now a time-stamped record of decisions, rewards, and observations that may have been generated under different regimes. It is also an intervention: the retrieved or summarized context changes what the frozen model attends to, and therefore changes the policy it implements. This is why external-memory work in language agents is relevant as a systems analogy even when it is not itself reinforcement learning: MemoryBank (Zhong et al., 2024), LongMem (Wang et al., 2023b), and MemGPT (Packer et al., 2023) make persistent memory explicit. Retrieval-augmented language models (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022), nearest-neighbor language models (Khandelwal et al., 2019), and in-context retrieval (Ram et al., 2023) make retrieval and evidence selection explicit. The same evidence-management problem appears in older sequential decision theory under different names: POMDP belief states (Kaelbling et al., 1998), contextual MDP latent contexts (Hallak et al., 2015), hidden-parameter MDP embeddings (Killian et al., 2017; Doshi-Velez & Konidaris, 2016), task-inference variables (Humplik et al., 2019; Ren et al., 2020), and Bayesian changepoint posteriors (Adams & MacKay, 2007) all encode the idea that recent observations should update beliefs about the current regime. Non-stationary ICRL differs in its engineering constraint: the belief update may be only an implicit computation inside a fixed sequence model, retrieval rule, or prompt summary.

Context therefore plays three roles at once. It is evidence about rewards, dynamics, goals, preferences, and uncertainty. It is memory, preserving information that may become useful again if a mode recurs. It is also a control input to a frozen policy, because changing the context can change behavior without changing weights. Non-stationary ICRL is hard because these roles do not always agree. The best evidence for the current mode may require discarding long-term memory; the best long-term memory may exceed the context budget; the most salient trajectory may be precisely the one that should no longer be trusted. Thus the core question is not whether the agent remembers history, but whether it can infer when history has stopped supporting the current policy. In a finite-budget system, this tradeoff can be written as a context-selection problem:

$$C_t^\star = \arg\max_{C \subseteq \mathcal{M}_{t-1},\ |C| \le b_t} \mathbb{E}[V_t^{\pi_\theta}(s_t \mid C)] - \lambda \operatorname{cost}(C), \tag{15}$$

where $\mathcal{M}_{t-1}$ is the memory available before acting, b_t is the context budget, and the cost term can represent tokens, retrieval calls, latency, or risk of stale evidence. Most existing systems approximate this objective implicitly; non-stationary evaluation should make the approximation visible.

5.2 Failure Modes

Table 8 summarizes common failure modes that are easy to miss in stationary evaluation.

Several of these failures come from context management rather than from the policy class itself. A sufficiently expressive model can still fail when its prompt is dominated by invalid evidence, when retrieval keys optimize surface similarity, or when compression removes a rare but decisive reward event. Conversely, a modest model can look robust if it receives clean, recent, mode-specific evidence. The policy and the context pipeline therefore have to be evaluated together.

The failure modes also explain why longer context is an ambiguous intervention. Increasing the context window can improve adaptation by making more evidence available, but it can also increase the fraction of invalid evidence, make attention more diffuse, and amplify high-salience pre-shift trajectories. The right comparison is therefore not "short context versus long context" in isolation. It is short context, full history, recent-only history, oracle-truncated history, learned retrieval, and adversarially contaminated retrieval under the same shift schedule. Only this comparison reveals whether scale is solving adaptation or merely hiding the cost of invalid context. This point is consistent with the broader ICL literature: training distributions (Chan et al., 2022), transformer function learning (Garg et al., 2022), gradient-free or probabilistic ICL mechanisms (Kirsch et al., 2022; Wies et al., 2023), demonstration ordering (Min et al., 2022), and contextual learning

Table 8: Failure modes induced by stale or invalid context.

Failure Mode	Mechanism	Symptom	Diagnostic Ablation
Stale Imitation	The model imitates high-return pre-shift trajectories after the reward or dynamics changed.	Strong performance before shift, slow or no recovery afterward.	Insert old optimal demonstrations into a post-shift prompt for decision-transformer (Chen et al., 2021), algorithm-distillation (Laskin et al., 2023), or suboptimal-ICRL (Dong et al., 2026) settings.
Context Inertia	Long context dominates new evidence.	New rewards are observed but behavior changes late.	Vary pre-shift context length while holding post-shift evidence fixed in AMAGO-style agents (Grigsby et al., 2024a;b), S4-style ICRL (Lu et al., 2023), or RELIC-style long-context agents (Elawady et al., 2024).
Mode Collapse	The agent maps several latent modes to one familiar behavior.	Recurring or mixed modes are handled as if stationary.	Evaluate mode-conditioned returns and confusion matrix for latent-belief meta-RL methods such as PEARL (Rakelly et al., 2019) and VariBAD (Zintgraf et al., 2020).
Retrieval Contamination	External memory retrieves superficially similar but invalid trajectories.	Retrieval helps in stationary tasks but hurts after interface or reward shift.	Compare similarity-only retrieval with validity-filtered retrieval in RA-DT (Schmied et al., 2024) and REGENT-style systems (Sridhar et al., 2024).
Compression Loss	Summaries discard the change evidence needed for adaptation.	Compressed context performs well on average but misses rare shifts.	Preserve raw post-shift events while compressing old context in memory systems (Zhong et al., 2024; Wang et al., 2023b; Packer et al., 2023), communication-based compression (Martinez-Lopez et al., 2025), or filtered-history training (Chen et al., 2025b).
Action-Token Mismatch	The model assumes old action semantics still hold.	Correct high-level plan but wrong low-level actions.	Permute, mask, or rescale actions at known change points (Chandak et al., 2020b; Sinii et al., 2024).
Over-Forgetting	The model discards old context that becomes useful again.	Slow recovery when a previous mode returns.	Use recurring-mode benchmarks and measure re-adaptation latency (Khetarpal et al., 2022; Steinparz et al., 2022; Schmied et al., 2024).

dynamics (Akyurek et al., 2024) can encourage in-context inference, memorization, or failure depending on how examples are clustered, repeated, ordered, or distributed across tasks. For ICRL, the analogous distributional feature is not only task diversity but whether context windows contain the right conflicts: pre-shift versus post-shift evidence, high-return stale trajectories versus low-return current evidence, and recurring modes separated by misleading intervening data.

Retrieval and long context make the problem sharper rather than eliminating it. Retrieval-augmented language systems (Izacard et al., 2022), retrieval-based RL systems (Goyal et al., 2022; Humphreys et al., 2022), and retrieval-augmented ICRL systems (Schmied et al., 2024; Sridhar et al., 2024) show that external memories can improve sample efficiency and reuse, but they also introduce a new decision variable: which memory should be treated as evidence for the current regime. Long-context studies similarly warn that simply appending more examples can dilute or distort in-context learning when the relevant evidence is far from the query or mixed with distractors (Li et al., 2024b; Bertsch et al., 2025). For non-stationary ICRL, this means retrieval precision and attention allocation should be reported as adaptation mechanisms, not only as engineering choices.

5.3 The Context Lifecycle

A useful system-level view is to give context a lifecycle:

1. **Write:** which interactions, demonstrations, and feedback are stored?
2. **Index:** how are stored items keyed for future retrieval?
3. **Retrieve:** which items are selected for the current decision?
4. **Compress:** how are long histories summarized within a finite context budget?

5. **Trust or weight:** after evidence is available to the policy, how much influence should each item have on action selection?
6. **Forget or isolate:** when should context be removed, downweighted, or stored in a mode-specific partition?

These entries should not be read as a mandatory serial architecture. Write, index, retrieve, compress, and forget or isolate are often discrete context-manager operations: a system can log an item, create a key, retrieve a subset, summarize a history, or move an item out of the active memory. Trust is different. It denotes the *influence assignment* given to available evidence once it reaches the decision computation. In an explicit system, trust may be implemented by a reliability score, uncertainty gate, value-fusion weight, safety filter, or regime posterior. In a plain transformer or recurrent policy, it may be an emergent property of attention, hidden-state dynamics, value prediction, and decoding rather than a separately exposed module. The lifecycle is therefore an analytic decomposition of responsibilities, not a claim that every ICRL agent contains six separate processing blocks.

This distinction clarifies the relation between retrieve and trust. Retrieval answers a candidate-selection question: which memories or summaries enter the context window at all? Trust answers an influence question: conditional on evidence being present, how strongly does it shape the action distribution? A retrieval-augmented agent can retrieve the right item and still fail by over-weighting a stale high-return trajectory; conversely, a long-context agent with no explicit retriever can still succeed if attention or value inference suppresses invalid evidence. For non-stationary ICRL, both questions should be reported when possible: selection quality through retrieval or context-ablation metrics, and influence quality through attention, gating, counterfactual prompting, value sensitivity, or stale-evidence stress tests. Under non-stationarity, every operation or influence facet in this lifecycle can become either an adaptation mechanism or a failure point. Figure 7 visualizes this lifecycle, and Figure 8 states the design question and non-stationary risk at each operation or facet.

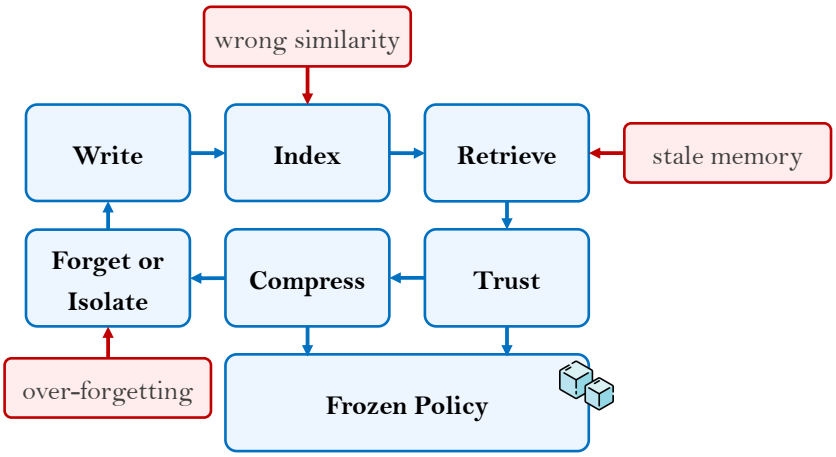

Figure 7: Context lifecycle. Under non-stationarity, every operation can improve adaptation or inject invalid evidence.

5.4 Implication for Survey Organization

The rest of the survey therefore organizes methods by how they manage context. Algorithm distillation learns from histories that contain improvement (Laskin et al., 2023). Supervised ICRL learns mappings from contextual examples to decisions (Lee et al., 2023; Lin et al., 2024; Huang et al., 2024a). Retrieval-augmented methods select external memories (Schmied et al., 2024; Sridhar et al., 2024; Goyal et al., 2022). Value-aware and Bayesian methods combine context with priors (Tarasov et al., 2025; Liu et al., 2026; Berkes et al., 2026). Model-based methods infer dynamics or planning computations in context (Son et al., 2025; Xu et al., 2026c). World-model and communication-based methods compress experience (Zhong et al., 2024; Wang et al., 2023b; Packer et al., 2023; Martinez-Lopez et al., 2025). Reflection-style agents use textual or symbolic summaries of feedback (Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2023a; Song et al., 2025).

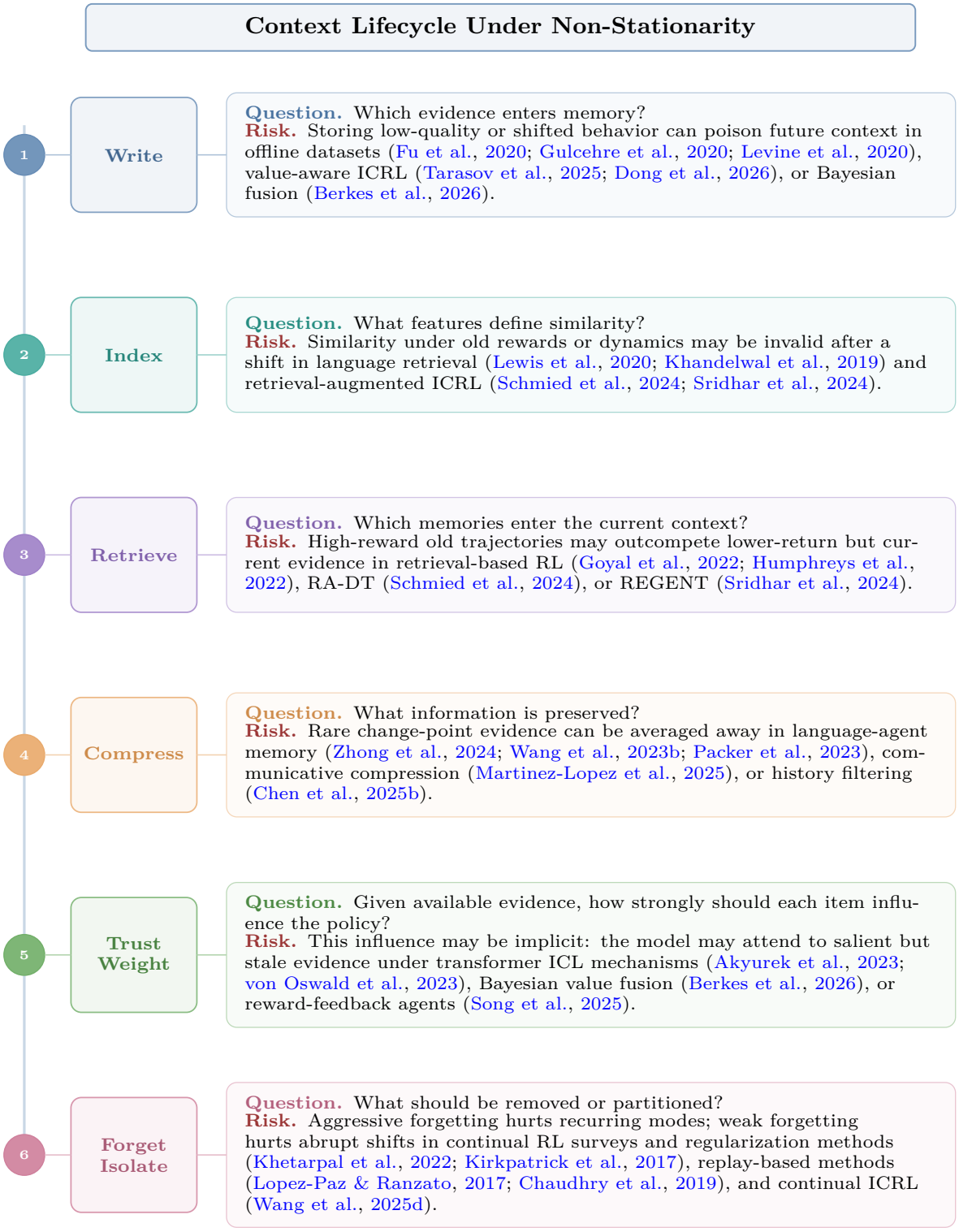

Figure 8: Context lifecycle operations and their non-stationary design questions. The layout follows a taxonomy-style view: each lifecycle operation is treated as a stage whose design decision determines how stale or shifted evidence can affect the policy.

For non-stationarity, the key question for each family is the same: what makes the context valid, and what happens when it is not?

This lifecycle view suggests a decomposition of responsibility. The write operation controls what evidence can ever be used. The index and retrieve operations control which evidence is considered similar or relevant enough to expose. The compress operation controls which rare events survive a finite budget. The trust facet controls how strongly available evidence affects the frozen policy, even when that weighting is implicit in the model rather than implemented as a separate module. The forget or isolate operation controls whether old evidence disappears, is downweighted, or remains available for future recurrence. A non-stationary ICRL paper need not solve all parts of the lifecycle, but it should state which responsibility its contribution improves, which parts are held fixed, and whether trust is explicitly modeled or only inferred from behavior.

6 Method Families Through the Lens of Context Management

Methods for ICRL differ less by their names than by where they place the burden of adaptation. Some rely on the sequence model to infer the right update from raw history. Others move part of the problem into retrieval, summarization, memory partitioning, value estimation, or model inference. For non-stationary settings, the useful organizing question is simple: when the world changes, what mechanism decides which context still matters?

Figure 9 summarizes the resulting landscape. The horizontal axis separates implicit history use from explicit context selection through retrieval, summaries, and memory policies. The vertical axis separates raw behavioral evidence from value- or model-aware evidence. The empty region is as informative as the occupied one: many systems scale context length or memory access without estimating validity, while value- and model-aware systems still depend on the evidence selected for them.

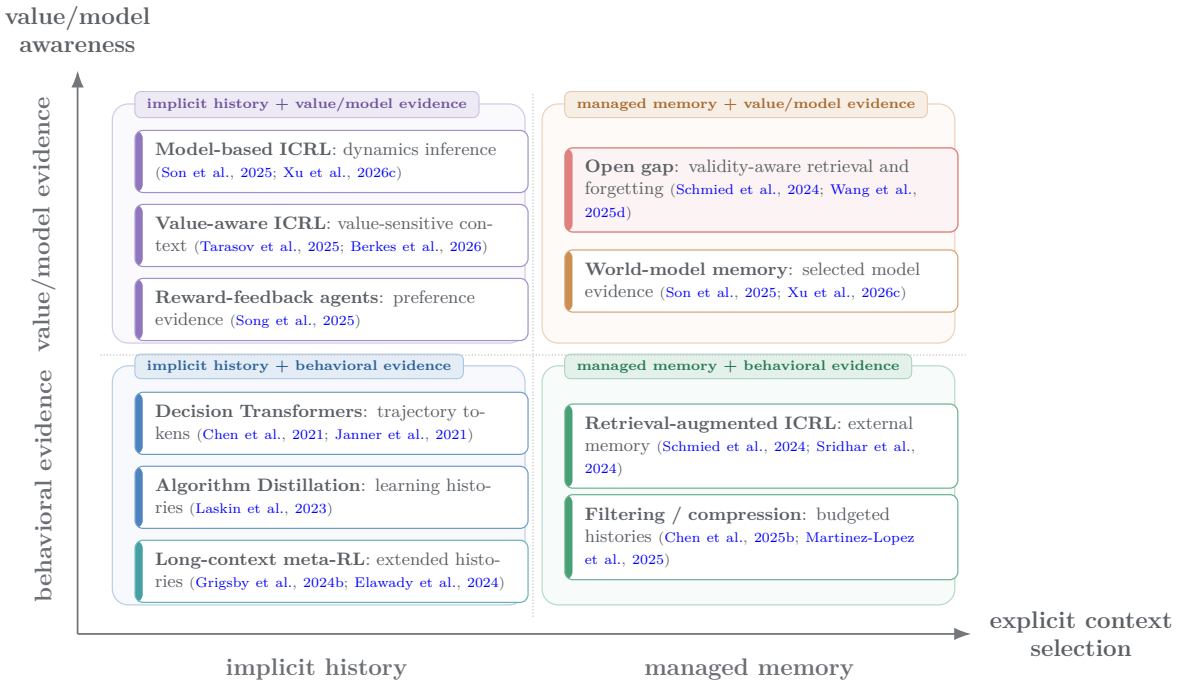

Figure 9: Method landscape organized by context-selection explicitness and value/model awareness.

6.1 Algorithm Distillation and Learning Histories

Algorithm distillation trains a sequence model on histories generated by a learning algorithm, with the aim that the trained model internalizes the algorithm's within-lifetime improvement (Laskin et al., 2023). For non-stationary ICRL, this family is important because learning histories can include exploration, failure, reward feedback, and policy improvement. If the pretraining histories include changes in task or reward, the model may learn a pattern of adaptation through context. If they are stationary, the model may instead learn that older high-return behavior should remain trusted. This distinction mirrors broader ICL evidence that models can shift between memorization, system identification, and genuine in-context learning depending on distributional burstiness (Chan et al., 2022), probabilistic or gradient-free ICL structure (Kirsch et al., 2022; Wies et al., 2023), example ordering (Min et al., 2022), and contextual learning dynamics (Akyurek et al., 2024).

The critical design variable is the distribution of histories. Histories generated by a strong learner on stationary tasks provide clean examples of improvement, but they may not teach change detection. Histories generated under weak data or noisy policies may better reflect deployment, but they can also create ambiguous supervision. Cross-episodic curricula (Shi et al., 2023), noise distillation (Zisman et al., 2024), contrastive context encoders (Wang et al., 2021a), and filtered learning histories (Chen et al., 2025b) all modify the his-

tory distribution to make within-context improvement easier to learn. From the non-stationary perspective, filtering should not merely remove low-return episodes; it should preserve transitions that reveal when old context became invalid. ICPRL adds a physically grounded variant of this idea by training a vision-language policy on multi-episode trial-and-error histories, so that later actions can condition on previous physical failures and successes without test-time weight updates (Xu et al., 2026c). The adjacent meta-exploration literature is useful here because it explicitly treats early actions as information-gathering, not only as return maximization, through exploration objectives (Stadie et al., 2018; Dorfman et al., 2020; Kamienny et al., 2020), decoupled exploration (Liu et al., 2021), hyperstate inference (Zintgraf et al., 2021), or first-explore strategies (Norman & Clune, 2024).

6.2 Supervised ICRL and Decision-Pretrained Transformers

Supervised pretraining can induce ICRL when the model is trained on datasets where context examples identify a latent decision problem (Lee et al., 2023; Lin et al., 2024). Decision Transformer (Chen et al., 2021), Trajectory Transformer (Janner et al., 2021), RvS (Emmons et al., 2022), prompting variants (Xu et al., 2022), generalized decision transformers (Furuta et al., 2022), chain-of-hindsight or agentic variants (Liu & Abbeel, 2023), and ICDT-style models (Huang et al., 2024a) provide the broader sequence-modeling substrate. These methods are attractive because they can use offline data and scale with standard supervised objectives. Recent scaling and routing variants, including multi-game decision transformers (Lee et al., 2022), online decision transformers (Zheng et al., 2022), hierarchical prompt transformers (Wang et al., 2025f), Vintix-style action-model formulations (Polubarov et al., 2025), and mixture-of-experts formulations (Wu et al., 2025), further test whether the same substrate can support broader task mixtures. Theoretical results on in-context policy improvement clarify why this family is more than sequence imitation: under structured conditions, a transformer can use trajectory context to implement a policy-improvement procedure rather than simply replaying the behavior policy (Liang & Lai, 2026). This gives supervised and decision-pretrained ICRL a principled bridge to classical policy iteration, but it also exposes the non-stationary missing assumption. Policy improvement is well grounded when the trajectories describe the same decision process; after reward, dynamics, or action-semantics shifts, the model must first infer which trajectories are still valid before any improvement step is meaningful. Their weakness under non-stationarity is objective mismatch. Next-action prediction can reward copying behavior from context even when that behavior is no longer optimal. Return conditioning can encode a desired outcome, but it does not by itself solve which past examples are relevant after a shift.

The non-stationary version of supervised ICRL should therefore include contrastive context designs. For example, prompts can contain pre-shift and post-shift trajectories with identical observations but different rewards. Training can require the model to attend to the most recent valid evidence, infer a latent mode, or predict change-point indicators. Without such designs, strong held-out-task performance may reflect static generalization rather than adaptation to changing environments. Demonstration-conditioned sequential decision studies show that task diversity, dataset size, and trajectory structure matter for in-context generalization, but those results still need explicit stale-context and shift tests before they establish non-stationary adaptation (Raparthy et al., 2024). The weak-label and suboptimal-data variants are particularly important for this survey. Offline ICRL rarely has perfect learning histories or optimal labels at scale; offline RL coverage issues (Levine et al., 2020), conservative and implicit value learning (Kumar et al., 2020; Kostrikov et al., 2022), return-conditioned value learning (Brandfonbrener et al., 2022), Q-learning-style ICRL (Tarasov et al., 2025), suboptimal-data ICRL (Dong et al., 2026), random-policy state-action distillation (Chen & Paternain, 2024), and TD-style ICRL (Wang et al., 2024a) can turn imperfect historical trajectories into useful adaptation signals, but only if the context window exposes which behavior was suboptimal under the current regime.

6.3 Efficient Sequence Backbones

The default ICRL backbone is a causal transformer, but long non-stationary lifetimes make the quadratic cost of attention a practical bottleneck. Structured state-space models, recurrent-memory transformers, gated recurrent attention, and Mamba-style sequence models therefore deserve explicit treatment. S4-style ICRL (Lu et al., 2023) and Decision S4 (David et al., 2023) show that state-space sequence models can

support long-horizon memory and sequence-based RL, while Decision Mamba (Ota, 2024; Huang et al., 2024b) and Mamba-based scaling of algorithm distillation (Beaussant & Mounsif, 2025) test selective state-space decision models. Related selective-state, recurrent-memory, and xLSTM-style models (Gu & Dao, 2023; Smith et al., 2023; Bulatov et al., 2022) and transformer-RL memory studies (Beck et al., 2024; Ni et al., 2024) test whether recurrent state can replace or complement transformer attention in decision modeling. For non-stationarity, the architectural question is not only speed. A compressed recurrent state can make old information cheap to carry, but it can also make invalid information hard to remove. Thus efficient backbones should be evaluated with the same stale-context, oracle-reset, and recurring-mode tests as transformer systems. Recent non-stationary ICL theory (Qin et al., 2026) makes this architectural point explicit: adaptive filtering (Sayed, 2011), gating (Yang et al., 2023b; Katsch, 2023), retention (Sun et al., 2023), and recurrent attention (Peng et al., 2024) can be interpreted as learnable forgetting factors when the target function changes over time.

6.4 Long-Context Meta-RL Agents

Long-context agents such as AdA-style systems (Bauer et al., 2023), AMAGO-style systems (Grigsby et al., 2024a;b), RELIC-style systems (Elawady et al., 2024), transformer meta-RL systems (Melo, 2022), and hierarchical meta-RL agents (Shala et al., 2024) use recurrent, transformer, or external-memory mechanisms to solve adaptive RL problems across long horizons and open-ended task spaces. They are natural candidates for non-stationary ICRL because they can integrate long histories and delayed feedback. However, long context is not automatically better. When the environment changes, a longer history increases both evidence and contamination. The relevant question becomes whether the model has learned attention patterns that separate current-mode evidence from stale-mode evidence. Large context, open-ended learning (Stooke et al., 2021), dataset efforts such as XLand and XLand-100B (Nikulin et al., 2025), broader generalist-agent sequence models (Reed et al., 2022; Polubarov et al., 2025; 2026), and episodic or long-term recall systems (Ritter et al., 2018; Pasukonis et al., 2022) make this question sharper because they scale learning histories far beyond toy settings while still requiring tests of whether scaled histories teach invalidation rather than only broad task recognition.

Long-context meta-RL also raises a reporting issue. If a model is trained online with RL across non-stationary episodes, then test-time adaptation may be partly due to learned recurrent dynamics rather than explicit context retrieval or prompting. This is still within our ICRL boundary when parameters are fixed at deployment, but papers should report the lifetime state, context window, reset rules, and whether memory persists across episodes.

6.5 Retrieval-Augmented ICRL

Retrieval-augmented methods externalize part of the context problem. Instead of fitting the entire history into the model's input, they retrieve trajectories, transitions, or summaries from memory in RA-DT (Schmied et al., 2024), REGENT (Sridhar et al., 2024), in-context Q-learning with retrieval (Xu et al., 2026a), and earlier retrieval-based RL systems (Goyal et al., 2022; Humphreys et al., 2022). This is one of the most promising families for non-stationarity because retrieval can, in principle, select mode-relevant memories and ignore stale ones. It is also one of the most fragile. If retrieval is based on surface similarity, it may select trajectories from the wrong reward regime. If retrieval is based on return, it may prefer high-return old behavior that is now invalid. If retrieval is based on recentness, it may fail when an old mode recurs.

For non-stationary ICRL, retrieval should be evaluated with validity metrics, not only downstream return. Precision of retrieved context, counterfactual performance without retrieval, and adversarial stale-memory tests are needed. The best retrieval key may combine observations, actions, rewards, inferred latent modes, uncertainty, and temporal metadata. This creates a bridge between ICRL and memory systems: retrieval is not just a scaling trick, but a learned or designed adaptation policy. The language-model retrieval literature is a useful warning because nearest-neighbor language models (Khandelwal et al., 2019), retrieval-augmented generation and pretraining (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022), in-context retrieval (Ram et al., 2023), and Atlas-style retrieval (Izacard et al., 2022) can improve prediction while leaving relevance and trust as separate design problems.

6.6 Value-Aware and Bayesian Context Use

Value-aware methods attempt to make ICRL less dependent on imitation of raw trajectories. Q-learning objectives can improve offline ICRL (Tarasov et al., 2025) by teaching the model value-sensitive behavior rather than pure behavior cloning, building on the broader offline RL lesson that conservative and implicit value regularization matter under support mismatch (Kumar et al., 2020; Kostrikov et al., 2022). Scalable in-context Q-learning (Liu et al., 2026) and compositional Q-learning (Xu et al., 2026a) extend this idea by conditioning value estimates on compact context or retrieved transitions. Bayesian-fusion approaches combine context with value priors, aiming to decide how much to trust contextual evidence under uncertainty (Berkes et al., 2026).

This family is especially relevant under reward and dynamics shifts. If the model estimates uncertainty over values, it can downweight context that conflicts with the current reward signal. If it only predicts next actions, it may reproduce old behavior even after feedback changes. The limitation is that value-aware objectives still need the right evidence. A Q-function conditioned on invalid retrieved context can still produce invalid actions. Thus value awareness and context validity are complementary: values help interpret context, while context management determines which evidence values are conditioned on.

6.7 Action-Space and Interface Adaptation

Action-space variation deserves separate treatment because it is both a scaling issue and a non-stationarity issue. Variable-action ICRL shows that models can be designed to generalize across discrete action spaces with different size, order, and semantic content (Sinii et al., 2024). For this survey, the non-stationary version is stricter: the action interface may change during a lifetime while old context still contains action tokens from the previous interface. This makes action embeddings, action descriptions, and interface metadata part of context validity rather than a mere implementation detail. Benchmarks that only switch rewards or tasks miss this failure mode.

6.8 Model-Based In-Context Planning

Model-based ICRL distills dynamics inference (Son et al., 2025) or planning computation (Xu et al., 2026c) into context-conditioned computation. Under dynamics shifts, this is a direct fit: the agent must infer how actions change the world from recent evidence and plan accordingly. Model-based context also gives a clearer diagnostic than action prediction. If a model predicts transitions correctly after a shift but still acts poorly, the failure may lie in planning. If it predicts old transitions after observing new ones, the failure is context adaptation. ICPRL is useful here because it separates an adaptive policy from a world model used for lookahead search in visual physics tasks (Xu et al., 2026c). For a non-stationary benchmark, the same architecture would need to show that the world-model-guided search reacts to changed physical parameters or action outcomes, rather than merely exploiting a static simulator prior. World-model work (Hafner et al., 2019), simulator-learning work (Yang et al., 2023a), Genie-style interactive generation (Bruce et al., 2024), and instructable simulators (Raad et al., 2024) provide adjacent mechanisms for this direction, but the survey treats them as relevant only when the learned model is used in a repeated decision loop with context-conditioned adaptation.

The challenge is compounding error. A model-based in-context planner may infer the wrong latent dynamics from stale history and then plan confidently under that wrong model. Non-stationary evaluation should therefore report both predictive accuracy and control performance across pre-shift, immediately post-shift, and recovered phases.

6.9 Context Compression and Communicative World Models

Finite context makes compression unavoidable. Communication- or world-model-based approaches can summarize interaction histories into compact language-agent memory (Zhong et al., 2024; Wang et al., 2023b; Packer et al., 2023), predictive video representations (Wu et al., 2024), or communicative world-model representations (Martinez-Lopez et al., 2025). Compression is useful under long horizons and recurring tasks,

but it can erase the rare evidence that marks a shift. The key question is not compression ratio alone, but preservation of decision-relevant change evidence. For example, a summary that preserves average reward may miss a sudden preference reversal; a summary that preserves the most recent transition may lose an older mode that will recur.

6.10 LLM and Foundation Agents as Extended ICRL Cases

LLM-based and foundation-agent systems are in scope only when the model participates in a sequential decision loop where context, actions, feedback, and later behavior are coupled. Recent work frames LLMs as in-context reinforcement learners under bandit feedback (Monea et al., 2024), reward feedback (Song et al., 2025), or online regret criteria (Park et al., 2024). In these settings, the frozen model weights are not the adaptation channel. The adaptation channel is the prompt, transcript, retrieved memory, reflection, executable skill, or summary that changes future actions after previous outcomes. ReAct-style agents (Yao et al., 2023) connect reasoning traces to environment actions, while Reflexion (Shinn et al., 2023), ExpeL (Zhao et al., 2024), Voyager (Wang et al., 2023a), and general computer control agents (Tan et al., 2025) store verbal reinforcement, experience summaries, executable skills, or interaction memories for later trials.

Non-stationarity makes this bridge case more than a systems detail. User preferences can drift, hidden instructions can change, reward feedback can be reinterpreted, tool schemas and website layouts can move, and old reflections can become stale policy hints. A memory may be semantically similar to the current situation but invalid because it was produced under a different reward function, tool interface, or user constraint. Natural-language summaries are especially risky because they compress many observations into declarative advice; after a shift, the advice can remain fluent and authoritative while becoming wrong. This is the LLM-agent analogue of stale trajectories in decision-transformer systems and stale retrieved transitions in retrieval-augmented ICRL.

The survey still treats these systems as extended cases rather than the core of ICRL. Many LLM-agent papers do not report scalar rewards, controlled action spaces, lifecycle curves, no-memory baselines, or context ablations. Success in an interactive language environment may reflect instruction following, prompt parsing, static task recognition, or retrieval of a near duplicate rather than reinforcement learning through context. Bandit studies provide an important caution: frozen LLMs often need explicit summaries (Krishnamurthy et al., 2024; Dai et al., 2024) or algorithmic scaffolding (Nie et al., 2025) to explore reliably in context. Multi-turn agent training methods such as PAPRIKA (Tajwar et al., 2025), ICAL (Sarch et al., 2024), ArCHer (Zhou et al., 2024), VAGEN (Wang et al., 2025c), SCoRe-style self-correction (Kumar et al., 2025), and cultural-accumulation systems (Cook et al., 2024) are therefore adjacent unless the experiment shows that repeated feedback changes future decisions through the context channel.

For a large-model result to support a non-stationary ICRL claim, the evidence should show post-shift behavioral change without parameter updates, compare memory policies, and expose stale-context failure. Useful baselines include no memory, last-k transcript, full transcript, summary memory, retrieval memory, oracle-current-rule context, and stale-memory adversaries. Small-scale reinforcement-tuned transformer studies that report adaptation to changing gridworld conditions are useful proof-of-concept evidence (Rentschler & Roberts, 2025), but a convincing large-model benchmark must also test whether verbal memories, tool histories, and retrieved examples remain valid under preference, instruction, reward, or interface shifts. Without environment interaction, scalar or preference feedback, and repeated decisions, LLM memory is better viewed as supervised or conversational in-context learning.

6.11 Method Matrix

Figure 10 condenses the preceding method discussion by mapping each family to its context mechanism, non-stationary strength, and main risk.

6.12 Cross-Family Lessons

Across these families, architecture is not the cleanest dividing line. Algorithm distillation and supervised ICRL ask the pretrained model to infer the meaning of context inside its forward pass. Retrieval and

Table 9: Large-model bridge cases for non-stationary ICRL.

Category	ICRL Entry Condition	Non-Stationary Risk
Decision Foundation Models / Generalist Sequence Agents	Trajectory context changes future actions with fixed parameters in broad decision-model settings (Reed et al., 2022; Polubarov et al., 2025; 2026).	Scale can mask stale-context failure when performance is reported only as broad task success.
LLM Reward-Feedback Agents	Feedback, reward, regret, or reflection changes later decisions in bandit or interactive settings (Monea et al., 2024; Song et al., 2025; Park et al., 2024).	Old verbal feedback or reflection can remain over-trusted after preference or reward shifts.
Tool-Use / Computer-Control Agents	Tool outcomes and action histories shape later tool calls or environment actions (Yao et al., 2023; Tan et al., 2025).	Tool APIs, website layouts, action schemas, or parser conventions can drift.
VLM / World-Model Agents	Visual observations and transition evidence update planning through context (Xu et al., 2026c; Martinez-Lopez et al., 2025).	Old physical, visual, or causal evidence can mislead after dynamics or observation shifts.
Long-Memory Agents	Retrieved memories, summaries, or executable skills influence later decisions (Zhong et al., 2024; Wang et al., 2023b; Packer et al., 2023; Wang et al., 2023a).	Similar memories may be invalid under the current regime, especially when summaries hide when evidence was collected.

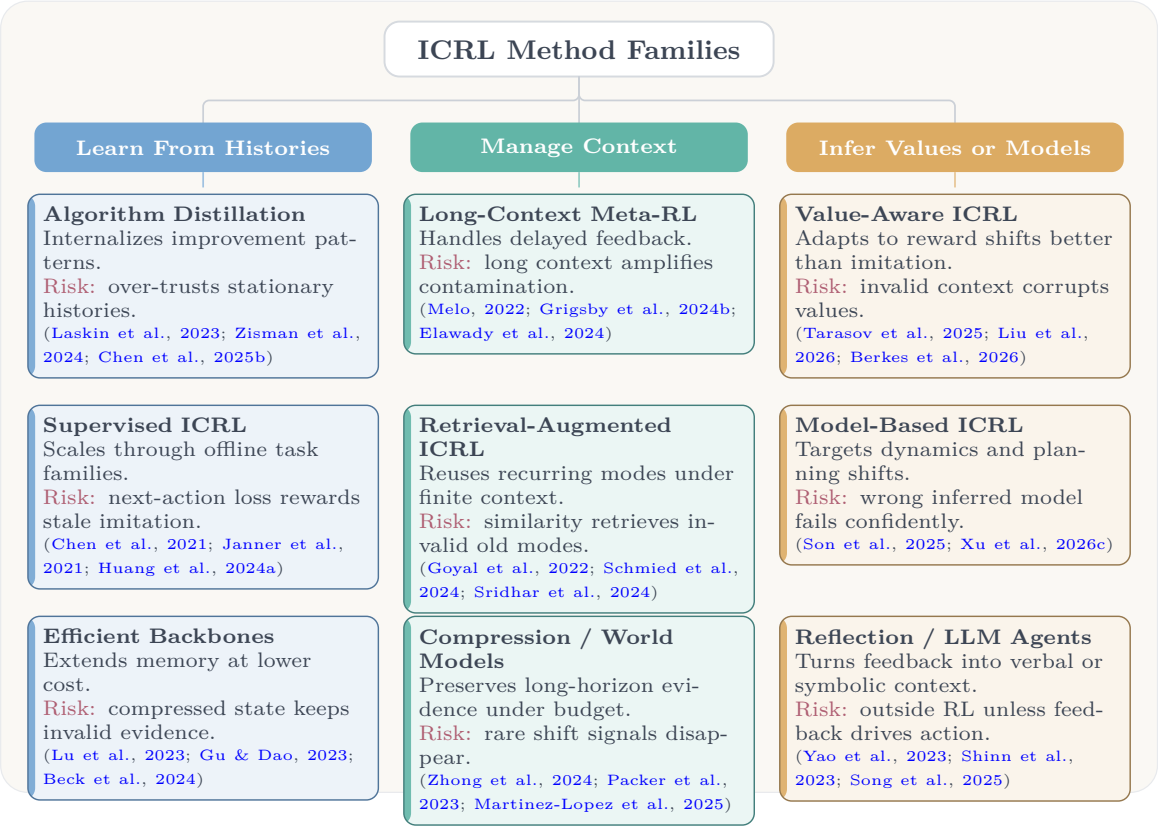

Figure 10: Method families organized by context-management mechanism. Each leaf states the adaptation advantage and the main non-stationary failure mode.

compression make the context pipeline more explicit: the model can adapt only to evidence that was selected and preserved. Value-aware and model-based methods change what the evidence means, treating context as information about values, dynamics, or uncertainty rather than as behavior to imitate. These choices are complementary under non-stationarity, but each needs different evidence. Table 10 turns this point into a standard for evaluating claims about each method family.

The common pattern is that useful systems need three things at once. Their training histories must contain invalidation events, otherwise old evidence is always presented as trustworthy. Their context managers need validity cues–time, shift likelihood, reward compatibility, uncertainty, or latent-mode metadata–instead of relying only on similarity. Their evaluations must test both forgetting and reuse. Aggressive forgetting can look good after an abrupt shift and fail when a mode recurs; preserving everything can look good on recurrence and fail after an irreversible preference change. The target is not more memory in the abstract, but controlled influence of memory.

Table 10: Required evidence for method families under non-stationarity.

Family	**Necessary Demonstration**	**Weak Evidence That Is Often Insufficient**
Algorithm Distillation	Learns to change behavior after context invalidation in AD (Laskin et al., 2023), contrastive context encoders (Wang et al., 2021a), or continual ICRL (Wang et al., 2025d), not only after ordinary trial-and-error.	Improvement on stationary held-out tasks.
Supervised ICRL	Uses conflicting prompt evidence according to current rewards or dynamics in supervised ICRL (Lee et al., 2023; Lin et al., 2024), TD-style ICRL (Wang et al., 2024a), or suboptimal-data ICRL (Dong et al., 2026).	Higher return with more demonstrations.
Efficient Sequence Backbones	Show that retained hidden state or compressed sequence state can be reset, gated, or corrected after shifts in S4-style models (Lu et al., 2023; David et al., 2023), Mamba-style models (Ota, 2024; Gu & Dao, 2023), or gated attention (Yang et al., 2023b).	Faster inference or longer context on stationary tasks.
Long-Context Meta-RL	Separates useful delayed evidence from stale early evidence in AMAGO-style systems (Grigsby et al., 2024a;b) and long-term memory settings (Pasukonis et al., 2022).	Better performance with longer memory.
Retrieval-Augmented ICRL	Retrieves valid memories when similar invalid memories exist in retrieval-based RL (Goyal et al., 2022; Humphreys et al., 2022), RA-DT (Schmied et al., 2024), or REGENT (Sridhar et al., 2024).	Retrieval beats no-retrieval on stationary tasks.
Value-Aware ICRL	Updates value estimates when reward semantics shift in Q-learning ICRL (Tarasov et al., 2025), scalable in-context Q-learning (Liu et al., 2026), or Bayesian value fusion (Berkes et al., 2026).	Better average offline return.
Model-Based ICRL	Revises inferred dynamics after transition changes in DICP (Son et al., 2025) or ICPRL (Xu et al., 2026c).	Accurate prediction before the shift.
Compression	Preserves rare change evidence under budget in communicative compression (Martinez-Lopez et al., 2025) or filtered histories (Chen et al., 2025b).	High compression ratio with small average loss.

7 Training Data and Objectives

7.1 Training Data as an Implicit Non-Stationarity Assumption

Every ICRL training pipeline encodes an assumption about how context should be interpreted. Stationary trajectories teach that old evidence remains relevant. Learning histories teach that behavior can improve across time. Mixed-task datasets teach that context identifies a latent task. Retrieval databases teach that similar past experience can help the present. Non-stationary training data must teach something more specific: the validity of context can change. This point is easy to miss if ICRL is treated only as offline RL at scale: offline RL datasets (Fu et al., 2020; Gulcehre et al., 2020; Agarwal et al., 2020), offline RL surveys (Levine et al., 2020; Prudencio et al., 2023), and offline sequence-modeling substrates (Chen et al., 2021; Janner et al., 2021; Emmons et al., 2022) provide the data substrate, but they do not automatically create within-context invalidation events.

This suggests a simple audit question for training data:

> Does the dataset contain examples in which previously useful context becomes invalid, and does the supervision reward the model for changing its behavior accordingly?

If the answer is no, then robust non-stationary adaptation is unlikely to emerge reliably from scale alone. The model may still generalize, but the benchmark should not interpret this as evidence of context-validity reasoning.

This audit question should be applied at the level of individual context windows, not only at the dataset level. A corpus may contain many tasks and still present each window as internally coherent. Such a corpus teaches task inference but not conflict resolution. To train non-stationary ICRL, windows should sometimes contain pre-shift and post-shift evidence together, with supervision or rewards that make the distinction consequential. The model must experience cases where the correct action is not the action most often seen, not the action seen in the highest-return old trajectory, and not necessarily the action associated with the nearest retrieved neighbor. General ICL studies make the same lesson visible in a simpler setting: example arrangement (Min et al., 2022), task diversity and burstiness (Chan et al., 2022), transformer function learning (Garg et al., 2022; Akyurek et al., 2023), gradient-free or probabilistic ICL mechanisms (Kirsch et al., 2022; Wies et al., 2023), and contextual learning dynamics (Akyurek et al., 2024) can determine whether a transformer learns from context or memorizes a distributional shortcut. For ICRL, dataset construction must additionally expose rewards, interventions, and control consequences.

7.2 Stationary Pretraining, Non-Stationary Testing

A common pattern is to pretrain on stationary tasks and test on related but changing tasks. This creates a useful stress test, but it should be interpreted carefully. Failure may indicate that the model never learned change detection. Success may indicate that the non-stationary benchmark is solvable through recency heuristics or static generalization. To separate these cases, evaluations should include context-order permutations, stale-context injections, repeated-mode tests, and ablations that remove post-shift reward evidence.

7.3 Non-Stationary Pretraining

Direct non-stationary pretraining can vary reward functions, dynamics, observations, action mappings, and task distributions across the lifetime. Work on lifelong ICRL (Xu et al., 2024) and continual ICRL (Wang et al., 2025d) moves in this direction. The main design choice is whether changes are random, curriculum-based, or adversarial. Randomized worlds (Wang et al., 2025a), contextual MDP families (Hallak et al., 2015), hidden-parameter MDPs (Killian et al., 2017; Doshi-Velez & Konidaris, 2016), open-ended or human-timescale task generators (Stooke et al., 2021; Bauer et al., 2023), unsupervised environment design (Dennis et al., 2020; Jiang et al., 2021b;a), PLR-style or evolving generators (Parker-Holder et al., 2022), and non-stationary MDP benchmark suites (Keplinger et al., 2025) can provide scale, but the generator must expose

the shift types the survey cares about. Otherwise, a large randomized dataset can still underrepresent action-semantics shifts, hidden preference changes, or recurring modes. Recent curriculum-design work makes this point concrete: generators can prioritize levels near the competence frontier (Rutherford et al., 2024), improve regret approximations for task discovery (Garcin et al., 2024), regularize environment design (Frans & Isola, 2023), or evolve the mechanics of the game itself (Earle & Togelius, 2024).

7.4 Objective Alignment

The training objective should align with the desired adaptation behavior. Behavior cloning rewards reproducing the data policy. Return-conditioned sequence modeling rewards action sequences associated with high returns in Decision Transformer (Chen et al., 2021), Trajectory Transformer (Janner et al., 2021), and RvS (Emmons et al., 2022). Q-learning-style objectives reward value-sensitive decisions in ICRL (Tarasov et al., 2025; Liu et al., 2026; Dong et al., 2026) and TD-style variants (Wang et al., 2024a). They can inherit useful lessons from conservative (Kumar et al., 2020), optimistic (Agarwal et al., 2020), implicit (Kostrikov et al., 2022), or return-conditioned offline value learning (Brandfonbrener et al., 2022). Model-based objectives reward predictive structure in DICP (Son et al., 2025) and ICPRL (Xu et al., 2026c). Contrastive or auxiliary objectives can reward change-point identification, latent-mode inference, or retrieval validity. Table 11 summarizes how common training choices shape expected behavior under non-stationarity.

Table 11: Training choices and their expected non-stationary behavior.

Training Choice	What It Encourages	Non-Stationary Benefit	Non-Stationary Risk
Stationary Expert Histories	Imitation of good behavior in trajectory models (Chen et al., 2021; Janner et al., 2021; Emmons et al., 2022) and prompting variants (Xu et al., 2022).	Strong in familiar stable tasks.	Over-trusts old demonstrations after shifts.
Learning Histories	Improvement from feedback over time in algorithm distillation (Laskin et al., 2023), cross-episodic curricula (Shi et al., 2023), noise distillation (Zisman et al., 2024), RELIC-style agents (Elawady et al., 2024), and ICPRL (Xu et al., 2026c).	Teaches adaptation trajectories.	May not teach invalidation unless histories shift.
Weak Or Mixed-Quality Data	Robustness to imperfect behavior policies in offline datasets (Fu et al., 2020; Gulcehre et al., 2020; Levine et al., 2020), offline value learning (Kumar et al., 2020; Kostrikov et al., 2022), random-policy distillation (Chen & Paternain, 2024), and value-aware ICRL (Tarasov et al., 2025; Dong et al., 2026; Berkes et al., 2026).	Forces value or feedback use.	Can confuse low quality with non-stationarity.
Procedural Task Generation	Broad task support and scalable diversity from open-ended environments (Stooke et al., 2021; Bauer et al., 2023), unsupervised environment design (Dennis et al., 2020; Jiang et al., 2021b;a), evolving or replay-guided generators (Parker-Holder et al., 2022), randomized MDPs (Wang et al., 2025a), XLand-MiniGrid (Nikulin et al., 2024), and Kinetix (Matthews et al., 2024b).	Can cover many latent modes.	Coverage depends entirely on generator design.
Retrieval Database Training	External memory use in RA-DT (Schmied et al., 2024) and REGENT (Sridhar et al., 2024).	Enables recurring-mode reuse.	Retrieval shortcuts can mask lack of adaptation.
Value-Aware Objectives	Action selection through inferred values in Q-learning ICRL (Tarasov et al., 2025), scalable in-context Q-learning (Liu et al., 2026), and compositional in-context Q-learning (Xu et al., 2026a).	Helps reward and preference shifts.	Still brittle if context selection is invalid.
Change-Point Auxiliaries	Explicit shift detection through Bayesian changepoints (Adams & MacKay, 2007), adaptive restarts (Auer et al., 2019), non-stationary contextual bandits (Chen et al., 2019), continual ICRL (Wang et al., 2025d), or dynamic-regret objectives (Chen et al., 2025a).	Improves recovery and interpretability.	Labels may be unavailable or oversimplified.
State-Space Sequence Training	Long-horizon history modeling with S4-style ICRL (Lu et al., 2023; David et al., 2023), Mamba-style models (Ota, 2024; Beaussant & Mounsif, 2025; Gu & Dao, 2023), or gated recurrent attention (Yang et al., 2023b).	Makes long lifetimes cheaper to model.	Hidden compression can make invalid evidence harder to audit.

7.5 Reporting Requirements

Papers should report the stationarity assumptions of both training and testing. At minimum, a non-stationary ICRL paper should specify which components can change, whether changes appear in training, whether changes are signaled, whether context persists across changes, and whether the agent has a finite context or retrieval budget. Without these details, it is difficult to know whether a method has learned non-stationary adaptation or has simply exploited a benchmark convention.

7.6 Data Construction Patterns

Several practical data-construction patterns follow from the survey's taxonomy. Contrastive windows pair two contexts that share observations but differ in rewards, dynamics, or action semantics. Shift-labeled windows include explicit change-point metadata, useful for supervised auxiliary losses and for measuring whether a model can use clean signals. Shift-unlabeled windows remove metadata and require inference from feedback. Recurring-mode windows force the model to store old evidence without letting it dominate every later decision. Contaminated windows insert invalid demonstrations or retrieved trajectories, making robustness to stale evidence part of the training objective.

These patterns can be mixed with existing ICRL training objectives. For behavior cloning, the key is to avoid rewarding stale imitation after a shift. For return conditioning, the key is to ensure that high desired returns cannot be achieved by repeating pre-shift actions. For Q-learning objectives, the key is to provide enough post-shift evidence for the value function to disambiguate regimes. For model-based objectives, the key is to evaluate predictions separately before and after changes. The training objective need not explicitly name "context validity", but it must create a learning signal where validity affects action quality.

8 Evaluation of Adaptation Beyond Memorization

8.1 Limits of Average Return

Average return is necessary, but it is a poor summary of adaptation. It collapses the lifetime into one number and can hide the exact interval where non-stationarity matters. One agent can score well by performing strongly before a shift and failing afterward. Another can recover after each shift but have lower steady-state return. For non-stationary ICRL, the curve itself is the evidence: behavior before the shift, damage immediately afterward, recovery, and performance when an old mode returns. This aligns with non-stationary bandits' move from static to dynamic comparators (Garivier & Moulines, 2011; Besbes et al., 2014; Li & Li, 2019) and with analogous non-stationary RL analyses (Lecarpentier & Rachelson, 2019; Cheung et al., 2020; Fei et al., 2020). Related MDP regret work (Wei & Luo, 2021) and deep RL generalization studies (Packer et al., 2018; Kirk et al., 2023) make the same warning concrete: held-out performance alone can hide shortcut learning. Figure 11 illustrates the lifecycle quantities that should be read from this curve.

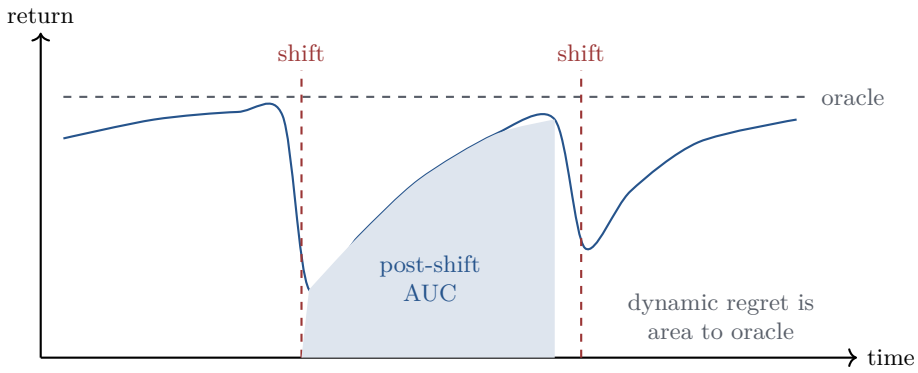

Figure 11: Lifecycle evaluation quantities for non-stationary ICRL. The shape of the curve is more informative than its average.

8.2 Core Metrics

Several metrics have to be read together because each exposes a different failure mode. For a shift at τ, post-shift area under the curve can be reported as

$$\mathrm{AUC}_{\tau:K} = \sum_{k=0}^{K-1} J_{\tau+k}^{\pi_\theta}, \tag{16}$$

while stale-context sensitivity can be measured by the drop between a valid context and a deliberately contaminated one:

$$S_{\mathrm{stale}} = J(\pi_\theta, g_{\mathrm{valid}}) - J(\pi_\theta, g_{\mathrm{valid}\cup\mathrm{stale}})\,. \tag{17}$$

To separate memory scale from adaptation quality, context efficiency can be normalized as

$$\eta_{\mathrm{ctx}} = \frac{J(\pi_\theta, g) - J(\pi_\theta, g_\emptyset)}{\mathrm{tokens}(C) + \rho\,\mathrm{retrievals}(C)}, \tag{18}$$

where ρ converts retrieval calls into a token-equivalent cost. Table 12 lists these metrics and explains what each one measures.

8.3 Benchmark Dimensions

A useful benchmark suite has to vary four dimensions at once: source of change, temporal pattern, observability, and context budget. The first includes reward, dynamics, observation, action semantics, task distribution, and behavior policy. The second includes abrupt, gradual, recurring, expanding, and adversarial changes. The third distinguishes explicit, partial, latent, delayed, and noisy evidence. The fourth covers

Table 12: Recommended metrics for non-stationary ICRL.

Metric	Measure	Use
Post-Shift AUC	Return area after a shift.	Summarizes damage and recovery (Wang et al., 2025d).
Recovery Time	Steps to near-oracle return.	Reports adaptation latency (Adams & MacKay, 2007; Wang et al., 2025d).
Dynamic Regret	Gap to a changing comparator.	Aligns ICRL with non-stationary regret theory (Besbes et al., 2014; Cheung et al., 2020; Chen et al., 2025a).
Stale-Context Sensitivity	Drop under invalid old context.	Detects over-reliance on stale evidence (Schmied et al., 2024; Berkes et al., 2026).
Context Efficiency	Return per context or memory cost.	Separates adaptation from long-context scale (Lu et al., 2023; Grigsby et al., 2024b; Sridhar et al., 2024).
Retrieval Utility	Gain from valid versus bad retrieval.	Audits memory selection (Goyal et al., 2022; Schmied et al., 2024; Sridhar et al., 2024).
Re-Adaptation Latency	Recovery when an old mode returns.	Tests recurrent-mode retention (Khetarpal et al., 2022; Abel et al., 2023).
Change-Detection Quality	Precision, recall, delay, calibration.	Evaluates inferred-shift reliability (Wang et al., 2025d; Chen et al., 2025a).

short context, long context, external retrieval, compressed memory, and no-memory baselines. Without this structure, a benchmark can easily support claims that are technically correct but much narrower than the label "non-stationary ICRL" suggests. Table 14 translates the factorial design into concrete diagnostic benchmark blocks.

Procedural environment families are especially useful for this purpose. XLand-MiniGrid (Nikulin et al., 2024) provides scalable JAX meta-RL environments with many task instances, and XLand-100B (Nikulin et al., 2025) turns this family into a large-scale ICRL learning-history dataset. Kinetix (Matthews et al., 2024b) provides an open-ended 2D physics-based control space with procedurally generated tasks. NS-Gym (Keplinger et al., 2025) is a more direct non-stationary MDP substrate, while PHYRE (Bakhtin et al., 2019), I-PHYRE (Li et al., 2024a), and DeepPHY (Xu et al., 2026b) provide interactive physical reasoning tasks where repeated attempts and visual dynamics can be repackaged into context-adaptation protocols (Xu et al., 2026c). These environments are not automatically non-stationary merely because they are procedural; they become non-stationary ICRL benchmarks when the evaluator constructs within-lifetime schedules that switch, drift, recur, or corrupt rewards, dynamics, observations, action interfaces, or task distributions. Offline datasets such as D4RL (Fu et al., 2020), RL Unplugged (Gulcehre et al., 2020), and Atari replay (Agarwal et al., 2020) can also be useful when they are repackaged into contextual windows with controlled policy quality, support mismatch, and post-shift evidence, but they should not be confused with non-stationary benchmarks unless the test protocol actually changes the task within the context lifetime.

8.4 Candidate Environment Families

Several existing environment families are useful substrates for non-stationary ICRL, even if they were not originally introduced under that label. Their value is not that they automatically solve the benchmark-design problem, but that they are programmable, procedurally diverse, memory-demanding, or already tied to continual-learning metrics. Table 13 groups these candidates by the kind of non-stationary ICRL evidence they can support.

The critical distinction is between an *environment substrate* and a *non-stationary ICRL protocol*. Procgen (Cobbe et al., 2020), MiniGrid (Chevalier-Boisvert et al., 2023), Craftax (Matthews et al., 2024a), MiniHack (Samvelyan et al., 2021), and Jumanji (Bonnet et al., 2024) are not evidence of non-stationary ICRL merely because they generate many tasks. They become relevant when the evaluator specifies a within-lifetime shift process, keeps old context visible after the shift, prevents online parameter updates, and measures recovery through context. Continual RL benchmarks such as CRLMaze (Lomonaco et al., 2020) and CORA (Powers et al., 2021) provide useful shift schedules and metrics, but they need the same care because many continual RL baselines assume online learning. An ICRL protocol should freeze the main policy parameters and compare raw-history, recent-only, retrieval, compression, oracle-change-point, and stale-context adversary variants.

Table 13: Candidate environment families for non-stationary ICRL evaluation.

Environment Family	Why It Is Useful	Non-Stationary Protocol to Add	Best Role in This Survey
AnyMDP / OmniRL	Randomized MDPs are generated by varying transitions and rewards, directly matching the formal ICRL setting of adapting to unseen decision processes (Wang et al., 2025a).	Switch or recur latent MDPs within a lifetime; vary reward and transition budgets; keep pre-shift context in the window; compare oracle restart with learned context filtering.	Controlled mechanism layer for reward, transition, and latent-mode shift.
Contextual / Hidden-Parameter MDP Generators	Contextual MDPs (Hallak et al., 2015) and hidden-parameter MDPs (Killian et al., 2017; Doshi-Velez & Konidaris, 2016; Perez et al., 2020) use latent variables to define rewards and dynamics while preserving shared structure.	Let the latent parameter drift, switch, or recur within a lifetime; hide the parameter from the agent and require inference from context.	Formal mechanism layer for latent-regime inference.
MiniGrid / BabyAI	MiniGrid (Chevalier-Boisvert et al., 2023) and BabyAI (Chevalier-Boisvert et al., 2019) are lightweight grid and instruction environments that are modular and easy to customize.	Change goals, object rules, door-key dependencies, instructions, observation masks, or action permutations after context has accumulated.	Minimal diagnostic layer for action semantics, instruction shift, and stale demonstrations.
Alchemy	Latent causal rules are procedurally resampled and require online inference, hypothesis testing, and action sequencing (Wang et al., 2021b).	Change the hidden chemistry during a lifetime, make old chemistry recur, or insert examples from incompatible chemistries into retrieval memory.	Latent-rule inference layer for context ambiguity and recurring modes.
XLand-MiniGrid / XLand-100B	XLand-MiniGrid (Nikulin et al., 2024) and XLand-100B (Nikulin et al., 2025) make broad ICRL pretraining and evaluation possible through scalable JAX meta-RL environments and large learning-history data.	Schedule task, layout, rule, reward, or action-interface changes across a single lifetime; hold context budget fixed across append-all and retrieval methods.	Large-scale meta-RL and learning-history layer.
Kinetix	Open-ended 2D physics-based control tasks provide procedural diversity with controllable physical structure (Matthews et al., 2024b).	Drift gravity, mass, joints, morphology, object layout, goals, or reward weights; test whether old physics evidence is downweighted.	Open-ended physics layer for dynamics and morphology-like shift without making robotics central.
PHYRE / I-PHYRE / DeepPHY	PHYRE (Bakhtin et al., 2019), I-PHYRE (Li et al., 2024a), and DeepPHY (Xu et al., 2026b) expose action-outcome evidence, visual dynamics, repeated attempts, and VLM-compatible observations relevant to world-model-guided ICRL (Xu et al., 2026c).	Change friction, gravity, object properties, action discretization, visual annotations, or reward rules across attempts; keep failed pre-shift attempts in context and measure whether the agent updates physical intuition.	Physically grounded dynamics and observation layer for world-model-guided ICRL.
Procgen	Procedurally generated games test visual and level generalization across many instances (Cobbe et al., 2020).	Change level generators, visual themes, object layouts, difficulty, or reward rules during a lifetime; include recurring generator families.	Visual procedural layer for observation and task-distribution shift.

Continued on next page

Table 13: Candidate environment families for non-stationary ICRL evaluation (continued).

Environment Family	Why It Is Useful	Non-Stationary Protocol to Add	Best Role in This Survey
Craftax	Fast JAX open-ended crafting tasks require exploration, long-term planning, memory, and adaptation to newly discovered situations (Matthews et al., 2024a).	Drift resource distributions, recipes, tool effects, biome rules, survival objectives, or inventory constraints; test old-recipe contamination.	Open-ended memory and planning layer, complementary to Kinetix.
MiniHack / NetHack Learning Environment	MiniHack (Samvelyan et al., 2021) is a sandbox for designing rich NetHack-based tasks, while NLE (Küttler et al., 2020) provides a stochastic, procedurally generated roguelike environment.	Change object semantics, dungeon families, monster/item distributions, reward shaping, or action effects; test whether retrieved dungeon memories remain valid.	Rich single-agent procedural layer for long-horizon memory and sparse rewards.
Jumanji	JAX environments for combinatorial and general decision problems are scalable and customizable in initial-state distribution and problem complexity (Bonnet et al., 2024).	Drift problem size, constraints, cost functions, instance distributions, or resource budgets; measure context efficiency under fixed retrieval cost.	Non-visual combinatorial layer for budget and constraint shift.
POPGym / POPGym Arcade	POPGym (Morad et al., 2023) explicitly stresses memory in POMDPs; POPGym Arcade (Wang et al., 2025e) provides pixel-based fully and partially observable variants and studies memory contamination by old observations.	Switch observability regimes, delay change evidence, inject stale observations, or alternate fully observable and partially observable modes.	Memory-validity layer for partial observability and context poisoning.
NS-Gym	NS-Gym (Keplinger et al., 2025) is a benchmark suite explicitly designed for non-stationary MDPs; it connects naturally to non-stationary reward and dynamics algorithms (Lecarpentier & Rachelson, 2019; Chandak et al., 2020c), latent-process algorithms (Luo et al., 2024), and policy-search settings (Pettet et al., 2024).	Freeze the agent after pretraining; expose regime changes only through interaction history; compare append-all, recent-only, oracle-change, retrieval, and learned context filters.	Direct bridge from non-stationary MDP evaluation to frozen-parameter ICRL.
CRLMaze	A 3D continual RL benchmark explicitly built around non-stationary environmental changes (Lomonaco et al., 2020).	Freeze the policy after pretraining and expose changes only through interaction history; compare ICRL context policies against online continual RL baselines.	Bridge case for explicit non-stationary RL environments.
CORA	A continual RL platform with benchmarks, baselines, and metrics for forgetting, plasticity, generalization, and forward transfer (Powers et al., 2021).	Recast Atari, Procgen, or NetHack task streams as frozen-parameter context-adaptation tests; report CORA-style forgetting and transfer with ICRL context ablations.	Protocol and metric bridge from continual RL to ICRL.
ScienceWorld / WebShop / ALFWorld	ScienceWorld (Wang et al., 2022), WebShop (Yao et al., 2022), and ALFWorld (Shridhar et al., 2020) expose feedback, tool/action histories, and task instructions in interactive language environments.	Drift user preferences, reward functions, hidden instructions, website/product distributions, or feedback reliability across trials.	Extended LLM-agent layer, included only when feedback drives repeated sequential decisions.

This table also clarifies what should be excluded from the core claim. Robotics-specific suites and controllers such as Meta-World (Yu et al., 2020a), Continual World (Wolczyk et al., 2022), LIBERO (Liu et al., 2023), RoboMimic (Mandlekar et al., 2021), Open X-Embodiment (Open X-Embodiment Collaboration et al., 2024), RoboMME (Dai et al., 2026), and humanoid whole-body-control systems (Ding et al., 2025) are informative for continual learning and generalist action modeling, but they are outside the main scope of this survey. They can be mentioned as adjacent benchmarks when discussing evaluation traditions, but the paper should not rely on them for its central non-stationary ICRL argument. Similarly, multi-agent suites should be avoided because they introduce strategic non-stationarity, for example when concurrently updating agents make the environment non-stationary from each individual agent's perspective (Luo et al., 2023), which is explicitly outside scope here.

Table 14: A benchmark blueprint for separating adaptation mechanisms.

Test Block	Controlled Change	Intended Capability	Shortcut It Rules Out
Reward Switch	Same observations and dynamics, new reward or preference in bandit settings (Garivier & Moulines, 2011; Besbes et al., 2014), safe-policy settings (Chandak et al., 2020a; Moeini et al., 2025a), or reward-feedback agents (Song et al., 2025).	Reinterpret feedback and stop stale imitation.	Static task recognition from observations.
Dynamics Drift	Same reward, changed transition kernel or latent physical parameters in non-stationary RL (Lecarpentier & Rachelson, 2019; Chandak et al., 2020c; Feng et al., 2023), DICP (Son et al., 2025), or ICPRL (Xu et al., 2026c).	Infer action-outcome evidence in context.	Copying old high-return action sequences.
Action Remapping	Same goal, permuted or rescaled action effects in changing-action-set settings (Chandak et al., 2020b) or variable-action ICRL (Sinii et al., 2024).	Detect interface change and relearn action semantics.	Treating action tokens as globally stable.
Latent Recurrence	Earlier mode returns after intervening modes in Alchemy-style tasks (Wang et al., 2021b), reactive or recurring-mode settings (Steinparz et al., 2022), retrieval-based RL (Goyal et al., 2022), RA-DT (Schmied et al., 2024), or REGENT (Sridhar et al., 2024).	Retrieve old valid memory without over-forgetting.	Recency-only adaptation.
Stale-Memory Attack	Insert high-confidence but invalid observations in POPGym-style memory tasks (Morad et al., 2023; Wang et al., 2025e), invalid demonstrations through history filtering or suboptimal-data tests (Chen et al., 2025b; Dong et al., 2026), or conflicting Bayesian evidence (Berkes et al., 2026).	Downweight misleading context.	Long-context or retrieval scale alone.
Delayed Feedback	Shift is only visible after long-horizon outcomes in S4-style ICRL (Lu et al., 2023), AMAGO-style agents (Grigsby et al., 2024a;b), RELIC-style agents (Elawady et al., 2024), or POPGym (Morad et al., 2023).	Maintain uncertainty and explore diagnostically.	Myopic reward matching.
Procedural Task Stream	Generate scheduled task, layout, physics, rule, or goal changes in XLand-MiniGrid/XLand-100B (Nikulin et al., 2024; 2025), Kinetix (Matthews et al., 2024b), Procgen (Cobbe et al., 2020), Craftax (Matthews et al., 2024a), MiniHack (Samvelyan et al., 2021), or Jumanji (Bonnet et al., 2024).	Test adaptation under controlled diversity and repeated shift seeds.	Overfitting to a small hand-written set of shifts.

8.5 Constructing Stale-Context Stress Tests

A non-stationary ICRL benchmark should be specified as a protocol over lifetimes, not only as a set of environment IDs. The core intervention is simple: create a period in which old context is genuinely useful, change the decision process, and then keep some of the old evidence visible so the agent must decide whether to trust it. This makes the failure modes in Table 8 operational. For example, to test stale imitation, the evaluator can collect high-return trajectories under regime m_0, switch to a regime m_1 with the same observations but different rewards or dynamics, and inject the m_0 trajectories into the post-shift context while also providing a small amount of valid m_1 evidence. If performance drops relative to the same post-shift context without the injected trajectories, the agent is not merely slow to learn; it is being actively pulled by stale evidence. This protocol connects continual ICRL evaluation (Wang et al., 2025d), history filtering and weak-data ICRL (Chen et al., 2025b; Dong et al., 2026), retrieval-augmented ICRL (Schmied et al., 2024; Sridhar et al., 2024), and non-stationary MDP or bandit stress tests (Besbes et al., 2014; Cheung et al., 2020; Fei et al., 2020).

A minimal protocol can be written as follows:

1. **Choose a base environment and shift family.** Specify which component changes: reward, transition, observation, action semantics, constraint, task distribution, or behavior policy (Lecarpentier & Rachelson, 2019; Chandak et al., 2020b; Moeini et al., 2025a).
2. **Generate a lifetime.** Run a pre-shift phase under m_0, a post-shift phase under m_1, and optionally a recurrence phase where m_0 returns (Wang et al., 2025d; Steinparz et al., 2022).
3. **Build paired contexts.** For the same post-shift query state, construct at least four contexts: no-context, valid-current context, valid-current plus stale context, and oracle-truncated context that removes pre-shift evidence.
4. **Freeze adaptation channels.** Keep policy parameters fixed and vary only the context policy: append-all, last-k, retrieval, compression, learned filter, oracle filter, and stale adversary (Moeini et al., 2025b; Schmied et al., 2024; Chen et al., 2025b).
5. **Measure lifecycle outcomes.** Report pre-shift return, immediate post-shift drop, recovery time, post-shift AUC, dynamic regret, stale-context sensitivity, and recurring-mode re-adaptation (Wang et al., 2025d; Chen et al., 2025a).
6. **Attribute the failure.** Use ablations to determine whether the bottleneck is change detection, retrieval, compression, influence weighting, forgetting, or policy capacity.

The most important design rule is to use paired evaluations. The valid and contaminated contexts should be identical except for the intervention being tested. Otherwise, a method can look robust simply because the contaminated run also has more useful post-shift evidence, or it can look fragile because the contaminated context changes budget, ordering, or formatting. When possible, the injected stale item should be strong under its original regime: a high-return demonstration, a confident value estimate, a retrieved near-neighbor, or a concise verbal rule. Weak stale evidence is too easy to ignore and underestimates the risk faced by long-context and retrieval-augmented agents. This paired-design principle is the context analogue of oracle-change and controlled-ablation baselines in changepoint, non-stationary regret, and ICRL reporting (Adams & MacKay, 2007; Chen et al., 2025a; Moeini et al., 2025b).

These interventions also define a practical "non-stationary ICRL-ready" benchmark checklist. A benchmark is ready for this survey's claims only if it can freeze the policy, expose the agent to within-lifetime shifts, keep old context available after the shift, inject or remove specific context items, control context budget, and report paired lifecycle metrics. It should also support oracle variants: oracle-current-mode context, oracle-change-point truncation, oracle retrieval, and oracle summary. Without these controls, a benchmark may still be useful for generalization or continual learning, but it cannot isolate whether a frozen policy adapts through context under stale evidence.

8.6 Baselines That Matter

The most informative baselines are often simple. A recent-only context baseline tests whether long-term memory is helping or hurting. A full-history baseline tests whether selective retrieval matters. A no-context baseline tests whether performance is static generalization. An oracle-change-point baseline estimates the value of knowing when to forget. A stale-context adversary estimates how badly the model can be misled by old evidence. For retrieval methods, random retrieval and similarity-only retrieval should be compared with any learned validity-aware retrieval method. The minimum reporting items in Table 16 are intended to make these baselines and ablations comparable across papers.

8.7 Distinguishing Adaptation from Memorization

Non-stationary evaluation should rule out three shortcuts. The first is memorization of task IDs or environment templates. The second is recency-only adaptation, where the model ignores long context and simply follows the last few rewards. The third is retrieval shortcutting, where a memory database contains near-duplicates of test tasks. Countermeasures include held-out generators, changed task order, unseen shift

Table 15: Concrete context-window interventions for stale-context stress tests.

Target Failure	Context Intervention	Paired Control	Diagnostic Signal
Stale Imitation	Inject high-return trajectories or demonstrations from the pre-shift reward/dynamics regime into the post-shift prompt. Keep observations similar so the old behavior looks relevant (Laskin et al., 2023; Chen et al., 2021; Dong et al., 2026).	Same post-shift context with the stale trajectories removed, replaced by neutral padding, or replaced by valid current-regime trajectories of matched length.	Large drop in post-shift AUC or higher dynamic regret indicates over-imitation of obsolete behavior (Wang et al., 2025d; Chen et al., 2025a).
Context Inertia	Vary the amount of pre-shift history while holding the amount of post-shift evidence fixed. Test short, medium, and long stale prefixes (Grigsby et al., 2024b; Elawady et al., 2024; Li et al., 2024b).	Oracle-truncated context that starts at the shift point, plus last-k context with the same token budget.	Recovery time increasing with stale-prefix length indicates that old context dominates new evidence.
Retrieval Contamination	Seed the memory database with near-neighbor trajectories from the wrong reward, dynamics, or action-interface regime, then query after the shift (Goyal et al., 2022; Schmied et al., 2024; Sridhar et al., 2024).	Validity-filtered retrieval, oracle-current-mode retrieval, and random retrieval with matched retrieval count.	Retrieval improves stationary return but hurts post-shift return, or selected items have low current-regime validity.
Compression Loss	Compress a mixed pre-shift/post-shift history so that the rare change evidence is easy to omit, such as one reward reversal or one transition anomaly (Zhong et al., 2024; Packer et al., 2023; Chen et al., 2025b).	Raw context containing the decisive post-shift event, and an oracle summary that preserves the event.	Compressed memory performs well on average but misses the shift or recovers late.
Action-Token Mismatch	Keep old action tokens in context after permuting, masking, rescaling, or renaming actions. The old trajectory should be high return before remapping (Chandak et al., 2020b; Sinii et al., 2024).	Same remapped environment with action descriptions, interface metadata, or oracle remapping supplied.	Correct high-level behavior but wrong low-level actions indicates stale action semantics.
Over-Forgetting	Use an $m_0 \to m_1 \to m_0$ schedule and remove or downweight the original m_0 evidence before recurrence (Khetarpal et al., 2022; Steinparz et al., 2022; Wang et al., 2025d).	Memory-isolated or oracle-regime context that preserves old m_0 evidence for reuse.	Good first recovery but poor recurrence performance indicates excessive forgetting.
Verbal Stale Memory	Insert an old natural-language rule, reflection, or reward explanation that was true before a preference or tool-schema shift (Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2023a; Song et al., 2025).	Same transcript with timestamped, contradicted, or retired memory; oracle-current-rule context.	The agent follows obsolete verbal advice despite contrary feedback.

Table 16: Minimum reporting checklist for non-stationary ICRL experiments.

Item	What to Report
Shift Specification	Which MDP component changes, when it changes, and whether the change is abrupt, gradual, or recurring (Adams & MacKay, 2007; Lecarpentier & Rachelson, 2019); regret and ICRL variants give complementary examples (Feng et al., 2023; Wang et al., 2025d).
Shift Observability	Whether the change is signaled, partially observable, latent, delayed, or noisy (Kaelbling et al., 1998; Hallak et al., 2015; Ghosh et al., 2021); supervised and Bayesian ICRL variants make this observable through context (Lee et al., 2023; Berkes et al., 2026).
Training Exposure	Whether similar changes appear during pretraining or only at test time (Xu et al., 2024; Wang et al., 2025d).
Context Policy	What is written, retrieved, compressed, forgotten, and reset (Schmied et al., 2024; Chen et al., 2025b).
Budget	Context length, retrieval count, memory size, latency budget, and computational budget (Grigsby et al., 2024b; Sridhar et al., 2024).
Baselines	No-context, recent-only, full-history, oracle-change-point, random-retrieval, and stale-retrieval baselines where applicable (Garivier & Moulines, 2011; Moeini et al., 2025b).
Lifecycle Curves	Return or regret before shift, immediately after shift, during recovery, and on recurring modes (Chen et al., 2025a; Wang et al., 2025d).
Ablations	Remove post-shift evidence, inject stale context, permute context order, and vary pre-shift context length (Chen et al., 2025b; Berkes et al., 2026).
Uncertainty	Variance across shift seeds, task sequences, and change frequencies (Besbes et al., 2014; Russac et al., 2019; Cheung et al., 2020); MDP and evaluation studies provide related variance sources (Fei et al., 2020; Feng et al., 2023; Kim & Tewari, 2020).

schedules, adversarial stale-memory insertion, and explicit tests of recurring modes where recency alone is insufficient.

8.8 A Practical Evaluation Stack

A practical stack should include at least three layers. The first layer is a controlled diagnostic layer. AnyMDP (Wang et al., 2025a), MiniGrid (Chevalier-Boisvert et al., 2023), BabyAI (Chevalier-Boisvert et al., 2019), and Alchemy-style tasks (Wang et al., 2021b) are appropriate here because the evaluator can exactly control which part of the decision process changes. This layer should be used for mechanism claims: reward switches, transition switches, action remapping, hidden rule changes, and context contamination. The experiments should be small enough that oracle baselines are meaningful, including oracle-change-point, oracle-retrieval, and oracle-current-mode policies. Generalization benchmarks should be used with the same caution: a held-out level or task family tests extrapolation, whereas non-stationary ICRL requires within-lifetime invalidation and recovery (Packer et al., 2018; Ghosh et al., 2021; Kirk et al., 2023).

The second layer is a scalable procedural layer. XLand-MiniGrid and XLand-100B (Nikulin et al., 2024; 2025), Kinetix (Matthews et al., 2024b), PHYRE/I-PHYRE/DeepPHY (Bakhtin et al., 2019; Li et al., 2024a; Xu et al., 2026b), Procgen (Cobbe et al., 2020), Craftax (Matthews et al., 2024a), MiniHack/NLE (Samvelyan et al., 2021; Küttler et al., 2020), and Jumanji (Bonnet et al., 2024) are natural candidates because they support many task instances while remaining programmable enough to impose explicit non-stationary schedules. This layer should test whether a method that works on clean diagnostics still works when observations are richer, horizons are longer, rewards are sparser, and the task distribution has many modes. It should also test context efficiency: if a method only succeeds by appending all experience, it may not be viable once the environment family becomes large.

The third layer is a memory and continual-protocol layer. POPGym (Morad et al., 2023), POPGym Arcade (Wang et al., 2025e), NS-Gym (Keplinger et al., 2025), CRLMaze (Lomonaco et al., 2020), and CORA (Powers et al., 2021) are useful here because they foreground memory, partial observability, non-stationary streams, and continual-learning metrics. This layer should measure whether the context policy can handle delayed evidence, old-observation poisoning, forward transfer, forgetting, and re-adaptation when old modes return. Unlike the first two layers, this layer is less about task diversity and more about diagnosing the lifetime of information.

An optional extended layer can include interactive language environments such as ScienceWorld (Wang et al., 2022), WebShop (Yao et al., 2022), and ALFWorld (Shridhar et al., 2020) when the paper explicitly studies

LLM reward-feedback agents. These environments are useful for testing preference drift, natural-language feedback, tool-use histories, and changing user instructions. However, they should remain extended cases unless the experiment contains repeated decisions, external feedback, and a clear mechanism by which future behavior improves through context. Otherwise, the result is better described as conversational memory or prompting rather than ICRL. For LLM-agent bridge cases, a non-stationary protocol should vary user preferences, reward feedback, hidden instructions, tool schemas, website distributions, or feedback reliability across repeated trials. Baselines should include no-memory, last-k context, full transcript, summary memory, retrieval memory, oracle-current-rule context, and stale-memory adversary variants. Metrics should include post-shift success rate, correction latency, invalid tool-call rate, stale-memory sensitivity, and preference compliance after drift.

Across all layers, the stack should be reported as a lifecycle evaluation rather than as a single leaderboard. The paper should state the shift source, temporal pattern, observability, context budget, and reset rule for every environment. It should also state whether the environment is intrinsically non-stationary or merely a programmable substrate used to construct a non-stationary protocol. This distinction is important for review: success on a broad procedural benchmark is not automatically success on non-stationary ICRL unless stale context is present and recovery must occur through context.

8.9 Interpreting Evaluation Outcomes

Evaluation should be diagnostic enough to explain why a method succeeds. High return with high stale-context sensitivity suggests that the method is powerful but brittle. Low stale-context sensitivity with poor recurring-mode performance suggests over-forgetting. Good recovery with poor dynamic regret may indicate that the method eventually adapts but pays a large cost after every change. Strong oracle-change-point performance and weak learned-change-point performance indicates that the main bottleneck is detection or context management rather than policy capacity.

The conclusion is that non-stationary ICRL evaluation should not end with a single return number. A convincing result must identify which failure mode was solved: recognizing that the regime changed, selecting evidence from the current regime, retaining useful evidence from older recurring regimes, or acting well after the relevant context has been selected. The strongest evidence is therefore a consistent pattern across return, dynamic regret, recovery time, stale-context sensitivity, and oracle gaps. If a method improves return but remains sensitive to stale context, the claim should be limited to better exploitation under favorable histories. If it closes the learned-to-oracle change-point gap, the claim can be about change detection or context filtering. If it improves recurring-mode performance without hurting post-shift recovery, the claim can be about useful long-term memory. This interpretation discipline keeps benchmark results tied to the central question of the survey: whether a frozen policy can use context to adapt under changing decision processes, not merely whether it performs well on a diverse task suite.

9 Theory and Conceptual Links

9.1 Existing Theoretical Coverage

Existing theory has begun to make ICRL more than an analogy between prompting and reinforcement learning. Supervised pretraining analyses show that decision transformers can implement in-context policies under structured task distributions (Lee et al., 2023; Lin et al., 2024). General ICL theory supplies the broader mechanisms: transformers can learn simple function classes in context (Garg et al., 2022; Akyurek et al., 2023), encode implicit models or gradient-descent-like computations inside the forward pass (von Oswald et al., 2023; Akyurek et al., 2024), and depend strongly on demonstration structure (Min et al., 2022; Wies et al., 2023). Recent work studies chain-of-thought or explicit iterative computation as a path toward in-context reinforcement learning (Xie et al., 2026). Other work shows that transformers can implement policy-improvement-like procedures in context (Liang & Lai, 2026). These results are important because they show that ICRL is not merely an empirical analogy: under certain assumptions, the architecture can represent update-like computations. Adjacent non-stationary ICL theory (Qin et al., 2026), while not reinforcement learning, also supports the relevance of recency, adaptive filtering (Sayed, 2011), gating (Yang et al., 2023b; Katsch, 2023), retention-style sequence models (Sun et al., 2023), and recurrent attention mechanisms (Peng et al., 2024) when the target function changes over time. Long-context ICL studies provide a complementary caution: increasing the number of examples can help only when the model can keep the relevant evidence accessible and distinguish it from distractors (Li et al., 2024b; Bertsch et al., 2025).

9.2 Added Challenges Under Non-Stationarity

Non-stationarity adds two theoretical problems that are easy to blur together. The comparator problem asks what the agent should be compared against when the optimal policy changes over time. Dynamic regret is the natural answer in many settings, including adversarial or bandit settings (Auer et al., 2002; Garivier & Moulines, 2011; Besbes et al., 2014), adaptive-window or weighted bandits (Auer et al., 2019; Russac et al., 2019; Li & Li, 2019), non-stationary MDPs (Cheung et al., 2020; Fei et al., 2020; Feng et al., 2023), refined MDP bounds (Wei & Luo, 2021; Mao et al., 2021; Domingues et al., 2021), and transformer-based analyses (Chen et al., 2025a). The validity problem asks which context items should influence the current policy. This part is less developed theoretically. Most in-context theory assumes that examples are sampled from a coherent latent task or distribution. Non-stationary ICRL requires theory for contexts that are finite, ordered, partially invalid, and sometimes contradictory. POMDP (Kaelbling et al., 1998), CMDP (Hallak et al., 2015), epistemic-POMDP (Ghosh et al., 2021), task-inference (Humplik et al., 2019), and HiP-MDP theory (Killian et al., 2017; Doshi-Velez & Konidaris, 2016) suggest one route: treat the current regime as latent and context as evidence about that regime. The missing piece is that an ICRL agent observes this evidence through a bounded and possibly learned context channel, so inference error and context-selection error become part of the control problem. One useful decomposition is to separate environmental movement from context-management error:

$$\mathrm{Reg}_T^{\mathrm{dyn}}(\pi_\theta, g) \le \underbrace{\mathcal{A}(T, B_T)}_{\text{move}} + \underbrace{\sum_{t=1}^{T} \epsilon_{\mathrm{ctx}}(t; g)}_{\text{ctx}} + \underbrace{\sum_{t=1}^{T} \epsilon_{\mathrm{model}}(t; \theta)}_{\text{model}} . \tag{19}$$

Here B_T is a variation budget: the movement term captures adaptation to environmental drift, the context term captures retrieval, compression, or stale-context mistakes, and the model term captures the approximation limits of the frozen decision model. This is not a theorem as stated; it is a target form for future theory that would make context-management assumptions explicit. The main gap is therefore an interface between ICRL theory, non-stationary RL regret analysis, and retrieval or memory theory: the target object is not only a changing MDP, but a changing MDP observed through a bounded context channel.

9.3 Open Theoretical Questions

Several questions would make the theory more precise:

- Which variation budgets or change-point assumptions make fixed-parameter context adaptation competitive with online updating?
- How should regret bounds scale with context length, memory size, retrieval error, and shift observability?
- When is stale context harmless because attention can suppress it, and when does it induce unavoidable lower bounds?
- How do value-aware and Bayesian context mechanisms alter sample complexity under reward or dynamics shift?
- Can recurring-mode memory be given a formal advantage over recency-only adaptation?

Table 17 maps existing theory topics to the non-stationary pieces that remain missing.

A useful first step is to make the underlying assumptions explicit:

- Under bounded variation in rewards or dynamics, theory could characterize how regret scales with context length.
- Under a finite set of recurring latent regimes, theory could identify when retrieval reduces re-adaptation latency.
- Under adversarial stale context, theory could characterize robustness guarantees for attention or retrieval policies.

These stylized settings would not capture the full complexity of deployment, but they would turn vague claims about "using context" into testable statements about when context helps and when it provably hurts.

Table 17: Theory topics and their relevance to non-stationary ICRL.

Theory Topic	Current Contribution	Missing Non-Stationary Piece
General ICL Theory	Explains function-class learning (Garg et al., 2022; Akyurek et al., 2023), implicit model fitting or forward-pass optimization (von Oswald et al., 2023; Kirsch et al., 2022), and learnability conditions (Wies et al., 2023).	Need sequential feedback, reward, and invalid-context extensions.
Supervised ICRL Theory	Explains when sequence models can map context to near-optimal decisions in supervised ICRL (Lee et al., 2023; Lin et al., 2024), TD-style ICRL (Wang et al., 2024a), or provable-emergence analyses of in-context temporal-difference learning (Wang et al., 2025b).	Context examples are often assumed coherent rather than stale or conflicting.
Policy Improvement in Context	Shows update-like computations can be represented by chain-of-thought ICRL (Xie et al., 2026), policy-improvement transformers (Liang & Lai, 2026), generalized decision transformers (Furuta et al., 2022), or agentic decision models (Liu & Abbeel, 2023).	Need finite-context and post-shift guarantees.
Dynamic Regret	Provides comparator for changing adversarial or bandit environments (Auer et al., 2002; Garivier & Moulines, 2011; Besbes et al., 2014), weighted bandits (Russac et al., 2019), non-stationary MDPs (Cheung et al., 2020; Fei et al., 2020; Feng et al., 2023), refined MDP bounds (Wei & Luo, 2021), and transformer analyses (Chen et al., 2025a).	Need bounds that include learned context construction and retrieval error.
Bayesian And Latent-Context Use	Frames context as evidence combined with POMDP or CMDP beliefs (Kaelbling et al., 1998; Hallak et al., 2015), HiP-MDP variables (Killian et al., 2017; Doshi-Velez & Konidaris, 2016), task-inference variables (Humplik et al., 2019), latent-belief meta-RL (Rakelly et al., 2019; Zintgraf et al., 2020), or Bayesian value fusion (Berkes et al., 2026).	Need misspecification and stale-evidence analysis.
Retrieval And Memory Theory	Studies language-agent memory (Zhong et al., 2024; Wang et al., 2023b; Packer et al., 2023), retrieval-based RL (Goyal et al., 2022; Humphreys et al., 2022), and retrieval-augmented ICRL (Schmied et al., 2024; Sridhar et al., 2024).	Need validity-aware retrieval under changing rewards, dynamics, and actions.

10 Design Principles and Open Problems

10.1 Design Principles

The preceding sections point to several design principles for non-stationary ICRL systems.

1. **Treat context as a managed resource.** Longer context is useful only when the system can identify what remains valid, a lesson shared by long-memory LLM systems (Zhong et al., 2024; Wang et al., 2023b; Packer et al., 2023), long-context ICL studies (Li et al., 2024b; Bertsch et al., 2025), efficient sequence models (Lu et al., 2023), and long-context ICRL agents (Elawady et al., 2024).
2. **Train on invalidation, not only adaptation.** Training histories should include moments where old evidence becomes wrong, not only stationary improvement traces from algorithm distillation (Laskin et al., 2023), noise or history filtering (Zisman et al., 2024; Chen et al., 2025b), continual or suboptimal ICRL (Wang et al., 2025d; Dong et al., 2026), or environment design (Dennis et al., 2020; Jiang et al., 2021b).
3. **Separate recency from relevance.** Recency is a useful heuristic in sliding-window and discounted methods (Garivier & Moulines, 2011; Russac et al., 2019; Trovo et al., 2020), but it fails under changepoints and recurring modes (Adams & MacKay, 2007).
4. **Expose retrieval to stale-memory tests.** Retrieval should be evaluated by validity, not only similarity or return, in retrieval-based RL (Goyal et al., 2022; Humphreys et al., 2022) and retrieval-augmented ICRL (Schmied et al., 2024; Sridhar et al., 2024).
5. **Report lifecycle curves.** A single average return hides post-shift failure and recovery in non-stationary MDPs (Cheung et al., 2020; Fei et al., 2020; Wei & Luo, 2021) and continual ICRL (Wang et al., 2025d).
6. **Include oracle context baselines.** Oracle-change-point baselines (Adams & MacKay, 2007) and general ICRL reporting baselines (Moeini et al., 2025b) estimate the headroom available from better context management.
7. **State the adaptation channel.** Papers should make clear whether adaptation occurs through prompt tokens, hidden state, external memory, retrieved trajectories, summaries, or online parameter updates.

10.2 Open Problems

Not every item in the research agenda is open in the same sense. Some topics already have direct starting points, such as continual ICRL benchmarks and metrics (Wang et al., 2025d), dynamic-regret analyses for non-stationary transformers (Chen et al., 2025a), history filtering (Chen et al., 2025b), retrieval-augmented ICRL (Schmied et al., 2024; Sridhar et al., 2024), and safe ICRL mechanisms (Moeini et al., 2025a; Kwon et al., 2026). Other topics are seeded only by adjacent literatures, such as changepoint detection, continual learning, retrieval, or safe RL, and still need to be reformulated under the fixed-parameter context-adaptation constraint. A third group is more genuinely open: the field lacks clear benchmarks, formal objectives, or accepted measurement protocols. Table 18 therefore distinguishes *seeded* directions from *partly seeded* and *truly open* ones.

Table 18: A status-aware roadmap for non-stationary ICRL.

Problem	Current Status	Actionable Next Step
Standardized Lifecycle Benchmarks	**Seeded.** Continual ICRL gives direct benchmark and metric examples, including post-shift recovery and lifecycle reporting (Wang et al., 2025d); broader benchmark reporting norms remain fragmented.	Consolidate a shared suite with declared shift source, temporal pattern, observability, context reset policy, stale-context ablations, and oracle-context headroom (Moeini et al., 2025b; Kirk et al., 2023; Wang et al., 2025d).

Table 18: A status-aware roadmap for non-stationary ICRL (continued).

Problem	Current Status	Actionable Next Step
Change-Point-Aware Context Policies	**Seeded but unresolved.** Changepoint methods and continual ICRL show useful templates (Adams & MacKay, 2007; Wang et al., 2025d), while transformer dynamic-regret work studies related adaptation objectives (Chen et al., 2025a).	Learn write/retrieve/forget rules from weakly labeled or unlabeled shifts, then evaluate detection delay, context repair, and post-shift return together.
Validity-Aware Retrieval	**Partly seeded.** Retrieval-augmented ICRL and retrieval-based RL exist (Goyal et al., 2022; Schmied et al., 2024; Sridhar et al., 2024), but most retrieval keys still do not estimate decision validity after shifts.	Retrieve by latent mode, uncertainty, reward fit, recency, and time metadata; report precision of valid retrieved items, not only downstream return.
Adaptive Forgetting And Memory Isolation	**Partly seeded.** Continual learning and lifelong memory offer replay, regularization, and isolation mechanisms (Khetarpal et al., 2022; Kirkpatrick et al., 2017; Zhong et al., 2024), but context-only agents need different guarantees.	Partition memory by inferred regime and test both irreversible shifts and recurring modes, where aggressive forgetting and indiscriminate retention fail in opposite ways.
Finite-Context Theory Under Stale Evidence	**Truly open.** ICL theory, non-stationary regret, and chain-of-thought ICRL theory provide components (Garg et al., 2022; Cheung et al., 2020; Xie et al., 2026), but not a theory of bounded context containing invalid examples.	Bound regret or sample complexity as a function of context length, retrieval error, stale-context mass, shift observability, and model approximation error.
Efficient Long-Memory Architectures	**Partly seeded.** S4, Mamba, and long-context agents address cost and memory length (Lu et al., 2023; Gu & Dao, 2023; Elawady et al., 2024), but rarely under matched invalidation tests.	Compare compressed-memory models under identical abrupt, gradual, and recurring shifts, with oracle-reset and stale-state ablations.
Action-Semantics Shift	**Partly seeded.** Variable-action ICRL and changing-action RL cover action-set variation (Chandak et al., 2020b; Sinii et al., 2024), but within-lifetime interface drift remains under-tested.	Test permutations, masks, rescaling, action-description changes, and interface remapping while old context still contains obsolete action tokens.
Context Poisoning	**Truly open for non-stationary ICRL.** Safety and Bayesian value-fusion work are relevant (Moeini et al., 2025a; Berkes et al., 2026), but adversarial stale memories are not yet a standard threat model.	Stress-test stale demonstrations, malicious retrievals, corrupted rewards, and misleading summaries; measure whether the agent can isolate harmful context without online parameter updates.
Verbal Memory Validity In LLM Agents	**Partly seeded.** Reflection and experience-library agents store natural-language lessons (Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2023a), but validity under preference, tool, or reward drift is rarely measured.	Timestamp, regime-tag, contradict, retire, or reactivate memories using reward and feedback evidence rather than semantic similarity alone.
Constraint And Safety Shifts	**Seeded.** Safe ICRL and barrier-style shielding provide mechanisms (Altman, 2021; Moeini et al., 2025a; Kwon et al., 2026), but shifting constraints require separate reward-validity and safety-validity tests.	Add safety filters that work without parameter updates and report reward, cost, and violation recovery after constraint changes.
Weak-Data Non-Stationarity	**Partly seeded.** Offline RL and value-aware ICRL address weak or suboptimal data (Fu et al., 2020; Kumar et al., 2020; Berkes et al., 2026), but changing data quality is usually not isolated from changing tasks.	Factor benchmarks by behavior-policy quality, reward shift, support mismatch, and uncertainty, so failures are not attributed only to non-stationarity.
Exploration Under Changing Feedback	**Partly seeded.** In-context exploration and LLM bandit work study diagnostic action choice (Krishnamurthy et al., 2024; Monea et al., 2024; Nie et al., 2025), but feedback drift adds a second inference problem.	Combine in-context exploration with shift-aware uncertainty summaries and report how quickly the agent collects disambiguating evidence after feedback changes.

Safety is a related but distinct pressure on this agenda. Safe ICRL extends the adaptation problem to constrained decision processes (Altman, 2021; Achiam et al., 2017; Satija et al., 2020), where context must help the agent improve reward while respecting cost budgets (Moeini et al., 2025a). Recent shielding work makes the same point more explicit under OOD deployment shifts: Q-Barrier (Kwon et al., 2026) suggests that a frozen safe-ICRL policy may need an action-level latent budget and cost filter, not only a longer history, to preserve a useful reward-safety tradeoff. The non-stationary version is harder because a context item may be reward-valid but safety-invalid after a constraint or cost shift. This reinforces the need to report validity with respect to rewards, dynamics, and constraints separately. Classical safe exploration (Moldovan & Abbeel, 2012), model-based safety guarantees (Berkenkamp et al., 2017), safe policy improvement (Laroche et al., 2019), and modern safe-RL benchmarks (Ji et al., 2023) provide useful components, but they still need to be recast under the fixed-parameter, context-only adaptation constraint.

10.3 A Research Agenda

The next phase of the field should move beyond the broad question of whether context helps. For non-stationary ICRL, the sharper questions are which context helps, when it helps, when it hurts, and how a frozen agent can know the difference. The roadmap should proceed in stages. The first stage is consolidation: seeded directions such as continual ICRL benchmarks (Wang et al., 2025d), safe ICRL (Moeini et al., 2025a; Kwon et al., 2026), and retrieval-augmented ICRL (Schmied et al., 2024; Sridhar et al., 2024) should be evaluated under a common lifecycle protocol rather than treated as isolated demonstrations. The second stage is translation: adjacent tools from changepoint detection, non-stationary regret, continual learning, retrieval, and safe RL should be rewritten in context-channel terms, specifying what is stored, retrieved, suppressed, or exposed to the frozen model. The third stage is genuinely open theory and stress testing: finite-context bounds, adversarial stale-context threat models, and validity-aware memory metrics should become standard targets.

Claims about non-stationary adaptation should specify the shift source, temporal pattern, observability regime, context policy, and training exposure. They should report post-shift curves and stale-context ablations, compare against simple context policies before attributing gains to architecture scale, and state whether a result is solving a new open problem or extending an already seeded line.

The agenda also requires a cleaner separation between the pretrained decision model, the context manager, and the memory store. Current papers often blur these components together. In stationary settings, that may be tolerable because most useful context points in the same direction. In non-stationary settings, the distinction is central. A strong policy with a weak context manager can fail, while a modest policy with a good validity filter can recover quickly.

10.4 Reporting Standards for a Mature Literature

A mature non-stationary ICRL literature would make claims at the right level of specificity. Instead of saying that a model handles non-stationarity, a paper would say that it handles reward/abrupt/partial shifts with a finite context budget, or dynamics/recurring/latent shifts with an external memory. Instead of reporting only average return, it would show lifecycle curves, recovery time, stale-context sensitivity, and oracle-context headroom. Instead of treating retrieval as a black-box improvement, it would report whether retrieved items are valid under the current regime. Instead of presenting long context as inherently beneficial, it would show when additional context helps, when it hurts, and what mechanism controls its influence.

This reporting standard is demanding, but it is aligned with the central promise of ICRL. If adaptation is supposed to happen without weight updates, then the evidence passed to the frozen model is the adaptation interface. The field should therefore scrutinize that interface with the same care that reinforcement-learning papers apply to rewards, dynamics, and exploration.

11 Conclusion

This survey has argued that non-stationary In-context reinforcement learning is not just ICRL evaluated on a larger or more diverse task distribution. It is the problem of adapting through context while deployed policy parameters remain fixed, even when accumulated evidence becomes obsolete, misleading, or useful again after a regime returns. The key difficulty is context validity: a fixed-parameter policy must infer which accumulated evidence still describes the current decision process and which evidence should be downweighted, retired, or reactivated. This view connects black-box meta-RL (Duan et al., 2016; Wang et al., 2016), decision sequence modeling (Chen et al., 2021; Lee et al., 2023), algorithm distillation (Laskin et al., 2023), retrieval-augmented agents (Schmied et al., 2024; Sridhar et al., 2024), and long-context adaptive systems (Elawady et al., 2024), but it shifts the organizing question from "does context help?" to "when should context be trusted?".

From this perspective, the survey made five contributions to the non-stationary ICRL problem. **First,** it defined the setting and separated it from adjacent notions such as stationary few-shot adaptation, domain randomization, online continual learning, standard offline RL, robotics-specific lifelong learning, and multi-agent strategic non-stationarity. The defining constraint is that adaptation occurs through context while the deployed policy parameters remain fixed. **Second,** it introduced a taxonomy that describes a non-stationary ICRL setting along three axes: what changes, how the change unfolds over time, and how observable the change is to the agent. This source–pattern–observability vocabulary is intended to prevent overbroad claims: reward specifications, transition kernels, action interfaces, observation channels, constraint models, and data sources create different problems, and abrupt, gradual, recurring, expanding, or adversarial changes require different evidence-management behavior. **Third,** the survey recast methods as context-management mechanisms rather than only as model families. Raw history conditioning, context filtering, retrieval, compression, recurrent memory, value-aware inference, model-based prediction, reflection, and forgetting are all ways of deciding which past evidence reaches the frozen decision policy and how strongly it should influence action. Under non-stationarity, these mechanisms are not engineering details; they are the adaptation channel. **Fourth,** the survey translated that view into training and evaluation requirements. Training data should contain invalidation, mixed pre-shift and post-shift histories, stale demonstrations, suboptimal or mismatched trajectories, recurring modes, and finite-context pressure (Chen et al., 2025b; Dong et al., 2026). Evaluation should then report lifecycle curves, post-shift recovery, dynamic regret, stale-context sensitivity, retrieval utility, oracle-context gaps, recurring-mode reuse, and finite-context degradation rather than only average return. Benchmarks such as procedural environments, continual RL suites, and memory-focused tasks become evidence for non-stationary ICRL only when they impose within-lifetime shifts and require recovery through context. **Fifth,** the survey identified the theoretical gap created by stale and conflicting context. Existing change-point methods (Adams & MacKay, 2007), non-stationary bandit regret (Besbes et al., 2014), non-stationary MDP regret (Cheung et al., 2020; Fei et al., 2020), and continual ICRL (Wang et al., 2025d) provide useful pieces, but they do not yet explain how regret, sample efficiency, or robustness should scale with context length, retrieval error, memory size, shift observability, and context poisoning. A mature theory of non-stationary ICRL should treat the context manager as part of the control system, because the agent acts through a bounded and imperfect evidence channel.

The practical conclusion is that future work should make its claim at the level of the shift it actually solves. A method that improves return under clean histories is not yet a non-stationary ICRL solution if it fails under stale context. A method that recovers after abrupt reward changes has not necessarily solved gradual dynamics drift or hidden action remapping. A method that forgets quickly may adapt after one-way change but fail when old regimes recur. Progress in non-stationary ICRL will come from systems that can detect change, select currently valid evidence, preserve reusable memories, reject obsolete context, and act well under a finite context budget. The central benchmark for the field is therefore not larger context alone, but disciplined evidence use across the lifetime of a changing decision process.

References

David Abel, Andre Barreto, Benjamin Van Roy, Doina Precup, Hado van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50377–50407, 2023.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, 2017.

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, 2020.

Ekin Akyurek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023.

Ekin Akyurek, Bailin Wang, Yoon Kim, and Jacob Andreas. In-context language learning: Architectures and algorithms. *arXiv preprint arXiv:2401.12973*, 2024.

Maruan Al-Shedivat, Trapit Bansal, Yuri Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. Continuous adaptation via meta-learning in nonstationary and competitive environments. *arXiv preprint arXiv:1710.03641*, 2017.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what not to forget. In *European Conference on Computer Vision*, pp. 139–154, 2018.

Eitan Altman. *Constrained Markov Decision Processes*. Routledge, 2021.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. In *SIAM Journal on Computing*, pp. 48–77, 2002.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pp. 138–158, 2019.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, 2019.

Jakob Bauer, Kate Baumli, Feryal M. P. Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, Karol Gregor, Edward Hughes, Sheleem Kashem, Maria Loks-Thompson, Hannah Openshaw, Jack Parker-Holder, Shreya Pathak, Nicolas Perez Nieves, Nemanja Rakicevic, Tim Rocktäschel, Yannick Schroecker, Satinder Singh, Jakub Sygnowski, Karl Tuyls, Sarah York, Alexander Zacherl, and Lei M. Zhang. Human-timescale adaptation in an open-ended task space. In *International Conference on Machine Learning*, 2023.

Dominik Baumann, Erfaun Noorani, Arsenii Mustafin, Xinyi Sheng, Bert Verbruggen, Arne Vanhoyweghen, Vincent Ginis, and Thomas B. Schön. Ergodicity in reinforcement learning. *arXiv preprint arXiv:2603.10895*, 2026.

Samuel Beaussant and Mehdi Mounsif. Scaling algorithm distillation for continuous control with mamba. *arXiv preprint arXiv:2506.13892*, 2025.

Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028*, 2023.

Maximilian Beck, Korbinian Poppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Gunter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107547–107603, 2024.

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, 2017.

Anaïs Berkes, Vincent Taboga, Donna Vakalis, David Rolnick, and Yoshua Bengio. In-context reinforcement learning through bayesian fusion of context and value prior. *arXiv preprint arXiv:2601.03015*, 2026.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. In-context learning with long-context models: An in-depth exploration. In *Association for Computational Linguistics*, pp. 12119–12149, 2025.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems*, 2014.

Zhenshan Bing, David Lerch, Kai Huang, and Alois Knoll. Meta-reinforcement learning in non-stationary and dynamic environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3476–3491, 2022.

Clément Bonnet, Daniel Luo, Donal Byrne, Shikha Surana, Sasha Abramowitz, Paul Duckworth, Vincent Coyette, Laurence I. Midgley, Elshadai Tegegn, Tristan Kalloniatis, Omayma Mahjoub, Matthew Macfarlane, Andries P. Smit, Nathan Grinsztajn, Raphael Boige, Cemlyn N. Waters, Mohamed A. Mimouni, Ulrich A. Mbou Sob, Ruan de Kock, Siddarth Singh, Daniel Furelos-Blanco, Victor Le, Arnu Pretorius, and Alexandre Laterre. Jumanji: A diverse suite of scalable reinforcement learning environments in JAX. In *International Conference on Learning Representations*, 2024.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George B. M. van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pp. 2206–2240, 2022.

David Brandfonbrener, Alberto Bietti, Jacob Buckman, Romain Laroche, and Joan Bruna. When does return-conditioned supervised learning work for offline reinforcement learning? *Advances in Neural Information Processing Systems*, 35:1542–1553, 2022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024.

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091, 2022.

Massimo Caccia, Jonas Mueller, Taesup Kim, Laurent Charlin, and Rasool Fakoor. Task-agnostic continual reinforcement learning: In praise of a simple baseline. *arXiv preprint arXiv:2205.14495*, 2022.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891, 2022.

Yash Chandak, Scott Jordan, Georgios Theocharous, Martha White, and Philip S. Thomas. Towards safe policy improvement for non-stationary mdps. *Advances in Neural Information Processing Systems*, 33: 9156–9168, 2020a.

Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, and Philip S. Thomas. Lifelong learning with a changing action set. In *AAAI Conference on Artificial Intelligence*, 2020b.

Yash Chandak, Georgios Theocharous, Shiv Shankar, Sridhar Mahadevan, Martha White, and Philip S. Thomas. Optimizing for the future in non-stationary mdps. In *International Conference on Machine Learning*, 2020c.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2019.

Baiyuan Chen, Shinji Ito, and Masaaki Imaizumi. Optimal dynamic regret by transformers for non-stationary reinforcement learning. *arXiv preprint arXiv:2508.16027*, 2025a.

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.

Weiqin Chen and Santiago Paternain. Random policy enables in-context reinforcement learning within trust horizons. *arXiv preprint arXiv:2410.19982*, 2024.

Weiqin Chen, Xinjie Zhang, Dharmashankar Subramanian, and Santiago Paternain. Filtering learning histories enhances in-context reinforcement learning. *arXiv preprint arXiv:2505.15143*, 2025b.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pp. 696–726, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, 2020.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. BabyAI: A platform to study the sample efficiency of grounded language learning. In *International Conference on Learning Representations*, 2019.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & Miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *arXiv preprint arXiv:2306.13831*, 2023.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International Conference on Machine Learning*, 2020.

Jonathan Cook, Chris Lu, Edward Hughes, Joel Z. Leibo, and Jakob Foerster. Artificial generational intelligence: Cultural accumulation in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024.

Yinpei Dai, Hongze Fu, Jayjun Lee, Yuejiang Liu, Haoran Zhang, Jianing Yang, Chelsea Finn, Nima Fazeli, and Joyce Chai. RoboMME: Benchmarking and understanding memory for robotic generalist policies. *arXiv preprint arXiv:2603.04639*, 2026.

Zhenwen Dai, Federico Tomasi, and Sina Ghiassian. In-context exploration-exploitation for reinforcement learning. In *International Conference on Learning Representations*, 2024.

Shmuel Bar David, Itamar Zimerman, Eliya Nachmani, and Lior Wolf. Decision S4: Efficient sequence-based RL via state spaces layers. In *International Conference on Learning Representations*, 2023.

Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre M. Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. In *Advances in Neural Information Processing Systems*, 2020.

Ziluo Ding, Haobin Jiang, Yuxuan Wang, Zhenguo Sun, Yu Zhang, Xiaojie Niu, Ming Yang, Weishuai Zeng, Xinrun Xu, and Zongqing Lu. JAEGER: Dual-level humanoid whole-body controller. *arXiv preprint arXiv:2505.06584*, 2025. doi: 10.48550/arXiv.2505.06584.

Omar Darwiche Domingues, Pierre Menard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 3538–3546, 2021.

Juncheng Dong, Moyang Guo, Ethan X. Fang, Zhuoran Yang, and Vahid Tarokh. In-context reinforcement learning from suboptimal historical data. *arXiv preprint arXiv:2601.20116*, 2026.

Ron Dorfman, Idan Shenfeld, and Aviv Tamar. Offline meta learning of exploration. *arXiv preprint arXiv:2008.02598*, 2020.

Finale Doshi-Velez and George Konidaris. Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. In *International Joint Conference on Artificial Intelligence*, pp. 1432–1440, 2016.

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Sam Earle and Julian Togelius. Autoverse: An evolvable game language for learning robust embodied agents. *arXiv preprint arXiv:2407.04221*, 2024.

Ahmad Elawady, Gunjan Chhablani, Ram Ramrakhya, Karmesh Yadav, Dhruv Batra, Zsolt Kira, and Andrew Szot. ReLIC: A recipe for 64k steps of in-context reinforcement learning for embodied AI. *arXiv preprint arXiv:2410.02751*, 2024.

Scott Emmons, Benjamin Eysenbach, Ilya Kostrikov, and Sergey Levine. RVS: What is essential for offline RL via supervised learning? In *International Conference on Learning Representations*, 2022.

Yingjie Fei, Zhuoran Yang, Zhaoran Wang, and Qiaomin Xie. Dynamic regret of policy optimization in non-stationary environments. In *Advances in Neural Information Processing Systems*, 2020.

Songtao Feng, Ming Yin, and Lin F. Yang. Non-stationary reinforcement learning under general function approximation. *arXiv preprint arXiv:2306.00861*, 2023.

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

Kevin Frans and Phillip Isola. Powderworld: A platform for understanding generalization via rich task distributions. In *International Conference on Learning Representations*, 2023.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks*, 2020.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062, 2019.

Hiroki Furuta, Yutaka Matsuo, and Shixiang Shane Gu. Generalized decision transformer for offline hindsight information matching. In *International Conference on Learning Representations*, 2022.

Samuel Garcin, James Doran, Shangmin Guo, Christopher G. Lucas, and Stefano V. Albrecht. Dred: Zero-shot transfer in reinforcement learning via data-regularised environment design. In *International Conference on Machine Learning*, 2024.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, 2022.

Aurelien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. In *International Conference on Algorithmic Learning Theory*, 2011.

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. *arXiv preprint arXiv:2107.06277*, 2021.

Anirudh Goyal, Abram Friesen, Andrea Banino, Theophane Weber, Nan Rosemary Ke, Adria Puigdomenech Badia, Arthur Guez, Mehdi Mirza, Peter C. Humphreys, Ksenia Konyushova, et al. Retrieval-augmented reinforcement learning. In *International Conference on Machine Learning*, pp. 7740–7765, 2022.

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

Jake Grigsby, Justin Sasek, Samyak Parajuli, Daniel Adebi, Amy Zhang, and Yuke Zhu. AMAGO-2: Breaking the multi-task barrier in meta-reinforcement learning with transformers. In *Advances in Neural Information Processing Systems*, 2024b.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Paine, Sergio Gomez, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, Gabriel Dulac-Arnold, Jerry Li, Mohammad Norouzi, Matthew Hoffman, Ofir Nachum, George Tucker, Nicolas Heess, and Nando de Freitas. RL unplugged: Benchmarks for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pp. 3929–3938, 2020.

Raia Hadsell, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, 2020.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pp. 2555–2565, 2019.

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

Nicolas Heess, Jonathan J. Hunt, Timothy P. Lillicrap, and David Silver. Memory-based control with recurrent neural networks. *arXiv preprint arXiv:1512.04455*, 2015.

Sili Huang, Jifeng Hu, Hechang Chen, Lichao Sun, and Bo Yang. In-context decision transformer: Reinforcement learning via hierarchical chain-of-thought. *arXiv preprint arXiv:2405.20692*, 2024a.

Sili Huang, Jifeng Hu, Zhejian Yang, Liwei Yang, Tao Luo, Hechang Chen, Lichao Sun, and Bo Yang. Decision mamba: Reinforcement learning via hybrid selective sequence modeling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 72688–72709, 2024b.

Peter Humphreys, Arthur Guez, Olivier Tieleman, Laurent Sifre, Theophane Weber, and Timothy Lillicrap. Large-scale retrieval for reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 20092–20104, 2022.

Jan Humplik, Alexandre Galashov, Leonard Hasenclever, Pedro A. Ortega, Yee Whye Teh, and Nicolas Heess. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424*, 2019.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, 2021.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Advances in Neural Information Processing Systems*, 2023.

Minqi Jiang, Michael Dennis, Jack Parker-Holder, Jakob N. Foerster, Edward Grefenstette, and Tim Rocktaschel. Replay-guided adversarial environment design. In *Advances in Neural Information Processing Systems*, 2021a.

Minqi Jiang, Edward Grefenstette, and Tim Rocktaschel. Prioritized level replay. In *International Conference on Machine Learning*, 2021b.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Pierre-Alexandre Kamienny, Matteo Pirotta, Alessandro Lazaric, Thibault Lavril, Nicolas Usunier, and Ludovic Denoyer. Learning adaptive exploration strategies in dynamic environments through informed policy regularization. *arXiv preprint arXiv:2005.02934*, 2020.

Tobias Katsch. Gateloop: Fully data-controlled linear recurrence for sequence modeling. *arXiv preprint arXiv:2311.01927*, 2023.

Nathaniel S. Keplinger, Baiting Luo, Iliyas Bektas, Yunuo Zhang, Kyle Hollins Wray, Aron Laszka, Abhishek Dubey, and Ayan Mukhopadhyay. NS-Gym: Open-source simulation environments and benchmarks for non-stationary markov decision processes. *arXiv preprint arXiv:2501.09646*, 2025.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Taylor W. Killian, Samuel Daulton, George Konidaris, and Finale Doshi-Velez. Robust and efficient transfer learning with hidden-parameter markov decision processes. In *Advances in Neural Information Processing Systems*, 2017.

Baekjin Kim and Ambuj Tewari. Randomized exploration for non-stationary stochastic linear bandits. In *Conference on Uncertainty in Artificial Intelligence*, pp. 71–80, 2020.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. General-purpose in-context learning by meta-learning transformers. In *NeurIPS Workshop on Meta-Learning*, 2022.

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

Akshay Krishnamurthy, Keegan Harris, Dylan J. Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? In *Advances in Neural Information Processing Systems*, 2024.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

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.

Heinrich Küttler, Nantas Nardelli, Alexander H. Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The NetHack learning environment. In *Advances in Neural Information Processing Systems*, 2020.

Minjae Kwon, Amir Moeini, Shangtong Zhang, and Lu Feng. Latent Q-barrier shielding for safe in-context reinforcement learning. *arXiv preprint arXiv:2605.25267*, 2026.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, 2019.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Stenberg Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. In *International Conference on Learning Representations*, 2023.

Erwan Lecarpentier and Emmanuel Rachelson. Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Jonathan N. Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

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*, 2022.

Timothee Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Diaz-Rodriguez. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information Fusion*, 58:52–68, 2020.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Shiqian Li, Kewen Wu, Chi Zhang, and Yixin Zhu. I-PHYRE: Interactive physical reasoning. In *International Conference on Learning Representations*, 2024a.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024b.

Wenzhe Li, Hao Luo, Zichuan Lin, Chongjie Zhang, Zongqing Lu, and Deheng Ye. A survey on transformers in reinforcement learning. *Transactions on Machine Learning Research*, 2023.

Yingying Li and Na Li. Online learning for markov decision processes in nonstationary environments: A dynamic regret analysis. In *American Control Conference*, pp. 1232–1237, 2019.

Haodong Liang and Lifeng Lai. Transformers provably implement in-context reinforcement learning with policy improvement. *arXiv preprint arXiv:2605.05755*, 2026.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *International Conference on Learning Representations*, 2024.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.

Evan Z. Liu, Aditi Raghunathan, Percy Liang, and Chelsea Finn. Decoupling exploration and exploitation for meta-reinforcement learning without sacrifices. In *International Conference on Machine Learning*, pp. 6925–6935, 2021.

Hao Liu and Pieter Abbeel. Emergent agentic transformer from chain of hindsight experience. In *International Conference on Machine Learning*, 2023.

Jinmei Liu, Fuhong Liu, Zhenhong Sun, Jianye Hao, Huaxiong Li, Bo Wang, Daoyi Dong, Chunlin Chen, and Zhi Wang. Scalable in-context Q-learning. In *International Conference on Learning Representations*, 2026.

Vincenzo Lomonaco, Karan Desai, Eugenio Culurciello, and Davide Maltoni. Continual reinforcement learning in 3d non-stationary environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 999–1008, 2020.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Chris Lu, Yannick Schroecker, Albert Gu, Emilio Parisotto, Jakob Foerster, Satinder Singh, and Feryal Behbahani. Structured state space models for in-context reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Baiting Luo, Yunuo Zhang, Abhishek Dubey, and Ayan Mukhopadhyay. Act as you learn: Adaptive decision-making in non-stationary markov decision processes. In *International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 1301–1309, 2024.

Hao Luo, Jiechuan Jiang, and Zongqing Lu. Model-based decentralized policy optimization. *arXiv preprint arXiv:2302.08139*, 2023.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning*, 2021.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Basar. Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In *International Conference on Machine Learning*, pp. 7447–7458, 2021.

Fernando Martinez-Lopez, Tao Li, Yingdong Lu, and Juntao Chen. In-context reinforcement learning via communicative world models. *arXiv preprint arXiv:2508.06659*, 2025.

Michael Matthews, Michael Beukman, Benjamin Ellis, Mikayel Samvelyan, Matthew Thomas Jackson, Samuel Coward, and Jakob Nicolaus Foerster. Craftax: A lightning-fast benchmark for open-ended reinforcement learning. In *International Conference on Machine Learning*, 2024a.

Michael Matthews, Michael Beukman, Chris Lu, and Jakob Foerster. Kinetix: Investigating the training of general agents through open-ended physics-based control tasks. *arXiv preprint arXiv:2410.23208*, 2024b.

Luckeciano C. Melo. Transformers are meta-reinforcement learners. In *International Conference on Machine Learning*, pp. 15340–15359, 2022.

Russell Mendonca, Xinyang Geng, Chelsea Finn, and Sergey Levine. Meta-reinforcement learning robust to distributional shift via model identification and experience relabeling. *arXiv preprint arXiv:2006.07178*, 2020.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022. doi: 10.18653/v1/2022.emnlp-main.759.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2018.

Amir Moeini, Minjae Kwon, Alper Kamil Bozkurt, Yuichi Motai, Rohan Chandra, Lu Feng, and Shangtong Zhang. Safe in-context reinforcement learning. *arXiv preprint arXiv:2509.25582*, 2025a.

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*, 2025b.

Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in markov decision processes. *arXiv preprint*, 2012.

Giovanni Monea, Antoine Bosselut, Kianté Brantley, and Yoav Artzi. LLMs are in-context bandit reinforcement learners. *arXiv preprint arXiv:2410.05362*, 2024.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. POPGym: Benchmarking partially observable reinforcement learning. In *International Conference on Learning Representations*, 2023.

Tianwei Ni, Benjamin Eysenbach, and Ruslan Salakhutdinov. Recurrent model-free rl can be a strong baseline for many pomdps. *arXiv preprint arXiv:2110.05038*, 2022.

Tianwei Ni, Michel Ma, Benjamin Eysenbach, and Pierre-Luc Bacon. When do transformers shine in rl? decoupling memory from credit assignment. *Advances in Neural Information Processing Systems*, 36, 2024.

Allen Nie, Yi Su, Bo Chang, Jonathan N. Lee, Ed H. Chi, Quoc V. Le, and Minmin Chen. EVOLvE: Evaluating and optimizing LLMs for in-context exploration. In *International Conference on Machine Learning*, 2025.

Alexander Nikulin, Vladislav Kurenkov, Ilya Zisman, Artem Agarkov, Viacheslav Sinii, and Sergey Kolesnikov. XLand-MiniGrid: Scalable meta-reinforcement learning environments in JAX. In *Advances in Neural Information Processing Systems*, 2024.

Alexander Nikulin, Ilya Zisman, Alexey Zemtsov, Viacheslav Sinii, Vladislav Kurenkov, and Sergey Kolesnikov. XLand-100B: A large-scale multi-task dataset for in-context reinforcement learning. In *International Conference on Learning Representations*, 2025.

Ben Norman and Jeff Clune. First-explore, then exploit: Meta-learning to solve hard exploration-exploitation trade-offs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 27490–27528, 2024.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, et al. Open X-embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation*, 2024.

Toshihiro Ota. Decision mamba: Reinforcement learning via sequence modeling with selective state spaces. *arXiv preprint arXiv:2403.19925*, 2024.

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

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Sindhu Padakandla. A survey of reinforcement learning algorithms for dynamically varying environments. *ACM Computing Surveys*, 2020.

Sindhu Padakandla, Prabuchandran K. J., and Shalabh Bhatnagar. Reinforcement learning in non-stationary environments. *Applied Intelligence*, 2020.

Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. Do llm agents have regret? a case study in online learning and games. *arXiv preprint*, 2024.

Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob N. Foerster, Edward Grefenstette, and Tim Rocktaschel. Evolving curricula with regret-based environment design. In *International Conference on Machine Learning*, 2022.

Jurgis Pasukonis, Timothy Lillicrap, and Danijar Hafner. Evaluating long-term memory in 3d mazes. *arXiv preprint arXiv:2210.13383*, 2022.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 2024.

Christian Perez, Felipe Petroski Such, and Theofanis Karaletsos. Generalized hidden parameter mdps: Transferable model-based rl in a handful of trials. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5403–5411, 2020.

Ava Pettet, Yunuo Zhang, Baiting Luo, Kyle Wray, Hendrik Baier, Aron Laszka, Abhishek Dubey, and Ayan Mukhopadhyay. Decision making in non-stationary environments with policy-augmented search. In *International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 2417–2419, 2024.

Andrei Polubarov, Nikita Lyubaykin, Alexander Derevyagin, Artyom Grishin, Igor Saprygin, Aleksandr Serkov, Mark Averchenko, Daniil Tikhonov, Maksim Zhdanov, Alexander Nikulin, Ilya Zisman, Albina Klepach, Alexey Zemtsov, and Vladislav Kurenkov. Vintix II: Decision pre-trained transformer is a scalable in-context reinforcement learner. In *International Conference on Learning Representations*, 2026. Poster; arXiv:2604.05112.

Andrey Polubarov, Nikita Lyubaykin, Alexander Derevyagin, Ilya Zisman, Denis Tarasov, Alexander Nikulin, and Vladislav Kurenkov. Vintix: Action model via in-context reinforcement learning. *arXiv preprint arXiv:2501.19400*, 2025.

Sam Powers, Eliot Xing, Eric Kolve, Roozbeh Mottaghi, and Abhinav Gupta. CORA: Benchmarks, baselines, and metrics as a platform for continual reinforcement learning agents. *arXiv preprint arXiv:2110.10067*, 2021.

Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10237–10257, 2023. doi: 10.1109/TNNLS.2023.3250269.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

Zhen Qin, Jiachen Jiang, and Zhihui Zhu. Learning to adapt: In-context learning beyond stationarity. *arXiv preprint arXiv:2604.10946*, 2026.

Maria Abi Raad, Arun Ahuja, Catarina Barros, Frederic Besse, Andrew Bolt, Adrian Bolton, Bethanie Brownfield, Gavin Buttimore, Max Cant, Sarah Chakera, et al. Scaling instructable agents across many simulated worlds. *arXiv preprint arXiv:2404.10179*, 2024.

Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International Conference on Machine Learning*, 2019.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. doi: 10.1162/tacl_a_00605.

Sharath Chandra Raparthy, Eric Hambro, Robert Kirk, Mikael Henaff, and Roberta Raileanu. Generalization to new sequential decision making tasks with in-context learning. In *International Conference on Machine Learning*, pp. 42138–42158, 2024.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, 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.

Hongyu Ren, Yuke Zhu, Jure Leskovec, Anima Anandkumar, and Animesh Garg. Ocean: Online task inference for compositional tasks with context adaptation. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1378–1387, 2020.

Micah Rentschler and Jesse Roberts. RL + transformer = a general-purpose problem solver. In *Proceedings of the 1st Workshop for Research on Agent Language Models*, pp. 401–410, 2025.

Samuel Ritter, Jane X. Wang, Zeb Kurth-Nelson, Siddhant M. Jayakumar, Charles Blundell, Razvan Pascanu, and Matthew Botvinick. Been there, done that: Meta-learning with episodic recall. In *International Conference on Machine Learning*, 2018.

Yoan Russac, Claire Vernade, and Olivier Cappe. Weighted linear bandits for non-stationary environments. *arXiv preprint arXiv:1909.09146*, 2019.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Alexander Rutherford, Michael Beukman, Timon Willi, Bruno Lacerda, Nick Hawes, and Jakob N. Foerster. No regrets: Investigating and improving regret approximations for curriculum discovery. In *Advances in Neural Information Processing Systems*, 2024.

Mikayel Samvelyan, Robert Kirk, Vitaly Kurin, Jack Parker-Holder, Minqi Jiang, Eric Hambro, Fabio Petroni, Heinrich Küttler, Edward Grefenstette, and Tim Rocktäschel. MiniHack the planet: A sandbox for open-ended reinforcement learning research. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Gabriel Sarch, Lawrence Jang, Michael J. Tarr, William W. Cohen, Kenneth Marino, and Katerina Fragkiadaki. VLM agents generate their own memories: Distilling experience into embodied programs of thought. In *Advances in Neural Information Processing Systems*, 2024.

Harsh Satija, Philip Amortila, and Joelle Pineau. Constrained markov decision processes via backward value functions. In *International Conference on Machine Learning*, 2020.

Ali H. Sayed. *Adaptive Filters*. John Wiley and Sons, 2011.

Thomas Schmied, Fabian Paischer, Vihang Patil, Markus Hofmarcher, Razvan Pascanu, and Sepp Hochreiter. Retrieval-augmented decision transformer: External memory for in-context RL. *arXiv preprint arXiv:2410.07071*, 2024.

Gresa Shala, Andre Biedenkapp, and Josif Grabocka. Hierarchical transformers are efficient meta-reinforcement learners. *arXiv preprint arXiv:2402.06402*, 2024.

Lucy Xiaoyang Shi, Yunfan Jiang, Jake Grigsby, Linxi Fan, and Yuke Zhu. Cross-episodic curriculum for transformer agents. In *Advances in Neural Information Processing Systems*, volume 36, pp. 13–34, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Viacheslav Sinii, Alexander Nikulin, Vladislav Kurenkov, Ilya Zisman, and Sergey Kolesnikov. In-context reinforcement learning for variable action spaces. In *International Conference on Machine Learning*, 2024.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023.

Jaehyeon Son, Soochan Lee, and Gunhee Kim. Distilling reinforcement learning algorithms for in-context model-based planning. In *International Conference on Learning Representations*, 2025.

Kefan Song, Amir Moeini, Peng Wang, Lei Gong, Rohan Chandra, Shangtong Zhang, and Yanjun Qi. Reward is enough: LLMs are in-context reinforcement learners. *arXiv preprint arXiv:2506.06303*, 2025.

Kaustubh Sridhar, Souradeep Dutta, Dinesh Jayaraman, and Insup Lee. REGENT: A retrieval-augmented generalist agent that can act in-context in new environments. *arXiv preprint arXiv:2412.04759*, 2024.

Bradly C. Stadie, Ge Yang, Rein Houthooft, Xi Chen, Yan Duan, Yuhuai Wu, Pieter Abbeel, and Ilya Sutskever. Some considerations on learning to explore via meta-reinforcement learning. *arXiv preprint arXiv:1803.01118*, 2018.

Christian Alexander Steinparz, Thomas Schmied, Fabian Paischer, Marius-Constantin Dinu, Vihang Prakash Patil, Angela Bitto-Nemling, Hamid Eghbal-zadeh, and Sepp Hochreiter. Reactive exploration to cope with non-stationarity in lifelong reinforcement learning. In *Conference on Lifelong Learning Agents*, pp. 441–469, 2022.

Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, Nat McAleese, Nathalie Bradley-Schmieg, Nathaniel Wong, Nicolas Porcel, Roberta Raileanu, Steph Hughes-Fitt, Valentin Dalibard, and Wojciech Marian Czarnecki. Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*, 2021.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

Fahim Tajwar, Yiding Jiang, Abitha Thankaraj, Sumaita Sadia Rahman, J. Zico Kolter, Jeff Schneider, and Ruslan Salakhutdinov. Training a generally curious agent. *arXiv preprint arXiv:2502.17543*, 2025.

Weihao Tan, Wentao Zhang, Xinrun Xu, Haochong Xia, Ziluo Ding, Boyu Li, Bohan Zhou, Junpeng Yue, Jiechuan Jiang, Yewen Li, Ruyi An, Molei Qin, Chuqiao Zong, Longtao Zheng, Yujie Wu, Xiaoqiang Chai, Yifei Bi, Tianbao Xie, Pengjie Gu, Xiyun Li, Ceyao Zhang, Long Tian, Chaojie Wang, Xinrun Wang, Börje F. Karlsson, Bo An, Shuicheng Yan, and Zongqing Lu. Cradle: Empowering foundation agents towards general computer control. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58658–58725. PMLR, 2025. URL https://proceedings.mlr.press/v267/tan25h.html.

Denis Tarasov, Alexander Nikulin, Ilya Zisman, Albina Klepach, Andrei Polubarov, Nikita Lyubaykin, Alexander Derevyagin, Igor Kiselev, and Vladislav Kurenkov. Yes, Q-learning helps offline in-context RL. *arXiv preprint arXiv:2502.17666*, 2025. Version 4, revised 25 May 2026.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *Advances in Neural Information Processing Systems*, 34:26980–26992, 2021.

Francesco Trovo, Stefano Paladino, Marcello Restelli, and Nicola Gatti. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 2020.

Bert Verbruggen, Arne Vanhoyweghen, and Vincent Ginis. Model-agnostic solutions for deep reinforcement learning in non-ergodic contexts. *arXiv preprint arXiv:2601.08726*, 2026.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, 2023.

Bernie Wang, Simon Xu, Kurt Keutzer, Yang Gao, and Bichen Wu. Improving context-based meta-reinforcement learning with self-supervised trajectory contrastive learning. *arXiv preprint arXiv:2103.06386*, 2021a.

Fan Wang, Pengtao Shao, Yiming Zhang, Bo Yu, Shaoshan Liu, Ning Ding, Yang Cao, Yu Kang, and Haifeng Wang. Towards large-scale in-context reinforcement learning by meta-training in randomized worlds. In *Advances in Neural Information Processing Systems*, 2025a.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Jane X. Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Jane X. Wang, Michael King, Nicolas Porcel, Zeb Kurth-Nelson, Tina Zhu, Charlie Deck, Peter Choy, Mary Cassin, Malcolm Reynolds, Francis Song, Gavin Buttimore, David P. Reichert, Neil Rabinowitz, Loic Matthey, Demis Hassabis, Alexander Lerchner, and Matthew Botvinick. Alchemy: A benchmark and analysis toolkit for meta-reinforcement learning agents. *arXiv preprint arXiv:2102.02926*, 2021b.

Jiuqi Wang, Ethan Blaser, Hadi Daneshmand, and Shangtong Zhang. Transformers learn temporal difference methods for in-context reinforcement learning. *arXiv preprint arXiv:2405.13861*, 2024a.

Jiuqi Wang, Rohan Chandra, and Shangtong Zhang. Towards provable emergence of in-context reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025b.

Kangrui Wang, Pingyue Zhang, Zihan Wang, Yaning Gao, Linjie Li, Qineng Wang, Hanyang Chen, Chi Wan, Yiping Lu, Zhengyuan Yang, Lijuan Wang, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, Yejin Choi, and Manling Li. VAGEN: Reinforcing world model reasoning for multi-turn VLM agents. In *Advances in Neural Information Processing Systems*, 2025c.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 2024b.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Conference on Empirical Methods in Natural Language Processing*, 2022.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74530–74543, 2023b.

Ye Wang, Kaiqian Cui, Tao Zhang, and Qin Jin. Towards unpredictable worlds: Continual in-context reinforcement learning in non-stationary environments. *OpenReview submission to ICLR 2026*, 2025d.

Zekang Wang, Zhe He, Borong Zhang, Edan Toledo, and Steven Morad. Investigating memory in RL with POPGym arcade. *arXiv preprint arXiv:2503.01450*, 2025e.

Zhe Wang, Haozhu Wang, and Yanjun Qi. Hierarchical prompt decision transformer: Improving few-shot policy generalization with global and adaptive guidance. In *Companion Proceedings of the ACM Web Conference*, pp. 520–529, 2025f. doi: 10.1145/3701716.3715233.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on Learning Theory*, pp. 4300–4354, 2021.

Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Maciej Wolczyk, Michal Zajac, Razvan Pascanu, Lukasz Kucinski, and Piotr Milos. Disentangling transfer in continual reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 6304–6317, 2022.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. iVideoGPT: Interactive VideoGPTs are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024.

Wenhao Wu, Fuhong Liu, Haoru Li, Zican Hu, Daoyi Dong, Chunlin Chen, and Zhi Wang. Mixture-of-experts meets in-context reinforcement learning. *arXiv preprint arXiv:2506.05426*, 2025.

Zixuan Xie, Xinyu Liu, Rohan Chandra, and Shangtong Zhang. Convergence and emergence of in-context reinforcement learning with chain of thought. *arXiv preprint arXiv:2605.07123*, 2026.

Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua B. Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *International Conference on Machine Learning*, 2022.

Qiushui Xu, Yuhao Huang, Yushu Jiang, Lei Song, Jinyu Wang, Wenliang Zheng, and Jiang Bian. In-context compositional Q-learning for offline reinforcement learning. In *International Conference on Learning Representations*, 2026a. arXiv:2509.24067.

TengYe Xu, Zihao Li, and Qinyuan Ren. Meta-reinforcement learning robust to distributional shift via performing lifelong in-context learning. In *International Conference on Machine Learning*, 2024.

Xinrun Xu, Pi Bu, Ye Wang, Börje F. Karlsson, Ziming Wang, Tengtao Song, Qi Zhu, Jun Song, Zhiming Ding, and Bo Zheng. DeepPHY: Benchmarking agentic VLMs on physical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34160–34168, 2026b. doi: 10.1609/aaai.v40i40.40711.

Xinrun Xu, Pi Bu, Ye Wang, Börje F. Karlsson, Ziming Wang, Tengtao Song, Qi Zhu, Jun Song, Shuo Zhang, Zhiming Ding, and Bo Zheng. ICPRL: Acquiring physical intuition from interactive control. *arXiv preprint arXiv:2603.13295*, 2026c.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023a.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023b.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*, 2023.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Avnish Narayan, Hayden Shively, Adithya Bellathur, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, 2020a.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, 2020b.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995, 2017.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *AAAI Conference on Artificial Intelligence*, 2024.

Qinqing Zheng, Amy Zhang, and Aditya Grover. Online decision transformer. In *International Conference on Machine Learning*, pp. 27042–27059, 2022.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In *International Conference on Machine Learning*, 2024.

Luisa M. Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A very good method for bayes-adaptive deep RL via meta-learning. In *International Conference on Learning Representations*, 2020.

Luisa M. Zintgraf, Leo Feng, Cong Lu, Maximilian Igl, Kristian Hartikainen, Katja Hofmann, and Shimon Whiteson. Exploration in approximate hyper-state space for meta reinforcement learning. In *International Conference on Machine Learning*, pp. 12991–13001, 2021.

Ilya Zisman, Vladislav Kurenkov, Alexander Nikulin, Viacheslav Sinii, and Sergey Kolesnikov. Emergence of in-context reinforcement learning from noise distillation. In *International Conference on Machine Learning*, 2024.