WHEN AGENTS CHANGE THEIR OWN TEST: PROSPECTIVE TARGET PRESERVATION UNDER CANDIDATE-INDUCED TASK EXPANSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-ended and self-improving (*autotelic*) agents increasingly help construct the tasks, tests, or subproblems on which progress is later judged. This creates a same-candidate dilemma: if candidate update U_t expands the evaluation population, may those candidate-caused additions change the objective used to decide whether U_t itself is accepted? We formalize *candidate-induced target endogeneity* and distinguish it from adaptive sampling, missing evaluation coverage (support failure), and intentionally performative objectives. Under *prospective* semantics, the current target distribution is fixed using only pre-candidate information; evidence may still be sampled adaptively and reweighted to that target, while candidate-caused tasks affect only later targets. We prove that any nontrivial same-round target change can reverse some bounded threshold decision even under exact, noiseless evaluation, so greater statistical precision cannot repair a change in the quantity being judged. Exact replay on 384 recorded LLM update-effect vectors yields false-promotion rates (acceptance under the post-candidate target despite rejection under the prospective target) of 73.4% and 54.1% in the primary and independently seeded confirmation phases when only the target semantics change. More importantly, a pre-specified capability-conditioned task-admission analysis on naturally trained Prioritized Level Replay policies in the Procgen benchmark finds 13/48 false promotions across 11/16 StarPilot seeds and, in an independently seeded replication, 36/192 across 29/64 Heist seeds. Eleven Heist reversals have prospective and post-target signs jointly separated from zero under our 95% certificate. Mean disagreement under same-size candidate-independent expansion is 3.2% and 2.5%, respectively. The Procgen CoinRun environment has no primary false promotions, providing a negative boundary. These results identify target timing as a distinct evaluation contract for self-expanding agents.

1 INTRODUCTION

Adaptive agents increasingly participate in constructing their own evaluation universe. Autotelic agents (that set, select, and pursue their own goals) build growing skill repertoires (Colas et al., 2022; Molinaro et al., 2026; Sigaud et al., 2023; Gaven et al., 2025); open-ended and unsupervised environment-design methods generate or evolve tasks in response to learner capability (Wang et al., 2019; Dennis et al., 2020; Parker-Holder et al., 2022); self-evolving benchmarks use models or agent trajectories to construct harder future tests (Wang et al., 2025; Guo et al., 2026); and long-horizon research agents use accept/reject gates before expanding strategies into downstream subproblems (Lin et al., 2026). These systems make evaluation adaptive in a deeper sense than ordinary selective sampling: the learner or candidate can influence which tasks exist, become admissible, or are considered relevant later.

This paper isolates one temporal question inside that broader trend. At round t, learner M_t is evaluated on a current task population $\mathcal{X}_t$. A candidate update U_t is proposed. The candidate may discover, generate, unlock, or certify additional tasks $G_t(U_t)$. If those additions enter the current evaluation population before the accept/reject decision (which we call *promotion*) is made, then the candidate partly changes the objective that judges its own survival. We call this *candidate-induced target endogeneity*. Which target is legitimate for that decision: the one declared before the candidate,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

or the one produced after the candidate's consequences are admitted? When the post-candidate target accepts a candidate that the prospective target rejects, we call the event a *false promotion*; the opposite reversal is a *false rejection*.

The question is upstream of ordinary bias correction. Active testing, off-policy evaluation (OPE), and high-confidence policy improvement (HCPI) estimate or certify a *specified target quantity* from selectively acquired data (Sawade et al., 2010; Kossen et al., 2021; Thomas et al., 2015). Automatic curriculum learning (ACL) and SAMPLR (a curriculum method designed to preserve ground-truth utility despite curriculum-induced distribution shift) similarly distinguish the distribution used to acquire or train on data from the objective ultimately being evaluated (Portelas et al., 2020; Jiang et al., 2022). These methods answer "how do we estimate the target?" once the target is fixed. They do not decide which same-round target is authorized when the candidate itself changes the task population.

Nor is every candidate-dependent target illegitimate. By a *target contract*, we mean the rule specifying which task distribution defines the current accept/reject decision and when that distribution is allowed to change. Performative learning deliberately evaluates decisions under induced distributions (Perdomo et al., 2020; Mandal et al., 2023). DeepFact's Audit-then-Score (AtS) protocol explicitly lets a challenger propose benchmark-label revisions that an auditor can accept before the challenger is scored (Huang et al., 2026). These are valid but different target contracts. Our scope is the common alternative in which current improvement is intended to be *prospective*: judge the candidate against a target fixed before its current consequences, then use those consequences to define later objectives.

We call this principle **prospective target preservation** (PTP). It implements the temporal contract

$$\boxed{\text{freeze} \;\rightarrow\; \text{acquire} \;\rightarrow\; \text{transport} \;\rightarrow\; \text{decide} \;\rightarrow\; \text{release}}.$$

The target is fixed before the current decision, not the learning process: acquisition may remain adaptive, and candidate-caused tasks may immediately influence the next round.

Our contributions are fourfold. **(1)** We formalize same-candidate target endogeneity separately from acquisition mismatch and missing acquisition support—target-relevant tasks that are never sampled—and characterize the prospective contract by requiring the current target to depend only on pre-candidate information. **(2)** We prove a target-invariance result: for any nontrivial change of current target and any interior promotion threshold, some bounded candidate effect reverses decision; hence no nontrivial target mutation can preserve all prospective decisions. **(3)** We isolate the mechanism with exact/noiseless replay and then test a pre-specified capability-conditioned task-admission rule on ordinary adaptive reinforcement-learning (RL) policies. An independent 64-seed Heist replication yields 36/192 false promotions across 29 seeds, including 11 high-precision certified cases. **(4)** We preserve negative boundaries: direct target sampling can be preferable, finite-budget transport can still err, CoinRun can remain unaffected, and preserving a declared target does not guarantee improvement on an unrelated downstream utility.

2 Related Work: Target contracts in adaptive evaluation

The closest literatures already contain most ingredients individually. Table 1 isolates the remaining distinction.

Dynamic benchmark theory is a particularly close structural ancestor: the current model can affect future benchmark data while an underlying population remains conceptually distinct (Shirali et al., 2023). Self-evolving benchmark work operationalizes automatic benchmark or task evolution at scale (Wang et al., 2025; Guo et al., 2026). DeepFact is closer temporally because the same challenger can trigger an audited benchmark correction before scoring (Huang et al., 2026); that protocol is appropriate precisely because correcting the benchmark label is part of its intended current semantics. Our question is complementary: when the evolving object is the *task population* and the declared promotion semantics are prospective, what prevents the candidate from moving its own current comparison scale?

The target contract is part of the problem definition. There is no estimator that can decide whether prospective, performative, or audited-reflexive semantics are "correct" in the abstract. Those contracts answer different questions. A prospective contract asks whether U_t improves the population

Table 1: Neighboring evaluation contracts. The contribution here is not dynamic benchmarking or importance weighting in general, but the prospective same-candidate rule for task-population expansion.

Line of work	What adapts?	Current candidate and current target	Relation to this paper
Active testing / OPE / HCPI (Kossen et al., 2021; Thomas et al., 2015)	acquired evidence / candidate estimate	evaluand fixed	solves estimation/certification once target is defined
ACL / SAMPLR (Portelas et al., 2020; Jiang et al., 2022)	curriculum / training distribution	target utility specified externally	establishes acquisition–target distinction
Open-ended / unsupervised environment design (UED) (Wang et al., 2019; Dennis et al., 2020; Parker-Holder et al., 2022)	tasks or environments evolve with learner capability	generated tasks drive training or capability acquisition	establishes adaptive task-population construction, but not which target may judge the same candidate that caused the expansion
Performative learning (Perdomo et al., 2020; Mandal et al., 2023)	deployed decision changes distribution	induced distribution is intentionally the objective	alternative, performative semantics
Dynamic and self-evolving benchmarks (Shirali et al., 2023; Wang et al., 2025; Guo et al., 2026)	future benchmark items/data	model/agent helps shape later benchmark	motivates evolving evaluation; no prospective promotion rule
DeepFact AtS (Huang et al., 2026)	benchmark labels/rationales	audited challenger evidence may revise benchmark before scoring	explicit reflexive contract for label correction, not task-population promotion
This paper	acquisition and task admission	current task target fixed pre-candidate; additions released after decision	prospective same-candidate contract

declared before its current consequences; a performative contract asks how U_t performs under the distribution it induces; an audited-reflexive contract may permit current benchmark correction when the candidate supplies evidence that the benchmark itself is wrong. Once the contract is declared, statistics can estimate its evaluand. Before it is declared, higher precision only estimates an ambiguous objective more precisely. This is why we treat target timing as an interface between objective definition and statistical evaluation.

To our knowledge, prior work has not formalized this prospective task-population promotion problem together with the target-invariance result below. We therefore avoid broader firstness claims about dynamic evaluation, decision-dependent distributions, benchmark correction, or self-evolving benchmarks.

3 Formal problem: when the candidate moves the target

Let $\mathcal{F}_t$ denote all information available before the current candidate is evaluated. The incumbent learner is M_t, the proposed update is U_t, and the current finite task universe is $\mathcal{X}_t$. Let $R(M, x)$ denote the expected evaluation score of learner M on task x, and let $M_t + U$ denote the learner obtained by applying candidate update U to M_t. The task-wise effect of the candidate is

$$d_t^U(x) = R(M_t + U, x) - R(M_t, x), \qquad |d_t^U(x)| \le 1.$$

A target distribution p_t specifies what counts as current improvement:

$$\Delta_{p,t}(U) = \mathbb{E}_{X \sim p_t}[d_t^U(X)], \qquad D_{p,t}(U) = \mathbf{1}\{\Delta_{p,t}(U) > \tau\}, \tag{1}$$

where $\Delta_{p,t}(U)$ is the candidate's mean effect under the target, τ is the pre-declared promotion threshold, and $\mathbf{1}\{\cdot\}$ is the indicator of acceptance. At evaluation draw i, evidence may instead be sampled from an adaptive acquisition distribution $q_{t,i}$. Candidate consequences $G_t(U)$ may additionally induce a candidate-dependent post-target distribution $r_t(U)$, or a mixture

$$\widetilde{p}_t(U) = (1 - \alpha)p_t + \alpha r_t(U), \qquad \alpha \in [0, 1],$$

where α is the target mass assigned to the candidate-induced component. We call the current decision *candidate-target-endogenous* when the target used to judge U_t is a nonconstant function of that same candidate or of consequences revealed only after proposing it. This is a property of the evaluation contract, not of the estimator.

Table 2: Distinct evaluation failures require distinct remedies. PTP addresses target timing; it does not replace coverage, statistical uncertainty control, or objective design.

Failure	What went wrong?	Appropriate remedy	What remains possible?
Target endogeneity	Candidate consequences change the objective used for the same candidate	Freeze the prospective current target; release additions afterward	Adaptive evidence collection and future target evolution
Acquisition mismatch	Evidence is sampled from $q \neq p$ while utility is defined under p	Sample directly from p or transport observations from q to p	Biased/adaptive acquisition when overlap holds
Missing support	Some target-relevant tasks have $p(x) > 0$ but $q(x) = 0$	Acquire additional coverage or introduce explicit modeling assumptions	No distribution-free recovery from q-only outcomes
Finite-sample error	The correct target is estimated imprecisely	More evaluation, variance control, or statistical certification	The target semantics remain unchanged
Objective misspecification	The declared p is not the utility the designer ultimately cares about	Redefine the target contract itself	PTP cannot make a poorly chosen objective desirable

Three different failure modes. Acquisition mismatch is $q_t \neq p_t$; support failure occurs when the target assigns positive probability to tasks that the acquisition distribution never samples; target endogeneity occurs when the current target itself changes from p_t to a candidate-dependent distribution such as $\widetilde{p}_t(U)$. These are different problems and require different remedies. Suppressing the round and candidate indices for readability, let Δ_r denote the exact candidate effect under any target distribution r, and let $\widehat{\Delta}_r$ be an estimate of that quantity. Then

$$\widehat{\Delta}_r - \Delta_p = (\widehat{\Delta}_r - \Delta_r) + (\Delta_r - \Delta_p), \tag{2}$$

where the first term is estimation error and the second is target-definition shift. Infinite data can remove the first term while leaving the second unchanged.

Table 2 summarizes the resulting diagnostic: PTP fixes only the target-timing component of the problem; the other rows remain genuine and orthogonal requirements.

Theorem 1 (Support limitation)**.** *If there is a task subset A with $p(A) > 0$ but $q(A) = 0$, then observations obtained only from q cannot determine even the sign of Δ_p for all bounded effect functions without additional modeling assumptions.*

This is the standard overlap limitation specialized to a promotion decision; proof is in Appendix A.

Proposition 1 (Exact same-candidate reversal)**.** *For $\widetilde{p}_U = (1-\alpha)p + \alpha r_U$,*

$$\Delta_{\widetilde{p}_U}(U) = (1-\alpha)\Delta_p(U) + \alpha\Delta_{r_U}(U).$$

Thus a candidate rejected by p can be accepted by $\widetilde{p}_U$ even when both expectations are known exactly.

The structural point is stronger than this mixture identity.

Theorem 2 (Universal prospective preservation forces target invariance)**.** *Let $p \neq r$ be distributions on finite $\mathcal{X}$ and let $\tau \in (-1, 1)$. There exists $d : \mathcal{X} \to [-1, 1]$ such that*

$$\mathbb{E}_p[d] < \tau < \mathbb{E}_r[d].$$

Consequently, any rule that preserves every *bounded prospective threshold decision must keep the current target unchanged.*

The proof constructs d from $\operatorname{sign}(r - p)$ and scales it around τ (Appendix A). The theorem is worst-case: it does not say every candidate reverses under every target change. It says that target mutation has no distribution-free decision-preservation guarantee, even under perfect measurement.

When is a particular candidate actually vulnerable? Theorem 2 is worst-case; for a fixed candidate the following bound gives an instance-specific stability condition. Define its prospective decision margin and task-wise effect range by

$$m_p(U) = |\Delta_p(U) - \tau|, \qquad L_U = \max_{x \in \mathcal{X}} d^U(x) - \min_{x \in \mathcal{X}} d^U(x).$$

Proposition 2 (Candidate-specific decision stability)**.** *Let*

$$\widetilde{p}_U = (1-\alpha)p + \alpha r_U, \qquad \alpha \in [0,1].$$

Then

$$|\Delta_{\widetilde{p}_U}(U) - \Delta_p(U)| \le \alpha L_U \,\mathrm{TV}(p, r_U), \tag{3}$$

where

$$\mathrm{TV}(p,r) = \frac{1}{2}\sum_x |p(x) - r(x)|$$

is total-variation distance. Consequently, the prospective and post-target decisions must agree whenever

$$m_p(U) > \alpha L_U \,\mathrm{TV}(p, r_U). \tag{4}$$

Equivalently, any decision reversal requires the prospective margin to be no larger than this candidate-specific target-drift budget.

The bound follows from $|\mathbb{E}_p f - \mathbb{E}_r f| \le (\max f - \min f)\mathrm{TV}(p,r)$ and $\Delta_{\widetilde{p}_U} - \Delta_p = \alpha(\Delta_{r_U} - \Delta_p)$. Thus reversal requires sufficient target movement, task-wise effect heterogeneity, and a small enough prospective margin—explaining why substantial mismatch need not produce reversal, as in CoinRun.

A common scale for comparing multiple candidates. The same issue arises when a system ranks several candidate updates rather than thresholding one candidate. Let

$$G_p(U,V) = \Delta_p(U) - \Delta_p(V)$$

be their prospective utility gap. If each candidate instead receives its own post-candidate target $\widetilde{p}_U$ and $\widetilde{p}_V$, then

$$\left|\left[\Delta_{\widetilde{p}_U}(U) - \Delta_{\widetilde{p}_V}(V)\right] - G_p(U,V)\right| \le \epsilon_U + \epsilon_V,$$

where

$$\epsilon_U = \alpha_U L_U \,\mathrm{TV}(p, r_U), \qquad \epsilon_V = \alpha_V L_V \,\mathrm{TV}(p, r_V).$$

Hence the prospective ranking is guaranteed to be preserved whenever

$$|G_p(U,V)| > \epsilon_U + \epsilon_V.$$

PTP makes this issue disappear at the target-definition level: candidates in the same round are evaluated against the same p_t. Candidate-specific targets may still be meaningful under explicitly performative semantics, but their numerical scores should not be interpreted as utilities on a common prospective scale.

4 Prospective target preservation

Under prospective semantics, the current target must be completely determined by information available before the candidate is judged. Formally,

$$\boxed{p_t \text{ is } \mathcal{F}_t\text{-measurable},}$$

meaning that constructing p_t cannot depend on consequences revealed only after proposing the current candidate. A practical protocol follows.

1. **Freeze**: declare p_t from $\mathcal{F}_t$ before current candidate consequences are admitted.
2. **Acquire**: collect evaluation evidence from predictable adaptive $q_{t,i}$.
3. **Transport**: provided that $q_{t,i}(x) > 0$ whenever $p_t(x) > 0$, reweight adaptively sampled evidence back to the fixed target:

$$\widehat{\Delta}_t^{\mathrm{PTP}} = \frac{1}{B}\sum_{i=1}^{B} \underbrace{\frac{p_t(X_i)}{q_{t,i}(X_i)}}_{\text{importance weight}} d_t^U(X_i), \tag{5}$$

   where B is the number of evaluation draws and $X_i \sim q_{t,i}$ is the task sampled at draw i.
4. **Decide**: apply the declared threshold or candidate comparison on the common target.

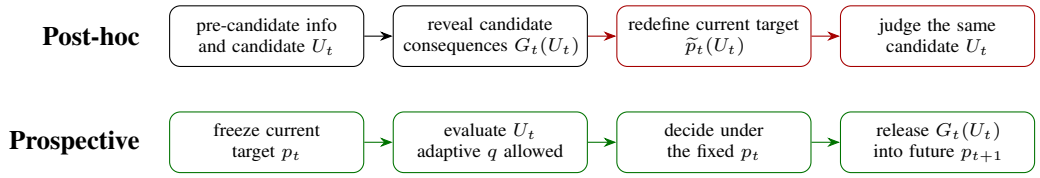

Figure 1: Two temporal contracts. Top: candidate-caused additions redefine the current target before the candidate is judged. Bottom: prospective target preservation fixes the current comparison and uses candidate-caused additions forward in time.

5. **Release**: admit $G_t(U_t)$ into $\mathcal{F}_{t+1}$, where it may alter p_{t+1}.

Theorem 3 (Correctness under prospective semantics)**.** *Fix round t. Suppose p_t and U_t are fixed before the current evaluation draws, each $q_{t,i}$ is chosen using only information available before draw i, $q_{t,i}(x) > 0$ whenever $p_t(x) > 0$, and d_t^U is bounded. Then Eq.* (5) *is conditionally unbiased for $\Delta_{p,t}(U_t)$. If, for some constant $\rho > 0$, $q_{t,i}(x) \geq \rho p_t(x)$ for every relevant task, then each importance weight is at most $1/\rho$ and the estimate is consistent. Releasing candidate-caused tasks only after the decision prevents them from changing the current estimand.*

The estimator is inherited importance sampling, not a claimed novelty. If sampling the target itself is inexpensive, direct target sampling (DIRECT-p) draws evaluation tasks directly from p_t and is simpler and slightly more accurate in our matched-budget controlled analysis. The novel object is the temporal contract that says which target the estimator is authorized to estimate.

Operational regimes and auditability. PTP freezes only the target version used for the current decision, not the learning process: candidates are judged against p_t, accepted discoveries can immediately shape p_{t+1}, and same-round candidates remain replayable and comparable under one criterion. Candidate-specific induced targets are instead appropriate when performative semantics are explicitly intended. If support under q is insufficient, neither timing nor importance weighting repairs the missing evidence. Figure 1 contrasts the two temporal contracts.

5 EVIDENCE

The experiments form an evidence ladder because no single study can establish mechanism, recurrence, and downstream consequences. SCER-1 isolates target mutation without acquisition noise; same-size random expansions test whether expansion alone explains the effect; the Procgen studies test recurrence across independently trained policies; CoinRun and StarPilot delimit when reversals do not certify; and the closed-loop study tests path dependence and an external utility that PTP does not optimize.

5.1 EXACT TARGET MUTATION WITH ACQUISITION REMOVED

Same-Candidate Endogeneity Replay (SCER-1) reuses all 384 task-wise effect vectors from our controlled LLM study on generated Python function-puzzle tasks but removes q-sampling entirely. Each candidate is scored exactly on the current archive. For each candidate, the candidate-linked component places mass only on tasks for which that candidate's recorded effect is positive; the post-target assigns fraction $\alpha = .20$ to this component and $1 - \alpha$ to the original target. The promotion threshold is $\tau = .02$. The candidate, its task-wise effect vector, and the prospective target are unchanged; only the target semantics differ. Table 3 reports the primary and independently seeded confirmation results.

Candidate-independent $r = p$ changes zero decisions. Expansion-mass, task-transfer, and threshold sensitivity preserve the qualitative effect (Appendix C). Figure 2 shows the exact score shifts.

5.2 NATURAL POLICIES WITH PROTOCOL-DEFINED CANDIDATE-INDUCED TASK ADMISSION

The stronger external test uses ordinary policies from Procgen, a benchmark of procedurally generated game levels, trained with the centralized Prioritized Level Replay (PLR) curriculum (Jiang et al.,

Table 3: SCER-1, exact/noiseless same-candidate target mutation. “0.5B” and “1.5B” denote solver sizes of 0.5 and 1.5 billion parameters. The 95% confidence intervals (CIs) use 50,000 bootstrap resamples of the independent replicate trajectories. The same-size random control is a secondary robustness analysis specified after the primary endpoint.“Expandable” means prospectively non-promotable candidates with at least one positive-effect task, so the candidate-linked expansion is nonempty.

Phase	False promotions	Rate (95% CI)	Among expandable candidates	Random-control rate
0.5B primary	116/158	73.4% [62.8,84.0]	116/148 = 78.4%	6.95%
1.5B confirm.	73/135	54.1% [42.9,63.2]	73/93 = 78.5%	4.70%

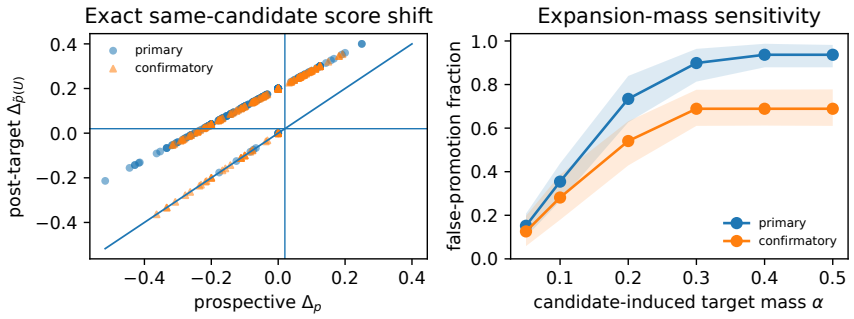

Figure 2: SCER-1. Left: exact prospective score versus candidate-linked post-target score for every candidate; “primary” denotes the 0.5B-parameter solver study and “confirmatory” the independently seeded 1.5B-parameter solver study. Right: false-promotion fraction versus candidate-induced target mass α. No acquisition sampling or finite-sample score estimation is involved.

2021; Cobbe et al., 2020). PLR adaptively prioritizes selected levels for replay during training. We test three Procgen environments: CoinRun, a platform-navigation task; Heist, a maze-navigation task requiring keys and locked passages; and StarPilot, an arcade-style space-navigation and combat task. The learner does *not* generate new Procgen levels. Instead, before outcomes are analyzed, we specify an admission rule on 200 existing levels: level identifiers 0–99 form the visible target and 100–199 form a latent reserve. The capability threshold is the 75th percentile of the *before-policy mean returns* on visible levels. Candidate update U_t unlocks reserve level z iff its after-policy mean return rises above that pre-candidate threshold while its before-policy mean return was below it. The prospective target is uniform over visible levels; the post-target is uniform over the visible levels plus the reserve levels unlocked by that same candidate. Thus policy training and returns are natural, while task admission is a controlled evaluation overlay representing candidate-induced expansion.

All checkpoint policies are evaluated with 50 paired episodes per level. A *false promotion* occurs when the prospective score is nonpositive but the candidate-expanded post-target score is positive. We call a reversal *joint-95 certified* when paired one-sided 97.5% confidence bounds place the prospective score below zero and the post-target score above zero; Appendix D gives the exact construction. A same-size candidate-independent control adds randomly chosen reserve levels using 10,000 draws per transition. Training seed is the independent statistical unit, and uncertainty for transition rates is computed by a seed-cluster bootstrap that resamples whole training seeds. Table 4 summarizes all four primary candidate-admission cohorts.

The 64-seed Heist result is the main prevalence estimate for this setup: 45.3% of independent seeds contain at least one reversal (29/64; Wilson binomial 95% confidence interval 33.7–57.4%), and 11 false promotions in 11 different seeds satisfy the conservative joint-95 certificate (prospective one-sided 97.5% upper bound < 0 and post-target one-sided 97.5% lower bound > 0). Several are far from zero (e.g., prospective/post scores $-0.354 \to +0.298$ and $-0.440 \to +0.169$). The observed transition false-promotion rate, 18.75%, is descriptively about 7.6$\times$ the 2.46% mean random-control disagreement. StarPilot independently shows recurrence at the cluster level, while CoinRun supplies an environment-level negative boundary. We therefore claim recurrence *within tested adaptive-learning regimes*, not a universal prevalence rate for open-ended agents.

Table 4: Candidate-induced task-admission results. "Affected" is the number of training seeds with at least one false promotion. Random is the mean disagreement probability under same-size candidate-independent reserve additions. The Heist64 evaluation protocol was specified before evaluating the independent replication seeds.

Environment	Study	Seeds	Transitions	False promotions	Affected seeds	Seed-cluster 95% CI	Certified / random control
CoinRun	initial cohort	8	24	0 (0.0%)	0/8	[0.0,0.0]%	0 / 0.00%
Heist	initial cohort	8	24	2 (8.3%)	1/8	[0.0,25.0]%	1 / 5.24%
StarPilot	independent cohort	16	48	13 (27.1%)	11/16	[16.7,37.5]%	0 / 3.19%
Heist	64-seed replication	64	192	36 (18.8%)	29/64	[13.5,24.5]%	11 / 2.46%

Why the effect is not mechanically forced. The unlock rule never inspects the visible-target promotion score, and every adjacent checkpoint transition is analyzed. A reversal additionally requires a nonpositive aggregate visible effect to become positive after candidate-linked admission. CoinRun has unlocks but zero primary reversals, same-size candidate-independent additions disagree only rarely, and 11 Heist transitions remain joint-95 certified. Thus expansion cardinality alone is insufficient: the effect depends on which tasks the candidate admits, its task-wise effect geometry, and its prospective margin, consistent with Proposition 2.

5.3 Acquisition mismatch and path dependence are separate

The original controlled LLM ecology isolates acquisition–target mismatch while keeping the target fixed. A Qwen3-4B generator maintains a validated Python-task archive; candidate learner updates are evaluated with Qwen2.5-Coder-0.5B-Instruct and, independently, Qwen2.5-Coder-1.5B-Instruct. Here p is the uniform archive target, S is the curriculum-selected skill subset, u_S is the uniform distribution over S, and the acquisition distribution is $q = .30p + .70u_S$. Each decision uses $B = 128$ sampled task effects and promotion threshold $\tau = .02$.

A *focused* candidate is a compact specialization update using at most eight demonstrations, emphasizing the curriculum-selected skill; a *balanced* control preserves the current memory and, when possible, appends up to two currently failed examples for rehearsal. We call a focused candidate *target-harmful* when its exact prospective target effect is at most $-.02$, and a balanced candidate *prospectively beneficial* when its exact target effect is at least $+.02$. Uncorrected q-weighted scoring promotes 59/93 and 55/95 target-harmful focused candidates, whereas prospective importance transport promotes 6/93 and 2/95; among prospectively beneficial balanced controls it retains 19/20 and 39/40 candidates. Mean estimator bias—estimated effect minus the exhaustive target effect, averaged across decisions—drops from $+0.3251$ to -0.00087 and from $+0.2399$ to $+0.00502$. Matched-budget DIRECT-p is slightly more target-accurate, as expected when direct target sampling is available.

A separate 200-trajectory, 120-round closed loop lets each judge's accept/reject decisions alter its future learner, archive, curriculum, and released frontier. Here *exact-target agreement* is the fraction of accept/reject decisions matching the decision obtained from exhaustive evaluation under the declared target, and a *target-harmful promotion* is an accepted update whose exact effect under that target is harmful. Prospective transport raises exact-target agreement from 54.7% to 79.7% and reduces target-harmful promotions from 41.73 to 15.55 per trajectory; DIRECT-p reaches 81.3% and 13.82. Yet a fixed external panel does *not* improve: the difference in area under the external-panel learning curve (AUC), prospective minus uncorrected, is -3.45×10^{-4} with 95% interval $[-4.35, -2.56] \times 10^{-4}$. This negative result is important: target preservation guarantees fidelity to the declared current semantics, not optimality of some other downstream objective.

6 Implications for self-expanding agent systems

PTP is most useful as a design test for systems in which evaluation and capability expansion interact. The question is not whether a system is called "open-ended," but whether an update being gated can influence the population used by that same gate. Three patterns make the issue concrete.

Self-evolving benchmarks. An agent trajectory can be used to construct harder benchmark tasks for subsequent agents (Guo et al., 2026). Under prospective semantics, benchmark version t remains the comparison scale for the candidate that helped reveal version $t + 1$; the generated tasks can be

released immediately afterward. This preserves both evolution and same-version comparability. It is not a prescription for benchmark systems whose explicit purpose is reflexive current correction.

Open-ended skills and curricula. An autotelic learner may discover goals that were previously absent or unreachable (Colas et al., 2022; Gaven et al., 2025). If a checkpoint gate asks whether the learner improved on the currently declared repertoire, newly discovered goals can enlarge the next repertoire without granting the discovering checkpoint a candidate-specific current target. Acquisition over the declared repertoire may still be highly nonuniform, which is the separate transport problem.

Agentic research and hierarchical decomposition. Long-horizon agent systems increasingly use gates before expanding accepted strategies into new downstream subproblems (Lin et al., 2026). PTP provides a general versioning principle for such architectures when the gate is meant to compare alternatives under a common pre-expansion criterion: judge first, then let the accepted branch change future obligations. This is an analogy and design implication, not a claim that the cited system currently exhibits the failure measured in our experiments. The same principle extends to candidate ranking: candidates evaluated against a common p_t remain comparable on one current utility scale, whereas candidate-specific post-targets can confound candidate quality with target movement; Section 3 gives the corresponding quantitative bound.

7 Scope, significance, and limitations

The same-candidate dilemma becomes operational when three conditions coincide: **(C1)** a consequential gate accepts or rejects a candidate; **(C2)** the candidate can change which tasks, goals, attacks, environments, benchmark items, or subproblems become admissible; and **(C3)** the intended current semantics are prospective. This pattern is broader than PLR. Self-evolving benchmarks and open-ended agents provide the task-expansion ingredient (Guo et al., 2026; Colas et al., 2022); research-agent pipelines provide explicit gates followed by downstream decomposition (Lin et al., 2026); dynamic benchmark and performative-learning work show that model-dependent data and induced distributions are legitimate objects of study (Shirali et al., 2023; Perdomo et al., 2020). We do not assert that any of those systems should use the prospective contract; the contract is appropriate only when the system's declared semantics require a common pre-candidate comparison.

The main limitations are equally explicit. First, the Procgen admission rule is a pre-specified capability-conditioned overlay on naturally trained policies, not literal agent generation of new levels. It establishes recurrence of the temporal mechanism under a natural learner plus controlled admission semantics; fully endogenous semantic goal invention remains future external validation. Second, Theorem 2 is structural/worst-case, not a prevalence theorem. Third, overlap and propensity observability remain assumptions for transport. Fourth, PTP does not design the globally best objective, and intentional performative objectives instantiate another valid contract. Finally, finite-budget evaluation can still make mistakes: the controlled studies retain nonzero harmful promotions after transport. The practical implication is therefore narrow but consequential: if candidate scores are meant to be comparable under a common current objective, task-population expansion must be versioned temporally. Candidate discoveries should affect future evaluation without silently changing the criterion that admits the candidate that produced them.

8 Conclusion

When autotelic agents can help construct the space over which progress is measured, evaluation requires an explicit temporal contract. Under prospective semantics, a candidate may change tomorrow's evaluation universe without rewriting the target that decides its survival today. Prospective target preservation formalizes that rule, standard transport handles adaptive evidence once the target is fixed, and Theorem 2 shows why target timing cannot be reduced to statistical precision. Exact replay isolates the mechanism, natural-policy task-admission analyses show recurrent false promotion in StarPilot and a large independent Heist replication, and CoinRun plus the downstream-utility negative result delimit the claim. The central lesson is simple: *first decide what counts as progress; then estimate it; only afterward let the candidate redefine what will count next.*

Reproducibility Statement

The supplementary package contains complete source and proofs; pre-specified protocols and analyzers for the exact same-candidate replay and Procgen task-admission studies; compact machine-readable outputs for the controlled LLM, StarPilot, cross-environment, and 64-seed Heist experiments; high-precision certificate records; claim-to-evidence and literature-audit documents. Large model weights, policy checkpoints, raw 50-episode panel arrays, and replay-event streams remain in the canonical internal archive and are intentionally not represented as included.

Ethics Statement

The controlled code-task experiments use synthetic programming tasks and locally executed language models; Procgen uses synthetic game environments. No human subjects, private user data, or sensitive personal datasets are involved. The principal integrity risk is methodological: adaptive evaluation can create misleading selection signals if acquisition, target definition, and candidate-caused task admission are conflated.

AI Use Statement

AI tools were solely used for literature discovery and writing refinement. The author checked literature-priority claims against primary or authoritative sources. All claims, proofs, code, citations, analyses, and the manuscript were done by the author, who takes responsibility for the submission.

References

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: A short survey. *Journal of Artificial Intelligence Research*, 74:1159–1199, 2022.

Loris Gaven, Thomas Carta, Clément Romac, Cédric Colas, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. MAGELLAN: Metacognitive predictions of learning progress guide autotelic LLM agents in large goal spaces. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, PMLR 267:18920–18953, 2025.

Dadi Guo, Tianyi Zhou, Dongrui Liu, Chen Qian, Qihan Ren, Shuai Shao, Zhiyuan Fan, Yi R. Fung, Kun Wang, Linfeng Zhang, and Jing Shao. Towards self-evolving agent benchmarks: Validatable agent trajectory via test-time exploration. In *International Conference on Learning Representations (ICLR)*, 2026.

Yukun Huang, Leonardo F. R. Ribeiro, Momchil Hardalov, Bhuwan Dhingra, Markus Dreyer, and Venkatesh Saligrama. DeepFact: Co-evolving benchmarks and agents for deep research factuality. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 34356–34386, 2026. doi:10.18653/v1/2026.acl-long.1586.

Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized level replay. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Minqi Jiang, Michael Dennis, Jack Parker-Holder, Andrei Lupu, Heinrich Küttler, Edward Grefenstette, Tim Rocktäschel, and Jakob Foerster. Grounding aleatoric uncertainty for unsupervised environment design. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Honghao Lin, David P. Woodruff, Yuan Deng, Jieming Mao, Song Zuo, and Vahab Mirrokni. Stellar Colosseum: A many-agent harness for long-horizon research in mathematics and theoretical computer science. arXiv:2609.15983, 2026.

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. Automatic curriculum learning for deep RL: A short survey. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

Christoph Sawade, Niels Landwehr, Steffen Bickel, and Tobias Scheffer. Active risk estimation. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 951–958, 2010.

Ali Shirali, Rediet Abebe, and Moritz Hardt. A theory of dynamic benchmarks. In *International Conference on Learning Representations (ICLR)*, 2023.

Olivier Sigaud, Gianluca Baldassarre, Cédric Colas, Stéphane Doncieux, Richard Duro, Pierre-Yves Oudeyer, Nicolas Perrin-Gilbert, and Vieri Giuliano Santucci. A definition of open-ended learning problems for goal-conditioned agents. arXiv:2311.00344, 2023.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Siyuan Wang, Zhuohan Long, Zhihao Fan, Xuanjing Huang, and Zhongyu Wei. Benchmark self-evolving: A multi-agent framework for dynamic LLM evaluation. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pp. 3310–3328, 2025.

Hongwei Yao, Yiming Liu, Meihui Chen, Jieling Chen, Zikun Chen, Yiling He, Wangze Ni, Cong Wang, and Kui Ren. ActBench: Self-evolving benchmark of behavioral safety in cowork agents. arXiv:2608.09476, 2026.

Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O. Stanley. POET: Open-ended coevolution of environments and their optimized solutions. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, pp. 142–151, 2019. doi:10.1145/3321707.3321799.

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 (NeurIPS)*, volume 33, pp. 13049–13061, 2020.

Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Evolving curricula with regret-based environment design. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, PMLR 162:17473–17498, 2022.

Gaia Molinaro, Anne G. E. Collins, Pierre-Yves Oudeyer, and Cédric Colas. Self-generated goals in the age of foundation models: Artificial autotelic agents and their implications for cognitive science. *Current Opinion in Behavioral Sciences*, 2026.

A Proofs and Additional Analysis

A.1 Uncorrected acquisition estimates the acquisition objective

Let $X_1, \ldots, X_B$ be independent and identically distributed (iid) draws from q, and define the uncorrected sample mean

$$\widehat{\Delta}_q^{\mathrm{uncorr}} = \frac{1}{B} \sum_{i=1}^{B} d(X_i).$$

Then

$$\mathbb{E}[\widehat{\Delta}_q^{\mathrm{uncorr}}] = \Delta_q, \qquad \mathbb{E}[\widehat{\Delta}_q^{\mathrm{uncorr}}] - \Delta_p = \sum_x (q(x) - p(x)) d(x).$$

Thus, if an acquisition-weighted score directly controls promotion, the system implements the acquisition objective whenever $q \neq p$. This is standard and is not claimed as a novel statistical result.

A.2 Proof of Theorem 1

Let A satisfy $p(A) > 0$ and $q(A) = 0$. Define two bounded effect functions

$$d_+(x) = \mathbf{1}\{x \in A\}, \qquad d_-(x) = -\mathbf{1}\{x \in A\}.$$

Because $q(A) = 0$, observations generated from either effect function under any number of draws from q are almost surely identical: both functions equal zero on the support of q. Yet

$$\mathbb{E}_p d_+ = p(A) > 0, \qquad \mathbb{E}_p d_- = -p(A) < 0.$$

Hence no procedure based only on q-supported observations can universally identify even the sign of target utility without assumptions relating unobserved and observed regions. This is the standard overlap limitation specialized to the promotion decision.

A.3 Proof of Proposition 1

Linearity gives

$$\Delta_{\widetilde{p}_U}(U) = (1-\alpha)\Delta_p(U) + \alpha\Delta_{r_U}(U).$$

The post-candidate judge accepts iff this quantity exceeds threshold τ, equivalently

$$\alpha\Delta_{r_U}(U) > \tau - (1-\alpha)\Delta_p(U).$$

There is no sampling term in this derivation. Therefore the decision can reverse even if both $\Delta_p(U)$ and $\Delta_{\widetilde{p}_U}(U)$ are known exactly. The issue is which estimand is authorized for the current decision.

A.4 Proof of Theorem 2

Because $r \neq p$, define

$$g(x) = \operatorname{sign}(r(x) - p(x)) \in \{-1, 0, 1\}.$$

Then

$$\delta := \mathbb{E}_r[g] - \mathbb{E}_p[g] = \sum_x (r(x) - p(x)) \operatorname{sign}(r(x) - p(x)) = \|r - p\|_1 > 0.$$

Let $m = (\mathbb{E}_p[g] + \mathbb{E}_r[g])/2$. Choose any

$$0 < a < \frac{1 - |\tau|}{1 + |m|}, \qquad b = \tau - am,$$

and define $d(x) = ag(x) + b$. Since $|g(x)| \leq 1$, $|d(x)| \leq |b| + a < 1$, so $d : \mathcal{X} \to [-1, 1]$. Moreover,

$$\mathbb{E}_p[d] = a\mathbb{E}_p[g] + \tau - am = \tau - \frac{a\delta}{2} < \tau,$$

$$\mathbb{E}_r[d] = a\mathbb{E}_r[g] + \tau - am = \tau + \frac{a\delta}{2} > \tau.$$

Thus every nontrivial current-target change reverses at least one admissible bounded threshold decision. The contrapositive gives the universal-preservation statement: a target rule that preserves every prospective decision cannot replace p by any $r \neq p$.

The theorem is deliberately worst-case. It does not state that every candidate reverses under every target change; it states that no nontrivial mutation of the current target can provide a distribution-free guarantee of preserving all prospective decisions.

A.5 Target-definition versus estimation error

Equation (2) is an exact identity for any target r and estimator $\widehat{\Delta}_r$. With unweighted samples from q, setting $r = q$ exposes acquisition-objective shift. With valid transport to p, $r = p$ and the target-definition term is zero. With a post-candidate target $r = \widetilde{p}(U)$, an arbitrarily precise estimator can drive the first term to zero while leaving $\Delta_{\widetilde{p}(U)} - \Delta_p$ unchanged. Missing support can make the first term non-identifiable for $r = p$. Definition error, acquisition error, support failure, and finite-sample error therefore require different remedies.

A.6 Proof of Theorem 3

Condition on the pre-draw history $\mathcal{H}_{i-1}$, containing $\mathcal{F}_t$, fixed candidate U_t, and previous evaluation draws. Predictability means $q_{t,i}$ is $\mathcal{H}_{i-1}$-measurable. Then

$$\mathbb{E}\left[\left.\frac{p_t(X_i)}{q_{t,i}(X_i)} d_t^U(X_i)\right| \mathcal{H}_{i-1}\right] = \sum_x q_{t,i}(x) \frac{p_t(x)}{q_{t,i}(x)} d_t^U(x)$$
$$= \sum_x p_t(x) d_t^U(x) = \Delta_{p,t}(U_t).$$

Averaging proves conditional unbiasedness of Eq. (5). If $q_{t,i}(x) \geq \rho p_t(x)$ and $|d_t^U(x)| \leq D$, then each weighted observation is bounded by D/ρ in magnitude. Standard martingale laws of large numbers give consistency; standard bounded martingale concentration yields finite-sample deviation control.

The procedural statement is separate. Because p_t is $\mathcal{F}_t$-measurable and candidate-caused tasks $G_t(U_t)$ are admitted only after the current decision, $G_t(U_t)$ is not an argument of the mapping that produced p_t. It may enter $\mathcal{F}_{t+1}$ and change p_{t+1}. Target freezing without transport can still estimate the wrong quantity from biased q data; transport without target freezing can accurately estimate a post-hoc target that was never the declared prospective objective.

A.7 A minimal self-expansion example

Suppose the current universe has n equally weighted tasks and candidate U has prospective mean $\Delta_p(U) = -\epsilon < 0$. The candidate causes m additional tasks to be admitted, each with measured effect $+1$, and a post-hoc judge becomes uniform over all $n + m$ tasks. Then

$$\Delta_{\text{post}}(U) = \frac{-n\epsilon + m}{n + m},$$

which is positive whenever $m > n\epsilon$. Thus even exact evaluation can reverse the current decision solely because favorable candidate-caused additions entered the target before the candidate was judged.

A.8 Candidate ranking and total variation

For candidates a, b, write $G_p = \Delta_p(a) - \Delta_p(b)$ and $G_q = \Delta_q(a) - \Delta_q(b)$. A ranking reversal occurs when $G_p G_q < 0$. Define the total-variation distance between two task distributions by

$$\text{TV}(p, q) = \frac{1}{2} \sum_x |p(x) - q(x)|.$$

If every candidate effect function has range at most L, meaning $\max_x d(x) - \min_x d(x) \leq L$, then

$$|\Delta_q(a) - \Delta_p(a)| \leq L \, \text{TV}(p, q).$$

If a_p^* and a_q^* maximize target and acquisition utility over the same finite candidate set, then

$$\Delta_p(a_p^*) - \Delta_p(a_q^*) \leq 2L \, \text{TV}(p, q).$$

This is a worst-case bound, not a predictor of realized sign reversals. CoinRun provides an empirical boundary: substantial $\text{TV}(p, q)$ can occur without harmful reversal.

B Literature and Priority Audit

Dynamic benchmarks, performative learning, self-evolving agent benchmarks, and benchmark correction all predate this paper. The original contribution is the *prospective task-population promotion contract*: a candidate may cause tasks to become admissible for future evaluation, but those same additions cannot alter the current target used for that candidate's own promotion if the declared semantics are prospective. Table 5 summarizes the closest neighboring evaluation contracts and the specific distinction retained by our PTP.

Table 5: Closest prior lines. "Grow" means that the model, agent, or evaluation process can change later task/data content; "Gate" means an explicit candidate accept/reject decision; and "same-candidate current revision" asks whether the candidate being judged can alter its own current benchmark or target before scoring.

Work	Adaptive evidence	Grow	Gate	Same-candidate current revision	Relation to PTP
Active risk/testing (Sawade et al., 2010; Kossen et al., 2021)	yes	no	model compare	no	Efficient/bias-corrected estimation of a specified risk; target already defined.
HCPI (Thomas et al., 2015)	logged	no	yes	no	Candidate certification is prior art; performance criterion is fixed.
ACL / SAMPLR (Portelas et al., 2020; Jiang et al., 2022)	yes	curriculum	optimization	no	Separates curriculum from target utility; does not govern candidate-caused current target expansion.
Autotelic / MAGELLAN (Colas et al., 2022; Gaven et al., 2025)	yes	goals	no same gate	no	Establishes evolving goal spaces and evaluation difficulty; focuses on competence/progress and goal choice.
Performative learning (Perdomo et al., 2020; Mandal et al., 2023)	not central	induced distribution	optimize	induced distribution is objective	Valid alternative contract: current objective may intentionally depend on the decision.
Dynamic benchmarks (Shirali et al., 2023)	yes	future benchmark	no same-update promotion	no	Model-dependent future benchmark data with a distinct underlying population.
Self-evolving agent benchmarks (?Guo et al., 2026; Yao et al., 2026)	trajectories	future tasks	benchmark evolution	no prospective gate formalized	Direct evidence that agent trajectories can construct harder future benchmark tasks.
DeepFact AtS (Huang et al., 2026)	evidence	labels/rationales	score after audit	yes, via audited correction	Explicit reflexive current contract for benchmark-label correction; not task-population promotion.
This paper	yes	task admission	yes	prohibited under prospective semantics	Formalizes pre-candidate target measurability and proves universal target invariance for prospective decisions.

Dynamic benchmarks and self-evolving benchmarks. Dynamic benchmark theory lets a current model shape future benchmark data while retaining a distinct underlying population (Shirali et al., 2023). Earlier dynamic-generation work reframes benchmark instances with multi-agent pipelines (**?**); Guo et al. operationalize trajectory-driven benchmark complexity evolution for agents (Guo et al., 2026); and ActBench evolves behavioral-safety cases from execution trajectories (Yao et al., 2026). These works make the evolving-evaluation setting concrete. Our narrower governed decision is different: whether candidate-caused task admission may change the target for the same candidate's current accept/reject decision.

DeepFact as an explicit alternative temporal contract. DeepFact's Audit-then-Score protocol allows a verifier's evidence to trigger audited label/rationale revisions before that verifier is scored (Huang et al., 2026). This is not a counterexample to the paper's conditional prescription; it shows why the contract must be stated explicitly. If correcting an erroneous benchmark label is part of current semantics, reflexive revision is appropriate. PTP concerns the distinct case in which current task-population improvement is intended to be prospective and comparable to a pre-candidate reference target.

Performative semantics. Performative prediction and performative RL study decision-dependent distributions as part of the objective (Perdomo et al., 2020; Mandal et al., 2023). Accordingly, we do not claim to introduce target endogeneity in general, nor that candidate-dependent objectives are invalid. PTP applies only after a system declares prospective current semantics.

Autotelic and open-ended learning. Autotelic agents can form their own skill sets and therefore make evaluation nontrivial (Colas et al., 2022); open-ended learning formalizes continuing novelty in goals/rewards/options (Sigaud et al., 2023); MAGELLAN studies competence and learning-progress prediction in evolving goal spaces (Gaven et al., 2025). These establish the parent setting. Our contribution isolates one evaluation subproblem inside it rather than attributing our exact question to those works.

C Experimental Protocols and Full Results

C.1 SCER-1: exact same-candidate target-endogeneity replay

Freeze and source data. SCER-1 was specified before its outcome computation. It uses all 384 scientific decision rows from the controlled evidence: eight primary replicates using the 0.5-billion-parameter solver and eight independently seeded confirmation replicates using the 1.5-billion-parameter solver, with 24 decisions per replicate. Each row contains the complete pre-candidate task-wise effect vector d_U, prospective target p, and promotion threshold $\tau = .02$. Integrity checks

require $\sum_x p(x) = 1$, $d_U(x) \in [-1, 1]$, and exact agreement between $p^\top d_U$ and recorded full-target utility.

Candidate-linked expansion. The controlled unlock distribution is

$$r_U(x) \propto \mathbf{1}\{d_U(x) > 0\}.$$

If no positive-effect task exists, the candidate induces no expansion. Otherwise $\widetilde{p}_U = (1-\alpha)p + \alpha r_U$. Primary values are $\alpha = .20$, $\tau = .02$, and induced-task transfer fidelity $\gamma = 1$. Sensitivity uses $\alpha \in \{.05, .10, .20, .30, .40, .50\}$, $\gamma \in \{.25, .50, .75, 1\}$, and $\tau \in \{0, .02, .05\}$. For $\gamma < 1$, induced tasks receive mean effect γ times the observed positive-region effect; this is a controlled transfer/difficulty stress, not a new observed model outcome.

The primary endpoint is the fraction of prospectively non-promotable candidates ($\Delta_p \leq \tau$) that become promotable under $\widetilde{p}_U$. PRIMARY and CONFIRMATORY remain separate. Uncertainty uses 50,000 bootstrap resamples of eight independent trajectory replicates. Because complete recorded effects are used, the primary comparison contains neither q-sampling nor finite evaluation noise. Table 6 gives the complete primary SCER-1 counts and rates.

Table 6: SCER-1 primary results at $\alpha = .20$, $\tau = .02$, $\gamma = 1$.

Metric	0.5B-parameter primary	1.5B-parameter confirmation
All candidates	192	192
Prospectively non-promotable	158	135
False promotions	116/158 (73.4%)	73/135 (54.1%)
Replicate-cluster 95% CI	[62.8,84.0]%	[42.9,63.2]%
Non-promotable with positive region	148	93
Conditional false promotions	116/148 (78.4%)	73/93 (78.5%)
All-candidate disagreement	116/192 (60.4%)	73/192 (38.0%)
Mean score shift	+0.2076	+0.1638
Candidate-independent $r = p$ disagreements	0	0

At $\gamma = 1, \tau = .02$, false-promotion rates across $\alpha = .05, .10, .20, .30, .40, .50$ are $15.2, 35.4, 73.4, 89.9, 93.7, 93.7\%$ for PRIMARY and $12.6, 28.1, 54.1, 68.9, 68.9, 68.9\%$ for CONFIRMATORY. At $\alpha = .20$, increasing γ from .25 to $.5, .75, 1$ changes rates from 19.0 to $37.3, 55.7, 73.4\%$ and from 14.1 to $28.1, 38.5, 54.1\%$. At $\alpha = .20, \gamma = 1$, $\tau = 0, .02, .05$ yields $75.6, 73.4, 62.7\%$ and $59.3, 54.1, 47.6\%$.

A candidate-independent $r = p$ control is an exact semantic no-op and yields zero disagreements. A symmetric negative-region expansion can analogously create false rejections. A same-size random-expansion diagnostic, specified as a secondary analysis after the primary endpoint, yields expected false-promotion probabilities 6.95% and 4.70%, far below candidate-linked rates; it is robustness rather than a confirmatory endpoint.

SCER-1 is direct evidence for the target-definition mechanism on recorded model-output effect vectors. It is not evidence that the original controlled task generator naturally created those candidate-linked additions.

C.2 Controlled evolving LLM task ecology

A Qwen3-4B generator proposes typed Python function puzzles, validated before archive entry. The learner is an in-context demonstration memory of capacity 20. Candidate updates are evaluated with Qwen2.5-Coder-0.5B-Instruct in the primary study and Qwen2.5-Coder-1.5B-Instruct in an independent confirmation. Each phase has eight trajectories and 24 decision rounds per trajectory (192 decisions).

At each decision the archive and target are fixed before candidate evaluation. Target p is uniform over the archive. If S denotes tasks carrying the curriculum-selected skill and u_S is uniform on S, acquisition uses

$$q = 0.30p + 0.70u_S,$$

so $q(x) \geq 0.30p(x)$ and importance weights are at most $3.\overline{3}$. Both judges receive the same $B = 128$ sampled effects and use strict promotion threshold $\widehat{\Delta} > .02$. Exhaustive full-archive evaluations provide retrospective target labels; they are not operational inputs to PTP.

Table 7 reports the controlled fixed-target acquisition results for both solver phases.

Table 7: Controlled fixed-target acquisition study. Intervals are trajectory-cluster bootstrap 95% intervals where available.

Metric	0.5B study	1.5B confirmation
Decisions	192	192
Target-harmful focused	93/120	95/120
Uncorrected harmful promotion	63.44% [47.83,74.73]	57.89% [51.00,64.44]
PTP transport harmful promotion	6.45% [0.96,15.91]	2.11% [0.00,5.32]
Absolute reduction	56.99 percentage points [38.89,71.57]	55.79 percentage points [49.49,61.05]
Target-beneficial balanced	20/72	40/72
PTP transport beneficial retention	95.0% [82.35,100]	97.5% [90.91,100]
Uncorrected mean estimator bias	+0.3251	+0.2399
PTP transport mean estimator bias	-0.00087	+0.00502
Uncorrected promoted target utility	-0.0597	-0.0278
PTP promoted target utility	+0.0600	+0.0749
PTP − uncorrected utility	+0.1197 [0.0750,0.1595]	+0.1026 [0.0894,0.1165]

The fixed-target study is deliberately separate from SCER-1. Here p is unchanged and error arises from using adaptive q as the objective; in SCER-1 acquisition is removed and error arises from changing p itself.

C.3 Matched-budget direct-target analysis

For each already measured effect vector, direct target sampling (DIRECT-p) draws the same $B = 128$ independent and identically distributed evaluations directly from p. Promotion probabilities are computed exactly from the finite effect vector rather than estimated with another model execution. In the 0.5B and 1.5B studies, expected harmful acceptance is 1.09% and 0.85%; beneficial recall is 92.5% and 93.9%; exact-target decision agreement is 93.6% and 95.2%. This is why it is recommend direct p sampling when it is inexpensive.

C.4 Closed-loop controlled simulator

The path-dependence experiment contains 200 paired independent trajectories of 120 rounds. Each judge controls its own learner state, archive, curriculum, accepted updates, and later candidate-caused frontier releases. Acquisition uses $q = .25p + .75u$ with $B = 64$. A fixed panel of 160 tasks is held out from promotion and used only as an external downstream metric. Table 8 summarizes the three closed-loop judges over 200 paired trajectories.

Table 8: Closed-loop controlled trajectories: means across 200 paired trajectories.

Judge	Target agreement	Harmful promotions	Accepted updates	External AUC
Uncorrected q	54.7%	41.73	85.53	0.499830
PTP transport	79.7%	15.55	38.53	0.499485
DIRECT-p	81.3%	13.82	34.61	0.499437

The paired bootstrap target-agreement gain of PTP over uncorrected evaluation is +25.0 percentage points (95% interval 24.1–25.9), and harmful promotions decrease by 26.18 per trajectory (25.32–27.05 fewer). In contrast, PTP-minus-uncorrected external AUC is -3.45×10^{-4} with interval $[-4.35, -2.56] \times 10^{-4}$. A 60-cell grid improves declared-target agreement in all nontrivial $q \neq p$ cells but improves external AUC in none of the 48 nontrivial cells. This is a deliberate negative result: correctness relative to a declared target does not imply better unrelated external utility.

C.5 NATURAL PLR ACQUISITION-TARGET AUDITS

Training uses Procgen easy mode with 200 level identifiers, CentralPLR, 12 parallel environments, 128 rollout steps, and 400,000 total training timesteps per seed. Policies are saved near 100k, 200k, 300k, and 398k. For each adjacent interval, q_t is the mean of actual full replay-probability vectors logged during the interval, while p is uniform over the same 200 levels. Before/after effects are evaluated on every level.

StarPilot discovery uses eight seeds (24 adjacent updates) and finds 2/24 $\Delta_q > 0, \Delta_p < 0$ reversals. Sixteen independent seeds yield 11/48 across 9/16 seeds. Re-evaluation with 50 episodes per level gives 10/48 across 8/16 seeds, two satisfying the original q/p joint certificate. In the cross-environment audit, CoinRun has 0/24 primary-identifiable reversals despite median $\mathrm{TV}(q, p) = .697$; Heist has 9/16 across 7/8 seeds, six jointly certified, with median $\mathrm{TV}(q, p) = .762$. These audits establish natural acquisition-target mismatch and heterogeneous candidate effects; they do not by themselves instantiate candidate-caused task-population expansion.

C.6 CANDIDATE-INDUCED TASK ADMISSION: INITIAL CROSS-ENVIRONMENT COHORT

The candidate-unlock protocol was specified before the corresponding high-precision panels were analyzed. For every checkpoint transition, level IDs 0–99 form a visible target and IDs 100–199 a latent reserve. The capability threshold is the 75th percentile of the *before-policy* visible-level mean returns. Candidate update U unlocks reserve level z iff the after-policy mean exceeds this pre-candidate threshold while the before-policy mean did not. The prospective target is uniform on the visible set; the post-target is uniform on visible plus candidate-unlocked reserve levels. The primary promotion threshold is zero.

On CoinRun, the pre-specified primary configuration yields 0/24 decision disagreements, zero affected seeds, and zero joint-95 events. On Heist it yields 2/24 false promotions across one of eight seeds, with one joint-95 event; mean unlocked-set size is 9.21, mean score shift is +0.293, and same-size candidate-independent random expansion disagrees with prospective evaluation with probability 5.24% on average.

Prespecified sensitivity does not justify replacing the primary result, but it tests brittleness. For Heist at $\tau = 0$ and the primary split, capability quantiles $.60, .75, .90$ yield 3/24, 2/24, and 1/24 disagreements. At capability quantile $.75$ and promotion threshold $\tau = 0$, alternative fixed splits 0–119/reserve 120–199 and even/odd yield 5/24 and 4/24 disagreements, respectively, versus 2/24 for the primary split. With the primary split and $q_{\mathrm{cap}} = .75$, thresholds $0, .02, .05$ yield 2/24, 4/24, and 7/24. CoinRun remains near-null over the grid, with isolated 1/24 cells under some secondary settings. These are sensitivity results, not independent replications.

C.7 STARPILOT CANDIDATE-ADMISSION COHORT

The same pre-specified primary rule (visible 0–99, reserve 100–199, capability quantile .75, threshold 0) is applied to the existing 16-seed StarPilot 50-episode panels, without retraining. Across 48 transitions, 13 become false promotions, zero become false rejections, and 11/16 seeds contain at least one false promotion. The seed-cluster bootstrap interval for the transition fraction is [16.7,37.5]%. Mean unlocked-set size is 10.29 and mean candidate-linked score shift is +0.1584. Same-size candidate-independent expansion (10,000 draws per transition) disagrees 3.19% of the time on average, and the mean candidate-minus-random score shift is +0.1588. No individual StarPilot event passes the stringent joint-95 certificate, so this study supports recurrence at the population/cluster level rather than individual-transition certification.

C.8 INDEPENDENT 64-SEED HEIST REPLICATION

The replication trains 64 independent Heist seeds, 1000–1063, with the same CentralPLR recipe: 400,000 total timesteps, 12 environments, 128 rollout steps, and checkpoints at 99,840; 199,680; 299,520; and 397,824. Before any policy-evaluation outcome from these seeds is observed, the evaluation protocol is specified: 200 levels, 50 paired episodes per level/checkpoint, visible IDs 0–99, reserve IDs 100–199, capability quantile .75, promotion threshold 0, and the same capability-crossing admission rule. All 256 checkpoint panels complete without failure; panel hashes are recorded

before the primary analyzer is run. Training seed is the independent unit. Table 9 summarizes the independently seeded Heist64 replication.

Table 9: Independent Heist64 candidate-admission result under the pre-specified primary rule.

Metric	Result
Independent seeds	64
Adjacent transitions	192
False promotions / false rejections	36 / 0
Transition false-promotion rate	18.75%
Seed-cluster bootstrap 95% interval	[13.54,24.48]%
Seeds with ≥ 1 false promotion	29/64 (45.31%)
Joint-95 false promotions	11 (11 distinct seeds)
Mean unlocked levels	8.01
Mean candidate-linked score shift	+0.2889
Mean same-size random disagreement	2.463%
Mean candidate-minus-random shift	+0.2862
Descriptive observed/random rate ratio	7.61

The 11 individually certified false promotions are shown in Table 10. The comparison is within a controlled task-admission semantics: the policies are naturally trained, but reserve Procgen levels are pre-existing rather than generated by the learner. Accordingly, the result supports recurrence in the tested Heist/CentralPLR regime, not a claim that 45% of open-ended agents in general exhibit target endogeneity.

Table 10: Heist64 joint-95 candidate-induced false promotions. $U_{97.5}$ and $L_{97.5}$ denote the one-sided 97.5% upper and lower confidence bounds, respectively. Every row satisfies $U_{97.5}(\Delta_{\mathrm{pre}}) < 0$ and $L_{97.5}(\Delta_{\mathrm{post}}) > 0$.

Seed	Transition	Δ_{pre}	$U_{97.5}$	Δ_{post}	$L_{97.5}$
1001	300k→398k	-.220	-.121	+.097	+.008
1009	300k→398k	-.256	-.177	+.266	+.179
1012	300k→398k	-.150	-.030	+.484	+.369
1013	200k→300k	-.228	-.116	+.185	+.073
1035	100k→200k	-.120	-.027	+.131	+.035
1045	200k→300k	-.354	-.260	+.298	+.205
1053	200k→300k	-.106	-.013	+.322	+.238
1054	200k→300k	-.156	-.069	+.118	+.020
1057	300k→398k	-.098	-.023	+.202	+.124
1059	300k→398k	-.216	-.088	+.146	+.023
1060	100k→200k	-.440	-.334	+.169	+.067

D Statistical Definitions and Certificates

Seed-cluster bootstrap. Transitions from one training trajectory are dependent. For the natural-policy studies, the cluster bootstrap resamples training seeds with replacement and recomputes the transition statistic over all transitions belonging to sampled seeds. The candidate-admission analyses use 50,000 seed-cluster resamples. These intervals describe uncertainty over the tested training-seed process under the specified environment/protocol; they are not confidence intervals for an unrestricted population of environments.

Candidate-admission joint-95 certificate. Each before/after checkpoint pair is evaluated with the same 50 episode indices per level. For the visible prospective target and candidate-expanded post-target, the analyzer forms paired episode-level score differences. A false promotion is labeled joint-95 only if a 97.5% one-sided Student-t upper confidence bound for the prospective score is below zero and a 97.5% one-sided lower bound for the post-target score is above zero. Bonferroni's inequality gives at most 5% joint error under the stated interval construction. The certificate is deliberately conservative and is applied transition by transition; it is not used to select or tune protocol settings.

Random-expansion control. For each natural candidate-admission transition, the analyzer draws 10,000 candidate-independent subsets from the reserve with the same cardinality as the candidate-linked unlocked set. It reports the probability that such a same-size random expansion changes the prospective decision and its mean score shift. This controls expansion size, not all possible semantics of random task construction. We therefore describe the candidate-linked/random comparison descriptively rather than as a universal causal risk ratio.

E Negative Results and Claim Boundaries

- **Direct target sampling can be preferable.** At equal evaluation budget, DIRECT-p is slightly more target-accurate than importance transport in the controlled task-effect analysis.
- **Target fidelity is not downstream optimality.** The closed-loop simulator improves decisions relative to the declared archive target while slightly worsening a separate fixed external panel.
- **CoinRun is a primary negative boundary.** Candidate-induced admission changes 0/24 primary decisions in the initial cross-environment cohort, and ordinary q/p reversal is likewise absent in its primary audit despite substantial acquisition-target total variation. A few secondary sensitivity cells contain 1/24 events; we do not summarize CoinRun as a universal zero.
- **StarPilot individual certificates are null for candidate admission.** The StarPilot cohort has strong seed-cluster recurrence but zero joint-95 candidate-admission events.
- **Finite-budget PTP transport can err.** The controlled studies still promote 6/93 and 2/95 target-harmful candidates.
- **The Heist64 prevalence claim is environment/protocol-specific.** The 29/64 affected-seed rate is not extrapolated to open-ended agents generally.
- **Procgen expansion is an evaluation overlay.** Reserve levels already exist; naturally trained candidate policies trigger their admission according to a rule specified independently of the evaluated outcomes. Literal semantic task generation by the learner remains a stronger external-validity target.
- **Performative/reflexive objectives can be valid.** PTP is conditional on prospective semantics; DeepFact-style audited current benchmark correction and performative objectives instantiate different legitimate contracts.
- **Historical priority** The priority claim is restricted to formalizing prospective task-population promotion with the universal target-invariance result; it does not claim firstness for dynamic or co-evolving evaluation.

F Artifact and Reproducibility Map

The supplementary archive contains: (i) complete anonymized manuscript source; (ii) SCER-1 code, protocol, primary summaries, sensitivity outputs, and figure; (iii) controlled LLM aggregate and replicate evidence sufficient to reproduce reported tables; (iv) cross-environment candidate-admission protocol/results and pre-specified sensitivity table; (v) StarPilot and 64-seed Heist analyzers, freeze records, summaries, and transition-level CSV files; (vi) compact summaries for the fixed-target acquisition, direct-p, closed-loop, and ordinary PLR audits; (vii) a claim-to-evidence ledger and literature-priority audit.

Large model weights, policy checkpoints, raw 50-episode panel arrays, replay-event streams, full datasets, caches, and virtual environments are omitted from the reviewer supplement. The 64-seed Heist archive retains pre-analysis hashes for all 256 high-precision panels, allowing the compact result files to be tied to the canonical evidence without shipping those bulky arrays. The manuscript uses PTP to avoid collision with the unrelated TRACE acronym (that we used it in our code as our algorithm name) used by an ICLR 2026 agent-benchmark method (Guo et al., 2026).