Idol Learning

Anonymous authors
Paper under double-blind review

Abstract

Rare successful trajectories can reflect either reproducible behavior or favorable environmental randomness. Directly reinforcing them risks consolidating lucky outcomes. We introduce Idol Learning, a probe-before-commit framework that treats a high-return trajectory as a behavioral hypothesis. After an ordinary learner update, Idol temporarily amplifies the trajectory's action likelihood, compares fresh reference and probe rollouts, and retains an evidence-dependent fraction of the same candidate displacement. This separates the strength of a diagnostic intervention from that of permanent learning. Under matched interaction budgets, controlled experiments show substantial gains with REINFORCE and earlier acquisition or improved final skill with PPO. Probe-strength sweeps distinguish useful temporary amplification from harmful unconditional commitment. However, stronger local masking reverses the gains: reward-based validation can accept updates that damage the intended skill. DeepSea shows earlier consolidation than successful-experience replay in a retrospectively selected discovery cohort; exploratory FrozenLake runs show higher mean online success with heterogeneous outcomes. These results support active validation of rare success when interventions expose informative behavioral differences, while identifying candidate construction, reward fidelity, and commitment calibration as limits to broader reliability.

1 Introduction

Sparse-reward reinforcement learning makes successful trajectories difficult to encounter and, once encountered, easy to underuse. A policy may produce a rewarding episode long before ordinary gradient updates make that behavior reliable. Episodic control addresses the need to assimilate valuable experience rapidly (Pritzel et al., 2017); self-imitation and successful-trajectory replay instead increase the influence of the agent's own good episodes on subsequent learning (Oh et al., 2018; Chaudhary et al., 2026). When informative successes are scarce, placing more weight on an exceptional trajectory is a natural response.

The difficulty is that the observed return is not a certificate of reproducible behavior: observed success is not always reproducible behavior. In a stochastic environment, a high-return episode can reflect effective actions, favorable environmental randomness, or both. Return-conditioned learning can mistake a lucky outcome for a reliable action sequence (Paster et al., 2022), while discovery and subsequent robustification need not be the same operation (Ecoffet et al., 2021). Amplifying a trajectory before checking what its actions cause may therefore make an accidental success more likely to be encoded in the policy.

Idol Learning treats a rare success as a behavioral hypothesis to test. Rather than immediately committing an additional amplified update from a rare success, we intervene on the policy and ask whether making that trajectory more likely causes future performance to improve. The resulting sequence is rare $\rightarrow$ probe $\rightarrow$ observe $\rightarrow$ commit: construct a deliberately strong but reversible candidate update, compare fresh behavior under the candidate with a reference policy, and use the observed response to decide how much of that same update to retain. Interventions have previously been used to resolve observational ambiguity in imitation learning (de Haan et al., 2019); here the question is whether a self-generated, high-return candidate produces a repeatable improvement as a whole, without identifying which individual actions were causal.

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.

Our contributions are fourfold. First, we introduce Idol Learning, a probe-before-commit mechanism that uses temporary policy interventions to evaluate additional updates proposed by rare successful experience before committing them. Second, controlled synthetic environments separate episode-level luck from locally corrupted success, exposing different ways in which a rewarded trajectory can mislead learning. Third, we show why an aggressive temporary intervention and an equally aggressive permanent update can have different effects, making validation and commit strength distinct design choices. Fourth, we study this mechanism with REINFORCE and PPO in synthetic tasks and with PPO in DeepSea and FrozenLake, reporting both beneficial regimes and clear limitations under local masking and the current probe construction.

2 Related Work

Episodic control rapidly exploits valuable experience (Blundell et al., 2016; Pritzel et al., 2017), while Go-Explore separates discovery from robustification (Ecoffet et al., 2021). Self-imitation and successful-trajectory replay reinforce an agent's good episodes through auxiliary updates (Oh et al., 2018; Li et al., 2022; Chaudhary et al., 2026). These methods motivate rapid consolidation, but observed return can reflect environmental luck rather than reproducible behavior (Paster et al., 2022). Results on outcome-conditioned imitation and exact success conditioning depend on their respective assumptions (Eysenbach et al., 2022; Russo, 2026); our setting instead concerns finite-sample evaluation of a trajectory-induced policy change.

Conservative and safe policy improvement moderate changes through bounds, interpolation, or baseline fallback (Kakade & Langford, 2002; Pirotta et al., 2013; Thomas et al., 2015; Laroche et al., 2019; Zhu & Matsubara, 2023). Candidate evaluation and the opportunity cost of conservative admission are also established concerns (Dann et al., 2023; Ma & Wu, 2026). Interventional imitation learning uses targeted interactions to resolve observational ambiguity (de Haan et al., 2019). Idol's specific choice is a deliberately amplified, reversible policy update triggered by rare self-generated success, followed by fresh evaluation that determines the retained fraction of that same update. It neither identifies causal actions within a trajectory nor inherits formal safe-improvement guarantees. Appendix C discusses additional trajectory-refinement and cross-domain candidate-validation work.

3 Idol Learning: Probe Before Commit

Construct a provisional intervention. Each round collects a main batch $\mathcal{B}$ and performs an ordinary REINFORCE or PPO update. Let θ denote the parameters after that update. The highest-return trajectory τ_c in the same batch proposes a further change. A standardized return-exceptionality score activates and scales the candidate; a zero return variance or zero likelihood gradient leaves it inactive. For valid state–action pairs, define

$$\ell_\theta(\tau_c) = \frac{1}{|\tau_c|} \sum_{t \in \tau_c} \log \pi_\theta(a_t \mid s_t), \qquad d = \frac{\nabla_\theta \ell_\theta(\tau_c)}{\|\nabla_\theta \ell_\theta(\tau_c)\|_2}, \qquad \theta_{\text{probe}} = \theta + s^* d. \tag{1}$$

The search first locates a displacement bound under parameter-norm and empirical forward-KL caps, with KL measured on main-batch states. If the lift at that bound meets the target, it bisects toward a smaller displacement; otherwise it uses the bound. This endpoint check can miss an attainable target when likelihood is nonmonotonic along the search direction. This intervention is intended to expose the candidate's behavioral consequences, not to prescribe the permanent update magnitude. Candidate construction changes neither optimizer state nor PPO's critic. Appendix A specifies activation, search limits, numerical cases, and all gate formulas.

Observe fresh behavior and commit. Independent new batches $\mathcal{D}_{\text{ref}} \sim \pi_\theta$ and $\mathcal{D}_{\text{probe}} \sim \pi_{\theta_{\text{probe}}}$ estimate the response of an episode-level outcome h:

$$\widehat{\Delta}_h = \frac{1}{n_{\text{probe}}} \sum_{\tau \in \mathcal{D}_{\text{probe}}} h(\tau) - \frac{1}{n_{\text{ref}}} \sum_{\tau \in \mathcal{D}_{\text{ref}}} h(\tau). \tag{2}$$

A gate maps these batches to $\lambda \in [0, 1]$. After restoring the saved reference parameters, Idol retains a fraction of the same candidate displacement:

$$\theta_{\text{new}} = \theta + \lambda(\theta_{\text{probe}} - \theta). \tag{3}$$

An inactive candidate or incomplete validation gives $\lambda = 0$; the preceding ordinary learner update remains. Reference/probe samples are not reused for optimizer updates. Validation evaluates the full probe policy; an intermediate policy obtained by partial commitment is not separately evaluated, and its benefit need not vary monotonically with the retained fraction.

Representative gates. Naive always commits an active candidate after complete validation, ignoring the evidence while paying the same sampling cost. Binary fully commits when the Welch-style standardized return response exceeds an entry threshold. Partial uses the same statistic but ramps the fraction from zero to one between entry and full-commit thresholds. Fisher instead applies a one-sided exact test to positive-return counts. LowerBound (LB) uses the unpooled Wald lower endpoint for the difference in positive-return rates and ramps commitment between two margins. The latter two use $Y = \mathbf{1}\{R > 0\}$, rather than return magnitude. These choices are not interchangeable when returns exceed one.

The gates are representative decision rules, not generally optimal tests or guarantees under repeated adaptive probing. Fresh reward evidence evaluates the policy change as a whole; it does not certify that a rewarded trajectory's individual actions express genuine skill. Probe informativeness and reward fidelity therefore remain prerequisites for useful validation.

4 Controlled Environments and Evaluation

The sequence task has H positions, one-hot position observations, and four actions. A hidden fixed target defines complete skill $S(\tau) = \mathbf{1}\{a_t = a_t^* \ \forall t\}$; uniform-policy skill is 4^{-H}. We vary terminal reward while keeping the task unchanged:

$$\begin{aligned} R_{\text{global}}(\tau) &= S(\tau) + Z, & Z &\sim \text{Bernoulli}(10.24 \cdot 4^{-H}), \\ R_{\text{mask}}(\tau) &= \mathbf{1}\{a_t = a_t^* \text{ or } M_t = 1 \ \forall t\}, & M_t &\overset{\text{iid}}{\sim} \text{Bernoulli}(q). \end{aligned} \tag{4}$$

Global luck produces episode-level false positives and permits return 2. Local masks waive individual correctness requirements: random-policy reward probability becomes $((1 + 3q)/4)^H$, although complete skill remains 4^{-H}. Neither learner nor gate observes luck events, masks, or correctness labels.

Both tasks use a zero-initialized linear policy. Unless a scan states otherwise, each method receives 430,080 environment steps per seed; Idol uses 256 main, 128 reference, and 128 probe episodes per round, including inactive rounds. Baselines spend all interactions on training. These comparisons match interaction budgets, not optimizer steps or computation. The default probe targets a nominal 75 reproductions in 128 probes from a uniform policy, with KL cap 1.6; strength scans use 3.2.

Our primary controlled-task metric is analytic complete-sequence skill, independent of reward corruption. Masked reward is reported separately. T90 denotes first attainment, not sustained success. Reported adjusted intervals belong to each experiment's separate comparison family; later same-seed extensions are not independent replications. Appendix B gives full task and optimization specifications, and Appendix D records cohorts, metrics, and comparison families.

5 Experiments

5.1 Validated consolidation with REINFORCE and PPO

The global-luck comparison uses 100 paired seeds per horizon for REINFORCE (FR-01), with Fisher and LB added in a later same-seed extension. Figure 1(a) shows validated gates exceeding both the base learner and Naive at every horizon. At $H = 6$, Partial raises mean

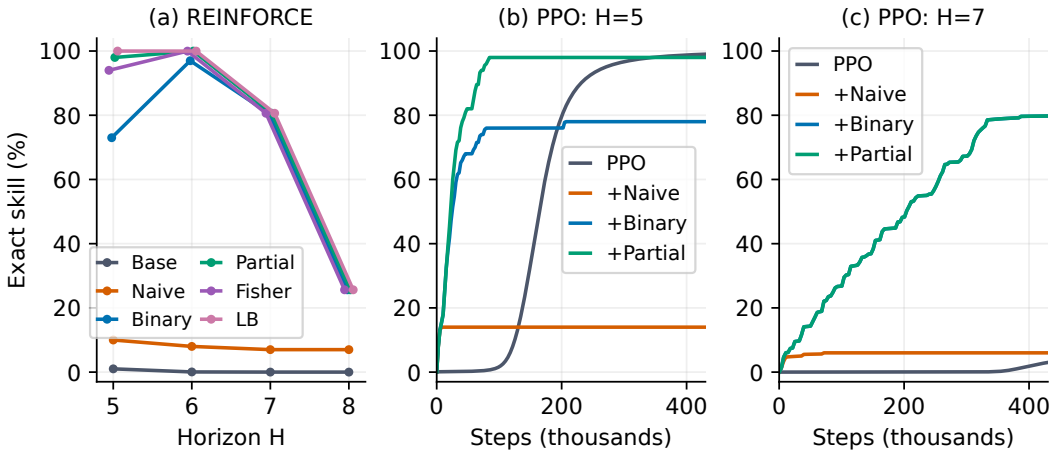


Figure 1: Global-luck consolidation at 430,080 steps per seed. (a) REINFORCE endpoints, 100 paired seeds per horizon; slight horizontal offsets expose overlapping gates. (b,c) PPO learning curves, all 50 paired seeds, including uniform initialization. PPO $H = 7$ and the Fisher/LB endpoints are later same-seed extensions. Curves show means, not confidence intervals; full results are in Appendix D.1.

Table 1: PPO global-luck final exact skill (%), 50 paired seeds per horizon. Δ is Partial minus PPO in percentage points, with its adjusted paired-bootstrap interval. H5/H6 use the original 20-comparison family; the later H7/H8 batch separately uses divisor 20. These do not form one joint confidence family. All methods use 430,080 environment steps.

H	PPO	+Naive	+Binary	+Partial	Δ [adjusted interval]
5	99.02	14.00	78.00	98.00	-1.02 $[-9.03, +1.03]$
6	84.84	8.00	95.98	99.98	$+15.14$ $[+6.38, +27.21]$
7	3.00	6.00	79.76	79.76	$+76.76$ $[+58.68, +92.51]$
8	0.0029	13.90	36.78	36.74	$+36.74$ $[+17.76, +56.80]$

exact skill from approximately 0.05% to 99.98%. All 16 Binary/Partial-versus-control intervals are positive under the replication's 20-comparison correction; the 16 corresponding Fisher/LB intervals are positive in their separate 28-comparison family. Strong unconditional imitation alone does not account for the gain.

Gate rankings depend on difficulty. Partial exceeds Binary at $H = 5$ by 25 percentage points, with adjusted interval $[13, 38]$, but endpoints nearly coincide at $H = 7, 8$. At $H = 8$, only 21/100 Binary and Partial seeds reach T90 and their median stays near the random-policy floor. A gate cannot consolidate an undiscovered correct sequence, and discovering one does not guarantee retention.

With PPO (FR-04; 50 paired seeds), Partial improves final skill at $H = 6$–8 (Table 1). At $H = 5$, its endpoint advantage is unsupported, but its mean curve rises earlier (Figure 1b): full-budget skill AUC is 91.84% versus PPO's 59.49%, a descriptive speed comparison. Binary instead ends below PPO at this horizon. At $H = 7$, both validated gates reach T90 in 39/50 seeds versus 0/50 for PPO; at $H = 8$, only 17/50 do so, and Partial's median is 0.0020%. The gains consequently do not imply uniform success across runs or gates.

5.2 Probe strength depends on the ambiguity of success

Independent 100-seed sweeps vary the nominal reproduction target m over 19 settings at $H = 6$, with a common KL cap of 3.2. Under global luck, Binary and Partial reach broad high-skill plateaus while Naive deteriorates (Figure 2a). An aggressive temporary intervention can thus be useful even when permanently applying it is harmful. These sweeps

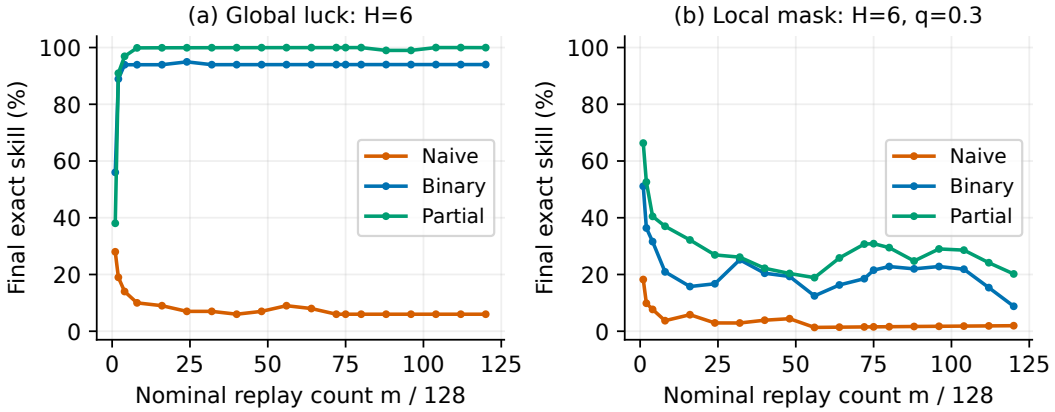


Figure 2: Probe-strength sweeps with REINFORCE: 100 seeds per setting, $H = 6$, 430,080 steps, KL cap 3.2. (a) Global luck, $p_{\text{luck}} = 0.0025$. (b) Local masking, $q = .3$. Both use Naive, Binary, and Partial, without a new Base run. The nominal count is calibrated at a uniform policy. Curves are descriptive means from separate cohorts.

support within-strength comparisons against Naive, not a cross-strength optimum or equivalence across a plateau.

Under local masking, the response reverses overall, although it is not monotone (Figure 2b). From $m = 1$ to 120, Partial's mean exact skill falls from 66.32% to 20.19%. The fraction of accepted active candidate policies with lower skill than the reference rises from 27.7% to 79.4%; this is distinct from the harm caused by a fractional commit. An imitated trajectory with k incorrect positions can still pass a fresh mask with probability q^k. Amplification can therefore reinforce behavior supported by reward but inconsistent with complete skill. Additional luck-rate and validation-batch sweeps show finite robustness: more validation can reduce harmful commits without consistently improving endpoints, and costs additional interactions (Appendices D.1.3–D.1.4).

5.3 Local masking separates reward from skill

FR-02 varies q at $H = 6$ with REINFORCE and 100 paired seeds. At $q = .3$, LB reaches 48.50% exact skill versus Base's 7.84%. At $q = .8$, however, the validated gates attain 57.59–62.28% masked reward while retaining only 0.10–0.23% exact skill, below Base's 2.20% (Figure 3a,b). Measuring reward alone would obscure this reversal.

PPO provides a stronger comparison (FR-05; 50 paired seeds per q). At $q = .1$, LB and PPO both reach T90 in every seed, while LB's skill AUC is 89.19% versus 60.40%. At $q = .3$ and .5, every tested Idol variant has lower mean final skill than PPO; even LB reaches only 61.35% and 53.07%, versus PPO's 98.68% and 98.24%. The corrected LB–PPO endpoint intervals are negative at both settings.

The all-seed outcomes are sharply split: at $q = .3$, 31/50 LB seeds finish above 90% skill and 19/50 below 1% (Figure 3c). Faster consolidation in a retrospectively chosen no-harm subgroup does not establish a population benefit. Selected failure-path analyses and oracle vetoes suggest that particular harmful commits can be decisive, but supply neither a deployable gate nor a randomized treatment comparison (Appendix D.2.3).

5.4 DeepSea and FrozenLake

In deterministic DeepSea with $N = 12$ (FR-06 v2), four methods each receive 51,200 episodes, or 614,400 environment steps, on 50 seeds. Before the first successful main rollout, neither SIL nor Idol adds an update beyond ordinary PPO, so their policies coincide with PPO at matched main-update rounds and seeds. Exactly the same 17 seeds discover a successful path in all methods. Figure 4(a) conditions retrospectively on this cohort: both

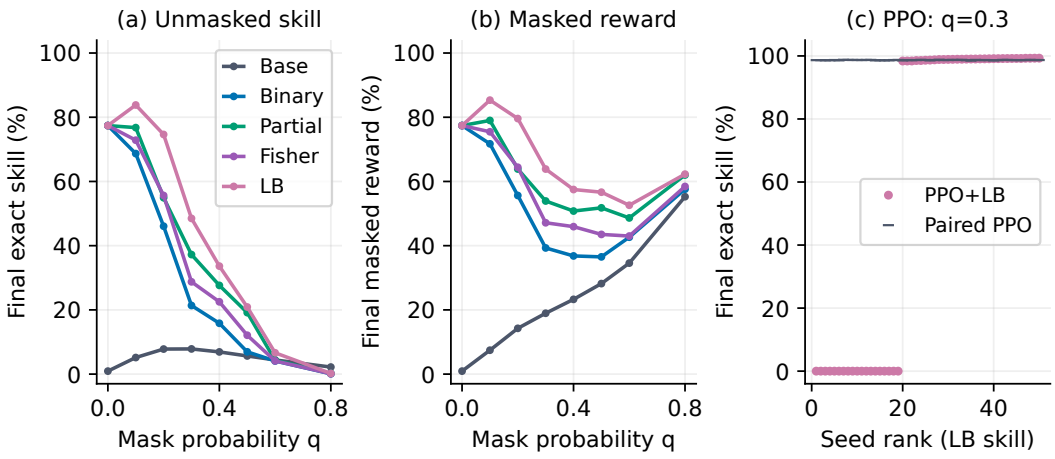


Figure 3: Local-mask limitations, $H = 6$, 430,080 steps, KL cap 1.6. (a,b) REINFORCE means over 100 paired seeds per q: complete skill and masked reward are different metrics. (c) All 50 PPO–LB endpoints at $q = .3$, ranked by LB outcome with paired PPO controls; ranking excludes no seeds. Full gate comparisons and post hoc failure diagnostics are in Appendix D.2.

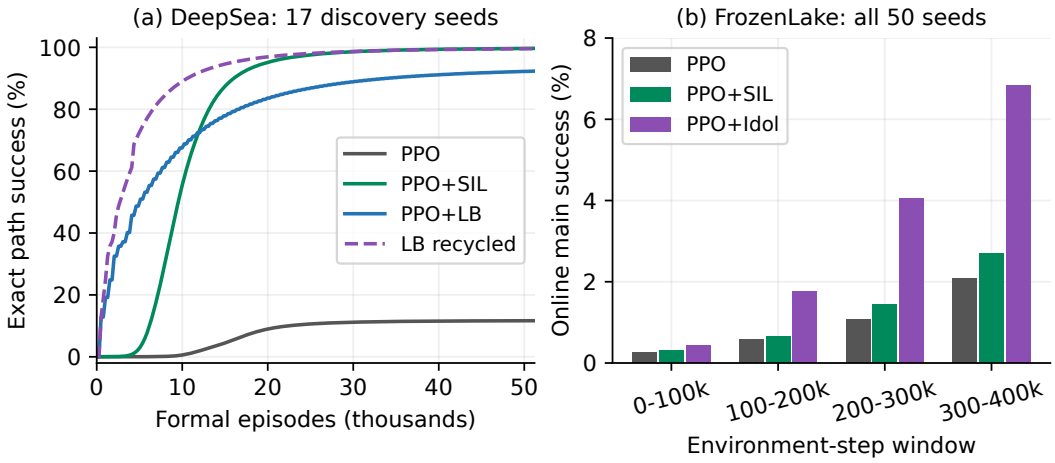


Figure 4: Benchmark outcomes with distinct evaluation protocols. (a) DeepSea analytic path success, conditioned on the same 17/50 main-discovery seeds; equal episode budgets and left-held checkpoints every 256 episodes. (b) FrozenLake online main-rollout success, all 50 seeds: complete batches are averaged within each step window per seed, then equally across seeds. PPO+Idol uses E4, selected through sequential 100k-step experiments on these same seeds: skip inactive validation, lower the LB admission margin from 0.02 to 0, then lower the full-commit margin from 0.05 to 0.001 (Appendix E.2). Neither panel is an independent confirmatory comparison.

LB variants consolidate earlier than PPO+SIL on the equal-episode axis. At the endpoint, PPO retains at least 90% success in 2/17 seeds, ordinary LB in 16/17, and SIL and recycled LB in 17/17. SIL's mean final success, 99.65%, is slightly above recycled LB's 99.57% and above ordinary LB's 92.31%. This is evidence of earlier post-discovery consolidation, not improved discovery or a superior endpoint. The other 33 seeds are excluded only from this explicitly conditional view.

The recycled variant reallocates inactive validation to ordinary PPO updates while preserving the episode budget. Its random stream also changes, so the endpoint coverage difference

from ordinary LB cannot be assigned to the extra updates alone. Appendix E.1 supplies the full protocol and SIL configuration.

FrozenLake uses the slippery 8×8 map (FR-07). Initial 100k-step diagnostics motivated E4, which skips inactive validation and lowers both admission and full-commit thresholds. Fresh 400k-step runs from initialization then compare PPO, SIL, and E4 on the same 50 seeds. Final-window online success averages 2.09%, 2.70%, and 6.82%, respectively (Figure 4b). E4 beats SIL on 30 seeds and loses on 20; its 3.78% median is well below its mean.

E4 was chosen after viewing those seeds' 100k outcomes, so this comparison is exploratory. These are training-rollout rates, not independent evaluations of frozen policies, and they do not separate lucky slip outcomes from repeatable behavior. All means continue rising through 400k steps; neither a plateau nor a ranking beyond this budget is established. Appendix E.2 preserves the tuning sequence, complete endpoint statistics, and window definitions.

6 Discussion and Limitations

Idol separates three design choices: constructing an informative intervention, allocating interactions to validation, and deciding how much to retain. A strong probe helps only when it exposes a behavioral difference relevant to the intended capability. Local masking violates this alignment: reproducible reward can support incorrect actions. Idol's direct parameter commit also lies outside PPO's clipped objective, so candidate caps do not ensure that an accepted update preserves accumulated skill.

Candidate construction becomes less direct in stochastic environments. Raising selected action likelihoods cannot control FrozenLake's slippery transitions or supply behavior at states absent from the trajectory. Weak probe responses, finite validation batches, and conservative retention may jointly limit useful learning. FrozenLake's admission and retention changes are consistent with this interpretation, but their same-seed tuning does not isolate its cause. Probes that accommodate deviations, combine trajectories, or amplify reusable behavior fragments require direct evaluation.

Validation also displaces training opportunities. Skipping inactive probes can recover budget, while larger validation batches can improve screening without improving final skill. No tested gate establishes an optimal allocation or commitment rule. Broader evaluation should freeze settings before testing independent seeds, measure frozen policies as well as online outcomes, and report failure rates alongside means. The current small policy classes and limited environments leave generalization to complex representations and longer behaviors unresolved.

7 Conclusion

Idol Learning treats rare success as a hypothesis to test through a strong, reversible policy intervention. Fresh behavioral evidence determines the retained fraction, separating diagnostic amplification from permanent learning. Across the tested settings, this can improve consolidation and accelerate acquisition, but benefits depend on informative probes and rewards aligned with the intended skill. The failures under local masking and the conditional or exploratory benchmark evidence delimit that claim. Better candidate construction, validation allocation, and commitment calibration remain central directions for making these gains reliable.

AI use statement

The initial research idea and problem formulation originated from the authors. Generative AI tools were subsequently used to assist with iterative research ideation and execution, including discussion and refinement of methodology, experimental design, implementation, and experimental analysis. Generative AI tools were also used for literature retrieval and

discovery, and to assist with drafting, restructuring, editing, and polishing parts of the manuscript.

All scientific decisions, experimental protocols, analyses, and final claims were reviewed and determined by the authors. Literature and citations identified or discussed with the assistance of generative AI were verified against the original sources. All AI-assisted code and manuscript content were reviewed by the authors. The authors take full responsibility for the final methodology, implementation, experimental results, claims, and contents of this work.

References

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In International Conference on Learning Representations, 2026. URL https://arxiv.org/abs/2507.19457.

Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z. Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. Model-free episodic control, 2016. URL https://arxiv.org/abs/1606.04460.

Gaurav Chaudhary, Laxmidhar Behera, and Washim Uddin Mondal. Match or replay: Self imitating proximal policy optimization, 2026. URL https://arxiv.org/abs/2603.27515.

Tianhong Dai, Hengyan Liu, and Anil Anthony Bharath. Episodic self-imitation learning with hindsight. Electronics, 9(10):1742, 2020. doi: 10.3390/electronics9101742. URL https://www.mdpi.com/2079-9292/9/10/1742.

Chris Dann, Mohammad Ghavamzadeh, and Teodor V. Marinov. Multiple-policy high-confidence policy evaluation. In Proceedings of the 26th International Conference on Artificial Intelligence and Statistics, volume 206 of Proceedings of Machine Learning Research, pp. 9470–9487, 2023. URL https://proceedings.mlr.press/v206/dann23a.html.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In Advances in Neural Information Processing Systems, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/947018640bf36a2bb609d3557a285329-Abstract.html.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. First return, then explore. Nature, 590:580–586, 2021. doi: 10.1038/s41586-020-03157-9. URL https://www.nature.com/articles/s41586-020-03157-9.

Benjamin Eysenbach, Soumith Udatha, Ruslan Salakhutdinov, and Sergey Levine. Imitating past successes can be very suboptimal. In Advances in Neural Information Processing Systems, volume 35, 2022. doi: 10.52202/068431-0438. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/284afdc2309f9667d2d4fb9290235b0c-Abstract-Conference.html.

Yijie Guo, Jongwook Choi, Marcin Moczulski, Shengyu Feng, Samy Bengio, Mohammad Norouzi, and Honglak Lee. Memory based trajectory-conditioned policies for learning from sparse rewards. In Advances in Neural Information Processing Systems, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/2df45244f09369e16ea3f9117ca45157-Abstract.html.

Yinsen Jia and Boyuan Chen. Temporal self-imitation learning, 2026. URL https://arxiv.org/abs/2606.19752.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In Proceedings of the Nineteenth International Conference on Machine Learning, pp. 267–274, 2002. URL https://hunch.net/~jl/projects/RL/RL.html.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pp. 3652–3661, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Yunfei Li, Tian Gao, Jiaqi Yang, Huazhe Xu, and Yi Wu. Phasic self-imitative reduction for sparse-reward goal-conditioned reinforcement learning. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 12765–12781, 2022. URL https://proceedings.mlr.press/v162/li22g.html.

Jing Liang, Hongyao Tang, Yi Ma, Yancheng He, Weixun Wang, Xiaoyang Li, Ju Huang, Wenbo Su, Jinyi Liu, Yan Zheng, Jianye Hao, and Bo Zheng. The mirage of optimizing training policies: Monotonic inference policies as the real objective for LLM reinforcement learning, 2026. URL https://arxiv.org/abs/2606.29526.

Qinzhen Ma and Ruihai Wu. When validation stops learning: Auditing update admission for continual embodied agents, 2026. URL https://arxiv.org/abs/2609.10873.

Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 3878–3887, 2018. URL https://proceedings.mlr.press/v80/oh18b.html.

Keiran Paster, Sheila McIlraith, and Jimmy Ba. You can't count on luck: Why decision transformers and RvS fail in stochastic environments. In Advances in Neural Information Processing Systems, volume 35, 2022. doi: 10.52202/068431-2824. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fe90657b12193c7b52a3418bdc351807-Abstract-Conference.html.

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In Proceedings of the 30th International Conference on Machine Learning, volume 28 of Proceedings of Machine Learning Research, pp. 307–315, 2013. URL https://proceedings.mlr.press/v28/pirotta13.html.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 2827–2836, 2017. URL https://proceedings.mlr.press/v70/pritzel17a.html.

Yulei Qin, Xiaoyu Tan, Zhengbao He, Gang Li, Haojia Lin, Zongyi Li, Zihan Xu, Yuchen Shi, Siqi Cai, Renting Rui, Shaofei Cai, Yuzheng Cai, Xuan Zhang, Sheng Ye, Ke Li, and Xing Sun. Learn the ropes, then trust the wins: Self-imitation with progressive exploration for agentic reinforcement learning. In International Conference on Learning Representations, 2026. URL https://arxiv.org/abs/2509.22601.

Daniel Russo. Success conditioning as policy improvement: The optimization problem solved by imitating success, 2026. URL https://arxiv.org/abs/2601.18175.

Yifei Shen, Bo Li, and Xinjie Zhang. Skillopt-lite: Better and faster agent self-evolution via one line of vibe, 2026. URL https://arxiv.org/abs/2607.03451.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pp. 2380–2388, 2015. URL https://proceedings.mlr.press/v37/thomas15.html.

Yang Tian, Rui Wang, Xumeng Wen, Junjie Li, Shizhao Sun, Lei Song, Jiang Bian, and Bo Zhao. PBSD: Privileged bayesian self-distillation for long-horizon credit assignment, 2026. URL https://arxiv.org/abs/2606.09348.

Tianshi Xu, Yuteng Chen, and Meng Li. CLEANER: Self-purified trajectories boost agentic reinforcement learning, 2026. URL https://arxiv.org/abs/2601.15141.

Runzhe Zhan, Yafu Li, Zhi Wang, Xiaoye Qu, Dongrui Liu, Jing Shao, Derek Wong, and Yu Cheng. ExGRPO: Learning to reason from experience. In International Conference on Learning Representations, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/15a296dc96baa12aaef4facdd391161f-Abstract-Conference.html.

Lingwei Zhu and Takamitsu Matsubara. Cautious policy programming: exploiting KL regularization for monotonic policy improvement in reinforcement learning. Machine Learning, 112:4527–4562, 2023. doi: 10.1007/s10994-023-06368-z. URL https://link.springer.com/article/10.1007/s10994-023-06368-z.

A Complete candidate construction and gate definitions

A.1 Provisional intervention

Each training round begins with a fresh main batch $\mathcal{B} = \{\tau_i\}_{i=1}^{B}$ and an ordinary REINFORCE or PPO update. We write θ for the policy parameters after this update. The candidate is constructed from the same main batch, so the subsequent comparison is between this updated reference policy and a further, temporary intervention. In the experiments, the frontier contains the highest-return trajectory τ_c (top-k selection with $k = 1$). Let $\bar{R}_{\mathcal{B}}$ and $s_{\mathcal{B}}$ be the main-batch mean and population standard deviation of episodic returns. The frontier strength and its activation factor are

$$F = \max\left\{0, \frac{R(\tau_c) - \bar{R}_{\mathcal{B}}}{s_{\mathcal{B}} + \varepsilon}\right\}, \qquad a_{\mathrm{F}} = \left[\frac{F - F_{\min}}{\sqrt{B-1} - F_{\min}}\right]_0^1, \tag{A1}$$

where $[x]_0^1 = \min(1, \max(0, x))$. We set $F = 0$ when the batch return variance is zero. A configured absolute return floor can additionally suppress a frontier whose return does not exceed that floor. If the frontier is active, a configured minimum may raise a_{F}; otherwise $a_{\mathrm{F}} = 0$ and no candidate update is made. The implementation treats the limiting case of one positive return in an otherwise binary zero-return batch as full activation despite numerical smoothing in F.

For the valid state–action pairs (s_t, a_t) in τ_c, define the mean action log likelihood and its normalized ascent direction at the post-update policy:

$$\ell_\vartheta(\tau_c) = \frac{1}{|\tau_c|} \sum_{t \in \tau_c} \log \pi_\vartheta(a_t \mid s_t), \qquad d = \frac{\nabla_\theta \ell_\theta(\tau_c)}{\|\nabla_\theta \ell_\theta(\tau_c)\|_2}. \tag{A2}$$

If the gradient is zero, the candidate is inactive. Otherwise, the candidate search varies only the policy parameters along $\theta + sd$, with $s \geq 0$. It targets a mean log-likelihood lift

$$L_{\mathrm{target}} = a_{\mathrm{F}} \min\{L_{\mathrm{configured}}, -\ell_\theta(\tau_c)\}, \tag{A3}$$

where the second term is an upper bound on the lift because an action log probability cannot exceed zero. It need not be attainable, particularly when the trajectory assigns different actions to the same state. The empirical forward KL is the average of $D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s_t) \| \pi_{\theta+sd}(\cdot \mid s_t))$ over all valid states in $\mathcal{B}$.

The implemented search first evaluates the parameter-norm cap. If its KL exceeds the configured cap, bisection locates a smaller allowed displacement bound. If the likelihood lift at that bound reaches the target, a second bisection searches toward a smaller displacement reaching it; otherwise the bound itself is used. This endpoint check is not a general feasibility test: likelihood can increase and then decrease along the direction, so a target missed at the bound may still be attainable inside the interval. The procedure therefore does not guarantee the smallest feasible displacement for arbitrary trajectories. The resulting candidate is

$$\theta_{\mathrm{probe}} = \theta + s^* d = U_{\alpha_{\mathrm{probe}}}(\theta, \tau_c), \qquad \alpha_{\mathrm{probe}} := s^*, \tag{A4}$$

where $U_{\alpha_{\mathrm{probe}}}$ denotes this candidate construction, rather than an additional optimizer step. The intervention is deliberately strong so that the behavior suggested by τ_c can have a measurable effect on fresh rollouts. Its purpose is diagnosis; its full magnitude is not presumed suitable for permanent learning. Candidate construction does not update the optimizer state or, for PPO, the critic.

A.2 Reference and post-intervention rollouts

We save the post-update policy parameters before the intervention and collect independent fresh batches from the reference and probe policies:

$$\mathcal{D}_{\mathrm{ref}} \sim \pi_\theta, \qquad \mathcal{D}_{\mathrm{probe}} \sim \pi_{\theta_{\mathrm{probe}}}. \tag{A5}$$

For an episode-level outcome h, the observed response is

$$\widehat{\Delta}_h = \frac{1}{n_{\mathrm{probe}}} \sum_{\tau \in \mathcal{D}_{\mathrm{probe}}} h(\tau) - \frac{1}{n_{\mathrm{ref}}} \sum_{\tau \in \mathcal{D}_{\mathrm{ref}}} h(\tau). \tag{A6}$$

The choices of h below are not interchangeable. Binary and Partial use episodic return R. Fisher and the lower-bound gate defined below use $Y(\tau) = \mathbf{1}\{R(\tau) > 0\}$, which marks a positive observed return. This reward-based signal is not an oracle label of underlying skill. Fresh rollouts estimate how the candidate changes measured performance under the available reward; they cannot by themselves certify that every rewarded action was correct.

The commit rule maps the two batches to $\lambda = g(\mathcal{D}_{\text{ref}}, \mathcal{D}_{\text{probe}}) \in [0, 1]$. Once the probe rollouts are complete, we first restore the saved reference parameters and then apply a fraction of the same candidate displacement:

$$\theta_{\text{new}} = \theta + \lambda(\theta_{\text{probe}} - \theta). \tag{A7}$$

An inactive candidate or incomplete validation receives $\lambda = 0$. The ordinary learner update that preceded candidate construction remains in place. Probe and commit strength are therefore distinct: a large, reversible intervention may provide a clear diagnostic while only a smaller fraction is retained. The validation batches evaluate the reference and full probe policies, not the intermediate policy produced by a partial commit; they do not establish monotonic improvement as the retained fraction changes.

A.3 Gate constructions

There is currently no established optimal choice of commit gate for this setting. We therefore consider several representative constructions based on thresholding, fractional commitment, and standard statistical comparisons, without claiming that any one is generally optimal. They share the same candidate construction, fresh validation, and fractional commitment. Apart from the diagnostic-free control, they use only the new validation rewards to determine λ. Thresholds are method parameters; their values are specified by the individual experimental configurations.

Naive (diagnostic-free control): For an active candidate with complete validation batches, Naive sets $\lambda = 1$ regardless of the measured response. It still collects reference and probe rollouts under the matched validation protocol, but does not use them to choose the update. It is therefore a control for committing the strong candidate without a gate.

Binary: Let $\bar{R}_j$, s_j^2, and n_j denote the sample mean, unbiased sample variance, and batch size for $j \in \{\text{ref}, \text{probe}\}$. Binary uses the Welch-style standardized return response

$$\widehat{\Delta}_R = \bar{R}_{\text{probe}} - \bar{R}_{\text{ref}}, \qquad z_R = \frac{\widehat{\Delta}_R}{\sqrt{\max\{s_{\text{probe}}^2/n_{\text{probe}} + s_{\text{ref}}^2/n_{\text{ref}}, \varepsilon^2\}}}. \tag{A8}$$

It makes a full commit when $z_R > z_{\min}$, and otherwise sets $\lambda = 0$. The score is a finite-sample decision statistic, not a claimed confidence guarantee.

Partial: Partial uses the same z_R and entry threshold as Binary, but scales the retained fraction between that threshold and a full-commit threshold $z_{\text{full}} > z_{\min}$:

$$\lambda_{\text{Partial}} = \begin{cases} 0, & z_R \le z_{\min}, \\ \min\left\{1, \left(\dfrac{z_R - z_{\min}}{z_{\text{full}} - z_{\min}}\right)^{\gamma_R}\right\}, & z_R > z_{\min}, \end{cases} \tag{A9}$$

with $\gamma_R > 0$. Thus a response above the entry threshold can justify a partial commit even before it reaches the full-commit threshold.

Fisher: Fisher binarizes the observed reward as $Y = \mathbf{1}\{R > 0\}$ and compares the two batches with a one-sided Fisher exact test for a higher positive-return rate under the probe. Let x_j be the number of positive returns in batch j, $M = x_{\text{ref}} + x_{\text{probe}}$, and $N = n_{\text{ref}} + n_{\text{probe}}$. Conditional on these margins, its upper-tail probability is

$$p_F = \sum_{k=x_{\text{probe}}}^{\min(n_{\text{probe}}, M)} \frac{\binom{M}{k}\binom{N-M}{n_{\text{probe}}-k}}{\binom{N}{n_{\text{probe}}}}. \tag{A10}$$

The gate sets $\lambda = 1$ if $p_F < \alpha_F$, and $\lambda = 0$ otherwise. This test evaluates the observed reward signal and does not identify genuine skill separately from reward corruption.

LowerBound (LB): With $\hat{p}_j = n_j^{-1} \sum_{\tau \in \mathcal{D}_j} Y(\tau)$, LB uses the unpooled Wald lower endpoint for the difference in positive-return rates:

$$L_Y = (\hat{p}_{\text{probe}} - \hat{p}_{\text{ref}}) - z_\alpha \sqrt{\frac{\hat{p}_{\text{probe}}(1-\hat{p}_{\text{probe}})}{n_{\text{probe}}} + \frac{\hat{p}_{\text{ref}}(1-\hat{p}_{\text{ref}})}{n_{\text{ref}}}}. \tag{A11}$$

For an entry margin δ, a full-commit margin $\delta_{\text{full}} > \delta$, a maximum fraction $\lambda_{\max} \leq 1$, and exponent $\gamma_L > 0$, its rule is

$$\lambda_{\text{LB}} = \begin{cases} 0, & L_Y \leq \delta, \\ \lambda_{\max} \left(\dfrac{L_Y - \delta}{\delta_{\text{full}} - \delta} \right)^{\gamma_L}, & \delta < L_Y < \delta_{\text{full}}, \\ \lambda_{\max}, & L_Y \geq \delta_{\text{full}}. \end{cases} \tag{A12}$$

The experiments use a square ramp when $\gamma_L = 2$. This plug-in bound is a gate construction, not a formal guarantee of safe improvement under repeated, adaptively selected probes. Its decision can still be misleading when positive reward fails to distinguish a useful trajectory from a locally corrupted one.

B Environment definitions and optimization protocol

We use two reward mechanisms on the same sequence-prediction task to distinguish rare discovery from the reliability of a rewarded trajectory. Both mechanisms preserve the action space, observations, target sequence, and fixed episode length. They change only how terminal reward is assigned, allowing us to study episode-level false positives and success that conceals individual errors separately.

B.1 Sequence prediction task

An episode has horizon H. At position $t \in \{1, \ldots, H\}$, the agent observes a one-hot encoding of its current position and chooses $a_t \in \{0, 1, 2, 3\}$. A fixed target sequence $a_1^*, \ldots, a_H^*$ is hidden from the policy. The episode terminates after the Hth action, and complete sequence accuracy is

$$S(\tau) = \mathbf{1}\{a_t = a_t^* \text{ for all } t = 1, \ldots, H\}. \tag{A13}$$

For a policy that samples each action uniformly, the probability of complete accuracy is

$$P_{\text{skill,random}} = \Pr(S = 1) = 4^{-H}. \tag{A14}$$

Thus increasing H directly controls the rarity of discovering a fully correct sequence. We measure complete skill by the probability of $S = 1$ under the policy, independently of the reward mechanism. In the implemented task, the target repeats $(0, 1, 2, 3)$; the global-luck rarity experiments use $H \in \{5, 6, 7, 8\}$, and the local-mask sweep uses $H = 6$.

B.2 Global luck: episode-level false success

In the first mechanism, a Bernoulli variable $Z \sim \text{Bernoulli}(p_{\text{luck}})$ is sampled independently of the actions at the end of each episode. With unit skill and luck rewards, the terminal return is

$$R_{\text{global}}(\tau) = S(\tau) + Z, \qquad p_{\text{luck}}(H) = 10.24 \cdot 4^{-H}. \tag{A15}$$

This is the standard luck setting outside the dedicated luck-rate sweep. For example, $p_{\text{luck}}(5) = 10.24/4^5 = 0.01$. An incorrect sequence can therefore receive a positive return solely because $Z = 1$. Such a false-positive trajectory may contain no useful information about the target sequence. If both the sequence and the luck draw succeed, the additive rule gives return 2; the false-positive case specifically has $S = 0$ and $Z = 1$. For a uniform random policy, its probability is $(1 - 4^{-H}) p_{\text{luck}}(H)$.

This setting directly tests Idol's policy-level diagnosis. Imitating a false-positive trajectory makes its actions more likely, but does not reliably increase the chance of reproducing the target sequence. A fresh reference-versus-probe comparison can instead ask whether the proposed change improves subsequent outcomes. The hidden luck event is used for analysis, never as an input to the policy or the commit gate.

B.3 Local mask: success with mixed-quality evidence

The second mechanism disables global luck. Independently at each position, a hidden mask $M_t \sim \text{Bernoulli}(q)$ waives the correctness requirement for that position. Actions and observations are unchanged; masks affect only the terminal reward:

$$R_{\text{mask}}(\tau) = \mathbf{1}\{a_t = a_t^* \text{ or } M_t = 1 \text{ for all } t = 1, \ldots, H\}. \tag{A16}$$

For a uniform random policy, one position satisfies this condition with probability $q + (1 - q)/4 = (1 + 3q)/4$. Independence across positions gives

$$P_{\text{reward,random}}(q) = \left(\frac{1+3q}{4}\right)^H, \qquad P_{\text{skill,random}} = 4^{-H}. \tag{A17}$$

Increasing q makes rewarded episodes more common without increasing the probability that a random trajectory is fully correct. Indeed, the fraction of rewarded random trajectories that are fully correct is $(1 + 3q)^{-H}$. A rewarded trajectory may therefore combine correct actions at unmasked positions with incorrect actions that were masked. This is a different ambiguity from episode-level luck: a single successful trajectory can contain both informative and misleading actions.

The local-mask experiments use $H = 6$ and $q \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8\}$. We distinguish the probability of receiving reward under the mask rule from the unmasked, complete-sequence skill probability. Neither the policy nor Idol's gate observes the masks or per-position correctness labels; these are retained only for diagnostic evaluation.

B.4 Policy and optimization protocol

Across both environments, the policy is a bias-free linear map from the one-hot position observation to four action logits. All policy weights start at zero, giving a uniform action distribution at every position. The REINFORCE learner used in FR-01–FR-03 updates this policy with Adam at learning rate 0.003, batch-centered episodic returns, entropy coefficient 0.01, discount $\gamma = 1$, and gradient-norm clipping at 1. The separate PPO learner used in FR-04 and FR-05 has a linear critic and an Adam optimizer with actor and critic learning rates 0.003 and 0.01. It uses $\gamma = 1$, GAE parameter 0.95, clipping ratio 0.2, up to four epochs with transition minibatches of 1,024, entropy coefficient 0.01, value coefficient 0.5, gradient-norm cap 1, and an epoch-level approximate-KL stop at 0.02.

Unless an ablation specifies otherwise, each Idol round selects the highest-return trajectory (top-1) from 256 main episodes and collects 128 fresh reference and 128 fresh probe episodes. From a uniform policy, the strong candidate is calibrated so that an ideal correct trajectory has a nominal expected 75 appearances among the 128 probe episodes. The target log-likelihood lifts for $H = 5, 6, 7, 8$ are $1.2794, 1.2972, 1.3099, 1.3195$; the forward-KL cap is 1.6, the parameter-norm cap is 16, and the minimum active frontier factor is 1. Binary and Partial enter above $z_R = 1.64$, with Partial reaching full commitment at $z_R = 3$. Section A.3 defines the gate rules.

The default budget is 430,080 environment steps per method and seed. Ordinary REINFORCE or PPO baselines spend all their interactions on training batches; Idol variants spend part of the same budget on reference and probe rollouts, which are not used for optimizer updates. For budget alignment, an Idol round still collects both validation batches when its main batch has no active candidate. That round makes no effective probe-policy intervention and sets $\lambda = 0$, but consumes the scheduled interactions. This fixed cost keeps the mechanism comparison simple. Skipping validation for inactive candidates could free interactions for additional main updates, so the observed same-budget improvement may understate a more efficient implementation; we do not measure that counterfactual. The comparisons match environment interactions, not optimizer updates. For $H = 6$, this yields 2,240 optimizer steps for PPO alone and 1,120 for PPO with Idol; FR-04's other horizons have different step counts because episode length varies.

C Extended related work

C.1 Rapid use of discovered experience

Model-free and neural episodic control use memory to act on valuable experience before a parametric learner has absorbed it Blundell et al. (2016); Pritzel et al. (2017). Go-Explore likewise separates discovering promising behavior from making it robust, including under stochasticity Ecoffet et al. (2021). In parametric learning, self-imitation learning (SIL) reuses past good decisions through an auxiliary off-policy objective Oh et al. (2018); episodic self-imitation with hindsight selects informative parts of relabeled episodes Dai et al. (2020); PAIR alternates online reinforcement learning with supervised learning from successful trajectories Li et al. (2022); and SIPP-Replay reuses self-encountered successes within a PPO-based method Chaudhary et al. (2026). Memory-based trajectory conditioning instead preserves diverse paths for subsequent exploration, partly to avoid overexploiting a narrow set of high-return trajectories Guo et al. (2020). More recent work emphasizes temporally efficient successful behaviors in manipulation and curriculum-guided self-imitation in language-model agents Jia & Chen (2026); Qin et al. (2026). These methods establish the value of rapid post-discovery exploitation. Idol's different question is whether a particular rare success should induce a permanent parametric update: it temporarily deploys a strong candidate, gathers new behavioral evidence, and only then decides what to retain.

C.2 When apparent success is unreliable

Return and behavioral quality can diverge. In stochastic offline reinforcement learning, Paster et al. show that return-conditioned policies may treat lucky returns as reliably attainable Paster et al. (2022). Outcome-conditioned imitation has also been shown to be suboptimal for some formulations Eysenbach et al. (2022), while exact success conditioning has a policy-improvement interpretation under a different set of assumptions Russo (2026). These results should not be conflated: our setting concerns a finite-sample, trajectory-induced candidate whose effect is measured after intervention. Within language-model reinforcement learning, ExGRPO selects past reasoning experiences using correctness and entropy; CLEANER purifies trajectories so that erroneous actions are not reinforced merely because a run eventually succeeds; and PBSD assigns turn-level credit even when a successful trajectory contains misleading actions Zhan et al. (2026); Xu et al. (2026); Tian et al. (2026). Idol does not infer which step of a rewarded trajectory was correct or causally necessary. It tests the downstream effect of amplifying the candidate update as a whole, then scales or rejects that update using fresh reference and probe rollouts. This also leaves a limitation: if the probe reward cannot distinguish genuine skill from a locally corrupted success, the gate can accept a harmful candidate.

C.3 Candidate evaluation and cautious updates

Conservative policy iteration and subsequent safe policy-iteration methods moderate policy changes using improvement bounds or interpolation Kakade & Langford (2002); Pirotta et al. (2013). High-confidence policy improvement and SPIBB address candidate reliability in batch settings, the latter falling back to a baseline in poorly supported regions Thomas et al. (2015); Laroche et al. (2019). Cautious policy programming further adjusts the degree of a policy update using an improvement lower bound Zhu & Matsubara (2023). The evaluation problem itself is also established: multiple-policy high-confidence evaluation studies how to estimate candidate returns efficiently from interaction data Dann et al. (2023). A recent audit of validation-gated continual learning shows that an overly conservative admission test can discard useful updates, making retained learning opportunities and interaction budget part of the evaluation Ma & Wu (2026). Idol does not claim that validation, partial updates, or online policy evaluation are new, nor does its finite-sample gate inherit the formal guarantees of these methods. Its distinctive design choice is to spend fresh interactions on a temporary, deliberately amplified update triggered by rare self-generated success, and to use the measured response to determine the permanent fraction of that same update.

C.4 Interventional diagnostics and candidate selection

Causal Confusion in Imitation Learning uses targeted interactions or expert queries to disambiguate causal hypotheses behind observational demonstrations de Haan et al. (2019). Candidate generation followed by evaluation and selective retention also appears in other domains: GEPA tests proposed prompt revisions, SkillOpt-Lite uses independent validation to gate agent-skill edits, and MIPU selectively accepts candidate updates to improve an LLM's inference-side policy under a training–inference mismatch Agrawal et al. (2026); Shen et al. (2026); Liang et al. (2026). These works make intervention and candidate validation important precedents, although their hypotheses, artifacts, and evaluation settings differ. Idol intervenes on a policy because a rare self-generated episode suggests a possible capability; it makes that behavior observable through a reversible candidate and uses the subsequent policy-level response, without stepwise causal identification, to decide whether and how strongly to commit.

D Extended controlled results and diagnostics

D.1 Global-luck experiments

We first study the global-luck sequence task of Section B.2. FR-01 uses REINFORCE under the shared policy and Idol protocol of Section B.4; its strength, luck, and validation-batch scans specify their departures below. FR-04 then replaces the base learner with PPO while retaining the global-luck task and the same interaction budget.

The primary outcome is the final policy's exact complete-sequence skill probability, computed analytically from its action probabilities, without either luck reward or a local mask. We also report the fraction of seeds whose official policy first reaches 90% exact skill within the budget (T90 attainment). T90 is a first crossing, not a claim that the policy remains above 90%. A non-learning uniform policy provides the analytic 4^{-H} floor. The equal-interaction, fixed-validation-cost comparison tests whether probing a strong candidate before commitment helps. Because the validated gates already substantially exceed the REINFORCE baseline in the core global-luck setting below, we retain this mechanism-focused protocol without reallocating inactive-round validation interactions to more training updates.

D.1.1 Rarity and commit gates

The original 20-seed strong-probe experiment established the working setting; we use its independent 100-seed replication (seeds 400–499) for the main comparison. This replication evaluates Base, Naive full commit, Binary, and Partial at $H = 5, 6, 7, 8$ under the standard luck rate of Eq. A15. A subsequent same-seed extension changes only the gate to Fisher or LB. It is a post-replication mechanism comparison, not part of the original preregistered set. Table A1 and Figure A1 show the results.

Table A1: Final exact complete-sequence skill probability (%) averaged over 100 seeds. The Fisher and LB columns come from the subsequent same-seed gate extension; all methods have equal environment-step budgets.

H	Base	Naive	Binary	Partial	Fisher	LB
5	1.016	10.000	73.000	97.999	93.986	99.987
6	0.053	8.000	96.978	99.976	99.972	99.973
7	0.009	7.000	80.617	80.617	80.616	80.616
8	0.002	6.998	25.635	25.635	25.635	25.635

Across all four horizons, each of the four validated gates exceeds both Base and Naive in mean final skill. All 16 preregistered Binary/Partial-versus-control paired intervals in the replication are positive after correction across its 20 comparisons; the corresponding 16 Fisher/LB-versus-control intervals are also positive after correction within the extension's separate 28-comparison family. These are two distinct inference families, not 32 comparisons

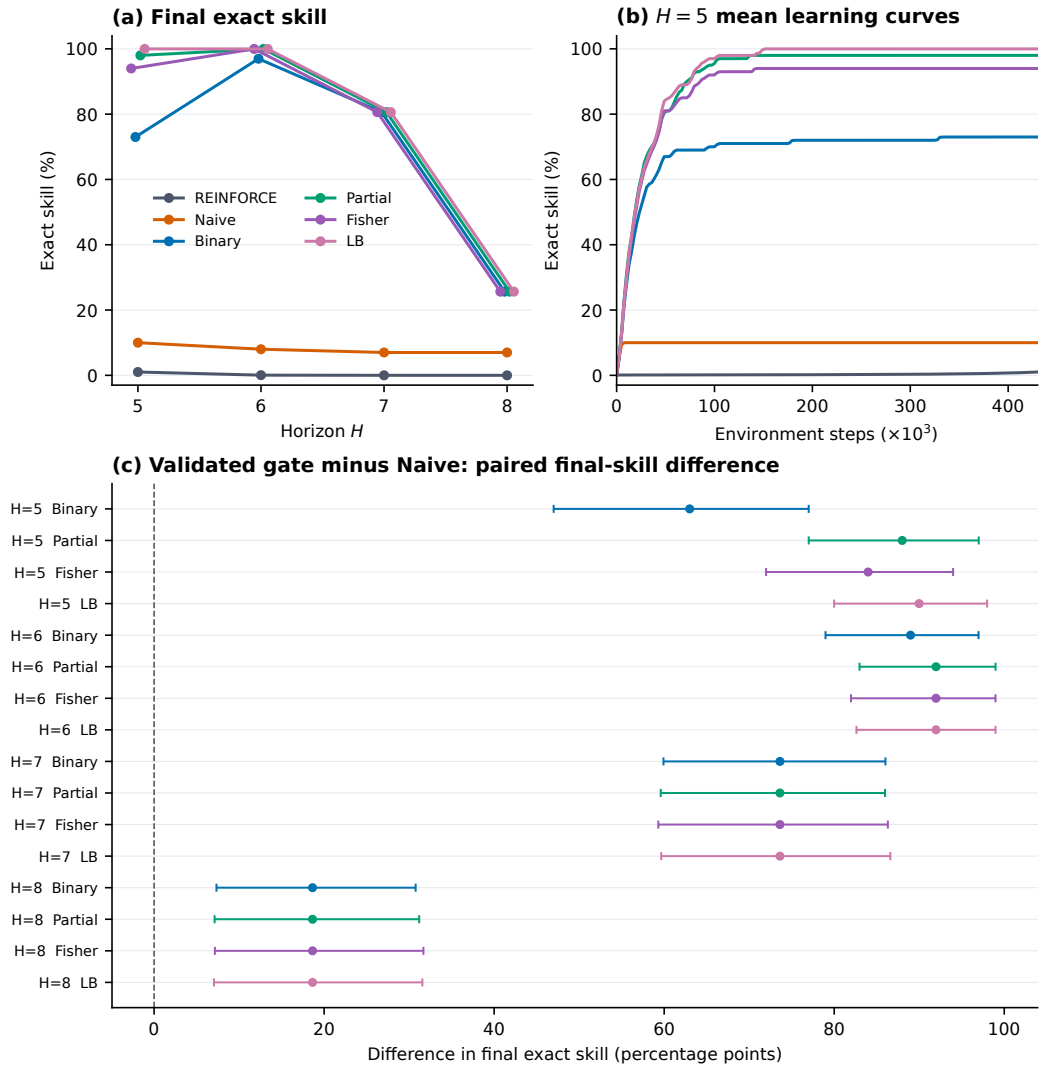


Figure A1: Global-luck rarity and gate comparison, using 100 paired seeds per horizon and 430,080 environment steps per method and seed. (a) Mean final exact skill. (b) Mean exact-skill learning curves at $H = 5$, evaluated at common 2,560-step checkpoints; Base is sampled every second generation, and the other methods every generation. (c) Paired final-skill advantage over Naive, with registered Bonferroni-adjusted bootstrap intervals. Binary and Partial intervals belong to the replication's 20-comparison family; Fisher and LB intervals belong to the later extension's separate 28-comparison family. Horizontal offsets in (a) expose overlapping curves and do not change H.

pooled into one test. The large gap from Naive shows why the strong candidate should be tested before commitment: stronger imitation alone is unreliable when a top-return episode can be lucky.

Gate differences are most visible at $H = 5$. Binary reaches 73.00% mean final skill, Partial 98.00%, Fisher 93.99%, and LB 99.99%. Partial exceeds Binary by 25.00 percentage points in the replication, with a corrected interval of approximately $[13.00, 38.00]$ points; Fisher exceeds Binary by 20.99 points in the later gate extension, with a corrected interval of approximately $[8.99, 33.99]$. LB makes no partial commits in this extension: accepted proposals jump past its full-commit threshold. Its near-perfect endpoint therefore does not establish a benefit from its square partial-commit ramp. At $H = 6$, performance approaches

saturation; at $H = 7, 8$, gate endpoints are nearly identical, so these data do not support a universal ranking among gates.

At $H = 5$, the mean learning curves in Figure A1(b) show the validated gates gaining exact skill within the first 100,000 interactions, while Naive levels off near 10% and Base remains near the random-policy floor. Increasing H also constrains what any gate can recover within the matched budget. At $H = 8$, Base encounters a complete sequence in 51 of 100 seeds yet learns almost no reliable skill. Binary and Partial discover one in 41 seeds, and only 21 reach T90. Their 25.63% mean final skill is driven by a minority of successful runs; the median remains near the random-policy floor. Discovery alone does not guarantee consolidation, and the gate cannot consolidate a correct sequence that was never discovered.

D.1.2 Probe-strength response

To examine candidate strength beyond the single FR-01 setting, we use the separately registered global-luck strength sweep (100 new seeds 500–599; $H = 6$, $p_{\text{luck}} = 0.0025$). Its 19 nominal strengths $m \in \{1, 2, 4, 8, 16, 24, 32, 40, 48, 56, 64, 72, 75, 80, 88, 96, 104, 112, 120\}$ specify the expected number of replays of an ideal selected trajectory among 128 probes at a uniform starting policy. This sweep uses a common KL cap of 3.2, whereas the FR-01 rarity and gate comparisons use 1.6; its $m = 75$ result is consequently not an exact rerun of the FR-01 setting. It compares Naive, Binary, and Partial without a new Base run.

Figure A2(a,b) shows that Binary reaches a broad final-skill plateau of roughly 94% from $m = 4$ onward, while Partial is near 99–100% from $m = 8$ onward. Naive falls from 28% at $m = 1$ to about 6% at $m = 120$, showing that stronger permanent imitation does not replace validation. Even after final skill levels off, the complete-budget checkpoint AUC improves as useful policies are learned earlier. The registered within-strength comparisons find Binary above Naive at all 19 strengths and Partial above Naive at $m \geq 2$ after correction over their 57-comparison family. The cross-m curves are descriptive; the experiment does not establish one optimal strength or formal equivalence across the plateau.

D.1.3 Luck prevalence

The initial luck-to-skill ratio is $R_0 = p_{\text{luck}} 4^H$. We vary it across seven values: 0, 1, 4, 10.24, 40.96, 163.84, and 655.36, while holding the strong candidate and LB rule fixed. This extension uses 50 paired seeds 800–849 at each of $H = 6, 7$, and compares Base, Naive, and LB at 430,080 steps per run. R_0 describes the uniform initial policy; as skill changes during training, the actual luck-to-skill ratio changes too.

Figure A2(c,d) shows Naive degrading as luck contamination increases. LB remains much stronger over a wide range, but it also weakens in the extreme regime: from about 99.97% at the standard $R_0 = 10.24$ to 53.98% at $R_0 = 655.36$ for $H = 6$, and from 78.94% to 25.99% for $H = 7$. At zero luck, Naive and LB coincide in this setting; a small amount of luck can supply additional rewarded candidates that the validation process subsequently exploits. This is a path effect of this reward-only frontier and gate, not evidence that luck itself improves the target sequence. At the most extreme ratio, LB still exceeds Base and Naive in the registered 42-comparison paired-bootstrap family, while its outcomes become highly bimodal across seeds. Its reward-based gate has a finite robustness range.

D.1.4 Validation batch size under high luck

Finally, an exploratory extension retains the same 50 seeds and LB gate at the two highest luck ratios, $R_0 \in \{163.84, 655.36\}$. It fixes the number of main training generations (140 for $H = 6$, 120 for $H = 7$) and increases each of the reference and probe batches from 128 to 256 or 512. The corresponding total budgets are 430,080, 645,120, and 1,075,200 environment steps. Thus panels (e,f) of Figure A2 compare validation resolution at fixed training opportunities, with unequal total interaction budgets.

Larger batches reduce immediately harmful commits in the extreme $R_0 = 655.36$ condition: from 53 to 6 at $H = 6$, and from 10 to 2 at $H = 7$. A better endpoint is established only for $H = 6$ at 512/512, which exceeds 128/128 by 22.00 percentage points (registered

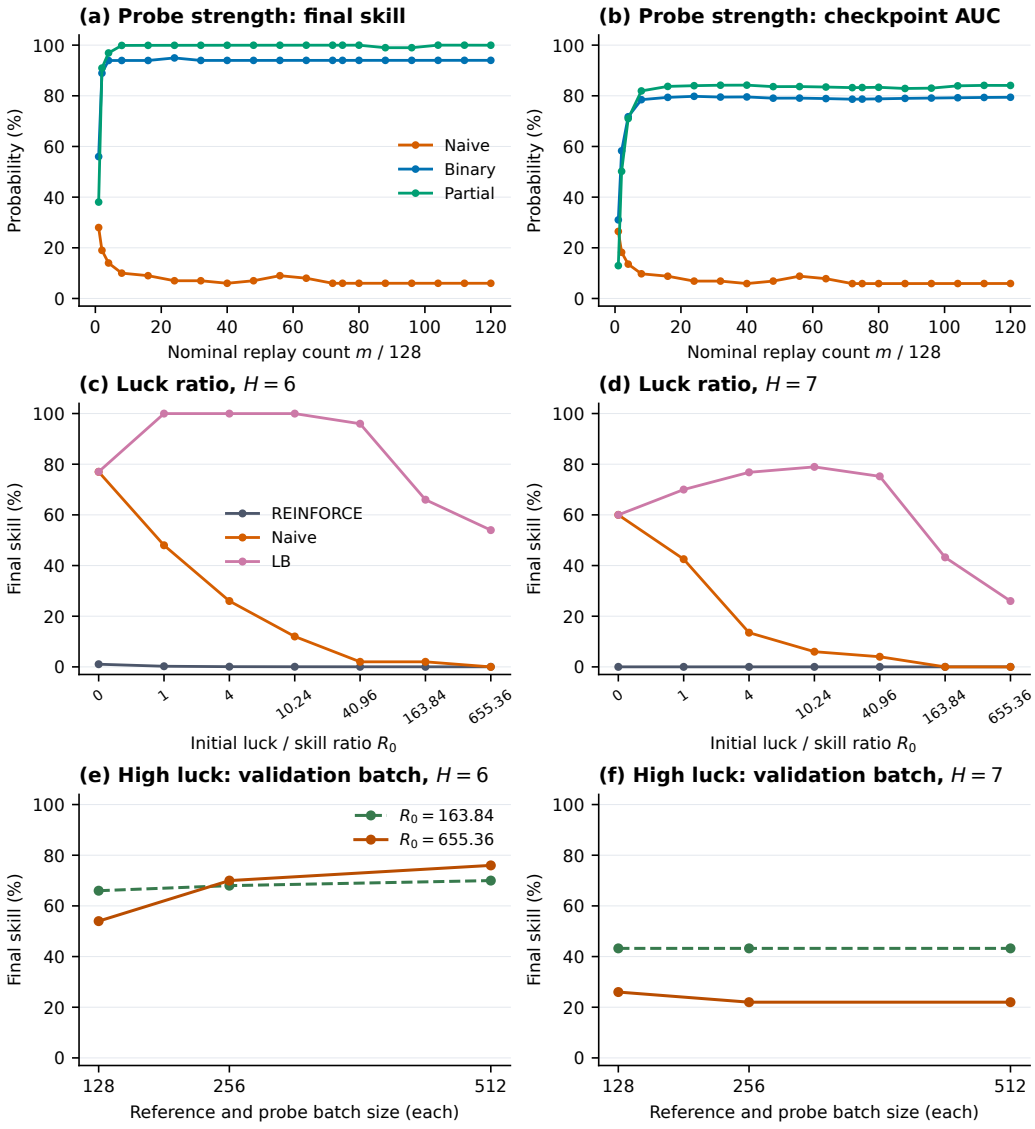


Figure A2: Global-luck parameter scans. (a,b) Probe strength m: final exact skill and full-budget checkpoint AUC at $H = 6$, 100 seeds per setting, 430,080 steps, KL cap 3.2; no new Base control was run. (c,d) Initial luck-to-skill ratio at $H = 6, 7$, 50 seeds per setting, 430,080 steps, KL cap 1.6. (e,f) High-luck reference and probe batch sizes (each shown on the horizontal axis) at fixed training generations, 50 reused seeds; total environment steps increase from 430,080 to 645,120 and 1,075,200. Lines display descriptive means across seeds. The three scans are separately registered cohorts and are not pooled.

12-comparison corrected interval $[4.01, 41.99]$). At $H = 7$, the larger batches do not improve the endpoint; the highest-ratio mean falls from 25.99% to 21.99%, with a corrected interval that includes zero. Better rejection of harmful candidates can therefore coexist with missed useful commits, and simply increasing validation samples does not remove the remaining discovery and gate-threshold limits.

D.1.5 Validation advantage with a PPO base learner

FR-04 asks whether validation-gated consolidation remains useful when the base learner changes from REINFORCE to PPO. At each $H = 5, 6, 7, 8$, it compares PPO alone with

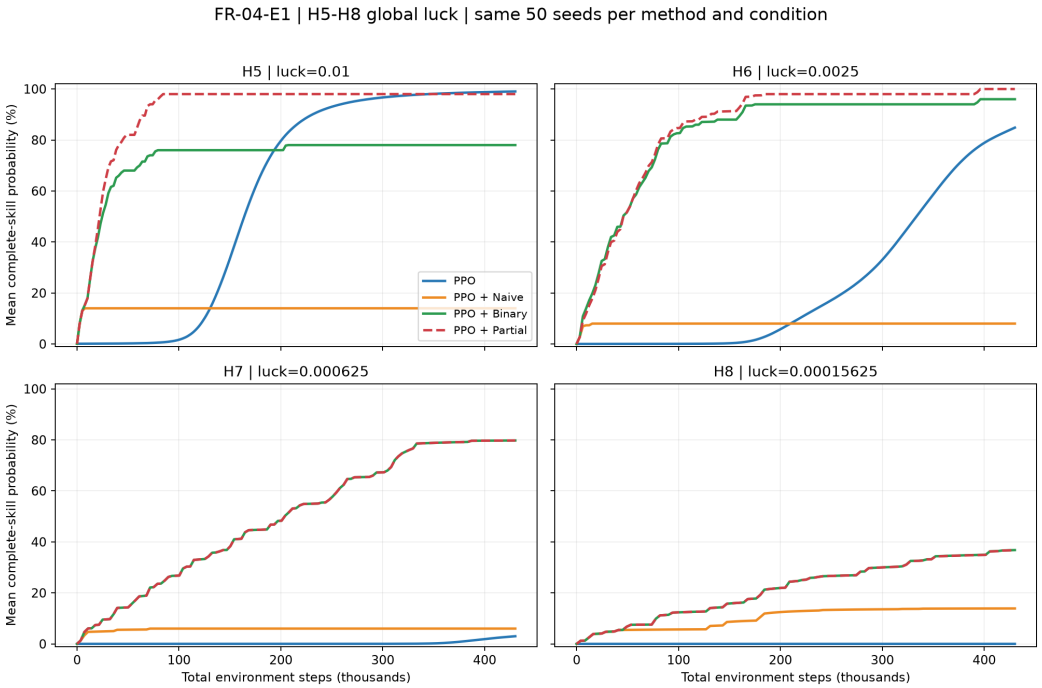


Figure A3: FR-04 mean exact-skill learning curves for PPO alone and three PPO–Idol variants at $H = 5$–8. Each condition uses the same 50 paired seeds and 430,080 environment steps per method; curves include all seeds and the uniform initial policy. H7/H8 were added after inspecting H5/H6 and are not an independent-seed replication.

PPO+Naive, PPO+Binary, and PPO+Partial on 50 paired seeds (700–749) and 430,080 environment steps per method and seed. All four arms use the same global-luck task and policy; the Idol arms spend half their episodes on fresh reference and probe rollouts instead of PPO training. Table 1 reports final exact skill, and Figure A3 shows the mean learning trajectories.

At $H = 6$, Partial raises final skill from PPO's 84.84% to 99.98% (+15.14 points, adjusted interval $[+6.38, +27.21]$); T90 attainment rises from 35/50 to 50/50 seeds. The advantage persists at $H = 7$: Partial and Binary both reach 79.76% against PPO's 3.00%, and 39/50 validated runs reach T90 against 0/50 PPO runs. At $H = 8$, Partial reaches 36.74% against PPO's 0.0029%, with a positive adjusted interval, but only 17/50 validated runs reach T90 and the Partial median is just 0.0020%. Thus the mean gain at the hardest horizon is driven by a minority of successful seeds.

The result is not uniform across difficulty or gates. At $H = 5$, Partial's endpoint is 98.00% against PPO's 99.02%, and its adjusted interval crosses zero. Its full-budget exact-skill AUC is nevertheless 91.84% against 59.49% for PPO, consistent with the earlier rise in Figure A3; this speed comparison is descriptive, without a registered AUC interval. Binary instead ends below PPO at $H = 5$ (78.00% versus 99.02%; adjusted Binary–PPO interval $[-39.04, -5.02]$). At $H = 7$, Binary and Partial coincide, so the result does not isolate an added benefit from partial commitment.

H5/H6 were collected first; H7/H8 reuse the same seeds after their results were inspected. Each batch retains its own registered comparison procedure, and their intervals should not be read as one jointly corrected family. FR-04 supports a validation advantage over PPO in several tested global-luck conditions, not a claim that every gate dominates PPO or a same-condition causal comparison between PPO and REINFORCE.

D.2 Local-mask experiments

We next use the local-mask reward of Section B.3 at $H = 6$, with global luck disabled. FR-02 varies the mask probability over the eight values specified there; FR-03 fixes $q = .3$ and scans probe strength; FR-05 tests $q = .1, .3, .5$ with PPO. All use the shared policy and default Idol protocol of Section B.4, except for the strength-scan cap stated below.

We evaluate both the policy's analytic probability of receiving reward under the mask rule and its analytic probability of predicting the complete target sequence without masks. T90 refers only to the latter. The hidden masks and per-position correctness labels are used for diagnostics, not supplied to the policy or gate. These two outcomes must be kept separate because increasing q can raise masked reward while weakening what a rewarded trajectory reveals about complete skill.

Sections D.2.1 and D.2.2 use REINFORCE, while Section D.2.3 uses the separately specified PPO learner. Each run retains the default environment-step budget unless a scan states otherwise.

D.2.1 Mask prevalence and commit gates

FR-02 tests whether commit rules can extract complete skill when rewarded trajectories contain masked errors. Its primary independent replication uses 100 seeds (300–399) at each of eight mask probabilities. The original Binary and Partial runs form FR-02-E2; the same-seed Base run (E2-R) and Fisher/LB gate runs (E2-G) were added afterward. We use their shared seeds for paired comparisons but retain their separate registered comparison families: 16 Partial–Binary comparisons in E2, 32 validated–Base comparisons in E2-R, and 64 gate comparisons in E2-G. Figure A4 shows both outcomes and the gate diagnostics.

At $q = 0$, all four gates finish at 77.40% exact skill, against 0.93% for Base. As the mask becomes common, the meaning of reward and exact skill diverges. At $q = .3$, mean final exact skill is 7.84% for Base, 21.37% for Binary, 37.23% for Partial, 28.74% for Fisher, and 48.50% for LB; the respective masked pass rates are 18.95%, 39.29%, 53.91%, 47.13%, and 63.86%. At $q = .8$, the gates attain 57.59–62.28% masked pass rates but only 0.10–0.23% exact skill, below Base's 2.20% exact skill. Masked reward alone would obscure this reversal.

The corrected E2 intervals favor Partial over Binary in exact skill at $q = .1, .3, .4, .5$ and in masked pass rate at $q = .1, .3, .4, .5, .6, .8$. The later Base extension supports Binary above Base in exact skill through $q = .3$, Partial through $q = .5$, but neither at $q = .6$; both are below Base at $q = .8$. The later gate extension supports LB above Partial for both outcomes at $q = .1, .2$. Fisher–Binary exact-skill intervals cross zero at every q, and Figure A4(d) does not establish a universal gate ranking. These claims use each batch's own Bonferroni-adjusted paired bootstrap family; the later extensions are mechanism checks rather than additional seeds for the E2 inference.

An active candidate has nonzero candidate displacement and enters the response gate. We define candidate-level FPR as the fraction of active trajectories with at least one incorrect action that receive any nonzero commit, and FNR as the fraction of active completely correct trajectories rejected. At $q = .3$, FPR is 343/2,952 for Binary, 531/6,724 for Partial, 317/3,661 for Fisher, and 330/5,973 for LB; all four FNR numerators are zero, with respective complete-candidate denominators 24, 41, 32, and 54. At $q = .8$, no active completely correct candidate occurs, so FNR is undefined rather than evidence of perfect sensitivity. FPR is a useful screening diagnostic, but it is not the rate of harmful policy commits: an incorrect selected trajectory may still improve exact skill after a partial commit, and the rate's denominator excludes inactive rounds. FPR also does not by itself order the gates' final skill.

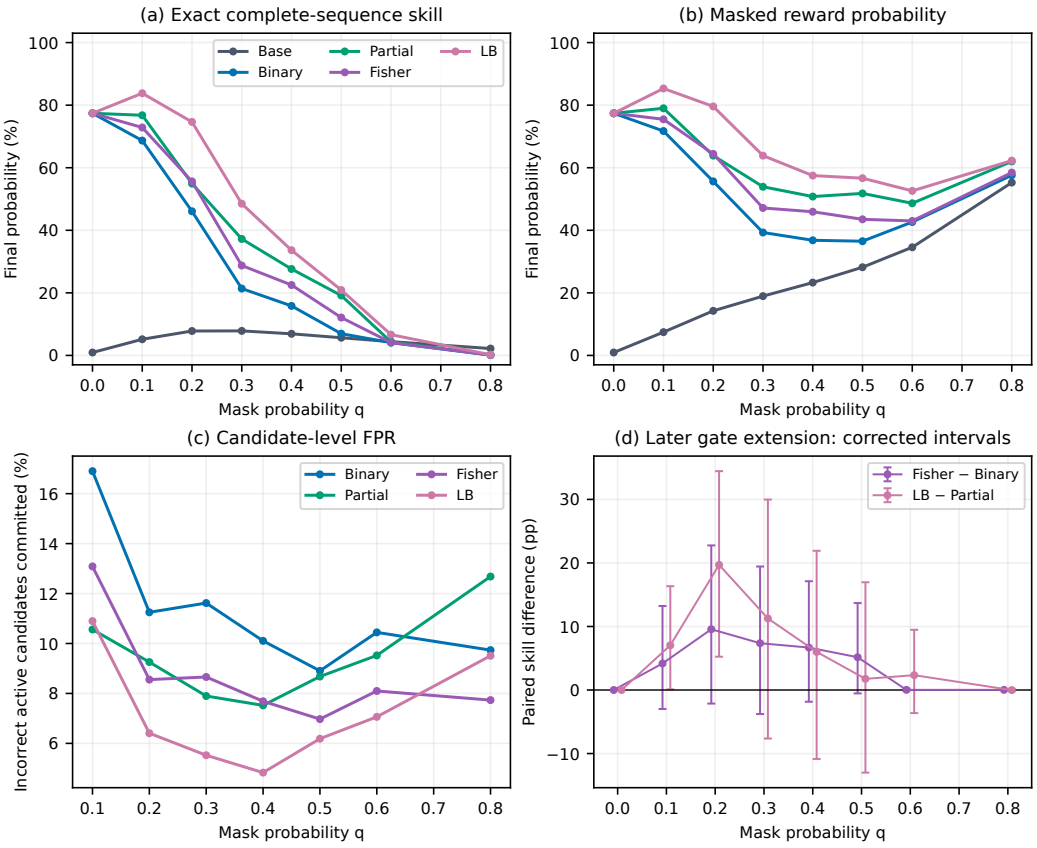


Figure A4: FR-02 local-mask comparison at $H = 6$, 100 paired seeds per q, 430,080 environment steps per run, and KL cap 1.6. (a) Final exact complete-sequence skill and (b) analytic masked pass probability. Base was added in E2-R; Fisher and LB were added in E2-G to the original Binary/Partial E2 cohort. (c) Incorrect active-candidate submission FPR; $q = 0$ has no incorrect active candidates and is omitted. (d) Later E2-G paired exact-skill differences with intervals corrected within its registered 64-comparison family. Curves in (a–c) show seed means, not confidence intervals.

D.2.2 Probe-strength response

FR-03-E2 fixes $q = .3$, uses 100 new seeds (600–699), and varies the candidate's nominal probe replay strength over the same 19 values used in Section D.1.2. It compares Naive, Binary, and Partial at 430,080 environment steps, without a new Base run. Its KL cap is 3.2, rather than FR-02's 1.6. Figure A5 reports exact skill, masked pass, full-budget left-held checkpoint AUC, and a candidate-policy diagnostic.

The current reward-difference gates perform worse overall at larger m, although the curves are far from monotone. From $m = 1$ to 120, Binary's mean final exact skill falls from 51.11% to 8.79%, and Partial's from 66.32% to 20.19%; Naive falls from 18.23% to 1.95%. The AUC and masked-pass curves retain intermediate rebounds. Within the experiment's registered 114-comparison family, Binary exceeds Naive on exact skill at 18 of 19 strengths and Partial at all 19; these within-strength comparisons do not test a cross-strength optimum. In particular, $m = 1$ is the highest observed Partial endpoint on this grid, but we do not infer that it is optimal.

Panel (d) asks a different question from FR-02 FPR. Among accepted active candidates, it counts candidate policies whose exact skill is below the pre-commit policy's exact skill. For Partial this fraction grows from 195/703 (27.7%) at $m = 1$ to 492/620 (79.4%) at

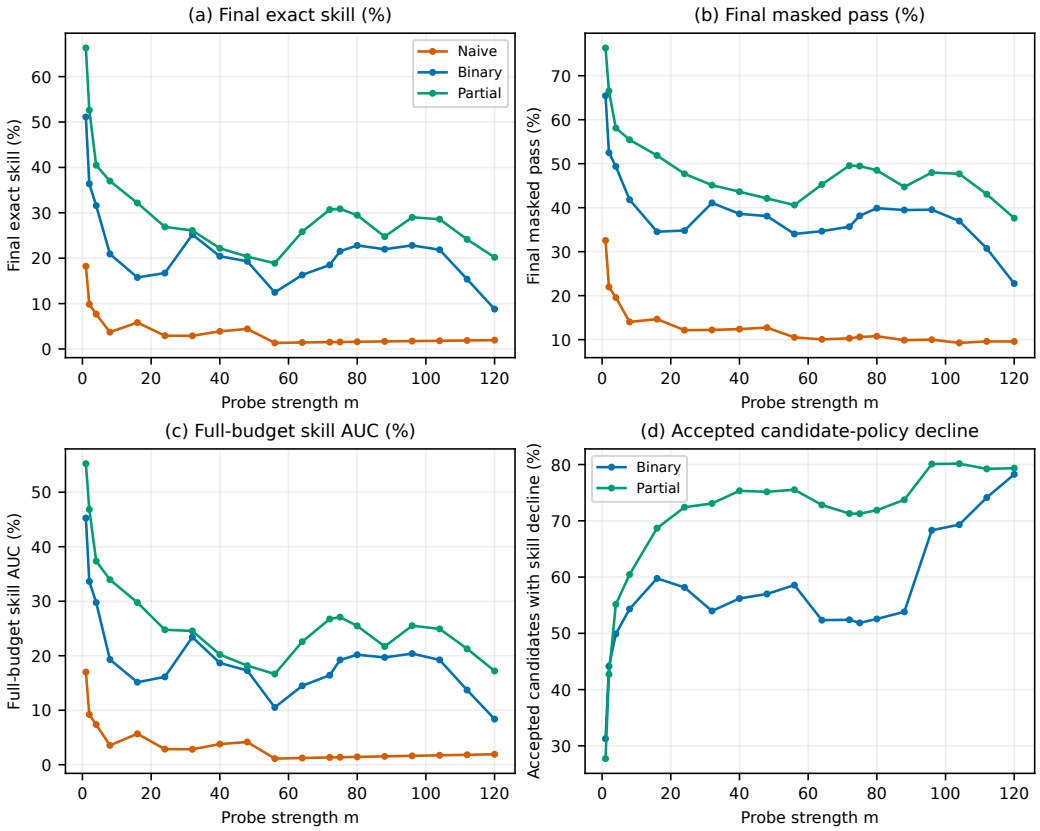


Figure A5: FR-03-E2 local-mask probe-strength sweep at $H = 6$, $q = .3$, 100 seeds per strength and method, 430,080 steps, and KL cap 3.2. (a) Final exact skill, (b) final masked pass probability, (c) full-budget exact-skill AUC using preceding checkpoint values over environment-step intervals, and (d) fraction of accepted active candidate policies with lower exact skill than the policy before submission. All curves are descriptive seed means or pooled event fractions; no cross-strength interval is added to the registered within-strength comparison family.

$m = 120$. It concerns a policy-level skill difference conditional on acceptance, rather than whether an incorrect trajectory passed the gate. A trajectory with k incorrect positions can pass an independent local mask with probability q^k; stronger imitation of such a selected trajectory can therefore increase masked reward evidence while moving the policy away from the complete target. This mechanism is consistent with the diagnostic but does not identify an untested replacement gate.

D.2.3 PPO: fast consolidation and catastrophic commits

FR-05 changes the ordinary learner to PPO while retaining the local-mask task and the fixed interaction accounting. Its formal comparison uses 50 paired seeds (750–799) at each of $q = .1, .3, .5$, with PPO alone and five PPO–Idol variants. Table A2 gives all six final exact-skill means. At $q = .1$, LB and PPO both reach T90 in 50/50 seeds. LB's final advantage is small, 98.9249% versus 98.7798%, but positive in the registered 72-comparison corrected interval: +0.1450 percentage points, $[+0.0188, +0.2694]$. Its full-budget trapezoidal skill AUC is 89.19% versus 60.40%; Figure A6(a) shows the earlier rise of mean exact skill under LB.

At $q = .3$, the 23 LB seeds with no observed harmful commit have AUC 86.64% against 66.12% for PPO on those same seeds, and final mean skill 98.88% against 98.69%. Fig-

Table A2: FR-05 final exact complete-sequence skill (%), mean over 50 paired seeds per mask probability. Every method uses 430,080 environment steps. PPO–Idol variants spend half the episodes on validation and have half as many optimizer steps as PPO.

q	PPO	PPO+Naive	PPO+Binary	PPO+Partial	PPO+Fisher	PPO+LB
.1	98.78	31.61	81.12	91.02	87.05	98.92
.3	98.68	3.97	37.63	49.53	45.53	61.35
.5	98.24	1.99	15.64	29.47	23.35	53.07

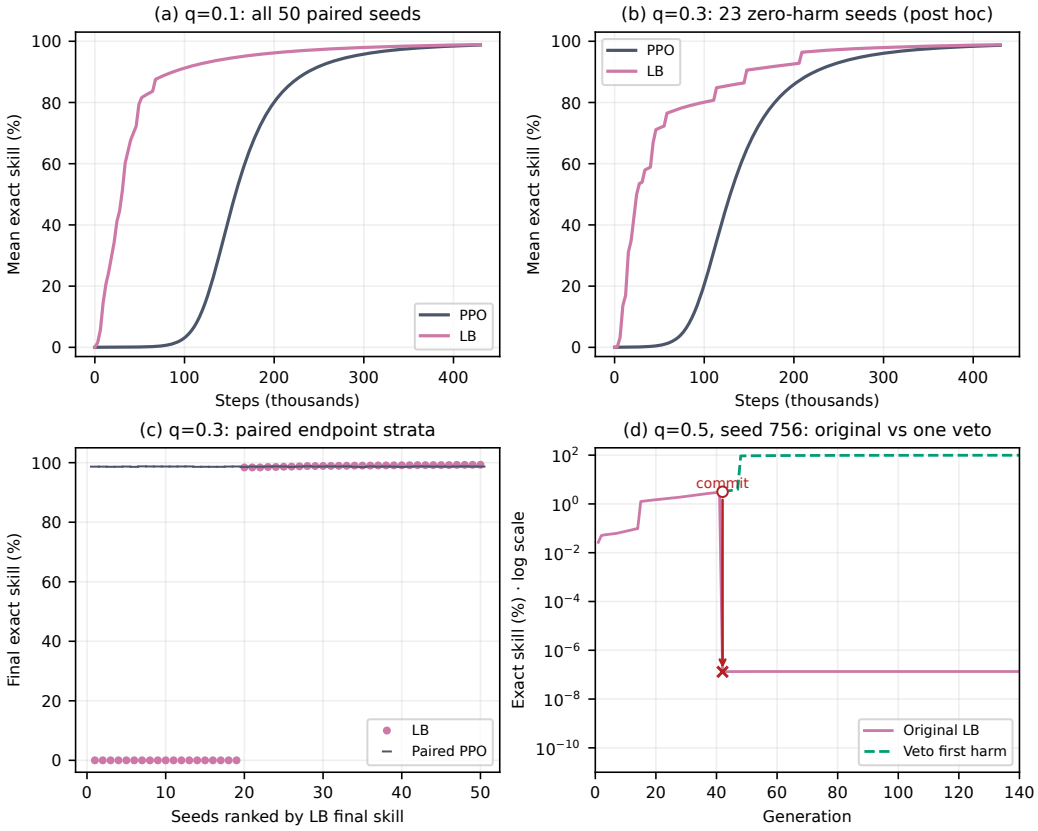


Figure A6: FR-05 PPO local-mask trajectories and endpoints: 50 paired seeds per q, 430,080 environment steps, KL cap 1.6 for candidate construction. (a) All 50 $q = .1$ seeds; (b) the 23 $q = .3$ LB seeds with zero observed harmful commits and their same-seed PPO controls, a post hoc conditional group. Mean exact-skill curves are sampled on a common 3,072-step grid. (c) Final $q = .3$ LB skill ranked by LB outcome, with paired PPO endpoints. (d) Original $q = .5$, seed 756 LB path and a separate post hoc, oracle-assisted rerun vetoing only its first harmful commit at generation 42 (dashed green). The open circle is after PPO's update but before commitment; the cross and arrow mark the original commit's immediate loss on a logarithmic ordinate. Panels (b–d) are descriptive diagnostics without new confirmatory comparisons.

ure A6(b) displays this post hoc conditional group, not a new randomized treatment contrast. For all 50 seeds, however, LB is below PPO: AUC 52.86% versus 65.93% and final skill 61.35% versus 98.68% at $q = .3$. At $q = .5$, the all-seed AUC is 35.45% versus 64.67%, and final skill 53.07% versus 98.24%. The corrected LB–PPO final-skill intervals are negative at both $q = .3$ and .5. Thus faster consolidation on successful LB paths does not translate into a population advantage as mask prevalence rises.

The average conceals a sharp split. At $q = .3$, 31/50 LB seeds end above 90% exact skill while 19/50 end below 1%; at $q = .5$ the counts are 27/50 and 23/50 (Figure A6c). The

$q = .3$ near-zero group contains 78 of the 86 immediately harmful LB commits. In 16 of these 19 failed seeds, the first harmful commit immediately cuts exact skill by at least a factor of 20. The eight $q = .3$ seeds with exactly one harmful commit that nevertheless reach high skill are early exceptions: all eight events occur by generation 12, and six cause less than 0.03 percentage points of immediate skill loss. The other two lose 0.22 and 0.86 points but recover with subsequent training. Thus a single harmful commit produces an immediate cliff in most of the eventual failures, while an early, recoverable loss need not prevent consolidation.

Figure A6(d) isolates a single event at $q = .5$, seed 756: after PPO's update in generation 42 and before LB's only harmful commit, exact skill is 3.1576%; the commit lowers it to approximately 0.00000013%. The open circle, cross, and arrow mark that within-generation policy change on a logarithmic scale. In a separate single-seed rerun, an oracle vetoes that first harmful commit and leaves subsequent LB decisions unchanged. The dashed comparison path then reaches 99.03% final exact skill, while the original remains near zero. This post hoc intervention illustrates the cost of this particular submission; it is neither a deployable gate nor a controlled claim that PPO is intrinsically more sensitive to harmful commits than REINFORCE.

A post hoc oracle veto provides a limited intervention check. From the original $q = .3$ LB run, we selected the 19 near-zero seeds after seeing their endpoints and reran each with only its first would-be harmful commit vetoed using the environment's analytic exact-skill oracle. Eleven of 19 then reach at least 90% final skill, while eight remain failures: vetoing one event rescues a majority of this selected failure group. The original 19 have mean final skill 0.000065%, versus 57.40% after veto. Because selection is retrospective and the oracle is unavailable to an ordinary gate, this is neither a deployable algorithm nor a fair all-seed PPO comparison; we do not attach a confirmatory interval to it.

E Benchmark protocols and supplementary results

E.1 DeepSea: consolidation after rare discovery

FR-06 v2 tests whether a discovered successful path can be retained in deterministic DeepSea with $N = 12$, two actions, and a fixed randomized action mapping (seed 42). A uniform policy completes the path with probability 2^{-12}. PPO, PPO+SIL, PPO+LowerBound, and its validation-budget-recycling variant each use 50 seeds (1000–1049), 51,200 formal episodes, and 614,400 environment steps per seed. The interaction budgets are equal, although the methods use different numbers of optimizer steps. The comparisons below use the corrected equal-budget FR-06 v2 registration and are descriptive rather than confirmatory.

Our PPO+SIL implementation uses positive-advantage auxiliary updates (Oh et al., 2018) with a replay capacity of 50,000 transitions, storing all transitions from episodes containing a positive reward. Once the buffer contains at least 12 transitions, each 256-episode main batch receives 10 SIL updates after its ordinary PPO update. Each auxiliary update samples up to 512 transitions with replacement, capped by the current buffer size, using priority exponent 0.6 and importance-weight exponent 0.1. Positive return advantages are capped at 1; the overall SIL loss weight is 0.1, and its value-loss coefficient is 0.01. The priority floor is 10^{-6}; losses are normalized by the positive-advantage sample count with a floor of 64, and the negative-log-probability cap is 100. SIL shares PPO's actor, critic, and Adam optimizer, with actor/critic learning rates 0.003/0.01, discount 1, entropy coefficient 0.01, and gradient-norm cap 1. These replay updates require no additional environment interactions.

For LowerBound, each candidate round uses 256 main episodes and, when validation is performed, 128 fresh reference and 128 probe episodes. A top-return main trajectory supplies the candidate. Its strong probe is calibrated to a nominal 75 reproductions among 128 probes from a uniform policy; this calibration does not imply 75 observed reproductions after PPO has changed the policy. The gate tests the one-sided lower bound on the difference in binary success rates, rejecting at or below 0.02, fully committing at or above 0.05, and using a quadratic partial fraction between these thresholds. To match PPO and SIL's 200 rounds

Table A3: FR-06 v2 outcomes conditional on the same 17 main-discovery seeds, at 51,200 episodes per method and seed. Final success is the analytic complete-path probability. T50/T90 are median algorithm generations from first main discovery to first crossing, calculated only among seeds reaching the respective threshold; generations cost 512 episodes for ordinary LowerBound and 256 for PPO/SIL, while the recycled variant has variable candidate-round cost.

Method	Final success (%)	Final $\geq 90\%$	T50 (gen.)	T90 (gen.)
PPO	11.66	2/17	60.5	93
PPO+SIL	99.65	17/17	32	56
PPO+LowerBound	92.31	16/17	4	37.5
LowerBound recycled	99.57	17/17	3	36

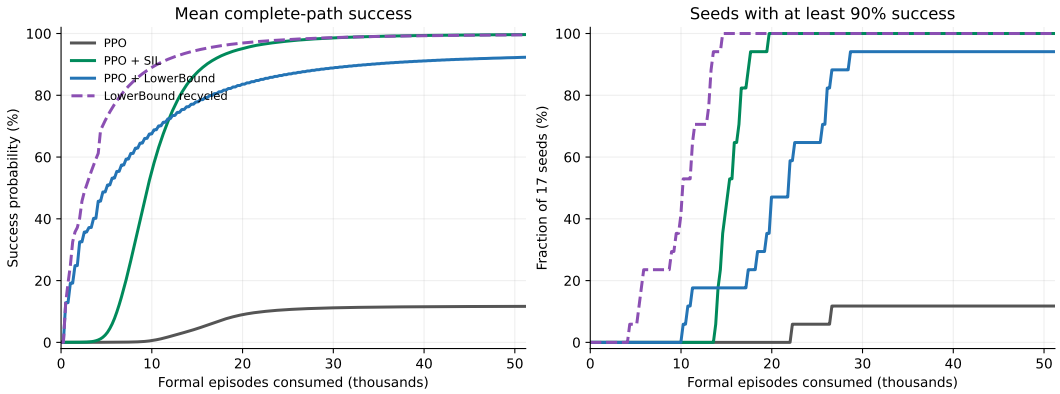


Figure A7: FR-06 v2 equal-episode trajectories on the fixed 17-seed main-discovery cohort: mean analytic complete-path success and the fraction of these seeds at or above 90% success. Curves use left-held policy checkpoints every 256 formal episodes, include the uniform initial policy, and end at 51,200 episodes. This retrospectively conditioned view is descriptive; it excludes the 33 seeds without main discovery and contains no significance intervals.

of 256 episodes, ordinary LowerBound contributes only its first 100 candidate rounds, each costing 512 episodes.

Only 17 of the 50 seeds encounter a complete successful path in a main rollout within the budget. These are the same 17 seeds in all four methods. We fix this retrospectively selected cohort for Table A3 and Figure A7 to ask how well each method consolidates a discovered path. The other 33 seeds did not supply that opportunity; excluding them does not establish a population-level advantage or improve the discovery rate.

At the common endpoint, PPO retains at least 90% complete-path success in only 2/17 discovered seeds, versus 16/17 for LowerBound and 17/17 for both SIL and recycled LowerBound. SIL's conditional mean final success is 99.65%, compared with 92.31% for ordinary LowerBound and 99.57% for the recycled variant. The latter therefore matches SIL's 17/17 endpoint coverage but does not exceed its mean endpoint. Figure A7 places learning on cumulative episodes: both LowerBound variants raise success earlier than SIL in this cohort, whereas the generation medians in Table A3 are not directly comparable as sample-efficiency measures.

The recycled variant changes how unused validation interactions are spent. During its 100 candidate rounds, an inactive candidate skips the scheduled 128 reference and 128 probe episodes. Each skipped pair reserves one 256-episode ordinary PPO update, executed after the candidate phase; every seed still consumes exactly 51,200 episodes. All 17 discovered seeds already exceed 90% success by the end of those 100 candidate rounds, so the tail adds no newly successful seed. One more seed succeeds than under ordinary LowerBound, but skipping inactive validation changes the subsequent reference/probe random stream. That 17/17 versus 16/17 difference cannot be attributed to the tail PPO updates. These results

show strong post-discovery consolidation in this fixed DeepSea setting, with no claim of a general ranking across environments.

E.2 FrozenLake: from conservative validation to stronger commitment

FR-07 moves to the standard stochastic 8×8 FrozenLake-v1 map, with slippery transitions and terminal success as the binary reward. PPO, PPO+SIL, and PPO+Idol use the same 50 seeds (1050–1099) and a target budget measured in actual environment steps; a final episode may cross that target by at most 199 steps. The 100,000-step experiments first exposed a weakness of the original LowerBound Idol settings. The final two complete main batches average 0.2344% success for Idol, below PPO's 0.3438% and SIL's 0.4297%. An inactive Idol candidate still spent reference and probe interactions in that original protocol: across the 50 seeds, 272 complete validations yielded only one positive, full commitment, and just 6.42 ordinary PPO main updates per seed on average. Successes are rare and potentially difficult to reproduce in this environment, which motivated more permissive use of promising candidates; these diagnostics do not by themselves establish that difficulty of reproduction as a causal explanation.

Table A4 traces the subsequent, same-seed mechanism changes. E2 skips reference/probe validation for inactive candidates, leaving more of the fixed interaction budget for ordinary PPO; its average main updates rise to 8.72, but its endpoint remains below SIL. E3 additionally lowers the positive-commit threshold from a 0.02 to a zero lower bound on the binary-success difference. It produces 13 positive commitments, yet none is full because its full-commit threshold remains 0.05 and the intermediate fraction is squared. E4 retains E3's admission threshold but lowers the full-commit threshold to 0.001. It still makes 13 positive commitments, now all full, and its final-two-batch success reaches 0.7813%. Thus admitting more candidates and committing a larger fraction of an admitted candidate are distinct changes: E3 supplies the former observation, while the larger observed gain appears with E4's stronger commitments. E2–E4 were chosen after inspecting preceding results on these same seeds, so this sequence is exploratory rather than a factorial causal test.

Table A4: FR-07 at a 100,000-environment-step target, 50 paired seeds. Counts pool all seeds; main PPO updates are per-seed means. Endpoint success is the equal-seed mean of each seed's last two complete online main batches. E2 skips inactive validation, E3 lowers the admission threshold, and E4 also lowers the full-commit threshold. All three modifications are post hoc same-seed diagnostics.

Idol setting	Full validations	Inactive skips	Positive/full commits	Main updates	Endpoint (%)
Original	272	0	1/1	6.42	0.2344
E2	157	229	0/0	8.72	0.2852
E3	162	220	13/0	8.64	0.3477
E4	156	222	13/13	8.56	0.7813

The 400,000-step comparison starts fresh runs from uniform initialization for PPO, PPO+SIL, and E4; it does not continue the 100,000-step checkpoints. All three conditions use the same 50 paired seeds and target budget. Since each method can consume a different number of interactions per generation, we summarize online main-rollout success against cumulative environment steps. Table A5 reports the mean within the final $(300{,}000, 400{,}000]$ window: 2.0859% for PPO, 2.7010% for SIL, and 6.8215% for E4. The same-seed E4–SIL mean difference is (+4.1205) percentage points, but E4 wins on only 30 of 50 seeds and loses on 20. Its 3.7760% median is substantially below its 6.8215% mean, reflecting heterogeneous outcomes rather than uniform seed-level superiority.

Table A5: FR-07 fresh 400,000-step runs on 50 paired seeds. The final-window statistic first averages complete online main batches per seed whose recorded cumulative environment steps lie in $(300{,}000, 400{,}000]$, then weights seeds equally. The final-two-batch statistic uses each seed's last two complete main batches. These are training-rollout outcomes, not independent evaluations of frozen policies.

Method	Final 100k mean (%)	Median (%)	Final two batches (%)	Main updates
PPO	2.0859	2.0578	2.6445	39.90
PPO+SIL	2.7010	2.7344	3.2852	34.38
PPO+Idol E4	6.8215	3.7760	7.7539	26.94

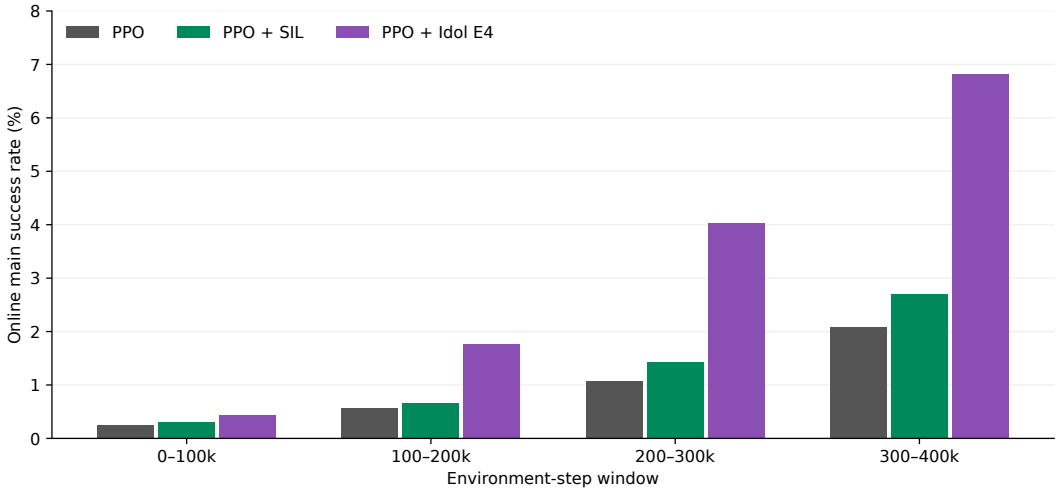


Figure A8: FR-07 fresh 400,000-step runs: online main-rollout success in four successive 100,000-environment-step windows, using the same 50 seeds per method. Within each window, complete main-batch rates are averaged per seed and then equally over seeds; cumulative steps are measured at the recorded batch checkpoint. The plot describes training outcomes and does not evaluate a frozen terminal policy.

Figure A8 shows E4's mean separation from the second 100,000-step window onward. All three window means continue to rise through 400,000 steps, so this budget does not show a clear learning plateau; the experiment does not determine their ranking beyond it. The E4 settings were selected after viewing the 100,000-step outcomes on these same 50 seeds, making the 400,000-step comparison descriptive rather than an independent confirmatory replication. FrozenLake's slip can also produce occasional successes: the recorded main-batch rate does not separate such events from reliably repeatable behavior, and no independent frozen-policy evaluation was run.